Reduced-bias estimation of some non-standard models

Asma Saleh

Principal Supervisor: Dr Ioannis Kosmidis
Subsidiary Supervisor: Professor Richard Chandler

A dissertation submitted for the degree of
Doctor of Philosophy

December 2020
Department of Statistical Science
University College London

To Mum and Dad

Acknowledgements

Words fail to express the level of gratitude and appreciation I have for my supervisor Dr Ioannis Kosmidis. Pursuing a PhD is not easy, I have experienced various emotions and events throughout this journey; some posed great, intimidating challenges and some were blissful. My supervisor made sure I overcame my fears and obstacles both personal and academic, and amplified my successful, joyful times. He taught me the importance and rewards of determination and perseverance. He provided emotional and psychological support when I doubted myself and questioned my capabilities as a researcher. He reassured me that I was stronger than I thought I was. He prioritised my well-being and happiness and could get me out of a hindering state of mind with ease, grace and patience. I would not have come this far academically if it wasn't for his sincere, constant care and aid. I can't thank you enough for helping me turn this dream into a reality. Without you I would not be able to say that I am proud of the person I am today.

I would like to thank UCL and EPSRC for the financial support during my PhD studies, and for the academics, especially Professor Richard Chandler for kindly serving as my subsidiary supervisor and Professor Christian Hennig for serving, jointly with Professor Richard Chandler, as the upgrade examiners, administrative staff and fellow PhD students for creating an everlasting supportive environment. I am also grateful to Dr Paul Northrop for reassuring me that this PhD is possible through hard work and diligence.

An ocean of thanks to my parents, sisters, brother and friends for a lifetime of love and support and for enduring my craziness with a smile. Finally, my most dearest and heart-felt thanks goes to my mother who was always ready to embrace me in times of need and give me the boost of energy and confidence I needed. She always reminded me of how proud she was of me for pursuing this PhD and gave me consistent words of encouragement. She has played the biggest role in pushing me towards my goals and dreams and tried her best to not let me break down and give up on this journey. I am perpetually grateful and appreciative of you. I love you, forever.

Declaration

I, Asma Saleh confirm that the work presented in this thesis is my own research. Where information has been derived from other sources, I confirm that this has been explicitly indicated in the thesis by references. The current thesis has not been submitted for examination at any other university than the University of London.

Abstract

There is a persisting interest in methods that reduce bias in the estimation of parametric models. There is already a wide range of methods that achieve that goal, with a few of them also delivering beneficial side effects. For example, the bias-reducing adjusted scores approach of Firth (1993) has been shown to always deliver finite estimates in models like logistic regression even when the maximum likelihood (ML) estimator takes infinite values. Other proposals (e.g. reduced-bias M estimation in Kosmidis and Lunardon (2020), and indirect inference of Kuk (1995)) have been shown to be able to reduce estimation bias even in cases where the model is partially specified, such as for general M-estimators. In this thesis, we examine the applicability, evaluate the performance and compare a range of bias reduction methods such as the bias-reducing adjusted score equations of Firth (1993), indirect inference and reduced-bias M estimation, in terms of their impact on estimation and inference, in well-used model classes in econometrics and statistics, which are beyond the various standard models that bias reduction methods have been used for before. In particular, we study the Heckit regression model which handles non-randomly selected samples where the observed range of the dependent variable is censored, i.e. it is only partially known whether it is above or below a fixed threshold. We also examine accelerated failure time models which are parametric survival models for censored lifetime observations. Finally, we consider two stratified models (see, Sartori, 2003) where interest lies in the estimation of a parameter in the presence of a set of nuisance parameters, whose dimension increases with the number of strata. The main challenge with these models is that even basic requirements, like consistency of the ML estimator, are not necessarily satisfied (see, Neyman and Scott, 1948). We focus on binomial matched pairs where the ML estimate of the parameter of interest may be infinite due to data separation. We propose a penalised version of the log-likelihood function based on adjusted responses which always results in a finite estimator of the log odds ratio. The probability limit of the penalised adjusted log-likelihood estimator is derived and it

is shown that in certain settings the ML, conditional and modified profile log-likelihood estimators drop out as special cases of the former estimator. It is found that for the models of censored data, Firth adjustments are not available in closed form whereas indirect inference and reduced-bias M estimation are applicable and are an improvement over traditional ML estimation.

Impact Statement

In this thesis we explore the problem of reducing the bias in the estimation of some non-standard models such as the Heckit model for non-randomly selected samples and the Weibull accelerated failure time model for censored lifetimes. The bias of the maximum likelihood estimator may be considerable for small samples and consequently impact the performance of inferences. The popular approach of bias-reducing adjusted scores of Firth (1993) is not always applicable and it involves cumbersome computations, while the method of reduced-bias M estimation of Kosmidis and Lunardon (2020) is simpler to implement and has wider applicability.

More specifically, we found that while the method of Firth (1993) is difficult to apply in the Heckit and Weibull accelerated failure time models, the reduced-bias M estimation method was easily implemented and effective in reducing the small sample bias of the ML estimator. This opens the door to further research on general Tobit and accelerated failure time models where the method of Kosmidis and Lunardon (2020) is applicable and has the potential of significant bias reduction.

The Heckit (Tobit II) model and Tobit models in general are broadly used in many fields for modelling censored and sample selected data. For instance, in economics Tobit models are used to analyse the relationship between household expenditure on durable goods and household incomes. In political sciences Tobit models can be used to investigate the role of various factors in European Union project selection fund receipt. Social scientists have also used these models to research on the effectiveness of attending one kind of school rather than another. On the other hand, accelerated failure time models are frequently encountered when modelling failure time data. For example, in biomedical and clinical studies these models are used to assess the efficacy and safety of a new chemotherapy in treatment of some advanced cancer. In engineering and reliability studies, accelerated failure time models are used to obtain information on the endurance of machine components subjected to life tests.

Other non-standard models that we consider are ones for stratified observations such as the binomial matched pairs model used in matched case-control studies, which are popular in biostatistics and epidemiology. The estimation of a parameter of interest in the presence of high-dimensional nuisance parameters is generally challenging because even the standard ML estimator is biased and inconsistent. Even though the methods of Firth (1993) and Kosmidis and Lunardon (2020) are applicable to the binomial matched pairs model, the former may be harder to apply in other stratified models. There is much scope to extend the framework of reduced-bias M estimation to stratified settings, and therefore potentially yield estimates with smaller bias.

Contents

Acknowledgements **i**

Declaration **ii**

Abstract **iii**

Impact Statement **v**

List of Tables **xvii**

List of Figures **xviii**

1 Introduction **1**

1.1 Bias reduction methods 1

1.2 Stratified models 4

1.3 Outline 6

2 Likelihood modifications and bias reduction methods **8**

2.1 Modifications of the likelihood function 8

2.1.1 Likelihood and related quantities 8

2.1.2 Exact conditional likelihood 9

2.1.3 Profile likelihood 10

2.1.4 Approximate conditional profile likelihood 10

2.1.5 Modified profile likelihood 11

2.1.6 Adjusted profile likelihood 15

2.2 Bias reduction methods 16

2.2.1 Preamble 16

2.2.2 Asymptotic bias correction 17

2.2.3 Indirect inference . . . 18
2.2.4 Firth's bias-reducing adjusted score equations . . . 21
2.2.5 Empirical bias-reducing adjusted estimating functions . . . 22

3 Heckit (Tobit II) selection model **26**
3.1 Introduction . . . 26
3.2 Description of the Heckit model . . . 29
3.3 Review of point estimation of the model parameters . . . 32
3.3.1 Maximum likelihood estimator . . . 32
3.3.2 Heckman's two step (Heckit) estimator . . . 34
3.4 Implicit bias reduction methods . . . 36
3.4.1 Firth's adjusted score equations method . . . 36
3.4.2 Empirical bias-reducing penalty . . . 43
3.4.3 Indirect inference . . . 44
3.5 Simulation study . . . 45
3.6 Analysis of female labor supply data . . . 56
3.7 Discussion and further work . . . 61
3.7.1 Summary . . . 61
3.7.2 Empirical bias-reducing penalty for Heckman two-step estimation 61
3.7.3 Bias reduction for Tobit V model . . . 67

4 Accelerated failure time model **69**
4.1 Introduction . . . 69
4.1.1 The hazard and survival functions . . . 71
4.1.2 Parametric models of the hazard function . . . 73
4.1.3 Censoring and the likelihood function . . . 76
4.2 Description of the model and maximum likelihood estimation . . . 78
4.3 Reduced-bias estimation methods . . . 80
4.3.1 Firth's adjusted score equations method . . . 80
4.3.2 Empirical bias-reducing penalty . . . 82
4.3.3 Indirect inference . . . 83
4.4 Simulation study . . . 84
4.5 Analysis of lung cancer survival data . . . 95
4.6 Discussion and further work . . . 98
4.6.1 Summary . . . 98

4.6.2 Bias reduction for frailty models . . . 98
4.6.3 Empirical bias-reducing penalty for a general accelerated failure time model . . . 99

5 Stratified models **102**
5.1 Introduction . . . 102
5.2 Matched gamma pairs model . . . 104
5.2.1 Maximum likelihood estimation . . . 104
5.2.2 Modified likelihood functions . . . 105
5.2.3 Reduced bias estimation of the parameter of interest . . . 109
5.2.4 Simulation study . . . 118
5.3 Binomial matched pairs model . . . 127
5.3.1 Review of point estimation of the log odds ratio . . . 127
5.3.1.1 Maximum likelihood . . . 127
5.3.1.2 Conditional maximum likelihood . . . 129
5.3.1.3 Modified profile maximum likelihood . . . 130
5.3.1.4 Firth's adjusted score equations method . . . 131
5.3.2 Binary matched pairs model . . . 132
5.3.3 Penalised likelihood based on adjusted responses method . . . 133
5.3.3.1 Probability limit of the penalised likelihood estimator based on adjusted responses . . . 134
5.3.4 Indirect inference estimation of the log odds ratio . . . 137
5.3.5 Complete enumeration study . . . 139
5.3.6 Analysis of crying babies data . . . 155
5.4 Discussion and further work . . . 160

6 Discussion and further work **162**
6.1 Summary of the thesis . . . 162
6.2 Further work . . . 164

Appendix A Algebraic derivations for the Heckit model **166**

Appendix B Algebraic derivations for the matched gamma pairs model **172**

List of Tables

3.1 Heckit observations with true correlation coefficient $\rho = 0.2$ and true variance $\sigma^2 = 1$. The true intercepts are $\beta_1 = 0.01$ and $\gamma_1 = 0.03$, while the true slopes are $\beta_2 = 0.7$ and $\gamma_2 = 0.9$. Second to sixth columns show the estimated bias and Monte Carlo simulation error (in parentheses) of maximum likelihood, Heckman two-step (Heckit), indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$, $n = 150$ and $n = 300$, with all entries multiplied by 10. 52

3.2 Heckit observations with true correlation coefficient $\rho = 0.2$ and true variance $\sigma^2 = 1$. The true intercepts are $\beta_1 = 0.01$ and $\gamma_1 = 0.03$, while the true slopes are $\beta_2 = 0.7$ and $\gamma_2 = 0.9$. Second to sixth columns show the average estimated standard error and empirical standard error (in parentheses) of maximum likelihood, Heckman two-step (Heckit), indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$, $n = 150$ and $n = 300$, with all entries multiplied by 10. The symbol __ indicates that the estimate is not available. 53

3.3 Heckit observations with true correlation coefficient $\rho = 0.2$ and true variance $\sigma^2 = 1$. The true intercepts are $\beta_1 = 0.01$ and $\gamma_1 = 0.03$, while the true slopes are $\beta_2 = 0.7$ and $\gamma_2 = 0.9$. Second to sixth columns show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 95% derived from Wald-type confidence intervals using the maximum likelihood, Heckman two-step, indirect inference, iRBM and eRBM estimates, for sample sizes $n = 50$, $n = 100$, $n = 150$ and $n = 300$ with all coverage probabilities multiplied by 100. . . 54

3.4 Heckit observations with true correlation coefficient $\rho = 0.2$ and true variance $\sigma^2 = 1$. The true intercepts are $\beta_1 = 0.01$ and $\gamma_1 = 0.03$, while the true slopes are $\beta_2 = 0.7$ and $\gamma_2 = 0.9$. Second to sixth columns show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 99% derived from Wald-type confidence intervals using the maximum likelihood, Heckman two-step, indirect inference, iRBM and eRBM estimates, for sample sizes $n = 50$, $n = 100$, $n = 150$ and $n = 300$ with all coverage probabilities multiplied by 100. . . 55
3.5 Labor Force Participation Equation for Married Women. Estimated regression coefficients and standard errors (in parentheses). 60
3.6 Wage Offer Equation for Married Women. Estimated regression coefficients and standard errors (in parentheses). 60

4.1 Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the estimated bias and Monte Carlo simulation standard error (in parentheses) of maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all entries multiplied by 10 and with overall censoring percentage of approximately 24%. 87
4.2 Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the estimated bias and Monte Carlo simulation standard error (in parentheses) of maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all entries multiplied by 10 and with overall censoring percentage of approximately 45%. 88
4.3 Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the average estimated standard error and empirical standard error (in parentheses) of maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all entries multiplied by 10 and with overall censoring percentage of approximately 24%. 89

4.4 Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the average estimated standard error and empirical standard error (in parentheses) of maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all entries multiplied by 10 and with overall censoring percentage of approximately 45%. 90
4.5 Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 95% derived from Wald-type confidence intervals using the maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all coverage probabilities multiplied by 100 and with overall censoring percentage of approximately 24%. 91
4.6 Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 95% derived from Wald-type confidence intervals using the maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all coverage probabilities multiplied by 100 and with overall censoring percentage of approximately 45%. 92
4.7 Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 99% derived from Wald-type confidence intervals using the maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all coverage probabilities multiplied by 100 and with overall censoring percentage of approximately 24%. 93

4.8 Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 99% derived from Wald-type confidence intervals using the maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all coverage probabilities multiplied by 100 and with overall censoring percentage of approximately 45%. 94
4.9 Fitted full Weibull accelerated failure time model to the lung cancer data using maximum likelihood (ML), indirect inference (II), iRBM and eRBM. 97
4.10 Fitted reduced Weibull accelerated failure time model to the lung cancer data using maximum likelihood (ML), indirect inference (II), iRBM and eRBM. 97

5.1 Matched gamma pairs. Estimators of ψ and their theoretical bias and variance functions. 117
5.2 Matched gamma pairs. Inference about common square root of product of means in q pairs of gamma observations with shape m. Numerical values of the theoretical biases of the estimators of ψ for various values of m and q, with all entries multiplied by 10. The parameter of interest is $\psi = 1$. The empty cells correspond to the fact that the bias of all estimators except $\tilde{\psi}_{ad}^{(E)}$ is independent of q. 120
5.3 Matched gamma pairs. Inference about common square root of product of means in q pairs of gamma observations with shape m. Numerical values of the theoretical variance of the estimators of ψ for various values of m and q, with all entries multiplied by 10. The parameter of interest is $\psi = 1$. 121
5.4 Matched gamma pairs. Inference about common square root of product of means in q pairs of gamma observations with shape m. Numerical values of the theoretical mean squared error of the estimators of ψ for various values of m and q, with all entries multiplied by 10. The parameter of interest is $\psi = 1$. 122

5.5 Matched gamma pairs. Inference about common square root of product of means in q pairs of gamma observations with shape m. Columns three to eight show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional profile and adjusted profile log-likelihood ratios, with all coverage probabilities multiplied by 100. The parameter of interest is $\psi = 1$. 123
5.6 Probability limits of profile, conditional, modified profile and adjusted log-likelihood based on adjusted responses estimators of the log odds ratio ψ in the binary matched pairs model when $m = 1$. For each ψ, the value in bold face corresponds to the limiting value closest to the truth if we ignore $\hat{\psi}_c$. 135
5.7 Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the bias and variance (in parentheses) of estimators derived from profile, conditional, modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods, with all entries multiplied by 10; seventh column show the bias and variance (in parentheses) of the indirect inference $\hat{\psi}_{a*}$ using the conditional model, with all entries multiplied by 10. 142
5.8 Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the bias and variance (in parentheses) of estimators derived from profile, conditional, modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods, with all entries multiplied by 10; seventh column show the bias and variance (in parentheses) of the indirect inference $\hat{\psi}_{a*}$ using the conditional model, with all entries multiplied by 10. 143
5.9 Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the bias and variance (in parentheses) of estimators derived from profile, conditional, modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods, with all entries multiplied by 10; seventh column show the bias and variance (in parentheses) of the indirect inference $\hat{\psi}_{a*}$ using the conditional model, with all entries multiplied by 10. 144

5.10 Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the bias and variance (in parentheses) of estimators derived from profile, conditional, modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods, with all entries multiplied by 10; seventh column show the bias and variance (in parentheses) of the indirect inference $\hat{\psi}_{a*}$ using the conditional model, with all entries multiplied by 10. 145
5.11 Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the bias and variance (in parentheses) of estimators derived from profile, conditional, modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods, with all entries multiplied by 10; seventh column show the bias and variance (in parentheses) of the indirect inference $\hat{\psi}_{a*}$ using the conditional model, with all entries multiplied by 10. 146
5.12 Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the bias and variance (in parentheses) of estimators derived from profile, conditional, modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods, with all entries multiplied by 10; seventh column show the bias and variance (in parentheses) of the indirect inference $\hat{\psi}_{a*}$ using the conditional model, with all entries multiplied by 10. 147
5.13 Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the coverage probability and average length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional, modified profile, penalized (Firth) and penalised based on adjusted responses log-likelihood ratios, with all coverage probabilities multiplied by 100. 148
5.14 Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the coverage probability and average length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional, modified profile, penalized (Firth) and penalised based on adjusted responses log-likelihood ratios, with all coverage probabilities multiplied by 100. 149

5.15 Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the coverage probability and average length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional, modified profile, penalized (Firth) and penalised based on adjusted responses log-likelihood ratios, with all coverage probabilities multiplied by 100. 150
5.16 Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the coverage probability and average length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional, modified profile, penalized (Firth) and penalised based on adjusted responses log-likelihood ratios, with all coverage probabilities multiplied by 100. 151
5.17 Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the coverage probability and average length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional, modified profile, penalized (Firth) and penalised based on adjusted responses log-likelihood ratios, with all coverage probabilities multiplied by 100. 152
5.18 Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the coverage probability and average length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional, modified profile, penalized (Firth) and penalised based on adjusted responses log-likelihood ratios, with all coverage probabilities multiplied by 100. 153
5.19 Binomial matched observations with true log odds ratio $\psi = 1$. The values of δ that minimize the bias of the penalised log-likelihood estimator based on adjusted responses for various values of q and m. δ was chosen from a set of 50 values ranging from 0.01 to 0.50. 154
5.20 The crying of babies data. 158
5.21 Crying babies real data example: estimates of ψ and its standard error (in parentheses) derived from profile, conditional, modified profile, penalised (Firth), penalised based on adjusted responses log-likelihoods and indirect inference, respectively. 158

5.22 Crying babies real data example: conditional bias and variance (in parentheses) of estimators as the true log odds ratio ψ varies; estimators derived from profile, conditional, modified profile, penalised (Firth), penalised based on adjusted responses log-likelihoods and indirect inference are denoted by $\hat{\psi}$, $\hat{\psi}_c$, $\hat{\psi}_{mp}$, $\hat{\psi}_*$, $\hat{\psi}_a$, $\tilde{\psi}_{a*}$, respectively, and all entries are multiplied by 10. For $\hat{\psi}_a$ and $\tilde{\psi}_{a*}$, the smallest bias value in each column is given in bold face. 159
5.23 Crying babies real data example: unconditional bias and variance (in parentheses) of estimators as the true log odds ratio ψ varies; estimators derived from modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods are denoted by $\hat{\psi}_{mp}$, $\hat{\psi}_*$, $\hat{\psi}_a$, respectively, and all entries are multiplied by 10. For $\hat{\psi}_a$, the smallest bias value in each column is given in bold face. 160

List of Figures

3.1 Plot of the inverse Mills ratio $m_2(\beta^\intercal x_i)$, where $\beta^\intercal x_i = \beta_1 + \beta_2 x_i$ with $\beta_1 = 0.01$ and $\beta_2 = 0.7$ and where $\beta^\intercal x_i$ varies in the range $[0.01, 0.71]$ in the top plot and $\beta^\intercal x_i$ varies in the range $[-6.99, 7.01]$ in the bottom plot. . 51

3.2 Histogram of the simulated values of the Heckit estimates for $n = 150$, after removing the excluded samples. 56

5.1 Matched gamma pairs. Plot of the theoretical bias of the various estimators of ψ for various values of m. 124

5.2 Matched gamma pairs. Plot of the theoretical variance of the various estimators of ψ for different combinations of m and q. 125

5.3 Matched gamma pairs. Plot of the theoretical mean squared error of the various estimators of ψ for different combinations of m and q. 126

5.4 Plot of the value of δ, that makes the penalised log-likelihood based on adjusted responses estimator, $\hat{\psi}_a$, consistent, against ψ for $m = 1$, $m = 3$, $m = 11$ and $m = 39$. 137

Chapter 1

Introduction

1.1 Bias reduction methods

The reduction of estimation bias in parametric statistical models has received considerable attention in the literature because misleading inferences may arise if the magnitude of the bias is large. Under the usual regularity conditions (see, Pace and Salvan, 1997, Section 3.4, pg. 89), the well known maximum likelihood (ML) estimator possesses optimal properties. In particular, it is consistent and asymptotically unbiased with a leading term in its bias expansion of order $O(n^{-1})$. However, for small sample sizes the first-order bias term of the ML estimator could be large enough to significantly impact the performance of inferential procedures. Several methods have been suggested in the literature for the removal of the $O(n^{-1})$ term of the ML estimator in regular parametric statistical models. Kosmidis (2014) identifies all known methods to reduce the bias of any suitably defined estimator, $\hat{\theta}$, not necessarily the ML estimator, as explicit or implicit attempts to approximate the solution of the equation

$$\tilde{\theta} = \hat{\theta} - B_{\hat{\theta}}(\theta), \tag{1.1}$$

with respect to a new estimator $\tilde{\theta}$, where $B_{\hat{\theta}}(\theta) = \mathrm{E}_{\theta}(\hat{\theta} - \theta)$ is the bias function.

The explicit methods, i.e. those that solve the equation $\tilde{\theta} = \hat{\theta} - B_{\hat{\theta}}(\hat{\theta})$ are based on a two step calculation where the bias term is first approximated, analytically or through simulation, then evaluated at $\hat{\theta}$, and then it is subtracted from the initial estimates. For example, when the initial estimator is the ML, the asymptotic bias correction approach approximates the bias function $B_{\hat{\theta}}(\theta)$ by the first term in the asymptotic expansion of

$B_{\hat{\theta}}(\theta)$ in decreasing powers of n, then evaluates it at the ML estimates. Efron (1975) showed that the resulting estimator $\tilde{\theta}$ has bias of asymptotic order $O(n^{-2})$ which is of smaller order than the $O(n^{-1})$ bias of the ML estimator. Obtaining analytical expressions for the first-order bias term of the ML estimator has been extensively studied in the literature, examples include Cox and Snell (1968) who derive an expression for the first-order bias term of the ML estimator in general parametric models and Schaefer (1983) who calculated the first-order bias terms for logistic regressions. The general matrix form of the first-order bias term of the ML estimator is given in Kosmidis and Firth (2010). Nevertheless, the asymptotic bias correction method cannot be applied when the ML estimates are infinite, which is common with models for categorical data, such as logistic regression models and models for censored lifetime data (Heinze and Schemper, 2002; Lin and Kim, 2020). Other popular methods for the approximation of the bias function of the ML estimator are through the jackknife and the bootstrap schemes (Quenouille, 1956; Efron and Tibshirani, 1993). These methods, however, can become computationally expensive if the ML estimator is not in closed form and inherit any of its pathologies such as that of infinite estimates.

An alternative family of estimators in regular parametric models was developed in Firth (1993) were the first-order bias term is removed from the asymptotic bias of the ML estimator by solving a set of adjusted score equations. This method has the advantage of not requiring the value of the ML estimate itself and thus it can still be applied even when the ML estimates are infinite. Firth (1993) considered the case of exponential families with canonical parametrisation, amongst others, and showed that for such models, the method is equivalent to maximising a penalised likelihood where the penalty function is the Jeffreys invariant prior. However, the asymptotic bias correction approach and the method of Firth (1993) have the disadvantage that they can only be applied if the first term in the asymptotic expansion of the bias function, $B_{\hat{\theta}}(\theta)$, of the ML estimator and other quantities such as the Fisher information matrix are available in closed form, a task which is difficult for many models.

Indirect inference (Guerrier et al., 2019) is an alternative, simulation-based procedure, that can be used for bias reduction of the ML estimator and of other estimators. Its simplest version proceeds similarly to the asymptotic bias correction except that we subtract from the ML estimator its full bias, instead of its first order bias term, and evaluate it at the new estimator $\tilde{\theta}$ which therefore becomes the solution of an implicit equation. In other words, the indirect inference estimator, $\tilde{\theta}$, solves the equation $\tilde{\theta} = \hat{\theta} - B_{\hat{\theta}}(\tilde{\theta})$. The

advantage of indirect inference over asymptotic bias correction is that it can be applied even if the bias of the ML estimator is not available in closed form, thus in principle it can be applied to any parametric model. For example, Kuk (1995) describes a simulation-based approach of implementing an iterative bias correction of the best linear unbiased prediction (BLUP) estimator, or any conveniently defined estimator, in generalised linear models with random effects, to yield estimates that are asymptotically unbiased and consistent. Of course indirect inference inherits the disadvantages of explicit methods such as asymptotic bias correction, jackknife and the bootstrap because it explicitly depends on the original estimator and it can be computationally expensive to calculate.

As can be seen from the above review, each method of bias reduction has its own advantages and disadvantages and there is no one method that produces better results than others. Nevertheless, all methods differ in terms of applicability and requirements for their computation. Kosmidis and Lunardon (2020) give an excellent classification of key bias reduction methods in the literature (Kosmidis and Lunardon, 2020, Table 1) in terms of the level of model specification, the way the methods approximate the bias, the type of the method according to the classification in Kosmidis (2014) and on the method's requirements in terms of computation of expectations, differentiation and access to the original estimator. Moreover, Kosmidis and Lunardon (2020) develop a novel method for the reduction of the asymptotic bias of M-estimators from general, unbiased estimating functions and call the new estimation method Reduced Bias M-estimation (RBM-estimation). Their method results in estimators with bias of lower asymptotic order than the original M-estimators and relies on empirical additive adjustments that depend only on the first two derivatives of the contributions to the unbiased estimating functions. Unlike the bias reducing adjusted scores approach in Firth (1993), the empirical adjustments do not require the computation of cumbersome expectations nor do they require the, potentially expensive, calculation of M-estimates from simulated samples, as is the case with simulation based bias reduction methods. In particular, RBM-estimation applies to models that are at least partially specified, does not rely on the original estimator $\hat{\theta}$ and uses an analytical approximation to the bias function $B_{\hat{\theta}}(\theta)$ that relies only on derivatives of the contributions to the estimating functions, a task which nowadays requires increasingly less analytical effort because of the availability of comprehensive automatic differentiation routines in popular computing environments. When the estimating functions are the components of the gradient of an objective function, as is the case in maximum likelihood estimation where the objective function is the log-likelihood, Kosmidis and Lunardon

(2020) show that bias reduction can always be achieved by the maximisation of an appropriately penalised version of the objective, unlike the method in Firth (1993) which does not always have a penalised likelihood interpretation. Moreover, they show that the RBM-estimators have the same asymptotic distribution and efficiency properties as the original M-estimators.

1.2 Stratified models

A challenging modelling setting is that of independent stratified observations (Sartori, 2003), with density or probability mass function

$$f_i(y_{ij}; \psi, \lambda_i),$$

where $i = 1, \ldots, q$ and $j = 1, \ldots, m_i$. The sample size is $n = \sum_{i=1}^{q} m_i$ and the unknown parameter is $\theta = (\psi^\top, \lambda)^\top$ where ψ is a p-dimensional parameter of interest and the nuisance parameter $\lambda = (\lambda_1, \ldots, \lambda_q)^\top$ has dimension q, with one parameter per stratum. We are interested in likelihood inference procedures for ψ when both the number of strata, q, and the stratum sample sizes, m_i, are allowed to increase to infinity. The log-likelihod function for θ based on independent observations $y_{11}, \ldots, y_{1m_1}, \ldots, y_{q1}, \ldots, y_{qm_q}$ is

$$l(\theta) = \sum_{i=1}^{q} \sum_{j=1}^{m_i} \ln f_i(y_{ij}; \psi, \lambda_i) = \sum_{i=1}^{q} l_i(\psi, \lambda_i).$$

The presence of a substantial number of nuisance parameters becomes a common obstacle that statisticians encounter in the theory of inference and estimation because, in general, the ML estimator does not yield consistent point estimates of ψ as the dimension of the nuisance parameter becomes large relative to the stratum sample size (for examples, see Neyman and Scott, 1948, who refer to ψ and λ as the structural and incidental parameters, respectively). This is known as the incidental parameter problem.

A way for eliminating nuisance parameters in estimation is to introduce the profile log-likelihood which is obtained by replacing the nuisance parameters in the log-likelihood with their ML estimates for a fixed value of the parameter of interest. The expected value of the profile score however is not zero in general, and estimates generated from the profile log-likelihood may be biased and inconsistent. See Example 1 of McCullagh and Tibshirani (1990) on the many normal means model where n pairs of independent

normal random variables are observed and the variance σ^2 is of interest while the means are nuisance. In this case the estimator of the variance generated from the profile log-likelihood is inconsistent. A score function that has zero expectation and whose variance equals the negative of its expected derivative matrix is said to be unbiased and information unbiased, respectively (Lindsay, 1982). According to this terminology, it is well known that the gradient of the log-likelihood function is both unbiased and information unbiased (see for example, Silvey, 1970, Chapter 2, pg. 36 and 40, for a proof). On the other hand, it can be shown that, in general, the score function computed from the profile log-likelihood is neither unbiased nor information unbiased (McCullagh and Tibshirani, 1990, Remark 2).

Since in problems with large numbers of nuisance parameters the profile log-likelihood is known to give inconsistent estimates, an alternative route is to work with modified profile log-likelihoods of the form $l_M(\psi) = l_p(\psi) + M(\psi)$, where $l_p(\psi)$ is the profile log-likelihood. Some examples of modified log-likelihood functions include the approximate conditional profile log-likelihood of Cox and Reid (1987), which requires orthogonality of the parameter of interest and the nuisance parameter, the modified profile log-likelihood of Barndorff-Nielson (1983) whose computation requires a sample space derivative and the adjusted profile log-likelihood of McCullagh and Tibshirani (1990), which is a simple adjustment of the profile log-likelihood so as to make it both asymptotically unbiased and information unbiased. The approximate conditional and modified profile log-likelihood functions are often simple to compute for exponential and composite group families and only involve the observed information matrix for the components of the nuisance parameter which is readily available from direct differentiation (Pace and Salvan, 1997, §4.7). On the other hand, the adjusted profile log-likelihood involves expectations which can be cumbersome to evaluate even for exponential family models. It has been shown through many examples that when the profile likelihood performs poorly, modified profile likelihoods can perform much better. For example, McCullagh and Tibshirani (1990) show that in the many normal means problem described above, the estimator of the variance derived from the profile likelihood is biased and inconsistent, while the same estimator derived from each of the approximate conditional, modified and adjusted profile likelihoods is unbiased and consistent. Lunardon (2018) showed that the bias reduction approach of Firth (1993) provides an inferential framework which is, from an asymptotic perspective, equivalent to that for modified profile log-likelihoods when dealing with nuisance parameters. The advantage of bias reduction of Firth (1993) over modified profile log-

likelihoods is that it can handle the problem of monotone likelihoods for stratified models with categorical responses. Nevertheless, the approach of Firth (1993) is not in general invariant under interest-respecting reparameterizations.

In this thesis, we propose a general simulation-based algorithm for indirect inference, motivated by that in Kuk (1995), which we adapt to nuisance parameter settings. In theory this method could be applied to reduce the bias of any suitably defined initial estimator of the parameter of interest in the presence of a set of nuisance parameters.

1.3 Outline

The recent framework for bias reduction (Kosmidis and Lunardon, 2020) motivates its application to non-standard models, such as those for censored observations, and its comparison to other well known methods in terms of reducing estimation bias especially in small sample sizes.

The current thesis is organised in the following way. In Chapter 2, we set up some notation and likelihood related quantities and we give a brief description of the various modifications of the likelihood function that have been proposed in the literature for improving estimation in stratified settings in the presence of nuisance parameters. We also review some of the explicit and implicit methods of bias reduction as in Kosmidis (2014), including the asymptotic bias correction, indirect inference and the bias-reducing adjusted score equations of Firth (1993). We describe the simulation-based algorithm for indirect inference, motivated by that in Kuk (1995), for nuisance parameter settings. We also describe the novel method of reduced-bias M estimation of Kosmidis and Lunardon (2020) which has wider applicability than previously proposed bias reduction methods.

In Chapter 3 we study the Heckit (Tobit II) model and review current methods for the estimation of its parameters. This model is non-standard because it is used to model non-randomly selected samples. For example, one observes market wage offers but has access to wage observations for only those who work. Since people who work are selected non-randomly from the population, estimating the effect of working on wages from the subpopulation who work may introduce sample selection bias because the wages of workers do not, in general, afford a reliable estimate of what non workers would have earned had they worked. The most common methods of estimation for this model in the literature are maximum likelihood and the Heckman two-step correction method (Heckman, 1976, 1979). We implement the methods of indirect inference and reduced-bias M estimation of

Kosmidis and Lunardon (2020) to this model and compare their performance in terms of bias reduction with ML estimation and the Heckman two-step method through simulation studies.

Chapter 4 focuses on the accelerated failure time model with the Weibull distribution when the lifetimes are right censored. We extend the indirect inference and empirical bias reducing adjustments method of Kosmidis and Lunardon (2020) to this model and compare their performance, through simulations, to ML estimation in terms of their frequentist properties.

In Chapter 5 we consider two stratified models, namely the matched gamma pairs model and the binomial matched pairs model. For the matched gamma pairs model, we review the profile, approximate conditional profile and modified profile log-likelihood methods of estimation of the parameter of interest which all yield biased and inconsistent estimates. We derive the adjusted profile log-likelihood and show that it yields an unbiased and consistent estimator of the parameter of interest which coincides with the estimator derived from indirect inference. For the binomial matched pairs model we review current methods of point estimation of the log odds ratio. These include the conditional and modified profile log-likelihoods (Barndorff-Nielson, 1983) and the bias reducing adjusted scores approach of Firth (1993). The former two methods however are known to inherit the problem of infinite estimates of the parameter of interest. We propose a penalised version of the log-likelihood function based on adjusted responses which always results in a finite estimator of the log odds ratio. The probability limit of the adjusted log-likelihood estimator is derived and it is shown that in certain settings the ML, conditional and modified profile log-likellihood estimators drop out as special cases of the former. We implement indirect inference to the adjusted log-likeihood estimator. The method of Firth (1993) also prevents infinite estimates of the log odds ratio and we compare its performance with our proposed adjustment of the log-likelihood, indirect inference and the current methods above through a complete enumeration study as in Lunardon (2018).

Finally, a summary of the main results and concluding remarks are given in Chapter 6, and we give some suggestions for further work in the area.

Appendices A and B include the algebraic derivations of several results presented in the main text.

Chapter 2

Likelihood modifications and bias reduction methods

2.1 Modifications of the likelihood function

In this section we define important likelihood-related quantities as in Pace and Salvan (1997, Chapter 1, §1.4.2) and we give a brief description of the class of linear exponential family models. We consider statistical models where the parameter space is partitioned into a parameter of interest and nuisance parameters. The more complex the structure of the nuisance parameters, the more attractive the possibility of basing inference on a likelihood function which only depends on the parameter of interest, thus eliminating the nuisance parameters altogether. This can be achieved by introducing pseudo-likelihoods which refer to any function of the data which depends only on the parameter of interest and which behaves, in some respects, as if it were a genuine likelihood (i.e. score with zero null expectation etc). We review some of the notions of pseudo-likelihoods such as the approximate conditional, modified and adjusted profile log-likelihoods as in Cox and Reid (1987), Barndorff-Nielson (1983) and McCullagh and Tibshirani (1990), respectively.

2.1.1 Likelihood and related quantities

Suppose that $\mathscr{Y} = (Y_1^\top, Y_2^\top, \ldots, Y_n^\top)^\top$ is a random n sample of independent d-vectors where each Y_i is distributed with density function $f_Y(y,\theta), y \in \mathscr{Y} \subseteq \mathbb{R}^d$, depending on a vector parameter $\theta = (\theta_1, \ldots, \theta_p)^\top \in \Theta \subseteq \mathbb{R}^p$. Suppose that the parameter θ can be partitioned as $\theta = (\psi^\top, \lambda^\top)^\top$ where $\psi = (\psi_1, \ldots, \psi_r)^\top$ is the parameter of interest and $\lambda = (\lambda_1, \ldots, \lambda_s)^\top$

is the nuisance parameter.

The *likelihood* and *log-likelihood* functions are defined, respectively, by

$$L(\theta) = L(\psi, \lambda) = \prod_{i=1}^{n} f_{Y_i}(y_i, \theta)\,,$$

$$l(\theta) = l(\psi, \lambda) = \ln L(\theta) = \sum_{i=1}^{n} \ln f_{Y_i}(y_i, \theta)\,.$$

Let the partial derivatives of the log-likelihood function be

$$\begin{aligned} l_r &= l_r(\theta) = \frac{\partial l(\theta)}{\partial \theta_r}\,, \\ l_{rs} &= l_{rs}(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s}\,, \end{aligned}$$

etc. These derivatives are called *likelihood quantities* because they are obtained from the likelihood function.

The *score* function is defined as $S(\theta) = \nabla_\theta l(\theta) = (l_1, \ldots, l_p)^\intercal$, where ∇_θ denotes the gradient of $l(\theta)$ with respect to θ.

The *observed information matrix*, $j(\theta)$, is

$$j = j(\theta) = -\nabla_\theta \nabla_\theta^\intercal l(\theta) = \begin{pmatrix} -l_{11} & \cdots & -l_{1p} \\ \vdots & \ddots & \vdots \\ -l_{p1} & \cdots & -l_{pp} \end{pmatrix}.$$

Furthermore, the *expected information* or *Fisher information matrix* is

$$i = i(\theta) = \mathrm{E}_\theta\{j(\theta)\}\,.$$

2.1.2 Exact conditional likelihood

Let u be a statistic such that the factorisation

$$f_Y(y; \psi, \lambda) = f_U(u; \psi, \lambda) f_{Y|U=u}(y; u, \psi) \tag{2.1}$$

holds. In other words, U is partially sufficient for λ. Provided that the likelihood factor which corresponds to $f_U(.)$ can be neglected, inference about ψ can be based on the

conditional model with density $f_{Y|U}(.)$. The corresponding likelihood function

$$L_c(\psi) = L_c(\psi; y|u) = f_{Y|U=u}(y; u, \psi) \tag{2.2}$$

is called the *conditional likelihood* based on conditioning on $U = u$.

2.1.3 Profile likelihood

Constructing conditional likelihoods is only possible when partially sufficient statistics are available for the nuisance parameter. Thus factorisations such as (2.1) are not always possible and arise, for example, in exponential and group families. An alternative in such cases is the profile likelihood.

The *overall maximum likelihood (ML) estimator* of θ is denoted by $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$ and is defined by $\hat{\theta} = \arg\max_{\psi,\lambda} l(\psi, \lambda)$. The ML estimator can be obtained by the solution of the score equations $S(\theta) = \nabla_\theta l(\theta) = 0$.

Let $\hat{\lambda}_\psi$ denote the constrained maximum likelihood estimate of λ for a fixed value of ψ and $\hat{\psi}_\lambda$ denote the constrained maximum likelihood estimate of ψ for a fixed value of λ.

The *profile log-likelihood* function for ψ is defined by

$$l_p(\psi) = \max_{\lambda} l(\psi, \lambda) = l(\psi, \hat{\lambda}_\psi)\,.$$

The constrained ML estimator of λ for fixed ψ is defined by $\hat{\lambda}_\psi = \arg\max_{\lambda} l(\psi, \lambda)$, i.e. is the solution in λ of $\partial l(\psi, \lambda)/\partial\lambda = 0$, while the overall ML estimator of ψ is defined by $\hat{\psi} = \arg\max_{\psi} l_p(\psi)$, i.e. is the solution in ψ of $\partial l_p(\psi)/\partial\psi = 0$. Therefore, the overall ML estimator of λ is given by $\hat{\lambda} = \hat{\lambda}_{\hat{\psi}}$.

2.1.4 Approximate conditional profile likelihood

Various modifications, have been proposed in the literature, aiming to improve profile likelihoods, in the sense that the expectation of the profile score is non zero, so estimation of the parameter of interest using profile likelihoods yields biased and inconsistent estimates, when the number of nuisance parameters grows relative to the stratum sample size. These modifications include the approximate conditional profile likelihood, the modified profile likelihood and the adjusted profile likelihood.

According to Cox and Reid (1987), the *approximate conditional profile log-likelihood* function for ψ in the case of a single parameter of interest ($r = 1$) is defined by

$$l_{cp}(\psi) = l(\psi, \hat{\lambda}_\psi) - \frac{1}{2}\ln\{\det j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\} \tag{2.3}$$

where $j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$ is the observed information per observation for the λ components and is given in terms of the negative of the second derivative of the log-likelihood function with respect to λ as

$$j_{\lambda\lambda}(\psi,\lambda) = -\frac{\partial^2 l(\psi,\lambda)}{\partial\lambda\partial\lambda^\intercal} = -\begin{pmatrix} \frac{\partial^2 l(\psi,\lambda)}{\partial\lambda_1^2} & \frac{\partial^2 l(\psi,\lambda)}{\partial\lambda_1\partial\lambda_2} & \cdots & \frac{\partial^2 l(\psi,\lambda)}{\partial\lambda_1\partial\lambda_s} \\ \frac{\partial^2 l(\psi,\lambda)}{\partial\lambda_2\partial\lambda_1} & \frac{\partial^2 l(\psi,\lambda)}{\partial\lambda_2^2} & \cdots & \frac{\partial^2 l(\psi,\lambda)}{\partial\lambda_2\partial\lambda_s} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\psi,\lambda)}{\partial\lambda_s\partial\lambda_1} & \frac{\partial^2 l(\psi,\lambda)}{\partial\lambda_s\partial\lambda_2} & \cdots & \frac{\partial^2 l(\psi,\lambda)}{\partial\lambda_s^2} \end{pmatrix}.$$

The properties of $l_{cp}(\psi)$ requires an orthogonal parametrization of ψ and λ, i.e. $\mathrm{E}(-\partial^2 l(\psi,\lambda)/\partial\psi\partial\lambda_j) = 0$ for $j = 1, 2, \ldots, s$. This is because the approximate conditional profile log-likelihood is an approximation to a likelihood with separable parameters.

Note that if we have a single nuisance parameter then the observed information per observation for λ is a scalar and is given by

$$j_{\lambda\lambda}(\psi,\lambda) = -\frac{d^2 l(\psi,\lambda)}{d\lambda^2}.$$

2.1.5 Modified profile likelihood

The modified profile log-likelihood can be though of as a correction for non-orthogonality by adding a penalty term to the approximate conditional profile log-likelihood function.

According to Barndorff-Nielson (1983), the *modified profile log-likelihood* function for ψ in the case of a single parameter of interest is defined by

$$l_{mp}(\psi) = l(\psi, \hat{\lambda}_\psi) - \frac{1}{2}\ln\{\det j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\} + \ln\{\det(d\hat{\lambda}/d\hat{\lambda}_\psi)\} \tag{2.4}$$

where $d\hat{\lambda}/d\hat{\lambda}_\psi$ is the matrix of partial derivatives of $\hat{\lambda}$ with respect to $\hat{\lambda}_\psi$, called the sample space derivative, i.e.

$$\frac{d\hat{\lambda}}{d\hat{\lambda}_\psi} = \frac{\partial\hat{\lambda}}{\partial\hat{\lambda}_\psi^\intercal} = \begin{pmatrix} \frac{\partial\hat{\lambda}_1}{\partial\hat{\lambda}_{1,\psi}} & \frac{\partial\hat{\lambda}_1}{\partial\hat{\lambda}_{2,\psi}} & \cdots & \frac{\partial\hat{\lambda}_1}{\partial\hat{\lambda}_{s,\psi}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\hat{\lambda}_s}{\partial\hat{\lambda}_{1,\psi}} & \frac{\partial\hat{\lambda}_s}{\partial\hat{\lambda}_{2,\psi}} & \cdots & \frac{\partial\hat{\lambda}_s}{\partial\hat{\lambda}_{s,\psi}} \end{pmatrix}.$$

Note that if we have a single nuisance parameter then the sample space derivative becomes a scalar. Furthermore, if $\hat{\lambda} = \hat{\lambda}_\psi$ then the matrix $\mathrm{d}\hat{\lambda}/\mathrm{d}\hat{\lambda}_\psi$ becomes the identity matrix and hence $l_{mp}(\psi) = l_{cp}(\psi)$. Moreover, the above definition does not require orthogonality of ψ and λ.

An alternative and often more convenient expression for (2.4) can be derived as follows (Severini, 2000, Section 9.3): Suppose that the log-likelihood function can be written in terms of the overall ML estimates $\hat{\psi}$ and $\hat{\lambda}$ and an ancillary statistic a, as $l(\psi,\lambda;\hat{\psi},\hat{\lambda},a)$, then by the definition of $\hat{\lambda}_\psi$ we know that

$$\left.\frac{\partial l(\psi,\lambda;\hat{\psi},\hat{\lambda},a)}{\partial\lambda}\right|_{\lambda=\hat{\lambda}_\psi} = 0,$$

and so differentiating the above with respect to $\hat{\lambda}$, using the multivariate chain rule for composite functions, yields

$$\frac{\partial\hat{\lambda}_\psi^\intercal}{\partial\hat{\lambda}}\left.\frac{\partial^2 l(\psi,\lambda;\hat{\psi},\hat{\lambda},a)}{\partial\hat{\lambda}_\psi^\intercal\partial\lambda}\right|_{\lambda=\hat{\lambda}_\psi} + \frac{\partial\hat{\lambda}}{\partial\hat{\lambda}}\left.\frac{\partial^2 l(\psi,\lambda;\hat{\psi},\hat{\lambda},a)}{\partial\hat{\lambda}^\intercal\partial\lambda}\right|_{\lambda=\hat{\lambda}_\psi} = 0.$$

Rearranging the above and taking the determinant we get

$$\begin{aligned} \det\left(\frac{\partial\hat{\lambda}_\psi^\intercal}{\partial\hat{\lambda}}\right) &= \det\left(\left.\frac{\partial^2 l(\psi,\lambda)}{\partial\hat{\lambda}^\intercal\partial\lambda}\right|_{\lambda=\hat{\lambda}_\psi}\right)\left[\det\left(-\left.\frac{\partial^2 l(\psi,\lambda)}{\partial\hat{\lambda}_\psi^\intercal\partial\lambda}\right|_{\lambda=\hat{\lambda}_\psi}\right)\right]^{-1} \\ &= \det\left(\left.\frac{\partial^2 l(\psi,\lambda)}{\partial\hat{\lambda}^\intercal\partial\lambda}\right|_{\lambda=\hat{\lambda}_\psi}\right)\left[\det\left(-\left.\frac{\partial^2 l(\psi,\lambda)}{\partial\lambda^\intercal\partial\lambda}\right|_{\lambda=\hat{\lambda}_\psi}\right)\right]^{-1} \\ &= \det\left(\left.\frac{\partial^2 l(\psi,\lambda)}{\partial\hat{\lambda}^\intercal\partial\lambda}\right|_{\lambda=\hat{\lambda}_\psi}\right)[\det j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)]^{-1}, \end{aligned} \tag{2.5}$$

where we have suppressed the dependence of the log-likelihood function on the ML estimates and the ancillary parameter, and where the second equality follows since differentiating the log-likelihood with respect to λ, substituting $\lambda = \hat{\lambda}_\psi$, then differentiating again with respect to $\hat{\lambda}_\psi$, is the same as differentiating the log-likelihood twice with respect to λ then substituting $\lambda = \hat{\lambda}_\psi$.

The modified profile log-likelihood function may therefore be written as

$$l_{mp}(\psi) = l(\psi, \hat{\lambda}_\psi) + \frac{1}{2}\ln\{\det j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\} - \ln\left\{\det\left(\frac{\partial^2 l(\psi, \hat{\lambda}_\psi)}{\partial\hat{\lambda}^\intercal \partial\lambda}\right)\right\}. \tag{2.6}$$

Since the log-likelihood is a function of sufficient statistics s and the parameter θ, it is well known that if the dimension of s and θ are equal, the ML estimator $\hat{\theta}$ is usually a one-to-one function of s and then $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$ is also sufficient. We can then write the log-likelihood as $l(\theta; \hat{\psi}, \hat{\lambda})$. If however, the maximum likelihood estimator is not sufficient then an ancillary statistic a is needed such that $(\hat{\theta}, a)$ is the sufficient statistic for the model and hence, we may write the log likelihood more generally as $l(\theta; \hat{\psi}, \hat{\lambda}, a)$.

The sample space derivative is therefore sometimes difficult to evaluate because in some cases in order to obtain the first term of (2.5), the log-likelihood function must be expressed as a function of $\hat{\psi}$, $\hat{\lambda}$ and an ancillary a so that it can be differentiated partially with respect to $\hat{\lambda}$, holding a fixed. Re-expressing the log-likelihood in this form is usually tough and is not possible for all models, hence an approximation to such derivative is needed.

The modified profile log-likelihood however, can be simplified for linear exponential families (Davison, 2003, Example 12.22). The likelihood function in a linear exponential family in canonical form can be written as

$$L(\psi, \lambda) \equiv h(y)\exp\left(\psi^\intercal t_1 + \lambda^\intercal t_2 - \kappa(\psi, \lambda)\right),$$

where there is no ancillary statistic and where $h(y)$ is a function of the data only, $t_1 = (t_{11}, \ldots, t_{1r})^\intercal$ and $t_2 = (t_{21}, \ldots, t_{2s})^\intercal$ are the sufficient statistics for the parameter of interest and the nuisance parameter, respectively. Furthermore, $\kappa(\psi, \lambda)$ is the logarithm of a normalisation factor, that ensures that the corresponding density $f_Y(y; \psi, \lambda)$ integrates to one, called the cumulant generating function and is given by

$$\kappa(\psi, \lambda) = \ln\left(\int_y h(y)\exp\left(\psi^\intercal t_1 + \lambda^\intercal t_2\right)\mathrm{d}y\right).$$

The log-likelihood function is

$$l(\psi,\lambda) = \psi^{\mathsf{T}} t_1 + \lambda^{\mathsf{T}} t_2 - \kappa(\psi,\lambda) + \text{constant}. \tag{2.7}$$

The overall ML estimates $\hat{\psi}$ and $\hat{\lambda}$ are solutions of the equations $t_1 - \kappa_\psi(\hat{\psi},\hat{\lambda}) = 0$ and $t_2 - \kappa_\lambda(\hat{\psi},\hat{\lambda}) = 0$, respectively where $\kappa_\psi(\psi,\lambda) = \partial\kappa(\psi,\lambda)/\partial\psi$ and $\kappa_\lambda(\psi,\lambda) = \partial\kappa(\psi,\lambda)/\partial\lambda$.

Therefore the log-likelihood may be written as

$$l(\psi,\lambda;\hat{\psi},\hat{\lambda}) = \psi^{\mathsf{T}}\kappa_\psi(\hat{\psi},\hat{\lambda}) + \lambda^{\mathsf{T}}\kappa_\lambda(\hat{\psi},\hat{\lambda}) - \kappa(\psi,\lambda),$$

and hence the first and second terms of (2.5) become, respectively

$$\begin{aligned}
\left.\frac{\partial^2 l(\psi,\lambda;\hat{\psi},\hat{\lambda})}{\partial\hat{\lambda}\partial\lambda}\right|_{\lambda=\hat{\lambda}_\psi} &= \frac{\partial}{\partial\hat{\lambda}}\left(\left.\frac{\partial l(\psi,\lambda;\hat{\psi},\hat{\lambda})}{\partial\lambda}\right|_{\lambda=\hat{\lambda}_\psi}\right) \\
&= \frac{\partial}{\partial\hat{\lambda}}\Big(\kappa_\lambda(\hat{\psi},\hat{\lambda}) - \kappa_\lambda(\psi,\hat{\lambda}_\psi)\Big) \\
&= \frac{\partial}{\partial\hat{\lambda}}\Big(\kappa_\lambda(\hat{\psi},\hat{\lambda})\Big) \\
&= \kappa_{\lambda\lambda}(\hat{\psi},\hat{\lambda}), \\
j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi;\hat{\psi},\hat{\lambda}) &= -\frac{\partial}{\partial\lambda}\left(\left.\frac{\partial l(\psi,\lambda;\hat{\psi},\hat{\lambda})}{\partial\lambda}\right|_{\lambda=\hat{\lambda}_\psi}\right) \\
&= -\frac{\partial}{\partial\lambda}\Big(\kappa_\lambda(\hat{\psi},\hat{\lambda}) - \kappa_\lambda(\psi,\lambda)\Big)\Big|_{\lambda=\hat{\lambda}_\psi} \\
&= -\frac{\partial}{\partial\lambda}\Big(-\kappa_\lambda(\psi,\lambda)\Big)\Big|_{\lambda=\hat{\lambda}_\psi} \\
&= \kappa_{\lambda\lambda}(\psi,\hat{\lambda}_\psi).
\end{aligned}$$

However, the second partial derivative of the cumulant generating function with respect to λ can be written as $\kappa_{\lambda\lambda}(\hat{\psi},\hat{\lambda}) = -\partial^2 l(\hat{\psi},\hat{\lambda};\hat{\psi},\hat{\lambda})/\partial\lambda^2 = j_{\lambda\lambda}(\hat{\psi},\hat{\lambda};\hat{\psi},\hat{\lambda})$, i.e. in terms of the observed information and similarly $\kappa_{\lambda\lambda}(\psi,\hat{\lambda}_\psi) = -\partial^2 l(\psi,\hat{\lambda}_\psi;\hat{\psi},\hat{\lambda})/\partial\lambda^2 = j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi;\hat{\psi},\hat{\lambda})$.

Therefore (2.5) reduces to

$$\det\left(\frac{\partial\hat{\lambda}_\psi}{\partial\hat{\lambda}}\right) = [\det j_{\lambda\lambda}(\hat{\psi},\hat{\lambda};\hat{\psi},\hat{\lambda})][\det j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi;\hat{\psi},\hat{\lambda})]^{-1}. \tag{2.8}$$

Thus substituting for the last term of (2.6), the modified profile log-likelihood function for ψ becomes

$$\begin{aligned} l_{mp}(\psi) &= l(\psi,\hat{\lambda}_\psi) + \frac{1}{2}\ln\{\det j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi;\hat{\psi},\hat{\lambda})\} - \ln\{\det j_{\lambda\lambda}(\hat{\psi},\hat{\lambda};\hat{\psi},\hat{\lambda})\} \\ &= l(\psi,\hat{\lambda}_\psi) + \frac{1}{2}\ln\{\det j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi;\hat{\psi},\hat{\lambda})\} + \text{constant}, \end{aligned} \tag{2.9}$$

where the term $\ln\{\det j_{\lambda\lambda}(\hat{\psi},\hat{\lambda};\hat{\psi},\hat{\lambda})\}$ has been neglected since it is independent of ψ.

2.1.6 Adjusted profile likelihood

The *profile log-likelihood score* function is defined by

$$U(\psi) = \frac{\partial l_p(\psi)}{\partial \psi}.$$

This score function is neither unbiased nor information unbiased (Lindsay, 1982; Silvey, 1970).

Now consider the case of a single parameter of interest and let $m(\psi)$ and $w(\psi)$ be two functions such that

$$m(\psi) = \mathrm{E}_{\psi,\hat{\lambda}_\psi}\{U(\psi)\}$$

$$w(\psi) = \left\{ -\mathrm{E}_{\psi,\hat{\lambda}_\psi}\left(\frac{\partial U(\psi)}{\partial \psi}\right) + \frac{\partial m(\psi)}{\partial \psi}\right\} \Big/ \mathrm{Var}_{\psi,\hat{\lambda}_\psi}\{U(\psi)\},$$

where expectations are computed at $\left(\psi,\hat{\lambda}_\psi^{\mathsf{T}}\right)^{\mathsf{T}}$ rather than at the true parameter value.

According to McCullagh and Tibshirani (1990), the *adjusted profile log-likelihood score* function is defined as

$$\widetilde{U}(\psi) = \Big\{U(\psi) - m(\psi)\Big\} w(\psi). \tag{2.10}$$

This means that the *adjusted profile log-likelihood* function for ψ is given by

$$l_{ap}(\psi) = \int^{\psi} \widetilde{U}(t)\,\mathrm{d}t. \tag{2.11}$$

In contrast to the profile score, the adjusted score function is asymptotically unbiased and information unbiased.

2.2 Bias reduction methods

2.2.1 Preamble

Suppose that interest lies in the estimation of the vector of parameters $\theta = (\psi^\intercal, \lambda^\intercal)^\intercal = (\theta_1, \ldots, \theta_p)^\intercal \in \Theta \subset \mathbb{R}^p$, as before, from data $y = (y_1, \ldots, y_n)^\intercal$ assumed to be realisations of a random quantity $Y = (Y_1, \ldots, Y_n)^\intercal$ distributed according to a parametric model G with density function $f_Y(y|x, \theta)$. The subscript n here is used as a measure of the information in the data and is usually the sample size. An *estimator* of θ, not necessarily the overall ML estimator, is a function $\hat{\theta} \equiv t\big(Y\big)$ and in the presence of observed data y the estimate would be $t\big(y\big)$.

The *bias* of an estimator $\hat{\theta}$ is defined by

$$B_{\hat{\theta}}(\theta) = g^*(\theta) - \theta\,,$$

where $g^*(\theta) = \mathrm{E}_\theta(\hat{\theta})$ and where the expectation is taken with respect to G which is the joint distribution function of the process that generated the data and where the symbol θ inside the brackets denotes that the corresponding function is evaluated at θ. An estimator whose bias is equal to zero for all values of the parameter θ is said to be *unbiased*.

It is well known that under standard regularity conditions, the overall ML estimator of θ, $\hat{\theta}$, has optimal properties. One important property is that $\hat{\theta}$ has bias of asymptotic order $O(n^{-1})$ which implies that this bias vanishes as $n \to \infty$ where n is the sample size. However, for small samples where n is finite the bias of $\hat{\theta}$ may have considerable magnitude. Biased estimators arise frequently in statistics and can result in misleading inferences if the magnitude of the bias is large (see for example, Kosmidis, 2014). Kosmidis (2014) identifies all known methods to reduce bias as attempts to approximate the solution of the equation

$$\tilde{\theta} = \hat{\theta} - B_{\hat{\theta}}(\theta)\,, \tag{2.12}$$

with respect to a new estimator $\tilde{\theta}$. These methods can be distinguished into *explicit* and *implicit* methods.

If both $B_{\hat{\theta}}(\theta)$ and θ were known then $\tilde{\theta}$ would be unbiased since

$$\begin{aligned} B_{\tilde{\theta}}(\theta) &= \mathrm{E}_\theta(\tilde{\theta}) - \theta \\ &= \mathrm{E}_\theta(\hat{\theta}) - B_{\hat{\theta}}(\theta) - \theta \\ &= B_{\hat{\theta}}(\theta) - B_{\hat{\theta}}(\theta) \\ &= 0, \end{aligned}$$

but of course, estimation would be unnecessary if θ was known, and moreover, the function $B_{\hat{\theta}}(\theta)$ usually is not available in closed form or G is unknown.

For many estimators $\hat{\theta}$ including the ML estimator, the bias function can be expanded in decreasing powers of n as

$$B_{\hat{\theta}}(\theta) = \mathrm{E}_\theta(\hat{\theta}) - \theta = \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + \frac{b_3(\theta)}{n^3} + O(n^{-4}), \tag{2.13}$$

where n is the sample size, θ is the true but unknown parameter value and b_t $(t = 1, 2, 3, \ldots)$ is an appropriate sequence of functions of θ.

Explicit methods rely on estimating the bias function $B_{\hat{\theta}}(\theta)$ at $\hat{\theta}$ and then subtracting it from $\hat{\theta}$ resulting in the new estimator $\tilde{\theta}$. The resulting equation is then solved with respect to $\hat{\theta}$.

2.2.2 Asymptotic bias correction

One approach to correct the bias of the overall ML estimator is to define a bias-corrected estimator

$$\tilde{\theta} = \hat{\theta} - \frac{b_1(\hat{\theta})}{n}, \tag{2.14}$$

where we have estimated the bias function $B_{\hat{\theta}}(\theta)$ in equation (2.12) by $b_1(\hat{\theta})/n$, the first-term in the right hand side of (2.13); the asymptotic expansion of the bias of the overall ML estimator, evaluated at $\hat{\theta}$. It may be shown (Pace and Salvan, 1997, Section 9.4) that $\tilde{\theta}$ has bias of asymptotic order $O(n^{-2})$ which is of smaller order than the $O(n^{-1})$ bias of $\hat{\theta}$.

The general form of the first-order bias term of the overall ML estimator $\hat{\theta}$ can be found in matrix notation in Kosmidis and Firth (2010) and is

$$\frac{b_1(\theta)}{n} = -\{i(\theta)\}^{-1} A(\theta), \tag{2.15}$$

where the function $A(\theta)$ is a p-dimensional vector with t-th component

$$A_t(\theta) = \frac{1}{2}\mathrm{tr}\left[\{i(\theta)\}^{-1}\left\{P_t(\theta) + Q_t(\theta)\right\}\right] \qquad (t = 1, \ldots, p)\,, \tag{2.16}$$

and where

$$P_t(\theta) = \mathrm{E}_\theta\left\{S(\theta)S(\theta)^\intercal S_t(\theta)\right\} \qquad (t = 1, \ldots, p)\,, \tag{2.17}$$

$$Q_t(\theta) = -\mathrm{E}_\theta\left\{j(\theta)S_t(\theta)\right\} \qquad (t = 1, \ldots, p)\,, \tag{2.18}$$

are higher order joint null moments of log-likelihood derivatives with $S_t(\theta)$ denoting the t-th log-likelihood derivative.

Implicit methods, on the contrary, rely on approximating the bias function but at the target estimator $\tilde{\theta}$ and then subtracting it from $\hat{\theta}$ resulting in the new estimator $\tilde{\theta}$. The resulting equation is then solved with respect to $\tilde{\theta}$ and hence, $\tilde{\theta}$ is the solution of an implicit equation.

2.2.3 Indirect inference

Indirect inference is a class of inferential procedures that appeared in the Econometrics literature in Gourieroux et al. (1993) and can be used for bias reduction. If the maximum likelihood estimator and its bias function can be written in closed form, then the simplest method of bias reduction via indirect inference relies on solving the equation

$$\tilde{\theta} = \hat{\theta} - B_{\hat{\theta}}(\tilde{\theta})\,, \tag{2.19}$$

with respect to $\tilde{\theta}$, which alternatively can be written as

$$\hat{\theta} = g^*(\tilde{\theta})\,. \tag{2.20}$$

When either there is a closed form solution for the ML estimator but not for its bias function or there is no closed form solution for neither the ML estimator nor its bias function, $B_{\hat{\theta}}(\theta)$ is approximated at $\tilde{\theta}$ through parametric bootstrap. Kuk (1995) independently produced the same idea for solving equation (2.19) by iteratively adjusting the estimator $\tilde{\theta}$ for reducing the bias in the estimation of generalised linear models with random effects.

In fact, Kuk (1995) describes a general method of adjusting any suitably defined initial estimator $\hat{\theta}$, where $\theta = (\theta_1, \ldots, \theta_p)^\intercal$ is a vector of unknown parameters, through iterative bias correction, to yield an estimator which is asymptotically unbiased and consistent.

The method of Kuk (1995) can be summarized as follows: suppose that $\hat{\theta}$ is some initial estimator of θ, not necessarily the ML estimator, obtained as the solution of a set of p estimating equations of the form

$$S(\theta) \equiv S(\theta; y) = 0\,. \tag{2.21}$$

Let the bias of $\hat{\theta}$ be given by

$$B_{\hat{\theta}}(\theta) = g(\theta) - \theta\,, \tag{2.22}$$

where $g(\theta) = \theta^*$ is defined implicitly by

$$\mathrm{E}_\theta\{S(\theta^*, Y)\} = 0\,. \tag{2.23}$$

To correct for the bias of $\hat{\theta}$, let $B^{(0)} = 0$ be an initial estimate of the bias of $\hat{\theta}$. Define

$$B^{(k+1)} = g(\hat{\theta} - B^{(k)}) - (\hat{\theta} - B^{(k)}) \tag{2.24}$$

as an updated estimate of the bias and

$$\tilde{\theta}^{(k+1)} = \hat{\theta} - B^{(k+1)} \tag{2.25}$$

the updated bias-corrected estimate of θ. Assuming that the limit of $B^{(k)}$ exists, we can let $k \to \infty$ in equation (2.24) to obtain

$$B = g(\tilde{\theta}) - (\hat{\theta} - B) \tag{2.26}$$

so that

$$\hat{\theta} = g(\tilde{\theta})\,. \tag{2.27}$$

The function $g(\theta) = \theta^*$, however, has no closed form solution in general, so Kuk (1995) proposes to approximate $g(\theta)$ by $g_R(\theta)$ where

$$g_R(\theta) = \frac{1}{R}\sum_{i=1}^{R} \hat{\theta}(y_i) \tag{2.28}$$

is the average of $\hat{\theta}$ over simulated samples and where $y_1, \ldots, y_R$ are simulated from the model with the parameters set at θ. Substituting g_R for g in equations (2.24) and (2.25),

we obtain

$$B_R^{(k+1)} = g_R(\hat{\theta} - B_R^{(k)}) - (\hat{\theta} - B_R^{(k)}) \tag{2.29}$$

as the Monte Carlo estimate of the bias of $\hat{\theta}$ at the $(k+1)$th iteration and

$$\tilde{\theta}^{(k+1)} = \hat{\theta} - B_R^{(k+1)} \tag{2.30}$$

the updated bias-corrected estimate of θ.

Suppose now that $g(\theta) = g^*(\theta)$ where $\hat{\theta}$ is the ML estimator of θ. In this case, observe that equations (2.24) become

$$B^{(k+1)} = \mathrm{E}_{\hat{\theta}-B^{(k)}}(\hat{\theta}) - (\hat{\theta} - B^{(k)}). \tag{2.31}$$

Letting $k \to \infty$ in (2.31) we therefore obtain

$$B = \mathrm{E}_{\tilde{\theta}}(\hat{\theta}) - (\hat{\theta} - B) \tag{2.32}$$

so that

$$\hat{\theta} = \mathrm{E}_{\tilde{\theta}}(\hat{\theta}). \tag{2.33}$$

So equation (2.20) that we aim to solve is simply equation (2.27) of Kuk (1995) where $g(\theta)$ is set to be $\mathrm{E}_{\theta}(\hat{\theta})$.

Consider now the case where the model parameter $\theta^{\mathsf{T}} = (\psi, \lambda^{\mathsf{T}})$ is partitioned into a scalar parameter of interest ψ and a vector of nuisance parameters λ, and we are interested in adjusting the bias of the ML estimator of ψ. We propose below a simulation-based algorithm for indirect inference estimation of the parameter of interest, ψ, motivated by that of Kuk (1995) above. Let $g(\psi) = \mathrm{E}_{\psi}(\hat{\psi})$ be approximated by

$$g_R(\psi) = \frac{1}{R}\sum_{i=1}^{R} \hat{\psi}(z_i) \tag{2.34}$$

where $z_1, \ldots, z_R$ are simulated from the model with the parameters set at $\theta = (\psi, \hat{\lambda})$. Replacing θ by ψ and $g_R(\theta)$ by $g_R(\psi)$ in equations (2.29) and (2.30), we obtain

$$\begin{aligned} B_R^{(k+1)} &= g_R(\hat{\psi} - B_R^{(k)}) - (\hat{\psi} - B_R^{(k)}) \\ &= g_R(\tilde{\psi}^{(k)}) - \tilde{\psi}^{(k)} \end{aligned} \tag{2.35}$$

as the Monte Carlo estimate of the bias of $\hat{\psi}$ at the $(k+1)$th iteration and

$$\begin{aligned}\tilde{\psi}^{(k+1)} &= \hat{\psi} - B_R^{(k+1)} \\ &= \hat{\psi} - g_R(\tilde{\psi}^{(k)}) + \tilde{\psi}^{(k)} \\ &= \hat{\psi} + \tilde{\psi}^{(k)} - \frac{1}{R}\sum_{i=1}^{R}\hat{\psi}(z_i) \end{aligned} \tag{2.36}$$

the updated bias-corrected estimate of ψ, where $z_1,\ldots,z_R$ are simulated from the model with the parameters set at $\theta = (\tilde{\psi}^{(k)}, \hat{\lambda})$. All that is required for the implementation of (2.36) is a routine for sampling from the probability model, and a routine for calculating the ML estimates of ψ and λ.

Note that if the initial estimate of the bias $B^{(0)}$ is set to zero then the initial estimate of $\tilde{\psi}$, $\tilde{\psi}^{(0)}$ coincides with $\hat{\psi}$. Note also how when g is set to be the expectation function, iterative bias correction simply reduces to iterative correction of the estimate $\tilde{\psi}$.

2.2.4 Firth's bias-reducing adjusted score equations

Firth (1993) explored an approach to bias reduction of the overall ML estimator by deriving an adjusted score function and hence showed that an estimator with $O(n^{-2})$ bias is obtained by solving an adjusted score equation in the form

$$S^*(\tilde{\theta}) = S(\tilde{\theta}) + A(\tilde{\theta}) = 0\,, \tag{2.37}$$

where $A(\theta)$ is allowed to depend on the data. Firth (1993) described two different alternatives of $A(\theta)$, denoted by $A^{(E)}(\theta)$ and $A^{(O)}(\theta)$, based on the expected and observed information matrix, respectively. The components of these two alternatives are given in matrix notation in Kosmidis and Firth (2010) and take the form

$$A_t^{(E)}(\theta) = A_t(\theta) \qquad (t = 1,\ldots,p)\,, \tag{2.38}$$

and

$$A_t^{(O)}(\theta) = j_t(\theta)\{i(\theta)\}^{-1}A^{(E)}(\theta) \qquad (t = 1,\ldots,p)\,, \tag{2.39}$$

where the $1 \times p$ vector $j_t(\theta)$ denotes the t-th row of $j(\theta)$ and where $A_t(\theta)$ is as given in (2.16).

The reason that this is an implicit method of bias reduction can be seen if equation

(2.37) is rewritten as

$$S(\tilde{\theta}) - \{i(\tilde{\theta})\}\frac{b_1(\tilde{\theta})}{n} = 0,$$

or equivalently as

$$\{i(\tilde{\theta})\}^{-1} S(\tilde{\theta}) = \frac{b_1(\tilde{\theta})}{n}, \tag{2.40}$$

which reveals that $\tilde{\theta}$ is an approximate solution to equation (2.12) because $B_{\hat{\theta}}(\theta)$ is approximated by $b_1(\theta)/n$ evaluated at $\theta = \tilde{\theta}$ and $\{i(\tilde{\theta})\}^{-1} S(\tilde{\theta})$ is the first term in the asymptotic expansion of $\hat{\theta} - \theta$ evaluated at $\theta = \tilde{\theta}$.

Note that (2.38) is simply the expected value of (2.39). Note also that in the case of an exponential family in canonical parametrization the observed information matrix $j(\theta)$ is independent of the data, so the $A^{(E)}(\theta)$ and $A^{(O)}(\theta)$ adjustments coincide.

2.2.5 Empirical bias-reducing adjusted estimating functions

All the methods of bias reduction discussed so far approximate the bias term in (2.12) either analytically or through simulation and assume either full or partial specification of the assumed underlying model G. Methods like asymptotic bias correction and Firth's bias-reducing adjusted score equation (Firth, 1993) approximate $B_{\hat{\theta}}(\theta)$ analytically and require access to log-likelihood derivatives and the computation of expectations of products of those under the assumed partial or full model G, respectively. For models with intractable or cumbersome likelihoods, these expectations are intractable or expensive to compute, and can be hard to derive even for simple models. On the other hand, indirect inference approximate the bias term by simulating samples from the assumed model and as a result, simulation-based methods are typically more computationally intensive than analytical methods. Finally, all the bias reduction methods reviewed so far, except for Firth's bias reducing adjusted scores approach, require the original estimator $\hat{\theta}$ and consequently they directly inherit any of the instabilities that $\hat{\theta}$ may have. For example, in logistic regression (Albert and Anderson, 1984) the maximum likelihood estimates may be infinite due to data separation. Then, simulation based methods of bias reduction like indirect inference cannot be applied because even if data separation did not occur for the original sample, there is always a positive probability that it will occur for at least one of the simulated samples.

Suppose that we observe the values $y_1, \ldots, y_n$ of a sequence of random vectors $Y_1, \ldots, Y_n$ with $y_i = (y_{i1}, \ldots, y_{ic_i})^\intercal \in \mathscr{Y} \subset \mathbb{R}^c$, with a sequence of covariate vectors $x_1, \ldots, x_n$ with

$x_i = (x_{i1}, \ldots, x_{iq_i})^\intercal \in \mathscr{X} \subset \mathbb{R}^q$. Let $Y = (Y_1^\intercal, \ldots, Y_n^\intercal)^\intercal$, and denote by X the set of $x_1, \ldots, x_n$. Suppose that we want to estimate the unknown parameter vector $\theta \in \Theta \subset \mathbb{R}^p$ using data $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$ through a vector of p estimating functions $\sum_{i=1}^n \omega^i(\theta) = \left(\sum_{i=1}^n \omega_1^i(\theta), \ldots, \sum_{i=1}^n \omega_p^i(\theta) \right)^\intercal$, where $\omega^i(\theta) = \omega(\theta, y_i, x_i)$ and $\omega_r^i(\theta) = \omega_r(\theta, y_i, x_i)$, $(r = 1, \ldots, p)$. The M-estimator $\hat{\theta}$ of θ results by the solution of the system of estimating equations

$$\sum_{i=1}^n \omega^i(\theta) = 0_p, \tag{2.41}$$

with respect to θ, where 0_p is a p-vector of zeros.

Assume for simplicity, the stronger modelling assumption that Y_i has a distribution function $F_i(y_i|x_i, \theta)$ and that the estimator $\hat{\theta}$ is the ML estimator which is the maximiser of the objective function $l(\theta) = \sum_{i=1}^n \ln f_i(y_i|x_i, \theta)$, where $f_i(y_i|x_i, \theta)$ is the joint density corresponding to $F_i(y_i|x_i, \theta)$. Then the estimating equations in (2.41) become

$$\sum_{i=1}^n \nabla \ln f_i(y_i|x_i, \theta) = 0_p, \tag{2.42}$$

assuming that the gradient exists in Θ.

Kosmidis and Lunardon (2020) derive an implicit reduced bias M-estimator (iRBM), $\tilde{\theta}$, with $O(n^{-3/2})$ bias under a set of assumptions (Kosmidis and Lunardon, 2020, §2.2) that results from the implicit solution of the empirical adjusted estimating equations

$$\sum_{i=1}^n \omega^i(\theta) + A(\theta) = 0_p, \tag{2.43}$$

where both $A(\theta) = A(\theta, Y, X)$ and its derivatives with respect to θ are $O_p(1)$ as n grows where n is a measure of information about θ, which is typically, the number of observations. Assuming that $Y_1, \ldots, Y_n$ are independent, the matrix form of the rth element of the vector of empirical bias-reducing adjustments reduces to

$$A_r(\theta) = -\text{trace}\{j(\theta)^{-1} d_r(\theta)\} - \frac{1}{2}\text{trace}\left[j(\theta)^{-1} e(\theta) \{j(\theta)^{-1}\}^\intercal u_r(\theta)\right], \tag{2.44}$$

where $j(\theta)$ is the matrix with sth row $-\sum_{i=1}^n \nabla \omega_s^i(\theta)$, $s = (1, \ldots, p)$, assumed invertible but not necessarily symmetric, $e(\theta) = \sum_{i=1}^n \{\omega^i(\theta)\}\{\omega^i(\theta)\}^\intercal$. In other words, the (s,t)th

element of $e(\theta)$ is

$$[e(\theta)]_{st} = \sum_{i=1}^{n} \omega_s^i(\theta)\omega_t^i(\theta), \tag{2.45}$$

$u_r(\theta) = \sum_{i=1}^n \nabla\nabla^\intercal \omega_r^i(\theta)$ and $d_r(\theta) = \sum_{i=1}^n \{\nabla\omega_r^i(\theta)\}\omega^i(\theta)$. In other words, the (s,t)th element of $d_r(\theta)$ is

$$[d_r(\theta)]_{st} = \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial\theta^s}\omega_r^i(\theta) \right\} \omega_t^i(\theta). \tag{2.46}$$

Note that $j(\theta)$, $e(\theta)$, $u_r(\theta)$ and $d_r(\theta)$ are all $p \times p$ matrices.

In the case of maximum likelihood estimation, where $\omega(\theta)$ is the gradient of the log-likelihood function, the above components of (2.44) reduce to $j(\theta) = -\sum_{i=1}^n \nabla\nabla^\intercal l_i(\theta)$, the observed information matrix, $e(\theta) = \sum_{i=1}^n \{\nabla l_i(\theta)\}\{\nabla^\intercal l_i(\theta)\}$, the matrix of outer products of log-likelihood derivatives, $u_r(\theta) = \sum_{i=1}^n \nabla\nabla^\intercal\{\partial l_i(\theta)/\partial\theta^r\}$, the hessian of the rth derivative of the log-likelihood and $d_r(\theta) = \sum_{i=1}^n \{\nabla(\partial l_i(\theta)/\partial\theta^r)\}\{\nabla l_i(\theta)\}$.

The iRBM-estimator $\tilde{\theta}$ is such that

$$\{i(\theta)\}^{1/2}(\tilde{\theta} - \theta) \xrightarrow{d} N_p(0_p, I_p), \tag{2.47}$$

when the M-estimator $\hat{\theta}$ is the ML estimator, where $i(\theta)$ is the expected (Fisher) information and I_p is the $p \times p$ identity matrix. As a result, the iRBM-estimator is asymptotically efficient, exactly as the ML and Firth's bias-reducing adjusted scores estimators are.

When estimation is through the maximisation of the log-likelihood function, Kosmidis and Lunardon (2020) (see §6.1) show that the empirical bias-reducing adjustments (2.44) is formally equivalent to the maximisation of a penalised log-likelihood function of the form

$$l(\theta) - \frac{1}{2}\text{trace}\{j(\theta)^{-1}e(\theta)\}, \tag{2.48}$$

assuming that the maximum exists. Note that the derivative of the penalty term above with respect to (the rth parameter of) θ is the $A_r(\theta)$ term in (2.44). Note also that since estimation is through the maximisation of the log-likelihood function, the iRBM-estimates can be computed using general numerical optimisation procedures for the maximisation of the penalised log-likelihood function (2.48), like those provided by the *optim* or *optimx* functions in R.

Under the same assumptions of §2.2, Kosmidis and Lunardon (2020) (see §12) give an explicit reduced bias M-estimator (eRBM-estimator) with $o(n^{-1})$ bias in general M-

estimation problems defined explicitly as

$$\theta^{\dagger} = \hat{\theta} + j(\hat{\theta})^{-1} A(\hat{\theta}), \tag{2.49}$$

where $\hat{\theta}$ is the M-estimator and where the rth component of $A(\theta)$ is as in expression (2.44) for the empirical bias-reducing adjustments. In contrast to the iRBM-estimator $\tilde{\theta}$ which requires only the estimating function contributions and their first derivatives, the eRBM-estimator $\theta^{\dagger}$ requires also the second derivatives of the estimating function contributions.

Chapter 3

Heckit (Tobit II) selection model

3.1 Introduction

Tobit models refer to regression models in which the range of the dependent variable is censored in some way. In a censored regression model only the value of the dependent variable is unknown, more specifically, it is only partially known whether it is above or below a fixed threshold or the value of another random variable, i.e. censored, while the value for the independent variable is still available. This is in contrast to truncated regression models which arise in cases where observations with values in the dependent variable below or above certain thresholds are excluded from the sample. This means that neither the dependent nor the independent variable is known so that whole observations are missing. Tobit models are censored regression models and are distinct from truncated regression models. A leading model used to deal with censored data is the Tobit I model proposed by Tobin (1958) which is a special case of the censored regression model when the dependent variable is censored from below at a threshold of zero. Tobin (1958) used this model to analyze household expenditure (purchases) on durable goods against income where the expenditure (the dependent variable) cannot be negative.

Ignoring censoring will generally lead to inconsistent and biased estimators. In the example above, because the expenditure is observed only when it is positive, conventional regression methods like ordinary least squares (OLS) estimators for the relationship between household expenditure and income are downwardly biased. Numerous applications of the Tobit I model have appeared in the economics literature such as the number of extramarital affairs (Fair, 1958) where the independent variables are sex age, number of years married, number of children, education, occupation and degree of religiousness.

Another example is the number of arrests (or convictions) per month after release from prison (Witte, 1980) where the independent variables are accumulated work release funds, number of months after release until first job, wage rate after release, age, race and drug use.

The classical Tobit I regression model describes the relationship between a censored continuous dependent variable and a vector of independent variables, where the statistical inference mainly focuses on the estimation of the regression parameter β and the variance σ^2 of the error term. Since the likelihood function of the Tobit I model is the product of the likelihood function of the probit regression model for binary responses relating to the dependent variable (Cameron and Trivedi, 2005, §14.3) and the likelihood function of the truncated normal regression model (Greene, 2004, §19.2), one can maximize the logarithm of the first of the two products to obtain the probit maximum likelihood estimator of β/σ (Amemiya, 1984, §4.1) which is consistent and asymptotically normally distributed. The maximization must be done by an iterative scheme such as Newton-Raphson (Amemiya, 1981, p.1495). However, one can only estimate the ratio β/σ by this method and not the regression parameter and standard deviation separately. Hence the probit ML estimator is not fully efficient because it ignores a part of the likelihood function that involves β and σ. The Tobit ML estimator however, defined as a solution of equating the score equations to zero using the full likelihood function of the Tobit I model, is strongly consistent and asymptotically normal (Amemiya, 1973). The score equations are nonlinear in the parameters and so must be solved iteratively. Amemiya (1973) also showed that the Tobit I likelihood function is not globally concave with respect to the original parameters β and σ^2. However, Olsen (1978) proved after a certain reparametrization the global concavity of the log-likelihood in the Tobit I model, which implies that a standard iterative method such as Newton-Raphson or the method of Fisher scoring always converges to the global maximum of the log-likelihood function. A major weakness of the Tobit ML estimator is its heavy reliance on distributional assumptions. If the underlying disturbances of the regression are either non normal or heteroskedastic (meaning that the variability of the disturbances is unequal across the range of values of the dependent variable), then the Tobit ML estimator is inconsistent (see Maddala, 1983, for a comprehensive discussion of Tobit models).

Variations of the Tobit I model can be produced by changing where and when censoring occurs. Amemiya (1984, p.30) and Amemiya (1985, p.384) classified these variations into five basic types (Tobit I - Tobit V), according to the form of the likelihood function.

Censored regression is a special case of a general problem known as sample selection. This means that observational studies are rarely based on pure random samples. Instead, a sample is, intentionally or unintentionally based in part on values taken by a dependent variable. Such samples are broadly defined as selected samples.

Sample selection bias may arise in particular for two reasons. First, there may be self selection, with the outcome of interest determined in part by individual choice of whether or not to participate in the activity of interest. Second, it can result from sample selection, with those who participate in the activity of interest being deliberately oversampled by analysts. In either case, parameter estimates may be inconsistent unless corrective measures are taken because consistency relies on relatively strong distributional assumptions (see Cameron and Trivedi, 2005, Chapter 16 for a a good description of Tobit and selection models). In the model of expenditures on durable goods described above, the consumer simultaneously decided whether or not to purchase a certain good and how much to spend on it. Alternatively, we could assume that these decisions are taken sequentially. First, the individual chooses whether or not to purchase the good. Subsequently the consumer determines how much they will spend. This formulation generalizes the simple Tobit I model by introducing a censoring latent variable that differs from the latent variable generating the outcome of interest so the resulting model comprises a participation (decision) equation and a resultant outcome equation. This model is referred to as the Tobit II or Heckit model.

There are many examples of self selection bias. One observes labor market wages for working women whose market wage exceeds the reservation wage at zero hours of work (Mroz, 1987). If the presence of children affects the work decision but does not affect market wages, regression evidence from selected samples of working women that women with children earn lower wages is not necessarily evidence that there is market discrimination against such women or that women with lower work experience earn less. Similarly, the wages of migrants do not, in general, afford a reliable estimate of what non migrants would have earned had they migrated. Comparisons of the wages of migrants with the wages of non migrants results in a biased estimate of the effect of migration.

As in the Tobit I model, consistent estimation of the ratio of the regression parameter and standard deviation of the participation equation in the Heckit model can be obtained by maximizing the probit part of the log-likelihood function. This estimator is also not fully efficient. Similarly, OLS leads to inconsistent estimation of the regression parameters of the participation and outcome equations unless the errors of the two

regression equations are uncorrelated. The most popular solutions for sample selection problems are based on Heckman (1976) where a two-step estimator, called the Heckit estimator, is proposed by combining a probit maximum likelihood procedure and a simple linear regression procedure. The method of Heckman (1976) was originally designed for the Tobit III model but it also applies to Tobit I - Tobit V, with some adjustments (Amemiya, 1984). The resulting estimators of the Heckman two-step method are consistent and asymptotically normally distributed (see Amemiya, 1984; Heckman, 1979, for a proof of the consistency and asymptotic normality of the Heckit estimators in the Tobit 1 model). When compared to the ML estimator which is also consistent and asymptotically normal, the Heckit estimator is simple to implement and is more robust in certain circumstances. For example, it remains consistent when the joint normality of the errors of the regression equations is not satisfied so it requires distributional assumptions weaker than that required for the ML estimator. However, the ML estimator enjoys greater efficiency compared to the Heckit estimator and the latter estimator of the square of the correlation coefficient, ρ^2, is not bounded by zero and one, compared to the limit of $[0,1]$ of the maximum likelihood estimator of ρ^2.

In this chapter we apply the indirect inference method (Kuk, 1995) and empirical bias reducing adjustments method (Kosmidis and Lunardon, 2020) to the special Heckit (Tobit II) model. We evaluate the performance of these estimation techniques through simulation and contrast them with the standard ML estimation and the Heckman two-step estimation procedure.

The chapter is organized as follows. In Section 3.2, we describe the Heckit model in full generality and review how the log-likelihood function is derived. We also review the maximum likelihood and Heckman two-step estimation methods in Section 3.3. In Section 3.4 we describe why the implementation of the Firth (1993) bias reduction method is not possible for this model and derive the necessary expressions for the implementation of the RBM-estimation. A simulation study is included in Section 3.5 and in Section 3.6 we apply the results to analyze a real data set of labor supply and compare the results of typically used estimation techniques to those from bias reduction.

3.2 Description of the Heckit model

Consider a random sample of n pairs of observations (y_i^S, y_i^O), $(i = 1, \ldots, n)$, where the first observation is binary, representing some sort of participation (or selection), and the

second observation represents the resulting outcome (Greene, 2012, Chapter 19, §19.3.4 and §19.5). Such a sample can be modelled by a selection equation and an outcome equation where the selection equation is (see Toomet and Henningsen, 2008, §2.1)

$$y_i^{S*} = \beta^\intercal x_i + \varepsilon_i^S, \tag{3.1}$$

$$y_i^S = \begin{cases} 0 & \text{if } y_i^{S*} < 0 \\ 1 & \text{if } y_i^{S*} \geq 0 \end{cases}. \tag{3.2}$$

In the above expressions, y_i^{S*} is the latent (unobserved) variable of the selection tendency for individual i, $\beta^\intercal = (\beta_1, \ldots, \beta_p)$ is a $1 \times p$ vector of parameters, x_i is a $p \times 1$ vector of explanatory regressors (independent variables), $\varepsilon_i^S \sim \mathcal{N}(0,1)$ and y_i^S is the observed binary value. Note that only the sign of y_i^{S*} is observed. The outcome equation is, then, given by

$$y_i^{O*} = \gamma^\intercal z_i + \varepsilon_i^O, \tag{3.3}$$

$$y_i^O = \begin{cases} 0 & \text{if } y_i^S = 0 \\ y_i^{O*} & \text{if } y_i^S = 1, \end{cases} \tag{3.4}$$

where y_i^{O*} is the latent outcome for individual i, $\gamma^\intercal = (\gamma_1, \ldots, \gamma_q)$ is a $1 \times q$ vector of parameters, z_i is a $q \times 1$ vector of explanatory regressors, $\varepsilon_i^O \sim \mathcal{N}(0, \sigma^2)$ and y_i^O is the observed outcome. Note that we observe the outcome only if the latent selection variable y_i^{S*} is positive. We assume the error terms follow a bivariate normal distribution with zero mean and correlation ρ

$$\begin{pmatrix} \varepsilon^S \\ \varepsilon^O \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix} \right), \tag{3.5}$$

where $\rho\sigma$ is the covariance. The variance of ε^S is set to be 1 because only the sign of y^{S*} is observed so the variance of ε^S cannot be estimated. The above formulation shows that we have a linear model with additive errors for each latent variable. The $n \times p$ regressor matrix X with x_i in its ith row and $n \times q$ matrix Z with z_i in its ith row are assumed to be of full rank (i.e. there is no exact linear relationship among any of the independent variables in the model) so that if all data were available, the parameters of each linear equation could be estimated by least squares. Note that X and Z may or may not have the

same columns. Note also that the Tobit I model is the special case $y^{S*} = y^{O*}$.

The likelihood function for the data is built up from two parts. The first part consists of the product of the probability of selection and the density conditional on selection for observations with $y_i^S = 1$. The second part is simply the probability of non selection for observations with $y_i^S = 0$. The Heckit model therefore has the likelihood function

$$\begin{aligned} L(\beta,\gamma,\sigma^2,\rho) &= \prod_{i=1}^{n} \left[f(y_i^O|y_i^S=1)\Pr(y_i^S=1)\right]^{y_i^S} \left[f(y_i^O|y_i^S=0)\Pr(y_i^S=0)\right]^{1-y_i^S} \\ &= \prod_{i=1}^{n} \left[f(y_i^O|y_i^{S*}\geq 0)\Pr(y_i^{S*}\geq 0)\right]^{y_i^S} \left[f(y_i^O|y_i^{S*}<0)\Pr(y_i^{S*}<0)\right]^{1-y_i^S} \\ &= \prod_{i=1}^{n} \left[f(y_i^O|y_i^{S*}\geq 0)\Pr(y_i^{S*}\geq 0)\right]^{y_i^S} \left[\Pr(y_i^{S*}<0)\right]^{1-y_i^S}, \end{aligned} \tag{3.6}$$

where the last equality follows since $f(y_i^O = 0|y_i^{S*} < 0) = 1$. The second product in (3.6) is simply $\Pr(y_i^{S*} < 0) = \Phi(-\beta^\intercal x_i)$, where $\Phi(\alpha)$ is the distribution function of the standard normal variable evaluated at α. The first product in (3.6) can be rewritten as

$$\begin{aligned} f(y_i^O|y_i^{S*}\geq 0)\Pr(y_i^{S*}\geq 0) &= \int_0^\infty f(y_i^O, y_i^{S*})\, dy_i^{S*} \\ &= \int_0^\infty f(y_i^{S*}|y_i^O) f(y_i^O)\, dy_i^{S*} \\ &= \int_0^\infty f(y_i^{S*}|y_i^{O*}=y_i^O) f(y_i^{O*}=y_i^O)\, dy_i^{S*}. \end{aligned} \tag{3.7}$$

Both densities in (3.7) have a normal distribution where $y_i^{O*} \sim \mathcal{N}(\gamma^\intercal z_i, \sigma^2)$ and $y_i^{S*}|y_i^{O*} = y_i^O \sim \mathcal{N}(\beta^\intercal x_i + \sigma^{-1}\rho(y_i^O - \gamma^\intercal z_i), 1-\rho^2)$ (see Greene, 2012, Appendix B, §B.9). The density $f(y_i^{O*} = y_i^O)$ goes outside the integral in (3.7) since it is independent of y_i^{S*} and can be written as $\sigma^{-1}\phi\left(\sigma^{-1}[y_i^O - \gamma^\intercal z_i]\right)$, where $\phi(\alpha)$ is the density function of the standard normal variable evaluated at α. Then, the integral $\int_0^\infty f(y_i^{S*}|y_i^{O*} = y_i^O)\, dy_i^{S*}$ is easily derived by using the substitution

$$t = \frac{y_i^{S*} - \beta^\intercal x_i - \sigma^{-1}\rho(y_i^O - \gamma^\intercal z_i)}{\sqrt{1-\rho^2}}, \tag{3.8}$$

and exploiting the definition of the distribution function $\Phi(\alpha)$ as an integral. The likeli-

hood about $\theta = (\beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_q, \sigma^2, \rho)$ reduces to

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \left[\Phi(-\beta^\intercal x_i)\right]^{1-y_i^S} \\ &\times \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i^O - \gamma^\intercal z_i)^2}{2\sigma^2}\right) \Phi\left(\frac{\beta^\intercal x_i + \sigma^{-1}\rho(y_i^O - \gamma^\intercal z_i)}{\sqrt{1-\rho^2}}\right)\right]^{y_i^S}, \end{aligned} \tag{3.9}$$

and, hence, the log-likelihood of the Heckit model is (see, Toomet and Henningsen, 2008, §2.1)

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \left\{ (1-y_i^S) \ln \Phi(-\beta^\intercal x_i) - \frac{1}{2} y_i^S \ln(2\pi) - \frac{1}{2} y_i^S \ln(\sigma^2) - \frac{y_i^S (y_i^O - \gamma^\intercal z_i)^2}{2\sigma^2} \right. \\ &+ \left. y_i^S \ln \Phi\left(\frac{\beta^\intercal x_i + \sigma^{-1}\rho(y_i^O - \gamma^\intercal z_i)}{\sqrt{1-\rho^2}}\right) \right\}. \end{aligned} \tag{3.10}$$

3.3 Review of point estimation of the model parameters

3.3.1 Maximum likelihood estimator

Since the Heckit model is fully parametric, it is straightforward to construct the maximum likelihood estimator. Let

$$m_1(\alpha) = \frac{\phi(\alpha)}{1 - \Phi(\alpha)}, \tag{3.11}$$

$$m_2(\alpha) = \frac{\phi(\alpha)}{\Phi(\alpha)}, \tag{3.12}$$

where m_1 and m_2 are known as inverse Mills ratios. Moreover, for simplicity in the forthcoming derivations, let

$$a_i = \beta^\intercal x_i, \tag{3.13}$$

$$b_i = \frac{\beta^\intercal x_i + \sigma^{-1}\rho(y_i^O - \gamma^\intercal z_i)}{\sqrt{1-\rho^2}}. \tag{3.14}$$

Then direct differentiation of the log-likelihood function (3.10) with respect to β, γ, σ and ρ yields the score equations

$$\nabla_\beta l(\theta) = X^\intercal \left[\frac{1}{\sqrt{1-\rho^2}} M_2 Y^S - M_1\left(\mathbf{1}_n - Y^S\right)\right], \tag{3.15}$$

$$\nabla_\gamma l(\theta) = Z^\intercal \left[\frac{1}{\sigma^2} \bar{Y}^O Y^S - \frac{\rho}{\sigma\sqrt{1-\rho^2}} M_2 Y^S \right], \tag{3.16}$$

$$\frac{\partial}{\partial \sigma} l(\theta) = \mathbf{1}_n^\intercal \left[-\frac{\rho}{\sigma^2\sqrt{1-\rho^2}} \bar{Y}^O M_2 Y^S - \frac{1}{\sigma} Y^S + \frac{1}{\sigma^3} (\bar{Y}^O)^2 Y^S \right], \tag{3.17}$$

$$\frac{\partial}{\partial \rho} l(\theta) = \frac{1}{\sigma(1-\rho^2)^{3/2}} \mathbf{1}_n^\intercal \left[\sigma\rho a M_2 Y^S + \bar{Y}^O M_2 Y^S \right], \tag{3.18}$$

where $\nabla_\beta l(\theta) = (\partial l(\theta)/\partial\beta_1, \ldots, \partial l(\theta)/\partial\beta_p)$ is the gradient of $l(\theta)$ with respect to β and similarly for $\nabla_\gamma l(\theta)$. The $(p+q+2)\times 1$ score equation is given by $S(\theta) = \left(\nabla_\beta l, \nabla_\gamma l, \partial l/\partial\sigma, \partial l/\partial\rho\right)^\intercal$.

Here, $\mathbf{1}_n$ is an n-vector of ones, Y^S is the $n\times 1$ vector of $Y_1^S, \ldots, Y_n^S$, $\bar{Y}^O = \text{diag}\{r_1, \ldots, r_n\}$ with $r_i = y_i^O - \gamma^\intercal z_i$, $i = 1, \ldots, n$, $a = \text{diag}\{a_1, \ldots, a_n\}$, $M_1 = \text{diag}\{m_1(a_1), \ldots, m_1(a_n)\}$ and $M_2 = \text{diag}\{m_2(b_1), \ldots, m_2(b_n)\}$. In obtaining these gradients and derivatives we use the results

$$\nabla_\beta a_i = x_i, \tag{3.19}$$

$$\nabla_\beta b_i = \frac{x_i}{\sqrt{1-\rho^2}}, \tag{3.20}$$

$$\nabla_\gamma b_i = -\frac{\rho z_i}{\sigma\sqrt{1-\rho^2}}, \tag{3.21}$$

$$\frac{\partial b_i}{\partial \sigma} = -\frac{\rho}{\sigma^2\sqrt{1-\rho^2}} \left(y_i^O - \gamma^\intercal z_i\right), \tag{3.22}$$

$$\frac{\partial b_i}{\partial \rho} = \frac{\rho\sigma a_i + \left(y_i^O - \gamma^\intercal z_i\right)}{\sigma(1-\rho^2)^{3/2}}. \tag{3.23}$$

These score equations are non linear and must be solved numerically, using for example the *selection* function from sampleSelection R package (see Toomet and Henningsen, 2008). Since the log-likelihood function of this model is not globally concave, one should use a good choice of initial values for the Newton-Raphson algorithm (see Toomet and Henningsen, 2008, §6.1), otherwise the algorithm may not converge or it may converge to a local instead of a global maximum.

The probit maximum likelihood estimation method for the Heckit model is derived by maximizing the probit part of the log-likelihood function. i.e. by maximizing the

logarithm of

$$L_p(\beta) = \prod_{i=1}^{n} \left[Pr(y_i^{S*} \geq 0)\right]^{y_i^S} \left[\Pr(y_i^{S*} < 0)\right]^{1-y_i^S}. \tag{3.24}$$

This yields the probit ML estimator of β which is consistent but inefficient. Note that in general, maximising the logarithm of the above likelihood yields the probit ML estimator of the scaled version of β, β/σ, however, we assume that the variance of the error term in the selection equation, ε^S, is one. The probit ML estimator of β is used as a first step in the Heckman's two step estimation method (see Section 3.3.2) which was proposed since maximum likelihood estimation was difficult to implement back then and which provides a complete set of consistent estimators of the model parameters.

3.3.2 Heckman's two step (Heckit) estimator

Since the Heckit model is by construction linear for the selection and outcome variables, one might consider ordinary least squares regression of the observed outcome Y^O on Z. This, however, leads to inconsistent parameter estimates of γ since the conditional mean $\mathrm{E}(Y^O|Z,X,Y^S=1)$ differs from $Z\gamma$ because $\mathrm{E}(\varepsilon^O|Z,X,Y^S=1)$ is non zero which is a necessary assumption for linear regression. Nonetheless, the expression for the conditional mean of $Y^O$was used by Heckman (1979) to motivate an alternative estimation procedure now known as Heckman's two step estimation method.

The conditional mean that applies to the observations in our sample may be written as (see Greene, 2012, §19.5 for a derivation which uses Theorem 19.5, p. 913 for the conditional moments of incidentally truncated bivariate normal distribution)

$$\begin{aligned}
\mathrm{E}(y_i^O|y_i^{S*}>0) &= \mathrm{E}(y_i^O|\varepsilon_i^S > -\beta^\intercal x_i) \\
&= \gamma^\intercal z_i + \mathrm{E}(\varepsilon_i^O|\varepsilon_i^S > -\beta^\intercal x_i) \\
&= \gamma^\intercal z_i + \rho\sigma m_1(-\beta^\intercal x_i) \\
&= \gamma^\intercal z_i + \rho\sigma m_2(\beta^\intercal x_i).
\end{aligned} \tag{3.25}$$

Alternatively, observe that $\varepsilon_i^O|\varepsilon_i^S \sim \mathcal{N}(\rho\sigma\varepsilon_i^S, \sigma^2(1-\rho^2))$ so $\varepsilon_i^O = \rho\sigma\varepsilon_i^S + \xi_i$, where $\xi_i \sim \mathcal{N}(0, \sigma^2(1-\rho^2))$ is independent of ε_i^S. This means that $\mathrm{E}(\varepsilon_i^O|\varepsilon_i^S > -\beta^\intercal x_i) = \rho\sigma\mathrm{E}(\varepsilon_i^S|\varepsilon_i^S > -\beta^\intercal x_i)$ so the truncated moments of the standard normal distribution can be used to find the latter expectation (see Cameron and Trivedi, 2005, §16.3.4 and §16.5.3).

It is now clear why OLS regression of Y^O on Z above leads to inconsistent estimation of γ as the expression (3.25) for the conditional mean includes the additional regressor $m_1(-X\beta)$, or equivalently since $\rho\sigma$ is not equal to zero (by assumption). Using Theorem 19.5 of Greene (2012) also yields the conditional variance

$$\begin{aligned} \text{Var}(y_i^O|y_i^{S*}>0) &= \text{Var}(y_i^O|\varepsilon_i^S > -\beta^\intercal x_i) \\ &= \text{Var}(\varepsilon_i^O|\varepsilon_i^S > -\beta^\intercal x_i) \\ &= \sigma^2\big(1-\rho^2\delta(-\beta^\intercal x_i)\big), \end{aligned} \tag{3.26}$$

where $\delta(a_i) = dm_1(a_i)/d(a_i) = m_1(a_i)[m_1(a_i)-a_i]$ is the derivative of the inverse Mills ratio m_1. The conditional variable $y_i^O|y_i^{S*}>0$ may therefore be written as

$$y_i^O|y_i^{S*}>0 = \gamma^\intercal z_i + \rho\sigma m_2(\beta^\intercal x_i) + \nu_i, \tag{3.27}$$

where ν_i is an error term with $\text{E}(\nu_i)=0$ and $\text{Var}(\nu_i) = \sigma^2\big(1-\rho^2\delta(-\beta^\intercal x_i)\big)$.

Note that even if we observe $\beta^\intercal x_i$ and hence $m_2(\beta^\intercal x_i)$, entering $m_2(X\beta)$ as a regressor in (3.27) would lead to unbiased but inefficient least squares estimators of γ and $\rho\sigma$. The inefficiency is a consequence of the heteroscedasticity of the error ν_i apparent from (3.26). Heckman's two step method can now be defined and it proceeds as follows

1. Obtain the probit maximum likelihood estimator of β, $\hat{\beta}_{prob}$, and use this estimator to compute $m_2(\hat{\beta}_{prob}^\intercal x_i)$ which is an estimate of the inverse Mills ratio.
2. Estimate γ and $\rho\sigma$ by least squares regression of Y^O on Z and $m_2(X\hat{\beta}_{prob})$.

The estimator $\hat{\beta}_{prob}$ and the OLS estimators of γ and $\rho\sigma$ from step 2, called the Heckit estimators, are all consistent, where consistent estimators of the individual parameters ρ and σ have also been derived in the literature (see Greene, 2012, p. 916). This method can be easily implemented using the *heckit* function from the sampleSelection R package.

Consistent estimators of the individual parameters ρ and σ can be obtained using

$$\hat{\sigma}^2 = \left(\frac{1}{n}\sum_{i=1}^n \hat{\nu}_i{}^2\right) + \left(\frac{1}{n}\sum_{i=1}^n \hat{\delta}_i\right)\left(\hat{\beta}^m\right)^2, \tag{3.28}$$

where $\hat{\nu}$ is the vector of residuals from the OLS estimation of (3.27), $\hat{\delta}_i = \hat{\delta}(\beta^\intercal x_i) = \hat{m_2}(\beta^\intercal x_i)\big[\hat{m_2}(\beta^\intercal x_i) + \hat{\beta}^\intercal x_i\big]$, and $\hat{\beta}^m = \widehat{\rho\sigma}$. A consistent estimator of the correlation between ε^S and ε^O can be obtained by $\hat{\rho} = \hat{\beta}^m/\hat{\sigma}$. Note that $\hat{\rho}$ can be outside the interval $[-1,1]$.

A consistent estimate of the variance-covariance matrix can be obtained by (see Greene, 2004, §19.5.3)

$$\widehat{\mathrm{Var}}\left[\hat{\gamma}, \hat{\beta}^{m}\right] = \hat{\sigma}^{2}\left[Z_{m}^{\intercal} Z_{m}\right]^{-1}\left[Z_{m}^{\intercal}\left(I_{n} - \hat{\rho}^{2}\hat{\Delta}\right) Z_{m} + R\right]\left[Z_{m}^{\intercal} Z_{m}\right]^{-1}, \tag{3.29}$$

where

$$R = \hat{\rho}^{2}\left(Z_{m}^{\intercal}\hat{\Delta}X\right)\widehat{\mathrm{Var}}\left[\hat{\beta}\right]\left(X^{\intercal}\hat{\Delta}Z_{m}\right), \tag{3.30}$$

where X is the matrix of all observations of x, Z_m is the matrix of all observations of z and $\hat{m}_2$, I_n is an identity matrix, $\hat{\Delta}$ is a diagonal matrix with ith diagonal element $\hat{\delta}_i$, and $\widehat{\mathrm{Var}}\left[\hat{\beta}\right]$ is the estimated variance-covariance matrix of the probit estimate of β. Any of the observed or expected information matrices of the probit likelihood could be used to compute the latter variance. Note that in R, the variance-covariance matrix of the heckit two-step estimators, using the *heckit* function, is only partially implemented with NAs in place of the unimplemented components. The unimplemented components are those of σ and ρ.

3.4 Implicit bias reduction methods

3.4.1 Firth's adjusted score equations method

The popular estimator of Firth (1993) for reducing the bias of the ML estimator of θ may be obtained through the solution of the adjusted score equation

$$S^{*}(\theta) = S(\theta) + A(\theta) = 0, \tag{3.31}$$

where $A(\theta)$ is $O_p(1)$ in magnitude as $n \to \infty$ and where the two alternatives of $A(\theta)$ described by Firth and are given by (2.38) and (2.39).

Differentiation of (3.15), (3.16), (3.17) and (3.18) yields the second order partial derivatives and Hessian matrices

$$\nabla_{\beta}\nabla_{\beta}^{\intercal} l(\theta) = X^{\intercal}\left\{ -\frac{1}{1-\rho^{2}} U M_{2}^{'} - (I_{n} - U) M_{1}^{'} \right\} X, \tag{3.32}$$

where $U = \mathrm{diag}\{Y_1^S, \ldots, Y_n^S\}$, $M_2^{'} = M_2(M_2 + b)$ where $b = \left[a + (\rho \bar{Y}^{O})/\sigma\right]/\sqrt{1-\rho^2}$, I_n is the $n \times n$ identity matrix and where $M_1^{'} = M_1(M_1 - a)$.

$$\nabla_\beta \nabla_\gamma^\intercal l(\theta) = \frac{\rho}{\sigma(1-\rho^2)} X^\intercal U M_2^{'} Z, \tag{3.33}$$

$$\frac{\partial}{\partial \sigma} \nabla_\beta l(\theta) = \frac{\rho}{\sigma^2(1-\rho^2)} X^\intercal M_2^{'} \bar{Y}^O Y^S, \tag{3.34}$$

$$\frac{\partial}{\partial \rho} \nabla_\beta l(\theta) = \frac{1}{(1-\rho^2)^{3/2}} X^\intercal \Big\{ -\frac{1}{\sigma\sqrt{1-\rho^2}} M_2^{'}(\rho\sigma a + \bar{Y}^O) Y^S + \rho M_2 Y^S \Big\}, \tag{3.35}$$

$$\nabla_\gamma \nabla_\gamma^\intercal l(\theta) = Z^\intercal U \Big\{ -\frac{1}{\sigma^2} I_n - \frac{\rho^2}{\sigma^2(1-\rho^2)} M_2^{'} \Big\} Z, \tag{3.36}$$

$$\frac{\partial}{\partial \sigma} \nabla_\gamma l(\theta) = Z^\intercal \Big\{ -\frac{2}{\sigma^3} \bar{Y}^O Y^S - \frac{\rho^2}{\sigma^3(1-\rho^2)} M_2^{'} \bar{Y}^O Y^S + \frac{\rho}{\sigma^2\sqrt{1-\rho^2}} M_2 Y^S \Big\}, \tag{3.37}$$

$$\frac{\partial}{\partial \rho} \nabla_\gamma l(\theta) = \frac{1}{\sigma(1-\rho^2)^{3/2}} Z^\intercal \Big\{ \frac{\rho}{\sigma\sqrt{1-\rho^2}} M_2^{'}(\rho\sigma a + \bar{Y}^O) Y^S - M_2 Y^S \Big\}, \tag{3.38}$$

$$\begin{aligned} \frac{\partial^2}{\partial(\sigma)^2} l(\theta) &= \mathbf{1}_n^\intercal \Big\{ -\frac{\rho^2}{\sigma^4(1-\rho^2)} M_2^{'} (\bar{Y}^O)^2 Y^S \\ &+ \frac{2\rho}{\sigma^3\sqrt{1-\rho^2}} \bar{Y}^O M_2 Y^S + \frac{1}{\sigma^2} Y^S - \frac{3}{\sigma^4} (\bar{Y}^O)^2 Y^S \Big\}, \end{aligned} \tag{3.39}$$

$$\frac{\partial^2}{\partial\rho\partial\sigma} l(\theta) = \mathbf{1}_n^\intercal \Big\{ \frac{\rho}{\sigma^3(1-\rho^2)^2} \bar{Y}^O M_2^{'}(\rho\sigma a + \bar{Y}^O) Y^S - \frac{1}{\sigma^2(1-\rho^2)^{3/2}} \bar{Y}^O M_2 Y^S \Big\}, \tag{3.40}$$

$$\frac{\partial^2}{\partial\rho^2} l(\theta) = \mathbf{1}_n^\intercal \Big\{ \frac{1}{\sigma(1-\rho^2)^{5/2}} \big[\sigma(2\rho^2+1)a + 3\rho\bar{Y}^O\big] M_2 Y^S - \Big(\frac{\sigma\rho a + \bar{Y}^O}{\sigma(1-\rho^2)^{3/2}} \Big)^2 M_2^{'} Y^S \Big\}. \tag{3.41}$$

In obtaining these derivatives and Hessian matrices we frequently used the following results

$$\frac{d\phi(\alpha)}{d\alpha} = -\alpha\phi(\alpha), \tag{3.42}$$

$$\delta_1(\alpha) = \frac{dm_1(\alpha)}{d\alpha} = m_1(\alpha)\big[m_1(\alpha) - \alpha\big], \tag{3.43}$$

$$\delta_2(\alpha) = \frac{dm_2(\alpha)}{d\alpha} = -m_2(\alpha)\big[m_2(\alpha) + \alpha\big], \tag{3.44}$$

$$(\nabla_\beta M_2) Y^S = -\frac{1}{\sqrt{1-\rho^2}} M_2^{'} U X, \tag{3.45}$$

$$(\nabla_\gamma M_2)Y^S = \frac{\rho}{\sigma\sqrt{1-\rho^2}} M_2' UZ, \tag{3.46}$$

$$\frac{\partial}{\partial\sigma} M_2 = \frac{\rho}{\sigma^2\sqrt{1-\rho^2}} M_2' \bar{Y}^O, \tag{3.47}$$

$$\frac{\partial}{\partial\rho} M_2 = -\frac{1}{\sigma(1-\rho^2)^{3/2}} M_2' (\rho\sigma a + \bar{Y}^O), \tag{3.48}$$

$$(\nabla_\beta M_1)(\mathbf{1}_n - Y^S) = M_1'(I_n - U)X, \tag{3.49}$$

$$(\nabla_\gamma \bar{Y}^O)Y^S = -UZ, \tag{3.50}$$

and where in (3.45), (3.46), (3.49) and (3.50) we made use of the relation $\text{diag}\{a\}b = \text{diag}\{b\}a$, where a and b are arbitrary $n \times n$ vectors.

The $(p+q+2)\times(p+q+2)$ observed information matrix becomes

$$j(\theta) = \begin{pmatrix} j_{\beta\beta} & j_{\beta\gamma} & j_{\beta\rho} & j_{\beta\sigma} \\ (j_{\beta\gamma})^\intercal & j_{\gamma\gamma} & j_{\gamma\rho} & j_{\gamma\sigma} \\ (j_{\beta\rho})^\intercal & (j_{\gamma\rho})^\intercal & j_{\rho\rho} & j_{\rho\sigma} \\ (j_{\beta\sigma})^\intercal & (j_{\gamma\sigma})^\intercal & (j_{\rho\sigma})^\intercal & j_{\sigma\sigma} \end{pmatrix}, \tag{3.51}$$

where $j_{\beta\beta} = -\nabla_\beta \nabla_\beta^\intercal l$, $j_{\beta\gamma} = -\nabla_\beta \nabla_\gamma^\intercal l$, $j_{\beta\rho} = -\partial(\nabla_\beta l)/\partial\rho$, $j_{\beta\sigma} = -\partial(\nabla_\beta l)/\partial\sigma$, $j_{\gamma\gamma} = -\nabla_\gamma \nabla_\gamma^\intercal l$, $j_{\gamma\rho} = -\partial(\nabla_\gamma l)/\partial\rho$, $j_{\gamma\sigma} = -\partial(\nabla_\gamma l)/\partial\sigma$, $j_{\rho\rho} = -\partial^2 l/\partial\rho^2$, $j_{\rho\sigma} = -\partial^2 l/\partial\rho\partial\sigma$, $j_{\sigma\sigma} = -\partial^2 l/\partial(\sigma)^2$.

To obtain the Fisher information matrix, $i(\theta)$, we need to calculate the expectation of the second order partial derivatives and gradients in $j(\theta)$. The expectations involved in the derivation of $i(\theta)$, $P_t(\theta)$ and $Q_t(\theta)$ are not all available in closed form and hence the method of Firth cannot be applied to this model. However, we can get access to a couple of these expectations using the score equations and the second and third Bartlett identities. We list below the analytic expressions for the expectations that are derived as solutions of various second and third Bartlett identities, and those as solutions of score equations, followed by a list of the remaining required expectations which are not available in closed form. We explain the requirements for the derivation of these unavailable expectations, which if obtained, make the Firth method applicable to this model.

1. The score equation $\mathrm{E}\big(\nabla_\beta l\big) = 0$, yields

$$\mathrm{E}\big(M_2 Y^S\big) = \sqrt{1-\rho^2}\, Q' \mathbf{1}_n, \tag{3.52}$$

where $Q^{'} = \text{diag}\{\phi(a_1), \ldots, \phi(a_n)\}$.

2. The score equation $\mathrm{E}\big(\partial l/\partial\rho\big) = 0$, yields

$$\mathrm{E}\big(\bar{Y}^{O} M_2 Y^{S}\big) = -\sigma\rho\sqrt{1-\rho^2}\, a Q^{'} \mathbf{1}_n. \tag{3.53}$$

3. The second Bartlett identity $-\mathrm{E}\big(\nabla_\beta \nabla_\beta^{\intercal} l\big) = \mathrm{E}\big[\big(\nabla_\beta l\big)\big(\nabla_\beta^{\intercal} l\big)\big]$, yields

$$\mathrm{E}\big(M_2 b Y^{S}\big) = (1-\rho^2) a Q^{'} \mathbf{1}_n. \tag{3.54}$$

4. The second Bartlett identity $-\mathrm{E}\big[\partial\big(\nabla_\beta l\big)/\partial\rho\big] = \mathrm{E}\big[\big(\nabla_\beta l\big)\big(\partial l/\partial\rho\big)\big]$, yields

$$\mathrm{E}\big(\bar{Y}^{O} M_2 b Y^{S}\big) = \rho\sigma(1-\rho^2)(I_n - a^2) Q^{'} \mathbf{1}_n. \tag{3.55}$$

5. The score equation $\mathrm{E}\big(\nabla_\gamma l\big) = 0$, yields

$$\mathrm{E}\big(\bar{Y}^{O} Y^{S}\big) = \sigma\rho Q^{'} \mathbf{1}_n. \tag{3.56}$$

6. The score equation $\mathrm{E}\big(\partial l/\partial\sigma\big) = 0$, yields

$$\mathrm{E}\big[(\bar{Y}^{O})^2 Y^{S}\big] = \sigma^2\big[Q - \rho^2 a Q^{'}\big]\mathbf{1}_n, \tag{3.57}$$

where $Q = \text{diag}\{\Phi(a_1), \ldots, \Phi(a_n)\}$.

7. The second Bartlett identity $-\mathrm{E}\big[\partial\big(\nabla_\beta l\big)/\partial\sigma\big] = \mathrm{E}\big[\big(\nabla_\beta l\big)\big(\partial l/\partial\sigma\big)\big]$, yields

$$\mathrm{E}\big[(\bar{Y}^{O})^2 M_2 Y^{S}\big] = \sigma^2\sqrt{1-\rho^2}\big[I_n - \rho^2(I_n - a^2)\big] Q^{'} \mathbf{1}_n. \tag{3.58}$$

8. The second Bartlett identity $-\mathrm{E}\big[\partial\big(\nabla_\gamma l\big)/\partial\sigma\big] = \mathrm{E}\big[\big(\nabla_\gamma l\big)\big(\partial l/\partial\sigma\big)\big]$, yields

$$\mathrm{E}\big[(\bar{Y}^{O})^3 Y^{S}\big] = \sigma^3\rho\big[3 - \rho^2(I_n - a^2)\big] Q^{'} \mathbf{1}_n. \tag{3.59}$$

9. The second Bartlett identity $-\mathrm{E}\big(\partial^2 l/\partial\rho^2\big) = \mathrm{E}\big[(\partial l/\partial\rho)^2\big]$, yields

$$\mathrm{E}\big[(\bar{Y}^{O})^2 M_2 b Y^{S}\big] = \sigma^2(1-\rho^2) a\big[I_n - \rho^2(3I_n - a^2)\big] Q^{'} \mathbf{1}_n. \tag{3.60}$$

10. The simultaneous solution of the two second Bartlett identities $-\mathrm{E}\big(\partial^2 l/\partial(\sigma)^2\big) =$

$\mathrm{E}\big[(\partial l/\partial\sigma)^2\big]$ and $-\mathrm{E}\big(\partial^2 l/\partial\sigma\partial\rho\big) = \mathrm{E}\big[(\partial l/\partial\sigma)(\partial l/\partial\rho)\big]$, yields

$$\mathrm{E}\big[(\bar{Y}^O)^3 M_2 Y^S\big] = \rho\sigma^3 a\sqrt{1-\rho^2}\big[\rho^2(3I_n - a^2) - 3I_n\big]Q'\mathbf{1}_n, \tag{3.61}$$

$$\mathrm{E}\big[(\bar{Y}^O)^4 Y^S\big] = \sigma^4\big[3Q - \rho^2 a\big[6I_n - \rho^2(3I_n - a^2)\big]Q'\big]\mathbf{1}_n. \tag{3.62}$$

11. The third Bartlett identity
$\mathrm{E}\big[\nabla_\beta \nabla_\beta^\intercal (\nabla_\beta l)\big] + 3\mathrm{E}\big[(\nabla_\beta \nabla_\beta^\intercal l)(\nabla_\beta l)\big] + \mathrm{E}\big[(\nabla_\beta l)(\nabla_\beta^\intercal l)(\nabla_\beta l)\big] = 0$, yields

$$\mathrm{E}\big(M_2 b^2 Y^S\big) = \sqrt{1-\rho^2}\big[\rho^2 + a^2(I_n - \rho^2)\big]Q'\mathbf{1}_n. \tag{3.63}$$

12. The third Bartlett identity $\mathrm{E}\big[\partial(\nabla_\beta \nabla_\beta^\intercal l)/\partial\rho\big] + \mathrm{E}\big[(\nabla_\beta \nabla_\beta^\intercal l)(\partial l/\partial\rho)\big] +$
$\mathrm{E}\big[(\nabla_\beta l)(\nabla_\beta^\intercal)(\partial l/\partial\rho)\big] + 2\mathrm{E}\big[(\nabla_\beta l)\{\partial(\nabla_\beta l)/\partial\rho\}\big] = 0$, yields

$$\mathrm{E}\big(\bar{Y}^O M_2 b^2 Y^S\big) = -\rho\sigma a\sqrt{1-\rho^2}\big[\rho^2(3I_n - a^2) - (2I_n - a^2)\big]Q'\mathbf{1}_n. \tag{3.64}$$

13. The third Bartlett identity $\mathrm{E}\big[\partial^2(\nabla_\beta l)/\partial\rho^2\big] + \mathrm{E}\big[(\nabla_\beta l)(\partial^2 l/\partial\rho^2)\big] +$
$\mathrm{E}\big[(\partial l/\partial\rho)^2(\nabla_\beta l)\big] + 2\mathrm{E}\big[(\partial l/\partial\rho)\{\partial(\nabla_\beta l)/\partial\rho\}\big] = 0$, yields

$$\mathrm{E}\big[(\bar{Y}^O)^2 M_2 b^2 Y^S\big] = \sigma^2\sqrt{1-\rho^2}\big[\rho^2(1-\rho^2)\big[3I_n - a^2(6I_n - a^2)\big] + a^2\big]Q'\mathbf{1}_n. \tag{3.65}$$

14. Using the moment generating function of the random variable $y_i^O | y_i^S = 1$ it may be shown that (see Appendix A for a full derivation),

$$\mathrm{E}\big[(\bar{Y}^O)^5 Y^S\big] = \rho\sigma^5\big\{\big[15 - \rho^2(10 - 3\rho^2)\big]I_n + \rho^2 a^2\big[2(5 - 3\rho^2)I_n + \rho^2 a^2\big]\big\}Q'\mathbf{1}_n, \tag{3.66}$$

$$\begin{aligned}\mathrm{E}\big[(\bar{Y}^O)^6 Y^S\big] = \sigma^6\big\{&15Q - \rho^2 a\big[5\rho^2(3 - 2\rho^2)a^2 \\ &+ 15\big[3 - \rho^2(3 - \rho^2)\big]I_n + \rho^4 a^4\big]Q'\big\}\mathbf{1}_n.\end{aligned} \tag{3.67}$$

15. Using (3.65), (3.58) and (3.61) (see Appendix A for a full derivation) it may be shown that

$$\mathrm{E}\big[(\bar{Y}^O)^4 M_2 Y^S\big] = \sigma^4\sqrt{1-\rho^2}\big\{\rho^4\big[3I_n - a^2(6I_n - a^2)\big] - 6\rho^2(1 - a^2) + 3I_n\big\}Q'\mathbf{1}_n. \tag{3.68}$$

16. Using (3.61) and (3.68) (see Appendix A for a full derivation) it may be shown that

$$\begin{aligned}\mathrm{E}\big[(\bar{Y}^O)^3 M_2 b Y^S\big] &= \rho\sigma^3\big\{\rho^2\big[a^2(9I_n - a^2) - 6I_n\big] \\ &\quad + \rho^4\big[3I_n - a^2(6I_n - a^2)\big] + 3(I_n - a^2)\big\}Q'\mathbf{1}_n. \end{aligned} \tag{3.69}$$

17. Using the final useful third Bartlett identity $\mathrm{E}\big[\big(\partial^2 l/\partial(\sigma)^2\big)\big(\partial l/\partial\rho\big)\big] + \mathrm{E}\big[\partial^3 l/\partial(\sigma)^2\partial\rho\big] + \mathrm{E}\big[\big(\partial l/\partial\sigma\big)^2\big(\partial l/\partial\rho\big)\big] + 2\mathrm{E}\big[\big(\partial l/\partial\sigma\big)\big(\partial^2 l/\partial\sigma\partial\rho\big)\big] = 0$ and the expectations (3.65), (3.61), (3.68) and (3.69) it may be shown that (see Appendix A for a full derivation)

$$\begin{aligned}\mathrm{E}\big[(\bar{Y}^O)^5 M_2 Y^S\big] &= -\rho a\sigma^5\sqrt{1-\rho^2}\big\{\rho^4\big[15I_n - a^2(10I_n - a^2)\big] \\ &\quad - 10\rho^2(3I_n - a^2) + 15I_n\big\}Q'\mathbf{1}_n. \end{aligned} \tag{3.70}$$

18. Using (3.70) (see Appendix A) we obtain

$$\begin{aligned}\mathrm{E}\big[(\bar{Y}^O)^4 M_2 b Y^S\big] &= \sigma^4 a\big\{\rho^4\big[33I_n - a^2(16I_n - a^2)\big] - 3\rho^2(7I_n - 2a^2) \\ &\quad - \rho^6\big[15I_n - a^2(10I_n - a^2)\big] + 3I_n\big\}Q'\mathbf{1}_n. \end{aligned} \tag{3.71}$$

The remaining expectations that are not available in analytic form are of the following random variables

1. $(M_2)^2 Y^S$
2. $(M_2)^3 Y^S$
3. $(M_2)^2 b Y^S$
4. $\bar{Y}^O (M_2)^2 Y^S$
5. $(\bar{Y}^O)^2 (M_2)^2 Y^S$
6. $\bar{Y}^O (M_2)^3 Y^S$
7. $(\bar{Y}^O)^2 (M_2)^3 Y^S$
8. $(\bar{Y}^O)^3 (M_2)^2 Y^S$
9. $(\bar{Y}^O)^3 (M_2)^3 Y^S$
10. $(\bar{Y}^O)^4 (M_2)^2 Y^S$
11. $\bar{Y}^O (M_2)^2 b Y^S$
12. $(\bar{Y}^O)^2 (M_2)^2 b Y^S$
13. $(\bar{Y}^O)^3 (M_2)^2 b Y^S$

It is not necessary though to approximate all of the above 13 expectations because

some of them may be obtained from others. For example, the third expectation above is defined as a linear combination of the first and fourth expectations (see (A.17) of Appendix A). The eleventh expectation is defined as a linear combination of the fourth and fifth expectations (see (A.18) of Appendix A). The twelfth expectation is defined as a linear combination of the fifth and eighth expectations (see (A.19) of Appendix A), while the thirteenth expectation above is defined as a linear combination of the eighth and tenth expectations (see (A.20) of Appendix A).

The expectation of the remaining nine random variables involves bivariate integrals. However, using the Law of iterated expectation (see Johnston and DiNardo, 1997, Appendix B.5, or see (A.1) of Appendix A), these bivariate integrals can be reduced to a univariate integral. For example, the ith component of the first integral in the above list may be written as

$$\begin{aligned}
\mathrm{E}_{Y_i^S,Y_i^O}\left[Y_i^S m_2^2(b_i)\right] &= \mathrm{E}_{Y_i^S}\left\{\mathrm{E}_{Y_i^O|Y_i^S}\left[Y_i^S m_2^2(b_i)\right]\right\} \\
&= \Pr(Y_i^S=0)\mathrm{E}_{Y_i^O|Y_i^S=0}\left[Y_i^S m_2^2(b_i)\right] \\
&+ \Pr(Y_i^S=1)\mathrm{E}_{Y_i^O|Y_i^S=1}\left[Y_i^S m_2^2(b_i)\right] \\
&= \Phi(a_i)\mathrm{E}_{Y_i^O|Y_i^S=1}\left[m_2^2(b_i)\right] \\
&= \Phi(a_i)\int_{-\infty}^{\infty}\frac{\phi^2(b_i)}{\Phi^2(b_i)}f(y_i^O|y_i^S=1)\,dy_i^O \\
&= \frac{1}{\sigma}\int_{-\infty}^{\infty}\frac{\phi^2(b_i)}{\Phi(b_i)}\phi\left(\frac{y_i^O-\gamma^{\intercal}z_i}{\sigma}\right)dy_i^O, \qquad (3.72)
\end{aligned}$$

where the third equality above follows since $\mathrm{E}_{Y_i^O|Y_i^S=0}\left[Y_i^S m_2^2(b_i)\right]=0$. In summary, the above decomposition implies that the required expectations are of the following form: $\mathrm{E}_{Y_i^O|Y_i^S=0}\left[m_2^{d_1}(b_i)b_i^{d_2}\left(y_i^O-\gamma^{\intercal}z_i\right)^{d_3}\right]$, where d_1,d_2 and d_3 are in the range $d_1\in\{2,3\}$, $d_2\in\{0,1\}$ and $d_3\in\{1,2,3,4\}$. The integral in (3.72) involves Gaussian functions and so do the remaining integrals in the above list. No closed form, of which we are aware of, exists for these Gaussian integrals and so numerical approximation is necessary which complicates the numerical implementation of the bias reduction methods in Firth (1993). Owen (1980) provides an extensive list of Gaussian-type integrals but none of them are helpful.

3.4.2 Empirical bias-reducing penalty

In contrast to Firth (1993), reducing the bias of the ML estimator through iRBM-estimation is straightforward and equivalent to the maximisation of the penalised function (2.48), where $j(\theta)$ is derived in Section 3.4.1 and the $(p+q+2)\times(p+q+2)$ matrix $e(\theta)$ takes the form

$$e(\theta) = \begin{pmatrix} e_{\beta\beta} & e_{\beta\gamma} & e_{\beta\rho} & e_{\beta\sigma} \\ (e_{\beta\gamma})^\intercal & e_{\gamma\gamma} & e_{\gamma\rho} & e_{\gamma\sigma} \\ (e_{\beta\rho})^\intercal & (e_{\gamma\rho})^\intercal & e_{\rho\rho} & e_{\rho\sigma} \\ (e_{\beta\sigma})^\intercal & (e_{\gamma\sigma})^\intercal & (e_{\rho\sigma})^\intercal & e_{\sigma\sigma} \end{pmatrix}, \tag{3.73}$$

where

$$\begin{aligned}
e_{\beta\beta} &= \sum_{i=1}^{n} \Big(\nabla_\beta l_i(\theta)\Big)\Big(\nabla_\beta^\intercal l_i(\theta)\Big) \\
&= X^\intercal\big[(I_n-U)(M_1)^2 + (1-\rho^2)^{-1}U(M_2)^2\big]X, && (3.74)\\
e_{\beta\gamma} &= \sum_{i=1}^{n} \Big(\nabla_\beta l_i(\theta)\Big)\Big(\nabla_\gamma^\intercal l_i(\theta)\Big) \\
&= \frac{1}{\sigma^2\sqrt{(1-\rho^2)}}X^\intercal U M_2\big[\bar{Y}^O - \sigma\rho(1-\rho^2)^{-1/2}M_2\big] && (3.75)\\
e_{\beta\rho} &= \sum_{i=1}^{n} \Big(\nabla_\beta l_i(\theta)\Big)\Big(\frac{\partial}{\partial\rho} l_i(\theta)\Big) \\
&= \frac{1}{\sigma(1-\rho^2)}X^\intercal(M_2)^2 U\big[\sigma\rho a + \bar{Y}^O\big]\mathbf{1}_n, && (3.76)\\
e_{\beta\sigma} &= \sum_{i=1}^{n} \Big(\nabla_\beta l_i(\theta)\Big)\Big(\frac{\partial}{\partial\sigma} l_i(\theta)\Big) \\
&= \frac{1}{\sigma^3\sqrt{(1-\rho^2)}}X^\intercal U M_2\big[-\sigma\rho\bar{Y}^O M_2 - \sigma^2 I_n + (\bar{Y}^O)^2\big]\mathbf{1}_n, && (3.77)\\
e_{\gamma\gamma} &= \sum_{i=1}^{n} \Big(\nabla_\gamma l_i(\theta)\Big)\Big(\nabla_\gamma^\intercal l_i(\theta)\Big) \\
&= \frac{1}{\sigma^4(1-\rho^2)}Z^\intercal U\big[(1-\rho^2)(\bar{Y}^O)^2 - 2\sigma\rho\sqrt{(1-\rho^2)}\bar{Y}^O M_2 + \sigma^2\rho^2(M_2)^2\big]Z, && (3.78)
\end{aligned}$$

$$
\begin{aligned}
e_{\gamma\rho} &= \sum_{i=1}^{n}\Big(\nabla_\gamma l_i(\theta)\Big)\Big(\frac{\partial}{\partial\rho}l_i(\theta)\Big)\\
&= \frac{1}{\sigma^3(1-\rho^2)^2}Z^\intercal M_2 U\Big[\big\{\sigma\rho a+\bar{Y}^O\big\}\big\{\sqrt{(}1-\rho^2)\bar{Y}^O-\sigma\rho M_2\big\}\Big]\mathbf{1}_n, \qquad (3.79)\\
e_{\gamma\sigma} &= \sum_{i=1}^{n}\Big(\nabla_\gamma l_i(\theta)\Big)\Big(\frac{\partial}{\partial\sigma}l_i(\theta)\Big)\\
&= \frac{1}{\sigma^5(1-\rho^2)}Z^\intercal U\Big[\sigma\rho\sqrt{(}1-\rho^2)M_2\big\{\sigma^2 I_n-2(\bar{Y}^O)^2\big\}\\
&+ (1-\rho^2)\bar{Y}^O\big\{(\bar{Y}^O)^2-\sigma^2 I_n\big\}+\sigma^2\rho^2\bar{Y}^O(M_2)^2\Big]\mathbf{1}_n, \qquad (3.80)\\
e_{\rho\rho} &= \sum_{i=1}^{n}\Big(\frac{\partial}{\partial\rho}l_i(\theta)\Big)\Big(\frac{\partial}{\partial\rho}l_i(\theta)\Big)\\
&= \frac{1}{\sigma^2(1-\rho^2)^3}\mathbf{1}_n^\intercal(M_2)^2U\big[\sigma\rho a(2\bar{Y}^O+\sigma\rho a)+(\bar{Y}^O)^2\big]\mathbf{1}_n, \qquad (3.81)\\
e_{\rho\sigma} &= \sum_{i=1}^{n}\Big(\frac{\partial}{\partial\rho}l_i(\theta)\Big)\Big(\frac{\partial}{\partial\sigma}l_i(\theta)\Big)\\
&= \frac{1}{\sigma^4(1-\rho^2)}\mathbf{1}_n^\intercal M_2 U\Big[-\sigma\rho\bar{Y}^O M_2\big\{\bar{Y}^O+\sigma\rho a\big\}-\sigma^2\rho\sqrt{(}1-\rho^2)a\\
&+ \sqrt{(}1-\rho^2)\bar{Y}^O\big\{(\bar{Y}^O)^2-\sigma^2+\sigma\rho a\bar{Y}^O\big\}\Big]\mathbf{1}_n, \qquad (3.82)\\
e_{\sigma\sigma} &= \sum_{i=1}^{n}\Big(\frac{\partial}{\partial\sigma}l_i(\theta)\Big)\Big(\frac{\partial}{\partial\sigma}l_i(\theta)\Big)\\
&= \frac{1}{\sigma^6(1-\rho^2)}\mathbf{1}_n^\intercal U\Big[\sigma\rho\bar{Y}^O M_2\big\{\sigma\rho\bar{Y}^O M_2+2\sigma^2\sqrt{(}1-\rho^2)I_n-2\sqrt{(}1-\rho^2)(\bar{Y}^O)^2\big\}\\
&+ \sigma^4(1-\rho^2)I_n+(1-\rho^2)(\bar{Y}^O)^2\big\{(\bar{Y}^O)^2-2\sigma^2 I_n\big\}\Big]\mathbf{1}_n. \qquad (3.83)
\end{aligned}
$$

The value of the eRBM estimator defined in (2.49) is readily available once the empirical bias-reducing penalty term in (2.48) is calculated. All that is needed is the numerical differentiation of this penalty term and its evaluation with the inverse observed information matrix at the ML estimates.

3.4.3 Indirect inference

The indirect inference estimator, $\tilde{\theta}$, is the solution of the equation

$$\tilde{\theta} = \hat{\theta} - B_{\hat{\theta}}(\tilde{\theta}), \qquad (3.84)$$

where $B_{\hat{\theta}}(\theta) = \mathrm{E}_{\theta}(\hat{\theta}) - \theta$ is the bias of the ML estimator evaluated at θ. Using (2.28), the Monte Carlo estimate of the bias function becomes

$$B_{\hat{\theta}}(\theta) = \frac{1}{R}\sum_{r=1}^{R} \hat{\theta}\left(y^{(r)}\right) - \theta, \tag{3.85}$$

where $y^{(1)}, \ldots, y^{(R)}$ are simulated from the Heckit model with the parameters set at θ. The solution of (3.84) is obtained iteratively where the indirect inference estimate of θ at the $(k+1)$-th iteration is given by

$$\tilde{\theta}^{(k+1)} = \hat{\theta} - \frac{1}{R}\sum_{r=1}^{R} \hat{\theta}\left(y^{(r)}\right) + \tilde{\theta}^{(k)}, \tag{3.86}$$

where $y^{(1)}, \ldots, y^{(R)}$ are simulated from the Heckit model with the parameters set at $\tilde{\theta}^{(k)}$ and where the initial estimate $\tilde{\theta}^{(0)}$ is chosen to be the ML estimate. The iterative process is then repeated until the difference of the components of $\tilde{\theta}^{(k+1)}$ and $\tilde{\theta}^{(k)}$ are all less than δ in absolute value at the current estimates, where δ is a small number.

3.5 Simulation study

In order to assess the finite sample performance of the maximum likelihood, Heckman two-step (Heckit), the indirect inference, iRBM and eRBM estimators we perform a simulation study where we compare their bias, Monte Carlo simulation error (calculated by dividing the square root of the estimated variance of the estimator by the square root of the number of simulations), average estimated standard error, empirical standard error, length and coverage probability of 95% and 99% confidence intervals for different sample sizes. Specifically, we simulated 3000 data sets from the Heckit model with true intercepts $\beta_1 = 0.01$ and $\gamma_1 = 0.03$, true slopes $\beta_2 = 0.7$ and $\gamma_2 = 0.9$, true correlation coefficient $\rho = 0.2$ and true variance $\sigma^2 = 1$. The explanatory variables for the selection equation and outcome equation were generated from the standard uniform distribution. The average censoring level was 36%. Four values of the sample size, n, were considered ranging from 50 to 300.

The maximum likelihood and Heckman two-step estimation were implemented by the sampleSelection package in R using the functions *selection* and *heckit*, respectively (see Toomet and Henningsen, 2008). The likelihood maximisation was performed using the maxLik package (see Henningsen and Toomet, 2011) which uses the Newton-Raphson

algorithm. Convergence issues may appear at the boundary of the parameter space (if $|\rho| \to 1$). Also the log-likelihood function is not globally concave so the model may not converge, or it may converge to a local maximum, if the initial values are not chosen well enough. The default starting values used by *selection* for the maximum likelihood estimation are obtained by Heckman's two-step estimation of the model. In our simulation we fix the initial values to be the true parameter values. This however, does not resolve convergence issues completely and $\hat{\rho}$ can still end up at the boundary of the parameter space.

The indirect inference estimates of σ and ρ can end up outside the boundary of the parameter space, i.e. $\sigma < 0$ and $\rho \notin (-1,1)$, which will be problematic when we simulate from the Heckit model at the current indirect inference estimates. We resolve this by applying indirect inference on the reparametrisation $\ln(\sigma)$ and $\ln[(1+\rho)/(1-\rho)]$ which allows both σ and ρ to diverge to infinity so that when we transform them back we avoid their boundary values. Another problem that may occur is that the variance-covariance matrix at the current indirect inference estimates may not be positive semidefinite. We correct for this by using a modified version of (3.86):

$$\tilde{\theta}^{(k+1)} = \varepsilon \left\{ \hat{\theta} - \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}\left(y^{(r)}\right) \right\} + \tilde{\theta}^{(k)}, \tag{3.87}$$

where we multiply the first two terms in the RHS of (3.86) by a small positive constant ε. We loop over ten values of ε, where $\varepsilon = (0.5)^p$, where $p = 0, \ldots, 10$ and we stop when the resulting indirect inference estimates of σ and ρ form a positive semi definite variance covariance matrix. We set the value of Monte Carlo replicates R to be 500 and δ is set to be 0.01. The indirect inference estimates of σ and ρ in Table 3.1 are not the bias corrected ones but the transformed version of them. The standard error of the indirect inference estimates is evaluated using the inverse of the observed information matrix evaluated at the indirect inference estimates, i,e, using the Hessian matrix which is the same method used for the standard error of the maximum likelihood estimates using the Newton-Raphson algorithm. Note that the final indirect inference estimate of ρ can still end up on the boundary of the parameter space just like the maximum likelihood estimate of ρ.

The maximisation of the empirical penalised log-likelihood function was performed using the *optimx* function in R (see Nash and Varadhan, 2011). Since the maximum likelihood estimate of ρ can be on the boundary of the parameter space, i.e. close to one,

and the heckit estimate of ρ, $\hat{\rho}_{hec}$, can be greater than one, the starting values for the optimisation of the empirical log-likelihood need to be case dependent. We consider four cases:

a) $|\hat{\rho}| > 0.99$ and $|\hat{\rho}_{hec}| < 1$

b) $|\hat{\rho}| > 0.99$ and $|\hat{\rho}_{hec}| > 1$

c) $|\hat{\rho}| < 0.99$ and $|\hat{\rho}_{hec}| < 1$

d) $|\hat{\rho}| < 0.99$ and $|\hat{\rho}_{hec}| > 1$,

and we consider six different starting values:

1. The ML estimates
2. The true parameter values
3. The Heckman two-step estimates
4. The ML estimates with the ML estimate of ρ replaced with 0.9
5. The ML estimates with the ML estimate of ρ replaced with the true parameter value of ρ
6. The ML estimates with the ML estimate of ρ replaced with the heckit estimate of ρ, $\hat{\rho}_{hec}$.

When case a is satisfied, we try *optimx* with each of the starting values 2,3,4,5 and 6, when case b is satisfied, we try each of the starting values 2,4 and 5 and when either case c or d is satisfied we try only the first starting value, i.e. the ML estimates. We specified that all available (and suitable) optimisation methods are used for *optimx* and we chose the iRBM-estimates from the optimisation method that converged and satisfied the kkt1 and kkt2 conditions where kkt1 checks whether the gradient at the final parameter estimates is small and kkt2 checks whether the Hessian at the final parameter estimates is positive definite, except when some of the parameters are on the boundary in which case the Hessian is allowed to be positive semi-definite. This means that we are automatically excluding any iRBM-estimates of ρ on the boundary of the parameter space. As with the indirect inference estimator, since we are trying to reduce the bias at the σ and ρ parameterisation, in order to avoid a constrained optimisation problem ($\sigma > 0$ and $|\rho| < 1$), we transform σ to $\ln(\sigma)$ and ρ to $\ln[(1+\rho)/(1-\rho)]$ so that we optimise the empirical penalised log-likelihood function on the real line. The estimates of $\ln(\sigma)$ and $\ln[(1+\rho)/(1-\rho)]$ from *optimx* are then transformed back and the resulting final estimates are the iRBM-estimates of σ and ρ.

The eRBM-estimates are computed directly by substituting the ML estimates into (2.49) and the standard errors of the iRBM-estimates and the eRBM-estimates are evalu-

ated using the Hessian matrix.

There are cases where the standard error of the maximum likelihood, indirect inference, iRBM and eRBM estimates is not available, because it is not possible to invert the observed information matrix or the estimate of ρ is on the boundary of the parameter space or close to it. Similarly, the standard error of the Heckman two-step estimates of σ and ρ are not available for all samples because the variance covariance matrix of the two-step estimators is only partially implemented in the sampleSelection package in R and NA is returned in place of the unimplemented components as mentioned in Toomet and Henningsen (2008, p.g.7). This means that we need to compute the coverage probability and length of confidence intervals under a convention. We assume that if the standard error for a particular parameter is not available then the confidence interval covers the true parameter value and that the interval has infinite length so we compute the median length over simulated samples.

For some samples the maximum likelihood estimation may fail and as a consequence the indirect inference and eRBM estimations will also fail. Moreover, the indirect inference estimator may not converge. The iRBM estimation can also fail. These samples are excluded from the summaries for all estimators so that no estimator is using more samples than others. The maximum likelihood, indirect inference and eRBM estimators fail for 12 and 2 samples for sample sizes $n = 50$ and $n = 100$, respectively. However, the indirect inference estimator does not converge for a small proportion of the simulated data for all sample sizes. The iRBM estimator fails for 314, 139, 98 and 103 samples for sample sizes $n = 50$, $n = 100$, $n = 150$ and $n = 300$, respectively. The samples that are excluded from the summaries are all of those for which the maximum likelihood, indirect inference, iRBM and eRBM estimators failed and for which the indirect inference estimator did not converge. In total, the percentage of excluded samples for $n = 50$, $n = 100$, $n = 150$ and $n = 300$ were 13.8%, 7.9%, 7.9% and 7.0%, respectively.

Table 3.1 reports the bias and Monte Carlo simulation error of estimators, Table 3.2 reports the average estimated standard error and empirical standard error, while Tables 3.3 and 3.4 report the coverage probability and median length of confidence intervals with nominal level 95% and 99%, respectively. Overall, we observe from Table 3.1 that all estimators of β_1 and β_2 are less biased than their ML estimator, for all sample sizes. The empirical standard errors, i.e. the square root of the variance, of all estimates of β_1 and β_2 are also decreased when compared to ML estimation, for all sample sizes. Compared to the ML estimator of γ_1, the indirect inference, iRBM and eRBM estimators are all less

biased for all sample sizes too, where the empirical standard error is only slightly inflated. In general, the indirect inference and iRBM estimators perform better than ML in terms of bias for most parameters for small sample sizes, while the eRBM estimator is less biased than ML for the intercept and slope parameters for all sample sizes considered. The Heckit estimator performs poorly for γ_1 and σ in terms of bias for small sample sizes which is due to the inflated average standard errors. This can be confirmed from the histogram of the simulated values of the Heckit estimates in Figure 3.2, since both estimates of γ_1 and σ can take very large values. Surprisingly, the performance of the reparameterised indirect inference estimates of σ and ρ is superior to those from maximum likelihood estimates in terms of bias while the empirical standard error is only slightly inflated. Note that Table 3.2 shows that the average estimated standard errors of the ML, indirect inference, iRBM and eRBM estimates are not close to (substantially smaller than) the corresponding empirical standard errors, this is because the sample sizes we consider are small and so the estimated asymptotic standard errors seem to underestimate finite sample variability. The 95% and 99% coverage probabilities in Tables 3.3 and 3.4 confirm the above results, in the sense that for any given sample size and parameter, the estimator whose coverage probability is closer to the nominal value is either the Heckit, indirect inference, iRBM or the eRBM estimator. In other words, one of these estimators performs better than ML for all sample sizes and all parameters considered. Moreover, the 95% coverage probability of the eRBM estimator is closer to the nominal level than the ML estimator for all parameters and all sample sizes. The 95% and 99% coverage rates are slightly lower than the nominal level since the empirical standard errors (i.e. the square root of the variance) exceed the average estimated standard errors (i.e. the latter are anti-conservative). The 95% and 99% coverage probability of the Heckit estimates of γ_1 are larger than the nominal level for all sample sizes, since the estimated standard errors are large. Moreover, the large Monte Carlo simulation error is due to the relatively small simulation size of 3000 data sets. It was not feasible to use more samples in this simulation study because of the expensive computation time. In particular, on a given Mac computer, the time taken to conduct inference for a single dataset with the ML, Heckit, indirect inference, iRBM and eRBM methods was 21 seconds, 15 seconds, 2-24 minutes, 5 minutes and 26 seconds, respectively. The computation time for the indirect inference method can range from 2 to 24 minutes depending on the number of iterations required before convergence, where 100 iterations took around 24 minutes for completion. This means that the precision of the conclusions made here is a result of the limited size of the simulation study and a

larger scale simulation would be beneficial.

One important note to add here is that the variable $m_2(\beta^\top x_i)$, the inverse Mills ratio, is a nonlinear function of $\beta^\top x_i$ that can be closely approximated by a linear function of $\beta^\top x_i$ over much of its range. This implies that if there is not much variability in $\beta^\top x_i$ then $m_1(\beta^\top x_i)$ can be closely approximated by a linear function of $\beta^\top x_i$ and so there is the potential for serious multicollinearity (see Leung and Yu, 2000, for an excellent review of the collinearity problems encountered in the two-step estimation method). Hence the standard errors of the Heckit estimates depend on the variation in the latent selection equation. More variation gives smaller standard errors for the Heckit estimates. Since $x_i \sim U[0,1]$, $\beta_1 + \beta_2 x_i \sim U[\beta_1, \beta_1 + \beta_2]$. Thus, the range of $\beta_1 + \beta_2 x_i$ in our simulation study is $[0.01, 0.71]$ which is very narrow. The top plot of Figure 3.1 is that of the inverse Mills ratio over the range $[0.01, 0.71]$, which shows that $m_2(\beta^\top x_i)$ is linear. This explains the huge standard errors and hence the high coverage probability of the Heckit estimates of γ_1. We can accommodate this problem by ensuring that there is substantial variation in the explanatory variable of the selection equation. So instead of generating the explanatory variables from the standard uniform distribution as we do in our simulation study we can change the support to $[-10, 10]$. The bottom plot of Figure 3.1 is that of the inverse Mills ratio over the range $[-6.99, 7.01]$, which shows that $m_2(\beta^\top x_i)$ is nonlinear.

This means that our simulation study can be extended where in addition to considering the effect of the correlation ρ of ε^S and ε^O we can consider the effect of the variability in the explanatory variable of the selection equation, and hence the degree of multicollinearity. Moreover, the simulation study can be further extended by considering different values of the true parameters in order to study whether the results are sensitive to the degree of censoring. For example, when $n = 50$, changing the value of β_1 from 0.01 to -0.5 and keeping all other parameter values fixed, the degree of censoring changes from 36% to 55%, while changing β_1 to 1 keeping all other parameter values fixed, reduces the degree of censoring to 9%.

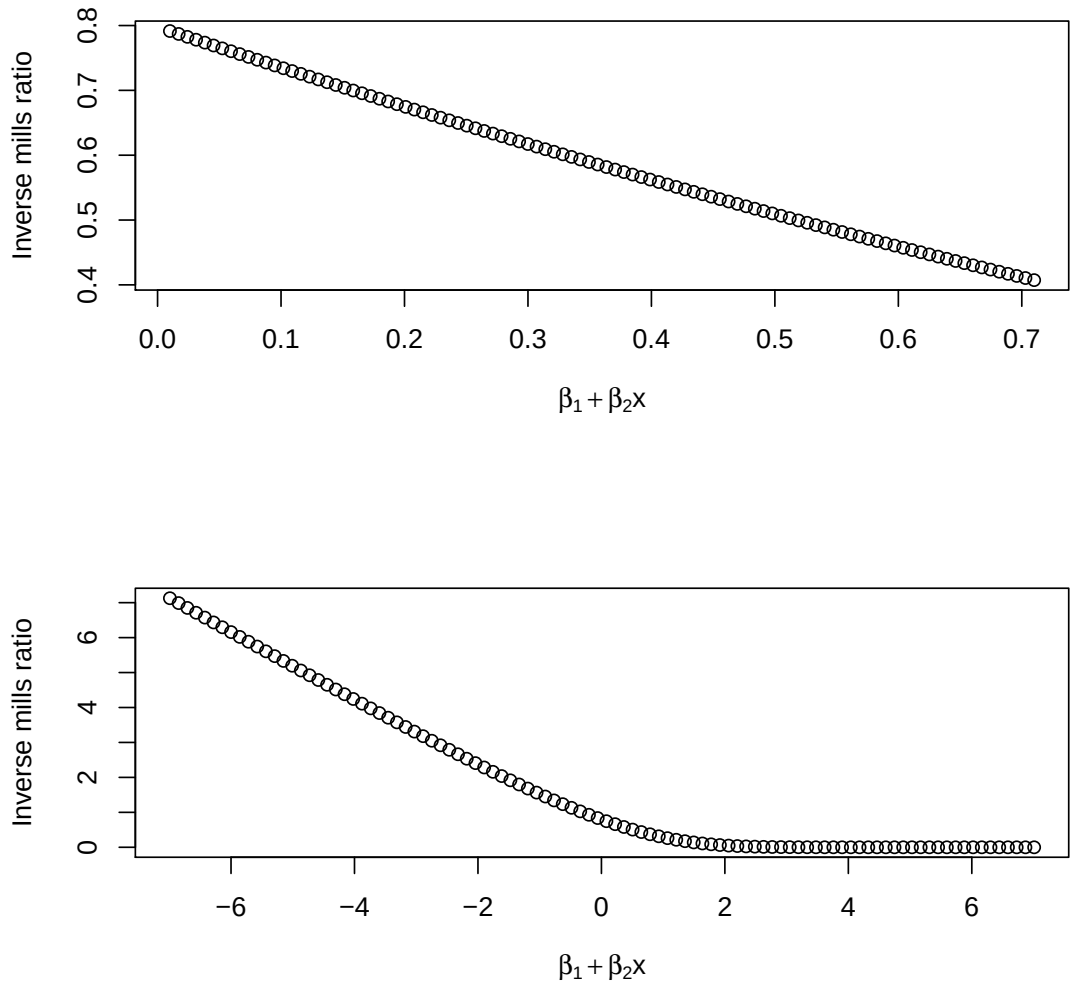

Figure 3.1: Plot of the inverse Mills ratio $m_2(\beta^\intercal x_i)$, where $\beta^\intercal x_i = \beta_1 + \beta_2 x_i$ with $\beta_1 = 0.01$ and $\beta_2 = 0.7$ and where $\beta^\intercal x_i$ varies in the range $[0.01, 0.71]$ in the top plot and $\beta^\intercal x_i$ varies in the range $[-6.99, 7.01]$ in the bottom plot.

Table 3.1: Heckit observations with true correlation coefficient $\rho = 0.2$ and true variance $\sigma^2 = 1$. The true intercepts are $\beta_1 = 0.01$ and $\gamma_1 = 0.03$, while the true slopes are $\beta_2 = 0.7$ and $\gamma_2 = 0.9$. Second to sixth columns show the estimated bias and Monte Carlo simulation error (in parentheses) of maximum likelihood, Heckman two-step (Heckit), indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$, $n = 150$ and $n = 300$, with all entries multiplied by 10.

	Maximum Likelihood	Heckman Two-Step	Indirect Inference	iRBM	eRBM
$n = 50$					
β_1	-0.384 (0.149)	-0.252 (0.147)	-0.269 (0.147)	-0.281 (0.146)	-0.314 (0.147)
β_2	1.301 (0.173)	0.837 (0.169)	0.860 (0.165)	0.580 (0.163)	0.888 (0.165)
γ_1	0.904 (0.161)	3.715 (4.663)	0.583 (0.170)	0.744 (0.155)	0.890 (0.164)
γ_2	-0.086 (0.175)	-0.052 (0.172)	0.299 (0.179)	-0.064 (0.173)	-0.079 (0.175)
σ	1.155 (0.154)	29.814 (6.063)	1.089 (0.155)	0.900 (0.150)	2.043 (0.165)
ρ	-1.598 (0.178)	-1.622 (0.213)	-1.309 (0.189)	-1.348 (0.166)	-1.610 (0.175)
$n = 100$					
β_1	-0.292 (0.131)	-0.174 (0.130)	-0.095 (0.128)	-0.155 (0.129)	-0.263 (0.131)
β_2	0.728 (0.130)	0.468 (0.129)	0.243 (0.126)	0.301 (0.125)	0.548 (0.128)
γ_1	0.546 (0.143)	-4.122 (4.649)	0.141 (0.152)	0.489 (0.143)	0.535 (0.143)
γ_2	-0.121 (0.145)	-0.125 (0.144)	0.232 (0.149)	-0.124 (0.145)	-0.127 (0.145)
σ	1.029 (0.144)	20.593 (5.382)	0.900 (0.144)	1.020 (0.145)	1.526 (0.151)
ρ	-1.012 (0.155)	-0.845 (0.183)	-0.653 (0.167)	-0.924 (0.155)	-1.026 (0.154)
$n = 150$					
β_1	-0.228 (0.127)	-0.158 (0.126)	-0.129 (0.126)	-0.105 (0.125)	-0.210 (0.127)
β_2	0.562 (0.116)	0.416 (0.116)	0.339 (0.115)	0.228 (0.113)	0.445 (0.115)
γ_1	0.366 (0.136)	-2.104 (2.056)	0.156 (0.143)	0.359 (0.137)	0.354 (0.137)
γ_2	-0.060 (0.133)	-0.101 (0.132)	0.017 (0.134)	-0.068 (0.133)	-0.059 (0.133)
σ	0.859 (0.142)	11.735 (2.657)	0.682 (0.140)	0.846 (0.142)	1.209 (0.146)
ρ	-0.768 (0.146)	-0.345 (0.165)	-0.540 (0.159)	-0.749 (0.147)	-0.780 (0.146)
$n = 300$					
β_1	-0.131 (0.121)	-0.087 (0.121)	-0.042 (0.120)	-0.048 (0.120)	-0.123 (0.121)
β_2	0.307 (0.105)	0.219 (0.105)	0.159 (0.104)	0.096 (0.103)	0.249 (0.104)
γ_1	0.252 (0.127)	-0.685 (0.260)	0.039 (0.133)	0.247 (0.127)	0.246 (0.127)
γ_2	-0.008 (0.121)	-0.021 (0.121)	0.012 (0.122)	-0.007 (0.121)	-0.008 (0.121)
σ	0.596 (0.136)	2.178 (0.297)	0.429 (0.134)	0.553 (0.136)	0.787 (0.139)
ρ	-0.566 (0.133)	-0.019 (0.142)	-0.198 (0.141)	-0.538 (0.134)	-0.573 (0.133)

Table 3.2: Heckit observations with true correlation coefficient $\rho = 0.2$ and true variance $\sigma^2 = 1$. The true intercepts are $\beta_1 = 0.01$ and $\gamma_1 = 0.03$, while the true slopes are $\beta_2 = 0.7$ and $\gamma_2 = 0.9$. Second to sixth columns show the average estimated standard error and empirical standard error (in parentheses) of maximum likelihood, Heckman two-step (Heckit), indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$, $n = 150$ and $n = 300$, with all entries multiplied by 10. The symbol __ indicates that the estimate is not available.

	Maximum Likelihood	Heckman Two-Step	Indirect Inference	iRBM	eRBM
$n = 50$					
β_1	3.638 (7.599)	3.754 (7.484)	3.550 (7.453)	3.623 (7.440)	3.549 (7.464)
β_2	6.330 (8.782)	6.520 (8.588)	6.060 (8.398)	6.229 (8.308)	6.154 (8.378)
γ_1	5.271 (8.192)	4035.060 (237.184)	3.629 (8.655)	5.059 (7.883)	4.429 (8.324)
γ_2	6.416 (8.903)	6.246 (8.734)	5.735 (9.094)	6.288 (8.813)	6.929 (8.896)
σ	2.281 (7.808)	__ (308.373)	1.367 (7.860)	2.049 (7.643)	2.875 (8.389)
ρ	5.276 (9.028)	__ (10.851)	2.300 (9.609)	5.150 (8.430)	9.046 (8.902)
$n = 100$					
β_1	2.474 (6.906)	2.591 (6.833)	2.425 (6.751)	2.482 (6.769)	2.492 (6.867)
β_2	4.384 (6.856)	4.648 (6.800)	4.253 (6.614)	4.397 (6.580)	4.469 (6.719)
γ_1	4.375 (7.497)	7790.362 (244.380)	2.809 (7.965)	3.893 (7.502)	6.053 (7.542)
γ_2	4.670 (7.617)	4.651 (7.575)	4.259 (7.816)	4.627 (7.604)	4.940 (7.611)
σ	1.774 (7.588)	__ (282.912)	1.049 (7.565)	1.515 (7.602)	2.761 (7.918)
ρ	4.718 (8.146)	__ (9.602)	1.740 (8.804)	4.006 (8.127)	6.613 (8.095)
$n = 150$					
β_1	2.034 (6.686)	2.114 (6.637)	1.980 (6.611)	2.025 (6.585)	2.070 (6.664)
β_2	3.574 (6.123)	3.750 (6.116)	3.440 (6.021)	3.549 (5.942)	3.682 (6.045)
γ_1	3.611 (7.167)	913.593 (108.057)	2.239 (7.538)	3.120 (7.176)	5.538 (7.203)
γ_2	3.671 (6.999)	3.666 (6.961)	3.379 (7.047)	3.632 (6.995)	3.865 (6.999)
σ	1.468 (7.443)	__ (139.654)	0.900 (7.375)	1.252 (7.453)	2.482 (7.679)
ρ	4.299 (7.690)	__ (8.668)	1.567 (8.342)	3.460 (7.728)	6.526 (7.660)
$n = 300$					
β_1	1.467 (6.405)	1.505 (6.377)	1.432 (6.342)	1.455 (6.342)	1.478 (6.397)
β_2	2.614 (5.540)	2.696 (5.533)	2.527 (5.482)	2.584 (5.439)	2.645 (5.506)
γ_1	2.919 (6.701)	18.519 (13.738)	1.709 (7.019)	2.377 (6.711)	4.729 (6.715)
γ_2	2.439 (6.408)	2.442 (6.401)	2.308 (6.421)	2.414 (6.409)	2.519 (6.408)
σ	1.065 (7.184)	__ (15.694)	0.696 (7.094)	0.891 (7.159)	1.769 (7.321)
ρ	3.873 (7.044)	__ (7.513)	1.516 (7.431)	2.915 (7.080)	6.472 (7.032)

Table 3.3: Heckit observations with true correlation coefficient $\rho = 0.2$ and true variance $\sigma^2 = 1$. The true intercepts are $\beta_1 = 0.01$ and $\gamma_1 = 0.03$, while the true slopes are $\beta_2 = 0.7$ and $\gamma_2 = 0.9$. Second to sixth columns show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 95% derived from Wald-type confidence intervals using the maximum likelihood, Heckman two-step, indirect inference, iRBM and eRBM estimates, for sample sizes $n = 50$, $n = 100$, $n = 150$ and $n = 300$ with all coverage probabilities multiplied by 100.

	Maximum Likelihood	Heckman Two-Step	Indirect Inference	iRBM	eRBM
$n = 50$					
β_1	91.8 (1.435)	94.6 (1.460)	95.5 (1.459)	94.9 (1.441)	94.1 (1.436)
β_2	91.1 (2.467)	94.2 (2.518)	94.8 (2.548)	94.5 (2.476)	93.1 (2.477)
γ_1	85.4 (1.918)	99.2 (4.020)	89.0 (1.610)	92.9 (1.977)	89.4 (Inf)
γ_2	91.6 (2.460)	92.5 (2.398)	94.2 (2.517)	94.0 (2.506)	95.6 (2.843)
$n = 100$					
β_1	93.0 (0.992)	95.0 (1.013)	94.7 (0.983)	95.0 (0.995)	94.0 (0.990)
β_2	92.4 (1.760)	94.6 (1.809)	95.2 (1.735)	95.0 (1.766)	93.4 (1.758)
γ_1	84.4 (1.552)	98.9 (2.958)	79.8 (1.133)	87.9 (1.514)	94.9 (Inf)
γ_2	93.5 (1.822)	94.0 (1.814)	92.8 (1.716)	94.4 (1.834)	95.3 (1.950)
$n = 150$					
β_1	93.6 (0.813)	94.5 (0.827)	94.0 (0.794)	94.7 (0.811)	94.1 (0.809)
β_2	94.1 (1.429)	95.2 (1.463)	94.9 (1.385)	95.4 (1.425)	95.2 (1.424)
γ_1	83.4 (1.285)	99.1 (2.294)	69.7 (0.882)	83.3,(1.185)	92.7 (Inf)
γ_2	93.8 (1.430)	93.8 (1.433)	92.1 (1.334)	94.0 (1.428)	94.8 (1.491)
$n = 300$					
β_1	94.8 (0.582)	95.4 (0.589)	94.6 (0.570)	94.9 (0.580)	94.7 (0.580)
β_2	92.3 (1.037)	95.9 (1.055)	94.9 (1.007)	95.6 (1.030)	95.0 (1.034)
γ_1	82.5 (1.056)	98.4 (1.625)	61.5 (0.667)	79.1 (0.910)	88.5 (2.105)
γ_2	94.3 (0.954)	94.4 (0.955)	93.4 (0.903)	94.3 (0.945)	94.9 (0.972)

Table 3.4: Heckit observations with true correlation coefficient $\rho = 0.2$ and true variance $\sigma^2 = 1$. The true intercepts are $\beta_1 = 0.01$ and $\gamma_1 = 0.03$, while the true slopes are $\beta_2 = 0.7$ and $\gamma_2 = 0.9$. Second to sixth columns show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 99% derived from Wald-type confidence intervals using the maximum likelihood, Heckman two-step, indirect inference, iRBM and eRBM estimates, for sample sizes $n = 50$, $n = 100$, $n = 150$ and $n = 300$ with all coverage probabilities multiplied by 100.

	Maximum Likelihood	Heckman Two-Step	Indirect Inference	iRBM	eRBM
$n = 50$					
β_1	97.5 (1.886)	99.2 (1.919)	99.2 (1.918)	99.2 (1.894)	98.6 (1.887)
β_2	97.6 (3.242)	99.2 (3.309)	98.9 (3.349)	98.6 (3.254)	97.7 (3.255)
γ_1	92.9 (2.520)	99.9 (5.283)	95.3 (2.116)	97.8 (2.598)	92.7 (Inf)
γ_2	97.0 (3.233)	97.8 (3.152)	98.0 (3.308)	98.4 (3.294)	98.8 (3.736)
$n = 100$					
β_1	98.4 (1.304)	99.4 (1.331)	98.8 (1.291)	99.0 (1.308)	98.4 (1.301)
β_2	97.4 (2.312)	99.3 (2.378)	98.6 (2.280)	98.4 (2.321)	97.8 (2.311)
γ_1	92.4,(2.039)	99.9 (3.887)	90.3 (1.489)	95.7 (1.990)	97.6 (Inf)
γ_2	98.4 (2.394)	98.6 (2.385)	98.1 (2.256)	98.7 (2.411)	98.7 (2.562)
$n = 150$					
β_1	98.3 (1.068)	99.1 (1.087)	98.6 (1.044)	98.9 (1.066)	98.3 (1.064)
β_2	98.2 (1.878)	98.8 (1.923)	98.3 (1.820)	98.4 (1.872)	98.1 (1.871)
γ_1	91.1 (1.689)	100.0 (3.015)	83.2 (1.160)	91.7 (1.558)	97.0 (Inf)
γ_2	98.3 (1.880)	98.5 (1.883)	97.9 (1.754)	98.4 (1.876)	98.8 (1.959)
$n = 300$					
β_1	98.9 (0.765)	99.2 (0.775)	98.9 (0.749)	99.1 (0.763)	98.8 (0.763)
β_2	98.9 (1.363)	99.1 (1.386)	98.9 (1.324)	99.0 (1.354)	98.6 (1.359)
γ_1	90.6,(1.388)	100.0 (2.135)	75.1 (0.876)	87.6 (1.195)	94.2 (2.767)
γ_2	98.4 (1.254)	98.5 (1.255)	98.1 (1.187)	98.4 (1.242)	98.6 (1.278)

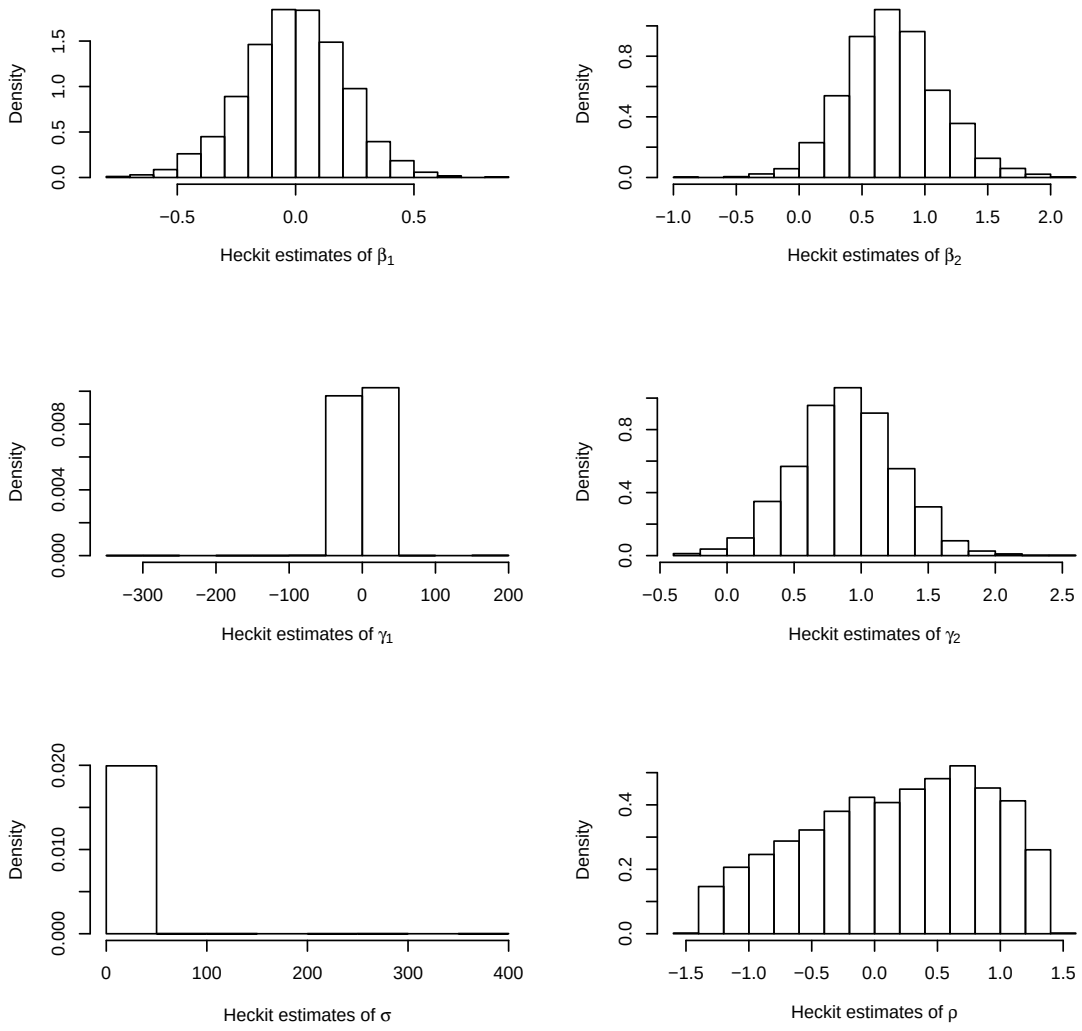

Figure 3.2: Histogram of the simulated values of the Heckit estimates for $n = 150$, after removing the excluded samples.

3.6 Analysis of female labor supply data

In this section we analyse Mroz (1987) data on female labor supply, which came from the University of Michigan Panel Study of Income Dynamics for the year 1975. The sample consists of 753 married white women between the ages of 30 and 60 in 1975, with 428 working at some time during the year, i.e. participating in the formal market (lfp $= 1$), while the remaining 325 observations are women who did not work for pay in

1975 (lfp $= 0$). The dependent variable, the wife's annual hours of work, is the product of the number of weeks the wife worked for money in 1975 and the average number of hours of work per week during the weeks she worked. The measure of the wage rate is the average hourly earnings, defined by dividing the total labor income of the wife in 1975 by the above measure of her hours of work.

Following Mroz (1987), we suppose that for these 428 individuals, the offered wage exceeded the reservation wage and, moreover, the unobserved effects in the two wage equations are correlated. As such, a wage equation based on the market data should account for the sample selection problem, due to unobservability of the wage offer for non working women. The labor force participation (described by the binary variable lfp, where lfp=1 if the individual works and 0 otherwise) equation is modelled by the quadratic polynomial (see Toomet and Henningsen, 2008, §5.1)

$$\text{lfp} = \beta_1 + \beta_2\,\text{age} + \beta_3\,\text{age}^2 + \beta_4\,\text{faminc} + \beta_5\,\text{kids} + \beta_6\,\text{educ} + \varepsilon^S, \tag{3.88}$$

where faminc is the family income in 1975 in dollars, kids is a binary variable which equals one if there are children under 18 in the household and educ is education measured in years of schooling. The wage equation is modelled by the quadratic polynomial (see Greene, 2012, Example 19.11)

$$\text{Wage} = \gamma_1 + \gamma_2\,\text{exper} + \gamma_3\,\text{exper}^2 + \gamma_4\,\text{educ} + \gamma_5\,\text{city} + \varepsilon^O, \tag{3.89}$$

where exper is labor market experience measured as the number of years the woman worked for money since her eighteenth birthday and city is a binary variable indicating that the individual lived in a large urban area.

Maximum likelihood, Heckman two-step and indirect inference estimates of the labor force participation and wage equations are shown in Tables 3.5 and 3.6 respectively. The fact that three variables are excluded from the wage offer equation is an assumption: we assume that, given the productivity factors, age, faminc and kids have no effect on the wage offer.

The differences between the maximum likelihood and indirect inference estimates in Tables 3.5 and 3.6 are small, with the estimate of ρ increasing in magnitude while that of σ decreasing for indirect inference. The differences are larger between maximum likelihood and Heckman two-step estimates as noted by Greene (2012) in Example 19.11. The indirect inference estimates were computed by using $R = 500$ Monte Carlo simulations

with $\delta = 0.001$ and the method converged after 32 iterations. Their standard errors were computed using the observed information matrix and are smaller than the standard errors from both maximum likelihood and Heckman two-step for all variables.

Note that, as mentioned in Toomet and Henningsen (2008, §5.1), the maximum likelihood and Heckman two-step estimated coefficients and standard errors are almost identical to the values published in Greene (2012), where the same maximum likelihood coefficients can be obtained with the Newton-Raphson (NR) maximisation method, while to obtain the published maximum likelihood standard errors we have to use the Berndt-Hall-Hall-Hausman (BHHH) method. This is because different ways of calculating the Hessian matrix may result in substantially different standard errors (Calzolari and Fiorentini, 1993). The NR algorithm uses exact analytic Hessian, while BHHH uses outer product approximation. In Tables 3.5 and 3.6 we use the NR method to obtain both the maximum likelihood estimates and standard errors.

The signs of the coefficients are the same across estimation methods, and the same evidence of statistical significance are found for all variables in each estimation method. Namely, the explanatory variables that have a strong effect on labor force participation are age, the presence of kids and education, while only education has a strong effect on the wage offer. The t-values of the two-sided t-tests of whether a variable is significantly different than zero were obtained by dividing the estimates by their standard errors. The p-values of the t-tests are very close in each estimation method for the labor force participation equation, being smallest for indirect inference for the age and kids variables. For example, the p-value for testing $H_0 : \beta_5 = 0$ from maximum likelihood, Heckman two-step and indirect inference are 5.68×10^{-4}, 6.38×10^{-4} and 3.61×10^{-4}, respectively. However, when testing $H_0 : \beta_6 = 0$ from maximum likelihood, Heckman two-step and indirect inference we find that the p-values are 4.30×10^{-5}, 2.19×10^{-5} and 3.91×10^{-5}, respectively being smallest for the Heckman two-step method. The p-value for testing $H_0 : \gamma_4 = 0$ is much smaller in maximum likelihood and indirect inference, being 7.33×10^{-10} and 4.36×10^{-11}, respectively, than in Heckman two-step where the p-value is only 3.56×10^{-5}. Overall, we observe that years of education has the highest effect on both labor force participation and the wage offer.

The t-test for sample selection bias in estimating the wage offer equation is based on the significance of the estimated coefficient of the inverse Mills ratio, $\widehat{\rho\sigma}$, which is equivalent to the t-test that ρ equals zero. The t-test for a sample selection problem in the Heckman two-step estimates fails to reject the hypothesis $H_0 : \rho\sigma = 0$ with p-

value 0.386. Surprisingly, the t-test for a selection effect in the maximum likelihood and indirect inference estimates, a nonzero ρ, also fails to reject the hypothesis $H_0 : \rho = 0$ with p-values 0.424 and 0.140 respectively, though the difference in the p-values is quite large.

The marginal effect of the regressors (independent variables) on y_i^O in the observed sample may be obtained by differentiating (3.25) and consists of two components. There is the direct effect on the mean of y_i^O, which is γ. In addition, for a particular independent variable, if it appears in the probability that y_i^{S*} is positive, which is given by $\Phi(\beta^\intercal x_i)$, then it will influence y_i^O through its presence in $m_1(\beta^\intercal x_i)$. The full effect of changes in a regressor that appears in both z_i and x_i on y_i^O is

$$\frac{\partial \mathrm{E}(y_i^O | y_i^{S*} > 0)}{\partial z_{ik}} = \gamma_k - \beta_k(\rho\sigma)\delta_1(-a_i), \tag{3.90}$$

where $\delta_1(-a_i) = dm_1(-a_i)/da_i = m_1(-a_i)[m_1(-a_i) + a_i]$ and where $a_i = \beta^\intercal x_i$ as in (3.13). The average marginal effect for education which appears as a regressor in both the labor force participation and wage equations is estimated as 0.480 for the maximum likelihood estimates, as 0.481 for the Heckman two-step estimates and estimated as 0.481 for the indirect inference estimates. These estimates were obtained by averaging over all observations, i.e. in (3.90) we replace a_i with $(1/n)\sum_{i=1}^n a_i$. The full effect of changes in a regressor that appears only in x_i on y_i^O is

$$\frac{\partial \mathrm{E}(y_i^O | y_i^{S*} > 0)}{\partial x_{ik}} = -\beta_k(\rho\sigma)\delta_1(-a_i). \tag{3.91}$$

Therefore, for the kids variable, which appears as a regressor only in the labor force participation equation, the average marginal effect is -0.293 for the Heckman two-step estimates and only -0.110 and -0.142 for the maximum likelihood and indirect inference estimates, respectively.

Table 3.5: Labor Force Participation Equation for Married Women. Estimated regression coefficients and standard errors (in parentheses).

	Maximum Likelihood	Heckman Two-Step	Indirect Inference
Independent Variables			
β_1	-4.120 (1.401)	-4.157 (1.402)	-4.153 (1.398)
β_2	0.184 (0.066)	0.185 (0.066)	0.187 (0.066)
β_3	-0.002 (0.001)	-0.002 (0.001)	-0.002 (0.001)
β_4	<0.001 (<0.001)	<0.001 (<0.001)	<0.001 (<0.001)
β_5	-0.451 (0.130)	-0.449 (0.131)	-0.466 (0.130)
β_6	0.095 (0.023)	0.098 (0.023)	0.095 (0.023)

Table 3.6: Wage Offer Equation for Married Women. Estimated regression coefficients and standard errors (in parentheses).

	Maximum Likelihood	Heckman Two-Step	Indirect Inference
Independent Variables			
γ_1	-1.963 (1.198)	-0.971 (2.059)	-1.837 (1.033)
γ_2	0.028 (0.062)	0.021 (0.062)	0.027 (0.059)
γ_3	<0.001 (0.002)	<0.001 (0.002)	<0.001 (0.002)
γ_4	0.457 (0.073)	0.417 (0.100)	0.452 (0.068)
γ_5	0.447 (0.316)	0.444 (0.316)	0.445 (0.304)
$(\rho\sigma)$		-1.098 (1.266)	
σ	3.108 (0.114)	3.200 (NA)	3.004 (0.103)
ρ	-0.132 (0.165)	-0.343 (NA)	-0.171 (0.116)

3.7 Discussion and further work

3.7.1 Summary

We have compared the performance of several estimators of the parameters of the Heckit (Tobit II) model with the already available ML and Heckman two-step estimation methods. We demonstrated through simulation studies that the indirect inference, iRBM and eRBM estimators reduce the bias in the ML estimator and improve the coverage probabilities of confidence intervals, for small sample sizes. The Heckman two-step estimator on the other hand, suffers from severe multicollinearity due to the fact that the inverse Mills ratio does not differ enough from a linear function, which is a consequence of the small variation in the explanatory variable of the selection equation. The collinearity problem can be alleviated by incorporating more variability in the selection equation, and it is desirable to investigate the performance of the bias reduction methods in this case and how they compare to the ML and Heckman two-step estimation methods. Our simulation study considered a single set of true parameter values but this could be easily extended to account for different values of the parameter space and how it affects estimation bias and inference. In particular, varying the true parameter values affects the degree of censoring and it is useful to extend the simulation study to account for different levels of censoring in the data. Moreover, the size of our simulation study could be considerably increased, to 10000 say, so that the Monte Carlo error is reduced, though this would be computationally expensive. Since the indirect inference estimator could in principle be applied to any well defined initial estimator, we may investigate the performance of indirect inference on the bias reduction of the Heckit estimates of the model parameters, especially to that of γ_1.

When applying bias reduction to real-world data sets, the indirect inference estimator can be expensive due to its computational complexity. However, the iRBM and eRBM estimators, on the other hand, are relatively easy to implement in general purpose packages.

3.7.2 Empirical bias-reducing penalty for Heckman two-step estimation

The Heckman two step estimator is a two step M-estimator since we can write it as a set of estimating equations. Moreover, the estimating equations are unbiased and hence we can apply iRBM-estimation to reduce the bias of the Heckman two step estimator by first applying iRBM-estimation to the probit ML estimating equation of β using the penalised log-likelihood function (2.48) and then applying iRBM-estimation to the least squares

estimating equations of γ and $\rho\sigma$ using (2.44).

The likelihood function of the probit model is given by

$$L_p(\beta) = \prod_{i=1}^{n} p_i^{y_i^S} (1-p_i)^{1-y_i^S}, \tag{3.92}$$

where $p_i = \Pr(y_i^S = 1) = \Pr(y_i^{S*} \geq 0) = \Phi(\beta^\intercal x_i)$, and the corresponding log-likelihood function is (see Cameron and Trivedi, 2005, §14.3.3)

$$l_p(\beta) = \sum_{i=1}^{n} \left\{ y_i^S \ln \Phi(\beta^\intercal x_i) + (1-y_i^S) \ln\left(1-\Phi(\beta^\intercal x_i)\right) \right\}. \tag{3.93}$$

The gradient of the probit log-likelihood function above is

$$\nabla_\beta l_p(\beta) = \sum_{i=1}^{n} \left\{ \frac{y_i^S - \Phi(\beta^\intercal x_i)}{\Phi(\beta^\intercal x_i)} x_i m_1(\beta^\intercal x_i) \right\}, \tag{3.94}$$

which is an unbiased estimating equation since $\mathrm{E}(y_i^S) = \Phi(\beta^\intercal x_i)$.

The OLS estimator of γ and $\xi = \rho\sigma$ minimises the sum of squared errors, v_i, given in (3.27) and is given by

$$\begin{aligned}
\sum_{i=1}^{n} v_i^2 &= \sum_{i=1}^{n} \left\{ y_i^O - \gamma^\intercal z_i - (\rho\sigma) m_1(-\beta^\intercal x_i) \right\}^2 \\
&= \sum_{i=1}^{n} \left\{ y_i^O - \gamma^\intercal z_i - (\rho\sigma) m_2(\beta^\intercal x_i) \right\}^2 \\
&= \sum_{i=1}^{n} \left\{ \left(y_i^O - \gamma^\intercal z_i\right)^2 - 2(\rho\sigma)\left(y_i^O - \gamma^\intercal z_i\right) m_2(\beta^\intercal x_i) \right. \\
&+ \left. (\rho\sigma)^2 \left(m_2(\beta^\intercal x_i)\right)^2 \right\},
\end{aligned} \tag{3.95}$$

where for simplicity we write y_i^O instead of $y_i^O | y_i^{S*} > 0$. The gradient with respect to γ and the partial derivative with respect to $\rho\sigma$ of (3.95), are given respectively by

$$\nabla_\gamma \left(\sum_{i=1}^{n} v_i^2 \right) = -2 \sum_{i=1}^{n} \left\{ \left(y_i^O - \gamma^\intercal z_i\right) - (\rho\sigma) m_2(\beta^\intercal x_i) \right\} z_i, \tag{3.96}$$

$$\frac{\partial}{\partial(\rho\sigma)} \left(\sum_{i=1}^{n} v_i^2 \right) = -2 \sum_{i=1}^{n} \left\{ \left(y_i^O - \gamma^\intercal z_i\right) - (\rho\sigma) m_2(\beta^\intercal x_i) \right\} m_2(\beta^\intercal x_i), \tag{3.97}$$

which are unbiased estimating equations since $\mathrm{E}(y_i^O | y_i^{S*} > 0) = \gamma^\intercal z_i + (\rho\sigma) m_2(\beta^\intercal x_i)$.

In terms of the notation of Section 2.2.5, $\theta = \big(\beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_q, \rho\sigma\big)$ so we have $p+q+1$ estimating equations given by

$$\sum_{i=1}^{n} \omega^i(\theta) = \left(\sum_{i=1}^{n} \omega_1^i(\beta), \sum_{i=1}^{n} \omega_2^i(\theta), \sum_{i=1}^{n} \omega_3^i(\theta) \right)^\intercal, \tag{3.98}$$

where

$$\sum_{i=1}^{n} \omega_1^i(\beta) = \left(\sum_{i=1}^{n} \omega_{11}^i(\beta), \ldots, \sum_{i=1}^{n} \omega_{1p}^i(\beta) \right)^\intercal, \tag{3.99}$$

$$\sum_{i=1}^{n} \omega_2^i(\theta) = \left(\sum_{i=1}^{n} \omega_{21}^i(\theta), \ldots, \sum_{i=1}^{n} \omega_{2q}^i(\theta) \right)^\intercal, \tag{3.100}$$

and where

$$\begin{aligned} \sum_{i=1}^{n} \omega_1^i(\beta) &= \sum_{i=1}^{n} \nabla_\beta l_p^i(\beta) \\ &= \nabla_\beta l_p(\beta) \\ &= X^\intercal M_1 Q^{-1} \big[(U - Q) \mathbf{1}_n \big], \end{aligned} \tag{3.101}$$

where X, M_1, U and $\mathbf{1}_n$ are as in Sections 3.3.1 and 3.4.1 and where $Q = \mathrm{diag}\{\Phi(a_1), \ldots, \Phi(a_n)\}$, $a_i = \beta^\intercal x_i$, $i = 1, \ldots, n$,

$$\begin{aligned} \sum_{i=1}^{n} \omega_2^i(\theta) &= \sum_{i=1}^{n} \nabla_\gamma v_i^2 \\ &= \nabla_\gamma \left(\sum_{i=1}^{n} v_i^2 \right) \\ &= -2 Z^\intercal \big[\bar{Y}^O - (\rho\sigma) M_2^a \big] \mathbf{1}_n, \end{aligned} \tag{3.102}$$

where $\bar{Y}^O$ is as in Section 3.3.1 and where $M_2^a = \mathrm{diag}\{m_2(a_1), \ldots, m_2(a_n)\}$,

$$\begin{aligned} \sum_{i=1}^{n} \omega_3^i(\theta) &= \sum_{i=1}^{n} \frac{\partial}{\partial(\rho\sigma)} v_i^2 \\ &= \frac{\partial}{\partial(\rho\sigma)} \left(\sum_{i=1}^{n} v_i^2 \right) \\ &= -2 \mathbf{1}_n^\intercal \big[\bar{Y}^O - (\rho\sigma) M_2^a \big] M_2^a \mathbf{1}_n. \end{aligned} \tag{3.103}$$

Since the estimation of β is through maximum likelihood, bias reduction through empirical bias-reducing adjustments is equivalent to the maximisation of the penalised function

$$l_p(\beta) - \frac{1}{2}\text{trace}\{j(\beta)^{-1}e(\beta)\}, \tag{3.104}$$

where

$$\begin{aligned}
j(\beta) &= -\sum_{j=1} \nabla_\beta \omega_1^i(\beta) \\
&= -\sum_{j=1} \nabla_\beta \nabla_\beta^\intercal l_p^i(\beta) \\
&= -\nabla_\beta \nabla_\beta^\intercal \sum_{i=1}^n l_p^i(\beta) \\
&= -\nabla_\beta \nabla_\beta^\intercal l_p(\beta) \\
&= -\sum_{i=1}^n \left\{ y_i^S x_i x_i^\intercal \left[\frac{\Phi(\beta^\intercal x_i)\delta_1(\beta^\intercal x_i) - m_1(\beta^\intercal x_i)\phi(\beta^\intercal x_i)}{\big(\Phi(\beta^\intercal x_i)\big)^2} \right] - x_i x_i^\intercal \delta_1(\beta^\intercal x_i) \right\} \\
&= X^\intercal \big[-UQ^{-2}(QM_1^{'} - M_1 Q^{'}) + M_1^{'} \big] X,
\end{aligned} \tag{3.105}$$

where $\delta_1(\beta^\intercal x_i) = m_1(\beta^\intercal x_i)\big[m_1(\beta^\intercal x_i) - \beta^\intercal x_i\big]$, $M_1^{'}$ is as in Section 3.4.1 and where $Q^{'} = \text{diag}\{\phi(a_1), \ldots, \phi(a_n)\}$,

$$\begin{aligned}
e(\beta) &= \sum_{i=1}^n \big\{\nabla_\beta l_p^i(\beta)\big\}\big\{\nabla_\beta^\intercal l_p^i(\beta)\big\} \\
&= \sum_{i=1}^n \left\{ \left[\frac{(y_i^S)^2 - 2y_i^S\Phi(\beta^\intercal x_i) + \big(\Phi(\beta^\intercal x_i)\big)^2}{\big(\Phi(\beta^\intercal x_i)\big)^2} \right] x_i x_i^\intercal \big(m_1(\beta^\intercal x_i)\big)^2 \right\} \\
&= X^\intercal M_1^2 \big[Q^{-2}U^2 - 2Q^{-1}U + I_n\big] X.
\end{aligned} \tag{3.106}$$

The $(q+1)\times(q+1)$ matrices $j(\gamma, \rho\sigma)$ and $e(\gamma, \rho\sigma)$ in (2.44) for the empirical bias-reducing adjustments of γ and $\rho\sigma$ are given by

$$\begin{aligned}
j(\gamma, \rho\sigma) &= \begin{pmatrix} j_{\gamma\gamma} & j_{\gamma(\rho\sigma)} \\ \big(j_{\gamma(\rho\sigma)}\big)^\intercal & j_{(\rho\sigma)(\rho\sigma)} \end{pmatrix}, \\
e(\gamma, \rho\sigma) &= \begin{pmatrix} e_{\gamma\gamma} & e_{\gamma(\rho\sigma)} \\ \big(e_{\gamma(\rho\sigma)}\big)^\intercal & e_{(\rho\sigma)(\rho\sigma)} \end{pmatrix},
\end{aligned}$$

where

$$\begin{aligned}
j_{\gamma\gamma} &= -\sum_{i=1}^{n} \nabla_\gamma \nabla_\gamma^\intercal(v_i^2) \\
&= -2\sum_{i=1}^{n} z_i z_i^\intercal \\
&= -2Z^\intercal Z, \qquad (3.107) \\
j_{\gamma(\rho\sigma)} &= -\sum_{i=1}^{n} \frac{\partial}{\partial(\rho\sigma)} \nabla_\gamma(v_i^2) \\
&= -2\sum_{i=1}^{n} z_i m_2(\beta^\intercal x_i) \\
&= -2Z^\intercal M_2^a \mathbf{1}_n, \qquad (3.108) \\
\left(j_{\gamma(\rho\sigma)}\right)^\intercal &= -2\mathbf{1}_n^\intercal M_2^a Z, \qquad (3.109) \\
j_{(\rho\sigma)(\rho\sigma)} &= -\sum_{i=1}^{n} \frac{\partial^2}{\partial(\rho\sigma)^2}(v_i^2) \\
&= -2\sum_{i=1}^{n} \left(m_2(\beta^\intercal x_i)\right)^2 \\
&= -2\mathbf{1}_n^\intercal \left(M_2^a\right)^2 \mathbf{1}_n, \qquad (3.110)
\end{aligned}$$

and where

$$\begin{aligned}
e_{\gamma\gamma} &= \sum_{i=1}^{n} \left\{\nabla_\gamma v_i^2\right\}\left\{\nabla_\gamma^\intercal v_i^2\right\} \\
&= 4\sum_{i=1}^{n} \left\{ \left[\left(y_i^O - \gamma^\intercal z_i\right)^2 - 2(\rho\sigma)\left(y_i^O - \gamma^\intercal z_i\right) m_2(\beta^\intercal x_i) + (\rho\sigma)^2 \left(m_2(\beta^\intercal x_i)\right)^2 \right] z_i z_i^\intercal \right\} \\
&= 4Z^\intercal \left[(\bar{Y}^O)^2 - 2(\rho\sigma)\bar{Y}^O M_2^a + (\rho\sigma)^2 \left(M_2^a\right)^2 \right] Z, \qquad (3.111) \\
e_{\gamma(\rho\sigma)} &= \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial(\rho\sigma)} v_i^2 \right\} \left\{ \nabla_\gamma v_i^2 \right\} \\
&= 4Z^\intercal \left(M_2^a\right)^2 \left[(\bar{Y}^O)^2 - 2(\rho\sigma)\bar{Y}^O M_2^a + (\rho\sigma)^2 \left(M_2^a\right)^2 \right] \mathbf{1}_n, \\
\left(e_{\gamma(\rho\sigma)}\right)^\intercal &= 4\mathbf{1}_n^\intercal \left[(\bar{Y}^O)^2 - 2(\rho\sigma) M_2^a \bar{Y}^O + (\rho\sigma)^2 \left(M_2^a\right)^2 \right] M_2^a Z, \qquad (3.112) \\
e_{(\rho\sigma)(\rho\sigma)} &= \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial(\rho\sigma)} v_i^2 \right\} \left\{ \frac{\partial}{\partial(\rho\sigma)} v_i^2 \right\} \\
&= 4\mathbf{1}_n^\intercal \left(M_2^a\right)^2 \left[(\bar{Y}^O)^2 - 2(\rho\sigma)\bar{Y}^O M_2^a + (\rho\sigma)^2 \left(M_2^a\right)^2 \right] \mathbf{1}_n. \qquad (3.113)
\end{aligned}$$

In addition, the matrices $u_r(\gamma,\rho\sigma)$, $u_{(\rho\sigma)}(\gamma,\rho\sigma)$, $d_r(\gamma,\rho\sigma)$ and $d_{(\rho\sigma)}(\gamma,\rho\sigma)$, $r = 1,\ldots,q$, can be written as

$$\begin{aligned}
u_r(\gamma,\rho\sigma) &= \begin{pmatrix} u_{r,\gamma\gamma} & u_{r,\gamma(\rho\sigma)} \\ \big(u_{r,\gamma(\rho\sigma)}\big)^{\mathsf{T}} & u_{r,(\rho\sigma)(\rho\sigma)} \end{pmatrix}, \\
u_{(\rho\sigma)}(\gamma,\rho\sigma) &= \begin{pmatrix} u_{(\rho\sigma),\gamma\gamma} & u_{(\rho\sigma),\gamma(\rho\sigma)} \\ \big(u_{(\rho\sigma),\gamma(\rho\sigma)}\big)^{\mathsf{T}} & u_{(\rho\sigma),(\rho\sigma)(\rho\sigma)} \end{pmatrix}, \\
d_r(\gamma,\rho\sigma) &= \begin{pmatrix} d_{r,\gamma\gamma} & d_{r,\gamma(\rho\sigma)} \\ \big(d_{r,\gamma(\rho\sigma)}\big)^{\mathsf{T}} & d_{r,(\rho\sigma)(\rho\sigma)} \end{pmatrix}, \\
d_{(\rho\sigma)}(\gamma,\rho\sigma) &= \begin{pmatrix} d_{(\rho\sigma),\gamma\gamma} & d_{(\rho\sigma),\gamma(\rho\sigma)} \\ \big(d_{(\rho\sigma),\gamma(\rho\sigma)}\big)^{\mathsf{T}} & d_{(\rho\sigma),(\rho\sigma)(\rho\sigma)} \end{pmatrix},
\end{aligned}$$

where

$$\begin{aligned}
u_{r,\gamma\gamma} &= \sum_{i=1}^{n} \nabla_\gamma \nabla_\gamma^{\mathsf{T}} \left(\frac{\partial}{\partial \gamma_r} v_i^2 \right) = 0, \\
u_{r,\gamma(\rho\sigma)} &= \sum_{i=1}^{n} \nabla_\gamma \left(\frac{\partial^2}{\partial(\rho\sigma)\partial\gamma_r} v_i^2 \right) = 0, \\
u_{r,(\rho\sigma)(\rho\sigma)} &= \sum_{i=1}^{n} \frac{\partial^2}{\partial(\rho\sigma)^2} \left(\frac{\partial}{\partial \gamma_r} v_i^2 \right) = 0, \\
u_{(\rho\sigma),\gamma\gamma} &= \sum_{i=1}^{n} \nabla_\gamma \nabla_\gamma^{\mathsf{T}} \left(\frac{\partial}{\partial(\rho\sigma)} v_i^2 \right) = 0, \\
u_{(\rho\sigma),\gamma(\rho\sigma)} &= \sum_{i=1}^{n} \nabla_\gamma \left(\frac{\partial^2}{\partial(\rho\sigma)^2} v_i^2 \right) = 0, \\
u_{(\rho\sigma),(\rho\sigma)(\rho\sigma)} &= \sum_{i=1}^{n} \left(\frac{\partial^3}{\partial(\rho\sigma)^3} v_i^2 \right) = 0.
\end{aligned}$$

Finally, the components of the $(q+1)\times(q+1)$ matrices $d_r(\gamma,\rho\sigma)$ and $d_{(\rho\sigma)}(\gamma,\rho\sigma)$ are given by

$$\begin{aligned}
d_{r,\gamma\gamma} &= \sum_{i=1}^{n} \left\{ \nabla_\gamma \left(\frac{\partial}{\partial \gamma_r} v_i^2 \right) \right\} \left\{ \nabla_\gamma v_i^2 \right\} \\
&= -4 \sum_{i=1}^{n} \left\{ \big(y_i^O - \gamma^{\mathsf{T}} z_i\big) - (\rho\sigma) m_2(\beta^{\mathsf{T}} x_i) \right\} z_{ir} z_i z_i^{\mathsf{T}} \\
&= -4 Z^{\mathsf{T}} T_r \big[\bar{Y}^O - (\rho\sigma) M_2^q \big] Z, \qquad (3.114)
\end{aligned}$$

where T_r has ith diagonal element z_{ir} and where

$$\begin{aligned}
d_{r,\gamma(\rho\sigma)} &= \sum_{i=1}^{n}\left\{\nabla_\gamma\left(\frac{\partial}{\partial\gamma_r}v_i^2\right)\right\}\left\{\frac{\partial}{\partial(\rho\sigma)}v_i^2\right\} \\
&= -4\sum_{i=1}^{n}\Big\{\big(y_i^O-\gamma^\intercal z_i\big)-(\rho\sigma)m_2(\beta^\intercal x_i)\Big\}z_{ir}z_i m_2(\beta^\intercal x_i) \\
&= -4Z^\intercal T_r M_2^a\big[\bar{Y}^O-(\rho\sigma)M_2^a\big]\mathbf{1}_n, \qquad (3.115)\\
\big(d_{r,\gamma(\rho\sigma)}\big)^\intercal &= -4\mathbf{1}_n^\intercal\big[\bar{Y}^O-(\rho\sigma)M_2^a\big]M_2^a T_r Z, \qquad (3.116)\\
d_{r,(\rho\sigma)(\rho\sigma)} &= \sum_{i=1}^{n}\left\{\frac{\partial^2}{\partial(\rho\sigma)\partial\gamma_r}v_i^2\right\}\left\{\frac{\partial}{\partial(\rho\sigma)}v_i^2\right\} \\
&= -4\sum_{i=1}^{n}\Big\{\big(y_i^O-\gamma^\intercal z_i\big)-(\rho\sigma)m_2(\beta^\intercal x_i)\Big\}z_{ir}\big(m_2(\beta^\intercal x_i)\big)^2 \\
&= -4\mathbf{1}_n^\intercal T_r\big(M_2^a\big)^2\big[\bar{Y}^O-(\rho\sigma)M_2^a\big]\mathbf{1}_n, \qquad (3.117)\\
d_{(\rho\sigma),\gamma\gamma} &= \sum_{i=1}^{n}\left\{\nabla_\gamma\left(\frac{\partial}{\partial(\rho\sigma)}v_i^2\right)\right\}\left\{\nabla_\gamma v_i^2\right\} \\
&= -4\sum_{i=1}^{n}\Big\{\big(y_i^O-\gamma^\intercal z_i\big)-(\rho\sigma)m_2(\beta^\intercal x_i)\Big\}z_i z_i^\intercal m_2(\beta^\intercal x_i) \\
&= -4Z^\intercal M_2^a\big[\bar{Y}^O-(\rho\sigma)M_2^a\big]Z, \qquad (3.118)\\
d_{(\rho\sigma),\gamma(\rho\sigma)} &= \sum_{i=1}^{n}\left\{\nabla_\gamma\left(\frac{\partial}{\partial(\rho\sigma)}v_i^2\right)\right\}\left\{\frac{\partial}{\partial(\rho\sigma)}v_i^2\right\} \\
&= -4\sum_{i=1}^{n}\Big\{\big(y_i^O-\gamma^\intercal z_i\big)-(\rho\sigma)m_2(\beta^\intercal x_i)\Big\}z_i\big(m_2(\beta^\intercal x_i)\big)^2 \\
&= -4Z^\intercal\big(M_2^a\big)^2\big[\bar{Y}^O-(\rho\sigma)M_2^a\big]\mathbf{1}_n, \qquad (3.119)\\
\big(d_{(\rho\sigma),\gamma(\rho\sigma)}\big)^\intercal &= -4\mathbf{1}_n^\intercal\big[\bar{Y}^O-(\rho\sigma)M_2^a\big]\big(M_2^a\big)^2 Z, \qquad (3.120)\\
d_{(\rho\sigma),(\rho\sigma)(\rho\sigma)} &= \sum_{i=1}^{n}\left\{\frac{\partial^2}{\partial(\rho\sigma)^2}v_i^2\right\}\left\{\frac{\partial}{\partial(\rho\sigma)}v_i^2\right\} \\
&= -4\sum_{i=1}^{n}\Big\{\big(y_i^O-\gamma^\intercal z_i\big)-(\rho\sigma)m_2(\beta^\intercal x_i)\Big\}\big(m_2(\beta^\intercal x_i)\big)^3 \\
&= -4\mathbf{1}_n^\intercal\big(M_2^a\big)^3\big[\bar{Y}^O-(\rho\sigma)M_2^a\big]\mathbf{1}_n^\intercal. \qquad (3.121)
\end{aligned}$$

3.7.3 Bias reduction for Tobit V model

A straightforward generalisation of the Tobit II model is the Roy (Tobit V) model, also called the switching regression model (see Toomet and Henningsen, 2008, §2.2). In the

Tobit II model the outcome variable for an individual might not be observed. Thus we observe y_i^O for individual i if $y_i^S = 1$ but may not observe y_i^O at all if $y_i^S = 0$. In the Roy model we have two outcome variables, where only one of them is observed, depending on the selection process. This model arises in many contexts like treatment effect and choice analysis. The selection equation in this model is

$$y_i^{S*} = \beta^\intercal x_i + \varepsilon_i^S, \tag{3.122}$$

$$y_i^S = \begin{cases} 0 & \text{if } y_i^{S*} < 0 \\ 1 & \text{if } y_i^{S*} \geq 0, \end{cases} \tag{3.123}$$

and the two outcome equations are

$$y_i^{O1*} = \gamma^\intercal z_i^{O1} + \varepsilon_i^{O1}, \tag{3.124}$$

$$y_i^{O2*} = \tau^\intercal z_i^{O2} + \varepsilon_i^{O2}, \tag{3.125}$$

$$y_i^O = \begin{cases} y_i^{O1*} & \text{if } y_i^S = 0 \\ y_i^{O2*} & \text{if } y_i^S = 1, \end{cases} \tag{3.126}$$

where y_i^{S*} is the latent (unobserved) variable of the selection tendency for individual i, y_i^S is the observed binary value and y_i^{O1*} and y_i^{O2*} are the two possible latent outcomes for individual i. The distribution of the errors is usually set to be the trivariate normal distribution. The Heckman two-step correction method can be easily extended to this model and implemented, alongside ML estimation, using the sampleSelection R package.

A possible direction for further research is the application of the bias reduction methods of indirect inference and empirical bias reducing adjustments to ML estimation in the Roy model.

Chapter 4

Accelerated failure time model

4.1 Introduction

Censored survival time data are frequently encountered in survival analysis. Survival analysis, also called duration analysis in economics, is used for data in the form of times from a well-defined time origin, for example entry to a study, until the occurrence of some particular event of interest, such as death of individuals in a clinical trial or failure in mechanical systems. Survival times however, are usually censored. This means that the event of interest has not been observed for some individuals before the end of the study. There are several types of censoring (see Kleinbaum and Klein, 2011, Chapter 1 for an excellent overview) but we only focus here on right-censored survival times which is encountered when the actual survival time of an individual is greater than that observed.

Two important classes of models used for survival data are the Cox proportional hazard (PH) models (Cox, 1972) (also called relative risk models) and accelerated failure time (AFT) models (Kalbfleisch and Prentice, 2002). The Cox proportional hazards model relates the hazard function (which is the instantaneous probability of failure or death) to covariates and assumes that the covariates have a multiplicative effect on the hazard without assuming any particular survival distribution for the data. The interpretation of this model is done using hazard ratios, defined as the ratio of the predicted hazard function under two different values of a covariate. A hazard ratio of one means that the covariate has no effect on the hazard of the event, while a hazard ratio greater than one or less than one means that the event of interest is more likely or less likely to occur, respectively, when the covariate increases. Consider comparing a new treatment with a standard one (two levels of a covariate), then the proportional hazards assumption simply means that

the hazard ratio for an individual on a new treatment to that for an individual on the standard treatment remains constant over time. The Cox proportional hazards model is a semiparametric model in that the distribution of the baseline hazard (i.e. the hazard that applies when all the covariates are equal to zero) is not specified (Cameron and Trivedi, 2005, Chapter 17). However, these models can be specialised if there is reason to assume that the baseline hazard follows a particular form. For example, assuming the hazard function to be the exponential hazard gives the exponential proportional hazards model.

An alternative to the Cox PH models is the parametric AFT models where a particular form of the survival distribution is assumed and where a direct relationship is specified between the covariates and the survival time. AFT models assume that the effect of a covariate is to accelerate or decelerate the survival time of an event by some constant. The interpretation of this model is done using survival time ratios, also knows as the acceleration factor. A survival time ratio greater than or less than one, means that the event of interest is less likely or more likely to happen, respectively. Several distributions have been used to model the survival time in AFT models, including the log-normal and log-logistic distributions (for a thorough discussion of those see Lawless, 2002, Chapter 6), however in this chapter we only consider the Weibull accelerated failure time model which is the only AFT model that satisfies the proportional hazards assumption. Inference procedures for the Weibull AFT model are based on the likelihood function and estimation of the regression parameters is achieved through maximum likelihood (Lawless, 2002, §6.3). The usual large sample likelihood theory applies to the maximum likelihood estimator of the Weibull AFT model as described in Kalbfleisch and Prentice (2002, §3.4), but in small samples with censored observations the maximum likelihood estimator could be substantially biased.

In this chapter we evaluate the performance of the indirect inference estimator (Kuk, 1995) and the empirical bias reducing adjusted log-likelihood estimator of Kosmidis and Lunardon (2020) in reducing the small sample bias of the ML estimator of the Weibull AFT model. To our knowledge, this has not been done before.

The chapter is organised as follows. In Section 4.2, we describe the Weibull accelerated failure time model and review the method of maximum likelihood estimation. We explain in Section 4.3 that the bias reduction method of Firth (1993) is not applicable to this model, because the required expressions for its derivation are not available in closed form and numerical approximations are necessary. The empirical bias reducing penalty of Kosmidis and Lunardon (2020) is also derived in this section and in Section 4.4, we

compare the performance of indirect inference and empirical bias reduction with ML estimation through a small scale simulation study. Finally, in Section 4.5, the above bias reduction methods are used for the analysis of lung cancer survival data (Kleinbaum and Klein, 2011, Chapter 7).

4.1.1 The hazard and survival functions

Let T be a continuous nonnegative random variable representing the waiting time of an individual from a homogeneous population (no explanatory variables) until the occurrence of a well defined event of interest. In survival analysis, we usually refer to the waiting time variable as "survival time", because it gives the time that an individual has survived over some period of time. We also refer to the event as "failure", because usually the event of interest is death, disease incidence, or some other negative individual experience, although the event of interest may, for instance, be recovery (e.g. return to work), in which case failure is a positive event. Major areas of application of survival data analysis are biomedical studies, social research, industrial life testing and economic research.

Biomedical examples include the carcinogenesis data from Pike (1966) which gives the times from insult with a carcinogen to mortality from vaginal cancer in rats, the Stanford heart transplant data from Crowley and Hu (1977) (see also, Kalbfleisch and Prentice, 2002, Appendix A, data set IV) who give survival times of potential heart transplant recipients from their date of acceptance into the Stanford heart transplant program, and the Veterans' Administration lung cancer data (Prentice, 1973; Kalbfleisch and Prentice, 2002, Appendix A, data set I) where males with advanced inoperable lung cancer were randomized to either a standard or test chemotherapy and the primary endpoint for therapy comparison was time to death (see also, Kalbfleisch and Prentice, 2002, Appendix A for various other medical data set examples). Kennan (1985) and Kiefer (1985) used survival analysis techniques in a social context to study the duration of strikes in U.S. manufacturing, measured in number of days from the start of strike. The accelerated life test data presented in Nelson and Hahn (1972) on the number of hours to failure of motorettes operating under various temperatures is yet another application in engineering research.

Suppose that T has probability density function (pdf) $f(t)$ and cumulative distribution function (cdf) $F(t) = \Pr(T \leq t)$, giving the probability that the event has occurred by duration t. It is often more convenient to work with the complement of the cdf, the

survival function

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= \int_t^{\infty} f(x)\,dx, \end{aligned} \tag{4.1}$$

which gives the probability that the event of interest has not occurred by duration t. Clearly, $S(t)$ is monotonically decreasing from one to zero since $F(t)$ is monotonically increasing from zero with $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$, that is theoretically, if time increased without limit, eventually nobody would survive.

An alternative characterisation of the distribution of T is given by the hazard function, or the conditional failure rate, which is the instantaneous rate per unit time at which failures occur, given that the individual has survived up to time t, defined as

$$\begin{aligned} \lambda(t) &= \lim_{dt\to 0} \frac{\Pr(t \le T < t + dt | T \ge t)}{dt} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{4.2}$$

Note that, the hazard function focuses on failing, that is, on the event occurring, in contrast to the survival function, which focuses on not failing. Thus, in some sense, the hazard function gives the opposite side of the information given by the survivor function. Note also that the scale for the hazard rate is not $(0,1)$, as for a probability, but rather $(0,\infty)$ and depends on the unit of time used. In contrast to the survivor function, the graph of $\lambda(t)$ does not have to start at one and decrease to zero, but rather can start anywhere and increase or decrease in any direction over time. In particular, for a specific value of t, the hazard function $\lambda(t)$ is always nonnegative, that is, equal to or greater than zero, and has no upper bound (see Kleinbaum and Klein, 2011, Chapter 1, for examples).

The survivor and hazard functions are also closely related and provide alternative but equivalent characterisations of the distribution of T, in particular given one, the other can be easily derived. Since from (4.1), $-f(t)$ is the derivative of $S(t)$, the hazard equals the change in log survivor function

$$\lambda(t) = -\frac{d \ln S(t)}{dt}. \tag{4.3}$$

Integrating the hazard with respect to t and introducing the boundary condition $S(0) = 1$, we can solve the above expression to obtain a formula for $S(t)$ as a function of the hazard

at all durations up to t:

$$S(t) = \exp\left(-\int_0^t \lambda(u)\,du\right). \tag{4.4}$$

The integral in the above equation is called the cumulative hazard function or integrated hazard function

$$\begin{aligned} \Lambda(t) &= \int_0^t \lambda(u)\,du \\ &= -\ln S(t), \end{aligned} \tag{4.5}$$

and is the sum of the risks faced between 0 and t. The above relations suggest that any nonnegative function $\lambda(t)$ that satisfies $\Lambda(t) < \infty$ for some $t > 0$ and $\Lambda(\infty) = \infty$, i.e. $S(\infty) = 0$ can be the hazard function of the continuous random variable T, and any distribution defined for $t \in [0,\infty)$ can serve as a survival distribution for T. In fact given a random variable W with a standard distribution in $(-\infty,\infty)$ we can generate a family of survival distributions by introducing location and scale changes of the form

$$\ln T = Y = \alpha + \sigma W, \tag{4.6}$$

which will be discussed further in the next subsection.

Another representation of the failure time distribution is the expectation of life, given by

$$\begin{aligned} \mathrm{E}(T) &= \int_0^\infty t f(t)\,dt \\ &= \int_0^\infty S(t)\,dt, \end{aligned} \tag{4.7}$$

where the second equality above is obtained by integrating by parts, and making use of the fact that $-f(t)$ is the derivative of $S(t)$, with limits $S(0) = 1$ and $S(\infty) = 0$. In words, the mean survival time is simply the integral of the survival function or the area under the survival curve.

4.1.2 Parametric models of the hazard function

In this section we discuss briefly the general properties of some of the standard parametric failure time models for homogeneous populations like the exponential and Weibull models which are often used in the literature. Johnson and Kotz (1970) and Kalbfleisch

and Prentice (2002, Chapter 2) provide a more detailed discussion of these distributions and of others that are frequently used such as the gamma, log-normal, log-logistic, generalised gamma and generalised F distributions. As before, let $T > 0$ be a random variable representing failure time and let $Y = \ln T = \alpha + \sigma W$ represent the log failure time. We summarise below the exponential and Weibull failure time distributions in terms of T, Y and W.

The exponential survival distribution is the simplest possible and is obtained by assuming a constant risk over time, so the hazard function is

$$\lambda(t) = \lambda > 0, \tag{4.8}$$

for all t. This is known as the memoryless property of the exponential distribution; the conditional probability of failure in a given short interval is the same regardless of when the observation is made. The corresponding survival and density functions of T are

$$\begin{aligned} S(t) &= \exp(-\lambda t), \\ f_T(t) &= \lambda \exp(-\lambda t). \end{aligned} \tag{4.9}$$

The density and survival function of $Y = \ln(T)$ are, respectively

$$f_Y(y) = \exp\left(y - \alpha - e^{y-\alpha}\right), \tag{4.10}$$

$$S(y) = \exp\left(-e^{y-\alpha}\right), \tag{4.11}$$

where $\alpha = -\ln(\lambda)$ and $-\infty < y < \infty$. In fact T has an exponential distribution with parameter λ, denoted $T \sim \text{Exp}(\lambda)$ if

$$Y = \ln(T) = \alpha + W, \tag{4.12}$$

where $\alpha = -\ln(\lambda)$ and W has a standard extreme value (minimum) distribution with density and survival functions given respectively by

$$f_W(w) = \exp\left(w - e^w\right), \tag{4.13}$$

$$S_W(w) = \exp\left(-e^w\right), \tag{4.14}$$

where $-\infty < w < \infty$. The moment generating function of W is

$$M_W(\theta) = \mathrm{E}\left(\mathrm{e}^{\theta \mathrm{W}}\right) = \Gamma(\theta+1), \tag{4.15}$$

where $\theta > -1$ and where

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x}\, dx \tag{4.16}$$

is the gamma function.

The exponential is a one-parameter distribution that is too restrictive in practice. A generalization commonly used in econometrics is the Weibull distribution which allows for a power dependence of the hazard on time. This yields the two-parameter Weibull distribution with hazard function

$$\begin{aligned} \lambda(t) &= \lambda^p p t^{p-1} \\ &= \lambda p (\lambda t)^{p-1}, \end{aligned} \tag{4.17}$$

for $\lambda, p > 0$. The log of the Weibull hazard is a linear function of log time with constant $p\ln(\lambda) + \ln(p)$ and slope $p-1$. Thus, this hazard is monotone decreasing for $p < 1$, reduces to the constant exponential hazard if $p = 1$, and increasing for $p > 1$. The survival and density functions of T are

$$\begin{aligned} S(t) &= \exp\left[-(\lambda t)^p\right], \\ f_T(t) &= \lambda p (\lambda t)^{p-1} \exp\left[-(\lambda t)^p\right]. \end{aligned} \tag{4.18}$$

The density and survival function of $Y = \ln(T)$ are, respectively

$$f_Y(y) = \sigma^{-1} \exp\left[\sigma^{-1}(y-\alpha) - e^{(y-\alpha)/\sigma}\right], \tag{4.19}$$

$$S_Y(y) = \exp\left[-e^{(y-\alpha)/\sigma}\right], \tag{4.20}$$

where $-\infty < y < \infty$, $\alpha = -\ln(\lambda)$ and $\sigma = p^{-1}$. In fact T has a Weibull distribution with parameters λ and p, denoted $T \sim W(\lambda, p)$ if

$$Y = \ln(T) = \alpha + \sigma W, \tag{4.21}$$

where $\alpha = -\ln(\lambda)$, $\sigma = p^{-1}$ and W has the extreme value (minimum) distribution (4.13).

Another generalization which includes the exponential distribution as a special case is the two parameter gamma distribution for T where $Y = \ln(T) = \alpha + W$, with W having the generalized extreme value distribution.

Other parametric models are also used such as the log-normal distribution for T which can be written in the form $Y = \ln(T) = \alpha + \sigma W$, where W has a standard normal distribution. The hazard function of the log-normal distribution increases from zero to reach a maximum and then decreases monotonically, approaching zero as t becomes large.

The gamma family can be further generalized by incorporating a scale parameter σ in the model for Y to give $Y = \ln(T) = \alpha + \sigma W$, where T and W have generalized gamma and generalized extreme value distributions, respectively. This three parameter model was introduced by Stacy (1962) and includes as special cases the exponential, the Weibull, the gamma and the log-normal (limiting case) distributions.

A good approximation to the log-normal distribution is the log-logistic distribution which is obtained for T if $Y = \ln(T) = \alpha + \sigma W$, where W has a standard logistic distribution.

The generalized F distribution for T incorporates all of the preceeding distributions as special or limiting cases if $Y = \ln(T) = \alpha + \sigma W$, where W is distributed as the log of an F variate. This model is particularly useful for discriminating between competing models such as the Weibull and log-logistic distributions for a given data set because it has the advantage of adapting to a wide variety of distributional shapes (see Kalbfleisch and Prentice, 2002, §3.8).

4.1.3 Censoring and the likelihood function

Failure time data are usually censored, in the sense that some individuals fail and therefore we know their exact survival time, whereas other individuals do not fail during their observation period and all we know is that their survival time exceeds the observation time. There are generally three reasons why censoring may occur: some individuals are still surviving at the time the study is terminated and the analysis is done, or contact with the individual is lost to follow up during the study period or individuals may be withdrawn or decide to withdraw from the study because of a worsening or improving prognosis.

As is intuitively apparent, some censoring mechanisms have the potential to introduce bias into the estimation of survival probabilities or into treatment comparisons. In practice data may be right censored, left censored or interval censored. Right censoring or censoring from above occurs when an individual is observed to fail after some time c, but

the actual time of failure is unknown and all we know is that failure occurs at some time in the interval (c,∞). Likewise, left censoring or censoring from below occurs if the individual is observed to fail at some time in the interval $(0,c)$ but the exact time is unknown. Interval censoring occurs when we only observe that the failure time falls within some interval (a,b), and it incorporates both right censoring and left censoring as special cases.

We focus our attention on right censoring as in most survival analysis literature. However, even with this restriction there are a variety of possible censoring mechanisms or assumptions, including independent censoring, random censoring and non informative censoring (Kleinbaum and Klein, 2011, Chapter 1, §XI). For standard survival analysis methods to be valid in the presence of censoring, the censoring mechanism needs to be one with independent censoring. In this chapter we assume random censoring which is a stronger assumption and more restrictive than independent censoring.

Consider survival studies in which n items or individuals are put on test and data of the form (t_i,δ_i,z_i), $i=1,\ldots,n$, are observed. Here δ_i is an indicator variable ($\delta_i=0$ if the ith item is censored; $\delta_i=1$ if the ith item failed), $t_i>0$ is the corresponding failure or censoring time, $z_i=(z_{i1},\ldots,z_{ip})^\intercal$ is a vector of covariates associated wth the ith individual which may contain information on treatment group, various physical measurements, and so on, where aspects of z_i are expected to be predictive of subsequent failure time, and the parameter vector $\theta=(\theta_1,\ldots,\theta_p)^\intercal$ is the covariate coefficients.

Let $\tilde{T}_i$ be the uncensored continuous failure time variable, so that the survivor function for the ith individual is $\Pr(\tilde{T}_i>t;\theta,z_i)=S(t;\theta,z_i)$ with the corresponding density $f(t;\theta,z_i)$. To obtain the likelihood function for θ, we need a probability model for the censoring mechanism. Assume that the censoring mechanism is random, specifically, assume that the censoring time C_i for the ith individual is a random variable with survivor and density functions G_i and g_i, respectively $(i=1,\ldots,n)$, and that given $z_1,\ldots,z_n$, the C_i's are stochastically independent of each other and of the independent failure times $\tilde{T}_1,\ldots,\tilde{T}_n$, i.e. $(\tilde{T}_i,C_i)$ are independent. So each individual in the sample has a time $\tilde{T}_i$ to failure and a time C_i to censoring. The observed data $(t_1,\delta_1),\ldots,(t_n,\delta_n)$ are realisations of the random variables

$$T_i = \min\left(\tilde{T}_i,C_i\right) \tag{4.22}$$

$$\delta_i = \begin{cases} 1 & \text{if } T_i=\tilde{T}_i\,(\tilde{T}_i\le C_i) \\ 0 & \text{if } T_i=C_i\,(\tilde{T}_i>C_i). \end{cases} \tag{4.23}$$

Thus T_i is the observed, possibly censored failure time.

Note that the random censorship model includes the special case of **Type I** censoring, where a sample of n individuals or units is followed for a fixed time, i.e. the censoring time is fixed in advance. This means that the number of units or individuals failing is random, but the total duration of the study is fixed.

Under the random censoring mechanism, it may be shown that (Kalbfleisch and Prentice, 2002, §3.2) the likelihood and log-likelihood functions on the data $(T_i = t_i, \delta_i, z_i)$, $i = 1, \ldots, n$, conditional on $z_1, \ldots, z_n$ are given by

$$L(\theta) = \prod_{i=1}^{n} f(t_i; \theta, z_i)^{\delta_i} S(t_i; \theta, z_i)^{1-\delta_i}, \tag{4.24}$$

$$l(\theta) = \sum_{i=1}^{n} \left\{ \delta_i \ln f(t_i; \theta, z_i) + (1 - \delta_i) \ln S(t_i; \theta, z_i) \right\}. \tag{4.25}$$

In fact the above likelihood and log-likelihood functions are more generally correct under the class of independent censoring mechanisms (Kalbfleisch and Prentice, 2002, §6.2).

The independent censorship model includes the special case of **Type II** censoring which occurs when a sample of n units or individuals are followed as long as necessary until d units have failed. For example, a clinical trial may end after d patients have died. This means that the number of failures or deaths is fixed in advance but the total duration of the study is random and cannot be known with certainty.

4.2 Description of the model and maximum likelihood estimation

Consider again data of the form $(\ln t_i, \delta_i, z_i)$, $(i = 1, \ldots, n)$, where δ_i is an indicator variable ($\delta_i = 0$ if the ith item is censored; $\delta_i = 1$ if the ith item failed), as in (4.23), and t_i is the corresponding failure or censoring time, as in (4.22). Accelerated failure time models (also called log-linear models) are ones where the covariates act additively on the log survival time

$$Y = \ln(T) = Z\beta + \sigma W, \tag{4.26}$$

where $\ln(T) = (\ln T_1, \ldots, \ln T_n)^\mathsf{T}$, $\beta^\mathsf{T} = (\beta_1, \ldots, \beta_p)$ is a $1 \times p$ vector of parameters, Z is an $n \times p$ matrix of covariates (explanatory variables) having rows $z_1, \ldots, z_n$, where $z_i = (z_{i1}, \ldots, z_{ip})^\mathsf{T}$ and Z has (i,t)th element z_{it}, σ is a scale constant and where $W = (Y - Z\beta)/\sigma$ is a standardised random error variable with density function $f_W(w)$ and survival function $S_W(w)$. We assume that W is independent of β given the covariates Z.

The density and survival function of the log survival time T can be expressed in terms of the density and survival function of W and are, respectively

$$f_{\ln T}(t) = \sigma^{-1} f_W(w), \tag{4.27}$$

$$S_{\ln T}(t) = S_W(w). \tag{4.28}$$

The log-likelihood function about β and σ for a general accelerated failure time model is given by

$$l(\beta,\sigma) = \sum_{i=1}^{n} \big\{ \delta_i \ln(\sigma^{-1}) + \delta_i \ln f(w_i) + (1-\delta_i) \ln S(w_i) \big\}. \tag{4.29}$$

Any of the parametric models for T discussed in Section 4.1.2 can be extended to allow parameters to depend on covariates and hence define an accelerated failure time model. In this chapter we consider the Weibull accelerated failure time model defined by choosing the Weibull distribution for T which makes the error term W have an extreme value distribution with density and survival function given respectively by (4.13) and (4.14). Only the scale parameter λ of the Weibull distribution is allowed to depend on the covariates and is given by $\lambda = \exp(-\alpha) = \exp(-Z\beta)$ while the shape parameter $p = \sigma^{-1}$ is fixed. The log-likelihood functions for the Weibull accelerated failure time model may be written as (Greene, 2012, §19.4.3.d)

$$l(\beta,\sigma) = \sum_{i=1}^{n} \big\{ \delta_i \big[w_i - \ln(\sigma) \big] - \exp(w_i) \big\}. \tag{4.30}$$

Differentiating the above log-likelihood with respect to β and σ, respectively yields

$$\begin{aligned} \nabla_\beta l(\theta) &= \sum_{i=1}^{n} \frac{z_i}{\sigma} \Big\{ -\delta_i + \exp(w_i) \Big\} \\ &= \frac{1}{\sigma} Z^\intercal \big[-\delta + W_e \mathbf{1}_n \big], \qquad (4.31) \\ \frac{\partial}{\partial\sigma} l(\theta) &= \sum_{i=1}^{n} \frac{1}{\sigma} \Big\{ w_i \big(\exp(w_i) - \delta_i \big) - \delta_i \Big\} \\ &= \frac{1}{\sigma} \mathbf{1}_n^\intercal \big[W(W_e \mathbf{1}_n - \delta) - \delta \big], \qquad (4.32) \end{aligned}$$

where Z is the $n \times p$ matrix with z_i as its i-th row, $\mathbf{1}_n$ is an n-vector of ones, δ is the $n \times 1$ vector of $\delta_1, \ldots, \delta_n$, $W = \mathrm{diag}\{w_1, \ldots, w_n\}$, where $w_i = (y_i - z_i^\intercal \beta)/\sigma$ and where

$W_e = \text{diag}\{\exp(w_1), \ldots, \exp(w_n)\}$. In obtaining (4.31) and (4.32) we use the results

$$\nabla_\beta w_i = -\frac{z_i}{\sigma}, \tag{4.33}$$

$$\frac{\partial w_i}{\partial \sigma} = -\frac{w_i}{\sigma}. \tag{4.34}$$

The above score equations (see Kalbfleisch and Prentice, 2002, §3.6, for the score equations of a general accelerated failure time model) have no closed form solution and must be solved numerically, for example, the *survreg* function from the survival R package (Therneau, 2021) provides the ML estimates.

4.3 Reduced-bias estimation methods

4.3.1 Firth's adjusted score equations method

The method of Firth (1993) is not applicable to the accelerated failure time model because the Fisher information matrix is not available in closed form. The latter matrix is not available unless the censoring process is fully specified and even when fully specified, the evaluation of the Fisher information requires numerical approximations. Under Type I or Type II censoring, Lawless (2002, §5.6) gives general expressions, in terms of expectations, for the components of the Fisher information matrix for a general accelerated failure time model, where the distribution of the failure time is not necessarily Weibull. Using the Law of iterated expectation and the conditional density $f(w|\delta)$, we derive below the components of the Fisher information matrix for the Weibull accelerated failure time model in terms of integrals, and explain how the score equations can be used to simplify these expressions.

Differentiating (4.31) and (4.32) yields the Hessian and second order partial derivatives

$$\nabla_\beta \nabla_\beta^\intercal l(\theta) = -\frac{1}{\sigma^2} Z^\intercal W_e Z, \tag{4.35}$$

$$\frac{\partial}{\partial \sigma} \nabla_\beta l(\theta) = \frac{1}{\sigma^2} Z^\intercal \big[\delta - W_e(W + I_n)\mathbf{1}_n\big], \tag{4.36}$$

$$\frac{\partial^2}{\partial(\sigma)^2} l(\theta) = \frac{1}{\sigma^2} \mathbf{1}_n^\intercal \big[(2W + I_n)\delta - W W_e(W + 2I_n)\mathbf{1}_n\big]. \tag{4.37}$$

The $(p+1)\times(p+1)$ observed information matrix becomes

$$j(\theta) = \begin{pmatrix} j_{\beta\beta} & j_{\beta\sigma} \\ (j_{\beta\sigma})^{\intercal} & j_{\sigma\sigma} \end{pmatrix}, \tag{4.38}$$

where $j_{\beta\beta} = -\nabla_\beta \nabla_\beta^{\intercal} l$, $j_{\beta\sigma} = -\partial(\nabla_\beta l)/\partial\sigma$, $j_{\sigma\sigma} = -\partial^2 l/\partial(\sigma)^2$ and where $(j_{\beta\sigma})^{\intercal} = -\sigma^{-2}\left[\delta^{\intercal} - \mathbf{1}_n^{\intercal}(W+I_n)W_e\right]Z$.

Assume that the censoring of the failure time data is according to Type I, i.e. $C_i = c_i$ is the prespecified censoring time for individual i and let $R_i = (\ln C_i - z_i^{\intercal}\beta)/\sigma$. Then the probability of failure is given by $\Pr(\delta_i = 1) = \Pr(T_i \leq C_i) = \Pr(W_i \leq R_i) = F(R_i) = 1 - \exp(-e^{R_i})$ and the conditional density of W_i is (Lawless, 2002, §5.1.1)

$$\begin{aligned} f(w_i|\delta_i = 1) &= \frac{f(w_i, \delta_i = 1)}{\Pr(\delta_i = 1)} \\ &= \frac{f(w_i)}{F(R_i)} \\ &= \frac{\exp(w_i - e^{w_i})}{1 - \exp(-e^{R_i})}, \end{aligned} \tag{4.39}$$

where $-\infty < w_i \leq R_i$. For simplicity, consider the components of the Fisher information matrix for β_1 which simplify to

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial \beta_1^2}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n} z_{1i}^2 \mathrm{E}(e^{w_i}), \tag{4.40}$$

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial \beta_1 \partial \sigma}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n} z_{1i} \mathrm{E}(w_i e^{w_i}), \tag{4.41}$$

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial (\sigma)^2}\right) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\Big\{2\mathrm{E}(\delta_i w_i) + \mathrm{E}(\delta_i) - \mathrm{E}(w_i^2 e^{w_i}) - 2\mathrm{E}(w_i e^{w_i})\Big\}. \tag{4.42}$$

However, the score equations $\mathrm{E}(\partial l_i/\partial\beta_1) = 0$ and $\mathrm{E}(\partial l_i/\partial\sigma) = 0$ imply respectively, that $\mathrm{E}(e^{w_i}) = \mathrm{E}(\delta_i) = \Pr(\delta_i = 1) = 1 - \exp(-e^{R_i})$ and $\mathrm{E}(w_i e^{w_i}) = \mathrm{E}(\delta_i w_i) + \mathrm{E}(\delta_i)$. Moreover, using the Law of iterated expectation, the expected value of $w_i e^{w_i}$ and $w_i^2 e^{w_i}$ may be

simplified. For example,

$$\begin{aligned}
\mathrm{E}_{W_i,\delta_i}\left(w_i e^{w_i}\right) &= \mathrm{E}_{\delta_i}\left\{\mathrm{E}_{W_i|\delta_i}\left(w_i e^{w_i}\right)\right\} \\
&= \Pr(\delta_i = 0)\mathrm{E}_{W_i|\delta_i=0}\left(w_i e^{w_i}\right) + \Pr(\delta_i = 1)\mathrm{E}_{W_i|\delta_i=1}\left(w_i e^{w_i}\right) \\
&= S(R_i)\left(R_i e^{R_i}\right) + F(R_i)\int_{-\infty}^{R_i} w_i e^{w_i}\frac{\exp(w_i - e^{w_i})}{F(R_i)}\,dw_i \\
&= R_i \exp\left(R_i - e^{R_i}\right) + \int_{-\infty}^{R_i} w_i \exp\left(2w_i - e^{w_i}\right) dw_i, \qquad (4.43)
\end{aligned}$$

where $S(R_i) = 1 - F(R_i)$ and similarly for the expected value of $w_i^2 e^{w_i}$. Using the above argument, (4.40), (4.41) and (4.42) simplify to

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial \beta_1^2}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n} z_{1i}^2\left[1 - \exp(-e^{R_i})\right], \qquad (4.44)$$

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial \beta_1 \partial \sigma}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n} z_{1i}\left\{R_i \exp\left(R_i - e^{R_i}\right) + \int_{-\infty}^{R_i} w_i \exp\left(2w_i - e^{w_i}\right) dw_i\right\}, \qquad (4.45)$$

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial (\sigma)^2}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left\{\left[1 - \exp(-e^{R_i})\right] + R_i^2 \exp\left(R_i - e^{R_i}\right) + \int_{-\infty}^{R_i} w_i^2 \exp\left(2w_i - e^{w_i}\right) dw_i\right\}. \qquad (4.46)$$

The above integrals are not available in closed form unless the sample is uncensored in which case we can let $C_i \to \infty$, i.e. $R_i \to \infty$, and the above integrals can be written in terms of Euler's constant. For a censored sample numerical integration is necessary.

4.3.2 Empirical bias-reducing penalty

The implementation of iRBM-estimation is straightforward and is equivalent to the maximisation of the penalised function (2.48), where $j(\theta)$ is derived in Section 4.3.1 and the $(p+1)\times(p+1)$ matrix $e(\theta)$ takes the form

$$e(\theta) = \begin{pmatrix} e_{\beta\beta} & e_{\beta\sigma} \\ (e_{\beta\sigma})^{\mathsf{T}} & e_{\sigma\sigma} \end{pmatrix}, \qquad (4.47)$$

where

$$
\begin{aligned}
e_{\beta\beta} &= \sum_{i=1}^{n}\Big(\nabla_{\beta} l_i(\theta)\Big)\Big(\nabla_{\beta}^{\intercal} l_i(\theta)\Big) \\
&= \frac{1}{\sigma^2} Z^{\intercal}\big[D^2 - 2DW_e + W_e^2\big]Z, \qquad (4.48) \\
e_{\beta\sigma} &= \sum_{i=1}^{n}\Big(\nabla_{\beta} l_i(\theta)\Big)\Big(\frac{\partial}{\partial\sigma} l_i(\theta)\Big) \\
&= \frac{1}{\sigma^2} Z^{\intercal}\Big[W_e\big\{-D(I_n+2W)+WW_e\big\}+D^2(I_n+W)\Big]\mathbf{1}_n, \qquad (4.49) \\
(e_{\beta\sigma})^{\intercal} &= \frac{1}{\sigma^2}\mathbf{1}_n^{\intercal}\Big[\big\{-(I_n+2W)D+W_eW\big\}W_e+(I_n+W)D^2\Big]Z, \qquad (4.50) \\
e_{\sigma\sigma} &= \sum_{i=1}^{n}\Big(\frac{\partial}{\partial\sigma} l_i(\theta)\Big)\Big(\frac{\partial}{\partial\sigma} l_i(\theta)\Big) \\
&= \frac{1}{\sigma^2}\mathbf{1}_n^{\intercal}\Big[WW_e\big\{WW_e-2D(I_n+W)\big\}+D^2\big\{W(2I_n+W)+I_n\big\}\Big]\mathbf{1}_n, \qquad (4.51)
\end{aligned}
$$

and where $D = \mathrm{diag}\{\delta_1,\ldots,\delta_n\}$.

The value of the eRBM estimator defined in (2.49) is readily available once the empirical bias-reducing penalty term in (2.48) is calculated. All that is needed is the numerical differentiation of this penalty term and its evaluation with the inverse observed information matrix at the ML estimates.

4.3.3 Indirect inference

The indirect inference estimator, $\tilde{\theta}$, is the solution of (2.19) and is obtained iteratively where the indirect inference estimate of θ at the $k+1$-th iteration is given by

$$
\tilde{\theta}^{(k+1)} = \hat{\theta} - \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}\big(y^{(r)}\big) + \tilde{\theta}^{(k)}, \qquad (4.52)
$$

where $y^{(1)},\ldots,y^{(R)}$ are simulated from the Weibull accelerated failure time model with the parameters set at $\tilde{\theta}^{(k)}$ and where the initial estimate $\tilde{\theta}^{(0)}$ is chosen to be the ML estimate. The iterative process is then repeated until the difference of the components of $\tilde{\theta}^{(k+1)}$ and $\tilde{\theta}^{(k)}$ are all less than δ in absolute value at the current estimates, where δ is a small number.

4.4 Simulation study

In order to assess the finite sample performance of the maximum likelihood, indirect inference, iRBM and eRBM estimators for the accelerated Weibull failure time model we conduct a simulation study to compare their bias, average estimated standard error, empirical standard error, length and coverage probability of 95% and 99% confidence intervals for different sample sizes. We focus on modelling with both categorical and continuous covariates. Let $\ln(t_i) = \beta_1 + z_{i2}\beta_2 + z_{i3}\beta_3 + \sigma w_i$ and assume z_{i2} is a continuous covariate generated from the standard normal distribution and z_{i3} is a binary covariate taking values 0 and 1 depending on group membership (e.g. new drug versus placebo). The log transformed Weibull survival times, t_i, were generated using the function *simEventData* from the R reda package (Wang and Yan, 2019) with survival and hazard functions

$$S(t_i) = \exp\left(- \mathrm{e}^{\left[\ln(t_i) - \beta_1 - z_{i2}\beta_2 - z_{i3}\beta_3\right]/\sigma} \right), \tag{4.53}$$

$$\lambda(t_i|z_i) = \rho(\lambda_{z_i} t_i)^{\rho}/t_i, \tag{4.54}$$

where $\lambda_{z_i} = \exp(-\beta_1 - z_{i2}\beta_2 - z_{i3}\beta_3)$ and where i denoted the subject $i = 1,\ldots,n$, $\beta_1 = 0.3$ and $\rho = 1/\sigma = 1.5$. The censoring times, c_i, were randomly generated from the exponential distribution with survival function $S(c_i) = \exp(-c_i/m)$, where m is the mean of the exponential distribution. We generated differing censoring levels by varying m to accommodate overall failure percentages of approximately 24% and 45%. Data sets were generated for $\beta_2 = 0.5$, $\beta_3 = 0.2$ and $\sigma = 0.667$ and sample sizes $n = 50, 100, 150$, resulting in a total of 6 scenarios. Failure indicators were defined as $\delta_i = 1$ if $\tilde{t}_i \leq c_i (t_i = \tilde{t}_i)$ and $\delta_i = 0$ if $\tilde{t}_i > c_i (t_i = c_i)$. There were an equal number of exposed $(z_{i3} = 1)$ and unexposed $(z_{i3} = 0)$ observations in each data set, and 5000 data sets were generated for each scenario.

The maximum likelihood estimates and their estimated standard errors were obtained using the R function *survreg* from the survival package. When *survreg* fails to converge, we use the R function *nlminb* to compute the ML estimates. The indirect inference estimate of σ can end up outside $(0,\infty)$ during iterations. In those cases it is not possible to simulate the R Monte Carlo samples from the accelerated Weibull failure time model at the current indirect inference estimates. We resolve this by using the reparametrisation $\ln(\sigma)$ which allows σ to diverge to infinity so that when we transform it back we avoid its boundary value. For this reason, the indirect inference estimate of σ is not the

bias corrected one but the transformed version of it. We set the value of Monte Carlo replicates R to be 500 and any samples for which the ML estimate does not converge are excluded from the Monte Carlo estimate of the mean of the ML estimates. We considered that the indirect inference estimator converged if the absolute difference in parameter estimates (over all parameters) between iterations was less than $\delta = 0.01$. The standard errors of the indirect inference estimates are evaluated using the inverse of the observed information matrix evaluated at the indirect inference estimates, i.e. using the Hessian matrix.

The iRBM-estimates are computed using the *optimx* function in R (Nash and Varadhan, 2011) for the maximisation of the empirical penalised log-likelihood (2.48). We specified that all available optimisation methods in *optimx* are used and we chose the estimates from the optimisation method that converged and satisfied the kkt1 and kkt2 conditions where kkt1 checks whether the gradient at the final parameter estimates is small and kkt2 checks whether the Hessian at the final parameter estimates is positive definite, except when some of the parameters are on the boundary in which case the Hessian is allowed to be positive semi-definite. Since we are trying to reduce the bias at the σ parametrisation, in order to avoid a constrained optimisation problem ($\sigma > 0$), we transform σ to $\ln(\sigma)$ so that we optimise the empirical penalised log-likelihood function on the real line. The estimate of $\ln(\sigma)$ from *optimx* is then transformed back using the exponential transformation and the resulting final estimate is the iRBM-estimate of σ. The eRBM-estimates are computed directly by substituting the ML estimates into (2.49) and the standard errors of the iRBM-estimates and the eRBM-estimates are evaluated using the Hessian matrix. We compute the coverage probability and length of confidence intervals under the following convention. We assume that if the standard error for a particular parameter is not available then the confidence interval covers the true parameter value and that the interval has infinite length so we compute the median length over simulated samples.

Tables 4.1 and 4.2 report the bias of estimators with overall censoring percentage of approximately 24% and 45%, respectively, Tables 4.3 and 4.4 report the average estimated standard error and empirical standard error with overall censoring percentage of approximately 24% and 45%, respectively, Tables 4.5 and 4.6 report the coverage probability and median length of confidence intervals with nominal level 95% with overall censoring percentage of approximately 24% and 45%, respectively and Tables 4.7 and 4.8 report the coverage probability and median length of confidence intervals with nominal level 99% with overall censoring percentage of approximately 24% and 45%, respectively.

All data sets converged and for an overall censoring of approximately 24%, the indirect inference and iRBM estimates are less biased than the ML estimates for all parameters for $n = 50$ and $n = 100$. For $n = 150$, the indirect inference and iRBM estimates perform better than ML in terms of bias for β_1, β_2 and σ. For an overall censoring of approximately 45%, the indirect inference and iRBM estimates are less biased than the ML estimates for all parameters for $n = 50$ and $n = 150$. For $n = 100$, the indirect inference and iRBM estimates perform better than ML in terms of bias for β_1, β_2 and σ. In general, the eRBM estimates do not improve the bias of the ML estimates except for the β_1 estimates which are less biased for all sample sizes considered and for both censoring levels. In fact, Tables 4.1 and 4.2 show that the eRBM estimates of the bias of σ are larger than the ML estimates of the bias of σ, i.e. eRBM does not reduce the bias in the estimation of the standard deviation. The average estimated standard errors are smaller than the empirical standard errors which means that the estimated asymptotic standard errors underestimate finite sample variability, and thus the 95% and 99% coverage rates are slightly low in general. For an overall censoring percentage of approximately 24%, the 95% coverage probabilities using the indirect inference, iRBM and eRBM estimators are closer to the nominal level than the 95% coverage probabilities using the ML estimator, for all parameters. For an overall censoring percentage of approximately 45%, the 95% coverage probabilities using the indirect inference and iRBM estimators are closer to the nominal level than the 95% coverage probabilities using the ML estimator. The 99% coverage probabilities using the indirect inference, iRBM and eRBM estimators are generally closer to the nominal level than those using the ML estimator for both censoring levels, except for $n = 150$, where the coverage probability of the ML estimator of β_2 is closer to the nominal level than the indirect inference, iRBM and eRBM estimators of β_1, for a 24% censoring level, and where the coverage probability of the ML estimator of β_3 is closer to the nominal level than the indirect inference, iRBM and eRBM for a 45% censoring level. We conclude that the performance of the indirect inference and iRBM estimators is superior to the ML estimator for the two censoring levels of 24% and 45% and for all parameters especially for sample sizes $n = 50$ and $n = 100$.

Table 4.1: Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the estimated bias and Monte Carlo simulation standard error (in parentheses) of maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all entries multiplied by 10 and with overall censoring percentage of approximately 24%.

	Maximum Likelihood	Indirect Inference	iRBM	eRBM
$n = 50$				
β_1	-0.123 (0.039)	-0.008 (0.038)	-0.048 (0.038)	-0.038 (0.039)
β_2	0.017 (0.034)	-0.009 (0.033)	0.002 (0.034)	0.095 (0.034)
β_3	-0.061 (0.052)	-0.021 (0.051)	-0.052 (0.051)	-0.085 (0.052)
σ	-0.284 (0.042)	0.001 (0.045)	-0.090 (0.044)	0.489 (0.052)
$n = 100$				
β_1	-0.082 (0.035)	-0.015 (0.034)	-0.029 (0.034)	-0.053 (0.035)
β_2	0.046 (0.031)	0.030 (0.031)	0.035 (0.031)	0.088 (0.031)
β_3	0.016 (0.045)	0.015 (0.045)	0.015 (0.045)	0.034 (0.045)
σ	-0.130 (0.043)	0.007 (0.044)	-0.016 (0.044)	0.261 (0.048)
$n = 150$				
β_1	-0.040 (0.033)	0.005 (0.033)	-0.004 (0.033)	-0.014 (0.033)
β_2	0.008 (0.030)	-0.001 (0.030)	0.002 (0.030)	0.034 (0.030)
β_3	0.006 (0.044)	0.008 (0.044)	0.008 (0.044)	0.009 (0.044)
σ	-0.092 (0.043)	-0.004 (0.044)	-0.011 (0.044)	0.157 (0.046)

Table 4.2: Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the estimated bias and Monte Carlo simulation standard error (in parentheses) of maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all entries multiplied by 10 and with overall censoring percentage of approximately 45%.

	Maximum Likelihood	Indirect Inference	iRBM	eRBM
$n = 50$				
β_1	-0.099 (0.041)	0.010 (0.040)	-0.044 (0.040)	0.091 (0.040)
β_2	0.115 (0.037)	0.036 (0.036)	0.081 (0.036)	0.227 (0.038)
β_3	-0.098 (0.056)	-0.051 (0.055)	-0.111 (0.056)	-0.130 (0.056)
σ	-0.323 (0.042)	-0.008 (0.046)	-0.116 (0.045)	0.854 (0.058)
$n = 100$				
β_1	-0.077 (0.035)	0.000 (0.035)	-0.031 (0.035)	-0.002 (0.035)
β_2	0.042 (0.032)	-0.003 (0.032)	0.020 (0.032)	0.098 (0.033)
β_3	-0.011 (0.048)	-0.026 (0.048)	-0.018 (0.048)	0.015 (0.048)
σ	-0.160 (0.043)	-0.006 (0.045)	-0.037 (0.044)	0.444 (0.050)
$n = 150$				
β_1	-0.050 (0.034)	-0.003 (0.034)	-0.019 (0.034)	0.011 (0.034)
β_2	0.032 (0.031)	0.009 (0.031)	0.014 (0.031)	0.066 (0.031)
β_3	-0.012 (0.045)	-0.009 (0.045)	-0.009 (0.045)	-0.009 (0.045)
σ	-0.103 (0.043)	-0.007 (0.044)	-0.016 (0.044)	0.290 (0.048)

Table 4.3: Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the average estimated standard error and empirical standard error (in parentheses) of maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all entries multiplied by 10 and with overall censoring percentage of approximately 24%.

	Maximum Likelihood	Indirect Inference	iRBM	eRBM
$n = 50$				
β_1	1.486 (2.752)	1.565 (2.702)	1.539 (2.719)	1.679 (2.725)
β_2	1.227 (2.382)	1.294 (2.368)	1.271 (2.378)	1.382 (2.423)
β_3	2.136 (3.646)	2.260 (3.619)	2.217 (3.640)	2.415 (3.675)
σ	0.807 (2.972)	0.896 (3.210)	0.867 (3.138)	1.080 (3.708)
$n = 100$				
β_1	1.041 (2.460)	1.067 (2.427)	1.063 (2.435)	1.107 (2.447)
β_2	0.872 (2.203)	0.895 (2.196)	0.891 (2.198)	0.929 (2.223)
β_3	1.524 (3.193)	1.566 (3.195)	1.559 (3.193)	1.626 (3.183)
σ	0.580 (3.027)	0.610 (3.140)	0.605 (3.122)	0.666 (3.364)
$n = 150$				
β_1	0.834 (2.351)	0.848 (2.329)	0.846 (2.334)	0.868 (2.339)
β_2	0.677 (2.105)	0.689 (2.101)	0.688 (2.102)	0.706 (2.117)
β_3	1.246 (3.083)	1.269 (3.082)	1.266 (3.081)	1.300 (3.082)
σ	0.472 (3.039)	0.487 (3.111)	0.486 (3.106)	0.515 (3.250)

Table 4.4: Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the average estimated standard error and empirical standard error (in parentheses) of maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all entries multiplied by 10 and with overall censoring percentage of approximately 45%.

	Maximum Likelihood	Indirect Inference	iRBM	eRBM
$n = 50$				
β_1	1.750 (2.879)	1.850 (2.830)	1.811 (2.854)	2.085 (2.819)
β_2	1.503 (2.593)	1.578 (2.549)	1.547 (2.580)	1.753 (2.653)
β_3	2.539 (3.947)	2.685 (3.912)	2.623 (3.950)	2.982 (3.987)
σ	0.935 (2.979)	1.047 (3.240)	1.007 (3.155)	1.457 (4.113)
$n = 100$				
β_1	1.217 (2.485)	1.253 (2.449)	1.244 (2.462)	1.334 (2.452)
β_2	1.060 (2.283)	1.087 (2.262)	1.080 (2.274)	1.153 (2.309)
β_3	1.811 (3.413)	1.861 (3.422)	1.849 (3.415)	1.977 (3.402)
σ	0.672 (3.035)	0.710 (3.162)	0.702 (3.137)	0.830 (3.560)
$n = 150$				
β_1	0.980 (2.411)	0.998 (2.389)	0.995 (2.396)	1.042 (2.384)
β_2	0.820 (2.178)	0.834 (2.168)	0.832 (2.171)	0.869 (2.194)
β_3	1.473 (3.186)	1.499 (3.184)	1.495 (3.183)	1.562 (3.186)
σ	0.543 (3.050)	0.563 (3.130)	0.561 (3.122)	0.623 (3.381)

Table 4.5: Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 95% derived from Wald-type confidence intervals using the maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all coverage probabilities multiplied by 100 and with overall censoring percentage of approximately 24%.

	Maximum Likelihood	Indirect Inference	iRBM	eRBM
$n = 50$				
β_1	92.68 (0.577)	94.06 (0.608)	93.74 (0.598)	95.08 (0.641)
β_2	92.78 (0.476)	94.36 (0.502)	93.74 (0.493)	94.74 (0.530)
β_3	92.84 (0.828)	94.18 (0.877)	93.84 (0.860)	95.20 (0.922)
σ	89.24 (0.313)	94.54 (0.348)	92.88 (0.337)	93.56 (0.390)
$n = 100$				
β_1	93.52 (0.407)	94.56 (0.417)	94.46 (0.415)	95.10 (0.431)
β_2	94.14 (0.340)	94.82 (0.349)	94.48 (0.348)	95.08 (0.362)
β_3	93.64 (0.596)	94.44 (0.612)	94.38 (0.610)	95.24 (0.633)
σ	92.58 (0.226)	95.30 (0.238)	94.88 (0.236)	93.04 (0.256)
$n = 150$				
β_1	94.38 (0.326)	94.70 (0.331)	94.64 (0.330)	95.14 (0.339)
β_2	94.44 (0.265)	95.10 (0.269)	95.14 (0.269)	95.30 (0.275)
β_3	94.34 (0.487)	94.66 (0.495)	94.70 (0.495)	95.12 (0.507)
σ	93.22 (0.184)	95.00 (0.190)	94.90 (0.190)	93.28 (0.200)

Table 4.6: Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 95% derived from Wald-type confidence intervals using the maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all coverage probabilities multiplied by 100 and with overall censoring percentage of approximately 45%.

	Maximum Likelihood	Indirect Inference	iRBM	eRBM
$n = 50$				
β_1	92.46 (0.674)	94.44 (0.714)	93.52 (0.698)	95.62 (0.785)
β_2	92.94 (0.577)	94.80 (0.608)	93.70 (0.595)	95.80 (0.668)
β_3	93.40 (0.982)	94.86 (1.040)	94.20 (1.016)	96.04 (1.132)
σ	88.22 (0.361)	94.62 (0.404)	92.16 (0.389)	95.62 (0.501)
$n = 100$				
β_1	93.68 (0.473)	94.74 (0.487)	94.46 (0.484)	95.50 (0.516)
β_2	94.14 (0.413)	94.80 (0.423)	94.54 (0.421)	95.74 (0.448)
β_3	93.86 (0.706)	94.62 (0.726)	94.40 (0.721)	95.64 (0.767)
σ	91.68 (0.262)	94.48 (0.277)	93.98 (0.274)	92.88 (0.316)
$n = 150$				
β_1	94.30 (0.383)	94.86 (0.390)	94.88 (0.389)	95.80 (0.407)
β_2	94.70 (0.321)	95.24 (0.326)	94.98 (0.325)	95.92 (0.339)
β_3	94.66 (0.576)	95.08 (0.587)	95.06 (0.585)	96.08 (0.611)
σ	93.50 (0.212)	95.60 (0.220)	95.50 (0.219)	93.40 (0.241)

Table 4.7: Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 99% derived from Wald-type confidence intervals using the maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all coverage probabilities multiplied by 100 and with overall censoring percentage of approximately 24%.

	Maximum Likelihood	Indirect Inference	iRBM	eRBM
$n = 50$				
β_1	97.92 (0.759)	98.60 (0.799)	98.40 (0.785)	98.62 (0.842)
β_2	98.02 (0.625)	98.60 (0.659)	98.28 (0.648)	98.68 (0.696)
β_3	97.92 (1.089)	98.62 (1.153)	98.38 (1.131)	98.66 (1.212)
σ	95.22 (0.411)	98.24 (0.457)	97.54 (0.442)	97.76 (0.512)
$n = 100$				
β_1	98.48 (0.535)	98.76 (0.548)	98.70 (0.546)	98.86 (0.566)
β_2	98.64 (0.447)	98.80 (0.459)	98.90 (0.457)	99.06 (0.475)
β_3	98.62 (0.783)	98.88 (0.805)	98.88 (0.801)	99.08 (0.832)
σ	97.60 (0.298)	98.74 (0.313)	98.52 (0.311)	98.64 (0.336)
$n = 150$				
β_1	98.78 (0.428)	98.98 (0.435)	98.88 (0.434)	99.04 (0.445)
β_2	99.02 (0.348)	99.16 (0.354)	99.12 (0.353)	99.24 (0.362)
β_3	98.74 (0.640)	98.88 (0.651)	98.88 (0.650)	98.94 (0.666)
σ	98.02 (0.242)	98.82 (0.250)	98.82 (0.249)	98.60 (0.263)

Table 4.8: Weibull accelerated failure time observations with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.2$ and true standard deviation $\sigma = 0.667$. Second to fifth columns show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 99% derived from Wald-type confidence intervals using the maximum likelihood, indirect inference, iRBM and eRBM estimators, for sample sizes $n = 50$, $n = 100$ and $n = 150$ with all coverage probabilities multiplied by 100 and with overall censoring percentage of approximately 45%.

	Maximum Likelihood	Indirect Inference	iRBM	eRBM
$n = 50$				
β_1	97.78 (0.886)	98.72 (0.938)	98.38 (0.918)	99.00 (1.032)
β_2	98.32 (0.758)	98.86 (0.798)	98.58 (0.782)	99.20 (0.878)
β_3	98.68 (1.291)	99.22 (1.367)	99.00 (1.335)	99.34 (1.488)
σ	94.26 (0.474)	98.26 (0.531)	97.22 (0.511)	98.30 (0.659)
$n = 100$				
β_1	98.56 (0.622)	98.80 (0.640)	98.78 (0.636)	99.08 (0.678)
β_2	98.78 (0.542)	99.12 (0.556)	98.98 (0.553)	99.22 (0.589)
β_3	98.38 (0.928)	98.76 (0.954)	98.60 (0.948)	99.06 (1.008)
σ	96.66 (0.344)	98.24 (0.363)	97.98 (0.360)	98.54 (0.415)
$n = 150$				
β_1	98.66 (0.504)	98.92 (0.513)	98.92 (0.511)	99.22 (0.534)
β_2	98.54 (0.421)	98.74 (0.428)	98.68 (0.427)	98.82 (0.446)
β_3	99.08 (0.757)	99.18 (0.771)	99.20 (0.769)	99.28 (0.802)
σ	97.84 (0.279)	98.84 (0.289)	98.60 (0.288)	98.98 (0.317)

4.5 Analysis of lung cancer survival data

We analyse the Veterans' Administration lung cancer data of Prentice (1973) (see also Kalbfleisch and Prentice, 2002, Appendix A, data set I), where 137 males with advanced lung cancer were randomly assigned to either a standard or test chemotherapy treatment $(\text{standard} = 0, \text{test} = 1)$. The endpoint for treatment comparison was time to death. Only 9 of the 137 survival times were censored. In addition to treatment (txt), the data include information on a number of covariates thought to be relevant to an individual's prognosis: the patient's performance status (ps) at diagnosis (which is a measure of general medical condition on a scale of 10 to 90; 10-30 completely hospitalised, 40-60 partial confinement, 70-90 able to care for self), the number of months from diagnosis of cancer to entry into the study (diagage), the age of the patient at diagnosis in years (age), whether the patient received prior therapy (prior) $(0 = \text{no}, 10 = \text{yes})$, and tumor cell-type, classified as being in one of four categories, squamous, small cell, adeno and large. Survival times are measured as days from the date of entry to the study.

Preliminary analysis by Lawless (2002, Example 6.3.3) suggests that a Weibull accelerated failure time model may be suitable. We fit this model to these data with eight regressor variables (full model) where

$$\begin{aligned} Z\beta &= \beta_1 + \beta_2(\text{ps}) + \beta_3(\text{diagage}) + \beta_4(\text{age}) + \beta_5(\text{prior}) \\ &+ \beta_6\text{I}(\text{cell-type = squamous}) + \beta_7\text{I}(\text{cell-type = smallcell}) \\ &+ \beta_8\text{I}(\text{cell-type = adeno}) + \beta_9(\text{trt}). \end{aligned} \tag{4.55}$$

The coefficients β_6, β_7 and β_8 measure differences between each of the cell-type squamous, small, adeno and the baseline cell-type large. Table 4.9 shows estimates and standard errors (se) of the maximum likelihood, indirect inference, iRBM and eRBM estimates, along with the asymptotic normal Z statistic (not to be confused with the covariates), used to test the hypotheses $H : \beta_j = 0$ via $Z_j = (\hat{\beta}_j - 0)/\text{se}(\hat{\beta}_j)$, treating Z_j as approximately $N(0,1)$ if H is true. The standard errors of the indirect inference, iRBM and eRBM estimates were computed using the Hessian matrix.

From Table 4.9, it seems that only the patient's performance status (ps) has a strong prognostic effect on survival time. This conclusion is true under all four estimation methods. There is, also, no apparent dependence of survival time on age or disease duration (diagage) before entry to the clinical trial. The difference between the ML estimates and

the indirect inference, iRBM and eRBM estimates are small, indicating that estimation bias has not been a major concern in this study. Overall, the estimated standard errors for the indirect inference and iRBM estimators of β are only slightly inflated compared to those from the ML estimator, while the estimated standard errors for the eRBM estimator are slightly smaller compared to ML estimation.

Next we fit a reduced Weibull accelerated failure time model with only (ps) as the covariate where

$$Z\beta = \beta_1 + \beta_2(\text{ps}), \tag{4.56}$$

and the resulting ML, indirect inference, iRBM and eRBM estimates are reported in Table 4.10 along with their estimated standard errors and Z statistics. The estimates of the parameters of the reduced model from different estimation methods are in close agreement and the estimated standard errors are only slightly inflated compared to ML estimation.

The adequacy of the exponential regression model relative to the Weibull model, using the eight regressor variables, was discussed in Kalbfleisch and Prentice (2002, §3.7.2) by testing the hypothesis $\sigma = 1$ and concluded that the ML estimate of σ under the full Weibull model $\hat{\sigma} = 0.9281$ and its estimated standard error of 0.0615 provides no evidence against the exponential model relative to the Weibull model with a Z statistic of 15.08. Testing the same hypothesis using the indirect inference, iRBM and eRBM estimation methods yields the same conclusion, i.e. provides no evidence against the exponential model. Testing the adequacy of the exponential model under the reduced model fitting also provides no evidence against the hypothesis $\sigma = 1$, for all estimation methods.

Table 4.9: Fitted full Weibull accelerated failure time model to the lung cancer data using maximum likelihood (ML), indirect inference (II), iRBM and eRBM.

	ML			II			iRBM			eRBM		
Parameter	Estimate	se	Z	Estimate	se	Z	Estimate	se	Z	Estimate	se	Z
β_1 (int)	3.0929	0.6836	4.52	3.1227	0.7166	4.36	3.0976	0.7129	4.35	3.0137	0.6544	4.60
β_2 (ps)	0.0301	0.0048	6.23	0.0296	0.0051	5.85	0.0297	0.0050	5.90	0.0306	0.0046	6.60
β_3 (diagage)	−0.0005	0.0084	−0.06	0.0007	0.0091	0.08	0.0001	0.0090	0.01	0.0002	0.0081	0.02
β_4 (age)	0.0061	0.0086	0.71	0.0060	0.0090	0.67	0.0062	0.0089	0.70	0.0060	0.0082	0.74
β_5 (prior)	−0.0044	0.0212	−0.21	−0.0053	0.0223	−0.24	−0.0043	0.0222	−0.20	−0.0046	0.0203	−0.23
β_6 (squamous)	0.3977	0.2547	1.56	0.3941	0.2672	1.47	0.4147	0.2656	1.56	0.3848	0.2429	1.58
β_7 (small)	−0.4285	0.2433	−1.76	−0.4375	0.2555	−1.71	−0.4101	0.2540	−1.61	−0.4169	0.2322	−1.80
β_8 (adeno)	−0.7350	0.2741	−2.68	−0.7257	0.2878	−2.52	−0.7284	0.2847	−2.56	−0.7303	0.2627	−2.78
β_9 (trt)	−0.2285	0.1868	−1.22	−0.2217	0.1957	−1.13	−0.2269	0.1949	−1.16	−0.2082	0.1790	−1.16
σ	0.9281	0.0615	15.08	0.9648	0.0679	14.21	0.9632	0.0676	14.24	0.8976	0.0565	15.88

Table 4.10: Fitted reduced Weibull accelerated failure time model to the lung cancer data using maximum likelihood (ML), indirect inference (II), iRBM and eRBM.

	ML			II			iRBM			eRBM		
Parameter	Estimate	se	Z	Estimate	se	Z	Estimate	se	Z	Estimate	se	Z
β_1 (int)	2.6447	0.2956	8.95	2.1277	0.3106	6.85	2.6536	0.2999	8.85	2.6428	0.3101	8.52
β_2 (ps)	0.0350	0.0048	7.26	0.0391	0.0049	7.92	0.0349	0.0049	7.13	0.0351	0.0051	6.93
σ	1.0225	0.0664	15.41	1.1667	0.0883	13.21	1.0343	0.0684	15.13	1.0678	0.0740	14.42

4.6 Discussion and further work

4.6.1 Summary

It has been shown that the implementation of the bias-reduction methods of indirect inference and empirical bias reducing penalised log-likelihood function for the Weibull accelerated failure time model can be easily performed. The above two methods are an improvement over the usual maximum likelihood in terms of bias and coverage probabilities of Wald type confidence intervals. These methods, although focused on the Weibull model, can be easily extended to other types of accelerated failure time models, such as the log-normal or log-logistic accelerated failure time models. Our simulation study can be extended by considering alternative values of β and σ. Moreover, the Wald type confidence intervals which use the normal approximations can be inaccurate for small samples, and so confidence intervals based on the likelihood ratio statistic or the parametric bootstrap procedures can perform better (are more accurate) in censored small or medium-size samples (Jeng and Meeker, 2000). In summary, we have demonstrated that the indirect inference estimator and the empirical bias-reducing penalised log-likelihood estimator (iRBM) provide improvement in reducing small sample bias over the traditional maximum likelihood in the Weibull accelerated failure time model for censored survival data.

4.6.2 Bias reduction for frailty models

The Weibull AFT model described in Section 4.2 (or any general AFT model) may be extended by considering a random frailty effect u_i (of expectation one) to account for heterogeneity between different items or individuals. These models can be written analogously to (4.26) in the form

$$Y = \ln(T) = u + Z\beta + \sigma W. \tag{4.57}$$

In other words, frailty is a random component designed to account for variability due to unobserved individual-level factors that is otherwise unaccounted for by the other independent variables (covariates) in the model. One way to do this is to model heterogeneity in the parametric model as described in Section 19.4.3.e of Greene (2004), by considering a survival function conditioned on the individual specific effect u_i and treat the survival function as $S(t_i|u_i)$. Then consider a model for the unobserved heterogeneity $f(u_i)$. Once

the frailty distribution $f(u_i)$ is chosen, the unconditional survival function is found by

$$S(t) = \int_0^\infty S(t|u) f(u)\, du. \tag{4.58}$$

The corresponding unconditional hazard function can then be found using the relationship (4.3). The gamma distribution with mean 1 and variance θ is a common choice for the distribution $f(u)$. The likelihood function with frailty effect can then be formulated using the unconditional probability density function which is the product of the unconditional hazard and survival functions. The likelihood is constructed in a similar manner to that described in Section 4.1.3 except that the unconditional probability density is used rather than $f(t)$ in (4.24). This means that there is one additional parameter to estimate, the variance of the frailty, θ. It is of interest to investigate the performance of the bias reduction methods such as indirect inference and the empirical bias reducing penalty in terms of reducing the estimation bias of θ in the frailty Weibull AFT model.

Furthermore, the Weibull accelerated failure time model that we considered in Section 4.2 has a scale parameter σ that does not depend on the covariates i.e. $\mathrm{Var}(Y_i|Z)$ was constant for all i. This assumption is sometimes unsuitable, and we may want to specify some form of dependency of σ on the covariates Z. A common choice is $\sigma(Z) = \exp(\tilde{Z}\gamma)$, since σ has to be nonnegative, where γ is a vector of parameters. The log-likelihood function for this model from a censored random sample is a direct generalisation of (4.29),

$$l(\beta, \gamma) = \sum_{i=1}^{n} \big\{ -\delta_i \ln(\sigma_i) + \delta_i \ln f(w_i) + (1-\delta_i) \ln S(w_i) \big\}, \tag{4.59}$$

where $w_i = (y_i - z_i^\intercal \beta)/\exp(\tilde{z}_i{}^\intercal \gamma)$. Lawless (2002, §6.4.2) give the first and second derivatives of the above log-likelihood for a general $f(w_i)$ and $S(w_i)$. The methods of indirect inference and empirical bias-reducing penalty we discussed in this chapter could be implemented to the above model of variable scale parameter and compared with ML estimation.

4.6.3 Empirical bias-reducing penalty for a general accelerated failure time model

The method of iRBM-estimation can in principle be applied to any accelerated failure time model, not just the Weibull one, by penalising the log-likelihood function (4.29). We derive below the components of the penalised log-likelihood (2.48) for a general density $f(w_i)$ and a general survival function $S(w_i)$. Differentiating (4.29) with respect to β and

σ, respectively gives (Kalbfleisch and Prentice, 2002, §3.6)

$$\begin{aligned}
\nabla_\beta l(\theta) &= \sum_{i=1}^{n} \frac{1}{\sigma} z_i a_i \\
&= \frac{1}{\sigma} Z^\top a \mathbf{1}_n, \qquad (4.60) \\
\frac{\partial}{\partial \sigma} l(\theta) &= \frac{1}{\sigma} \sum_{i=1}^{n} \big(w_i a_i - \delta_i \big) \\
&= \frac{1}{\sigma} \mathbf{1}_n^\top \big[Wa - D \big] \mathbf{1}_n, \qquad (4.61)
\end{aligned}$$

where $a = \text{diag}\{a_1, \ldots, a_n\}$, $D = \text{diag}\{\delta_1, \ldots, \delta_n\}$ and where

$$a_i = -\delta_i \frac{d \ln f(w_i)}{dw_i} + (1 - \delta_i) \lambda(w_i), \qquad (4.62)$$

where $\lambda(w_i) = f(w_i)/S(w_i)$ as defined in (4.2) and (4.3). The $(p+1) \times (p+1)$ matrices $j(\theta)$ and $e(\theta)$ are given by

$$j(\theta) = \begin{pmatrix} j_{\beta\beta} & j_{\beta\sigma} \\ (j_{\beta\sigma})^\top & j_{\sigma\sigma} \end{pmatrix} \quad \text{and} \quad e(\theta) = \begin{pmatrix} e_{\beta\beta} & e_{\beta\sigma} \\ (e_{\beta\sigma})^\top & e_{\sigma\sigma} \end{pmatrix},$$

where

$$\begin{aligned}
j_{\beta\beta} &= -\nabla_\beta \nabla_\beta^\top l(\theta) \\
&= \sum_{i=1}^{n} \frac{1}{\sigma^2} A_i z_i z_i^\top \\
&= \frac{1}{\sigma^2} Z^\top A Z, \qquad (4.63) \\
j_{\beta\sigma} &= -\frac{\partial}{\partial \sigma} \nabla_\beta l(\theta) \\
&= \frac{1}{\sigma^2} \sum_{i=1}^{n} z_i \big(w_i A_i + a_i \big) \\
&= \frac{1}{\sigma^2} Z^\top \big[WA + a \big] \mathbf{1}_n, \qquad (4.64)
\end{aligned}$$

$$
\begin{aligned}
j_{\sigma\sigma} &= -\frac{\partial^2}{\partial(\sigma)^2} l(\theta) \\
&= \frac{1}{\sigma^2} \sum_{i=1}^{n} \left\{ w_i^2 A_i + 2 a_i w_i - \delta_i \right\} \\
&= \frac{1}{\sigma^2} \mathbf{1}_n^\intercal \left[W^2 A + 2aW - D \right] \mathbf{1}_n, \qquad (4.65)
\end{aligned}
$$

where

$$
\begin{aligned}
A_i &= \frac{da_i}{dw_i} \\
&= -\delta_i \frac{d^2 \ln f(w_i)}{dw_i^2} + (1-\delta_i) \left[\lambda(w_i) \frac{d \ln f(w_i)}{dw_i} + \lambda^2(w_i) \right], \qquad (4.66)
\end{aligned}
$$

and where $A = \text{diag}\{A_1 \ldots, A_n\}$. In addition,

$$
\begin{aligned}
e_{\beta\beta} &= \frac{1}{\sigma^2} Z^\intercal a^2 Z, \qquad (4.67) \\
e_{\beta\sigma} &= \frac{1}{\sigma^2} Z^\intercal a \left[Wa - D \right] \mathbf{1}_n, \qquad (4.68) \\
\left(e_{\beta\sigma} \right)^\intercal &= \frac{1}{\sigma^2} \mathbf{1}_n^\intercal \left[aW - D \right] aZ, \qquad (4.69) \\
e_{\sigma\sigma} &= \frac{1}{\sigma^2} \mathbf{1}_n^\intercal \left[Wa - D \right]^2 \mathbf{1}_n. \qquad (4.70)
\end{aligned}
$$

Chapter 5

Stratified models

5.1 Introduction

Consider inference about a scalar parameter, ψ, of interest based on models with m_i observations from each of q independent strata in the presence of nuisance (incidental) parameters λ_i, $i = (1,\ldots,q)$, one for each stratum. Such models are known as stratified models (Sartori, 2003), and it is well known, since Neyman and Scott (1948), that the maximum likelihood estimator, derived from the profile log-likelihood, is not, in general, a consistent estimator of ψ as the dimension of the nuisance parameter increases while the stratum sample size is kept fixed; this is known as the incidental parameter problem.

It is possible to solve this problem, in some cases when the model has a particular structure, like in exponential families in canonical form, using conditional or marginal log-likelihoods (Pace and Salvan, 1997, §4.4 and §4.5). However, these are not always available and so an alternative is to work with the approximate conditional profile log-likelihood of Cox and Reid (1987) or the modified profile log-likelihood of Barndorff-Nielson (1983) which involve only the information matrix for nuisance parameters. The modified profile log-likelihood can also be difficult to compute in some models because it requires the calculation of a sample space derivative. The adjusted profile log-likelihood of McCullagh and Tibshirani (1990) is another simple alternative which adjusts the profile log-likelihood score function so that it is unbiased and information unbiased. Another method that have been used to reduce the bias of the maximum likelihood estimator in stratified settings in the presence of a large number of nuisance parameters is the adjusted score equations approach of Firth (1993).

In this Chapter we consider two stratified models, namely the matched gamma pairs

model (see Sartori, 2003, Example 4) in Section 5.2, and the binomial matched pairs model (see Sartori, 2003; Lunardon, 2018, Example 3 and §4.1, respectively) in Section 5.3, and compare and contrast different methods of estimation of the parameter of interest. In the matched gamma pairs model, there is no exact conditional or marginal log-likelihood for the parameter of interest, and Sartori (2003) compared the profile log-likelihood and the modified profile log-likelihood of Barndorff-Nielson (1983), which coincides with the approximate conditional profile log-likelihood of Cox and Reid (1987), through a simulation study and showed that the modified profile log-likelihood leads to superior inference especially for large numbers of nuisance parameters. Cox and Reid (1992) showed that the maximum modified profile log-likelihood estimator of ψ is less biased than the maximum likelihood estimator. We review these methods of estimation and further compare them with the adjusted profile log-likelihood estimator of McCullagh and Tibshirani (1990) and show that the latter is exactly unbiased and consistent. We derive the asymptotic bias corrected estimator, the adjusted score equations estimator of Firth (1993) and the indirect inference estimator of Kuk (1995), of the parameter of interest ψ, and show that the latter coincides exactly with the unbiased adjusted profile log-likelihood estimator of McCullagh and Tibshirani (1990). We find that for a fixed stratum size, the asymptotic bias corrected estimator produces a substantial improvement over the modified profile log-likelihood estimator in terms of bias especially when the stratum sample size is small. The expected and observed adjusted score equations estimators defined in (2.38) and (2.39) are less biased than the maximum likelihood estimator but in comparison to the other estimators, the difference in bias is negligible for large numbers of nuisance parameters.

For the binomial matched pairs model, we review in Section 5.3.1 the profile, conditional, modified profile and Firth (1993) penalised likelihood estimators of ψ. Section 5.3.2 is a review of current methods of estimation in the special case of the binary matched pairs model. Since the ML estimator of ψ can be infinite, we propose in Section 5.3.3 a penalised log-likelihood function based on adjusted responses that always yields finite point estimates of the parameter of interest. In Section 5.3.3.1, we derive the probability limit of that estimator while in Section 5.3.4, the indirect inference procedure is described and applied to reduce the bias of the penalised maximum likelihood estimator based on adjusted responses. The exact properties of the above estimators are obtained through complete enumeration as in Lunardon (2018) where no simulation is required. Results and discussion are given in Section 5.3.5 and in Section 5.3.6 we analyse a real data set.

5.2 Matched gamma pairs model

Consider the independent exponential random variables $Y_{ij1} \sim \text{Exp}\big(\lambda_i/\psi\big)$ and $Y_{ij2} \sim \text{Exp}\big(1/(\psi\lambda_i)\big)$, $i = 1,\ldots,q$, with rates λ_i/ψ and $1/(\psi\lambda_i)$ or scales (means) ψ/λ_i and $\psi\lambda_i$, respectively, as in Sartori (2003, Example 4). This implies that $Y_{i1} = \sum_{j=1}^{m} Y_{ij1} \sim \text{Gamma}\big(m, \lambda_i/\psi\big)$ and $Y_{i2} = \sum_{j=1}^{m} Y_{ij2} \sim \text{Gamma}\big(m, 1/(\psi\lambda_i)\big)$. In other words, (Y_{i1}, Y_{i2}) are considered as matched gamma pairs with shape parameter m and rates λ_i/ψ and $1/(\psi\lambda_i)$, respectively. Note that Y_{i1} and Y_{i2} have expected values $m\psi/\lambda_i$ and $m\psi\lambda_i$, respectively. The stratum sample size is $2m$ and so the sample size is $n = \sum_{i=1}^{q} 2m = 2qm$. The parameter of interest, ψ, is the inverse square root of the product of the rates, while the nuisance parameters are the square root of the ratio of the rates.

5.2.1 Maximum likelihood estimation

The full log-likelihood function for the above matched gamma pairs model, up to an additive constant is

$$l(\psi, \lambda_i) = -2mq\ln(\psi) - \frac{1}{\psi}\sum_{i=1}^{q}\big(\lambda_i y_{i1} + (y_{i2}/\lambda_i)\big). \tag{5.1}$$

The partial derivative of the full log-likelihood with respect to λ_i is

$$\frac{\partial l(\psi, \lambda_i)}{\partial \lambda_i} = -\frac{y_{i1}}{\psi} + \frac{y_{i2}}{\psi\lambda_i^2}.$$

Hence, equating the above to zero and solving for λ_i we find that the constrained ML estimator of λ_i is $\hat{\lambda}_{i,\psi} = \big(y_{i2}/y_{i1}\big)^{1/2}$. The profile log-likelihood for ψ, obtained by substituting $\hat{\lambda}_{i,\psi}$ in (5.1) is

$$l_p(\psi) = -2mq\ln(\psi) - \frac{2}{\psi}\sum_{i=1}^{q}\big(y_{i1}y_{i2}\big)^{1/2}. \tag{5.2}$$

The partial derivative of $l_p(\psi)$ with respect to ψ is

$$\frac{\partial l_p(\psi)}{\partial \psi} = -\frac{2mq}{\psi} + \frac{2}{\psi^2}\sum_{i=1}^{q}\big(y_{i1}y_{i2}\big)^{1/2}.$$

Equating the above to zero and solving for ψ we find that the overall ML estimator of ψ is (Sartori, 2003, Example 4)

$$\hat{\psi} = \frac{1}{mq} \sum_{i=1}^{q} \left(y_{i1} y_{i2}\right)^{1/2}. \tag{5.3}$$

Using integration and exploiting the probability density functions of the Gamma$\big((m+1/2), \lambda_i/\psi\big)$ and Gamma$\big((m+1/2), 1/(\psi\lambda_i)\big)$ variates, we find that the expected value of $\left(Y_{i1}Y_{i2}\right)^{1/2}$ is

$$\mathrm{E}\Big(\left(Y_{i1}Y_{i2}\right)^{1/2}\Big) = \frac{\psi\big(\Gamma(m+1/2)\big)^2}{\big(\Gamma(m)\big)^2}. \tag{5.4}$$

Hence as noted in Cox and Reid (1992, Example 3), $\mathrm{E}\big(\hat{\psi}\big) = \psi\big(\Gamma(m+1/2)\big)^2/m\big(\Gamma(m)\big)^2 \neq \psi$ so $\hat{\psi}$ is a biased estimator of ψ. Moreover, $\mathrm{E}\big(\hat{\psi}\big) \nrightarrow \psi$ as $q \to \infty$ so $\hat{\psi}$ is also inconsistent. Note that when $m = 1$, we have a pair of exponential matched pairs with $\mathrm{E}\big(\hat{\psi}\big) = \psi\pi/4$.

Similarly, using integration and exploiting the probability density functions of the Gamma$\big((m+1/2), \lambda_i/\psi\big)$ and Gamma$\big((m+1/2), 1/(\psi\lambda_i)\big)$ variates again, and using the identity $\Gamma(z) = (z-1)\cdot\Gamma(z-1)$, $\forall z \in \mathbb{R}$, we find that

$$\mathrm{E}\big(Y_{i1}Y_{i2}\big) = \psi^2 m^2. \tag{5.5}$$

Combining (5.4) and (5.5) we obtain

$$\mathrm{Var}\Big(\left(Y_{i1}Y_{i2}\right)^{1/2}\Big) = \frac{\psi^2\Big(m^2\big(\Gamma(m)\big)^4 - \big(\Gamma(m+1/2)\big)^4\Big)}{\big(\Gamma(m)\big)^4}. \tag{5.6}$$

When $m = 1$, the variance of the ML estimator of the parameter of interest for the matched exponential pairs becomes $\mathrm{Var}\big(\hat{\psi}\big) = \psi^2\big(1-\pi^2\big)/16q$.

5.2.2 Modified likelihood functions

To derive the approximate conditional profile log-likelihood we first check orthogonality of ψ and λ_i then calculate the term $j_{\lambda\lambda}$ as follows: for all $i = 1, \ldots, q$ we have

$$\frac{\partial l(\psi, \lambda_i)}{\partial \lambda_i} = \frac{1}{\psi}\Big(\frac{1}{\lambda_i^2} y_{i2} - y_{i1}\Big),$$

and

$$\frac{\partial^2 l(\psi,\lambda_i)}{\partial \lambda_i^2} = -\frac{2}{\psi\lambda_i^3} y_{i2}\,.$$

Hence, $\mathrm{E}\big(-\partial^2 l(\psi,\lambda_i)/\partial\psi\partial\lambda_i\big) = \mathrm{E}\big(\{(Y_{i2}/\lambda_i^2) - Y_{i1}\}/\psi^2\big) = \big(\{\mathrm{E}(Y_{i2})/\lambda_i^2\} - \mathrm{E}(Y_{i1})\big)/\psi^2$ $= \big((m\psi\lambda_i/\lambda_i^2) - (m\psi/\lambda_i)\big)/\psi^2 = \big((m\psi/\lambda_i) - (m\psi/\lambda_i)\big)/\psi^2 = 0$, i.e. ψ and λ_i are orthogonal. For all $i, j = 1,\ldots,q$, $i \neq j$ we have

$$\frac{\partial^2 l(\psi,\lambda_i)}{\partial \lambda_i \partial \lambda_j} = 0\,.$$

Therefore,

$$j_{\lambda\lambda}(\psi,\lambda_i) = \frac{2}{\psi}\begin{pmatrix} \frac{y_{12}}{\lambda_1^3} & 0 & 0 & \cdots & 0 \\ 0 & \frac{y_{22}}{\lambda_2^3} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{y_{q2}}{\lambda_q^3} \end{pmatrix}$$

and so $\det j_{\lambda\lambda}(\psi,\lambda_i) = (2/\psi)^q \prod_{i=1}^q (y_{i2}/\lambda_i^3)$ since if α is a scalar and A is a $q \times q$ matrix, then $\det(\alpha \cdot A) = \alpha^q \det(A)$. The approximate conditional profile log-likelihood for ψ is (Sartori, 2003, Example 4)

$$\begin{aligned} l_{cp}(\psi) &= l_p(\psi) - \frac{1}{2}\ln\{\det j_{\lambda\lambda}(\psi,\hat{\lambda}_{i,\psi})\} \\ &= -2mq\ln(\psi) - \frac{2}{\psi}\sum_{i=1}^{q}\big(y_{i1}y_{i2}\big)^{1/2} + \frac{q}{2}\ln(\psi)\,. \end{aligned}$$

The partial derivative of the conditional profile log-likelihood with respect to ψ is given by

$$\frac{\partial l_{cp}(\psi)}{\partial \psi} = -\frac{(4mq-q)}{2\psi} + \frac{2}{\psi^2}\sum_{i=1}^{q}\big(y_{i1}y_{i2}\big)^{1/2}.$$

Equating this to zero and solving for ψ we find that the conditional profile log-likelihood estimator of ψ is (Sartori, 2003, Example 4)

$$\begin{aligned}\hat{\psi}_{cp} &= \frac{4}{4mq-q}\sum_{i=1}^{q}\left(y_{i1}y_{i2}\right)^{1/2} \\ &= \frac{4m}{4m-1}\hat{\psi}. \end{aligned} \tag{5.7}$$

However, $\hat{\psi}_{cp}$ is still biased and inconsistent since $\mathrm{E}\big(\hat{\psi}_{cp}\big) = \big((4m)/(4m-1)\big)\mathrm{E}\big(\hat{\psi}\big) = 4\psi\big(\Gamma(m+1/2)\big)^2/(4m-1)\big(\Gamma(m)\big)^2 \neq \psi$ and also $\mathrm{E}\big(\hat{\psi}_{cp}\big) \nrightarrow \psi$ as $q \to \infty$. Note that when $m=1$, $\hat{\psi}_{cp} = 4\hat{\psi}/3$, $\mathrm{E}\big(\hat{\psi}_{cp}\big) = \psi\pi/3$ and $\mathrm{Var}\big(\hat{\psi}_{cp}\big) = \psi^2\big(1-\pi^2\big)/9q$. So if $q \to \infty$ both estimators $\hat{\psi}$ and $\hat{\psi}_{cp}$ are inconsistent but $\hat{\psi}_{cp}$ has substantially smaller bias than $\hat{\psi}$ as noticed by Cox and Reid (1992).

Since $\hat{\lambda}_{i,\psi}$ is not a function of ψ, $\hat{\lambda}_i = \hat{\lambda}_{i,\hat{\psi}} = \big(y_{i2}/y_{i1}\big)^{1/2} = \hat{\lambda}_{i,\psi}$. Therefore we can immediately deduce that $l_{mp}(\psi) = l_{cp}(\psi)$, as noted in Section 2.1.5.

To derive the adjusted profile log-likelihood in (2.11) we first evaluate the adjusted score function and then integrate as follows:

$$U(\psi) = \frac{\partial l_p(\psi)}{\partial\psi} = -\frac{2mq}{\psi} + \frac{2}{\psi^2}\sum_{i=1}^{q}\left(y_{i1}y_{i2}\right)^{1/2},$$

$$\mathrm{E}(U) = -\frac{2mq}{\psi} + \frac{2q\big(\Gamma(m+1/2)\big)^2}{\psi\big(\Gamma(m)\big)^2} = \frac{-2mq\big(\Gamma(m)\big)^2 + 2q\big(\Gamma(m+1/2)\big)^2}{\psi\big(\Gamma(m)\big)^2},$$

$$\mathrm{Var}(U) = \frac{4q\Big(m^2\big(\Gamma(m)\big)^4 - \big(\Gamma(m+1/2)\big)^4\Big)}{\psi^2\big(\Gamma(m)\big)^4},$$

$$\frac{\partial U}{\partial\psi} = \frac{2mq}{\psi^2} - \frac{4}{\psi^3}\sum_{i=1}^{q}\left(y_{i1}y_{i2}\right)^{1/2},$$

$$\frac{\partial\mathrm{E}(U)}{\partial\psi} = \frac{2mq\big(\Gamma(m)\big)^2 - 2q\big(\Gamma(m+1/2)\big)^2}{\psi^2\big(\Gamma(m)\big)^2},$$

$$-\mathrm{E}\left(\frac{\partial U}{\partial\psi}\right) = -\frac{2mq}{\psi^2} + \frac{4q\big(\Gamma(m+1/2)\big)^2}{\psi^2\big(\Gamma(m)\big)^2} = \frac{-2mq\big(\Gamma(m)\big)^2 + 4q\big(\Gamma(m+1/2)\big)^2}{\psi^2\big(\Gamma(m)\big)^2}.$$

So the adjustments turn out to be

$$m(\psi) = \frac{-2mq\big(\Gamma(m)\big)^2 + 2q\big(\Gamma(m+1/2)\big)^2}{\psi\big(\Gamma(m)\big)^2},$$

$$w(\psi) = \frac{\big(\Gamma(m+1/2)\big)^2\big(\Gamma(m)\big)^2}{2\Big(m^2\big(\Gamma(m)\big)^4 - \big(\Gamma(m+1/2)\big)^4\Big)}.$$

The adjusted profile log-likelihood for ψ is therefore given by

$$\begin{aligned} l_{ap}(\psi) &= \int \left(\frac{2}{\psi^2}\sum_{i=1}^{q}\big(y_{i1}y_{i2}\big)^{1/2} - \frac{2q\big(\Gamma(m+1/2)\big)^2}{\psi\big(\Gamma(m)\big)^2}\right)\cdot w(\psi)\,dt \\ &= \frac{-\big(\Gamma(m+1/2)\big)^2\big(\Gamma(m)\big)^2}{\psi\Big(m^2\big(\Gamma(m)\big)^4 - \big(\Gamma(m+1/2)\big)^4\Big)}\sum_{i=1}^{q}\big(y_{i1}y_{i2}\big)^{1/2} \\ &- \frac{q\big(\Gamma(m+1/2)\big)^4}{\Big(m^2\big(\Gamma(m)\big)^4 - \big(\Gamma(m+1/2)\big)^4\Big)}\ln(\psi). \end{aligned}$$

The partial derivative of the adjusted profile log-likelihood with respect to ψ is given by

$$\begin{aligned} \frac{\partial l_{ap}(\psi)}{\partial\psi} &= \frac{\big(\Gamma(m+1/2)\big)^2\big(\Gamma(m)\big)^2}{\psi^2\Big(m^2\big(\Gamma(m)\big)^4 - \big(\Gamma(m+1/2)\big)^4\Big)}\sum_{i=1}^{q}\big(y_{i1}y_{i2}\big)^{1/2} \\ &- \frac{q\big(\Gamma(m+1/2)\big)^4}{\psi\Big(m^2\big(\Gamma(m)\big)^4 - \big(\Gamma(m+1/2)\big)^4\Big)}. \end{aligned}$$

Equating the above to zero and solving for ψ we find that the adjusted profile log-likelihood estimator of ψ is given by

$$\begin{aligned} \hat{\psi}_{ap} &= \frac{\big(\Gamma(m)\big)^2}{q\big(\Gamma(m+1/2)\big)^2}\sum_{i=1}^{q}\big(y_{i1}y_{i2}\big)^{1/2} \\ &= \frac{m\big(\Gamma(m)\big)^2}{\big(\Gamma(m+1/2)\big)^2}\hat{\psi}. \end{aligned} \tag{5.8}$$

In contrast to the ML and conditional profile log-likelihood estimators of ψ, $\hat{\psi}_{ap}$ is both unbiased and consistent since $\mathrm{E}\big(\hat{\psi}_{ap}\big) = \Big(m\big(\Gamma(m)\big)^2/\big(\Gamma(m+1/2)\big)^2\Big)\mathrm{E}\big(\hat{\psi}\big) = \psi$, and

$\mathrm{Var}\left(\hat{\psi}_{ap}\right) = \psi^2\left(m^2\left(\Gamma(m)\right)^4 - \left(\Gamma(m+1/2)\right)^4\right)\Big/ q\left(\Gamma(m+1/2)\right)^4 \to 0$ as $q \to \infty$. When $m = 1$, $\hat{\psi}_{ap} = 4\hat{\psi}/\pi$, $\mathrm{E}\left(\hat{\psi}_{ap}\right) = \psi$ and $\mathrm{Var}\left(\hat{\psi}_{ap}\right) = \psi^2\left(1-\pi^2\right)/q\pi^2$.

5.2.3 Reduced bias estimation of the parameter of interest

The second order derivatives of $l(\psi, \lambda_i)$ are given by

$$\begin{aligned}
\frac{\partial^2 l(\psi,\lambda_i)}{\partial \psi^2} &= \frac{2mq}{\psi^2} - \frac{2}{\psi^3}\sum_{i=1}^{q}\left(\lambda_i y_{i1} + (y_{i2}/\lambda_i)\right) \\
&= -\frac{2}{\psi^3}\sum_{i=1}^{q}\left[\lambda_i\left(y_{i1} - \frac{m\psi}{\lambda_i}\right) + \frac{1}{\lambda_i}\left(y_{i2} - \lambda_i m\psi\right) + m\psi\right], \\
\frac{\partial^2 l(\psi,\lambda_i)}{\partial \psi \partial \lambda_i} &= \frac{1}{\psi^2}\left(y_{i1} - \frac{y_{i2}}{\lambda_i^2}\right), \\
\frac{\partial^2 l(\psi,\lambda_i)}{\partial \lambda_i^2} &= -\frac{2y_{i2}}{\psi\lambda_i^3}, \\
\frac{\partial^2 l(\psi,\lambda_i)}{\partial \lambda_i \partial \lambda_j} &= 0,
\end{aligned}$$

and their expected values are

$$\begin{aligned}
\mathrm{E}\left(\frac{\partial^2 l(\psi,\lambda_i)}{\partial \psi^2}\right) &= -\frac{2mq}{\psi^2}, \\
\mathrm{E}\left(\frac{\partial^2 l(\psi,\lambda_i)}{\partial \psi \partial \lambda_i}\right) &= 0, \\
\mathrm{E}\left(\frac{\partial^2 l(\psi,\lambda_i)}{\partial \lambda_i^2}\right) &= -\frac{2m}{\lambda_i^2}, \\
\mathrm{E}\left(\frac{\partial^2 l(\psi,\lambda_i)}{\partial \lambda_i \partial \lambda_j}\right) &= 0.
\end{aligned}$$

The observed information matrix is therefore given by

$$j(\psi,\lambda_i) = \begin{pmatrix}
-z & \frac{1}{\psi^2}\left(\frac{y_{12}}{\lambda_1^2} - y_{11}\right) & \cdots & \frac{1}{\psi^2}\left(\frac{y_{q2}}{\lambda_q^2} - y_{q1}\right) \\
\frac{1}{\psi^2}\left(\frac{y_{12}}{\lambda_1^2} - y_{11}\right) & \frac{2y_{12}}{\psi\lambda_1^3} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\frac{1}{\psi^2}\left(\frac{y_{q2}}{\lambda_q^2} - y_{q1}\right) & 0 & \cdots & \frac{2y_{q2}}{\psi\lambda_q^3}
\end{pmatrix},$$

where z is the second order derivative of $l(\psi,\lambda_i)$ with respect to ψ, and the Fisher information matrix is

$$i(\psi,\lambda_i) = \begin{pmatrix} \frac{2mq}{\psi^2} & 0 & \cdots & 0 \\ 0 & \frac{2m}{\lambda_1^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{2m}{\lambda_q^2} \end{pmatrix}.$$

Since $i(\psi,\lambda_i)$ is diagonal, the inverse fisher information matrix is given by

$$\{i(\psi,\lambda_i)\}^{-1} = \begin{pmatrix} \frac{\psi^2}{2mq} & 0 & \cdots & 0 \\ 0 & \frac{\lambda_1^2}{2m} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\lambda_q^2}{2m} \end{pmatrix}.$$

The score function is given by

$$\begin{aligned} S(\psi,\lambda_i) &= \begin{pmatrix} -\frac{2mq}{\psi} + \frac{1}{\psi^2}\sum_{i=1}^{q}\left[\lambda_i y_{i1} + (y_{i2}/\lambda_i)\right] \\ \frac{1}{\psi}\left(\frac{y_{12}}{\lambda_1^2} - y_{11}\right) \\ \vdots \\ \frac{1}{\psi}\left(\frac{y_{q2}}{\lambda_q^2} - y_{q1}\right) \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\psi^2}\sum_{i=1}^{q}\left[\lambda_i\left(y_{i1} - \frac{m\psi}{\lambda_i}\right) + \frac{1}{\lambda_i}\left(y_{i2} - \lambda_i m\psi\right)\right] \\ \frac{1}{\psi}\left(\frac{y_{12}}{\lambda_1^2} - y_{11}\right) \\ \vdots \\ \frac{1}{\psi}\left(\frac{y_{q2}}{\lambda_q^2} - y_{q1}\right) \end{pmatrix}. \end{aligned}$$

To make presentation simpler, let

$$\begin{aligned} x &= S_\psi(\psi,\lambda_i) = \partial l(\psi,\lambda_i)/\partial\psi, \\ w_i &= S_{\lambda_i}(\psi,\lambda_i) = \partial l(\psi,\lambda_i)/\partial\lambda_i, \end{aligned}$$

where $i = 1\ldots,q$.

Using the central moments of the gamma distribution, a bit of algebra and independence of the observations (see Appendix B), it may now be verified that (2.17) yields for the ψ and λ_i components of $P(\psi,\lambda_i)$, $i=1,\ldots,q$, the following two matrices, respectively

$$
\begin{aligned}
P_\psi(\psi,\lambda_i) &= \mathrm{E}\{S(\psi,\lambda_i)S(\psi,\lambda_i)^\top S_\psi(\psi,\lambda_i)\} \\
&= \mathrm{E}\begin{pmatrix} x^3 & x^2w_1 & \cdots & x^2w_q \\ x^2w_1 & xw_1^2 & \cdots & xw_1w_q \\ \vdots & \vdots & \ddots & \vdots \\ x^2w_q & xw_1w_q & \cdots & xw_q^2 \end{pmatrix} \\
&= \begin{pmatrix} \frac{4mq}{\psi^3} & 0 & \cdots & 0 \\ 0 & \frac{4m}{\psi\lambda_1^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{4m}{\psi\lambda_q^2} \end{pmatrix}, \\
P_{\lambda_i}(\psi,\lambda_i) &= \mathrm{E}\{S(\psi,\lambda_i)S(\psi,\lambda_i)^\top S_{\lambda_i}(\psi,\lambda_i)\} \\
&= \mathrm{E}\begin{pmatrix} x^2w_i & xw_1w_i & \cdots & xw_qw_i \\ xw_1w_i & w_1^2w_i & \cdots & w_1w_qw_i \\ \vdots & \vdots & \ddots & \vdots \\ xw_qw_i & w_1w_qw_i & \cdots & w_q^2w_i \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 & \cdots & \frac{4m}{\psi\lambda_i^2} & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ \frac{4m}{\psi\lambda_i^2} & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix}.
\end{aligned}
$$

Moreover, the two matrices that are required for the calculation of (2.18) for the ψ and λ_i components of $Q(\psi,\lambda_i)$, are given respectively by (see Appendix B)

$$
\begin{aligned}
Q_{\psi}(\psi,\lambda_i) &= -\mathrm{E}\{j(\psi,\lambda_i)S_{\psi}(\psi,\lambda_i)\} \\
&= -\mathrm{E}\begin{pmatrix} -zx & \frac{1}{\psi}xw_1 & \cdots & \frac{1}{\psi}xw_q \\ \frac{1}{\psi}xw_1 & \frac{2}{\psi\lambda_1^3}y_{12}x & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\psi}xw_q & 0 & \cdots & \frac{2}{\psi\lambda_q^3}y_{q2}x \end{pmatrix} \\
&= -\begin{pmatrix} \frac{4mq}{\psi^3} & 0 & \cdots & 0 \\ 0 & \frac{2m}{\psi\lambda_1^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{2m}{\psi\lambda_q^2} \end{pmatrix}, \\
Q_{\lambda_i}(\psi,\lambda_i) &= -\mathrm{E}\{j(\psi,\lambda_i)S_{\lambda_i}(\psi,\lambda_i)\} \\
&= -\mathrm{E}\begin{pmatrix} -zw_i & \frac{1}{\psi}w_1w_i & \cdots & \frac{1}{\psi}w_qw_i \\ \frac{1}{\psi}w_1w_i & \frac{2}{\psi\lambda_1^3}y_{12}w_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\psi}w_qw_i & 0 & \cdots & \frac{2}{\psi\lambda_q^3}y_{q2}w_i \end{pmatrix} \\
&= -\begin{pmatrix} 0 & 0 & \cdots & \frac{2m}{\psi\lambda_i^2} & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ \frac{2m}{\psi\lambda_i^2} & 0 & \cdots & \frac{2m}{\lambda_i^3} & \cdots & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix},
\end{aligned}
$$

where $i = 1, \ldots, q$. Therefore, for the ψ component of $A(\psi, \lambda_i)$, (2.16) yields

$$\begin{aligned} A_{\psi}(\psi, \lambda_i) &= \frac{1}{2}\mathrm{tr}\big[\{i(\psi, \lambda_i)\}^{-1}\{P_{\psi}(\psi, \lambda_i) + Q_{\psi}(\psi, \lambda_i)\}\big] \\ &= \frac{1}{2}\mathrm{tr}\begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\psi} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\psi} \end{pmatrix} \\ &= \frac{q}{2\psi}, \end{aligned}$$

while for the λ_i component, $i = (1, \ldots, q)$, (2.16) becomes

$$\begin{aligned} A_{\lambda_i}(\psi, \lambda_i) &= \frac{1}{2}\mathrm{tr}\big[\{i(\psi, \lambda_i)\}^{-1}\{P_{\lambda_i}(\psi, \lambda_i) + Q_{\lambda_i}(\psi, \lambda_i)\}\big] \\ &= \frac{1}{2}\mathrm{tr}\begin{pmatrix} 0 & 0 & \cdots & \frac{\psi}{q\lambda_i^2} & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ \frac{1}{\psi} & 0 & \cdots & -\frac{1}{\lambda_i} & \cdots & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix} \\ &= -\frac{1}{2\lambda_i}. \end{aligned}$$

Thus the first-order bias term of the overall maximum likelihood estimator $\hat{\theta} = (\hat{\psi}, \hat{\lambda}_i)$ in (2.15) takes the form

$$\begin{aligned} \frac{b_1(\psi, \lambda_i)}{n} &= -\{i(\psi, \lambda_i)\}^{-1} A(\psi, \lambda_i) \\ &= \begin{pmatrix} -\frac{\psi^2}{2mq} & 0 & \cdots & 0 \\ 0 & -\frac{\lambda_1^2}{2m} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{\lambda_q^2}{2m} \end{pmatrix} \begin{pmatrix} \frac{q}{2\psi} \\ -\frac{1}{2\lambda_1} \\ \vdots \\ -\frac{1}{2\lambda_q} \end{pmatrix} \\ &= \left(-\frac{\psi}{4m}, \frac{\lambda_1}{4m}, \ldots, \frac{\lambda_q}{4m}\right)^{\mathsf{T}}, \end{aligned}$$

and so the asymptotic bias-corrected estimator of ψ is given by

$$\begin{aligned}
\tilde{\psi}_{as} &= \hat{\psi} - \frac{b_1(\hat{\psi})}{n} \\
&= \frac{1}{mq}\sum_{i=1}^{q}\left(y_{i1}y_{i2}\right)^{1/2} + \frac{\hat{\psi}}{4m} \\
&= \frac{1}{mq}\sum_{i=1}^{q}\left(y_{i1}y_{i2}\right)^{1/2} + \frac{1}{4m^2q}\sum_{i=1}^{q}\left(y_{i1}y_{i2}\right)^{1/2} \\
&= \frac{4m+1}{4m^2q}\sum_{i=1}^{q}\left(y_{i1}y_{i2}\right)^{1/2} \\
&= \frac{4m+1}{4m}\hat{\psi}.
\end{aligned}$$

The asymptotic bias-corrected estimator, $\tilde{\psi}_{as}$, is biased and inconsistent as an estimator of ψ since $\mathrm{E}(\tilde{\psi}_{as}) = (4m+1)\psi\Big(\Gamma(m+1/2)\Big)^2/\Big(4m^2\big(\Gamma(m)\big)^2\Big) \neq \psi$ and also $\mathrm{E}(\tilde{\psi}_{as}) \nrightarrow \psi$ as $q \to \infty$. When $m = 1$, $\hat{\psi}_{as} = 5\hat{\psi}/4$, $\mathrm{E}\big(\hat{\psi}_{as}\big) = 5\psi\pi/16$ and $\mathrm{Var}\big(\hat{\psi}_{as}\big) = 25\psi^2\big(1 - \pi^2\big)/256q$. This shows that $\tilde{\psi}_{as}$ has smaller bias than both $\hat{\psi}$ and $\hat{\psi}_{cp}$.

The indirect inference estimator of ψ, $\tilde{\psi}_{II}$, is the solution of the implicit equation (2.19) and is given by

$$\begin{aligned}
\tilde{\psi}_{II} &= \hat{\psi} - B_{\tilde{\psi}_{II}}(\hat{\psi}) \\
&= \frac{1}{mq}\left(\sum_{i=1}^{q}\left(y_{i1}y_{i2}\right)^{1/2}\right) - \tilde{\psi}_{II}\left(\frac{\big(\Gamma(m+1/2)\big)^2 - m\big(\Gamma(m)\big)^2}{m\big(\Gamma(m)\big)^2}\right).
\end{aligned}$$

Solving the above for $\tilde{\psi}_{II}$ shows that it coincides with $\hat{\psi}_{ap}$ exactly, hence the indirect inference estimator of ψ is unbiased and consistent.

The two alternative adjustments to the score function $S(\psi, \lambda_i)$ given by (2.38) and (2.39) are calculated respectively as

$$A^{(E)}(\psi, \lambda_i) = \begin{pmatrix} \frac{q}{2\psi} \\ -\frac{1}{2\lambda_1} \\ \vdots \\ -\frac{1}{2\lambda_q} \end{pmatrix},$$

$$A^{(O)}(\psi,\lambda_i)=\begin{pmatrix} -\frac{q}{2\psi}+\frac{1}{4m\psi^2}\sum_{i=1}^{q}\left[3\lambda_i y_{i1}+(y_{i2}/\lambda_i)\right] \\ -\frac{1}{4m\psi\lambda_1^2}(y_{12}+\lambda_1^2 y_{11}) \\ \vdots \\ -\frac{1}{4m\psi\lambda_q^2}(y_{q2}+\lambda_q^2 y_{q1}) \end{pmatrix}.$$

To obtain the expected adjusted score estimator of ψ we need to solve simultaneously $S_\psi(\psi,\lambda_i)+A_\psi^{(E)}(\psi,\lambda_i)=0$ and $S_{\lambda_i}(\psi,\lambda_i)+A_{\lambda_i}^{(E)}(\psi,\lambda_i)=0$, which are given respectively by

$$q(1-4m)\psi+2\sum_{i=1}^{q}\left[\lambda_i y_{i1}+(y_{i2}/\lambda_i)\right]=0, \tag{5.9}$$

$$2\lambda_i^2 y_{i1}+\psi\lambda_i-2y_{i2}=0. \tag{5.10}$$

The positive root of the quadratic equation (5.10) for λ_i is given by

$$\tilde{\lambda}_i^{(E)}=\frac{-\psi+(\psi^2+16y_{i1}y_{i2})^{1/2}}{4y_{i1}}. \tag{5.11}$$

Substituting (5.11) in (5.9) we find that the expected adjusted score estimator of ψ, $\tilde{\psi}_{ad}^{(E)}$, is the root of the implicit equation

$$q(1-4m)\psi+\sum_{i=1}^{q}(\psi^2+16y_{i1}y_{i2})^{1/2}=0, \tag{5.12}$$

with no closed form solution. The alternative set of equations $S_\psi(\psi,\lambda_i)+A_\psi^{(O)}(\psi,\lambda_i)=0$ and $S_{\lambda_i}(\psi,\lambda_i)+A_{\lambda_i}^{(O)}(\psi,\lambda_i)=0$ simplify to

$$2mq(1+4m)\psi-\sum_{i=1}^{q}\left[(3+4m)\lambda_i y_{i1}+(1+4m)(y_{i2}/\lambda_i)\right]=0, \tag{5.13}$$

$$(4m+1)\lambda_i^2 y_{i1}-(4m-1)y_{i2}=0. \tag{5.14}$$

The positive root of the quadratic equation (5.14) for λ_i is given by

$$\tilde{\lambda}_i^{(O)}=\left(\frac{(4m-1)y_{i2}}{(4m+1)y_{i1}}\right)^{1/2}. \tag{5.15}$$

Substituting (5.15) in (5.13) we find that the observed adjusted score estimator of ψ is

given by

$$\begin{aligned}\tilde{\psi}_{ad}^{(O)} &= \frac{(3+4m)(4m-1)+(4m+1)^2}{2mq(4m+1)^{3/2}(4m-1)^{1/2}}\sum_{i=1}^{q}\left(y_{i1}y_{i2}\right)^{1/2} \\ &= \left(\frac{16m^2+8m-1}{(4m+1)^{3/2}(4m-1)^{1/2}}\right)\hat{\psi}. \end{aligned} \tag{5.16}$$

$\tilde{\psi}_{ad}^{(O)}$ however is a biased estimator of ψ since

$$\mathrm{E}(\tilde{\psi}_{ad}^{(O)}) = \left(\frac{\left(16m^2+8m-1\right)\left(\Gamma(m+1/2)\right)^2}{m(4m+1)^{3/2}(4m-1)^{1/2}\left(\Gamma(m)\right)^2}\right)\psi, \tag{5.17}$$

and is inconsistent since $\mathrm{E}(\tilde{\psi}_{ad}^{(O)}) \nrightarrow \psi$ as $q \to \infty$ since $\mathrm{E}(\tilde{\psi}_{ad}^{(O)})$ is independent of q. When $m = 1$, $\tilde{\psi}_{ad}^{(O)} = 23\hat{\psi}/\sqrt{375}$, $\mathrm{E}\left(\tilde{\psi}_{ad}^{(O)}\right) = 23\psi\pi/(4\sqrt{375})$ and $\mathrm{Var}\left(\tilde{\psi}_{ad}^{(O)}\right) = 529\psi^2\left(1-\pi^2\right)/(6000q)$. This shows that $\tilde{\psi}_{ad}^{(O)}$ has smaller bias than $\hat{\psi}$ while $\hat{\psi}_{cp}$ is less biased than $\tilde{\psi}_{ad}^{(O)}$.

We summarise, in Table 5.1, the bias and variance of the estimators of ψ which are available in closed form.

Table 5.1: Matched gamma pairs. Estimators of ψ and their theoretical bias and variance functions.

	Estimator	Bias	Variance
$\hat{\psi}$	$\frac{1}{mq}\sum_{i=1}^{q}\left(y_{i1}y_{i2}\right)^{1/2}$	$\left(\frac{\left(\Gamma(m+1/2)\right)^2-m\left(\Gamma(m)\right)^2}{m\left(\Gamma(m)\right)^2}\right)\psi$	$\left(\frac{m^2\left(\Gamma(m)\right)^4-\left(\Gamma(m+1/2)\right)^4}{qm^2\left(\Gamma(m)\right)^4}\right)\psi^2$
$\hat{\psi}_{cp}=\hat{\psi}_{mp}$	$\left(\frac{4m}{4m-1}\right)\hat{\psi}$	$\left(\frac{4\left(\Gamma(m+1/2)\right)^2-(4m-1)\left(\Gamma(m)\right)^2}{(4m-1)\left(\Gamma(m)\right)^2}\right)\psi$	$\left(\frac{16m^2\left(\Gamma(m)\right)^4-16\left(\Gamma(m+1/2)\right)^4}{q(4m-1)^2\left(\Gamma(m)\right)^4}\right)\psi^2$
$\hat{\psi}_{ap}=\hat{\psi}_{II}$	$\left(\frac{m\left(\Gamma(m)\right)^2}{\left(\Gamma(m+1/2)\right)^2}\right)\hat{\psi}$	0	$\left(\frac{m^2\left(\Gamma(m)\right)^4-\left(\Gamma(m+1/2)\right)^4}{q\left(\Gamma(m+1/2)\right)^4}\right)\psi^2$
$\tilde{\psi}_{as}$	$\left(\frac{4m+1}{4m}\right)\hat{\psi}$	$\left(\frac{(4m+1)\left(\Gamma(m+1/2)\right)^2-4m^2\left(\Gamma(m)\right)^2}{4m^2\left(\Gamma(m)\right)^2}\right)\psi$	$\left(\frac{m^2(4m+1)^2\left(\Gamma(m)\right)^4-(4m+1)^2\left(\Gamma(m+1/2)\right)^4}{16qm^4\left(\Gamma(m)\right)^4}\right)\psi^2$
$\tilde{\psi}_{ad}^{(O)}$	$\left(\frac{16m^2+8m-1}{(4m+1)^{3/2}(4m-1)^{1/2}}\right)\hat{\psi}$	$\left(\frac{\left(16m^2+8m-1\right)\left(\Gamma(m+1/2)\right)^2-m(4m+1)^{3/2}(4m-1)^{1/2}\left(\Gamma(m)\right)^2}{m(4m+1)^{3/2}(4m-1)^{1/2}\left(\Gamma(m)\right)^2}\right)\psi$	$\left(\frac{\left(16m^2+8m-1\right)^2\left[m^2\left(\Gamma(m)\right)^4-\left(\Gamma(m+1/2)\right)^4\right]}{qm^2(4m+1)^3(4m-1)\left(\Gamma(m)\right)^4}\right)\psi^2$

5.2.4 Simulation study

We report, in Tables 5.2, 5.3 and 5.4 the numerical values of the theoretical bias, variance and mean squared error, respectively, of all the estimators of ψ for $m \in \{1,3,5,8\}$ and $q \in \{4,8,16,32,64,128\}$, when the true value of ψ is one. The corresponding values in Tables 5.2, 5.3 and 5.4 for the expected adjusted score estimator of ψ, $\tilde{\psi}_{ad}^{(E)}$, are based on a simulation study of 10000 matched gamma samples where the true parameter of interest is $\psi = 1$ and the true nuisance parameters, λ, are fixed to be a vector of consecutive, equi-spaced values between 1 and 5.

The adjusted profile log-likelihood estimator, $\hat{\psi}_{ap}$, which coincides with the indirect inference estimator, $\hat{\psi}_{II}$, are the only estimators with zero bias, while the estimator with the largest bias in magnitude is the ML estimator, $\hat{\psi}$. The bias of all other estimators is less than that of $\hat{\psi}$ for all values of m, and converges to zero as m increases. As noted theoretically for $m = 1$, excluding the unbiased estimators $\hat{\psi}_{ap}$ and $\hat{\psi}_{II}$, we find that the asymptotic bias-corrected estimator ψ, $\tilde{\psi}_{as}$ is the least biased estimator, followed by the conditional and modified profile log-likelihood estimators, $\hat{\psi}_{cp}$ and $\hat{\psi}_{mp}$, leaving the observed adjusted score estimator $\tilde{\psi}_{ad}^{(O)}$ as the estimator of ψ with the largest bias. The above conclusion is true for all values of m considered. The expected adjusted score estimator, $\tilde{\psi}_{ad}^{(E)}$, is less biased than $\hat{\psi}$ for all m and less biased than $\tilde{\psi}_{ad}^{(O)}$ for $m \in \{3,5,8\}$, but compared to the other estimators of ψ it does not perform better in terms of bias. We conclude that in terms of estimators resulting from modifications of the likelihood function, the adjusted profile log-likelihood estimator is exactly unbiased, while all other estimators achieve bias reduction of the ML estimator.

Even though $\hat{\psi}$ has the largest bias in magnitude, Table 5.3 shows that it possesses the least amount of variance amongst all estimators of ψ. The variance of all estimators decreases to zero as the number of strata and stratum sample size increase, independently or together. This means that while all estimators of ψ reduce the bias of $\hat{\psi}$, they inflate the variance a little. Table 5.4 shows that, for a fixed value of m and increasing q, the mean squared error of all estimators, except that of $\hat{\psi}$ becomes very close to each other and decreases towards zero at a rate higher than the mean squared error of $\hat{\psi}$ does. On the other hand, fixing q and increasing m, the mean squared error of all estimators including that of $\hat{\psi}$ decrease towards zero and becomes closer to each other. The largest difference between the mean squared error of $\hat{\psi}$ and the other estimators is observed when m is very small and q is very large.

All of the above results may also be deduced from Figures 5.1, 5.2 and 5.3 which

show plots of the theoretical biases, variances and mean squared errors of the estimators in Table 5.1 for a different combination of values of the stratum sample size m and the number of strata q than those considered in Tables 5.2, 5.3 and 5.4. Finally, we simulated 10000 matched gamma pairs with true $\psi = 1$ and obtained, in Table 5.5, the coverage probability and median length of 95% confidence intervals for ψ based on the chi-squared approximation to the distribution of the profile, conditional profile and adjusted profile log-likelihood ratios, denoted by W, W_{cp} and W_{ap}, respectively. We observe that for $m = 1$ and $m = 2$ the coverage probability derived from W_{ap} is closer to the nominal level than those derived from W_{cp} for all values of q. For the remaining values of m, W_{ap} performs better than W_{cp} for most values of q or gives a coverage probability that is very close to that given by W_{cp}.

In terms of the incidental parameter problem, all estimators of ψ, except for $\hat{\psi}_{ap} = \hat{\psi}_{II}$, are biased and inconsistent when m is fixed and q is allowed to increase to infinity. This means that even though these estimators are less biased than the ML estimator, they do not solve the incidental parameter problem. However, the adjusted profile log-likelihood estimator which coincides with the indirect inference estimator is exactly unbiased and consistent when m is fixed and q is allowed to increase to infinity, thus we have found a consistent estimator which is the only estimator that solves the incidental parameter problem of Neyman and Scott (1948) in the matched gamma pairs model.

Table 5.2: Matched gamma pairs. Inference about common square root of product of means in q pairs of gamma observations with shape m. Numerical values of the theoretical biases of the estimators of ψ for various values of m and q, with all entries multiplied by 10. The parameter of interest is $\psi = 1$. The empty cells correspond to the fact that the bias of all estimators except $\tilde{\psi}_{ad}^{(E)}$ is independent of q.

		$q=4$	$q=8$	$q=16$	$q=32$	$q=64$	$q=128$
$m=1$	$\hat{\psi}$	-2.1460					
	$\hat{\psi}_{cp}$=$\hat{\psi}_{mp}$	0.4720					
	$\hat{\psi}_{ap}$=$\hat{\psi}_{II}$	0.0000					
	$\tilde{\psi}_{as}$	-0.1825					
	$\tilde{\psi}_{ad}^{(E)*}$	1.4608	1.5264	1.5579	1.5348	1.5442	1.5486
	$\tilde{\psi}_{ad}^{(O)}$	-0.6717					
$m=3$	$\hat{\psi}$	-0.7961					
	$\hat{\psi}_{cp}$=$\hat{\psi}_{mp}$	0.0406					
	$\hat{\psi}_{ap}$=$\hat{\psi}_{II}$	0.0000					
	$\tilde{\psi}_{as}$	-0.0291					
	$\tilde{\psi}_{ad}^{(E)*}$	0.1018	0.0993	0.0977	0.0920	0.0860	0.0892
	$\tilde{\psi}_{ad}^{(O)}$	-0.1127					
$m=5$	$\hat{\psi}$	-0.4869					
	$\hat{\psi}_{cp}$=$\hat{\psi}_{mp}$	0.0138					
	$\hat{\psi}_{ap}$=$\hat{\psi}_{II}$	0.0000					
	$\tilde{\psi}_{as}$	-0.0113					
	$\tilde{\psi}_{ad}^{(E)*}$	0.0308	0.0313	0.0336	0.0223	0.0267	0.0284
	$\tilde{\psi}_{ad}^{(O)}$	-0.0441					
$m=8$	$\hat{\psi}$	-0.3075					
	$\hat{\psi}_{cp}$=$\hat{\psi}_{mp}$	0.0052					
	$\hat{\psi}_{ap}$=$\hat{\psi}_{II}$	0.0000					
	$\tilde{\psi}_{as}$	-0.0046					
	$\tilde{\psi}_{ad}^{(E)*}$	0.0149	0.0187	0.0179	0.0094	0.0121	0.0117
	$\tilde{\psi}_{ad}^{(O)}$	-0.0181					

*The bias values reported for the expected adjusted score estimator, $\tilde{\psi}_{ad}^{(E)}$, are based on a simulation study of 10000 samples. The parameter of interest is $\psi = 1$ and the nuisance parameter λ is fixed to be a vector of consecutive, equi-spaced values between 1 and 5.

Table 5.3: Matched gamma pairs. Inference about common square root of product of means in q pairs of gamma observations with shape m. Numerical values of the theoretical variance of the estimators of ψ for various values of m and q, with all entries multiplied by 10. The parameter of interest is $\psi = 1$.

		$q=4$	$q=8$	$q=16$	$q=32$	$q=64$	$q=128$
$m=1$	$\hat{\psi}$	0.9579	0.4789	0.2395	0.1197	0.0599	0.0299
	$\hat{\psi}_{cp}$=$\hat{\psi}_{mp}$	1.7029	0.8514	0.4257	0.2129	0.1064	0.0532
	$\hat{\psi}_{ap}$=$\hat{\psi}_{II}$	1.5528	0.7764	0.3882	0.1941	0.0971	0.0485
	$\tilde{\psi}_{as}$	1.4967	0.7483	0.3742	0.1871	0.0935	0.0468
	$\tilde{\psi}_{ad}^{(E)\dagger}$	2.0705	1.0256	0.5201	0.2596	0.1296	0.0648
	$\tilde{\psi}_{ad}^{(O)}$	1.3512	0.6756	0.3378	0.1689	0.0845	0.0422
$m=3$	$\hat{\psi}$	0.3822	0.1911	0.0956	0.0478	0.0239	0.0119
	$\hat{\psi}_{cp}$=$\hat{\psi}_{mp}$	0.4549	0.2274	0.1137	0.0569	0.0284	0.0142
	$\hat{\psi}_{ap}$=$\hat{\psi}_{II}$	0.4512	0.2256	0.1128	0.0564	0.0282	0.0141
	$\tilde{\psi}_{as}$	0.4486	0.2243	0.1121	0.0561	0.0280	0.0140
	$\tilde{\psi}_{ad}^{(E)\dagger}$	0.4637	0.2330	0.1165	0.0568	0.0289	0.0145
	$\tilde{\psi}_{ad}^{(O)}$	0.4411	0.2205	0.1103	0.0551	0.0276	0.0138
$m=5$	$\hat{\psi}$	0.2375	0.1188	0.0594	0.0297	0.0148	0.0074
	$\hat{\psi}_{cp}$=$\hat{\psi}_{mp}$	0.2632	0.1316	0.0658	0.0329	0.0164	0.0082
	$\hat{\psi}_{ap}$=$\hat{\psi}_{II}$	0.2625	0.1312	0.0656	0.0328	0.0164	0.0082
	$\tilde{\psi}_{as}$	0.2619	0.1309	0.0655	0.0327	0.0164	0.0082
	$\tilde{\psi}_{ad}^{(E)\dagger}$	0.2644	0.1317	0.0650	0.0329	0.0164	0.0082
	$\tilde{\psi}_{ad}^{(O)}$	0.2602	0.1301	0.0650	0.0325	0.0163	0.0081
$m=8$	$\hat{\psi}$	0.1514	0.0757	0.0378	0.0189	0.0095	0.0047
	$\hat{\psi}_{cp}$=$\hat{\psi}_{mp}$	0.1613	0.0806	0.0403	0.0202	0.0101	0.0050
	$\hat{\psi}_{ap}$=$\hat{\psi}_{II}$	0.1611	0.0806	0.0403	0.0201	0.0101	0.0050
	$\tilde{\psi}_{as}$	0.1610	0.0805	0.0402	0.0201	0.0101	0.0050
	$\tilde{\psi}_{ad}^{(E)\dagger}$	0.1617	0.0788	0.0397	0.0197	0.0100	0.0050
	$\tilde{\psi}_{ad}^{(O)}$	0.1605	0.0803	0.0401	0.0201	0.0100	0.0050

†The variance values reported for the expected adjusted score estimator, $\tilde{\psi}_{ad}^{(E)}$, are based on a simulation study of 10000 samples for various values of m and q. The parameter of interest is $\psi = 1$ and the nuisance parameter λ is fixed to be a vector of consecutive, equi-spaced values between 1 and 5.

Table 5.4: Matched gamma pairs. Inference about common square root of product of means in q pairs of gamma observations with shape m. Numerical values of the theoretical mean squared error of the estimators of ψ for various values of m and q, with all entries multiplied by 10. The parameter of interest is $\psi = 1$.

		$q=4$	$q=8$	$q=16$	$q=32$	$q=64$	$q=128$
$m=1$	$\hat{\psi}$	1.4184	0.9395	0.7000	0.5803	0.5204	0.4905
	$\hat{\psi}_{cp}$=$\hat{\psi}_{mp}$	1.7252	0.8737	0.4480	0.2351	0.1287	0.0755
	$\hat{\psi}_{ap}$=$\hat{\psi}_{II}$	1.5528	0.7764	0.3882	0.1941	0.0971	0.0485
	$\tilde{\psi}_{as}$	1.5000	0.7517	0.3775	0.1904	0.0969	0.0501
	$\tilde{\psi}_{ad}^{(E)\ddagger}$	2.2839	1.2585	0.7628	0.4952	0.3680	0.3046
	$\tilde{\psi}_{ad}^{(O)}$	1.3964	0.7207	0.3829	0.2140	0.1296	0.0873
$m=3$	$\hat{\psi}$	0.4456	0.2545	0.1589	0.1112	0.0873	0.0753
	$\hat{\psi}_{cp}$=$\hat{\psi}_{mp}$	0.4550	0.2276	0.1139	0.0570	0.0286	0.0144
	$\hat{\psi}_{ap}$=$\hat{\psi}_{II}$	0.4512	0.2256	0.1128	0.0564	0.0282	0.0141
	$\tilde{\psi}_{as}$	0.4487	0.2244	0.1122	0.0562	0.0281	0.0141
	$\tilde{\psi}_{ad}^{(E)\ddagger}$	0.4647	0.2340	0.1174	0.0577	0.0297	0.0153
	$\tilde{\psi}_{ad}^{(O)}$	0.4423	0.2218	0.1115	0.0564	0.0288	0.0151
$m=5$	$\hat{\psi}$	0.2612	0.1425	0.0831	0.0534	0.0386	0.0311
	$\hat{\psi}_{cp}$=$\hat{\psi}_{mp}$	0.2632	0.1316	0.0658	0.0329	0.0165	0.0082
	$\hat{\psi}_{ap}$=$\hat{\psi}_{II}$	0.2625	0.1312	0.0656	0.0328	0.0164	0.0082
	$\tilde{\psi}_{as}$	0.2619	0.1310	0.0655	0.0327	0.0164	0.0082
	$\tilde{\psi}_{ad}^{(E)\ddagger}$	0.2645	0.1318	0.0651	0.0330	0.0165	0.0083
	$\tilde{\psi}_{ad}^{(O)}$	0.2604	0.1303	0.0652	0.0327	0.0165	0.0083
$m=8$	$\hat{\psi}$	0.1608	0.0851	0.0473	0.0284	0.0189	0.0142
	$\hat{\psi}_{cp}$=$\hat{\psi}_{mp}$	0.1613	0.0807	0.0403	0.0202	0.0101	0.0050
	$\hat{\psi}_{ap}$=$\hat{\psi}_{II}$	0.1611	0.0806	0.0403	0.0201	0.0101	0.0050
	$\tilde{\psi}_{as}$	0.1610	0.0805	0.0402	0.0201	0.0101	0.0050
	$\tilde{\psi}_{ad}^{(E)\ddagger}$	0.1617	0.0788	0.0397	0.0197	0.0101	0.0050
	$\tilde{\psi}_{ad}^{(O)}$	0.1606	0.0803	0.0402	0.0201	0.0101	0.0050

‡The mean squared error values reported for the expected adjusted score estimator, $\tilde{\psi}_{ad}^{(E)}$, are based on a simulation study of 10000 samples for various values of m and q. The parameter of interest is $\psi = 1$ and the nuisance parameter λ is fixed to be a vector of consecutive, equi-spaced values between 1 and 5.

Table 5.5: Matched gamma pairs. Inference about common square root of product of means in q pairs of gamma observations with shape m. Columns three to eight show the coverage probability and median length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional profile and adjusted profile log-likelihood ratios, with all coverage probabilities multiplied by 100. The parameter of interest is $\psi = 1$.

		$q = 4$	$q = 8$	$q = 16$	$q = 32$	$q = 64$	$q = 128$
$m = 1$	W	99.82 (4.71)	99.99 (4.37)	100.00 (4.30)	100.00 (4.29)	100.00 (4.31)	100.00 (4.35)
	W_{cp}	95.50 (2.01)	95.53 (1.30)	95.09 (0.88)	94.55 (0.61)	93.29 (0.42)	90.83 (0.30)
	W_{ap}	94.75 (1.82)	95.21 (1.19)	95.04 (0.81)	95.17 (0.56)	95.05 (0.39)	95.16 (0.27)
$m = 2$	W	99.52 (1.65)	99.72 (1.42)	99.98 (1.34)	100.00 (1.31)	100.00 (1.28)	100.00 (1.27)
	W_{cp}	95.31 (1.14)	95.31 (0.78)	95.59 (0.54)	95.34 (0.38)	94.65 (0.27)	94.79 (0.19)
	W_{ap}	95.03 (1.12)	95.10 (0.76)	95.28 (0.53)	95.10 (0.37)	94.85 (0.26)	94.90 (0.18)
$m = 3$	W	98.95 (0.96)	99.07 (0.68)	99.23 (0.51)	99.46 (0.42)	99.67 (0.38)	99.94 (0.35)
	W_{cp}	95.02 (0.88)	94.74 (0.61)	94.83 (0.42)	95.23 (0.30)	95.05 (0.21)	94.89 (0.15)
	W_{ap}	94.91 (0.87)	94.73 (0.60)	94.92 (0.42)	95.27 (0.30)	94.82 (0.21)	94.74 (0.15)
$m = 4$	W	97.25 (0.83)	95.77 (0.60)	93.31 (0.46)	90.23 (0.38)	89.75 (0.34)	93.34 (0.32)
	W_{cp}	94.60 (0.74)	94.77 (0.52)	95.07 (0.36)	95.31 (0.25)	95.15 (0.18)	95.01 (0.13)
	W_{ap}	94.51 (0.74)	94.73 (0.51)	95.05 (0.36)	95.22 (0.25)	95.10 (0.18)	94.77 (0.13)
$m = 5$	W	96.72 (1.16)	97.21 (1.03)	99.03 (0.95)	99.77 (0.92)	100.00 (0.90)	100.00 (0.90)
	W_{cp}	94.78 (0.66)	95.34 (0.46)	95.44 (0.32)	94.97 (0.23)	95.15 (0.16)	94.86 (0.11)
	W_{ap}	94.76 (0.65)	95.29 (0.45)	95.40 (0.32)	95.04 (0.22)	95.12 (0.16)	94.77 (0.11)
$m = 8$	W	99.92 (3.45)	100.00 (3.33)	100.00 (3.28)	100.00 (3.25)	100.00 (3.24)	100.00 (3.24)
	W_{cp}	94.89 (0.51)	95.37 (0.36)	95.05 (0.25)	95.12 (0.18)	94.82 (0.12)	95.02 (0.09)
	W_{ap}	94.88 (0.51)	95.36 (0.36)	94.95 (0.25)	95.17 (0.18)	94.85 (0.12)	95.00 (0.09)

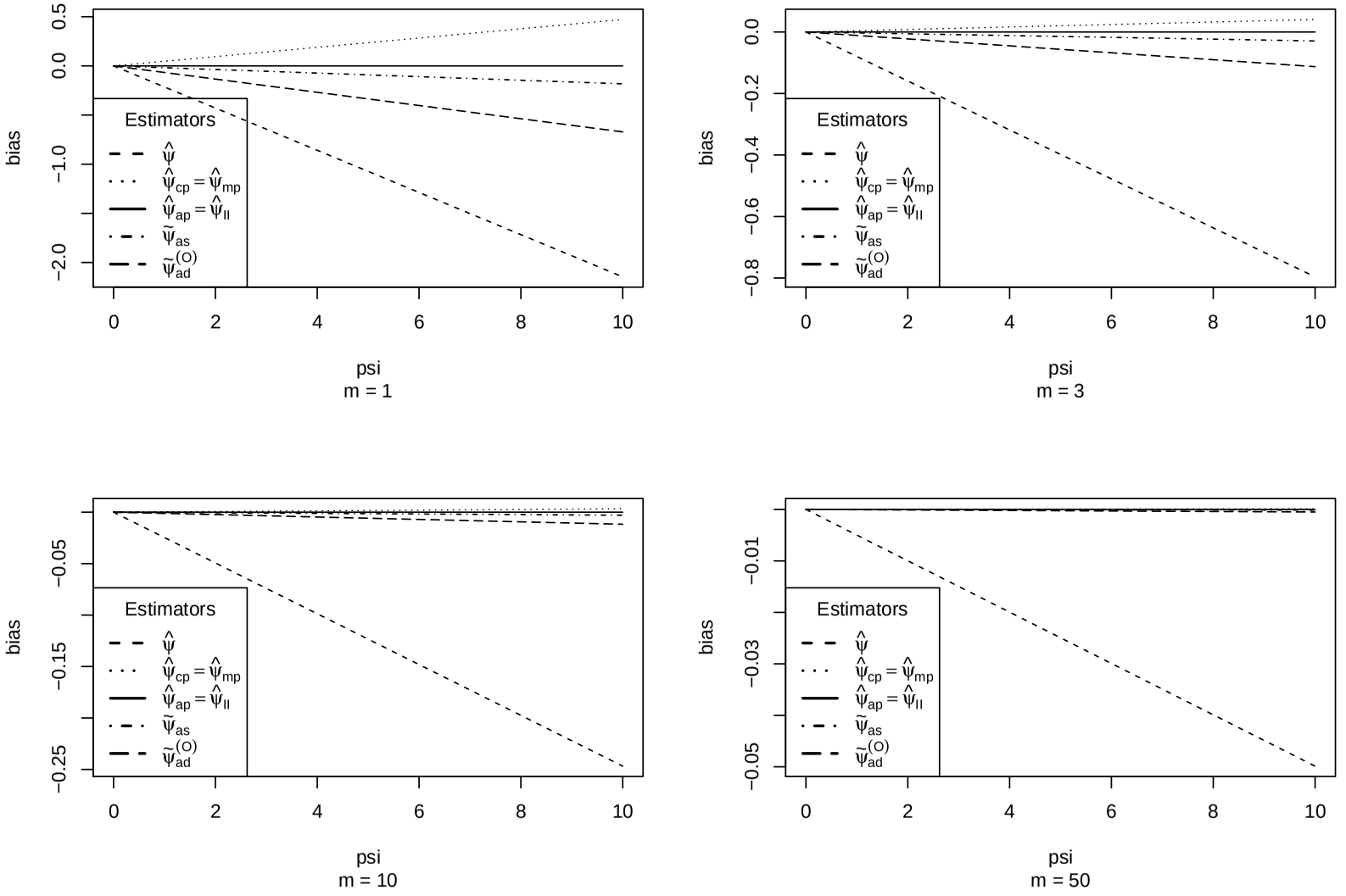

Figure 5.1: Matched gamma pairs. Plot of the theoretical bias of the various estimators of ψ for various values of m.

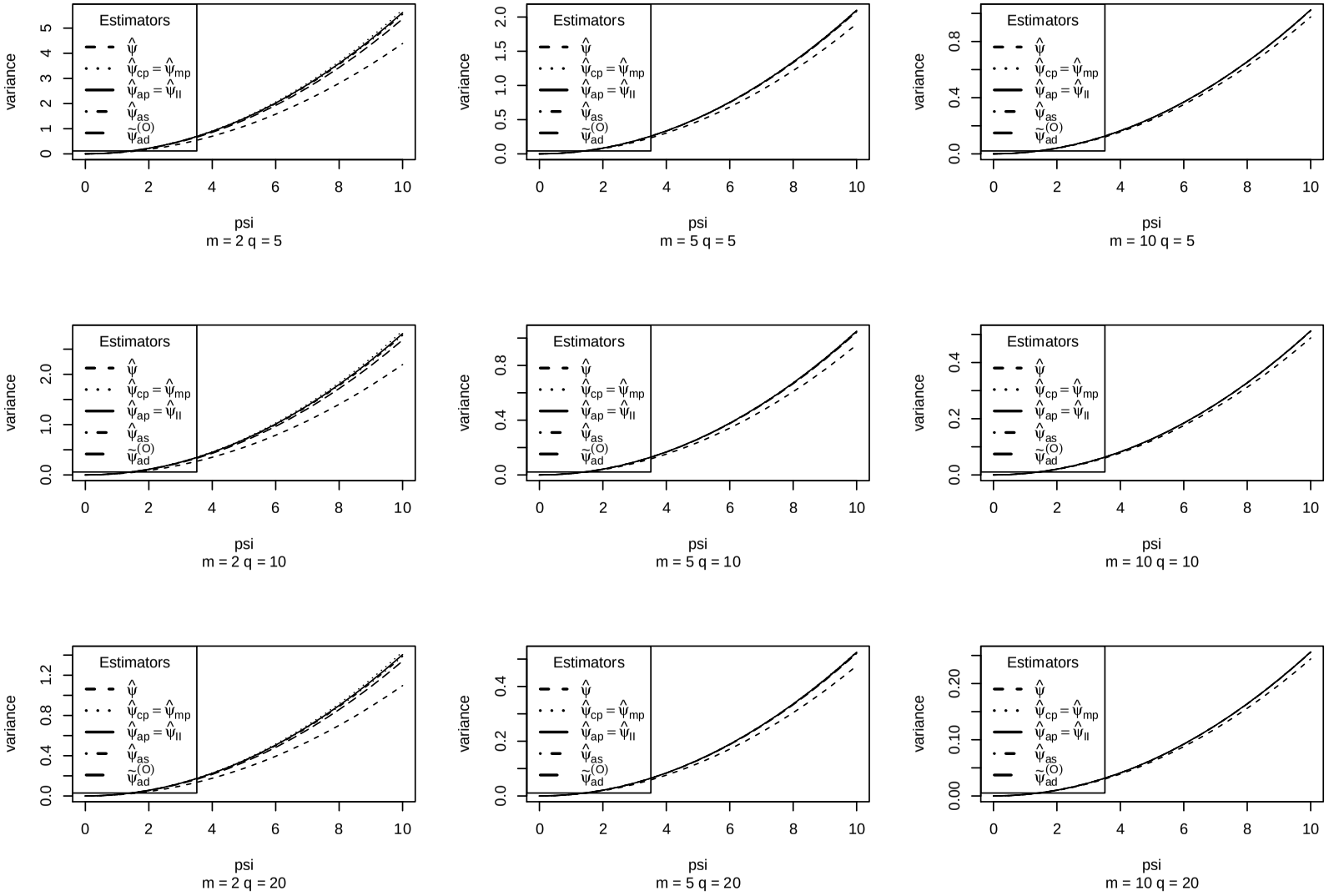

Figure 5.2: Matched gamma pairs. Plot of the theoretical variance of the various estimators of ψ for different combinations of m and q.

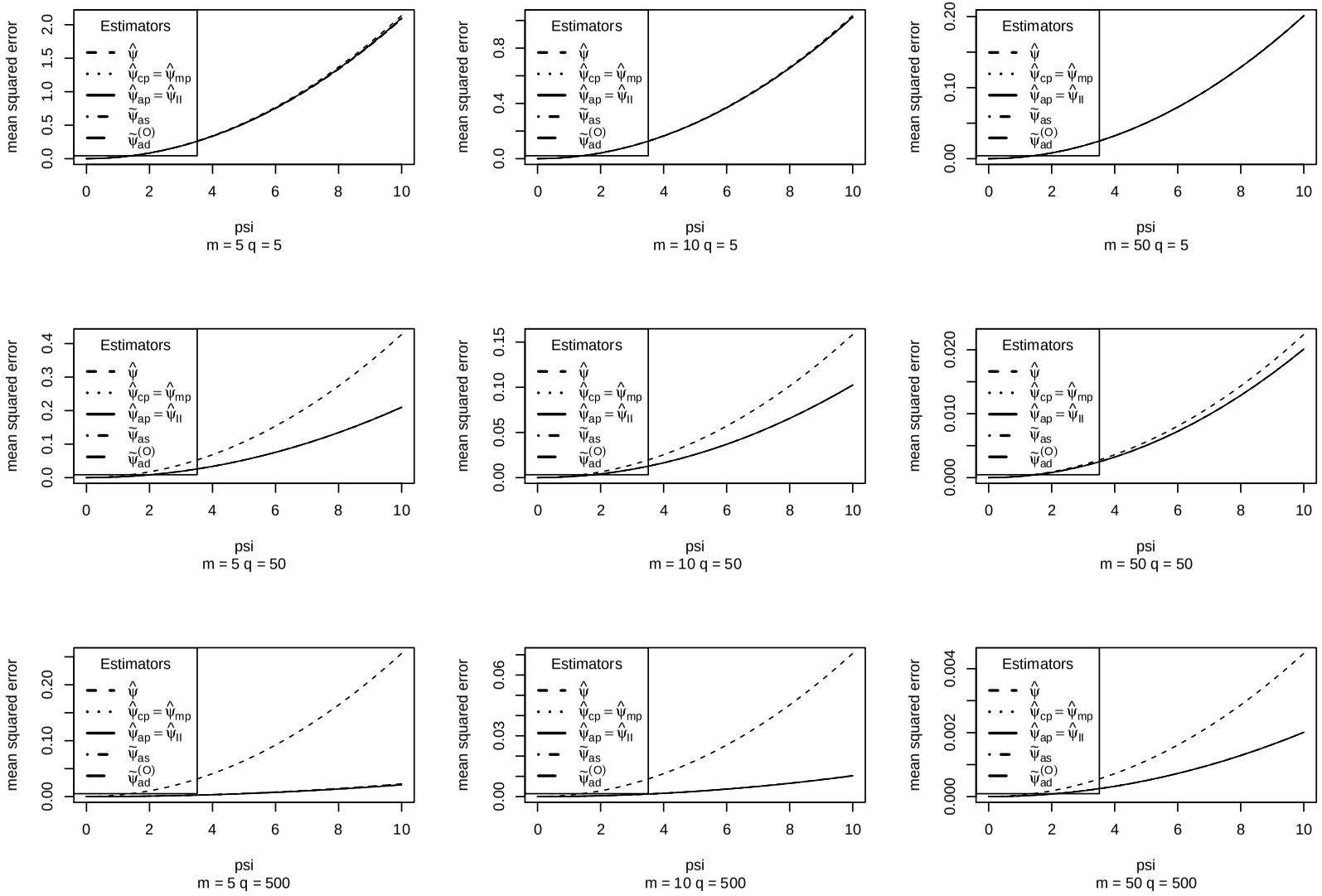

Figure 5.3: Matched gamma pairs. Plot of the theoretical mean squared error of the various estimators of ψ for different combinations of m and q.

5.3 Binomial matched pairs model

Consider a series of q independent pairs of independent binomial random variables Y_{i1}, Y_{i2}, with $Y_{i1} \sim \text{Bi}(1, \pi_{i1})$ and $Y_{i2} \sim \text{Bi}(m_i, \pi_{i2})$ as in Lunardon (2018, §4). Let the success probabilities satisfy

$$\pi_{i1} = \frac{\exp(\psi + \lambda_i)}{1 + \exp(\psi + \lambda_i)}, \tag{5.18}$$

and

$$\pi_{i2} = \frac{\exp(\lambda_i)}{1 + \exp(\lambda_i)}, \tag{5.19}$$

where $\psi = \ln\{\pi_{i1}/(1-\pi_{i2})\} - \ln\{\pi_{i2}/(1-\pi_{i2})\}$, the log odds ratio, is the parameter of interest and $\lambda_i = \ln\{\pi_{i2}/(1-\pi_{i2})\}$ is the nuisance parameter, $i = 1, \ldots, q$. This is a stratified setting as in Sartori (2003) where the sample size is $n = \sum_{i=1}^{q} m_i$ and where q is the number of strata and m_i is the ith stratum sample size. We will refer to this model as the binomial matched pairs model and to the model with $m_i = m = 1$ as the binary matched pairs model.

This model often arises in case-control studies in medical contexts, where y_{i1} and y_{i2} may represent for example the numbers of exposed or experimental persons among one case and m_i controls in the ith stratum and where interest lies in studying the influence of some risk factor or the effect of some treatment (see Cox, 1970, §1.2 for more examples).

5.3.1 Review of point estimation of the log odds ratio

Several estimators of the common log odds ratio ψ have been proposed in the literature. These include the Mantel and Haenszel, empirical logit and Birch estimators (see Breslow, 1981; Gart, 1971, for a review of these estimators and of their properties). In this section, however, we only consider estimators of the log odds ratio that depend on the data only through the sufficient statistic.

5.3.1.1 Maximum likelihood

The log-likelihood function for $\theta = (\psi, \lambda_1, \ldots, \lambda_q)^\intercal$ for the above binomial matched pairs model is (Lunardon, 2018, §4.1)

$$l(\theta) = \sum_{i=1}^{q} \psi y_{i1} + \sum_{i=1}^{q} \lambda_i (y_{i1} + y_{i2}) - \sum_{i=1}^{q} \left[\ln\{1 + \exp(\psi + \lambda_i)\} + m_i \ln\{1 + \exp(\lambda_i)\} \right]. \tag{5.20}$$

This is a linear full exponential family in canonical form (Pace and Salvan, 1997, §5) with jointly sufficient statistics $t = \sum_{i=1}^{q} y_{i1}$ and $s_i = y_{i1} + y_{i2}$ for ψ and λ_i, respectively. Throughout this section, we consider for simplicity $m_i = m$ with totals $s_i = (m+1)/2$ as in Sartori (2003, Example 3). In this setting, the constrained maximum likelihood estimator of λ_i for a fixed value of ψ, denoted by $\hat{\lambda}_{i,\psi}$, will be identical for all $i = 1, \ldots, q$ and so we set $\hat{\lambda}_{i,\psi} = \hat{\gamma}_\psi$. Equivalently, denote the constrained maximum likelihood estimator of ψ for a fixed value of γ by $\hat{\psi}_\gamma$. The score equations for the log-likelihood function with respect to γ and ψ are respectively

$$\frac{m+1}{2} - \frac{\exp(\psi+\gamma)}{1+\exp(\psi+\gamma)} - m\frac{\exp(\gamma)}{1+\exp(\gamma)} = 0, \tag{5.21}$$

$$t - q\frac{\exp(\psi+\gamma)}{1+\exp(\psi+\gamma)} = 0. \tag{5.22}$$

The solution of (5.22) is $\hat{\psi}_\gamma = \ln\{t/(q-t)\} - \gamma$, and on substituting this in (5.21) and solving for γ we get the maximum likelihood estimator of the nuisance parameter $\hat{\gamma} = \ln\left[\{q(m+1)-2t\}/\{q(m-1)+2t\}\right]$. Substituting $\hat{\gamma}$ in $\hat{\psi}_\gamma$ we get the maximum likelihood estimator of the parameter of interest

$$\hat{\psi} = \ln\left(\frac{t\{q(m-1)+2t\}}{\{q-t\}\{q(m+1)-2t\}}\right). \tag{5.23}$$

Note that when $t = 0$ or $t = q$, $\hat{\psi}$ is $-\infty$ or $+\infty$, respectively.

Using the weak law of large numbers, Slutsky's theorem and the Continuous mapping theorem (Florescu, 2014, §7), we find that $\hat{\psi}$ converges in probability to $\psi + \ln[\{(m+1)\exp(\psi)+m-1\}/\{(m-1)\exp(\psi)+m+1\}]$ as $q \to \infty$, and so it is inconsistent. When m is also allowed to increase to ∞, $\hat{\psi}$ will tend to ψ. This means that $\hat{\psi}$ will be consistent only when both m and q diverge.

Given that the totals s_i are fixed, the maximum likelihood estimator of ψ depends on the data only through the sufficient statistic $T = \sum_{i=1}^{q} Y_{i1}$ and so its bias and variance can be calculated exactly using

$$\mathrm{E}_\psi\{\hat{\psi}(T)\} = \sum_{t=1}^{q-1} \hat{\psi}(t)\frac{\mathrm{pr}(T=t|S_i=s_i)}{1-\mathrm{pr}(T=0|S_i=s_i)-\mathrm{pr}(T=q|S_i=s_i)}, \tag{5.24}$$

$$\mathrm{var}_\psi\{\hat{\psi}(T)\} = \mathrm{E}_\psi[\{\hat{\psi}(T)\}^2] - [\mathrm{E}_\psi\{\hat{\psi}(T)\}]^2. \tag{5.25}$$

5.3.1.2 Conditional maximum likelihood

The conditional log-likelihood function is based on the distribution of Y_{i1} given $S_i = s_i$ in each stratum. Davison (1988) and Gart (1970) noted that the conditional density of Y_{i1} given S_i is

$$\text{pr}(Y_{i1} = y_{i1}|S_i = s_i) = \frac{\binom{1}{y_{i1}}\binom{m_i}{s_i - y_{i1}}\exp(\psi y_{i1})}{\sum_{u=0}^{\min(1,s_i)}\binom{1}{u}\binom{m_i}{s_i-u}\exp(\psi u)}. \tag{5.26}$$

This is the noncentral hypergeometric distribution which is obtained by rewriting the left hand side of (5.26) as $\Pr(Y_{i1} = y_{i1}, S_i = s_i)/\Pr(S_i = s_i) = \Pr(Y_{i1} = y_{i1})\Pr(Y_{i2} = s_i - y_{i1})/\Pr(Y_{i1} + Y_{i2} = s_i)$, by independence of Y_{i1} and Y_{i2}, and noting that $\Pr(Y_{i1} + Y_{i2} = s_i) = \sum_{u=0}^{s_i}\Pr(Y_{i1} = u)\Pr(Y_{i2} = s_i - u)$. Because y_{i1} can only be 0 or 1, the right hand side of (5.26) can be further simplified to

$$\text{pr}(Y_{i1} = y_{i1}|S_i = s_i) = \left[\frac{s_i\exp(\psi)}{m_i + 1 + s_i\{\exp(\psi) - 1\}}\right]^{y_{i1}}\left[\frac{m_i - s_i + 1}{m_i + 1 + s_i\{\exp(\psi) - 1\}}\right]^{1-y_{i1}}. \tag{5.27}$$

This shows that $Y_{i1}|S_i$ has a Bernoulli distribution with success probability the first term inside the bracket of the right hand side of (5.27). Taking the logarithm of the product of (5.27) gives the conditional log-likelihood function which simplifies to

$$l_c(\psi) = \sum_{i=1}^{q}\psi y_{i1} - \sum_{i=1}^{q}\ln\left[m_i + 1 + s_i\{\exp(\psi) - 1\}\right]. \tag{5.28}$$

The score equation for the conditional log-likelihood function is

$$t - \sum_{i=1}^{q}\frac{s_i\exp(\psi)}{m_i + 1 + s_i\{\exp(\psi) - 1\}} = 0. \tag{5.29}$$

Letting $m_i = m$ and $s_i = (m+1)/2$, (5.28) simplifies to

$$l_c(\psi) = \psi t - q\ln\{\exp(\psi) + 1\}, \tag{5.30}$$

and (5.29) simplifies to

$$t - q\frac{\exp(\psi)}{\exp(\psi) + 1} = 0, \tag{5.31}$$

so the solution of the latter gives the conditional maximum likelihood estimator

$$\hat{\psi}_c = \ln\left(\frac{t}{q-t}\right). \tag{5.32}$$

When $t = 0$ or $t = q$, $\hat{\psi}_c$ is $-\infty$ or $+\infty$, respectively. In the setting of Lunardon (2018), i.e. when $m_i = m$ and $s_i = (m+1)/2$, the success probability of the Bernoulli random variable $Y_{i1}|S_i$ simplifies to $\pi = \exp(\psi)/\{\exp(\psi)+1\}$. The distribution of the sufficient statistic T given S_i is therefore Binomial with denominator q and success probability π. The conditional distribution of T can also be obtained using the convolution method following Butler and Stephens (2017, §2). This will be particularly useful for general m_i and s_i where the Binomial conditional distribution of T no longer holds. In fact, the conditional distribution of T will be Poisson binomial.

By noting that T converges in probability to $q\pi$ by the weak law of large numbers, we find that $\ln\{t/(q-t)\} \xrightarrow{p} \psi$ by Slutsky's theorem and the Continuous mapping theorem, so $\hat{\psi}_c$ is consistent. As the conditional maximum likelihood estimator depends on the data only through the sufficient statistic, its bias and variance can be calculated using (5.24) and (5.25) but replacing $\hat{\psi}(T)$ by $\hat{\psi}_c(T)$.

5.3.1.3 Modified profile maximum likelihood

Davison (2003, §12) showed that for a linear exponential family in canonical form, the modified profile log-likelihood function of Barndorff-Nielson (1983) reduces to (see Section 2.1.5)

$$l_{mp}(\psi) = l(\psi, \hat{\lambda}_\psi) + \frac{1}{2}\ln\{\det j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\}, \tag{5.33}$$

where $l(\psi, \hat{\lambda}_\psi)$ is the profile log-likelihood obtained from (5.20) by substituting the constrained maximum likelihood estimator of λ and where $j_{\lambda\lambda}(\psi, \lambda)$ is the observed information per observation for the λ components and is given by the negative of the second derivative of the log-likelihood function with respect to λ. In the setting $m_i = m$ and $s_i = (m+1)/2$, $j_{\lambda\lambda}(\psi, \lambda)$ becomes the $q \times q$ matrix with ith diagonal element

$$-\frac{\partial^2 l(\psi,\gamma)}{\partial \gamma^2} = \frac{\exp(\psi+\gamma)}{\big(1+\exp(\psi+\gamma)\big)^2} + m\frac{\exp(\gamma)}{\big(1+\exp(\gamma)\big)^2}, \tag{5.34}$$

and zero elsewhere, and where we observed in Section 5.3.1.1 that the solution of (5.21), $\hat{\lambda}_\psi = \hat{\gamma}_\psi$, is not available in closed form. This means that there is no closed form expres-

sion for (5.33) and we calculate the maximum modified profile log-likelihood estimator, $\hat{\psi}_{mp}$, numerically and evaluate its bias and variance using (5.24) and (5.25), respectively, by replacing $\hat{\psi}(T)$ with $\hat{\psi}_{mp}(T)$.

5.3.1.4 Firth's adjusted score equations method

When θ is the canonical parameter of an exponential family model like in the model considered here, Firth (1993) showed that the adjusted score equations estimator of θ, obtained as the solution of (2.37), is equivalent to the maximiser of the penalised log-likelihood function

$$l_*(\theta) = l(\theta) + \frac{1}{2}\ln\{\det i(\theta)\}, \tag{5.35}$$

where $i(\theta) = \mathrm{E}\{j(\theta)\}$ is the Fisher information matrix. In the setting $m_i = m$ and $s_i = (m+1)/2$, the second order partial derivatives of $l(\psi, \lambda_i)$ are

$$\frac{\partial^2 l(\psi,\lambda_i)}{\partial \psi^2} = -\sum_{i=1}^{q} \frac{\exp(\psi+\lambda_i)}{\big(1+\exp(\psi+\lambda_i)\big)^2} \tag{5.36}$$

$$\frac{\partial^2 l(\psi,\lambda_i)}{\partial \psi \partial \lambda_i} = -\frac{\exp(\psi+\lambda_i)}{\big(1+\exp(\psi+\lambda_i)\big)^2} \tag{5.37}$$

$$\frac{\partial^2 l(\psi,\lambda_i)}{\partial \lambda_i^2} = -\frac{\exp(\psi+\lambda_i)}{\big(1+\exp(\psi+\lambda_i)\big)^2} - m\frac{\exp(\lambda_i)}{\big(1+\exp(\lambda_i)\big)^2}. \tag{5.38}$$

Since the above derivatives do not depend on the data, the Fisher information matrix coincides with the observed information and is given by

$$i(\psi,\lambda_i) = \begin{pmatrix} \sum_{i=1}^{q} V_{i1} & V_{11} & \cdots & V_{q1} \\ V_{11} & (V_{11}+V_{12}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ V_{q1} & 0 & \cdots & (V_{q1}+V_{q2}) \end{pmatrix}, \tag{5.39}$$

where $V_{i1} = \big(\exp(\psi+\lambda_i)\big)/\big(1+\exp(\psi+\lambda_i)\big)^2$ and $V_{i2} = m\big(\exp(\lambda_i)\big)/\big(1+\exp(\lambda_i)\big)^2$, $i = 1,\ldots,q$. The determinant of $i(\psi,\lambda_i)$ is obtained using the identity (see Magnus and Neudecker, 2007, Chapter 1, p.g. 28)

$$\text{If D is invertible, then:} \qquad \det\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(D)\det(A - BD^{-1}C), \tag{5.40}$$

where A, B, C, and D are matrices of dimension $n \times n$, $n \times m$, $m \times n$, and $m \times m$ respectively, and simplifies to

$$\det i(\psi, \lambda_i) = \left(\prod_{i=1}^{q} (V_{i1} + V_{i2}) \right) \left(\sum_{i=1}^{q} \frac{V_{i1} V_{i2}}{V_{i1} + V_{i2}} \right). \tag{5.41}$$

Therefore the penalty function that needs to be added to the log-likelihood function is

$$\frac{1}{2} \ln\{\det i(\psi, \lambda_i)\} = \frac{1}{2} \sum_{i=1}^{q} \ln(V_{i1} + V_{i2}) + \frac{1}{2} \ln \left(\sum_{i=1}^{q} \frac{V_{i1} V_{i2}}{V_{i1} + V_{i2}} \right). \tag{5.42}$$

The score equations for the penalised log-likelihood of Firth (1993), $l_*(\psi, \lambda_i)$, with respect to λ_i and ψ involve cumbersome expressions and have no closed form solution so the penalised log-likelihood of Firth (1993) estimator of ψ, denoted by $\hat{\psi}_*$ is obtained numerically. The bias and variance of $\hat{\psi}_*$ are calculated using (5.24) and (5.25), respectively.

5.3.2 Binary matched pairs model

The binary matched pairs model is a special case of the binomial matched pairs model when $m = 1$. This implies that in the setting of Lunardon (2018), $s_i = 1$, and so $a = d = 0$, where a, b, c and d denote the number of pairs of the form $(0,0), (0,1), (1,0)$ and $(1,1)$ respectively, with $a + b + c + d = q$. Note that $\sum_{i=1}^{q} y_{i1} = c + d$, $\sum_{i=1}^{q} y_{i2} = b + d$ and $\sum_{i=1}^{q} (y_{i1} + y_{i2}) = b + c + 2d$. We call pairs of the form $(0,0)$ and $(1,1)$ concordant, while pairs of the form $(0,1)$ and $(1,0)$ are called discordant. In this case, $\hat{\gamma}_\psi = -\psi/2$ and so the profile log-likelihood for ψ is (see Davison, 2003, Example 12.23)

$$l_p(\psi) = \psi t - 2q \ln\{1 + \exp(\psi/2)\}, \tag{5.43}$$

which is maximised at

$$\begin{aligned} \hat{\psi} &= 2 \ln \left(\frac{t}{q - t} \right) \\ &= 2 \ln \left(\frac{c}{b} \right). \end{aligned} \tag{5.44}$$

Alternatively, $\hat{\psi}$ can be obtained by substituting $m = 1$ in (5.23). Davison (2003) showed that $\hat{\psi}$ converges in probability to 2ψ as $q \to \infty$, thus it is inconsistent.

Only discordant pairs enter the conditional log-likelihood and it is given by

$$l_c(\psi) = c\psi - (b+c)\ln\{\exp(\psi)+1\}, \tag{5.45}$$

which is maximised at

$$\hat{\psi}_c = \ln\left(\frac{c}{b}\right), \tag{5.46}$$

which converges in probability to ψ as $q \to \infty$, as noted by Davison (2003, Example 12.23), so it is consistent.

Substituting $\hat{\gamma}_\psi = -\psi/2$ in (5.34) gives $j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) = 2\big(\exp(\psi/2)\big)/\big(1+\exp(\psi/2)\big)^2$, and in this case Davison (2003, Example 12.23) showed that

$$l_{mp}(\psi) = \frac{1}{4}\psi(q+4t) - 3q\ln\{1+\exp(\psi/2)\}, \tag{5.47}$$

which is maximized at

$$\begin{aligned} \hat{\psi}_{mp} &= 2\ln\left(\frac{q+4t}{5q-4t}\right) \\ &= 2\ln\left(\frac{b+5c}{c+5b}\right), \end{aligned} \tag{5.48}$$

where the latter converges in probability to $2\ln\left[\{1+5\exp(\psi)\}/\{5+\exp(\psi)\}\right]$ as $q \to \infty$. Note that when $c = 0$ or $b = 0$, $\hat{\psi}_{mp}$ is $2\ln(1/5)$ or $2\ln(5)$, respectively, i.e. $\hat{\psi}_{mp}$ is finite. Although $\hat{\psi}_{mp}$ is inconsistent, Davison (2003) showed that it is less biased than $\hat{\psi}$.

5.3.3 Penalised likelihood based on adjusted responses method

In order to avoid infinite estimates of ψ, as is the case with $\hat{\psi}$, $\hat{\psi}_c$ and $\hat{\psi}_{mp}$ (for $m \neq 1$), when all of the y_{i1} observations are zero or one, we propose to adjust the log-likelihood function by adding a small number $\delta > 0$ to each success, y_{i1} and y_{i2}, and to each failure, $1 - y_{i1}$ and $m_i - y_{i2}$. The penalised log-likelihood function based on adjusted responses for $\theta = (\psi, \lambda_1, \ldots, \lambda_q)^\intercal$ becomes

$$\begin{aligned} l_a(\theta) &= \sum_{i=1}^{q} \psi(y_{i1}+\delta) + \sum_{i=1}^{q} \lambda_i(y_{i1}+y_{i2}+2\delta) \\ &- \sum_{i=1}^{q} \left[(1+2\delta)\ln\{1+\exp(\psi+\lambda_i)\} + (m_i+2\delta)\ln\{1+\exp(\lambda_i)\}\right]. \end{aligned} \tag{5.49}$$

When $m_i = m$ and $s_i = (m+1)/2$, the score equations for the above log-likelihood function with respect to γ and ψ simplify respectively to

$$\frac{m+1+4\delta}{2} - (1+2\delta)\frac{\exp(\psi+\gamma)}{1+\exp(\psi+\gamma)} - (m+2\delta)\frac{\exp(\gamma)}{1+\exp(\gamma)} = 0, \tag{5.50}$$

$$t + q\delta - q(1+2\delta)\frac{\exp(\psi+\gamma)}{1+\exp(\psi+\gamma)} = 0, \tag{5.51}$$

where we set the constrained penalised maximum likelihood estimator of λ_i, based on adjusted responses, for a fixed value of ψ, denoted by $\hat{\lambda}_{i,\psi,a}$, to be $\hat{\gamma}_{\psi,a}$ because it will be identical for all $i = 1,\ldots,q$. The simultaneous solution of (5.50) and (5.51) give the penalised maximum likelihood estimators of γ and ψ, based on adjusted responses

$$\hat{\gamma}_a = \ln\left[\frac{q(m+1+2\delta)-2t}{q(m-1+2\delta)+2t}\right], \tag{5.52}$$

$$\hat{\psi}_a = \ln\left[\frac{(t+q\delta)\{q(m-1+2\delta)+2t\}}{\{q(m+1+2\delta)-2t\}\{q(1+\delta)-t\}}\right]. \tag{5.53}$$

Note that when $\delta = 0$, $\hat{\psi}_a = \hat{\psi}$. Note also that when $t = 0$ or $t = q$, $\hat{\psi}_a$ is finite, while when $t = q/2$, $\hat{\psi}_a = 0$. When $m = 1$,

$$\hat{\psi}_a = 2\ln\left[\frac{t+q\delta}{q(1+\delta)-t}\right]. \tag{5.54}$$

Since $\hat{\psi}_a$ depends on the data only through the sufficient statistic t, its bias and variance are computed using (5.24) and (5.25), respectively.

5.3.3.1 Probability limit of the penalised likelihood estimator based on adjusted responses

In this section we obtain the probability limit of the penalised log-likelihood estimator based on adjusted responses and derive the relationship that δ should satisfy in order to make this estimator consistent. We also show how the modified profile log-likelihood estimator (when $m = 1$) and the conditional log-likelihood estimator can be recovered for particular values of δ.

When $m = 1$, as $q \to \infty$, $\hat{\psi}_a$ converges in probability to

$$2\ln\left(\frac{\delta\{\exp(\psi)+1\}+\exp(\psi)}{\delta\{\exp(\psi)+1\}+1}\right), \tag{5.55}$$

while for a general m, we find that as $q \to \infty$, $\hat{\psi}_a$ converges in probability to

$$\ln\left(\frac{\left[\delta\{\exp(\psi)+1\}+\exp(\psi)\right]\left[\{m-1+2\delta\}\{\exp(\psi)+1\}+2\exp(\psi)\right]}{\left[\delta\{\exp(\psi)+1\}+1\right]\left[\{m+1+2\delta\}\{\exp(\psi)+1\}-2\exp(\psi)\right]}\right). \tag{5.56}$$

Similar to Table 12.3 of Davison (2003), Table 5.6 compares the limiting values of $\hat{\psi}$, $\hat{\psi}_c$, $\hat{\psi}_{mp}$ and $\hat{\psi}_a$ when $m = 1$ for a set of values of ψ ranging from 0 to 5 and a set of values of δ ranging from 0.05 to 0.50. We note that for any given ψ, there exists a value of δ for which the limit of $\hat{\psi}_a$ is closer to the truth than $\hat{\psi}_{mp}$. In other words, there is evidence that there exists a δ value such that $\hat{\psi}_a$ converges to the truth faster than $\hat{\psi}_{mp}$. These values of δ decrease as the true value of ψ increase. Observe also that when $\delta = 0.25$, the limiting value of $\hat{\psi}_a$ coincides with that of $\hat{\psi}_{mp}$. In fact substituting $\delta = 0.25$ in (5.55), we find that $\hat{\psi}_a$ converges in probability to the same limit of $\hat{\psi}_{mp}$ given in Section 5.3.2. This means that the penalised log-likelihood estimator based on adjusted responses recovers the modified profile log-likelihood estimator when $m = 1$ and $\delta = 0.25$, i.e. $\hat{\psi}_a = \hat{\psi}_{mp}$.

Table 5.6: Probability limits of profile, conditional, modified profile and adjusted log-likelihood based on adjusted responses estimators of the log odds ratio ψ in the binary matched pairs model when $m = 1$. For each ψ, the value in bold face corresponds to the limiting value closest to the truth if we ignore $\hat{\psi}_c$.

ψ		0	0.5	1	1.5	2.0	2.5	3	4	5
Limit of $\hat{\psi}$		0	1	2	3	4	5	6	8	10
Limit of $\hat{\psi}_c$		0	0.5	1	1.5	2	2.5	3	4	5
Limit of $\hat{\psi}_{mp}$		0	0.66	1.27	1.81	2.24	2.56	2.79	3.05	3.16
Limit of $\hat{\psi}_a$	0.05	0	0.91	1.79	2.63	3.41	4.09	4.66	5.44	5.83
	0.10	0	0.83	1.62	2.36	3.00	3.52	3.93	4.43	**4.65**
	0.15	0	0.76	1.49	2.14	2.69	3.12	3.44	**3.82**	3.97
	0.20	0	0.71	1.37	1.96	2.44	2.81	**3.08**	3.38	3.51
	0.25	0	0.66	1.27	1.81	2.24	**2.56**	2.79	3.05	3.16
	0.30	0	0.62	1.19	1.68	**2.07**	2.36	2.56	2.79	2.88
	0.35	0	0.58	1.12	1.57	1.93	2.19	2.37	2.57	2.65
	0.40	0	0.55	1.05	**1.47**	1.81	2.05	2.21	2.39	2.46
	0.45	0	0.52	**0.99**	1.39	1.70	1.92	2.07	2.24	2.30
	0.50	0	**0.49**	0.94	1.32	1.60	1.81	1.95	2.10	2.16

When $m = 1$, in order to make $\hat{\psi}_a$ consistent we need to equate the ratio inside the logarithm of (5.55) with $\sqrt{\exp(\psi)}$ which simplifies to the equation

$$\big[\exp(\psi) - 1\big]\big[\delta^2\{\exp(\psi)\}^2 + \{2\delta^2 - 1\}\{\exp(\psi)\} + \delta^2\big] = 0. \tag{5.57}$$

When $\psi = 0$, there is no adjustment because there is no bias so we consider the positive solution of the quadratic equation $[\delta^2\{\exp(\psi)\}^2 + \{2\delta^2 - 1\}\{\exp(\psi)\} + \delta^2] = 0$ in terms of δ which simplifies to

$$\delta = \frac{\sqrt{\exp(\psi)}}{1 + \exp(\psi)}. \tag{5.58}$$

Substituting (5.58) into (5.54) gives us an implicit equation in $\hat{\psi}_a$ which when solved numerically gives the same estimate as $\hat{\psi}_c$. This means that the value of δ that achieves consistency of $\hat{\psi}_a$ is the one that recovers $\hat{\psi}_c$. This is disadvantageous because we inherit exactly the same problems with conditional log-likelihood (i.e. infinite estimates) if we attempt to tune δ to make $\hat{\psi}_a$ consistent. The value of δ in terms of t and q that recovers $\hat{\psi}_c$ is obtained by equating (5.54) with $\hat{\psi}_c$ and simplifies to

$$\delta = \sqrt{\frac{t(3qt - 2t^2 - q^2)}{q^2(2t - q)}}. \tag{5.59}$$

Observe that when $t = 0$ or $t = q$, $\delta = 0$ and so $\hat{\psi}_a = \hat{\psi}$, while when $t = q/2$, δ is infinite.

For a general m, the relationship that δ should satisfy in order to make $\hat{\psi}_a$ consistent is found by equating the ratio inside the logarithm of (5.56) with $\exp(\psi)$ which simplifies to the equation

$$\begin{aligned} \delta\big\{\exp(\psi)\big\}^3\big\{m - 1 + 2\delta\big\} \quad &- \quad \big\{\exp(\psi)\big\}^2\big\{2 + \delta(1 - m - 2\delta)\big\} \\ &+ \quad \big\{\exp(\psi)\big\}\big\{2 + \delta(1 - m - 2\delta)\big\} - \delta(m - 1 + 2\delta) = 0, \end{aligned} \tag{5.60}$$

with no closed form solution. The value of δ in terms of t, q and m that recovers $\hat{\psi}_c$ satisfies the equation

$$2q^2\delta^2(q - 2t) + q^2\delta\big\{q(m - 1) + 2t(1 - m)\big\} - 2t(q^2 - 3tq + 2t^2) = 0. \tag{5.61}$$

Figure 5.4 shows a plot of δ, the root of (5.60), against ψ for $m = 1$, $m = 3$, $m = 11$ and

$m = 39$. This plot shows that there is a scaled logistic relationship between δ and ψ and that the best choice of δ lies in the range $0 < \delta < 0.5$. Given the true value of ψ, as m increases, the value of δ that makes $\hat{\psi}_a$ consistent decreases to zero. This is expected because we know from standard asymptotic theory that the maximum likelihood estimator is asymptotically consistent.

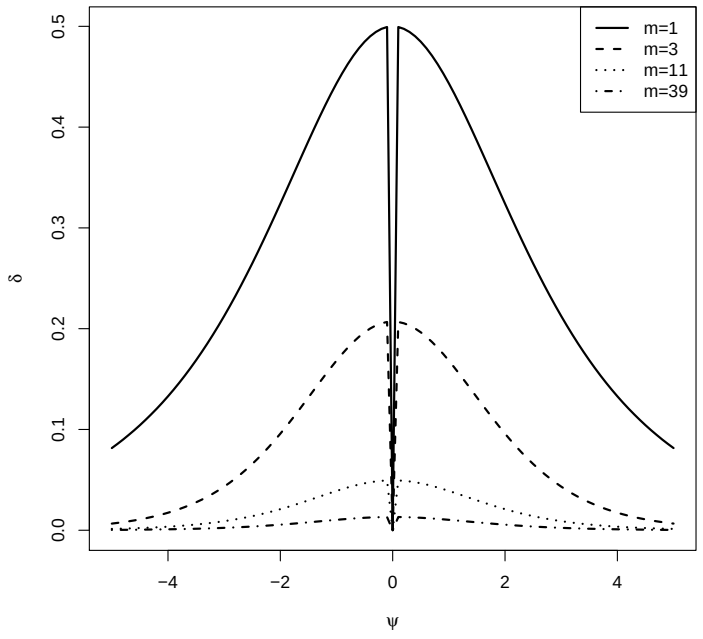

Figure 5.4: Plot of the value of δ, that makes the penalised log-likelihood based on adjusted responses estimator, $\hat{\psi}_a$, consistent, against ψ for $m = 1$, $m = 3$, $m = 11$ and $m = 39$.

5.3.4 Indirect inference estimation of the log odds ratio

Suppose that $\hat{\psi}$ is some initial estimator of ψ, not necessarily the maximum likelihood estimator, then the simplest method of bias reduction of $\hat{\psi}$ via indirect inference relies on solving the equation

$$\tilde{\psi} = \hat{\psi} - \mathrm{B}_{\hat{\psi}}(\tilde{\psi}, \lambda), \tag{5.62}$$

with respect to $\tilde{\psi}$ where $\mathrm{B}_{\hat{\psi}}(\tilde{\psi}, \lambda) = \mathrm{E}_{\tilde{\psi},\lambda}(\hat{\psi}) - \tilde{\psi}$ is the bias function of $\hat{\psi}$ evaluated at $\tilde{\psi}$ and $\lambda = \lambda_1, \ldots, \lambda_q$, as described in Section 2.2.3. We call $\tilde{\psi}$ the indirect inference

estimator of ψ. Alternatively, (5.62) can be written as

$$\hat{\psi} = \mathrm{E}_{\tilde{\psi},\lambda}(\hat{\psi}). \tag{5.63}$$

Since we want to reduce the bias of $\hat{\psi}_a$ when $m_i = m$ and $s_i = (m+1)/2$, our indirect inference estimator $\tilde{\psi}$ is the solution of

$$\hat{\psi}_a = \mathrm{E}_{\tilde{\psi},\gamma}(\hat{\psi}_a). \tag{5.64}$$

There are two possible estimators of γ to use to plug in $\mathrm{E}_{\tilde{\psi},\gamma}(\hat{\psi}_a)$: the penalised maximum likelihood estimator based on adjusted responses, $\hat{\gamma}_a$, or the constrained penalised maximum likelihood estimator based on adjusted responses, $\hat{\gamma}_{\psi,a}$. As the expectation of $\hat{\psi}_a$ can be obtained using complete enumeration, two versions of $\tilde{\psi}$, $\tilde{\psi}_{a1}$ and $\tilde{\psi}_{a2}$, can be defined

$$\hat{\psi}_a(T) = \sum_{u=0}^{q} \hat{\psi}_a(u)\mathrm{Pr}(T = u; \tilde{\psi}_{a1}, \hat{\gamma}_a), \tag{5.65}$$

$$\hat{\psi}_a(T) = \sum_{u=0}^{q} \hat{\psi}_a(u)\mathrm{Pr}(T = u; \tilde{\psi}_{a2}, \hat{\gamma}_{\psi,a}), \tag{5.66}$$

where $\mathrm{Pr}(T = u; \tilde{\psi}_{a1}, \hat{\gamma}_a)$ and $\mathrm{Pr}(T = u; \tilde{\psi}_{a2}, \hat{\gamma}_{\psi,a})$ are the unconditional density of the sufficient statistic T which is binomial with denominator q and success probabilities $\exp(\tilde{\psi}_{a1} + \hat{\gamma}_a)/\left\{1 + \exp(\tilde{\psi}_{a1} + \hat{\gamma}_a)\right\}$ and $\exp(\tilde{\psi}_{a2} + \hat{\gamma}_{\psi,a})/\left\{1 + \exp(\tilde{\psi}_{a2} + \hat{\gamma}_{\psi,a})\right\}$, respectively. A third version of $\tilde{\psi}$, $\tilde{\psi}_{a*}$, can be obtained by using the conditional density $\mathrm{pr}(T = u|S_i; \psi)$ which is binomial with denominator q and success probability $\exp(\psi)/\{\exp(\psi) + 1\}$,

$$\hat{\psi}_a(T) = \sum_{u=0}^{q} \hat{\psi}_a(u)\mathrm{Pr}(T = u|S_i; \tilde{\psi}_{a*}). \tag{5.67}$$

No closed form solution exists for any of the above three estimators, so we solve numerically and calculate their expectation and variance using (5.24) and (5.25). For $t \in \{0, q\}$, $\tilde{\psi}_{a1}$, $\tilde{\psi}_{a2}$ and $\tilde{\psi}_{a*}$ have no solution. In fact the expectations in the right hand side of (5.65), (5.66) and (5.67) are all bounded below by $\hat{\psi}_a(0)$ and above by $\hat{\psi}_a(q)$. This means that when the binary observations are all zero or one the indirect inference estimator is not defined which is unfortunate because even though we overcome the problem of infinite estimates at $t = 0$ and $t = q$ by introducing a penalised likelihood estimator based on adjusted responses, when we attempt to reduce the bias of the later the same problem

appears again at those boundary values of t.

5.3.5 Complete enumeration study

In this section we reproduce the complete enumeration study in Lunardon (2018, Table 1) which compares the finite sample bias and variance of estimators derived from profile, conditional, modified profile and penalized (Firth (1993)) likelihoods, denoted by $\hat{\psi}$, $\hat{\psi}_c$, $\hat{\psi}_{mp}$ and $\hat{\psi}_*$, respectively. We enrich this study by adding the penalised maximum likelihood estimator based on adjusted responses, $\hat{\psi}_a$, and the indirect inference estimator based on $\hat{\psi}_a$ using the conditional model, denoted by $\tilde{\psi}_{a*}$, for a set of 20 values of δ ranging from 0.05 to 1.00. The comparison in Lunardon (2018, Table 1) also assesses the coverage probability and length of 95% confidence intervals for ψ based on the chi-squared approximation to the distribution of $W_*(\psi)$ and to the distributions of the profile, conditional and modified profile log-likelihood ratios, denoted by $W(\psi)$, $W_c(\psi)$ and $W_{mp}(\psi)$, respectively. We extend this comparison by adding the coverage probability and length of 95% confidence intervals for ψ based on the chi-squared approximation to the distribution of the penalised log-likelihood ratio based on adjusted responses, denoted by $W_a(\psi)$. The exact bias and variance of estimators and the exact coverage probability and average length of confidence intervals is obtained through complete enumeration because the estimators and confidence intervals all depend on the sufficient statistic $T = \sum_{i=1}^{q} Y_{i1}$ and so the distribution of T given $S_1 = s_1, \ldots, S_q = s_q$ can be computed numerically following Butler and Stephens (2017). These summaries however, are computed only for $t \in \{1, \ldots, q-1\}$ since when $t \in \{0, q\}$, $\hat{\psi}$, $\hat{\psi}_c$ and $\hat{\psi}_{mp}$ (for $m \neq 1$) are infinite.

Tables 5.7-5.12 report the bias and variance of estimators, while Tables 5.13-5.18 report the coverage probability and average length of confidence intervals with nominal level 95% when the true log odds ratio ψ is unity, with $m \in \{1, 3, 11, 39\}$ and $q \in \{30, 100, 1000\}$.

Overall, for fixed δ the numerical value of the bias and variance of the estimator $\hat{\psi}_a$ decrease as m increases. In many cases, however this means that the bias becomes more negative, i.e. the magnitude of the bias increases with m. This suggests that for any combination of q and m, there exists a particular value of δ, above which the bias of $\hat{\psi}_a$ does not improve. In fact for any combination of m and q, there exists a value of δ such that $\hat{\psi}_a$ has minimum bias which is smaller than the bias of the estimators $\hat{\psi}$, $\hat{\psi}_c$, $\hat{\psi}_{mp}$ and $\hat{\psi}_*$; for example, for $q = 30$, $m = 1$ this optimal δ in terms of bias is 0.45, for the combinations $q = 30, m = 39$ and $q \geq 100, m \geq 11$, the optimal value of δ that

gives minimum bias becomes smaller than 0.05. For $q = 30$, at the optimum δ value, the estimator $\hat{\psi}_a$ has smaller bias and variance than $\hat{\psi}_c$. This is also true for other values of q, except that it is not very clear from Tables 5.7-5.12 because we consider a specific set of values for δ; for example, for $q = 100$ and $m = 11$, at $\delta = 0.045$, $\hat{\psi}_a$ has bias and variance -0.03 and 0.47, respectively (both multiplied by 10), while for $q = 1000$ and $m = 1$, at $\delta = 0.444$, $\hat{\psi}_a$ has bias and variance 0.00 and 0.04, respectively (both multiplied by 10). This optimal value of δ decrease as m increases, but for a fixed m, as q increase above 30, the optimum δ value remain constant. This behaviour can be seen more clearly from Table 5.19 where the optimum value of δ that minimizes the bias of $\hat{\psi}_a$ was chosen from a finer set of δ values. For the values of m and q chosen in Table 5.19, the effect of fixing m and allowing q to increase is a larger optimal δ, while fixing q and increasing m decreases this optimal δ value. However, the pattern in Table 5.19 suggest that as both q and m are allowed to diverge, the optimal δ value becomes close to zero and this makes intuitive sense because we know that $\delta = 0$ gives rise to the maximum likelihood estimator $\hat{\psi}$ which is asymptotically unbiased when both m and q tend to ∞.

Nevertheless, the bias results for the estimator $\tilde{\psi}_{a*}$ show a marked improvement over $\hat{\psi}$, $\hat{\psi}_{mp}$ and $\hat{\psi}_c$, for all values of q and m. When $\tilde{\psi}_{a*}$ is compared with $\hat{\psi}_*$, the former has smaller bias and variance for $m = 1$ and $m = 3$ for all values of q. When compared with $\hat{\psi}_a$, the bias of $\tilde{\psi}_{a*}$ is reduced for all values of δ except the optimal one. The interesting result to note here is that as δ increases, the bias and variance of $\tilde{\psi}_{a*}$ approach that of $\hat{\psi}_c$ for all combinations of q and m considered. To wrap up this comparison of estimators, if we were to choose a δ based on $\hat{\psi}_{a*}$, then it will be the one that is close to (but not equal to) zero because it will give the smallest bias and variance. However, $\hat{\psi}_{a*}$ does not perform better than $\hat{\psi}_a$ all the time. In fact, if we were to choose a δ based on $\hat{\psi}_a$, then it will be the optimal δ that minimizes the bias of $\hat{\psi}_a$ because the bias and variance of $\hat{\psi}_a$ at this optimal δ is smaller than that of any other estimator in the table for any combination of q and m. Having said that, in practice we are only given a data set and we don't know the particular δ for that data set so $\tilde{\psi}_{a*}$ will be a good choice because its bias and variance are very competitive regardless of the value of δ.

Concerning the coverage probability and average length of 95% confidence intervals, Lunardon (2018) noted that intervals derived from $W_{mp}(\psi)$ and $W_*(\psi)$ are consistent with those from $W_c(\psi)$ for $m \geq 3$. The coverage probability and average length of confidence intervals derived from $W_a(\psi)$ show an improvement over those derived from $W_c(\psi)$ for particular values of δ (shown in bold face in Tables 5.13-5.18). We considered 20 values

of δ ranging from 0.31 to 0.50 for $m = 1$, 0.06 to 0.25 for $m = 3$ and 0.01 to 0.20 for $m = 11$ and $m = 39$. For those particular values of δ in bold face, the coverage probability derived from $W_a(\psi)$ is closer to the nominal coverage of 95% than that derived from $W_c(\psi)$. This agrees with the fact that $\hat{\psi}_a$ performs better than $\hat{\psi}_c$ in terms of bias and variance for some optimal value of δ. As the value of δ increase the average length of confidence intervals derived from $W_a(\psi)$ decreases which is expected because the variance (and hence standard error) of $\hat{\psi}_a$ becomes smaller as δ becomes larger.

In conclusion it has been shown how, in the binomial matched pairs model, finite estimates of the log odds ratio are produced in cases where the observations are either all equal to zero or all equal to one by penalising the log-likelihood function through the additive adjustment of a tuning parameter $\delta > 0$ to each success and failure. This value δ was then used as a parameter that could be tuned to improve the bias and/or variance of the penalised log-likelihood estimator based on adjusted responses. The indirect inference method was applied to further reduce the bias of $\hat{\psi}_a$. The method can be applied in principle to any parametric model. We are also free to use estimators other than $\hat{\psi}_a$ as initial estimates. Indeed the setting considered here where m_i and s_i are fixed to m and $(m+1)/2$, respectively is a special case and perhaps a large scale simulation study would be useful to account for different stratum sizes and totals.

Table 5.7: Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the bias and variance (in parentheses) of estimators derived from profile, conditional, modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods, with all entries multiplied by 10; seventh column show the bias and variance (in parentheses) of the indirect inference $\hat{\psi}_{a*}$ using the conditional model, with all entries multiplied by 10.

	$\hat{\psi}$	$\hat{\psi}_{\mathbf{c}}$	$\hat{\psi}_{\mathbf{mp}}$	$\hat{\psi}_{*}$	$\hat{\psi}_{\mathbf{a}}$	$\tilde{\psi}_{\mathbf{a}*}$
δ			$q = 30, m = 1$			
0.05					8.49 (5.58)	0.08 (1.80)
0.10					6.65 (4.25)	0.17 (1.85)
0.15					5.16 (3.39)	0.22 (1.88)
0.20					3.94 (2.78)	0.26 (1.90)
0.25					2.91 (2.33)	0.29 (1.91)
0.30					2.03 (1.99)	0.31 (1.92)
0.35					1.27 (1.72)	0.33 (1.92)
0.40					0.60 (1.50)	0.34 (1.93)
0.45					0.01 (1.33)	0.35 (1.93)
0.50	10.89 (7.87)	0.44 (1.97)	2.91 (2.33)	2.76 (2.27)	-0.52 (1.18)	0.36 (1.94)
0.55					-0.99 (1.06)	0.37 (1.94)
0.60					-1.42 (0.95)	0.38 (1.94)
0.65					-1.81 (0.87)	0.38 (1.95)
0.70					-2.16 (0.79)	0.39 (1.95)
0.75					-2.49 (0.72)	0.39 (1.95)
0.80					-2.79 (0.66)	0.40 (1.95)
0.85					-3.06 (0.61)	0.40 (1.95)
0.90					-3.32 (0.57)	0.40 (1.95)
0.95					-3.55 (0.53)	0.41 (1.95)
1.00					-3.77 (0.49)	0.41 (1.96)
δ			$q = 30, m = 3$			
0.05					2.26 (2.30)	0.15 (1.82)
0.10					1.24 (1.84)	0.22 (1.87)
0.15					0.41 (1.53)	0.27 (1.89)
0.20					-0.29 (1.30)	0.30 (1.91)
0.25					-0.88 (1.13)	0.32 (1.92)
0.30					-1.40 (0.99)	0.34 (1.93)
0.35					-1.85 (0.88)	0.35 (1.93)
0.40					-2.25 (0.79)	0.37 (1.94)
0.45					-2.61 (0.71)	0.37 (1.94)
0.50	3.56 (3.06)	0.44 (1.97)	0.69 (1.82)	0.48 (1.72)	-2.94 (0.64)	0.38 (1.95)
0.55					-3.23 (0.59)	0.39 (1.95)
0.60					-3.50 (0.54)	0.39 (1.95)
0.65					-3.74 (0.50)	0.40 (1.95)
0.70					-3.97 (0.46)	0.40 (1.95)
0.75					-4.18 (0.43)	0.41 (1.95)
0.80					-4.37 (0.40)	0.41 (1.95)
0.85					-4.56 (0.37)	0.41 (1.96)
0.90					-4.72 (0.35)	0.41 (1.96)
0.95					-4.88 (0.33)	0.42 (1.96)
1.00					-5.03 (0.31)	0.42 (1.96)

Table 5.8: Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the bias and variance (in parentheses) of estimators derived from profile, conditional, modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods, with all entries multiplied by 10; seventh column show the bias and variance (in parentheses) of the indirect inference $\hat{\psi}_{a*}$ using the conditional model, with all entries multiplied by 10.

	$\hat{\psi}$	$\hat{\psi}_{\mathbf{c}}$	$\hat{\psi}_{\mathbf{mp}}$	$\hat{\psi}_{*}$	$\hat{\psi}_{\mathbf{a}}$	$\tilde{\psi}_{\mathbf{a}*}$
δ			$q = 30, m = 11$			
0.05					0.08 (1.62)	0.11 (1.81)
0.10					-0.85 (1.26)	0.19 (1.85)
0.15					-1.60 (1.02)	0.24 (1.88)
0.20					-2.22 (0.85)	0.27 (1.90)
0.25					-2.74 (0.73)	0.30 (1.91)
0.30					-3.19 (0.63)	0.32 (1.92)
0.35					-3.58 (0.55)	0.34 (1.93)
0.40					-3.92 (0.49)	0.35 (1.93)
0.45					-4.22 (0.44)	0.36 (1.94)
0.50	1.28 (2.23)	0.44 (1.97)	0.46 (1.94)	0.09 (1.74)	-4.49 (0.40)	0.37 (1.94)
0.55					-4.73 (0.36)	0.38 (1.94)
0.60					-4.95 (0.33)	0.39 (1.95)
0.65					-5.15 (0.30)	0.39 (1.95)
0.70					-5.34 (0.28)	0.40 (1.95)
0.75					-5.50 (0.26)	0.40 (1.95)
0.80					-5.66 (0.24)	0.40 (1.95)
0.85					-5.80 (0.22)	0.41 (1.95)
0.90					-5.94 (0.21)	0.41 (1.96)
0.95					-6.06 (0.20)	0.41 (1.96)
1.00					-6.17 (0.18)	0.42 (1.96)
δ			$q = 30, m = 39$			
0.05					-0.52 (1.46)	0.09 (1.80)
0.10					-1.44 (1.12)	0.17 (1.85)
0.15					-2.18 (0.90)	0.23 (1.88)
0.20					-2.79 (0.74)	0.26 (1.90)
0.25					-3.31 (0.62)	0.29 (1.91)
0.30					-3.75 (0.53)	0.31 (1.92)
0.35					-4.13 (0.46)	0.33 (1.93)
0.40					-4.47 (0.41)	0.34 (1.93)
0.45					-4.77 (0.36)	0.36 (1.94)
0.50	0.68 (2.04)	0.44 (1.97)	0.44 (1.96)	0.03 (1.74)	-5.03 (0.32)	0.37 (1.94)
0.55					-5.27 (0.29)	0.37 (1.94)
0.60					-5.48 (0.26)	0.38 (1.95)
0.65					-5.68 (0.24)	0.39 (1.95)
0.70					-5.85 (0.22)	0.39 (1.95)
0.75					-6.02 (0.20)	0.40 (1.95)
0.80					-6.17 (0.19)	0.40 (1.95)
0.85					-6.30 (0.17)	0.40 (1.95)
0.90					-6.43 (0.16)	0.41 (1.95)
0.95					-6.55 (0.15)	0.41 (1.96)
1.00					-6.66 (0.14)	0.41 (1.96)

Table 5.9: Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the bias and variance (in parentheses) of estimators derived from profile, conditional, modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods, with all entries multiplied by 10; seventh column show the bias and variance (in parentheses) of the indirect inference $\hat{\psi}_{a*}$ using the conditional model, with all entries multiplied by 10.

	$\hat{\psi}$	$\hat{\psi}_{\mathbf{c}}$	$\hat{\psi}_{\mathbf{mp}}$	$\hat{\psi}_{*}$	$\hat{\psi}_{\mathbf{a}}$	$\tilde{\psi}_{\mathbf{a}*}$
δ			$q = 100, m = 1$			
0.05					8.08 (1.57)	0.02 (0.52)
0.10					6.36 (1.23)	0.04 (0.52)
0.15					4.96 (0.99)	0.06 (0.52)
0.20					3.78 (0.82)	0.07 (0.52)
0.25					2.79 (0.69)	0.08 (0.52)
0.30					1.93 (0.59)	0.08 (0.52)
0.35					1.19 (0.51)	0.09 (0.52)
0.40					0.53 (0.45)	0.09 (0.53)
0.45					-0.05 (0.40)	0.09 (0.53)
0.50	10.24 (2.11)	0.12 (0.53)	2.79 (0.69)	2.74 (0.68)	-0.57 (0.35)	0.10 (0.53)
0.55					-1.03 (0.32)	0.10 (0.53)
0.60					-1.46 (0.29)	0.10 (0.53)
0.65					-1.84 (0.26)	0.10 (0.53)
0.70					-2.19 (0.24)	0.10 (0.53)
0.75					-2.51 (0.22)	0.11 (0.53)
0.80					-2.81 (0.20)	0.11 (0.53)
0.85					-3.08 (0.18)	0.11 (0.53)
0.90					-3.33 (0.17)	0.11 (0.53)
0.95					-3.56 (0.16)	0.11 (0.53)
1.00					-3.78 (0.15)	0.11 (0.53)
δ			$q = 100, m = 3$			
0.05					2.05 (0.66)	0.04 (0.52)
0.10					1.09 (0.54)	0.06 (0.52)
0.15					0.30 (0.45)	0.07 (0.52)
0.20					-0.37 (0.38)	0.08 (0.52)
0.25					-0.95 (0.33)	0.09 (0.52)
0.30					-1.45 (0.29)	0.09 (0.53)
0.35					-1.89 (0.26)	0.10 (0.53)
0.40					-2.29 (0.23)	0.10 (0.53)
0.45					-2.64 (0.21)	0.10 (0.53)
0.50	3.23 (0.84)	0.12 (0.53)	0.48 (0.51)	0.42 (0.50)	-2.96 (0.19)	0.10 (0.53)
0.55					-3.25 (0.18)	0.10 (0.53)
0.60					-3.52 (0.16)	0.11 (0.53)
0.65					-3.76 (0.15)	0.11 (0.53)
0.70					-3.99 (0.14)	0.11 (0.53)
0.75					-4.19 (0.13)	0.11 (0.53)
0.80					-4.39 (0.12)	0.11 (0.53)
0.85					-4.57 (0.11)	0.11 (0.53)
0.90					-4.73 (0.10)	0.11 (0.53)
0.95					-4.89 (0.10)	0.11 (0.53)
1.00					-5.04 (0.09)	0.11 (0.53)

Table 5.10: Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the bias and variance (in parentheses) of estimators derived from profile, conditional, modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods, with all entries multiplied by 10; seventh column show the bias and variance (in parentheses) of the indirect inference $\hat{\psi}_{a*}$ using the conditional model, with all entries multiplied by 10.

	$\hat{\psi}$	$\hat{\psi}_{\mathbf{c}}$	$\hat{\psi}_{\mathbf{mp}}$	$\hat{\psi}_{*}$	$\hat{\psi}_{\mathbf{a}}$	$\tilde{\psi}_{\mathbf{a}*}$
δ			$q = 100, m = 11$			
0.05					-0.13 (0.46)	0.03 (0.52)
0.10					-0.99 (0.36)	0.05 (0.52)
0.15					-1.70 (0.30)	0.06 (0.52)
0.20					-2.30 (0.25)	0.07 (0.52)
0.25					-2.80 (0.21)	0.08 (0.52)
0.30					-3.24 (0.19)	0.09 (0.52)
0.35					-3.62 (0.16)	0.09 (0.53)
0.40					-3.95 (0.15)	0.09 (0.53)
0.45					-4.25 (0.13)	0.10 (0.53)
0.50	0.96 (0.61)	0.12 (0.53)	0.15 (0.52)	0.05 (0.51)	-4.51 (0.12)	0.10 (0.53)
0.55					-4.75 (0.11)	0.10 (0.53)
0.60					-4.97 (0.10)	0.10 (0.53)
0.65					-5.17 (0.09)	0.11 (0.53)
0.70					-5.35 (0.08)	0.11 (0.53)
0.75					-5.52 (0.08)	0.11 (0.53)
0.80					-5.67 (0.07)	0.11 (0.53)
0.85					-5.81 (0.07)	0.11 (0.53)
0.90					-5.94 (0.06)	0.11 (0.53)
0.95					-6.07 (0.06)	0.11 (0.53)
1.00					-6.18 (0.06)	0.11 (0.53)
δ			$q = 100, m = 39$			
0.05					-0.72 (0.41)	0.03 (0.52)
0.10					-1.58 (0.32)	0.05 (0.52)
0.15					-2.29 (0.26)	0.06 (0.52)
0.20					-2.87 (0.22)	0.07 (0.52)
0.25					-3.37 (0.18)	0.08 (0.52)
0.30					-3.80 (0.16)	0.08 (0.52)
0.35					-4.17 (0.14)	0.09 (0.53)
0.40					-4.50 (0.12)	0.09 (0.53)
0.45					-4.79 (0.11)	0.10 (0.53)
0.50	0.36 (0.55)	0.12 (0.53)	0.12 (0.53)	0.01 (0.51)	-5.05 (0.10)	0.10 (0.53)
0.55					-5.29 (0.09)	0.10 (0.53)
0.60					-5.50 (0.08)	0.10 (0.53)
0.65					-5.69 (0.07)	0.10 (0.53)
0.70					-5.87 (0.07)	0.11 (0.53)
0.75					-6.03 (0.06)	0.11 (0.53)
0.80					-6.18 (0.06)	0.11 (0.53)
0.85					-6.31 (0.05)	0.11 (0.53)
0.90					-6.44 (0.05)	0.11 (0.53)
0.95					-6.56 (0.05)	0.11 (0.53)
1.00					-6.67 (0.04)	0.11 (0.53)

Table 5.11: Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the bias and variance (in parentheses) of estimators derived from profile, conditional, modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods, with all entries multiplied by 10; seventh column show the bias and variance (in parentheses) of the indirect inference $\hat{\psi}_{a*}$ using the conditional model, with all entries multiplied by 10.

	$\hat{\psi}$	$\hat{\psi}_{\mathbf{c}}$	$\hat{\psi}_{\mathbf{mp}}$	$\hat{\psi}_{*}$	$\hat{\psi}_{\mathbf{a}}$	$\tilde{\psi}_{\mathbf{a}*}$
δ			$q = 1000, m = 1$			
0.05					7.93 (0.15)	0.00 (0.05)
0.10					6.25 (0.12)	0.00 (0.05)
0.15					4.88 (0.10)	0.01 (0.05)
0.20					3.72 (0.08)	0.01 (0.05)
0.25					2.74 (0.07)	0.01 (0.05)
0.30					1.90 (0.06)	0.01 (0.05)
0.35					1.16 (0.05)	0.01 (0.05)
0.40					0.51 (0.04)	0.01 (0.05)
0.45					-0.07 (0.04)	0.01 (0.05)
0.50	10.02 (0.20)	0.01 (0.05)	2.74 (0.07)	2.74 (0.07)	-0.59 (0.04)	0.01 (0.05)
0.55					-1.05 (0.03)	0.01 (0.05)
0.60					-1.47 (0.03)	0.01 (0.05)
0.65					-1.85 (0.03)	0.01 (0.05)
0.70					-2.20 (0.02)	0.01 (0.05)
0.75					-2.52 (0.02)	0.01 (0.05)
0.80					-2.81 (0.02)	0.01 (0.05)
0.85					-3.08 (0.02)	0.01 (0.05)
0.90					-3.34 (0.02)	0.01 (0.05)
0.95					-3.57 (0.02)	0.01 (0.05)
1.00					-3.79 (0.01)	0.01 (0.05)
δ			$q = 1000, m = 3$			
0.05					1.97 (0.06)	0.00 (0.05)
0.10					1.04 (0.05)	0.01 (0.05)
0.15					0.26 (0.04)	0.01 (0.05)
0.20					-0.40 (0.04)	0.01 (0.05)
0.25					-0.97 (0.03)	0.01 (0.05)
0.30					-1.47 (0.03)	0.01 (0.05)
0.35					-1.91 (0.03)	0.01 (0.05)
0.40					-2.30 (0.02)	0.01 (0.05)
0.45					-2.65 (0.02)	0.01 (0.05)
0.50	3.12 (0.08)	0.01 (0.05)	0.40 (0.05)	0.40 (0.05)	-2.97 (0.02)	0.01 (0.05)
0.55					-3.26 (0.02)	0.01 (0.05)
0.60					-3.53 (0.02)	0.01 (0.05)
0.65					-3.77 (0.01)	0.01 (0.05)
0.70					-3.99 (0.01)	0.01 (0.05)
0.75					-4.20 (0.01)	0.01 (0.05)
0.80					-4.39 (0.01)	0.01 (0.05)
0.85					-4.57 (0.01)	0.01 (0.05)
0.90					-4.74 (0.01)	0.01 (0.05)
0.95					-4.89 (0.01)	0.01 (0.05)
1.00					-5.04 (0.01)	0.01 (0.05)

Table 5.12: Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the bias and variance (in parentheses) of estimators derived from profile, conditional, modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods, with all entries multiplied by 10; seventh column show the bias and variance (in parentheses) of the indirect inference $\hat{\psi}_{a*}$ using the conditional model, with all entries multiplied by 10.

	$\hat{\psi}$	$\hat{\psi}_{\mathbf{c}}$	$\hat{\psi}_{\mathbf{mp}}$	$\hat{\psi}_{*}$	$\hat{\psi}_{\mathbf{a}}$	$\tilde{\psi}_{\mathbf{a}*}$
δ			$q = 1000, m = 11$			
0.05					-0.20 (0.04)	0.00 (0.05)
0.10					-1.05 (0.04)	0.00 (0.05)
0.15					-1.74 (0.03)	0.01 (0.05)
0.20					-2.33 (0.02)	0.01 (0.05)
0.25					-2.82 (0.02)	0.01 (0.05)
0.30					-3.26 (0.02)	0.01 (0.05)
0.35					-3.63 (0.02)	0.01 (0.05)
0.40					-3.96 (0.01)	0.01 (0.05)
0.45					-4.26 (0.01)	0.01 (0.05)
0.50	0.85 (0.06)	0.01 (0.05)	0.04 (0.05)	0.03 (0.05)	-4.52 (0.01)	0.01 (0.05)
0.55					-4.76 (0.01)	0.01 (0.05)
0.60					-4.98 (0.01)	0.01 (0.05)
0.65					-5.17 (0.01)	0.01 (0.05)
0.70					-5.35 (0.01)	0.01 (0.05)
0.75					-5.52 (0.01)	0.01 (0.05)
0.80					-5.67 (0.01)	0.01 (0.05)
0.85					-5.81 (0.01)	0.01 (0.05)
0.90					-5.95 (0.01)	0.01 (0.05)
0.95					-6.07 (0.01)	0.01 (0.05)
1.00					-6.18 (0.01)	0.01 (0.05)
δ			$q = 1000, m = 39$			
0.05					-0.80 (0.04)	0.00 (0.05)
0.10					-1.64 (0.03)	0.00 (0.05)
0.15					-2.33 (0.03)	0.01 (0.05)
0.20					-2.90 (0.02)	0.01 (0.05)
0.25					-3.39 (0.02)	0.01 (0.05)
0.30					-3.82 (0.02)	0.01 (0.05)
0.35					-4.19 (0.01)	0.01 (0.05)
0.40					-4.51 (0.01)	0.01 (0.05)
0.45					-4.80 (0.01)	0.01 (0.05)
0.50	0.25 (0.05)	0.01 (0.05)	0.01 (0.05)	0.00 (0.05)	-5.06 (0.01)	0.01 (0.05)
0.55					-5.29 (0.01)	0.01 (0.05)
0.60					-5.50 (0.01)	0.01 (0.05)
0.65					-5.70 (0.01)	0.01 (0.05)
0.70					-5.87 (0.01)	0.01 (0.05)
0.75					-6.03 (0.01)	0.01 (0.05)
0.80					-6.18 (0.01)	0.01 (0.05)
0.85					-6.32 (0.01)	0.01 (0.05)
0.90					-6.44 (0.00)	0.01 (0.05)
0.95					-6.56 (0.00)	0.01 (0.05)
1.00					-6.67 (0.00)	0.01 (0.05)

Table 5.13: Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the coverage probability and average length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional, modified profile, penalized (Firth) and penalised based on adjusted responses log-likelihood ratios, with all coverage probabilities multiplied by 100.

	W	W_c	W_{mp}	W_*	W_a
δ			$q = 30, m = 1$		
0.31					92.0 (1.7)
0.32					92.0 (1.7)
0.33					92.0 (1.7)
0.34					92.0 (1.6)
0.35					92.0 (1.6)
0.36					96.2 (1.6)
0.37					96.2 (1.6)
0.38					96.2 (1.6)
0.39					96.2 (1.6)
0.40	57.8 (2.4)	96.2 (1.7)	85.0 (1.8)	93.0 (1.7)	96.2 (1.6)
0.41					96.2 (1.6)
0.42					96.2 (1.6)
0.43					96.2 (1.5)
0.44					96.2 (1.5)
0.45					96.2 (1.5)
0.46					97.8 (1.5)
0.47					97.8 (1.5)
0.48					97.8 (1.5)
0.49					**95.6 (1.5)**
0.50					**95.6 (1.5)**
δ			$q = 30, m = 3$		
0.06					92.0 (1.7)
0.07					92.0 (1.7)
0.08					92.0 (1.7)
0.09					92.0 (1.7)
0.10					92.0 (1.7)
0.11					96.2 (1.6)
0.12					96.2 (1.6)
0.13					96.2 (1.6)
0.14					96.2 (1.6)
0.15	85.0 (1.9)	96.2 (1.7)	96.2 (1.7)	96.2 (1.6)	96.2 (1.6)
0.16					96.2 (1.6)
0.17					96.2 (1.6)
0.18					96.2 (1.5)
0.19					97.8 (1.5)
0.20					**95.6 (1.5)**
0.21					**95.6 (1.5)**
0.22					**95.6 (1.5)**
0.23					**95.6 (1.5)**
0.24					**95.6 (1.5)**
0.25					**95.6 (1.5)**

Table 5.14: Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the coverage probability and average length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional, modified profile, penalized (Firth) and penalised based on adjusted responses log-likelihood ratios, with all coverage probabilities multiplied by 100.

	W	W_c	W_{mp}	W_*	W_a
δ			$q = 30, m = 11$		
0.01					96.2 (1.7)
0.02					96.2 (1.7)
0.03					96.2 (1.7)
0.04					96.2 (1.6)
0.05					96.2 (1.6)
0.06					93.9 (1.6)
0.07					93.9 (1.6)
0.08					**95.6 (1.6)**
0.09					**95.6 (1.5)**
0.10	92.0 (1.8)	96.2 (1.7)	96.2 (1.7)	93.9 (1.7)	**95.6 (1.5)**
0.11					**95.6 (1.5)**
0.12					**95.6 (1.5)**
0.13					**95.6 (1.5)**
0.14					96.1 (1.5)
0.15					96.1 (1.4)
0.16					96.1 (1.4)
0.17					96.1 (1.4)
0.18					96.1 (1.4)
0.19					91.7 (1.4)
0.20					91.8 (1.4)
δ			$q = 30, m = 39$		
0.01					96.2 (1.7)
0.02					93.9 (1.7)
0.03					93.9 (1.6)
0.04					93.9 (1.6)
0.05					**95.6 (1.6)**
0.06					**95.6 (1.6)**
0.07					**95.6 (1.5)**
0.08					**95.6 (1.5)**
0.09					**95.6 (1.5)**
0.10	96.2 (1.7)	96.2 (1.7)	96.2 (1.7)	93.9 (1.6)	**95.6 (1.5)**
0.11					96.1 (1.5)
0.12					96.1 (1.4)
0.13					96.1 (1.4)
0.14					91.7 (1.4)
0.14					91.7 (1.4)
0.16					91.8 (1.4)
0.17					91.8 (1.4)
0.18					91.8 (1.4)
0.19					91.8 (1.3)
0.20					91.8 (1.3)

Table 5.15: Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the coverage probability and average length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional, modified profile, penalized (Firth) and penalised based on adjusted responses log-likelihood ratios, with all coverage probabilities multiplied by 100.

	W	W_c	W_{mp}	W_*	W_a
δ			$q = 100, m = 1$		
0.31					88.4 (0.9)
0.32					88.4 (0.9)
0.33					88.4 (0.9)
0.34					92.3 (0.9)
0.35					92.3 (0.9)
0.36					91.8 (0.9)
0.37					**94.6 (0.9)**
0.38					**94.6 (0.9)**
0.39					**94.6 (0.9)**
0.40	15.0 (1.3)	94.6 (0.9)	77.3 (1.0)	77.3 (1.0)	**94.6 (0.9)**
0.41					95.7 (0.9)
0.42					95.7 (0.8)
0.43					95.7 (0.8)
0.44					96.9 (0.8)
0.45					96.9 (0.8)
0.46					95.7 (0.8)
0.47					96.4 (0.8)
0.48					96.4 (0.8)
0.49					96.4 (0.8)
0.50					96.4 (0.8)
δ			$q = 100, m = 3$		
0.06					88.4 (0.9)
0.07					88.4 (0.9)
0.08					88.4 (0.9)
0.09					91.8 (0.9)
0.10					91.8 (0.9)
0.11					94.6 (0.9)
0.12					94.6 (0.9)
0.13					94.6 (0.9)
0.14					95.7 (0.9)
0.15	69.9 (1.0)	94.6 (0.9)	93.9 (0.9)	95.7 (0.9)	95.7 (0.9)
0.16					95.7 (0.9)
0.17					96.9 (0.8)
0.18					95.7 (0.8)
0.19					95.7 (0.8)
0.20					96.4 (0.8)
0.21					96.4 (0.8)
0.22					94.7 (0.8)
0.23					**95.1 (0.8)**
0.24					**95.1 (0.8)**
0.25					**95.1 (0.8)**

Table 5.16: Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the coverage probability and average length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional, modified profile, penalized (Firth) and penalised based on adjusted responses log-likelihood ratios, with all coverage probabilities multiplied by 100.

	W	W_c	W_{mp}	W_*	W_a
δ			$q = 100, m = 11$		
0.01					93.9 (0.9)
0.02					93.9 (0.9)
0.03					95.7 (0.9)
0.04					94.6 (0.9)
0.05					95.7 (0.9)
0.06					95.7 (0.9)
0.07					94.7 (0.8)
0.08					94.7 (0.8)
0.09					**95.1 (0.8)**
0.10	91.8 (0.9)	94.6 (0.9)	94.6 (0.9)	94.6 (0.9)	92.6 (0.8)
0.11					92.8 (0.8)
0.12					92.8 (0.8)
0.13					92.8 (0.8)
0.14					89.5 (0.8)
0.15					89.5 (0.8)
0.16					89.5 (0.8)
0.17					85.0 (0.8)
0.18					85.0 (0.8)
0.19					85.0 (0.7)
0.20					79.4 (0.7)
δ			$q = 100, m = 39$		
0.01					94.6 (0.9)
0.02					95.7 (0.9)
0.03					95.7 (0.9)
0.04					94.7 (0.9)
0.05					94.7 (0.8)
0.06					**95.1 (0.8)**
0.07					92.6 (0.8)
0.08					92.8 (0.8)
0.09					92.8 (0.8)
0.10	93.9 (0.9)	94.6 (0.9)	94.6 (0.9)	94.6 (0.9)	89.5 (0.8)
0.11					89.5 (0.8)
0.12					85.0 (0.8)
0.13					85.0 (0.8)
0.14					85.0 (0.8)
0.15					79.4 (0.8)
0.16					79.4 (0.7)
0.17					79.4 (0.7)
0.18					72.5 (0.7)
0.19					72.5 (0.7)
0.20					64.7 (0.7)

Table 5.17: Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the coverage probability and average length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional, modified profile, penalized (Firth) and penalised based on adjusted responses log-likelihood ratios, with all coverage probabilities multiplied by 100.

	W	W_c	W_{mp}	W_*	W_a
δ			$q = 1000, m = 1$		
0.31					34.4 (0.3)
0.32					39.8 (0.3)
0.33					48.2 (0.3)
0.34					56.7 (0.3)
0.35					62.2 (0.3)
0.36					70.0 (0.3)
0.37					77.0 (0.3)
0.38					81.1 (0.3)
0.39					86.3 (0.3)
0.40	0.0 (0.4)	95.0 (0.3)	6.3 (0.3)	6.3 (0.3)	89.1 (0.3)
0.41					92.3 (0.3)
0.42					94.6 (0.3)
0.43					95.5 (0.3)
0.44					96.1 (0.3)
0.45					96.0 (0.3)
0.46					95.7 (0.3)
0.47					**94.8 (0.3)**
0.48					92.5 (0.3)
0.49					90.5 (0.3)
0.50					88.0 (0.3)
δ			$q = 1000, m = 3$		
0.06					34.4 (0.3)
0.07					42.5 (0.3)
0.08					53.9 (0.3)
0.09					62.2 (0.3)
0.10					70.0 (0.3)
0.11					77.0 (0.3)
0.12					83.0 (0.3)
0.13					87.7 (0.3)
0.14					91.3 (0.3)
0.15	4.1 (0.3)	95.0 (0.3)	91.2 (0.3)	91.2 (0.3)	**94.5 (0.3)**
0.16					95.8 (0.3)
0.17					96.1 (0.3)
0.18					**95.5 (0.3)**
0.19					94.0 (0.3)
0.20					92.5 (0.3)
0.21					89.3 (0.3)
0.22					85.0 (0.3)
0.23					81.5 (0.3)
0.24					75.3 (0.3)
0.25					68.1 (0.3)

Table 5.18: Binomial matched observations with true log odds ratio $\psi = 1$. Second to sixth columns show the coverage probability and average length (in parentheses) of confidence intervals with nominal level 95% derived from profile, conditional, modified profile, penalized (Firth) and penalised based on adjusted responses log-likelihood ratios, with all coverage probabilities multiplied by 100.

	W	W_c	W_{mp}	W_*	W_a
δ			$q = 1000, m = 11$		
0.01					86.2 (0.3)
0.02					91.2 (0.3)
0.03					94.1 (0.3)
0.04					95.4 (0.3)
0.05					**94.9 (0.3)**
0.06					92.4 (0.3)
0.07					87.9 (0.3)
0.08					81.5 (0.3)
0.09					75.3 (0.3)
0.10	79.1 (0.3)	95.0 (0.3)	95.0 (0.3)	95.0 (0.3)	65.6 (0.3)
0.11					54.6 (0.3)
0.12					43.3 (0.3)
0.13					35.1 (0.2)
0.14					25.1 (0.2)
0.15					16.9 (0.2)
0.16					12.0 (0.2)
0.17					7.2 (0.2)
0.18					4.0 (0.2)
0.19					2.4 (0.2)
0.20					1.2 (0.2)
δ			$q = 1000, m = 39$		
0.01					**95.0 (0.3)**
0.02					94.1 (0.3)
0.03					91.3 (0.3)
0.04					86.5 (0.3)
0.05					79.5 (0.3)
0.06					70.6 (0.3)
0.07					57.4 (0.3)
0.08					46.1 (0.3)
0.09					35.1 (0.3)
0.10	93.6 (0.3)	95.0 (0.3)	95.0 (0.3)	95.0 (0.3)	25.1 (0.2)
0.11					16.9 (0.2)
0.12					10.6 (0.2)
0.13					6.2 (0.2)
0.14					2.9 (0.2)
0.15					1.4 (0.2)
0.16					0.6 (0.2)
0.17					0.3 (0.2)
0.18					0.1 (0.2)
0.19					0.0 (0.2)
0.20					0.0 (0.2)

Table 5.19: Binomial matched observations with true log odds ratio $\psi = 1$. The values of δ that minimize the bias of the penalised log-likelihood estimator based on adjusted responses for various values of q and m. δ was chosen from a set of 50 values ranging from 0.01 to 0.50.

m	$q = 4$	$q = 8$	$q = 12$	$q = 16$	$q = 20$	$q = 24$	$q = 28$	$q = 32$
1	0.05	0.37	0.43	0.45	0.45	0.45	0.45	0.45
3	0.01	0.12	0.17	0.18	0.18	0.18	0.18	0.18
5	0.01	0.06	0.10	0.11	0.11	0.11	0.11	0.11
7	0.01	0.04	0.08	0.08	0.08	0.08	0.08	0.08
9	0.01	0.03	0.06	0.07	0.07	0.07	0.06	0.06
11	0.01	0.02	0.05	0.06	0.06	0.06	0.05	0.05
13	0.01	0.02	0.05	0.05	0.05	0.05	0.05	0.05
15	0.01	0.01	0.05	0.05	0.05	0.05	0.04	0.04

5.3.6 Analysis of crying babies data

In this section we illustrate the methods discussed above in a general setting by providing a real-data example with different stratum sizes. We re-analyse the crying of babies data set given in Cox (1970, Example 1.2) which we include in Table 5.20. The data come from an experiment intended to assess the effectiveness of rocking motion on the crying of babies and were collected according to a matched case-control design with one case and m_i controls per stratum, where $i = 1, \ldots, 18$ and m_i takes on various values from 5 to 9. On each of 18 days babies not crying at a specified time in a hospital were served as subjects. On each day one baby chosen at random formed the experimental group and the remainder were controls. The binary response was whether the baby was crying or not at the end of a specified period. In Table 5.20, not crying is taken as a "success" and the observed numbers y_{i2} and y_{i1} are therefore the numbers of babies in the two groups not crying. The number of non crying babies in the experimental group is $t = 15$.

The estimates of the log odds ratio ψ and its standard error are reported in Table 5.21. Davison (1988) obtained the maximum likelihood and conditional maximum likelihood estimates while Lunardon (2018) obtained the estimates of ψ derived from the modified profile and penalised (Firth) log-likelihoods. We found the penalised based on adjusted responses log-likelihood and indirect inference estimates for 20 values of δ ranging from 0.05 to 1.00. As the true log odds ratio ψ is unknown, it is difficult to decide which estimator should be preferred however, there is an important observation to note. As δ approaches 1, the indirect inference estimates of ψ approach the conditional log-likelihood estimates and the standard errors of $\hat{\psi}_{a*}$ approach that of $\hat{\psi}_c$. This observation has been noted before in the complete enumeration study in terms of bias and variance in the specific setting where the ith stratum sample size was fixed. Even though this observation has not been proved analytically, the results based on the crying babies data show that, at least numerically, it may also hold for a general binary matched pairs with different stratum sizes.

The standard errors of the estimates in Table 5.21, except for $\hat{\psi}_c$, are obtained using the Fisher information matrix evaluated at the given estimate of ψ. In particular, for the maximum likelihood estimator the standard error is obtained using $\sqrt{\text{diag}\{i^{-1}(\hat{\psi}, \hat{\lambda}_1, \ldots, \hat{\lambda}_{18})\}}$, where i is the Fisher information matrix. For the conditional maximum likelihood estimator the standard error is obtained using $\sqrt{(-\partial^2 l_c(\psi)/\partial\psi^2)^{-1}}$, evaluated at $\hat{\psi}_c$. For the modified profile likelihood estimator, $\hat{\psi}_{mp}$, the standard error is obtained using

$\sqrt{(\partial^2 l_{mp}(\psi)/\partial\psi^2)^{-1}}$, evaluated at $\hat{\psi}_{mp}$, i.e. using the Hessian. For the penalised (Firth) and penalised based on adjusted responses likelihood estimators, the standard errors are obtained using $\sqrt{\text{diag}\{i^{-1}(\hat{\psi}_*, \hat{\lambda}_{1,*}, \ldots, \hat{\lambda}_{18,*})\}}$ and $\sqrt{\text{diag}\{i^{-1}(\hat{\psi}_a, \hat{\lambda}_{1,a}, \ldots, \hat{\lambda}_{18,a})\}}$, respectively. Finally, for the indirect inference estimator the standard error is obtained using $\sqrt{(-\partial^2 l_c(\psi)/\partial\psi^2)^{-1}}$, evaluated at $\tilde{\psi}_{a*}$. However according to Kuk (1995), to obtain the estimated standard error for the indirect inference estimator we need a further correction of the Fisher information using sandwich estimators of the variance based on the Godambe information because the second Bartlett identity no longer holds. Similarly, a Godambe information matrix would be better to use for the estimated standard error of the modified profile likelihood estimator. The estimated standard error reported in Lunardon (2018) for $\hat{\psi}_*$ is obtained using the Hessian matrix and is slightly different to our result.

To assess the reliability of the estimators we compute their actual bias and variance, conditioning on the observed totals as in Lunardon (2018, Table 2), for a set of values of ψ ranging from -3 to 3. The conditional distribution of the sufficient statistic $T = \sum_{i=1}^{18} Y_{i1}$, which represents the total number of babies not crying in the experimental group, is the distribution of the sum of independent Bernoulli random variables with different probabilities. To calculate this distribution we use the R function *dkbinom* which gives the mass function of the sum of k independent Binomial random variables, with possibly different probabilities. This function implements the convolution algorithm of k binomials described in Butler and Stephens (2017).

The results reported in Table 2 of Lunardon (2018) are incorrect because Lunardon (2018) computes the conditional bias and variance of $\hat{\psi}_c$, $\hat{\psi}$ and $\hat{\psi}_{mp}$ for $t \in \{1, \ldots, 17\}$ but without rescaling and normalizing the conditional distribution of T. The corrected conditional summaries are reported in Table 5.22 with the addition of the penalised log-likelihood estimator based on adjusted responses, $\hat{\psi}_a$, and the indirect inference estimator, $\hat{\psi}_{a*}$, for various values of δ. We observe that for $\psi \in \{-2, -1, 0, 1, 2\}$ there exists a value of δ in the range $0.01 < \delta < 0.20$ such that $\hat{\psi}_a$ is less biased than any of $\hat{\psi}$, $\hat{\psi}_c$, $\hat{\psi}_{mp}$ and $\hat{\psi}_*$. In fact the effect of increasing the absolute value of ψ is to decrease the optimal δ value in terms of the bias of $\hat{\psi}_a$. For $\psi = 3$, the maximum likelihood estimator seems to be the least biased but this is due to the effect of infinite estimates at $t = 0$ and $t = 18$. In other words, removing these infinite estimates significantly lowers the average of the estimates for $t \in \{1, \ldots, 17\}$. This is also true for the conditional log-likelihood estimator, $\hat{\psi}_c$.

Overall, for $\psi \in \{-2,-1,0,1\}$ the indirect inference estimator is an improvement over the penalised log-likelihood estimator based on adjusted responses in terms of bias for all values of δ considered. We notice that there exists a δ value in the range $0.01 < \delta < 0.20$ such that $\hat{\psi}_{a*}$ is less biased than $\hat{\psi}_a$. In fact the indirect inference estimator for $\psi \in \{-2,-1,0,1\}$ is competitive for all δ values considered, which makes it the best choice amongst estimators. However, for $\psi \in \{-3,2,3\}$ the best choice of δ is the largest possible, in this case $\delta = 1$. It is worth noting that as in the complete enumeration study in Tables 5.7-5.12, it is also the case here that we observe that the bias and variance of the indirect inference estimator approaches that of the conditional log-likelihood estimator as δ increase to 1.

Table 5.23 reports the unconditional bias and variance of the modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihood estimators for $t \in \{0,\ldots,18\}$. Even though $\hat{\psi}_*$ has smaller bias and variance than $\hat{\psi}_{mp}$ for all value of ψ, we again observe as in Table 5.22 that there exists a value of δ such that $\hat{\psi}_a$ is less biased than $\hat{\psi}_*$ for all values of ψ. We may conclude that $\hat{\psi}_a$ is preferable over $\hat{\psi}_*$ as its variance is also decreased at the optimal δ value for all values of ψ considered (except for $\psi = 0$). The relationship between the optimal δ value and the true value of ψ coincides with that in the conditional case (Table 5.22), i.e. the optimal δ, in terms of bias, decreases as the absolute value of ψ increases. Note that the indirect inference estimator has no solution at $t = 0$ and $t = 18$, and so it is excluded from the unconditional summaries in Table 5.23.

Table 5.20: The crying of babies data.

Day i	No. of control babies m_i	No. not crying y_{i2}	No. of experimental babies	No. not crying y_{i1}
1	8	3	1	1
2	6	2	1	1
3	5	1	1	1
4	6	1	1	0
5	5	4	1	1
6	9	4	1	1
7	8	5	1	1
8	8	4	1	1
9	5	3	1	1
10	9	8	1	0
11	6	5	1	1
12	9	8	1	1
13	8	5	1	1
14	5	4	1	1
15	6	4	1	1
16	8	7	1	1
17	6	4	1	0
18	8	5	1	1

Table 5.21: Crying babies real data example: estimates of ψ and its standard error (in parentheses) derived from profile, conditional, modified profile, penalised (Firth), penalised based on adjusted responses log-likelihoods and indirect inference, respectively.

δ	$\hat{\psi}$	$\hat{\psi}_{\mathbf{c}}$	$\hat{\psi}_{\mathbf{mp}}$	$\hat{\psi}_{*}$	$\hat{\psi}_{\mathbf{a}}$	$\tilde{\psi}_{\mathbf{a}*}$
0.05					1.174 (0.656)	1.157 (0.669)
0.10					0.986 (0.602)	1.184 (0.674)
0.15					0.841 (0.561)	1.201 (0.676)
0.20					0.726 (0.529)	1.212 (0.678)
0.25					0.633 (0.502)	1.220 (0.680)
0.30					0.555 (0.480)	1.226 (0.681)
0.35					0.490 (0.461)	1.230 (0.682)
0.40					0.435 (0.445)	1.234 (0.682)
0.45					0.387 (0.430)	1.237 (0.683)
0.50	1.432 (0.734)	1.256 (0.686)	1.270 (0.689)	1.156 (0.666)	0.345 (0.417)	1.239 (0.683)
0.55					0.309 (0.405)	1.241 (0.684)
0.60					0.277 (0.395)	1.243 (0.684)
0.65					0.249 (0.385)	1.245 (0.684)
0.70					0.224 (0.376)	1.246 (0.684)
0.75					0.201 (0.368)	1.247 (0.684)
0.80					0.181 (0.361)	1.248 (0.685)
0.85					0.162 (0.354)	1.248 (0.685)
0.90					0.146 (0.347)	1.249 (0.685)
0.95					0.131 (0.341)	1.250 (0.685)
1.00					0.117 (0.335)	1.251 (0.685)

Table 5.22: Crying babies real data example: conditional bias and variance (in parentheses) of estimators as the true log odds ratio ψ varies; estimators derived from profile, conditional, modified profile, penalised (Firth), penalised based on adjusted responses log-likelihoods and indirect inference are denoted by $\hat{\psi}$, $\hat{\psi}_c$, $\hat{\psi}_{mp}$, $\hat{\psi}_*$, $\hat{\psi}_a$, $\tilde{\psi}_{a*}$, respectively, and all entries are multiplied by 10. For $\hat{\psi}_a$ and $\tilde{\psi}_{a*}$, the smallest bias value in each column is given in bold face.

			ψ						
			-3	-2	-1	0	1	2	3
$\hat{\psi}$			-8.97 (9.56)	-5.59 (7.60)	-2.02 (4.48)	0.37 (4.40)	2.47 (6.25)	2.30 (5.71)	-2.53 (2.98)
$\hat{\psi}_c$			-0.05 (4.18)	-0.69 (4.01)	-0.12 (3.00)	0.36 (3.30)	0.88 (4.65)	-0.71 (4.06)	-6.36 (2.06)
$\hat{\psi}_{mp}$			-0.79 (3.89)	-1.39 (4.22)	-0.42 (3.23)	0.34 (3.42)	1.00 (4.75)	-0.51 (4.13)	-6.12 (2.10)
$\hat{\psi}_*$			0.56 (3.35)	-0.62 (3.77)	-0.17 (2.94)	0.07 (3.02)	-0.12 (3.75)	-2.71 (2.95)	-9.04 (1.42)
		0.01	-7.51 (7.96)	-4.99 (6.83)	-1.88 (4.27)	0.22 (4.17)	1.86 (5.56)	1.02 (4.77)	**-4.27 (2.41)**
		0.02	-6.23 (6.77)	-4.44 (6.21)	-1.75 (4.07)	0.08 (3.95)	1.31 (4.99)	**-0.10 (4.06)**	-5.77 (2.00)
		0.03	-5.09 (5.86)	-3.94 (5.68)	-1.62 (3.88)	**-0.06 (3.75)**	0.80 (4.53)	-1.08 (3.52)	-7.06 (1.69)
		0.04	-4.07 (5.14)	-3.46 (5.23)	-1.49 (3.71)	-0.18 (3.57)	0.33 (4.15)	-1.95 (3.10)	-8.20 (1.46)
		0.05	-3.13 (4.56)	-3.01 (4.84)	-1.37 (3.55)	-0.30 (3.41)	**-0.10 (3.82)**	-2.74 (2.75)	-9.22 (1.27)
		0.06	-2.28 (4.08)	-2.59 (4.51)	-1.26 (3.41)	-0.41 (3.26)	-0.50 (3.54)	-3.47 (2.47)	-10.14 (1.12)
$\hat{\psi}_a$	δ	0.07	-1.49 (3.68)	-2.20 (4.21)	-1.15 (3.27)	-0.51 (3.12)	-0.87 (3.29)	-4.13 (2.24)	-10.97 (1.00)
		0.08	-0.76 (3.35)	-1.82 (3.94)	-1.04 (3.14)	-0.61 (2.99)	-1.22 (3.07)	-4.73 (2.04)	-11.73 (0.90)
		0.09	**-0.08 (3.06)**	-1.46 (3.71)	-0.94 (3.02)	-0.71 (2.86)	-1.55 (2.88)	-5.30 (1.87)	-12.43 (0.82)
		0.10	0.55 (2.81)	-1.12 (3.50)	-0.84 (2.90)	-0.80 (2.75)	-1.86 (2.71)	-5.82 (1.72)	-13.07 (0.75)
		0.15	3.22 (1.96)	**0.37 (2.69)**	-0.37 (2.43)	-1.18 (2.28)	-3.18 (2.07)	-7.99 (1.21)	-15.70 (0.50)
		0.20	5.28 (1.46)	1.59 (2.16)	**0.03 (2.06)**	-1.48 (1.92)	-4.21 (1.65)	-9.63 (0.92)	-17.65 (0.37)
		1.00	17.07 (0.19)	9.53 (0.37)	3.28 (0.45)	-2.73 (0.41)	-9.52 (0.29)	-17.70 (0.13)	-26.96 (0.05)
		0.01	2.72 (2.60)	0.51 (3.00)	0.13 (2.64)	**-0.01 (2.83)**	-0.62 (3.30)	-3.75 (2.46)	-10.41 (1.17)
		0.02	2.26 (2.93)	0.37 (3.17)	0.11 (2.67)	0.02 (2.88)	-0.40 (3.57)	-3.16 (2.83)	-9.55 (1.39)
		0.03	1.97 (3.14)	0.28 (3.28)	0.10 (2.69)	0.04 (2.92)	-0.25 (3.74)	-2.80 (3.06)	-9.04 (1.52)
		0.04	1.75 (3.29)	0.20 (3.36)	0.09 (2.71)	0.07 (2.95)	-0.14 (3.87)	-2.54 (3.22)	-8.68 (1.61)
		0.05	1.57 (3.41)	0.14 (3.43)	0.08 (2.72)	0.08 (2.97)	-0.05 (3.96)	-2.34 (3.34)	-8.40 (1.68)
		0.06	1.43 (3.50)	0.08 (3.48)	0.07 (2.74)	0.10 (2.99)	**0.02 (4.03)**	-2.19 (3.43)	-8.19 (1.73)
$\tilde{\psi}_{a*}$	δ	0.07	1.31 (3.57)	0.03 (3.53)	0.06 (2.75)	0.11 (3.01)	0.08 (4.09)	-2.05 (3.51)	-8.02 (1.78)
		0.08	1.21 (3.63)	**-0.01 (3.57)**	0.06 (2.76)	0.13 (3.03)	0.14 (4.15)	-1.94 (3.57)	-7.87 (1.81)
		0.09	1.12 (3.69)	-0.04 (3.60)	0.05 (2.77)	0.14 (3.04)	0.18 (4.19)	-1.85 (3.62)	-7.75 (1.84)
		0.10	1.05 (3.73)	-0.08 (3.63)	0.04 (2.78)	0.15 (3.06)	0.23 (4.23)	-1.77 (3.66)	-7.64 (1.86)
		0.15	0.77 (3.88)	-0.21 (3.74)	0.02 (2.83)	0.19 (3.11)	0.38 (4.35)	-1.48 (3.80)	-7.27 (1.93)
		0.20	0.59 (3.96)	-0.29 (3.80)	**0.00 (2.86)**	0.23 (3.14)	0.48 (4.43)	-1.30 (3.87)	-7.06 (1.97)
		1.00	**0.05 (4.16)**	-0.62 (3.99)	-0.09 (2.98)	0.35 (3.28)	0.83 (4.63)	**-0.78 (4.04)**	**-6.44 (2.06)**

Table 5.23: Crying babies real data example: unconditional bias and variance (in parentheses) of estimators as the true log odds ratio ψ varies; estimators derived from modified profile, penalised (Firth) and penalised based on adjusted responses log-likelihoods are denoted by $\hat{\psi}_{mp}$, $\hat{\psi}_{*}$, $\hat{\psi}_{a}$, respectively, and all entries are multiplied by 10. For $\hat{\psi}_{a}$, the smallest bias value in each column is given in bold face.

			ψ						
			-3	-2	-1	0	1	2	3
$\hat{\psi}_{mp}$			-2.65 (6.86)	-1.50 (4.52)	-0.42 (3.23)	0.35 (3.44)	1.46 (6.60)	4.26 (17.04)	7.10 (21.60)
$\hat{\psi}_{*}$			-1.13 (5.82)	-0.72 (4.03)	-0.17 (2.95)	0.07 (3.03)	0.18 (4.55)	0.18 (7.49)	-1.48 (7.45)
$\hat{\psi}_{a}$	δ	0.030	-7.87 (12.84)	-4.09 (6.29)	-1.62 (3.89)	**-0.05 (3.77)**	1.13 (5.47)	2.07 (8.88)	1.13 (8.79)
		0.040	-6.46 (10.22)	-3.60 (5.71)	-1.49 (3.71)	-0.18 (3.58)	0.63 (4.87)	0.75 (6.98)	**-1.30 (6.37)**
		0.045	-5.83 (9.26)	-3.36 (5.46)	-1.43 (3.63)	-0.24 (3.50)	0.40 (4.63)	**0.17 (6.30)**	-2.31 (5.56)
		0.050	-5.24 (8.46)	-3.14 (5.23)	-1.37 (3.56)	-0.30 (3.42)	**0.17 (4.41)**	-0.36 (5.74)	-3.22 (4.91)
		0.060	-4.18 (7.19)	-2.71 (4.83)	-1.26 (3.41)	-0.41 (3.26)	-0.25 (4.03)	-1.32 (4.86)	-4.80 (3.95)
		0.100	**-0.82 (4.39)**	-1.21 (3.68)	-0.84 (2.90)	-0.79 (2.75)	-1.67 (3.00)	-4.24 (2.96)	-9.28 (2.08)
		0.150	2.17 (2.84)	**0.30 (2.81)**	-0.37 (2.43)	-1.18 (2.28)	-3.03 (2.25)	-6.77 (1.93)	-12.85 (1.23)
		0.200	4.42 (2.04)	1.53 (2.24)	**0.03 (2.06)**	-1.48 (1.93)	-4.08 (1.78)	-8.62 (1.39)	-15.33 (0.83)

5.4 Discussion and further work

For the matched gamma pairs model, we compared several estimators of the parameter of interest and derived the asymptotic bias corrected and adjusted profile log-likelihood estimators and showed that the latter coincides with the indirect inference estimator exactly and that it is unbiased. It was shown that the asymptotic bias corrected estimator was a substantial improvement over the modified profile log-likelihood estimator. We also established that the adjusted profile log-likelihood estimator is consistent when the stratum sample size is fixed while the dimension of the nuisance parameter is allowed to increase to infinity. This solved the incidental parameter problem of Neyman and Scott (1948) for this model.

The performance of the estimators considered can be extended by considering other values for the true parameter of interest than one. Possible future work for this model is to derive the empirical bias-reducing adjusted estimators iRBM and eRBM of Kosmidis and Lunardon (2020) and compare their performance with the other estimators.

For the binomial matched pairs model, we evaluated the performance of a new penalised log-likelihood estimator of the log odds ratio which is based on an additive adjustment, $\delta > 0$, to the responses so as to avoid infinite estimates which is inherited by the maximum likelihood and conditional likelihood estimators. We calculated the probability

limit of this estimator and showed that the maximum likelihood, conditional and modified profile log-likelihood estimators, when $m = 1$ for the latter, can be retrieved from this new estimator for certain values of δ. It was found that indirect inference estimation based on the new estimator is competitive for a wide range of values of δ.

It is worth investigating numerically whether, for a general value of m, there exists a δ that recovers the modified profile log-likelihood estimator of ψ from the penalised log-likelihood estimator based on adjusted responses, because this will imply that $\hat{\psi}_{mp}$ could be retrieved from $\hat{\psi}_a$ for any value of m not just in the special case of the binary matched pairs model where $m = 1$. One future direction is to investigate the performance of the estimators of the log odds ratio outside the setting of Lunardon (2018) for a general m_i and s_i. This is possible since in this general setting the distribution of the sufficient statistic $T = \sum_{i=1}^{q} y_{i1}$ can be obtained using the convolution method of Butler and Stephens (2017) and is Poisson binomial. Obtaining the probability limit of the indirect inference estimator in the setting of Lunardon (2018) is desired. The complete enumeration study in Section 5.3.5 may be expanded by considering other values of ψ, e.g. $\psi \in \{-3,-2,-1,0,1,2,3\}$. Just like in the matched gamma pairs, the iRBM and eRBM estimators may be derived for this model and compared and contrasted with the others. Finally, yet another possible future direction would be to investigate the performance of an alternative adjustment to the log-likelihood function where a small number $\delta > 0$ is added to each success but subtracted from each failure.

Chapter 6

Discussion and further work

6.1 Summary of the thesis

The current thesis examines the applicability, evaluates the performance and compares various bias reduction methods proposed in the literature, such as the bias-reducing adjusted score equations of Firth (1993), indirect inference of Kuk (1995) and reduced-bias M estimation of Kosmidis and Lunardon (2020), in terms of their impact on estimation and inference, in some parametric non-standard models used in econometrics and statistics, such as those used for sample selected, censored or stratified observations.

In particular, we studied the Heckit regression model in Chapter 3, which is also referred to as the Tobit II model, which handles non-randomly selected samples where the observed range of the dependent variable is censored. We reviewed the methods of maximum likelihood and Heckman two-step estimation and derived the empirical bias-reducing adjustments of Kosmidis and Lunardon (2020). We compared the performance of these estimation methods alongside the indirect inference method of Kuk (1995) through a simulation study. The parameter setting we chose for our simulation study implied that the Heckman two-step estimator suffered from sever multicollinearity and that the indirect inference, iRBM and eRBM estimators were useful in reducing the bias of the ML estimator and improving the coverage probabilities of Wald-type confidence intervals, when the sample size is small. While the adjusted score equations approach of Firth (1993) was not possible to implement for the Heckit model because it required numerical approximation of integrals, the indirect inference and reduced-bias M estimation methods were easier to implement though they are computationally expensive and may not converge for all samples. Chapter 4 focuses on the Weibull accelerated failure time

model, where the adjusted score equations approach of Firth (1993) is not applicable, and examines the performance of indirect inference and empirical bias-reducing adjustments, which were straightforward to implement and converged for all samples, and compares them with the standard ML estimation method. It was found that, when the censoring level is moderate, the indirect inference and the empirical bias-reducing penalised log-likelihood estimator (iRBM) reduce the small sample bias of the ML estimator and improve the coverage probability of Wald-type confidence intervals. The results of these two chapters opens the door to further research on general Tobit and accelerated failure time models where the methods of indirect inference of Kuk (1995) and reduced-bias M estimation of Kosmidis and Lunardon (2020) are applicable and have the potential of significant bias reduction. The Tobit models are broadly used in many fields, such as econometrics and political and social sciences, for modelling censored and sample selected data. Accelerated failure time models are frequently encountered when modelling failure time data such as in biomedicine and even in engineering and reliability studies.

Finally, Chapter 5 considers two stratified models, the matched gamma pairs model and the binomial matched pairs model and investigates the performance of current methods in the literature for handling the estimation of a scalar parameter of interest in the presence of a set of nuisance (incidental) parameters whose dimension becomes large relative to the stratum sample size. The challenge with such models is that the maximum likelihood estimator of the parameter of interest, is not in general consistent. This is well known as the incidental parameter problem since Neyman and Scott (1948). The methods considered in the literature for handling nuisance parameters ranges from those that modify the profile log-likelihood function such as the approximate conditional and modified profile log-likelihoods of Cox and Reid (1987) and Barndorff-Nielson (1983), respectively, to those that modify the initial estimating equation like the approach of McCullagh and Tibshirani (1990) and Firth (1993), and others that, in contrast, solve the biased estimating equation without modification and then adjust the bias of the resulting estimator as in the indirect inference approach of Kuk (1995). The empirical bias-reducing adjustment of Kosmidis and Lunardon (2020) on the hand, can be thought of as either a method that adjusts the initial estimating function or as a method that adjusts the log-likelihood function, when estimation is through maximum likelihood. For the matched gamma pairs model we reviewed the profile, approximate conditional profile and modified profile log-likelihood methods of estimation of the parameter of interest which all yield biased and inconsistent estimates. We derived the adjusted profile log-likelihood of McCullagh and

Tibshirani (1990) and showed that it yields an unbiased and consistent estimator of the parameter of interest. The indirect inference estimator was also derived which was identical to the adjusted profile log-likelihood estimator of McCullagh and Tibshirani (1990). This solved the incidental parameter problem of Neyman and Scott (1948) for the matched gamma pairs model. The case of the binomial matched pairs model was more challenging because the ML, exact conditional and modified profile log-likelihood estimators of the parameter of interest, the log odds ratio, may be infinite. We proposed a penalised version of the log-likelihood function based on adjusted responses which always results in a finite estimator of the log odds ratio. The probability limit of this new estimator is derived and it was shown that in certain settings the ML, exact conditional and modified profile log-likelihood estimators drop out as special cases of the former. The indirect inference estimation method was then used to reduce the bias of the penalised maximum likelihood estimator based on adjusted responses. The performance of all these methods was compared through a complete enumeration study and it was found that the indirect inference estimator was very competitive in terms of bias reduction in the estimation of the log odds ratio when the stratum sample size is fixed while allowing the dimension of the nuisance parameter to increase. The binomial matched pairs model, and in particular the special case of the binary matched pairs model, is very popular in biostatistics and matched case-control studies, and even though the methods of Firth (1993) and Kosmidis and Lunardon (2020) are applicable to this model, the former may be harder to apply in other stratified models. This gives scope to extend the framework of reduced-bias M estimation of Kosmidis and Lunardon (2020) to stratified settings where bias reduction is beneficial.

6.2 Further work

In this section we list briefly some topics for future research in the area of Tobit, accelerated failure time, and stratified models, most of which have already been mentioned in more detail in the discussion and further work sections of each chapter.

1. Expand the simulation study in Chapter 3 for the Heckit model to include a case where the Heckman two-step estimator does not suffer from multicollinearity problems and examine the performance of indirect inference of Kuk (1995) and empirical reduced-bias M estimation of Kosmidis and Lunardon (2020) in comparison to the ML and Heckman two-step estimation methods. The simulation study may also

be expanded by considering the sensitivity of the results to the degree of censoring in the data.

2. Implement the empirical bias-reducing adjusted estimating functions for the Heckman two-step estimator of the Heckit (Tobit II) model. Since the Heckman two-step estimator can suffer from serious collinearity problems as noticed in our simulation study in Chapter 3, and hence give rise to biased estimates with huge standard errors, it is of interest to propose methods that have the potential of effectively reducing estimation bias when collinearity problems arise. As shown in Section 3.7.2, we derived the empirical bias-reducing penalty for adjusting the Heckman two-step estimator and it only remains to optimise the resulting empirical bias-reducing penalised log-likelihood of Kosmidis and Lunardon (2020) which can be easily implemented in R.
3. Study the performance of bias reduction methods such as the indirect inference of Kuk (1995) and empirical reduced-bias M estimation of Kosmidis and Lunardon (2020) for general Tobit models, such as the Roy (Tobit V) model.
4. Extend the methods of indirect inference and empirical bias-reducing penalty to the frailty Weibull accelerated failure time model.
5. Implement indirect inference and empirical bias-reducing adjusted estimating functions in general accelerated failure time models with other failure time distributions, such as the log-normal or log-logistic distributions. The empirical bias-reducing adjustments have been derived in Section 4.6.3 and can be easily implemented numerically in R.
6. Develop statistical software for the implementation of the empirical bias-reducing penalty for the Heckit and Weibull accelerated failure time models.
7. Explore further the properties of the penalised log-likelihood estimator of the log odds ratio based on adjusted responses and the indirect inference estimator in a more general setting outside that of Lunardon (2018).
8. Consider alternative adjustments to the log-likelihood function of the binomial matched pairs model. One example is that were a small number is added to each success but subtracted from each failure.
9. Examine the performance of bias reduction methods for handling nuisance parameters in stratified accelerated failure time models.
10. Investigate the performance of the reduced-bias estimator from empirically adjusted estimating functions for general M estimation in other stratified settings.

Appendix A

Algebraic derivations for the Heckit model

In this Appendix, we show how some of the expectations in Section 3.4.1 are obtained.

The expectation of the two random variables $y_i^S\left[y_i^O-\gamma^\intercal z_i\right]^5$ and $y_i^S\left[y_i^O-\gamma^\intercal z_i\right]^6$ in (3.66) and (3.67) involves bivariate integrals. However, using the Law of iterated expectation (see Johnston and DiNardo, 1997, Appendix B.5), these two expectations can be reduced to univariate integrals. For example,

$$\begin{aligned}
\mathrm{E}_{Y_i^S,Y_i^O}\left[Y_i^S\left(Y_i^O-\gamma^\intercal z_i\right)^5\right] &= \mathrm{E}_{Y_i^S}\left\{\mathrm{E}_{Y_i^O|Y_i^S}\left[Y_i^S\left(Y_i^O-\gamma^\intercal z_i\right)^5\right]\right\} \\
&= \Pr(Y_i^S=0)\mathrm{E}_{Y_i^O|Y_i^S=0}\left[Y_i^S\left(Y_i^O-\gamma^\intercal z_i\right)^5\right] \\
&+ \Pr(Y_i^S=1)\mathrm{E}_{Y_i^O|Y_i^S=1}\left[Y_i^S\left(Y_i^O-\gamma^\intercal z_i\right)^5\right] \\
&= \Phi(a_i)\mathrm{E}_{Y_i^O|Y_i^S=1}\left[\left(Y_i^O-\gamma^\intercal z_i\right)^5\right] \\
&= \Phi(a_i)\int_{-\infty}^{\infty}\left(y_i^O-\gamma^\intercal z_i\right)^5 f(y_i^O|y_i^S=1)\,dy_i^O, \qquad \text{(A.1)}
\end{aligned}$$

where the third equality above follows since $\mathrm{E}_{Y_i^O|Y_i^S=0}\left[Y_i^S\left(Y_i^O-\gamma^\intercal z_i\right)^5\right]=0$. The integral in (A.1) can be obtained by exploiting the moments of the random variable $Y^O|Y^S=1$. The moment generating function of the conditional random variable $Y^O|Y^S=1$ is derived

in Bierens (2007, Appendices 1 and 2, p.10-14) and is given by

$$
\begin{aligned}
M_{Y^O|Y^S=1}(t) &= \int_{-\infty}^{\infty} \exp(ty^O) \frac{\phi\left[(y^O-\gamma^\intercal Z)/\sigma\right]}{\sigma\Phi(\beta^\intercal X)} \Phi\left(\frac{\beta^\intercal X+\rho(y^O-\gamma^\intercal Z)/\sigma}{\sqrt{1-\rho^2}}\right) dy^O \\
&= \frac{\exp(t\gamma^\intercal Z)}{\Phi(\beta^\intercal X)} \int_{-\infty}^{\infty} \exp(t\sigma u)\phi(u)\Phi\left(\frac{\beta^\intercal X+\rho u}{\sqrt{1-\rho^2}}\right) du \\
&= \frac{\exp[t\gamma^\intercal Z+(t^2\sigma^2)/2]}{\Phi(\beta^\intercal X)} \int_{-\infty}^{\infty} \phi(u-t\sigma)\Phi\left(\frac{\beta^\intercal X+\rho u}{\sqrt{1-\rho^2}}\right) du \\
&= \frac{\exp[t\gamma^\intercal Z+(t^2\sigma^2)/2]}{\Phi(\beta^\intercal X)} \int_{-\infty}^{\infty} \phi(u^*)\Phi\left(\frac{\beta^\intercal X+\rho(u^*+t\sigma)}{\sqrt{1-\rho^2}}\right) du^*, \qquad \text{(A.2)}
\end{aligned}
$$

where in the second equality we make the substitution $u = (y^O - \gamma^\intercal Z)/\sigma$, in the third equality we make use of the relation $\exp(t\sigma u)\phi(u) = \exp(t^2\sigma^2/2)\phi(u-t\sigma)$ and in the last equality we make the substitution $u^* = u - t\sigma$. Using the product rule of differentiation and the identity (see Bierens, 2007, Appendix 1 for a derivation)

$$
\int_{-\infty}^{\infty} \phi(x)\phi(a+bx)\,dx = \frac{\phi\left(a/\sqrt{1+b^2}\right)}{\sqrt{1+b^2}}, \qquad \text{(A.3)}
$$

the first partial derivative of the moment generating function in (A.2) may be written as

$$
\frac{\partial}{\partial t} M_{Y^O|Y^S=1}(t) = (\gamma^\intercal Z + t\sigma^2) M_{Y^O|Y^S=1}(t) + \rho\sigma \exp[t\gamma^\intercal Z + (t^2\sigma^2)/2] \frac{\phi(\beta^\intercal X + \rho t\sigma)}{\Phi(\beta^\intercal X)}, \qquad \text{(A.4)}
$$

and hence the expectation (first moment) of $Y^O|Y^S = 1$ is

$$
\begin{aligned}
\mathrm{E}\left[Y^O|Y^S=1\right] &= \frac{\partial}{\partial t} M_{Y^O|Y^S=1}(t)\bigg|_{t=0} \\
&= \gamma^\intercal Z + \rho\sigma \frac{\phi(\beta^\intercal X)}{\Phi(\beta^\intercal X)} \\
&= \gamma^\intercal Z + \rho\sigma m_2(\beta^\intercal X). \qquad \text{(A.5)}
\end{aligned}
$$

Higher order moments of $Y^O|Y^S = 1$ may be obtained by successive differentiation of (A.4) and after some algebra the second, third, fourth, fifth and sixth moments have the form

$$\begin{aligned}
\mathrm{E}\big[(Y^O)^2|Y^S=1\big] &= \sigma^2+(\gamma^\intercal Z)^2+2\rho\sigma(\gamma^\intercal Z)m_2(\beta^\intercal X)-\rho^2\sigma^2(\beta^\intercal X)m_2(\beta^\intercal X)\\
&= \sigma^2+\big[\gamma^\intercal Z+\rho\sigma m_2(\beta^\intercal X)\big]^2-\rho^2\sigma^2(\beta^\intercal X)m_2(\beta^\intercal X)\\
&\quad -\rho^2\sigma^2\big[m_2(\beta^\intercal X)\big]^2, \qquad \text{(A.6)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\big[(Y^O)^3|Y^S=1\big] &= 3\sigma^2(\gamma^\intercal Z)+(\gamma^\intercal Z)^3+\rho\sigma^3(3-\rho^2)m_2(\beta^\intercal X)+3\rho\sigma(\gamma^\intercal Z)^2m_2(\beta^\intercal X)\\
&\quad -3\rho^2\sigma^2(\beta^\intercal X)(\gamma^\intercal Z)m_2(\beta^\intercal X)+\rho^3\sigma^3(\beta^\intercal X)^2m_2(\beta^\intercal X), \qquad \text{(A.7)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\big[(Y^O)^4|Y^S=1\big] &= 3\sigma^4+6\sigma^2(\gamma^\intercal Z)^2+4\rho\sigma(\gamma^\intercal Z)^3m_2(\beta^\intercal X)-\rho^4\sigma^4(\beta^\intercal X)^3m_2(\beta^\intercal X)\\
&\quad +4\rho\sigma^3(3-\rho^2)(\gamma^\intercal Z)m_2(\beta^\intercal X)-3\rho^2\sigma^4(2-\rho^2)(\beta^\intercal X)m_2(\beta^\intercal X)\\
&\quad +(\gamma^\intercal Z)^4-6\rho^2\sigma^2(\beta^\intercal X)(\gamma^\intercal Z)^2m_2(\beta^\intercal X)\\
&\quad +4\rho^3\sigma^3(\gamma^\intercal Z)(\beta^\intercal X)^2m_2(\beta^\intercal X), \qquad \text{(A.8)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\big[(Y^O)^5|Y^S=1\big] &= 10\sigma^2(\gamma^\intercal Z)^3+(\gamma^\intercal Z)^5+\rho\sigma^5\big[15-\rho^2(10-3\rho^2)\big]m_2(\beta^\intercal X)\\
&\quad -15\rho^2\sigma^4(2-\rho^2)(\beta^\intercal X)(\gamma^\intercal Z)m_2(\beta^\intercal X)+15\sigma^4(\gamma^\intercal Z)\\
&\quad -5\rho^4\sigma^4(\gamma^\intercal Z)(\beta^\intercal X)^3m_2(\beta^\intercal X)-10\rho^2\sigma^2(\beta^\intercal X)(\gamma^\intercal Z)^3m_2(\beta^\intercal X)\\
&\quad +10\rho^3\sigma^3(\beta^\intercal X)^2(\gamma^\intercal Z)^2m_2(\beta^\intercal X)+\rho^5\sigma^5(\beta^\intercal X)^4m_2(\beta^\intercal X)\\
&\quad +2\rho^3\sigma^5(5-3\rho^2)(\beta^\intercal X)^2m_2(\beta^\intercal X)+5\rho\sigma(\gamma^\intercal Z)^4m_2(\beta^\intercal X),\\
&\quad +10\rho\sigma^3(3-\rho^2)(\gamma^\intercal Z)^2m_2(\beta^\intercal X) \qquad \text{(A.9)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\big[(Y^O)^6|Y^S=1\big] &= 12\rho^3\sigma^5(5-3\rho^2)(\gamma^\intercal Z)(\beta^\intercal X)^2m_2(\beta^\intercal X)+6\rho\sigma(\gamma^\intercal Z)^5m_2(\beta^\intercal X)\\
&\quad +20\rho\sigma^3(3-\rho^2)(\gamma^\intercal Z)^3m_2(\beta^\intercal X)-5\rho^4\sigma^6(3-2\rho^2)(\beta^\intercal X)^3m_2(\beta^\intercal X)\\
&\quad -15\rho^4\sigma^4(\gamma^\intercal Z)^2(\beta^\intercal X)^3m_2(\beta^\intercal X)-15\rho^2\sigma^2(\beta^\intercal X)(\gamma^\intercal Z)^4m_2(\beta^\intercal X)\\
&\quad -15\rho^2\sigma^6\big[3-\rho^2(3-\rho^2)\big](\beta^\intercal X)m_2(\beta^\intercal X)-\rho^6\sigma^6(\beta^\intercal X)^5m_2(\beta^\intercal X)\\
&\quad +20\rho^3\sigma^3(\beta^\intercal X)^2(\gamma^\intercal Z)^3m_2(\beta^\intercal X)+6\rho^5\sigma^5(\beta^\intercal X)^4(\gamma^\intercal Z)m_2(\beta^\intercal X)\\
&\quad +6\rho\sigma^5\big[15-\rho^2(10-3\rho^2)\big](\gamma^\intercal Z)m_2(\beta^\intercal X)+(\gamma^\intercal Z)^6\\
&\quad -45\rho^2\sigma^4(2-\rho^2)(\beta^\intercal X)(\gamma^\intercal Z)^2m_2(\beta^\intercal X)\\
&\quad +15\sigma^6+45\sigma^4(\gamma^\intercal Z)^2+15\sigma^2(\gamma^\intercal Z)^4. \qquad \text{(A.10)}
\end{aligned}$$

Using (A.5)-(A.9) it may be shown after some algebra that

$$\int_{-\infty}^{\infty}\left(y_i^O-\gamma^\intercal z_i\right)^5 f(y_i^O|y_i^S=1)\,dy_i^O=\rho\sigma^5\Big\{\left[15-\rho^2(10-3\rho^2)\right] +\rho^2a_i^2\left[2(5-3\rho^2)+\rho^2a_i^2\right]\Big\}m_2(a_i), \tag{A.11}$$

where $a_i=\beta^\intercal x_i$ and similarly using (A.5)-(A.10) it may be shown that

$$\int_{-\infty}^{\infty}\left(y_i^O-\gamma^\intercal z_i\right)^6 f(y_i^O|y_i^S=1)\,dy_i^O=\sigma^6\Big\{15-\rho^2a_i\Big[5\rho^2(3-2\rho^2)a_i^2 +15\left[3-\rho^2(3-\rho^2)\right]+\rho^4a_i^4\Big]m_2(a_i)\Big\}. \tag{A.12}$$

Multiplying (A.11) and (A.12) by $\Phi(a_i)$ we obtain respectively the expectation of $y_i^S\left[y_i^O-\gamma^\intercal z_i\right]^5$ and $y_i^S\left[y_i^O-\gamma^\intercal z_i\right]^6$.

The expectation of $y_i^S m_2(b_i)\left[y_i^O-\gamma^\intercal z_i\right]^4$ can be obtained by exploiting the definition of b_i in (3.14) to rewrite the expectation of $y_i^S b_i^2 m_2(b_i)\left[y_i^O-\gamma^\intercal z_i\right]^2$ as follows

$$\begin{aligned}\mathrm{E}\left\{y_i^S b_i^2 m_2(b_i)\left[y_i^O-\gamma^\intercal z_i\right]^2\right\}&=\frac{a_i^2}{(1-\rho^2)}\mathrm{E}\left\{y_i^S m_2(b_i)\left[y_i^O-\gamma^\intercal z_i\right]^2\right\}\\&+\frac{2\rho a_i}{\sigma(1-\rho^2)}\mathrm{E}\left\{y_i^S m_2(b_i)\left[y_i^O-\gamma^\intercal z_i\right]^3\right\}\\&+\frac{\rho^2}{\sigma^2(1-\rho^2)}\mathrm{E}\left\{y_i^S m_2(b_i)\left[y_i^O-\gamma^\intercal z_i\right]^4\right\}.\end{aligned} \tag{A.13}$$

Substituting (3.65), (3.58) and (3.61) in (A.13) and rearranging we obtain the required expectation.

By exploiting the definition of b_i in (3.14), the expectation of $y_i^S b_i m_2(b_i)\left[y_i^O-\gamma^\intercal z_i\right]^3$ can be rewritten as

$$\begin{aligned}\mathrm{E}\left\{y_i^S b_i m_2(b_i)\left[y_i^O-\gamma^\intercal z_i\right]^3\right\}&=\frac{a_i}{\sqrt{1-\rho^2}}\mathrm{E}\left\{y_i^S m_2(b_i)\left[y_i^O-\gamma^\intercal z_i\right]^3\right\}\\&+\frac{\rho}{\sigma\sqrt{1-\rho^2}}\mathrm{E}\left\{y_i^S m_2(b_i)\left[y_i^O-\gamma^\intercal z_i\right]^4\right\}.\end{aligned} \tag{A.14}$$

Substituting (3.61) and (3.68) in the above we get the required expectation.

Using (3.61) and (3.68) and the definition of b_i, the expectation of $y_i^S b_i^2 m_2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]^3$ can be written as

$$\begin{aligned}\mathrm{E}\big\{y_i^S b_i^2 m_2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]^3\big\} &= \frac{\rho^2}{\sigma^2(1-\rho^2)}\mathrm{E}\big\{y_i^S m_2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]^5\big\} \\ &+ \frac{\rho\sigma^3 a_i}{\sqrt{1-\rho^2}}\big\{2\rho^4\big[3-a_i^2(6-a_i^2)\big] - \rho^2\big[12-a_i^2(15-a_i^2)\big] \\ &+ 3(2-a_i^2)\big\}\phi(a_i), \end{aligned} \tag{A.15}$$

while using (3.68) and the definition of b_i, the expectation of $y_i^S b_i m_2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]^4$ can be written as

$$\begin{aligned}\mathrm{E}\big\{y_i^S b_i m_2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]^4\big\} &= \frac{\rho}{\sigma\sqrt{1-\rho^2}}\mathrm{E}\big\{y_i^S m_2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]^5\big\} \\ &+ \sigma^4 a_i\big\{\rho^4\big[3-a_i^2(6-a_i^2)\big] - 6\rho^2(1-a_i^2) + 3\big\}\phi(a_i). \end{aligned} \tag{A.16}$$

Substituting (3.65), (A.15), (3.69), (A.16) and (3.68) in the third Bartlett identity $\mathrm{E}\big[(\partial^2 l_i/\partial(\sigma^2)^2)(\partial l_i/\partial\rho)\big] + \mathrm{E}\big[\partial^3 l_i/\partial(\sigma^2)^2\partial\rho\big] + \mathrm{E}\big[(\partial l_i/\partial\sigma^2)^2(\partial l_i/\partial\rho)\big] + 2\mathrm{E}\big[(\partial l_i/\partial\sigma^2)(\partial^2 l_i/\partial\sigma^2\partial\rho)\big] = 0$, we obtain an equation in one unknown which is the expectation of $y_i^S m_2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]^5$.

The expectation of $y_i^S b_i m_2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]^4$ is easily obtained by substituting (3.70) in (A.16).

The expectations of $y_i^S b_i m_2^2(b_i)$, $y_i^S b_i m_2^2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]$, $y_i^S b_i m_2^2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]^2$ and $y_i^S b_i m_2^2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]^3$ may be written respectively as

$$\mathrm{E}\big\{y_i^S b_i m_2^2(b_i)\big\} = \frac{a_i}{\sqrt{1-\rho^2}}\mathrm{E}\big\{y_i^S m_2^2(b_i)\big\} + \frac{\rho}{\sigma\sqrt{1-\rho^2}}\mathrm{E}\big\{y_i^S m_2^2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]\big\}, \tag{A.17}$$

$$\begin{aligned}\mathrm{E}\big\{y_i^S b_i m_2^2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]\big\} &= \frac{a_i}{\sqrt{1-\rho^2}}\mathrm{E}\big\{y_i^S m_2^2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]\big\} \\ &+ \frac{\rho}{\sigma\sqrt{1-\rho^2}}\mathrm{E}\big\{y_i^S m_2^2(b_i)\big[y_i^O - \gamma^\intercal z_i\big]^2\big\}, \end{aligned} \tag{A.18}$$

$$\mathrm{E}\{y_i^S b_i m_2^2(b_i)[y_i^O - \gamma^\intercal z_i]^2\} = \frac{a_i}{\sqrt{1-\rho^2}}\mathrm{E}\{y_i^S m_2^2(b_i)[y_i^O - \gamma^\intercal z_i]^2\} + \frac{\rho}{\sigma\sqrt{1-\rho^2}}\mathrm{E}\{y_i^S m_2^2(b_i)[y_i^O - \gamma^\intercal z_i]^3\}, \quad \text{(A.19)}$$

$$\mathrm{E}\{y_i^S b_i m_2^2(b_i)[y_i^O - \gamma^\intercal z_i]^3\} = \frac{a_i}{\sqrt{1-\rho^2}}\mathrm{E}\{y_i^S m_2^2(b_i)[y_i^O - \gamma^\intercal z_i]^3\} + \frac{\rho}{\sigma\sqrt{1-\rho^2}}\mathrm{E}\{y_i^S m_2^2(b_i)[y_i^O - \gamma^\intercal z_i]^4\}. \quad \text{(A.20)}$$

Appendix B

Algebraic derivations for the matched gamma pairs model

In this appendix we derive the expectations required for the calculation of the $P(\psi,\lambda)$ and $Q(\psi,\lambda)$ matrices of Firth (1993) adjusted score equations in Section 5.2.3.

Since $Y_{i1} \sim \text{Gamma}(m, \lambda_i/\psi)$ and $Y_{i2} \sim \text{Gamma}\big(m, 1/(\psi\lambda_i)\big)$ using the rate parametrization, $\text{E}[Y_{i1} - (m\psi/\lambda_i)] = \text{E}[Y_{i2} - m\psi\lambda_i] = 0$. The second central moments of Y_{i1} and Y_{i2} are given respectively by $m\psi^2/\lambda_i^2$ and $m\psi^2\lambda_i^2$. The third central moments of Y_{i1} and Y_{i2} are $2m\psi^3/\lambda_i^3$ and $2m\psi^3\lambda_i^3$, respectively.

$$\begin{aligned}
\text{E}(x^3) &= \frac{1}{\psi^6}\sum_{i=1}^{q}\text{E}\left\{\left[\lambda_i\left(Y_{i1} - \frac{m\psi}{\lambda_i}\right) + \frac{1}{\lambda_i}\left(Y_{i2} - \lambda_i m\psi\right)\right]^3\right\} \\
&= \frac{1}{\psi^6}\sum_{i=1}^{q}\text{E}\left\{\lambda_i^3\left(Y_{i1} - \frac{m\psi}{\lambda_i}\right)^3 + \frac{1}{\lambda_i^3}\left(Y_{i2} - \lambda_i m\psi\right)^3\right\} \\
&= \frac{1}{\psi^6}\sum_{i=1}^{q} 4m\psi^3 \\
&= \frac{4qm}{\psi^3},
\end{aligned}$$

Where the second equality above follows because the expectation of the other terms of the cubic are null. Now for ease of presentation of the forthcoming expressions, let $c_{i1} :=$

$y_{i1} - (m\psi/\lambda_i)$ and $d_{i2} := y_{i2} - \lambda_i m\psi$. For $r = 1, \ldots, q$,

$$\begin{aligned} x^2 w_r &= \frac{1}{\psi^5}\left(\frac{1}{\lambda_r^2} y_{r2} - y_{r1}\right)\left(\sum_{i=1}^{q}\left[\lambda_i c_{i1} + \frac{1}{\lambda_i} d_{i2}\right]\right)^2 \\ &= \frac{1}{\psi^5}\left(\frac{1}{\lambda_r^2} d_{r2} - c_{r1}\right)\left(\sum_{i=1}^{q}\left[\lambda_i c_{i1} + \frac{1}{\lambda_i} d_{i2}\right]\right)^2 \\ &= \frac{1}{\psi^5}\left(\frac{1}{\lambda_r^2} d_{r2} - c_{r1}\right)\left(\sum_{i=1}^{q}\left[\lambda_i c_{i1} + \frac{1}{\lambda_i} d_{i2}\right]^2\right. \\ &+ \left.\sum_{i=1}^{q}\sum_{\substack{j=1 \\ j\neq i}}^{q}\left[\lambda_i c_{i1} + \frac{1}{\lambda_i} d_{i2}\right]\left[\lambda_j c_{j1} + \frac{1}{\lambda_j} d_{j2}\right]\right). \end{aligned}$$

The expectation of the second term of the last equality above is zero since $i \neq j$. If $i \neq r$ then the expectation of the first term is also zero. If $i = r$ then the expectation of the first term is $\mathrm{E}\left[(D_{r2}^3/\lambda_r^4) - \lambda_r^2 C_{r1}^3\right]/\psi^5 = \left[(2m\psi^3\lambda_r^3/\lambda_r^4) - \lambda_r^2(2m\psi^3/\lambda_r^3)\right]/\psi^5 = 0$. For $r, s = 1, \ldots, q$,

$$\begin{aligned} x w_r w_s &= \frac{1}{\psi^4}\left(\frac{1}{\lambda_r^2} y_{r2} - y_{r1}\right)\left(\frac{1}{\lambda_s^2} y_{s2} - y_{s1}\right)\left(\sum_{i=1}^{q}\left[\lambda_i c_{i1} + \frac{1}{\lambda_i} d_{i2}\right]\right) \\ &= \frac{1}{\psi^4}\left(\frac{1}{\lambda_r^2} d_{r2} - c_{r1}\right)\left(\frac{1}{\lambda_s^2} d_{s2} - c_{s1}\right)\left(\sum_{i=1}^{q}\left[\lambda_i c_{i1} + \frac{1}{\lambda_i} d_{i2}\right]\right). \end{aligned}$$

When $r \neq s$, $\mathrm{E}(x w_r w_s) = 0$. When $r = s$ but is not equal to i, $\mathrm{E}(x w_r^2) = 0$. When $i = r = s$, $\mathrm{E}(x w_r^2) = \mathrm{E}\left[(D_{r2}^3/\lambda_r^5) + \lambda_r C_{r1}^3\right]/\psi^4 = \left[(2m\psi^3\lambda_r^3/\lambda_r^5) + \lambda_r(2m\psi^3/\lambda_r^3)\right]/\psi^4 = 4m/(\psi\lambda_r^2)$. For $r, s, t = 1, \ldots, q$,

$$w_r w_s w_t = \frac{1}{\psi^3}\left(\frac{1}{\lambda_r^2} d_{r2} - c_{r1}\right)\left(\frac{1}{\lambda_s^2} d_{s2} - c_{s1}\right)\left(\frac{1}{\lambda_t^2} d_{t2} - c_{t1}\right).$$

If $r \neq s \neq t$ or any two of r, s and t are equal, then the expectation of the above reduces to zero. If $r = s = t$ then $\mathrm{E}(w_r^3) = \mathrm{E}\left[(D_{r2}^3/\lambda_r^6) - C_{r1}^3\right]/\psi^3 = \left[(2m\psi^3\lambda_r^3/\lambda_r^6) - (2m\psi^3/\lambda_r^3)\right]/\psi^3 = 0$.

$$-zx = \frac{2}{\psi^5}\left(\sum_{i=1}^{q}\sum_{j=1}^{q}\left[\lambda_i c_{i1} + \frac{1}{\lambda_i} d_{i2} + m\psi\right]\left[\lambda_j c_{j1} + \frac{1}{\lambda_j} d_{j2}\right]\right).$$

If $i \neq j$, $\mathrm{E}(-zx) = 0$. If $i = j$, $\mathrm{E}(-zx) = 2\sum_{i=1}^{q} \mathrm{E}\left[\lambda_i^2 C_{i1}^2 + (D_{i2}^2/\lambda_i^2)\right]/\psi^5 = 4mq/\psi^3$. For

$r = 1, \ldots, q$,

$$\frac{1}{\psi} x w_r = \frac{1}{\psi^4} \left(\frac{1}{\lambda_r^2} d_{r2} - c_{r1} \right) \left(\sum_{i=1}^{q} \left[\lambda_i c_{i1} + \frac{1}{\lambda_i} d_{i2} \right] \right).$$

$\mathrm{E}(x w_r / \psi) = 0$ whether $i = r$ or not.

$$\begin{aligned}
\mathrm{E}\left(\frac{2}{\psi \lambda_r^3} x y_{r2} \right) &= \frac{2}{\psi^3 \lambda_r^3} \mathrm{E}\left\{ \sum_{i=1}^{q} y_{r2} \left[\lambda_i \left(y_{i1} - \frac{m\psi}{\lambda_i} \right) + \frac{1}{\lambda_i} \left(y_{i2} - \lambda_i m \psi \right) \right] \right\} \\
&= \frac{2}{\psi^3 \lambda_r^3} \sum_{i=1}^{q} \mathrm{E}\left[\frac{1}{\lambda_i} y_{r2} \left(y_{i2} - \lambda_i m \psi \right) \right] \\
&= \frac{2}{\psi^3 \lambda_r^3} \mathrm{E}\left[\frac{1}{\lambda_r} y_{r2} \left(y_{r2} - \lambda_r m \psi \right) \right] \\
&= \frac{2m}{\psi \lambda_r^2},
\end{aligned}$$

where the third equality follows since the expectation of the term inside the sum is non zero only when $i = r$ and where we used the fact that the second raw moment of Y_{i2} is $m(m+1)\psi^2 \lambda_i^2$.

$$-z w_r = \frac{2}{\psi^4} \left(\frac{1}{\lambda_r^2} d_{r2} - c_{r1} \right) \left(\sum_{i=1}^{q} \left[\lambda_i c_{i1} + \frac{1}{\lambda_i} d_{i2} + m\psi \right] \right).$$

The expected value of the above is null whether $i = r$ or not.

$$\frac{1}{\psi} w_r w_s = \frac{1}{\psi^3} \left(\frac{1}{\lambda_r^2} d_{r2} - c_{r1} \right) \left(\frac{1}{\lambda_s^2} d_{s2} - c_{s1} \right).$$

When $r \neq s$, $\mathrm{E}(w_r w_s / \psi) = 0$. When $r = s$, $\mathrm{E}(w_r w_s / \psi) = \mathrm{E}\left[(D_{r2}^2 / \lambda_r^4) + C_{r1}^2 \right] / \psi^3 = 2m / \psi \lambda_r^2$, for $r = 1, \ldots, q$.

$$\frac{2}{\psi \lambda_r^3} y_{r2} w_s = \frac{2}{\psi^2 \lambda_r^3} y_{r2} \left(\frac{1}{\lambda_s^2} y_{s2} - y_{s1} \right).$$

If $r \neq s$, the expectation of the above is zero by independence. If $r = s$, $2\mathrm{E}\left[Y_{r2} W_s \right] / (\psi^2 \lambda_r^3) = 2\mathrm{E}\left[(Y_{r2}^2 / \lambda_r^2) - Y_{r2} Y_{r1} \right] / (\psi^2 \lambda_r^3) = 2m / \lambda_r^3$.

Bibliography

Albert, A. and J. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*(1), 1–10.

Amemiya, T. (1973). Regression analysis when the dependent variable is tuncated normal. *Econometrica 41*(6), 997–1016.

Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature 19*(4), 1483–1536.

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics 24*(1-2), 3–61.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge, Massachusetts.

Barndorff-Nielson, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika 70*(2), 343–365.

Bierens, H. J. (2007). Maximum likelihood estimation of Heckman's sample selection model. *Unpublished manuscript, Pennsylvania State University*.

Breslow, N. (1981). Odds ratio estimators when the data are sparse. *Biometrika 68*(1), 73–84.

Butler, K. and M. A. Stephens (2017). The distribution of a sum of independent binomial random variables. *Methodol. Comp. Appl. Prob. 19*, 557–571.

Calzolari, G. and G. Fiorentini (1993). Alternative covariance estimators of the standard tobit model. *Economics Letters 42*(1), 5–13.

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society 49*(1), 1–39.

Cox, D. R. and N. Reid (1992). A note on the difference between profile and modified profile likelihood. *Biometrika 79*(2), 408–11.

Cox, D. R. and E. J. Snell (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological 30*, 248–275.

Crowley, J. and M. Hu (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association 72*(357), 27–36.

Davison, A. C. (1988). Approximate conditional inference in generalized linear models. *Journal of the Royal Statistical Society 50*(3), 445–461.

Davison, A. C. (2003). *Statistical Models*. Cambridge: Cambridge University Press.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *The Annals of Statistics 3*, 1189–1217.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Fair, R. C. (1958). A theory of extramarital affairs. *Journal of Political Economy 86*(1), 45–61.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38.

Florescu, I. (2014). *Probability and Stochastic Processes*. John Wiley and Sons.

Gart, J. J. (1970). Point and interval estimation of the common odds ratio in the combination of 2×2 tables with fixed marginals. *Biometrika 57*(3), 471–475.

Gart, J. J. (1971). The comparison of proportions: A review of significance tests, confidence intervals and adjustments for stratification. *Int. Statist. Rev. 39*(2), 148–169.

Gourieroux, C., A. Monfort, and E. Renault (1993). Indirect inference. *Journal of Applied Econometrics 8*, 85–118.

Greene, W. (2004). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *The Econometrics Journal 7*, 98–119.

Greene, W. H. (2012). *Econometric Analysis* (7th ed.). Prentice Hall.

Guerrier, S., E. Dupuis-Lozeron, Y. Ma, and M.-P. Victoria-Feser (2019). Simulation-based bias correction methods for complex models. *Journal of the American Statistical Association 114*(525), 146–157.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement 5*(4), 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine 21*, 2409–2419.

Henningsen, A. and O. Toomet (2011). maxlik: A package for maximum likelihood estimation in R. *Computational Statistics 26*(3), 443–458.

Jeng, S. and W. Q. Meeker (2000). Comparisons of approximate confidence interval procedures for type I censored data. *Technometrics 42*(2), 135–148.

Johnson, N. L. and S. Kotz (1970). *Distributions in Statistics: Continuous Univariate Distributions* (2nd ed.), Volume 1. New York: John Wiley and Sons.

Johnston, J. and J. DiNardo (1997). *Econometric Methods*. McGraw-Hill.

Kalbfleisch, J. and R. Prentice (2002). *The Statistical Analysis of Failure Time Data* (2nd ed.). New York: John Wiley and Sons.

Kennan, J. (1985). The duration of contract strikes in U.S. manufacturing. *Journal of Econometrics 28*, 5–28.

Kiefer, N. (1985). Econometric analysis of duration data. *Journal of Econometrics 28*(1), 1–169.

Kleinbaum, D. G. and M. Klein (2011). *Survival Analysis: A Self-Learning Text* (3rd ed.). New York: Springer.

Kosmidis, I. (2014). Bias in parametric estimation: reduction and useful side-effects. *WIREs Computational Statistics 6*, 185–196.

Kosmidis, I. and D. Firth (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics 4*, 1097–1112.

Kosmidis, I. and N. Lunardon (2020). Empirical bias-reducing adjustments to estimating functions. *arXiv:2001.03786*, 1–34.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society 57*(2), 395–407.

Lawless, J. (2002). *Statistical Models and Methods for Lifetime Data*. John Wiley and Sons.

Leung, S. F. and S. Yu (2000). Collinearity and two-step estimation of sample selection models: problems, origins, and remedies. *Computational Economics 15*, 173–199.

Lin, H. M., W. J. M. and H. Y. Kim (2020). Firth adjustment for Weibull current-status survival analysis. *Communications in Statistics - Theory and Methods 49*(18), 4587–4602.

Lindsay, B. (1982). Conditional score functions: some optimality results. *Biometrika 69*(3), 503–512.

Lunardon, N. (2018). On bias reduction and incidental parameters. *Biometrika 105*(1), 233–238.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge University Press.

Magnus, J. R. and H. Neudecker (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (3rd ed.). John Wiley and Sons.

McCullagh, P. and R. Tibshirani (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society 52*(2), 325–344.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica 55*(4), 765–799.

Nash, J. C. and R. Varadhan (2011). Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software 43*(9), 1–14.

Nelson, W. B. and G. J. Hahn (1972). Linear estimation of a regression relationship from censored data: 1. simple methods and their applications (with discussion). *Technometrics 14*, 247–276.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrika 16*(1), 1–32.

Olsen, R. J. (1978). Note on the uniqueness of the maximum likelihood estimator for the Tobit model. *Econometrica 46*(5), 1211–1215.

Owen, D. (1980). A table of normal integrals. *Communications in Statistics: Simulation and Computation B9*(4), 389–419.

Pace, L. and A. Salvan (1997). *Principles of Statistical Inference: from a Neo-Fisherian Perspective*. Singapore: World Scientific Publishing Co.

Pike, M. C. (1966). A method of analysis of certain class of experiments in carcinogenesis. *Biometrics 22*, 142–161.

Prentice, R. L. (1973). Exponential survivals with censoring and explanatory variables. *Biometrika 60*(2), 279–288.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika 43*(3/4), 353–360.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika 90*(3), 533–549.

Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine 2*, 71–78.

Severini, T. A. (2000). *Likelihood Methods in Statistics*, Volume 22. Oxford University Press.

Silvey, S. D. (1970). *Statistical Inference*. Penguin: Harmondsworth.

Stacy, E. W. (1962). A generalisation of the gamma distribution. *The Annals of Mathematical Statistics 33*, 1187–1192.

Therneau, T. M. (2021). *A Package for Survival Analysis in R*. R package version 3.2-10.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica 26*(1), 24–36.

Toomet, O. and A. Henningsen (2008). Sample selection models in R: Package sampleSelection. *Journal of Statistical Software 27*(7), 1–23.

Wang, W., F. H. and J. Yan (2019). *reda: Recurrent Event Data Analysis*. R package version 0.5.2.

Witte, A. D. (1980). Estimating the economic model of crime with individual data. *Quarterly Journal of Economics 94*(1), 57–84.