Multilevel structural equation models for the interrelationships between multiple dimensions of childhood socioeconomic circumstances, partnership stability and midlife health

Yajing Zhu

Department of Statistics

London School of Economics and Political Science

A thesis submitted to the Department of Statistics of the London School of Economics for the degree of Doctor of Philosophy, London.

September 2018

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 87,251 words.

Statement of conjoint work

I confirm that parts of Chapters 3 and 4 were adpated into a paper entitled "A General 3-Step Maximum Likelihood Approach to Estimate the Effects of Multiple Latent Categorical Variables on a Distal Outcome" jointly co-authored with my supervisors, Prof. Fiona Steele and Prof. Irini Moustaki, and published in Structural Equation Modeling: A Multidisciplinary Journal.

I confirm that parts of Chapters 5-7 were adapted into a paper entitled "A structural equation model for the interrelationships between multiple dimensions of childhood socioeconomic circumstances, partnership stability and midlife health" jointly co-authored with my supervisors, Prof. Fiona Steele and Prof. Irini Moustaki, and is under review for the Journal of the Royal Statistical Society: Series A (Statistics in Society).

Yajing Zhu
September 2018

Acknowledgements

This work is dedicated to my parents, for their continuous love throughout all these years and, for supporting my decision to pursue a doctorate study in social statistics.

I am deeply grateful to my supervisors, Prof. Fiona Steele and Prof. Irini Moustaki for their continuous support for and patience with me, both academically and emotionally throughout the years. I also thank Prof. Chris Skinner for his valuable feedback during my upgrade. This thesis is not possible without their wisdom, guidance, direction and encouragement. I would also like to thank them for being honourable scholars with academic integrity, who are passionate about their work and have a deep love for the society. These qualities have provided and will continue to offer inspirations for me to appreciate and embrace challenges.

I am also greatly indebted to my husband, Vladimir, for his never-ending patience with me throughout the process, for his knowledge and humour that have generated multiple exciting discussions and debates that have strengthened my passion about the work, and for his unconditional love and understanding during ups and downs.

My thank-you also goes to my colleagues at the Dept. of Statistics at the London School of Economics, without whom I would not have enjoyed such a friendly and stimulating research environment; and to my friends who have tolerated my complaints, frustrations, being inconsiderate and yet still stand beside me.

Next, I would like to acknowledge the financial support provided by the London School of Economics, Department of Statistics scholarship. I have also received continuous support from the administrative team of the Dept. of Statistics, whom have made it possible for me to concentrate on my studies.

Finally, to reflect my deep love for traditional Chinese poetry, I offer all the best wishes to my future self by quoting the poet Li Bai from the Tang Dynasty:

"长风破浪会有时，直挂云帆济沧海(唐–李白–行路难)"

(translated as: "Someday, with my sail piercing the clouds, I will mount the wind, break the waves, and traverse the vast, rolling sea.")

Abstract

Recent studies have contributed to understanding of the mechanisms behind the association between childhood circumstances and later life. It has been hypothesized that experiences in childhood operate through influencing trajectories of life events and functional changes in health-related behaviours that can mediate the effects of childhood socioeconomic circumstances (SECs) on later health. Using data from the 1958 British birth cohort, we propose a multilevel structural equation modelling (SEM) approach to investigate the mediating effects of partnership stability, an example of life events in adulthood.

Childhood circumstances are abstract concepts with multiple dimensions, each measured by a number of indicators over four childhood waves (at ages 0, 7, 11 and 16). Latent class models are fitted to each set of these indicators and the derived categorical latent variables characterise the patterns of change in four dimensions of childhood SECs. To relate these latent variables to a distal outcome, we first extend the 3-step maximum likelihood (ML) method to handle multiple, associated categorical latent variables and investigate sensitivity of the proposed estimation approach to departures from model assumptions.

We then extend the 3-step ML approach to estimate models with multiple outcomes of mixed types and at different levels in a hierarchical data structure. The final multilevel SEM is comprised of latent class models and a joint regression model that relates these categorical latent variables to partnership transitions in adulthood and midlife health, while allowing for informative dropout. Most likely class memberships are treated as imperfect measurements of the latent classes. Life events (e.g. partnership transitions), distal outcomes (e.g. midlife health) and dropout indicators are viewed as items of one or more individual-level latent variables. To account for endogeneity and indirect associations, the effects of childhood SECs on partnership transitions for ages 16-50 and distal health at age 50 are jointly modelled by allowing for a residual association across equations due to shared but differential influences of time-invariant unobservables on each response. Finally, sensitivity analyses are performed to investigate the extent to which the specifications of the dropout model influence the estimated effects of childhood SECs on midlife health.

Table of contents

List of figures **10**

List of tables **12**

List of symbols **20**

1 Introduction **1**

1.1 Motivation . . . 1

1.2 Research objective . . . 2

2 Literature Review **6**

2.1 Introduction . . . 6

2.2 Measures of childhood circumstances . . . 6

2.2.1 General health . . . 6

2.2.2 Biomedical . . . 7

2.2.3 Education and cognitive ability . . . 8

2.2.4 Socioeconomic situation . . . 8

2.2.5 Environmental . . . 9

2.2.6 Behavioural and diet . . . 10

2.2.7 Summary . . . 10

2.3 Measures of adult health outcomes . . . 11

2.3.1 Mortality . . . 11

2.3.2 Presence of physical disease . . . 11

2.3.3 General health problems . . . 13

2.3.4 Mental health . . . 13

2.3.5 Health-related behaviour . . . 13

2.3.6 Summary . . . 14

2.4 Pathways linking childhood circumstances and adult health outcomes . . . 14

2.4.1 Confounders . . . 14

2.4.2 Mediators 16

2.4.3 More complex structures 16

2.5 Statistical methods 17

2.5.1 Regression analysis 18

2.5.2 Structural equation modelling 20

2.5.3 Missing data 22

3 Data and measures 25

3.1 The 1958 National Child Development Study 25

3.2 Measures of childhood socioeconomic circumstances 28

3.3 Missing data and the predictors of missingness 30

3.4 Partnership histories and risk factors of experiencing each event 32

3.5 Measures of midlife health and the risk factors 36

4 Latent class analysis of childhood socioeconomic circumstances and a distal outcome 39

4.1 Introduction 39

4.2 Overview of latent class analysis 41

4.2.1 Methodology 41

4.2.2 Application to health studies 45

4.3 Review of the 1-step and step-wise methods: a single categorical latent variable 46

4.3.1 Modal class method 47

4.3.2 The modified BCH approach 48

4.3.3 The Lanza-Tan-Bray approach 49

4.3.4 The 3-step method 50

4.3.5 Comparison of methods: Evidence from simulation studies 52

4.4 Methods to relate multiple categorical latent variables to a distal outcome . 53

4.5 Simulation study 58

4.5.1 Design and population parameters 58

4.5.2 Study 1: all model assumptions are satisfied 60

4.5.3 Study 2: violation of the normality assumption about H 62

4.5.4 Study 3: violation of the conditional independence assumption about H 68

4.5.5 Summary of simulation results 71

4.6 Application of 3-step ML approach to a midlife health outcome 73

5 Event history analysis of the effects of childhood socioeconomic circumstances on partnership formation and dissolution 78

5.1 Introduction . . . 78

5.2 Review of event history analysis . . . 79

5.2.1 Challenges of using event history data . . . 79

5.2.2 Notation and definitions . . . 80

5.2.3 Continuous-time event history model . . . 81

5.2.4 Discrete-time event history model . . . 81

5.2.5 Modelling recurrent events . . . 83

5.2.6 Modelling correlated processes . . . 84

5.3 Research objective . . . 86

5.4 Review of the application of EHA . . . 88

5.5 A general 3-step ML approach for multilevel EHA . . . 90

5.5.1 Model in step 3 . . . 90

5.5.2 Estimation in step 3 . . . 91

5.6 Simulation study . . . 94

5.6.1 Data generation . . . 94

5.6.2 Simulation results . . . 97

5.7 Application of 3-step ML approach to partnership events . . . 100

5.7.1 EHA of the time to first partnership formation . . . 101

5.7.2 EHA of the time to partnership dissolution . . . 104

5.7.3 Joint modelling of partnership dissolution and formation . . . 107

5.7.4 Summary of results . . . 109

6 Joint modelling of distal health outcomes and event histories predicted by categorical latent variables 112

6.1 Introduction . . . 112

6.2 Literature review: relating event history data to external variables . . . 114

6.2.1 External variables as covariates . . . 114

6.2.2 Joint models . . . 115

6.2.3 Connection to our research questions . . . 117

6.3 A general 3-step ML approach for structural equation models . . . 118

6.3.1 Model in Step 3 . . . 118

6.3.2 Estimation . . . 121

6.3.3 Simulation study . . . 122

6.4 Application of 3-step ML approach to a structural equation model . . . 128

7 Missing data 134

7.1 Introduction 134

7.2 Review of missing data mechanisms 135

7.3 Review of models for informative dropout 137

7.3.1 Selection models 137

7.3.2 Pattern mixture models 141

7.3.3 Comparison between models 142

7.4 Modelling informative dropout in the multilevel SEM 143

7.4.1 Modelling approaches for outcomes with non-ignorable missingness 144

7.4.2 Modelling approaches for outcomes with non-ignorable missingness and predictors with missing values 147

7.4.3 Estimation 154

7.5 Application of 3-step ML approach to an SEM with dropout 155

7.5.1 Description of the analysis sample 155

7.5.2 Predictors in the dropout model 157

7.5.3 Multilevel SEM with dropout 159

8 Conclusion 166

8.1 Methodological contributions 166

8.2 Substantive impacts 169

8.3 Limitations and future work 171

References 175

Appendix A Results of latent class analysis on four childhood socio-economic concepts 189

A.1 Introduction 189

A.2 The choice of the number of classes in the latent class analysis 189

A.3 More on simulation studies: no modelling assumptions are violated. 194

A.3.1 Simulation results for a continuous H 194

A.3.2 Simulation results for a binary H 198

A.4 More simulation studies: modelling assumptions are violated 202

A.5 Syntax for selected simulation studies 210

A.5.1 Latent GOLD program for the 3-step approach in Study 1 (2-LV) . 210

A.5.2 Mplus program for Study 3 that investigates the change in the number of classes needed when there is local dependence of items and the outcome, using the 1-step approach (1-LV) 211

A.5.3 Mplus program for Study 3 that investigates the change in the number of classes needed when there is local dependence of items and the outcome, using the 3-step approach (1-LV) 212

A.6 Syntax for the empirical study 213

Appendix B Proof of estimation bias due to endogeneity 214

Appendix C Simulation study to demonstrate bias in estimates due to endogenous predictor 217

Appendix D Syntax for the general 3-step ML approach for event history outcomes 219

Appendix E Joint modelling of the time to first partnership formation and partnership dissolution 221

Appendix F Input data structure and syntax for the SEM in Chapter 6 224

Appendix G Input data structure and syntax for the SEM in Chapter 7 229

G.1 The input data structure 229

G.2 LatentGOLD syntax for the application 230

List of figures

1.1 General framework for the relationship between childhood circumstances, life events in adulthood and midlife health. 3

2.1 Confounding in the general form of Figure 1.1. Y is a health outcome, X is a childhood SEC and Z is a confounder. 15

2.2 Mediation in the general form of Figure 1.1. Y is a health outcome, X is a childhood SEC and W is a mediator. 16

2.3 More complex structures that involve both confounders and mediators. . . . 17

4.1 3-step maximum likelihood method for a simplification of a model with two categorical latent variables C_1 and C_2 and the external covariate $X^{(H)}$ predicting a distal outcome H. M_1 and M_2 are modal classes derived from fitting individual latent class models for two distinct sets of indicators (not shown on the figure for clarity). The curved arrow between two latent variables indicates the existence of association. 54

6.1 A path diagram of a simplification of the multilevel SEM defined by Equations (6.3) and (6.4) with two categorical latent variables and a factor structure for the individual-level random effects. C_1 and C_2 are the latent variables for childhood SECs, M_1 and M_2 are the most likely class memberships derived in step 1 for each dimension of the childhood SECs and H_i is midlife health. $\mathbf{X}_i^{(P)}$ and $\mathbf{X}_i^{(H)}$ denotes the covariate set for partnership events and distal health respectively. The upper half of the node with stacked rectangles is the outcome variable in the partnership submodels. The lower half contains summaries of the partnership history that are used as predictors in the health submodel. 122

A.1 Plots of AIC, BIC and ssa-BIC after performing individual LCA of four different childhood measures 191

A.2 Estimated probabilities of each grouped categories of each childhood measure, condtional on class membership 193

List of tables

3.1 Overview of all NCDS sweeps with information on the type of survey administered (adapted from Goodman et al., 2011) 27

3.2 Percentage distribution of childhood measures by age. 30

3.3 Missing data patterns in waves 5-8 for partnership records (in percentages). 31

3.4 Distributions of predictors of the dropout probabilities ($N = 14,309$). For categorical predictors, percentages are reported; for continuous covariates, mean and standard deviations (SD) are reported. 32

3.5 Life-table estimates of quantiles of the number of years to first partnership and time to separation, by partnership type ($N = 7,313$). 33

3.6 Distributions of time-varying predictors of first partnership formation based on the discrete-time file with 6-month intervals ($N = 7,313$). 34

3.7 Distributions of predictors of recurrent partnership dissolutions based on the discrete-time file with 6-month intervals ($N = 7,313$). 35

3.8 Percentages of individuals in each category of the original 5-point self-reported measure of midlife health, and of the derived binary health measure ($N = 7,313$). 37

3.9 Distributions of categorical and continuous predictors of the midlife health ($N = 7,313$). For categorical predictors, percentage distributions are reported; for continuous predictors, mean and the standard deviation (SD) are reported. Percentage time spent single is calculated as the proportion of total number of months spent in between relationships over the total number of months observed before dropout. 38

4.1 Expected frequency (cell counts) for two categorical latent variables 55

4.2 Proportion in each latent class generated for all scenarios considered in the simulation study. 59

4.3 Study 1: Simulation results when all model assumptions are satisfied (N=2000; entropy set to 0.7 (high) and 0.4 (low); 500 replications). 62

4.4 Study 2a: Simulation results for excess kurtosis (N=200, 2000; 500 replications). 64
4.5 Study 2b: Simulation results for skewness (N=200, 2000; 500 replications). 65
4.6 Simulation results for the number of classes when there is within-class non-normality of H. Results are reported for the percentage of replications for which each model has the minimum ssaBIC value and for which each model is rejected or not based on the BLRT. The selected models are bolded. ssaBIC=sample size adjusted BIC; BLRT=bootstrap likelihood ratio test (average p-values across 500 replications in brackets). 67
4.7 Simulation results for the number of classes when there is local dependence (1-step approach). Results are reported in (%) and selected models are bolded. ssaBIC=sample size adjusted BIC; BLRT=bootstrap likelihood ratio test (average p-values across 500 replications in brackets). 70
4.8 Study 3: Estimated coefficients for fitting a 4-class model when there is local dependence between H and $Y_{10}^{(C)}$. 71
4.9 Most likely class membership for each aspect of childhood SECs derived from fitting latent class models to each of the four sets of repeated composite measures of childhood SECs. Entropy is also reported where values close to one indicate good class separation. 74
4.10 Comparison of effects (in log-odds) of four dimensions of childhood SECs on midlife health using the 3-step ML, modal class and 1-step approaches. Standard errors (SE) are reported in brackets. 76
4.11 Estimated associations between four latent aspects of socioeconomic circumstances in childhood using the 3-step ML approach. Standard errors (SE) are reported in brackets. 77

5.1 Examples of the generated event histories for three individuals 97
5.2 Estimated effects of predictors on logit hazard of an event: high entropy (0.8) and small sample size ($N = 500$) 98
5.3 Estimated effects of predictors on logit hazard of an event: high entropy (0.8) and large sample size ($N = 2000$) 99
5.4 Estimated effects of predictors on logit hazard of an event: low entropy (0.4) and small sample size ($N = 500$) 99
5.5 Estimated effects of predictors on logit hazard of an event: low entropy (0.4) and large sample size ($N = 2000$) 100
5.6 Effects of predictors on the logit hazard of first partnership formation 103
5.7 Covariate effects on the logit hazard of partnership dissolution 106

5.8 Joint modelling for time to first partnership formation and subsequent dissolutions . . . 111

6.1 Simulation results (in log-odds) for the 3-step procedure applied to the SEM with an event history submodel and a distal health submodel: high entropy (0.8) and small sample size ($N = 500$) . . . 125

6.2 Simulation results (in log-odds) for the 3-step procedure applied to the SEM with an event history submodel and a distal health submodel: high entropy (0.8) and large sample size ($N = 2000$) . . . 126

6.3 Simulation results (in log-odds) for the 3-step procedure applied to the SEM with an event history submodel and a distal health submodel: low entropy (0.4) and small sample size ($N = 500$) . . . 127

6.4 Simulation results (in log-odds) for the 3-step procedure applied to the SEM with an event history submodel and a distal health submodel: low entropy (0.4) and large sample size ($N = 2000$) . . . 128

6.5 Estimated coefficients (in log-odds) and standard errors from the structural equation model (using the 3-step approach) for the effects of four categorical latent summaries of childhood SECs on a binary general health status at age 50, mediated by partnership transitions in adulthood. . . . 132

7.1 Simulation results for three scenarios of missing not at random (consider only models with time-invariant predictors): comparison of three modelling approaches (CCA, MI (FCS) and multilevel SEM). Estimates for parameters in the model for H_i are reported. The first three columns report relative bias (in %) and last column reports coverage rates of the nominal 95% confidence interval for the estimates for the multilevel SEM. . . . 152

7.2 Relative bias (%) for the second type of misspecification of the dropout submodel, where a common predictor for all processes is omitted from the dropout submodel. Results are shown for each of the MNAR scenarios. . . 154

7.3 Chi-squared tests for association between each of the four dimensions of childhood socioeconomic situations and having complete or incomplete partnership histories. . . . 156

7.4 Estimated effects of covariates in childhood waves on the log-odds of dropout during adulthood (standard errors in brackets). . . . 158

7.5 Estimated covariate effects on the log-odds of having poor health in midlife (standard errors in brackets). SEM (Model 1) refers to model (7.12) with a scalar D_i being the dropout outcome. SEM (Model 2) refers to model (7.13) with a vector $\mathbf{D}_i$ being the dropout outcome. 163
7.6 Estimated covariate effects on the log-odds of dropout at each wave (standard errors in brackets). 164
7.7 Effects of childhood SECs on partnership transitions. Results are presented for respective SEMs before (see also Table 6.5) and after including the dropout submodel (specifically, SEM (Model 2)). 165

A.1 Measure 1: father/male head social class (2.92% has missing data in all childhood waves) 190
A.2 Measure 2: financial hardship (2.23% has missing data in all childhood waves) 190
A.3 Measure 3: material hardship (12.52% has missing data in all childhood waves) 190
A.4 Measure 4: family structure (0.93% has missing data in all childhood waves) 191
A.5 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy=0.8) 195
A.6 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy=0.7) 195
A.7 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy=0.55) 196
A.8 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy=0.4) 196
A.9 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy = 0.8 in both LCAs) . 197
A.10 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy = 0.7 and 0.4, respectively in two LCAs) 197
A.11 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy = 0.4 in both LCAs) . 198
A.12 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a binary distal outcome (entropy=0.8). 199
A.13 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a binary distal outcome (entropy=0.7). 199
A.14 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a binary distal outcome (entropy=0.55). 200

A.15 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a binary distal outcome (entropy=0.4). 200

A.16 Simulation results for estimating a model with 2 associated latent categorical variables, a covariate and a binary distal outcome (entropy = 0.8 in both LCAs).201

A.17 Simulation results for estimating a model with 2 associated latent categorical variables, a covariate and a binary distal outcome (entropy = 0.4 in both LCAs).201

A.18 Study 2c: Simulation results for bimodality (N=200,2000; 500 replications). 203

A.19 Study 2a: Simulation results for excess kurtosis (500 replications): log-linear ω parameters. 204

A.20 Study 2b: Simulation results for skewness (500 replications): log-linear ω parameters. 205

A.21 Summary of results for class-specific effects on the distal outcome across all situations investigated, using the 1-step, 3-step ML and 3-step BCH approaches: Bimodality 206

A.22 Summary of results for class-specific effects on the distal outcome across all situations investigated, using the 1-step, 3-step ML and 3-step BCH approaches: Skewness 207

A.23 Summary of results for class-specific effects on the distal outcome across all situations investigated, using the 1-step, 3-step ML and 3-step BCH approaches: Kurtosis 208

A.24 Summary of results for class-specific effects on the distal outcome across all situations investigated, using the 1-step, 3-step ML and 3-step BCH approaches: Conditional dependence (1 latent variable) 209

C.1 Simulation study of the direction of estimation bias in the presence of endogenous variables 218

E.1 Joint model Part I: covariate effects on the logit hazard of first partnership formation 222

E.2 Joint model Part II: covariate effects on the logit hazard of partnership dissolution 223

F.1 An example input data structure under the discrete-time setting 225

F.2 Description of variables in the syntax for empirical analysis 225

G.1 An example of the stacked input data structure. 230

G.2 Description of variables in the syntax for empirical analysis 231

List of symbols

Greek Symbols

δ Censoring indicator

ωs Parameters in the log-linear model for the association between categorical latent variables

σ_u^2 Variance of the random effects

Superscripts

(D) Superscript for quantities related to missingness (or dropout)

(F) Superscript for quantities related to first partnership formation

(H) Superscript for quantities related to health

(S) Superscript for quantities related to partnership dissolutions

Other Symbols

$h_0(t)$ Baseline hazard function. Variant: α_t

C_q A latent class variable for dimension q of the childhood socioeconomic circumstances

D_i A single binary indicator for the occurrence of dropout across waves

$F(t)$ Cumulative distribution function

H Distal outcome (e.g. health)

$h(t)$ Hazard function, hazard rate. Variants: h_{ti} and h_{tij}

i Index for individuals, $i = 1, ..., N$

j Index for episodes of a repeatable event, $j = 1, ...J_i$

K_q Total number of categories for C_q

k_q Index for the latent class, $k_q = 1,...,K_q$

$\lambda^{(D)}$ Coefficient of the random effect term u_i for dropout

$\lambda^{(F)}$ Coefficient of the random effect term u_i for first partnership formation

$\lambda^{(H)}$ Coefficient of the random effect term u_i for distal health

$\lambda^{(S)}$ Coefficient of the random effect term u_i for recurrent partnership dissolutions

M_q Most likely class membership (modal class) for dimension q of the childhood socioeconomic circumstances

N Total number of individuals

n_{tij} Exposure in each period t of episode j of individual i

R Total number of waves of a survey

r Index for the wave of a survey, $r = 1,...R$

s Index for the discrete-time interval $[s, s+1)$ in the partnership formation process, $s = 1,...,S_i$

$S(t)$ Survival function

t Event or censored time. Also used to index time intervals $[t, t+1)$ in the discrete-time model

u_i Individual-specific time-invariant unobservables, $u_i \sim N(0, \sigma_u^2)$

$\mathbf{D}_i$ An $R \times 1$ vector of binary dropout indicators for individual i

$X^{(H)}$ Predictor of the distal outcome H

$\mathbf{X}_{tij}$ A set of time-invariant and time-varying predictors for duration

$\mathbf{Y}_q^{(C)}$ A $P_q \times 1$ vector of categorical indicators for the categorical latent variable C_q

$\mathbf{Y}^{(D)}$ A vector of unobserved longitudinal outcomes

$\mathbf{Y}^{(O)}$ A vector of observed longitudinal outcomes

$\mathbf{Y}$ A vector of (observed and unobserved) longitudinal outcomes

y_i, y_{ij} Duration variables for a non-repeatable event and a recurrent event

y_{tij} Binary (or grouped binary) response variable for the expanded person-episode-period file

Acronyms / Abbreviations

AIC	Akaike information criterion
ANOVA	Analysis of variance
BCH	Bolck-Croon-Hagenaars
BCS70	1970 British cohort study
BIC	Bayesian information criterion
BMI	Body Mass Index
CCA	Complete case analysis
CHD	Coronary heart disease
EHA	Event history analysis
EM	Expectation-maximisation
FEC	Forced expiratory vital capacity
FEV_1	Forced expiratory volume
FIML	Full information maximum likelihood
GMM	Growth mixture model
LCA	Latent class analysis
LCGA	Latent class growth analysis
LRT	Likelihood ratio test
LV	Latent variable
MC	Modal class approach
MCMC	Markov Chain Monte Carlo

MI	Multiple imputation
NCDS	National Child Development Study
OLS	Oridinary least square
PC	Pseudo class approach
PH	Proportional hazard
SD	Standard deviation
SEC	Socioeconomic circumstances
SEM	Structural equation model
SEP	Socio-economic position
SE	Standard error
ssaBIC	Sample size adjusted BIC

Chapter 1

Introduction

1.1 Motivation

In recent years, there has been growing interest in the sources of social and health inequalities. A better understanding of life course influences can be particularly beneficial to policies that aim to reduce these inequalities due to demographic and socioeconomic circumstances (SECs) in early life. Due to the availability of rich data collected from large-scale longitudinal (in particular, birth cohort) studies, recent methodological developments in life-course epidemiology have been made to test the influence of environmental risk factors on health outcomes at different life stages (see Galobardes et al. (2006) and Ben-Shlomo et al. (2016) for a comprehensive review). Aside from genetic reasons, previous research has found that disadvantage in multiple aspects of childhood, such as poor parental socioeconomic circumstances and housing conditions, family financial difficulties and parental relationship instability, is associated with poor physical (e.g. cardiovascular disease) and mental (e.g. depression) health in later life (e.g. Kelly-Irving et al., 2013; Poulton et al., 2002).

Some recent studies have also contributed to understanding of the mechanisms behind such associations by uncovering behavioural, physical and psychosocial pathways that mediate the effects of childhood SECs on midlife or later health (e.g. Ben-Shlomo et al., 2016; Cohen et al., 2010; Lacey et al., 2014). In particular, it has been hypothesized that experiences in childhood operate through influencing trajectories of life course events and functional changes in health-related behaviours that can mediate the effects of childhood SECs on later health. This thesis considers, as an example, the mediating effects of one of these life events: the experiences of partnership formation and dissolution up to midlife. Many studies have found that family characteristics during childhood have a significant impact on the timing of such events. For example, individuals from families of a relatively high social position tend to delay forming the first partnership (e.g. Berrington and Diamond, 2000; Steele et al., 2006) while individuals who have experienced parental separation during childhood are at increased risk of partnership dissolution themselves (Kiernan and Cherlin, 1999; Steele et al., 2006). Moreover, previous research has found a negative association between partnership instability and well-being (e.g. Brewer and Nandi, 2014; Maughan and Taylor, 2001).

We use the rich life course data collected in the 1958 National Child Development Study (NCDS) in Great Britain, including histories of partnership events collected for ages 16 to 50. To investigate the extent to which the effects of childhood SECs on midlife health are mediated by partnership experiences, a number of methodological challenges need to be addressed. First, multiple aspects of childhood SECs are of substantive interest, and information from repeated measurements over four childhood waves (ages 0, 7, 11 and 16) should be combined while accounting for missing data and measurement error (the latter is often caused by using a single indicator or ad-hoc summaries). Second, most previous research has used simple forms of structural equation models (SEMs) with only observed variables (including multiple mediators and confounders) to explore the pathways linking childhood circumstances to later health (e.g. Gale et al., 2009; Pakpahan et al., 2017). Few studies have additionally introduced latent variables to capture trajectories of development while accounting for measurement error (e.g. Green et al., 2012; Hagger-Johnson et al., 2011). Studies that have used event history data have been primarily interested in identifying risk factors of time-to-event outcomes, rather than relating the information captured by event histories to distal outcomes such as later health. Finally, attrition is a common issue in cohort studies and the reasons for dropout are likely to be associated with the outcomes of interest.

Considering the need to incorporate information from different life stages and domains (e.g. in socioeconomic and health regimes), this research explores the mechanisms underlying the effects of childhood circumstances on adult health in midlife. Potential pathways linking childhood circumstances, experiences of partnership formation and dissolution (examples of life events in adulthood) and midlife health are jointly considered. We aim to develop a general modelling framework that can address the outlined statistical challenges simultaneously, while making full use of the available longitudinal data to address the substantive questions.

1.2 Research objective

The conceptual modelling framework is illustrated in Figure 1.1 where childhood circumstances are related to the life events and midlife health. We then divide the overall research question into four main parts where each builds upon the previous one in terms of the longitudinal information incorporated and model complexity. The resulting sub-research-questions and the modelling objectives are proposed as follows.

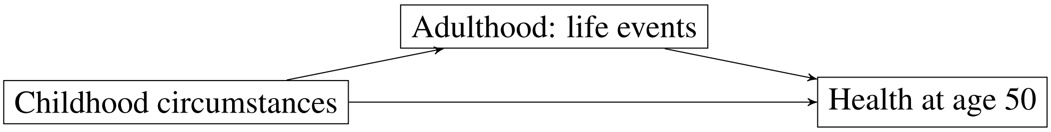


Fig. 1.1 General framework for the relationship between childhood circumstances, life events in adulthood and midlife health.

RQ_1: How should we measure childhood circumstances?

Childhood circumstances (socioeconomic situations) are key covariates in this research. In light of the large number of childhood variables in the dataset, a common practice is to include one or more measures (e.g. social class, financial situations) collected at different points in time as predictors in the model for various outcomes of interest. Repeated measures are seldom used in previous work, which is mainly because measurements are not consistent over time. Moreover, socioeconomic situations cover a wide range of aspects so that decisions need to be made about how to incorporate available information in the longitudinal dataset.

In this research, we aim to construct consistent repeated measures for each of the major aspects of childhood social and economic environment that have often been considered in previous research. For each aspect, derived repeated measures are treated as imperfect measurements of a latent time-invariant construct and a latent class analysis is performed to differentiate groups of people based on their longitudinal patterns of change over the set of repeated measures. The resulting latent summaries of childhood circumstances (categorical latent variables) are then related to partnership transitions, midlife health and the dropout tendency in later modelling stages.

RQ_2: How do childhood circumstances influence partnership transitions between ages 16 and 50?

Family situations during childhood have been shown in the previous literature to have an impact on moving into and out of partnerships (e.g. Berrington and Diamond, 2000; Kiernan and Cherlin, 1999). However, results on the direction and significance of such influences are mixed. We intend to offer more insights into this by taking into account the longitudinal information on childhood situations.

We can estimate an event history model that relates the derived latent categorical variables of childhood SECs to partnership transitions. Two main issues may lead to biased estimates of childhood effects: the potential misclassification in the latent class analysis and the presence of endogenous covariates in the model for partnership dissolution. Age at the start of partnership, a variable that is highly related to the outcome of the first partnership formation process (age at the first partnership), may be endogenous with respect to partnership dissolutions. Specifically, there exists unmeasured variables that are associated with the decision to form the first partnership early and also to separate early. If not accounted for, the shared influences of unobservables can lead to biased estimates of the effect of this endogeneous covariate and other covariates that are associated with it.

To address the first issue, methods to correct for misclassification of individuals into classes based on the most likely class membership (modal class) are discussed. The relative performance of these approaches in cases with multiple latent categorical variables and a distal outcome is evaluated using an extensive simulation study for multiple scenarios. To address the second issue, joint models for first partnership formation and recurrent dissolutions are considered.

RQ_3: Are there any mediating effects of partnership transitions on health at age 50?

RQ_3 builds upon RQ_2 to allow for summaries of an individual's partnership history (e.g. change in partnership status) to influence midlife health. Partnership transitions are therefore considered as both a predictor (in the partnership–health model) and an outcome (in the childhood SEC–partnership model). Also, we note that summaries of partnership experiences across 34 years (from age 16 to 50) may be endogenous with respect to midlife health.

We specify a structural equation model that jointly models partnership transitions and midlife health, with childhood circumstances (categorical latent variables) as common predictors. Partnership outcomes are therefore outcomes in one equation and predictors in another equation. To the best of our knowledge, this has never been attempted before as time-to-event data are generally used only as response variables in most health or social studies.

RQ_4: Are there any residual interrelationships between partnership experiences, dropout propensity and midlife health?

RQ_4 builds upon RQ_3 by accounting for the attrition in the data. Attrition (or dropout) is a common issue in cohort studies and leads to a loss of individuals over time. To address this, we model the dropout process simultaneously with partnership transitions

> and distal health under a relaxed assumption of non-ignorable dropout. We aim to investigate the impacts of the dropout mechanism on key parameters of interest: the effects of childhood SECs (primary interest) and partnership transitions (secondary interest) on midlife health and, if such effects are sensitive to the specification of the dropout model.

The remainder of the thesis is organised as follows. After a comprehensive review of the existing work on measures of childhood socioeconomic situations and adult health in Chapter 2, a description of the longitudinal data used in this research is provided in Chapter 3. Chapter 4 discusses the modelling strategy that relates multiple childhood SECs to a single distal outcome. An extension of the model to duration data is discussed in Chapter 5 where childhood SECs are related to partnership transitions (first formation and recurrent dissolutions). Chapter 6 describes a multilevel structural equation model where childhood SECs and summaries of partnership experiences are linked to midlife health. The treatment of missing data is discussed in Chapter 7. Finally, Chapter 8 concludes the thesis by summarising the main methodological and substantive contributions and, providing directions for future work.

Chapter 2

Review of previous research on childhood circumstances and adult health

2.1 Introduction

In the field of epidemiology, the impact of childhood experiences on later health outcomes has been widely studied. In birth cohort studies, a large number of childhood measurements are taken when respondents are below the age of 18. Following Mensah and Hobcraft (2008), these indicators during childhood can be roughly grouped into the following categories: general health situation, biomedical measures (e.g. height, weight), cognition and academic performance, family socioeconomic situation and behavioural-related issues.

After a discussion of childhood and adult measures in Section 2.2 and Section 2.3, potential pathways through which childhood circumstances influence adult health are reviewed in Section 2.4. The statistical methods used to explore their associations are summarised in Section 2.5.

2.2 Measures of childhood circumstances

2.2.1 General health

Previous studies have used various measures of general health during childhood, such as an ordinal indicator for general well-being (from excellent to poor) of participants by the age of 17 (Elo, 1998), a categorical indicator for the existence of infectious long-term health issues and a binary record of having physical accidents and admission to hospital at age 7 (Jones et al., 2009). Examples of repeated measures from the 1958 National Child Development Study (NCDS) in Britain (also referred to as the 1958 British cohort study) include the number of mental and physical chronic conditions at ages 7 and 16 (Case et al., 2005), a binary indicator for the presence of health problems at ages 7, 11 and 16, as well as their different combinations (i.e. presence of health issues at both 7, 11 or 11 and 16 years, see Jackson (2010)).

2.2.2 Biomedical

Aside from indicators of general health, the use of biomedical measurements of height, weight and mental health in the postnatal and childhood period has been prevalent in the previous literature. Birth weight has been used in a number of studies using birth cohort data, for example the 1958 British cohort (Bartley et al., 1994; Mensah and Hobcraft, 2008; Potter and Ulijaszek, 2013), the 1946 British National Survey of Health and Development (Orfei et al., 2008) and four Swedish cohorts from 1914, 1918, 1922 and 1930 (Andersson et al., 2001).

The second most commonly studied measures are the body mass index (BMI) and height. Using data from the 1972 Bogalusa longitudinal heart study in the United States where participant records were collected at the ages 5, 6, 7, 8 and 19-23, Freedman et al. (2001) used repeated measures of BMI and height during childhood and included the minimum BMI, as well as the age at which minimum BMI was reached as explanatory variables. Barker et al. (2005) calculated children's BMI from repeated measures of height and weight starting from birth up to two years of age and annually thereafter until age 11. Repeated BMI measures at ages 2 and 11 were grouped into three quantiles and combinations of them (e.g. the first quantile of BMI at age 2 and the third quantile BMI at age 11) were included as predictors of coronary disease in midlife (Torfadottir et al., 2012). Using data from the 1946 and 1958 British birth cohort studies, Orfei et al. (2008) considered the postnatal height growth as the percentage of height measured at age 7 versus that in midlife.

Other biomedical measures of physical health include genotypes related to fat mass, obesity and children's gestational age at birth (Kaakinen et al., 2010), records of the presence of nine health conditions at ages 7, 11 and 16 (Goodman et al., 2011) and breastfeeding history (Jackson, 2010; Potter and Ulijaszek, 2013) from the 1958 British cohort study. In terms of indicators for mental health, psychological examination results have been used, together with general health questions in surveys. For instance, Goodman et al. (2011) used malaise scores from the medical questionnaire and repeated measurements from parent reports to create a binary indicator for the presence of moderate to severe mental problems before age 16.

2.2.3 Education and cognitive ability

With regard to measures of childhood cognition and academic performance, results from cognitive tests have been widely used. In an analysis of the 1970 British cohort study (BCS70), Feinstein and Bynner (2004) used a measure based on cognitive tests commissioned at ages 5 and 10, comprised of reading, general ability and picture language tests. From the 1958 British cohort study, Chandola et al. (2006a) considered the general ability test scores measured at age 11. For both the 1958 and 1970 British cohorts, Gale et al. (2009) and Henderson et al. (2012) used a general summary of cognitive ability derived from a principal components analysis while Jackson (2010) used the mean reading and mathematics score recorded at ages 7 and 11 and Power et al. (2010) standardised the general ability score at age 11. Similar measurements of cognitive ability were considered by Hagger-Johnson et al. (2011) using the Aberdeen children of 1950s study and Gale et al. (2009) using four cohort studies (the 1936 Aberdeen birth cohort, 1921 Lothian birth cohort, 1946 and 1958 British cohort studies). Other education-related factors during childhood include parental and self-expectations of a child's further education or job prospects and the number of passes in O-level and A-level tests by age 18 (Jackson, 2010).

2.2.4 Socioeconomic situation

In general, socioeconomic circumstances (SEC) can be regarded as summaries of the social and economic characteristics of an individual, their household or family and the associated community (Entwisle and Astone, 1994). Due to the difficulty in identifying perfect measures for such abstract constructs, several practical recommendations on possible measures have been proposed. The major categories of measurements are: parental occupational status, education level and household income. It has been suggested that occupation of the principal income contributor in the household, regardless of the gender should be recorded (Hauser, 1994). Where data are available, information about both parents should be collected, including their labour force status (e.g. whether he/she is working part-time, full-time or self-employed) and the highest level of education achievement, for a better understanding of the economic situation of the household (Entwisle and Astone, 1994; Hauser, 1994). To transform various job categories into SEC, Nakao and Treas (1992) designed a socioeconomic index (continuous score) based on categories of occupation recorded in the 1980 census.

An alternative is to use the classification of social class proposed in the Registrar General of 1934 (General, 1933), which employed a five-class system (class I to V were coded "professional", "managerial", "skilled", "semi-skilled" and "unskilled"). Hauser (1994) also suggested recording the total amount of income in the previous year rather than the inaccurate measure of volatile short-term income.

Specifically, the following measures have been considered in multiple studies. Social class, socioeconomic position or status (SEP) categorised by the five-class system (General, 1933) are the most commonly used measures. For example, Frankel et al. (1999) considered the SEP from the 1937 British Carnegie survey of diet and health (Boyd Orr cohort) where its classification was based on the occupation of the head male in the family (also used by Kaakinen et al. (2010)). Other indicators include dummies for each social class (Frankel et al., 1999), a binary indicator for father's social class to distinguish manual and non-manual occupations (Kelly-Irving et al., 2013; Maggs et al., 2008) and a general indicator using both father and grandfather's social class under the five-class classification system of the Registrar General of 1934 (Jackson, 2010).

The financial situation during childhood is another common measure of socioeconomic situation. It can be represented by several indicators such as parental income, working status and housing situation. Using the 1958 British cohort study, Jackson (2010) considered the monthly family income in intervals recorded at age 16, transformed into a continuous scale by taking the logarithm of the midpoint of each bracket. Lane et al. (2013) used the record of family total income before the birth of the child. Bartley (2003) studied the effects of childhood financial adversity level on midlife lung function. In this work, financial adversity by the age of 11 was a summary of three variables: whether the father was in a low occupation category or unemployed at the child's birth, the assessment of a health visitor about the financial adversity level of a household, and if the participant had free school meals at age 11. For parental working status, Jackson (2010) created a dummy variable at each measurement year for whether the mother was employed outside the home. For housing situation, Hagger-Johnson et al. (2011) considered housing tenure and created an overcrowding index calculated as the number of people per room (also used by Kelly-Irving et al. (2013)).

2.2.5 Environmental

During childhood, participants are exposed to a positive or negative family environment depending on parents' health-related behaviour during pregnancy and childhood, as well as their parenting styles. Some studies label these indicators as "childhood adverse experience" (e.g. Jones et al., 2009; Kelly-Irving et al., 2013). For example, using the 1958 British cohort study, Jones et al. (2009) created binary indicators for parental divorce or separation, alcohol problems and the presence of financial difficulties in childhood. Using the same cohort, Power and Jefferis (2002) considered an indicator of maternal smoking during pregnancy (also considered by Case et al., 2005; Jackson, 2010; Kaakinen et al., 2010; Lane et al., 2013). Kelly-Irving et al. (2013) used an indicator of whether a child had experienced physical neglect, been put into social care or had parents with mental illness (also considered by Lane et al., 2013) by age 16 and parental alcohol problems measured at age 7. In terms of parenting behaviour, Lane et al. (2013) considered positive parenting (an ordinal scale with 10 levels) and the four-category parenting style (i.e. authorities, neglectful, authoritarian and permissive) measured at age 5.

2.2.6 Behavioural and diet

This last category summarises measurements that are related to children's behavioural problems (reported by parents or teachers), existence of physical activities and dietary habits. Kaakinen et al. (2010) took into account children's level of participation in activities and the presence of drinking and smoking behaviour at age 14. Using the 1966 Northern Finland birth cohort, Green et al. (2011) studied the population of cancer survivors during childhood for which their level of physical activity and general function were measured. Several post-cancer symptoms related to pain and anxiety were also considered as predictors of adult obesity (Green et al., 2011). Sigurdsson et al. (2002) devised a single binary summary (whether or not motor impairment exists) of five measurements of motor skills (e.g. coordination and hand control) recorded at ages 7 and 11 for the 1958 British birth cohort. Regarding dietary habits (such as the frequency of milk intake considered by Torfadottir et al., 2012), multiple measurements have been widely discussed in papers published between 1975 and 1998 (Parsons et al., 1999).

2.2.7 Summary

In this section, we have discussed a number of aspects of childhood circumstances. Among these measures, we are particularly interested in childhood socioeconomic circumstances because of its multidimensional complexity, its substantive importance for later health and the fact that this area is methodologically understudied. Also, the dataset that we use for this thesis (described in Chapter 3) contains repeated measures of multiple dimensions of socioeconomic circumstances on four occasions in childhood, providing rich information for our modelling interest. Other childhood measures that have been discussed above are either not recorded consistently in childhood waves, or contain a large amount of missing data. In analyses discussed in later chapters, we therefore focus on developing models to extract information from childhood SECs. While relating these socioeconomic measures to later health, some of the remaining childhood measures (e.g. biomedical measures) are considered as control variables. We also note that although the proposed methodologies (discussed in later chapters) are illustrated using childhood SECs, to accommodate other research questions, they can be applied or adapted to other childhood measures reviewed in this chapter, should data be available.

2.3 Measures of adult health outcomes

Measurements of health outcomes in adulthood (over age 18) vary greatly across studies. In general, they can be categorised into five major groups, namely mortality, presence of physical disease, general health problems, mental health and health-related behaviour.

2.3.1 Mortality

Mortality is a commonly used response variable in various analyses. Frankel et al. (1999) analysed records of death due to coronary heart disease (CHD) and stroke by 1997 from the National Health Service Central Register. The administrative data were then linked to the participants from the 1937 British Boyd Orr cohort study to obtain childhood measures of health. A similar approach, using the same administrative data, was adopted by McCarron et al. (2002) to trace participants studying at Glasgow University during 1948-1968 who had a record of death in midlife from causes related to cardiovascular and respiratory disease.

2.3.2 Presence of physical disease

Cancer

Apart from records of mortality in administrative data, records of the diagnosis of adult disease form another major group of health outcomes. It is also common for studies to consider more than one adult health outcome. In the case of cancer, Andersson et al. (2001) used the cancer registrar's record until 1998 in their study of four Swedish birth cohorts. The registry recorded the time of the first diagnosis of all types of cancer and tumours. Similarly, Torfadottir et al. (2012) retrieved the information about the diagnosis of prostate cancer, as well as its severity for each recorded individual from the Icelandic cancer registry. In the 1958 British birth cohort study, self-reported records of cancer were collected at ages 46 and 50 and a binary indicator for having cancer (or not) by the age of 50 was created to be the response variable (Kelly-Irving et al., 2013).

Heart problems

In the case of heart problems, hospital admission records of CHD when participants were in their 40s and 50s were considered using the 1934-1944 Helsinki cohort (Barker et al., 2005). Indicators of the presence of CHD between ages 34 and 53 was also used by Park et al. (2013) in a study of three British birth cohorts (1946, 1958 and 1970). Other records available in these studies include that of the existence of long-standing illness and medical supervision. As alternatives to these direct measurements of heart disease, Power et al. (2010) focused on the risk factors of heart disease measured at age 45 for the 1958 British birth cohort. Available measures include BMI, circumferences of the waist, blood pressure and biomedical indicators associated with the presence of diabetes (cholesterol and triglyceride level as well as HbA1c).

Obesity and diabetes

BMI is a commonly used measure of obesity. Parsons et al. (1999) reviewed papers from 1975 to 1998 and obesity was defined mostly in terms of fatness, leanness, their changes over time and percentiles of BMI. Freedman et al. (2001) used the BMI measure (also collected from the study of childhood cancer survivors by Green et al. (2011)) from the United States Bogalusa Heart Study when participants were between ages 19 and 23. Power and Jefferis (2002) transformed the raw repeated measures of BMI at ages 23 and 33 from the 1958 British cohort study into binary indicators of obesity (see also Chandola et al. (2006a,b) and Potter and Ulijaszek (2013) who also considered BMI at age 42). Continuous measures of raw BMI were used by Chandola et al. (2006a,b) in order to explore the progress of weight gain across the life course. BMI at age 31 was used in an analysis of the 1966 Northern Finland Birth cohort (Kaakinen et al., 2010). For diabetes, Park et al. (2013) considered the number of medical supervision cases and that of hospital admission as indicators for having Type II diabetes.

Lung dysfunction

Other indicators of adult health, such as lung function in midlife have been mostly measured by FEV_1 (forced expiratory volume, with a reduced value in the presence of lung disease) and FVC (forced expiratory vital capacity, often measured jointly with FEV_1 as an indicator for lung function) in the 1946 (Orfei et al., 2008) and 1958 British cohort studies (Bartley et al., 2012; Orfei et al., 2008).

2.3.3 General health problems

Aside from the measures of a specific disease, general indicators for health in adulthood have been widely used in previous research. Elo (1998) considered a general measure of self-assessed health status in adulthood ranging from excellent to poor (also used by Henderson et al. (2012)) and a binary indicator for the presence of health problems, including chronic heart disease, stroke and lung disease. The 1958 British cohort study has questions about the self-reported general health status. For example, Case et al. (2005) used repeatedly measured health status at ages 23, 33 and 42 and treated them as separate outcomes. Mensah and Hobcraft (2008) used the record of self-reported long-standing illness and general health status measured at around age 30 from two British cohort studies (1958 and 1970). Chronic widespread pain was considered by Jones et al. (2009) at ages 42 and 45. The record of long-term sick leave after age 33 was used by Henderson et al. (2012) for three British birth cohorts. Ploubidis et al. (2015) considered a set of biomarkers of health status at ages 44-46 from the 1958 British cohort study. All biomarkers were collected in the medical survey and covered issues related to respiratory, metabolic function and inflammatory problems.

2.3.4 Mental health

Another major group of health problems, in addition to issues related to physical health, are concerned with mental health. The 1958 and 1970 British cohort studies contain a clinical assessment of malaise which has been often taken as a medical indicator for mental health (e.g. Feinstein and Bynner, 2004; Mensah and Hobcraft, 2008). It was administered when participants were roughly age 30. Gale et al. (2009) studied four birth cohorts (1921 Lothian birth cohort, 1936 Aberdeen birth cohort, 1946 and 1958 British cohorts) and scores from the professional Warwick-Edinburgh scale of mental health were collected at around the ages of 74, 87, 50 and 60 years, respectively. Answers to the question related to the self-reported suffering of hypertension were also used in a pooled analysis of the 1946, 1958 and 1970 British birth cohorts (Park et al., 2013).

2.3.5 Health-related behaviour

In addition to the above-mentioned health outcomes, researchers have also considered the health-endangering behaviours in adulthood as outcome variables (including smoking, alcohol consumption and illegal drug use). Self-reported smoking behaviour (categorical smoking status together with binary or ordinal measures of frequency) was collected in the 1958 and 1970 British cohort studies when participants were in their early 30s (Feinstein and Bynner, 2004; Gale et al., 2009). Maggs et al. (2008) used repeated measures of the quantity of alcohol use per week at ages 23 and 33 and a binary measure of harmful alcohol consumption at age 42 from the 1958 British cohort study. Using the same cohort study, illegal use of different types of drug over the past year was considered (White et al., 2012).

2.3.6 Summary

Among the health outcomes reviewed above, self-reported general health state at age 50 is used to illustrate our proposed methodology (discussed in later chapters). This type of midlife health outcome has been prevalent in previous health studies and is often available in multiple datasets (including the one that we use in this thesis) as a general measure of overall health, including both physical and mental well-being. From a methodological viewpoint, health state is often measured on a categorical scale. The models that we develop for this type of variable can therefore be generalised to other health outcomes straightforwardly. Moreover, the modelling framework that we develop in Chapter 6 can be extended to include multiple health outcomes (possibly of mixed types) or repeated measures of a health outcome on multiple occasions. This may be an interest for future research where a particular set of health outcomes are of interest.

2.4 Pathways linking childhood circumstances and adult health outcomes

While exploring the relationship of interest, different pathways through which childhood situations may affect adult health have been proposed and tested. It should be noted that the labelling of confounders and mediators in this section is based on the setting of Figure 1.1. For example, if a confounder used in one previous study is more suited to be a mediator in our research, it will be grouped into the mediator category. Such re-labelling is clearly described below.

2.4.1 Confounders

Confounding is a major issue in epidemiology studies. A confounder is defined to be a risk factor of the outcome (e.g. disease or health-related outcomes) that is associated with, but is not an intermediate result of the exposure (the key risk factor of interest) (Salas et al., 1999). A simple example of confounding is depicted in Figure 2.1 where X is the exposure (a childhood SEC measure in our case) and Y is the outcome variable (a health outcome in our case). Z is defined as a confounder if it is a risk factor of outcome Y and also has an impact on exposure X.

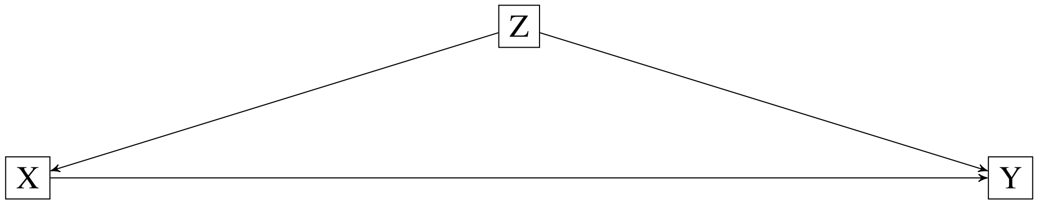


Fig. 2.1 Confounding in the general form of Figure 1.1. Y is a health outcome, X is a childhood SEC and Z is a confounder.

In a regression of Y on X, neglecting the confounder leads to omitted variable bias and hence spurious associations (i.e. biased point estimates) may be concluded. A large amount of past literature has attempted to address this problem by including potential confounders as additional explanatory variables in regression models.

To model the relationship between childhood SECs and health in adulthood, one set of variables that has been commonly considered as confounders is the SEC of parents, including parental occupation and social class (e.g. Case et al., 2005; McCarron et al., 2002; Orfei et al., 2008), education levels (e.g. Case et al., 2005; Kelly-Irving et al., 2013) and the financial situation (e.g. Case et al., 2005; Elo, 1998; White et al., 2012) measured by household overcrowding and the receipt of state benefits. Most of the parental SEC measures used in these studies are similar to those described in Section 2.2.4.

Other potential confounders are physical measures on the parents, for example height (Case et al., 2005), BMI (Chandola et al., 2006b; Power and Jefferis, 2002), maternal age at child's birth (Kaakinen et al., 2010; Kelly-Irving et al., 2013), hypertension (Kaakinen et al., 2010), as well as the disease history in the family (Torfadottir et al., 2012). Other pre-exposure measurements are also commonly seen in previous papers. For example, Andersson et al. (2001) used measurements of maternal protein level and parity (also considered by Kaakinen et al. (2010); Power and Jefferis (2002)). Another example is the health-related behaviour of parents, such as the smoking behaviour of the mother during pregnancy (Kaakinen et al., 2010; Kelly-Irving et al., 2013; Lane et al., 2013).

Other confounders that have often been controlled for in models for health outcomes include measurements collected at birth of the respondent or before the time of measurement of the key exposure. For example, quantiles of birth year have been used to control for cohort effects (McCarron et al., 2002). Mental health measured by anxiety level at age 11 was conditioned on while estimating the effect of intelligence at age 11 on midlife illegal drug use (White et al., 2012). Ploubidis et al. (2015) used the measurements of education, self-reported health status and other medical assessment results at age 23 as confounders in an analysis of the effects of partnership status trajectories for ages 23-42 on health state at ages 42-44.

2.4.2 Mediators

In order to understand the relationship between the measurements in childhood and midlife health outcomes, intermediate variables (mediators) have often been considered. Conditioning on these variables, researchers have attempted to uncover the direct and indirect effects of early-life exposure on adult health conditions (MacKinnon et al., 2002). Figure 2.2 is a simple illustration of a mediation structure. Again, X and Y are the exposure and outcome variables, respectively. W is an intermediate outcome of X such that the effect of X on Y is mediated through W. We therefore obtain two pathways, the direct effect of X on Y controlling on W and the indirect effect of X on Y through W. W is therefore defined as a mediator in this relationship. In this simple mediating structure, adding up the direct and indirect effects gives the total effect of X on Y (for continuous W and Y), assuming a collection of assumptions for causal analysis is valid (Stone, 1993). In our study of the effect of childhood SECs on midlife health, W indicates life events in adulthood (e.g. partnership, employment and fertility events). To illustrate the methodology, we consider in later analyses partnership events, i.e. the formation and dissolution events as examples of mediators (mediating processes, to be precise).

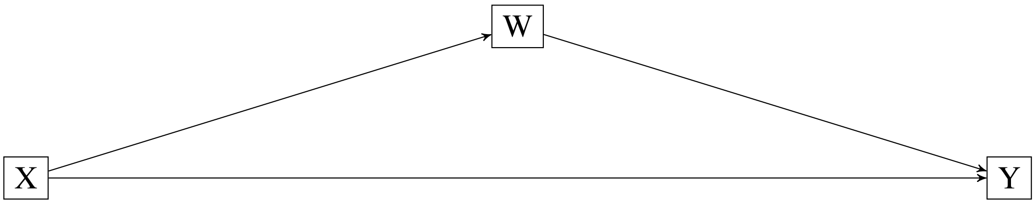
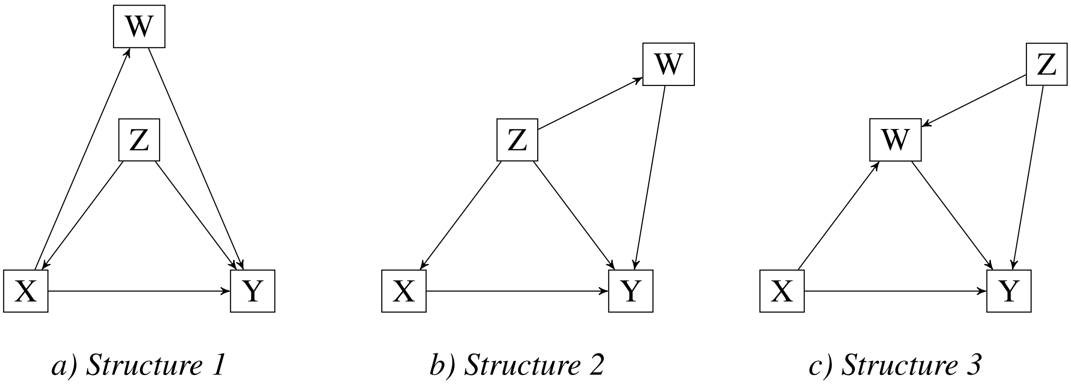


Fig. 2.2 Mediation in the general form of Figure 1.1. Y is a health outcome, X is a childhood SEC and W is a mediator.

2.4.3 More complex structures

To address the research questions summarised in Section 1.2, a complex structure is required considering the existence of both confounders and mediators. Examples of different structures are presented in Figure 2.3, in which Z denotes a confounder, W denotes a mediator, X denotes the exposure and Y denotes the outcome. Structure 1 illustrates a situation with Z being the confounder and W being the mediator but without intermediate confounding or mediation in the X to Y relationship. In Structure 2, W is a mediator in the Z to Y relationship and is not influenced by X. In Structure 3, W is a mediator in the X to Y relationship and Z is a confounder in the W to Y relationship.

Our research question (see Figure 1.1) can be addressed via a joint consideration of such structures. Specifically, the interrelationships between childhood SECs (X), partnership events (W) and midlife health (Y) are investigated by assuming a structure that is a combination of structures 1 and 3 depicted in Figure 2.3, where Z can represent observed confounders such as childhood health state or unmeasured confounders (in this case, squares should be replaced by circles to indicate latency), such as personal characteristics. Chapters 4 to 7 are devoted to the discussion of modelling strategies, the necessity, the selection and the treatment of these variables. We note that this hypothesised structure is motivated by findings from previous literature and we have assumed that intermediate mediators (i.e. W in structure 2) are not present in the modelling framework. Alternative structures can be explored by modifying the models proposed in later chapters and the comparison of different structures themselves constitute a separate research topic. These are beyond the scope of this thesis but are interesting fields for future research.



a) Structure 1 *b) Structure 2* *c) Structure 3*

Fig. 2.3 More complex structures that involve both confounders and mediators.

2.5 Statistical methods to model relationships between childhood circumstances and adult health

The increased availability of longitudinal datasets allows for the specification of complex models to empirically test the hypothesis on the mechanism through which childhood experiences influence midlife health via social experiences in adulthood (Kuh et al., 2003). Statisticians and epidemiologists have made multiple attempts to address this problem, or at least part of it. Two of the most prevalent methods that have been used in the past literature are forms of regression analysis and more advanced methods that can be formulated as structural equation models (e.g. growth curve models and latent class analysis).

2.5.1 Regression analysis

Multiple linear regression

Multiple linear regression is commonly used when the outcome variable is continuous, such as raw BMI, height and weight. In the linear setting, the outcome variable is regressed on a set of explanatory variables, often taken at different stages of life in longitudinal studies. Linear association can be easily tested and adjustments for gender, age group, ethnicity and residence at birth are commonly made to control for omitted variable bias (Bender, 2009). It is a traditional approach in epidemiology studies and hence a large number of studies used this method. We list here only some recent ones as examples. For instance, Freedman et al. (2001) considered the association between "age at which minimum BMI was achieved" and "adult BMI level", and tested whether this relationship was sensitive to changes in childhood BMI levels at different ages. Orfei et al. (2008) fitted separate models for two measures of lung function and Gale et al. (2009) estimated the effect of cognitive ability on mental health in midlife. It should be noted that Maggs et al. (2008) also employed a hierarchical modelling approach by including predictors of adult health at different ages step by step (starting with background SEC variables and then adding sets of childhood, adolescent and early adulthood variables).

Logistic regression

When the outcome variable is not continuous, such as the presence of a disease or death, logistic regression methods have been adopted (e.g. Clayton et al., 1993; Elo, 1998; White et al., 2012). For ordered outcomes, such as the level of health status, the cumulative logit model has been fitted in many studies (e.g. Elo, 1998; Jackson, 2010; Park et al., 2013).

Survival analysis

In many cases, the outcome variable of interest is the duration of time from when an individual becomes at risk of an event until an event occurs (e.g. diagnosis of disease or death). Event history data exhibit two main features that should be handled properly in the model. First, event times are not fully observed. For individuals who have not experienced the event at the end of the observation period, the time-to-event is right-censored. These individuals cannot be simply excluded as they do not form a random sample of the population (most of them have a low hazard of experiencing an event); event times cannot be treated as complete as this understates the true duration. Second, there may exist time-dependent covariates and we are interested to know how the risk of experiencing an event at time t is influenced by the value of such covariates at that time. Survival analysis (also termed event history analysis) has been widely used in previous studies to model such data. The questions that we are interested to answer are, for example, if a person has not experienced the event of interest by time t, what is the risk that this particular person would experience that event later; and if there are other factors that influence the timing of this event (either time-invariant, such as gender, or time-varying, such as employment status), how do they affect the probability of the occurrence of the event (e.g. Klein and Moeschberger, 2003; Steele et al., 2004). To model event history data, continuous-time and discrete-time models have both been popular in the literature. Among the continuous-time approaches, the Cox proportional hazard (PH) model that does not require a specification of the baseline hazard (Blossfeld and Gotz, 2001) is the most widely used and there have been multiple applications of the Cox PH model to health problems (e.g. Frankel et al., 1999; McCarron et al., 2002; Torfadottir et al., 2012). Commonly used time-invariant predictors include the height measured at baseline time and individual characteristics such as gender (Frankel et al., 1999). In terms of time-varying covariates, Torfadottir et al. (2012) considered milk intake levels measured repeatedly at different stages of life (adolescent, midlife and present). In social research, as event times are typically recorded retrospectively to the nearest month or year (i.e. measured in time intervals), a discrete-time setting is more natural. After data restructuring, regression models for binary or binomial responses can be fitted to interval-censored data. Discrete-time models have been commonly considered in epidemiology studies (e.g. DeWit et al., 2000; Reardon et al., 2002) and social research (e.g. Lillard et al., 1995; Steele et al., 2005). A comprehensive review of discrete-time event history models can be found in Allison (1984) and theories and concepts relevant to this approach are summarised in Section 5.2.

Methods for repeated measures

When the outcome of interest is measured repeatedly over time, fitting separate multiple regressions for each outcome variable or a multivariate multiple regression are two common approaches. For example, Torfadottir et al. (2012) fitted two Cox PH models to estimate the effects of milk intake and other covariates on two stages of prostate cancer. Power and Jefferis (2002) studied how the odds of obesity at five ages (7, 11, 16, 23, 33) varies over time due to maternal smoking using the multivariate approach and Kelly-Irving et al. (2013) modelled the repeated measures of self-reported cancer at ages 33, 42, 46 and 50 as outcome variables, predicted by a set of explanatory variables collected in the 1958 British cohort study.

Summary of limitations

Apart from the many advantages of the approaches discussed above, they suffer from several limitations. To be specific, although all the above methods could help to examine the associations between exposures during childhood and health outcomes in adulthood, they cannot provide information about the underlying pathways through which childhood effects transmit over the life course. For longitudinal studies in particular, measurements are taken at different stages of life and hence there exists correlation between these measurements for the same participant. In most of the studies mentioned above, repeated measures were simply included as predictors in the regression model (Freedman et al., 2001; Gale et al., 2012). As they are highly correlated, standard errors may be overstated, which can lead to spurious conclusions about the significance of effects of interest. These repeated measures are also likely to be subject to measurement error and have missing values, which may be additional sources of bias in estimates. The predictors and the outcome variable in the regression model may also share common influences (confounders), either observed (but not controlled for) or unobserved. Together with the potential existence of mediators, these issues can lead to spurious conclusions about the relationship of interest.

Some studies have considered this issue, but simply adjusted for confounders and mediators by including extra predictors in the model, often ignoring temporal ordering of the quantities (e.g. Gale et al., 2012; Power et al., 2010). This is only a partial remedy for omitted variable bias and as more variables are introduced into the model, we may have more endogenous predictors and a greater amount of measurement error in covariates. Structural equation modelling (SEM) is a preferred tool that can mitigate estimation bias by fitting a set of equations simultaneously under a hypothesized construct and therefore temporal ordering of the predictors can be accounted for. The measurement model accounts for the measurement error in a set of variables that capture a broad phenomena (latent variables), and the structural model reveals the underlying relationship between these latent constructs. Many researchers have used the forms of SEM described below to incorporate confounding and mediation structures as well as latent variables in order to explore various potential pathways underlying the associations of interest.

2.5.2 Structural equation modelling

SEM with only observed variables – Path analysis

Case et al. (2005) studied the influence of childhood SEC on self-reported health status in midlife using an SEM with all observed variables and discrete outcome variables. Education level during adolescence and SEC and health status in young adulthood were regressed on childhood variables. SEC and health status in mid-age were then regressed on health status and SEC in early adulthood, education in adolescence and childhood variables. In this analysis, adolescent and early adulthood variables acted as mediators. However, more pathways can be considered. For example, adolescent education level could impact SEC and health in early adulthood, but such pathways were neglected in the model. With repeated measures available in their studies, it is also possible to use growth curve modelling or latent class analysis to examine trajectories or patterns of changes in SEC and health over time.

Kaakinen et al. (2010) modelled the effect of the genotype of fat mass and obesity on the BMI progression across a lifetime (measured at birth, ages 14 and 31), adjusted for maternal health and SEC, health-related behaviour at ages 14 and 31 (these covariates are attached only to the BMI variable at the corresponding age). One limitation of this study is that predictors of BMI at an early life stage (such as the health-related smoking, drinking behaviour and dietary habits) could have a lasting impact on BMI at a later age. This can be tested via additional pathways. Including in the model potential mediators such as health-related drinking and smoking behaviour in early adulthood, as well as life events such as separation and child bearing (for the female population) may also be considered in the model.

SEM with latent variables

Despite the widespread use of path analysis, the variables used in the model are considered as perfect measurements of different characteristics. In previous studies, most health measures were self-reported and path analysis with only observed variables neglects the potential measurement error. Considering this, latent variables are often introduced into the model when there exists a multi-dimensional concept (e.g. socioeconomic circumstances, overall health) that cannot be perfectly measured. Many studies have employed the SEM approach with latent variables that are measured by multiple indicators. For instance, Hagger-Johnson et al. (2011) studied the effects of childhood circumstances on adult health using the Aberdeen children of the 1950s cohort. Two latent constructs for childhood situations were designed: latent childhood SEC (measured by five indicators of family socioeconomic background at birth) and latent intelligence at age 11 (measured by four indicators of verbal, mathematics and language ability). They also tested another SEM where indicators of the first latent variable were treated instead as predictors of latent childhood SEC. In both models, education level in midlife and the latent intelligence at age 11 were regarded as mediators. Green et al. (2011) specified an SEM with a latent variable for physical function measured by five indicators to study the influence of radiation treatment, physical activities, lifestyle, cancer-related pain and anxiety on the presence of adult obesity for a population of child cancer survivors. The variables included in the SEM were those that had a significant association with the obesity outcome in adulthood from a standard univariate logistic regression analysis.

To handle longitudinal data, growth mixture models (GMM) have been commonly used. Instead of assuming a homogeneous population, GMM assume a heterogeneous population that consists of multiple unobserved population classes, each with different patterns of growth (characterised by different sets of growth parameters). The latent class growth analysis (LCGA) is a special case of GMM that estimates an average growth curve for each sub-population. In GMM, individual variation (in the form of random effects on the intercept or slope terms or both) around the average growth curve are allowed. Hagenaars and McCutcheon (2002, Chapter 10) contains a review of the methodology of GMM and multiple applications of the model to health studies.

2.5.3 Missing data

Missing data is a common problem for cohort studies. Following Little and Rubin (1987), the mechanism leading to missing data can be classified as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). MCAR is the most restrictive assumption where the probability of missingness is assumed to be independent of both the outcome (which has missing values) and the observed covariates. The weaker assumption of MAR assumes that the missingness is independent of the outcome conditional on observed covariates. Both MCAR and MAR are forms of "ignorable" missing data mechanisms where the outcomes of interest can be modelled without an explicit account of the missing data mechanism (i.e. parameters for the analysis model and the missingness are distinct). MNAR is the most general assumption where the missingness probability is related to the missing values. In the following, we briefly review different treatments of the missing data under these assumptions.

Under the assumption of MCAR, listwise deletion is the most commonly used method (e.g. Næss et al., 2007; Turrell et al., 2007) where records with missing data either in the outcome or the predictors are excluded from the analysis sample, leading to a complete-case analysis. The reduction in the sample size can inflate the standard errors and, more importantly, the remaining individuals may be different from those who were excluded. The more relaxed assumption of MAR requires a broader consideration of statistical tools to handle missingness, of which imputation and maximum likelihood methods are the most popular in the literature. Simple imputation methods include substituting means or conditional means from a single regression for missing values. In addition to generating biased point estimates, this also leads to an underestimation of standard errors as the uncertainty in imputed values is ignored. To remedy this, the multiple imputation (MI) method generates random variation in the imputation process and the analysis model is fitted to each of the multiply imputed datasets with results pooled across datasets. MI has been widely used in previous studies (e.g. Jackson, 2010; Kelly-Irving et al., 2013; Power et al., 2010) using the Markov Chain Monte Carlo (MCMC) methods and fully conditional specifications. However, a number of arbitrary decisions are required to perform MI. For example, the imputation model needs to be compatible with the analysis model and it is preferrable to include in the imputation model additional variables (that are highly correlated with the outcome) to make the MAR assumption more plausible. The choice of such auxillary variables and their interactions is, however, arbitrary. The specification of the imputation model can be error-prone when multiple variables need to be imputed. This can lead to biased estimates, particularly when the proportion of missing data is large. Also, a large number of datasets need to be used to reduce the variability of estimates but generating too many datasets can be time-consuming, especially when the sample size is large and the model of interest is complex. Finally, different random-number seeds lead to different results from MI. Under the MAR assumption, Seaman and White (2013) discussed an alternative "inverse propensity weighting" (IPW) method and compared its performance with MI. Unlike MI that specifies the joint distribution of missing data conditional on observed variables, IPW models the probability of being completely observed and individuals with complete data are weighted by the inverse of their probability of having complete information. With only one model to be specified, compared with MI, IPW is less prone to misspecification error. However, it has been noted that IPW is generally less efficient than MI because MI incorporates information from individuals with partially missing data. Maximum likelihood methods, under the MAR assumption, have often been adopted in the literature (e.g. Chandola et al., 2006a; Gale et al., 2009; Ploubidis et al., 2015) due to its improved efficiency (e.g. MI requires a large number of imputations to achieve full efficiency), no need to specify additional models (e.g. imputation models) that are subject to misspecification, a greater flexibility to model repeated measures (e.g. by estimating mixed models) and the ability to handle data of more complex structures (e.g. by estimating structural equation models). Maximum likelihood (ML) approach estimates the parameters by maximising the observed-data likelihood, where observations with missing values are integrated out.

Another advantage of the ML approach is its ability to estimate models under the MNAR assumption and it has been adopted in various studies with different specification of the models for missing data (e.g. Attanasio and Emmerson, 2003; Washbrook et al., 2014). Under MNAR, the missing data mechanism needs to be modelled simultaneously with the model for the primary outcome as the probability of missingness is related to the missing values. A large number of models could be accounted for by the ML approach, including generalised linear models and models with latent variables. For the Heckman selection model that accounts for informative selection (Heckman, 1977), in addition to the original two-step approach, the ML approach is an alternative estimation method.

As it is impossible to test, both empirically and theoretically, if the missing data mechanism is correctly specified, to offer some protection on the parameter estimates, Molenberghs et al. (2014) advised to perform the sensitivity analysis to check if the estimates are sensitive to the specification of the model for missingness. Treatments of the missing data are discussed in Chapter 7.

Chapter 3

Data and measures

This chapter begins with an overview of the design of the longitudinal study used in this research (Section 3.1). Section 3.2 describes the derivation of repeated measures of each dimension of childhood socioeconomic circumstances (SECs). Section 3.3 presents the missing data patterns in adult waves where partnership information was collected, and the corresponding predictors of probabilities of missingness. Partnership history data and the predictors of the formation and dissolution processes are summarised in Section 3.4. The midlife health measures and the risk factors considered in the model are presented in Section 3.5.

3.1 The 1958 National Child Development Study

The National Child Development Study (NCDS) is a longitudinal birth cohort study that began in 1958. The target population (18,558) contains all individuals born (including stillbirths) in one week of March 1958 in England, Scotland and Wales. Immigrants to Great Britain are included only in the first four sweeps of NCDS (up to age 16) and post-16 immigrants are entirely excluded in further sweeps. Temporary emigrants from the Great Britain are included in the entire study. This cohort has been traced at ages 7, 11, 16, 23, 33, 42, 46, 50 and 55 and followed up until death or permanent emigration from Great Britain. Including also the data collection at birth and a biomedical sweep at age 44-45, the total number of sweeps comes to 11. In all childhood waves (from birth to age 16), the survey interviewed cohort members' parents (to collect, for example, information on the socioeconomic situation of the family), teachers (to collect measures of cognitive development), school doctors and cohort members themselves (since age 7). The adulthood sweeps (from age 23 onwards) cover a wide range of topics, including physical (health-related, mostly self-reported) and mental well-being, life events (e.g. employment and partnership changes) and cognitive developments. In particular, a medical survey was administered in 2002-2003, with over 9,300 cohort members assessed by qualified nurses on a range of health indicators (e.g. level of cortisol and HDL cholesterol). In all questionnaires, some questions were asked of the same respondents repeatedly to update the information at different sweeps. These repeated measures have not been widely used in previous studies, but will be taken full advantage of in this research.

Table 3.1 is adapted from Goodman et al. (2011) and summarises the topics covered in each sweep, along with the amount of missing data. The rich information contained in this longitudinal study makes it useful to study concepts across different domains (e.g. physical, social and health) at different stages in life (child, youth, early-adult and midlife) and their interrelationships over time.

In addition to the published sweeps, event history data on partnership, employment and fertility are available after linking relevant survey responses across sweeps. In this thesis, we use partnership transitions as an example of life events mentioned in Figure 1.1. The partnership history includes information about the start and end date (in months) of co-residential partnerships, their type (marriage or cohabitation), the reason for separation (if end date exists) and the number of partners. These data were collected retrospectively at each adulthood survey using self-completion questionnaires and face-to-face interviews. Based on the work of Di Salvo and Smith (1995) and Kallis (2005) linking the partnership history from age 16 to 42, the continuous history has been extended up to age 50.

Table 3.1 Overview of all NCDS sweeps with information on the type of survey administered (adapted from Goodman et al., 2011)

Sweep	0	1	2	3	4	5	6	Medical	7	8	9
Year	(1958)	(1965)	(1969)	(1974)	(1981)	(1991)	(2000)	(2002-3)	(2004-5)	(2008-9)	(2013)
Age	Birth	7	11	16	23	33	42	44-45	46	50	55
Sample[a]	17,415	15,425	15,337	14,654	12,537	11,469	11,419	9,377	9,534	9,790	9,137
Missing[b]	6.2	16.9	17.4	21.0	32.4	38.2	38.5	49.5	48.6	47.2	50.8
Surveys	Mother Medical	Parents School Tests Medical CM[c]	Parents School Tests Medical CM	Parents School Tests Medical CM Census	CM Census	CM Partner Mother[d] Offspring	CM	CM Bio measure Blood Saliva	CM	Tests CM Biography	CM

[a] Sample = achieved sample size (at least one survey instrument partially completed)

[b] Missing = percentages of wave non-response out of the target population (N=18,558)

[c] CM = cohort member

[d] Mother = CM or partner

3.2 Measures of childhood socioeconomic circumstances

Childhood is defined in the 1958 NCDS as the period before age 16 (1958-1974). Early questionnaires are mostly focused on family situations with similar questions asked in all childhood waves. Based on the previous literature on the potential childhood factors that contribute to adult health issues (reviewed in Section 2.2) and the available information in the dataset, there are four key measures of interest: social class of the family, levels of financial difficulty, levels of material hardship and family structure. As some indicators are not recorded consistently in all childhood sweeps, certain modifications and re-constructions of the variables are necessary, in order to obtain repeated measures of these childhood characteristics over time for future modelling. This is guided by previous research as described below. At each time point, a composite categorical index is derived for each dimension of childhood socioenoomic circumstances, using multiple variables that capture similar characteristics. These composite indices form repeated measures for each childhood SEC dimension across four childhood waves. We aim to summarise the pattern of development of these characteristics over time and to use these summaries as predictors of outcomes in later life.

The social class measure is based on the occupation of the male head of the cohort member's family. The original coding in the four childhood waves followed the UK official guidance (General Registrar's Social Class) in 1951, 1960, 1966, 1970. Referring to Kuh et al. (2003), Chandola et al. (2006a,b), Schoon et al. (2003) and Case and Paxson (2011), repeated measures are derived as follows. In general, six ordinal categories are maintained, denoting unemployed, unskilled, partially skilled, skilled, managerial and professional. For cases with a single mother, the maternal grandfather's social class is used if available; otherwise, occupation is coded as missing. Cohort members with fathers in the armed forces are placed in the missing category as the classification of their social class depends on their military rank, and there is no unified approach to link them with social class. In addition, the unemployment category includes those families where the male head is either unemployed or sick.

The measure of financial difficulty is coded following the approach of Elliott and Lawrence (2014); Sacker et al. (2002); Schoon et al. (2003) using multiple indicators. Instead of constructing one indicator for the entire childhood, we create a binary indicator for each sweep in childhood. At the birth sweep, the indicator takes value 1 if a cohort member's father is in a low social class. At age 7, this indicator takes value 1 if the father is in the last two social class categories or has requested supplementary benefits or claims to be in financial difficulty. At ages 11 and 16, value 1 is assigned if there is at least one positive answer to the questions related to financial hardship (being a recipient of free school meals, being a recipient of benefits and father belonging to the last two social class categories).

Note that benefits include official supplementary benefits, unemployment support and family income support (Schoon et al., 2003).

The third measure, material hardship is an ordinal variable with five categories (from low to high) derived from a summary of yes or no questions related to the following four aspects: the existence of overcrowding, no full sole use of household amenities, not owning the property and recipient of support benefits. Answers to these four questions were collected repeatedly at ages 7, 11 and 16 (no question is available at birth sweep) and combined to form consistent repeated measures of material hardship following Schoon et al. (2003).

The last measure of family structure is constructed as a nominal variable with five categories following Hobcraft (1998). The five situations from poor to good are as follows: in care or in foster care or in other similar situations, cared for by other blood relatives, cared for by a single parent, cared for by step parents and cared for by joint parents. To be more specific, the single parent category includes individuals cared for by natural parents who are divorced or separated. The step parents category includes individuals cared for by one natural parent and one step parent. The joint parent category includes individuals cared for by two natural or adoptive parents. The above classification is partially supported by the definitions and practices used by the British Association of Adoption and Fostering.

Table 3.2 displays the proportions of cohort members in each category of each measure. Comparison with the statistics from published studies cited above supports the validity of the construction of these childhood socioeconomic measures. The distribution of these measures is presented based on the achieved sample size at each wave. From this table, it is obvious that the proportion of missing data increases significantly over time in all measures, the treatment of missing data in childhood waves (discussed in the next chapter) is therefore important for modelling practices.

Table 3.2 Percentage distribution of childhood measures by age.

Measure	Category	0 years	7 years	11 years	16 years
Social Class	Unemployed	0.06	2.17	3.14	3.51
	Unskilled	9.28	5.48	4.33	3.44
	Partially skilled	11.82	15.72	13.93	10.00
	Skilled	55.10	50.02	42.86	37.00
	Managerial	12.68	13.45	15.30	14.17
	Professional	4.16	4.85	4.70	3.85
	Missing	6.91	8.32	15.75	28.03
Financial Difficulty	Yes	9.34	14.72	20.87	17.34
	No	83.76	71.44	67.48	61.03
	Missing	6.91	13.85	11.65	21.63
Material Hardship	Low	–[a]	23.67	29.26	30.04
	Low to medium	–	23.65	26.18	26.28
	Medium	–	22.17	23.84	17.59
	Medium to high	–	7.53	8.14	3.19
	High	–	1.32	1.27	0.20
	Missing	–	21.66	11.31	22.70
Family Structure	Others[b]	–	0.52	0.67	0.91
	Blood relatives	–	0.30	0.35	0.43
	Single parent	3.85	3.62	4.81	7.82
	Step parents	0.19	1.69	3.39	3.53
	Joint parents	95.90	88.92	80.82	66.96
	Missing	0.06	4.95	9.96	20.36
Achieved sample size (N)		17,415	15,425	15,337	14,654

[a] dash = not available

[b] Others = in care/in foster care/in other situations

3.3 Missing data and the predictors of missingness

The merged partnership history dataset (1974-2008) records all co-residential partnership episodes which are over one month in length. Partnerships ending within one month are also recorded for completeness. A detailed description of the dataset is available in Hancock et al.

(2011). After excluding cohort members who died before age 16 ($N = 881$) and individuals who have records with coding error or erroneous partnership information ($N = 3,368$), our sample consists of $14,309$ individuals, of whom $7,313$ have complete partnership histories. The remaining $6,996$ individuals have incomplete partnership histories for ages 16-50, i.e. they are not present at wave 8 (age 50).

We have full partnership histories for all individuals who were present at the age 50 wave, regardless of whether they missed one or more previous adult waves: for individuals who did not respond at a given wave, their history was updated at the next wave. Therefore only individuals who dropout (and do not return) have incomplete histories. Table 3.3 presents a summary of wave non-response and dropout in partnership records. The missing data patterns show that the majority of dropouts occur at wave 5 (age 33). Models for dropout include a set of fully observed early-life demographic covariates, following Hawkes and Plewis (2006). These include gender, logarithm of mother's age at delivery, mother's smoking behaviour and a set of characteristics measured at age 16: the BMI, reading and maths scores (log-transformed) and scores on the Rutter's behaviour scale (dichotomised). Mother's smoking behaviour is a binary variable derived from an ordinal variable that records the severity of smoking behaviour after four months of pregnancy. Behavioural problem at age 16 is a binary variable derived from the 24-item Rutter's behaviour measure where values larger than 9 indicates behaviour problems. Descriptive statistics for these variables are summarised in Table 3.4. More discussions on the choice of these explanatory variables are available in Chapter 7.

Table 3.3 Missing data patterns in waves 5-8 for partnership records (in percentages).

Missing type[a]	W5 (age 33)	W6 (age 42)	W7 (age 46)	W8 (age 50)
$P\&R$	52.8[b]	58.9	42.4	49.1
$P\&\bar{R}$	5.1	5.5	11.4	6.6
$\bar{P}$	42.1	35.6	46.2	44.3
Summary of dropout (incomplete history: $N = 6,996/N = 14,309$)				
Individuals	$N = 4,586$ (65.6)[c]	$N = 397$ (5.7)	$N = 1,307$ (18.7)	$N = 706$ (10.1)

[a] $P\&R$ denotes present and has partnership record,
$P\&\bar{R}$ present and no partnership record,
and $\bar{P}$ not present.

[b] Percentages are computed as a fraction of the $N = 14,309$ cohort members.

[c] Percentages are computed as a fraction of the $N = 6,996$ cohort members.

Table 3.4 Distributions of predictors of the dropout probabilities ($N = 14,309$). For categorical predictors, percentages are reported; for continuous covariates, mean and standard deviations (SD) are reported.

Categorical predictors	Percentages
Time of dropout	
Age 33	65.6
Age 42	5.7
Age 46	18.7
Age 50	10.1
Gender	
Male	50.9
Female	49.1
Mother’s smoking behaviour	
No	69.7
Yes	30.3
Behavioural problem at age 16	
No	89.1
Yes	10.9
Continuous predictors	Mean (SD)
log(mother’s age at birth)	3.3 (0.2)
log(BMI at age 16)	3.0 (0.1)
log(reading score at age 16)	3.2 (0.3)
log(math score at age 16)	2.4 (0.5)

3.4 Partnership histories and risk factors of experiencing each event

After grouping monthly duration data into 6-month intervals, the individuals with complete histories ($N = 7,313$) contributed 340,883 records to the process of first partnership formation and 451,639 records to the process of partnership dissolutions. The two partnership events of interest are the first partnership (either marriage or cohabitation) formation and subsequent partnership dissolution. Table 3.5 gives descriptive statistics for the distribution of the number of years to each event of interest using the life-table methods of Section 5.2.2. Note that the number of years until first partnership formation is calculated from the baseline age of 16.

This preliminary analysis shows that a quarter of the cohort members entered into marriage within 5 years (before age 21) and three quarters of individuals were married by age 27. The hazard of forming the first partnership peaked at age 24 for marriage unions and at age 21 for cohabitation unions. In terms of the timing of separation, roughly a quarter of the married couples in the sample separated after 34 years compared to only 3 years for cohabitors. After 34 years of follow-up, 41% of cohabitors remained in the co-residential relationship. The peak hazard of separation occurred after 4 years and 21 years of marriage and in the first 2 years of cohabitation.

Table 3.5 Life-table estimates of quantiles of the number of years to first partnership and time to separation, by partnership type ($N = 7,313$).

Event	Survival rate		
	75%	50%	25%
Formation[a]			
Marriage	7	8	11
Cohabitation	7	11	19
Dissolution[b]			
Marriage	––[c]	––[c]	––[c]
Cohabitation	3	17	––[d]

[a] Duration = time to first partnership formation from age 16

[b] Duration = time to partnership dissolution from the start of each partnership

[c] 75.2% were still married after 34 years

[d] 41.3% were still in co-residential cohabitation after 33 years

In terms of the predictors in the models for partnership transitions, in addition to childhood SEC variables, other explanatory variables, including the type of partnership, age at the start of partnership, education level, the number of pre-school children and the number of previous partners are considered. The choice of these variables is guided by results from previous studies (e.g. Ermisch and Francesconi, 2000; Ploubidis et al., 2015; Steele et al., 2005). Of these variables, the first two are directly available in the partnership history. Of the remaining variables, the level of education is defined as the number of years in post-16 full-time education. It is an ordinal time-varying variable that is derived from the information collected retrospectively at ages 42 and 33. The number of pre-school children is derived from the information about fertility collected at sweep 6 (age 42) where the total number of biological and pre-school (≤ 5 years old) children of the cohort member can be derived for each time interval. Should there be missing data in time-varying predictors (for individuals with complete partnership histories), we impute the data by using information collected in earlier and later waves. For example, to impute missing values in "the number of years in post-16 full-time education", we assume an individual is in continuous full-time education until the highest level is obtained. The number of previous partners is a binary variable where value 1 denotes "at least one partner" in the relationship history. Descriptive statistics for these explanatory variables included in the final models for partnership transitions are summarised in Table 3.6 and Table 3.7. The distributions of the explanatory variables are comparable with those reported by Steele et al. (2005) and Berrington and Diamond (2000).

Table 3.6 Distributions of time-varying predictors of first partnership formation based on the discrete-time file with 6-month intervals ($N = 7,313$).

Predictors	Percentages
Age	
[16,19)	24.7
[19,22)	21.5
[22,25)	16.5
[25,28)	13.0
[28,31)	11.1
[31,34)	8.7
[34,37)	1.3
[37,40)	0.9
[40,43)	0.8
43+	1.5
Number of post-compulsory years of education	
0	56.9
1	12.7
2	12.2
3-5	11.7
6+	6.5

Table 3.7 Distributions of predictors of recurrent partnership dissolutions based on the discrete-time file with 6-month intervals ($N = 7,313$).

Predictors	Percentages
Duration (years)[a]	
[1,3)	18.5
[3,6)	14.2
[6,9)	12.5
[9,12)	11.3
[12,15)	10.4
[15,18)	9.5
[18,21)	8.3
[21,24)	6.8
[24,27)	5.0
27+	3.5
Partnership type	
Cohabitation	12.6
Marriage	87.4
Number of post-compulsory years of education[a]	
0	61.1
1	9.9
2	10.5
3-5	9.6
6+	8.9
Number of pre-school children[a]	
0	52.6
1	16.3
2	23.1
3+	8.1
Number of previous partners	
0	58.7
1+	41.3
Age at the start of relationship (years)[a]	
< 20	15.7
$[20,25)$	47.7
$[25-30)$	22.0
$[30,35)$	9.2
35+	5.4

[a] Time-varying variables

3.5 Measures of midlife health and the risk factors

The NCDS dataset contains several health outcomes recorded at age 50, including general health, reasons for illness in the past 12 months and health-related behaviours that have been reviewed in Section 2.3. However, to minimise the amount of missing data and for the interest of having a comprehensive health measure, we derive the midlife health outcome from the variable "HLTHGEN" collected at wave 8 (age 50). It is a self-reported measure of health recorded on a 5-point scale with categories excellent, very good, good, fair and poor. To reduce measurement error due to subjective interpretations of adjacent categories (e.g. the boundary between good and very good is unclear), we group the first three categories into "good health" (coded 0) and the last two into "poor health" (coded 1), leading to a binary measure of midlife health. Results of the grouping are summarised in Table 3.8.

In addition to the childhood SECs, the choice of predictors for health (summarised in Table 3.9) is informed by previous studies (e.g. Ploubidis et al., 2015; Potter and Ulijaszek, 2013). As a control for health status in childhood, in addition to gender, we derive a binary indicator for overweight at age 16 using WHO cut-offs where BMI $(kg/m^2) > 25.0$ denotes "overweight" (coded 1) (WHO, 2000). The limited choice of control variables is explained below. First, we intended to control for more variables related to early health before age 16 but they have a large amount of missing values that requires imputation. As they are possibly related to childhood SECs (latent variables), the imputation procedure is not straightforward. Indicators for childhood SECs may be used as auxiliary variables to impute early health variables but these indicators themselves contain missing values. Therefore, in addition to the validity of the imputation model, the practicality is also of concern (discussed by Seaman and White (2013)). Second, variables that are collected between ages 16 and 50 are not considered as control variables in the health model because they could be potential mediators or intermediate confounders of the relationship between childhood SECs, partnership transitions and midlife health. Examples of such variables are health state, education attainment and fertility status at age 23 and 33. For a proper causal analysis, more discussions about the structural relationship between quantities of interest are required by seeking expert knowledge in applied fields but this is beyond the scope of this thesis.

To relate partnership transitions to midlife health, three summaries of partnership experiences are derived from the 34-year partnership history data. The total number of partners before age 50 is coded as 0, 1, 2 and 3+. The percentage of time spent single for ages 16-50 is computed as the fraction of time spent between partnerships over the 34 years of follow-up. Age at the first partnership, an outcome of the history of first partnership formation, is a continuous variable that is centered around its median and then log-transformed. For individuals who have never had a partner over the 34 years, the value is set at zero. It is essentially an interaction of the variable "have partnered or not" and "the age at first partnership". It also allows for a clear interpretation of the effect of the percentage of time spent in a single status on midlife health, by allowing for comparison among those who started the first partnership at the same age.

Table 3.8 Percentages of individuals in each category of the original 5-point self-reported measure of midlife health, and of the derived binary health measure $(N = 7,313)$.

Original categories	Percentages
Excellent	19.4
Very good	33.6
Good	29.3
Fair	12.3
Poor	5.4
Grouped categories	
Good	82.3
Poor	17.7

Table 3.9 Distributions of categorical and continuous predictors of the midlife health $(N = 7,313)$. For categorical predictors, percentage distributions are reported; for continuous predictors, mean and the standard deviation (SD) are reported. Percentage time spent single is calculated as the proportion of total number of months spent in between relationships over the total number of months observed before dropout.

Categorical predictors	Percentages
Overweight at age 16	
No	57.0
Yes	43.0
Gender	
Male	50.9
Female	49.1
Total number of partners	
0	2.0
1	37.6
2	8.9
3+	51.5
Continuous predictors	Mean (SD)
Percentage time spent single	0.3 (0.2)
log(age at first partnership)	0.3 (1.3)

Chapter 4

Latent class analysis of childhood socioeconomic circumstances and a distal outcome

4.1 Introduction

It has been common to use measures of childhood socioenomic circumstances (SECs) at one or multiple time points during childhood as predictors for partnership transitions and later health. For instance, parental social class defined by father or male head's social class at birth, ages 7, 11 and 16 have been considered as predictors of partnership transitions (Aassve et al., 2006b) and physical well-being in midlife (Kelly-Irving et al., 2013). The experience of parental divorce by age 16 (a binary indicator) has been found to be associated with physical disease, such as cancer (Kelly-Irving et al., 2013), and partnership decisions (Aassve et al., 2006b). Financial hardship in childhood is often constructed as a composite indicator using multiple items for living conditions collected at birth and ages 7 and 11. Previous studies have noted their importance in relation to partnership transitions (Berrington and Diamond, 2000) and physical health (e.g. Bartley et al., 2012; Kelly-Irving et al., 2013) in midlife.

However, these studies have a few limitations. First, including only the covariates covering one or two aspects of childhood experiences that are measured at one or two time points in the survey provides a limited summary of the family's socioeconomic situation in childhood, especially when repeated measures are available in the survey. Second, treating repeated measures of childhood situations at different time points as separate predictors in the regression model can lead to inflated standard errors as these predictors may be highly correlated. They may also be endogenous: as the measurements are taken on the same subject, they may be correlated with the individual-level unobservables that are also predictive of the outcome. In addition, it is inevitable that missing values exist in one or more predictors. Simple treatment of the missing values such as listwise deletion can also lead to biased estimates of coefficients or spurious conclusions about the significance of covariate effects. Imputation methods, such as the multiple imputation (MI) can be used instead but it does not help to reduce the dimension of childhood data.

In this study, we are interested in summarising the pattern of changes in four dimensions of SECs in childhood, i.e. parental social class, financial difficulty, material hardship and family structure, and relate these summaries to distal outcomes of interest. To derive these summaries of the change in values of repeated measures with minimal loss of information, latent class models are employed (see Hagenaars and McCutcheon (2002)). The advantages of this approach are threefold. First, we can take full advantage of the repeated measures available in the 1958 NCDS childhood waves and take into account the patterns of change in different aspects of socioeconomic situation over childhood. Second, we can provide further insights into the associations between childhood circumstances over time (represented by longitudinal typologies) and the distal outcomes. Third, latent class analysis uses all available information in the data by using the method of full information maximum likelihood estimation (FIML). Individuals with missing values in repeated measures will not be excluded as long as the information is not missing in all repeated measures. FIML also has several advantages over the other commonly used MI approach as FIML produces deterministic results while results from the MI can vary with respect to the number of replications in the imputation procedure, the choice of the random number seed and the specification of the imputation model (particularly when multiple variables need to be imputed). Moreover, as latent class models are embedded in the structural equation framework, further extensions, such as relating latent variables to observed outcomes and other latent variables can be easily allowed for.

The development of four categorical latent summaries of four aspects of childhood situations is partly inspired by Ploubidis et al. (2015) who used a continuous latent summary of the general childhood situation with living conditions, parental social class measured at ages 7, 11 and 16 as observed indicators. Although their focus was the effects of partnership trajectories on midlife health, this latent summary of childhood situations was included in the structural equation model to account for confounding in midlife health. However, we note that their latent construct for childhood circumstances is a very general summary of indicators taken at different times that measure different aspects in childhood. Considering the difficulty in interpreting this latent variable and our particular interest in childhood measures, four categorical latent variables are considered in this research, reflecting the patterns of change in four different aspects of the childhood socioeconomic situation over time.

The remainder of the chapter is organised as follows. After an overview of latent class analysis methodology and its application to health studies in Section 4.2, step-wise methods to estimate the effect of a single categorical latent variable on a distal outcome are reviewed in Section 4.3. In Section 4.4, an extension to the 3-step method is proposed for models with multiple, and possibly associated categorical latent variables. This is followed by an extensive simulation study discussed in Section 4.5 to assess the relative performance of three methods using different simulated settings for cases with multiple categorical latent variables and a distal outcome. Finally, Section 4.6 describes an application of latent class analysis to each of the four sets of repeated measures of childhood circumstances, as well as an analysis of the effects of childhood SECs on midlife health (at age 50) using these three methods.

4.2 Overview of latent class analysis

4.2.1 Methodology

Latent class analysis (LCA) is a technique used mainly for data classification and reduction. In the longitudinal setting with repeated measures, the latent classes can be treated as a summary of patterns of change in values of repeated measures over time. The main idea is that there exists a set of indicators (observed variables) which measure the underlying latent variable with measurement error. The associations between observed variables can therefore be decomposed to that captured by latent variables as well as by residual terms.

Denote by C_q the latent class variable for childhood SEC dimension q $(q = 1,...,Q)$ with classes indexed by k_q $(k_q = 1,...,K_q)$. Each latent class variable is measured by a $P_q \times 1$ vector of observed categorical indicators, denoted by $\mathbf{Y}_q^{(C)}$ where p $(p = 1,...,P_q)$ indexes the indicators, each with W categories (indexed by $w = 1,...,W$). Note that we focus on repeated measures of the latent class, hence for each latent construct, its indicators have the same number of categories. $\mathbf{Y}_q^{(C)}$ is often referred to as the response pattern, where superscript (C) denotes "childhood". A key assumption of LCA is that conditional on the class membership, indicators are independent. Equivalently, the within-class covariance of the indicators is zero. We first consider a model with one latent class variable, dropping the subscript q for simplicity. The conditional independence assumption implies

$$P(\mathbf{Y}^{(C)}|C) = \prod_{p=1}^{P} P(Y_p^{(C)}|C).$$

The marginal probability of the observed patterns can be written as

$$P(\mathbf{Y}^{(C)}) = \sum_{k=1}^{K} P(C=k) \prod_{p=1}^{P} P(Y_p^{(C)}|C=k), \tag{4.1}$$

which shows that the joint distribution of response patterns is a mixture of all class-specific distributions, each of which is weighted by the class proportion. For nominal indicators, the model is a multinomial logit regression written as:

$$\log\left(\frac{P(Y_p^{(C)}=w|C=k)}{P(Y_p^{(C)}=W|C=k)}\right) = \alpha_{wpk}, \tag{4.2}$$

where $w = 1,...,W-1$ as we set $\alpha_{Wpk} = 0$ for the reference category W. For ordinal indicators, a cumulative logit model can be specified as

$$\log\left(\frac{P(Y_p^{(C)}\leq w|C=k)}{P(Y_p^{(C)}> w|C=k)}\right) = \alpha'_{wpk}. \tag{4.3}$$

Finally, a multinomial logit model can be defined for the distribution of the categorical latent variable C as

$$\log\left(\frac{P(C=k)}{P(C=K)}\right) = \beta_k, \tag{4.4}$$

where we set $\beta_K = 0$ for the reference class K.

As the categorical latent variable C summarises the response patterns of individuals based on the observed indicators, a standard approach to cluster people into distinct subgroups is to use the modal class membership defined as the class with the highest posterior probability,

$$P\left(C=k|\mathbf{Y}^{(C)}\right) = \frac{P(\mathbf{Y}^{(C)}|C=k)P(C=k)}{\sum_{k=1}^{K} P(\mathbf{Y}^{(C)}|C=k)P(C=k)}. \tag{4.5}$$

The model can also be easily extended to include covariates that predict $\mathbf{Y}^{(C)}$ or C in the right hand-side of (4.2) to (4.4). Parameter estimation is often realised via the iterative Expectation-Maximization (EM) algorithm (Dempster et al., 1977), the Newton-Raphson method (especially for the last few iterations) or Bayesian methods such as Markov Chain Monte Carlo (MCMC) (Hagenaars and McCutcheon, 2002). As the EM algorithm is widely available in most software packages, we now briefly describe the technical steps of parameter estimation.

For an individual i ($i = 1, ..., N$), the marginal probability of their response pattern is

$$P(\mathbf{Y}_i^{(C)}) = \sum_{k=1}^{K} P(C_i = k) P(\mathbf{Y}_i^{(C)} | C_i = k), \tag{4.6}$$

which gives the log-likelihood of the observed data

$$l(\mathbf{Y}^{(C)}) = \sum_{i=1}^{N} \log P(\mathbf{Y}_i^{(C)}). \tag{4.7}$$

By substituting (4.6) into (4.7), the observed-data log-likelihood can be re-expressed as

$$l(\mathbf{Y}^{(C)}) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} P(\mathbf{Y}_i^{(C)} | C_i = k) P(C_i = k). \tag{4.8}$$

Note that for maximum likelihood estimation, (4.8) needs to be maximised subject to a "penalty term" that the sum of prior probabilities $P(C_i = k)$ is equal to 1. More technical details of the estimation procedure is set out in Section 6.5 of Bartholomew et al. (2011) for binary measurements and in Sections 6.10–6.12 for ordinal and nominal measurements with more than two categories. Key steps that are relevant to our proposed approach are summarised below.

Denote by $\boldsymbol{\alpha}_k$ the class-specific parameters to be estimated in the measurement model (4.2) or (4.3). To obtain the maximum likelihood estimates, the derivative of (4.8) with respect to $\boldsymbol{\alpha}_k$ is

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\alpha}_k} &= \sum_{i=1}^{N} \frac{P(C_i = k)}{\sum_{k=1}^{K} P(\mathbf{Y}_i^{(C)} | C_i = k) P(C_i = k)} \frac{\partial P(\mathbf{Y}_i^{(C)} | C_i = k)}{\partial \boldsymbol{\alpha}_k} \\
&= \sum_{i=1}^{N} \frac{P(C_i = k) P(\mathbf{Y}_i^{(C)} | C_i = k)}{\sum_{k=1}^{K} P(\mathbf{Y}_i^{(C)} | C_i = k) P(C_i = k)} \cdot \frac{1}{P(\mathbf{Y}_i^{(C)} | C_i = k)} \cdot \frac{\partial P(\mathbf{Y}_i^{(C)} | C_i = k)}{\partial \boldsymbol{\alpha}_k} \\
&= \sum_{i=1}^{N} P(C_i = k | \mathbf{Y}_i^{(C)}) \cdot \frac{\partial \log P(\mathbf{Y}_i^{(C)} | C_i = k)}{\partial \boldsymbol{\alpha}_k};
\end{aligned} \tag{4.9}$$

and the derivative with respect to $\boldsymbol{\beta}_k$ is

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\beta}_k} &= \sum_{i=1}^{N} \frac{P(\mathbf{Y}_i^{(C)} | C_i = k)}{\sum_{k=1}^{K} P(\mathbf{Y}_i^{(C)} | C_i = k) P(C_i = k)} \frac{\partial P(C_i = k)}{\partial \boldsymbol{\beta}_k} \\
&= \sum_{i=1}^{N} P(C_i = k | \mathbf{Y}_i^{(C)}) \cdot \frac{\partial \log P(C_i = k)}{\partial \boldsymbol{\beta}_k}.
\end{aligned} \tag{4.10}$$

Therefore maximizing the log-likelihood for a mixture model (4.8) is achieved by maximizing the weighted log-likelihood of component distributions, where the weights also depend on model parameters. As the class membership is unobserved, it is treated as a missing value in the EM algorithm. An initial guess of parameters is first used to compute $P(C_i = k|\mathbf{Y}_i^{(C)})$ using (4.5). Substituting these quantities into (4.9) and (4.10) allows us to maximize the weighted log-likelihood and obtain updated parameters. The iterative process continues until convergence is reached. Examples of such convergence criterion include the difference in the parameter estimates and that of log-likelihood function in consecutive iterations to be less than a threshold. Latent class models can now be estimated in a number of software packages, including Stata 15.1 (StataCorp, 2017), Mplus (all versions) (Muthén and Muthén, 2017) and LatentGOLD (all versions) (Vermunt and Magidson, 2015). One drawback of this iterative process is that the likelihood of the mixture distribution (4.8) can have multiple local maxima, which may be remedied by running the algorithms with multiple random starts (Dempster et al., 1977).

To estimate the standard errors of the parameter estimates, we take the square roots of diagonal elements of the asymptotic covariance matrix. Denote by $\boldsymbol{\theta}$ the set of parameters (αs, βs) to be estimated in the model. We take the second partial derivative of the log-likelihood function (4.8) with respect to each parameter and obtain the Hessian matrix

$$\mathscr{H}(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{j_1} \partial \boldsymbol{\theta}_{j_2}},$$

where $j_1, j_2 \in \{1, ..., K\}$. For maximum likelihood estimators, we can obtain the asymptotic covariance matrix of parameters by inverting the observed Fisher information matrix $\mathscr{I}(\boldsymbol{\theta})$, which contains the negatives of derivatives in $\mathscr{H}(\boldsymbol{\theta})$. Denote by $\hat{\boldsymbol{\theta}}$ the set of maximum likelihood estimate of parameters, the asymptotic covariance matrix is

$$\text{Var}(\hat{\boldsymbol{\theta}}) = \mathscr{I}^{-1}(\hat{\boldsymbol{\theta}}).$$

As $P(Y_p^{(C)}|C = k)$ and $P(C = k)$ are functions of the parameters, asymptotic standard errors could be derived using the delta method with first-order Taylor expansion (see more details in Garre and Vermunt (2006)). Specifically,

$$\text{SE}\left(P(Y_p^{(C)} \leq w|C = k)\right) = \sqrt{\frac{\partial P(Y_p^{(C)} \leq w|C = k)}{\partial \alpha_{pwk}} \text{diag}(\mathscr{I}^{-1}) \frac{\partial P(Y_p^{(C)} \leq w|C = k)'}{\partial \alpha_{pwk}}},$$

$$\text{SE}\big(P(C = k)\big) = \sqrt{\frac{\partial P(C = k)}{\partial \beta_k} \text{diag}(\mathscr{I}^{-1}) \frac{\partial P(C = k)'}{\partial \beta_k}}.$$

In terms of the use of statistics to choose the number of latent classes to be retained in the model, Nylund et al. (2007) found via simulation studies that the Bayesian information criterion (BIC), among all information criteria (e.g. Akaike information criterion and sample size adjusted BIC), and the bootstrap likelihood ratio test perform the best across different settings. Note that "best" refers to the ability to recover the true and the most parsimonious model. Other likelihood ratio tests that compare nested models, such as the standard log-likelihood ratio test (LRT) and the empirical likelihood ratio test (Lo et al., 2001) are also useful guidelines. In addition, entropy is a commonly reported statistic that measures the degree of class separation. The relative size of each class and whether the substantive meaning of classes is interpretable are also considered. For analyses of longitudinal data, patterns of state change (for categorical or nominal indicators) can be plotted over time to assess the number of clearly separated classes. A joint consideration of all these factors is necessary as the goodness-of-fit statistics and entropy alone do not offer a clear cut-off point for the number of classes to be retained (detailed discussions are available in Asparouhov and Muthén (2012) and Jung and Wickrama (2008)).

4.2.2 Application to health studies

The standard LCA for longitudinal data can be extended to latent class growth analysis (LCGA), which assumes that each class has a distinct set of growth parameters for the group-mean growth trajectory, without allowing for within-group variance. To allow for unobserved heterogeneity within groups, growth mixture models can be employed that introduce random effects into the specification of the group-mean trajectory (Jung and Wickrama, 2008). This broad set of growth mixture models has been widely applied in studies of health. For example, Chandola et al. (2006a) examined the association of intelligence at age 11 and BMI growth during adulthood by specifying a structural equation model (SEM) that includes a latent variable for parental socioenomic circumstances and a growth curve model for BMI progression between ages 16 and 42. Before estimating the SEM to test the mediating and confounding pathways, a logistic regression was used to model obesity in adulthood, and potential confounders and mediators were adjusted step by step. This model can be extended, for example, by considering other potential mediators in early adulthood, such as smoking and drinking behaviour, as well as life events (e.g. partnership transitions) because one may expect them to influence changes in BMI.

In other applications, Lane et al. (2013) considered a growth mixture model to uncover patterns of the growth trajectory of BMI between ages 1 and 11 (using seven repeated measures). They also studied how family socioeconomic circumstances, measured by household total income before child birth, could impact the classification of individuals based on the growth pattern of BMI. This relationship was hypothesized to be mediated by depression and smoking during pregnancy, as well as parenting style after birth. One key limitation of this work is that few factors were considered. Other variables such as diet and education during early childhood also have important influences on the change in BMI during the analysis period. In addition, the family circumstance is a rather broad and vague concept, which may not be adequately captured by a single measure of family income. One possible remedy is to introduce a latent variable as a summary of situations in the family. For example, Ploubidis et al. (2015) explored the relationship between patterns of change in partnership status between ages 23 and 42 (represented by a categorical latent variable) and health biomarkers in midlife (collected from the medical survey between ages 44 and 46). Parental and early childhood SECs, health status, as well as the health and education level at age 23, were treated as confounders. The model of Ploubidis et al. (2015) suggests a cumulative effect of patterns of change in partnership status during the entire period from age 23 to 42 on health at age 45 but the classification of the pattern is based on the most likely class membership that is subject to classification error.

4.3 Review of the 1-step and step-wise methods: a single categorical latent variable

In general, methods to estimate the effect of a single categorical latent variable on a distal outcome (denoted by H for health) can be categorised into two major groups: the 1-step approach and various step-wise approaches. The 1-step approach simultaneously estimates the measurement model and the regression model of H on C, essentially treating H as an additional indicator for C. Parameter estimates, including item response probabilities and regression coefficients, are obtained by jointly maximizing the log-likelihood of response patterns and the distal outcome. There are three main advantages of the 1-step approach. First, it is more efficient compared to step-wise approaches that may introduce additional uncertainty between steps; second it allows for more flexible model structures, such as models with direct effects of covariates on indicators and the distal outcome; and, third, it is straightforward to account for residual association between H and the vector $\mathbf{Y}^{(C)}$, beyond that captured by class membership (Bakk et al., 2013). However, the 1-step approach has received criticism over the past ten years regarding the practicality and validity of the simultaneous estimation and the requirement for additional distributional assumptions about H. Vermunt (2010) noted the burden of having to re-estimate the entire model should one decide to add or delete covariates in the measurement model. He also pointed out a more serious issue that the inclusion of a distal outcome into the measurement model creates an unintended circular relationship in that the latent class C that is supposed to explain H is also determined partly by H (also discussed by Bakk and Vermunt (2016)). If there are multiple distal outcomes, the shift in the latent class proportions can be severe, especially when a large number of classes are retained or when class separation is poor. Moreover, by treating H as an indicator for C, the 1-step approach requires additional assumptions. For instance, a continuous H should be normally distributed within classes (Asparouhov and Muthén, 2014a; Bakk et al., 2013; Bakk and Vermunt, 2016).

Various step-wise approaches have been proposed in recent years in order to preserve the latent classes from the measurement model for $\mathbf{Y}^{(C)}$. The key difference from the 1-step approach is that in step-wise approaches the measurement model is estimated separately, with parameters from this step carried forward in later analyses that involve external variables. Commonly used approaches are the modal class approach, the modified Bolck, Croon and Hagenaars (BCH) approach (proposed by Vermunt (2010), developed from Bolck et al. (2004)), the 3-step maximum likelihood (ML) approach (with modal or proportional assignments) and the Lanza-Tan-Bray (LTB) approach (Lanza et al., 2013). In the following, we summarise the key concepts and restrictions of each method in situations where the latent variable predicts the distal outcome. The methods are described for a single latent variable, as in previous research. We consider the extension to multiple latent variables in the next section.

4.3.1 Modal class method

After performing the latent class analysis, the modal class assignment is saved for each individual based on their posterior probability of being in each class. The class membership is often used in further studies as a known nominal covariate. However, treating modal classes as observed variables in analyses ignores the uncertainty of classification (i.e. misclassification). For example, in a 2-class model, an individual with posterior probabilities of 0.49 and 0.51 of being in class 1 and 2 will be assigned to class 2. Misclassification may lead to biased estimates of class effects on other quantities of interest as such individuals are forced into a class. In addition, by treating the estimated class membership as observed, standard errors in the model are underestimated, which can lead to incorrect statistical inference about the significance of class effects.

A modification of the modal class approach is the pseudo class approach (Clark and Muthén, 2009). Instead of assigning individuals to classes with certainty, individuals are now assigned to classes randomly sampled from the multinomial distribution based on the posterior probability of being in each class. It is similar to the multiple imputation method used to handle missing data (Little and Rubin, 1987). Treating the classes as missing, multiple random draws (20 is recommended by Wang et al. (2005)) of class membership are made, which allows for the uncertainty in class membership. The model of interest is then fitted to each of the 20 datasets with imputed class variables. Mean effects and the associated standard errors can be computed following the proposition of Little and Rubin (1987) and Schafer (1997) as the procedure is analogous to a standard multiple imputation approach. Note that mean effects are computed as the averaged point estimates of the coefficients across 20 models. The variance of estimated effects is a combination of the variance within random draws and that across the random draws. The simulation study by Clark and Muthén (2009) shows that for the one categorical latent variable case, when there is clear separation of classes (with high values of entropy), the pseudo class method performs well. However, a recent paper by Asparouhov and Muthén (2014a) found that the pseudo class approach performs poorly when the value of entropy of the latent class model is as low as 0.5. When it comes to modelling the relationship between H and the modal class membership (denoted by $M\,(M = 1,\ldots,K)$), one needs to take into account that the joint density of (H, M) is different from (H, C) and therefore a correction is needed.

4.3.2 The modified BCH approach

To study the effects of latent predictors on a distal outcome, i.e. the conditional distribution of H given C, one needs to first establish the relationship between (H, M) and (H, C). Translating equation (9) of Bolck et al. (2004) into our context, we have

$$P(M = m,H) = \sum_{k=1}^{K} P(C = k,H)P(M = m|C = k), \tag{4.11}$$

which can be expressed in terms of the conditional relationship of H given C as

$$P(M = m,H) = \sum_{k=1}^{K} P(H|C = k)P(C = k)P(M = m|C = k), \tag{4.12}$$

which is equivalent to a latent class model with two indicators (M, H). Equations (4.11) and (4.12) also imply that to obtain $P(H|C = k)$, one needs to adjust for the misclassification probability $P(M = m|C = k)$. Also worth noting is that to obtain (4.11) and (4.12), it is implied that M depends only on $\mathbf{Y}^{(C)}$ (as M is only determined in the measurement model) and we assume $\mathbf{Y}^{(C)}$ is conditionally independent of H given C.

The modified BCH approach (Vermunt, 2010) originated from the BCH approach proposed by Bolck et al. (2004). It was first developed for latent class models with covariates before being extended by Bakk et al. (2013) to the situation where H is a distal outcome. The weighted pseudo log-likelihood function is:

$$l = \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} \log P(H_i|C=k)P(C=k), \tag{4.13}$$

where i indexes subjects, $w_{ik} = \sum_{r=1}^{K} p_{im} d_{mk}$ and $p_{im} = P(M=m|\mathbf{Y}^{(C)} = \mathbf{y}_i^{(C)})$ with m and k $(m,k \in \{1,...,K\})$ indexing the modal and latent classes, respectively. Note that d_{mk} represents an element of the inverted $K \times K$ matrix D of the misclassification probabilities $P(M=m|C=k)$. A detailed description of this approach is available in Vermunt (2010) and Bakk et al. (2013).

Using the modified BCH approach, the latent class solution derived from step 1 remains unchanged. It is also robust to violations of within-class normality assumption for a continuous H and the constant variance across classes (Asparouhov and Muthén, 2014b; Bakk and Vermunt, 2016). However, estimation problems may arise for a categorical H when negative cell frequencies are obtained in the weighted cross-classification of categorical H and C (Asparouhov and Muthén, 2014b; Bakk et al., 2013). This problem may be exacerbated when there are multiple distal outcomes. Bakk and Vermunt (2016) also show that standard errors are underestimated when the sample size is small and class separation is poor.

4.3.3 The Lanza-Tan-Bray approach

The LTB approach was first proposed by Lanza et al. (2013) as a method that preserves the latent class solution estimated from the measurement model. To estimate the class-specific means for a continuous outcome H, they first estimate $P(C|H)$, treating H as a predictor of C. Bayes' theorem is then applied to obtain $P(H|C)$ by using the kernel density to approximate $P(H)$. In a later modification by Asparouhov and Muthén (2014a), the sample distribution of H is used which produces similar results to using the kernel approximation. However, it suffers from several limitations. First, the measurement model cannot include covariates as otherwise both the covariates and the outcomes will be predicted by C after Bayes' transformation (Asparouhov and Muthén, 2014a). Second, when the within-class distribution of H has outliers, the estimated mean of H is severely biased (Bakk et al., 2014). They also find that for continuous H the LTB approach produces heavy bias when heterogeneity across classes is not accounted for. Despite these limitations, the performance of the LTB for categorical H is satisfactory. However, it may not be generalised to multiple distal outcomes, unless they are strictly conditionally independent given C. Further details about the LTB approach, including standard error estimation can be found in Asparouhov and Muthén (2014a) and Bakk and Vermunt (2016).

4.3.4 The 3-step method

The 3-step ML approach is based on the idea of the modified BCH approach, which was first proposed by Vermunt (2010) to account for misclassification while estimating the effect of covariates on the class membership. It was further developed by Bakk et al. (2013) and Asparouhov and Muthén (2014a) to handle situations with a distal outcome H. The procedure of parameter estimation includes three steps:

Step 1: Perform a LCA without the distal outcome and its predictor (denoted by $X^{(H)}$). Save the posterior probability of being in each class (4.5) and the most likely class membership (i.e. modal class) (M) for each individual.

Step 2: Calculate the misclassification probabilities (4.14).

Step 3: Include the modal class in a model for the distal outcome. Treat the modal class as an imperfect measurement of the categorical latent variable, with measurement error captured by the misclassification probabilities calculated in step 2.

After performing a standard latent class analysis, we obtain for each response pattern the modal class M and posterior probabilities of being in each class (calculated using (4.5) in Section 4.2). The second step involves deriving the association of C and its imperfect measurement M (Asparouhov and Muthén, 2014a; Bakk et al., 2013). Using Bayes' theorem, we have

$$\begin{aligned} P(M=m|C=k) &= \sum_{\mathbf{Y}^{(C)}} P(M=m,\mathbf{Y}^{(C)}|C=k) \\ &= \sum_{\mathbf{Y}^{(C)}} \frac{P(C=k|M=m,\mathbf{Y}^{(C)})\cdot P(M=m,\mathbf{Y}^{(C)})}{P(C=k)}, \end{aligned} \tag{4.14}$$

In the numerator of (4.14), we have

$$\begin{aligned} P(M=m,\mathbf{Y}^{(C)}) &= P(M=m|\mathbf{Y}^{(C)})\cdot P(\mathbf{Y}^{(C)}), \\ P(C=k|M=m,\mathbf{Y}^{(C)}) &= P(C=k|\mathbf{Y}^{(C)}), \end{aligned} \tag{4.15}$$

assuming conditional independence of the modal and latent class given $\mathbf{Y}^{(C)}$. In fact, this assumption comes naturally from the measurement model as C is a function of $\mathbf{Y}^{(C)}$. Conditional on $\mathbf{Y}^{(C)}$, M does not contain any extra information. Similarly, we have

$$P(M=m|C=k,\mathbf{Y}^{(C)}) = P(M=m|\mathbf{Y}^{(C)}). \tag{4.16}$$

Using an alternative factorisation of $P(M=m|C=k)$ and plugging in (4.16), similar to equation (5) in Bakk et al. (2013) we have

$$\begin{aligned} P(M=m|C=k) &= \sum_{\mathbf{Y}^{(C)}} P(M=m|\mathbf{Y}^{(C)},C=k)\cdot P(\mathbf{Y}^{(C)}|C=k) \\ &= \sum_{\mathbf{Y}^{(C)}} P(M=m|\mathbf{Y}^{(C)})\cdot P(\mathbf{Y}^{(C)}|C=k), \end{aligned} \tag{4.17}$$

where $P(\mathbf{Y}^{(C)}|C=k)$ can be derived using (4.2) or (4.3) from the standard latent class model in the first step.

In the third step, a regression analysis is performed where $P(M=m|C=k)$ is fixed at the values calculated from (4.17) in step 2. This keeps the latent class solutions derived from step 1 intact when the distal outcome H is included in the model.

Taking the binary H as an example, we then specify a logit model as

$$\text{logit}\left(P(H=1|C=k,X^{(H)})\right) = \tau_k + \lambda X^{(H)}, \tag{4.18}$$

where τ_k denotes the class-specific intercept with τ_K set to zero for the reference category. λ is the effect of covariate $X^{(H)}$ on the distal outcome that is assumed to be constant across classes. Class-specific effects can be allowed by using λ_k.

The corresponding log-likelihood of the observed data can be written as

$$l(H,M|X^{(H)};\boldsymbol{\tau},\lambda) = \sum_{i=1}^{N}\log\sum_{k=1}^{K} P(H_i,M_i|X_i^{(H)},C_i=k)P(C_i=k), \tag{4.19}$$

where $\boldsymbol{\tau}=\{\tau_1,...,\tau_K\}$. Assuming conditional independence of the distal outcome H and modal class M given the latent class C gives

$$\begin{aligned} P(H_i=1,M_i|X_i^{(H)},C_i=k) &= P(H_i=1|M_i,X_i^{(H)},C_i=k)\cdot P(M_i|X_i^{(H)},C_i=k) \\ &= P(H_i=1|X_i^{(H)},C_i=k)\cdot P(M_i|C_i=k), \end{aligned} \tag{4.20}$$

where $P(M_i|X_i^{(H)},C_i=k)$ can be reduced to $P(M_i|C_i=k)$ as $X^{(H)}$ is not included in step 1. The conditional independence assumption of $P(H|M,C)=P(H|C)$ follows from the fact that

$P(M|C, \mathbf{Y}^{(C)}) = P(M|\mathbf{Y}^{(C)})$ (discussed above). As H and M are both functions of C, given the class membership, the association between them is fully accounted for.

The key advantage of the 3-step ML approach is that it preserves the class solution in the measurement model and that the efficiency of the 3-step ML approach is close to that of the 1-step approach. Similar to the modified BCH method, the misclassification probabilities obtained from step 1 are carried forward in subsequent analyses. One limitation is that although step 1 estimates the measurement model separately from H, the inclusion of H in step 3 may lead to a change in the latent class proportions when, for a continuous H the within-class distribution of H is bimodal (Asparouhov and Muthén, 2014a; Bakk and Vermunt, 2016). The 3-step ML approach also understates the standard errors by treating the misclassification probabilities in step 2 as observed, rather than estimated from step 1. A standard error correction method was proposed by Bakk et al. (2014) that takes into account this additional source of variation.

4.3.5 Comparison of methods: Evidence from simulation studies

A number of simulation studies have been conducted to compare the performance of different methods in a range of situations including varying entropy levels and sample sizes. Vermunt (2010) considered a latent class model with covariates that predict class membership. He finds that the modified BCH and 3-step ML approaches result in slightly downward-biased estimates, while 1-step estimates have a slight upward bias (averaging across all scenarios with varying entropy levels and sample sizes). It has also been noted that when the sample size is small and entropy is low (0.36), the 1-step, the modified BCH, and the 3-step ML approaches all fail, although estimates from the 1-step approach are less biased than those from the latter two approaches, especially for large samples ($N = 10,000$). One possible explanation is that at low entropy levels, the differences between classes are overstated, which leads to an underestimation of the classification error (Bakk et al., 2014; Vermunt, 2010). Standard errors are severely underestimated using the modified BCH approach, although using a sandwich variance estimator provides a slight correction. Both the 3-step ML and 1-step approaches give average estimated standard errors (SE) that are close to the standard deviation of the parameter estimates across replications; the former SE is slightly underestimated while the latter is slightly overestimated. The 3-step ML approach is also shown to be roughly as efficient as the 1-step approach.

From studies that considered latent class models with latent variables as predictors of a distal outcome, we can conclude the following. When all necessary model assumptions hold, the sample size is large and class separation is good (entropy > 0.6), all methods perform well with small bias, correct SEs and good coverage. When the sample size is small and entropy is low (< 0.6), all methods can fail with either large bias or poor coverage (e.g. Asparouhov and Muthén, 2014b; Lanza et al., 2013), although the 1-step approach slightly outperforms other methods (Asparouhov and Muthén, 2014a,b).

The robustness of each method to departures from normality has been investigated for continuous H. When H follows a bimodal distribution within classes, the class proportions may be affected for both the 1-step and 3-step ML approaches, which then leads to heavily biased estimates of the effects of C on H (Asparouhov and Muthén, 2014a,b; Bakk and Vermunt, 2016). It has also been noted by Asparouhov and Muthén (2014a) that when estimates for class proportions from step 1 are used as starting values in step 3 (instead of random sets of starting values), the latent class solution remains unchanged. However, in a later investigation of the impact of a bimodal H, the class proportions from step 1 and step 3 differ significantly in almost all replications when entropy is 0.7 (Asparouhov and Muthén, 2014b). The modified BCH approach provides unbiased estimates but poor coverage (below 90%) across all entropy levels, particularly when entropy is low. This is mainly because the weights (w_{ik}s) depend on the misclassification error, which has a higher variability when class separation is unclear (Asparouhov and Muthén, 2014b). When H has a medium or low degree of bimodality, the LTB approach results in large bias when class separation is low, and unbiased point estimates coupled with poorer coverage when class separation is high, compared with the 3-step ML approach that allows for unequal class-specific residual variances (Asparouhov and Muthén, 2014b; Bakk and Vermunt, 2016). Bakk and Vermunt (2016) also showed that both the modified BCH and the 3-step ML approaches are insensitive to unequal class-specific variances because they can explicitly allow for unequal variances, while the LTB approach produces large bias.

Considering the robustness, efficiency, interpretability and the potential for generalisation to more complex model structures with possibly mixed types of distal outcomes that are measured at different levels in a hierarchical structure, the 3-step ML approach is particularly appealing. Further extensions and investigations of this approach are therefore the focus of this thesis.

4.4 Methods to relate multiple categorical latent variables to a distal outcome

When there is more than one categorical latent variable, extensions to standard step-wise approaches are required. The consideration of such an extension was first suggested by Bolck et al. (2004). Bakk et al. (2013) also discussed briefly an application of the 3-step maximum likelihood (ML) approach with two categorical latent variables where one predicts the other. Our proposed model has a more flexible specification that allows for an association between the latent variables through a log-linear model. Similar structures can be found in the structural equation modelling literature. For example, Muthén (2001) discussed a confirmatory latent class analysis of a two-wave panel study, with two associated latent class variables for antisocial behaviour; and mixture growth modelling with repeated measures of two related latent variables capturing fundamental individual differences, where each class has a unique set of growth parameters. In social research, it is common to have more than one latent predictor and researchers may wish to treat these as categorical (e.g. socioeconomic situations). In the structural equation modelling framework, Ploubidis et al. (2015) considered an application with two latent variables, with a causal relationship, that jointly predict distal outcomes (using the modal class approach). Bauldry et al. (2016) fitted a model that estimates the effects of two associated continuous latent summaries of perceptions of physical and personality attractiveness on education attainment (using the 1-step approach).

We now consider an extension of the 3-step method for models with multiple categorical latent variables. The method is described for two categorical latent variables (see Figure 4.1). Extensions to include multiple categorical variables are straightforward and four latent variables are considered in the application in Section 4.6.

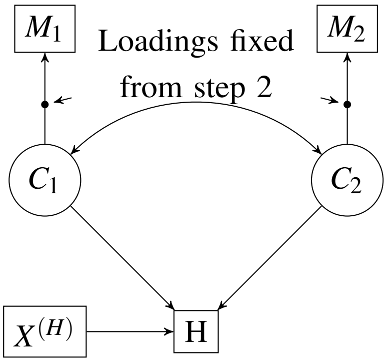


Fig. 4.1 3-step maximum likelihood method for a simplification of a model with two categorical latent variables C_1 and C_2 and the external covariate $X^{(H)}$ predicting a distal outcome H. M_1 and M_2 are modal classes derived from fitting individual latent class models for two distinct sets of indicators (not shown on the figure for clarity). The curved arrow between two latent variables indicates the existence of association.

The associated posterior probabilities of being in each class, as well as the misclassification probabilities, are calculated in the second step for use in the last step. For a general setting (without specifying the direction of the association between C_1 and C_2), a log-linear model can be specified. As the latent variables are both categorical, we consider the cross-classification of C_1 and C_2 (shown in Table 4.1 assuming two classes for each categorical latent variable). Denote by k_1 and k_2 $(k_1, k_2 \in \{1,2\})$ the class index for each latent variable and $\mu_{k_1k_2}$ the expected frequency in each cell. We assume cell counts $\overset{\text{i.i.d}}{\sim}$ Poisson($\mu_{k_1k_2}$).

Table 4.1 Expected frequency (cell counts) for two categorical latent variables

	Counts	C_2 1	2	Total
C_1	1	μ_{11}	μ_{12}	$\mu_{1.}$
	2	μ_{21}	μ_{22}	$\mu_{2.}$
	Total	$\mu_{.1}$	$\mu_{.2}$	N

We model the association by specifying a log-linear model with a two-way interaction between C_1 and C_2,

$$\log(\mu_{k_1k_2}) = \omega_0 + \omega_{k_1}^{(C_1)} + \omega_{k_2}^{(C_2)} + \omega_{k_1k_2}^{(C_1C_2)}, \tag{4.21}$$

where ω_0 is the intercept, $\omega_{k_1}^{(C_1)}$ and $\omega_{k_2}^{(C_2)}$ are the main effects of latent variables C_1 and C_2 and $\omega_{k_1k_2}^{(C_1C_2)}$ is their interaction effect. Latent class memberships of C_1 and C_2, respectively can be obtained from measurement models in step 1, which are carried forward into step 3, leaving four cell frequencies to estimate. (4.21), however, contains nine free parameters to be estimated. For model identification, five constraints need to be imposed and we set

$$\omega_2^{(C_1)} = \omega_2^{(C_2)} = \omega_{12}^{(C_1C_2)} = \omega_{21}^{(C_1C_2)} = \omega_{22}^{(C_1C_2)} = 0, \tag{4.22}$$

if category 2 is taken as the reference for each latent variable.

These ωs are related to the expected frequency counts in Table 4.1 by

$$\begin{aligned}
\mu_{11} &= \exp(\omega_0 + \omega_1^{(C_1)} + \omega_1^{(C_2)} + \omega_{11}^{(C_1C_2)}),\\
\mu_{12} &= \exp(\omega_0 + \omega_1^{(C_1)}),\\
\mu_{21} &= \exp(\omega_0 + \omega_1^{(C_2)}),\\
\mu_{22} &= \exp(\omega_0),
\end{aligned} \tag{4.23}$$

where the parameters (ωs) can be interpreted as log-odds or log-odds ratio via

$$\begin{aligned}\omega_1^{(C_1)} &= \log\left(\frac{\mu_{12}}{\mu_{22}}\right),\\ \omega_1^{(C_2)} &= \log\left(\frac{\mu_{21}}{\mu_{22}}\right),\\ \omega_{11}^{(C_1C_2)} &= \log\left(\frac{\mu_{11}/\mu_{12}}{\mu_{21}/\mu_{22}}\right).\end{aligned} \tag{4.24}$$

The association between categorical latent variables C_1 and C_2 can be measured by the odds ratio $\exp\left(\omega_{11}^{(C_1C_2)}\right)$, where the numerator (μ_{11}/μ_{12}) represents the odds of $C_2 = 1$ for $C_1 = 1$ and the denominator (μ_{21}/μ_{22}) represents the odds of $C_2 = 1$ for $C_1 = 2$. If an odds ratio is above (below) 1, this indicates a positive (negative) association between C_1 and C_2: compared to those assigned to $C_1 = 2$, people assigned to $C_1 = 1$ have a higher (lower) chance to be in the first class of C_2. If the odds ratio is 1, this indicates no association between C_1 and C_2.

After specifying the association between the latent variables, the last step is to fit a regression model to estimate the effects of C_1, C_2 and $X^{(H)}$ on the distal outcome H. Extending (4.18) to a model with two categorical latent variables, we write

$$\operatorname{logit}\left(P(H_i = 1|C_{1i} = k_1, C_{2i} = k_2, X_i^{(H)})\right) = \tau_0 + \sum_{q=1}^{2} \tau_{C_q,k_q} I(C_{qi} = k_q) + \lambda X_i^{(H)}, \tag{4.25}$$

where τ_{C_1,k_1} and τ_{C_2,k_2} are the coefficients of dummy variables for C_1 and C_2 respectively, with the last category of each taken as the reference, i.e. $\tau_{C_1,2} = \tau_{C_2,2} = 0$. (4.25) does not consider the interaction effect of C_1 and C_2 on H but this extension can be easily allowed for.

The resulting log-likelihood of the observed data that is to be maximised in step 3 is

$$\begin{aligned} l(H, M_1, M_2|X^{(H)}; \boldsymbol{\tau}, \lambda) &= \sum_{i=1}^{N} \log P(H_i, M_{1i}, M_{2i}|X_i^{(H)}) \qquad (4.26)\\ &= \sum_{i=1}^{N} \log \sum_{k_2=1}^{K_2-1} \sum_{k_1=1}^{K_1-1} P(H_i, M_{1i}, M_{2i}|X_i^{(H)}, C_{1i} = k_1, C_{2i} = k_2)\\ &\qquad \cdot P(C_{1i} = k_1, C_{2i} = k_2), \end{aligned}$$

where $\boldsymbol{\tau} = \{\tau_0,\ \tau_{C_1,k_1},\ \tau_{C_2,k_2}\}$ $(k_1, k_2 = \{1,2\})$ and we set $\tau_{C_1,K_1} = \tau_{C_2,K_2} = 0$.

Applying the chain rule, the first probability term can be written as

$$\begin{aligned}P(H_i,M_{1i},M_{2i}|X_i^{(H)},C_{1i}=k_1,C_{2i}=k_2) = & P(H_i|X_i^{(H)},C_{1i}=k_1,C_{2i}=k_2,M_{1i},M_{2i}) \\ & \cdot P(M_{1i}|M_{2i},X_i^{(H)},C_{1i}=k_1,C_{2i}=k_2) \\ & \cdot P(M_{2i}|X_i^{(H)},C_{1i}=k_1,C_{2i}=k_2). \qquad (4.27)\end{aligned}$$

Again, for the two categorical latent variable case, we need a similar conditional independence assumption as that in the one latent variable case. We assume that M_1 and M_2 do not impact the conditional distribution of H given C_1 and C_2. This is equivalent to

$$P(H|C_1,C_2) = P(H|C_1,C_2,M_1,M_2). \qquad (4.28)$$

In addition, as in the first step, we perform separate latent class analyses for C_1 and C_2. An implicit assumption is required that C_2 does not influence the conditional distribution of $M_1|C_1$ and that C_1 does not influence the conditional distribution of $M_2|C_2$ (defined by (4.17)).

With the above assumptions, (4.27) can be reduced to the product of three probabilities

$$\begin{aligned}P(H_i,M_{1i},M_{2i}|X_i^{(H)},C_{1i}=k_1,C_{2i}=k_2) = & P(H_i|X_i^{(H)},C_{1i}=k_1,C_{2i}=k_2) \\ & \cdot P(M_{1i}|C_{1i}=k_1) \\ & \cdot P(M_{2i}|C_{2i}=k_2), \qquad (4.29)\end{aligned}$$

where the first term on the right-hand side can be computed from (4.25) in step 3 and the other two terms are fixed at values computed from (4.17) in the second step.

The second probability term in (4.26) can be derived from the log-linear model for latent variables (4.21-4.23), where $P(C_1=k_1,C_2=k_2) = \frac{\mu_{k_1k_2}}{N}$.

The extension of the 3-step approach to include multiple categorical latent variables allows for a flexible specification of the association between latent variables. Note that in step 3, we are essentially estimating a latent class model where Ms are imperfect measurements for Cs with fixed loadings; and the loading for the path towards H is freely estimated. Asymptotic standard errors can be estimated using the formula derived in Section 6.6 of Bartholomew et al. (2011) or the parametric bootstrap method. As our model is not overly complex, the former approach is used throughout the estimation.

4.5 Simulation study

As noted in Section 4.3.5, the few studies that have compared the performance of the 3-step maximum likelihood approach and simultaneous 1-step methods have found that their performance is often similar when modelling assumptions are valid. We build upon earlier work by conducting a simulation study to further investigate the relative performance of the two methods under departures from two key model assumptions (within-class normality of continuous H and conditional independence of $Y^{(C)}$s and H) and for the extension to two categorical latent variables. For both investigations, we are concerned with potential bias of coefficients for the latent variables in the model for H, as well as a more fundamental problem where the number of classes that are needed to capture the association among the $Y^{(C)}$s may be altered. Extending the work of Bakk and Vermunt (2016) that examined the robustness of the 3-step ML approach for bimodal and heterogeneous class-specific distributions of H, we consider the performance of the general 3-step ML approach and the 1-step approach under other forms of non-normality, i.e. skewness and excess kurtosis. These two forms of non-normality are common in practice and may not be well captured by a finite mixture of normal components. Previous research has found that non-normality of H can lead to both biased coefficients and shifted class proportions for both methods (Asparouhov and Muthén, 2014a; Bakk and Vermunt, 2016). We build on the literature by investigating whether non-normality of H affects the number of classes needed in the measurement model for the 1-step approach. We anticipate that additional (spurious) classes may be required to capture the distribution of H. The second investigation of the impact of local dependence of H on both methods has not been considered in previous research. We anticipate that if such dependence is not accounted for, both methods can give biased estimates and the 1-step approach may even identify spurious classes.

4.5.1 Design and population parameters

We generate data from models with a distal outcome H. Both continuous and binary H are considered. For simplicity, we do not include covariates but extensions are straightforward. We compare the performance of the 1-step and 3-step methods in a number of scenarios where all assumptions are met (Study 1), when the normality assumption about a continuous H is violated in various ways (Study 2) and when the conditional independence assumption is violated (Study 3). Note that assumptions considered in studies 2 and 3 are common to both methods. We generate data from a measurement model with ten dichotomous indicators, where the first five measure C_1 and the latter five measure C_2. Taking the second class as the reference for both latent variables, class 1 (2) of C_1 and C_2 gives high (low) response probabilities for all five indicators. Latent variables are generated from the log-linear model specified by (4.21-4.24) of Section 4.3.4. We set $\omega_1^{(C_1)} = \omega_1^{(C_2)} = 0.7$ and $\omega_{11}^{(C_1C_2)} = 0$ for the setting of independent categorical latent variables and $\omega_{11}^{(C_1C_2)} = -0.5$ for the associated case (negative association). Table 4.2 shows the distribution of categorical latent variables generated using the above specification.

Table 4.2 Proportion in each latent class generated for all scenarios considered in the simulation study.

Independent C_1, C_2		C_2 1	2	Total
C_1	1	0.38	0.14	0.52
	2	0.31	0.17	0.48
	Total	0.69	0.31	1
Associated C_1, C_2		**C_2 1**	**2**	**Total**
C_1	1	0.33	0.27	0.60
	2	0.27	0.13	0.40
	Total	0.60	0.40	1

Similar to previous research (Asparouhov and Muthén, 2014a; Bakk et al., 2014; Bakk and Vermunt, 2016), varying sample sizes are considered and we manipulate the entropy levels through class-specific thresholds in the measurement model. For each item we set the class-specific thresholds in the measurement model to be $\alpha_{1pk} = -\alpha_{2pk}$ (following (4.2) and (4.3) in Section 4.2.1). The value of these thresholds is modified to obtain different levels of entropy (degree of class separation). Specifically, to obtain entropy values of 0.8, 0.7, 0.55 and 0.4, α_{1pk} is set to 1.5, 1.25, 1, 0.75, accordingly.

The distal outcome H is generated from the model $H = \beta_0 + \beta_1 I(C_1 = 1) + \beta_2 I(C_2 = 1) + \varepsilon$ where $\varepsilon \sim N(0,1)$ and from the model $\text{logit}(P(H = 1|C_1, C_2)) = \beta_0 + \beta_1 I(C_1 = 1) + \beta_2 I(C_2 = 1)$ for a binary H (note that $I(\cdot)$ denotes the indicator function). Other parameter values for the population model specific to each scenario are provided in the corresponding sub-sections below. In the 3-step ML approach, as results from previous studies show that modal and proportional assignments of individuals to classes lead to similar parameter estimates in step 3 (Bakk et al., 2013; Bakk and Vermunt, 2016; Vermunt, 2010), and that the proportion of misclassified observations is smaller using the modal assignment, we use modal assignments in step 1. In order to mimic empirical studies, in the 1-step approach we use 100 sets of random starting values and impose parameter constraints (e.g. greater or less than zero) on log-linear parameters for latent class allocations. These constraints are mainly set to avoid potential label switching in the class allocation of mixture models. We generate 500 replications in each study. The latent class models for the two sets of binary items are estimated separately in Mplus 7.31 (Muthén and Muthén, 2017); the modal class assignments and misclassification probabilities are then exported to Latent GOLD 5.0 (Vermunt and Magidson, 2015) for step 3 of the estimation procedure. The reported summary statistics are relative bias (%), average standard error across replications (SE), standard deviation of estimates across replications (SD) and 95% coverage rates. We compare SE and SDs to check if the SEs are estimated unbiased. The codes for selected simulation studies are included in Appendix A.5.

4.5.2 Study 1: all model assumptions are satisfied

Simulations are carried out for combinations of low and high entropy levels, sample sizes of 500, 2000 and 10,000, and for correlated and independent latent variables. As it is more likely to encounter cases with associated latent variables, we present results for correlated C_1 and C_2 (results for independent latent variables are very similar and the models converge more quickly). As results are fairly similar for all values of N (except for gains in estimation efficiency when sample sizes are large), we present results only for $N = 2000$. For continuous H, coefficients are set at $\beta_0 = 3$, $\beta_1 = 2$, $\beta_2 = -1.5$; for binary H, $\beta_0 = 1.2$, $\beta_1 = 1$, $\beta_2 = -1.5$. Similar to the results obtained for the 3-step ML approach with one latent variable (Asparouhov and Muthén, 2014a; Muthén and Muthén, 2017), Table 4.3 presents results from a realistic scenario where the class separation is good in one measurement model but poor in the other. Results show that both the 3-step ML and 1-step approaches give unbiased estimates and excellent coverage for almost all parameters when model assumptions hold. The 3-step ML approach gives a slightly lower coverage for $\omega_1^{(C_2)}$ due to its underestimated SE. This is probably because the entropy of the measurement model for C_2 is low, and a large amount of variability in the estimated misclassification error (computed in step 2 and carried forward into the last step) is ignored in step 3. In the following studies, we investigate the relative performance of these two methods in scenarios where the model assumptions are violated in various ways.

In Appendix A, we have also included results from estimating the model using the modal class approach. As it has been often used in the literature, it serves as a good reference point. The additional simulation study produces an interesting but unexpected finding. We initially expect a good performance of the modal class approach when the separation of classes is clear in both LCAs (i.e. using the most likely class membership as an observed covariate to predict distal outcomes). However, tables in Appendix A.3 show that this is not true, both for continuous and binary distal outcomes. Estimates are heavily biased and coverage rates are far below the nominal 95%, even when values of the entropy of both LCAs are above 0.7. Among the studies relating latent class variables to distal outcomes, Asparouhov and Muthén (2014a) reported similar results (see their Table 6) while evaluating the effect of one categorical latent variable on a distal outcome. They noticed that the pseudo-class approach performs poorly (with large bias and low coverage rates) even with a clear class separation (entropy = 0.7). Note that when the value of entropy is high, pseudo class and modal class assignments are similar as the class assignment is based primarily on posterior probabilities of being in each class. As our model allows for more than one categorical latent variable, the failure of the modal class approach in the single latent class variable case becomes more obvious (with a zero coverage rate for some parameters). We also find that as the values of entropy become smaller, the standard errors estimated using the modal class approach tend to be underestimated (differ more from SD). From the extended simulation study presented in Appendix A.3 with an exogenous covariate X predicting the distal outcome, we also find that contrary to what we observe in the continuous H case, the modal class approach produces a biased estimate for the effect of X on the binary H (β_3) and a poor confidence interval coverage (see tables in Appendix A.3.2), even when entropy level is high (i.e. good class separation).

In addition to the above scenarios, simulations are also performed for models with higher values of entropy in order to find situations where the traditional modal class approach is appropriate. Simulations are performed by setting values of entropy at 0.85, 0.88, 0.90 and 0.92, respectively, keeping the other population parameters unchanged. We find that when the values of entropy are both above 0.9, the modal class approach gives unbiased estimates with close to 95% coverage but it is rare in real applications to have such clear class separations.

Table 4.3 Study 1: Simulation results when all model assumptions are satisfied (N=2000; entropy set to 0.7 (high) and 0.4 (low); 500 replications).

Continuous H		1-step				3-step			
Parameters	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
$\beta_1\,(C_1)$	2.00	−0.03	0.07	0.07	0.94	−0.01	0.04	0.04	0.95
$\beta_2\,(C_2)$	−1.50	0.35	0.07	0.07	0.93	0.02	0.01	0.01	0.93
$\omega_1^{(C_1)}$	0.70	−0.87	0.08	0.06	0.94	0.01	0.07	0.08	0.94
$\omega_1^{(C_2)}$	0.70	−2.26	0.14	0.08	0.94	−1.05	0.08	0.11	0.81
$\omega_{11}^{(C_1C_2)}$	−0.50	2.28	0.11	0.08	0.93	−1.03	0.09	0.09	0.95
Binary H		1-step				3-step			
Parameters	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
$\beta_1\,(C_1)$	1.00	0.49	0.20	0.20	0.96	0.05	0.08	0.08	0.97
$\beta_2\,(C_2)$	−1.50	0.43	0.19	0.20	0.97	0.10	0.14	0.14	0.94
$\omega_1^{(C_1)}$	0.70	−0.11	0.07	0.07	0.97	0.01	0.08	0.09	0.94
$\omega_1^{(C_2)}$	0.70	0.06	0.10	0.11	0.96	0.01	0.09	0.13	0.83
$\omega_{11}^{(C_1C_2)}$	−0.50	0.00	0.08	0.09	0.97	−0.02	0.11	0.11	0.95

Bias (%)=(Estimate-True)/True × 100%

4.5.3 Study 2: violation of the normality assumption about H

We are mainly concerned with the situation where the continuous distal outcome H is non-normal, but we fit a standard finite mixture model assuming within-class normality. The normality assumption is common to both the 1-step and 3-step ML approaches. In the simultaneous 1-step approach, H is treated as an additional indicator for the latent variables. We therefore hypothesise that compared to the 3-step ML approach where the measurement model in step 1 is estimated separately from H, the 1-step approach is more sensitive to non-normal within-class distributions. Through the simulation study, we evaluate the relative performance of the two methods when the within-class distribution of continuous H exhibits skewness, excess kurtosis, and bimodality, respectively. The results for bimodality are given in Appendix A.4 as they are similar to results for the single latent variable case (Bakk et al., 2014; Bakk and Vermunt, 2016) but with slightly lower coverage in situations with poor class separation. We also conduct simulations using the modified BCH approach (detailed results in Appendix A.4) as previous research found its robustness to violations of distributional assumptions (Asparouhov and Muthén, 2014b; Bakk and Vermunt, 2016), hence serving as a good reference point.

We simulate non-normality by generating H from a mixture of non-normal and normal distributions. We first focus on the impact of these forms of non-normality on the main coefficients of interest, i.e. β_1, β_2 (estimates for ωs are shown in Appendix A.4). Next we investigate whether the number of classes needed in the mixture model can be influenced by non-normality. For each form of non-normality, entropy is fixed at the same value for each latent variable: 0.7 (high) and 0.4 (low). Parameters of the log-linear model for latent variables are set at values (see Section 4.5.1) that generate the following proportions for cells in the cross-classification of C_1 and C_2 (hereafter referred to as class patterns): 0.33 for [C_1=1, C_2=1], 0.27 for [1,2] and [2,1] and 0.13 for [2,2]. Across all sub-studies, we generate sample sizes of N=200 and 2000, where 200 can be regarded as a (very) small sample.

Study 2a: Excess kurtosis

In this scenario, we generate H from the model $H = 3 + 2I(C_1 = 1) - 1.5I(C_2 = 1) + \varepsilon$, where ε is drawn from a student-t distribution with 7 degrees of freedom (excess kurtosis =2) for class patterns [1,2] and [2,1], but from $N(0,1)$ for the other two class patterns. Selected results are summarised in Table 4.4. We find that the 3-step ML approach provides unbiased estimates in all situations apart from small N and low entropy, where there is an obvious change in the class proportions from step 1 to step 3 (see Table A.19 in Appendix A for estimates of ωs). For both the 3-step ML and 1-step approaches, performance is best for large N and clear class separation. However, when class separation is poor, the 1-step approach produces large bias. This is probably because when class separation is poor, class memberships are heavily influenced by the distal outcome H. Comparing the results of the 3-step ML approach with that of the BCH approach in Table A.23, the performance of the two approaches is similar across all scenarios investigated, but we noticed the large bias in BCH estimates in scenarios with a small sample size and a low entropy value.

Table 4.4 Study 2a: Simulation results for excess kurtosis (N=200, 2000; 500 replications).

Parameters		1-step				3-step			
	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
N=200, High entropy									
$\beta_1\,(C_1)$	2.00	−4.48	0.30	0.73	0.90	0.60	0.19	0.19	0.94
$\beta_2\,(C_2)$	−1.50	−7.80	0.30	0.60	0.92	−0.30	0.19	0.19	0.96
N=200, Low entropy									
$\beta_1\,(C_1)$	2.00	−8.43	0.59	0.77	0.89	−14.38	0.33	0.44	0.84
$\beta_2\,(C_2)$	−1.50	−11.45	0.59	0.63	0.88	−25.14	0.36	0.79	0.85
N=2000, High entropy									
$\beta_1\,(C_1)$	2.00	0.03	0.09	0.09	0.95	0.96	0.06	0.06	0.94
$\beta_2\,(C_2)$	−1.50	−0.09	0.09	0.09	0.97	1.57	0.06	0.06	0.95
N=2000, Low entropy									
$\beta_1\,(C_1)$	2.00	−52.03	0.14	1.72	0.72	−2.68	0.10	0.10	0.92
$\beta_2\,(C_2)$	−1.50	55.01	0.15	1.29	0.70	−3.21	0.11	0.11	0.92

Study 2b: Skewness

We now consider H with a skewed distribution for class pattern [1 2] and [2 1] but a normal distribution for the other class patterns. In the simulation, we generate residual ε from the log-normal distribution with zero mean. We generate a right-skewed H from the model $H = 3+2I(C_1=1)-1.5I(C_2=1)+\varepsilon$ (skewness= 5.0) for class pattern [1 2] and for class pattern [2 1], we generate a left-skewed H from the model $H = 3+2I(C_1=1)-1.5I(C_2=1)-\varepsilon$ (skewness= -5.0). For other class patterns, $\varepsilon \sim N(0,1)$. To capture the heterogeneity of the data, it is standard practice to fit a finite mixture model assuming within-class normality. The results are presented in Table 4.5. In general, although both approaches give biased estimates, the 3-step ML approach yields less bias than the 1-step approach when the within-class distribution of H is skewed. When class separation is clear the general 3-step ML approach produces estimates with a relative bias slightly over 5%, even when the sample size is small. However, when class separation is poor, class assignment in step 1 is shifted in step 3 (see Table A.20 in Appendix A for the heavily biased estimates of the ω parameters), which can partly explain the biased estimates for βs. It should also be noted that for the general 3-step ML approach alone, skewness can lead to heavier bias than bimodality and excess kurtosis. Clark and Muthén (2009) showed that kurtosis can be approximated as a quadratic function of skewness and hence if the distribution of the data is highly skewed, it also has severe excess kurtosis, which can exacerbate the bias in parameter estimates. Comparing results with the BCH approach (see Table A.22), the BCH approach outperforms the 3-step ML approach in terms of relative bias and coverage of the estimated coefficients for almost all combinations of entropy levels and sample sizes tested, with the exception of the case where the class separation is poor and sample size is small ($N = 200$).

Table 4.5 Study 2b: Simulation results for skewness (N=200, 2000; 500 replications).

Parameters		1-step				3-step			
	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
N=200, High entropy									
$\beta_1\,(C_1)$	2.00	−65.44	0.38	1.64	0.57	−7.31	0.27	0.23	0.90
$\beta_2\,(C_2)$	−1.50	66.67	0.45	1.44	0.60	−5.66	0.27	0.24	0.94
N=200, Low entropy									
$\beta_1\,(C_1)$	2.00	−71.66	1.32	2.15	0.58	−22.73	0.46	0.49	0.56
$\beta_2\,(C_2)$	−1.50	−24.77	1.34	2.16	0.72	−41.29	0.46	0.49	0.76
N=2000, High entropy									
$\beta_1\,(C_1)$	2.00	−84.37	0.13	1.74	0.47	−7.25	0.09	0.07	0.63
$\beta_2\,(C_2)$	−1.50	72.31	0.13	1.23	0.54	−6.22	0.09	0.07	0.82
N=2000, Low entropy									
$\beta_1\,(C_1)$	2.00	−150.11	0.12	0.68	0.03	−28.89	0.14	0.33	0.20
$\beta_2\,(C_2)$	−1.50	145.00	0.14	0.38	0.03	−38.09	0.14	0.33	0.25

Number of classes

In the above simulations, we observe that in some situations the 1-step approach produces heavily biased estimates (including those for the ω parameters shown in Appendix A). As noted earlier, it is also possible that for the 1-step approach non-normality of H may lead to an increase in the number of latent classes needed to fit the data (Bauer, 2007). This is a major concern in empirical studies because a change in the number of classes may alter their interpretation and hence bias their estimated effects on the distal outcome. For the 3-step ML approach, the measurement model is estimated without H in step 1 such that the true number of classes should be obtained (provided the assumptions in step 1 hold).

To investigate the impact of non-normality on the number of classes required in the 1-step approach, data are generated from a model with only one categorical latent variable (1-LV) with four classes. The 2-LV model with two independent classes for each variable can be viewed as a 1-LV model with four classes. Linking with previous scenarios, the same class proportions are generated, corresponding to four class patterns in the 2-LV model but with no correlation. For data generated from a 4-class model, models with 3-5 classes are fitted and the sample size adjusted BIC (ssaBIC) and p-values from the bootstrap likelihood ratio test (BLRT) are obtained for each, following the recommendation of Nylund et al. (2007). All parameter values in the population model for each scenario remain the same as in the 2-LV model.

Table 4.6 shows the percentage of replications for which each model has the minimum BIC value and for which each model is rejected or not based on the BLRT. If the number of classes is unaltered, we expect values in the ssaBIC column close to 100% for the 4-class model and close to 0% for the 3 and 5-class models. We expect the percentage of replications rejecting the null hypothesis be roughly 95% based on BLRT for tests of 2-class vs 3-class and 3-class vs 4-class models and close to 5% for the test of the 4-class vs 5-class model (p-values presented in brackets are averaged across replications).

Of the three types of within-class non-normality of H considered, skewness is particularly troublesome as the selection rate based on ssaBIC is the lowest (below 80%). When N=2000, we also observe that ssaBIC identifies more classes than there truly exist (ssaBIC agrees with BLRT when class separation is poor). When the sample size is small (N=200), we focus on BLRT as Nylund et al. (2007) show that ssaBIC performs poorly in such situations. When class separation is poor (low entropy), fewer classes are needed for all types of non-normality (according to BLRT). However, when the class separation is good, regardless of the sample size, none of the three types of non-normality of H influence the ability of the 1-step approach to correctly identify the number of classes.

Combining the results from investigations of bias in the coefficient estimates and of the ability to extract the true number of classes when the within-class normality assumption is violated, it is obvious that, although the assumption is common to both the 1-step and the 3-step ML methods, the 1-step approach is more sensitive than the 3-step ML approach to all forms of non-normality. Both methods perform the worst when there is within-class skewness in H, but the 1-step approach is also likely to alter the number of classes needed to fit the data (consistent with the findings of Bauer (2007)). Moreover, both methods perform poorly when sample size is small and the class separation is poor, as expected.

Table 4.6 Simulation results for the number of classes when there is within-class non-normality of H. Results are reported for the percentage of replications for which each model has the minimum ssaBIC value and for which each model is rejected or not based on the BLRT. The selected models are bolded. ssaBIC=sample size adjusted BIC; BLRT=bootstrap likelihood ratio test (average p-values across 500 replications in brackets).

	Entropy	ssaBIC (%)			BLRT (%)		
Scenarios, N=200		3-class	4-class	5-class	3-class (H_0:2-class, H_1:3-class)	4-class (H_0:3-class, H_1:4-class)	5-class (H_0:4-class, H_1:5-class)
Bimodality	High	0	**94**	6	90(0.04)	**100(0.00)**	65(0.30)
	Low	11	**79**	10	40(0.43)	41(0.32)	16(0.58)
Excess kurtosis	High	0	**93**	7	96(0.03)	**100(0.00)**	65(0.29)
	Low	12	**79**	9	41(0.41)	29(0.37)	12(0.60)
Skewness	High	0	**79**	21	98(0.02)	**99(0.01)**	64(0.29)
	Low	4	**76**	20	72(0.20)	63(0.18)	27(0.50)
Scenarios, N=2000							
Bimodality	High	0	**100**	0	100(0.00)	**100(0.00)**	18(0.73)
	Low	0	**84**	16	100(0.00)	**100(0.00)**	30(0.56)
Excess kurtosis	High	0	**98**	2	100(0.00)	**100(0.00)**	31(0.64)
	Low	0	**97**	3	100(0.00)	**100(0.00)**	24(0.60)
Skewness	High	0	48	**52**	100(0.00)	**100(0.00)**	62(0.36)
	Low	0	5	**95**	100(0.00)	100(0.00)	**97(0.03)**

4.5.4 Study 3: violation of the conditional independence assumption about H

In addition to the above evaluations, we are interested in studying the impact on the number of classes extracted and the parameter estimates of violation of the assumption that H and the $Y^{(C)}$s are conditionally independent given the latent variable C. This has not been studied in previous research but the conditional independence assumption is common to both the 1-step and 3-step ML approaches, i.e. $P(\mathbf{Y}^{(C)}, H|C) = P(\mathbf{Y}^{(C)}|C)P(H|C)$ (see Bakk et al. (2013)). Compared to the 1-step approach, one obvious advantage of the 3-step ML approach is that it is not subject to the change in the number of classes when local dependence of H is present as the decision of the number of classes to be retained is made in step 1, without H. We also anticipate that if such residual dependence is not accounted for, the 1-step approach will produce more biased estimates for the relationship between categorical latent variables and H than the 3-step ML approach, as a wrong model with insufficient or extra classes will be estimated. Study 3 investigates the relative performance of both approaches for different entropy levels and sample sizes.

For ease of illustration, we consider a 1-LV model with four classes and continuous H. Data were generated from a model with class proportions of 0.30, 0.25, 0.25, 0.20. We then generate ten binary indicators conditional on class membership from a logit measurement model. We consider the same high and low entropy situations by manipulating the thresholds in the measurement model described earlier, for sample sizes N=500 and 2000. We increase the small sample size from 200 (used in Study 2) because we have reformulated the 2-LV model with two classes for each variable as a 1-LV model with four independent classes. The larger sample size of N=500 helps to avoid boundary solutions due to small classes. Next, to induce local dependence between item $Y_{10}^{(C)}$ and H, we introduce an additional continuous random variable $u \sim N(0,4)$. Note that conditional independence between all $Y^{(C)}$s is still valid so that the measurement model is correctly specified. In addition, we set the class-specific variance of ε to be 4, 3, 2, 1 for C=1, 2, 3, 4, respectively and the corresponding class-specific means are 3.5 (β_1), 5 (β_2), 1.5 (β_3) and 3 (β_4). The data are then analysed using both the 1-step and the 3-step ML approaches using the DU3STEP command in Mplus. Note that we employ a slightly different parameterisation to that in earlier simulations as we are restricted by the technicalities of the program. The parameters estimated are means of H in each latent class rather than contrasts with a reference category.

We first check if the number of classes needed is altered in this scenario using the 1-step approach. The results presented in Table 4.7 show that when conditional independence holds the 1-step approach tends to identify fewer classes when the class separation is unclear and especially when sample size is small; these findings are consistent with those of Nylund et al.

(2007). Second, both ssaBIC and BLRT indicate that even when the conditional independence assumption between H and items in the measurement model is violated for only one item, there is a tendency to extract additional (spurious) classes, irrespective of the level of class separation. When there is local dependence between H and an indicator, the percentage of times that ssaBIC favours the $(K+1)$–class model over the correct K-class model increases. For example, in the high entropy case with $N = 500$, ssaBIC suggests a correct 4-class model in 89% of replications when conditional independence holds, which decreases to 61% when the assumption breaks down. Similarly, the percentage of times that ssaBIC suggests the 5-class model increases from 13% when conditional independence holds, to 39% when the assumption breaks down.

In addition, we observe that when class separation is unclear or sample size is small, there is greater disagreement between the ssaBIC and BLRT statistics. Our findings are particularly worrying for empirical studies as the assumption of conditional independence between the distal outcome and items that measure the latent variable is rather strong. If such local dependence is not accounted for in the model, we anticipate that the problem discussed above will be exacerbated when H is correlated with more than one item.

We next examine the impact of departures from conditional dependence on the estimated coefficients when a model with the correct number of classes is fitted using both the 1-step and 3-step ML approaches. The simulations are performed in Mplus using the DU3STEP command that allows for unequal class-specific variances. The results are reported in Table 4.8. For illustrative purposes, we only present results for parameters β_3 and β_4 (class-specific means of H for $C = 3$ and $C = 4$) as they are the most biased among all βs.

Table 4.7 Simulation results for the number of classes when there is local dependence (1-step approach). Results are reported in (%) and selected models are bolded. ssaBIC=sample size adjusted BIC; BLRT=bootstrap likelihood ratio test (average p-values across 500 replications in brackets).

N	Scenario	ssaBIC(%)			BLRT(%)		
High entropy		3-class	4-class	5-class	3-class	4-class	5-class
					(H_0:2-class, H_1:3-class)	(H_0:3-class, H_1:4-class)	(H_0:4-class, H_1:5-class)
500	Independence[a]	0	**89**	13	100(0.00)	**100(0.00)**	1(0.82)
	Dependence[b]	0	**61**	39	100(0.00)	**99(0.00)**	14(0.61)
2000	Independence	0	**100**	0	100(0.00)	**100(0.00)**	2(0.77)
	Dependence	0	2	**98**	100(0.00)	100(0.00)	**100(0.00)**
Low entropy							
500	Independence	**64**	32	4	68(0.12)	7(0.65)	1(0.82)
	Dependence	31	**52**	17	64(0.13)	26(0.41)	4(0.70)
2000	Independence	12	**88**	0	100(0.00)	**100(0.00)**	2(0.80)
	Dependence	0	18	**82**	100(0.00)	100(0.00)	**97(0.01)**

[a] Independence refers to independence between H and $Y^{(C)}$s;

[b] Dependence refers to residual correlation between H and $Y_{10}^{(C)}$.

Table 4.8 Study 3: Estimated coefficients for fitting a 4-class model when there is local dependence between H and $Y_{10}^{(C)}$.

Parameters	1-step				3-step			
	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
N=500, High entropy								
β_3	−11.35	0.32	0.32	0.92	−6.69	0.27	0.30	0.89
β_4	−5.93	0.35	0.35	0.92	−3.54	0.29	0.31	0.93
N=500, Low entropy								
β_3	−59.52	0.59	0.65	0.52	9.01	0.40	0.76	0.70
β_4	−9.70	0.68	0.93	0.66	−6.68	0.44	0.92	0.60
N=2000, High entropy								
β_3	−11.54	0.15	0.16	0.79	−10.98	0.13	0.14	0.72
β_4	−6.04	0.16	0.16	0.82	−6.07	0.14	0.14	0.75
N=2000, Low entropy								
β_3	−60.00	0.30	0.36	0.14	−10.95	0.21	0.48	0.62
β_4	−10.67	0.52	0.70	0.64	−10.64	0.24	0.63	0.50

Clearly, regardless of the entropy level, the 1-step approach is more sensitive to the violation of the conditional independence assumption than the 3-step ML approach; the latter produces estimates with at most 11% relative bias. The 1-step method performs particularly poorly when entropy is low. This is expected as forcing the 1-step approach to estimate a 4-class model (when the 5-class model is a better fit) leads to changes in the interpretation of classes, resulting in larger bias in the estimates. The same rationale also applies to the observation that for the 3-step ML approach, the relative bias does not seem to reduce when the sample size increases, and regardless of the entropy level. Comparing with the results from the BCH approach (see Table A.24), we find a similar performance for 3-step ML and BCH, except for the scenario with poor class separation and large sample size, where the modified BCH approach produces greater bias.

4.5.5 Summary of simulation results

The simulation studies show that when all model assumptions are satisfied, the 1-step and 3-step ML approaches perform equally well. When model assumptions are violated, the estimates from both methods are subject to bias, although the 3-step ML approach is less sensitive. Specifically, when there is within-class non-normality for a continuous H, skewness of H is shown to be the worst form of non-normality for both approaches, compared to bimodality and excess kurtosis. Moreover, the results confirmed a major drawback of the 1-step approach as it not only alters the class proportions (also shown in Asparouhov and Muthén (2014a) and Bakk and Vermunt (2016)), but also changes the number of classes needed to capture the association among indicators, particularly at low entropy levels. When there is local dependence between H and the indicators for the latent variables, the 1-step approach leads to greater bias than the 3-step ML approach. This is mainly explained by a tendency to extract too many classes when there is residual correlation between H and the $Y^{(C)}$s. It should be noted that the extraction of pseudo classes is not necessarily wrong from a theoretical point of view, but one needs to question the validity of such extra classes, which may not be interpretable.

Comparing results of the 3-step ML approach with the BCH approach, in general, we do not observe a consistently better performance of the modified BCH approach in situations where model assumptions are violated, except for the case where the conditional distribution of H is skewed. If in applications of 3-step ML, a substantial shift in classification from step 1 to step 3 is observed, the general 3-step ML approach may not be appropriate and further developments of the BCH approach for more than two latent variables could be helpful in this situation. However, in addition to the severe underestimation of standard errors in the BCH approach (see Tables A.21 to A.24), Bakk and Vermunt (2016) also noted the presence of negative cell frequencies for the BCH approach in an application with a categorical distal outcome and poor class separation. Overall, the development of the 3-step approach is more promising as it is more easily generalised to multilevel models for longitudinal and other forms of clustered data (more discussions see Chapters 5 to 7).

Regarding the impact of manipulating design factors (i.e. entropy and sample size) on the amount of bias of the general 3-step ML approach, we find that in cases where distributional assumptions are violated, low entropy levels (when sample size is fixed) and small sample size (when entropy is fixed) lead to poor estimates. In the case where there is local dependence between the distal outcome H and an indicator $Y^{(C)}$, the performance of the 3-step ML approach is similar at high and low entropy levels for fixed sample size. When entropy is fixed, the 3-step approach tends to produce greater bias as sample size increases. This could be explained by the fact that class proportions in step 3 of the 3-step approach are influenced by the inclusion of H and such influence is more obvious in larger sample sizes.

There are several issues that have not been addressed or discussed in this simulation study. First, in the three simulation studies, we have assumed a measurement model where the latent class solution is not influenced by the outcome H. This is natural when we are interested in an H that is temporally distal to the indicators $\mathbf{Y}^{(C)}$, as is common in longitudinal studies. However, it is possible that H is an important indicator that helps to identify the latent classes, for example when H and the $Y^{(C)}$s are measured contemporaneously. In this case, the true data-generating model would take the form of the 1-step model, and we would expect that the 3-step ML approach that excludes H from the measurement model would lead to incorrect latent class solutions. Second, although simulation results suggest that the 1-step approach tends to extract extra classes when local dependence exists, it should be noted that the approach is also flexible enough to allow for additional pairwise association between H and the $Y^{(C)}$s without introducing additional classes. In contrast, as H is only introduced in the last step of the 3-step ML approach, it is less straightforward to adapt this approach to account for local dependence. Third, as we have shown several limitations of the 3-step ML approach when model assumptions do not hold, further research is required that modifies the current approach to improve its robustness.

4.6 Application: a study of the effects of childhood socioeconomic circumstances on midlife health

We now illustrate the general 3-step ML approach in an analysis of the effects of four categorical latent variables, capturing different aspects of childhood socioeconomic situations (SECs), on midlife health. The data are taken from the 1958 British cohort study that contains four waves of childhood information (collected at ages 0, 7, 11 and 16). The distal outcome H is a binary indicator for midlife health state, derived from the five-category self-reported health status. We consider repeated measures of four aspects of childhood SECs: social class (father or male head's occupation), financial difficulty, material hardship and family structure. The construction of these indicators is guided by the works of Hobcraft (1998), Schoon et al. (2003) and Chandola et al. (2006a). More details of the construction and distribution of these repeated measures and the health outcome are available in Section 3.2. Syntax for the empirical study is included in Section A.6 of Appendix A.

In step 1, we fit separate latent class models to each set of repeated measures. The results of the latent class models for childhood SECs are summarised in Table 4.9, where descriptive labels for each class are based on an examination of the pattern of change in categories of repeated measures over four childhood waves. For each latent variable, the choice of the number of classes is based on a combination of goodness-of-fit statistics and tests (including the AIC, sample-size adjusted BIC, entropy, and the Lo-Mendell-Rubin and bootstrap likelihood ratio tests), the proportion in each class (ensuring at least 10% in each category) and interpretation of the classes. Further details are provided in Section A.2 of Appendix A. Table 4.9 below reports the percentages of individuals with missing data on all items. Note that we label the nominal classes in an ordinal way to aid interpretations. The assigned orders are based on the examination of the estimated probabilities of each grouped categories of childhood measures across four childhood waves, conditional on class memberships (see Figure A.2 of Appendix A). In step 2, misclassification probabilities are calculated and treated as fixed loadings in step 3.

Table 4.9 Most likely class membership for each aspect of childhood SECs derived from fitting latent class models to each of the four sets of repeated composite measures of childhood SECs. Entropy is also reported where values close to one indicate good class separation.

Childhood measure	Entropy	Modal class allocation (%)			
		Missing(%)	1	2	3
Social class[a]	0.727	2.9	21.0 (high)	51.0 (medium)	25.1 (low)
Financial difficulty	0.700	2.2	80.7 (low)	17.1 (high)	
Material hardship	0.790	12.5	28.9 (low)	30.7 (high)	27.9 (medium)
Family structure/union	0.916	0.9	9.3 (unstable)	89.7 (stable)	

[a] Father or male head social class

In step 3, latent variables are related to outcomes of interest, the binary health state at age 50. Note that we consider a binary health indicator as 1) a binary outcome is of substantive interest because health state (e.g. self-reported general health state) is often measured on a categorical scale, of which a binary variable is the simplest form and, 2) for a continuous H, non-normal distributions within classes may cause the 3-step ML approach to produce biased estimates (shown in simulation studies described in Section 4.5.3). As true class memberships are latent, this assumption of within-class normality is empirically untestable. For a binary outcome, however, the key assumption of the model for the 3-step ML approach is the validity of the logit-linear specification. As this is a fundamental assumption that is related to the functional form of the model, compared to the normality assumption for the continuous outcome, it is more likely to hold as one often carefully discusses the choice of functional specifications of a model. Also included in the model are the control variables: a binary indicator of overweight at age 16 (based on the WHO cut-off of BMI ($kg/m^2 > 25.0$)) and gender. The model is fitted using the 3-step ML method where the approach discussed in

Section 4.4 is extended to include more than two categorical latent variables. Results for a comparison of 3-step ML, modal class and 1-step approaches are summarised in Table 4.10.

The estimation approaches yield different results. In particular, the 3-step ML approach finds that low levels of social class is significantly (at the 5% level) associated with poor midlife health. The 1-step approach finds this association is stronger but only significant at the 10% level. Also, on the effect of high levels of financial difficulty on midlife health, the 3-step approach finds a significant association but the 1-step approach finds the association non-significant. Results of the modal class approach are mostly similar to the 3-step approach. One exception is that the modal class approach finds family structure significantly associated with poor health but the 3-step approach find the association non-significant. We also find that in general, standard errors in the modal class approach are smaller than those from the 3-step and 1-step approaches. This is mainly because in the modal class approach, the categorical latent variables are treated as known in the regression model for H, hence the uncertainty is underestimated. As the results from the 3-step approach and the 1-step approach are not the same, we expect some degree of violation of model assumptions (an implication of the simulation studies described in Section 4.5.3). For a binary distal outcome, this could be the presence of the conditional dependence between the indicators for the categorical latent variables and the distal health given the class membership. Future analyses should take this into account.

Table 4.10 Comparison of effects (in log-odds) of four dimensions of childhood SECs on midlife health using the 3-step ML, modal class and 1-step approaches. Standard errors (SE) are reported in brackets.

Predictors	3-step ML		Modal class		1-step	
	Est.	(SE)	Est.	(SE)	Est.	(SE)
Intercept	−2.36**	(0.09)	−2.36**	(0.08)	−2.36**	(0.10)
Male (ref.=female)	−0.05	(0.06)	−0.04	(0.06)	−0.05	(0.06)
Overweight at age 16[a] (ref.= No)	0.25**	(0.07)	0.26**	(0.07)	0.24**	(0.07)
Childhood circumstances						
Social class [b] (ref.=High)						
Low	0.40**	(0.19)	0.45**	(0.11)	0.66*	(0.19)
Medium	0.32**	(0.11)	0.31**	(0.09)	0.36**	(0.10)
Financial difficulty (ref.=Low)						
High	0.53**	(0.21)	0.47**	(0.09)	0.15	(0.39)
Material hardship (ref.=Low)						
Medium	0.33**	(0.11)	0.31**	(0.08)	0.32**	(0.09)
High	0.35**	(0.12)	0.41**	(0.09)	0.41**	(0.11)
Family structure (ref.=Stable)						
Unstable	0.08	(0.13)	0.21**	(0.11)	0.12	(0.13)

** $p < 0.05$, * $p < 0.1$

[a] Binary indicator for overweight at age 16

[b] Father or male head social class

In addition to the above findings, the extended 3-step ML approach also finds significant associations between the four latent summaries of childhood socioeconomic circumstances, except for the medium-level social class and high-level financial difficulty. This relationship is represented by the log-odds ratios in Table 4.11. An example of how to interpret the numbers is provided as follows. In the first three rows, the log-odds ratios indicate that compared to those children with fathers in the high social class, children with fathers in the low social class have a higher tendency to experience adverse financial and housing situations.

Table 4.11 Estimated associations between four latent aspects of socioeconomic circumstances in childhood using the 3-step ML approach. Standard errors (SE) are reported in brackets.

Association		Log-odds ratio	SE
Low social class	High financial difficulty	1.62**	0.80
Low social class	Medium material hardship	2.49**	0.14
Low social class	High material hardship	3.88**	0.26
Low social class	Unstable family structure	−3.02**	0.80
Medium social class	High financial difficulty	0.45	0.48
Medium social class	Medium material hardship	1.36**	0.06
Medium social class	High material hardship	3.31**	0.21
Medium social class	Unstable family structure	0.27**	0.12
High financial difficulty	Medium material hardship	0.64**	0.25
High financial difficulty	High material hardship	3.12**	0.25
High financial difficulty	Unstable family structure	4.77**	0.80
Medium material hardship	Unstable family structure	0.49**	0.11
High material hardship	Unstable family structure	−0.28**	0.14

$^{**}p < 0.05$, $^{*}p < 0.1$

Substantively, we conclude that controlling for early health (overweight at age 16) and gender, individuals from families with material disadvantages (e.g. father from low levels of social class) are significantly more likely to be in a poor health state at age 50. The effect of unstable family structure, however, is non-significant. This finding invokes two related questions that require further investigation. First, would the significant effects of childhood variables on midlife health disappear or persist when the information in adulthood, in particular, life events (over 34 years of follow-up) is considered? Second, for the non-significant childhood variables, are the effects (if there are any) truly non-significant or are they influenced by life events (e.g. partnership experiences) in adulthood? Recalling the mediation framework proposed in Figure 1.1 of Chapter 1, the non-significance may be a result of combining effects (e.g. conditional on the experiences of life events) of opposite directions, a dominant non-significant direct or indirect effect or both. In the next chapter, we extend the 3-step ML approach to models that relate multiple categorical latent variables to life events. We use partnership transitions across 34 years of follow-up as an example.

Chapter 5

Event history analysis of the effects of childhood socioeconomic circumstances on partnership formation and dissolution

5.1 Introduction

In this chapter, we consider partnership transitions as an example of the life events mentioned in Section 1.2, that are hypothesized to mediate the effects of childhood socioeconomic circumstances (SEC, captured by four categorical latent variables) on midlife health. To model this mediation structure, we first need to extend the methodology proposed in Chapter 4 by relating multiple categorical latent variables to partnership outcomes (i.e. time-to-event data). Event history analysis (EHA) is often used to model time-to-event data to identify potential risk factors for the timing of events. Typically, event histories are derived from information about the timing of the events of interest over the life course. Examples include partnership, employment and fertility histories where the first (i.e. partnership history) is our main interest with the event being entry into partnership or separation. It is also common for individuals to experience such events more than once over the study period, leading to a multilevel structure. Event history data are often collected retrospectively in cohort or panel studies (e.g. the 1958 NCDS and the UK Household Longitudinal Study in Britain). In the 1958 NCDS, cohort members are asked to recall the timing of events to the nearest year and month. Specifically, they were asked at age 33 to give information about all previous partnerships since age 16, including the start and end dates, the type of partnership and many other details; similar information was collected at ages 42, 46 and 50. More details are available in Section 3.4.

After an overview of the methodology in Section 5.2, research objectives are stated in Section 5.3. This is followed by Section 5.4 that reviews previous studies exploring the relationship of childhood SECs and partnership events using EHA. In Section 5.5, the general 3-step maximum likelihood (ML) approach proposed in Chapter 4 is extended for the analysis of multilevel event history data. Simulation studies are conducted in Section 5.6 to evaluate the performance of the extended approach. Finally, Section 5.7 discusses an application of the 3-step ML method to estimate the effects of four latent summaries of childhood SEC on the risk of partnership formation and dissolution.

5.2 Review of event history analysis

In this review, the following key terms are used throughout. First, the response variable in the event history analysis is the duration until the occurrence of an event. In the presence of repeated events (e.g. partnership dissolutions), an individual may have multiple episodes where an episode is defined as the period starting from the time when an individual becomes at risk of experiencing an event (enters the risk set) to the time when an event or censoring occurs. Specifically, in the analysis of time to partnership dissolution, an individual enters the risk set at the start of a partnership. For individuals who have experienced separation more than once in the observation period, duration of a partnership is re-calculated when a new episode starts. A detailed discussion of the modelling techniques of event history data is available in Hosmer and Lemeshow (1999) and Steele et al. (2005). The following sections set out the key notation and concepts relevant to our study.

5.2.1 Challenges of using event history data

There are two main challenges when dealing with event history data from the birth cohort study. The first issue is that the time-to-event data are not fully observed as not all individuals have experienced the event of interest by the end of the observation period. These individuals' event times are right-censored (incomplete), which is a common form of censoring in social and health research. A simple treatment of such data would be to drop censored observations from the analysis. Apart from reducing the sample size, this approach can lead to biased estimates as the excluded sample consists mostly of individuals with a low risk of experiencing the event, and therefore long durations (Steele et al., 2005). The methods discussed in Section 5.2.3 and Section 5.2.4 can handle right-censored data. One common assumption is that the censoring time is independent of the event time. This assumption is sensible in studies where the censoring time depends on the interview time, which is not related to the event time. Another challenge of dealing with event history data is the need to incorporate time-varying covariates. For example, in the case of partnership histories, we may be interested in the effect of the number of pre-school children at time t on the risk of marriage dissolution at that time.

5.2.2 Notation and definitions

Denote by T a random variable for the event time. An important quantity in EHA is the hazard function which is defined as

$$h(t) = \lim_{\delta t \to 0} \frac{p(t \leq T < t + \delta t | T \geq t)}{\delta t},$$

where $h(t)$ represents the hazard function (or rate) of having an event within the time interval $(t, t + \delta t)$, conditional on not having experienced the event before time t. The other two related functions that are also important in EHA are $S(t)$, the survival function, which is the probability of not having experienced the event before t, and $F(t)$, the cumulative distribution function, which is the probability of experiencing the event before t. They are related quantities as

$$\begin{aligned} S(t) &= P(T \geq t), \\ F(t) &= 1 - S(t) \\ &= P(T < t). \end{aligned}$$

To examine the distribution of event times and obtain non-parametric estimates of the above functions, a life table can be constructed, after grouping duration into intervals $[t, t+1)$. Estimates of the hazard, survival and cumulative distribution function are

$$\begin{aligned} \hat{h}(t) &= \frac{\text{No. of events during}[t, t+1)}{\text{No. of people in the risk set} - 0.5 \times \text{No. of censored cases during}(t, t+1)}, \\ \hat{S}(t) &= \prod_{v=1}^{t-1} \left(1 - \hat{h}(v)\right), \\ \hat{F}(t) &= 1 - \hat{S}(t), \end{aligned}$$

where the denominator of $\hat{h(t)}$ is adjusted for $0.5\times$ the number of censored cases under the assumption of uniformly distributed censoring times within $[t, t+1)$.

5.2.3 Continuous-time event history model

Event times may be treated as either continuous or discrete. In continuous time a general proportional hazard (PH) model takes the form

$$h(t|\mathbf{X}) = h_0(t)f(\mathbf{X}), \tag{5.1}$$

where $f(\mathbf{X})$ denotes a function of a vector of time-invariant covariates $\mathbf{X}$ and $h_0(t)$ denotes the baseline hazard rate, which is the hazard when $\mathbf{X}$ takes value $\mathbf{0}$. The PH assumption implied by (5.1) is that the effects of $\mathbf{X}$ on the hazard are constant over time. If the form of the baseline hazard function is specified, subject to assumptions about the distribution of event times, (5.1) becomes a parametric model. For example, the assumption of exponentially distributed event times implies a constant hazard rate and a Weibull or Gamma distribution implies a monotonically increasing or decreasing hazard rate (see Allison (1984) for more discussions).

Another approach that has been widely used in applied research is the semi-parametric Cox proportional hazard model (Cox, 1972). In the Cox model, (5.1) becomes

$$h(t|\mathbf{X}) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{X}), \tag{5.2}$$

where $\boldsymbol{\beta}$ denotes the coefficient vector. It is semi-parametric as it does not require any assumptions about the distribution of event times and therefore does not require specification of $h_0(t)$. However, it assumes multiplicative effects of $\mathbf{X}$ on the hazard ratio. The assumption that covariate effects are constant over time can be relaxed and time-varying covariates can be incorporated by expanding each duration record, assuming the time-varying covariates are constant within each time interval.

5.2.4 Discrete-time event history model

In social surveys, such as the NCDS, details of previous partnerships are usually collected retrospectively and respondents are asked to recall the timing of events to the nearest year and month, resulting in discretely measured durations. It is therefore more natural to use the discrete-time approach to model interval-censored data. There are three major advantages of using a discrete-time over a continuous-time setting. First, the inclusion of time-varying covariates is straightforward in a discrete-time model. Second, the response variable in the discrete-time event history model is the binary event indicator so that standard methods for binary data may be used. Third, the traditional continuous-time approach assumes tied events (cooccurence of events) are not present, which can be inappropriate when the duration is measured in broad time units such as months (e.g. Singer and Willett, 2003; Steele, 2011; Steele et al., 2005). Considering all these advantages and its straightforwardness to allow for and test for non-proportional hazards, this thesis will focus on discrete-time EHA.

Before performing the analysis, duration data for a non-repeatable event need to be restructured (expanded) to a person-period file by splitting the duration into multiple time intervals. For each time interval t, let y_i denote the event or censored time for an individual i ($i = 1,...,N$). Let δ_i denote the censoring indicator, which takes value 1 if y_i is right-censored, and 0 if y_i is fully observed. In addition, we define a binary response variable y_{ti} such that

$$y_{ti} = \begin{cases} 1, & y_i = t, \delta_i = 0 \\ 0, & y_i = t, \delta_i = 1 \\ 0, & y_i > t. \end{cases}$$

A logit model can be formulated, for example, to model the binary y_{ti},

$$\text{logit}\,(h_{ti}) = \alpha_t + \boldsymbol{\beta}'\mathbf{X}_{ti}, \tag{5.3}$$

where $h_{ti} = P(y_{ti} = 1 | y_{t'<t,i} = 0)$, α_t denotes the logit baseline hazard function, $\mathbf{X}_{ti}$ denotes a vector of time-varying and time-invariant covariates and $\boldsymbol{\beta}$ is a vector of the corresponding effects. Alternatively, other link functions, such as the complementary log-log are also commonly used in the literature (Hedeker et al., 2000). For the specification of α_t, a piecewise function (different α_t for different intervals) and polynomials of t are the most commonly considered. Although the covariate effects are assumed constant over time in (5.3), non-proportional effects can be easily allowed by including the interaction of α_t and $\mathbf{X}_{ti}$ in the model. Because of its flexibility, the discrete-time approach has been widely applied in previous research (e.g. Reardon et al., 2002; Steele et al., 2005, 2006).

To use the discrete-time method, one needs to restructure the data, which can be computationally inefficient when the time interval is narrow and the observation period is long. Alternatively, small time intervals can be grouped together (e.g. monthly data into quarterly or yearly data) if the assumption of a constant covariate value and hazard rate within the grouped time interval can be justified. When durations are grouped, the exposure time in each grouped interval needs to be considered. For example, suppose an individual i experiences an event in the 15^{th} month. Instead of creating 15 monthly records for this individual, the event time can be grouped into 12-month intervals. This individual then contributes 2 records to the analysis file. In the first record, the individual is exposed to 12 months with $y_{ti} = 0$ and in the second record, the individual is exposed to 3 months with $y_{ti} = 1$. y_{ti} can now be considered as a response variable that follows a binomial distribution with denominator n_{ti} equal to the exposure time (Steele, 2011; Steele et al., 2004).

5.2.5 Modelling recurrent events

So far we have discussed the model for non-repeatable events (e.g. first entry into partnership). To handle duration data for recurrent events (e.g. partnership dissolutions), one approach is to model the duration of each episode separately. This may lead to biased estimates as the event times for each episode share the same subject-specific influences that may not be fully captured by existing covariates (Liu et al., 2004).

A more appropriate approach is to jointly model all repeated episodes. The unobserved individual-level time-invariant effect should be included in this joint model as it influences the hazard of experiencing the event in all episodes. Such unobserved variables can be accounted for by including an individual-level random variable u_i. Another benefit of joint modelling is that we can evaluate whether effects of some covariates on the hazard of an event persist in all episodes. This can be achieved by interacting dummies of the order of the event with these covariates in the model.

In terms of the formulation of a joint model, we can consider the repeated events being nested within each individual, which leads to a two-level model with episodes (level 1) nested within individuals (level 2). Such models can be estimated using methods for multilevel binary responses, which are now implemented in all mainstream statistics packages.

For recurrent events, denote by y_{ij} the event or censored time, where $j\,(j=1,\ldots,J_i)$ indexes the episode. We define a corresponding binary response variable y_{tij} such that

$$y_{tij} = \begin{cases} 1 & y_{ij} = t, \text{uncensored} \\ 0 & y_{ij} = t, \text{censored} \\ 0 & y_{ij} > t. \end{cases}$$

Denote by $h_{tij} = P(y_{tij} = 1 | y_{t'<t,ij} = 0)$ the hazard of separation in time interval $[t, t+1)$ of episode j. A discrete-time model for recurrent events (e.g. partnership dissolutions) can be formulated as

$$\text{logit}\left(h_{tij}\right) = \alpha_t + \boldsymbol{\beta}'\mathbf{X}_{tij} + u_i, \tag{5.4}$$

where $u_i \sim N(0, \sigma_u^2)$ and $\mathbf{X}_{tij}$ denotes a vector of time-invariant and time-varying covariates whose effects can vary across episodes.

Similar to the modelling approach for non-repeatable events, given the long observation period, the monthly duration data are aggregated to reduce the size of the discrete-time dataset. The likelihood function for this discrete-time model for recurrent events is equivalent to that for a model for multilevel binary data where y_{tij} follows a binomial distribution with denominator n_{tij} equal to the exposure time in each aggregated time interval, respectively (Steele et al., 2005).

With the random effect u_i (also termed shared frailty in Vaupel et al. (1979)) in the model, the $\boldsymbol{\beta}$s need to be interpreted with caution. For instance, suppose we have a single binary covariate, in a model without random effects. In this model $\exp(\boldsymbol{\beta})$ is the population-average odds ratio contrasting the odds of experiencing an event at time t for the two categories of X. In the model with shared frailty, however, $\exp(\boldsymbol{\beta})$ is the individual-specific odds ratio, or a comparison of odds for individuals with the same value of u_i. The predicted odds with $u_i = 0$ is not the population-average odds, but the median odds, because of the non-linearity of the logistic function. To estimate the population-average odds, u_i needs to be integrated out. Detailed discussions of population-average and subject-specific effects are available in, for example, Hu et al. (1998), Zeger et al. (1988) and Hardin and Hilbe (2002).

Note that to account for the effect of individual-level unobservables, u_i can also be included in the model (5.3) but it may not be identified without repeated events (multiple episodes for an individual).

5.2.6 Modelling correlated processes

The previous sections have discussed methods to model one event history (with a single event or repeated events) in the observation period. Event histories are, however, often associated with one another. For instance, the presence and age of children (outcomes of the fertility history) have been shown to influence the hazard of marriage dissolution (from the partnership history) (e.g. Lillard, 1993; Steele et al., 2005). Other examples arise when considering the potential influence of partnership formations (e.g. age at the start of each parntership) on partnership stability. A simple approach to model such relationships is to include the outcome of one process as a covariate in a model for another process. This may lead to model misspecification and hence biased estimates if this covariate and the outcome in the model have shared unobserved risk factors. In general, it is not possible to control for all shared influences, some of which can be unobserved individual-specific characteristics. Consider model (5.4) in Section 5.2.5, where $\mathbf{X}_{tij}$ contains outcomes from another correlated process. In this case, some of the covariates in $\mathbf{X}_{tij}$ may be endogenous, i.e. correlated with the random effects u_i. Such endogeneity can lead to biased estimates (Steele, 2011; Steele et al., 2006). Appendix C describes a simulation study that shows the magnitude and direction of bias in the presence of endogenous predictors. This accompanies a theoretical investigation provided in Appendix B.

To handle endogeneity, a joint model of associated processes can be employed (Steele et al., 2005). To illustrate this for the discrete-time model, suppose we want to explore the relationship between partnership formations (process (1)) and dissolutions (process (2)), where time intervals for each process are indexed by t_1 $(t_1 = 1, \dots, T_{1ij})$ and t_2 $(t_2 = 1, \dots, T_{2ij})$, respectively. We can stack response variables into a vector $(y^{(1)}_{t_1ij}, y^{(2)}_{t_2ij})$, where in this example, $y^{(1)}_{t_1ij}$ is a binary indicator of union formation and $y^{(2)}_{t_2ij}$ is a binary indicator of partnership separation. Note that these two variables are defined following Section 5.2.5. The joint model can be formulated as

$$\begin{aligned}
\operatorname{logit}\left(h^{(1)}_{t_1ij}\right) &= \alpha^{(1)}_t + \boldsymbol{\beta}^{(1)'}_1 \mathbf{X}^{(1)}_{t_1ij} + \boldsymbol{\beta}^{(1)'}_2 \mathbf{V}^{(2)}_{t_1ij} + u^{(1)}_i, \\
\operatorname{logit}\left(h^{(2)}_{t_2ij}\right) &= \alpha^{(2)}_t + \boldsymbol{\beta}^{(2)'}_1 \mathbf{X}^{(2)}_{t_2ij} + \boldsymbol{\beta}^{(2)'}_2 \mathbf{V}^{(1)}_{t_2ij} + u^{(2)}_i,
\end{aligned} \tag{5.5}$$

where $h^{(1)}_{t_1ij}$ and $h^{(2)}_{t_2ij}$ denote the respective hazard functions for each process, $\mathbf{X}^{(1)}_{t_1ij}$ and $\mathbf{X}^{(2)}_{t_2ij}$ denote the vectors of predictors for each process and $\mathbf{V}^{(1)}_{t_1ij}$ and $\mathbf{V}^{(2)}_{t_2ij}$ denote the vectors of prior outcomes of the two processes by time t_1 and t_2, respectively. In model (5.5), the effect of outcomes of the partnership formation history (e.g. the age at the start of partnership) due to the start of time interval t_2 on the risk of separation at that time is represented by $\boldsymbol{\beta}^{(2)}_2$; similarly, the effect of outcomes of the marriage history up to the start of time interval t_1 (e.g. marital status, or the number of previous partners) on the risk of forming a new partnership is represented by $\boldsymbol{\beta}^{(1)}_2$. Note that to fix concepts, we illustrate the modelling approach for the formation and dissolution processes, that are partnership events. In social research, other life events, such as those in employment and fertility histories can also be associated.

The two equations in model (5.5) are estimated simultaneously by allowing for correlation between random effects. It is common to assume normally distributed individual-specific random effects, where $(u^{(1)}_i, u^{(2)}_i) \sim N(0, \Omega)$ and Ω is a random effect covariance matrix with variance $\sigma^{2(1)}$, $\sigma^{2(2)}$ for each process and $\sigma^{2(12)}$ for the covariance. For instance, in our example of the partnership formation and dissolution processes, a positive estimate of $\sigma^{(12)}$ indicates that those with an above-average risk of entering into a partnership also have a higher risk of separation. More specifically, suppose there exists a true positive influence of the early union formation on the risk of dissolutions. If the joint model finds a positive correlation between the random effects of the two processes, it suggests that if we ignore the shared unmeasured influences on the two processes by fitting a single-process model only, we will overestimate the positive effect of union formation (or previous unions) on dissolutions. This is because in the "formed a partnership" category, there is an over-representation of people who have a higher risk of separation, which inflates the risk of separation for people in the "formed a partnership" status. More details on the multi-process model and its advantages over the single-process model are available in Lillard (1993) and Steele et al. (2005).

Turning to estimation, model (5.5) is essentially a multilevel bivariate response model for the stacked column of binary responses $(y^{(1)}_{t_1ij}, y^{(2)}_{t_2ij})$. We next define dummies s_1 and s_2 for each process. The corresponding covariates included in each equation are then multiplied by s_1 and s_2. In the random part of the model, we fit individual-level random effects to s_1 and s_2, allowing for an unstructured correlation. Ideally, to handle the potential endogenous predictors in $\mathbf{V}^{(1)}_{t_1ij}$ and $\mathbf{V}^{(2)}_{t_2ij}$, the instrumental variable approach should be explored. Instrument variables are highly associated with the endogenous variables but not with the residuals; but for our study, they can be difficult to find. Fortunately for us, data are clustered in a hierarchical structure and the existence of repeated events may assist identification of the random effects parameters (Lillard, 1993; Liu et al., 2004).

The joint modelling approach has been widely applied in social research in the last decade. For example, Lillard (1993) used a simultaneous equation modelling approach to explore the relationship between the timing of marital dissolution and child-bearing within marriage. Aassve et al. (2006b) estimated models for birth events, union formation and dissolution, as well as employment and non-employment events simultaneously. A non-zero covariance between the individual-specific random effects in these processes was allowed to account for the association between the timing of these life events. Building upon these findings, Goldstein et al. (2004) jointly modelled the transition between two partnership status (in and out of partnership). To account for the shared unobserved influences in two processes, Steele et al. (2005, 2006) used the simultaneous equation modelling approach to jointly model the timing of partnership transitions and child-bearing.

5.3 Research objective

In this chapter, we aim to address RQ_2 (see Section 1.2): How do childhood circumstances influence partnership transitions between ages 16 and 50 ? This is one pathway in the full structural equation model that helps to understand whether the influence of childhood circumstances on health in later life can be partially explained by their own social situations over the life course. In the next chapter, we will extend the model to allow for effects of both childhood circumstances and partnership experiences on heath at age 50. Keeping the potential endogeneity issue in mind, to answer RQ_2 we fit models for the time to the formation of first co-residential partnership and for the time to partnership dissolution simultaneously. Such a joint model is necessary as in the dissolution model, the age at the start of partnership is an endogenous covariate partly determined by the outcome variable in the formation model (i.e. age at the first partnership). We can also infer the effects of childhood SECs on partnership formation, in addition to that on partnership dissolution from fitting this joint model.

RQ_2 can be further broken into the following sub-questions that need to be addressed:

1) How are different aspects of childhood SECs associated with the timing of entry into the first partnership?

2) How are different aspects of childhood SECs associated with the risk of partnership dissolution?

3) Are there shared unobserved characteristics influencing the timing of first partnership formation and recurrent dissolutions? (e.g. do cohort members whose unobserved characteristics place them at a higher risk to start their first partnership early also tend to separate from their partners early?) If yes, what are the impacts on the estimated effects of childhood SECs on partnership dissolutions?

Endogeneity and unobserved heterogeneity

From this section onwards, the concepts of endogeneity and unobserved heterogeneity are often mentioned. Both concepts are commonly discussed in the literature in sociology, economics, epidemiology and statistics but can sometimes cause confusion due to their close relationship. We give a brief discussion of their relationships below. Detailed discussions on these topics are available in Wooldridge (2010, Chapters 15,16) and Elwert and Winship (2014).

Endogeneity is a broad term that mainly refers to the problem caused by explanatory variables being correlated with the error term in a regression model, and hence violating the assumption of standard regression models. There are three main sources of endogeneity: model misspecification or omitted variables, measurement error and simultaneity (i.e. reverse causality). These sources are not perfectly distinct. For example, the problem of measurement error could be framed as an omitted variable problem and omitted variables that are correlated with a predictor in a regression model are commonly referred to as unobserved confounders.

Unobserved heterogeneity simply refers to variation between individuals in the outcome of interest that is due to omitted individual-specific characteristics. In this thesis, unobserved heterogeneity is captured in statistical models by an individual-level random effect. If the outcome and any of its observed predictors have shared unmeasured time-invariant influences, we have an endogeneity problem because those predictors will be correlated with the random effect.

5.4 Review of the application of EHA to explore the relationship between childhood SEC and partnership histories

There is a substantial literature that aims to identify childhood risk factors of entry into and moving out of co-residential partnership. It has been found in multiple studies that family characteristics during childhood have a significant impact on the timing of such events.

Among the family characteristics that are considered as predictors of partnership events, results on the effects of parental socioeconomic situations are mixed. For the event of entering into partnership, it has been found that those from families of relatively high social position tend to delay first partnership formation (e.g. Berrington and Diamond (2000) and Berrington (2003) using the 1958 NCDS data up to age 33; Wiik (2009) using the Norwegian family survey). However, such effects have been found in other studies to be non-significant for partnership formation (Aassve et al., 2006a). In the case of partnership dissolution, some British studies showed that people from families with more advantaged social backgrounds are prone to early separation (Kiernan and Cherlin, 1999; Steele et al., 2005, 2006) while other studies in the US and Norway found the opposite (Bumpass et al., 1991; Lyngstad, 2006). In addition, parental occupation has been shown by some researchers to have no significant impact on the tendency to separate among married women (Aassve et al., 2006a; Steele et al., 2005, 2006).

In previous studies, family financial situation has also been commonly used as an indicator of parental SEC. For instance, Wiik (2009) found a negative effect of family wealth on the risk of early marriage in Norway. However, this factor does not seem to influence the risk of moving out of partnership (Bumpass et al., 1991; Kiernan and Cherlin, 1999; Lyngstad, 2006).

Other factors during childhood that have been supported by previous studies to be influences on partnership decisions in adulthood include the living conditions and parental education level. For instance, Berrington and Diamond (2000) reported that women living in public-rented housing were prone to start cohabitation early. Wiik (2009) concluded that children from families with well-educated parents tended to start their partnership (either cohabitation of marriage) late. However, results on the effects of parental education level on the risk of partnership dissolution are mixed. For instance, Bumpass et al. (1991) found that mother's education level did not affect the stability of their child's first marriage in the US, while Lyngstad (2006) found a high level of parental education tended to lower children's risk of marriage dissolution in Norway.

In addition to these results, experiences of parental relationship breakdown during childhood have also been shown in various studies to be associated with partnership events. Researchers found that in Britain children who grew up in divorced families tended to enter into partnership early (Aassve et al., 2006a; Berrington and Diamond, 2000; Steele et al., 2006). Wiik (2009) also found that parental divorce lowered the risk of entering into marriage, but boosted the risk of forming cohabitation for the first time among Norwegians. Regarding the event of moving out of partnership, results from previous research in the US, Britain and Finland are consistent, concluding that the experience of family breakdown during childhood is associated with higher risk of partnership dissolution (e.g. Amato, 1996; Kiernan and Cherlin, 1999; Steele et al., 2006).

Although the association between childhood socioeconomic background and the timing of partnership formation and dissolution have been widely recognised in the literature, the pathways that underlie such associations are not yet clear. Potential mediators have been suggested in the literature, such as the age at the start of first partnership and the type of relationship (cohabitation or marriage). However, these mediation pathways have not been explicitly tested in a structural equation model. Previous studies have suggested that respondents from rich families tend to have better education and hence a delayed entry into their first partnership (Blossfeld and Huinink, 1991). It has been shown that the experience of parental separation in childhood is related to a higher chance of early cohabitation in the first partnership (Cherlin et al., 1995). For example, a positive influence of parental divorce on their children's risk of separation has been shown to be partly mediated by the age at and type of first partnership from studies in the US and Norway (Amato, 1996; Bumpass et al., 1991; Kiernan and Cherlin, 1999), where the direct effect still remains significant. By controlling for more indicators of the family background, such as housing tenure and parental social class, a similar conclusion was drawn by Berrington and Diamond (2000) that the family demographic influence on children's relationship stability is significant for early partnership formation (between ages 16 and 23). For respondents who start their first partnership late (between ages 23 and 32), time-invariant parental effects are largely accounted for by respondents' own characteristics such as age and educational level at the start of partnership. In a recent study using NCDS data up to age 42, Ploubidis et al. (2015) examined the direct effects of childhood situations on partnership transitions between ages 23 and 42, as well as the indirect effects mediated by health at age 23. Although their main interest was the effects of patterns of partnership transitions on midlife health, a significant selection effect into partnership due to childhood situations and health in early life (age 23) was found.

In this chapter, our main focus is on using EHA to explore the effects of childhood circumstances, which were not the key measures of interest in previous studies, on the timing of partnership formation and dissolution. Instead of using only one or two simple childhood indicators, summaries of the SECs during childhood (see Chapter 4) that consider the longitudinal information available in childhood waves, are included in the model as latent predictors.

5.5 An extension of the general 3-step ML approach to multilevel event history models

After discretisation, the response variable that is created from the time to first partnership formation is simply a single-level binary variable. However, to estimate the effects of childhood SECs on the tendency to dissolve a partnership, the proposed general 3-step ML approach of Chapter 4 needs to be extended to consider recurrent separation events. This leads to a multilevel binary response that can be fitted by model (5.4) in step 3, where durations of episodes from the same individual may be dependent. As episodes are nested within individuals, this implies that individual-level characteristics, including the latent summaries of childhood circumstances, are level-2 variables. Using the 3-step ML approach, we first estimate separate latent class models for each categorical latent predictor to obtain modal classes and the misclassification probabilities (as in Chapter 4). We then focus on developing the specification and estimation strategies for the multilevel event history model in step 3.

5.5.1 Model in step 3

We now describe a random effects discrete-time event history model where durations are predicted by multiple and possibly associated categorical latent variables. The proposition under the latent variable modelling framework is inspired by early work of Muthén and Masyn (2005) that proposed a latent class model for non-repeatable discrete-time event history outcomes where the latent classes account for unobserved heterogeneity, and Masyn (2009) who models the low-frequency repeatable event times by allowing for a subject-specific unobservable to account for dependency between durations within an individual. However, neither study considered categorical latent predictors. The model we propose can model high-frequency repeatable event times of a heterogeneous population, allows for categorical latent predictors and corrects for potential misclassification in the modal class estimation. This model can also be extended to handle more complex modelling interests.

For example, one may want to relate the categorical latent predictors and the event histories to multivariate distal outcomes. As an illustration, building on the model (4.25) proposed in Chapter 4 for a single distal health outcome, we extend (5.4) to a simplification of a model with two categorical latent predictors C_{1i} and C_{2i}, i.e.

$$\operatorname{logit}\left(h_{tij}\right) = \alpha_t + \boldsymbol{\beta}'\mathbf{X}_{tij} + \sum_{q=1}^{2}\sum_{k_q=1}^{K_q-1} \tau_{C_q,k_q} I(C_{qi} = k_q) + u_i, \tag{5.6}$$

where τ_{C_1,k_1} and τ_{C_2,k_2} are the coefficients of dummy variables for C_1 and C_2 respectively, with the last category of each taken as the reference, i.e. $\tau_{C_q,K_q}(q = 1,2) = 0$. This is essentially a mixture model with two latent class variables measured by survival outcomes (Lin et al., 2002; Masyn, 2009; Muthén and Masyn, 2005). The joint distribution of C_{1i} and C_{2i} is specified by a log-linear model, as proposed in (4.21) of Chapter 4. Fitting a binary response model with a logit link to $y_{tij}\,(t = 1,2,...,T_{ij})$, the conditional density function for an individual i can be derived under the assumption that conditional on u_i, durations of episodes from the same individual are independent. We write:

$$f\left(\mathbf{y}_i|\mathbf{X}_i, C_{1i}, C_{2i}, u_i\right) = \prod_{t=1}^{T_i}\prod_{j=1}^{J_i} h_{tij}^{y_{tij}}(1-h_{tij})^{(1-y_{tij})}, \tag{5.7}$$

where the vector $\mathbf{y}_i$ denotes the response vector across episodes and time intervals for each individual i and $\mathbf{X}_i$ denotes the corresponding vector of observed predictors.

5.5.2 Estimation in step 3

Estimation of parameters in latent class models (steps 1 and 2) has been discussed in detail in Section 4.2.1. We now focus on estimating parameters in step 3 of the model where categorical latent variables are linked to a discrete-time event history outcome and the modal classes and misclassification probabilities have been derived from steps 1 and 2.

Following the methodology proposed in Section 4.3.4 where the misclassification probability (4.14) and the complete data log-likelihood for an outcome H (4.26) are derived, we can extend the likelihood function to allow for heterogeneity in event times. Let $\boldsymbol{\xi}_i = (u_i, C_{1i} = k_1, C_{2i} = k_2)$ be a vector of all latent variables (both continuous and discrete). If they are all observed, the complete data log-likelihood is written as:

$$
\begin{aligned}
l &= \sum_{i=1}^{N} \log f(\mathbf{y}_i, M_{1i}, M_{2i}, \boldsymbol{\xi}_i | \mathbf{X}_i) \\
&= \sum_{i=1}^{N} \left[\log f_1(\mathbf{y}_i, M_{1i}, M_{2i} | \mathbf{X}_i, \boldsymbol{\xi}_i) + \log \Psi(\boldsymbol{\xi}_i) \right],
\end{aligned} \tag{5.8}
$$

where $\Psi(\boldsymbol{\xi}_i)$ denotes the joint distribution of the latent variables, $\mathbf{X}_i$ is a vector of all observed covariates for individual i and (M_{1i}, M_{2i}) are the modal classes derived in step 1 (further details are given in Section 4.3.4).

Assuming that $C_1 \perp\!\!\!\perp M_2|C_2$, $C_2 \perp\!\!\!\perp M_1|C_1$ and the conditional independence of manifest variables ($\mathbf{Y}_i^{(C)}$) and the distal outcome ($\mathbf{y}_i$) given the latent variables, we have

$$
f_1(\mathbf{y}_i, M_{1i}, M_{2i} | \mathbf{X}_i, \boldsymbol{\xi}_i) = f(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\xi}_i) P(M_{1i}|C_{1i}) P(M_{2i}|C_{2i}), \tag{5.9}
$$

where $P(M_{1i}|C_{1i})$ and $P(M_{2i}|C_{2i})$ are known quantities computed in step 2.

In this model, we assume predictors $\mathbf{X}_i$ are independent with $\boldsymbol{\xi}_i$. It is analogous to a common assumption to avoid endogeneity in the econometrics literature (discussed in Wooldridge (2010, Chapters 15,16)). If we also assume u_i is independent of (C_{1i}, C_{2i}), (5.8) can be decomposed into five terms after substitution of (5.9) to give

$$
l = \sum_{i=1}^{N} \left[\log f(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\xi}_i) + \log P(M_{1i}|C_{1i}) + \log P(M_{2i}|C_{2i}) + \log P(C_{1i}, C_{2i}) + \log \phi(u_i) \right], \tag{5.10}
$$

where $\phi(\cdot)$ is the standard normal density.

As the conditional distribution of $\mathbf{y}_i$ depends on the latent variables $\boldsymbol{\xi}_i$, (5.10) can be maximised using the EM algorithm (Dempster et al., 1977). In the E-step, the expected score function is computed where the expectation is taken with respect to the posterior distribution of $\boldsymbol{\xi}_i$ given all observed data, i.e. $g(\boldsymbol{\xi}_i | \mathbf{y}_i, \mathbf{X}_i, M_{1i}, M_{2i})$. In the M-step, parameter updates are obtained by applying a root-finding algorithm to the function in the E-step. Following the derivation procedure set out in McLachlan and Krishnan (2007), we describe the estimation procedure in more details.

1) Estimation of ω parameters in the log-linear model
Denote by $\boldsymbol{\theta}_1 = (\omega_0, \omega_{k_1}^{(C_1)}, \omega_{k_2}^{(C_2)}, \omega_{k_1 k_2}^{(C_1 C_2)})$, the set of all ω parameters. The individual contribution to the expected score function of $\boldsymbol{\theta}_1$ is:

$$
E\left[S_i(\boldsymbol{\theta}_1)\right] = \sum_{k_2=1}^{K_2-1} \sum_{k_1=1}^{K_1-1} \int_{u_i} S_i(\boldsymbol{\theta}_1) g(\boldsymbol{\xi}_i | \mathbf{y}_i, \mathbf{X}_i, M_{1i}, M_{2i}) du_i, \tag{5.11}
$$

where

$$S_i(\boldsymbol{\theta}_1) = \frac{\partial \text{log} P(C_{1i}, C_{2i})}{\partial \boldsymbol{\theta}_1},$$

and

$$g(\boldsymbol{\xi}_i | \mathbf{y}_i, \mathbf{X}_i, M_{1i}, M_{2i}) = \frac{f(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\xi}_i) P(M_{1i} | C_{1i}) P(M_{2i} | C_{2i}) \Psi(\xi_i)}{f(\mathbf{y}_i, M_{1i}, M_{2i} | \mathbf{X}_i)}.$$

Note that the relationship between C_{1i} and C_{2i} is specified by (4.21) of Chapter 4. In the M-step, we need to solve $\sum_{i=1}^{N} E\left[S_i(\boldsymbol{\theta}_1)\right] = 0$. Integrals in (5.11) can be approximated using, for example, Monte Carlo methods (Browne et al., 2001; Sammel et al., 1997) or Gaussian-Hermite quadrature (Lesaffre and Spiessens, 2001) which replaces the integral with a weighted summation over u_i.

2) Estimation of parameters in the survival model
Denote by $\boldsymbol{\theta}_2 = (\alpha_t, \boldsymbol{\beta}, \tau_{k_1}^{(C_1)}, \tau_{k_2}^{(C_2)}, \sigma_u)$ the set of parameters in the survival model. The individual contribution to the expected score function of $\boldsymbol{\theta}_2$ is

$$E\left[S_i(\boldsymbol{\theta}_2)\right] = \sum_{k_2=1}^{K_2-1} \sum_{k_1=1}^{K_1-1} \int_{u_i} S_i(\boldsymbol{\theta}_2) g(\boldsymbol{\xi}_i | \mathbf{y}_i, \mathbf{X}_i, M_{1i}, M_{2i}) du_i, \tag{5.12}$$

where

$$S_i(\boldsymbol{\theta}_2) = \frac{\partial \text{log} f(\mathbf{y}_i | \mathbf{X}_i, \boldsymbol{\xi}_i)}{\partial \boldsymbol{\theta}_2} + \frac{\partial \phi(u_i)}{\partial \boldsymbol{\theta}_2},$$

and

$$\phi(u_i) = 1/\sqrt{2\pi\sigma_u^2} \exp\left[-u_i^2/(2\sigma_u^2)\right];$$

the latter is only related to σ_u. Similarly, solving $\sum_{i=1}^{N} E\left[S_i(\boldsymbol{\theta}_2)\right] = 0$ requires the approximation of the integral in (5.12). Estimation with higher dimensions of the latent variables (either discrete or continuous) can be computationally expensive.

3) Steps in the part of EM algorithm for step 3 of the extended 3-step ML approach

   (a) Generate initial estimates for all parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ in step 3.

   (b) E-step: compute $E\left[S_i(\boldsymbol{\theta}_1)\right]$ and $E\left[S_i(\boldsymbol{\theta}_2)\right]$ given in (5.11) and (5.12).

   (c) M-step: solve $\sum_{i=1}^{N} E\left[S_i(\boldsymbol{\theta}_1)\right] = 0$ and $\sum_{i=1}^{N} E\left[S_i(\boldsymbol{\theta}_2)\right] = 0$ to obtain updated estimates of parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

   (d) Repeat steps 2 and 3 until convergence is reached.

4) Standard errors
Asymptotic standard errors can be obtained by computing the information matrix $I(\boldsymbol{\theta})$ at

the maximum likelihood estimates, and taking diagonal elements of the inverse of $I(\hat{\boldsymbol{\theta}})$. Specifically, for an individual i, the contribution to Fisher's information matrix is

$$I_i(\boldsymbol{\theta}) = -E\left[\frac{\partial S_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right].$$

An alternative is to use the bootstrapped standard errors when the asymptotic formulas are not feasible. This option is available in many software packages and has been discussed in detail in Section 6.6 of Bartholomew et al. (2011).

The model in step 3 described above is fitted in LatentGOLD 5.1 (Vermunt and Magidson, 2015). Gauss-Hermite quadrature is used to approximate the integrals in the expected score function. The inverse Fisher information is used to compute the asymptotic standard errors. In terms of the convergence of the EM algorithm, two criteria are implemented in LatentGOLD. We set tolerance level of the total of the absolute relative changes in all parameters at 10^{-3} and the change in consecutive log-likelihoods at 10^{-8}.

5.6 Simulation study

5.6.1 Data generation

A simulation study was conducted to investigate the performance of the proposed multilevel discrete-time event history model under the general latent variable framework. This model allows for repeatable event times that are predicted by multiple associated categorical latent variables and also accounts for unobserved between-subject heterogeneity. The scenarios we have considered include a combination of different entropy levels and sample sizes, as in Section 4.5.

We first generate synthetic discrete-time event histories for repeatable events, taking into account the common influence of subject-specific u_i on the durations of all episodes (indexed by j) of an individual i. For simplicity, we consider a linear baseline logit hazard function, one individual-specific continuous covariate X_i, two categorical latent variables C_{1i} and C_{2i} (each measured by a separate set of indicators) and one time-varying continuous covariate X_{tij} as predictors of discrete-time binary event indicator y_{tij}.

We start by generating binary C_{1i} and C_{2i} from the log-linear model specified in (4.21-4.24), Section 4.3.4. Note that C_{1i} and C_{2i} could contain more than two categories but this may lead to fewer observations in each cell of the classification table and longer computational time. For illustration purposes and to avoid boundary estimates, we consider two categories for each of the C_{1i} and C_{2i}. We set $\omega_1^{(C_1)} = 0.7$, $\omega_1^{(C_2)} = 0.4$ and $\omega_{11}^{(C_1C_2)} = -0.5$, i.e.

latent variables are negatively correlated. The corresponding latent class proportions are $P(C_{1i} = 1, C_{2i} = 1) = 0.288$, $P(C_{1i} = 1, C_{2i} = 2) = 0.318$, $P(C_{1i} = 2, C_{2i} = 1) = 0.236$, $P(C_{1i} = 2, C_{2i} = 2) = 0.158$. We then generate 10 indicators $\{Y_{1i}^{(C)}, \ldots, Y_{10i}^{(C)}\}$, where the first 5 measure C_{1i} and the second 5 measure C_{2i}. Corresponding to high (0.8) and low (0.4) entropy values, we set the thresholds α_{1dk} (following the notation of Section 4.5.1) to 1.5 and 0.75, accordingly.

Next, discrete time-to-event data for repeatable events are generated from a simplified form of the model (5.6), i.e.

$$\text{logit}(h_{tij}) = \alpha_0 + \alpha_1 t + \beta_1 X_i + \beta_2 X_{tij} + \tau_{C_1,1} I(C_{1i} = 1) + \tau_{C_2,1} I(C_{2i} = 1) + u_i, \qquad (5.13)$$

where $\tau_{C_1,2} = \tau_{C_2,2} = 0$ and $I(\cdot)$ is the indicator function.

Instead of using the calendar-time (time elapsed since the entry into the first risk set), we employ the gap-time approach to define time intervals for recurrent events, i.e. the clock is reset to zero once an event occurs (Kelly and Lim, 2000). Generation of continuous-time event histories is often featured in medical studies using Cox's proportional hazard models (e.g.Austin, 2012; Bender et al., 2005) but details of generating discrete-time data for repeatable events have not been provided in previous research involving simulation studies for discrete-time models. We provide a pseudo algorithm (Algorithm 1) to describe the data generation process.

Denote by y_{ij} the duration of episode j of individual i, D_i the non-informative censoring time, T the maximum number of time intervals per person (set T equal across individuals) and $\boldsymbol{\theta}$ the whole set of true parameter values. In total, 500 replications of samples of sizes

$N = (500, 2000)$ are generated. Examples of the event histories of three individuals from one generated dataset are presented in Table 5.1.

Algorithm 1: Algorithm to generate event times from a random effects discrete-time event history model with (latent) categorical predictors

```
1  Input $\boldsymbol{\theta}$, $T$
2  Generate $X_i \sim N(0,1)$, $u_i \sim N(0,1)$, $X_{tij} \sim N(0,1)$, $C_{1i}$ and $C_{2i}$ from the log-linear
     model
3  Generate $T$ time intervals ($t = 1,2,...,T$) and $t_{ij} = t$ to update gap-time
4  Set episode index $j = 1$ for all time intervals of $i$
5  Generate discrete censoring time $D_i \sim$ uniform[1, $T$]
6  Generate $y_{ij}$ and $y_{tij}$ from model (5.13)
7  if $t_{ij} = min\{t : y_{tij} = 1, t \leq T\}$ then
8      $j = j+1$ and reset $t_{ij} = t - j$ in intervals where $t > t_{ij}$
9      Return to line 6
10 else
11     Break, $i = i+1$
12 end
13 Replace $y_{tij} =$ . in intervals where $t > D_i$ to account for censoring.
```

Table 5.1 Examples of the generated event histories for three individuals

$i = 1$	Experienced more than one event, censored									
Calendar Time t	1	2	3	4	5	6	7	8	9	10
Gap-time t_{ij}	1	2	3	1	2	3	1	2	3	1
Episode j	1	1	1	2	2	2	3	3	3	4
D_i	1	1	1	1	1	1	1	1	1	1
y_{tij}	0	.	.	.	.	.	.	.	.	.
$i = 2$	Experienced more than one event, not censored									
Calendar Time t	1	2	3	4	5	6	7	8	9	10
Gap-time t_{ij}	1	1	2	1	1	1	1	1	2	1
Episode j	1	2	2	3	4	5	6	7	7	8
D_i	10	10	10	10	10	10	10	10	10	10
y_{tij}	1	0	1	1	1	1	1	0	1	1
$i = 3$	Never experienced any event, censored									
Calendar Time t	1	2	3	4	5	6	7	8	9	10
Gap-time t_{ij}	1	2	3	4	5	6	7	8	9	10
Episode j	1	1	1	1	1	1	1	1	1	1
D_i	3	3	3	3	3	3	3	3	3	3
y_{tij}	0	0	0	.	.	.	.	.	.	.

5.6.2 Simulation results

We examined the performance of the proposed methodology for the general 3-step ML approach and its implementation in LatentGOLD 5.1 by fitting models to the simulated data. Results for 500 replications are summarised in Tables 5.2 to 5.5 for four settings with different combinations of sample sizes ($N = 500, 2000$) and entropy values (0.8 and 0.4) of the measurement models (syntax is given in Appendix D). We observe that in high entropy cases (good classification of individuals in the measurement models), estimates are unbiased (relative bias < 5%) and have good coverage (close to the nominal 95%) regardless of the sample size. When the class separation is poor and sample size is small, estimates of all parameters in the model are slightly biased and standard errors are understated. In particular, the variance of the random effects (σ_u^2) is poorly identified (Table 5.4) and this affects estimates of the regression coefficients. This is mainly because we are estimating a complex model with a combination of discrete and continuous level-2 unobservables. More observations (a larger N) and longer event histories (more level-1 measurements) are needed to better identify the model. Table 5.5 shows that even at the low entropy level, when sample size is increased to $N = 2000$, bias in the estimates and standard errors is greatly reduced. Note that across all simulated scenarios we still observe some parameters with slightly underestimated standard errors. This is mainly because of the neglected uncertainty in the estimated misclassification probabilities (derived after steps 1 and 2) in step 3. The underestimation is particularly obvious when the classification is poor as a large amount of variability in misclassification probabilities is ignored in step 3.

Table 5.2 Estimated effects of predictors on logit hazard of an event: high entropy (0.8) and small sample size ($N = 500$)

Parameters	True	Relative bias (%)	SE	SD	95% Coverage
β_0	−2.00	−0.04	0.24	0.23	0.96
$\beta_1(t)$	1.50	0.30	0.11	0.11	0.95
$\beta_2(X_i)$	1.50	0.54	0.10	0.11	0.95
$\beta_3(X_{tij})$	−0.50	0.56	0.06	0.06	0.95
$\tau_{C_1,1}$	2.50	0.18	0.19	0.19	0.94
$\tau_{C_2,1}$	−1.00	1.30	0.18	0.18	0.96
σ_u^2	1.00	−0.57	0.23	0.24	0.94
$\omega_1^{(C_1)}$	0.70	−0.67	0.16	0.17	0.95
$\omega_1^{(C_2)}$	0.40	1.16	0.17	0.18	0.94
$\omega_{11}^{(C_1C_2)}$	−0.50	−0.46	0.23	0.22	0.97

Relative bias (%)= (Estimate-True) / True×100%

Table 5.3 Estimated effects of predictors on logit hazard of an event: high entropy (0.8) and large sample size ($N = 2000$)

Parameters	True	Relative bias (%)	SE	SD	95% Coverage
β_0	−2.00	−0.54	0.12	0.12	0.94
$\beta_1(t)$	1.50	−0.15	0.06	0.06	0.94
$\beta_2(X_i)$	1.50	−0.11	0.05	0.05	0.94
$\beta_3(X_{tij})$	−0.50	−0.20	0.03	0.03	0.96
$\tau_{C_1,1}$	2.50	−0.15	0.09	0.09	0.95
$\tau_{C_2,1}$	−1.00	0.41	0.09	0.09	0.95
σ_u^2	1.00	−0.11	0.11	0.11	0.95
$\omega_1^{(C_1)}$	0.70	−0.77	0.08	0.09	0.94
$\omega_1^{(C_2)}$	0.40	−0.28	0.09	0.09	0.95
$\omega_{11}^{(C1C_2)}$	−0.50	−1.24	0.11	0.12	0.93

Relative bias (%)= (Estimate-True)/ True×100%

Table 5.4 Estimated effects of predictors on logit hazard of an event: low entropy (0.4) and small sample size ($N = 500$)

Coefficients	True	Relative bias (%)	SE	SD	95% Coverage
β_0	−2.00	−2.83	0.32	0.43	0.85
$\beta_1(t)$	1.50	0.82	0.11	0.12	0.95
$\beta_2(X_i)$	1.50	0.48	0.11	0.12	0.94
$\beta_3(X_{tij})$	−0.50	0.99	0.06	0.06	0.96
$\tau_{C_1,1}$	2.50	−4.76	0.27	0.32	0.88
$\tau_{C_2,1}$	−1.00	−1.73	0.34	0.34	0.94
σ_u^2	1.00	27.07	0.36	0.43	0.86
$\omega_1^{(C_1)}$	0.70	−4.09	0.33	0.55	0.77
$\omega_1^{(C_2)}$	0.40	−9.30	0.36	0.60	0.78
$\omega_{11}^{(C1C_2)}$	−0.50	−11.42	0.52	0.51	0.97

Relative bias (%)= (Estimate-True)/ True×100%

Table 5.5 Estimated effects of predictors on logit hazard of an event: low entropy (0.4) and large sample size ($N = 2000$)

Coefficients	True	Relative bias (%)	SE	SD	95% Coverage
β_0	−2.00	−1.82	0.16	0.19	0.87
$\beta_1(t)$	1.50	0.06	0.06	0.05	0.95
$\beta_2(X_i)$	1.50	−0.05	0.06	0.06	0.94
$\beta_3(X_{tij})$	−0.50	0.29	0.03	0.03	0.95
$\tau_{C_1,1}$	2.50	−1.27	0.12	0.13	0.94
$\tau_{C_2,1}$	−1.00	1.50	0.16	0.18	0.94
σ_u^2	1.00	5.07	0.17	0.19	0.93
$\omega_1^{(C_1)}$	0.70	−3.01	0.16	0.25	0.79
$\omega_1^{(C_2)}$	0.40	0.03	0.17	0.25	0.83
$\omega_{11}^{(C1C_2)}$	−0.50	−3.39	0.26	0.26	0.96

Relative bias (%)= (Estimate-True)/ True×100%

5.7 Application: a study of the effects of childhood socioeconomic circumstances on partnership formation and dissolution

In the original dataset, partnership durations are recorded to the nearest year and month. Following the rationale described in Section 5.2.4, we use discrete-time EHA. Time-to-event data are first restructured into grouped time intervals of 6-month, rather than monthly intervals to reduce the size of the analysis file (e.g. Steele et al., 2005). A simple example of the data restructuring process is illustrated as follows. Suppose there is an individual i who has experienced partnership dissolution after 15 months. This partnership episode will have three records in the person-period file, with exposure times n_{tij} of 6, 6 and 3 months for each record, respectively. We then create a vector y_{tij} = (0, 0, 1) as the response variable that indicates the occurrence of partnership dissolution during each 6-month interval of episode j. The underlying assumption for grouping intervals is that the hazard of experiencing an event and values of time-varying covariates are constant within each grouped interval. The restructured person-episode-period event history file is then merged with the childhood dataset that contains repeated measures of childhood SECs. In this application, we focus on individuals ($N = 7,313$) with complete partnership histories (i.e. provided valid responses at wave 8 at age 50) for ages 16-50. This also allows for a comparison of results from models in Chapter 6, where summaries of the partnership history over 34 years of follow-up are used as predictors of the distal health outcome. More details on partnership data derived from the survey are available in Section 3.3 and Section 3.4. The methods to handle missing data (mostly due to dropouts) in the analysis are discussed in Chapter 7.

The following analyses are based on the restructured person-episode-period file (in 6-month intervals). To model the timing of first partnership formation and recurrent dissolution, we consider model (5.6) introduced in Section 5.5.1 but extend it to include four categorical latent predictors. Both submodels have adjusted for exposure time n_{tij} as not all episodes are observed for a full 6-month interval. Briefly, to allow for exposure in a logit model, y_{tij} is modelled as a binomial outcome (grouped binary) with denominator n_{tij}. All models in this application are specified in LatentGOLD 5.1. A set of explanatory variables are considered and a full description is included in Section 3.3 and Section 3.4.

5.7.1 EHA of the time to first partnership formation

For the event of first partnership formation, $N = 7,313$ cohort members contribute a total of 340,883 six-month person-period records after discretisation. To address the research question set out in Section 5.3, we first fit a logit model for the time to first partnership formation, predicted by four latent dimensions of childhood SECs. Reasons for considering only the first partnership event are set out as follows. First, it has been shown that influences of childhood situations are significant on the first formation, but not the formation of subsequent partnerships (Berrington and Diamond, 2000). Second, in later stages of the research which involves joint modelling (Section 5.7.3) of partnership transitions, the formation process is modelled to adjust for the endogenous predictor "age at the start of a partnership" in the dissolution model. Among the literature reviewed in Section 5.4, most previous studies that modelled the time to first entry into partnership did not consider joint modelling of correlated processes. Among the few studies that have considered joint models for recurrent formation (i.e. entry into each partnership) and dissolution events, childhood effects were not their main interest. For example, Steele et al. (2006) made the distinction between the risk of first and subsequent partnership transitions by allowing for the interaction of duration variables with indicators of previous partners, which is not the main interest of our research. Therefore in this analysis, we model the time to first partnership formation only.

To allow for more flexible duration effects on the baseline hazard, we created time dummies for the grouped 3-year intervals, leading to a piecewise constant specification of the logit baseline hazard. Note that the analysis file is still based on the 6-month intervals; the 3-year interval is only used to simplify duration effects. Lengths of exposure time within these intervals are accounted for in the model, which has been discussed in Section 5.2.4.

After fitting separate latent class models for each dimension of childhood circumstances, modal classes as well as the misclassification probabilities are retained. These estimates are then carried forward to the next step which involves fitting a model for the time to first partnership. Noting that the event of first partnership formation is non-repeatable, we fit a single-level model to estimate the effects of latent childhood SECs on the timing of first partnership formation. We first employ the traditional modal class approach and then the general 3-step approach, using the methods described in Section 5.2.4 for a binary response. The 1-step approach is not considered as the latent class membership may be heavily distorted due to the multilevel duration data.

Results in Table 5.6 show that the estimated coefficients of observed predictors are similar for both approaches. Coefficients of the duration (age) terms indicate that the baseline tendency of entering into first partnership is the highest during ages 22 to 25. In terms of the effect of time-varying education level, we find that compared to those who leave education at or before age 16, individuals with a higher education level are significantly more inclined to delay the entry into first partnership, particularly those with 3-5 years of post-16 full-time education (i.e. in universities). This is consistent with earlier findings of Berrington and Diamond (2000).

With regards to the effects of the childhood SECs, the estimated coefficients differ for the two approaches. Using the modal class approach we find that cohort members growing up in families with fathers in the low or medium social class, higher material hardship and parental relationship instability tend to hasten entry into first partnership (relative to their counterparts with more favourable SECs). Similar findings have also been reported in Steele et al. (2005). In contrast, consistent with the results of Aassve and Billari (2006) and Aassve et al. (2006b), using the 3-step ML approach, we find that the effect of having a father in the low social class is non-significant and that the impact of high financial difficulty on the tendency of early partnership formation is larger in magnitude (estimated log-odds of 0.20 from the extended 3-step ML approach versus 0.11 from the modal class approach).

Table 5.6 Effects of predictors on the logit hazard of first partnership formation

Covariates	Modal class		General 3-step	
	Estimate	SE	Estimate	SE
Intercept	−6.13**	0.05	−6.11**	0.05
Age[a] (ref=16-19)				
19-22	1.55**	0.05	1.54**	0.05
22-25	2.01**	0.05	1.99**	0.05
25-28	1.86**	0.05	1.85**	0.05
28-31	1.73**	0.06	1.71**	0.06
31-34	1.64**	0.07	1.60**	0.07
34-37	1.20**	0.10	1.22**	0.09
37-40	0.94**	0.12	0.91**	0.12
40-43	0.39**	0.16	0.38**	0.16
43+	0.04	0.13	0.07	0.13
Number of years in post-16 full-time education[a] (ref.=0)				
1	−0.15**	0.04	−0.16**	0.04
2	−0.20**	0.04	−0.20**	0.04
3-5	−0.37**	0.04	−0.36**	0.04
6+	−0.15**	0.05	−0.15**	0.05
Latent categorical predictors				
Social class[b] (ref.=High)				
Low	0.10**	0.04	0.04	0.08
Medium	0.11**	0.03	0.13**	0.04
Financial difficulty (ref. =Low)				
High	0.11**	0.04	0.20**	0.09
Material hardship (ref.=Low)				
Medium	0.04	0.03	0.03	0.04
High	0.06	0.04	0.03	0.05
Family structure (ref. =Stable)				
Unstable	0.14**	0.05	0.10*	0.06

**$p < 0.05$,*$p < 0.1$

[a] time-varying covariate

[b] Father or male head social class

5.7.2 EHA of the time to partnership dissolution

For the event of partnership dissolution, those cohort members with complete partnership histories contribute a total of 451,639 six-month person-episode-period records after discretisation. For this event, duration is defined as the time to dissolution since entering into a partnership (either cohabitation or marriage). As the separation events are recurrent, we consider a two-level random intercept model (5.4) to estimate the effects of childhood circumstances on the risk of partnership dissolution, allowing for the influence of individual-level unobservables captured by a random effect term u_i. Similar to the model for the partnership formation process, dummies for grouped 3-year intervals are created to specify a piecewise constant logit baseline hazard. The age at the start of the partnership, partnership type, the number of previous partners, the number of pre-school children and education level are included as risk factors for dissolution. We have also tested if childhood effects on partnership separation differ for the first and subsequent partnerships. This is following a hypothesis that the decision to separate in second and higher-order partnerships can be influenced by the previous experience of separation, and thus the effects of childhood background on the dissolution risk for later partnerships may be absorbed by previous partnership outcomes (also discussed in Steele et al. (2006)). This hypothesis is tested by including the interactions between a binary indicator of the existence of previous partners and the childhood variables in the model. However, as all interaction terms are non-significant, they are omitted from the model. This suggests that childhood effects on partnership dissolutions do not vary for recurrent partnerships. This two-level random intercept model with categorical latent predictors is estimated first using the modal class approach, and then the general 3-step ML approach. Results are summarised in Table 5.7 where we only report significant estimates of the duration effects to save space.

Similar to what we observe in the formation model, both approaches produce similar estimates for the effects of the observed predictors on the risk of partnership dissolution. From the duration terms (number of years in partnership), compared to couples who have been together for less than 3 years, the hazard of separation is significantly lower for couples in relationship for 15-18 years but the highest for those in a partnership for 21-27 years. Compared to married couples, cohabitors are at a significantly higher risk of early separation. People with previous partnerships have 38% $(1-\exp(-0.48))$ lower odds of separation compared to their never-partnered counterparts. Individuals who started their partnership late also tend to have a significantly more stable partnership than those who entered into partnership before age 20 (similar to the findings of Steele et al. (2005)). Turning to time-varying covariates, we find that of all education levels considered, only those with over 6 years of post-16 education experience have a significantly lower risk of separation (see results of the 3-step ML approach). Regarding the presence of pre-school children in relationships, consistent with the results reported in Steele et al. (2005), a significant stabilizing effect of the existence of young children is found: couples with at least one pre-school child have a significantly lower chance to separate compared to those without. For the effects of latent childhood circumstances, both the modal class and 3-step ML approaches find the influence of social class, and financial difficulty on the risk of partnership dissolution non-significant. Different from the modal class approach, the 3-step ML approach confirms only a significant and negative influence of high levels of material hardship on the hazard of separation. Consistent with findings in most of the previous studies (e.g. Aassve et al., 2006a; Amato, 1996; Kiernan and Cherlin, 1999), our results suggest that cohort members from unstable families (e.g. who have experienced paternal separation) have a 35% $(\exp(0.30)-1)$ higher odds of separation than their counterparts from more stable family backgrounds. The magnitude of this effect is larger than that estimated from the modal class approach (log-odds of 0.30 versus 0.23). The significant estimate of σ_u can be interpreted as follows: ceteris paribus, an individual with an unobserved factor that is one standard deviation above the mean has a 1.72 times $(\exp(1)-1)$ higher odds of separation in any 6-month interval than an average person (with $u_i = 0$).

Table 5.7 Covariate effects on the logit hazard of partnership dissolution

Covariates	Modal class		General 3-step	
	Estimate	SE	Estimate	SE
Intercept	−6.54**	0.11	−6.56**	0.10
Number of years in partnership (ref.=3)				
15	−0.27**	0.12	−0.28**	0.11
21	0.30**	0.12	0.32**	0.11
24	0.53**	0.12	0.56**	0.12
27	0.23	0.14	0.28**	0.14
Partnership type[a] (ref.=Marriage)				
Cohabitation	1.86**	0.06	1.87**	0.06
Number of previous partners[a] (ref.=0)				
1+	−0.48**	0.06	−0.48**	0.06
Age at the start of partnership (ref.=less than 20)				
20-25	−0.50**	0.07	−0.51**	0.07
25-30	−0.78**	0.09	−0.79**	0.09
30-35	−0.93**	0.11	−0.93**	0.10
35+	−1.03**	0.11	−1.04**	0.11
Number of years in post-16 full-time education [a] (ref.=0)				
1	0.03	0.09	0.04	0.09
2	−0.10	0.09	−0.06	0.09
3-5	0.04	0.09	0.04	0.09
6+	−0.15	0.10	−0.19*	0.10
Number of pre-school children[a](ref.=0)				
1	−0.46**	0.08	−0.47**	0.08
2	−0.43**	0.08	−0.45**	0.08
3+	−0.36**	0.11	−0.36**	0.11
Latent categorical predictors				
Social class[b] (ref.=High)				
Low	−0.08	0.09	0.01	0.15
Medium	−0.04	0.07	−0.04	0.08
Financial difficulty (ref.=Low)				
High	0.01	0.09	−0.09	0.17
Material hardship (ref.=Low)				
Medium	−0.12*	0.07	−0.12	0.08
High	−0.24**	0.07	−0.25**	0.09
Family structure (ref.=Stable)				
Unstable	0.23**	0.09	0.30**	0.10
Random effect variance σ_u^2	1.01**	0.10	1.00**[c]	0.10

$^{**}p < 0.05, ^{*}p < 0.1$

[a] time-varying covariate; [b] Father or male head social class

[c] Likelihood ratio test for $H_0 : \sigma_u = 0$ gives $\chi_1^2 = 308.99, p < 0.001$

5.7.3 Joint modelling of partnership dissolution and formation

In order to investigate the third research question set out in Section 5.3, we estimate models for the timing of first partnership formation and the subsequent dissolutions simultaneously (model (5.5) in Section 5.2.6), allowing for $\mathrm{corr}(u_i^{(1)}, u_i^{(2)}) \neq 0$. Note that the subscript j in the first equation can be dropped as there are no repeated events for the first partnership formation.

Rationale for the joint model

Before fitting model (5.5), the rationale underlying this joint modelling approach is summarised below. We hypothesize that those who tend to form their first partnership early and those who are prone to separate early share similar intrinsic characteristics which are not fully captured by the covariates. This could bias the estimates of childhood effects in the dissolution model, if childhood circumstances influence both entry into first partnership and dissolution. To be more specific, childhood covariates, which depend only on children's own characteristics, may be considered as exogenous with respect to the timing of both the formation and dissolution of partnerships. However, age at the start of each partnership, a variable that is highly dependent on the outcome of the formation model (age at the start of first partnership), is also commonly considered as a predictor of partnership dissolution. It can be endogenous if the formation and dissolution processes share unmeasured influences. Neglecting endogeneity of age at the start of partnership in the dissolution model can lead to a biased estimate of its effect on the risk of partnership dissolution (see Rabe-Hesketh and Skrondal (2012) for a comprehensive account of the problem of endogeneity bias). Consequently, the estimated effects of variables that are associated with age at the start of partnership (e.g. childhood SECs, also shown by Kiernan and Cherlin (1999)) can also be biased.

To correct for endogeneity-induced biased estimates, two-stage least squares and joint modelling are two well-established methods (Rabe-Hesketh and Skrondal, 2012). The latter is more straightforward and flexible for handling both continuous and categorical response types. Allowing for the correlation of individual random effects in the equation for the time to first partnership and that for the time to subsequent dissolutions can account for shared influences of unmeasured time-invariant characteristics.

Simulation study

Appendix C contains the results of a simulation study designed to assess the influence of the endogenous predictor, age at the start of partnership, on the hazard of separation. The simulation study considers a simple example to assess bias in estimated coefficients for a continuous outcome y_2 (e.g. time to partnership dissolution) and a continuous endogenous predictor y_1 (age at the start of each partnership in the current example). Each variable is only observed once for an individual, leading to a single-level model. More formally, we consider the following data generating process:

$$\begin{aligned} y_1 &= \alpha_0 X + \alpha_1 Z + u_1, \\ y_2 &= \beta_0 X + \beta_1 y_1 + u_2, \end{aligned} \tag{5.14}$$

where X is a binary predictor with a negative influence on y_1 ($\alpha_0 < 0$) and a positive influence on y_2 ($\beta_0 > 0$). We also set a positive relationship between y_1 and y_2 ($\beta_1 > 0$). u_1 and u_2 represent individual-level unobserved characteristics that follow a bivariate normal distribution with non-zero correlation. Z is introduced as an instrumental variable for model identification. To ensure it satisfies the requirements of an instrument, $\text{corr}(Z, u_1) = \text{corr}(Z, u_2) = 0$ and $\text{corr}(y_1, Z) = +/-0.9$.

Based on the simulations results, we conclude the following:

(a) If the correlation of the individual-level residuals across the two equations is positive (i.e. those who tend to start a partnership late also tend to stay longer with their subsequent partners), without joint modelling, the positive effects on the timing of dissolution of the endogenous age variable and the childhood variables (e.g. family structure) that are associated with it will be overstated.

(b) Conversely, if the correlation of individual-level residuals is negative (i.e. those who tend to start a partnership late tend to separate early from their partners in subsequent relationships), without joint modelling, the positive effects on the timing of dissolution of the endogenous age variable and the childhood variables (e.g. family structure) that are associated with the endogenous variable will be understated.

An empirical explanation of the upward bias of the positive effect of age at the start of partnership on the time to dissolution is straightforward. $\text{corr}(u_1, u_2) > 0$ implies that among those who partner late (high y_1), there is an over-representation of individuals whose unmeasured characteristics place them at a low risk of dissolution ($u_2 > 0$). If we assume $\text{corr}(u_1, u_2) = 0$, the presence of individuals with $u_2 > 0$ among those with high y_1 pushes up the partnership duration (y_2), leading to an overstatement of the positive effect of age at the start of partnership (y_1) on y_2. The direction of biased effects of X on y_2 is difficult to explain (as it depends on the value of more than one parameter) but can be shown mathematically (see Appendix B for details).

Findings from the joint model

Following the procedure set out in Section 5.2.6, we estimate the formation and dissolution models jointly, allowing for a correlation between $u_i^{(1)}$ and $u_i^{(2)}$ to be freely estimated. The main interest, the effects of childhood SECs on partnership transitions are summarised in Table 5.8 and the remaining results are reported in Appendix E. We have several interesting findings. First, the joint model fails to converge when the covariance matrix for the random effects is freely estimated. For identification purposes, we therefore factorise the random effects by setting $u_i^{(2)} = u_i$, $u_i^{(1)} = \lambda^{(F)} u_i$ which reduces the number of random effect parameters from 3 to 2 and assumes a common set of individual-level unobservables u_i with differential effects on the risk of formation $(\lambda^{(F)})$ and dissolution (1, fixed). In this application, we find a significant $\lambda^{(F)}$ (-0.28) and the estimated effects of most covariates (see Table E.2 for the effects of observed predictors) in the joint model are consistently larger in magnitude than those estimated from separate analyses. This is in line with the simulation results for a scenario with a negative residual correlation between equations. A negative $\lambda^{(F)}$ also indicates a negative relationship between the timing of first parternship formation and subsequent partnership dissolutions. This finding partially agrees with the results in Steele et al. (2006) for the cohabitors but it seems to contradict with findings of Aassve and Billari (2006) where a positive correlation between the formation and dissolution process is confirmed. We provide two possible explanations here. First, previous studies investigated the effects of recurrent formation and dissolution of unions while our analysis considers only the first partnership formation events and subsequently recurrent dissolution events. The overall positive correlation between processes found in these studies may be explained by the positive associations among the second and higher orders of events. Second, in line with our findings, it is likely that individuals who enter the first relationship early also have a higher propensity to favour longer partnerships due to their latent interest in being in partnerships (i.e. shorter time in single status).

5.7.4 Summary of results

After all the above discussions, the research questions set out in Section 5.3 could be answered. We focus on the results from the 3-step ML approach because we have a large dataset and good classification of individuals in step 1 (i.e. high entropy values in each of the measurement models). Simulation studies (Section 5.6) have shown that under these circumstances (assuming model assumptions are satisfied), estimates from the 3-step ML approach are superior to those from the modal class approach. We outline the key findings as follows.

1) How are different aspects of childhood SECs associated with the timing of entry into the first partnership?

We find that individuals from families with fathers in the medium social class, high financial difficulty and unstable family structure tend to enter into first partnership earlier than their counterparts from more advantaged socioeconomic backgrounds. Effects of other aspects of childhood SECs are non-significant.

2) How are different aspects of childhood SECs associated with the risk of partnership dissolution?

   Among the childhood SECs considered, we find that children from families with high material hardship (e.g. poor housing situations) have a lower risk of separation while unstable family structure in childhood is a significant risk factor for relationship break-downs. Further investigations also find that such effects do not tend to vary for recurrent partnerships. In addition, individual heterogeneity in the tendency to separate is present in this sample, suggesting that there exist influences of unobserved individual-level characteristics that are not accounted for by the observed predictors.

3) Are there shared unobserved characteristics influencing the timing of first partnership formation and recurrent dissolutions? If yes, what are the implications?

   Joint modelling of the time to first partnership and the time to recurrent partnership dissolution finds a significant and negative $\lambda^{(F)}$, indicating a negative association between the hazard of forming the first partnership and dissolving subsequent unions due to the shared influence of a common set of individual-specific unobservables.

Table 5.8 Joint modelling for time to first partnership formation and subsequent dissolutions

Covariates	Modal class		General 3-step	
	Estimate	SE	Estimate	SE
Formation model				
Intercept	−6.18**	0.06	−6.17**	0.06
Latent categorical predictors				
Social class[a] (ref.=High)				
Low	0.11**	0.05	0.04	0.08
Medium	0.12**	0.04	0.13**	0.04
Financial difficulty (ref. =Low)				
High	0.13**	0.05	0.24**	0.10
Material hardship (ref.=Low)				
Medium	0.04	0.03	0.04	0.04
High	0.07*	0.04	0.04	0.05
Family structure (ref. =Stable)				
Unstable	0.12**	0.05	0.14**	0.06
Dissolution model				
Intercept	−6.33**	0.12	−6.36**	0.12
Latent categorical predictors				
Social class[a] (ref.=High)				
Low	−0.10	0.09	−0.02	0.15
Medium	−0.06	0.07	−0.05	0.08
Financial difficulty (ref. =Low)				
High	−0.01	0.09	−0.11	0.18
Material hardship (ref.=Low)				
Medium	−0.13**	0.07	−0.13	0.08
High	−0.25**	0.08	−0.27**	0.10
Family structure (ref. =Stable)				
Unstable	0.23**	0.09	0.29**	0.11
$\lambda^{(F)}$	−0.28**	0.07	−0.28**	0.06
Random effect variance σ_u^2	1.07**	0.05	1.15**	0.11

** $p < 0.05$,* $p < 0.1$

[a] Father or male head social class

Chapter 6

Joint modelling of distal health outcomes and event histories predicted by categorical latent variables

6.1 Introduction

In this chapter, we investigate the extent to which the effects of childhood SECs on midlife health are mediated by one of the life events mentioned in the conceptual framework of Section 1.2: the experiences of partnership formation and dissolution up to midlife. A number of methodological challenges need to be addressed. After obtaining summaries of four aspects of childhood SECs (see Chapter 4) and linking them to partnership outcomes (see Chapter 5), we now further extend the model by relating these two pieces of information collected at different life stages (throughout childhood and, across 34 years of follow-up in adulthood) to a distal health outcome. Most previous research has used simple forms of structural equation models (SEMs) with only observed variables (including multiple mediators and confounders) to explore the pathways linking childhood SECs to later health. Few studies have additionally introduced latent variables to capture trajectories of development while accounting for measurement error (e.g. Hagger-Johnson et al., 2011). A more detailed review of the pathways proposed in previous studies is available in Section 2.4 and Section 2.5. As noted in Section 2.5, studies that have used event history data have been primarily interested in identifying risk factors of time-to-event outcomes rather than relating the information captured by event histories to distal outcomes such as later health. In this chapter, we develop a general modelling framework that can address these challenges simultaneously, while making full use of the available longitudinal data. We propose a multilevel SEM that can incorporate information from different life stages and domains and capture the dependencies between childhood socioeconomic experiences, partnership transitions in adulthood and midlife health.

The main contributions and advantages of our model can be summarised as follows. First, the 3-step ML method of Chapter 4 (also in Zhu et al. (2017)) for estimating the effects of multiple associated categorical latent variables on a distal outcome is generalised to handle mixed outcomes measured at different levels in a hierarchical structure. The method is applied in an analysis of the effects of latent variables for four dimensions of childhood SECs on a binary health outcome and duration data on recurrent partnership transitions. Second, partnership histories are used to define not only outcomes that are predicted by childhood SECs but predictors of later health, by deriving summary variables of partnership stability for ages 16-50 (e.g. the total number of partners and percentage of time spent single). Third, the effects of childhood SECs on partnership formation and dissolution hazards for ages 16-50 and distal health at age 50 are jointly modelled by allowing for a residual association across equations via a continuous latent variable (random effect) capturing unobserved time-invariant characteristics with differential effects on each response. This residual association helps to 1) mitigate the problem of endogeneity due to correlation between individual-specific unobservables and the summaries of partnership events included as predictors in the health model, 2) further allow for an indirect relationship between partnership stability and midlife health, in addition to that captured by the summary variables and, 3) account for the additional dependence between partnership transitions and midlife health given the latent classes. The proposed SEM belongs to the general latent variable modelling frameworks of Skrondal and Rabe-Hesketh (2004) and Asparouhov and Muthén (2011), and hence has the same generalizability. More specifically, life events (e.g. partnership transitions) and distal outcomes (e.g. midlife health) can be viewed as items of one or more individual-level latent variables. Different specifications of the structural model can address various research questions where the underlying association of individual-level variables (rather than their manifestations with measurement errors) is of primary interest.

The rest of this chapter is organised as follows. Section 6.2 reviews statistical models that relate event history data to external variables, which is a key component model of the proposed SEM. This review helps to motivate our method described in Section 6.3, which further extends the proposed 3-step ML approach of Chapters 4 and 5 to the multilevel structural equation modelling framework. Section 6.4 describes the analysis of the effects of four associated dimensions of childhood circumstances on midlife health, mediated by partnership transitions across 34 years of follow-up.

6.2 Literature review: relating event history data to external variables

In many cohort studies, data often have complex structures (e.g. multilevel, mixed types) and have been collected at different times, for example, situations measured repeatedly in childhood or in adulthood, event histories (often collected retrospectively at one or more time points and later linked) and some outcomes such as health measured only once in adulthood. Bearing in mind the main research objective set out in Section 6.1 and to make full use of the data, in this section, we focus our discussion on relating event history data to external variables. External variables include those measured before (e.g. repeated measures of early-life situations or dropout indicators), during (e.g. repeated measures of health or another event history) and after (e.g. distal outcomes measured once only) the event history. There are two main treatments of these variables in the literature: 1) as predictors of the time-to-event data; 2) as outcomes that are modelled jointly with event times. The following sections provide an overview of models that have been used in the literature and discuss connections with our proposed approach.

6.2.1 External variables as covariates

Recently, a major interest of medical studies has been to make better inference for the effects of repeated measures of relevant medical indicators (e.g. biomarker or responses to drugs) on a time-to-event outcome (e.g. time to death) (e.g. Faucett and Thomas, 1996; Proust-Lima and Taylor, 2009). For this purpose, repeated measures are often used to define time-varying covariates in the event history model. As repeated measures are collected intermittently while the time-to-event outcome contains information at more regular time intervals, this is essentially a missing data issue for the predictors. Multiple methods have been proposed to mitigate this problem but some could lead to heavily biased estimates. For instance, simple imputation of the longitudinal data (i.e. repeated measures) using either the last-observation-carry-forward approach or smoothing artificially increases measurement error and ignores the existence of earlier repeated measures (Carpenter et al., 2004). Alternatively, fitting a regression model for the repeated measures and using the predicted values as time-varying predictors for the time-to-event data ignores the uncertainty in the predicted values of longitudinal variables (Little and Rubin, 2014). In addition, treating repeated measures as time-dependent covariates can lead to model misspecification if they are related to the time-to-event outcome through unmeasured characteristics. This is a standard endogeneity problem that can lead to biased estimates (Steele, 2011; Steele et al., 2006).

6.2.2 Joint modelling of external variables and event times

The above mentioned problems have given rise to the joint modelling approach, which performs simultaneous estimation of the model for the external variable and that for the time-to-event outcome. The covariance structure of the residuals across the submodels can be specified according to different modelling hypotheses. This approach has been widely applied in social research. Lillard et al. (1995), one of the earliest users, estimated a joint model for pre-marital cohabitation (binary, repeated measures) and the time to marital dissolution, testing for selection of divorce-prone people into pre-marital cohabitation by allowing for residual correlation between the two submodels. More discussion of such models is available in Blossfeld and Gotz (2001). Estimating a joint model has several advantages. First, it improves modelling efficiency when submodels share common observed predictors. Second, it handles potential endogeneity in the model of primary interest through explicitly allowing for residual correlations across submodels and hence, more accurate estimates. Finally, we can examine the extent to which the relationship between the external variables and the time-to-event data is due to influences of shared unmeasured characteristics of individuals. A comprehensive review of various specifications of joint models is available in Hickey et al. (2016) and Tsiatis and Davidian (2004). In the following, we briefly summarise the key modelling components in the literature that inform the development of our model, which is described in detail in Section 6.3.

The joint model is comprised of two submodels, where the model for the external variable is linked with that for the time-to-event data. In medical studies, it is common to have repeated measures on external variables, which can be analysed by models for longitudinal data (longitudinal submodels). One possible specification is a joint latent class model where the association between two submodels is accounted for by a categorical latent variable, assuming that individuals within the same class share the same parameter values (Proust-Lima and Taylor, 2009). Other proposals for connecting the two submodels focus on specifying the association structure of the residual terms. We use the superscript (H) for longitudinal and (S) for survival submodels, respectively.

Submodel for longitudinal data

For the longitudinal submodel, we consider one measure of interest for illustration purposes but the extension to multivariate outcomes is straightforward. Denote by $H_{vi}\,(v = 1,\ldots,T)$ the measurement of the longitudinal outcome H (e.g. repeated measures for health) at time v for individual $i\,(i = 1,...,N)$. To account for dependencies between repeated measures within an individual, $u_i^{(H)}$, an individual-specific random effect is included in the model. We fit a generalized linear mixed model for H_{vi}:

$$g\{E[H_{vi}]\} = \boldsymbol{\beta}'\mathbf{X}_{vi}^{(H)} + u_i^{(H)}, \tag{6.1}$$

where $g(\cdot)$ denotes the link function and $\mathbf{X}_{vi}^{(H)}$ denotes a vector of time-varying and time-invariant predictors for the longitudinal outcome.

Submodel for time-to-event data

For the time-to-event submodel, we consider the multilevel event history model (5.4), allowing for recurrent events and for the influence of individual-specific unobservables through $u_i^{(S)}$. Denote by $h_{tij}^{(S)}$ the hazard in time interval $[t, t+1)$ in episode j of individual i. The model can be written as:

$$\text{logit}\left(h_{tij}^{(S)}\right) = \alpha_t + \boldsymbol{\beta}'\mathbf{X}_{tij}^{(S)} + u_i^{(S)}, \tag{6.2}$$

where α_t is a duration effect and $\mathbf{X}_{tij}^{(S)}$ is a vector of time-varying and time-invariant covariates for hazard of an event. Note that $\mathbf{X}_{tij}^{(S)}$ can contain elements in $\mathbf{X}_{vi}^{(H)}$ or quantities related to the outcome H_{vi}.

To account for endogeneity of $\mathbf{X}_{tij}^{(S)}$ in (6.2) due to shared influences of time-invariant unobservables, an association structure for $(u_i^{(H)}, u_i^{(S)})$ can be specified. For example, it can be of a functional form with distributional assumptions (Ibrahim et al., 2004; Luo and Wang, 2014; Musoro et al., 2015; Rizopoulos, 2011). In the shared random effects framework, $(u_i^{(H)}, u_i^{(S)})$ can be specified as random effects with a joint variance-covariance matrix be freely estimated (Rizopoulos, 2011). In addition to these parametric approaches, to avoid potential incorrect inference caused by misspecification of the distribution of random effects, non-parametric approaches have also been advocated. For example, the random effects may be assumed to follow a smooth conditional density with no jumps, kinks or oscillations (Song et al., 2002), which accounts for a wide range of distributions that can be multi-modal, skewed or fat-tailed.

Among these approaches, the shared random effects framework is the most appropriate for our substantive interest (more discussion is provided in Section 6.2.3). The association structure could be specified by assuming $(u_i^{(H)}, u_i^{(S)}) \sim N(0, \Omega_u)$, which has several advantages. First, it allows for a flexible association between two submodels through an unstructured variance-covariance matrix Ω_u. The sign of the covariance term $\sigma^{(HS)}$ indicates the direction of the relationship between H_{vi} and the time-to-event outcome due to individual-specific unobservables. Interpretation of the random part is similar to what has been discussed in Section 5.2.6. Second, allowing $\sigma^{(HS)} \neq 0$ accounts for potential endogeneity if the covariate and the outcome have shared influences of time-invariant unobservables. Third, depending on the modelling hypotheses, variations within the shared random effects framework can be employed, for example, by reparameterisations of Ω_u through factorisation (i.e. $u_i^{(S)} = \lambda u_i^{(H)}$). Finally, this joint model is specified within the framework of a multilevel SEM as random effects are essentially level-2 latent variables (see Skrondal and Rabe-Hesketh (2004) and Muthén and Asparouhov (2008)). The joint model (6.1, 6.2) can be estimated in several software packages, e.g. stjm in Stata (Crowther et al., 2016), frailtypack in R (Rondeau et al., 2012), Mplus and LatentGOLD. Models can be estimated using maximum likelihood where the integrals of the random effects in the likelihood function are approximated using numerical approximation methods (e.g. Gaussian quadratures) or Bayesian methods (e.g. MCMC).

6.2.3 Connection to our research questions

In medical studies, it is common to have repeated measures of bio-markers of health conditions, with measurements collected either after prescription of drugs or during routine body check, as well as the timing of a terminal event. Consequently, in the literature on joint modelling of external variables and survival outcomes, medical researchers usually focus on the time-to-event submodel because the occurrence of an event (e.g. death) is the endpoint of the study. Our research, however, has a different setting. The cohort study used in the research, the 1958 NCDS contains event histories of partnership transitions for 34 years and several measures of midlife health at age 50. Recall the substantive interest of the influence of childhood circumstances on midlife health (a distal outcome) through partnership stability (an event history outcome), where the distal health outcome is our endpoint of study. Therefore, our key modelling interest is to treat partnership stability both as an outcome in the childhood-partnership relationship, but also a predictor in the partnership-midlife health relationship. To the best of our knowledge, such models have not been discussed in the methodological literature. Relating to the research objective set out in Section 6.1, outcomes of the event history, for example, summaries of partnership experiences could be used as predictors of later health outcomes. We therefore specify a structural equation model that is composed of an event history submodel for partnership transitions and a submodel for distal health. To handle potential endogeneity of partnership summaries in the health submodel due to shared influences of unobserved individual-specific characteristics, we could employ the shared random effects framework (reviewed above) by allowing for correlation between residuals. Such a joint model can also be easily generalised in two directions, capitalizing on previous research on the corresponding submodels. For the time-to-event submodel, models for multiple event histories or for competing risks could be considered (e.g. Berrington and

Diamond, 2000; Steele, 2011; Steele et al., 2004) and summary information of the event histories could be derived and treated as predictors of the distal outcome. For the submodel for distal outcomes, more than one, possibly associated variables of mixed types can be considered.

6.3 A general 3-step ML approach for structural equation models

We now further extend the 3-step ML approach to structural equation models, where the measurement models for latent dimensions of childhood SECs are not estimated simultaneously with the rest of the model. Similar to the method described in Section 5.5, in steps 1 and 2, we fit separate latent class models to each set of composite indices for each dimension of childhood SECs. Modal classes and misclassification probabilities are saved. These quantities feed into the SEM in step 3, that relates latent childhood SECs to the distal health outcome and allow the effects to be mediated by partnership transitions. In the following subsections, we first discuss the modelling framework in step 3 and then investigate the performance of our extended 3-step ML approach across several simulated scenarios.

6.3.1 Model in Step 3

In steps 1 and 2, we fit separate latent class models to each measurement vector $\mathbf{Y}_q^{(C)}$ $(q = 1,...,4)$ for latent dimensions of childhood SECs and save the modal classes and misclassification probabilities. These quantities are inputs in the estimation of the SEM at step 3 where the latent childhood SECs (C_qs) are related to time-to-event outcomes and midlife health. In step 3, we estimate a joint model for partnership transitions and midlife health where C_q is included as a predictor in each submodel. Specifically, the model in step 3 contains three submodels – two discrete-time event history submodels for the formation and dissolution processes and the health submodel – that are connected by allowing for a non-zero correlation between random effects.

Submodels for partnership formation and dissolution

Following the analyses in Chapter 5, we focus on the time to first partnership formation and time to (possibly recurrent) partnership dissolutions in our analyses. The partnership data have a hierarchical structure with partnership episodes nested within individuals. A joint model of the formation and dissolution processes is considered because a key predictor in the dissolution model, age at the start of the partnership, is highly related to the outcome of the partnership formation model (i.e. age at the start of first partnership), and may thus be endogenous. We allow for a residual association between the formation and dissolution processes by including an individual-level random effect u_i in both models.

Building on the specification of models described in Section 5.5.1, to prepare for the discrete-time event history analysis, the duration to first partnership formation of individual i is expanded to S_i records indexed by s and, for the dissolution model, the duration of each partnership episode j $(j = 1,...,J_i)$ is expanded to T_{ij} records indexed by t. Denote by $y_i^{(F)}$ and $y_{ij}^{(S)}$ the event or censored time for formation and separation events. We define the corresponding binary responses $y_{si}^{(F)}$ and $y_{tij}^{(S)}$ such that

$$y_{si}^{(F)} = \begin{cases} 1 & y_i^{(F)} = s, \text{uncensored} \\ 0 & y_i^{(F)} = s, \text{censored} \\ 0 & y_i^{(F)} > s. \end{cases} \qquad y_{tij}^{(S)} = \begin{cases} 1 & y_{ij}^{(S)} = t, \text{uncensored} \\ 0 & y_{ij}^{(S)} = t, \text{censored} \\ 0 & y_{ij}^{(S)} > t. \end{cases}$$

Denote by $h_{si}^{(F)} = P(y_{si}^{(F)} = 1|y_{s'<s,i}^{(F)} = 0)$ the hazard of the first entry into partnership in the time interval $[s, s+1)$ and $h_{tij}^{(S)} = P(y_{tij}^{(S)} = 1|y_{t'<t,ij}^{(S)} = 0)$ the hazard of separation in time interval $[t, t+1)$ of episode j. Given the long observation period, the monthly duration data are aggregated to six-month intervals to reduce the size of the discrete-time dataset. The likelihood function for this discrete-time model is equivalent to the model for multilevel binary data where $y_{si}^{(F)}$ and $y_{tij}^{(S)}$ follow binomial distributions with denominator $n_{si}^{(F)}$ and $n_{tij}^{(S)}$ equal to the exposure time in each time interval, respectively (Steele et al., 2005).

We specify a generalized linear model for the binomial responses using a logit link function:

$$\begin{aligned} \text{logit}\left(h_{si}^{(F)}\right) &= \alpha_s^{(F)} + \boldsymbol{\beta}^{(F)'}\mathbf{X}_{si}^{(F)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1} \tau_{C_q,k_q}^{(F)} I(C_{qi} = k_q) + \lambda^{(F)}u_i \\ \text{logit}\left(h_{tij}^{(S)}\right) &= \alpha_t^{(S)} + \boldsymbol{\beta}^{(S)'}\mathbf{X}_{tij}^{(S)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1} \tau_{C_q,k_q}^{(S)} I(C_{qi} = k_q) + \lambda^{(S)}u_i, \end{aligned} \tag{6.3}$$

where $\alpha_s^{(F)}$ and $\alpha_t^{(S)}$ are functions of time which specify the baseline hazard (assumed to be piecewise constant here), the **X**s are event-specific vectors of predictors with coefficients $\boldsymbol{\beta}$, $I(C_{qi} = k_q)$ is a dummy variable for class k of latent childhood SEC variable q, with coefficient τ_{C_q,k_q} (taking category K_q as the reference category), and $\lambda^{(F)}$ and $\lambda^{(S)}$ are coefficients of the individual random effects u_i. The implicit assumption for this model is that both event hazards are influenced by a common set of individual-specific unobservables but differential effects of u_i are allowed through the λ parameters. Model (6.3) is similar to the multiprocess model (without categorical latent variables) that has been discussed in previous research to model correlated processes with endogenous predictors (e.g. Aassve et al., 2006a).

Submodel for midlife health

We consider a scenario where the distal health outcome is the endpoint that is collected later than the event histories and only once per individual. The aim is to relate latent childhood SECs and summaries of partnership experience to midlife health. For simplicity, we consider a single health outcome denoted by H_i. The extension to more than one health outcome is straightforward, and involves constructing a multivariate response vector.

Denote by $\mathbf{X}_i^{(H)}$ a vector of exogenous health-related predictors and $\mathbf{Z}_i^{(P)}$ a vector of summary measures of partnership stability. Examples of such summaries include the total number of partners, percentage of time spent single, the latest partnership status or the longitudinal profile of partnership transitions (Ploubidis et al., 2015). For an outcome following an exponential family distribution, a generalized linear model can be specified for $E(H_i|C_q, \mathbf{X}_i^{(H)}, \mathbf{Z}_i^{(P)}, u_i)$ with appropriate link function. Note that this model extends the earlier model described in Section 4.4 with an addition of $\mathbf{Z}_i^{(P)}$ in the covariate set. The self-reported health status in midlife is typically measured on a binary or ordinal scale. Suppose that we have a binary measure (coded 1 for poor health), then a logit model for $P_i^{(H)} = P(H_i = 1)$ can be written:

$$\operatorname{logit}\left(P_i^{(H)}\right) = \alpha_0^{(H)} + \boldsymbol{\beta}_1^{(H)'}\mathbf{Z}_i^{(P)} + \boldsymbol{\beta}_2^{(H)'}\mathbf{X}_i^{(H)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(H)} I(C_{qi} = k_q) + \lambda^{(H)}u_i, \quad (6.4)$$

where $\lambda^{(H)}$ is the effect of the individual random effect u_i on the log-odds of poor health.

$\mathbf{Z}_i^{(P)}$ is endogenous if there are shared unmeasured influences on partnership transitions and health. If these unmeasured variables are time-invariant, there will be a correlation between elements in $\mathbf{Z}_i^{(P)}$ and u_i. Suppose, for example, that individuals prone to poor health tend also to have less stable relationships and that $Z_i^{(P)}$ (scalar for the sake of illustration) is a measure of the overall partnership experience. In that case we expect $\operatorname{corr}(Z_i^{(P)}, u_i) > 0$ and, if ignored, a positive $\beta_1^{(H)}$ will be overstated and the effects of variables that are associated with $Z_i^{(P)}$ (e.g. childhood circumstances) may also be biased. To handle endogeneity due to shared unmeasured individual-level influences, we allow for non-zero residual correlations across equations. One option is to specify different but correlated random effects for each outcome. However, including multiple correlated random effects may be computationally infeasible and the model may be poorly identified when there is only one health outcome and few recurrent events. An alternative, which we adopt, is to include the same u_i across equations, but with a different coefficient for each outcome. Our specification has several advantages. First, the joint model defined by Equations (6.3) and (6.4) can be viewed as a factor model where the time-to-event outcomes and distal health are essentially indicators for the latent variable u_i. The loading for u_i in the dissolution submodel, $\lambda^{(S)}$, is fixed at one for scaling and identification purposes while the remaining loadings, $\lambda^{(F)}$ and $\lambda^{(H)}$, are freely estimated. Second, the interpretation of $\lambda^{(H)}$ is informative. If high values of H_i indicate poor health, then $\lambda^{(H)} > 0$ suggests that people whose unobserved time-invariant characteristics put them at high risk of dissolution tend also to have poorer health in later life. Finally, Section 4.5.5 and Zhu et al. (2017) noted that the estimates from the 3-step approach may be biased due to local dependence (i.e. outcomes are dependent conditional on latent classes). The inclusion of u_i can help to mitigate this problem by capturing the additional associations between outcomes due to unmeasured time-invariant characteristics.

Figure 6.1 is a graphical illustration of a simplification of the model defined by Equations (6.3) and (6.4), in step 3 of the 3-step procedure, that imposes a factor structure on the random effects. For clarity of presentation, we present a model with two associated categorical latent variables, each measured by a distinct set of repeated measures. The step 1 latent class models for C_1 and C_2 are not depicted for simplicity but we note that modal classes (M_1, M_2) and misclassification probabilities derived in steps 1 and 2 are carried forward to step 3.

The proposed multilevel SEM combines a multilevel generalized linear model (Skrondal and Rabe-Hesketh, 2004) with a multilevel factor model (Goldstein, 2010) with both continuous and categorical level 2 (individual-level) factors. Such a specification allows for both a direct (via $\mathbf{Z}_i^{(P)}$ and the latent Cs) and indirect (via the individual-level unobservables u_i) association between partnership transitions, childhood SECs and midlife health.

6.3.2 Estimation

The proposed SEM with the three submodels defined by Equations (6.3) and (6.4) can be estimated using the 3-step maximum likelihood approach both manually and automatically (using the step3 option) in LatentGOLD (Vermunt and Magidson, 2015). Manual and automatic estimation using the 3-step procedure for models with one categorical latent variable is also possible in Mplus 7.1 (Muthén and Muthén, 2017) and later versions. Latent class models in step 1 are estimated using full information maximum likelihood with the iterative EM algorithm (Dempster et al., 1977) to estimate the latent class models (Hagenaars and McCutcheon, 2002; Vermunt, 2010). Before carrying out the estimation in step 3, the input dataset needs to be restructured. Denote by $\mathbf{Y}_i = (H_i, \mathbf{y}_i^{(F)}, \mathbf{y}_i^{(S)})$ a stacked response

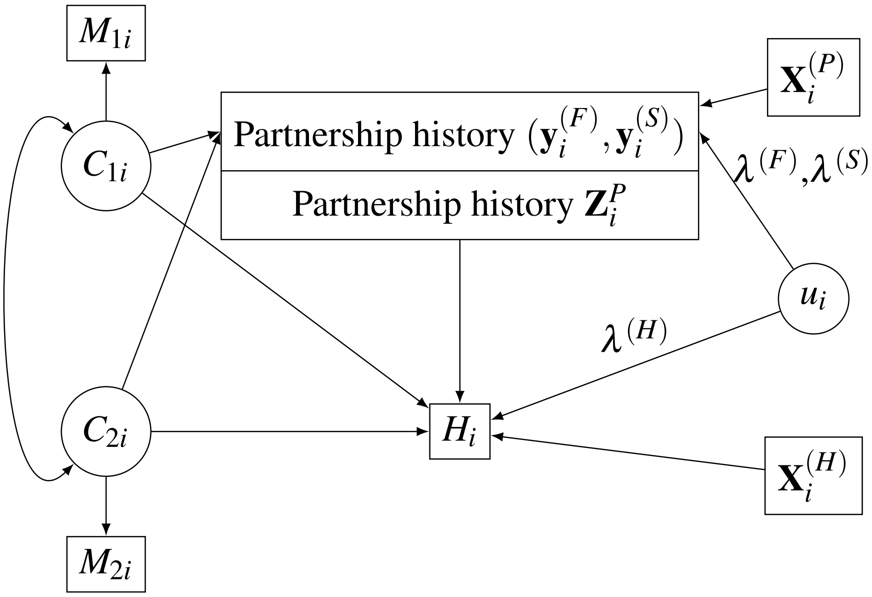

M_{1i}
C_{1i}
Partnership history $(\mathbf{y}_i^{(F)}, \mathbf{y}_i^{(S)})$
Partnership history $\mathbf{Z}_i^P$
$\mathbf{X}_i^{(P)}$
$\lambda^{(F)}, \lambda^{(S)}$
u_i
$\lambda^{(H)}$
C_{2i}
H_i
$\mathbf{X}_i^{(H)}$
M_{2i}

Fig. 6.1 A path diagram of a simplification of the multilevel SEM defined by Equations (6.3) and (6.4) with two categorical latent variables and a factor structure for the individual-level random effects. C_1 and C_2 are the latent variables for childhood SECs, M_1 and M_2 are the most likely class memberships derived in step 1 for each dimension of the childhood SECs and H_i is midlife health. $\mathbf{X}_i^{(P)}$ and $\mathbf{X}_i^{(H)}$ denotes the covariate set for partnership events and distal health respectively. The upper half of the node with stacked rectangles is the outcome variable in the partnership submodels. The lower half contains summaries of the partnership history that are used as predictors in the health submodel.

vector for partnership outcomes and binary midlife health. A binomial logit model is then fitted to the single stacked response vector with denominator $(1, \mathbf{n}_i^{(F)}, \mathbf{n}_i^{(S)})$. Explanatory variables can be included in the model by interacting them with four binary indicators that index each response. If H_i is not binary, a different link function in the health submodel is required. In that case, we can define a bivariate response vector where midlife health H_i is expanded to the same length of partnership outcomes. For H_i, only one record per individual is retained and the remaining entries (both for H_i and its predictors) are set to zero. The part of the data with both the outcome and the predictors set to zero does not contribute to the likelihood function of the model. Examples of the data structure for a binary health outcome and LatentGOLD syntax for the application are available in Appendix F.

6.3.3 Simulation study

Data generation

Before applying the proposed joint model to real data, a simulation study is conducted to evaluate the performance of the methodology under various conditions, in particular, with different combinations of sample sizes $N = (500, 2000)$ and entropy values (High, Low). In this simulation, "High" entropy is defined as 0.8 and "Low" as 0.4. This setting for four scenarios (with combined sample size and entropy values) is the same as that described in Section 4.5.1 of Chapter 4. For illustration purposes, we consider a simple example with one event history outcome of recurrent events (partnership dissolution) and a single binary distal outcome (midlife health) where both outcomes are predicted by two associated categorical latent variables C_1 and C_2, each with two categories.

Data are generated from the 3-step procedure. First, we generate binary latent variables C_1 and C_2 from the log-linear model (4.21) with a negative association (i.e. $\omega_{11}^{(C_1C_2)} = -0.5$). Five measurements for each C are then generated for high and low entropy values (more details in Section 4.5.1). Next, multilevel discrete-time event history data are generated from the second equation (but with two categorical latent variables) of the event history submodels (6.3), where the hazard is predicted by a time-invariant covariate $X_i^{(S)} \sim N(0,1)$, a time-varying covariate $X_{tij}^{(S)} \sim N(0,1)$, and C_1 and C_2. Note that the baseline hazard function is simplified to be a linear function of t. The model also allows for the influence of individual-specific random effects $u_i \sim N(0, \sigma_u^2)$. For more details on the data-generating algorithm and the simulation set-up for the event history submodels, see Section 5.6.1. In total, $T = 10$ time intervals are generated for each individual. Time intervals are not grouped in the simulated dataset and a Bernoulli distribution is assumed for the binary response. We next generate a distal health outcome. As health measures (e.g. general health status, previous history of hospital admissions) in later waves of the 1958 NCDS are mostly categorical variables, for simplicity, we generate a binary health outcome from the logit model (6.4) (but with two categorical latent variables). Distal health (H_i) is predicted by a covariate $X_i^{(H)} \sim N(0,1)$, a summary indicator of the event history $Z_i^{(P)}$ (total number of episodes per individual i), C_1 and C_2. Also included in the health submodel is the time-invariant random effect u_i where a differential effect on health $(\lambda^{(H)})$ is allowed for. The joint model used to generate the data is:

$$\text{logit}\left(h_{tij}^{(S)}\right) = \beta_0 + \beta_1 t + \beta_2 X_i^{(S)} + \beta_3 X_{tij}^{(S)} + \sum_{q=1}^{2} \tau_{C_q,1}^{(S)} I(C_{qi} = 1) + u_i, \tag{6.5}$$

$$\text{logit}\left(P(H_i = 1)\right) = \alpha_0 + \alpha_1 X_i^{(H)} + \alpha_2 Z_i^{(P)} + \sum_{q=1}^{2} \tau_{C_q,1}^{(H)} I(C_{qi} = 1) + \lambda^{(H)} u_i, \tag{6.6}$$

where $u_i \sim N(0, \sigma_u^2)$ and the identification constraints are $\tau_{C_1,2}^{(S)} = \tau_{C_2,2}^{(S)} = \tau_{C_1,2}^{(H)} = \tau_{C_2,2}^{(H)} = 0$.

We generate a total of 500 imputed datasets for each of the four combinations of sample sizes and entropy values. Models are estimated by the 3-step ML procedure (using the step3 option with ml as a sub-option) available in LatentGOLD 5.1. In this procedure, information on the modal classes and misclassification probabilities is automatically merged with the original simulated datasets after the first two steps, for ease of access in step 3.

Simulation results

Following suggestions in Burton et al. (2006) and similar to the statistics reported in Chapter 4, we use the following key statistics to evaluate the performance of the 3-step method for the joint model. Table 6.1 to Table 6.4 present the results of estimating the SEM described above. In addition to the coefficients and standard errors, we also report the estimate of $\omega_{11}^{(C_1C_2)}$, which reflects the association between the categorical latent variables in the log-linear model (see details in Section 4.4).

As expected, for most parameters, relative bias is less than 1% and all less than 5% for the scenario with good classification (high entropy) and large sample size ($N = 2000$). Average standard errors are very close to the standard deviations and the coverage for all parameters is close to the nominal level of 95%. However, when sample size is small ($N = 500$) and entropy is low (0.4), estimates of the coefficients in the health submodel are particularly worrying, with relative bias above 10%. The large bias in coefficients is essentially a scaling effect due to the biased estimate for σ_u^2 (17.83% relative bias) because the magnitude of coefficients depends on the magnitude of the variance of random effects in a random effects generalised linear model (Yatchew and Griliches, 1985). We also find that standard errors are heavily underestimated (some by 50% from a comparison of the SE and SD) but if we look closely, when sample size increases from 500 to 2000, despite having poor classification in the measurement models, both the point estimates and standard errors improve. Across all scenarios investigated, we also find that estimates of coefficients in the event history submodel are in general less biased than those in the health submodel. Such a difference is particularly obvious in less satisfactory situations where either the sample size is small or the entropy is low (or both). This is partly due to the use of multiple observations per individual for the event history outcome while the health outcome is observed only once for each individual. Note that the identification of the health submodel (a random effects model) relies on the event history submodel where there exists more than one indicator (i.e. level-1 units) for the individual-level continuous latent variable u_i. In summary, the simulation study shows that to estimate an SEM with such a complex structure, with multiple individual-level categorical latent variables and one continuous latent variable, it is preferable to have large sample sizes and also a large number of level-1 units for model identification. More types or repeated measures of health outcomes (i.e. more lower-level indicators for higher-level latent variables) may also be useful. One should also be aware of the trade-off between the complexity of the model and computational time, discussed in Snijders and Bosker (2000).

Table 6.1 Simulation results (in log-odds) for the 3-step procedure applied to the SEM with an event history submodel and a distal health submodel: high entropy (0.8) and small sample size ($N = 500$)

Parameter	True	Relative bias (%)	SE	SD	95% Coverage
		Event history submodel			
β_0	−2.00	0.34	0.23	0.22	0.96
$\beta_1\,(t)$	1.50	0.03	0.11	0.11	0.96
$\beta_2\,(X_i^{(S)})$	1.50	0.11	0.10	0.10	0.95
$\beta_3\,(X_{tij}^{(S)})$	−0.50	0.28	0.06	0.06	0.96
$\tau_{C_1,1}^{(S)}$	2.50	0.27	0.18	0.18	0.95
$\tau_{C_2,1}^{(S)}$	−1.00	−1.07	0.17	0.18	0.95
		Health submodel			
α_0	−3.00	5.49	1.03	2.15	0.86
$\alpha_1\,(X_i^{(H)})$	−1.50	6.19	0.50	1.14	0.83
$\alpha_2\,(Z_i^{(P)})$	0.50	6.11	0.17	0.35	0.86
$\tau_{C_1,1}^{(H)}$	−2.00	4.15	0.91	1.55	0.88
$\tau_{C_2,1}^{(H)}$	1.50	2.24	0.81	1.39	0.90
$\lambda^{(H)}$	1.50	0.55	1.95	2.33	0.71
σ_u^2	1.00	−1.87	0.22	0.21	0.94
$\omega_{11}^{(C_1C_2)}$	−0.50	−0.46	0.23	0.22	0.97

Relative bias (%)= (Estimate-True) / True×100%

Table 6.2 Simulation results (in log-odds) for the 3-step procedure applied to the SEM with an event history submodel and a distal health submodel: high entropy (0.8) and large sample size ($N = 2000$)

Parameter	True	Relative bias (%)	SE	SD	95% Coverage
		Event history submodel			
β_0	−2.00	0.25	0.12	0.12	0.95
$\beta_1\,(t)$	1.50	0.26	0.06	0.05	0.96
$\beta_2\,(X_i^{(S)})$	1.50	0.07	0.05	0.05	0.96
$\beta_3\,(X_{tij}^{(S)})$	−0.50	0.25	0.03	0.03	0.96
$\tau_{C_1,1}^{(S)}$	2.50	0.01	0.09	0.09	0.95
$\tau_{C_2,1}^{(S)}$	−1.00	−0.19	0.09	0.09	0.95
		Health submodel			
α_0	−3.00	1.42	0.30	0.30	0.96
$\alpha_1\,(X_i^{(H)})$	−1.50	1.89	0.15	0.15	0.96
$\alpha_2\,(Z_i^{(P)})$	0.50	1.02	0.05	0.05	0.97
$\tau_{C_1,1}^{(H)}$	−2.00	1.34	0.25	0.24	0.97
$\tau_{C_2,1}^{(H)}$	1.50	1.64	0.22	0.22	0.96
$\lambda^{(H)}$	1.50	3.10	0.29	0.29	0.96
σ_u^2	1.00	0.06	0.11	0.11	0.94
$\omega_{11}^{(C_1C_2)}$	−0.50	−1.24	0.11	0.12	0.93

Relative bias (%)= (Estimate-True) / True×100%

Table 6.3 Simulation results (in log-odds) for the 3-step procedure applied to the SEM with an event history submodel and a distal health submodel: low entropy (0.4) and small sample size ($N = 500$)

Parameter	True	Relative bias (%)	SE	SD	95% Coverage
		Event history submodel			
β_0	-2.00	-1.91	0.31	0.38	0.88
$\beta_1\,(t)$	1.50	0.24	0.11	0.11	0.96
$\beta_2\,(X_i^{(S)})$	1.50	0.33	0.11	0.11	0.95
$\beta_3\,(X_{tij}^{(S)})$	-0.50	0.03	0.06	0.06	0.96
$\tau_{C_1,1}^{(S)}$	2.50	-3.65	0.25	0.27	0.94
$\tau_{C_2,1}^{(S)}$	-1.00	-2.95	0.31	0.32	0.94
		Health submodel			
α_0	-3.00	11.24	0.81	1.47	0.97
$\alpha_1\,(X_i^{(H)})$	-1.50	13.20	0.41	0.95	0.95
$\alpha_2\,(Z_i^{(P)})$	0.50	11.16	0.13	0.24	0.98
$\tau_{C_1,1}^{(H)}$	-2.00	12.49	0.66	1.14	0.98
$\tau_{C_2,1}^{(H)}$	1.50	12.86	0.57	0.89	0.97
$\lambda^{(H)}$	1.50	22.51	0.81	1.60	0.95
σ_u^2	1.00	17.83	0.32	0.40	0.89
$\omega_{11}^{(C_1C_2)}$	-0.50	-11.84	0.52	0.51	0.97

Relative bias (%)= (Estimate-True) / True×100%

Table 6.4 Simulation results (in log-odds) for the 3-step procedure applied to the SEM with an event history submodel and a distal health submodel: low entropy (0.4) and large sample size ($N = 2000$)

Parameter	True	Relative bias (%)	SE	SD	95% Coverage
		Event history submodel			
β_0	−2.00	−0.66	0.15	0.17	0.92
$\beta_1\,(t)$	1.50	0.09	0.06	0.05	0.96
$\beta_2\,(X_i^{(S)})$	1.50	−0.07	0.06	0.06	0.94
$\beta_3\,(X_{tij}^{(S)})$	−0.50	0.30	0.03	0.03	0.95
$\tau_{C_1,1}^{(S)}$	2.50	−0.52	0.12	0.12	0.94
$\tau_{C_2,1}^{(S)}$	−1.00	−0.79	0.15	0.15	0.94
		Health submodel			
α_0	−3.00	0.21	0.43	0.42	0.93
$\alpha_1\,(X_i^{(H)})$	−1.50	−0.24	0.20	0.20	0.92
$\alpha_2\,(Z_i^{(P)})$	0.50	−0.57	0.07	0.06	0.92
$\tau_{C_1,1}^{(H)}$	−2.00	−1.37	0.39	0.38	0.93
$\tau_{C_2,1}^{(H)}$	1.50	0.37	0.36	0.35	0.96
$\lambda^{(H)}$	1.50	−3.70	0.41	0.41	0.90
σ_u^2	1.00	3.13	0.15	0.17	0.93
$\omega_{11}^{(C_1C_2)}$	−0.50	−3.39	0.26	0.26	0.96

Relative bias (%)= (Estimate-True) / True × 100%

6.4 Application: a study of the effects of childhood SECs on midlife health, mediated by partnership transitions in adulthood

After a discussion of the performance of the 3-step ML approach for estimating a multilevel SEM across four scenarios in a simulation study, we apply the methodology to a real example. Recalling the research objective set out in Section 6.1, we are interested in understanding how childhood SECs influence midlife health and to what extent the effects are mediated by partnership transitions (which is one example of the life events throughout adulthood). The specification of the SEM allows us to further explore if there is additional indirect association between partnership experiences and later health, that are due to shared influences of unobserved time-invariant characteristics of cohort members that are not explained by the variables derived from the partnership history.

The outcome variables (partnership transitions and the midlife health) and their respective set of predictors are described in Chapter 3, along with descriptive statistics. The rationale for the choice of each covariate set is set out in Section 3.4 for the partnership transitions and Section 3.5 for the midlife health. The proposed SEM that is fitted to the data is defined by Equations (6.3) and (6.4) and the analysis sample contains $N = 7,313$ individuals with complete partnership histories (i.e. provided valid responses on partnership information at wave 8, age 50).

To better understand the mechanism through which childhood SECs influence health in midlife, we build the SEM gradually. Model 1 is a baseline model where midlife health is predicted by childhood SECs, controlling for gender and early health state at age 16 (i.e. the model described in Section 4.6). Model 2 extends Model 1 by bringing in partnership outcomes (formation and dissolution). It jointly models the partnership transitions and midlife health, predicted by childhood SECs, where summaries of partnership experiences are included as predictors for distal health. Model 2 further allows $\lambda^{(H)}$ and $\lambda^{(F)}$ to be unconstrained to account for endogenous summaries of partnership stability in the health submodel and the endogenous "age at the start of partnership" in the dissolution submodel because formation and dissolution processes may have unmeasured shared influences of u_i. We also fix $\lambda^{(S)}$ at 1 and let σ_u^2 be freely estimated, as discussed in previous sections. Note that we have also considered the intermediate model that is a simpler version of Model 2 where $\lambda^{(H)}$ is constrained at zero (i.e. residual association between the health and partnership submodels are zero). Results are very similar to those from Model 2 (mostly because in Model 2, $\lambda^{(H)}$ is non-significant). For the convenience of comparing models discussed in this chapter and in Chapter 7 with similar specifications for the random effects, we report below the results from the more flexible Model 2.

Estimation of the entire SEM starts with estimating four latent class models where the categorical latent variables summarise the longitudinal information (composite indices as repeated measures in waves 0-3 in childhood) on four dimensions of childhood SECs: father or male head's social class, material hardship, financial difficulty and family structure. More details on the composite indices and the results of the latent class models are given in Section 3.2 and Section 4.6, respectively. Modal classes and misclassification probabilities derived from each of the latent class models are retained and used in the model of step 3. Following the methodology discussed in Section 6.3.1, in step 3, we estimate an SEM composed of three submodels: the distal health submodel, the partnership formation submodel and the partnership dissolution submodel. Residual correlation across three models is allowed for via a common set of individual-specific unobservables with differential effects on each outcome. The entire SEM is estimated in LatentGOLD 5.1 using maximum likelihood estimation where the random effects are integrated out using numerical Hermite-Gaussian quadrature. The 3-step procedure is implemented manually and the syntax is provided in Appendix F).

Table 6.5 presents the results for fitting the multilevel SEM using the extended 3-step approach. Results for the modal class approach are not reported as it can produce spurious results due to misclassifications (shown in the simulation study of Chapter 4). In terms of the application, in Section 4.6, we fitted a model to investigate the effects of childhood SECs on midlife health, which is essentially the Model 1 described above. Results suggest that despite having good class separations in the latent class models (step 1), the estimated effects of childhood SECs do differ if the modal class approach is used, even for the simplest model. Considering the limitation of the modal class approach and the need to have a clear interpretation of the results, the following discussions are based on findings from the 3-step ML approach. After steps 1 and 2, we find that entropy values for all of the latent class models are above 0.7, indicating good classifications (Section 4.6). This, coupled with the large sample size provides certain reassurance of the good performance of the general 3-step approach. We outline the key findings as follows.

Controlling for gender and early physical health measured at age 16 (represented by the indicator of overweight), Model 1 finds significant effects of three dimensions of childhood SECs – parental social class, financial difficulty and material hardship – on midlife health while the effect of family structure is non-significant. The direction of these significant parameters show that cohort members from families with unfavourable conditions on the first three aspects (i.e. with material disadvantage) are significantly more likely to be in a poor health state at age 50.

Results from Models 2 offer more insights on the mediating effects of partnership transitions. Conditional on individual-specific unobservables and holding other variables in the model constant, experiences of material difficulties (the first three dimensions of childhood SECs) seem to be both directly and indirectly (by increasing the risk of partnership instability) related to a higher likelihood of poor midlife health. The effect of family structure on midlife health seems to operate only indirectly by influencing partnership transitions which in turn affect later health. Specifically, cohort members who have experienced unstable family structures in childhood form their first partnerships earlier and have a higher risk of separation across relationships. From the health submodel (see the first panel of Table 6.5), individuals who have formed their first partnership late tend to have a lower risk to develop health issues at age 50. Moreover, conditional on the age at first partnership, cohort members who have spent a longer time single between ages 16 and 50 have an increased chance of poor health in midlife. These results suggest that unstable family structure is positively related to a higher chance of poor midlife health and the influence is only indirect. In terms of the residual associations, the significant $\lambda^{(F)}$ found in Chapter 5 is non-significant in this analysis. We caution that the interpretation of u_i is slightly different for the two models where in the model specified in Chapter 5, u_i denotes a common set of individual-level unobservables that are predictive of partnership transitions. The analysis in this chapter treats u_i as a common set of unmeasured time-invariant characteristics that explains the residual association between partnership transitions and midlife health. These two sets of unobservables are not necessarily the same, hence the interpretation of the $\lambda^{(F)}$ may differ accordingly.

We also note that these findings are for individuals with a complete partnership history that only form approximately half of full analysis sample $(N = 14,309)$. Are individuals who drop out from the study systematically different from the complete-history individuals? To what extent are our conclusions sensitive to the modelling of the dropout mechanisms? We discuss these questions in the next chapter where we further extend the proposed multilevel SEM by incorporating models for the missing data (in particular, dropouts).

Table 6.5 Estimated coefficients (in log-odds) and standard errors from the structural equation model (using the 3-step approach) for the effects of four categorical latent summaries of childhood SECs on a binary general health status at age 50, mediated by partnership transitions in adulthood.

Covariates	Model 1		Model 2	
	Est.	(SE)	Est.	(SE)
Submodel for midlife health[a]				
Intercept	−2.36**	(0.09)	−2.70**	(0.15)
Male (ref.= Female)	−0.05	(0.06)	−0.04	(0.06)
Overweight at age 16[b] (ref.= No)	0.25**	(0.07)	0.26**	(0.07)
Childhood circumstances				
Social class [c](ref.=High)				
Low	0.40**	(0.19)	0.46**	(0.11)
Medium	0.32**	(0.11)	0.31**	(0.09)
Financial difficulty (ref.=Low)				
High	0.53**	(0.21)	0.46**	(0.09)
Material hardship (ref.=Low)				
Medium	0.33**	(0.11)	0.32**	(0.08)
High	0.35**	(0.12)	0.41**	(0.09)
Family structure (ref.=Stable)				
Unstable	0.08	(0.13)	0.14	(0.13)
Partnership experience				
Total number of partners before age 50 (ref. =1)				
0			−0.12	(0.32)
2			0.04	(0.13)
3+			0.15	(0.23)
Age at first partnership[d]			−0.13**	(0.05)
Percentage time spent single[e]			1.08**	(0.37)
Random effect parameters				
σ_u^2			0.93**	(0.10)
$\lambda^{(H)}$			−0.20	(0.15)
$\lambda^{(F)}$			0.03	(0.09)

Continued on next page

Table 6.5 – continued from previous page

Covariates	Model 1		Model 2	
	Est.	(SE)	Est.	(SE)
Submodel for first entry into partnership				
Social class[c] (ref.=High)				
Low			0.11**	(0.04)
Medium			0.10**	(0.03)
Financial difficulty (ref.=Low)				
High			0.03	(0.04)
Material hardship (ref.=Low)				
Medium			0.04	(0.03)
High			0.05	(0.03)
Family structure (ref.=Stable)				
Unstable			0.15**	(0.04)
Submodel for partnership dissolution				
Social class[c] (ref.=High)				
Low			−0.13*	(0.07)
Medium			−0.04	(0.06)
Financial difficulty (ref.=Low)				
High			0.05	(0.07)
Material hardship (ref.=Low)				
Medium			−0.09*	(0.05)
High			−0.14**	(0.06)
Family structure (ref.=Stable)				
Unstable			0.23**	(0.07)

**$p < 0.05$,*$p < 0.1$

[a] General health at age 50 (binary).

[b] Binary indicator (with WHO cut-off) for overweight at age 16.

[c] Father or male head social class.

[d] Age at first relationship is centred at median and log-transformed.

[e] Percentage of time spent in single status during ages 16-50.

Chapter 7

Missing data

7.1 Introduction

Missing data is a common issue in longitudinal studies. In our analysis of the data from the 1958 NCDS, joint modelling of the partnership experiences with the distal health outcome is of interest but not all individuals have a complete partnership history for ages 16-50. Restriction of the analysis sample to those individuals with a complete history results in a loss of nearly 50% of those cohort members present at age 16. Will the selection of the complete-history individuals in the analysis sample bias the estimated effects of childhood SECs and partnership transitions on midlife health for the entire population? Is the dropout process independent (i.e. ignorable) or not (i.e. non-ignorable) of the outcomes of interest? If the latter is true, ignoring the dropout process can lead to biased estimates in the model of primary interest (midlife health) and secondary interest (time to partnership dissolution) as the complete-history individuals are systematically different from those with incomplete partnership histories. In this chapter, we explicitly model the selection process by including an additional equation for dropout in the multilevel structural equation model proposed in Chapter 6. The main questions we intend to address in this chapter are: 1) Is there an indication of a non-ignorable dropout process? 2) If dropout is informative, how does it impact the effects of childhood SECs and partnership experiences on later health? 3) Are estimates of the effects of childhood SECs sensitive to specifications of the dropout model? It is natural to model the missing data mechanism simultaneously with partnership transitions and distal health, hence extending the model proposed in Chapter 6.

Consistent with the estimation of the models proposed in previous chapters, this chapter discusses likelihood-based approaches to handle missing data. Sections 7.2 and 7.3 review the classification of missing data mechanisms of Little and Rubin (1987) and the development of methodology that handles missing data corresponding to each mechanism. Section 7.4

discusses the specifications of the dropout model in relation to the main outcomes of interest. In Section 7.5, we describe an application of the models discussed in this chapter to investigate the extent to which the dropout mechanism influences the inferences about the interrelationships between multiple dimensions of childhood SECs, partnership experience and later health.

7.2 Review of missing data mechanisms

We first review the general missing data mechanisms, with a particular focus on the relationship with our empirical application. Using the terminology of Little and Rubin (1987), we group the missing data mechanism into three categories: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). For each subject i, denote by $r\,(r = 1,2,\ldots,R)$ the wave of the survey where R is the total number of waves, $\mathbf{Y}_i^{(O)}$ a vector of observed outcomes across waves, $\mathbf{Y}_i^{(D)}$ a vector of missing (unobserved) outcomes, and $\mathbf{D}_i = \{D_{ri}\}$ a vector of dummy variables indicating wave non-response. In the following discussion, we refer to $(\mathbf{Y}_i^{(O)}, \mathbf{Y}_i^{(D)}, \mathbf{D}_i)$ the full set of responses and denote by $f(\mathbf{Y}_i^{(O)}, \mathbf{Y}_i^{(D)}, \mathbf{D}_i)$ the corresponding joint probability density function. For maximum likelihood estimation, $f(\mathbf{Y}_i^{(O)}, \mathbf{D}_i)$ is a key component in the observed data likelihood. For simplicity of illustration, covariates are suppressed in the following discussion. The individual contribution to the observed likelihood function can be written as:

$$\begin{aligned} f(\mathbf{Y}_i^{(O)}, \mathbf{D}_i) &= \int_{\mathbf{Y}_i^{(D)}} f(\mathbf{Y}_i^{(O)}, \mathbf{Y}_i^{(D)}, \mathbf{D}_i) d\mathbf{Y}_i^{(D)} \\ &= \int_{\mathbf{Y}_i^{(D)}} f(\mathbf{D}_i | \mathbf{Y}_i^{(O)}, \mathbf{Y}_i^{(D)}) f(\mathbf{Y}_i^{(O)}, \mathbf{Y}_i^{(D)}) d\mathbf{Y}_i^{(D)}, \end{aligned} \tag{7.1}$$

where the unobserved outcomes are integrated out.

Depending on the assumption about the missing data mechanism, the conditional density function of missingness, i.e. $f(\mathbf{D}_i | \mathbf{Y}_i^{(O)}, \mathbf{Y}_i^{(D)})$ can be further simplified.

MCAR

The missing data mechanism is termed "missing completely at random" if the probability of non-response is independent of both the observed and the missing outcomes (Little and Rubin, 1987). More specifically,

$$f(\mathbf{D}_i | \mathbf{Y}_i^{(O)}, \mathbf{Y}_i^{(D)}) = f(\mathbf{D}_i), \tag{7.2}$$

which results in

$$\begin{aligned} f(\mathbf{Y}_i^{(O)}, \mathbf{D}_i) &= \int_{\mathbf{Y}_i^{(D)}} f(\mathbf{D}_i) f(\mathbf{Y}_i^{(O)}, \mathbf{Y}_i^{(D)}) d\mathbf{Y}_i^{(D)}, \\ &= f(\mathbf{D}_i) f(\mathbf{Y}_i^{(O)}). \end{aligned} \tag{7.3}$$

MAR

The missing data mechanism is termed "missing at random" if the probability of non-response is related to the observed outcomes $(\mathbf{Y}_i^{(O)})$ and conditional on this information, it is independent of the missing outcomes $(\mathbf{Y}_i^{(D)})$ (Little and Rubin, 1987), i.e.

$$f(\mathbf{D}_i | \mathbf{Y}_i^{(O)}, \mathbf{Y}_i^{(D)}) = f(\mathbf{D}_i | \mathbf{Y}_i^{(O)}), \tag{7.4}$$

leading to

$$\begin{aligned} f(\mathbf{Y}_i^{(O)}, \mathbf{D}_i) &= \int_{\mathbf{Y}^{(D)}} f(\mathbf{D}_i | \mathbf{Y}_i^{(O)}) f(\mathbf{Y}_i^{(O)}, \mathbf{Y}_i^{(D)}) d\mathbf{Y}_i^{(D)}, \\ &= f(\mathbf{D}_i | \mathbf{Y}_i^{(O)}) f(\mathbf{Y}_i^{(O)}). \end{aligned} \tag{7.5}$$

Under MCAR and MAR, parameters of the model for observed outcomes and the models for missingness are distinct such that inferences for the primary outcomes of interest (i.e. $\mathbf{Y}_i^{(O)}$) are not influenced by the missingness process. These two missing data mechanisms are therefore termed "ignorable missing".

MNAR

If the distributions of the missingness indicator and the outcomes share common parameters, the conditional density function $f(\mathbf{D}_i | \mathbf{Y}_i^{(O)}, \mathbf{Y}_i^{(D)})$ cannot be simplified, leading to the mechanism termed "missing not at random". Consequently, maximum likelihood estimation for parameters in the model for $\mathbf{Y}_i^{(O)}$ cannot ignore the missing data process, a situation referred to as "non-ignorable missing". MNAR is particularly common in longitudinal studies with repeated measures. For example, in studies of human migration, failure to respond to the survey could be because of loss to follow up after a change of residence; in studies of health outcomes, individuals with poor health may tend to gradually leave the study over time.

The estimation of models assuming ignorable missing data has been well-developed and includes the full information maximum likelihood, Bayesian approaches, multiple imputation, inverse propensity weighting (e.g. the weighted generalised estimation equations method) for marginal models and more robust methods that combine the direct likelihood method with the inverse probability-weighted approaches, to cite a few examples. A detailed discussion of these approaches is available in Molenberghs et al. (2014, Chapters 2-5). However, it is impossible, both empirically and theoretically, to assess the validity of the assumption of ignorable missingness due to infinite alternative model specifications. It is for this reason that we consider effects of different specifications of MNAR models on estimates of the parameters of substantive interest.

For longitudinal data, there are two main types of missingness: monotone missingness (i.e. individuals drop out at a given time point and never return to the study) and non-monotone missingness (i.e. individuals are not present at a particular time point but later return to the study). For the partnership history obtained from the 1958 NCDS, the partnership information is recorded at each wave and is used to update each individual's partnership history. This updating feature of the data collection process eliminates intermittent non-responses on partnership experiences. For example, if an individual has information recorded in waves 1 and 3 but not in wave 2, the updating procedure fills in the missing data in wave 2 at the wave 3 interview. Therefore in this study, an individual's partnership history is incomplete (censored) only when he or she dropped out of the study and never returned, i.e. monotone missingness. We therefore focus on modelling approaches for the dropout process, in particular, non-ignorable (informative) dropout.

7.3 Review of models for informative dropout

In this section, we review models to handle dropout in longitudinal data. We discuss the strengths and limitations of each approach and justify the method that we will focus on in the rest of this chapter. Following Little (1995), models developed under the MNAR assumption can be grouped into three categories: selection models, pattern mixture models and shared parameter models. This classification is based on how the joint density function of the missingness mechanism and the primary outcome is factorised. We will focus on the first two main groups of models. The shared parameter model can be regarded as a natural extension to these models by introducing additional dependencies between the missing data process and the distribution of outcomes through latent variables (e.g by allowing for correlated residuals or random effects). For simplicity of the presentation, process-specific covariates are suppressed in the specification of the following models.

7.3.1 Selection models

The specification of a selection model was first described by Heckman (1977) in a study to correct for bias in estimates due to non-random sample selection in a cross-sectional setting. It has been adapted for modelling non-ignorable dropout in longitudinal studies with continuous and categorical outcomes (e.g. Follmann and Wu, 1995; Molenberghs et al., 1997). The contribution of individual i to the joint distribution of the missing data pattern and the observed outcomes is factorised into two parts, the marginal distribution of $\mathbf{Y}_i^{(O)}$ and the conditional distribution of $\mathbf{D}_i|\mathbf{Y}_i^{(O)}$, i.e.

$$P(\mathbf{Y}_i^O, \mathbf{D_i}) = P(\mathbf{Y}_i^{(O)})P(\mathbf{D}_i|\mathbf{Y}_i^{(O)}). \tag{7.6}$$

An obvious advantage of the selection model is that the first component of (7.6) is the marginal distribution of the observed outcomes, which is also the component in the complete-case likelihood. This makes (7.6) a natural factorisation of the joint density that helps to explore the impact of missingness on the key parameters of interest. There are two main specifications of the relationship between the selection mechanism and outcomes: the Diggle-Kenward-type model where the probability of missingness is directly dependent on the past and current values of the outcome; and the shared-parameter-type model where the connection between the missing data process and the outcomes is indirect, through shared parameters. In the following, we present models for the standard repeated measures set-up and Section 7.4 discusses how to adapt the framework to our setting.

Diggle-Kenward model

For individual i define $\mathbf{Y}_i = \{Y_{ri} : r = 1,...,R\}$ the set of observed and unobserved outcomes of interest and $\mathbf{D}_i = \{D_{ri} : r = 1,...,R\}$ a vector of binary indicators for drop out at each wave. Define $D_{ri} = 1$ if dropout occurs at wave r (i.e. an individual is present at $r-1$ but not r), $D_{ri} = 0$ if he is present at wave r or censored and, missing values for the time periods after the dropout time. For example, in a study with $R = 4$, an individual who drops out at the third wave has a 4×1 vector for missingness $(0,0,1,.)$, which is of the wide form data structure or $(0,0,1)$ for a long-form structure. Note that the outcome at r is unobserved should an individual drop out at r. Diggle and Kenward (1994) specified a logit model for the non-ignorable dropout process where previous and current outcomes are used as predictors in the dropout model. A simple form of the model with one lagged outcome can be written:

$$\text{logit}\big(P(D_{ri} = 1)\big) = \alpha_r + \beta_1 Y_{ri} + \beta_2 Y_{r-1,i}, \tag{7.7}$$

where α_r is a function of time and Y_{ri} is unobserved if $D_{ri} = 1$.

The joint model for dropout (7.7) and for the outcomes $\mathbf{Y}_i$ can be estimated via the maximum likelihood approach. Specifically, the likelihood requires integrating out the unobserved $\mathbf{Y}_i^{(D)}$. A main advantage of this model is that $\beta_1 = 0$ and $\beta_2 \neq 0$ indicates MAR while $\beta_1 \neq 0$ indicates MNAR. However, we note that estimation of this model can be computationally heavy, particularly for the numerical integrations. For longitudinal data, developments in generalised linear models have been reviewed in detail by Ibrahim and Molenberghs (2009) and Molenberghs et al. (2014). Applications using forms of the Diggle-Kenward model to handle non-ignorable dropout are discussed in Muthén et al. (2011) for growth curve models and in Washbrook et al. (2014) to model the change of residence with panel data. Model (7.7) could be further extended to have covariates or more lagged outcomes (Diggle and Kenward, 1994). Despite its attraction to directly relate the current and previous outcome values to the probability of dropout, there is one obvious drawback. This model imposes a strong assumption on the selection mechanism both in terms of the functional form (i.e. the mean and variance structures of the dropout model) and distributional assumptions. Therefore, it is not possible to conclude whether the missing data mechanism is MNAR or MAR based on the results of hypothesis tests of β_1 and β_2 because they are dependent on a specific set of modelling assumptions.

Heckman-type model

Heckman (1977) proposed a joint model for the outcome and the missing data process, allowing for an indirect association through correlated residuals. We begin with a description of the original model which was developed for a cross-sectional setting. Let D_i be a binary missing data indicator that takes value 1 if an individual i is missing and 0 otherwise. Suppose Y_i is a continuous outcome that is unobserved if $D_i = 1$. The simplified full model for the outcome and missing data process is:

$$Y_i = \begin{cases} \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_{1i} & D_i = 0 \\ . & D_i = 1 \end{cases} \tag{7.8}$$

where $\varepsilon_{1i} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2_{\varepsilon_1})$.

The selection mechanism is modelled using a probit regression, i.e.

$$\begin{cases} D_i & = I(D_i^* > 0) \\ D_i^* & = \mathbf{Z}_i'\boldsymbol{\delta} + \varepsilon_{2i} \end{cases} \tag{7.9}$$

where $I(\cdot)$ is the indicator function, $\varepsilon_{2i} \overset{\text{i.i.d}}{\sim} N(0,1)$ and $\text{corr}(\varepsilon_{1i}, \varepsilon_{2i}) = \rho$. For identification, the covariate sets $\mathbf{X}_i$ and $\mathbf{Z}_i$ can have overlapping elements but cannot be identical. Specifically $\mathbf{Z}$ should contain at least one element, commonly termed an instrumental variable in the econometrics literature, that is not in $\mathbf{X}$. This is referred to as the exclusion restriction

(Puhani, 2000). Estimation of the joint model of (7.8) and (7.9) can be achieved using a two-stage approach by first fitting a probit model for missingness, obtaining the fitted value of D_i, computing a function of the fitted value of D_i (referred to as the inverse Mill's ratio) and including that quantity as an additional covariate in the model for Y_i. Another more efficient way is to use maximum likelihood estimation where the models (7.8) and (7.9) are estimated simultaneously.

The Heckman selection model has been extended to panel data (Wooldridge, 2010, Chapter 17). Establishing the link with structural equation models, Rabe-Hesketh et al. (2004) proposed a latent variable approach that allows for non-normal outcomes. The application of Heckman-type models has also been prevalent in fields outside of econometrics (see Attanasio and Emmerson, 2003; Bärnighausen et al., 2011; Washbrook et al., 2014). This model has two main limitations. First, for the missing data process, it is difficult to find valid and strong instruments that are only predictive of non-response but not the primary outcome of interest. Second, it is unclear if and how values of the outcome influence the missingness probability, although one could argue that having the outcome directly influence the tendency to dropout does not always have a substantive justification.

Shared-parameter model

Shared parameter models were first discussed by Wu and Carroll (1988) based on the idea that an individual's response probability and value of the outcome are dependent on a common set of individual-specific characteristics, some of which are unobserved. Assuming that the process of missingness and outcomes are dependent on separate but potentially correlated sets of unobserved individual-specific characteristics, represented by a vector of individual random effects $\mathbf{u}_i = (u_{1i}, u_{2i})$, the joint probability density function can be factorised as:

$$f(\mathbf{Y}_i, \mathbf{D}_i, \mathbf{u}_i) = f(\mathbf{Y}_i|u_{1i}) f(\mathbf{D}_i|u_{2i}) f(\mathbf{u}_i), \tag{7.10}$$

where (u_{1i}, u_{2i}) follows a bivariate normal distribution with a non-zero correlation.

To a degree, the Heckman selection model is also a shared-parameter model where the models for the missing data process and the outcomes have correlated residuals. There are also other forms where the shared parameters are introduced into the model. For example, Beunckens et al. (2008) extended the models of Diggle-Kenward and Wu and Carroll (1988) to a mixture model with a latent class variable measured by a set of missing data indicators. Both the outcome trajectories and the missingness have class-specific distributions, which breaks the direct relationship between Y_{ri} and D_{ri} in the model (7.7). In simulation studies, Muthén et al. (2011) found that the model of Beunckens et al. (2008) has a lower BIC and is computationally less expensive as there is no need to integrate out unobserved outcomes.

Sensitivity analysis
As mentioned above, parametric selection models impose strong assumptions on the model for dropout and these assumptions are untestable due to an unlimited number of alternatives with various functional forms and distributions. A remedy for this is to conduct a sensitivity analysis by assessing the impact on estimated coefficients in the primary model of interest of using different specifications of the model for the dropout process. Another possibility is to check for the presence of influential subjects by introducing subject-specific perturbations in the model parameters. In the model (7.7), for example, β_1 can be replaced by β_{1i} to allow for variations across individuals. For further details see Verbeke et al. (2001) and Molenberghs et al. (2014).

7.3.2 Pattern mixture models

Pattern mixture models assume that individuals belong to subgroups with pre-defined outcome processes (e.g. Ekholm and Skinner, 1998; Glynn et al., 1986; Little, 1995). The joint distribution of outcomes and missingness can be factorised into a marginal distribution of missing patterns and a conditional distribution of outcomes given the missingness pattern, i.e.

$$P(\mathbf{Y}_i, \mathbf{D}_i) = P(\mathbf{D}_i)P(\mathbf{Y}_i|\mathbf{D}_i). \tag{7.11}$$

Therefore the pattern mixture model is essentially a mixture of different response models weighted by the dropout pattern.

Different forms of pattern-mixture models and their estimation have been discussed by Little (1995), Hogan and Laird (1997) and, for binary and longitudinal data, Ekholm and Skinner (1998). Pattern mixture models have an intrinsic identifiability problem and restrictions on parameters in the conditional distribution of $\mathbf{Y}_i|\mathbf{D}_i$ across dropout patterns need to be imposed (Molenberghs et al., 2014). Specifically, consider a study where individuals are measured for R waves. For individuals who drop out at r, denote by $P_r(\cdot)$ the corresponding joint density of outcomes. The complete-data distribution of outcomes under the pattern mixture specification can be factorised as:

$$P_r(Y_{1i}, ..., Y_{Ri}) = P_r(Y_{1i}, ..., Y_{r-1,i})P_r(Y_{ri}, ..., Y_{Ri}|Y_{1i}, ..., Y_{r-1,i}),$$

where the first term can be modelled using the observed data while the second term is unidentified. One solution is to assume that the distribution of the unobserved outcomes is the same as their complete counterparts, i.e.

$$P_r(Y_{ri}|Y_{1i}, ..., Y_{r-1,i}) = P_{R+1}(Y_{ri}|Y_{1i}, ..., Y_{r-1,i}),$$

where $P_{R+1}(\cdot)$ indicates the density of outcomes for individuals who do not drop out before the end of the study. Alternatively, we could assume the equality of neighbouring densities, i.e.

$$P_r(Y_{ri}|Y_{1i},...,Y_{r-1,i}) = P_{r+1}(Y_{ri}|Y_{1i},...,Y_{r-1,i}).$$

However, this assumption is untestable and there exist many alternative restrictions. Another inconvenience of the pattern-mixture model is that parameters (in particular, regression coefficients) of the marginal distribution of the outcomes are not readily available, but need to be averaged across missing data patterns (i.e. dropout times). The standard errors also need to be constructed, for example, following the steps described in Little (1994) for normally distributed data and Ekholm and Skinner (1998) for binary data. From a technical viewpoint, the estimation procedure could also be computationally expensive if the number of missing data patterns is large or if more general models are considered (e.g. with random effects). Also, in the pattern mixture model, it is assumed that individuals with the same missing pattern are in the same "class" with the same distribution of $\mathbf{Y}_i$, which can be regarded as a strong assumption. An extension proposed by Roy (2003) assumes the outcome is predicted by a latent class, that is a mixture of missing indicators, i.e. the D_{ri}s are included as predictors of the latent class variable. Muthén et al. (2011) adapted this model by having two associated latent class variables: one measured by the dropout pattern and the other by the outcome process. The main idea is to separate out the "dropout classes" and the "outcome classes" that are implicitly combined in the original model of Roy (2003).

7.3.3 Comparison between models

As discussed in (Diggle et al., 2002, Chapter 13), each model, either in the class of selection models or the pattern mixture models, has its own merits and disadvantages. Given that each model makes untestable assumptions, it is more helpful to use the application to motivate the choice of model, rather than base the choice solely on theoretical grounds. Essentially, both types of model are reduced forms of the saturated model where $\mathbf{Y}_i$, $\mathbf{D}_i$ and $\mathbf{u}_i$ are all somehow connected. The choice of where to place the constraints, potentially motivated by the application, leads to the distinctions between these models.

In our context, the primary outcome is midlife health status, the intermediate outcome is the time to partnership dissolution and the missingness is due to dropout in the collection of partnership histories. We are mainly interested in answering two research questions: 1) What causes the dropout? In particular, are childhood SECs associated with dropout? 2) Do assumptions about the dropout mechanism influence the parameter estimates in the health model? As the MAR and MNAR mechanisms are impossible to distinguish empirically, assuming MAR, as in many previous studies in this field, may be overly simplistic. We therefore adopt a more general framework that takes into account MNAR. In order to compare results from the models with and without consideration of the dropout process, a natural choice is the selection model where the parameters of interest can be directly obtained. Also, the identification constraints are more straightforward than those needed in the pattern mixture model. Another approach is to incorporate the indirect associations between the dropout mechanism, the partnership process and midlife health, due to unobserved individual-specific characteristics. This motivates our focus on selection models with shared parameters. To address question 2), sensitivity analyses can be conducted by making alternative assumptions about the functional form of the dropout model.

7.4 Modelling informative dropout in the multilevel SEM

Health outcomes at age 50 were not recorded for individuals who were not present at the time of data collection (i.e. dropout at or before age 50). Are those individuals with a complete partnership history (i.e. present at age 50) systematically different from their counterparts with partial or no partnership information in important but unmeasured characteristics? Among the models for informative dropout reviewed in Section 7.3, the general framework of selection models is appealing as our main focus is on inferences about parameters in the model for the primary outcome, rather than for subpopulations with the same missing data patterns (as in the pattern mixture model). In the following discussion, in addition to individuals with complete partnership histories, we also retain individuals with partial partnership information until dropout.

Motivated by the work of Washbrook et al. (2014) who developed a bivariate probit model and a Diggle-Kenward type model (referred to as a "direct dependence" model) for the joint modelling of missing and outcome processes, we consider two forms of selection: indirect selection on time-invariant unobservables; and direct selection on the primary outcome. The former type of selection is allowed for via residual correlation between the health model and the selection model. The direct selection model, on the other hand, allows the probability of dropout to depend directly on health outcomes measured at the dropout time. The imposed causal relationship is supported by substantive evidence from several studies that have suggested a direct relationship between midlife health state and the tendency to dropout from a longitudinal study (Zhou et al., 2014, Chapter 7). In addition to the test of the hypothesised relationship between dropout and later health, these different model specifications also serve as a means of sensitivity analysis. It has been widely discussed in the selection model literature that the estimated coefficients in the primary models of interest could be sensitive to the specification of both the dropout model and its relationship with the main outcomes (Diggle et al., 2002, Chapter 13). In this section, we propose three different specifications of the dropout model. The first two are models with indirect selection through latent variables, where the first model uses a single dropout indicator while the second uses a vector of discrete-time dropout indicators. The last model allows for direct selection.

The consideration of covariate exclusion restrictions is also necessary for model identification and the stability of the estimator to misspecification of the error distribution. This has been discussed by various researchers, including a comprehensive account in Wooldridge (2010, Chapter 17) and in Washbrook et al. (2014). The most widely discussed exclusion criteria in the Heckman-type selection model requires including variables that are predictive of the selection probability, but not the outcome of interest. For the application of the direct dependence model to panel data, identification using instruments in the primary model of interest (i.e. variables that are only predictive of the outcome but not the probability of dropout) has also been suggested by Hirano et al. (2001) and Washbrook et al. (2014). The inclusion of both instruments in the selection model and the outcome model may improve parameter estimates but invalid instruments, coupled with potential model misspecification (e.g. parametric assumptions about the error distribution) may lead to unstable or even biased estimates.

All models described below consist of the three equations for partnership formation, dissolution and midlife health from models (6.3,6.4) in Chapter 6, with the categorical latent variables summarising childhood SECs and other observed outcome-specific covariates. We now extend this SEM to include a fourth equation for the dropout process.

7.4.1 Modelling approaches for outcomes with non-ignorable missingness

Following the notation used in previous chapters, denote by $\mathbf{Y}_i^{(F)} = \{Y_{si}^{(F)} : s = 1, ..., S_i\}$ a vector of binary time-to-event outcomes for the first partnership formation, $\mathbf{Y}_i^{(D)} = \{Y_{tij}^{(D)} : j = 1, 2, ...J_i;\ t = 1, ..., T_{J_i}\}$ a vector of time-to-separation for recurrent partnerships. The corresponding discrete-time hazards of partnership formation and dissolution are $\mathbf{h}_{si}^{(F)}$ and $\mathbf{h}_{tij}^{(D)}$, as defined in Section 6.3.1. $\mathbf{Z}_i^{(P)}$ is a vector of summary variables derived from the partnership history that predicts H_i, the midlife health state. For simplicity, we consider a binary outcome since health is often measured on categorical scales and commonly reduced to a binary variable in analysis. We next define variables in the dropout model. Using information on the data collection process of the 1958 NCDS, we notice that only the individuals who are present at wave 8 (age 50) could have complete partnership histories. This is because records of partnership experience at each wave are updated using the last available record (more details see Chapter 3). We therefore consider two types of dropout indicators: a summary indicator D_i and an indicator for dropout at wave r, D_{ri}. Let D_i be a binary variable for complete (0) or incomplete (1) partnership histories for ages 16-50 with $P_i^{(D)} = P(D_i = 1)$.

In the models below, covariates specific to a process are indicated using the equation-specific superscripts. Latent childhood SECs, which are predictors of all processes, are denoted by C_{qi} for the q^{th} $(q = 1, \dots, 4)$ categorical latent variable with K_q categories.

Model 1: indirect selection (D_i)

In this model, we allow for dependency of the partnership outcomes, midlife health and dropout on a common set of time-invariant unobservables. The primary models of interest are specified using a logit link for the probability of experiencing the corresponding partnership event or being in the poor health state, similar to previous specifications in Section 6.3.1. We model the probability of having an incomplete partnership history ($D_i = 1$) by fitting a logit model. In this system of four models, a common set of time-invariant unobservables (captured by individual-level random effects u_i) is included in each submodel to account for shared unmeasured influences, where a differential influence (through λs) is allowed for. Specifically, denote by subscripts s, t for time intervals, we write

$$\begin{cases} \text{logit}\left(h_{si}^{(F)}\right) = \alpha_s^{(F)} + \boldsymbol{\beta}^{(F)\prime}\mathbf{X}_{si}^{(F)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(F)} I(C_{qi} = k_q) + \lambda^{(F)}u_i \\ \text{logit}\left(h_{tij}^{(S)}\right) = \alpha_t^{(S)} + \boldsymbol{\beta}^{(S)\prime}\mathbf{X}_{tij}^{(S)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(S)} I(C_{qi} = k_q) + u_i \\ \text{logit}\left(P_i^{(H)}\right) = \alpha_0^{(H)} + \boldsymbol{\beta}_1^{(H)\prime}\mathbf{Z}_i^{(P)} + \boldsymbol{\beta}_2^{(H)\prime}\mathbf{X}_i^{(H)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(H)} I(C_{qi} = k_q) + \lambda^{(H)}u_i \\ \text{logit}\left(P_i^{(D)}\right) = \alpha_0^{(D)} + \boldsymbol{\beta}^{(D)\prime}\mathbf{X}_i^{(D)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(D)} I(C_{qi} = k_q) + \lambda^{(D)}u_i \end{cases} \tag{7.12}$$

where superscripts (F) and (S) denote partnership formation and dissolution processes, respectively, and $u_i \sim N(0, \sigma_u^2)$ is an individual-specific random effect for a common set of individual-level unobservables.

In model (7.12), because u_i is common to all outcomes, non-zero coefficients $\lambda^{(F)}$, $\lambda^{(D)}$ and $\lambda^{(H)}$ imply that missingness is non-ignorable as it is associated with the health outcome beyond that captured by observed quantities (note that $\lambda^{(S)}$ for the dissolution submodel is fixed at 1 for identification). Focusing on the residual part of the model, the significance of $\lambda^{(D)}$ signals a selection effect in the tendency of partnership dissolution. For example, $\lambda^{(D)} > 0$ indicates that cohort members with unobserved characteristics that place them at a higher probability to dropout tend also to be more likely to have unstable partnerships. This model fits naturally within the current multilevel SEM framework using the shared parameter specification but the information on dropout time is not used. This leads to our Model 2, where the outcome in the dropout model is time to dropout.

Model 2: indirect selection (D_{ri})

We now employ a slightly different specification of the dropout model where the dropout outcome is coded as a discrete time-to-dropout indicator D_{ri} for $r = 1$ (wave 5 at age 33), 2 (wave 6 at age 42), 3 (wave 7 at age 46) and 4 (wave 8 at age 50). Information collected at these four adulthood waves in the NCDS contribute to the linked partnership history. The updating feature of the partnership history data indicates that if an individual is not present at wave r but present at wave $r+1$, the partnership situation between waves $r-1$ and $r+1$ will be filled in. To model the dropout probability at each wave, the commonly used logit link is applied. Wave-specific dummies, indexed by subscript r are used to specify a piecewise-constant logit baseline hazard. To further allow for time-varying effects of predictors, interactions of wave dummies with covariates can be included. Considering the complexity of the current model with four categorical latent variables and random effects, we assume time-invariant effects for simplicity. Denote by $P_{ri}^{(D)} = P(D_{ri} = 1 | D_{r'<r,i} = 0)$ and $\alpha_r^{(D)}$ the wave-specific logit baseline hazard. Modifying the fourth equation in model (7.12), the full SEM that jointly models the outcomes of interest and the dropout mechanism is

$$\begin{cases} \text{logit}\left(h_{si}^{(F)}\right) = \alpha_s^{(F)} + \boldsymbol{\beta}^{(F)'}\mathbf{X}_{si}^{(F)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(F)} I(C_{qi} = k_q) + \lambda^{(F)} u_i \\ \text{logit}\left(h_{tij}^{(D)}\right) = \alpha_t^{(D)} + \boldsymbol{\beta}^{(D)'}\mathbf{X}_{tij}^{(D)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{k_q}^{(D)} I(C_{qi} = k_q) + u_i \\ \text{logit}\left(P_i^{(H)}\right) = \alpha_0^{(H)} + \boldsymbol{\beta}_1^{(H)'}\mathbf{Z}_i^{(P)} + \boldsymbol{\beta}_2^{(H)'}\mathbf{X}_i^{(H)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(H)} I(C_{qi} = k_q) + \lambda^{(H)} u_i \\ \text{logit}\left(P_{ri}^{(D)}\right) = \alpha_r^{(D)} + \boldsymbol{\beta}^{(D)'}\mathbf{X}_i^{(D)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(D)} I(C_{qi} = k_q) + \lambda^{(D)} u_i \end{cases} \tag{7.13}$$

Model 3: direct selection

So far, Models 1 and 2 allow only for indirect selection through unobserved individual-specific characteristics. It is also possible that the time to dropout depends directly on the health state at the time of interview. We now allow for a direct selection mechanism by imposing a causal relationship between primary outcomes of interest and the probability of dropout, similar to the Diggle-Kenward model (Diggle and Kenward, 1994). With this specification, the random effect is not needed in the dropout submodel as we allow for the dropout probability at wave r to directly depend on the previous partnership outcomes and the health situation at wave r. For simplicity, we also assume time-invariant effects of covariates but the model can be extended straightforwardly to allow for time-varying effects by including interactions between the wave dummies and the covariates of interest. Similar to previous practices, modifying the fourth equation in model (7.13) gives us the full SEM written as:

$$\begin{cases} \text{logit}\left(h_{si}^{(F)}\right) = \alpha_s^{(F)} + \boldsymbol{\beta}^{(F)'}\mathbf{X}_{si}^{(F)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(F)} I(C_{qi}=k_q) + \lambda^{(F)}u_i \\ \text{logit}\left(h_{tij}^{(D)}\right) = \alpha_t^{(D)} + \boldsymbol{\beta}^{(D)'}\mathbf{X}_{tij}^{(D)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(D)} I(C_{qi}=k_q) + u_i \\ \text{logit}\left(P_i^{(H)}\right) = \alpha_0^{(H)} + \boldsymbol{\beta}_1^{(H)'}\mathbf{Z}_i^{(P)} + \boldsymbol{\beta}_2^{(H)'}\mathbf{X}_i^{(H)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(H)} I(C_{qi}=k_q) + \lambda^{(H)}u_i \\ \text{logit}\left(P_{ri}^{(D)}\right) = \alpha_r^{(D)} + \boldsymbol{\beta}_1^{(D)}\mathbf{Z}_{r-1,i}^{(P)} + \beta_2 H_{ri} + \boldsymbol{\beta}_3^{(D)'}\mathbf{X}_{r-1,i}^{(D)} + \sum_{q=1}^{4}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q}^{(D)} I(C_{qi}=k_q) \end{cases} \tag{7.14}$$

Consistent with the specification of models for other time-to-event outcomes in Chapter 6, model (7.14) assumes a wave-specific logit baseline hazard $(\alpha_r^{(D)})$. $\mathbf{Z}_{r-1,i}^{(P)}$ is a vector of summaries of the partnership history up to wave $r-1$ and H_{ri} denotes health state at wave r. Note that when dropout occurs at wave r, H_{ri} is missing and hence needs to be integrated out in the likelihood-based estimation. Finally, in addition to childhood SECs, we include in the model $\mathbf{X}_{r-1,i}^{(D)}$, a vector of time-varying and time-invariant covariates collected at wave $r-1$.

In this model, the significance of β_2 suggests a direct effect of health on the likelihood of dropout. The dropout submodel in model (7.14) can be further extended to allow simultaneously for direct and indirect selection. For example, although not on modelling missing data, Steele et al. (2013) considered both types of selection by specifying a joint model for employment transitions and health, allowing for the influence of earlier health on subsequent employment transitions and later health, and also correlated residuals across equations.

Before proceeding to the discussion of estimation, we note that although the above-mentioned models serve as a tool for sensitivity analysis, alternative specifications are unlimited, both in terms of the functional form of the model, the parametric assumptions about the distribution of error and random effects, and the structure of residuals across equations in the SEM. Future work is needed to understand the effects of different combinations of these factors on the bias and variance of estimated coefficients.

7.4.2 Modelling approaches for outcomes with non-ignorable missingness and predictors with missing values

The models discussed above assume a scenario where the outcomes have missing values (due to dropout) where the covariate set (predictors) is complete. However, we note that for dropouts, not only the outcome, but also time-varying predictors shall contain missing values (time-invariant predictors will not be influenced). This causes a series of problems if we work with a sample with both complete and incomplete observations. Among the individuals with incomplete partnership histories, in the partnership submodels, event indicators for partnership transitions contain missing values as censoring time is essentially the dropout time (assumed informative); it also generates missing data in the corresponding time-varying predictors of partnership events (e.g. number of pre-school children, type of partnership and age at the start of each partnership).

Turning to the health submodel, for individuals who have dropped out early, the health state at age 50 is unavailable and so are the partnership summaries (elements of $\mathbf{Z}_i^{(P)}$), which are derived from the 34-year partnership history. One option is to impute covariates with missing values (i.e. $\mathbf{Z}_i^{(P)}$ and $\mathbf{X}_i^{(P)}$) where $\mathbf{X}_i^{(P)}$ denotes the combined set of time-varying covariates in the two partnership submodels. However, these variables are probably associated with childhood SECs, the latent variables in the model, which complicates the imputation procedure. For instance, one may use indicators for childhood SECs, or the modal classes as proxies to impute covariates with missing values but these proxies are subject to measurement error and may themselves contain missing values. As there are a number of covariates that require imputation, misspecification of the imputation model for the multiple imputation procedure may lead to greater bias. More importantly, the degree of misspecification is difficult to assess empirically. Moreover, the imputation of time-varying predictors (e.g. age at the start of each partnership, number of pre-school children) of partnership events can be challenging, both methodologically and practically, as these variables are nested within each time period of an episode of each individual. Therefore, for the individuals who drop out early, we need to impute, for example, the total number of episodes, the duration of each episode and the time gap between episodes before generating time-varying quantities that respects the temporal order. The hierarchical data structure creates a higher degree of complexity of the imputation procedure for a scenario with missingness both in outcomes and time-varying predictors. Further research on this topic would require the proposition of modelling approaches and simulation studies, which is beyond the scope of this thesis. Despite these difficulties (mainly with regards to missing values in time-varying predictors), we make a first attempt in this chapter by conducting a small simulation study to evaluate the performance of different approaches proposed in the literature to handle missing data both in covariates and outcomes due to dropout. In this attempt, we assume the model for partnership transitions does not have time-varying predictors and the time-invariant predictors have missing values.

Methods to handle missing covariate values have been discussed in Bartlett et al. (2014) and White and Carlin (2010). They found that in situations with a complete outcome and incomplete covariates, the complete-case analysis (CCA) and the multiple imputation (MI) methods are both consistent if missingness is completely at random. For non-MCAR scenarios, CCA is consistent if the missingness is 1) independent of both the outcome and the covariate with missing values (MAR) or 2) conditionally independent of the outcome given the covariate with missing values (MNAR); MI is consistent if the assumption of MAR about the incomplete covariate is valid. In our study, we assume the missing data mechanism is non-ignorable and that, beyond the association captured by fully observed covariates, the missingness is directly or indirectly (through u_i) associated with the outcomes of interest. Consequently, both MI, or any other methods based on the assumption of MAR, such as the inverse-propensity-weighting approach, and CCA will produce biased estimates. An alternative is to leave these partnership summaries as missing but model the missingness mechanism explicitly and jointly with the main outcomes of interest.

The following simulation study aims to investigate the performance of multiple approaches that handle the situation where we have both level-1 (time units) and level-2 (individual-level) data that are subject to non-ignorable missingness. As a result of dropouts, some covariates and the outcomes collected at a later time point are both missing. Denote by $\boldsymbol{\theta}$ the set of true population parameters, i the subject indicator, u_i the subject-specific random effect and (Z_{1i}, Z_{2i}) the fully observed covariates. We generate outcomes data for five waves from (7.15) in Algorithm 2. Consider the dropout indicators generated from Equation (7.17) in Algorithm 2, as an example. Denote by $\mathbf{D}_i = \{D_{ri} : r = 1, ..., 5\}$ a vector of discrete-time dropout indicators, H_i the binary health outcome, X_{ri} a binary variable indicating the presence of a partner during the time interval $[r, r+1)$, X_i the total number of partners and V_i a continuous variable that is highly associated with Y_i (polychoric correlation coefficient $= 0.7$). Algorithm 2 summarises the data generating process.

Algorithm 2: Simulation study: data generating process for a longitudinal setting where both the outcome and a subset of covariates are missing from the dropout time onwards. The missing data mechanism is non-ignorable.

```
Input $\boldsymbol{\theta}$
Generate $(Z_{1i}, Z_{2i}) \sim BVN(\binom{0}{0}, \begin{pmatrix}1 & \\ 0.5 & 1\end{pmatrix})$ and $u_i \sim N(0,1)$
Generate binary $X_{ri}$ from the model:
    $\text{logit}\,(P(X_{ri}=1)) = \beta_0^{(P)} + \beta_1^{(P)} Z_{1i} + u_i, r = 1, \ldots, 5$    (7.15)
    Generate $X_i = \sum_{r=1}^{5} X_{ri}$
Generate $H_i$ from the model:
    $\text{logit}\,(P(H_i=1)) = \beta_0^{(H)} + \beta_1^{(H)} X_i + \beta_2^{(H)} Z_{1i} + \lambda^{(H)} u_i,$    (7.16)
    Generate $V_i = H_i + \varepsilon_i$, where $\varepsilon_i \sim N(0,1)$
Generate the discrete-time dropout indicator $D_{ri}$ from the model:
    $\text{logit}\,(P(D_{ri}=1)) = \beta_0^{(D)} + \beta_1^{(D)} H_i + \beta_2^{(D)} Z_{1i} + \beta_3^{(D)} Z_{2i} + \beta_4^{(D)} X_i + \lambda^{(D)} u_i.$    (7.17)
    such that an individual that drops out at $r = 3$, for example, is assigned a vector of
    missing indicators $\mathbf{D}_i = (0, 0, 1, ., .)$.
Generate $D_i = I(D_{ri} = 1, \text{for any } r)$, a single binary indicator for missingness.
Set $X_{ri}$ at missing for the dropout time and onwards.
Set $X_i$ and $H_i$ at missing if $D_{ri} = 1$ for any $r$.
```

In the generated dataset, Z_{1i}, Z_{2i}, V_i and the single binary indicator D_i are fully observed while H_i, X_i and X_{ri} are partially observed with missing values. The data are generated assuming that the non-ignorable missingness arises from both a direct association between the outcome H_i, and an indirect association through the shared unobserved factors captured by u_i. We set $N = 15,000$ (close to the size of the empirical data), $\beta_0^{(P)} = \beta_0^{(H)} = \beta_0^{(D)} = \beta_3^{(D)} = 1$, $\beta_1^{(P)} = \beta_2^{(H)} = \beta_2^{(D)} = 0.2$ and $\lambda^{(D)} = \lambda^{(H)} = 1.2$ where the parameters for the missing mechanism varies for each scenario (discussed below). Motivated by White and Carlin (2010), three potential scenarios of MNAR are considered in this study:

1. Missingness depends on Z_1 and Z_2 and u (i.e. $\beta_1^{(D)} = \beta_4^{(D)} = 0$). MNAR is only driven by the indirect residual associations through u.

2. Missingness depends on Z_1, Z_2, X and u (i.e.$\beta_1^{(D)} = 0$).

3. Missingness depends on Z_1, Z_2, H, X and u.

If not explicitly reset to zero, we set $\beta_1^{(D)} = \beta_4^{(D)} = 0.5$. For each scenario, the corresponding parameters lead to roughly 40% missing values.

We next consider three modelling approaches in the presence of missing data:

CCA: Complete-case analysis for H_i (individuals with missing values on X_i or H_i are excluded)

MI (Full conditional specification): Missing values in X_i and H_i are jointly imputed following the multiple imputation procedure (Little and Rubin, 1987) with a full conditional specification that conditions on a set of complete covariates Z_{1i}, Z_{2i} and V_i. Note that V_i is not present in the analysis model and hence is used as an auxillary variable in the imputation model.

Multilevel SEM: The outcomes are stacked into a response vector $(\mathbf{D}_i, H_i, \mathbf{X}_i)$ and all models are estimated simultaneously, allowing for the shared influence of random effects u_i with differential effects on each outcome. Note that for model identification, we constrain the random effect coefficient at 1 in the submodel for $\mathbf{X}_i$. Full information maximum likelihood estimation is performed to estimate this SEM.

For each scenario, 50 replications of $N = 15,000$ are generated and all models are estimated in LatentGOLD 5.1 (Vermunt and Magidson, 2015) with maximum likelihood approach. Results are reported in Table 7.1. Note that we do not report the coverage rates for other approaches because they produce biased point estimates.

Table 7.1 Simulation results for three scenarios of missing not at random (consider only models with time-invariant predictors): comparison of three modelling approaches (CCA, MI (FCS) and multilevel SEM). Estimates for parameters in the model for H_i are reported. The first three columns report relative bias (in %) and last column reports coverage rates of the nominal 95% confidence interval for the estimates for the multilevel SEM.

Parameters	CCA	MI (FCS)	SEM	Coverage (SEM)
Scenario 1: MNAR (Missingness depends on Z_1 and Z_2 and u)				
$\beta_0^{(H)}$	57.26	5.83	−0.15	95.1
$\beta_1^{(H)}\,(X_i)$	22.13	−4.39	1.28	94.9
$\beta_2^{(H)}\,(Z_{1i})$	−25.39	−31.49	−2.61	93.9
$\lambda^{(H)}$			2.81	95.4
Scenario 2: MNAR (Missingness depends on Z_1, Z_2, X and u)				
$\beta_0^{(H)}$	52.42	22.13	4.30	95.2
$\beta_1^{(H)}\,(X_i)$	24.42	8.65	2.58	95.1
$\beta_2^{(H)}\,(Z_{1i})$	−21.46	−21.77	−2.33	94.9
$\lambda^{(H)}$			−1.02	93.8
Scenario 3: MNAR (M depends on Z_1, Z_2, H, X and u)				
$\beta_0^{(H)}$	51.23	14.95	4.60	95.3
$\beta_1^{(H)}\,(X_i)$	20.93	2.51	3.02	94.8
$\beta_2^{(H)}\,(Z_{1i})$	−24.34	−25.78	−5.24	93.6
$\lambda^{(H)}$			−1.16	95.1

Relative bias (%)= (Estimate-True) / True×100%

Coverage rate (%) is computed for the nominal 95% confidence interval

Across all scenarios, the joint modelling approach is superior to all other methods. As expected, the CCA and MI approaches produce poor estimates across all scenarios investigated. This can be explained by the association between the missingness and the outcome H_i that is beyond that captured by observed variables. In terms of the performance of the multilevel SEM, as expected, across all three scenarios, the relative bias for all parameters in the health submodel are mostly below 5%. The use of information on the longitudinal D_{ri}s and X_{ri}s helps the identification of the variance of random effects, which is particularly important for non-linear models for H_i (e.g. logit in our case) because the estimated coefficients are interrelated through the variance of u_i (even for the coefficients of independent covariates).

It should also be noted that in this simulation study, all submodels are correctly specified (i.e. recover the parameters of the data generating model). Various forms of model misspecification, such as omitted variables that are not captured by u_i and having more than one covariate correlated with the random effect u_i can lead to biased estimates. As it is not possible to consider all forms of misspecification, it has been advised by White and Carlin (2010) to conduct sensitivity analyses to provide some reassurance on inferences. We also note that for model identification purposes, it is beneficial to have multiple time-varying outcomes (i.e. X_{ri} and D_{ri}) nested within individuals (Lillard, 1993). Next, we investigate the effects of misspecification of the dropout submodel on the parameters in the health submodel. We assume the specifications of models for quantities of primary interest (X_{ri} and H_i) are correctly specified and only the dropout submodel is misspecified. We consider two types of misspecification: 1) Z_2, a variable that only predicts the dropout probabilities is omitted from the dropout model and 2) Z_1, a common predictor for all processes is omitted from the dropout submodel. Fitting the multilevel SEM to the simulated data, we find that under the first type of misspecification, parameters in the health submodel are correctly recovered with relative bias less than 5%. Under the second type of misspecification, simulation results (see Table 7.2) suggest that across all scenarios $\text{var}(u_i)$ is correctly recovered as is the random effect coefficient $\lambda^{(H)}$. However, the coefficient of Z_1 in the model for health is heavily biased while the coefficients for other covariates suffer from a relative bias of less than 13%. The bias is most severe in scenario 1 where the missingness does not depend on summaries of the partnership experience and the health outcome.

The findings from the simulation study have two main implications for empirical analysis when there are no or completely observed time-varying predictors. First, if the models are correctly specified, jointly modelling of the missingness and outcomes of interest using maximum likelihood estimation yields unbiased estimates and correct standard errors. Second, predictors that are common to the missingness process and the primary outcomes should be included in the specification of the dropout model. Neglecting such predictors may lead to a substantial bias in the effects of these covariates on the primary outcomes of interest.

Table 7.2 Relative bias (%) for the second type of misspecification of the dropout submodel, where a common predictor for all processes is omitted from the dropout submodel. Results are shown for each of the MNAR scenarios.

Parameters in the health submodel	Scenario 1	Scenario 2	Scenario 3
$\beta_0^{(H)}$	12.7	8.3	1.0
$\beta_1^{(H)}\,(X_i)$	6.1	4.7	-3.3
$\beta_2^{(H)}\,(Z_{1i})$	50.1	8.0	36.3
$\lambda^{(H)}$	-3.3	-5.8	-5.0
$\text{var}(u_i)$	1.5	3.1	2.8

Relative bias=(Estimate-True)/True×100%

The above analyses serve as a first attempt to understand the performance of different modelling approaches to handle clustered data with missing data in both the outcome and predictors, while restricting our attention on models with time-invariant predictors only. The simulation study also contains a sensitivity analysis to evaluate to what extent the estimates are sensitive to the misspecification of the dropout submodel. In our application, however, time-varying predictors are included in the partnership submodels and they have missing data due to dropouts. As an alternative solution, we limit our interest on making inferences on the relationship between childhood circumstances, partnership transitions and midlife health for the population with complete partnership histories ($N = 7,313$). We use the likelihood-based approach to jointly model the dropout tendency (for the entire population of $N = 14,309$) and the outcomes of interest that are completely observed.

7.4.3 Estimation

The proposed SEM with the four submodels defined by Equations (6.3) and (6.4) and one of the dropout models described in Section 7.4.1 can be estimated using the 3-step maximum likelihood approach, both manually and automatically (using the step3 option) in LatentGOLD 5.1 (Vermunt and Magidson, 2015). The estimation procedure is similar to that described in Chapter 6 except now an additional equation is included into the modelling framework. Denote by $\mathbf{Y}_i = (H_i, \mathbf{y}_i^{(F)}, \mathbf{y}_i^{(S)}, \mathbf{D}_i)$ a stacked response vector for the binary midlife health, partnership outcomes and the dropout process where for Model 1, $\mathbf{D}_i$ is replaced by a scalar D_i. A binomial logit model is then fitted to the single stacked response vector with denominator $(1, \mathbf{n}_i^{(F)}, \mathbf{n}_i^{(S)}, 1)$ for Model 1 and with $(1, \mathbf{n}_i^{(F)}, \mathbf{n}_i^{(S)}, \mathbf{1})$ for Models 2 and 3. Explanatory variables can be included in the model by interacting them with four binary indicators that index each response. For individuals who dropout before age 50, elements of the vector $\mathbf{D}_i$ after the dropout time are coded as missing values. The estimation of Model 2 therefore requires extra dimensions of integration (in addition to that for the random effect) in the E-step of the EM algorithm, which can be computationally expensive if $\mathbf{D}_i$ has a high dimension. Fortunately for this particular study, there are only four adulthood waves. In terms of estimating Model 3, to obtain the marginal log-likelihood for the observed outcomes, both the missing values of H_{ri} (for individuals who dropped out) and the random effects need to be integrated out. The simplex algorithm (Nelder and Mead, 1965) was used for a continuous outcome, where the integral was approximated numerically (Diggle and Kenward, 1994). Ibrahim and Molenberghs (2009) proposed a more efficient Monte Carlo sampling scheme where samples of the random effect and missing response values (both unobserved quantities) are drawn from their joint distribution conditional on the observed outcomes.

In this chapter, all three models are estimated using the EM algorithm with Monte Carlo integration where the dimension of integration varies for each individual. A detailed description of the steps involved are available in Ibrahim and Molenberghs (2009). They also recommended building a not-too-complex model for the missing mechanism to ensure identifiability of the model.

7.5 Application: a study of the effects of childhood socioeconomic circumstances on midlife health, mediated by partnership transitions with non-ignorable dropout

7.5.1 Description of the analysis sample

The data used in this chapter were collected in the NCDS 1958. A detailed description of the dataset, the key variables and the linked partnership histories between ages 16 and 50 is available in Chapter 3. The analysis sample consist of $N = 14,309$ individuals, of whom $N = 7,313$ have complete partnership histories for ages 16-50 and $N = 6,996$ have dropped out before age 50. Previous analyses have implicitly assumed that data are MCAR and in this chapter, we explore alternative missing mechanisms and in particular, MNAR. We focus on modelling the missing data process during adulthood waves because the reasons for missingness during childhood waves (ages 0, 7, 11 and 16) are likely to differ from those in later waves (recorded at age 23 onwards) when cohort members themselves, rather than their carers, are the survey respondents.

The completeness of individual responses to partnership information has been classified into three types in the original dataset: present in the survey and has partnership records, present in the survey but without partnership records, and not present in the survey. The first two types both belong to the "present" category. Note that we focus on waves starting from wave 5 because at age 33, cohort members were asked for details of their co-residential relationships between ages 16 and 33. We next define the timing of dropout in light of the updating feature of partnership records. The term dropout usually refers to the situation where an individual does not respond to the survey at a given wave and does not return thereafter. In the linked partnership history with updates from the latest interview, the timing of dropout is therefore the first time when a consecutive "not present in the survey" is coded up to wave 8 (age 50).

Before modelling, we first perform a simple test of the MCAR assumption that has been implicitly made in the complete-case analyses of Chapter 6. We assess the independence between our key measures of interest (the four dimensions of childhood SECs) and the probability of having a complete or incomplete partnership history in adulthood. Note that for simplicity, modal classes are used as a summary of childhood situations up to age 16. Chi-squared tests are conducted to test for an association between two nominal variables i.e. the binary indicator for a complete history and each of the categorical childhood SEC variables. If the dropout is truly completely at random with respect to childhood SECs, tests should not reject the null hypothesis of no association.

From Table 7.3, it is clear that all four dimensions of childhood socioeconomic situations are significantly associated with the probability of having complete or incomplete partnership histories, providing evidence against the MCAR assumption. Although it is not possible to test for MAR due to the inability to account for all observed variables, their interactions and different distributional assumptions, the above tests at least confirm the need to focus on alternative missing data mechanisms.

Table 7.3 Chi-squared tests for association between each of the four dimensions of childhood socioeconomic situations and having complete or incomplete partnership histories.

Covariates	Test statistics
Male head's social class	$\chi^2 = 78.25, d.f. = 2,$ p-value < 0.001
Financial difficulty	$\chi^2 = 95.15, d.f. = 1,$ p-value < 0.001
Material hardship	$\chi^2 = 125.62, d.f. = 2,$ p-value < 0.001
Family structure	$\chi^2 = 23.69, d.f. = 1,$ p-value < 0.001

7.5.2 Predictors in the dropout model

Before fitting a multilevel SEM with the dropout submodel to the analysis sample, we first need to identify fully observed covariates to be included in the model for the dropout process. Hawkes and Plewis (2006) showed that there is a large drop in the response rate in the NCDS between waves 3 and 4. To minimise the missingness in predictors for dropout during adulthood, we consider only variables in the childhood waves (up to age 16, waves 0-3) as from age 23 (wave 4) onwards, the respondents are the cohort members themselves, rather than their parents or carers as in previous waves. Given the large number of variables available across four childhood waves, the use of machine-learning techniques, for example LASSO for variable selection (Tibshirani, 1996), was considered. However, such methods are infeasible due to the existence of missing values (item non-response) in the information recorded for ages 0-16. Instead, we draw on related literature by choosing covariates (without heavy missingness) that have been found to be significantly related to the probability of dropout in adulthood waves. For example, Hawkes and Plewis (2006) using the NCDS, Rich et al. (2013) using the MCS and Mostafa and Wiggins (2014) using the BCS70 have highlighted the importance of parental SECs, individuals' education level and health situation collected during childhood waves on the dropout tendency. We also select informative predictors by fitting a set of logit models for different dropout outcomes: defining a binary response for incomplete partnership history and separate binary responses for the timing of dropout. Based on the results from these preliminary analyses, the covariates retained in the final dropout model are gender, maternal situations recorded at birth (mother's age at delivery, mother smoking within 4 months of delivery), cohort members' own situations in childood (BMI at age 16, behavioural problems as measured by the Rutter scale, reading and maths scores in exams recorded at age 16), and socioeconomic circumstances in the family for ages 0-16 that are summarised by four categorical latent variables (more details see Section 3.2 and Section 4.6). Table 7.4 presents the results from estimating a set of logit models for the probability of dropout at each adulthood wave, using the general 3-step ML approach to account for misclassification error in childhood SECs. We also report the area under the receiver operating characteristic (ROC) curve (Hosmer and Lemeshow, 2013, Chapter 5) to show the discrimination of the fitted logistic model. The curve is a plot of the sensitivity (proportion of observations with $D_{ri} = 1$ that have a predicted probability larger than the cut-off point) against 1-specificity (specificity is estimated as the proportion of observations with $D_{ri} = 0$ with predicted probability below the cut-off point) across cut-off points ranging from 0 to 1. For a random allocation, the area under the ROC curve is 0.5. Higher values indicate a higher degree of discrimination.

Table 7.4 Estimated effects of covariates in childhood waves on the log-odds of dropout during adulthood (standard errors in brackets).

Covariates	Incomplete history		Dropout at 33		Dropout at 42		Dropout at 46		Dropout at 50	
	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)
Intercept	2.19**	(0.62)	0.98	(0.69)	−4.14**	(1.77)	−0.47	(1.01)	−2.67**	(1.34)
Male (ref.=Female)	0.16**	(0.04)	0.30**	(0.04)	−0.38**	(0.11)	−0.05	(0.06)	−0.13	(0.08)
Maternal age[a] (years)	−0.24**	(0.09)	−0.23**	(0.10)	0.47*	(0.26)	−0.15	(0.15)	−0.29	(0.20)
Maternal smoking[b]	−0.02	(0.04)	−0.10	(0.04)	0.09	(0.11)	0.10	(0.06)	0.05	(0.09)
log(BMI at age 16)	−0.18	(0.17)	−0.28	(0.19)	0.01	(0.49)	0.01	(0.28)	0.23	(0.36)
Behaviour at age 16 (ref.=No problem)[c]	0.26**	(0.06)	0.19**	(0.06)	0.38**	(0.15)	−0.08	(0.10)	0.27**	(0.12)
Reading at age 16	−0.21**	(0.07)	−0.09	(0.07)	−0.15	(0.17)	−0.24**	(0.10)	0.06	(0.15)
Maths at age 16	−0.28**	(0.04)	−0.18**	(0.05)	−0.18	(0.11)	−0.23**	(0.07)	−0.12	(0.09)
Social class (ref.=High)										
Low	0.13**	(0.06)	0.17**	(0.07)	−0.27	(0.18)	0.06	(0.10)	0.09	(0.14)
Medium	0.10**	(0.05)	0.14**	(0.06)	−0.16	(0.15)	−0.03	(0.09)	0.15	(0.11)
Financial difficulty (ref.=Low)										
High	0.22**	(0.06)	0.28**	(0.06)	0.33**	(0.15)	−0.12	(0.09)	−0.13	(0.13)
Material hardship (ref.=Low)										
Medium	0.05	(0.05)	0.02	(0.05)	0.02	(0.15)	0.05	(0.08)	0.13	(0.1)
High	0.18**	(0.05)	0.08	(0.06)	0.22	(0.15)	0.19**	(0.09)	0.16	(0.11)
Family structure (ref.=Stable)										
Unstable	0.19**	(0.06)	0.25**	(0.07)	0.33**	(0.16)	−0.01	(0.10)	−0.37**	(0.16)
Area under the ROC curve	0.79		0.76		0.71		0.66		0.56	

$**p < 0.05, *p < 0.1$

[a] Recorded at the birth wave; [b] Binary indicator for maternal smoking within 4 months of delivery.

[c] Binary indicator for behavioural problems, scored originally using the Rutter scale.

From Table 7.4 we find that male cohort members are significantly more likely than females to drop out at early waves, while females tend to drop out at later waves. Among the parental influences, cohort members with young mothers are more likely to provide incomplete partnership information (i.e. a higher likelihood to drop out) while the smoking behaviour of mothers does not influence respondents' participation in the study. Physical health (measured by BMI at age 16) does not affect the likelihood of dropout but cognitive development (reading and maths scores) significantly influences dropout. Specifically, cohort members with high scores in maths and reading at age 16 are less likely to drop out at adulthood waves. Also, as expected, respondents with behaviour problems recorded at age 16 are significantly more likely to drop out at almost all waves, although the effect is negligible at wave 7 (age 46). In terms of SECs in childhood (at ages 0-16), it is clear that individuals from families with unfavourable socioeconomic situations have a significantly higher chance to drop out. One particular factor is family structure. Individuals born in families with an unstable family structure have a significantly higher likelihood to drop out than their counterparts from stable families. This is particularly clear in earlier waves (by age 42) but if followed up until then, they tend to remain in the study in future waves.

Also note that with the exclusion restriction for identification of the joint model in mind, we include in the dropout model variables that are only predictive of the dropout process, but not directly related to the primary outcomes of interest (i.e. distal health). These variables include mother's age at birth and cohort member's academic scores and behaviour at age 16. Because these features are recorded early in life and capture characteristics that are probably unrelated to health, it is sensible to expect that any direct influence of these quantities on health at age 50 is negligible after 34 years.

7.5.3 The multilevel SEM with different specifications of the dropout submodel

We next fit the multilevel SEM that consists of two equations for partnership transitions, one equation for the distal health and one equation for the dropout process proposed in Section 7.4.1, using the general 3-step ML approach. Three specifications of the dropout submodel are considered. Estimation of the first two dropout models is carried out in LatentGOLD 5.1 which can perform the 3-step procedure with more than one categorical latent variable. Estimation of the Diggle-Kenward-type model using the 3-step feature is only feasible in Mplus using the manual 3-step procedure. To our knowledge, for models with more than one latent variable, Mplus is only able to implement the 3-step approach for estimating latent transition models where the interest is in modelling a change in class memberships over time. As a compromise, we attempted to estimate the SEM with only one categorical latent variable but the model did not converge after running for over 10 days. To further investigate whether non-convergence is due to model complexity or is data-dependent, we conducted a small simulation study for a simple version of the proposed SEM where the health outcome at age 50 is jointly modelled with the dropout probability at each wave. We use the last equation of (7.14) to specify the dropout model where the probability of dropout at each wave is predicted by a categorical latent variable for childhood SECs and repeated measures of health at four adulthood waves. We generated 50 datasets with a medium entropy value of 0.6.

In the first attempt, we generated a 4×1 vector of binary outcomes (for repeated measures of health in adulthood) from the logistic model and a 4×1 vector of binary discrete-time dropout indicators from a standard Diggle-Kenward model with one lagged outcome as a predictor. Included in the covariate set for the dropout process are a two-category latent class variable (measured by five binary indicators) and previous and current health measurements. Values of the health outcomes are set to missing from the dropout time onwards. To allow for full control of the estimation process, the manual 3-step procedure was implemented and the full model was estimated using the EM algorithm with Monte Carlo integration. Unfortunately the model did not converge, but the investigation of log files (which record the updated values of the log-likelihood and of each parameter at each iteration step) shows the parameter that is unidentified is the effect of the outcome on the dropout probability at the same time. We suspect the failure is probably due to the numerical integration required for the Diggle-Kenward model, but future work is needed to understand the reason for non-convergence.

To further simplify the situation, instead of having binary outcomes, in the second attempt we generated continuous health outcomes at each adulthood wave from a multivariate normal distribution and re-estimated the SEM using the same estimation method. The true parameters of interest are all successfully recovered with relative bias below 5%, mean standard errors slightly below the standard deviations (partly because of the small number of replications used) and coverage rates that are close to the nominal 95%. All these results suggest that the 3-step procedure can be used to fit Diggle-Kenward-type models with continuous outcomes (at least for models with one categorical latent variable), where the MNAR is driven by a direct association between the dropout probability and the distal outcome. As in our application the distal outcome is the binary midlife health state, we are only able to fit Models 1 and 2 using the general 3-step approach (data structure and syntax are provided in Appendix G). Focusing on the primary interest of the effects of childhood SECs on midlife health, estimated parameters and the associated standard errors in the health submodel, as well as the random effect parameters in the system of equations are summarised in Table 7.5. In particular, we report results from the SEM ignoring dropout and for two specifications of the dropout submodel. The corresponding estimated coefficients for the dropout submodel are summarised in Table 7.6. We present findings below that help to answer the research questions set out in Section 7.1.

1) Is there any indication of a non-ignorable dropout process? Are estimates of the effects of childhood SECs sensitive to the specification of the dropout model?

   From the multilevel SEM described in this chapter, we find a significant non-zero $\lambda^{(D)}$, indicating a non-ignorable (informative) dropout process as the dropout probability is indirectly associated with the partnership experiences and the distal health state, via a common set of individual-specific unobservables. Of the two specifications of the dropout submodel, the second model is preferred. Incorporating information on the time at dropout can help to better identify the random effects parameters, and hence all other parameters of interest due to the scaling effect of the variance parameter in the random effects logit model (Yatchew and Griliches, 1985).

   Turning to the interpretation of the association between processes, after adjusting for covariates, we find that individuals whose unobserved characteristics place them at a higher risk to drop out early also have generally a higher hazard of separating from a partner at any time for ages 16-50, but are less likely to be in a poor health state in midlife. A possible example of such an unobserved characteristics is "work ambition". Holding other covariates constant, these individuals may be very busy or often move around, making them difficult to have stable relationships and therefore have a higher risk to separate. These individuals may also be more difficult for interviewers to contact or have a higher refusal probability, leading to a higher risk of dropout. Moreover they may be more motivated to pursue a better quality of life, for example by exercising regularly and eating healthily, in order to maximise their achievement at work, and hence are more likely to stay in good health in midlife (i.e. lower risk of poor midlife health).

   Although we are not able to estimate the Diggle-Kenward model for a categorical distal outcome using the 3-step ML approach, we successfully estimated the two dropout models proposed above. The estimated coefficients of the childhood SECs and the associated standard errors in the health submodel are similar across the two specifications of the dropout process. Nonetheless, we note that as the imposed form of the dropout model is empirically unverifiable, the sensitivity tests that can be conducted are limited. It may be more useful to perform more of such tests by considering alternative sets of model assumptions.

2) If the dropout process is informative, how does it impact the effects of childhood SECs and partnership experiences on midlife health?

The estimates in the last two columns of Table 7.6 show that the highest hazard of dropout occurs at age 33. If they do not drop out at age 33, the tendency to drop out is the lowest in wave 6 (age 42). Regarding the characteristics of cohort members, we find that male individuals with unfavourable socioeconomic situations in childhood, poor cognitive development (low reading or maths scores) and behavioural problems are significantly more likely to dropout.

Turning to the health submodel, a comparison of the estimated coefficients in the first and the remaining two columns of Table 7.5 shows that the significance of the effects of all four dimensions of childhood SECs on midlife health remains unchanged with slightly reduced magnitudes. The effect of unstable family structure on midlife health remains non-significant after introducing the dropout model. These results suggest that the inclusion of the dropout model has limited influence on the effects of childhood SECs on midlife health. In terms of the effects of partnership experience, most results remain unchanged but in Model 2, we find that compared with those with one partner for ages 16-50, cohort members with more than three partners are significantly more likely to be a in a poor health state in midlife.

In general, the results suggest that after accounting for non-ignorable dropout, cohort members growing up in families with unfavourable circumstances in the first three dimensions of childhood SECs: father's social class, financial difficulty and material hardship (i.e. material disadvantage) are expected to have a higher chance for being in poor midlife health. Combined with the results from the simplest Model 1 of Section 6.4 (without conditional on random effects and partnership experiences), we conclude that the effects of disadvantages in material dimensions of the childhood socioceonomic circumstances both directly and indirectly increase the likelihood of poor midlife health. Unstable family structure, on the other hand, seems to have an indirect effect on midlife health, through influencing partnership experiences (see results in Table 7.7). Specifically, individuals from families with unstable family structures are significantly more likely to partner early and separate with their partners throughout the 34 years of follow-up. As unstable partnership experiences tend to increase the likelihood of poor midlife health, these individuals are also subject to this exposure. Note that the interpretation of direct and indirect effects should be with caution, under the assumption of no unmeasured confounders. It is, however, possible to have residual confounding for partnership transitions and midlife health that is not accounted for by time-invariant random effects. More discussions are included in Section 8.3.

Table 7.5 Estimated covariate effects on the log-odds of having poor health in midlife (standard errors in brackets). SEM (Model 1) refers to model (7.12) with a scalar D_i being the dropout outcome. SEM (Model 2) refers to model (7.13) with a vector $\mathbf{D}_i$ being the dropout outcome.

Health submodel	SEM (No dropout)		SEM (Model 1)		SEM (Model 2)	
Predictors	Est.	(SE)	Est.	(SE)	Est.	(SE)
Intercept	−2.70**	(0.146)	−2.93**	(0.197)	−3.06**	(0.226)
Male (ref.= Female)	−0.04	(0.062)	−0.05	(0.064)	−0.05	(0.063)
Overweight at age 16[a] (ref.= No)	0.26**	(0.065)	0.26**	(0.070)	0.26**	(0.070)
Childhood socioeconomic circumstances						
Social class (ref. =High)						
Low	0.46**	(0.112)	0.45**	(0.115)	0.44**	(0.116)
Medium	0.31**	(0.092)	0.30**	(0.097)	0.30**	(0.097)
Financial difficulty (ref.=Low)						
High	0.46**	(0.093)	0.44**	(0.098)	0.42**	(0.099)
Material hardship (ref.=Low)						
Medium	0.32**	(0.079)	0.32**	(0.089)	0.32**	(0.119)
High	0.41**	(0.086)	0.40**	(0.095)	0.39**	(0.096)
Family structure (ref.=Stable)						
Unstable	0.14	(0.128)	0.19	(0.114)	0.17	(0.165)
Summaries of partnership experiences						
Total number of partners before age 50 (ref. =1)						
0	−0.12	(0.322)	−0.17	(0.320)	−0.13	(0.318)
2	0.04	(0.130)	0.14	(0.133)	0.18	(0.136)
3+	0.15	(0.232)	0.36	(0.247)	0.41*	(0.244)
Age at first partnership[b]	−0.13**	(0.048)	−0.13**	(0.047)	−0.13**	(0.047)
Percentage time spent single	1.08**	(0.371)	1.22**	(0.379)	1.26**	(0.380)
Random effects parameters						
σ_u^2	0.93**	(0.097)	1.00**	(0.041)	1.32**	(0.104)
$\lambda^{(D)}$			1.06**	(0.104)	1.25**	(0.124)
$\lambda^{(H)}$	−0.20	(0.147)	−0.40**	(0.168)	−0.44**	(0.163)
$\lambda^{(F)}$	0.03	(0.088)	−0.01	(0.029)	−0.05**	(0.024)
$\lambda^{(S)}$	1		1		1	

[a] Binary indicator for overweight at age 16.

[b] Age at first partnership is centred at median.

Table 7.6 Estimated covariate effects on the log-odds of dropout at each wave (standard errors in brackets).

Dropout submodel	Separate estimation		SEM (Model 1)		SEM (Model 2)	
Predictors	Est.	(SE)	Est.	(SE)	Est.	(SE)
Intercept	2.19**	(0.623)	2.60**	(0.754)	1.10	(0.693)
Wave (ref. =Age 33)						
Age 42					−2.10**	(0.065)
Age 46					−0.44**	(0.074)
Age 50					−0.80**	(0.093)
Male (ref.=Female)	0.16**	(0.036)	0.18**	(0.045)	0.19**	(0.043)
Maternal age[a]	−0.24**	(0.091)	−0.30**	(0.111)	−0.27**	(0.102)
Maternal smoking[b]	−0.02	(0.040)	−0.03	(0.048)	−0.04	(0.044)
log(BMI at age 16)	−0.18	(0.169)	−0.15	(0.204)	−0.12	(0.188)
Behaviour at age 16 (ref.=No problem)[c]	0.26**	(0.059)	0.30**	(0.072)	0.25**	(0.066)
Reading at age 16	−0.21**	(0.073)	−0.29**	(0.088)	−0.27**	(0.079)
Maths at age 16	−0.28**	(0.044)	−0.34**	(0.054)	−0.30**	(0.049)
Childhood socio-ecnomic circumstances						
Social class (ref. =High)						
Low	0.13**	(0.062)	0.16**	(0.076)	0.18**	(0.074)
Medium	0.10**	(0.049)	0.12**	(0.061)	0.13**	(0.060)
Financial difficulty (ref.=Low)						
High	0.22**	(0.057)	0.27**	(0.07)	0.30**	(0.068)
Material hardship (ref.=Low)						
Medium	0.05	(0.047)	0.06	(0.058)	0.05	(0.057)
High	0.18**	(0.051)	0.23**	(0.063)	0.20**	(0.062)
Family structure (ref.=Stable)						
Unstable	0.19**	(0.063)	0.24**	(0.078)	0.30**	(0.076)

[a] Maternal age was recorded at the birth wave when the cohort member was born.

[b] Maternal smoking is a binary indicator for smoking within 4 months of delivery, recorded at the birth wave.

[c] Binary indicator for behavioural problems derived from the 26-item Rutter's scale. A cut-off of score 9 is applied where higher values indicate "behavioural problems".

Table 7.7 Effects of childhood SECs on partnership transitions. Results are presented for respective SEMs before (see also Table 6.5) and after including the dropout submodel (specifically, SEM (Model 2)).

Childhood SECs	SEM (no dropout submodel)		SEM (Model 2)	
	Est.	(SE)	Est.	(SE)
First entry into partnership				
Social class[a] (ref.=High)				
Low	0.11**	(0.04)	0.11**	(0.04)
Medium	0.10**	(0.03)	0.11**	(0.03)
Financial difficulty (ref.=Low)				
High	0.03	(0.04)	0.03	(0.04)
Material hardship (ref.=Low)				
Medium	0.04	(0.03)	0.04	(0.03)
High	0.05	(0.03)	0.05	(0.03)
Family structure (ref.=Stable)				
Unstable	0.15**	(0.04)	0.15**	(0.03)
Partnership dissolution				
Social class[a] (ref.=High)				
Low	-0.13*	(0.07)	-0.07	(0.07)
Medium	-0.04	(0.06)	0.01	(0.05)
Financial difficulty (ref.=Low)				
High	0.05	(0.07)	0.17**	(0.07)
Material hardship (ref.=Low)				
Medium	-0.09*	(0.05)	-0.09	(0.06)
High	-0.14**	(0.06)	-0.11*	(0.06)
Family structure (ref.=Stable)				
Unstable	0.23**	(0.07)	0.29**	(0.07)

**$p < 0.05$, *$p < 0.1$.

[a] Father or male head social class.

Chapter 8

Conclusion

This chapter summarises the key contributions of this thesis to the existing literature. The primary substantive research question is about understanding the impacts of childhood socioeconomic circumstances on midlife health, and if such influences are mediated by social processes (e.g. life events) in adulthood. Methodological contributions are discussed in Section 8.1 and the impact of the proposed modelling framework on the substantive understanding of the subject of interest is summarised in Section 8.2. Section 8.3 includes a discussion of the limitations of the research and proposals for future work.

8.1 Methodological contributions

Childhood socioeconomic circumstances (SECs) are abstract concepts and consist of multiple dimensions. In the four childhood waves of the NCDS (conducted at ages 0, 7, 11 and 16), multiple indicators are available, capturing the status in multiple dimensions of childhood SECs at each time point. Our interest is to summarise the information across waves to characterise the overall pattern of each dimension of childhood SECs over time, taking into account measurement error and minimising the loss of information due to missing values, and then relating these SECs to temporally distal outcomes. Guided by the commonly used measures of childhood experiences in previous studies, we identified four associated dimensions of childhood SECs: father or male head's social class, material hardship, financial difficulties and family structure. At each wave, we construct a composite index for each aspect of childhood SECs, using the corresponding set of variables capturing this particular dimension. For each dimension, these composite indices form the repeated measures. Motivated by the applications of latent class growth analysis (Hagenaars and McCutcheon, 2002, Chapter 10), in Chapter 4, we derive four associated latent categorical summaries of different dimensions of childhood SECs using the respective repeated measures in childhood waves. The typology or patterns of change in childhood SECs, represented by the categorical latent variables, are more comprehensive summaries of childhood situations than the SECs measured at a single time point. Fitting a latent class model for each SEC also maximises the use of childhood longitudinal data with missing values, as the models are estimated using full information maximum likelihood (ML) under the missing at random assumption.

Relating the latent class variables to temporally distal outcomes is of substantive interest. Such outcomes include life events (e.g. partnership transitions) and midlife health and we start by considering a single distal outcome (more details see the latter half of Chapter 4). The 3-step maximum likelihood approach is used to account for the misclassification error in the modal classes (most likely class membership) derived from latent class models. We then extend the 3-step ML approach for models with one categorical latent variable to multiple, and possibly associated, latent variables. Simulation studies are performed to assess the robustness of the approach to violations of model assumptions, including various degrees of the within-class non-normality for a continuous distal outcome and conditional dependence between items and the distal outcome. Results show that when all model assumptions are satisfied, the 1-step (simultaneous estimation of the measurement and the regression model) and 3-step ML approaches perform equally well. When model assumptions are violated, the estimates from both the 1-step and 3-step methods are subject to bias, although the 3-step ML approach is less sensitive. Specifically, when there is within-class non-normality for a continuous distal, within-class skewness of the distal outcome is shown to be the worst form of non-normality for both approaches, compared to bi-modality and excess kurtosis. Moreover, the results confirmed a major drawback of the 1-step approach, as it not only alters the class proportions (see also Asparouhov and Muthén (2014a) and Bakk and Vermunt (2016)), but also changes the number of classes needed to capture the association among indicators in the latent class model, particularly at low entropy levels (poor class separation). When there is local dependence between the distal outcome and the indicators for the latent class variable, the 1-step approach leads to greater bias than the 3-step ML approach. This is mainly explained by a tendency to extract too many classes when there is residual correlation between the distal outcome and the indicators. It should be noted that the extraction of pseudo classes is not necessarily wrong from a theoretical point of view, but one needs to question the validity of such extra classes, which might not be interpretable.

We next extend the 3-step ML approach to estimate a multilevel structural equation model (SEM) where the latent class variables can be related to clustered outcomes of mixed types (see more discussions in Chapter 5 and Chapter 6). Of particular interest are the outcomes of partnership transitions (duration data, nested within individuals), midlife health (individual-level) and dropout (binary indicators, nested within individuals). Partnership histories over 34 years (from age 16 to 50) are used in the analysis, serving both as outcomes (in one equation) and as predictors (in a another equation), which has rarely been considered in previous studies. It is of substantive interest to simultaneously estimate the effects of childhood SECs on time-to-event outcomes and the cumulative effects of partnership stability over 34 years on midlife health. To relate the partnership experiences to midlife health, three summary variables are derived that capture the overall partnership stability over 34 years. The proposed SEM therefore consists of latent class models and regression models, where the latter includes submodels for partnership formation and dissolutions and a submodel for midlife health. In a joint model, we allow for residual association across equations via the correlated random effects. To reduce the computational complexity, we impose a factor structure on the correlation. This residual association also helps to 1) mitigate the problem of endogeneity due to correlation between individual-specific unobservables and the summaries of partnership events included as predictors in the model for midlife health, 2) further allow for an indirect relationship between partnership stability and midlife health, in addition to that captured by the summary variables of partnership experiences and 3) account for the additional dependence between partnership transitions and midlife health given the latent classes.

Dropout is a common problem in cohort studies and missing values in outcomes need to be accounted for. Because it is impossible to distinguish the missing at random (MAR) mechanism from the missing not at random (MNAR) mechanism, we consider a more flexible MNAR assumption (i.e. informative dropout) where the missingness probabilities are related to the outcomes of interest. The proposed SEM is therefore extended to include an additional equation to model the dropout time where the dropout probabilities and the outcomes of interests have shared influences of a common set of individual-specific unobservables (more details in Chapter 7). This is achieved by allowing for a residual association across four equations (two equations for partnership transitions, one equation for midlife health and one equation for dropout). We also conduct a sensitivity analysis to assess whether the effects of childhood SECs on midlife health are sensitive to the specification of the dropout submodel. In addition to the two forms of indirect association between dropout probabilities and distal outcomes, we also consider a direct association in the form of a Diggle-Kenward model. However, due to the complexity of the SEM and the large dataset, we are only able to estimate models with indirect associations. In terms of the effects of misspecification of the dropout submodel on key parameters of interest, simulation results show that in the dropout submodel, the omission of common predictors (confounders) of the dropout tendency and midlife health can produce a large bias in the estimates of the parameters of interest in the health submodel.

This suggests that the dropout submodel should include all possible predictors that may be shared by the primary outcomes of interest.

The proposed SEM belongs to the general latent variable modelling frameworks of Skrondal and Rabe-Hesketh (2004) and Asparouhov and Muthén (2012), and hence has the same generalizability. More specifically, life events (e.g. partnership transitions), distal outcomes (e.g. midlife health) and the dropout indicators can be viewed as items of one or more individual-level latent variables. Different specifications of the structural model can accommodate various research questions where the underlying association of individual-level variables (rather than their manifestations with measurement errors) is of primary interest.

8.2 Substantive impacts

This thesis aims to address four main research questions set out in Chapter 1. RQ_1 mainly contains methodological questions and our contributions have been summarised in Section 8.1. The substantive findings from models developed to answer $RQs\, 2-4$ are summarised as follows.

RQ_2 How do childhood circumstances influence partnership stability between ages 16 and 50?

For the first partnership formation, results from the final model (SEM (Model 2) of Table 7.7) show that individuals from families with fathers in medium and low social classes or with an unstable family structure (e.g. who experienced parental separation) are significantly more likely to enter into the first partnership earlier than their counterparts from more favourable family backgrounds. Effects of other aspects of childhood SEC are not significant.

For partnership dissolutions (SEM (Model 2) of Table 7.7), we find that among the four childhood SECs considered, children from families with high financial difficulty or with an unstable family structure have a significantly higher tendency to separate from a partner. Those from families with a high degree of material hardship (e.g. poor housing conditions) are less likely to experience separation in adulthood. Further investigations show that such effects do not tend to vary for recurrent partnerships. In addition, individual heterogeneity in the tendency to separate is present in this sample, suggesting that there exist common influences of unobserved individual-level characteristics on the dissolution hazard for all episodes within an individual.

In terms of the residual association, we find a significant and negative $\lambda^{(F)}$ (see Table 7.5, SEM (Model 2)), indicating a negative association between the hazard of forming the first partnership and dissolving subsequent unions due to a common set of individual-specific unobservables.

RQ_3 Do partnership experiences mediate the effects of childhood SECs on health at age 50?

Controlling for gender and BMI at age 16, the results of the simplest Model 1 in Chapter 6 show significant effects on midlife health of the three dimensions of childhood SECs that relate to material difficulties – parental social class, financial difficulty and material hardship – with cohort members from families with unfavourable conditions on these aspects significantly more likely to be in a poor health state at age 50. The direct effect of family structure is non-significant.

Conditional on individual-specific unobservables, the experience of material difficulties is both directly and indirectly (by increasing the risk of partnership instability) related to a higher likelihood of poor midlife health. In contrast, the effect of family structure on midlife health operates only indirectly by influencing partnership transitions, which in turn affect later health. Specifically, cohort members who have experienced unstable family structures in childhood form their first partnerships earlier and have a higher risk of separation across relationships. From the health submodel (see Table 7.5), individuals who have formed their first partnership late tend to have a lower risk to develop health issues at age 50. Moreover, conditional on the age at first partnership, cohort members who have spent a longer time single between ages 16 and 50 have an increased chance of poor health in midlife. We also find that, compared with individuals who have only formed one partnership, those who have over three relationships are significantly more likely to have poor midlife health. These results suggest that unstable family structure is positively related to a higher chance of poor midlife health, but the influence is only indirect through partnership transitions.

The results reveal that partnership transitions have an important role to play in the relationship between childhood SECs and later health. We also find direct effects of material disadvantages in childhood on midlife health, which may carry policy implications and requires further research.

RQ_4 Are there any residual interrelationships between partnership experiences, dropout propensity and midlife health?

Estimates for the random effects coefficients are all significant at the 5% significance level. This confirms the additional association between partnership transitions, midlife health and the dropout tendency that are due to unobserved time-invariant characteristics. We also find that dropout is non-ignorable as it is indirectly associated with midlife health and partnership transitions via individual-specific unobservables.

8.3 Limitations and future work

The NCDS contains a large number of measurements for childhood SECs and early health across four childhood waves. In this study, our choice of indicators was guided by previous research and hence a selective set of indicators was used to construct composite indices that capture the respective abstract dimension of the childhood SEC. Alternatively, the choice of childhood measures can be facilitated by more efficient automatic variable selection techniques, such as LASSO and LARS that have been developed in the field of machine learning (Efron et al., 2004; Tibshirani, 1996). One should, however, note the ignorance of the meaning of variables in machine learning approaches. Moreover, the capability of these tools to account for missing values in predictors, and the presence of latent variables needs to be carefully investigated.

The model proposed in this thesis could be extended in several ways. First, we have considered a small set of control variables for midlife health. To include other variables that are indicators for early health in childhood, missing values need to be imputed. As they may be associated with childhood SECs (latent variables), the imputation procedure requires careful consideration. Although one option would be to use the modal classes for the SEC variables as auxiliary variables in the imputation model, they are subject to classification errors, particularly when class separation is poor. Alternatively, indicators for the childhood SECs may be used but they contain missing values themselves. In terms of the mediation mechanism, other confounders that have not been considered in the SEM should also be discussed. The variables that are confounders in the "partnership–midlife health" may include education attainment, health states and health-related behaviours (such as drinking and smoking) in early adulthood (at ages 23 and 33); variables that are confounders in the "childhood SECs–partnership" may include childhood health states, behavioural problems and examination scores. To examine the impact of the inclusion of these variables on the key parameters of interest (i.e. effects of childhood SECs on midlife health), analyses could be performed in terms of an extensive sensitivity analysis. Note that these variables may be potentially intermediate confounders that sit on the causal pathway, which can further complicate the problem.

Second, a key component in our estimation procedure for the proposed SEM is the 3-step approach that relates latent categorical variables to mixed types of multilevel distal outcomes. However, the misclassification probabilities calculated in step 2 after fitting latent class models are assumed known and carried forward in estimation of the structural regression model in step 3. Estimation uncertainty in step 1 is therefore ignored. This does not seem to be a big problem in our analysis because the entropy levels for each of the four latent class models are high (above 0.7), indicating a good class separation.

However, if the class separation is poor, we would have a higher degree of uncertainty in the misclassification probabilities. If this is ignored in step 3, the overall uncertainty in the estimated parameters of interest can be understated, producing spurious inferences. The recent proposition of a two-step approach by Bakk and Kuha (2017) accounts for the variation in the estimated misclassification probabilities in step 1, that were previously treated as known in the regression model.

Third, regarding the substantive use of our proposed multilevel SEM, possible extensions of the 3-step ML approach may lie in these areas. For example, the 3-step ML approach can be extended to model more than one distal outcome, e.g. distinguishing measurements for mental and physical health, which may be of mixed types. Additional assumptions may be necessary and methodological research are needed to evaluate the extent to which the estimates are sensitive to the violation of these assumptions. The 3-step ML approach can also be extended to handle a scenario where the distal outcome is a latent construct measured by a number of indicators. In this thesis, we have considered a single health outcome derived from the self-reported general health state at age 50. If measurement error in self-reported health states is a concern, one may introduce a latent variable into the model that summaries the broad phenomena (or the longitudinal trajectory) captured by multiple indicators (or repeated measures) of general health. This is motivated by the fact that later life health situations are often self-reported and multiple measures or biomarkers may be available in the survey. However, the performance of the 3-step approach for extensions involving a regression of a latent health variable on latent childhood SECs requires further investigation.

Fourth, dropout is a common issue in cohort studies which leads to missing data both in the outcomes and predictors. It can be particularly troublesome for models with time-varying predictors because when the outcome is missing, these predictors are missing for the same time units. For our modelling framework, imputing time-varying predictors has both methodological and practical challenges due to the following reasons: 1) These predictors are probably related to childhood SECs, which are latent; 2) Time-varying predictors for partnership transitions, for example, are nested within each time period of each episode of each individual. To impute these predictors, we need to impute the time periods first (which is essentially imputing duration data, the outcome of EHA); 3) If we impute these predictors using wave-specific models, the gap (over 3 years) is too wide for us to assume they are constant over the gap. 4) There are many time-varying predictors that are subject to missing values (e.g. age at the start of each partnership – episode level, type of partnership – episode level, number of pre-school children – period-level) and misspecification of the imputation model can generate even greater bias. Considering this, we have restricted our attention to the subsample with complete partnership histories. To assess the impact of missingness both in outcomes and time-invariant predictors, we have performed a simulation study that shows that under three different missing data mechanisms (for non-ignorable missingness), joint modelling of the event history data (subject to missingness), distal outcomes (subject to missingness) and dropout can produce unbiased point estimates and good standard errors. Should time-varying predictors (subject to missingness) exist, alternative modelling approaches should be considered. For example, Roy and Lin (2005) proposed a joint model for the time-varying covariates, primary outcomes of interest and the dropout indicators, where a lagged transition model was suggested for time-varying covariates. Independence between time-varying covariates is assumed, conditional on their history and the complete time-invariant covariates. Estimation is based on the joint integrated quasilikelihood when the missingness is informative. For survival analysis, Dupuy et al. (2006) discussed the theoretical properties of the maximum likelihood estimators for the joint estimation of the Cox model for time-to-event data and a multivariate model for longitudinal (time-varying) covariates where time-dependent covariates are subject to missingness. Joint estimation of longitudinal and survival models has been prevalent in the field of biostatistics (Faucett and Thomas, 1996; Proust-Lima and Taylor, 2009). Extensions are required, however, for the 3-step ML approach to handle clustered data with such complexity.

While a number of measures of health at different stages in life are available, for simplicity, the current multilevel SEM models the midlife health controlling for one measure of general health at age 16. It is possible to extend our model to accommodate multiple repeated measures of health conditions for ages 16-50. Specifically, an additional model can be included to model the progression of health using self-reported measures at age 16, 23, 33, 42, 46, 50, but one should note that health records are measured with error and recorded at large age gaps. To connect this model with our existing framework, we could borrow strength from the literature in medical studies for the joint modelling of survival data and multivariate repeated measures of health. Substantively, the extended model can capture the dynamic relationship between childhood SECs, partnership (or employment and fertility) experiences, and the progression of health. Such a model would allow for the interrelationship between processes and the potential dynamic effect of early health on partnership and employment transitions. For example, individuals with poor early health may have less stable relationship and employment experiences, leading to higher mental distress and poor midlife health. Within the processes modelled, unstable relationships at time points $t-1$ can lead to poor health and unemployment at t, which in turn may result in increased risk of partnership dissolution at time t and $t+1$. Should the change in these social processes or health states be of research interest, the 3-step approach can be extended to latent transition analysis.

Nylund-Gibson et al. (2014) discussed an application where two latent variables are causally related; further extensions can allow these latent variables to predict distal outcomes.

Finally, we discuss the possibility of extending our model to a causal model for policy recommendations. The current research loosely defines the partnership experiences as "mediators". As we did not investigate alternative causal structures underlying the substantive quantities of interest, the research findings should be interpreted with caution as no strictly causal inferences are made. So far, our model has attempted to empirically test the framework where the influence of childhood SECs on midlife health is transmitted over the life course. Should the causal inferences be of interest for policy insights (e.g. to decompose the total effects of childhood SECs into direct and indirect effects and to understand to what extent can intervention benefit midlife health), we need to investigate the existence of other mediators in addition to the experience of different life events (e.g. partnership, employment and fertility transitions), the intermediate constructs that mediate the relationship between partnership experiences and midlife health (e.g. health-related behaviours, including smoking, drinking and diets), or those that mediate the relationship between childhood SECs and partnership experiences. Potential confounders should also be examined, as well as the validity of the additional statistical assumptions necessary for making causal inferences. The potential outcome framework (facilitated by directed acyclic graphs) is another alternative modelling approach that rigorously defines the causal effects by comparing the quantities of interest in two hypothetical worlds (Pearl, 2001). It should be noted that this direction of future work can be challenging. There may be a number of mediators and intermediate confounders that are of mixed types. The mediators of interest may be time-to-event outcomes, or latent variables and the exposure, childhood SECs are latent categorical variables, posing challenges for algebraic derivations of the counterfactual-based causal effects and the decomposition of the total effects into causal definitions of direct and indirect effects. A general modelling framework that explores the use of structural equation models for causal analysis where the mediators are latent and measured with error has been proposed by Muthén and Muthén (2017) but further research is required to jointly address the concerns set out above.

References

Aassve, A. and Billari, Francesco C.and Spéder, Z. (2006). Societal transition, policy changes and family formation: Evidence from Hungary. *European Journal of Population / Revue européenne de Démographie*, 22(2):127–152.

Aassve, A., Burgess, S., Propper, C., and Dickson, M. (2006a). Employment, family union and childbearing decisions in Great Britain. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):781–804.

Aassve, A., Burgess, S. M., Dickson, M., and Propper, C. (2006b). Modelling poverty by not modelling poverty: an application of a simultaneous hazards approach to the uk. *LSE STICERD Research Paper No. CASE106.*

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data: First Edition*. Sage, New York.

Amato, P. R. (1996). Explaining the intergenerational transmission of divorce. *Journal of Marriage and the Family*, 58(3):628–640.

Andersson, S., Bengtsson, C., Hallberg, L., Lapidus, L., Niklasson, A., Wallgren, A., and Hulthen, L. (2001). Cancer risk in swedish women: the relation to size at birth. *British Journal of Cancer*, 84(9):1193–1198.

Asparouhov, T. and Muthén, B. (2012). Using Mplus TECH11 and TECH14 to test the number of latent classes. *Mplus Web Notes*, 14:22.

Asparouhov, T. and Muthén, B. (2014a). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3):329–341.

Asparouhov, T. and Muthén, B. (2014b). Auxiliary variables in mixture modeling: Using the BCH method in Mplus to estimate a distal outcome model and an arbitrary secondary model. *Mplus Web Notes*, 21:1–22.

Asparouhov, T. and Muthén, B. O. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In Hox, J. and Roberts, J. K., editors, *Handbook of Advanced Multilevel Analysis*, chapter 2, pages 15–40. Routledge, New York.

Attanasio, O. P. and Emmerson, C. (2003). Mortality, health status, and wealth. *Journal of the European Economic Association*, 1(4):821–850.

Austin, P. C. (2012). Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29):3946–3958.

Bakk, Z. and Kuha, J. (2017). Two-step estimation of models between latent classes and external variables. *Psychometrika*, pages 1–22.

Bakk, Z., Oberski, D. L., and Vermunt, J. K. (2014). Relating latent class assignments to external variables: standard errors for correct inference. *Political Analysis*, 22(4):520–540.

Bakk, Z., Tekle, F. B., and Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, 43(1):272–311.

Bakk, Z. and Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(1):20–31.

Barker, D. J., Osmond, C., Forsén, T. J., Kajantie, E., and Eriksson, J. G. (2005). Trajectories of growth among children who have coronary events as adults. *New England Journal of Medicine*, 353(17):1802–1809.

Bärnighausen, T., Bor, J., Wandira-Kazibwe, S., and Canning, D. (2011). Correcting hiv prevalence estimates for survey nonparticipation using heckman-type selection models. *Epidemiology*, pages 27–35.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach, 3rd Edition*. John Wiley & Sons, Chichester.

Bartlett, J. W., Carpenter, J. R., Tilling, K., and Vansteelandt, S. (2014). Improving upon the efficiency of complete case analysis when covariates are mnar. *Biostatistics*, 15(4):719–730.

Bartley, M. (2003). Commentary: Relating social structure and health. *International Journal of Epidemiology*, 32(6):958–960.

Bartley, M., Kelly, Y., and Sacker, A. (2012). Early life financial adversity and respiratory function in midlife: a prospective birth cohort study. *American Journal of Epidemiology*, 175(1):33–42.

Bartley, M., Power, C., Blane, D., Smith, G. D., and Shipley, M. (1994). Birth weight and later socioeconomic disadvantage: evidence from the 1958 british cohort study. *BMJ*, 309(6967):1475–1478.

Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*, 42(4):757–786.

Bauldry, S., Shanahan, M. J., Russo, R., Roberts, B. W., and Damian, R. (2016). Attractiveness compensates for low status background in the prediction of educational attainment. *PLoS One*, 11(6):e0155313.

Ben-Shlomo, Y., Cooper, R., and Kuh, D. (2016). The last two decades of life course epidemiology, and its relevance for research on ageing. *International Journal of Epidemiology*, 45(4):973–988.

Bender, R. (2009). Introduction to the use of regression models in epidemiology. In Verma, M., editor, *Cancer Epidemiology*, pages 179–195. Humana Press, Totowa.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723.

Berrington, A. (2003). Change and continuity in family formation among young adults in britain. S3RI Applications and Policy Working Papers A03/04, Southampton Statistical Sciences Research Institute.

Berrington, A. and Diamond, I. (2000). Marriage or cohabitation: A competing risks analysis of first-partnership formation among the 1958 british birth cohort. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(2):127–151.

Beunckens, C., Molenberghs, G., Verbeke, G., and Mallinckrodt, C. (2008). A latent-class mixture model for incomplete longitudinal gaussian data. *Biometrics*, 64(1):96–105.

Blossfeld, H.-P. and Gotz, R., editors (2001). *Techniques of event history modeling: New approaches to casual analysis: 2rd Edition*. Psychology Press.

Blossfeld, H.-P. and Huinink, J. (1991). Human capital investments or norms of role transition? How women's schooling and career affect the process of family formation. *American Journal of Sociology*, 97(1):143–168.

Bolck, A., Croon, M., and Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, 12(1):3–27.

Brewer, M. and Nandi, A. (2014). Partnership dissolution: how does it affect income, employment and well-being? ISER Working Paper Series 2014-30, Institute for Social and Economic Research, University of Essex.

Browne, W. J., Goldstein, H., and Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, 1(2):103–124.

Bumpass, L. L., Sweet, J. A., and Cherlin, A. (1991). The role of cohabitation in declining rates of marriage. *Journal of Marriage and the Family*, pages 913–927.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292.

Carpenter, J., Kenward, M., Evans, S., and White, I. (2004). Last observation carry-forward and last observation analysis. *Statistics in Medicine*, 23(20):3241–3242.

Case, A., Fertig, A., and Paxson, C. (2005). The lasting impact of childhood health and circumstance. *Journal of Health Economics*, 24(2):365–389.

Case, A. and Paxson, C. (2011). The long reach of childhood health and circumstance: Evidence from the Whitehall II study. *The Economic Journal*, 121(554):183–204.

Chandola, T., Clarke, P., Morris, J., and Blane, D. (2006a). Pathways between education and health: a causal modelling approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(2):337–359.

Chandola, T., Deary, I. J., Blane, D., and Batty, G. D. (2006b). Childhood IQ in relation to obesity and weight gain in adult life: the National Child Development (1958) study. *International Journal of Obesity*, 30(9):1422–1432.

Cherlin, A. J., Kiernan, K. E., and Chase-Lansdale, P. L. (1995). Parental divorce in childhood and demographic outcomes in young adulthood. *Demography*, 32(3):299–318.

Clark, S. L. and Muthén, B. (2009). Relating latent class analysis results to variables not included in the analysis. www.statmodel.com/download/relatinglca.pdf. Online; accessed 30 October 2015.

Clayton, D., Hills, M., and Pickles, A. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.

Cohen, S., Janicki-Deverts, D., Chen, E., and Matthews, K. A. (2010). Childhood socioeconomic status and adult health. *Annals of the New York Academy of Sciences*, 1186(1):37–55.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34:187–220.

Crowther, M. J., Andersson, T. M., Lambert, P. C., Abrams, K. R., and Humphreys, K. (2016). Joint modelling of longitudinal and survival data: incorporating delayed entry and an assessment of model misspecification. *Statistics in Medicine*, 35(7):1193–1209.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–38.

DeWit, D. J., Adlaf, E. M., Offord, D. R., and Ogborne, A. C. (2000). Age at first alcohol use: a risk factor for the development of alcohol disorders. *American Journal of Psychiatry*, 157(5):745–750.

Di Salvo, P. and Smith, K. (1995). National Child Development Study: NCDS housing event histories. Date note 1, City University, London.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data: First Edition*. Oxford University Press, Oxford.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, pages 49–93.

Dupuy, J.-F., Grama, I., Mesbah, M., et al. (2006). Asymptotic theory for the Cox model with missing time-dependent covariate. *The Annals of Statistics*, 34(2):903–924.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Ekholm, A. and Skinner, C. (1998). The muscatine children's obesity data reanalysed using pattern mixture models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(2):251–263.

Elliott, J. and Lawrence, J. (2014). Refining childhood social class measures in the 1958 british cohort study. CLS Cohort Studies 2014/1, Institute of Education, UCL.

Elo, I. T. (1998). Childhood conditions and adult health: Evidence from the Health and Retirement study. PARC Working Paper Series WPS 98-03, Population Aging Research Center, University of Pennsylvania.

Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53.

Entwisle, D. R. and Astone, N. M. (1994). Some practical guidelines for measuring youth's race/ethnicity and socioeconomic status. *Child Development*, pages 1521–1540.

Ermisch, J. and Francesconi, M. (2000). The effect of parents' employment on children's educational attainment. Technical report, Institute for Social and Economic Research, University of Essex.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, 15(15):1663–1685.

Feinstein, L. and Bynner, J. (2004). The importance of cognitive development in middle childhood for adulthood socioeconomic status, mental health, and problem behavior. *Child development*, 75(5):1329–1339.

Follmann, D. and Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics*, pages 151–168.

Frankel, S., Smith, G. D., and Gunnell, D. (1999). Childhood socioeconomic position and adult cardiovascular mortality: the Boyd Orr Cohort. *American Journal of Epidemiology*, 150(10):1081–1084.

Freedman, D., Kettel Khan, L., Serdula, M., Srinivasan, S., and Berenson, G. (2001). BMI rebound, childhood height and obesity among adults: the Bogalusa Heart Study. *International Journal of Obesity & Related Metabolic Disorders*, 25(4).

Gale, C. R., Cooper, R., Craig, L., Elliott, J., Kuh, D., Richards, M., Starr, J. M., Whalley, L. J., and Deary, I. J. (2012). Cognitive function in childhood and lifetime cognitive change in relation to mental wellbeing in four cohorts of older people. *PloS one*, 7(9):e44860.

Gale, C. R., Johnson, W., Deary, I. J., Schoon, I., and Batty, G. D. (2009). Intelligence in girls and their subsequent smoking behaviour as mothers: the 1958 National Child Development Study and the 1970 British Cohort Study. *International Journal of Epidemiology*, 38(1):173–181.

Galobardes, B., Smith, G. D., and Lynch, J. W. (2006). Systematic review of the influence of childhood socioeconomic circumstances on risk for cardiovascular disease in adulthood. *Annals of Epidemiology*, 16(2):91–104.

Garre, F. G. and Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika*, 33(1):43–59.

General, R. (1933). *Census of England and Wales, 1931: Ecclesiastical Areas (England)*. HM Stationery Office.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing inferences from self-selected samples*, pages 115–142. Springer.

Goldstein, H. (2010). *Multilevel Statistical Models, 4th Edition*. John Wiley & Sons, New York.

Goldstein, H., Pan, H., and Bynner, J. (2004). A flexible procedure for analyzing longitudinal event histories using a multilevel model. *Understanding Statistics*, 3(2):85–99.

Goodman, A., Joyce, R., and Smith, J. P. (2011). The long shadow cast by childhood physical and mental problems on adult life. *Proceedings of the National Academy of Sciences*, 108(15):6032–6037.

Green, D. M., Cox, C. L., Zhu, L., Krull, K. R., Srivastava, D. K., Stovall, M., Nolan, V. G., Ness, K. K., Donaldson, S. S., Oeffinger, K. C., et al. (2011). Risk factors for obesity in adult survivors of childhood cancer: a report from the Childhood Cancer Survivor Study. *Journal of Clinical Oncology*, 30(3):246–255.

Green, D. M., Cox, C. L., Zhu, L., Krull, K. R., Srivastava, D. K., Stovall, M., Nolan, V. G., Ness, K. K., Donaldson, S. S., Oeffinger, K. C., et al. (2012). Risk factors for obesity in adult survivors of childhood cancer: a report from the Childhood Cancer Survivor Study. *Journal of Clinical Oncology*, 30(3):246.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge University Press, Cambridge.

Hagger-Johnson, G., Batty, G. D., Deary, I. J., and Von Stumm, S. (2011). Childhood socioeconomic status and adult health: comparing formative and reflective models in the Aberdeen Children of the 1950s Study (prospective cohort study). *Journal of Epidemiology and Community Health*, 65(11):1024–1029.

Hancock, M., Elliott, J., Johnson, J., and Bukodi, E. (2011). National child development study, partnership histories 1974-2008. A guide to datasets, Centre for Longitudinal Studies, Institute of Education, University of London.

Hardin, J. W. and Hilbe, J. M. (2002). *Generalized Estimating Equations (GEE): 1st Edition*. Wiley Online Library, London.

Hauser, R. M. (1994). Measuring socioeconomic status in studies of child development. *Child Development*, 65(6):1541–1545.

Hawkes, D. and Plewis, I. (2006). Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):479–491.

Heckman, J. J. (1977). Sample selection bias as a specification error (with an application to the estimation of labor supply functions). Technical report, National Bureau of Economic Research, Cambridge, Massachusetts, USA.

Hedeker, D., Siddiqui, O., and Hu, F. B. (2000). Random-effects regression analysis of correlated grouped-time survival data. *Statistical Methods in Medical Research*, 9(2):161–179.

Henderson, M., Richards, M., Stansfeld, S., and Hotopf, M. (2012). The association between childhood cognitive ability and adult long-term sickness absence in three British birth cohorts: a cohort study. *BMJ open*, 2(2):e000777.

Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, 16(1):117.

Hirano, K., Imbens, G. W., Ridder, G., and Rubin, D. B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69(6):1645–1659.

Hobcraft, J. (1998). Intergenerational and life-course transmission of social exclusion: Influences and childhood poverty, family disruption and contact with the police. STICERD Research Paper No. CASE015, London School of Economics, London.

Hogan, J. W. and Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, 16(3):239–257.

Hosmer, D. W. and Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modelling of Time to Event Data: 1st Edition*. John Wiley & Sons, New York.

Hosmer, D. W. and Lemeshow, S. (2013). *Applied logistic regression: Second edition*. John Wiley & Sons, New York.

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., and Pentz, M. A. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology*, 147(7):694–703.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statistica Sinica*, 14(3):863–883.

Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test*, 18(1):1–43.

Jackson, M. I. (2010). A life course perspective on child health, cognition and occupational skill qualifications in adulthood: Evidence from a British cohort. *Social Forces*, 89(1):89–116.

Jones, G. T., Power, C., and Macfarlane, G. J. (2009). Adverse events in childhood and chronic widespread pain in adult life: Results from the 1958 British Birth Cohort Study. *Pain*, 143(1):92–96.

Jung, T. and Wickrama, K. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, 2(1):302–317.

Kaakinen, M., Läärä, E., Pouta, A., Hartikainen, A.-L., Laitinen, J., Tammelin, T. H., Herzig, K.-H., Sovio, U., Bennett, A. J., Peltonen, L., et al. (2010). Life-course analysis of a fat mass and obesity-associated (FTO) gene variant and body mass index in the Northern Finland Birth Cohort 1966 using structural equation modeling. *American Journal of Epidemiology*, 172(6):653–665.

Kallis, C. (2005). Cls cohort studies data note 5: Partnership histories in NCDS5 and NCDS6. Technical report, Institute of Education, University of London.

Kelly, P. J. and Lim, L. L. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine*, 19(1):13–33.

Kelly-Irving, M., Lepage, B., Dedieu, D., Lacey, R., Cable, N., Bartley, M., Blane, D., Grosclaude, P., Lang, T., and Delpierre, C. (2013). Childhood adversity as a risk for cancer: findings from the 1958 British birth cohort study. *BMC Public Health*, 13(767).

Kiernan, K. E. and Cherlin, A. J. (1999). Parental divorce and partnership dissolution in adulthood: evidence from a British cohort study. *Population Studies*, 53(1):39–48.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data. 2nd Edition*. Springer, New York.

Kuh, D., Ben-Shlomo, Y., Lynch, J., Hallqvist, J., and Power, C. (2003). Life course epidemiology. *Journal of Epidemiology and Community Health*, 57(10):778.

Lacey, R. E., Bartley, M., Pikhart, H., Stafford, M., and Cable, N. (2014). Parental separation and adult psychological distress: an investigation of material and relational mechanisms. *BMC Public Health*, 14(1):272.

Lane, S. P., Bluestone, C., and Burke, C. T. (2013). Trajectories of BMI from early childhood through early adolescence: SES and psychosocial predictors. *British Journal of Health Psychology*, 18(1):66–82.

Lanza, S. T., Tan, X., and Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(1):1–26.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):325–335.

Lillard, L. A. (1993). Simultaneous equations for hazards: Marriage duration and fertility timing. *Journal of Econometrics*, 56(1-2):189–217.

Lillard, L. A., Brien, M. J., and Waite, L. J. (1995). Premarital cohabitation and subsequent marital. *Demography*, 32(3):437–457.

Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97(457):53–65.

Little, R. J. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3):471–483.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Liu, L., Wolfe, R. A., and Huang, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics*, 60(3):747–756.

Lo, Y., Mendell, N. R., and Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88(3):767–778.

Luo, S. and Wang, J. (2014). Bayesian hierarchical model for multiple repeated measures and survival data: an application to Parkinson's disease. *Statistics in Medicine*, 33(24):4279–4291.

Lyngstad, T. H. (2006). Why do couples with highly educated parents have higher divorce rates? *European Sociological Review*, 22(1):49–60.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1):83.

Maggs, J. L., Patrick, M. E., and Feinstein, L. (2008). Childhood and adolescent predictors of alcohol use and problems in adolescence and adulthood in the National Child Development Study. *Addiction*, 103(1):7–22.

Masyn, K. E. (2009). Discrete-time survival factor mixture analysis for low-frequency recurrent event histories. *Research in Human Development*, 6(2-3):165–194.

Maughan, B. and Taylor, A. (2001). Adolescent psychological problems, partnership transitions and adult mental health: an investigation of selection effects. *Psychological Medicine*, 31(2):291–305.

McCarron, P., Okasha, M., McEwen, J., and Smith, G. D. (2002). Height in young adulthood and risk of death from cardiorespiratory disease: a prospective study of male former students of Glasgow University, Scotland. *American Journal of Epidemiology*, 155(8):683–687.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Mensah, F. and Hobcraft, J. (2008). Childhood deprivation, health and development: associations with adult health in the 1958 and 1970 British prospective birth cohort studies. *Journal of Epidemiology and Community Health*, 62(7):599–606.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). *Handbook of Missing Data Methodology: 1st Edition.* CRC Press, London.

Molenberghs, G., Kenward, M. G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, 84(1):33–44.

Mostafa, T. and Wiggins, D. (2014). Handling attrition and non-response in the 1970 british cohort study. Technical report, Institute of Education, University of London.

Musoro, J. Z., Geskus, R. B., and Zwinderman, A. H. (2015). A joint model for repeated events of different types and multiple longitudinal outcomes with application to a follow-up study of patients after kidney transplant. *Biometrical Journal*, 57(2):185–200.

Muthén, B. and Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G., editors, *Longitudinal Data Analysis*, chapter 6, pages 143–165. Chapman and Hall/CRC, New York.

Muthén, B., Asparouhov, T., Hunter, A. M., and Leuchter, A. F. (2011). Growth modeling with non-ignorable dropout: Alternative analyses of the STAR*D antidepressant trial. *Psychological Methods*, 16(1):17–33.

Muthén, B. and Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics*, 30(1):27–58.

Muthén, B. O. (2001). Latent variable mixture modeling. In Marcoulides, G. A. and Schumacker, R. E., editors, *New developments and techniques in structural equation modeling*, pages 21–54.

Muthén, L. K. and Muthén, B. O. (2017). *Mplus User's Guide. Eighth Edition.* Muthén & Muthén, Los Angeles, CA.

Næss, Ø., Strand, B. H., and Smith, G. D. (2007). Childhood and adulthood socioeconomic position across 20 causes of death: a prospective cohort study of 800,000 Norwegian men and women. *Journal of Epidemiology & Community Health*, 61(11):1004–1009.

Nakao, K. and Treas, J. (1992). The 1989 socioeconomic index of occupations: Construction from the 1989 occupational prestige scores. Technical report, National Opinion Research Center, Chicago.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4):535–569.

Nylund-Gibson, K., Grimm, R., Quirk, M., and Furlong, M. (2014). A latent transition mixture model using the three-step specification. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3):439–454.

Orfei, L., Strachan, D. P., Rudnicka, A. R., and Wadsworth, M. (2008). Early influences on adult lung function in two national British cohorts. *Archives of disease in childhood*, 93(7):570–574.

Pakpahan, E., Hoffmann, R., and Kröger, H. (2017). The long arm of childhood circumstances on health in old age: Evidence from SHARELIFE. *Advances in Life Course Research*, 31:1–10.

Park, M. H., Sovio, U., Viner, R. M., Hardy, R. J., and Kinra, S. (2013). Overweight in childhood, adolescence and adulthood and cardiovascular risk in later life: pooled analysis of three British birth cohorts. *PLoS One*, 8(7).

Parsons, T. J., Power, C., Logan, S., and Summerbelt, C. (1999). Childhood predictors of adult obesity: a systematic review. *International Journal of Obesity*, 23.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc.

Ploubidis, G. B., Silverwood, R. J., DeStavola, B., and Grundy, E. (2015). Life-course partnership status and biomarkers in midlife: Evidence from the 1958 British birth cohort. *American Journal of Public Health*, 105(8):1596–1603.

Potter, C. M. and Ulijaszek, S. J. (2013). Predicting adult obesity from measures in earlier life. *Journal of Epidemiology and Community Health*, 67(12):1032–1037.

Poulton, R., Caspi, A., Milne, B. J., Thomson, W. M., Taylor, A., Sears, M. R., and Moffitt, T. E. (2002). Association between children's experience of socioeconomic disadvantage and adult health: a life-course study. *The Lancet*, 360(9346):1640–1645.

Power, C. and Jefferis, B. J. (2002). Fetal environment and subsequent obesity: a study of maternal smoking. *International Journal of Epidemiology*, 31(2):413–419.

Power, C., Jefferis, B. J., and Manor, O. (2010). Childhood cognition and risk factors for cardiovascular disease in midadulthood: the 1958 British birth cohort study. *American Journal of Public Health*, 100(1):129–136.

Proust-Lima, C. and Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics*, 10(3):535–549.

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14(1):53–68.

Rabe-Hesketh, S. and Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata, Volumes I and II, Third Edition*. Stata Press, Texas.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, 69(2):167–190.

Reardon, S. F., Brennan, R. T., and Buka, S. L. (2002). Estimating multi-level discrete-time hazard models using cross-sectional data: Neighborhood effects on the onset of adolescent cigarette use. *Multivariate Behavioral Research*, 37(3):297–330.

Rich, C., Cortina-Borja, M., Dezateux, C., Geraci, M., Sera, F., Calderwood, L., Joshi, H., and Griffiths, L. J. (2013). Predictors of non-response in a UK-wide cohort study of children's accelerometer-determined physical activity using postal methods. *BMJ Open*, 3(3).

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Rondeau, V., Mazroui, Y., and Gonzalez, J. R. (2012). frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47(4):1–28.

Roy, J. (2003). Modeling longitudinal data with non-ignorable dropouts using a latent dropout class model. *Biometrics*, 59(4):829–836.

Roy, J. and Lin, X. (2005). Missing covariates in longitudinal data with informative dropouts: Bias analysis and inference. *Biometrics*, 61(3):837–846.

Sacker, A., Schoon, I., and Bartley, M. (2002). Social inequality in educational achievement and psychosocial adjustment throughout childhood: magnitude and mechanisms. *Social Science & Medicine*, 55(5):863–880.

Salas, M., Hotman, A., and Stricker, B. H. (1999). Confounding by indication: an example of variation in the use of epidemiologic terminology. *American Journal of Epidemiology*, 149(11):981–983.

Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):667–678.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data: 1st Edition*. Chapman and Hall/CRC, London.

Schoon, I., Sacker, A., and Bartley, M. (2003). Socio-economic adversity and psychosocial adjustment: a developmental-contextual perspective. *Social Science & Medicine*, 57(6):1001–1015.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295.

Sigurdsson, E., Van Os, J., and Fombonne, E. (2002). Are impaired childhood motor skills a risk factor for adolescent anxiety? Results from the 1958 UK birth cohort and the National Child Development Study. *American Journal of Psychiatry*, 159(6):1044–1046.

Singer, J. D. and Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press, New York.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman and Hall/CRC, New York.

Snijders, T. A. and Bosker, R. J. (2000). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling: First Edition*. Sage, London.

Song, X., Davidian, M., and Tsiatis, A. A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58(4):742–753.

StataCorp, L. (2017). Stata statistical software: Release 15.

Steele, F. (2011). Multilevel discrete-time event history analysis with applications to the analysis of recurrent employment transitions. *Australian & New Zealand Journal of Statistics*, 53(1):1–20.

Steele, F., French, R., and Bartley, M. (2013). Adjusting for selection bias in longitudinal analyses using simultaneous equations modeling: the relationship between employment transitions and mental health. *Epidemiology*, 24(5):703–711.

Steele, F., Goldstein, H., and Browne, W. (2004). A general multilevel multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. *Statistical Modelling*, 4(2):145–159.

Steele, F., Kallis, C., Goldstein, H., and Joshi, H. (2005). The relationship between childbearing and transitions from marriage and cohabitation in britain. *Demography*, 42(4):647–673.

Steele, F., Kallis, C., and Joshi, H. (2006). The formation and outcomes of cohabiting and marital partnerships in early adulthood: the role of previous partnership experience. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):757–779.

Stone, R. (1993). The assumptions on which causal inferences rest. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 455–466.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 267–288.

Torfadottir, J. E., Steingrimsdottir, L., Mucci, L., Aspelund, T., Kasperzyk, J. L., Olafsson, O., Fall, K., Tryggvadottir, L., Harris, T. B., Launer, L., et al. (2012). Milk intake in early life and risk of advanced prostate cancer. *American Journal of Epidemiology*, 175(2):144–153.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Turrell, G., Lynch, J. W., Leite, C., Raghunathan, T., and Kaplan, G. A. (2007). Socioeconomic disadvantage in childhood and across the life course and all-cause mortality and physical function in adulthood: evidence from the Alameda County Study. *Journal of Epidemiology & Community Health*, 61(8):723–730.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M. G. (2001). Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics*, 57(1):7–14.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4):450–469.

Vermunt, J. K. and Magidson, J. (2015). *LG-Syntax User's Guide: Manual for Latent GOLD 5.0 Syntax Module*. Statistical Innovations Inc, Belmont, MA.

Wang, C.-P., Hendricks Brown, C., and Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association*, 100(471):1054–1076.

Washbrook, E., Clarke, P. S., and Steele, F. (2014). Investigating non-ignorable dropout in panel studies of residential mobility. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(2):239–266.

White, I. R. and Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29(28):2920–2931.

White, J. W., Gale, C. R., and Batty, G. D. (2012). Intelligence quotient in childhood and the risk of illegal drug use in middle-age: the 1958 National Child Development Survey. *Annals of Epidemiology*, 22(9):654–657.

WHO (2000). Obesity: preventing and managing the global epidemic. Report of a WHO Consultation (WHO Technical Report Series) 2000-894, World Health Organization.

Wiik, K. A. (2009). 'you'd better wait!'—socio-economic background and timing of first marriage versus first cohabitation. *European Sociological Review*, 25(2):139–153.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data: 2nd Edition*. MIT press, Cambridge,MA.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, pages 175–188.

Yatchew, A. and Griliches, Z. (1985). Specification error in probit models. *The Review of Economics and Statistics*, pages 134–139.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060.

Zhou, X.-H., Zhou, C., Lui, D., and Ding, X. (2014). *Applied Missing Data Analysis in the Health Sciences*. John Wiley & Sons, Hoboken.

Zhu, Y., Steele, F., and Moustaki, I. (2017). A general 3-step maximum likelihood approach to estimate the effects of multiple latent categorical variables on a distal outcome. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(5):643–656.

Appendix A

Results of latent class analysis on four childhood socio-economic concepts

A.1 Introduction

This appendix accompanies Chapter 4. Section A.2 presents the rationale behind the choice of the number of classes to be retained from estimating a latent class model. More simulation results for comparing the 3-step approach with the modified BCH and the modal class approaches are summarised for situations where modelling assumptions are not violated (Section A.3) and when they are (Section A.4). Codes for selected simulation studies are included in Section A.5. Codes for the empirical study is available in Section A.6.

A.2 The choice of the number of classes in the latent class analysis

The following tables show the judgement process of the number of classes to be retained in the analysis. In the tables, ssa-BIC stands for sample-size adjusted BIC and the last two columns report the p-value of Lo-Mendell-Rubin and bootstrap likelihood ratio test (comparing the likelihood of a model with $j-1$ classes versus that of a model with j classes), respectively. Figure A.1 plots different information criteria from four LCAs. Considering the relative proportion of each class (at least more than 10% in each category), interpretation of each class and the above statistics, three, two, three and two classes are retained for each childhood measure.

Table A.1 Measure 1: father/male head social class (2.92% has missing data in all childhood waves)

Classes	Log-likelihood	AIC	BIC	ssa-BIC	Entropy	LMR test	BLR test
1	-69553.7	139147.5	139302.8	139239.2	1	-	-
2	-62159.7	124401.3	124719.7	124589.4	0.836	0	0
3	-59279.7	118683.3	119164.8	118967.8	0.727	0	0
4	-57949.6	116065.2	116709.8	116446.0	0.763	0	0
5	-57203.0	114614.0	115421.6	115091.1	0.741	0	0
6	-56955.8	114161.5	115132.2	114735.0	0.755	0	0
7	-56830.0	113952.0	115085.8	114621.8	0.698	0	0

Table A.2 Measure 2: financial hardship (2.23% has missing data in all childhood waves)

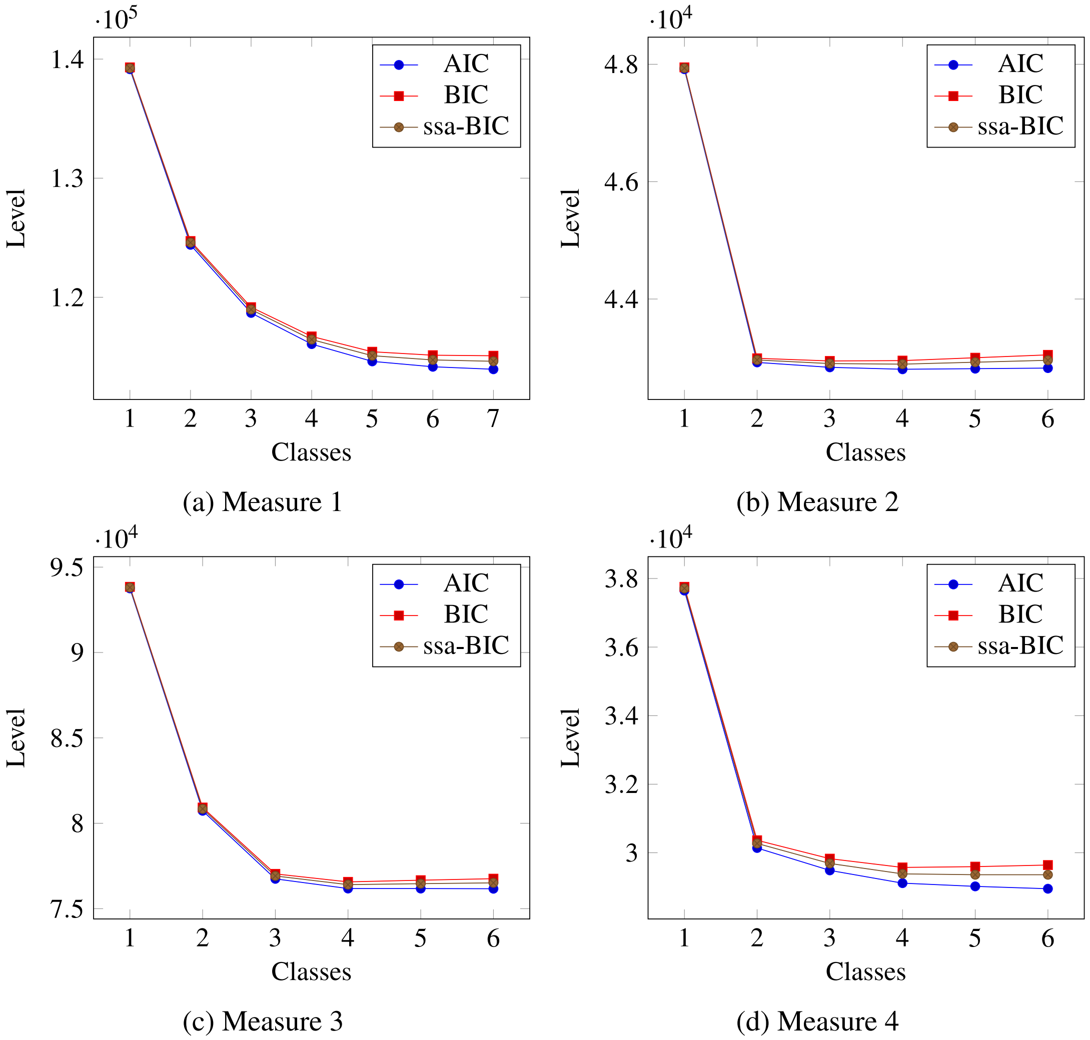
Classes	Log-likelihood	AIC	BIC	ssa-BIC	Entropy	LMR test	BLR test
1	-23954.1	47916.2	47947.3	47934.6	1	-	-
2	-21450.9	42919.9	42989.8	42961.2	0.700	0	0
3	-21403.9	42835.8	42944.5	42900.0	0.549	0	0
4	-21382.5	42802.9	42950.4	42890.1	0.494	0	0
5	-21382.5	42812.9	42999.3	42923.0	0.548	0	0
6	-21382.5	42822.9	43048.1	42955.9	0.334	0	0

Table A.3 Measure 3: material hardship (12.52% has missing data in all childhood waves)

Classes	Log-likelihood	AIC	BIC	ssa-BIC	Entropy	LMR test	BLR test
1	-46864.2	93752.4	93844.2	93806.0	1	-	-
2	-40340.3	80730.5	80921.8	80842.3	0.829	0	0
3	-38334.6	76745.2	77035.9	76915.1	0.790	0	0
4	-38038.1	76178.2	76568.3	76406.3	0.780	0	0
5	-38024.5	76177.1	76666.6	76463.2	0.706	0	0
6	-38006.3	76166.6	76755.6	76510.9	0.740	0	0

Table A.4 Measure 4: family structure (0.93% has missing data in all childhood waves)

Classes	Log-likelihood	AIC	BIC	ssa-BIC	Entropy	LMR test	BLR test
1	-18809.7	37647.5	37756.2	37711.7	1	-	-
2	-15040.5	30139.0	30364.2	30272.0	0.916	0	0
3	-14699.0	29486.0	29827.7	29687.9	0.928	0	0
4	-14497.8	29113.6	29571.8	29384.3	0.938	0	0
5	-14436.0	29020.0	29594.7	29359.5	0.946	0	0
6	-14386.3	28950.6	29641.7	29358.9	0.949	0	0



(a) Measure 1

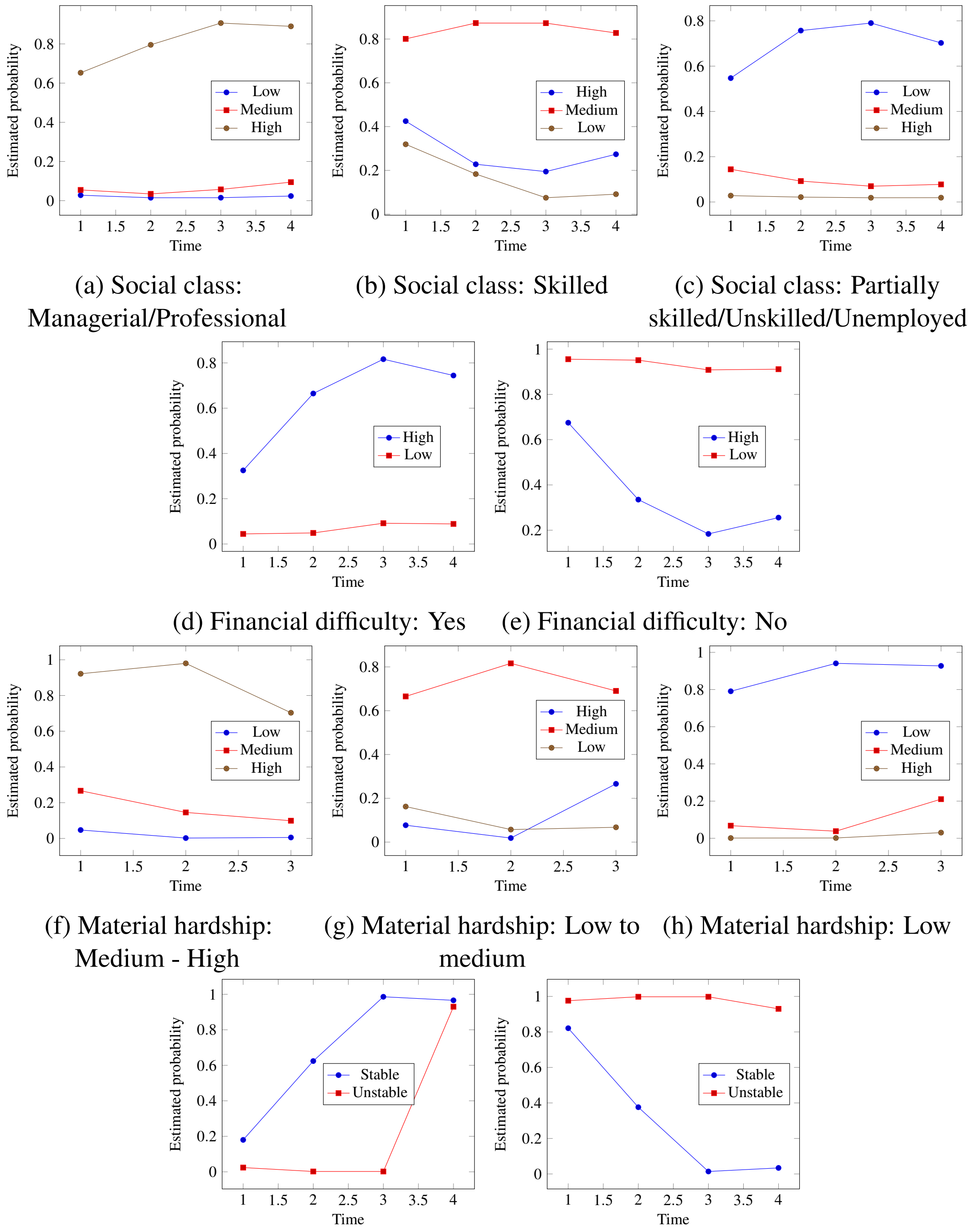
(b) Measure 2

(c) Measure 3

(d) Measure 4

Fig. A.1 Plots of AIC, BIC and ssa-BIC after performing individual LCA of four different childhood measures

In order to observe the patterns of change in each measure of the childhood SECs, for each categorical childhood measure, we plot the estimated conditional probability of having responses on each category given the class membership (not included here as this is an initial step). This reflects the probability of individuals in a particular class to endorse a certain category of the item. For each childhood measure, categories with high probabilities of endorsement are grouped. Figure A.2 shows the estimated probability of endorsing grouped categories (indicated clearly in sub-captions) given the class membership. These categories and labels of classes are available in Table 3.2 (Chapter 3) and Table 4.9 (Chapter 4). To illustrate the interpretation of these sub-figures, we use the first row as an example. The first picture shows that compared to those in the "Medium" and "Low" social class, individuals in the "High" social class have higher probabilities of being in managerial/professional categories of all four of the repeated measures of male head social class.



(a) Social class: Managerial/Professional

(b) Social class: Skilled

(c) Social class: Partially skilled/Unskilled/Unemployed

(d) Financial difficulty: Yes

(e) Financial difficulty: No

(f) Material hardship: Medium - High

(g) Material hardship: Low to medium

(h) Material hardship: Low

(i) Family structure: Joint parents

(j) Family structure: Other situations - step parents

Fig. A.2 Estimated probabilities of each grouped categories of each childhood measure, condtional on class membership

A.3 More on simulation studies: no modelling assumptions are violated.

A.3.1 Simulation results for a continuous H

Population parameters are set at values described in Chapter 4. We consider more scenarios here and present results for the modal class approach for reference as it has been commonly used in the literature. We also add a continuous covariate X in the models to mimic the real-world scenario. To evaluate the relative performance of estimation methods, the following statistics are calculated, following suggestions in Burton et al. (2006). First, bias is calculated as the raw difference between the mean of the estimated values across replications and the true values. Standard errors (SE), which are calculated as the average of estimated standard errors across replications, and standard deviations (SD) of mean estimates are also computed. A comparison of SE and SD is for the purpose to see if SEs are estimated unbiased. Third, the coverage rate is calculated as percentage of 95% confidence intervals across 500 replications of varied sample sizes ($N = 500$, 2000 and 10,000) that contain the true value of the parameter. Except for gains in estimation efficiency when sample sizes are large, results are rather similar across these sample sizes. We present below the results for $N = 2000$.

In addition to the results presented in the main text, we also find that in spite of the poor performance of the modal class approach in estimating most parameters, the estimated effect of the covariate X on H (β_3 for X) is unbiased in all situations but this is not true when a binary distal outcome is considered. In terms of the performance of the 3-step ML method, it produces unbiased estimates, close to 95% coverage rates, and good estimates for SEs (close to SDs) in all situations. As the class separation becomes unclear (with low values of entropy), the 3-step ML method substantially outperforms the modal class approach in terms of all evaluation criteria reported in the following tables.

Table A.5 to Table A.8 present the results of estimating a model with two independent latent categorical variables, a continuous covariate and a continuous distal outcome.

Table A.5 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy=0.8)

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.03	0.03	0.96	0.04	0.02	0.03	0.56
β_1 (C_1)	0.00	0.02	0.02	0.95	-0.22	0.02	0.02	0.00
β_2 (C_2)	0.00	0.03	0.03	0.95	0.17	0.02	0.02	0.00
β_3 (X)	0.00	0.01	0.01	0.94	0.00	0.01	0.01	0.94
Structural								
$\omega_1^{(C_1)}$	0.00	0.04	0.04	0.96				
$\omega_1^{(C_2)}$	0.00	0.04	0.04	0.95				
$\omega_{11}^{(C_1C_2)}$	0.00	0.05	0.05	0.96				

Table A.6 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy=0.7)

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.09	0.10	0.93	0.07	0.03	0.03	0.21
β_1 (C_1)	0.00	0.08	0.09	0.95	0.37	0.02	0.02	0.00
β_2 (C_2)	0.00	0.09	0.09	0.95	0.27	0.02	0.03	0.00
β_3 (X)	0.00	0.04	0.04	0.94	0.00	0.01	0.01	0.96
Structural								
$\omega_1^{(C_1)}$	0.00	0.05	0.06	0.95				
$\omega_1^{(C_2)}$	0.00	0.05	0.06	0.95				
$\omega_{11}^{(C_1C_2)}$	0.00	0.09	0.09	0.95				

Table A.7 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy=0.55)

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.03	0.04	0.88	0.11	0.03	0.03	0.02
β_1 (C_1)	0.00	0.03	0.03	0.95	−0.58	0.03	0.03	0.00
β_2 (C_2)	0.00	0.03	0.03	0.94	0.43	0.03	0.03	0.00
β_3 (X)	0.00	0.01	0.01	0.97	0.00	0.01	0.01	0.97
Structural								
$\omega_1^{(C_1)}$	0.00	0.05	0.06	0.92				
$\omega_1^{(C_2)}$	0.00	0.06	0.06	0.90				
$\omega_{11}^{(C_1C_2)}$	−0.01	0.07	0.07	0.96				

Table A.8 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy=0.4)

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.04	0.06	0.84	0.14	0.03	0.06	0.16
β_1 (C_1)	−0.01	0.04	0.04	0.96	−0.83	0.03	0.07	0.00
β_2 (C_2)	0.01	0.04	0.05	0.94	0.62	0.03	0.05	0.00
β_3 (X)	0.00	0.01	0.01	0.97	0.00	0.01	0.01	0.97
Structural								
$\omega_1^{(C_1)}$	0.00	0.07	0.10	0.84				
$\omega_1^{(C_2)}$	0.00	0.07	0.11	0.81				
$\omega_{11}^{(C_1C_2)}$	0.01	0.09	0.10	0.94				

Results for estimating a model with two associated latent categorical variables are summarised in Table A.9 to Table A.11. In this simulation study, situations in which the latent variables are positively and negatively associated are simulated ($\omega_{11}^{C_1C_2} = +/-0.5$). As estimated values are fairly close in both cases, only the results for estimating a model with negatively associated C_1 and C_2 are reported.

Table A.9 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy = 0.8 in both LCAs)

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.03	0.04	0.95	0.09	0.03	0.05	0.21
β_1 (C_1)	0.00	0.03	0.03	0.94	−0.48	0.03	0.03	0.00
β_2 (C_2)	0.00	0.03	0.03	0.96	0.39	0.03	0.03	0.00
β_3 (X)	0.00	0.01	0.01	0.94	0.00	0.01	0.01	0.94
Structural								
$\omega_1^{(C_1)}$	0.00	0.05	0.06	0.92				
$\omega_1^{(C_2)}$	0.01	0.05	0.06	0.94				
$\omega_{11}^{(C_1C_2)}$	−0.01	0.07	0.07	0.94				

Table A.10 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy = 0.7 and 0.4, respectively in two LCAs)

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.04	0.04	0.95	0.09	0.04	0.07	0.26
β_1 (C_1)	0.00	0.04	0.04	0.95	−0.52	0.03	0.03	0.00
β_2 (C_2)	0.00	0.01	0.01	0.93	0.59	0.03	0.05	0.00
β_3 (X)	0.00	0.07	0.08	0.94	0.00	0.01	0.01	0.94
Structural								
$\omega_1^{(C_1)}$	0.00	0.07	0.08	0.94				
$\omega_1^{(C_2)}$	0.01	0.08	0.11	0.81				
$\omega_{11}^{(C_1C_2)}$	−0.01	0.09	0.09	0.95				

Table A.11 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a continuous distal outcome (entropy = 0.4 in both LCAs)

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.05	0.07	0.81	0.09	0.04	0.11	0.54
β_1 (C_1)	0.00	0.04	0.04	0.93	−0.79	0.04	0.08	0.00
β_2 (C_2)	0.01	0.05	0.05	0.93	0.63	0.04	0.05	0.00
β_3 (X)	0.00	0.01	0.01	0.95	0.00	0.01	0.01	0.94
Structural								
$\omega_1^{(C_1)}$	0.00	0.09	0.13	0.82				
$\omega_1^{(C_2)}$	0.01	0.09	0.14	0.82				
$\omega_{11}^{(C_1C_2)}$	0.00	0.11	0.12	0.93				

A.3.2 Simulation results for a binary *H*

We now consider a binary distal outcome and perform the simulation following the procedure described in Section 4.5.1 in Chapter A. Results are summarised in Table A.12 to Table A.15 for models with independent latent categorical variables and in Table A.16 to Table A.17 for models with associated latent variables ($\omega_{11}^{(C_1C_2)} = -0.5$).

Table A.12 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a binary distal outcome (entropy=0.8).

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.07	0.07	0.95	−0.08	0.06	0.06	0.76
β_1 (C_1)	−0.01	0.07	0.07	0.96	−0.13	0.06	0.06	0.34
β_2 (C_2)	0.00	0.07	0.07	0.96	0.20	0.06	0.06	0.09
β_3 (X)	0.00	0.04	0.05	0.93	−0.05	0.04	0.04	0.77
Structural								
$\omega_1^{(C_1)}$	0.00	0.04	0.05	0.95				
$\omega_1^{(C_2)}$	0.00	0.04	0.04	0.95				
$\omega_{11}^{(C_1C_2)}$	0.00	0.06	0.06	0.95				

Table A.13 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a binary distal outcome (entropy=0.7).

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.07	0.07	0.95	−0.13	0.06	0.06	0.41
β_1 (C_1)	0.00	0.07	0.07	0.96	−0.21	0.06	0.06	0.03
β_2 (C_2)	0.00	0.08	0.08	0.96	0.33	0.06	0.06	0.00
β_3 (X)	0.00	0.04	0.05	0.94	−0.08	0.04	0.04	0.53
Structural								
$\omega_1^{(C_1)}$	0.00	0.05	0.05	0.93				
$\omega_1^{(C_2)}$	0.00	0.05	0.05	0.93				
$\omega_{11}^{(C_1C_2)}$	0.00	0.06	0.06	0.95				

Table A.14 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a binary distal outcome (entropy=0.55).

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.09	0.09	0.93	−0.19	0.06	0.06	0.07
β_1 (C_1)	0.00	0.08	0.08	0.96	−0.32	0.05	0.05	0.00
β_2 (C_2)	0.00	0.09	0.09	0.95	0.50	0.06	0.06	0.00
β_3 (X)	0.00	0.05	0.05	0.94	−0.11	0.04	0.04	0.22
Structural								
$\omega_1^{(C_1)}$	0.00	0.06	0.07	0.92				
$\omega_1^{(C_2)}$	0.00	0.06	0.07	0.93				
$\omega_{11}^{(C_1C_2)}$	0.00	0.08	0.08	0.96				

Table A.15 Simulation results for estimating a model with 2 latent categorical variables, a covariate and a binary distal outcome (entropy=0.4).

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.11	0.11	0.94	−0.22	0.06	0.13	0.24
β_1 (C_1)	−0.02	0.11	0.11	0.96	−0.46	0.06	0.06	0.00
β_2 (C_2)	0.03	0.12	0.12	0.94	0.69	0.06	0.07	0.00
β_3 (X)	−0.01	0.05	0.05	0.93	−0.16	0.04	0.04	0.03
Structural								
$\omega_1^{(C_1)}$	−0.01	0.10	0.13	0.87				
$\omega_1^{(C_2)}$	0.00	0.13	0.13	0.87				
$\omega_{11}^{(C_1C_2)}$	0.01	0.14	0.14	0.95				

Table A.16 Simulation results for estimating a model with 2 associated latent categorical variables, a covariate and a binary distal outcome (entropy = 0.8 in both LCAs).

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.09	0.09	0.95	−0.14	0.06	0.07	0.40
β_1 (C_1)	0.01	0.08	0.08	0.96	−0.28	0.06	0.06	0.00
β_2 (C_2)	0.00	0.09	0.09	0.95	0.42	0.06	0.06	0.00
β_3 (X)	0.00	0.04	0.05	0.94	−0.07	0.04	0.04	0.57
Structural								
$\omega_1^{(C_1)}$	0.00	0.06	0.07	0.94				
$\omega_1^{(C_2)}$	0.01	0.06	0.07	0.94				
$\omega_{11}^{(C_1C_2)}$	−0.01	0.07	0.07	0.95				

Table A.17 Simulation results for estimating a model with 2 associated latent categorical variables, a covariate and a binary distal outcome (entropy = 0.4 in both LCAs).

Parameters	3-step				Modal class			
	Bias	SE	SD	Coverage	Bias	SE	SD	Coverage
Regression								
β_0 (Intercept)	0.00	0.12	0.13	0.94	−0.16	0.09	0.11	0.57
β_1 (C_1)	0.00	0.13	0.13	0.97	−0.46	0.07	0.07	0.00
β_2 (C_2)	0.00	0.14	0.15	0.95	0.64	0.08	0.09	0.00
β_3 (X)	0.00	0.05	0.05	0.95	−0.13	0.04	0.04	0.11
Structural								
$\omega_1^{(C_1)}$	0.00	0.13	0.16	0.88				
$\omega_1^{(C_2)}$	0.01	0.13	0.16	0.88				
$\omega_{11}^{(C_1C_2)}$	0.00	0.17	0.17	0.96				

A.4 More simulation studies: modelling assumptions are violated

According to Bakk and Vermunt (2016), the 3-step ML approach with one latent variable provides approximately unbiased estimates at medium and low degrees of bimodality. We use a similar specification for the model with two latent variables but extend earlier work by generating the residual ε from a bimodal mixture distribution: $0.5N(-1,0.5)+0.5N(1,0.5)$ (Bakk and Vermunt, 2016) for class pattern [1 2] and from $N(0,1)$ for other class patterns. Results are presented for four combinations of sample size and entropy levels.

We observe that in low entropy cases, both the 1-step and the general 3-step ML approaches produce biased estimates (greater than 5% in terms of relative bias) and poor coverage, irrespective of the sample size, although the bias is much smaller for the 3-step ML approach. In high entropy cases, both the 1-step and the 3-step ML approaches produce unbiased estimates with close to the nominal 95% coverage. In this study, the degree of bimodality is medium (according to Bakk and Vermunt (2016)) and our results for the 3-step ML approach are broadly consistent with their findings [1]. It is also worth noting that when class separation is poor, estimates for parameters in the log-linear model are biased. This agrees with earlier findings of Muthén and Muthén (2017) and Bakk and Vermunt (2016) that the class proportions are shifted in Step 3 when the normality assumption does not hold.

[1] Note that due to the unavailability of the automatic 3-step module for two or more LVs in Mplus, we manually monitor the shift in classifications after Step 1 and Step 3. Replications with substantial classification shifts (i.e. over 20% of observations have different classification patterns of $[C_1,C_2]$ in Step 2 and in Step 3) are discarded. For example, for the case of $N=2000$ and high entropy, we retained only 130/500 replications. The same issue has also been reported in Asparouhov and Muthén (2014b). For other forms of violation of distributional assumptions, we encounter the same problem of the shifted class solution. Similar manual assessments are employed.

Table A.18 Study 2c: Simulation results for bimodality (N=200,2000; 500 replications).

Parameters		1-step				3-step			
	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
N=200, High entropy									
β_1	2.00	−6.39	0.28	0.73	0.93	−1.26	0.18	0.18	0.93
β_2	−1.50	−5.23	0.28	0.56	0.92	−1.49	0.18	0.19	0.93
$\omega_1^{(C_1)}$	0.70	−2.97	0.26	0.27	0.93	−1.94	0.28	0.28	0.96
$\omega_1^{(C_2)}$	0.70	−1.43	0.27	0.29	0.91	−0.18	0.28	0.29	0.94
$\omega_{11}^{(C_1C_2)}$	−0.50	−0.34	0.32	0.33	0.89	−6.43	0.36	0.36	0.97
N=200, Low entropy									
β_1	2.00	−57.17	0.43	1.91	0.64	−15.76	0.31	0.42	0.80
β_2	−1.50	−33.39	0.40	1.47	0.55	−5.76	0.34	0.44	0.84
$\omega_1^{(C_1)}$	0.70	−29.01	0.34	0.56	0.63	6.69	0.74	0.87	0.77
$\omega_1^{(C_2)}$	0.70	0.07	0.38	0.60	0.66	−2.12	0.73	0.84	0.73
$\omega_{11}^{(C_1C_2)}$	−0.50	5.80	0.44	0.59	0.66	−8.24	0.93	0.96	0.90
N=2000, High entropy									
β_1	2.00	−0.51	0.09	0.09	0.95	−0.62	0.06	0.05	0.95
β_2	−1.50	0.31	0.08	0.09	0.94	−0.72	0.06	0.06	0.96
$\omega_1^{(C_1)}$	0.70	−0.19	0.08	0.09	0.94	−2.09	0.09	0.09	0.94
$\omega_1^{(C_2)}$	0.70	0.26	0.09	0.08	0.96	−1.06	0.09	0.09	0.95
$\omega_{11}^{(C_1C_2)}$	−0.50	−0.18	0.11	0.11	0.95	−3.90	0.11	0.11	0.95
N=2000, Low entropy									
β_1	2.00	−52.76	0.13	1.84	0.67	1.02	0.09	0.25	0.80
β_2	−1.50	−21.55	0.13	1.20	0.62	12.62	0.11	0.18	0.51
$\omega_1^{(C_1)}$	0.70	−33.63	0.12	0.32	0.63	−35.66	0.15	0.24	0.50
$\omega_1^{(C_2)}$	0.70	−10.43	0.13	0.33	0.67	−22.91	0.14	0.21	0.68
$\omega_{11}^{(C_1C_2)}$	−0.50	−7.04	0.16	0.25	0.78	−67.62	0.20	0.26	0.52

Bias (%)=(Estimate-True)/True × 100%

Table A.19 Study 2a: Simulation results for excess kurtosis (500 replications): log-linear ω parameters.

Parameters		1-step				3-step			
	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
N=200, High entropy									
$\omega_1^{(C_1)}$	0.70	−3.27	0.63	0.27	0.92	−6.66	0.27	0.28	0.95
$\omega_1^{(C_2)}$	0.70	−3.25	0.27	0.29	0.91	−8.30	0.28	0.29	0.94
$\omega_{11}^{(C_1C_2)}$	−0.50	2.00	0.32	0.34	0.88	−18.71	0.36	0.36	0.96
N=200, Low entropy									
$\omega_1^{(C_1)}$	0.70	−2.17	0.48	0.50	0.83	90.53	0.76	3.15	0.94
$\omega_1^{(C_2)}$	0.70	−2.30	0.46	0.60	0.68	85.15	0.77	3.20	0.94
$\omega_{11}^{(C_1C_2)}$	−0.50	10.72	0.56	0.61	0.71	105.99	19.27	3.43	0.99
N=2000, High entropy									
$\omega_1^{(C_1)}$	0.70	−0.05	0.08	0.09	0.95	−6.30	0.09	0.09	0.91
$\omega_1^{(C_2)}$	0.70	−0.04	0.09	0.09	0.95	−6.81	0.09	0.09	0.90
$\omega_{11}^{(C_1C_2)}$	−0.50	−0.34	0.11	0.11	0.94	−15.31	0.11	0.11	0.88
N=2000, Low entropy									
$\omega_1^{(C_1)}$	0.70	−26.13	0.11	0.34	0.67	−9.35	0.15	0.21	0.80
$\omega_1^{(C_2)}$	0.70	−26.00	0.12	0.35	0.68	−11.69	0.15	0.23	0.76
$\omega_{11}^{(C_1C_2)}$	−0.50	15.90	0.17	0.22	0.83	−27.51	0.21	0.21	0.88

Bias (%)=(Estimate-True)/True × 100%

Table A.20 Study 2b: Simulation results for skewness (500 replications): log-linear ω parameters.

Parameters		1-step				3-step			
	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
N=200, High entropy									
$\omega_1^{(C_1)}$	0.70	−41.49	0.17	0.35	0.57	−3.27	0.28	0.28	0.95
$\omega_1^{(C_2)}$	0.70	−42.03	0.16	0.37	0.53	−2.04	0.28	0.29	0.95
$\omega_{11}^{(C_1C_2)}$	−0.50	−15.18	0.23	0.39	0.65	−9.74	0.36	0.36	0.96
N=200, Low entropy									
$\omega_1^{(C_1)}$	0.70	48.43	0.70	0.78	0.78	827.31	0.75	32.94	0.85
$\omega_1^{(C_2)}$	0.70	32.81	0.69	0.84	0.72	823.43	0.75	32.86	0.83
$\omega_{11}^{(C_1C_2)}$	−0.50	−27.62	0.82	0.92	0.78	180.85	0.49	0.43	0.93
N=2000, High entropy									
$\omega_1^{(C_1)}$	0.70	−53.74	0.05	0.31	0.39	−4.83	0.09	0.09	0.92
$\omega_1^{(C_2)}$	0.70	−52.38	0.05	0.30	0.41	−4.29	0.09	0.09	0.93
$\omega_{11}^{(C_1C_2)}$	−0.50	−6.32	0.10	0.25	0.94	−8.71	0.11	0.11	0.93
N=2000, Low entropy									
$\omega_1^{(C_1)}$	0.70	−990.00	0.02	0.29	0.03	39.44	0.38	0.37	0.15
$\omega_1^{(C_2)}$	0.70	−989.00	0.02	0.30	0.03	38.13	0.38	0.38	0.18
$\omega_{11}^{(C_1C_2)}$	−0.50	167.90	0.32	0.51	0.29	124.56	0.51	0.66	0.18

Bias (%)=(Estimate-True)/True × 100%

Table A.21 Summary of results for class-specific effects on the distal outcome across all situations investigated, using the 1-step, 3-step ML and 3-step BCH approaches: Bimodality

		1-step				3-step ML				BCH			
Scenarios	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
N=200, High entropy													
$\beta_1(C_1)$	2.00	−6.39	0.28	0.73	0.93	−1.26	0.18	0.18	0.93	−3.34	0.19	0.53	0.91
$\beta_2(C_2)$	−1.50	−5.23	0.28	0.56	0.92	−1.49	0.18	0.19	0.93	−0.44	0.2	0.27	0.94
N=200, Low entropy													
$\beta_1(C_1)$	2.00	−57.17	0.43	1.91	0.64	−15.76	0.31	0.42	0.80	−51.96	0.39	1.56	0.68
$\beta_2(C_2)$	−1.50	−33.39	0.40	1.47	0.55	−5.76	0.34	0.44	0.84	−53.68	0.41	1.29	0.68
N=2000, High entropy													
$\beta_1(C_1)$	2.00	−0.51	0.09	0.09	0.95	−0.62	0.06	0.05	0.95	2.23	0.06	0.06	0.90
$\beta_2(C_2)$	−1.50	0.31	0.08	0.09	0.94	−0.72	0.06	0.06	0.96	1.90	0.06	0.06	0.93
N=2000, Low entropy													
$\beta_1(C_1)$	2.00	−52.76	0.13	1.84	0.67	1.02	0.09	0.25	0.80	−5.43	0.13	0.51	0.92
$\beta_2(C_2)$	−1.50	−21.55	0.13	1.20	0.62	12.62	0.11	0.18	0.51	−4.49	0.15	0.34	0.94

Table A.22 Summary of results for class-specific effects on the distal outcome across all situations investigated, using the 1-step, 3-step ML and 3-step BCH approaches: Skewness

		1-step				3-step ML				BCH			
Scenarios	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
N=200, High entropy													
$\beta_1(C_1)$	2.00	−65.44	0.38	1.64	0.57	−7.31	0.27	0.23	0.90	−3.77	0.29	0.60	0.92
$\beta_2(C_2)$	−1.50	66.67	0.45	1.44	0.60	−5.66	0.27	0.24	0.94	−0.59	0.29	0.36	0.95
N=200, Low entropy													
$\beta_1(C_1)$	2.00	−71.66	1.32	2.15	0.58	−22.73	0.46	4.91	0.56	−53.70	0.51	1.60	0.67
$\beta_2(C_2)$	−1.50	−24.77	1.34	2.16	0.72	−41.29	0.46	4.87	0.76	−52.51	0.52	1.31	0.69
N=2000, High entropy													
$\beta_1(C_1)$	2.00	−84.37	0.13	1.74	0.47	−7.25	0.09	0.07	0.63	2.66	0.09	0.09	0.92
$\beta_2(C_2)$	−1.50	72.31	0.13	1.23	0.54	−6.22	0.09	0.07	0.82	1.90	0.09	0.16	0.94
N=2000, Low entropy													
$\beta_1(C_1)$	2.00	−150.11	0.12	0.68	0.03	−28.89	0.14	0.33	0.20	−5.86	0.17	0.52	0.92
$\beta_2(C_2)$	−1.50	145.00	0.14	0.38	0.03	−38.09	0.14	0.33	0.25	−4.34	0.18	0.35	0.95

Table A.23 Summary of results for class-specific effects on the distal outcome across all situations investigated, using the 1-step, 3-step ML and 3-step BCH approaches: Kurtosis

		1-step				3-step ML				BCH			
Scenarios	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
N=200, High entropy													
$\beta_1(C_1)$	2.00	−4.48	0.30	0.73	0.90	0.60	0.19	0.19	0.94	−3.43	0.20	0.53	0.92
$\beta_2(C_2)$	−1.50	−7.80	0.30	0.60	0.92	−0.30	0.19	0.19	0.96	−1.44	0.21	0.28	0.95
N=200, Low entropy													
$\beta_1(C_1)$	2.00	−8.43	0.59	0.77	0.89	−14.38	0.33	0.44	0.84	−52.40	0.40	1.56	0.68
$\beta_2(C_2)$	−1.50	−1.45	0.59	0.63	0.88	−25.14	0.36	0.79	0.85	−54.38	0.42	1.28	0.68
N=2000, High entropy													
$\beta_1(C_1)$	2.00	0.03	0.09	0.09	0.95	0.96	0.06	0.06	0.94	2.28	0.06	0.07	0.87
$\beta_2(C_2)$	−1.50	−0.09	0.09	0.09	0.97	1.57	0.06	0.06	0.95	2.17	0.06	0.07	0.93
N=2000, Low entropy													
$\beta_1(C_1)$	2.00	−52.03	0.14	1.72	0.72	−2.68	0.10	0.10	0.92	−5.28	0.14	0.51	0.92
$\beta_2(C_2)$	−1.50	55.01	0.15	1.29	0.70	−3.21	0.11	0.11	0.92	−3.87	0.15	0.35	0.95

Table A.24 Summary of results for class-specific effects on the distal outcome across all situations investigated, using the 1-step, 3-step ML and 3-step BCH approaches: Conditional dependence (1 latent variable)

			1-step				3-step ML				BCH		
Scenarios	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
N=500, High entropy													
$\beta_3(C_3)$	1.50	−11.35	0.32	0.32	0.92	−6.69	0.27	0.30	0.89	−7.64	0.27	0.30	0.89
$\beta_4(C_4)$	3.00	−5.93	0.35	0.35	0.92	−3.54	0.29	0.31	0.93	−3.99	0.29	0.31	0.91
N=500, Low entropy													
$\beta_3(C_3)$	1.50	−59.52	0.59	0.65	0.52	9.01	0.40	0.76	0.70	−7.20	0.43	0.66	0.79
$\beta_4(C_4)$	3.00	−9.70	0.68	0.93	0.66	−6.68	0.44	0.92	0.60	−9.76	0.44	0.67	0.73
N=2000, High entropy													
$\beta_3(C_3)$	1.50	−11.54	0.15	0.16	0.79	−10.98	0.13	0.14	0.72	−13.04	0.14	0.15	0.69
$\beta_4(C_4)$	3.00	−6.04	0.16	0.16	0.82	−6.07	0.14	0.14	0.75	−6.58	0.14	0.14	0.72
N=2000, Low entropy													
$\beta_3(C_3)$	1.50	−60.00	0.30	0.36	0.14	−10.95	0.21	0.48	0.62	−21.22	0.26	0.51	0.63
$\beta_4(C_4)$	3.00	−10.67	0.52	0.70	0.64	−10.64	0.24	0.63	0.50	−11.37	0.26	0.59	0.52

A.5 Syntax for selected simulation studies

A.5.1 Latent GOLD program for the 3-step approach in Study 1 (2-LV)

```
[[init]] /*create multiple lgs files*/
iterators = rep;
rep = 1:500;
filename = "simlgs_[[rep]].lgs";
outputDirectory = "C:\2lcsim\simlgs";
[[/init]]
//LG5.0//
version = 5.0
infile ’C:\2lcsim\simas\2lc_[[rep]].txt’ quote = single
model
options
output parameters=first standarderrors
write=’C:\2lcsim\simlgs\output_[[rep]].txt’;
variables
dependent M1 nominal 2, M2 nominal 2, Z nominal;
independent X numeric;
latent lat1 nominal 2 coding=2, lat2 nominal 2 coding=2;
equations
lat1<-1; lat2<-1;
/*D~wei and F~wei are P(M|C) dervied from separate LCA in Step 1*/
/* These values are inserted into {}s in the last line.*/
M1<- (D~wei) 1| lat1;    M2<- (F~wei) 1| lat2;
Z <- 1+ lat1 + lat2 + X; lat1<->lat2;
D={}; F={};
```

A.5.2 Mplus program for Study 3 that investigates the change in the number of classes needed when there is local dependence of items and the outcome, using the 1-step approach (1-LV)

```
title: highentropy2000_H0:3-class; H1:4-class;
DATA:FILE = "C:\dat\high2000_list.dat";
TYPE = MONTECARLO;
VARIABLE:
NAMES = y1-y10 Z L; %ys=item, Z=distal outcome
USEVARIABLES = y1-y10 Z;
CATEGORICAL=y1-y10;
classes = lat(4); !4-class model
analysis:
type = mixture;processors=8;Starts=0;
model:
%overall%
[lat#1*2.282];[lat#2*1.526];[lat#3*1.526];
%lat#1%
[y1$1-y10$1*-1.5];[Z*3.5] (p1);
%lat#2%
[y1$1-y5$1*1.5];[y6$1-y10$1*-1.5];[Z*5] (p2);
%lat#3%
[y1$1-y5$1*-1.5];[y6$1-y10$1*1.5];[Z*1.5] (p3);
%lat#4%
[y1$1-y10$1*1.5];[Z*3] (p4);
MODEL CONSTRAINT:
New(b0*3);b0 = p4;
New(b1*0.5);b1 = p1-p4;
New(b2*2);b2 = p2-p4;
New(b3*-1.5);b3=p3-p4;
OUTPUT: TECH14; %request BLRT results
SAVEDATA: results = high2000_1step.txt;
```

A.5.3 Mplus program for Study 3 that investigates the change in the number of classes needed when there is local dependence of items and the outcome, using the 3-step approach (1-LV)

```
title: highentropy2000_H0:3-class; H1:4-class;
DATA:FILE = "C:\dat\high2000_list.dat";
TYPE = MONTECARLO;
VARIABLE:
NAMES = y1-y10 Z L;
USEVARIABLES=y1-y10 Z;
AUXILIARY=Z(DU3STEP);
CATEGORICAL = y1-y10;
CLASSES = lat(4);
ANALYSIS: type = mixture; starts = 0; PROCESSORS = 8;
MODEL:
%OVERALL%
[lat#1*0.405];[lat#2*0.223];[lat#3*0.223];
%lat#1%
[y1$1-y10$1*-1.5];
%lat#2%
[y1$1-y5$1*1.5];[y6$1-y10$1*-1.5];
%lat#3%
[y1$1-y5$1*-1.5];[y6$1-y10$1*1.5];
%lat#4%
[y1$1-y10$1*1.5];
SAVEDATA: results = high2000_3step.txt;
```

A.6 Syntax for the empirical study

Latent GOLD syntax for applying the 3-step ML approach to the empirical example

```
options
output parameters=effect standarderrors probmeans=posterior profiler
classification ParameterCovariances frequencies bivariateresiduals iterationdetails;
variables
dependent m1 nominal 3, m2 nominal 2, m3 nominal 3, m4 nominal 2, y nominal coding=first;
independent gender nominal coding=2, male nominal coding=2, ovwt16 nominal coding=2;
latent l1 nominal 3 coding=3, l2 nominal 2 coding=2,
l3 nominal 3 coding=3, l4 nominal 2 coding=2;
equations
l1<-1;l2<-1;l3<-1;l4<-1;
m1<- (C~wei) 1| l1; m2<- (D~wei) 1| l2; m3<- (E~wei) 1| l3; m4<- (F~wei) 1| l4;
y<- 1+l1 + l2 + l3 + l4 + male + ovwt16;
l1<->l2;l1<->l3;l1<->l4; l2<->l3;l2<->l4; l3<->l4;
C={0.824 0.169 0.007
0.061 0.910 0.028
0.008 0.093 0.899};
D={0.734 0.266 0.032 0.968};
E={0.854 0.097 0.049
0.055 0.943 0.002
0.053 0.003 0.944};
F={0.837 0.163 0.003 0.997};
```

Appendix B

Proof of estimation bias due to endogeneity

This appendix accompanies Chapter 5. The following equations prove the existence and direction of bias due to the endogenous y_1 ($corr(u_1,u_2) \neq 0$). We use the same model considered in the simulation study described in Chapter 5 (see Appendix C):

$$\begin{aligned} y_1 &= \alpha_0 X + \alpha_1 Z + u_1, \\ y_2 &= \beta_0 X + \beta_1 y_1 + u_2, \end{aligned} \tag{B.1}$$

where Z denotes the instrumental variable, $(u_1,u_2) \sim MVN(0,\Omega)$ with $\Omega = \begin{pmatrix} \sigma_{u_1}^2 & \\ \sigma_{u_{12}} & \sigma_{u_2}^2 \end{pmatrix}$. As an instrumental variable, the following conditions need to be satisfied,

$$\begin{aligned} corr(Z,u_1) &= 0 \\ corr(Z,u_2) &= 0 \\ corr(Z,y_1) &\neq 0. \end{aligned} \tag{B.2}$$

If we model both equations for y_1 and y_2, the estimates of main interest, $\hat{\beta}_0$ and $\hat{\beta}_1$ will be unbiased as after rearrangement, we have

$$\begin{aligned} y_2 &= \beta_0 X + \beta_1(\alpha_0 X + \alpha_1 Z + u_1) + u_2 \\ &= (\beta_0 + \beta_1\alpha_0)X + (\beta_1\alpha_1)Z + (\beta_1 u_1 + u_2) \\ &= \tilde{\beta}_0 X + \tilde{\beta}_1 Z + \tilde{u}. \end{aligned} \tag{B.3}$$

It is obvious from (B.3) that Z and X are both not correlated with $\tilde{u}$ such that ordinary least square (OLS) estimation can yield unbiased estimates for $\tilde{\beta_0}$, $\tilde{\beta_1}$, and $\sigma^2_{\tilde{u}}$. As (B.1) is self-identified, $\hat{\alpha_0}$ and $\hat{\alpha_1}$ can be estimated using the OLS for the equation of y_1. After rearrangement, all parameters can be estimated and estimates are unbiased.

However, if we establish a model only for y_2, ignoring the endogeneity of y_1, OLS will be still used in standard software packages to obtain estimates for β_0 and β_1, which gives us

$$\begin{aligned}\widehat{\beta_0} &= \frac{\sum y_1^2 \cdot \sum Xy_2 - \sum Xy_1 \cdot \sum y_1y_2}{\sum X^2 \cdot \sum y_1^2 - \sum Xy_1 \cdot \sum Xy_1},\\ \widehat{\beta_1} &= \frac{\sum X^2 \cdot \sum y_1y_2 - \sum Xy_1 \cdot \sum Xy_2}{\sum X^2 \cdot \sum y_1^2 - \sum Xy_1 \cdot \sum Xy_1}.\end{aligned} \tag{B.4}$$

Note that for simplicity, $\sum$ denotes a summation across all individual observations.

For $\hat{\beta_1}$, after re-arranging the denominator, we have

$$D = S_X^2 + S_{y_1}^2 + NS_X^2\bar{y_1}^2 + N\bar{X}^2S_{y_1}^2 - S_{Xy_1} - 2NS_{Xy_1}\bar{X}\bar{y_1}, \tag{B.5}$$

where D is short for denominator, N denotes the total number of observations, S_X^2 and $S_{y_1}^2$ denote the sample variances and S_{Xy_1} denotes sample covariance. In (B.5), the first four terms are positive and in our simulation study (see Appendix C), with the setting of $\alpha_0 < 0$, $\beta_0 > 0$ and the generated y_1 with $\bar{y_1} < 0$, X with $\bar{X} > 0$, we have $S_{Xy_1} < 0$. The dominant part in (B.5) is positive (N is large).

In the numerator, we substitute the neglected model of y_1 into (B.4) to obtain the size of bias when modelling y_2 alone. After re-arranging the numerator, we have

$$Nm_1 = \beta_1 \cdot D + \sum X^2 \sum y_1u_2, \tag{B.6}$$

where Nm_1 is short for the numerator of β_1 and $\sum y_1u_2 = S_{u_1u_2}$, where $S_{u_1u_2}$ denotes the sample covariance of random effects. Putting both the numerator and denominator together, we have

$$\hat{\beta_1} = \beta_1 + \frac{\sum X^2 \cdot S_{u_1u_2}}{D}. \tag{B.7}$$

Let ρ denote the correlation of random effects u_1 and u_2, we have the following conclusion about $\hat{\beta_1}$: if $\rho(u_1, u_2) > 0 (< 0)$, β_1 is overestimated (underestimated).

Turning to the estimation of β_0, the denominator remains unchanged (D) and after re-arranging the numerator of $\widehat{\beta_0}$, we have

$$Nm_0 = \beta_0 \cdot D - \sum Xy_1 \cdot S_{u_1u_2}, \tag{B.8}$$

which gives us

$$\hat{\beta_0} = \beta_0 - \frac{\sum X y_1 \cdot S_{u_1 u_2}}{D}. \tag{B.9}$$

As $\alpha_0 < 0$, plugging the equation for y_1 into (B.9) leads to $\sum X y_1 < 0$ (under the assumption that X and Z are independent and that X is independent of u_1). Therefore we draw the following conclusions about $\hat{\beta_0}$: if $\rho(u_1, u_2) > 0 (< 0)$, β_0 is overestimated (underestimated).

However, in addition to the above comments, the following issues should be noted:

1. If $\alpha_0 > 0$, the sign of the denominator can be difficult to tell.

2. Should X and Z be correlated, its correlation and the relative size of α_0 and α_1 may lead to $\sum X y_1$ taking different signs and therefore affect the conclusion on the direction of estimation bias.

3. In general, the existence of bias is confirmed but the direction of bias is not clear for the above reasons. Regardless of this, in the interest of the empirical application considered in this research, the current setting of α_s and β_s follows the direction of significant childhood effects on the hazard of partnership formation and dissolution using results from Section 5.7.1 and Section 5.7.2. Comments on the direction of bias based on the current setting are valid.

Appendix C

Simulation study to demonstrate bias in estimates due to endogenous predictor

This Appendix accompanies Chapter 5, Section 5.7.3. The following simulation study explores the direction of bias for a continuous outcome and a continuous endogenous predictor. We consider the following formulation:

$$y_1 = \alpha_0 X + \alpha_1 Z + u_1, \tag{C.1}$$

$$y_2 = \beta_0 X + \beta_1 y_1 + u_2, \tag{C.2}$$

where X is a binary predictor with a negative influence on y_1 ($\alpha_0 < 0$) and a positive influence on y_2 ($\beta_0 > 0$). We also set a positive relationship between y_1 and y_2 ($\beta_1 > 0$). u_1 and u_2 are random variables following a bivariate normal distribution which represent the individual-level unobserved characteristics for each equation. Z is introduced as an instrumental variable for model identification.To ensure it satisfies the requirements of an instrument, $\text{corr}(Z,u_1) = \text{corr}(Z,u_2) = 0$, $\text{corr}(y_1,Z) = +/-0.9$.

We set $\beta_0 > 0$, $\beta_1 > 0$, $\alpha_0 < 0$ and true parameter values in the simulation study are set towards the same direction of the estimated coefficients in Chapter 5, Section 5.7.3. 500 datasets of 500 observations are simulated. The correlation of u_1 and u_2 are set at high values in order to test the amount of bias in estimates if the endogeneity of y_1 is not account for. Results are summarised in Table C.1 with associated standard errors in brackets.

Table C.1 Simulation study of the direction of estimation bias in the presence of endogenous variables

Parameter	True value	Single model for y_2		Joint model for y_1, y_2	
		$\rho > 0$	$\rho < 0$	$\rho > 0$	$\rho < 0$
α_0	-3			-3.00 (0.07)	-3.00 (0.07)
α_1	2			2.00 (0.04)	2.00 (0.05)
β_0	2.5	3.04 (0.97)	1.96 (0.09)	2.50 (0.09)	2.50 (0.10)
β_1	3	3.18 (0.02)	2.82 (0.18)	3.00 (0.02)	3.00 (0.02)

$\rho = \text{corr}(u_1, u_2) = +/-0.9$

From the simulation study, it is obvious that the joint model takes care of the endogenous y_1 and the estimates are no longer biased with smaller standard errors than those estimated from the single model. Without joint modelling, when $\text{corr}(u_1, u_2) > 0$ ($\text{corr}(u_1, u_2) < 0$), the estimates are generally biased upwards (downwards).

Appendix D

Syntax for the general 3-step ML approach for event history outcomes

This Appendix accompanies Chapter 5, Section 5.7.3. We provide an example syntax to estimate the proposed model based on simulated datasets using the general 3-step approach.

```
[[init]] /*template files for .lgs files in steps 1 and 2*/
        iterators = rep;
        rep = 1:500;
        filename = "step12_scenario1_[[rep]].lgs";
        outputDirectory = "C:\2lcsim\scenario1";
[[/init]]
//LG5.1//        version = 5.1
        infile 'C:\2lcsim\scenario1\2lc_[[rep]].txt' quote = single
        model
                options
                        output parameters=first standarderrors
                                probmeans=posterior profile bivariateresiduals;
                        outfile 'scenario1_classification.sav' classification
                        /*output classification probabilities*/
                        keep y_tij t x z id;  /* keep original data*/
        variables
                dependent  /* Us are manifest variables in the latent class models*/
                        u1 nominal 2 coding=first,u2 nominal 2 coding=first,
                        u3 nominal 2 coding=first, u4 nominal 2 coding=first,
                        u5 nominal 2 coding=first, u6 nominal 2 coding=first,
                        u7 nominal 2 coding=first, u8 nominal 2 coding=first,
                        u9 nominal 2 coding=first, u10 nominal 2 coding=first;
                latent
                        C1 nominal 2 coding=2,
                        C2 nominal 2 coding=2;
        equations
                C1<-(a)1;C2<-(b)1;
                u1-u5<-1+(-)C1;    /*avoid label-switching problem*/
                u6-u10<-1+(-)C2;
                a~=0.7;b~=0.4;
        end model

[[init]] /*template files for .lgs files in step 3*/
        iterators = rep;
        rep = 1:500;
        filename = "step3_scenario1_[[rep]].lgs";
        outputDirectory = "C:\2lcsim\scenario1";
[[/init]]
//LG5.1// version = 5.1
        infile 'scenario1_classification.sav'
                /*data saved after steps 1 and 2*/
        model
        options
                output parameters=first standarderrors
                        ParameterCovariances probmeans=posterior
                        profile bivariateresiduals
                write='C:\2lcsim\scenario1\est_[[rep]].txt';
                        /*output estimated coefficients*/
                step3 ml modal; /*request 3-step procedure*/
                        /*use modal classes from the saved dataset*/
                keep y_tij t x z id;  /* keep original data*/
        variables
                groupid id; /*level-2 group specification*/
                independent
                        t numeric, x numeric, z numeric;
                dependent
                        y_tij nominal 2 coding=first;
                latent
                        GClass1 group nominal posterior=(C1#1 C1#2) coding=2,
                        GClass2 group nominal posterior=(C2#1 C2#2) coding=2,
                        GCFactor2 group continuous;
        equations
                        GCFactor2; /*continuous level-2 random effects term*/
                        GClass1<-1;GClass2<-1;
                        y_tij<- 1+t+x+z+GClass1+GClass2+(1)GCFactor2;
                        GClass1<->GClass2; /*allow for association*/
        end model
```

Appendix E

Joint modelling of the time to first partnership formation and partnership dissolution

This Appendix accompanies Chapter 5, Section 5.7.3 and provides the remaining part of the results are not reported in the main chapter. A joint model is fitted for the time to first partnership formation and the time to recurrent partnership dissolution, allowing for the influence of a common set of individual-level unobserved characteristics u_i. After factorisation, the random effects term for the formation process is $\lambda^{(F)}u_i$ while for the dissolution process, we have u_i, with the coefficient fixed at one.

Tables E.1 and E.2 shows the estimated coefficients and standard errors from fitting a joint model using modal class approach and the general 3-step approach.

Table E.1 Joint model Part I: covariate effects on the logit hazard of first partnership formation

Covariates	Modal class		General 3-step	
	Estimate	SE	Estimate	SE
Age (ref=16-19)				
19-22	1.56**	0.05	1.54**	0.05
22-25	2.05**	0.06	2.02**	0.05
25-28	1.96**	0.07	1.92**	0.06
28-31	1.80**	0.08	1.75**	0.07
31-34	1.54**	0.10	1.46**	0.08
34-37	0.88**	0.14	0.84**	0.12
37-40	0.67**	0.16	0.60**	0.14
40-43	0.21	0.20	0.17	0.18
43+	−0.29	0.18	−0.31**	0.15
Number of years in post-16 full-time education [a] (ref.=0)				
1	−0.19**	0.05	−0.18**	0.05
2	−0.20**	0.05	−0.20**	0.04
3-5	−0.38**	0.05	−0.37**	0.04
6+	−0.19**	0.05	−0.19**	0.05

**$p < 0.05$,*$p < 0.1$

[a] time-varying covariate

Table E.2 Joint model Part II: covariate effects on the logit hazard of partnership dissolution

Covariates	Modal class		General 3-step	
	Estimate	SE	Estimate	SE
Number of years in partnership (ref.=3)				
21[a]	0.41**	0.12	0.42**	0.12
24	0.51**	0.13	0.54**	0.13
Partnership type[b] (ref.=Marriage)				
Cohabitation	1.93**	0.06	1.96**	0.06
Number of previous partners[b] (ref.=0)				
1+	−0.53**	0.07	−0.48**	0.07
Age at the start of partnership (ref.=less than 20)				
20-25	−0.49**	0.09	−0.56**	0.09
25-30	−0.64**	0.14	−0.81**	0.16
30-35	−0.75**	0.18	−0.92**	0.19
35+	−0.84**	0.20	−1.07**	0.21
Number of years in post=16 full-time education [b] (ref.=0)				
1	−0.06	0.09	−0.02	0.09
2	−0.03	0.09	0.03	0.09
3-5	0.03	0.09	0.04	0.09
6+	−0.31**	0.10	−0.27**	0.10
Number of pre-school children[b](ref.=0)				
1	−0.64**	0.09	−0.61**	0.09
2	−0.59**	0.09	−0.56**	0.09
3+	−0.54**	0.12	−0.50**	0.12

**$p < 0.05$,*$p < 0.1$

[a] Only significant duration effects are reported

[b] Time-varying covariates

Appendix F

Input data structure and syntax for the SEM in Chapter 6

This Appendix accompanies Chapter 6. Following the description in Section 6.3.2, we provide an example of input data structure Table F.1 for a binary distal outcome. Instead of the commonly used stacked structure, we provide an alternative data structure which can be also convenient for outcomes of other types (e.g. continuous, ordinal) that require separate link functions from those used for the event history outcomes. Next, we show a LatentGOLD syntax to estimate the proposed model using the manual 3-step maximum likelihood approach for the application described in Section 6.4, with four latent categorical predictors for health.

The following variables are used in the example of input data structure (Table F.1). ID is an identifier for cohort member. y_{tij} is binary response variable derived in the discrete-time event history. H is a binary response variable indicting health condition at age 50. $X^{(S)}$ and $X^{(H)}$ are corresponding continuous covariates in the event history submodel and the health submodel.

Table F.1 An example input data structure under the discrete-time setting

ID	t	y_{tij}	H	$X^{(S)}$	$X^{(H)}$
1	1	0	1	1.5	0.5
1	2	0	0	1.5	0
1	3	1	0	1.5	0
1	1	0	0	1.5	0
1	2	0	0	1.5	0
2	1	0	0	0.8	-0.2
2	2	1	0	0.8	0
3	1	0	1	-1.5	1.5
3	2	0	0	-1.5	0
3	3	0	0	-1.5	0

The variables listed in Table F.2 are used in the syntax to estimate an SEM for the empirical analysis where four level-2 latent categorical variables and summaries of the partnership history predict a binary health outcome.

Table F.2 Description of variables in the syntax for empirical analysis

Variables	Description
cmid	Individual identifier
$m1-m4$	Modal classes derived in Step 1 (level-2)
y	Binomial outcome with exposure s_{nt} (level-1)
gnhlth50b (H)	Binary health outcome at level-2
s1	Binary indicator for the health submodel
$s1^*$	All predictors for the health situation
s2	Binary indicator for the parnterhip formation model
$s2^*$	All predictors for the tendency of partnership formation
s3	Binay indicator for the partnership dissolution submodel
$s3^*$	All predictors for the risk of partnership dissolution
GCFactor2	Level-2 latent continuous variable
$l1-l4$	Latent categorical variables

```
//LG5.1//
version = 5.1
infile ’C:\lgformdisRQ3_ses.txt’ delim = 0x09 quote = single
```

```
model
options
algorithm
        tolerance=1e-008 emtolerance=0.01 emiterations=250 nriterations=50;
bayes
        categorical=2 variances=2 latent=2;
startvalues
        seed=0 sets=10 tolerance=1e-005 iterations=50;
quadrature
        nodes=10;

missing
        includeall;
output
      parameters=first standarderrors ParameterCovariances
        write='C:\results.txt'; /*ouput estimates and standard errors*/
variables
        groupid cmid;

        dependent

        m1 nominal 3, m2 nominal 2, m3 nominal 3,m4 nominal 2,
        y binomial exposure=s_nt, gnhlth50b nominal 2;

        independent
        s1 nominal coding=first,
        s1_ovwt16 numeric, s1_totptn0 nominal coding=first,s1_totptn2 nominal coding=first,
        s1_totptn3 nominal coding=first,s1_age1stctr numeric,
        s1_timesingle numeric,

        s2 nominal coding=first,
        s2_pwt2 nominal coding=first, s2_pwt3 nominal coding=first, s2_pwt4 nominal coding=first,
         s2_pwt5 nominal coding=first, s2_pwt6 nominal coding=first, s2_pwt7 nominal coding=first,
        s2_pwt8 nominal coding=first, s2_pwt9 nominal coding=first, s2_pwt10 nominal coding=first,

        s2_edu2 nominal coding=first, s2_edu3 nominal coding=first, s2_edu4 nominal coding=first,
        s2_edu5 nominal coding=first, s3 nominal coding=first,
        s3_pwt2 nominal coding=first, s3_pwt3 nominal coding=first, s3_pwt4 nominal coding=first,
        s3_pwt5 nominal coding=first, s3_pwt6 nominal coding=first, s3_pwt7 nominal coding=first,
        s3_pwt8 nominal coding=first, s3_pwt9 nominal coding=first, s3_pwt10 nominal coding=first,

        s3_ptype2 nominal coding=first, s3_edu2 nominal coding=first,
        s3_edu3 nominal coding=first, s3_edu4 nominal coding=first, s3_edu5 nominal coding=first,
        s3_pre2 nominal coding=first, s3_pre3 nominal coding=first, s3_pre4 nominal coding=first,
        s3_NPTN2 nominal coding=first, s3_agrl2 nominal coding=first,
        s3_agrl3 nominal coding=first, s3_agrl4 nominal coding=first, s3_agrl5 nominal coding=firs

latent
        l1 nominal 3 coding=3, l2 nominal 2 coding=2,
        l3 nominal 3 coding=3, l4 nominal 2 coding=2, GCFactor2 group continuous;
equations
        l1<-1;
        l2<-1;
        l3<-1;
        l4<-1;

        m1<- (C~wei) 1| l1;
        m2<- (D~wei) 1| l2;
        m3<- (E~wei) 1| l3;
        m4<- (F~wei) 1| l4;

    /*freely estimate the variance of random effects*/
        GCFactor2;

    /*(1) means coefficient is fixed at 1*/
        y<-s2+s2_pwt2+s2_pwt3+s2_pwt4+s2_pwt5+s2_pwt6+s2_pwt7
                +s2_pwt8+s2_pwt9+s2_pwt10
                +s2*l1+s2*l2+s2*l3+s2*l4+s2*GCFactor2
                 +s2_edu2+s2_edu3+s2_edu4+s2_edu5
                +s3+s3_pwt2+s3_pwt3+s3_pwt4+s3_pwt5+s3_pwt6+s3_pwt7+s3_pwt8+s3_pwt9+s3_pwt10
                +s3*l1+s3*l2+s3*l3+s3*l4+ s3_edu2+s3_edu3+s3_edu4+s3_edu5
                +s3_pre2+s3_pre3+s3_pre4+s3_NPTN2+s3_ptype2
                +s3_agrl2+s3_agrl3+s3_agrl4+s3_agrl5+(1)s3*GCFactor2;

         gnhlth50b<-s1+s1_ovwt16+s1_totptn0+s1_totptn2+s1_totptn3+s1_timesingle+s1_age1stctr
                +s1*l1+s1*l2+s1*l3+s1*l4+s1*GCFactor2;

        /*allow for association between latent categorical variables*/
        l1<->l2;l1<->l3;l1<->l4;
        l2<->l3;l2<->l4;
        l3<->l4;

        /*misclassification error computed in Step 2,*/
    /*after estimating separate latent class models in Step 1*/
C={0.824 0.169 0.007
   0.061 0.910 0.028
   0.008 0.093 0.899};
D={0.734 0.266
```

```
        0.032 0.968};
E={0.854 0.097 0.049
   0.055 0.943 0.002
   0.053 0.003 0.944};
F={0.837 0.163
   0.003 0.997};
end model
```

Appendix G

Input data structure and syntax for the SEM in Chapter 7

This Appendix accompanies Section 7.5.3 of Chapter 7. We first provide the data structure (Section G.1) for estimation and then the LatentGOLD syntax.

G.1 The input data structure

The data are restructured such that partnership formation and dissolution outcomes, midlife health outcome and time-to-dropout outcomes are stacked. For the partnership transitions, each row contains a record for a 6-month interval of a particular episode (including individual ID, a set of covariates and the event and censoring indicators). For the health outcome, each row is assigned to an individual, with records of the health outcome, summaries of partnership experiences and other individual-specific characteristics. For the missing data outcomes, each individual has four rows where each row records the information at each of the four adult waves, including the wave indicator and a set of demographic and early-life variables. The outcomes are recorded as discrete time-to-dropout indicators.

An example of the expanded and stacked input data structure is illustrated in Table G.1. To connect the covariates with their respective outcomes in the estimation, a binary indicator (s) is created for each submodel. Interactions of s variables with each of the covariates are included in the analysis model. ID is an identifier for cohort member. y is stacked binary response vector that contains time-to-event outcomes (partnership events, dropout events) and the midlife health. Note that for data stored in long form, a dropout vector becomes $D_i = (0,0,1)$. t is the indicator for 6-month intervals in models for partnership transitions and the indicator for waves in the dropout submodel. $X^{(D)}$, $X^{(H)}$, $X^{(F)}$ and $X^{(S)}$ are corresponding covariates in each submodel. We define that $s1$=I(health submodel), $s2$=I(partnership formation submodel), $s3$=I(partnership dissolution submodel) and $s0$=I(dropout submodel). Note that for simplicity, our example only considers time-invariant covariates. For models of partnership transitions and the dropout probability, time-varying variables are also used in our analysis.

Table G.1 An example of the stacked input data structure.

ID	t	y	$X^{(F)}$	$X^{(S)}$	$X^{(H)}$	$X^{(D)}$	s0	s1	s2	s3
1	1	0	1.2	0.5	0.6	1.5	1	0	0	0
1	2	0	1.2	0.5	0.6	1.5	1	0	0	0
1	3	0	1.2	0.5	0.6	1.5	1	0	0	0
1	4	1	1.2	0.5	0.6	1.5	1	0	0	0
1	1	1	1.2	0.5	0.6	1.5	0	1	0	0
1	1	0	1.2	0.5	0.6	1.5	0	0	1	0
1	2	0	1.2	0.5	0.6	1.5	0	0	1	0
1	3	1	1.2	0.5	0.6	1.5	0	0	1	0
1	1	0	1.2	0.5	0.6	1.5	0	0	0	1
1	2	0	1.2	0.5	0.6	1.5	0	0	0	1
1	3	0	1.2	0.5	0.6	1.5	0	0	0	1
1	4	0	1.2	0.5	0.6	1.5	0	0	0	1
1	5	1	1.2	0.5	0.6	1.5	0	0	0	1

G.2 LatentGOLD syntax for the application

The variables listed in Table G.2 are used in the syntax to estimate an multilevel SEM for the empirical analysis where four level-2 latent categorical variables and summaries of the partnership history predict a binary health outcome.

Table G.2 Description of variables in the syntax for empirical analysis

Variables	Description
cmid	Individual identifier
$m1-m4$	Modal classes derived in Step 1 (level-2)
s_{nt}	Exposure in months. For the health and dropout submodels, it is set to 1.
y	Stacked binomial and binary outcome with exposure s_{nt}
$s0$	Binary indicator for the dropout submodel
$s0^{*}$	All predictors for the missing data mechanism
$s1$	Binary indicator for the midlife health submodel
$s1^{*}$	All predictors for the midlife health
$s2$	Binary indicator for the parnterhip formation submodel
$s2^{*}$	All predictors for the tendency of partnership formation
$s3$	Binay indicator for the partnership dissolution submodel
$s3^{*}$	All predictors for the risk of partnership dissolution
GCFactor2	Level-2 latent continuous variable (random effect)
$l1-l4$	Latent categorical variables

As the models proposed in the chapter are developed step by step (i.e. nested), we present the syntax for the most complex Model 2 with a dropout submodel.

```
//LG5.1//
version = 5.1

/*Load data into LG*/
infile 'C:\Users\zhuy18\drop3.txt' delim = 0x09 quote = single


model
options
/*Set the number of PC cores for analysis*/
maxthreads=7;
algorithm
tolerance=1e-008 emtolerance=0.01 emiterations=1000 nriterations=1000;
bayes
categorical=2 variances=2 latent=2;
/*10 random sets of starting values for each parameter are used*/
startvalues
seed=0 sets=10 tolerance=1e-005 iterations=50;
quadrature
nodes=10;
```

```
/*Do not exclude records with missing outcome values*/
missing
includedependent;
output
parameters=first standarderrors iterationdetails;

variables
/*Set level-2 identifier*/
groupid cmid;
/*Outcomes*/
dependent
m1 nominal 3, m2 nominal 2, m3 nominal 3,m4 nominal 2,
y binomial exposure=s_nt;
/*Covariate set */
/*After interacting ’s’ variables with each covariate*/
independent
/*Dropout submodel*/
s0 nominal coding=first,
s0_male nominal coding=first,s0_mumsmoke nominal coding=first,s0_mumage numeric,
s0_logbmi16 numeric, s0_rutt16 nominal coding=first,s0_read16 numeric,
s0_math16 numeric,
s0_time2 nominal coding=first, s0_time3 nominal coding=first, s0_time4 nominal coding=first,
s0_timesingle numeric, s0_logage1stctr numeric,
/*Health submodel*/
s1 nominal coding=first,
s1_ovwt16 nominal coding=first, s1_totptn0 nominal coding=first,
s1_totptn2 nominal coding=first,
s1_totptn3 nominal coding=first,s1_timesingle numeric,s1_logage1stctr numeric,
s1_m11 nominal coding=first, s1_m12 nominal coding=first, s1_m21 nominal coding=first,
s1_m31 nominal coding=first, s1_m32 nominal coding=first, s1_m41 nominal coding=first,
/*Partnership formation submodel*/
s2 nominal coding=first,
s2_pwt2 nominal coding=first, s2_pwt3 nominal coding=first, s2_pwt4 nominal coding=first,
s2_pwt5 nominal coding=first, s2_pwt6 nominal coding=first, s2_pwt7 nominal coding=first,
s2_pwt8 nominal coding=first, s2_pwt9 nominal coding=first, s2_pwt10 nominal coding=first,
s2_edu2 nominal coding=first, s2_edu3 nominal coding=first, s2_edu4 nominal coding=first,
s2_edu5 nominal coding=first,
/*Partnership dissolution submodel*/
s3 nominal coding=first,
s3_pwt2 nominal coding=first, s3_pwt3 nominal coding=first, s3_pwt4 nominal coding=first,
s3_pwt5 nominal coding=first, s3_pwt6 nominal coding=first, s3_pwt7 nominal coding=first,
s3_pwt8 nominal coding=first, s3_pwt9 nominal coding=first, s3_pwt10 nominal coding=first,
s3_ptype2 nominal coding=first, s3_edu2 nominal coding=first,
s3_edu3 nominal coding=first, s3_edu4 nominal coding=first, s3_edu5 nominal coding=first,
s3_pre2 nominal coding=first, s3_pre3 nominal coding=first, s3_pre4 nominal coding=first,
s3_NPTN2 nominal coding=first, s3_agrl2 nominal coding=first,
s3_agrl3 nominal coding=first, s3_agrl4 nominal coding=first, s3_agrl5 nominal coding=first;
/*Define latent variables*/
latent
l1 nominal 3 coding=3, l2 nominal 2 coding=2,
l3 nominal 3 coding=3, l4 nominal 2 coding=2,
GCFactor2 group continuous;

/*Specifications of Model 3*/
equations
/*Freely estimate the variance of random effects*/
GCFactor2;
l1<-1;
l2<-1;
l3<-1;
l4<-1;
m1<- (C~wei) 1| l1;
m2<- (D~wei) 1| l2;
m3<- (E~wei) 1| l3;
m4<- (F~wei) 1| l4;
/*(1) means coefficient is fixed at 1*/
y<-s2+s2_pwt2+s2_pwt3+s2_pwt4+s2_pwt5+s2_pwt6+s2_pwt7
+s2_pwt8+s2_pwt9+s2_pwt10+s2_edu2+s2_edu3+s2_edu4+s2_edu5
+s2*l1+s2*l2+s2*l3+s2*l4+s2*GCFactor2
+s3+s3_pwt2+s3_pwt3+s3_pwt4+s3_pwt5+s3_pwt6+s3_pwt7+s3_pwt8+s3_pwt9+s3_pwt10
+s3_edu2+s3_edu3+s3_edu4+s3_edu5
+s3_pre2+s3_pre3+s3_pre4+s3_NPTN2+s3_ptype2
+s3_agrl2+s3_agrl3+s3_agrl4+s3_agrl5
+s3*l1+s3*l2+s3*l3+s3*l4+(1)s3*GCFactor2
+s1+s1_ovwt16+s1_totptn0+s1_totptn2+s1_totptn3+s1_timesingle+s1_logage1stctr
+s1*l1+s1*l2+s1*l3+s1*l4+s1*GCFactor2
+s0+s0_time2+s0_time3+s0_time4
+s0_male+s0_mumsmoke+s0_mumage
+s0_logbmi16+s0_rutt16+s0_read16+s0_math16
+s0*l1+s0*l2+s0*l3+s0*l4
+s0*GCFactor2;
/*Allow for association between latent categorical variables*/
l1<->l2;l1<->l3;l1<->l4;
l2<->l3;l2<->l4;
l3<->l4;
/*misclassification error computed in Step 2,*/
/*after estimating separate latent class models in Step 1*/
C={0.824 0.169 0.007
0.061 0.910 0.028
0.008 0.093 0.899};
D={0.734 0.266 0.032 0.968};
E={0.854 0.097 0.049
0.055 0.943 0.002
0.053 0.003 0.944};
F={0.837 0.163
0.003 0.997};
end model
```

Outputs from this analysis are tedious but we report the iteration details below.

```
Iteration Details
1  -1733861.3059239858   -1733833.6088785846          250688
2  -1733870.3828115778   -1733842.6938904596         1633820
3  -1733846.6588695655   -1733818.9681174560         1304151
8  -1733856.4200588253   -1733828.7269226671          918135
9  -1733857.2397525713   -1733829.5368320588         1261654
10  -1733858.1206868228   -1733830.4246911164          391714
4  -1733855.9038073951   -1733828.2045238742          683639
5  -1733880.0925004676   -1733852.4023928693          601210
6  -1733872.1666846022   -1733844.4683980744         2448701
7  -1733863.1902829485   -1733835.4889357542         2677084
1  -1733814.2679846210   -1733786.4374404261         1304151
Iteration           log-posterior        log-likelihood       criterion
0  -1733814.2679846212   -1733786.4374404263
EM         5  -1733814.0810119943   -1733786.2469908863      0.04320942
EM        10  -1733813.9289523473   -1733786.0919409944      0.03998211
EM        15  -1733813.7948007477   -1733785.9549561716      0.03712827
EM        20  -1733813.6761829760   -1733785.8336595139      0.03456968
EM        25  -1733813.5710834935   -1733785.7260276929      0.03225710
EM        30  -1733813.4777844076   -1733785.6303339188      0.03015222
EM        35  -1733813.3948142016   -1733785.5450979599      0.02822586
EM        40  -1733813.3209074736   -1733785.4690462337      0.02645513
EM        45  -1733813.2549724071   -1733785.4010793380      0.02482154
EM        50  -1733813.1960640168   -1733785.3402452748      0.02330988
EM        55  -1733813.1433620192   -1733785.2857172769      0.02190764
EM        60  -1733813.0961524190   -1733785.2367753480      0.02060392
EM        65  -1733813.0538121504   -1733785.1927908615      0.01938946
EM        70  -1733813.0157961552   -1733785.1532135988      0.01825625
EM        75  -1733812.9816265462   -1733785.1175608775      0.01719739
EM        80  -1733812.9508834090   -1733785.0854083211      0.01620679
EM        85  -1733812.9231970177   -1733785.0563820477      0.01527886
EM        90  -1733812.8982412247   -1733785.0301520329      0.01440881
EM        92  -1733812.8910613842   -1733785.0226121563      0.00999651
```

```
Newton     1  -1733812.6568408129   -1733784.7645299169      0.97381569
Newton     2  -1733812.6550888985   -1733784.7603087286      0.01657667
Newton     3  -1733812.6550887532   -1733784.7602860134      0.00010411
Newton     4  -1733812.6550887527   -1733784.7602860110      0.00000001
Standard errors
Preparing output

Wall clock time=623.423123  CPU=183.918939
```