Università
degli Studi
di Padova

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXIX

Reducing the Impact of Bias in Likelihood Inference for Prominent Model Settings

Coordinatore del Corso: Prof. Monica Chiogna

Supervisore: Prof. Nicola Sartori

Co-supervisore: Dr. Ioannis Kosmidis

Dottoranda: Claudia Di Caterina

31 Gennaio 2017

Abstract

The existence of bias in inferential procedures based on the likelihood function has given rise to a great deal of research in the statistical literature. The magnitude of such bias plays a crucial role in estimation: if large, misleading conclusions on the quantities of interest are likely to be drawn. This is a matter of serious concern when the available sample size is small to moderate or when the model under study does not meet the regularity conditions for usually reliable maximum likelihood inference. In the present thesis, we attempt to reduce the impact of bias in both these circumstances, by following distinct paths. For finite-sample problems, we propose a convenient way to refine Wald-type inference in regression settings through asymptotic bias correction of the z-statistic. Such approach stems from the intuition of seeing that pivot as the estimator of a model reparametrization. For non-regular problems, with special focus on scenarios characterized by the presence of incidental parameters, we suggest a strategy to extend the current range of applications of the modified profile likelihood. This solution, founded on Monte Carlo simulation, is versatile enough to cope with several nonstandard modeling frameworks for grouped data.

Sommario

L'esistenza di distorsione nelle procedure inferenziali basate sulla funzione di verosimiglianza ha dato origine ad un grande flusso di ricerca nella letteratura statistica. L'entità di tale distorsione detiene un ruolo cruciale nel processo di stima: se grande, può portare a conclusioni fuorvianti sulle quantità di interesse. Tale questione è oggetto di particolare preoccupazione quando la numerosità campionaria è modesta o quando il modello oggetto di studio non rispetta le condizioni di regolarità necessarie ad ottenere risultati affidabili tramite le usuali techniche di massima verosimiglianza. In questa tesi, si tenta di ridurre l'impatto della distorsione in entrambe le circostanze, seguendo vie differenti. Per problemi in campioni di modesta grandezza, viene proposto un modo pratico per migliorare l'inferenza condotta col test di Wald in modelli di regressione. Tale approccio, incentrato sulla correzione asintotica della distorsione della statistica utilizzata, deriva dall'intuizione di guardare ad essa come allo stimatore di una riparametrizzazione del modello. Per problemi non regolari di stima caratterizzati dalla presenza di parametri incidentali, si suggerisce una strategia volta ad estendere il campo di applicazione della verosimiglianza profilo modificata. La versatilità di questa soluzione, fondata sulla simulazione Monte Carlo, permette di trattare vari modelli complessi per dati raggruppati.

To Cri,

the best example of dedication I could have.

Acknowledgements

Coming to the end of my PhD was not an easy task, but I was fortunate enough to be supported in many respects throughout these years. And given my poor aptitude for writing, especially in a language which is not mine, I apologize since now because the next lines will probably not be able to properly express my gratitude.

First, I would like to thank the people without whom this dissertation could not exist. Let me begin with my supervisor, Nicola Sartori, who not only was an invaluable mentor but also was so patient to bear the various down phases I experienced during my graduate studies. Then my co-supervisor, Ioannis Kosmidis, whose knowledge of statistics was essential for guiding me in the research process. I am also very grateful to Professor Alessandra Salvan, both for how she taught me the statistical subjects I love most and for her tangible help in deriving some crucial results of the thesis. Even the technical suggestions by the two external evaluators, Alastair Young and Geert Dhaene, fundamentally improved the scientific value of this final manuscript.

After so much time spent there, the Department of Statistical Sciences in Padova is now a place where I feel home. Among those that made this possible, a special mention goes to Tommy and Bruno, two statisticians who greatly contributed also to my personal growth. Nonetheless, the first reason why I am glad I was a PhD student is Lucia. She is no more a colleague to me, she is one of the closest friends. I am deeply thankful to her, the only human being with whom I can share work issues as openly as private thoughts.

At this point, I need to thank the people who perhaps do not understand much of statistics, but have been equally important for the conclusion of my educational path. I owe my family more than what is possible to describe with words and I really hope this thesis makes them proud. My mother, particularly, is well aware of the effort I had to put for pursuing such goal. I want her to be sure that I immensely appreciate the effort she in turn had to put for standing by my side in the hardest moments.

Honestly, I do not believe I would be writing these acknowledgments if my friends had not been there. Above all, I will always be grateful to Eli and Carlo for the remote assistance they supplied me while I was, not just physically, far away. Two other friends prevented me from getting lost in London: Jess and Long. I say thanks to them and, generally, to “my persons”, the persons I really missed during that period. They may not know what bias is, yet they are the best method to reduce mine.

Contents

List of Figures

List of Tables

Introduction

Overview

The general notion of bias, i.e. systematic distortion in mean of some quantity, is unquestionably central to statistics. Both researchers and practitioners are concerned with the problem of bias in estimation, since reliability of inferential conclusions is closely tied to its magnitude.

The frequency-decision paradigm is the one which gives more emphasis to the unbiasedness of a statistical procedure. In that theoretical framework, the purpose of such property is twofold: it is both a fundamental criterion to restrict the class of potential inference techniques in order to find the optimal one and a condition to guarantee the impartiality of one method with respect to the various parameter values (Lehmann and Romano, 2006, Section 1.5).

From a Fisherian viewpoint, unbiasedness of usual statistical procedures in regular models (see, for example, Pace and Salvan, 1997, Section 3.4, for a characterization of regularity conditions) is ensured only asymptotically. Indeed, likelihood-based quantities are generally biased when the sample is of small or even moderate size. Particularly, inferential inaccuracies caused by the bias of the maximum likelihood (ML) estimator have given rise to an ongoing stream of research on disparate ways for reducing it. In the related body of literature, the rich diversity of approaches admits to be classified according to several aspects.

A useful distinction can be made between methods for bias correction and methods for bias reduction (Kosmidis, 2007). Techniques belonging to the first category foresee the derivation of a bias-corrected estimator by subtracting from the ML one a suitable approximation of its bias. One popular manner to estimate such bias is via bootstrap resampling (Efron, 1979). Asymptotic corrective procedures require instead the analytical expression of the leading term in the asymptotic bias expansion of the ML estimator. For a broad family of regular scenarios, this was obtained by Cox and Snell (1968), in their investigation of higher-order properties of residuals in parametric models. On the grounds of that finding, Anderson and Richardson (1979) and Schaefer (1983) computed specific formulae for bias correction in the logistic regression. Later on, Efron (1975) studied the bias-corrected estimator derived upon evaluation of the first-order bias at the ML estimate. Such quantity was shown both to have bias of smaller order than the classical ML estimator and to be second-order efficient (refer also to Section 9.4.3 in Pace and Salvan, 1997). Asymptotic bias correction was also successively applied by Cook *et al.* (1986) in nonlinear regressions with normal errors and by Cordeiro and McCullagh (1991) in the context of generalized linear models (Nelder and Wedderburn, 1972).

The main advantage of bias-correction methods is the simplicity of their implementation, once an approximation to the bias is available. Yet, such procedures also suffer from one serious limitation: bias-corrected estimates inherit the instabilities of ML ones. This represents a critical drawback in situations with categorical responses, where there is a positive probability that the ML estimator is infinite. Among others, Bull *et al.* (2002) and Kosmidis and Firth (2009) examined the topic. Besides that, asymptotic bias correction poses an additional problem because is performable only when the first term in the bias expansion of the ML estimator may be expressed in closed form. For some models this exercise is tiresome, if not impracticable.

The class of bias-reduction methods differentiates from the former in one crucial respect: these techniques do not depend directly on the ML estimator. To some extent, they can be interpreted as bias-preventive (Kosmidis, 2007). In fact, a new estimator is obtained in such a way that its bias is known to be asymptotically smaller than that of the ML one. Eminent examples of bias-reduced estimators are based on modified score functions (Warm, 1989). Formalization of such approach in regular settings is owed to Firth (1993), who setup a general methodological framework for finding first-order unbiased estimators by solution of an adjusted score equation. This procedure has proved to be notably useful when dealing with models for discrete dependent variable. More precisely, empirical evidence in Heinze and Schemper (2002) and Zorn (2005) indicated that bias-reduced estimates in logistic regressions are always finite, even in cases where ordinary ML estimates are not. The desirable attributes of the bias-reduced estimator under a number of categorical-response scenarios were also investigated by Bull *et al.* (2007) and Kosmidis and Firth (2011). Nevertheless, such technique shares a defect with asymptotic bias-corrective methods: in order to be implemented, not only the score function and the Fisher information need to be explicitly available, but also the first-order bias of the ML estimator.

The discussion above has focused on methods to reduce the impact of the finite-sample bias of the ML estimator in regular estimation problems. However, if the model under study does not fulfill the standard conditions, even the habitual asymptotic unbiasedness can fail. This happens, for instance, in models where the dimension of the parameter space increases along with the sample size. Inappropriateness of ML inferential procedures due to such deviation from regularity was first brought to light by Neyman and Scott (1948) and in fact is well-known in the statistical and econometric literatures as Neyman & Scott or incidental parameters problem (Lancaster, 2000).

As pointed out by Kosmidis (2014), in these circumstances reduction of bias may be achieved by means of the modified profile likelihood function (Barndorff-Nielsen, 1983). Indeed, the implicated adjustment to the profile likelihood eliminates the highest-order term in the asymptotic bias expansion of the profile score. Since such part can get considerably large in models subject to Neyman & Scott problems (McCullagh and Tibshirani, 1990), that inferential instrument has been found especially suited for drawing trustworthy conclusions on the parameter of interest in the presence of many nuisance parameters. Specifically, when data are collected in clusters and each incidental component is related to a group in the sample, the modified profile likelihood delivers estimators with improved properties. In a way, respecting the preceding lines of reasoning, this function may then be thought of as a bias-reduction method in the two-index asymptotic setting (Sartori, 2003).

Main contributions of the thesis

For providing a motivation of our work, the previous section was essentially dedicated to depict the prominence of the role of bias in likelihood-based inference. To this aim, only few of the numerous attempts to limit its effects on estimation have been cited. In this thesis, two separate routes toward the reduction of bias are taken.

On one hand, we tackle the typical bias of likelihood quantities in small-to-moderate samples. This task, as already seen, has been extensively carried out in the past with reference to the ML estimator via bias-corrective and bias-reducing methods. However, the biased quantity considered here is a statistic. Precisely, it is the Wald z-statistic, largely used in regression contexts to test the significance of one predictor's influence. The original idea in this first analysis lies in looking at the z-statistic as at an estimator of a model reparametrization. Such expedient allows to obtain a convenient closed-form expression of its first-order bias for performing asymptotic bias correction.

On the other hand, we address the asymptotic bias (meaning inconsistency) of likelihood quantities in non-regular problems. In particular, the attention is turned to the erroneous inferences supplied by the profile likelihood when incidental parameters are present. The employment of the modified profile likelihood and of its approximation proposed by Severini (1998b) has already proven fruitful in several models for clustered data, where problems of Neyman & Scott can be severe. Nonetheless, the difficulty of their computation prevents these two functions from being fully exploited. Our main contribution in this regard is to propose a new strategy to calculate the modified profile likelihood even in nonstandard modeling frameworks. Such recommended solution, based on Monte Carlo simulations, is simple and widely applicable.

The rest of the current dissertation is organized in the following way. In Chapter 1, we set up some of the notation which is used throughout the thesis and we outline the general features of likelihood-related quantities, highlighting their connection with the bias issue. Special heed is paid to the strengths and weaknesses of the Wald pivot and to those functions used for making inference on the component of interest in the global parameter. Section 1.4.3 closes the chapter by giving an account of the properties of the modified profile likelihood under the two-index asymptotic scenario.

In Chapter 2, an approach for enhancing the normal approximation to the null distribution of the z-statistic in small-to-moderate-sized samples is suggested. Such procedure basically consists in the correction of the moments of the combinant. Section 2.2 investigates the validity of this strategy in specific single-parameter models. A more general proposal to derive an adjusted z-test in regression settings is put forward in Section 2.3. Characteristics of the associated location adjusted z-statistic are studied both analytically and empirically by Sections 2.4 and 2.5, in the relevant context of generalized linear models. Section 2.6 delineates open topics and traces future avenues of research in the area.

Chapter 3 is dedicated to demonstrate how the domain of applicability of the modified profile likelihood is expanded by taking advantage of simulation. In Section 3.2 we present the Monte Carlo approximation to Severini's function, with particular mention to the great generality of its implementation. In the remaining parts of the chapter, the helpfulness of this solution is illustrated through simulation experiments considering fairly complex modeling assumptions. In more detail, Section 3.3 deals with an econometric model for dependent observations, Section 3.4 discusses inference on binary datasets with missing values and Section 3.5 examines a regression scenario for censored survival data. We make some final remarks in Section 3.6, where also the agenda for further investigations is established.

Chapter 1

Likelihood-based inference in the presence of bias

1.1 Likelihood and related quantities

Let $\mathcal{F} = \left\{p_Y(y;\theta), \theta \in \Theta \subseteq \mathrm{I\!R}^k\right\}$ be a family of probability density functions for the random variable Y which varies in the sample space $\mathcal{Y}$. Such class of models is indexed by the parameter θ, taking values in the compact non-empty set Θ. The random variable Y describes the available data, which in basic settings can be expressed as $y = (y_1, \ldots, y_n)$, with n representing the total sample size. Obviously, complexity of the experimental design can be higher and more than one index might be convenient to identify the units in the sample. Circumstances like the latter will be considered in Section 1.4 and better investigated in Chapter 3. Note that, here and henceforth, in order to avoid clutter we omit the transpose symbol acting on vectors unless such an omission could result in ambiguity. Furthermore, the theory in this first part of the thesis is presented referring to absolutely continuous and independent random variables, but all results apply in fact also to the discrete case and to more general frameworks where the information supplied by the data increases along with the sample size n.

The likelihood function for θ takes the form

$$L(\theta) = L(\theta; y) = p_Y(y;\theta),$$

and the associated log-likelihood function is simply its logarithm, i.e. $l(\theta) = l(\theta; y) = \log L(\theta)$. The maximum likelihood (ML) estimate for model $\mathcal{F}$ can then be defined as $\hat\theta = \hat\theta(y) = \arg\max_{\theta\in\Theta} l(\theta)$. With a slight abuse of notation, we shall also use $\hat\theta = \hat\theta(Y)$ to denote the corresponding random variable, known as ML estimator, since the specific meaning will always be evident by the context. Inferential techniques resulting from the likelihood are founded on the probability distributions of the random variable $l(\theta; Y)$ and of its related quantities, for θ fixed and y varying in $\mathcal{Y}$ according to some density $p_Y(y; \tilde{\theta})$ in $\mathcal{F}$, where $\tilde{\theta} \in \Theta$ is a parameter value not necessarily equal to θ. On the same lines as Pace and Salvan (1997, Section 1.4), when considering $\tilde{\theta} = \theta$ we will speak of null distribution and of null moments for a certain likelihood-based quantity. Furthermore, symbols such as $P_\theta(\cdot), E_\theta(\cdot)$ and $\mathrm{Var}_\theta(\cdot)$ shall indicate the event probability, expected value and variance, respectively, computed with reference to $p_Y(y; \theta)$.

An important feature of the log-likelihood function is its invariance to the parametrization of the model. In particular, if $\omega = \omega(\theta)$ is a one-to-one infinitely differentiable smooth function from $\Theta \subseteq \mathbb{R}^k$ to $\Omega \subseteq \mathbb{R}^k$, the log-likelihood under the transformation ω is $l^\Omega(\omega) = l^\Theta(\theta(\omega))$, where $l^\Theta(\theta)$ is the log-likelihood in the parametrization θ and $\theta(\omega)$ is the inverse function of $\omega(\theta)$. From this follows the important property of equivariance under reparametrization of the ML estimate, implying that $\hat{\omega} = \omega(\hat{\theta})$ and $\hat{\theta} = \theta(\hat{\omega})$.

Now assume that, possibly after a change in the parametrization of the model, the global k-dimensional parameter θ can be partitioned into (ψ, λ), where ψ is the parameter of interest having dimension k_0 and λ is the nuisance component, of dimension $k - k_0$. Given that in this case it is possible to write $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$, let $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$ be the constrained ML estimate of θ, with $\hat{\lambda}_\psi$ indicating the ML estimate of λ for a fixed value of ψ. In such situations, one desirable property of a statistical procedure is invariance under interest-respecting parametrization (Pace and Salvan, 1997, Section 4.2.4). This principle advocates that inferential conclusions for ψ obtained in respect of the original parametrization $\theta = (\psi, \lambda)$ are compatible with those derived for the component of interest $\tau = \tau(\psi)$ under the interest-preserving transformation $\omega = (\tau, \zeta)$, where τ is one-to-one and $\zeta = \zeta(\psi, \lambda)$.

The full k-dimensional score vector is $l_\theta = \partial l(\theta)/\partial\theta$, whereas the observed information and the Fisher expected information $k \times k$ matrices are defined as $j(\theta) = -l_{\theta\theta} = -\partial^2 l(\theta)/(\partial\theta\partial\theta^{\mathrm{T}})$ and $i(\theta) = E_\theta\{j(\theta)\}$, respectively. Partial derivatives of $l(\theta)$ with respect to specific subset components of the global parameter θ will be consequently denoted by $l_\lambda = \partial l(\theta)/\partial\lambda$, $l_{\psi\lambda} = \partial^2 l(\theta)/(\partial\psi\partial\lambda^{\mathrm{T}})$, $l_{\psi\lambda\lambda} = \partial l_{\psi\lambda}/\partial\lambda$ and so forth. We shall also express the null expectations of these likelihood quantities and of their products as $\nu_\lambda = E_\theta(l_\lambda)$, $\nu_{\psi\lambda\lambda} = E_\theta(l_{\psi\lambda\lambda})$, $\nu_{\lambda,\psi\lambda} = E_\theta(l_\lambda l_{\psi\lambda})$, etc. In dealing instead with a generic function $g = g(\theta)$, the notation $g_{/\psi}$ will usually be adopted for its first-order partial derivative with respect to ψ. However, whenever the argument of differentiation is clear enough, we shall prefer the simpler symbols $g', g'', \ldots,$ to indicate the derivatives of g. Moreover, in the sequel, expressions such as $j_{\psi\psi}$ or $i_{\lambda\lambda}$ will be helpful for denoting blocks of $j(\theta)$ and $i(\theta)$ related to the coordinates of the corresponding suitable component of θ. In order to index blocks of the inverse matrices $j(\theta)^{-1}$ and $i(\theta)^{-1}$, superscripts like those in $j^{\psi\psi}$ and $i^{\lambda\lambda}$ shall be used.

Hypothesis testing and interval estimation for the unknown parameter of interest derived from the likelihood function are tasks generally performed by taking advantage of first-order asymptotic results regarding fundamental likelihood quantities (Reid, 2003). Specifically, the log-likelihood ratio statistic, which takes the form

$$W = W(\psi) = 2\big\{l(\hat{\theta}) - l(\hat{\theta}_\psi)\big\},$$

and its two other asymptotically equivalent versions, i.e. the score statistic

$$W_u = W_u(\psi) = l_\psi\big(\hat{\theta}_\psi\big)^{\mathrm{T}} i^{\psi\psi}\big(\hat{\theta}_\psi\big) l_\psi\big(\hat{\theta}_\psi\big)$$

and the Wald statistic

$$W_e = W_e(\psi) = \big(\hat{\psi} - \psi\big)^{\mathrm{T}} i^{\psi\psi}\big(\hat{\theta}_\psi\big)^{-1}\big(\hat{\psi} - \psi\big), \tag{1.1}$$

all have $\chi^2_{k_0}$ asymptotic null distribution under standard regularity conditions on the model $\mathcal{F}$ (see, e.g., Pace and Salvan, 1997, Section 3.4). In addition, when ψ is scalar, one is allowed to rely on the corresponding signed versions of the combinants:

$$\begin{aligned} Z = Z(\psi) &= \operatorname{sgn}\big(\hat{\psi} - \psi\big)\sqrt{W}, \\ Z_u = Z_u(\psi) &= l_\psi\big(\hat{\theta}_\psi\big)\sqrt{i^{\psi\psi}\big(\hat{\theta}_\psi\big)}, \\ Z_e = Z_e(\psi) &= \big(\hat{\psi} - \psi\big)\big/\sqrt{i^{\psi\psi}\big(\hat{\theta}_\psi\big)}, \end{aligned} \tag{1.2}$$

that for large n have null $N(0,1)$ distribution (Skovgaard, 1989).

In what follows, special attention will be given to the Wald statistic, since inferential procedures based on it are perhaps the ones most affected by the presence of bias in the ML estimation of θ (see Section 1.2.3).

1.2 The Wald statistic

1.2.1 Null distribution

Asymptotic results about the null distribution of the combinants introduced in Section 1.1 stem from limit theorems of probability theory whose validity depends on the amount of information available for the study. If we define $i_1(\theta) = \lim_{n\to+\infty} i(\theta)/n$ to be the average limit information in a sample of n independent observations, the central limit theorem gives

$$\frac{l_\theta}{\sqrt{n}} \xrightarrow{d} N_k\big(0, i_1(\theta)\big). \tag{1.3}$$

Such large-sample null distribution of the score represents also the starting point for obtaining that of the Wald statistic. In the rest of the section, this derivation will be briefly reviewed for the simpler case where the interest is on the full parameter θ, following what reported in Section 3.4.1 of Pace and Salvan (1997). Expressions (1.1) and (1.2), when $k = k_0$, can be thus rewritten as

$$\begin{aligned} W_e(\theta) &= (\hat\theta - \theta)^{\mathrm{T}} i(\theta)(\hat\theta - \theta), \\ Z_e(\theta) &= (\hat\theta - \theta)\sqrt{i(\theta)}. \end{aligned} \tag{1.4}$$

In order to learn how comparable results shall be obtained for the pivots $W_e(\psi)$ and $Z_e(\psi)$ in the presence of a nuisance component λ, the reading of Section 9.3 in Cox and Hinkley (1974) is highly recommended.

Let us first consider the case $k = 1$, to further simplify the present exposition. Assume that $\mathcal{F}$ is a regular model and $\hat\theta$ is a consistent solution of the likelihood equation $l_\theta = 0$. Then, the score function admits the Taylor expansion about the value θ

$$0 = l_\theta(\hat\theta) = l_\theta + (\hat\theta - \theta) l_{\theta\theta} + O_p(1).$$

Recalling that $l_{\theta\theta} = -j(\theta)$, a simple manipulation of the previous expression gives

$$\frac{l_\theta}{\sqrt{n}} = \sqrt{n}(\hat\theta - \theta)\frac{j(\theta)}{n} + O_p\big(n^{-1/2}\big).$$

Since, by a law of large numbers, $j(\theta)/n \xrightarrow{p} i_1(\theta)$, it is possible to rearrange the terms and write

$$\sqrt{n}(\hat\theta - \theta) = i_1(\theta)^{-1}\frac{l_\theta}{\sqrt{n}} + O_p\big(n^{-1/2}\big).$$

Now, exploiting the well known properties of the normal distribution and the limiting result (1.3) about the score in the one-parameter case, it is evident that

$$\sqrt{n}(\hat\theta - \theta) \xrightarrow{d} N\big(0, i_1(\theta)^{-1}\big), \tag{1.5}$$

and this immediately leads to the null asymptotic $N(0,1)$ distribution of the pivotal quantity $Z_e(\theta)$ in (1.4). Furthermore, thanks to the consistency of the ML estimator, we have $ni_1(\hat{\theta}) \xrightarrow{p} i(\theta)$, causing also the combinant $(\hat{\theta} - \theta)\sqrt{i(\hat{\theta})}$ to be approximately standard normally distributed when n tends to infinity.

If $k > 1$, the result reported in (1.5) holds with a k-dimensional normal limit distribution. This implies that $W_e(\theta)$ has asymptotic χ^2_k null distribution, as well as the same statistic which estimates the null Fisher information by $i(\hat{\theta})$. We finally stress that, when the partition $\theta = (\psi, \lambda)$ is adopted, by similar arguments concerning the null asymptotic properties of the constrained ML estimator (Pace and Salvan, 1997, p. 145) it is possible to use $i^{\psi\psi}(\hat{\theta})$ in place of $i^{\psi\psi}(\hat{\theta}_\psi)$ in both formulae (1.1) and (1.2), still mantaining the large-sample distribution of those original combinants.

1.2.2 Advantages and disadvantages

As anticipated in Section 1.1, the three versions of the likelihood ratio combinant are asymptotically equivalent. More formally, provided that the asymptotic order of magnitude of the absolute error in an rth-order approximation to a random variable is $O_p(n^{-r/2})$, their equivalence in probability holds to the first order and the reciprocal relationships

$$W = W_u + O_p\big(n^{-1/2}\big),$$
$$W = W_e + O_p\big(n^{-1/2}\big)$$

apply (Pace and Salvan, 1997, Section 3.4.1). Such pivotal quantities are generally used to build confidence regions for or to test hypotheses about the parameter of interest ψ. The signed statistics $Z(\psi_0)$, $Z_u(\psi_0)$ and $Z_e(\psi_0)$ are particularly useful when $k_0 = 1$ and $H_0\colon \psi = \psi_0$ is tested against the one-sided alternative H_1: $\psi > \psi_0$ or H_1: $\psi < \psi_0$.

Forasmuch as accuracy in the approximation to the null distribution of the three test statistics is the same, the choice between them has to be made on the basis of different criteria. With respect to its competitors, straightforward interpretation and extremely simple implementation are definitely the main strengths of the Wald pivot. Indeed, when testing $H_0\colon\psi = \psi_0$, the latter's formulation consists in a direct comparison between the estimated value and the hypothetical one, taking also the error of such estimation into account. Furthermore, since the block $i^{\psi\psi}$ of the information matrix in formulae (1.1) and (1.2) can be evaluated at the global estimate $\hat{\theta}$ without affecting the asymptotic properties of the combinants (see Section 1.2.1), $W_e(\psi_0)$ and $Z_e(\psi_0)$ require only the unconstrained model fitting, which represents common practice for any basic statistical software. On the contrary, both the likelihood ratio and the score tests need the ML estimate under H_0 to be computed. Finally, particularly in regression settings where one is interested in the construction of confidence intervals for scalar coefficients, the inversion of Z_e is particularly convenient.

All the reasons listed above justify somehow the extensive use of the Wald statistic in general applications, despite the drawbacks associated with it. One of them is undoubtedly the lack of interest-respecting parametrization invariance: inferential conclusions based on W_e or Z_e depend upon the way the collection of probability distributions $\mathcal{F}$ is indexed. Among the consequences, we have that the observed significance level of the Wald test can be different when derived for testing $H_0\colon \tau = \tau_0 = \tau(\psi_0)$ instead of $H_0\colon \psi = \psi_0 = \psi(\tau_0)$, unless τ_0 is a linear transformation of ψ_0. Conversely, W, W_u and the corresponding signed versions are invariant. For a deeper discussion pertaining this matter, the reader can refer to Section 1.3 in Barndorff-Nielsen (1988) and Section 2.11 in Pace and Salvan (1997).

Another aspect which makes the Wald statistic less appealing for inference, especially with respect to the likelihood ratio pivot, lies in the fact that its expression does not account for the curvature of the log-likelihood function. Hence, confidence regions and tests based on W_e are reasonably accurate if $l(\hat{\theta}_\psi)$ is almost quadratic around its maximum, but those based on W are much more adequate if the log-likelihood has alternatively a pronounced asymmetrical shape, as commonly occurs when the sample size is small to moderate. In the most extreme cases, inverting the Wald statistic can even lead to confidence regions including values of ψ outside the parameter space.

Moreover, when one is concerned with testing a simple null hypothesis about the component of interest, adopting the score pivot may be preferable to avoid the computation of the global estimate $\hat{\psi}$, which happens to be quite demanding if the unrestricted model has a complex form or k_0 is significantly large. Nonetheless it is important noticing that the quantities l_ψ and $i^{\psi\psi}$ appearing in the formulation of W_u and Z_u still have to be obtained starting from the original complete likelihood of the model.

Here the focus was put on mainly practical advantages and disadvantages relating to the employment of the Wald combinant for making inference on the parameter ψ. In Chapter 2, we will suggest a way to improve the quality of such inferential conclusions by preserving at the same time those features which make W_e so suited for statistical applications. Of course, other grounds of comparison between the three pivots defined in Section 1.1 might have been considered: the non-null distribution under the alternative hypothesis, playing a crucial role in decision theory, is probably the most popular. A review of results concerning the power of those tests is outside the scope of this work, but relevant references on the topic certainly are Cox and Hinkley (1974), in particular Chapter 5 and Chapter 9 for derivation of both exact and asymptotic optimal properties, Section 3.5 in Pace and Salvan (1997) and Chapter 4 in Young and Smith (2005).

1.2.3 The effect of bias in hypothesis testing

It is well known (see, e.g, Kosmidis, 2014) that the bias of the ML estimator under standard regularity conditions can be expanded in decreasing powers of n as

$$E_\theta(\hat{\theta} - \theta) = \frac{b(\theta)}{n} + \frac{b_2(\theta)}{n^2} + \frac{b_3(\theta)}{n^3} + O(n^{-4}), \tag{1.6}$$

where all functions in the sequence $b(\theta), b_2(\theta), b_3(\theta), \dots,$ are of asymptotic order $O(1)$. Expression (1.6) clearly suggests that such bias is vanishing for $n \to +\infty$, thus the ML estimator is asymptotically unbiased.

However, unbiasedness of $\hat{\theta}$ is generally lost in finite samples and this presence of bias in the estimation of the unknown parameter can significantly affect the adequacy of ordinary statistical procedures. As an example, in Kosmidis (2014) it is illustrated how in Beta regressions a remarkable bias of the ML estimator for the dispersion parameter can lead to inaccurate Wald-type inference for the parameters of interest, even if the latter are estimated with sufficient precision. In this framework, and usually in generalized linear models, the downward-biased estimate of the dispersion parameter enters multiplicatively in the denominator of the Wald statistic, yielding to excessively narrow confidence intervals and anti-conservative tests for the regression coefficients.

More broadly speaking, the general expression of the pivot itself conveys the intuition that bias in point estimation has also consequences on Wald-based inferential conclusions. Consider for illustration a problem of hypothesis testing at an approximate significance level α when $k_0 = 1$ and $H_0 \colon \psi = \psi_0$. The Wald test statistic for the scalar parameter of interest in this case may be formulated as

$$Z_e(\psi_0) = \frac{\hat{\psi} - \psi_0}{\sqrt{i^{\psi\psi}(\hat{\theta})}} \tag{1.7}$$

and is also named z-statistic after its standard normal asymptotic distribution under the null hypothesis (see Section 1.2.1). In the same spirit, the corresponding test is typically referred to as the z-test. Note that under H_0 it is not assigned a specific value to the nuisance component, hence the hypothesis is said to be composite and might be equivalently expressed as $H_0 \colon \theta = \theta_0$ with $\theta_0 = (\psi_0, \lambda) \in \Theta_0 \subseteq \mathbb{R}^{k-1}$. In such occasions, the exact size of the z-test is defined as $\breve{\alpha} = \sup_{\theta \in \Theta_0} P_\theta(y \in \mathcal{Y}_R)$, where $\mathcal{Y}_R \subset \mathcal{Y}$ is the region of the sample space that leads to reject the null, depending on the alternative hypothesis and on the given value of α (Pace and Salvan, 1997, Section 3.5.3).

By looking at (1.7), it does not seem so illogical to speculate on the possibility that the farther the ML estimate $\hat{\theta}$ from θ_0, the farther the null moments of the Wald statistic from those of the $N(0,1)$ random variable. Indeed, the approximation to the distribution of $Z_e(\psi_0)$ can be particularly poor in small samples where the bias of the ML estimator is noticeable, as already pointed out. This, in its turn, may cause the exact size $\breve{\alpha}$ of the z-test to differ considerably from the nominal level α.

In order to alleviate such problems in inferences based on the approximate normality of $\hat{\theta}$, a rich stream of statistical literature touched upon in the Introduction has been devoted to propose useful techniques for reducing the bias of estimators. Once again, we refer to Kosmidis (2014) for a thorough review of such so-called implicit and explicit methods, which share the purpose of deriving a new estimator whose bias is asymptotically smaller than that of the original one. Specifically, when applied to improve the finite properties of $\hat{\theta}$, all the various procedures deliver an estimator with bias of asymptotic order $o(n^{-1})$. A quite natural way to conduct a more reliable test on the component of interest consists then in using the same statistic (1.7), but with $\hat{\theta}$ replaced by the corresponding bias-corrected estimate.

To show how this strategy can be effective, a simple simulation study may be performed. First, a dataset is generated starting from the covariates x_{i1} and x_{i2} ($i = 1, \ldots, 15$), independently drawn from the $N(1,1)$ distribution. Responses y_i are then simulated under the assumption $Y_i \sim \Gamma(\phi^{-1}, \vartheta_i)$, where $\vartheta_i = (\phi\mu_i)^{-1}$ and $E_\theta(Y_i) = \mu_i = \exp(\beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2})$, with $\theta = (\beta_1, \beta_2, \beta_3, \phi) = (1, 1, 2, 0.5)$. Notice that here ϕ is the nuisance parameter controlling the dispersion in the dependent variable, since $\mathrm{Var}_\theta(Y_i) = \phi\mu_i^2$. A Gamma regression model can now be fitted on the data by ML method, and standard deviations of the parameter estimates are obtained using the square root of the diagonal elements in the inverse Fisher information matrix. Table 1.1 shows these ML estimates with corresponding estimated standard errors and 0.95 Wald confidence intervals. In addition, a parametric bootstrap based on 5000 replicates

TABLE 1.1: ML fit of the Gamma regression model with log-link and Wald 0.95 confidence intervals for the parameters.

	Estimate	Estimated Standard Error	0.95 Confidence Interval
β_1	0.361	0.250	(-0.128, 0.851)
β_2	1.507	0.170	(1.174, 1.839)
β_3	1.859	0.165	(1.535, 2.183)
ϕ	0.223	0.079	(0.069, 0.377)

(Efron and Tibshirani, 1993, Section 6.5) is implemented to estimate the bias of the components of $\hat{\theta}$; such values result equal to -0.006, -0.006, -0.007 and -0.043 for $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ and $\hat{\phi}$, respectively. Following the argument expressed above, in order to check whether the Wald intervals are shorter than expected due to the significant downward bias of the dispersion parameter, we can compare their empirical coverages to those obtained by inversion of the Wald pivots which instead employ the bias-corrected ML estimates. Coverage estimation is performed through a study based on 5000 simulations, each using 5000 bootstrap replications to derive the bias. Results for confidence levels 0.90, 0.95 and 0.99 are reported in Table 1.2 and confirm that the use of estimates corrected via bootstrap helps the Wald-type intervals to approach the larger nominal coverage.

Table 1.2: Empirical coverages of individual confidence intervals based on the Wald statistic and on the Wald statistic which uses bias-corrected estimates of the model parameters.

	Wald			Wald with Bias-corrected Estimates		
	0.90	0.95	0.99	0.90	0.95	0.99
β_1	0.828	0.900	0.962	0.869	0.926	0.976
β_2	0.834	0.893	0.961	0.861	0.921	0.973
β_3	0.828	0.892	0.961	0.865	0.913	0.975

Different methods to refine the quality of first-order inference in finite samples have been subjects of much of the classical research. Whereas the approach adopted in Table 1.2 attempts to design more accurate tests by adjusting the ML estimator, such other techniques focus directly on the test statistic as a whole. In particular, the task has been addressed following two main avenues: the first consists in obtaining a new test statistic whose null distribution is closer to the limiting one, the second consists in obtaining a new reference distribution which better approximates the test statistic's exact null distribution.

The pioneer of the first strategy is Bartlett (1937), who introduced a special correction for the likelihood ratio statistic which was later generalized to regular problems by Lawley (1956). A similar methodology was employed by Cordeiro and Ferrari (1991) to derive Bartlett-type corrections for statistics other than the likelihood ratio, including score and Wald. Both Bartlett and Bartlett-type corrections are aimed at bringing the exact size of asymptotic tests closer to the nominal level, yet in the second case the adjustment is commonly a function of the unmodified statistic itself. A recent overview of this kind of corrections to the Wald test can be found in Section 3.4 of Cordeiro and Cribari-Neto (2014). Moreover, even the renowned t variable for testing hypotheses about the population mean was considered for a modification intended to account for skewed distributions of the data (Johnson, 1978).

The procedure based on the concept of prepivoting (Beran, 1987) perhaps combines the two possible solutions. Indeed, prepivoting is defined as the transformation of a statistic by the cumulative distribution function of its bootstrap null distribution. The prepivoted test, obtained comparing this new statistic to a suitable quantile of the $U(0, 1)$ distribution, has usually a smaller asymptotic order of error in level than the original one. Beran (1988) also showed that Bartlett's adjustment to likelihood ratio tests can be regarded as an analytical approximation to such prepivoted test. Whilst being originally thought for enhancing the accuracy of confidence regions, the approach based on bootstrap resampling was then reformulated by Hall and Martin (1988) so as to deal with several statistical problems under the same unifying theoretical framework. More recent developments of the topic concern the employment of weighted bootstrap iterations to make prepivoting more efficient. Theoretical and practical benefits of this modified procedure are well illustrated by Lee and Young (2003) and Young (2003).

The idea which focuses on deriving a more apposite reference distribution for some standard test statistic has principally leaned on the use of Edgeworth and saddlepoint approximations. Statistical applications of such techniques were discussed by, among others, Barndorff-Nielsen and Cox (1979) and Reid (1988). Furthermore, this general type of approach for improving hypothesis testing is particularly popular in the misspecified and composite likelihood literatures. Under those scenarios, the likelihood ratio statistic has been shown to have an unconventional limiting null distribution, corresponding with that of a linear combination of independent χ^2 random variables. Readers interested in this material should consult Kent (1982), Varin *et al.* (2011) and references therein.

In Chapter 2, a new attempt to improve the adequacy of first-order inference based on the Wald pivot will be presented. The proposed method, involving a correction to the usual z-statistic, belongs to the first class of techniques described above, but stems from the general idea of bias reduction discussed at the beginning of this section. Indeed, as will be later described, such adjustment may be conveniently obtained by exploiting the asymptotic bias expansion of the ML estimator.

1.3 Treatment of nuisance parameters

1.3.1 Introduction

More and more often nowadays, researchers are concerned about drawing inferential conclusions only about some aspects of the phenomenon under study, which are captured during the modelling phase by the partial component ψ of the global parameter. Whenever this is the case, to work with a likelihood function depending just on this component of interest seems advisable, especially if the configuration of the nuisance parameter is complex and no loss of information about ψ takes place. In the statistical theory, such valuable tool is called pseudo-likelihood, since it behaves in some respects as a genuine likelihood but may be not deduced from a density function. Under regularity conditions, pseudo-likelihoods usually share with $L(\theta)$ useful properties like, for instance, zero null expectation of the score function, approximate normality of maximum likelihood estimators and χ^2-asymptotically null distributed log-likelihood ratio statistics.

Marginal and conditional likelihoods (Pace and Salvan, 1997, Section 4.3) are classical examples of pseudo-likelihoods which are in fact proper likelihoods. Specifically, they derive from a statistical model defined as a reduction of the original one. As long as the order of information in this simplified model remains equal to $O(n)$, it can be shown that usual asymptotic results about likelihood quantities apply (Severini, 2000, Chapter 8). However, outside the class of exponential and group family models, these particular pseudo-likelihoods are either impossible or computationally cumbersome to obtain. This drawback makes then arise the need for a more general approach, described in the next section.

1.3.2 Profile likelihood

In parametric models, one simple idea to define a pseudo-likelihood function for the parameter of interest is to replace λ in the original likelihood expression with some consistent estimate. When this substitution is done with $\hat{\lambda}_\psi$, the ML estimate of the nuisance component for fixed ψ introduced in Section 1.1, the ensuing function is the profile likelihood $L_P(\psi) = L(\psi, \hat{\lambda}_\psi)$.

Although not a genuine likelihood, $L_P(\psi)$ has several interesting traits which can be taken advantage of in order to make inference about ψ. First of all, the maximum profile likelihood estimate computed by maximization of $L_P(\psi)$ coincides with the component relating to the parameter of interest of the overall ML estimate, i.e. $\hat{\psi}$. Furthermore, the profile log-likelihood ratio statistic is equal to the one built from $L(\theta)$ for testing hypotheses on ψ when λ is unknown. In mathematical notation, one can write

$$W_P = W_P(\psi) = 2\big\{l_P(\hat{\psi}) - l_P(\psi)\big\} = 2\big\{l(\hat{\psi},\hat{\lambda}) - l(\psi,\hat{\lambda}_\psi)\big\} = W,$$

where, as usual, $l_P(\psi) = \log L_P(\psi)$ and $l_P(\hat{\psi}) = \log L_P(\hat{\psi})$. Similarly, we have that $W_{uP} = W_u$ and $W_{eP} = W_e$. The same relationships obviously hold for the one-sided versions of the combinants, Z_P, Z_{uP} and Z_{eP}. Another relevant feature of the profile likelihood concerns the profile observed information. It is easy to show that (see, for example, Section 4.6 of Pace and Salvan, 1997)

$$j_P(\psi) = -\frac{\partial^2}{\partial\psi\partial\psi^{\mathrm{T}}} l\big(\hat{\theta}_\psi\big) = -\big[l_{\psi\psi}\big(\hat{\theta}_\psi\big) - l_{\psi\lambda}\big(\hat{\theta}_\psi\big)\big\{l_{\lambda\lambda}\big(\hat{\theta}_\psi\big)\big\}^{-1} l_{\lambda\psi}\big(\hat{\theta}_\psi\big)\big],$$

and therefore $j_P(\psi)^{-1} = j^{\psi\psi}(\hat{\theta}_\psi)$, the ψ-block in the inverse of the full observed information matrix evaluated at the restricted ML estimate. It is finally noteworthy that even $L_P(\psi)$ enjoys the property of invariance with respect to interest-preserving transformations (see Section 1.1).

The peculiarities of the profile likelihood just presented undoubtedly motivate its leading position among pseudo-likelihoods. Indeed, the standard practice for conducting statistical analyses when also characteristics of not immediate interest need to be accounted for is to base inference on $L_P(\psi)$. Nevertheless, the fact that this pseudo-likelihood is not directly derived from a density function does have some consequences. In general, $l_P(\psi)$ does not satisfy the Bartlett identities (Bartlett, 1953, Section 2): even in regular cases, for instance, the null expectation of the profile score function is not zero. More specifically, DiCiccio *et al.* (1996) proved the validity of the equation

$$E_\theta\big(l_{P/\psi}\big) = -\rho_\psi + O\big(n^{-1}\big),$$

where the dominant bias term ρ_ψ is of order $O(1)$ and takes the form

$$\rho_\psi = \big(i_{\psi\lambda} i_{\lambda\lambda}^{-1} \nu_{\lambda\lambda,\lambda} - \nu_{\psi\lambda,\lambda}\big) i_{\lambda\lambda}^{-1} - \frac{1}{2}\big(\nu_{\psi\lambda\lambda} - i_{\psi\lambda} i_{\lambda\lambda}^{-1} \nu_{\lambda\lambda\lambda}\big) i_{\lambda\lambda}^{-1}\,. \tag{1.8}$$

McCullagh and Tibshirani (1990) pointed out that, when the dimension of λ is large relative to the sample size, such bias may even critically affect the accuracy of ordinary asymptotic results. In the next sections, some of the adjustments to the profile likelihood proposed for reducing the order of its score bias shall be examined.

1.3.3 Adjusted profile likelihoods

The inferential issues associated with the use of $l_P(\psi)$ can be ascribed to the lack of knowledge about λ. In particular, acting as the nuisance component were known and equal to $\hat{\lambda}_\psi$ is not sensible if the data do not contain a sufficient amount of information about it. During the last decades, various modified forms of the profile likelihood have been developed with the intention of taking into consideration the uncertainty implied by the estimation of λ.

Loosely speaking, a typical expression for the logarithmic version of some adjusted profile likelihood $L_A(\psi)$ is simply

$$l_A(\psi) = \log L_A(\psi) = l_P(\psi) + A(\psi), \tag{1.9}$$

where the adjustment term $A(\psi)$ represents a suitable smooth function having derivatives of order $O_p(1)$. Several propositions have been put forward as plausible formulations of such term; despite having been obtained from different perspectives, all of them generally introduce a correction able to reduce the bias of the profile score. In fact, one can see that

$$E_\theta(A_{/\psi}) = \rho_\psi + O\big(n^{-1}\big).$$

Yet, within the usual asymptotic framework, this correction does not translates into enhanced formal properties of quantities related to $l_A(\psi)$: the log-likelihood ratio pivot has still a $\chi^2_{k_0}$ null approximate distribution to the first order and the rate of convergence of the corresponding adjusted ML estimator to the true parameter ψ remains of order $O_p\big(n^{-1/2}\big)$. Nonetheless, statistical procedures descending from adjusted profile likelihoods are typically more reliable than those from $L_P(\psi)$, especially when $k - k_0$ is large (see, e.g., DiCiccio and Stern, 1994). The most extreme situation where the number of nuisance parameters grows with the sample size deserves special attention and will be closely discussed in Section 1.4.

A prominent example of adjustment is surely the one proposed by Cox and Reid (1987). The major quality of their approximate conditional log-likelihood is the simplicity, as it only requires the computation of quantities delivered as output by standard numerical procedures for fitting the constrained model. Its expression is

$$l_{AC}(\psi) = l_P(\psi) - \frac{1}{2}\log\big|j_{\lambda\lambda}\big(\hat{\theta}_\psi\big)\big|,$$

so this function may be viewed as a sort of penalized profile log-likelihood, where the penalty serves to account for the knowledge about λ as the component of interest varies.

Unfortunately, the employment of this adjusted version of $l_P(\psi)$ is restricted to models where ψ and λ are orthogonal, meaning $i_{\psi\lambda} = 0$. Such a parametrization is definitely useful from a practical point of view, but exists for any value of ψ only when $k_0 = 1$. Furthermore, even if the parameter of interest is scalar, orthogonality between the components of θ can be quite hard to achieve (Pace and Salvan, 1997, Section 4.7). Lastly, another disadvantage connected with the use of $l_{AC}(\psi)$ lies in its lack of exact invariance under interest-respecting parametrizations.

As already emphasized, a variety of expressions for the modification term $A(\psi)$ is available in the literature in addition to that just described. References to these different proposals are Fraser and Reid (1988, 1989), McCullagh and Tibshirani (1990), DiCiccio and Stern (1993), Stern (1997) and Pace and Salvan (2006), to name but a few. This thesis will instead deal extensively with the modified profile likelihood, which represents in most respects the ideal pseudo-likelihood and is presented below.

1.3.4 Modified profile likelihood and its approximations

In 1983 Barndorff-Nielsen introduced a new method to reduce the score bias of $l_P(\psi)$. Further developments of such approach were then published in his later papers of 1988, 1994 and 1995. The modified profile log-likelihood is defined as

$$\begin{aligned} l_M(\psi) &= l_P(\psi) + M(\psi) \\ &= l_P(\psi) - \frac{1}{2}\log\left|j_{\lambda\lambda}(\hat\theta_\psi)\right| + \log D(\psi), \end{aligned} \tag{1.10}$$

where

$$D(\psi) = \left|\frac{\partial\hat\lambda_\psi}{\partial\hat\lambda}\right| = \frac{\left|j_{\lambda\lambda}(\hat\theta_\psi)\right|}{\left|l_{\lambda;\hat\lambda}(\hat\theta_\psi)\right|}. \tag{1.11}$$

The quantity $l_{\lambda;\hat\lambda}(\hat\theta_\psi) = \partial l(\hat\theta_\psi;\hat\theta,a)/(\partial\lambda\partial\hat\lambda^{\mathrm{T}})$ is called sample space derivative, because the log-likelihood is differentiated with regard to some ML estimate. Note that here a stands for ancillary statistic, either exact or approximate, in the meaning provided by Section 2.8 of Pace and Salvan (1997); therefore a has, at least approximately, a probability distribution independent of θ and $(\hat\theta, a)$ is minimal sufficient since summarizes all and only the relevant information in the data.

The reasons why $L_M(\psi) = \exp\{l_M(\psi)\}$ has a central role in the class of adjusted profile likelihoods are numerous. For instance, in contrast with Cox and Reid's modification, it is invariant under interest-preserving transformations and does not require to find an orthogonal partition of the overall parameter. Perhaps more importantly, it was originally conceived as an highly accurate approximation to proper likelihoods for

ψ, such as conditional or marginal ones, whenever they exist (Barndorff-Nielsen and Cox, 1994, Section 8.2). More favourable attributes of the modified profile likelihood are investigated in DiCiccio *et al.* (1996) and Severini (1998a).

On the other hand, applicability of the inferential tool developed by Barndorff-Nielsen is limited by the necessity of specifying some ancillary statistic a, so that the term $D(\psi)$ can be computed. This results straightforward in full exponential models, where the ML estimate is a sufficient statistic itself, and in transformation models, where the configuration ancillary is available. However, outside these particular families one usually has to resort to approximate solutions.

In the case of orthogonality between ψ and λ, the relation $\hat{\lambda}_\psi = \hat{\lambda} + O_p\big(n^{-1}\big)$ holds (Pace and Salvan, 1997, Section 4.7) and consequently $D(\psi) = 1 + O_p\big(n^{-1}\big)$, so the term $\log D(\psi)$ in (1.10) can be in some sense neglected, like in Section 9.5.2 of Severini (2000). This entails that $l_M(\psi)$ and $l_{AC}(\psi)$ are asymptotically equivalent to the second order, as one can also write $l_M(\psi) = l_{AC}(\psi) + \log D(\psi)$. Based on what previously said, though, the approximation of the modified profile likelihood via the function proposed by Cox and Reid comes at the price of exact invariance.

When parameters are not orthogonal, the calculation of $D(\psi)$ cannot be avoided and thus approximating somehow the sample space derivative in (1.11) is generally needed. To this aim, covariances, empirical covariances or tangential directions to an approximate ancillary may be used. All these methods return invariant adjustments of $L_P(\psi)$ which differ from the original modified profile likelihood by the asymptotic order $O_p\big(n^{-1}\big)$ when the component of interest stays in the moderate deviation region, i.e. $\psi = \hat{\psi} + O_p\big(n^{-1/2}\big)$ (Severini, 2000, Section 9.5).

The first technique was initially suggested by Skovgaard (1996) to approximate the modification of directed log-likelihood ratio tests defined by Barndorff-Nielsen (1986), but its specific application to the modified profile likelihood dates back to Severini (1998b). In broad terms, such approach allows the approximation of sample space derivatives by covariances between particular components of the score function. According to this general principle, $l_{\lambda;\hat{\lambda}}(\theta)$ can be considered asymptotically equivalent to the quantity $I_{\lambda\lambda}(\theta;\hat{\theta})$, where

$$I_{\lambda\lambda}(\theta;\hat{\theta}) = E_{\hat{\theta}}\big\{l_\lambda(\theta) l_\lambda(\hat{\theta})^{\mathrm{T}}\big\}.$$

Substitution of $I_{\lambda\lambda}(\hat{\theta}_\psi;\hat{\theta})$ for the sample space derivative in (1.11) and simple manipulation of formula (1.10) yield to Severini's approximate version of the modified profile log-likelihood:

$$\begin{aligned} l_{\widetilde{M}}(\psi) &= l_P(\psi) + \widetilde{M}(\psi) \\ &= l_P(\psi) + \frac{1}{2}\log\left|j_{\lambda\lambda}(\hat{\theta}_\psi)\right| - \log\left|I_{\lambda\lambda}(\hat{\theta}_\psi;\hat{\theta})\right|. \end{aligned} \tag{1.12}$$

The function $L_{\widetilde{M}}(\psi) = \exp\{l_{\widetilde{M}}(\psi)\}$ is probably the most popular approximation to the modified profile likelihood. In fact, it has proved to be a solid statistical tool for drawing precise inferences on the parameter of interest in models not necessarily belonging to exponential or group families. Severini's proposal is also the main object under analysis in Chapter 3, where a procedure to handle even quite complex sampling and/or modelling assumptions will be illustrated.

Here, we have limited ourselves to give explicit formulation of $l_{\widetilde{M}}(\psi)$, yet of course other expressions of approximate modified profile log-likelihood exist; see, e.g., Barndorff-Nielsen (1994) for the case $k_0 = 1$. Moreover, we remark that one detailed exposition concerning the three approximation methods mentioned above is Section 9.5 in Severini (2000).

1.4 Reducing the bias of the profile score for independent clustered data

1.4.1 Introduction

The first ones who characterized the now famous incidental parameters problem were Neyman and Scott (1948). Such a name refers in particular to the scalar components of λ which increase with the sample size. In these situations, regularity of the model is not met and usual first-order approximations for inferences on the so-called structural parameter ψ fail. Among others, also Portnoy (1988), Pierce and Peters (1992) and Lancaster (2000) dealt with this topic.

The same issue is commonly found when units in the sample are organized in many distinct clusters and the dimension of the nuisance component is assumed to be dependent on the total number of groups; this last part of the chapter is indeed dedicated to models of this type, known in the econometric literature under the name of fixed effects models (see also Section 3.1).

Before proceeding, it is worth highlighting the fact that an ordinary asymptotic setting like that studied so far, where approximation errors are expressed only in terms of powers of the total sample size n, does not enable to mathematically formalize the inferential superiority of the modified profile likelihood with respect to $L_P(\psi)$. Thus, in the sequel, a two-index asymptotic setting shall be introduced for deeper comprehension of the theoretical results about the refined properties of $L_M(\psi)$ and its approximations contained in Section 1.4.3.

1.4.2 Notation and setup

Let us consider parametric statistical models for independent and clustered data taking form

$$Y_{it} \sim p_{Y_{it}}(y_{it}; \psi, \lambda_i), \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T_i. \tag{1.13}$$

The total sample size is $n = \sum_{i=1}^{N} T_i$ and the nuisance component is $\lambda = (\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^N$. Notice also that definition (1.13) is appropriate to include in the specification regression models, where one can write $p_{Y_{it}}(y_{it}; \psi, \lambda_i) = p_{Y_{it}}(y_{it}; \psi, \lambda_i, x_{it})$, where x_{it} are known covariates. For the sake of simplicity but without loss of generality, assume that $\psi \in \mathbb{R}$ and $T_i = T$ for every i, with $n = TN$. In particular, the second requirement of balanced groups might be relaxed in such a way as to consider situations where $T_i = K_i T$, for $A \leq K_i \leq B$ and with A, B positive and finite numbers (Sartori, 2003).

The two-index asymptotic setting, named $(T \times N)$-asymptotics, permits both the number of clusters, N, and the cluster sample size, T, to tend to infinity. Standard asymptotic theory in fact applies when the number of incidental parameters N is fixed, but if instead N increases and T does not, Neyman & Scott problems are likely to be observed, since $N = O(n)$. Specifically, the latter circumstance can be reproduced in the context of $(T \times N)$-asymptotics simply by letting N go to infinity much faster than T.

The log-likelihood for model (1.13) may conveniently be expressed by

$$l(\theta) = \sum_{i=1}^{N} l^i(\theta) = \sum_{i=1}^{N} l^i(\psi, \lambda_i),$$

where

$$l^i(\psi, \lambda_i) = \sum_{t=1}^{T} \log p_{Y_{it}}(y_{it}; \psi, \lambda_i)$$

is the log-likelihood function related to the ith cluster, assumed to be regular in the usual sense. Separability of $l(\theta)$ with respect to incidental parameters is a direct consequence of independence among clusters. Similarly, as $\hat{\lambda}_\psi$ comes from the solution of N

independent likelihood equations, the profile log-likelihood admits to be written as

$$l_P(\psi) = l\big(\psi, \hat{\lambda}_\psi\big) = \sum_{i=1}^{N} l^i\big(\psi, \hat{\lambda}_{i\psi}\big) = \sum_{i=1}^{N} l_P^i(\psi). \tag{1.14}$$

After a standard expansion of the profile score for the ith cluster $l^i_{P/\psi}$ (see Sartori, 2003, and references therein for further computational details), it is fairly simple to show that

$$E_\theta\big(l^i_{P/\psi}\big) = -\rho^i_\psi + O\big(T^{-1}\big), \tag{1.15}$$

where ρ^i_ψ is of asymptotic order $O(1)$ and has the same structure as the quantity in (1.8). Now, it may be immediately checked that the major impediment to adequate ML inferences in the presence of clustered data with group-specific nuisance parameters has to do with the accumulation of the score bias across clusters. More explicitly, by combining equations (1.14) and (1.15), one shall conclude with little difficulty that in this case the leading term in the expected value of the profile score $l_{P/\psi}$ equals $-\sum_{i=1}^{N} \rho^i_\psi$ and hence is, asymptotically, of order $O(N)$.

1.4.3 Results in the two-index asymptotic setting

The present part revisits the various $(T \times N)$-asymptotic properties of the profile and modified profile likelihood, derived by Sartori (2003) under the model hypotheses stated in Section 1.4.2. It is important to specify that such results were in fact obtained referring to any general adjusted profile log-likelihood $l_A(\psi)$ as defined in (1.9), with adjustment term satisfying two key requirements. Specifically,

$$A(\psi) = \sum_{i=1}^{N} A^i(\psi), \tag{1.16}$$

where $A^i(\psi)$ is a suitable smooth function, having derivatives of order $O_p(1)$ whose null expected value is such that

$$E_\theta\big(A^i_{/\psi}\big) = \rho^i_\psi + O\big(T^{-1}\big). \tag{1.17}$$

In plain words, the modification for each group needs to eliminate the leading term of the ith profile score bias in order to be effective.

Even though both adjustments $M(\psi)$ and $\widetilde{M}(\psi)$ in (1.10) and (1.12), repectively, may be shown to enjoy properties (1.16) and (1.17), for clarity purpose the following theoretical results are presented with reference to the modified profile log-likelihood

$l_M(\psi)$ only. We stress here that all of them can actually be extended to Severini's approximation $l_{\widetilde{M}}(\psi)$.

The first notable findings pertain to the $(T \times N)$-asymptotic distribution of the profile and modified profile score statistics

$$W_{uP} = l^2_{P/\psi}(\psi)/j_P(\psi),$$
$$W_{uM} = l^2_{M/\psi}(\psi)/j_M(\psi),$$

where $j_M(\psi) = -\partial^2 l_M(\psi)/(\partial\psi\partial\psi^{\mathrm{T}})$ is the modified profile observed information. In particular, it was shown that W_{uP} has the usual χ^2_1 asymptotic distribution as long as $N = o(T)$, meaning if the number of clusters increases at a slower rate than the sample size in every cluster. On the other hand, W_{uM} is asymptotically χ^2_1-distributed when $N = o(T^3)$, meaning if the number of clusters grows slower than the cube of the cluster size. Therefore, the condition to be satisfied by $L_M(\psi)$ is weaker than that to be satisfied by $L_P(\psi)$. To put it simply, whenever N increases faster than T, but not faster than T^3, W_{uM} has the ordinary approximate distribution, while this cannot be guaranteed for W_{uP}. Moreover, even in situations when both pivots are χ^2-distributed, W_{uM} may be proved to have a smaller upper bound of the approximation error. Conclusions do not change if expected informations $i_P(\psi)$ and $i_M(\psi)$ are used to compute the score statistics in place of $j_P(\psi)$ and $j_M(\psi)$, respectively.

Formal acknowledgment of the better consistency properties of the modified profile likelihood estimator in the two-index scenario is certainly another important result. Under the $(T \times N)$-asymptotics, denoting by $\hat{\psi}_M$ the maximizer of $l_M(\psi)$, both $\hat{\psi}$ and $\hat{\psi}_M$ are consistent in so far as N and T go to infinity, no matter what the relative behaviour of the indexes is. Nevertheless, the rate of convergence to the true parameter value changes according to the mutual relationship between N and T. Indeed, by expanding the likelihood equations associated with $l_P(\psi)$ and $l_M(\psi)$ around ψ, it is not difficult to see that

$$\hat{\psi} = \psi + O_p\big(n^{-1/2}\big)$$

if $N = o(T)$, with $\hat{\psi} = \psi + O_p\big(T^{-1}\big)$ otherwise, and

$$\hat{\psi}_M = \psi + O_p\big(n^{-1/2}\big)$$

if $N = o(T^3)$, with $\hat{\psi}_M = \psi + O_p\big(T^{-2}\big)$ otherwise. Hence, whether the number of groups increases faster than the cluster size, the ML estimator $\hat{\psi}$ may converge to ψ at a slower rate with respect to $\hat{\psi}_M$.

Interestingly, the three popular likelihood-based combinants, both those deriving from $L_P(\psi)$ and those deriving from $L_M(\psi)$, are first-order asymptotically equivalent even in the two-index setting. One may write the profile and modified profile log-likelihood ratio statistics as

$$\begin{aligned} W_P &= 2\big\{l_P(\hat{\psi}) - l_P(\psi)\big\}, \\ W_M &= 2\big\{l_M(\hat{\psi}_M) - l_M(\psi)\big\}, \end{aligned}$$

and the profile and modified profile Wald statistics as

$$\begin{aligned} W_{eP} &= (\hat{\psi} - \psi)^2 j_P(\psi), \\ W_{eM} &= (\hat{\psi}_M - \psi)^2 j_M(\psi), \end{aligned}$$

where, as usual, the observed information might be replaced by its expectation. Formally, equivalence to the first order in the $(T \times N)$-asymptotics for the statistics related to the profile likelihood is expressed, when $N = o(T)$, by equations

$$\begin{aligned} W_{eP} &= W_{uP}\big\{1 + O_p\big(n^{-1/2}\big)\big\}, \\ W_P &= W_{eP}\big\{1 + O_p\big(n^{-1/2}\big)\big\}, \end{aligned}$$

otherwise the same hold with relative approximation errors of order $O_p\big(T^{-1}\big)$. It is perhaps helpful underlining that in these cases we speak of relative error because the order actually considered is the one of the absolute error divided by the quantity to be approximated. Similarly, it can be found that

$$\begin{aligned} W_{eM} &= W_{uM}\big\{1 + O_p\big(n^{-1/2}\big)\big\}, \\ W_M &= W_{eM}\big\{1 + O_p\big(n^{-1/2}\big)\big\}, \end{aligned}$$

if $N = o(T^3)$, otherwise equivalence is achieved to relative order $O_p\big(T^{-1}\big)$. Roughly speaking, when one of the three pivots has the χ^2 asymptotic distribution, the other two are equivalent to it with a relative error of order $O_p\big(n^{-1/2}\big)$ for both $l_P(\psi)$ and $l_M(\psi)$, as can be shown to happen in standard asymptotics. The crucial point here is that the sufficient condition to obtain such distribution for the quantities based on the modified profile likelihood is less stringent than the one applying to the profile likelihood. Ultimately, to conclude this survey of asymptotic results, it is correct to highlight that the same $(T \times N)$-properties of $L_P(\psi)$ and $L_M(\psi)$ can be derived by considering formulations of the score and Wald pivots which involve the information evaluated at the appropriate estimator of ψ.

In terminating the part devoted to the treatment of incidental parameters, an approach alternative to the profile likelihood and its modifications within the same frequentist paradigm of inference is worth quoting. Particularly, we refer to the integrated likelihood of Severini (2007), where elimination of the nuisance parameters occurs via integration with respect to some carefully selected density of λ. Such function was proved to be asymptotically equivalent to the modified profile likelihood in general frameworks and to benefit of analogue $(T \times N)$-properties in the two-index context for clustered data just examined (De Bin *et al.*, 2015).

Chapter 2

Adjusted z-tests

2.1 Introduction

In this chapter we propose a method to adjust the z-statistic for a scalar parameter of interest, like the one defined in formula (1.7). Specifically, this is done having in mind the goal of enhancing the quality of Wald-type inference, which is particularly unsatisfactory when carried out on samples of small-to-moderate size, without undermining the merits connected with the outstanding ease of its implementation (see Section 1.2.2).

The reader will beyond doubt notice that, in what follows, the case $k = 1$ is initially treated separately. Indeed, at least in the first place, the modification of the Wald pivot suggested for this special circumstance differs from our general proposition, mainly because of the higher complexity of the problem when $k > 1$. Nevertheless, we find that part of the thesis particularly meaningful for its function of motivating the basic idea behind the methodology used. Such idea essentially consists in raising the extent of testing accuracy by correcting the z-statistic in order to make its null moments closer to those of the reference standard normal random variable.

Thus, the next section will deal with the rather uncommon yet interesting single-parameter setting, which also gives the chance to explicitly derive the relevant properties of the adjusted Wald combinant and compare them to those of its standard version. Later on, the general location adjustment for the case $k \geq 1$ will be presented and its theoretical features in the situation of a scalar global parameter investigated. Lastly, closing considerations will be anticipated by a special mention to the importance of improving z-tests in the context of generalized linear models, illustrating also the performance of the location adjusted z-statistic through some simulation results.

2.2 Motivation of the study: one-parameter models

2.2.1 Notation and setup

For a random sample $y = (y_1, \dots, y_n)$ of independent observations, assume a very elementary parametric statistical model defined as

$$Y_i \sim p_{Y_i}(y_i; \theta), \qquad \theta \in \Theta \subseteq \mathrm{I\!R}, \qquad i = 1, \dots, n. \tag{2.1}$$

In the presence of a unique parameter θ, let us adopt the convenient power notation (Pace and Salvan, 1997, p. 344) to indicate products of log-likelihood derivatives and their expected values. For instance, we will write

$$\begin{aligned} l_r = l_r(\theta) &= \frac{\partial^r l(\theta)}{\partial \theta^r}, \\ \nu_{r_1,\dots,r_m} = \nu_{r_1,\dots,r_m}(\theta) &= E_\theta\big(l_{r_1} \cdots l_{r_m}\big), \qquad m \geq 1. \end{aligned}$$

If one is interested in making inference on the scalar parameter and needs to verify the simple null hypothesis $H_0\colon \theta = \theta_0$, for some value $\theta_0 \in \mathrm{I\!R}$, the most widespread choice is to conduct a z-test. As already seen, such a test relies on the popular Wald z-statistic, which in this case may have the two asymptotically equivalent formulations:

$$\mathring{T} = \mathring{T}(\theta_0) = (\hat{\theta} - \theta_0)\mathring{\nu}_{1,1}^{1/2}, \tag{2.2}$$

$$\widehat{T} = \widehat{T}(\theta_0) = (\hat{\theta} - \theta_0)\hat{\nu}_{1,1}^{1/2}, \tag{2.3}$$

where $\mathring{\nu}_{1,1} = i(\theta_0)$ and $\hat{\nu}_{1,1} = i(\hat{\theta})$ are the expected information under model (2.1) evaluated at the hypothesized value and at the ML estimate, respectively. Notice, in particular, that formula (1.7) coincides with $\widehat{T}$ when $k = k_0 = 1$, yet here the subset of the parameter space compatible with the null Θ_0 has only one element and the exact size of the z-test equals then $\breve{\alpha} = P_{\theta_0}(y \in \mathcal{Y}_R)$, where $\mathcal{Y}_R$ is the rejection region introduced in Section 1.2.3.

When all usual regularity conditions are satisfied by model (2.1), according to what shown in Section 1.2.1, both $\mathring{T}$ and $\widehat{T}$ are approximately $N(0, 1)$-distributed under H_0. For this reason, a z-test generally consists in comparing to the quantiles of the standard normal distribution the observed value of the Wald statistic used. However, it is well known that such limiting result is reliable only if the sample size n is large enough. When this is not the case, inferential conclusions drawn from z-tests can be misleading.

For the purpose of fixing the aforementioned problem, in the following we will derive the null mean and variance of $\mathring{T}$ and $\widehat{T}$; such quantities, as will be seen, play in fact a primary role in the correction of the Wald test statistic applicable when $k = 1$.

2.2.2 Cumulants of the Wald statistics

Cumulants of the pivotal quantities (2.2) and (2.3) are tightly connected with those of the ML estimator. In particular, $\mathring{T}$, which evaluates the standard error of the ML estimator at the fixed hypothetical value θ_0, simply consists of a linear transformation of $\hat{\theta}$. The statistical literature hosts a rich variety of results regarding the theoretical features of the ML estimator. Just to cite a few, in 1977 Shenton and Bowman, expanding their previous work of 1963, derived the first four moments of the distribution of $\hat{\theta}$ to orders $O\big(n^{-2}\big), O\big(n^{-3}\big), O\big(n^{-3}\big)$ and $O\big(n^{-4}\big)$, respectively; later, Peers and Iqbal (1985) obtained also asymptotic expansions for the cumulants of $\hat{\theta}$ till the fourth order, in the case of vector parameter.

In order to perform the adjustment in this simple setting, only the first two cumulants of $\mathring{T}$ and $\widehat{T}$ will be needed. For computing those, good starting points are the expansions of $E_\theta(\hat{\theta}-\theta)^r$ for $r = 1, \dots, 4$, where the order of asymptotic approximation can be chosen according to the result

$$E_\theta(\hat{\theta} - \theta)^r = \begin{cases} O\big(n^{-r/2}\big) & \text{if } r \text{ is even} \\ O\big(n^{-(r+1)/2}\big) & \text{if } r \text{ is odd,} \end{cases} \tag{2.4}$$

which are implied by (9.30) and (9.36) in Sections 9.2 and 9.3, respectively, of Pace and Salvan (1997). Such expansions for the scalar case were derived using the procedure described in Section 9.4 of Pace and Salvan (1997) and take the form:

$$E_\theta(\hat{\theta} - \theta) = \frac{\nu_3 + 2\nu_{1,2}}{2\nu_{1,1}^2} \overset{\bullet\bullet}{+} O\big(n^{-2}\big), \tag{2.5}$$

$$\begin{aligned} E_\theta(\hat{\theta} - \theta)^2 = {} & \frac{1}{\nu_{1,1}} \overset{\bullet\bullet}{+} \frac{\nu_4 - \nu_{1,1}^2 + 3\nu_{1,3} + 3\nu_{2,2} + 2\nu_{1,1,2}}{\nu_{1,1}^3} \\ & + \frac{11\nu_3^2 + 36\nu_3\nu_{1,2} + 24\nu_{1,2}^2}{4\nu_{1,1}^4} \overset{\bullet\bullet}{+} O\big(n^{-3}\big), \end{aligned} \tag{2.6}$$

$$E_\theta(\hat{\theta} - \theta)^3 = \frac{7\nu_3 + 12\nu_{1,2}}{2\nu_{1,1}^3} \overset{\bullet\bullet}{+} O\big(n^{-3}\big), \tag{2.7}$$

$$E_\theta(\hat{\theta} - \theta)^4 = \frac{3}{\nu_{1,1}^2} \overset{\bullet\bullet}{+} O\big(n^{-3}\big), \tag{2.8}$$

where for ease of reading, here and elsewhere in this chapter, we write $\overset{\bullet\bullet}{+}$ every time the terms which follow are asymptotically smaller for an order $O_p\bigl(n^{-1}\bigr)$ than are the preceding terms in the formula. Likewise, the symbols $\overset{\bullet}{+}$ and $\overset{\bullet\bullet\bullet}{+}$ will be used to indicate a fall of order $O_p\bigl(n^{-1/2}\bigr)$ and of $O_p\bigl(n^{-3/2}\bigr)$, respectively, adopting the same convenient notation as in Chapter 9 of Pace and Salvan (1997). Note that simplification of expressions (2.5)–(2.8) was achieved by exploiting Bartlett's identities and well-known relations between cumulants and central moments of a distribution (see, e.g, Pace and Salvan, 1997, p. 83). Another useful formula to bear in mind when doing this kind of calculations is the one reported in Stern (2006), which directly links the mean of a product of log-likelihood derivatives to its asymptotic order. Namely, $\nu_{r_1,\dots,r_m} = O\bigl(n^{m-\lfloor (m_1+1)/2 \rfloor}\bigr)$, where m_1 is the number of elements in the subscript partition equal to 1 such that $0 \leq m_1 \leq m$ and $\lfloor x \rfloor$ denotes the integer part of x.

Let us now start with $\mathring{T}$, defined in (2.2). Such combinant was already studied by Pfanzagl (1973), who obtained a two-term Edgeworth expansion (Hall, 1992, Chapter 2) for the null distribution under fulfilment of mild regularity conditions. Using expansions (2.5) and (2.6), it is immediate to derive approximations to the first two moments of the statistic under H_0. Specifically, we can write

$$E_{\theta_0}\bigl(\mathring{T}\bigr) = \frac{\mathring{\nu}_3 + 2\mathring{\nu}_{1,2}}{2\mathring{\nu}_{1,1}^{3/2}} \overset{\bullet\bullet}{+} O\bigl(n^{-3/2}\bigr) = \mathring{E}_1\bigl(\mathring{T}\bigr) \overset{\bullet\bullet}{+} O\bigl(n^{-3/2}\bigr), \tag{2.9}$$

$$E_{\theta_0}\bigl(\mathring{T}^2\bigr) = \frac{2\mathring{\nu}_{1,1,2}}{\mathring{\nu}_{1,1}^2} \overset{\bullet\bullet}{+} \frac{\mathring{\nu}_4 + 3\mathring{\nu}_{1,3} + 3\mathring{\nu}_{2,2}}{\mathring{\nu}_{1,1}^2} + \frac{11\mathring{\nu}_3^2 + 36\mathring{\nu}_3\mathring{\nu}_{1,2} + 24\mathring{\nu}_{1,2}^2}{4\mathring{\nu}_{1,1}^3} \overset{\bullet\bullet}{+} O\bigl(n^{-2}\bigr), \tag{2.10}$$

stressing that $\mathring{E}_1\bigl(\mathring{T}\bigr) = O\bigl(n^{-1/2}\bigr)$. Thus, the null variance is equal to:

$$\begin{aligned} \mathrm{Var}_{\theta_0}\bigl(\mathring{T}\bigr) &= \frac{6\mathring{\nu}_{2,2} - \mathring{\nu}_{1,1,1,1}}{3\mathring{\nu}_{1,1}^2} \overset{\bullet\bullet}{+} \frac{2\mathring{\nu}_4 + 5\mathring{\nu}_{1,3}}{3\mathring{\nu}_{1,1}^2} + \frac{7\mathring{\nu}_{1,2}^2 + 14\mathring{\nu}_{1,2}\mathring{\nu}_{1,1,1} + 5\mathring{\nu}_{1,1,1}^2}{2\mathring{\nu}_{1,1}^3} \overset{\bullet\bullet}{+} O\bigl(n^{-2}\bigr) \\ &= \mathring{V}_1\bigl(\mathring{T}\bigr) \overset{\bullet\bullet}{+} \mathring{V}_2\bigl(\mathring{T}\bigr) \overset{\bullet\bullet}{+} O\bigl(n^{-2}\bigr), \end{aligned} \tag{2.11}$$

where $\mathring{V}_1\bigl(\mathring{T}\bigr)$ and $\mathring{V}_2\bigl(\mathring{T}\bigr)$ are the quantities of order $O(1)$ and $O\bigl(n^{-1}\bigr)$, respectively, in the expansion. Since the first-order asymptotic variance of $\mathring{T}$ was shown to be 1 in Section 1.2.1, the last expression might look a bit odd. However, it is possible to see that there is no contradiction between the two results, because in fact the only term in $\mathring{V}_1\bigl(\mathring{T}\bigr)$ which is $O(1)$ equals 1. Indeed, using the fourth Bartlett's identity, we have

$$\begin{aligned} \mathring{V}_1\bigl(\mathring{T}\bigr) &= -\frac{2\mathring{\nu}_4 + 8\mathring{\nu}_{1,3} + 12\mathring{\nu}_{1,1,2} + 3\mathring{\nu}_{1,1,1,1}}{3\mathring{\nu}_{1,1}^2} \\ &= 1 \overset{\bullet\bullet}{+} O\bigl(n^{-1}\bigr), \end{aligned}$$

as the validity of relations $\mathring{\nu}_{1,1,2} = -\mathring{\nu}_{1,1}^2 + O(n)$ and $\mathring{\nu}_{1,1,1,1} = 3\mathring{\nu}_{1,1}^2 + O(n)$ can easily be checked.

Computing the same cumulants for the z-statistic $\widehat{T}$ in (2.3) demands a bit more effort. The complication in doing so is given by the fact that the Fisher information is evaluated at the ML estimate, and hence needs to be expanded itself about the null value θ_0. To the best of our knowledge, no publication has dealt with this specific matter up to now. The statistic equivalent to $\widehat{T}$ in the multiparameter setting was considered by Hayakawa and Puri in 1985. For the case $k = 1$, Taniguchi (1991) obtained the Edgeworth expansion of the χ_1^2 distribution of $\widehat{T}^2$ for a wide class of stochastic processes, while dos Santos and Cordeiro (1999) focused on the Bartlett-type correction of $\widehat{T}^2$ in exponential family models. Moreover, Stafford (1992) derived the first four cumulants of the z-statistic formulated yet with the observed information in place of its expected value. Unfortunately, despite the undeniable relevance to different extents of all these works to our problem, we could not find a manner to take advantage of the results therein; thus, the necessary steps to calculate the moments of $\widehat{T}$ shall be detailed below.

The procedure begins with the application of the stochastic Taylor formula (Pace and Salvan, 1997, Section 9.3.1) to $\hat{\nu}_{1,1}$. In particular, recalling that $\nu_{1,1}$ is of order $O(n)$ as well as its derivatives, it is fairly simple to get the following asymptotic expansion around θ_0:

$$\begin{aligned}\hat{\nu}_{1,1} &= \mathring{\nu}_{1,1}\left[1 \overset{\bullet}{+} \left\{-(\hat{\theta}-\theta_0)\frac{\mathring{\nu}_2+\mathring{\nu}_{1,2}}{\mathring{\nu}_{1,1}}\right\} \overset{\bullet}{+} \left\{-(\hat{\theta}-\theta_0)^2\frac{\mathring{\nu}_4+2\mathring{\nu}_{1,3}+\mathring{\nu}_{2,2}+\mathring{\nu}_{1,1,2}}{2\mathring{\nu}_{1,1}}\right\}\right] \overset{\bullet}{+} O_p\big(n^{-1/2}\big)\\ &= \mathring{\nu}_{1,1}\big(1 \overset{\bullet}{+} A_1 \overset{\bullet}{+} A_2\big) \overset{\bullet}{+} O_p\big(n^{-1/2}\big),\end{aligned}$$

where $A_1 = O_p\big(n^{-1/2}\big)$ and $A_2 = O_p\big(n^{-1}\big)$. Then, we have that

$$\begin{aligned}\hat{\nu}_{1,1}^{1/2} &= \mathring{\nu}_{1,1}^{1/2}\big(1 \overset{\bullet}{+} A_1 \overset{\bullet}{+} A_2\big)^{1/2} \overset{\bullet}{+} O_p\big(n^{-1}\big)\\ &= \mathring{\nu}_{1,1}^{1/2}\left(1 \overset{\bullet}{+} \frac{A_1}{2} \overset{\bullet}{+} \frac{A_2}{2} - \frac{A_1^2}{8}\right) \overset{\bullet}{+} O_p\big(n^{-1}\big),\end{aligned}$$

where the second equality results from the popular Maclaurin series $(1+x)^{1/2} = 1 + x/2 - x^2/8 + o(x^2)$, with $x = A_1 + A_2$. Lastly, the asymptotic expansion for $\widehat{T}$ takes the form

$$\begin{aligned}\widehat{T} = (\hat{\theta}-\theta_0)\hat{\nu}_{1,1}^{1/2} &= (\hat{\theta}-\theta_0)\mathring{\nu}_{1,1}^{1/2}\left(1 \overset{\bullet}{+} \frac{A_1}{2} \overset{\bullet}{+} \frac{A_2}{2} - \frac{A_1^2}{8}\right) \overset{\bullet}{+} O_p\big(n^{-3/2}\big)\\ = \mathring{T}\left(1 \overset{\bullet}{+} \frac{A_1}{2} \overset{\bullet}{+} \frac{A_2}{2} - \frac{A_1^2}{8}\right) \overset{\bullet}{+} O_p\big(n^{-3/2}\big).\end{aligned}$$

At this stage, expansions for the moments of order one and two of the distribution under H_0 of $\widehat{T}$ may be found with no great difficulty, employing formulae (2.6)–(2.10). Specifically,

$$E_{\theta_0}\big(\widehat{T}\big) = \frac{\mathring{\nu}_{1,2}}{2\mathring{\nu}_{1,1}^{3/2}} \overset{\bullet\bullet}{+} O\big(n^{-3/2}\big) = \mathring{E}_1\big(\widehat{T}\big) \overset{\bullet\bullet}{+} O\big(n^{-3/2}\big), \tag{2.12}$$

$$E_{\theta_0}\big(\widehat{T}^2\big) = \frac{3\mathring{\nu}_{2,2} + \mathring{\nu}_{1,1,2}}{2\mathring{\nu}_{1,1}^2} \overset{\bullet\bullet}{+} \left\{-\left(\frac{\mathring{\nu}_4}{2\mathring{\nu}_{1,1}^2} + \frac{3\mathring{\nu}_3^2 + 2\mathring{\nu}_3\mathring{\nu}_{1,2}}{4\mathring{\nu}_{1,1}^3}\right)\right\} \overset{\bullet\bullet}{+} O\big(n^{-2}\big),$$

where, in parallel with what seen in equation (2.9), $\mathring{E}_1\big(\widehat{T}\big) = O\big(n^{-1/2}\big)$. This allows the expansion of the second null cumulant of (2.3) to be written as

$$\begin{aligned}\mathrm{Var}_{\theta_0}\big(\widehat{T}\big) &= \frac{15\mathring{\nu}_{2,2} - \mathring{\nu}_{1,1,1,1}}{12\mathring{\nu}_{1,1}^2} \overset{\bullet\bullet}{+} \left\{-\left(\frac{7\mathring{\nu}_4 + 4\mathring{\nu}_{1,3}}{12\mathring{\nu}_{1,1}^2} + \frac{22\mathring{\nu}_{1,2}^2 + 16\mathring{\nu}_{1,2}\mathring{\nu}_{1,1,1} + 3\mathring{\nu}_{1,1,1}^2}{4\mathring{\nu}_{1,1}^3}\right)\right\} \overset{\bullet\bullet}{+} O\big(n^{-2}\big)\\ &= \mathring{V}_1\big(\widehat{T}\big) \overset{\bullet\bullet}{+} \mathring{V}_2\big(\widehat{T}\big) \overset{\bullet\bullet}{+} O\big(n^{-2}\big),\end{aligned} \tag{2.13}$$

being, as usual, $\mathring{V}_1\big(\widehat{T}\big) = O(1)$ and $\mathring{V}_2\big(\widehat{T}\big) = O\big(n^{-1}\big)$. Furthermore, along the line of reasoning used earlier for $\mathring{T}$, one can also prove that $\mathring{V}_1\big(\widehat{T}\big) = 1 \overset{\bullet\bullet}{+} O\big(n^{-1}\big)$.

It is probably noteworthy that expressions for the asymptotic approximations to the null cumulants of $\mathring{T}$ and $\widehat{T}$ can be remarkably simplified if one wishes to exclusively refer the results to exponential families with canonical parameter $\theta \in \Theta$ (Pace and Salvan, 1997, p. 176). Indeed, as in this framework log-likelihood derivatives of order higher than 1 do not depend on the data, we have that $l_r = \nu_r$ for every $r \geq 2$. This implies that formulae (2.9) and (2.11) for $\mathring{T}$ reduce to

$$E_{\theta_0}\big(\mathring{T}\big) = \frac{\mathring{\nu}_3}{2\mathring{\nu}_{1,1}^{3/2}} \overset{\bullet\bullet}{+} O\big(n^{-3/2}\big), \tag{2.14}$$

$$\mathrm{Var}_{\theta_0}\big(\mathring{T}\big) = 1 \overset{\bullet\bullet}{+} \frac{\mathring{\nu}_4}{\mathring{\nu}_{1,1}^2} + \frac{5\mathring{\nu}_3^2}{2\mathring{\nu}_{1,1}^3} \overset{\bullet\bullet}{+} O\big(n^{-2}\big), \tag{2.15}$$

while (2.12) and (2.13) for $\widehat{T}$ become

$$E_{\theta_0}\big(\widehat{T}\big) = O\big(n^{-3/2}\big), \tag{2.16}$$

$$\mathrm{Var}_{\theta_0}\big(\widehat{T}\big) = 1 \overset{\bullet\bullet}{+} \left\{-\left(\frac{\mathring{\nu}_4}{2\mathring{\nu}_{1,1}^2} + \frac{3\mathring{\nu}_3^2}{4\mathring{\nu}_{1,1}^3}\right)\right\} \overset{\bullet\bullet}{+} O\big(n^{-2}\big). \tag{2.17}$$

Therefore, in canonical exponential family models evaluating the expected information at the ML estimate instead of the true parameter value has the appreciable consequence of centering the null distribution of the Wald combinant closer to 0.

2.2.3 Location and scale correction of the Wald statistics

What has been obtained in the last section will now be helpful for adjusting $\mathring{T}$ and $\widehat{T}$ in such a way as to get new pivots whose finite-sample null distribution agrees better with that of a standard normal random variable.

One possible strategy to pursue this objective is imitating the system of mean and variance correction adopted in DiCiccio and Stern (1994) to construct more accurate asymptotic combinants based on the signed root of the likelihood ratio test for a scalar parameter of interest ψ. The same methodology was later employed by Stern (2006), who considered statistics derived from the general objective function of an M-estimator within a certain statistical class.

Consequently, by reference to expansions (2.9), (2.11), (2.12) and (2.13) for the null cumulants of the unmodified Wald pivotal quantities (2.2) and (2.3) introduced in Section 2.2.1, the location-scale adjusted z-statistics in the single-parameter case may be defined as

$$\mathring{T}^{(ls)} = \mathring{T}^{(ls)}(\theta_0) = \frac{\mathring{T} - \mathring{E}_1\big(\mathring{T}\big)}{\big\{\mathring{V}_1\big(\mathring{T}\big)+\mathring{V}_2\big(\mathring{T}\big)\big\}^{1/2}}, \tag{2.18}$$

$$\widehat{T}^{(ls)} = \widehat{T}^{(ls)}(\theta_0) = \frac{\widehat{T} - \mathring{E}_1\big(\widehat{T}\big)}{\big\{\mathring{V}_1\big(\widehat{T}\big)+\mathring{V}_2\big(\widehat{T}\big)\big\}^{1/2}}, \tag{2.19}$$

given that $\mathring{V}_1\big(\mathring{T}\big)+\mathring{V}_2\big(\mathring{T}\big) > 0$ and $\mathring{V}_1\big(\widehat{T}\big)+\mathring{V}_2\big(\widehat{T}\big) > 0$, respectively. Whenever one of such requirements is not complied with for some particular pair (θ_0, n), only the correction in mean is performed instead. Therefore, just in these situations, we shall rely on the pivot with simpler form $\mathring{T}^{(l)} = \mathring{T} - \mathring{E}_1\big(\mathring{T}\big)$ or $\widehat{T}^{(l)} = \widehat{T} - \mathring{E}_1\big(\widehat{T}\big)$, respectively.

Now, it is not too involving to prove that the mean and variance of both the proposed combinants resemble more closely those of the reference standard normal distribution when the null hypothesis is true. To start, let us compare the first two cumulants of $\mathring{T}$ and $\mathring{T}^{(ls)}$. From results of the previous part, we have learned that for the standard z-statistic those quantities can be expressed by

$$\begin{aligned} E_{\theta_0}\big(\mathring{T}\big) &= O\big(n^{-1/2}\big), \\ \mathrm{Var}_{\theta_0}\big(\mathring{T}\big) &= \mathring{V}_1\big(\mathring{T}\big) \overset{\bullet\bullet}{+} \mathring{V}_2\big(\mathring{T}\big) \overset{\bullet\bullet}{+} O\big(n^{-2}\big) \\ &= 1 \overset{\bullet\bullet}{+} O\big(n^{-1}\big). \end{aligned}$$

With the purpose to derive similar expressions for the corresponding location-scale adjusted pivot formulated in (2.18), it is useful to write:

$$\begin{aligned}\mathring{T}^{(ls)} &= \big\{\mathring{T} - \mathring{E}_1\big(\mathring{T}\big)\big\}\big\{\mathring{V}_1\big(\mathring{T}\big)+\mathring{V}_2\big(\mathring{T}\big)\big\}^{-1/2}\\ &= \big\{\mathring{T} - \mathring{E}_1\big(\mathring{T}\big)\big\}\mathring{V}_1\big(\mathring{T}\big)^{-1/2}\left\{1 \overset{\bullet\bullet}{+} \frac{\mathring{V}_2\big(\mathring{T}\big)}{\mathring{V}_1\big(\mathring{T}\big)}\right\}^{-1/2}\\ &= \big\{\mathring{T} - \mathring{E}_1\big(\mathring{T}\big)\big\}\mathring{V}_1\big(\mathring{T}\big)^{-1/2}\left[1 \overset{\bullet\bullet}{+} \left\{ - \frac{\mathring{V}_2\big(\mathring{T}\big)}{2\mathring{V}_1\big(\mathring{T}\big)}\right\} \overset{\bullet\bullet}{+} O\big(n^{-2}\big)\right],\end{aligned}$$

where the last equality sign applies because $(1+x)^{-1/2} = 1 - x/2 - 3x^2/8 + o(x^2)$, with $x = \mathring{V}_2\big(\mathring{T}\big)/\mathring{V}_1\big(\mathring{T}\big) = O\big(n^{-1}\big)$. Denoting by $\mathring{v}(\mathring{T})$ the ratio $\mathring{V}_2\big(\mathring{T}\big)/\mathring{V}_1\big(\mathring{T}\big)$ and proceeding with the calculations, one finally obtains

$$\mathring{T}^{(ls)} = \mathring{V}_1\big(\mathring{T}\big)^{-1/2}\left[\mathring{T} \overset{\bullet}{+} \big\{ - \mathring{E}_1\big(\mathring{T}\big)\big\} \overset{\bullet}{+} \left\{ - \mathring{T}\frac{\mathring{v}\big(\mathring{T}\big)}{2}\right\}\right] \overset{\bullet}{+} O_p\big(n^{-3/2}\big).$$

Asymptotic expansions for the null expected value and variance of the location-scale adjusted z-statistic $\mathring{T}^{(ls)}$ are then:

$$\begin{aligned}E_{\theta_0}\big(\mathring{T}^{(ls)}\big) &= \mathring{V}_1\big(\mathring{T}\big)^{-1/2}\left\{E_{\theta_0}\big(\mathring{T}\big) - \mathring{E}_1\big(\mathring{T}\big) - E_{\theta_0}\big(\mathring{T}\big)\frac{\mathring{v}\big(\mathring{T}\big)}{2}\right\} + O\big(n^{-3/2}\big) = O\big(n^{-3/2}\big),\\ \mathrm{Var}_{\theta_0}\big(\mathring{T}^{(ls)}\big) &= \mathring{V}_1\big(\mathring{T}\big)^{-1}\left\{E_{\theta_0}\big(\mathring{T}^2\big)+\mathring{E}_1\big(\mathring{T}\big)^2-E_{\theta_0}\big(\mathring{T}^2\big)\mathring{v}\big(\mathring{T}\big)-2E_{\theta_0}\big(\mathring{T}\big)\mathring{E}_1\big(\mathring{T}\big)\right\}+O\big(n^{-3/2}\big)\\ &= \mathring{V}_1\big(\mathring{T}\big)^{-1}\left[\left\{\mathring{V}_1\big(\mathring{T}\big)+\mathring{V}_2\big(\mathring{T}\big)+\mathring{E}_1\big(\mathring{T}\big)^2\right\}\left\{1-\mathring{v}\big(\mathring{T}\big)\right\}-\mathring{E}_1\big(\mathring{T}\big)^2\right] + O\big(n^{-3/2}\big)\\ &= 1 \overset{\bullet\bullet\bullet}{+} O\big(n^{-3/2}\big).\end{aligned}$$

Moreover, provided the fact that when $H_0\colon \theta = \theta_0$ is true the relations

$$\begin{aligned}E_{\theta_0}\big(\widehat{T}\big) &= O\big(n^{-1/2}\big),\\ \mathrm{Var}_{\theta_0}\big(\widehat{T}\big) &= 1 \overset{\bullet\bullet}{+} O\big(n^{-1}\big)\end{aligned}$$

are valid, by essentially following the steps just reviewed in reference to $\mathring{T}$, it can be shown that the same expansions are valid for the cumulants of $\widehat{T}^{(ls)}$ reported in (2.19):

$$\begin{aligned}E_{\theta_0}\big(\widehat{T}^{(ls)}\big) &= O\big(n^{-3/2}\big),\\ \mathrm{Var}_{\theta_0}\big(\widehat{T}^{(ls)}\big) &= 1 \overset{\bullet\bullet\bullet}{+} O\big(n^{-3/2}\big).\end{aligned}$$

As pointed out by Stafford (1992), the adequacy of the normal approximation to the exact null distributions of competing combinants can be assessed by contrasting their cumulants with the corresponding values for the $N(0,1)$ random variable. In our case, the leading terms in the expansions for the mean and variance of all the pivots are equal to 0 and 1, respectively; the comparison must hence regard the remaining non-zero terms, which represent departure from normality. As the asymptotic orders of such remainders are smaller for $\mathring{T}^{(ls)}$ and $\widehat{T}^{(ls)}$, in principle one would expect these corrected z-statistics to provide a better tool for inference on small-to-moderate-sized samples with respect to $\mathring{T}$ and $\widehat{T}$. Next, we will analyze the behaviour of such various pivotal quantities in some specific single-parameter settings, so that to evaluate the foundation of this conjecture.

2.2.4 Special modeling frameworks

Exponential model

Let us assume that independent observations in the sample $y = (y_1, \ldots, y_n)$ are drawn from the exponential distribution defined by

$$Y_i \sim Exp\big(e^{\theta}\big), \qquad \theta \in \mathrm{I\!R}, \qquad i = 1, \ldots, n, \tag{2.20}$$

where $E_\theta(Y_i) = \mu_i = e^{-\theta} > 0$. The log-likelihood function for θ is simply $l(\theta) = n\theta - n\bar{y}e^{\theta}$, where $\bar{y} = \sum_{i=1}^{n} y_i/n$ is the sample mean. From this quantity, it is immediate to derive the score $l_1 = n - n\bar{y}e^{\theta}$ and the ML estimate $\hat{\theta} = -\log\bar{y}$, as well as the expected information, which here does not depend on the parameter. Indeed, we can write $\nu_{1,1} = \mathring{\nu}_{1,1} = \hat{\nu}_{1,1} = n$. As a consequence, according to formulations (2.2) and (2.3), in order to test the hypothesis $H_0\colon \theta = \theta_0$ we can use the z-statistic

$$\mathring{T} = \widehat{T} = -\sqrt{n}(\log\bar{y} + \theta_0).$$

Now, by employing formulae (2.9) and (2.11) to derive expressions for $\mathring{E}_1\big(\mathring{T}\big)$, $\mathring{V}_1\big(\mathring{T}\big)$ and $\mathring{V}_2\big(\mathring{T}\big)$, the corresponding location-scale adjusted z-statistic can be calculated. Observe that in this case, as $\mathring{T}$ and $\widehat{T}$ coincide, the same pivot results when one uses instead definitions in (2.12) and (2.13) of the quantities $\mathring{E}_1\big(\widehat{T}\big)$, $\mathring{V}_1\big(\widehat{T}\big)$ and $\mathring{V}_2\big(\widehat{T}\big)$. In formal notation, we get

$$\mathring{T}^{(ls)} = \widehat{T}^{(ls)} = -\frac{\sqrt{n}(\log\bar{y} + \theta_0) + (2\sqrt{n})^{-1}}{\big(1 + \frac{1}{2n}\big)^{1/2}}.$$

The great simplicity of model (2.20) allows to compute the exact distributions of the two versions of the Wald z-statistic and compare them with the standard normal. Indeed, the only thing we need to know is that $\overline{Y} = \sum_{i=1}^{n} Y_i/n \sim \Gamma(n, ne^{\theta_0})$ under the null hypothesis. Then $\mathring{T}$ and $\mathring{T}^{(ls)}$ are just transformations of this random variable, whose null density may be found with ease.

In Figure 2.1, it is possible to appreciate the effectiveness of the location-scale adjustment in this framework: the cumulative distribution function (CDF) of $\mathring{T}^{(ls)} = \widehat{T}^{(ls)}$ is closer to that of the $N(0,1)$ than the CDF of the unmodified z-statistic. Moreover, such discrepancy remains quite visible when the sample increases in size. Note also that these plots can be referred to every value $\theta_0 \in \mathbb{R}$, as the null probability density functions of the combinants do not depend on the true parameter.

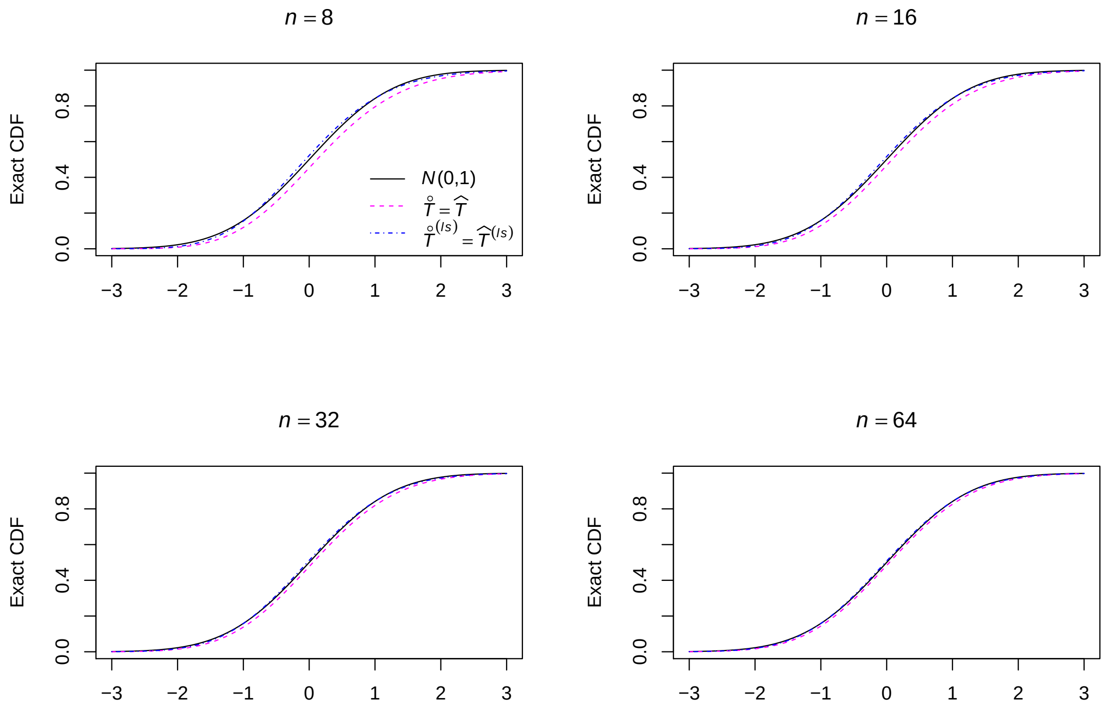

FIGURE 2.1: Comparison under the exponential model of the null CDFs of $\mathring{T} = \widehat{T}$ and $\mathring{T}^{(ls)} = \widehat{T}^{(ls)}$ to that of the $N(0,1)$, for any $\theta_0 \in \mathbb{R}$ and for various sample sizes n.

Under these assumptions, one may immediately see that the signed version of the score statistic is

$$Z_u = Z_u(\theta_0) = \sqrt{n}\big(1 - \bar{y}e^{\theta_0}\big).$$

Since its exact distribution follows directly from that of $\overline{Y}$ too, the performance of our adjusted Wald-type pivot may also be assessed with regard to this other popular likelihood-based combinant. Figure 2.2, specifically, displays such comparison: even in these pictures, the normal approximation looks more appropriate when used for the null CDF of the location-scale adjusted z-statistic than for that of Z_u, especially for smaller n.

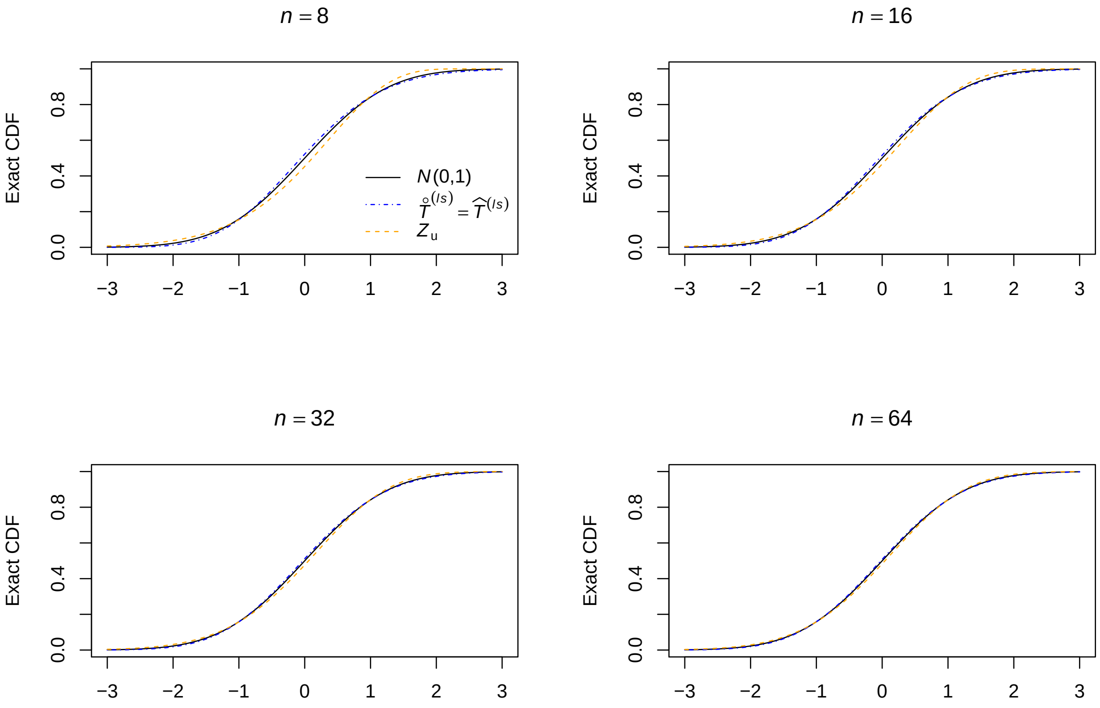

FIGURE 2.2: Comparison under the exponential model of the null CDFs of $\mathring{T} = \widehat{T}$ and Z_u to that of the $N(0,1)$, for any $\theta_0 \in \mathrm{I\!R}$ and for various sample sizes n.

Poisson model

The one-parameter Poisson log-linear model for independent units $y_1, \dots, y_n$ may be specified as

$$Y_i \sim Pois\big(e^{\theta}\big), \qquad \theta \in \mathrm{I\!R}, \qquad i = 1, \dots, n, \tag{2.21}$$

with $E_\theta(Y_i) = \mu_i = e^{\theta} > 0$. In this case, the log-likelihood and score functions can be written as $l(\theta) = n\bar{y}\theta - ne^{\theta}$ and $l_1 = n\bar{y} - ne^{\theta}$, respectively. The ML estimate is then equal to $\hat{\theta} = \log \bar{y}$, while the Fisher information is $\nu_{1,1} = ne^{\theta}$. Hence, the two versions (2.2) and (2.3) of the Wald test statistic for $H_0 \colon \theta = \theta_0$ now differ and take the forms

$$\mathring{T} = \sqrt{n}e^{\theta_0/2}(\log \bar{y} - \theta_0),$$
$$\widehat{T} = \sqrt{n\bar{y}}(\log \bar{y} - \theta_0).$$

At this point, it is convenient to recognize that model (2.21) belongs to a canonical exponential family, as the logarithmic function was chosen for connecting the mean of the ith Poisson random variable, μ_i, to the parameter θ, i.e. $\log \mu_i = \theta$ $(i = 1, \dots, n)$. Such consideration makes possible the employment of the simplified expressions (2.14)–(2.17) to obtain the modifications of $\mathring{T}$ and $\widehat{T}$. Specifically, we get the following location-scale adjusted z-statistics generally described in (2.18) and (2.19):

$$\mathring{T}^{(ls)} = \frac{\sqrt{n}e^{\theta_0/2}(\log \bar{y} - \theta_0) - (2\sqrt{n})^{-1}e^{-\theta_0/2}}{\left(1 + \frac{3e^{-\theta_0}}{2n}\right)^{1/2}},$$
$$\widehat{T}^{(ls)} = \frac{\sqrt{n\bar{y}}(\log \bar{y} - \theta_0)}{\left(1 - \frac{e^{-\theta_0}}{4n}\right)^{1/2}}.$$

Recalling that, by the previous assumptions, $n\overline{Y}$ is a Poisson random variable with expectation equal to ne^{θ}, we can calculate once again the exact distribution of the four pivots when H_0 is true for checking whether the quality of the normal approximation changes in the different cases. In doing so, some precautions need to be taken. First of all, it is important to notice that in this setting the probability of observing an infinite ML estimate is positive for any value of the parameter θ; in particular, $\hat{\theta} = -\infty$ when all the units in the sample equal 0. By looking at the expressions of the various z-statistics, it is not difficult to see that in such situations we can write $\mathring{T} = \mathring{T}^{(ls)} = -\infty$ and $\widehat{T} = \widehat{T}^{(ls)} = 0$, due to the well-known results

$$\lim_{x\to 0} \log x = -\infty \qquad \text{and} \qquad \lim_{x\to 0} x^{1/2} \log x = 0. \tag{2.22}$$

Furthermore, in the computation of the distribution of $\widehat{T}^{(ls)}$, one must pay attention to the possibility that the bracketed quantity in the denominator is not strictly positive. When this occurs, according to what defined in Section 2.2.3, under such canonical exponential model we shall have $\widehat{T}^{(ls)} = \widehat{T}^{(l)} = \widehat{T}$ for every $\theta_0 \in \mathbb{R}$.

Both the discreteness of the problem and the dependence on θ_0 of the null distributions of the combinants suggest to analyze their behaviour by means of the pictures in Figure 2.3. Here, for each of the competitors, the exact coverage of the confidence interval obtained by inversion of the z-statistic for testing H_0 at level $\alpha = 0.05$ versus the alternative $H_1\colon \theta \neq \theta_0$ is plotted against the values of θ_0. In all panels, the theoretical coverage probability 0.95 is indicated by the horizontal red line to facilitate interpretation. By looking at the various plots, a first comment to be made concerns perhaps the discrepancy in coverage recorded for lower values of θ_0 between the pairs $\mathring{T}$, $\mathring{T}^{(ls)}$ and $\widehat{T}$, $\widehat{T}^{(ls)}$. This must be ascribed to the two distinct values the couples take when the ML estimate is not finite. Such an event, more likely when the true parameter is small, leads indeed to different conclusions of the test according to the statistic used: H_0 is rejected if one employs $\mathring{T}$ or $\mathring{T}^{(ls)}$, whereas it is accepted otherwise. Based on the pictures, the adoption of $\widehat{T}$ or $\widehat{T}^{(ls)}$ appears generally more advisable, as it results in better coverage properties even for larger values of θ_0.

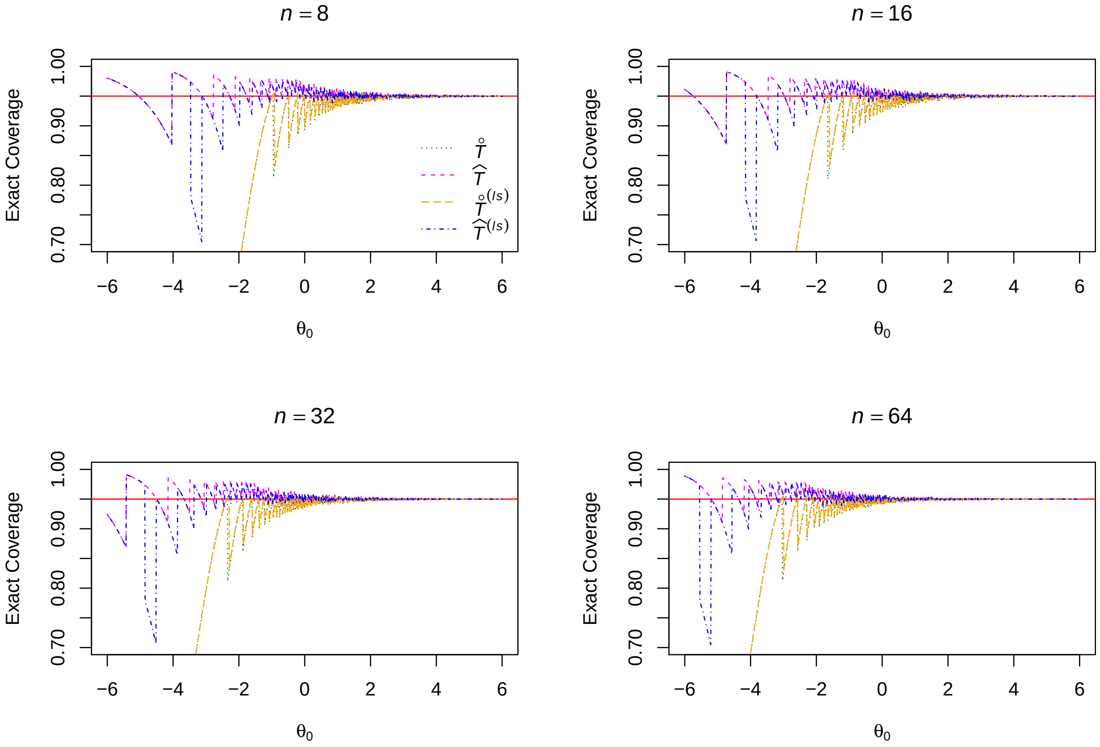

FIGURE 2.3: Exact coverage probabilities under the Poisson model for the two-sided interval at confidence level 0.95 based on pivots $\mathring{T}$, $\widehat{T}$, $\mathring{T}^{(ls)}$ and $\widehat{T}^{(ls)}$. Values are shown as a function of $\theta_0 \in \mathbb{R}$ and for various sample sizes n.

Turning now to consider the main object of our study, Figure 2.3 tells us that improvements generated by the location-scale correction of $\mathring{T}$ and $\widehat{T}$ are surely not as unquestionable as in the exponential case. More in detail, the adjustment of $\mathring{T}$ seems somewhat helpful for alleviating the excessive liberality of the corresponding test, while the use of $\widehat{T}^{(ls)}$ commonly generates lower exact coverage probabilities with respect to $\widehat{T}$. One can rightly argue that such probabilities are clearly not closer to the nominal level for lower values of the true parameter. The reason of this visible drop in coverage is in fact that, for specific combinations of θ_0 and n, the denominator of $\widehat{T}^{(ls)}$ approaches 0 and the whole test statistic becomes very large in absolute value, bringing about the rejection of H_0. Nevertheless, this inconvenient behaviour is observed for a range of θ_0 which shifts to more and more negative parameter values as n grows. In the remaining region, especially around $\theta_0 = 0$, the confidence interval based on $\widehat{T}^{(ls)}$ appears instead to be at least as accurate as that based on its classical counterpart.

In case of independent Poisson-distributed random variables, the signed versions of the score and log-likelihood ratio statistics are expressed by

$$Z_u = \sqrt{n} e^{-\theta_0/2} \big(\bar{y} - e^{\theta_0} \big),$$
$$Z = Z(\theta_0) = \operatorname{sign}(\log \bar{y} - \theta_0) \sqrt{2n \big\{ \bar{y}(\log \bar{y} - \theta_0) - \big(\bar{y} - e^{\theta_0} \big) \big\}},$$

respectively. Since both formulations correspond to simple functions of $\bar{y}$ and under model (2.21) we have $n\overline{Y} \sim Pois(ne^{\theta_0})$ when H_0 is true, also exact coverage probabilities of the two-tailed confidence intervals resulting by inversion of Z_u and Z may be checked for better evaluating the performance of our suggested modification to the Wald pivot. This is possible in Figure 2.4, where such coverages are directly contrasted with that based on $\widehat{T}^{(ls)}$. The indication offered by the plots here is again not as clear as in the exponential model, but the score pivot looks the most recommendable for testing H_0 with the various sample sizes considered. To conclude, keeping in mind that the prime scope of our proposition is to improve Wald-type inference without complicating too much the original procedure, we can say that in the one-parameter Poisson model the location-scale adjusted z-statistic does not always serve the purpose.

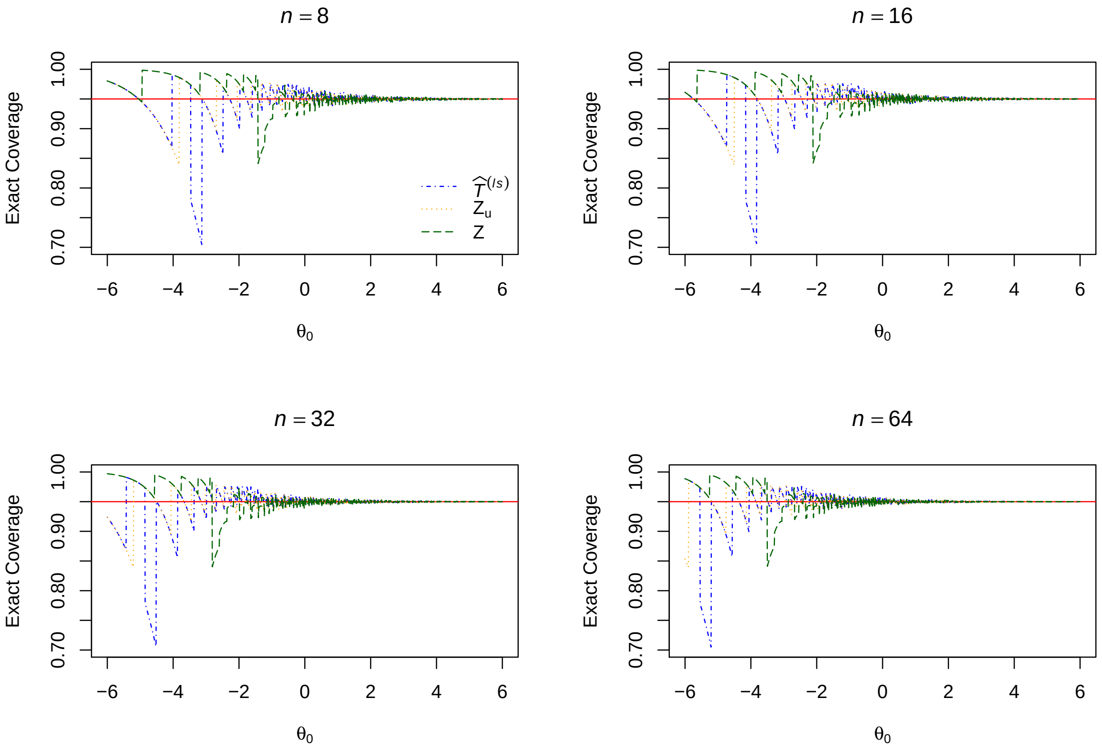

FIGURE 2.4: Exact coverage probabilities under the Poisson model for the two-sided interval at confidence level 0.95 based on pivots $\mathring{T}^{(ls)}$, Z_u and Z. Values are shown as a function of $\theta_0 \in \mathbb{R}$ and for various sample sizes n.

Logistic model

Consider a sample y with independent binary realizations $y_1, \ldots, y_n$ of the following distribution:

$$Y_i \sim Bern\left(\frac{e^{\theta}}{1+e^{\theta}}\right), \qquad \theta \in \mathbb{R}, \qquad i = 1, \ldots, n, \tag{2.23}$$

with $E_\theta(Y_i) = \mu_i = e^\theta/\big(1+e^\theta\big) \in (0,1)$. The log-likelihood function for these data results equal to $l(\theta) = n\bar{y}\theta - n\log\big(1+e^\theta\big)$ and its differentiation with respect to the scalar parameter θ delivers the score $l_1 = n\bar{y} - ne^\theta/\big(1+e^\theta\big)$. By solving the likelihood equation $l_1 = 0$, one straightforwardly obtains $\hat{\theta} = \log\big\{\bar{y}/(1-\bar{y})\big\}$. Moreover, the expected information can be shown to take the form $\nu_{1,1} = ne^\theta/\big(1+e^\theta\big)^2$. Using these results, it is possible to find the following expressions of the classical Wald z-statistics defined in (2.2) and (2.3):

$$\begin{aligned}
\mathring{T} &= \frac{\sqrt{n}e^{\theta_0/2}}{1+e^{\theta_0}}\left(\log\frac{\bar{y}}{1-\bar{y}} - \theta_0\right),\\
\widehat{T} &= \sqrt{n\big(\bar{y}-\bar{y}^2\big)}\left(\log\frac{\bar{y}}{1-\bar{y}} - \theta_0\right).
\end{aligned}$$

If model (2.23) holds, one may write $\theta = \log\big\{\mu_i/(1-\mu_i)\big\} = \text{logit}(\mu_i)$ and so the link function between the parameter and the mean of Y_i $(i = 1, \ldots, n)$ is canonical (McCullagh and Nelder, 1989, Section 2.2.4). This permits to employ formulae (2.14)–(2.17) to derive the location-scale corrections to $\mathring{T}$ and $\widehat{T}$ presented in (2.18) and (2.19). Ultimately, we get

$$\begin{aligned}
\mathring{T}^{(ls)} &= \frac{\frac{\sqrt{n}e^{\theta_0/2}}{1+e^{\theta_0}}\big(\log\frac{\bar{y}}{1-\bar{y}} - \theta_0\big) - \frac{e^{\theta_0}-1}{2\sqrt{n}e^{\theta_0/2}}}{\big(1+\frac{3e^{\theta_0}-2+3e^{-\theta_0}}{2n}\big)^{1/2}},\\
\widehat{T}^{(ls)} &= \frac{\sqrt{n(\bar{y}-\bar{y}^2)}\big(\log\frac{\bar{y}}{1-\bar{y}} - \theta_0\big)}{\big(1-\frac{e^{\theta_0}+2+e^{-\theta_0}}{4n}\big)^{1/2}}.
\end{aligned}$$

As considered before, the exact distributions of the four combinants need to be computed in order to evaluate their relative performance in terms of coverage properties exhibited by the associated 0.95 two-sided confidence intervals. Because $n\overline{Y}$ is a binomial random variable of indexes n and $e^\theta/(1+e^\theta)$, even in this case such calculation is not challenging, but requires to consider the fact that the ML estimate can also take infinite values. Specifically, $\hat{\theta} = -\infty$ $(+\infty)$ if all units in the sample equal 0 (1). By applying the popular limiting results reported in (2.22) to the current expressions of the z-statistics, it is easy to prove that when $\hat{\theta} = \pm\infty$, we have $\mathring{T} = \mathring{T}^{(ls)} = \pm\infty$ and $\widehat{T} = \widehat{T}^{(ls)} = 0$. Furthermore, it should be recalled that the distribution of $\widehat{T}^{(ls)}$ must be derived under the condition pertaining to the existence of its expression. In the habitual way, if the quantity between parentheses at the denominator is not strictly positive, we refer to the distribution of $\widehat{T}^{(l)}$ instead.

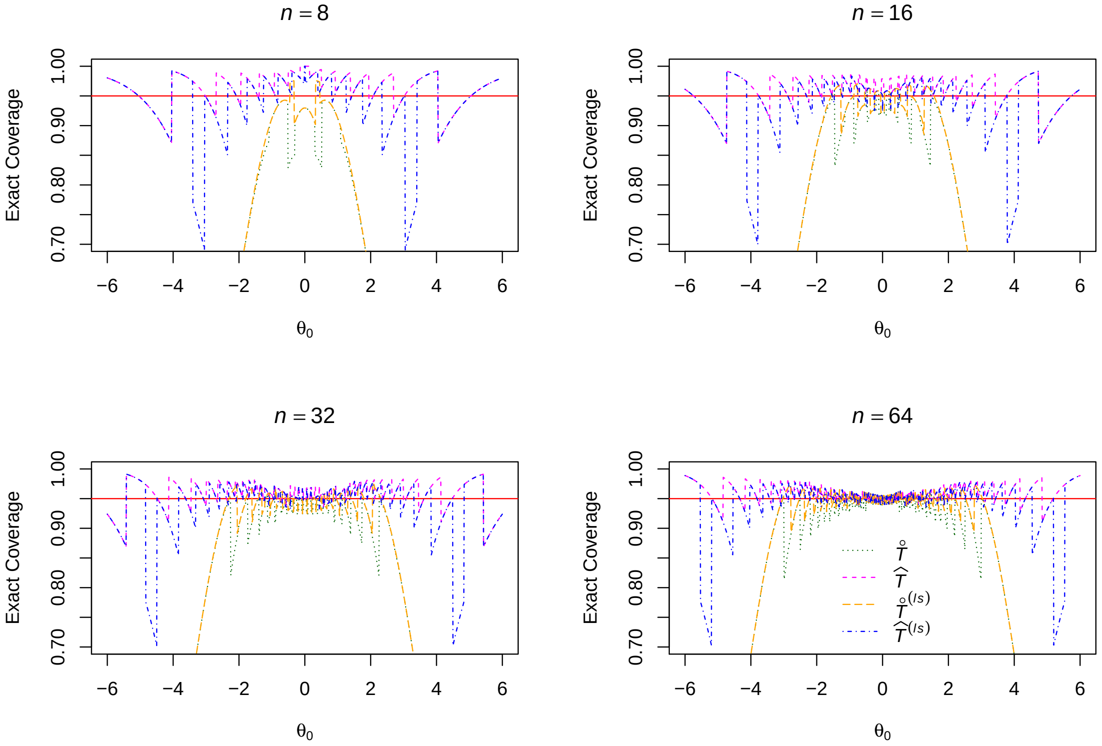

FIGURE 2.5: Exact coverage probabilities under the logistic model for the two-sided interval at confidence level 0.95 based on pivots $\mathring{T}$, $\widehat{T}$, $\mathring{T}^{(ls)}$ and $\widehat{T}^{(ls)}$. Values are shown as a function of $\theta_0 \in \mathbb{R}$ and for various sample sizes n.

Exact coverages of the confidence intervals based on $\mathring{T}$, $\mathring{T}^{(ls)}$, $\widehat{T}$ and $\widehat{T}^{(ls)}$ for varying θ_0 and several sample sizes n can then be seen in Figure 2.5. The decision in hypothesis testing implied by the presence of an infinite ML estimate for the various pivots is now revealed in the plots as the absolute value of the true parameter increases. Similarly to the Poisson setting, the disagreement between the conclusions of the test based on $\mathring{T}$ or $\mathring{T}^{(ls)}$ and the test based on $\widehat{T}$ or $\widehat{T}^{(ls)}$ is indeed testified by the different trend of the corresponding coverage probabilities for extreme values of θ_0. In outline it seems that, also for the logistic model, $\widehat{T}$ and $\widehat{T}^{(ls)}$ are generally more reliable tools for inference. In this framework, correcting the expectation and variance of the z-statistics looks especially profitable, even when the sample size is quite large: both intervals related to $\mathring{T}^{(ls)}$ and $\widehat{T}^{(ls)}$ have coverages remarkably closer to 0.95 than their regular version. The sole exception being made for isolated cases where the denominator of $\widehat{T}^{(ls)}$ approaches 0, recognisable in the various panels of Figure 2.5 by the two symmetrical spikes in the coverage curve of its associated confidence interval. Unlike what seen for the Poisson model, though, such problem tends to arise just for a very specific set of $|\theta_0|$ values at each n, and this set moves farther away from 0 when the sample size increases. Therefore, such complication does not look serious enough to impair the overall positive performance of $\widehat{T}^{(ls)}$.

Both exact distributions of the remaining likelihood-based pivotal quantities are easy to obtain for model (2.23). The expressions of the signed versions of the score and likelihood ratio statistics are indeed transformations of the sample mean $\bar{y}$ too, namely

$$Z_u = \sqrt{n}\frac{1+e^{\theta_0}}{e^{\theta_0/2}}\left(\bar{y} - \frac{e^{\theta_0}}{1+e^{\theta_0}}\right),$$

$$Z = \text{sign}\left(\log\frac{\bar{y}}{1-\bar{y}} - \theta_0\right)\sqrt{2n\left\{\bar{y}\left(\log\frac{\bar{y}}{1-\bar{y}} - \theta_0\right) + \log(1-\bar{y}) + \log\left(1+e^{\theta_0}\right)\right\}}.$$

Exact coverage probabilities of the corresponding two-tailed intervals at confidence level 0.95 are plotted in Figure 2.6 along with those referred to $\widehat{T}^{(ls)}$, so as to offer a more complete picture of the relative inferential adequacy delivered by the modified Wald combinant in this framework. One more time, the images seem to suggest the use of

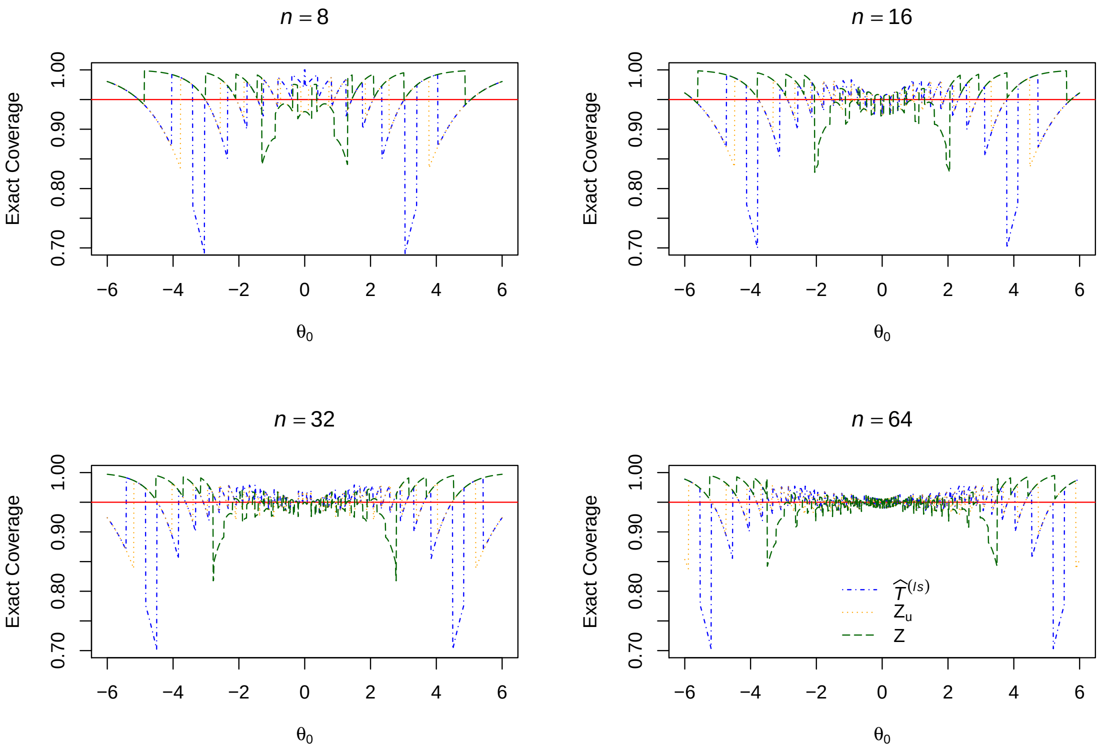

FIGURE 2.6: Exact coverage probabilities under the logistic model for the two-sided interval at confidence level 0.95 based on pivots $\mathring{T}^{(ls)}$, Z_u and Z. Values are shown as a function of $\theta_0 \in \mathbb{R}$ and for various sample sizes n.

the score statistic for drawing correct conclusions on the unknown parameter. Yet we can observe that, aside from those regions of the panels where the aforesaid instability caused by values close to 0 of its denominator manifests itself, $\widehat{T}^{(ls)}$ generally proves to behave reasonably well with respect to both Z_u and Z.

The present section has shown how the idea of correcting the moments of the z-statistic to better match those of the standard normal distribution may be successful in some single-parameter models. In fact, not only in most cases was accuracy of Wald-based inferential procedures improved, but also their essential simplicity was maintained. In the next part of this chapter, the same approach will be reformulated in such a way as to cope also with more complex scenarios.

2.3 Adjusting z-tests in regression settings

2.3.1 Notation and setup

Let us now introduce a standard regression model, where the mean of the dependent variable is related to a set of covariates through some specified function. To formalize the problem, consider a random sample $y = (y_1, \ldots, y_n)$ of independent observations from the generic distribution

$$Y_i \sim p_{Y_i}(y_i; \theta, x_i), \qquad \theta \in \Theta \subseteq \mathrm{I\!R}^k, \qquad i = 1, \ldots, n, \tag{2.24}$$

where $x_i = (x_{i1}, \ldots, x_{ik_0})$ is the k_0-dimensional vector of fixed covariates for the ith unit and the global parameter can be partitioned as $\theta = (\psi, \lambda)$. In particular, let the component of interest $\psi = \beta = (\beta_1, \ldots, \beta_{k_0}) \in \mathrm{I\!R}^{k_0}$ be the vector of scalar regression coefficients, while $\lambda = (\lambda_1, \ldots, \lambda_{k-k_0}) \in \Lambda \subseteq \mathrm{I\!R}^{k-k_0}$ contains the remaining unknown quantities supposed by the model (e.g. dispersion/precision parameters as defined in Section 2.5.1). It is then possible to link the mean of the ith response variable with the corresponding k_0 so-called regressors in x_i as:

$$E_\theta(Y_i) = \mu_i = h\biggl(\sum_{j=1}^{k_0} \beta_j x_{ij}\biggr), \qquad i = 1, \ldots, n, \tag{2.25}$$

where h is some suitably smooth function typically selected according to the support of Y_i. Notice that modeling frameworks like those considered in the last section are in fact special cases of this more general scenario. Indeed, specification (2.1) follows straightforwardly from (2.24) by setting $k = k_0 = 1$ with $x_i = 1$ for every $i = 1, \ldots, n$

and by choosing an appropriate function h. Below, we shall use the notation defined in Section 1.1 to refer to the usual likelihood quantities.

In regression settings one of the most common ways to investigate the effect of a specific covariate, accounting for all the others, on the dependent variable is via z-tests. The procedure for testing H_0: $\beta_j = \beta_{0j}$ ($j = 1, \ldots, k_0$) is the same as the one exposed in Section 2.2.1 for models with scalar global parameter. However, here the Wald z-statistic for the jth coefficient which is standard output of many statistical software takes the form

$$\widehat{T}^j = T^j(\hat{\theta}; \beta_{0j}) = \frac{\hat{\beta}_j - \beta_{0j}}{\sqrt{\kappa_j(\hat{\theta})}}, \tag{2.26}$$

where κ_j indicates the (j, j)th element in the block $i^{\beta\beta}$ of the inverse Fisher information matrix. Clearly, $\widehat{T}^j = \widehat{T}$ defined in (2.3) if $k = 1$. We stress that in the current context the standard error of $\hat{\beta}_j$ is usually evaluated at the global ML estimate so that to avoid fitting the restricted model under the null hypothesis, which might be time-consuming for large datasets and/or in the presence of many parameters.

As repeatedly emphasized in the preceding parts, the $N(0, 1)$ distribution can be a very poor approximation for the null behaviour of the pivot (2.26) in small-to-moderate-sized samples. Moreover, in multiple regression models the failure of such asymptotic result may occur also whether k is large relative to n (see, for example, McCullagh and Nelder, 1989, Section 6.2.4). Thus, in the same vein as what already suggested for the one-parameter case, the next section will present a convenient procedure to enhance Wald-type inferences while allowing the overall parameter to be multidimensional.

2.3.2 Location adjusted z-statistic

The Wald combinant in (2.26) is undoubtedly not as easy to deal with as its analogue (2.3) in the setting with scalar parameter is. In particular, the explicit computation of the former's cumulants is tedious and results in expressions that are much less handy than those reported in Section 2.2.3. Consequently, under the present regression scenario an alternative approach for obtaining the quantities required to perform the moments correction of the z-statistic might be desirable.

The backbone of the insight behind the modification of the Wald pivot we are going to propose is seeing the function

$$T^j = T^j(\theta; \beta_{0j}) = \frac{\beta_j - \beta_{0j}}{\sqrt{\kappa_j(\theta)}} \tag{2.27}$$

as a non-singular transformation of the full parameter θ and identifying $\widehat{T}^j$ in (2.26) as its ML estimator. Then, similarly to $\hat{\theta}$, $\widehat{T}^j$ may be considered to suffer from finite-sample bias, which one can try to reduce by applying, for instance, the standard technique for asymptotic bias correction described by Efron (1975, Remark 11, p. 1214).

In order to derive a general formula for the bias of the z-statistic, assume T^j in (2.27) is at least three times differentiable in the argument θ. Given the consistency of the ML estimator, the Taylor expansion of $\widehat{T}^j - T^j$ about θ, written by adopting the Einstein summation convention, is

$$\begin{aligned} T^j(\hat{\theta};\beta_{0j}) - T^j(\theta;\beta_{0j}) = (\hat{\theta}^s - \theta^s)T^j_s(\theta;\beta_{0j}) + \frac{1}{2}(\hat{\theta}^s - \theta^s)(\hat{\theta}^t - \theta^t)T^j_{st}(\theta;\beta_{0j}) \\ + \frac{1}{6}(\hat{\theta}^s - \theta^s)(\hat{\theta}^t - \theta^t)(\hat{\theta}^u - \theta^u)T^j_{stu}(\theta;\beta_{0j}) + O_p\big(n^{-3/2}\big), \end{aligned} \tag{2.28}$$

with $T^j_s(\theta;\beta_{0j}), T^j_{st}(\theta;\beta_{0j})$ and $T^j_{stu}(\theta;\beta_{0j})$ gradient, hessian and third derivative, respectively, of function (2.27) $(s,t,u = 1,\dots,k)$, all of order $O\big(n^{1/2}\big)$. Then the following expression ensues straightforwardly from taking expectations in both sides of (2.28) and applying result (2.4), as done in Remark 3 of Kosmidis and Firth (2010, Section 4.3):

$$\begin{aligned} E_\theta\big\{T^j(\hat{\theta};\beta_{0j}) - T^j(\theta;\beta_{0j})\big\} &= B^s(\theta)T^j_s(\theta;\beta_{0j}) + \frac{1}{2}\xi^{s,t}(\theta)T^j_{st}(\theta;\beta_{0j}) + O\big(n^{-3/2}\big) \\ &= B_{T^j}(\theta;\beta_{0j}) + O\big(n^{-3/2}\big), \end{aligned} \tag{2.29}$$

where $B^s(\theta)$ is such that $E_\theta\big(\hat{\theta}^s - \theta^s\big) = B^s(\theta) + o\big(n^{-1}\big)$ and $\xi^{s,t}(\theta)$ is the (s,t)th element of $i(\theta)^{-1}(s,t = 1,\dots,k)$. The first term in the asymptotic bias expansion of $\widehat{T}^j = T^j(\hat{\theta};\beta_{0j})$ may thus be estimated by $B_{T^j}(\hat{\theta};\beta_{0j})$, so that to define the location adjusted z-statistic in regression settings as

$$\widehat{T}^{j,*} = T^{j,*}(\hat{\theta};\beta_{0j}) = \widehat{T}^j - B_{T^j}(\hat{\theta};\beta_{0j}). \tag{2.30}$$

Henceforth, we will refer to the test based on $\widehat{T}^{j,*}$ as the adjusted z-test. Note that the advantage of viewing $\widehat{T}^j$ as an estimator of a transformation of θ lies in the simplicity of the procedure to derive its bias. Indeed, $B_{T^j}(\theta;\beta_{0j})$ in formula (2.29) depends only on quantities which are normally computed with no effort in regression frameworks.

The importance of our expedient justifies the choice of considering for correction Wald pivots which use the expected information matrix to approximate the standard error of $\hat{\beta}_j$. On this basis, the reparametrization trick is in fact readily applicable, as data enter the expression only through the ML estimates. We are also aware that definition (2.30) does not completely agree with what recommended for one-parameter models. As the primary objective is approaching the null distribution of the z-statistic to the $N(0,1)$, in that case the correction was sensibly performed by using its moments under H_0. In the general scenario under analysis, the composite null hypothesis admits the specification $H_0\colon \theta = \theta_0$ with $\theta_0 = (\beta_1, \ldots, \beta_{0j}, \ldots, \beta_{k_0}, \lambda_1, \ldots, \lambda_{k-k_0}) \in \Theta_0 \subseteq \mathrm{I\!R}^{k-1}$, so the null expected value of (2.26) can be expressed as

$$E_{\theta_0}(\widehat{T}^j) = B_{T^j}(\theta_0; \beta_{0j}) + O\big(n^{-3/2}\big).$$

The most natural estimator of θ_0, now partially unknown, is obviously $\hat{\theta}_{\beta_{0j}}$, thus in principle the adjustment in location should be accomplished via $B_{T^j}(\hat{\theta}_{\beta_{0j}}; \beta_{0j})$. The decision to lean rather on $B_{T^j}(\hat{\theta}; \beta_{0j})$ is taken with the aim of keeping the computational cost of classical Wald-type procedures unchanged, by avoiding the constrained maximization of the log-likelihood function. However, such a resolution rests also on practical grounds: simulation results not shown here have not detected sensitive improvements in the general performance of the adjusted z-test when evaluation of the bias at the constrained ML estimate is preferred. In closing, we acknowledge that a scale correction of the Wald combinant is not being considered in this multiparameter setting because of the difficulty implicit in the derivation of a convenient expression for the variance of $\widehat{T}^j$.

2.4 Location adjusted z-statistic when $k = 1$

2.4.1 Asymptotic results

In this part, focus is put back on models with scalar parameter θ. Such special framework is indeed particularly suitable to illustrate in a clear and effective way some general features of the location adjusted z-statistic defined in (2.30). Consider again assumptions of model (2.1) and the notation adopted in Section 2.2.1. For coherence of exposition, let us express the location adjusted z-statistic for single-parameter models as

$$\widehat{T}^* = \widehat{T}^*(\theta_0) = \widehat{T} - B_T(\hat{\theta}; \theta_0). \tag{2.31}$$

In fact, it is not complicated to see that formulation (2.30) when $k = 1$ reduces to (2.31), recalling the definition of the Wald pivot $\widehat{T}$ given in (2.3). Now, provided that the first-order bias of the ML estimate $\hat{\theta}$ can be written in power notation as

$$B(\theta) = \frac{\nu_3 + 2\nu_{1,2}}{2\nu_{1,1}^2} = O\big(n^{-1}\big) \tag{2.32}$$

(see, e.g., Pace and Salvan, Example 9.11, p. 360), one may employ formula (2.29) to derive the expression for the bias of $\widehat{T}$. In particular, after some computational steps detailed in Appendix it results

$$B_T(\theta;\theta_0)=\frac{\nu_{1,2}}{2\nu_{1,1}^{3/2}}-\frac{\theta-\theta_0}{8}\left(\frac{3\nu_3^2+8\nu_3\nu_{1,2}+5\nu_{1,2}^2}{\nu_{1,1}^{5/2}}+\frac{2\nu_4+4\nu_{1,3}+2\nu_{2,2}+2\nu_{1,1,2}}{\nu_{1,1}^{3/2}}\right),$$

which is of order $O\big(n^{-1/2}\big)$, like its derivatives. Notice that such expression is consistent with what obtained in Section 2.2.2, since

$$B_T(\theta_0;\theta_0)=\frac{\mathring{\nu}_{1,2}}{2\mathring{\nu}_{1,1}^{3/2}}=\mathring{E}_1\big(\widehat{T}\big).$$

In order to evaluate some asymptotic properties of $\widehat{T}^*$ as defined in (2.31), it is helpful to observe that a valid asymptotic expansion around θ_0 for the estimated correction in mean of the Wald combinant is

$$B_T(\hat{\theta};\theta_0)=\frac{\mathring{\nu}_{1,2}}{2\mathring{\nu}_{1,1}^{3/2}}\overset{\bullet}{+}(\hat{\theta}-\theta_0)B_T'(\theta_0;\theta_0)\overset{\bullet}{+}O_p\big(n^{-3/2}\big).$$

As a consequence, we can conclude that

$$E_{\theta_0}\big(\widehat{T}^*\big)=E_{\theta_0}\big(\widehat{T}\big)-E_{\theta_0}\big\{B_T(\hat{\theta};\theta_0)\big\}=O\big(n^{-3/2}\big),$$

and thus the efficacy of the location adjustment of the z-statistic is established. Let us now try to study the behaviour of the null variance of $\widehat{T}^*$ by similar argument as that used for the bias-corrected ML estimator in Section 9.4.3 of Pace and Salvan (1997). Firstly, it is quite simple to find that

$$\begin{aligned}\mathrm{Var}_{\theta_0}\big(\widehat{T}^*\big)&=\mathrm{Var}_{\theta_0}\big\{\widehat{T}-B_T(\hat{\theta};\theta_0)\big\}\\&=\mathrm{Var}_{\theta_0}\big(\widehat{T}\big)+\mathrm{Var}_{\theta_0}\big\{B_T(\hat{\theta};\theta_0)\big\}-2\mathrm{Cov}_{\theta_0}\big\{\widehat{T},B_T(\hat{\theta};\theta_0)\big\}\\&=\mathrm{Var}_{\theta_0}\big(\widehat{T}\big)-2\mathrm{Cov}_{\theta_0}\big\{\widehat{T},B_T(\hat{\theta};\theta_0)\big\}\overset{\bullet\bullet}{+}O\big(n^{-2}\big).\end{aligned}\tag{2.33}$$

Secondly, with some reasonable effort, expression (2.13) may be rewritten as

$$\mathrm{Var}_{\theta_0}\big(\widehat{T}\big)=\frac{\mathring{\nu}_{2,2}}{\mathring{\nu}_{1,1}^2}\overset{\bullet\bullet}{+}\frac{2B_T'(\theta_0;\theta_0)}{\mathring{\nu}_{1,1}^{1/2}}-2\{B_T(\theta_0;\theta_0)\}^2\overset{\bullet\bullet}{+}O\big(n^{-2}\big)\tag{2.34}$$

and the null covariance between $\widehat{T}$ and $B_T(\hat{\theta};\theta_0)$ can be expanded in the following way:

$$\mathrm{Cov}_{\theta_0}\big\{\widehat{T},B_T(\hat{\theta};\theta_0)\big\}=\frac{B_T'(\theta_0;\theta_0)}{\mathring{\nu}_{1,1}^{1/2}}\overset{\bullet\bullet}{+}O\big(n^{-2}\big).\tag{2.35}$$

Finally, by replacement of the quantities (2.34) and (2.35) in formula (2.33), the variance of the location adjusted z-statistic is

$$\mathrm{Var}_{\theta_0}\big(\widehat{T}^*\big) = \frac{\mathring{\nu}_{2,2}}{\mathring{\nu}_{1,1}^2} - 2\{B_T(\theta_0;\theta_0)\}^2 \overset{\bullet\bullet}{+} O\big(n^{-2}\big),$$

where, as one expects, the term of order $O(1)$ is equal to 1, since $\nu_{2,2} = \nu_{1,1}^2 + O(n)$ through well-known relations between cumulants and central moments. It is therefore easy to see that the comparison between the $O\big(n^{-1}\big)$ terms in the variances of $\widehat{T}$ and $\widehat{T}^*$ depends on the function $B_T'(\theta_0;\theta_0)$, which in Appendix is shown to take the form

$$B_T'(\theta_0;\theta_0) = -\left(\frac{\mathring{\nu}_4 - \mathring{\nu}_{2,2} - \mathring{\nu}_{1,1,2}}{4\mathring{\nu}_{1,1}^{3/2}} + \frac{3\mathring{\nu}_3^2 + 2\mathring{\nu}_3\mathring{\nu}_{1,2} - \mathring{\nu}_{1,2}^2}{8\mathring{\nu}_{1,1}^{5/2}}\right). \tag{2.36}$$

Unfortunately, there seems to exist no general indication about the sign of such expression, so the relative variance properties of the two pivots need to be evaluated on a case by case basis.

As usual, the special class of exponential families with canonical parameter θ offers the chance to further simplify the present scenario. In particular, one can straightforwardly obtain that in those models

$$B_T(\theta;\theta_0) = -\frac{\theta-\theta_0}{8}\left(\frac{3\nu_3^2}{\nu_{1,1}^{5/2}} + \frac{2\nu_4}{\nu_{1,1}^{3/2}}\right)$$

and

$$B_T'(\theta_0;\theta_0) = -\left(\frac{3\mathring{\nu}_3^2}{8\mathring{\nu}_{1,1}^{5/2}} + \frac{\mathring{\nu}_4}{4\mathring{\nu}_{1,1}^{3/2}}\right).$$

We highlight that the only quantity with ambiguous sign in the last expression is $\mathring{\nu}_4$. Thus, for example, a useful observation might be that the term of order $O\big(n^{-1}\big)$ in the variance of $\widehat{T}^*$ is smaller than that of the unmodified combinant $\widehat{T}$ if $\mathring{\nu}_4 \leq -3\mathring{\nu}_3^2/(2\mathring{\nu}_{1,1})$.

2.4.2 Inference on a binomial proportion

To give some insight on the practical use of $\widehat{T}^*$ in a realistic setting, the problem of inference on a binomial proportion may now be discussed. Indeed, such one-parameter model has often been considered in the literature, primarily due to issues associated with the erratic coverage properties of Wald-type confidence intervals (Brown *et al.*, 2001).

Let the sample y consist of n independent units y_i drawn from the Bernoulli distribution

$$Y_i \sim Bern(\theta), \qquad \theta \in (0,1), \qquad i = 1, \ldots, n. \tag{2.37}$$

The log-likelihood of the model is $l(\theta) = n\bar{y} \log\big\{\theta/(1-\theta)\big\} + n\log(1-\theta)$, thus the score function equals $l_1 = n(\bar{y} - \theta)/\big\{\theta(1-\theta)\big\}$. Moreover, it is easy to prove that the ML estimate $\hat{\theta} = \bar{y}$ is unbiased and the expression of the Fisher information takes the form $\nu_{1,1} = n/\{\theta(1-\theta)\}$.

Consider a statistical test regarding the proportion θ which involves the null hypothesis H_0: $\theta = \theta_0$ and the alternative H_1: $\theta \neq \theta_0$, for some specific value $\theta_0 \in (0,1)$. Several pivotal quantities are available to address this inferential problem. Specifically, it is not hard to see that under assumptions (2.37) the standard Wald z-statistics (2.2) and (2.3) are, respectively,

$$\mathring{T} = \sqrt{n}\frac{\bar{y} - \theta_0}{\sqrt{\big\{\theta_0(1-\theta_0)\big\}}},$$
$$\widehat{T} = \sqrt{n}\frac{\bar{y} - \theta_0}{\sqrt{\big\{\bar{y}(1-\bar{y})\big\}}}.$$

Expansions (2.11) and (2.13) for their variances cannot be written in a succinct form, since the expectations of log-likelihood derivatives implicated are not as simple as those for models in Section 2.2.4. Hence in this case we shall not report the location-scale adjusted z-statistics $\mathring{T}^{(ls)}$ in (2.18) and $\widehat{T}^{(ls)}$ in (2.19) explicitly. On the other hand, the general expression of the location adjusted z-statistic (2.31) here becomes:

$$\widehat{T}^* = \sqrt{n}\frac{\bar{y} - \theta_0}{\sqrt{\bar{y}(1-\bar{y})}} - \frac{4\bar{y}^2 - \bar{y} - 8\bar{y}^2\theta_0 + 8\bar{y}\theta_0 - 3\theta_0}{8\big\{\bar{y}(1-\bar{y})\big\}^{3/2}}.$$

Another sort of modified Wald combinant which is extremely popular in the research area dedicated to interval estimation for binomial proportions is the one recommended by Agresti and Coull (1998), namely

$$\widetilde{T} = \widetilde{T}(\theta_0) = \sqrt{\tilde{n}}\frac{\tilde{y} - \theta_0}{\sqrt{\big\{\tilde{y}(1-\tilde{y})\big\}}}, \qquad \tilde{y} = \frac{n\bar{y} + 2}{n+4}, \;\; \tilde{n} = n + 4.$$

Evidently, as happens with $\widehat{T}$, the latter basic expression makes $\widetilde{T}$ particularly adequate to construct confidence intervals for the unknown parameter by inversion. Unlike the standard Wald pivotal quantity, though, the proposal of Agresti and Coull (1998) has exhibited appreciable coverage properties in small to moderate samples, representing thus a valid benchmark to judge the effectiveness of our method. For what concerns the rest of the likelihood combinants, it may be effortlessly shown that the one-sided version of the score statistic Z_u coincides with $\mathring{T}$, whereas the signed root of the log-likelihood ratio statistic can be written as

$$Z = \text{sign}(\bar{y} - \theta_0)\sqrt{2n\big\{(1-\bar{y})\log(1-\bar{y}) + \bar{y}\log(\bar{y}/\theta_0) - (1-\bar{y})\log(1-\theta_0)\big\}}.$$

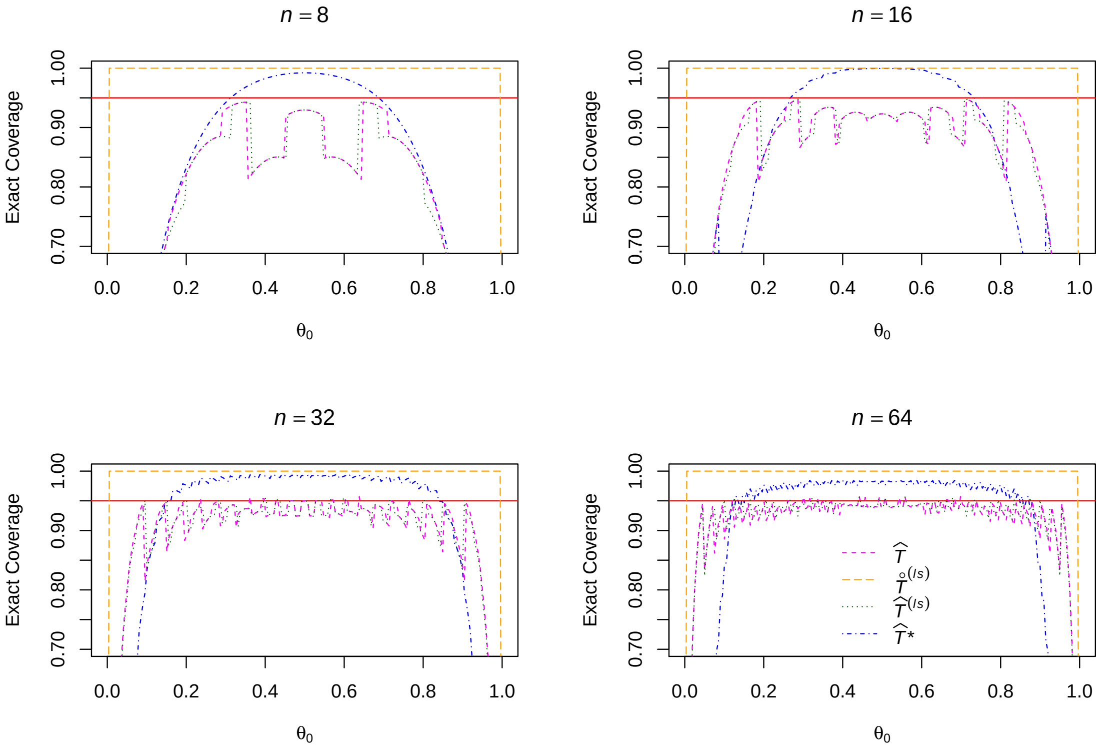

FIGURE 2.7: Exact coverage probabilities under the binomial model for the two-sided interval at confidence level 0.95 based on pivots $\widehat{T}$, $\mathring{T}^{(ls)}$, $\widehat{T}^{(ls)}$ and $\widehat{T}^{*}$. Values are shown as a function of $\theta_0 \in (0, 1)$ and for various sample sizes n.

Adopting the procedure described in Section 2.2.4, exact coverage probabilities of confidence intervals built from the variety of pivots above can be obtained and compared. In Figure 2.7 it is possible to visualize in the usual way the relative testing performance of $\widehat{T}$, the two location-scale adjusted z-statistics and $\widehat{T}^{*}$. We observe that the simultaneous correction in mean and variance of the Wald combinants under model (2.37) appears not to be advisable. In greater detail, $\mathring{T}^{(ls)}$ always leads to accept H_0 because the quantity at the denominator in its expression remains too large for any couple (θ_0, n), while $\widehat{T}^{(ls)}$ is not able to enhance the coverage properties of the standard Wald pivot. The performance of the adjusted z-test is not especially satisfying either and, surprisingly, seems to deteriorate as the sample size increases from $n = 8$ to $n = 16$ for values of θ_0 around 0.5. An additional matter to be explored is the unusual smoothness of its related coverage curve. Nevertheless, $\widehat{T}^*$ might be considered generally more reliable than its standard version when $n = 8$. We shall then proceed with a further evaluation of its testing properties with regard to the other statistics involved in the analysis. Panels of Figure 2.8 allow to contrast the actual coverage of the confidence interval derived by inverting the location adjusted z-statistic with those ensuing from the score, likelihood ratio and Agresti and Coull combinants, respectively. As can be seen, despite the exceptional simplicity of its formulation, $\widetilde{T}$ proves to be the pivotal quantity ensuring the highest general accuracy in inference. In this second comparison, even for the smallest sample size, the performance of $\widehat{T}^*$ does not look as solid as those of its competitors.

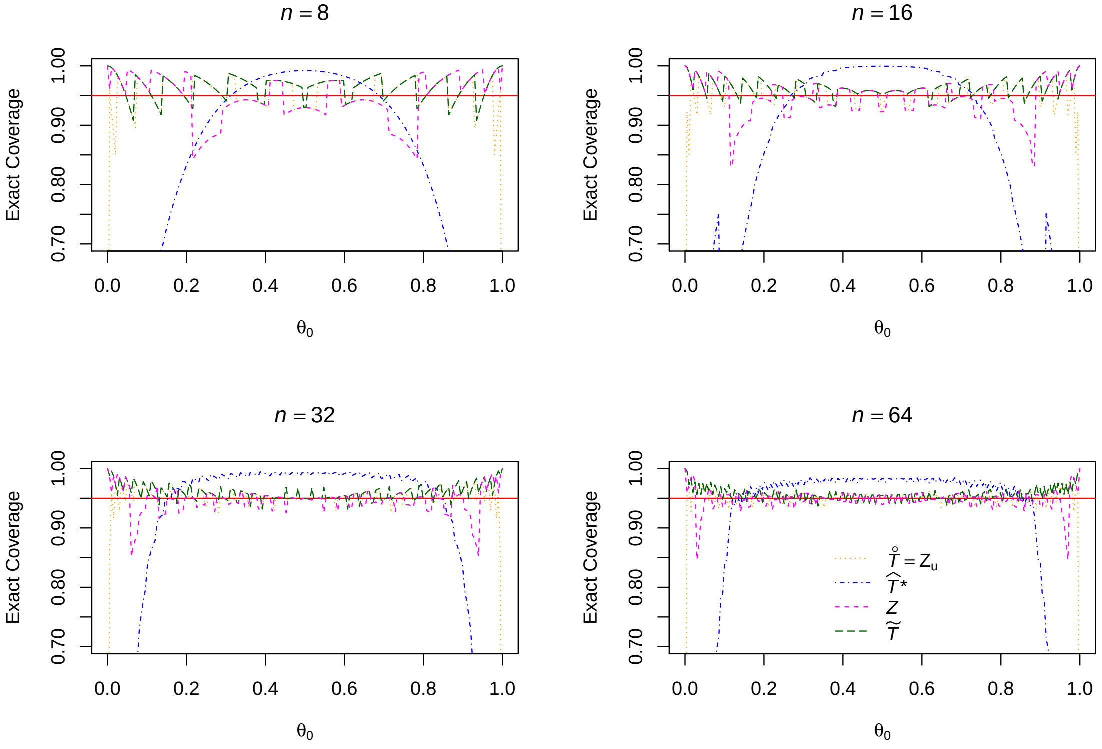

FIGURE 2.8: Exact coverage probabilities under the binomial model for the two-sided interval at confidence level 0.95 based on pivots $\widehat{T}^*$, $\mathring{T} = Z_u$, $\widetilde{T}$ and Z. Values are shown as a function of $\theta_0 \in (0, 1)$ and for various sample sizes n.

2.5 Generalized linear models

2.5.1 Introduction

In regression settings, a prime position among parametric statistical specifications is enjoyed by generalized linear models (GLMs). Popularized by McCullagh and Nelder (1989), such class of models was originally introduced as a flexible tool for relaxing the basic assumptions of classical linear regressions in order to allow the dependent variable to have both a distribution other than normal and a variance depending on the covariates (Nelder and Wedderburn, 1972). According to the general setup defined in Section 2.3.1, under standard hypotheses of a GLM the response Y_i $(i = 1, \ldots, n)$ follows a probability distribution belonging to an exponential dispersion family (Jørgensen, 1987) such that

$$g(\mu_i) = h^{-1}(\mu_i) = \sum_{j=1}^{k_0} \beta_j x_{ij} = \eta_i,$$

$$\mathrm{Var}_\theta(Y_i) = \phi V(\mu_i),$$

where μ_i and h were defined in (2.25) and V is the so-called variance function. Commonly, g and η_i are known as link function and linear predictor, respectively, whereas ϕ is the dispersion parameter, sometimes expressed as $\phi(\lambda) = 1/\lambda$, with λ named precision parameter.

GLMs are particularly relevant for our study in a number of respects. Indeed, a considerable stream of research has focused on the analysis of bias in the estimation of such models. The first noticeable result in this field was achieved by Cordeiro and McCullagh (1991), who provided general expressions for the first-order biases of the ML estimators, illustrating in addition a simple algorithm to derive bias-corrected estimates. Subsequently, Cordeiro and Barroso (2007) went further by obtaining the term of order $O\big(n^{-2}\big)$ in the bias expansion of the estimators and defined third-order bias-corrected estimates. Removal of the leading bias term by adjustment of the score vector was discussed instead in Kosmidis and Firth (2009). Specifically, the authors gave a necessary and sufficient condition for the existence of a penalized likelihood interpretation of that method in GLMs.

However, point estimation has not been the only topic of interest within this family of models during the years. In fact, the non-gaussian and possibly discrete nature of the dependent variable poses a significant challenge to the accuracy of the usual asymptotic approximations for inferences in the moderate sample situation. Consequently, as many times remarked, undesirable side-effects are likely to be observed both in interval estimation and hypothesis testing. Such question was approached, for instance, by Sun *et al.* (2000), who proposed to correct confidence bands for the mean response μ_i by applying a Cornish-Fisher expansion (Pace and Salvan, 1997, Section 10.6) to the distribution of the ML estimator. From our perspective, another interesting work on the subject is that of Xu and Gupta (2005), where improvement of confidence regions in GLMs was reached upon a modification of the Wald statistic which accounts for non-normality of the response and finiteness of the sample.

To summarize, both the substantial need of enhancing first-order inferential procedures and the immediate availability of closed-form expressions for the bias of ML estimators make the GLMs framework the perfect statistical environment for the employment of our suggested location adjusted z-statistic. More specifically, let us emphasize the fact that all quantities appearing in the general correction formula (2.29) are obtainable without difficulties under this scenario, not just the first-order bias of $\hat{\theta}$. In fact, calculation of the derivatives of the function T^j defined in (2.27) is markedly facilitated by the simple general form taken by the expected information matrix in a GLM (see, e.g., Pace and Salvan, 1997, p. 239). This aspect has led quite naturally to develop the `R` package `brglm2` (Kosmidis, 2016), which automatically implements computations for obtaining the location adjusted z-statistic along with other methods of bias reduction in GLMs.

2.5.2 Performance of the location adjusted z-statistic

In this last part of the chapter, some illustrations about the testing performance of the location adjusted z-statistic are supplied. The effects of the proposed correction in mean are assessed via simulation in experimental settings belonging to the class of GLMs. Such evaluation involves not only the unadjusted z-statistic as first term of comparison, but also other pivotal quantities typically adopted for asymptotic inference. To be consistent with the notation defined for the two versions of Wald pivot, Z_{uP}^j and Z_P^j denote the signed profile score and log-likelihood ratio statistics introduced in Section 1.3.2 which correspond to the usual null hypothesis on the jth regression coefficient. Besides these combinants, we consider also the second-order accurate modified root of the profile log-likelihood ratio statistic in its variant derived by Skovgaard (1996). Particularly, the latter takes the form

$$Z_P^{*,j} = Z_P^j + \frac{1}{Z_P^j} \log \frac{Z_P^j}{\widetilde{U}^j}, \tag{2.38}$$

where

$$\widetilde{U}^j = \big[\widehat{S}^{-1}\widehat{Q}\big]_j \big|j(\hat{\theta})\big|^{1/2} \big|i(\hat{\theta})^{-1}\big| \big|\widehat{S}\big| \big|j_{-jj}(\hat{\theta}_{\beta_{0j}})\big|^{-1/2}$$

approximates the analogue term containing sample space derivatives in the original formulation of Barndorff-Nielsen (1986, 1991). Note that $[\,\cdot\,]_j$ indicates the jth coordinate of the related vector, $\widehat{S} = \mathrm{Cov}_{\hat{\theta}}\big\{l_\theta(\hat{\theta}), l_\theta(\hat{\theta}_{\beta_{0j}})\big\}, \widehat{Q} = \mathrm{Cov}_{\hat{\theta}}\big\{l_\theta(\hat{\theta}), l(\hat{\theta}) - l(\hat{\theta}_{\beta_{0j}})\big\}$ and $j_{-jj}(\hat{\theta}_{\beta_{0j}})$ is the $(k-1)\times(k-1)$ matrix formed by deleting the jth row and jth column of the observed information evaluated at the restricted ML estimate. In the next experiments, the `R` package `likelihoodAsy` (Bellio and Pierce, 2015) is used for computing $Z_P^{*,j}$ as reported in (2.38).

Gamma regression

A first simulation study can be set up as follows: starting from $n = 8$, for every ith unit, covariates x_{i1} and x_{i2} $(i = 1, \ldots, n)$ are generated as independent realizations of a $N(1, 1)$. The corresponding observed dependent variable y_i in each of the 2000 simulated datasets is then randomly drawn from a $\Gamma(\phi^{-1}, \vartheta_i)$ distribution with dispersion parameter $\phi = 0.5$ and rate $\vartheta_i = (\phi\mu_i)^{-1}$, where $\mu_i = \exp(\beta_{01} + \beta_{02}x_{i1} + \beta_{03}x_{i2})$ with $\beta_{01} = 1$, $\beta_{02} = 1$ and $\beta_{03} = 2$. On every sample, the composite null hypothesis $H_0\colon \beta_j = \beta_{0j}$ $(j = 1, 2, 3)$ is tested versus the two-tailed alternative taking the other regressors into account and using several pivots, so that empirical rejection probabilities of the corresponding tests at significance levels $\alpha = 0.01, 0.05$ can be estimated. This procedure is repeated for $n = 16, 32, 64$, but instead of generating a new set of covariates every time, the same x_{i1} and x_{i2} $(i = 1, \ldots, 8)$ are used for adjacent blocks of 8 units. Results of the study are available in Table 2.1, which displays estimated rejection probabilities for tests based on the standard Wald statistic $\widehat{T}^j$, the location adjusted z-statistic $\widehat{T}^{j,*}$, the one-sided profile score statistic Z^j_{uP}, the signed root of the profile log-likelihood ratio statistic Z^j_P and its modification $Z^{*,j}_P$ $(j = 1, 2, 3)$. As may be seen, for small values of n (especially $n = 8, 16$) the adjusted z-test has empirical rejection probabilities much closer to α than the classical version, and does also better than the test associated with the log-likelihood ratio combinant. Among first-order tests, Z^j_{uP} appears to have the best general performance, even comparable to the second-order accurate $Z^{*,j}_P$. Not surprisingly, such discrepancies tend to disappear as the sample size grows.

A more realistic scenario is considered in the next simulation experiment, involving the *clotting* dataset (McCullagh and Nelder, 1989, p. 300). The data record observations of $n = 18$ mean clotting times in seconds of blood (y) for nine percentage concentrations of normal plasma (x_1) and two lots of clotting agent $(x_2 = 1, 2)$. Assuming $Y_1, \ldots, Y_n$ are independent $\Gamma(\phi^{-1}, \vartheta_i)$-distributed random variables with $\vartheta_i = (\phi\mu_i)^{-1}$ and $\mu_i = \exp(\beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i1} x_{i2})$ $(i = 1, \ldots, n)$, a Gamma regression with log link is fitted to the data and 2000 samples of size n are simulated under the ML fit, namely with $\theta = \hat{\theta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\phi})$. Similarly as before, to test $H_0 : \beta_j = \beta_{0j} = \hat{\beta}_j$ $(j = 1, 2, 3, 4)$ while accounting for the other covariates in the model, the usual statistics are computed on every dataset. Table 2.2 reports empirical rejection probabilities of the associated two-sided tests at theoretical levels $\alpha = 0.01, 0.05$. For each regression coefficient, the adjusted z-test results in rejection probabilities closer to α than its standard variant. Furthermore, the normal Q-Q plots in Figure 2.9 illustrate how the adjustment in location enhances the normal approximation to the null distribution of the z-statistic when testing $H_0\colon \beta_4 = \beta_{04}$. Table 2.2 gives also evidence that, although

Table 2.1: Empirical rejection probabilities at nominal levels $\alpha = 0.01, 0.05$ of the two-sided tests related to $\widehat{T}^j$, its location adjusted version $\widehat{T}^{j,*}$, the profile score statistic Z^j_{uP}, the profile likelihood ratio statistic Z^j_P and its modification $Z^{j,*}_P$ ($j = 1, 2, 3$) in the Gamma regression model, estimated by a study based on 2000 simulated datasets of size $n = 8, 16, 32, 64$.

	$\alpha = 0.01$					$\alpha = 0.05$				
$n = 8$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.109	0.040	0.015	0.051	0.014	0.178	0.096	0.074	0.135	0.060
$j = 2$	0.113	0.048	0.004	0.062	0.015	0.199	0.105	0.068	0.147	0.072
$j = 3$	0.107	0.046	0.005	0.057	0.016	0.200	0.099	0.066	0.144	0.064
$n = 16$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.043	0.026	0.015	0.027	0.015	0.107	0.068	0.062	0.087	0.057
$j = 2$	0.046	0.020	0.008	0.023	0.009	0.112	0.071	0.057	0.083	0.057
$j = 3$	0.039	0.020	0.006	0.024	0.011	0.116	0.068	0.051	0.081	0.051
$n = 32$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.023	0.013	0.010	0.014	0.010	0.072	0.058	0.051	0.061	0.054
$j = 2$	0.022	0.014	0.008	0.013	0.011	0.076	0.059	0.048	0.061	0.049
$j = 3$	0.024	0.017	0.011	0.018	0.013	0.074	0.056	0.043	0.061	0.045
$n = 64$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.020	0.016	0.013	0.014	0.018	0.071	0.063	0.058	0.062	0.065
$j = 2$	0.014	0.013	0.009	0.011	0.012	0.061	0.052	0.049	0.056	0.050
$j = 3$	0.014	0.011	0.008	0.010	0.009	0.063	0.056	0.050	0.058	0.053

not performing as well as $Z^{j,*}_P$, $\widehat{T}^{j,*}$ is always preferable to the profile likelihood ratio statistic and seems even more reliable than Z^j_{uP} when the nominal size equals 0.05.

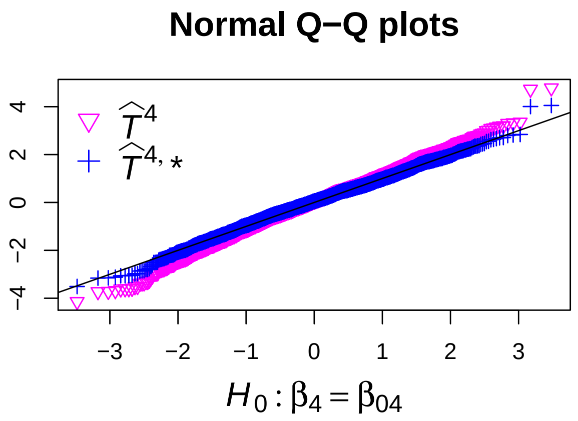

Figure 2.9: Normal Q-Q plots based on 2000 values of $\widehat{T}^4$ and $\widehat{T}^{4,*}$ computed under the null hypothesis $H_0 \colon \beta_4 = \beta_{04}$ in the *clotting* example.

Table 2.2: Empirical rejection probabilities at nominal levels $\alpha = 0.01, 0.05$ of the two-tailed tests related to $\widehat{T}^j$, $\widehat{T}^{j,*}$, Z^j_{uP}, Z^j_P and $Z^{j,*}_P$ $(j = 1, 2, 3, 4)$ in the *clotting* example. The figures are based on a simulation study with 2000 replications.

	$\alpha = 0.01$					$\alpha = 0.05$				
	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.036	0.016	0.006	0.023	0.006	0.106	0.059	0.070	0.089	0.051
$j = 2$	0.039	0.015	0.010	0.023	0.008	0.108	0.060	0.071	0.088	0.052
$j = 3$	0.035	0.015	0.010	0.024	0.008	0.092	0.056	0.064	0.076	0.046
$j = 4$	0.034	0.014	0.010	0.019	0.008	0.105	0.054	0.067	0.082	0.045

Given the impressively accurate behaviour exhibited by the location adjusted z-statistic in the last setting, it is worth checking whether a correction in scale might be helpful to further improve z-testing. A parametric bootstrap based on 1000 replicates has thus been employed to estimate the variance of $\widehat{T}^{j,*}$ $(j = 1, 2, 3, 4)$ on each simulated sample. Then, by standard implementation of the scale adjustment, the bootstrap scale-corrected z-statistic $\widehat{T}^{j,*}_{boot}$ has been obtained. Estimated rejection probabilities of the corresponding test at level $\alpha = 0.01, 0.05$ can be found in Table 2.3, which aids comparison with the best performers of the previous analysis. The scale correction of the location adjusted z-statistic surely succeeds in enhancing the agreement of the empirical rejection probability of the z-test to its nominal level, especially when $\alpha = 0.05$. Moreover, it might be of interest to note that empirical rejection probabilities based on $\widehat{T}^{j,*}$ are always larger than those based on $\widehat{T}^{j,*}_{boot}$, hence the variance of the location adjusted z-statistic must exceed 1. To conclude, the adoption of bootstrap certainly adds some computational burden to the Wald procedure, yet appears to assure a performance of the location adjusted z-statistic comparable to second-order tests within the Gamma regression framework.

Table 2.3: Empirical rejection probabilities at nominal levels $\alpha = 0.01, 0.05$ of the two-tailed tests related to $\widehat{T}^{j,*}$, $\widehat{T}^{j,*}_{boot}$, Z^j_{uP} and $Z^{j,*}_P$ $(j = 1, 2, 3, 4)$ in the *clotting* example. Figures are based on a simulation study with 2000 replications and 1000 bootstrap iterations.

	$\alpha = 0.01$				$\alpha = 0.05$			
	$\widehat{T}^{j,*}$	$\widehat{T}^{j,*}_{boot}$	Z^j_{uP}	$Z^{j,*}_P$	$\widehat{T}^{j,*}$	$\widehat{T}^{j,*}_{boot}$	Z^j_{uP}	$Z^{j,*}_P$
$j = 1$	0.016	0.011	0.006	0.006	0.059	0.051	0.070	0.051
$j = 2$	0.015	0.014	0.010	0.008	0.060	0.053	0.071	0.052
$j = 3$	0.015	0.014	0.010	0.008	0.056	0.048	0.064	0.046
$j = 4$	0.014	0.012	0.010	0.008	0.054	0.047	0.067	0.045

Poisson log-linear model

Consider now the following simulation setting. For each $i = 1, \ldots, 8$, covariates x_{i1} and x_{i2} are independently drawn from the $N(0, 1)$ and $Bern(0.6)$ distributions, respectively. Responses y_i are thus generated as realizations of Poisson random variables with mean $\mu_i = \exp(\beta_{01} + \beta_{02}x_{i1} + \beta_{03}x_{i2})$, where $\beta_{01} = 1$, $\beta_{02} = 1$ and $\beta_{03} = 2$. Datasets of larger size $n = 16, 32, 64$ are also created using the same original set of covariates. Rejection probabilities of the usual tests for $H_0 : \beta_j = \beta_{0j}$ $(j = 1, 2, 3)$ at several significance levels are then estimated by means of 5000 iterations for each sample size n. Table 2.4 presents such results for theoretical values of $\alpha = 0.01, 0.05$, whereas Table 2.5 deals with greater nominal levels $\alpha = 0.1, 0.2$. Under this scenario, the number of simulation trials is increased because less variation in testing performance may be observed among the various statistics. For instance, unlike what seen in the studies concerning the Gamma regression, here $Z_P^{j;*}$ is not outclassing the other competitors. Moreover, even the standard Wald test proves to be quite reliable, thus the room for refinement due to the location adjustment is not as large as before. Nevertheless, the experiment suggests that some profitable effects are still appreciable, especially as α grows and also for moderate values of n.

2.6 Discussion and further work

The fundamental idea behind this chapter, introduced in Section 2.1, has been to improve first-order Wald inference on small-to-moderate samples in regression settings by adjusting the null moments of the z-statistic. Because such a method is not guaranteed to succeed in increasing the overall agreement between the null distribution of the pivot and the standard normal distribution, several scenarios were taken into consideration to verify the actual usefulness of this approach.

Section 2.2 dealt with some motivating examples of our research. In simple frameworks with scalar global parameter, obtaining explicit asymptotic expansions for the mean and variance of the z-statistic was shown to be not so demanding. The location-scale adjustment seems particularly effective in the exponential case: the normal approximation to the distribution of the adjusted z-statistic is critically improved with respect to that of the ordinary version and is even more accurate than that of the score statistic, for all the sample sizes considered. In the Poisson setting the location-scale adjusted z-statistic performs in a dubious way, while under the logistic model the corresponding test confirmed to be typically more reliable than the ordinary one, although limitations of its performance connected with the correction in variance cannot be denied.

Table 2.4: Empirical rejection probabilities at nominal levels $\alpha = 0.01, 0.05$ of the two-sided tests related to $\widehat{T}^j$, its location adjusted version $\widehat{T}^{j,*}$, the profile score statistic Z^j_{uP}, the profile likelihood ratio statistic Z^j_P and its modification $Z^{j,*}_P$ ($j = 1, 2, 3$) in the Poisson log-linear model, estimated by a study based on 5000 simulated datasets of size $n = 8, 16, 32, 64$.

	$\alpha = 0.01$					$\alpha = 0.05$				
$n = 8$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.011	0.011	0.011	0.011	0.012	0.048	0.050	0.049	0.053	0.053
$j = 2$	0.009	0.010	0.010	0.010	0.012	0.048	0.049	0.049	0.051	0.053
$j = 3$	0.009	0.009	0.009	0.010	0.014	0.047	0.048	0.049	0.048	0.053
$n = 16$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.009	0.009	0.010	0.010	0.014	0.043	0.044	0.044	0.046	0.048
$j = 2$	0.008	0.008	0.008	0.007	0.011	0.045	0.046	0.046	0.045	0.049
$j = 3$	0.010	0.010	0.010	0.011	0.014	0.047	0.047	0.047	0.046	0.049
$n = 32$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.008	0.008	0.008	0.008	0.012	0.052	0.053	0.053	0.051	0.057
$j = 2$	0.008	0.008	0.008	0.007	0.013	0.044	0.044	0.044	0.046	0.050
$j = 3$	0.010	0.011	0.011	0.010	0.016	0.048	0.048	0.048	0.047	0.054
$n = 64$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.010	0.009	0.010	0.010	0.016	0.044	0.044	0.045	0.045	0.051
$j = 2$	0.013	0.013	0.013	0.012	0.018	0.052	0.052	0.052	0.052	0.058
$j = 3$	0.012	0.012	0.012	0.012	0.016	0.047	0.047	0.047	0.048	0.052

In Section 2.3 a convenient way to implement the location adjustment of the z-statistic under general regression scenarios was presented. The core intuition of viewing the combinant as an estimator of a reparametrization permits the proposed approach to enjoy the simplicity of original Wald-type inference. Indeed, the necessary ingredients to compute the location-adjusted z-statistic are easily obtainable from standard output of routines for fitting regression models. As a result, the computational effort implied by the procedure is equal to that implied by classical z-testing. We remark also that the same basic technique may be adopted to adjust z-statistics which use the observed information for the estimates' standard errors.

In Section 2.4 advantage was taken again of the single-parameter setting in order to study some theoretical properties of the location adjusted z-statistic and to evaluate its testing performance in a realistic situation. The asymptotic comparison between the two versions of the z-statistic did not resulted in a comprehensive pattern of difference in variability. For sure this analysis deserves to be further developed, both analytically and

TABLE 2.5: Empirical rejection probabilities at nominal levels $\alpha = 0.1, 0.2$ of the two-sided tests related to $\widehat{T}^j$, its location adjusted version $\widehat{T}^{j,*}$, the score statistic Z^j_{uP}, the likelihood ratio statistic Z^j_P and its modification $Z^{j,*}_P$ ($j = 1, 2, 3$) in the Poisson log-linear model, estimated by a study based on 5000 simulated datasets of size $n = 8, 16, 32, 64$.

	$\alpha = 0.1$					$\alpha = 0.2$				
$n = 8$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.100	0.101	0.102	0.105	0.106	0.206	0.208	0.207	0.212	0.210
$j = 2$	0.099	0.101	0.101	0.101	0.103	0.193	0.195	0.194	0.197	0.198
$j = 3$	0.095	0.098	0.097	0.102	0.104	0.193	0.196	0.194	0.198	0.201
$n = 16$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.092	0.094	0.093	0.092	0.097	0.193	0.194	0.194	0.193	0.197
$j = 2$	0.090	0.091	0.090	0.092	0.096	0.185	0.187	0.186	0.185	0.191
$j = 3$	0.096	0.097	0.097	0.099	0.101	0.198	0.200	0.199	0.200	0.202
$n = 32$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.100	0.100	0.100	0.101	0.106	0.197	0.198	0.197	0.198	0.204
$j = 2$	0.096	0.097	0.097	0.096	0.101	0.197	0.198	0.197	0.198	0.202
$j = 3$	0.099	0.100	0.099	0.102	0.108	0.202	0.204	0.203	0.204	0.210
$n = 64$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$	$\widehat{T}^j$	$\widehat{T}^{j,*}$	Z^j_{uP}	Z^j_P	$Z^{j,*}_P$
$j = 1$	0.092	0.092	0.092	0.093	0.100	0.194	0.194	0.194	0.196	0.202
$j = 2$	0.099	0.099	0.099	0.099	0.105	0.199	0.199	0.199	0.200	0.206
$j = 3$	0.093	0.094	0.093	0.093	0.097	0.192	0.192	0.192	0.191	0.196

empirically. Within the problem of inference on a binomial proportion, the behaviour of the location adjusted z-statistic was not found as satisfying as in the one-parameter models examined in Section 2.2.4. Determining whether the presence of a bounded parameter space may reduce the efficacy of the suggested approach appears then helpful.

Section 2.5 was devoted instead to the location adjustment of z-statistics in GLMs. Such prominent modeling framework is in fact especially suited to the application of our method. Among the practical aspects which contribute to further ease the steps of calculation, the existence of closed-form expressions for the bias of ML estimators is probably the most notable (Cordeiro and McCullagh, 1991).

The performance of the adjusted z-test in this context was illustrated through some simulation studies. Results relating to the Gamma regression are very remarkable: the location adjusted z-statistic always exhibits more adequate rejection probabilities than its direct competitor. For smaller samples, the adjusted z-test seems even more reliable than the profile likelihood ratio test and, in some cases, than the profile score test. Notice that, contrary to our proposal, both the latter require the constrained ML fit under the null hypothesis in order to be obtained. The testing accuracy of the location adjusted z-statistic was also shown to be comparable to that of higher-order tests when a bootstrap is employed for correcting its scale. Beyond any doubt, the bootstrap implementation makes the method much more intensive from a computational standpoint. It would certainly be preferable to find a simpler way to perform the scale adjustment of the z-statistic, similar to that used for centering its location.

Under the Poisson log-linear model, simulation evidence in support of the better performance of the location adjusted z-statistic was not as strong as for the Gamma regression case. However, the minor discrepancies in the empirical rejection probabilities of the two variants of the z-test allow to conclude that the adjustment in location is rather effective in this setting as well.

Of course, both the findings and the limitations of our study give rise to the need for further work into this subject. Some open problems have already been mentioned above, but there are more questions still left unanswered. Below, we delineate the main future directions of research in the form of a list:

i) Elaborate on the analysis in Section 2.4.1 by comparing the variances of the standard and location adjusted z-statistics in special simple model settings, like those of Section 2.2.4.

ii) Extend the variance analysis in Section 2.4.1 to the case of multidimensional parameter.

iii) Derive asymptotic (e.g. Edgeworth, Cornish-Fisher) expansions for the distributions of the standard and location adjusted z-statistics to formally establish whether the normal approximation is improved by the adjustment in location.

iv) Develop a power analysis to compare the distributions of the standard and location adjusted z-statistics under the alternative hypothesis.

v) Perform other Monte Carlo experiments, involving both real and simulated datasets, to empirically test the relative performance in the GLMs framework of the standard and location adjusted z-statistics, even with regard to the other likelihood-based pivots considered in Section 2.5.2. In particular, consider Poisson and binomial distributions of the response variable.

vi) Derive the location adjustment and empirically test the relative performance of the standard and location adjusted z-statistics under general regression scenarios, like the Cox proportional hazards and Beta regression models.

vii) Explore the possibility of implementing a fairly simple scale adjustment of the z-statistic along with the proposed correction in location.

viii) Investigate ways to adopt the same general approach with other test statistics, e.g. log-likelihood ratio or score statistics.

ix) Consider the potential application of the methodology suggested to p-values and/or rejection probabilities of the z-statistic, rather than to the pivot itself. In fact, at a given significance level of the test, such quantities may be viewed in their turn as model reparametrizations.

Chapter 3

Monte Carlo modified profile likelihood for clustered data

3.1 Introduction

The modified profile likelihood (MPL) (Barndorff-Nielsen, 1983) was introduced as prime example among adjusted profile likelihoods in Section 1.3.4. Unfortunately, the great beneficial impact of its employment can be directly observed only within the families of full exponential and composite group models, where the explicit derivation of an ancillary statistic is either unnecessary or practically possible.

In Chapter 1, we saw that the approximation owed to Severini (1998b) to this pseudo-likelihood function helps to overcome most of those computational difficulties, leaning on expected values asymptotically equivalent to the sample space derivatives involved in the original version of the MPL. Such expedient has thus sensitively extended the scope of this inferential instrument. Nevertheless, it is not complicated to check that covariances between score components like those present in Severini's modification may still not be readily available for a number of statistical problems.

The increasing complexity of phenomena nowadays dealt with is probably the main reason of the unquestioned current dissemination in all applied areas of clustered data, also known as grouped data, longitudinal data, stratified data or panel data (Hsiao, 2007). In Section 1.4 emphasis was placed on the fact that, due to their singular structure, datasets under those denominations are typically analyzed through statistical models intrinsically connoted by the incidental parameters problem. This character, more specifically, has to do with the usual choice of capturing the unobserved heterogeneity across groups via cluster-specific nuisance parameters, commonly named individual effects. Specifications of such type, especially popular in econometrics, are referred to as fixed effects models, in opposition to the so-called random effects models. The latter, as their title suggests, on one hand enable to get around Neyman & Scott problems by considering those group features as random variables, on the other introduce quite serious complications. To cite a few, the selection of some suitable underlying distribution for the implicit individual effects and the assumption of their incorrelation with the regressors (Lancaster, 2000). The last rather unrealistic postulate, in particular, drives the most widespread decision to opt for fixed effects models, which do not constrain the dependence of the distinguishing cluster-related traits on covariates.

The special significance held by the MPL for clustered data is then apparent. Based on what shown in Section 1.4 with reference to the basic setup (1.13) of fixed effects models, this adjustment to the profile likelihood can considerably refine ordinary inferential accuracy in samples where the total number of groups is much larger than the single cluster size. It would thus be useful to test whether similar results are retained in the presence of nonstandard modeling and/or sampling assumptions. To such aim, in the next section an automatic method to compute Severini's MPL even in those unconventional situations will be presented.

3.2 Monte Carlo approximation to Severini's modified profile likelihood

Consider, as done in Section 1.4.2, clustered observations subdivided in N groups of balanced size T. The hypothesis of independence among distinct clusters remains valid, yet here sampling units within groups are allowed to be correlated with each other. Hence, a general model with incidental parameters is now better expressed by

$$Y_{it}|X_{it} = x_{it} \sim p_{Y_{it}|X_{it}}(y_{it}|x_{it};\psi,\lambda_i), \qquad i=1,\ldots,N, \quad t=1,\ldots,T, \tag{3.1}$$

to accommodate also dynamic specifications where the index t runs over consecutive time periods and the temporal evolution of the dependent variable is explained by including in the vector of covariates x_{it} previously recorded responses belonging to the same cluster.

The version of the MPL proposed by Severini to approximate the original function of Barndorff-Nielsen can be found in (1.12). Under the assumption of independent groups, we have $\widetilde{M}(\psi) = \sum_{i=1}^{N} \widetilde{M}_i(\psi)$ where

$$\widetilde{M}_i(\psi) = \frac{1}{2}\log j_{\lambda_i\lambda_i}(\hat{\theta}_\psi) - \log I_{\lambda_i\lambda_i}(\hat{\theta}_\psi;\hat{\theta}), \qquad i=1,\ldots,N. \tag{3.2}$$

The quantity $j_{\lambda_i \lambda_i}$ is simply the (i,i)th element in the diagonal block $j_{\lambda\lambda}$ of the observed information, while $I_{\lambda_i \lambda_i}(\hat{\theta}_\psi; \hat{\theta}) = E_{\hat{\theta}}\big\{l_{\lambda_i}(\hat{\theta}_\psi) l_{\lambda_i}(\hat{\theta})\big\}$ is the (i,i)th element in the diagonal matrix of expected values $I_{\lambda\lambda}(\hat{\theta}_\psi; \hat{\theta})$.

Section 3.1 already anticipated that, for a variety of reasons related to the complexity of the model under study, a closed-form expression of $I_{\lambda_i \lambda_i}(\hat{\theta}_\psi; \hat{\theta})$ cannot be always obtained. When this happens, a possible strategy consists in rather using the following Monte Carlo approximation based on R replicates:

$$I^*_{\lambda_i \lambda_i}(\hat{\theta}_\psi; \hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} l^r_{\lambda_i}(\hat{\theta}_\psi) l^r_{\lambda_i}(\hat{\theta}), \qquad i = 1, \dots, N, \tag{3.3}$$

where $l^r_{\lambda_i}$ is the scalar partial score computed for the rth sample $y^r = (y^r_{it})$ $(r = 1, \dots, R)$ randomly generated from the ML fit of model (3.1), thus by setting $(\psi, \lambda) = (\hat{\psi}, \hat{\lambda})$. Note that calculation of $I^*_{\lambda_i \lambda_i}(\hat{\theta}_\psi; \hat{\theta})$ only requires to derive the score function l_{λ_i} and to simulate from the assumed model, with no need of additional fitting. Indeed, $\hat{\theta}$ and $\hat{\theta}_\psi$ in (3.3) are the estimates derived from the observed data. This makes the procedure far less computationally expensive than a standard bootstrap. Moreover, the execution time of such solution is not particularly influenced by the value of T and the number of replications R usually does not need to exceed 500 for a reasonably adequate estimation of ψ, as attested by sensitivity analyses not reported here.

The principal advantage of this Monte Carlo strategy is by all means its potential broad applicability. Already experimented by Bartolucci *et al.* (2016), it allowed the MPL of Severini to prove its competitiveness with econometric inferential methods in estimating dynamic fixed effects models for binary panel data. In what follows, we will make use of the same technique in order to calculate $l_{\widetilde{M}}(\psi)$ and verify its superiority with respect to usual ML procedures under different special scenarios. Of course, the focus shall be on models with incidental parameters for which explicit formulation of (3.2) is either impossible or too demanding.

For ease of reference, from now on Severini's approximation to the MPL computed by Monte Carlo simulation will be called Monte Carlo MPL and denoted by $L_{\widetilde{M}^*}(\psi)$. The corresponding log-likelihood function is then $l_{\widetilde{M}^*}(\psi) = \log L_{\widetilde{M}^*}(\psi) = l_P(\psi) + \widetilde{M}^*(\psi)$, where the modification term may be written as

$$\widetilde{M}^*(\psi) = \sum_{i=1}^{N} \widetilde{M}^*_i(\psi) = \sum_{i=1}^{N} \left\{ \frac{1}{2} \log j_{\lambda_i \lambda_i}(\hat{\theta}_\psi) - \log I^*_{\lambda_i \lambda_i}(\hat{\theta}_\psi; \hat{\theta}) \right\}, \tag{3.4}$$

with $I^*_{\lambda_i \lambda_i}(\hat{\theta}_\psi; \hat{\theta})$ defined in (3.3).

3.3 Nonstationary AR(1) model

3.3.1 Setup and background

The first object of our analysis belongs to the class of linear dynamic models for continuous panel data, largely employed in the field of econometrics. Specifically, let us consider the nonstationary version of the first-order autoregressive specification

$$Y_{it}|Y_{i,t-1} = y_{i,t-1} \sim N(\lambda_i + \rho y_{i,t-1}, \sigma^2), \qquad i = 1, \dots, N, \quad t = 1, \dots, T, \tag{3.5}$$

with $y_0 = (y_{10}, \dots, y_{N0})$ vector of unrestricted and given initial conditions. Here, the structural parameter is $\psi = (\rho, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$ and $\lambda = (\lambda_1, \dots, \lambda_N) \in \mathbb{R}^N$ represents the nuisance component of individual effects. The lack of stationarity of the stochastic process Y_{it} in each group implies the temporal variation of its mean or its autocovariance function, i.e. the covariance of the response with itself at pairs of time points. As a consequence, the autoregressive parameter ρ is left free to equal or exceed unity and the fixed vector y_0 does not need to meet any specific requirement, so that the likelihood function is expressed by conditioning on these N starting values. In order to facilitate the presentation, both exogenous covariates and further lagged responses $y_{i,t-l}\,(l > 1)$ are excluded from the set of model regressors; however, no additional difficulties would be encountered in applying the proposed methodology otherwise.

The incidental parameters problem occurring in the analogue stationary AR(1) model has been addressed in the statistical literature several times. Particularly Cruddas *et al.* (1989) proved that, if the first two moments of the process are assumed to stay constant over time, an accurate marginal likelihood for ψ not only exists but also is asymptotically equivalent to the first modification of $L_P(\psi)$ introduced by Barndorff-Nielsen. Furthermore, in Example 1 of Bartolucci *et al.* (2016) it is shown how Severini's MPL, obtained upon orthogonal interest-preserving transformation, coincides in fact with the conditional approximate likelihood of Cox and Reid. Not surprisingly, also econometricians showed interest in this issue and produced a proliferation of possible solutions to improve standard ML inference in general fixed effects dynamic models for panel data. Among the most successful are, for instance, the instrumental-variable (Hsiao, 2003, Section 4.3.3.c) and the generalized method of moments (Arellano and Bond, 1991) estimators for ψ. One latest proposition which also allows for a multivariate response is the bias-corrected estimator of Dhaene and Jochmans (2016), specially tailored for macroeconomic settings with $N = O(T)$.

Here, though, a great deal of attention is paid to the nonstationarity assumption of model (3.5). Indeed, analytical derivation of $I_{\lambda_i\lambda_i}(\hat{\theta}_\psi; \hat{\theta})$ in this case would be possible but quite tedious, and Monte Carlo approximation dramatically reduces the amount of effort demanded to use Severini's modification. Moreover, we are specifically concerned with datasets where T is much smaller than N, meaning with situations where $l_P(\psi)$ exhibits its worst performance. Even estimation of ψ under these conditions was already investigated applying procedures alternative to the MPL. By way of example, inference in general autoregressions of order l was thoroughly examined in Dhaene and Jochmans (2014). As the bias of $l_P(\psi)$ in such models was not found to depend on the incidental parameters, an adjusted profile log-likelihood was obtained through integration of the recentered score function. While exploring the various connections of their work with past publications on the topic, the authors gave evidence of the equivalence existing between their solution and that of Lancaster (2002, Section 3) when $l = 1$. From a purely statistical perspective, the latter proposed a Bayesian strategy grounded on the preliminary orthogonalization of λ to the structural component. This served to integrate out the individual effects from the likelihood, so as to derive a marginal posterior density with consistent mode for ψ. Such ensuing posterior distribution, besides being a special case of the adjustment to the profile likelihood prescribed by Dhaene and Jochmans (2014), was also proved to be the Bayesian counterpart of Cox and Reid's approximate conditional likelihood. Another approach to make inference on the autoregressive parameter in (3.5), diverse in essence but equivalent in substance to the last one, was later adopted by De Bin *et al.* (2015). Their results, obtained in a frequentist fashion via the integrated likelihood of Severini (2007), in fact closely agree with the findings in both Lancaster (2002) and Dhaene and Jochmans (2014).

3.3.2 Monte Carlo modified profile likelihood

It is easy to see that, under the hypothesis of independent groups, the log-likelihood of model (3.5) conditioned on the initial vector y_0 is

$$l(\theta) = -\frac{NT}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \lambda_i - \rho y_{i,t-1})^2. \tag{3.6}$$

Differentiation with respect to the ith incidental parameter leads to the scalar partial score function

$$l_{\lambda_i}(\theta) = l_{\lambda_i}(\psi, \lambda_i) = \frac{1}{\sigma^2}\sum_{t=1}^{T}(y_{it} - \lambda_i - \rho y_{i,t-1}), \qquad i = 1, \ldots, N,$$

and subsequent solution to the ith component of the likelihood equation $l_{\lambda_i}(\theta) = 0$ delivers the following constrained ML estimate of λ_i:

$$\hat{\lambda}_{i\psi} = \bar{y}_i - \rho\bar{y}_{i,-1} = \hat{\lambda}_{i\rho}, \tag{3.7}$$

where $\bar{y}_i = \sum_{t=1}^{T} y_{it}/T$ and $\bar{y}_{i,-1} = \sum_{t=0}^{T-1} y_{it}/T$. Clearly, the profile log-likelihood $l_P(\psi)$ is then obtained by replacement of λ_i with $\hat{\lambda}_{i\rho}$ in expression (3.6) for each $i = 1, \dots, N$. The next quantity needed for computing Severini's modification is immediately available from the derivative of the ith partial score with regard to λ_i, namely

$$j_{\lambda_i\lambda_i}(\hat{\theta}_\psi) = \frac{T}{\sigma^2}, \qquad i = 1, \dots, N,$$

whereas $I_{\lambda_i\lambda_i}(\hat{\theta}_\psi; \hat{\theta})$ requires more elaboration. The ML estimate of λ_i simply equals $\hat{\lambda}_i = \hat{\lambda}_{i\hat{\rho}} = \bar{y}_i - \hat{\rho}\bar{y}_{i,-1}$, where we have that

$$\hat{\rho} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}y_{i,t-1} - T\sum_{i=1}^{N}\bar{y}_i\bar{y}_{i,-1}}{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t-1}^2 - T\sum_{i=1}^{N}\bar{y}_{i,-1}^2} \tag{3.8}$$

is the ordinary least squares (OLS) estimate of the autoregressive parameter. Then, by adding and subtracting the same quantity $\rho\bar{y}_{i,-1}$, one can write

$$\begin{aligned}\hat{\lambda}_i &= \bar{y}_i - \rho\bar{y}_{i,-1} + \rho\bar{y}_{i,-1} - \hat{\rho}\bar{y}_{i,-1} \\ &= \hat{\lambda}_{i\rho} - (\hat{\rho} - \rho)\bar{y}_{i,-1}. \end{aligned} \tag{3.9}$$

Exploiting this last result with the aim of calculating $I_{\lambda_i\lambda_i}(\hat{\theta}_\psi; \hat{\theta})$, let us express the partial score evaluated at the constrained ML estimate in a more convenient way. In particular, we begin from

$$\begin{aligned} l_{\lambda_i}(\hat{\theta}_\psi) &= \frac{1}{\sigma^2}\sum_{t=1}^{T}\big(y_{it} - \hat{\lambda}_{i\rho} - \rho y_{i,t-1}\big) \\ &= \frac{1}{\sigma^2}\sum_{t=1}^{T}\big(y_{it} - \hat{\lambda}_{i\rho} + \hat{\lambda}_i - \hat{\lambda}_i - \rho y_{i,t-1} + \hat{\rho}y_{i,t-1} - \hat{\rho}y_{i,t-1}\big), \end{aligned} \tag{3.10}$$

where the second equality holds because we simultaneously sum to and subtract from the bracketed part both $\hat{\lambda}_i$ and $\hat{\rho}y_{i,t-1}$. Now, since manipulating (3.9) leads to

$$\hat{\lambda}_{i\rho} = \hat{\lambda}_i + (\hat{\rho} - \rho)\bar{y}_{i,-1},$$

by substitution of the latter expression in (3.10) it is not hard to obtain

$$\begin{aligned}
l_{\lambda_i}(\hat{\theta}_\psi) &= \frac{1}{\sigma^2}\left\{\sum_{t=1}^{T}\big(y_{it}-\hat{\lambda}_i-\hat{\rho}y_{i,t-1}\big)+T\big(\hat{\lambda}_i-\hat{\lambda}_{i\rho}\big)+\sum_{t=1}^{T}\big(\hat{\rho}-\rho\big)y_{i,t-1}\right\}\\
&= \frac{1}{\sigma^2}\big\{\hat{\sigma}^2 l_{\lambda_i}(\hat{\theta})+T\big(\hat{\lambda}_i-\hat{\lambda}_{i\rho}\big)+T\big(\hat{\rho}-\rho\big)\bar{y}_{i,-1}\big\}.
\end{aligned}$$

Then, the necessary expectation results equal to a linear function of ρ, and specifically to

$$\begin{aligned}
I_{\lambda_i\lambda_i}(\hat{\theta}_\psi;\hat{\theta}) &= E_{\hat{\theta}}\big\{l_{\lambda_i}(\hat{\theta}_\psi)l_{\lambda_i}(\hat{\theta})\big\}\\
&= \frac{1}{\sigma^2}E_{\hat{\theta}}\Big[\big\{\hat{\sigma}^2 l_{\lambda_i}(\hat{\theta})+T\big(\hat{\lambda}_i-\hat{\lambda}_{i\rho}\big)+T\big(\hat{\rho}-\rho\big)\overline{Y}_{i,-1}\big\}l_{\lambda_i}(\hat{\theta})\Big] \qquad (3.11)\\
&= \frac{1}{\sigma^2}\big\{\hat{\sigma}^2\widehat{E}_1+T\big(\hat{\rho}-\rho\big)\widehat{E}_2\big\},
\end{aligned}$$

with $\widehat{E}_1 = E_{\hat{\theta}}\big\{l^2_{\lambda_i}(\hat{\theta})\big\}$ e $\widehat{E}_2 = E_{\hat{\theta}}\big\{\overline{Y}_{i,-1}l_{\lambda_i}(\hat{\theta})\big\}$. Notice that, as the expected value (3.11) is computed with reference to the distribution $p_Y(y;\hat{\theta})$, the last equality sign applies because the quantities $\hat{\theta} = (\hat{\psi},\hat{\lambda})$ and $\hat{\theta}_\psi = (\psi,\hat{\lambda}_\psi)$ must be considered given and $E_{\hat{\theta}}\big\{l_{\lambda_i}(\hat{\theta})\big\} = 0$.

Although possible in principle, the analytical calculation of $\widehat{E}_1$ and $\widehat{E}_2$ is not straightforward in practice. Conversely, for the reasons discussed in Section 3.2, estimating $I_{\lambda_i\lambda_i}(\hat{\theta}_\psi;\hat{\theta})$ via Monte Carlo simulation represents an easily implementable solution. The MPL of Severini can then be employed to make inference on ψ in the autoregression for nonstationary panel data by replacing such expectation in its ith group-specific adjustment term (3.2) with the following empirical mean:

$$I^*_{\lambda_i\lambda_i}(\hat{\theta}_\psi;\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\left[\left\{\frac{1}{\sigma^2}\sum_{t=1}^{T}\big(y^r_{it}-\hat{\lambda}_{i\rho}-\rho y^r_{i,t-1}\big)\right\}\left\{\frac{1}{\hat{\sigma}^2}\sum_{t=1}^{T}\big(y^r_{it}-\hat{\lambda}_i-\hat{\rho}y^r_{i,t-1}\big)\right\}\right], \qquad (3.12)$$

where y^r_{it} $(i = 1,\dots,N, t = 1,\dots,T)$ is generated by model (3.5) with $(\psi,\lambda) = (\hat{\psi},\hat{\lambda})$, but the starting vector is kept unchanged, namely $y^r_0 = y_0$ for each $r = 1,\dots,R$. It can be worthwhile adding that, in this specific case, one alternative strategy for obtaining (3.11) could foresee analogue Monte Carlo approximations to $\widehat{E}_1$ and $\widehat{E}_2$, which may be derived just once because they involve $\hat{\theta}$ only. However, the overall computational cost of this procedure would be the same as that entailed by using (3.12), since the whole expected value in (3.11) would still need to be calculated for any different value of ψ.

3.3.3 Computational aspects

The global ML estimate $\hat{\theta}$ can be easily obtained in closed-form by applying the OLS method to the linear autoregression with normally distributed errors corresponding to the specification (3.5). As a consequence, the ML estimate for the variance parameter σ^2 is expressed by

$$\hat{\sigma}^2 = \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{\left(y_{it} - \hat{\lambda}_i - \hat{\rho} y_{i,t-1}\right)^2}{NT}, \tag{3.13}$$

where formulations of $\hat{\rho}$ and $\hat{\lambda}_i$ result directly from (3.8). On the contrary, maximization of $l_{\widetilde{M}^*}(\psi)$ to find the estimate $\hat{\psi}_{\widetilde{M}^*}$ usually has to be performed by means of numerical algorithms and estimated standard errors are obtained using the second derivative of the function at its maximum. Under this particular scenario, nevertheless, it is more convenient to derive $\hat{\sigma}^2_{\widetilde{M}^*}$ by evaluation of the explicit constrained estimate

$$\hat{\sigma}^2_{\rho,\widetilde{M}^*} = \hat{\sigma}^2_{\widetilde{M}^*}(\rho) = \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{\left(y_{it} - \hat{\lambda}_{i\rho} - \rho y_{i,t-1}\right)^2}{N(T-1)}$$

at $\hat{\rho}_{\widetilde{M}^*}$, i.e. the scalar solution to the optimization problem with objective function $l^{\rho}_{\widetilde{M}^*}(\rho) = l_{\widetilde{M}^*}\big(\rho, \hat{\sigma}^2_{\rho,\widetilde{M}^*}\big)$. Observe that, similarly, also $l_P(\psi)$ can be further profiled in order to get $l^{\rho}_P(\rho) = l_P(\rho, \hat{\sigma}^2_{\rho})$, where $\hat{\sigma}^2_{\rho}$ takes the form equivalent to (3.13), but with estimates $\hat{\rho}$ and $\hat{\lambda}_i$ replaced by ρ and $\hat{\lambda}_{i\rho}$ as in (3.7), respectively.

According to expression (3.11), for values of the autoregressive parameter beyond a certain threshold depending on $\hat{\rho}$ the expectation $I_{\lambda_i\lambda_i}(\hat{\theta}_\psi; \hat{\theta})$ is negative and $l_{\widetilde{M}}(\psi)$ is not computable, paralleling the integrated likelihood of De Bin *et al.* (2015). Therefore, in its turn, even the approximate expectation $I^*_{\lambda_i\lambda_i}(\hat{\theta}_\psi; \hat{\theta})$ can be smaller than or equal to 0 for not very large values of ρ. A potentially undefined modification term obviously poses a problem for the numerical optimization of $l^{\rho}_{\widetilde{M}^*}(\rho)$. In addition, as will emerge more clearly from the plots available in Section 3.3.4, the Monte Carlo MPL is found to reach its global maximum as $\rho \to +\infty$ for any sample size, in accordance with the distinct functions for inference on ψ studied in Lancaster (2002), Dhaene and Jochmans (2014) and De Bin *et al.* (2015). On such grounds, we choose to maximize $l^{\rho}_{\widetilde{M}^*}(\rho)$ by performing a one-dimensional search in a real bounded interval Υ through the algorithm implemented by the `R` function `optimize`. Specifically, adopting the same notation as Lancaster's (2002), $\Upsilon = (-\rho_l, \rho_u)$ with $\rho_l = \rho_u = 1.5$, since in general applications the autoregressive parameter is hardly observed to lie outside these extremes. The estimate resulting from local maximization of $l_{\widetilde{M}^*}(\psi)$ in this framework is then uniquely defined as $\hat{\psi}_{\widetilde{M}^*} = \big(\hat{\rho}_{\widetilde{M}^*}, \hat{\sigma}^2_{\widetilde{M}^*}\big)$, where $\hat{\rho}_{\widetilde{M}^*} = \arg\max_{\rho\in\Upsilon} l^{\rho}_{\widetilde{M}^*}(\rho)$ and $\hat{\sigma}^2_{\widetilde{M}^*} = \hat{\sigma}^2_{\hat{\rho}_{\widetilde{M}^*},\widetilde{M}^*}$. We refer to

Dhaene and Jochmans (2014) for a careful discussion about the conditions under which consistency of the local maximizer of their adjusted profile log-likelihood is achieved.

3.3.4 Simulation studies and numerical examples

In the present section, the accuracy of the Monte Carlo MPL in drawing inferences on ψ is assessed with regard to that of the standard profile likelihood through a series of simulations. More in detail, two main experiments based on $S = 2000$ iterations are performed, both considering datasets with $T = 4, 8, 16$ and $N = 250, 500, 1000$. The performance of $l_P(\psi)$ and $l_{\widetilde{M}^*}(\psi)$ is examined in respect of bias (B), median bias (MB), root mean squared error (RMSE) and median absolute error (MAE) of the corresponding estimators. Precisely, with specific reference to $\hat{\rho}$ we compute

$$\begin{aligned}
\mathrm{B}_{\hat{\rho}} &= \sum_{s=1}^{S} \left(\hat{\rho}^s - \rho\right)/S, \\
\mathrm{MB}_{\hat{\rho}} &= \left(\hat{\rho}^{(S/2)} + \hat{\rho}^{(S/2+1)}\right)/2 - \rho, \\
\mathrm{RMSE}_{\hat{\rho}} &= \sqrt{\sum_{s=1}^{S} \left(\hat{\rho}^s - \rho\right)^2/S}, \\
\mathrm{MAE}_{\hat{\rho}} &= \left(|\hat{\rho} - \rho|^{(S/2)} + |\hat{\rho} - \rho|^{(S/2+1)}\right)/2,
\end{aligned}$$

where ρ is the value of the autoregressive parameter used to simulate the S datasets, $\hat{\rho}^s$ is its ML estimate on the sth sample ($s = 1, \ldots, S$) and $x^{(s)}$ denotes the sth element in the vector of order statistics $(x^{(1)}, \ldots, x^{(S)})$, with $x^{(s_1)} \leq x^{(s_2)}$ for $s_1 < s_2$. Obviously, homologous quantities are obtained for $\hat{\sigma}^2, \hat{\rho}_{\widetilde{M}^*}$ and $\hat{\sigma}^2_{\widetilde{M}^*}$. The empirical standard deviation (SD) of the various estimates is also reported. In the habitual way, considering again $\hat{\rho}$ for illustration, one may write

$$\mathrm{SD}_{\hat{\rho}} = \sum_{s=1}^{S} \left(\hat{\rho}^s - \bar{\hat{\rho}}\right)^2/(S-1), \qquad \bar{\hat{\rho}} = \sum_{s=1}^{S} \hat{\rho}^s/S.$$

In addition, the ratio SE/SD of $\hat{\rho}$ and $\hat{\rho}_{\widetilde{M}^*}$, where SE stands for the average over simulations of likelihood-based estimated standard errors, and empirical coverages of 0.95 Wald confidence intervals (CI) for ρ are shown. Note that, like remarked by Bartolucci *et al.* (2016), the large values of N examined here ensure adequacy of the quadratic approximation around the maximum of both $l_P(\psi)$ and $l_{\widetilde{M}^*}(\psi)$, hence the generally more accurate coverages derived by inversion of the log-likelihood ratio statistic would be in this case substantially identical.

The two simulation setups differ only in the true value of the autoregressive parameter set to generate the samples from model (3.5): in the first $\rho = 0.5$, while in the second $\rho = 0.9$. For what concerns the remaining parameters, the conditional variance of the response variable is $\sigma^2 = 1$ and the individual effects are independently drawn from a $N(1,1)$ distribution, following the example of Lancaster (2002). In every simulated dataset, all N initial observations in the vector y_0 are fixed equal to 0 with no loss of generality, since this is equivalent to interpret each y_{it} as $y_{it} - y_{i0}$ and each λ_i as $\lambda_i - y_{i0}(1-\rho)$ $(t = 1, \ldots, T, i = 1, \ldots, N)$ (Lancaster, 2002). Lastly, the number of Monte Carlo replicates employed to compute $l_{\widetilde{M}^*}(\psi)$ is $R = 500$.

Table 3.1: Inference on $\rho = 0.5$ in the nonstationary AR(1) model for panel data. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
250	4	$l_P(\psi)$	-0.186	-0.186	0.025	0.187	0.186	0.879	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.020	0.018	0.037	0.042	0.028	0.921	0.915
	8	$l_P(\psi)$	-0.114	-0.115	0.018	0.116	0.115	0.921	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.002	0.002	0.020	0.020	0.013	0.989	0.942
	16	$l_P(\psi)$	-0.070	-0.070	0.013	0.071	0.070	0.960	0.000
		$l_{\widetilde{M}^*}(\psi)$	-0.000	0.000	0.014	0.014	0.009	1.002	0.944
500	4	$l_P(\psi)$	-0.184	-0.183	0.017	0.184	0.183	0.896	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.018	0.019	0.025	0.031	0.022	0.952	0.881
	8	$l_P(\psi)$	-0.113	-0.113	0.013	0.114	0.113	0.902	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.002	0.002	0.014	0.014	0.010	0.972	0.943
	16	$l_P(\psi)$	-0.069	-0.069	0.009	0.069	0.069	0.983	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.000	0.000	0.009	0.009	0.007	1.029	0.959
1000	4	$l_P(\psi)$	-0.187	-0.187	0.013	0.187	0.187	0.879	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.019	0.018	0.019	0.026	0.019	0.923	0.795
	8	$l_P(\psi)$	-0.115	-0.115	0.009	0.115	0.115	0.919	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.002	0.002	0.010	0.010	0.007	0.987	0.948
	16	$l_P(\psi)$	-0.070	-0.070	0.007	0.070	0.070	0.935	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.000	0.000	0.007	0.007	0.005	0.977	0.940

Inferential results for ρ and σ^2 of the first study are displayed in Tables 3.1 and 3.2, respectively. Similar comments as in Bartolucci *et al.* (2016) can be made. In all cases, no significant differences between bias and median bias of the same estimator are observed but the improvement determined by using the Monte Carlo MPL in this sense is undeniable. Consistently with the theory, the bias does not vary with N but decreases as T increases, whereas the root mean squared error depends on both indexes. Empirical coverage probabilities of confidence intervals for the autoregressive parameter based on $l_{\widetilde{M}^*}(\psi)$ are generally accurate, with larger departures from the nominal level occurring when $T = 4$. Such conspicuous refinements to the poor interval estimation supplied by $l_P(\psi)$ mainly stem from bias reduction. Yet some correction in curvature also takes place, being SE/SD for the Monte Carlo MPL typically closer to 1 than for the ordinary profile likelihood.

TABLE 3.2: Inference on $\sigma^2 = 1$ in the nonstationary AR(1) model for panel data with $\rho = 0.5$. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE
250	4	$l_P(\psi)$	-0.300	-0.301	0.036	0.303	0.301
		$l_{\widetilde{M}^*}(\psi)$	0.013	0.011	0.060	0.062	0.041
	8	$l_P(\psi)$	-0.147	-0.148	0.029	0.150	0.148
		$l_{\widetilde{M}^*}(\psi)$	-0.000	-0.001	0.035	0.035	0.024
	16	$l_P(\psi)$	-0.071	-0.071	0.022	0.074	0.071
		$l_{\widetilde{M}^*}(\psi)$	-0.001	-0.001	0.023	0.023	0.016
500	4	$l_P(\psi)$	-0.299	-0.299	0.026	0.300	0.299
		$l_{\widetilde{M}^*}(\psi)$	0.013	0.013	0.043	0.045	0.029
	8	$l_P(\psi)$	-0.147	-0.148	0.020	0.148	0.148
		$l_{\widetilde{M}^*}(\psi)$	-0.000	-0.001	0.024	0.024	0.017
	16	$l_P(\psi)$	-0.070	-0.070	0.015	0.072	0.070
		$l_{\widetilde{M}^*}(\psi)$	-0.000	-0.000	0.017	0.017	0.011
1000	4	$l_P(\psi)$	-0.300	-0.299	0.018	0.301	0.299
		$l_{\widetilde{M}^*}(\psi)$	0.013	0.014	0.030	0.033	0.022
	8	$l_P(\psi)$	-0.147	-0.147	0.015	0.147	0.147
		$l_{\widetilde{M}^*}(\psi)$	0.001	0.000	0.018	0.018	0.012
	16	$l_P(\psi)$	-0.070	-0.070	0.011	0.071	0.070
		$l_{\widetilde{M}^*}(\psi)$	-0.000	-0.000	0.012	0.012	0.008

Tables 3.3 and 3.4 illustrate instead results of the simulation experiment run with a true value of ρ approaching the boundaries of the stationary region $(-1, 1)$, particularly $\rho = 0.9$. Relative behaviours of the two methods for estimating the structural component are basically in line with those analyzed in the previous study. Perhaps, one may argue that here the general improvements originating from the employment of

$l_{\widetilde{M}^*}(\psi)$ are somewhat milder than when the autoregressive parameter is farther away from nonstationariety. This observation can be referred both to bias and, mostly, to empirical coverages of Wald confidence intervals for ρ. Nonetheless, the quality of MPL-based inference remains unquestionably higher than that achieved through standard ML techniques.

TABLE 3.3: Inference on $\rho = 0.9$ in the nonstationary AR(1) model for panel data. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
250	4	$l_P(\psi)$	-0.130	-0.130	0.018	0.131	0.130	0.894	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.022	0.021	0.028	0.036	0.024	0.899	0.871
	8	$l_P(\psi)$	-0.051	-0.051	0.008	0.052	0.051	0.922	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.003	0.003	0.010	0.010	0.007	0.976	0.933
	16	$l_P(\psi)$	-0.022	-0.023	0.004	0.023	0.023	0.957	0.001
		$l_{\widetilde{M}^*}(\psi)$	0.000	0.000	0.005	0.005	0.003	1.003	0.950
500	4	$l_P(\psi)$	-0.128	-0.128	0.013	0.129	0.128	0.905	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.021	0.020	0.019	0.028	0.021	0.928	0.774
	8	$l_P(\psi)$	-0.050	-0.050	0.006	0.050	0.050	0.933	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.003	0.003	0.007	0.007	0.005	0.980	0.928
	16	$l_P(\psi)$	-0.022	-0.022	0.003	0.022	0.022	0.957	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.001	0.001	0.003	0.003	0.002	1.001	0.946
1000	4	$l_P(\psi)$	-0.131	-0.131	0.009	0.131	0.131	0.895	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.021	0.021	0.014	0.025	0.021	0.909	0.612
	8	$l_P(\psi)$	-0.051	-0.051	0.004	0.051	0.051	0.923	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.003	0.003	0.005	0.006	0.004	0.969	0.884
	16	$l_P(\psi)$	-0.022	-0.022	0.002	0.022	0.022	0.923	0.000
		$l_{\widetilde{M}^*}(\psi)$	0.001	0.001	0.002	0.002	0.002	0.968	0.930

Figures 3.1 and 3.2 graphically show the different tendencies of the functions described in Section 3.3.3, meaning $l_P^\rho(\rho)$ and $l_{\widetilde{M}^*}^\rho(\rho)$, in their relative version. Specifically, quantities in the former figure are referred to samples generated from model (3.5) with $\rho = 0.5$, while those in the latter are computed starting from datasets simulated by fixing $\rho = 0.9$. These plots substantially confirm the results of simulations discussed so far. In each of them, the maximum of the profile log-likelihood is significantly smaller than the true value of the autoregressive parameter, corresponding to the vertical line. For this main reason, such value never belongs to the 0.95 confidence region defined

by inversion of the profile likelihood ratio statistic and marked by the horizontal line. This may also be attributed to the accentuated curvature of $l_P^\rho(\rho)$. Conversely, the local maximization of the Monte Carlo MPL yields to adequate both point and interval estimation of ρ. The unusual trend of $l_{\widetilde{M}^*}^\rho(\rho)$, whose global maximizer lies at infinity, was already anticipated in Section 3.3.3 and can now be directly checked. Indeed, the absence of restrictions on the initial conditions y_{i0} $(i = 1, \ldots, N)$ causes the Monte Carlo MPL to be re-increasing, sometimes already in the stationary parameter region (Dhaene and Jochmans, 2014). Quite interestingly, especially for small values of T and larger values of ρ, $l_{\widetilde{M}^*}^\rho(\rho)$ may also be everywhere increasing. Two representations of this event with positive probability are given by Figure 3.3.

Table 3.4: Inference on $\sigma^2 = 1$ in the nonstationary AR(1) model for panel data with $\rho = 0.9$. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE
250	4	$l_P(\psi)$	-0.297	-0.298	0.036	0.299	0.298
		$l_{\widetilde{M}^*}(\psi)$	0.021	0.017	0.062	0.066	0.043
	8	$l_P(\psi)$	-0.144	-0.145	0.029	0.147	0.145
		$l_{\widetilde{M}^*}(\psi)$	0.002	0.000	0.035	0.035	0.024
	16	$l_P(\psi)$	-0.070	-0.070	0.021	0.074	0.070
		$l_{\widetilde{M}^*}(\psi)$	-0.001	-0.001	0.023	0.023	0.016
500	4	$l_P(\psi)$	-0.295	-0.295	0.026	0.297	0.295
		$l_{\widetilde{M}^*}(\psi)$	0.020	0.019	0.044	0.048	0.032
	8	$l_P(\psi)$	-0.144	-0.144	0.020	0.145	0.144
		$l_{\widetilde{M}^*}(\psi)$	0.001	0.001	0.024	0.025	0.017
	16	$l_P(\psi)$	-0.069	-0.069	0.015	0.071	0.069
		$l_{\widetilde{M}^*}(\psi)$	0.000	0.000	0.017	0.017	0.011
1000	4	$l_P(\psi)$	-0.296	-0.296	0.018	0.297	0.296
		$l_{\widetilde{M}^*}(\psi)$	0.021	0.021	0.031	0.037	0.026
	8	$l_P(\psi)$	-0.144	-0.144	0.015	0.144	0.144
		$l_{\widetilde{M}^*}(\psi)$	0.003	0.002	0.018	0.018	0.012
	16	$l_P(\psi)$	-0.070	-0.070	0.011	0.070	0.070
		$l_{\widetilde{M}^*}(\psi)$	0.000	0.000	0.012	0.012	0.008

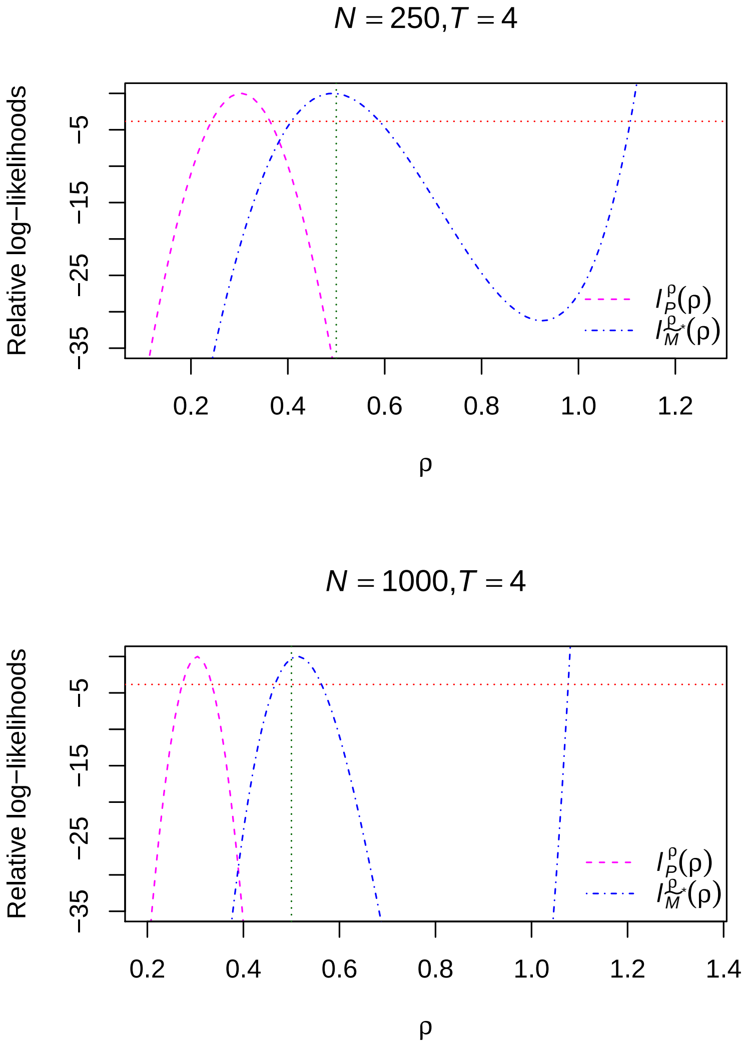

FIGURE 3.1: Relative log-likelihoods for two datasets generated under the nonstationary AR(1) model with $\rho = 0.5$. The vertical line indicates the true value of the autoregressive parameter, while the horizontal line gives the 0.95 confidence intervals for ρ based on the profile and modified profile log-likelihood ratio statistics.

3.4 Models for binary data with missing values

3.4.1 Introduction

These days missing data are the rule rather than the exception in quantitative research analysis. It comes then as no surprise that such a great deal of literature has been produced on the topic since the early 1970s, when opportunities given by the technological developments in computer science could be fruitfully seized.

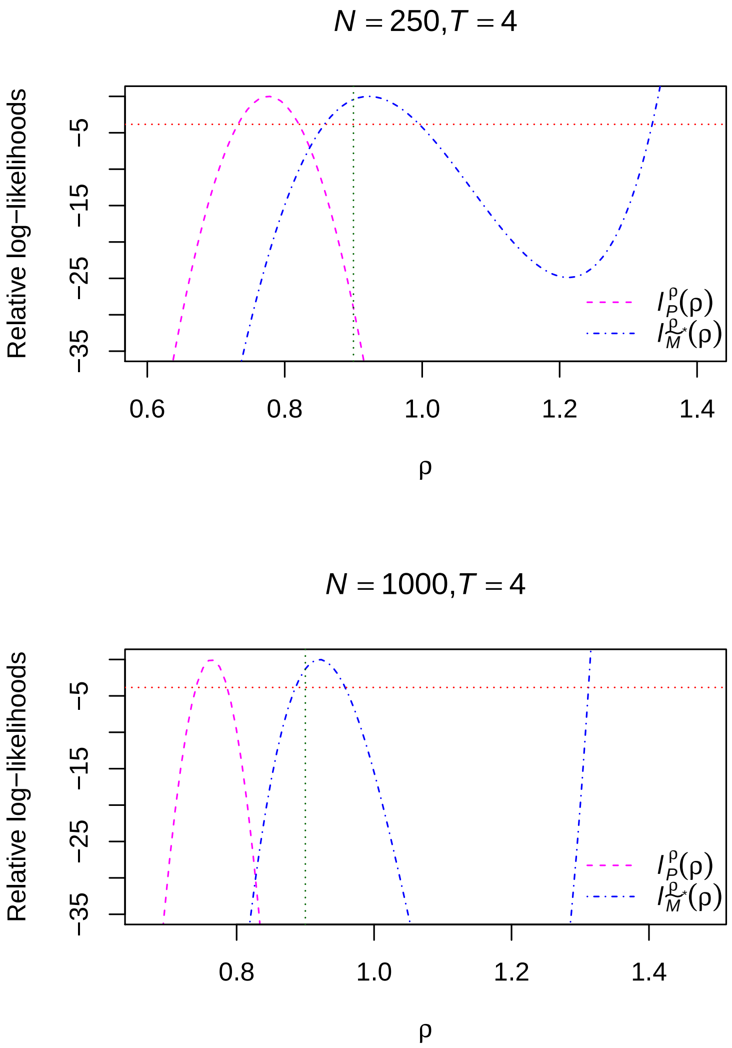

FIGURE 3.2: Relative log-likelihoods for two datasets generated under the nonstationary AR(1) model with $\rho = 0.9$. The vertical line indicates the true value of the autoregressive parameter, while the horizontal line gives the 0.95 confidence intervals for ρ based on the profile and modified profile log-likelihood ratio statistics.

The lacking registration of some data in one study may occur in a multiplicity of ways. According to Little and Rubin (2002, Sections 1.2 and 1.3) the classification of missing values can be based on two main criteria: pattern of missingness and mechanism of missingness. The former essentially describes which data are observed and which are not. For instance, one usually speaks of univariate missing data whether missingness is confined to a single recorded variable and of multivariate missing data otherwise. Under the same framework, a further distinction which is useful in regression settings is made between incomplete predictors and/or incomplete outcomes. Missing-data patterns are

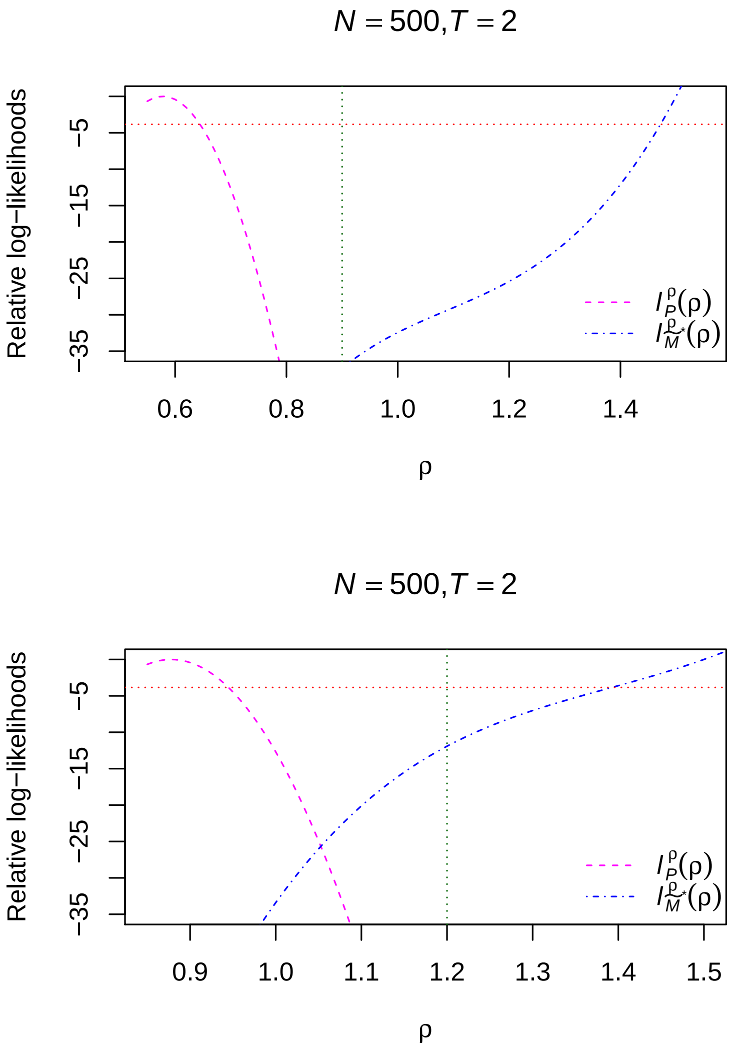

FIGURE 3.3: Relative log-likelihoods for two datasets generated under the nonstationary AR(1) model with $\rho = 0.9$ and $\rho = 1.2$, respectively. The vertical line indicates the true value of the autoregressive parameter, while the horizontal line gives the 0.95 confidence intervals for ρ based on the profile and modified profile log-likelihood ratio statistics.

a matter of particular importance for clustered observations. In longitudinal studies collecting information on a set of cases repeatedly over time, like clinical trials or panel surveys, a typical issue is indeed attrition, due to subjects dropping out prior to the end of the follow-up occasions and not coming back. Such pattern of missingness is said monotone to be distinguished from the general or arbitrary ones, when intermittent observations may arise instead. For additional examples of incomplete-data patterns, interested readers may also consult Schafer and Graham (2002).

To put it simply, mechanisms leading to incomplete datasets appertain to the relationship between measured variables and the probability of missing data (Baraldi and Enders, 2010). This concept found its first mathematical formalization in the seminal paper by Rubin (1976), who explicitly treated the missing values as realizations of a random variable with some probability distribution. Such an approach enabled the author to develop the categorisation of data still in use today: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The first case naively supposes the missingness probability to be completely unrelated to the data values, missing or not. More realistically, with an MAR mechanism missingness is instead allowed to depend on the observed entries of the dataset. Alternatively, when the probability of missing observations depends also on values that are unobserved, the data are called MNAR. A meticulous elucidation of these definitions can be read in Mealli and Rubin (2015), where an extended typology of missing-data mechanisms is also presented.

Inferential procedures to handle missing values in the estimation of statistical models must be selected taking into account both the pattern and the mechanism of missingness occurring in the study. Notice that the latter substantially corresponds to an assumption imposed by the analyst, which most of the times is empirically untestable (Baraldi and Enders, 2010). Formulation of this hypothesis requires extreme care because the true nature of the underlying missingness generation process deeply affects the validity of inferential results obtained with the numerous missing-data methods. A complete taxonomy of such techniques, along with many helpful references, is reported in Section 1.4 of Little and Rubin (2002). Accessible overviews of traditional and modern strategies for coping with partially observed data are instead Schafer and Graham (2002) and Baraldi and Enders (2010).

As throughout this thesis, here we only consider estimation procedures for incomplete datasets directly depending on the likelihood function. In general, along the above lines of argument, this model-based methodology asks to specify both a distribution for the data with usual full parameter $\theta \in \Theta$ and a mechanism for the missing values indexed by, say, $\gamma \in \Gamma$. However, a fundamental result in Rubin (1976) entails that the weakest sufficient conditions under which it is appropriate to ignore the missing-data mechanism when conducting likelihood inferences on θ are two: missingness at random of the data and distinctness between θ and γ, in the sense that the joint parameter space must be expressible as $\Theta \times \Gamma$. This motivates the terminology which refers to MCAR and MAR as ignorable mechanisms and to MNAR as nonignorable (Little and Rubin, 1987).

When data are MNAR and hence a model for the missingness process has to be formulated, different approaches can be adopted (Little and Rubin, 2002, Section 15.1). Nevertheless, the main distinction lies between so-called selection models and pattern-mixture models (Fitzmaurice *et al.*, 2008, Chapter 18). To better describe the difference, let us consider independent possibly missing clustered observations y_{it} and define the corresponding missingness indicators M_{it} such that $M_{it} = 1$ if y_{it} is unobserved and $M_{it} = 0$ otherwise ($i = 1, \ldots, N, t = 1, \ldots, T$). From a likelihood-related standpoint, the joint distribution of Y_{it} and M_{it} in some global parametrization φ needs to be specified and the manner in which it is factorized discriminates between the classes of nonignorable models. Particularly, selection models assume a marginal distribution for Y_{it} and a conditional distribution of M_{it} given Y_{it}, so that

$$p_{Y_{it},M_{it}}(y_{it}, m_{it}; \varphi, x_{it}) = p_{Y_{it}}(y_{it}; \theta, x_{it}) p_{M_{it}|Y_{it}}(m_{it}|y_{it}; \gamma, x_{it}), \tag{3.14}$$

with $\varphi = (\theta, \gamma)$; rather, pattern-mixture models explicitly assign some marginal distribution to M_{it} and one conditional distribution to Y_{it} given M_{it}, obtaining the factorization

$$p_{Y_{it},M_{it}}(y_{it}, m_{it}; \varphi, x_{it}) = p_{M_{it}}(m_{it}; \delta, x_{it}) p_{Y_{it}|M_{it}}(y_{it}|m_{it}; \omega, x_{it}),$$

where $\varphi = (\omega, \delta)$. Each of these modeling frameworks has its own benefits and drawbacks, thus the choice is usually made according to the special context of analysis. In wide generality, selection models appear more sensible in situations of ignorable missingness; for a comprehensive discussion on the topic, see Michiels *et al.* (1999), Section 18.3 in Fitzmaurice *et al.* (2008) and references therein.

3.4.2 Computational methods

Computationally speaking, in moderately complex models for incomplete datasets with general patterns, maximization of the log-likelihood function incorporating all the available information is quite an arduous task. Indeed this function, named observed log-likelihood, often involves integrals or summations over the distribution of the missing data which are hardly tractable.

It is well-known that the iterative EM algorithm (Dempster *et al.*, 1977) is a possibly advantageous strategy for ML estimation whenever data either are partially not observed or may be viewed as such. In fact, this approach is pervasive in the literature of missing data, and many extensions to the original version have been posited to tackle specific combinations of pattern and mechanism of missingness. Other than those examined in Section 8.5 of Little and Rubin (2002), it might be worth quoting a few more proposals somehow related with the studies reported in the following sections. Firstly, given that the focus here is on maximization of profile and adjusted profile likelihoods, a due reference is made to the work of Kim and Taylor (1995), who presented the general EM routine to be applied under linear restrictions on the parameters. As for particular missing-data problems, Ibrahim *et al.* (1999a) and Ibrahim *et al.* (1999b) generalized the EM algorithm for handling MAR and MNAR covariates, respectively, under regression scenarios. Both solutions rely on a Monte Carlo implementation of the EM procedure (Wei and Tanner, 1990) and on a Gibbs sampler with adaptive rejection region (Gilks and Wild, 1992) for reasons of computational efficiency. Another strategy is that of Sinha and Maiti (2008), who developed an EM-type algorithm for the specific analysis of matched case-control data with nonignorable missing exposure. Targeting instead the missingness of the dependent variable, Ibrahim and Lipsitz (1996) used a weighted EM procedure in binomial regressions with MNAR response, while Fitzmaurice *et al.* (1994) considered EM estimation of models for MAR binary missing clustered data. The stochastic EM algorithm for managing arbitrary patterns of nonignorable missingness in the outcome of longitudinal studies was used instead by Gad and Ahmed (2006). Lastly, one relevant contribution in this research area was recently provided by Yang and Kim (2016), who approximated the observed log-likelihood for MAR data by importance sampling in every EM iteration.

Obviously, ML estimation in missing-data problems can be performed by numerical iterative algorithms alternative to the EM (Little and Rubin, 2002, Section 8.1). Among the variety of examples hosted by the literature, we shall recall the Nelder-Mead simplex method (Nelder and Mead, 1965) employed in Troxel *et al.* (1998a) and Troxel *et al.* (1998b) for optimization purposes in presence of arbitrarily MNAR clustered observations. Furthermore, both Parzen *et al.* (2006) and Sinha *et al.* (2011) carried out maximization of a pseudo-likelihood by the popular Newton-Raphson algorithm. Interestingly, their approach can be interpreted as semiparametric in spirit, because it avoids defining some joint distribution for the binary longitudinal data with nonignorable missingness and non-monotone patterns. As a result, the function to be optimized is much more computationally tractable.

On a general note, the application of the EM algorithm notoriously eases numerical complexities linked with the direct maximization of the observed log-likelihood when the assumed distribution of the data belongs to the class of exponential families. However, the basic iterative process does not automatically deliver estimated standard errors of ML estimates and might converge very slowly if the portion of missing information is large (Little and Rubin, 2002, Section 8.1). Some aforesaid variants of the original procedure manage to fix these issues, but at the expense of simplicity in implementation. For certain, a universal best solution to maximize the log-likelihood in problems with incomplete observations is impossible to prescribe, thus every situation needs to be assessed individually. Before closing, it is important to point out that nonignorable missing-data models must be carefully fitted regardless of the method employed, because the available information may often be insufficient to estimate some parameters (Ibrahim *et al.*, 2001).

3.4.3 Binary regressions with missing response

In this section, special attention is given to possibly missing clustered binary observations. Several regression models for such kind of data have been reviewed and compared in, for example, Fitzmaurice *et al.* (1995), with specific reference to nonignorable dropouts. By contrast, here we examine arbitrary patterns of missingness and not only MNAR mechanisms, yet the key points of that work apply also to these situations. Furthermore, until otherwise stated, covariates are considered given and entirely observed.

Adopting the typical factorization of selection models defined in (3.14), for independent observations y_{it} one can write the marginal distribution

$$Y_{it} \sim Bern(\pi_{it}), \quad \pi_{it} = \pi_{it}(\theta) = F(\lambda_i + \beta x_{it}), \quad i = 1, \dots, N, \quad t = 1, \dots, T, \tag{3.15}$$

with F some suitable cumulative distribution function, whereas the conditional model for the missingness indicator introduced in Section 3.4.1 may be expressed by

$$M_{it} | Y_{it} = y_{it} \sim Bern(\zeta_{it}), \qquad i = 1, \dots, N, \quad t = 1, \dots, T, \tag{3.16}$$

where $\zeta_{it} \in (0, 1)$. Specifically, choosing a canonical link as done in Diggle and Kenward (1994) and denoting by logit^{-1} the distribution function of the logistic random variable, the following general formulation is attributed to ζ_{it}:

$$\zeta_{it} = \zeta_{it}(\gamma) = P(M_{it} = 1 | Y_{it} = y_{it}) = \text{logit}^{-1}(\gamma_1 + \gamma_2 x_{it} + \gamma_3 y_{it}). \tag{3.17}$$

The parameter of interest in the joint model described by (3.15)–(3.17) coincides with the unique regression coefficient $\beta \in \mathbb{R}$, and the usual incidental parameters are grouped in $\lambda = (\lambda_1, \dots, \lambda_N) \in \mathbb{R}^N$, so that $\theta = (\beta, \lambda) \in \mathbb{R}^{N+1}$. As further nuisance component, we also have the coefficients in the logistic model for the indicator of missingness, $\gamma = (\gamma_1, \gamma_2, \gamma_3) \in \mathbb{R}^3$, thus the overall parameter here is $\varphi = (\theta, \gamma) \in \mathbb{R}^{N+4}$. The structural component common to all groups in the sample is finally defined as $\psi = (\beta, \gamma) \in \mathbb{R}^4$.

Once again, for the purposes of this discussion it is sufficient to envisage only one predictor, but extensions of the forthcoming analysis to cases with multiple regressors are straightforward. We also stress that, although not contemplated here, substitution of a cluster-specific intercept for γ_1 in (3.17) might be deemed appropriate.

According to the assumption we make about the mechanism which generates the missing values, it is possible to identify different relations between the missingness probability and the variables in the study. Such relations, in their turn, translate into constraints on the model parameters (Parzen *et al.*, 2006). Particularly, since here covariates are nonrandom and their distribution is not modeled, from specification (3.17) follows that data can be either MCAR, when $\gamma_3 = 0$, or MNAR otherwise (Baker, 1995).

The primary objective of this part is to see whether Monte Carlo simulation effectively improves the performance of the MPL by Severini when making inference on ψ in situations with missing data. Indeed, models like (3.15) for complete observations y_{it} were already investigated in Bellio and Sartori (2003), who showed how analytically deriving $\widetilde{M}(\psi)$ in order to consistently estimate ψ when N is much larger than T. Unfortunately, the presence of missing values creates trouble in the explicit calculation of the adjustment term. Generally, the expectation therein should be evaluated with regard to the joint distribution $p(y_{it}, m_{it}; \hat{\varphi}, x_{it})$, taking also the missing-data mechanism into account, but the correct way of doing so is not without ambiguity. More specifically, in the light of the arguments made by Kenward and Molenberghs (1998), one expects to be allowed to neglect the missingness process only when data are MCAR. Thus, even in this setting, we shall see how the Monte Carlo strategy can easily overcome the computational difficulties experienced during the use of the MPL.

Let us now obtain the necessary likelihood quantities for drawing inferences on the parameter of interest under the most general MNAR framework. For the sake of clarity, denote by y^{obs} the observed entries of $y = (y_{it})$ and by y^{mis} the remaining missing components. As highlighted in Section 6.2 of Little and Rubin (2002), the actual data consist of y^{obs} and the indicators of missingness $m = (m_{it})$. The observed likelihood is then obtained by summing over y^{mis} the joint probability distribution of $Y = (Y^{obs}, Y^{mis})$ and M. Precisely, one can write

$$L(\varphi) = L(\varphi; y^{obs}, m) = \sum_{y^{mis}} p_Y\big(y^{obs}, y^{mis}; \theta\big) p_{M|Y}\big(m|y^{obs}, y^{mis}; \gamma\big),$$

where the presence of fixed covariates is ignored for succinctness. In our case, since the groups of observations are independent, the corresponding MNAR log-likelihood may be written as usual in the additive form $l(\varphi) = \sum_{i=1}^{N} l^i(\varphi)$ and its maximizer is the global ML estimate $\hat{\varphi}$. By assumptions (3.15)–(3.17), it is not too difficult to derive the expression for the ith summand:

$$l^i(\varphi) = \sum_{t=1}^{T} \biggl[m_{it} \log \left\{ (1-\pi_{it})\zeta_{it}^0 + \pi_{it}\zeta_{it}^1 \right\} \tag{3.18}$$
$$+ (1-m_{it}) \bigl\{ y_{it} \log \pi_{it} + (1-y_{it}) \log(1-\pi_{it}) + \log(1-\zeta_{it}) \bigr\} \biggr],$$

where $\zeta_{it}^0 = \text{logit}^{-1}(\gamma_1 + \gamma_2 x_{it})$ and $\zeta_{it}^1 = \text{logit}^{-1}(\gamma_1 + \gamma_2 x_{it} + \gamma_3)$. Notice that $l^i(\varphi)$ is substantially divided in two parts: the first accounts for the missing observations y^{mis} and the second for the recorded y^{obs}. After one differentiation with respect to the ith incidental parameter λ_i, we get the partial score function

$$l_{\lambda_i}(\varphi) = \sum_{t=1}^{T} \left\{ m_{it} \log \frac{f_{it}(\zeta_{it}^1 - \zeta_{it}^0)}{\pi_{it}\zeta_{it}^1 + (1-\pi_{it})\zeta_{it}^0} + (1-m_{it}) \frac{(y_{it}-\pi_{it}) f_{it}}{\pi_{it}(1-\pi_{it})} \right\}, \tag{3.19}$$

where $f_{it} = f_{it}(\theta) = \partial F(\lambda_i + \beta x_{it})/\partial \lambda_i$ and the separate contribution of unobserved and observed data is still evident. Then, differentiating one more time and changing the sign of the obtained derivative lead to

$$j_{\lambda_i\lambda_i}(\varphi) = \sum_{t=1}^{T} \Biggl[m_{it} \left\{ \frac{f'_{it}}{f_{it}} - \frac{(\zeta_{it}^1 - \zeta_{it}^0) f_{it}}{\pi_{it}\zeta_{it}^1 + (1-\pi_{it})\zeta_{it}^0} \right\}$$
$$+ (1-m_{it})(y_{it}-\pi_{it}) \left\{ \frac{f'_{it} - f_{it}^2}{\pi_{it}(1-\pi_{it})} - \frac{f_{it}(1-2\pi_{it})}{\pi_{it}^2(1-\pi_{it})^2} \right\} \Biggr], \tag{3.20}$$

where $f'_{it} = f'_{it}(\theta) = \partial^2 F(\lambda_i + \beta x_{it})/\partial \lambda_i^2$. The solution to the ith component of the likelihood equation $l_{\lambda_i}(\varphi) = 0$ can be found numerically, and we denote it by $\hat{\lambda}_{i\psi}$. Substituting this value for λ_i in (3.18) permits to obtain the MNAR profile log-likelihood as $l_P(\psi) = \sum_{i=1}^N l_P^i(\psi)$. Defined the full constrained ML estimate $\hat{\varphi}_\psi$ in the conventional way, the same replacement in equation (3.20) gives instead $j_{\lambda_i\lambda_i}(\hat{\varphi}_\psi)$.

At this stage, we are left with the computation of $I_{\lambda_i\lambda_i}(\hat{\varphi}_\psi; \hat{\varphi}) = E_{\hat{\varphi}}\bigl\{ l_{\lambda_i}(\hat{\varphi}_\psi) l_{\lambda_i}(\hat{\varphi}) \bigr\}$. For this model, the intricacy of such task not only has practical but also conceptual origins. Understanding how to take this expected value over the unconditional sampling distribution, using the terminology of Kenward and Molenberghs (1998), is not that obvious. In fact, the joint distribution of (Y_{it}, M_{it}) was not specified directly, but divided in the two factors (3.15) and (3.16). Viceversa, the Monte Carlo solution presented in Section 3.2 may be applied quite plainly even in these circumstances. Particularly, the approximation (3.3) in the MNAR case takes the form

$$I^*_{\lambda_i\lambda_i}(\hat{\varphi}_\psi;\hat{\varphi}) = \frac{1}{R}\sum_{r=1}^{R} l^r_{\lambda_i}(\hat{\varphi}_\psi) l^r_{\lambda_i}(\hat{\varphi}), \qquad i = 1,\dots,N, \tag{3.21}$$

where $l^r_{\lambda_i}$ is the score of the rth partially observed sample y^r_{it} $(r = 1,\dots,R)$ obtained in two steps: first, a complete dataset $y^{r,C}_{it}$ is simulated under model (3.15) with $\theta = \hat{\theta}$ and second, some entries in this dataset are deleted and considered missing according to the specification (3.16) with MNAR probability $\zeta_{it} = \zeta_{it}(\hat{\gamma}) = \text{logit}^{-1}(\hat{\gamma}_1 + \hat{\gamma}_2 x_{it} + \hat{\gamma}_3 y^{r,C}_{it})$. Note that $\hat{\psi} = (\hat{\theta},\hat{\gamma})$ is the global maximizer of the MNAR profile log-likelihood which also takes the generation process of missingness into consideration. Therefore, by such procedure, the average of the score products over the R incomplete samples y^r_{it} properly estimates the unconditional expectation required.

Before proceeding, it seems worthwhile making a few more comments about the general formula (3.18). Supposing an ignorable MCAR missing-data mechanism by imposing $\gamma_3 = 0$ in (3.17) yields clearly to $\zeta^0_{it} = \zeta^1_{it} = \zeta_{it} = \text{logit}^{-1}(\gamma_1 + \gamma_2 x_{it})$, and hence (3.18) simplifies in

$$l^i(\varphi) = \sum_{t=1}^{T} \big[m_{it}\log\zeta_{it} + (1-m_{it})\big\{y_{it}\log\pi_{it} + (1-y_{it})\log(1-\pi_{it}) + \log(1-\zeta_{it})\big\}\big].$$

Since our interest is only on the parameter β and ζ_{it} does not carry any useful information about it, we can rely on the equivalent function

$$l^i(\theta) = l^i(\theta; y^{obs}) = \sum_{t:\, y_{it}\in y^{obs}} \big\{y_{it}\log\pi_{it} + (1-y_{it})\log(1-\pi_{it})\big\}, \tag{3.22}$$

which is the ordinary group-related log-likelihood in binary regressions computed only on the recorded data. Indeed, when the missingness mechanism is MCAR, a complete-case analysis discarding units with missing values is unbiased, as the wholly observed cases are basically a random sample from the reference population (Little and Rubin, 2002, Section 3.2). For this specific model, it is also fully efficient because θ and γ are distinct, provided that the full parameter space is $\Phi = \mathbb{R}^{N+1} \times \mathbb{R}^2 = \Theta \times \Gamma$ (Little and Rubin, 2002, p. 120). This means that likelihood inference can be conducted disregarding the process which generates the missing observations. As a major implication for our study, the expected value involved in Severini's MPL may be derived from the conditional distribution of Y_{it} given $M_{it} = 0$. Specifically, it can be effortlessly shown (Bellio and Sartori, 2003) that in situations like this such expectation has the following closed-form expression:

$$I_{\lambda_i\lambda_i}(\hat{\theta}_\beta;\hat{\theta}) = \sum_{t:\, y_{it}\in y^{obs}} \frac{f_{it}(\hat{\theta}_\beta) f_{it}(\hat{\theta})}{\{1-\pi_{it}(\hat{\theta}_\beta)\}\pi_{it}(\hat{\theta}_\beta)}, \qquad i=1,\ldots,N, \tag{3.23}$$

where estimates $\hat{\theta} = (\hat{\beta},\hat{\lambda})$ and $\hat{\theta}_\beta = (\beta,\hat{\lambda}_\beta)$ descend from ordinary ML inference on the parameter of interest β via the MCAR profile log-likelihood $l_P(\beta)$ based on (3.22). Furthermore, inasmuch as under the hypothesis of ignorable missingness it is possible to utilize the function $l(\theta)$ with components (3.22), the general Monte Carlo approximation reported in (3.21) admits to be reformulated in the MCAR case as

$$I^*_{\lambda_i\lambda_i}(\hat{\theta}_\beta;\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R} l^r_{\lambda_i}(\hat{\theta}_\psi) l^r_{\lambda_i}(\hat{\theta}), \qquad i=1,\ldots,N, \tag{3.24}$$

where $l^r_{\lambda_i} = \sum_{t:\, y_{it}\in y^{obs}} (y^r_{it}-\pi_{it}) f_{it}/\{\pi_{it}(1-\pi_{it})\}$ is the score of the incomplete sample y^r_{it} simulated by the two-step procedure above but with an important difference: now $\hat{\theta}$ results from the maximization of $l(\theta)$, while $\hat{\gamma} = (\hat{\gamma}_1,\hat{\gamma}_2)$ is obtained by a separate ML fit of the logistic regression (3.17) subject to the constraint $\gamma_3 = 0$, with the missingness indicator as dependent variable and the covariate x_{it} as unique predictor.

In the sequel, the utility of Monte Carlo approximation in the presence of incomplete data will be evaluated through simulation experiments referring to binary regressions with different missingness processes. Specifically, objects of comparison shall be the unadjusted profile log-likelihood (either the MCAR $l_P(\beta)$ or the MNAR $l_P(\psi)$), the modification proposed by Severini $l_{\widetilde{M}}(\beta)$ that ignores the missing values and is computed analytically by formula (3.23) and the Monte Carlo MPL that accounts for some presumed missingness mechanism. In order to avoid confusion, its MCAR variant employing the estimate (3.24) will be denoted by $l_{\widetilde{M}^*}(\beta)$, whereas $l_{\widetilde{M}^*}(\psi)$ shall indicate the MNAR MPL with habitual expectation approximated by (3.21).

Logistic regression: simulation studies

The first part of analyses is performed supposing a logit link between the mean of the response and the predictors, meaning $F = \text{logit}^{-1}$ in model (3.15). Pairing this assumption with that of an MCAR mechanism brings about the equality

$$\begin{aligned} I_{\lambda_i\lambda_i}(\hat{\theta}_\beta;\hat{\theta}) &= \sum_{t:\, y_{it}\in y^{obs}} \big[1-\pi_{it}(\hat{\theta})\big]^2 \\ &= \sum_{t:\, y_{it}\in y^{obs}} \big[1-\text{logit}^{-1}(\hat{\lambda}_i+\hat{\beta}x_{it})\big], \qquad i=1,\ldots,N, \end{aligned}$$

whose right-hand side does not depend on the parameter of interest. Hence the only part of Severini's modification term relevant to estimating β is $\frac{1}{2}\log|j_{\lambda\lambda}(\hat{\theta}_\beta)|$ and one can write

$$\begin{aligned}\widetilde{M}(\beta) &= \frac{1}{2}\sum_{i=1}^{N}\log\left[\sum_{t:\, y_{it}\in y^{obs}} \pi_{it}(\hat{\theta}_\beta)\{1-\pi_{it}(\hat{\theta}_\beta)\}\right] \\ &= \frac{1}{2}\sum_{i=1}^{N}\log\left[\sum_{t:\, y_{it}\in y^{obs}} \text{logit}^{-1}(\hat{\lambda}_{i\beta}+\beta x_{it})\{1-\text{logit}^{-1}(\hat{\lambda}_{i\beta}+\beta x_{it})\}\right]. \qquad (3.25)\end{aligned}$$

It is also simple to show that in such a setting the score component related to the ith incidental parameter equals

$$\begin{aligned} l_{\lambda_i}(\theta) &= \sum_{t:\, y_{it}\in y^{obs}} \{y_{it} - \pi_{it}(\theta)\} \\ &= \sum_{t:\, y_{it}\in y^{obs}} \{y_{it} - \text{logit}^{-1}(\lambda_i + \beta x_{it})\}, \qquad i = 1, \dots, N, \end{aligned}$$

thus the expression of the MCAR Monte Carlo estimate $I^*_{\lambda_i\lambda_i}(\hat{\theta}_\beta;\hat{\theta})$ follows immediately from the previous formula and (3.24). Loosely speaking, if observations are MCAR, $l_{\widetilde{M}}(\beta)$ and $l_{\widetilde{M}^*}(\beta)$ take the same forms as in general logistic regressions for panel data with no missing values, yet are computed only on the complete units. The numerical maximization of both functions may then be automatically implemented by the `R` package `panelMPL` (Bellio and Sartori, 2015), after some minor manipulation of the code which enables to manage also unbalanced group sizes.

For the reasons extensively discussed earlier, a correct analytical formulation of Severini's MPL is not available when missingness in the data is hypothesized to be non-ignorable. On the contrary, $\widetilde{M}^*(\psi)$ can be calculated via Monte Carlo simulation as indicated in (3.21). All the quantities appearing therein are very easy to derive in the logistic case and their specific expressions are not included here for brevity purposes only. Turning to examine the optimization step in the MNAR scenario, even though the model under analysis belongs to an exponential family and would be suitable for an EM-type routine, the functions $l_P(\psi)$ and $l_{\widetilde{M}^*}(\psi)$ are directly maximized numerically by the Nelder-Mead algorithm. This decision may be motivated by several arguments. The first has to do with the form of the observed log-likelihood in regressions where the possibly missing response is binary; indeed, such function is not as computationally intractable as commonly is when dealing with continuous data (Gad and Ahmed, 2006). Secondly, our independence assumption avoids the specification of relationships among observations that would introduce structural parameters of not direct interest to be estimated (Troxel *et al.*, 1998b). Moreover, this choice permits not to worry about the considerable percentage of missing values in the data and the calculation of standard errors, as always estimated by means of the second numerical derivative of the maximized function. Notice that in the MNAR case the argument $\psi = (\beta, \gamma)$ of the objective functions to be optimized has dimension equal to 4, whereas in the MCAR case β is scalar. The higher complexity in the maximization problem is reflected by longer execution times and numerical instabilities, especially in the estimation of γ and its variance. Both Baker (1995) and Ibrahim and Lipsitz (1996) came across issues of this kind while fitting similar nonignorable missing-data models for binary responses. The authors attribute such problems to the lack of information in the sample about the parameters ruling the missingness process, which may then result not identifiable. At the suggestion of Parzen *et al.* (2006), to further facilitate the estimation phase one might try modeling in a simpler manner the nonignorable mechanism; yet, in our case, dropping γ_1 and/or γ_2 in (3.17) does not appear very sensible.

Before going through the details of the experiments run, it is worth recalling that, as is common practice for binary longitudinal regressions, the optimization stage needs to be anticipated by the omission of non-informative groups (Bellio and Sartori, 2003) from the sample under analysis. In missing-data situations, whatever the supposed mechanism, the clusters which cannot contribute to estimate β are those with $y_{it}^{obs} = 0$ or $y_{it}^{obs} = 1$ for every $t = 1, \ldots, T$ and those which are totally unobserved, i.e. where $y_{it} = y_{it}^{mis}$ for each $t = 1, \ldots, T$ $(i = 1, \ldots, N)$.

Let us now describe the basic setup of the simulation studies. The two principal settings are recognisable according to the model used to select the missing values in the experimental datasets. In both of them, the covariate x_{it} is simulated by means of independent draws from the standard normal distribution, while intercepts λ_i $(i = 1, \ldots, N)$ are obtained as $\lambda_i = \sum_{t=1}^{T} x_{it}/T + u_i$, where $u_i \sim N(0, 1)$. The values of the structural components in (3.15) and (3.17) for generating the $S = 2000$ samples with MCAR observations are set equal to $\beta = 1$, $\gamma_1 = -0.5$ and $\gamma_2 = 0.3$. Rather, simulation of the MNAR data is carried out with $\beta = 2$, $\gamma_1 = -1$, $\gamma_2 = 0.3$ and $\gamma_3 = 2$. The true values of γ are chosen in such a way as to observe a percentage of missing observations in the resulting datasets varying between 40% and 50%. Changing the value of the regression coefficient in the second framework seems instead to mitigate the computational instabilities associated with the estimation of γ. One possible explanation for this finding is that, with the fixed nonignorable probability of missing data, a larger value of β serves to maintain the portion of informative clusters comparable to the MCAR case, reducing so the lack of knowledge about the missingness process. Tables 3.5, 3.6 and 3.7 show results of the series of simulations conducted in the context of logistic regressions. Performances of the compared inferential functions are reported by computing measures of accuracy analogue to those described for the autoregressive model in Section 3.3.4. In the study considering an underlying MCAR mechanism, dimensions of the simulated datasets correspond to different combinations of $T = 4, 6, 10$ and $N = 50, 100, 250$. One may directly contrast the behaviour of the likelihoods built under the correct MCAR hypothesis by looking at Table 3.5. The latter visibly certifies the inadequacy of inference

Table 3.5: Inference on $\beta = 1$ in the logistic regression for MCAR longitudinal data. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\beta)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\beta)$	0.793	0.666	0.771	1.106	0.672	0.669	0.732
		$l_{\widetilde{M}}(\beta)$	0.138	0.111	0.351	0.377	0.233	1.024	0.969
		$l_{\widetilde{M}^*}(\beta)$	0.138	0.110	0.360	0.385	0.231	1.000	0.969
	6	$l_P(\beta)$	0.793	0.666	0.771	1.106	0.672	0.669	0.732
		$l_{\widetilde{M}}(\beta)$	0.097	0.069	0.271	0.288	0.178	0.991	0.962
		$l_{\widetilde{M}^*}(\beta)$	0.098	0.070	0.271	0.288	0.178	0.990	0.961
	10	$l_P(\beta)$	0.243	0.228	0.235	0.338	0.234	0.857	0.807
		$l_{\widetilde{M}}(\beta)$	0.038	0.031	0.184	0.188	0.116	0.972	0.945
		$l_{\widetilde{M}^*}(\beta)$	0.039	0.032	0.184	0.188	0.115	0.972	0.943
100	4	$l_P(\beta)$	0.684	0.631	0.466	0.828	0.631	0.746	0.534
		$l_{\widetilde{M}}(\beta)$	0.098	0.089	0.236	0.255	0.160	1.060	0.965
		$l_{\widetilde{M}^*}(\beta)$	0.098	0.088	0.236	0.256	0.159	1.059	0.964
	6	$l_P(\beta)$	0.436	0.413	0.277	0.517	0.413	0.813	0.542
		$l_{\widetilde{M}}(\beta)$	0.073	0.064	0.181	0.195	0.130	1.009	0.951
		$l_{\widetilde{M}^*}(\beta)$	0.073	0.065	0.181	0.195	0.129	1.009	0.948
	10	$l_P(\beta)$	0.229	0.220	0.168	0.284	0.220	0.859	0.658
		$l_{\widetilde{M}}(\beta)$	0.028	0.021	0.132	0.135	0.089	0.969	0.947
		$l_{\widetilde{M}^*}(\beta)$	0.029	0.023	0.132	0.135	0.089	0.969	0.946
250	4	$l_P(\beta)$	0.634	0.612	0.297	0.701	0.612	0.731	0.199
		$l_{\widetilde{M}}(\beta)$	0.079	0.071	0.158	0.176	0.117	1.004	0.934
		$l_{\widetilde{M}^*}(\beta)$	0.080	0.072	0.158	0.177	0.115	1.003	0.934
	6	$l_P(\beta)$	0.412	0.401	0.183	0.451	0.401	0.789	0.207
		$l_{\widetilde{M}}(\beta)$	0.059	0.055	0.121	0.134	0.089	0.975	0.928
		$l_{\widetilde{M}^*}(\beta)$	0.059	0.055	0.121	0.135	0.089	0.973	0.926
	10	$l_P(\beta)$	0.225	0.222	0.105	0.249	0.222	0.883	0.326
		$l_{\widetilde{M}}(\beta)$	0.027	0.026	0.084	0.088	0.057	0.991	0.940
		$l_{\widetilde{M}^*}(\beta)$	0.028	0.026	0.084	0.088	0.058	0.992	0.938

on β deriving by the employment of the profile likelihood in this incidental parameters setting. The introduction of the modification term, either explicitly calculated or approximated by Monte Carlo simulation with $R = 500$, conspicuously refines both point estimation and the actual coverage of Wald confidence intervals. The overall effects of the adjustment to $l_P(\beta)$ are essentially equivalent to those viewed in Section 3.3.4. Yet the most important evidence supplied here by Table 3.5 is the absence of the need to take the MCAR mechanism into consideration when computing Severini's MPL. Indeed, the performance of $l_{\widetilde{M}}(\beta)$ is substantially identical to that of $l_{\widetilde{M}^*}(\beta)$ for all the sample sizes considered. This confirms what argued by Kenward and Molenberghs (1998).

Inference on the same MCAR datasets can also be made via the functions $l_P(\psi)$ and $l_{\widetilde{M}^*}(\psi)$, which assume a general nonignorable model of missingness. Experimental outcomes of such analysis, presented in Table 3.6, are of doubtful interpretation. Contrary to expectations, the global accuracy of the MNAR Monte Carlo MPL appears to worsen as the group size raises. More precisely, the bias of its estimator is even higher than

TABLE 3.6: Inference on $\beta = 1$ in the logistic regression for MCAR longitudinal data. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	0.628	0.500	0.840	1.048	0.540	0.571	0.775
		$l_{\widetilde{M}^*}(\psi)$	-0.002	-0.105	0.513	0.513	0.292	0.636	0.869
	6	$l_P(\psi)$	0.318	0.261	0.503	0.595	0.340	0.661	0.781
		$l_{\widetilde{M}^*}(\psi)$	-0.117	-0.150	0.336	0.356	0.253	0.693	0.752
	10	$l_P(\psi)$	0.165	0.153	0.280	0.325	0.214	0.756	0.818
		$l_{\widetilde{M}^*}(\psi)$	-0.187	-0.192	0.249	0.311	0.218	0.669	0.653
100	4	$l_P(\psi)$	0.510	0.449	0.527	0.733	0.458	0.628	0.681
		$l_{\widetilde{M}^*}(\psi)$	-0.078	-0.114	0.276	0.287	0.198	0.798	0.862
	6	$l_P(\psi)$	0.264	0.263	0.350	0.438	0.307	0.646	0.697
		$l_{\widetilde{M}^*}(\psi)$	-0.190	-0.193	0.220	0.290	0.214	0.669	0.627
	10	$l_P(\psi)$	0.181	0.183	0.200	0.269	0.198	0.775	0.716
		$l_{\widetilde{M}^*}(\psi)$	-0.262	-0.220	0.225	0.345	0.227	0.498	0.460
250	4	$l_P(\psi)$	0.459	0.438	0.381	0.597	0.440	0.559	0.469
		$l_{\widetilde{M}^*}(\psi)$	-0.133	-0.133	0.212	0.250	0.164	0.624	0.691
	6	$l_P(\psi)$	0.257	0.259	0.266	0.370	0.277	0.589	0.544
		$l_{\widetilde{M}^*}(\psi)$	-0.222	-0.184	0.189	0.292	0.189	0.484	0.471
	10	$l_P(\psi)$	0.205	0.210	0.124	0.240	0.210	0.803	0.430
		$l_{\widetilde{M}^*}(\psi)$	-0.344	-0.257	0.212	0.405	0.257	0.302	0.164

TABLE 3.7: Inference on $\beta = 2$ in the logistic regression for MNAR longitudinal data. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	10	$l_P(\psi)$	0.559	0.445	0.794	0.971	0.492	0.547	0.782
		$l_{\widetilde{M}}(\beta)$	-0.082	-0.102	0.330	0.340	0.227	1.053	0.938
		$l_{\widetilde{M}^*}(\psi)$	-0.105	-0.180	0.630	0.638	0.340	0.526	0.746
	20	$l_P(\psi)$	0.211	0.195	0.291	0.359	0.225	0.820	0.859
		$l_{\widetilde{M}}(\beta)$	-0.097	-0.106	0.222	0.242	0.168	1.009	0.904
		$l_{\widetilde{M}^*}(\psi)$	-0.087	-0.084	0.255	0.270	0.176	0.888	0.883
	30	$l_P(\psi)$	0.144	0.137	0.201	0.247	0.162	0.891	0.879
		$l_{\widetilde{M}}(\beta)$	-0.112	-0.121	0.182	0.214	0.151	0.975	0.870
		$l_{\widetilde{M}^*}(\psi)$	-0.029	-0.036	0.174	0.177	0.119	0.996	0.941
100	10	$l_P(\psi)$	0.442	0.402	0.519	0.681	0.434	0.581	0.677
		$l_{\widetilde{M}}(\beta)$	-0.104	-0.117	0.232	0.254	0.182	1.035	0.909
		$l_{\widetilde{M}^*}(\psi)$	-0.196	-0.239	0.423	0.466	0.317	0.543	0.642
	20	$l_P(\psi)$	0.189	0.177	0.202	0.277	0.184	0.837	0.808
		$l_{\widetilde{M}}(\beta)$	-0.110	-0.111	0.161	0.195	0.140	1.001	0.867
		$l_{\widetilde{M}^*}(\psi)$	-0.077	-0.084	0.171	0.188	0.129	0.962	0.914
	30	$l_P(\psi)$	0.148	0.144	0.140	0.203	0.147	0.904	0.796
		$l_{\widetilde{M}}(\beta)$	-0.116	-0.121	0.125	0.170	0.129	0.996	0.820
		$l_{\widetilde{M}^*}(\psi)$	-0.028	-0.032	0.121	0.124	0.085	1.027	0.950
250	10	$l_P(\psi)$	0.440	0.409	0.348	0.561	0.410	0.537	0.498
		$l_{\widetilde{M}}(\beta)$	-0.131	-0.136	0.141	0.192	0.146	1.067	0.859
		$l_{\widetilde{M}^*}(\psi)$	-0.203	-0.244	0.344	0.399	0.281	0.434	0.541
	20	$l_P(\psi)$	0.198	0.193	0.133	0.239	0.193	0.788	0.558
		$l_{\widetilde{M}}(\beta)$	-0.117	-0.119	0.101	0.155	0.123	0.983	0.760
		$l_{\widetilde{M}^*}(\psi)$	-0.070	-0.075	0.118	0.137	0.093	0.886	0.884
	30	$l_P(\psi)$	0.136	0.134	0.088	0.161	0.134	0.909	0.609
		$l_{\widetilde{M}}(\beta)$	-0.126	-0.125	0.078	0.148	0.125	1.003	0.636
		$l_{\widetilde{M}^*}(\psi)$	-0.034	-0.035	0.075	0.083	0.056	1.028	0.929

that of the ML one if $T = 10$ and the empirical coverage of Wald confidence intervals, always far below the nominal level, falls dramatically when T grows. This last issue has also to do with the systematic underestimation of the estimates' variability, which seems exacerbated by the increasing number of within-cluster units. On the opposite $l_P(\psi)$ exhibits the habitual behaviour, proving to be more reliable than $l_{\widetilde{M}^*}(\psi)$ for interval estimation when $T = 6, 10$ for every value of N. Notice that, while the MCAR profile log-likelihood in Table 3.5 results less adequate for inference on β than its MNAR counterpart in Table 3.6, $l_{\widetilde{M}}(\beta)$ and $l_{\widetilde{M}^*}(\beta)$ neglecting the missing data process are typically much superior to $l_{\widetilde{M}^*}(\psi)$. Therefore, in this logistic setting, one might claim that unnecessary additional parameters to be estimated bring more harm than good and the Monte Carlo MPL accounting for MNAR data is not robust to a simpler MCAR true mechanism. The causes are unclear and surely merit further investigation.

A different picture is offered instead by Table 3.7, which refers to the second experiment based on datasets generated with MNAR observations. Here, for reasons that will soon be explained, the previous values of N are associated with larger group sizes, i.e. $T = 10, 20, 30$. Classical ML inference through the MNAR profile log-likelihood is found critically imprecise, especially in terms of the ensuing estimator's bias, even when $T = 30$. Yet the most significant simulation outcome concerns the relative pattern of inferential results reached by the two versions of the MPL considered. Quite interestingly, for any given number of clusters, as T increases the performance of $l_{\widetilde{M}}(\beta)$ deteriorates whereas that of $l_{\widetilde{M}^*}(\psi)$ improves, in sharp contrast to what non- above. Probably, for smaller T (we also tried $T = 4, 6$ like in the preceding study) the amount of information carried by the data is not adequate to properly estimate the correct nonignorable missingness mechanism. Therefore, accounting for it via Monte Carlo simulation has the only effect to degrade the quality of inferences drawn. In particular, this appears to be mostly due to underestimation of variability in the estimates resulting by the maximization of the MNAR MPL. Indeed, the numerical instabilities formerly mentioned are more present when the cluster size is small. The fact that, conversely, the MPL by Severini leads to worse results for large T may seem counterintuitive. A possible motivation is that incompleteness of the data is more perceived in larger groups and thus the harmful impact of the wrong MCAR assumption reveals itself as T grows. In outline, one may conclude that, if the units in the clusters are not many, the analytical version of the MPL by Severini is preferable even when the underlying mechanism of missingness should not be ignored. The convenience of the Monte Carlo strategy may instead be appreciated when groups are large and the process generating the missing values is suspected to be nonrandom. As a final note, we observe that a fairer assessment on the overall performance of Severini's $l_{\widetilde{M}}(\beta)$ and of the MNAR Monte Carlo MPL $l_{\widetilde{M}^*}(\psi)$ could be made by checking their inferential behaviours also in the presence of MCAR datasets with larger clusters, by analogy with the latter experiment about nonignorable missingness.

Probit regression: simulation studies

Suppose now that specifications (3.15)–(3.17) hold with $F = \Phi$, where Φ is the cumulative distribution function of the standard normal random variable. Even in probit regressions for clustered binary data y_{it} ($i = 1, \ldots, N$, $t = 1, \ldots, T$) an explicit formulation for Severini's adjustment exists. As in the former case, if the unobserved values are presumed to be MCAR that same expression can be computed on the available units. Specifically, denoting by ϕ the probability density function of the $N(0, 1)$, the expectation (3.23) simply becomes

$$I_{\lambda_i\lambda_i}(\hat{\theta}_\beta; \hat{\theta}) = \sum_{t\colon y_{it} \in y^{obs}} \frac{\phi\big(\hat{\lambda}_{i\beta} + \beta x_{it}\big)\phi\big(\hat{\lambda}_i + \hat{\beta} x_{it}\big)}{\big\{1 - \Phi\big(\hat{\lambda}_{i\beta} + \beta x_{it}\big)\big\}\Phi\big(\hat{\lambda}_{i\beta} + \beta x_{it}\big)}, \qquad i = 1, \ldots, N. \tag{3.26}$$

Under these hypotheses, it is immediate to show that the ith partial score function may be expressed as

$$l_{\lambda_i}(\theta) = \sum_{t\colon y_{it} \in y^{obs}} \frac{\big\{y_{it} - \Phi\big(\lambda_i + \beta x_{it}\big)\big\}\phi\big(\lambda_i + \beta x_{it}\big)}{\Phi\big(\lambda_i + \beta x_{it}\big)\big\{1 - \Phi\big(\lambda_i + \beta x_{it}\big)\big\}}, \qquad i = 1, \ldots, N, \tag{3.27}$$

and $j_{\lambda_i\lambda_i}(\theta)$ is readily derived by changing sign to its first derivative with respect to λ_i. Using (3.26) and (3.27), it is then possible to obtain both $l_{\widetilde{M}}(\beta)$ in closed form and $l_{\widetilde{M}^*}(\beta)$ as described in (3.24), which postulate an MCAR missingness model. Furthermore, observe that in the present probit regression framework the formula of the standard profile log-likelihood $l_P(\beta)$ follows directly from (3.22) with $\pi_{it} = \Phi(\lambda_i + \beta x_{it})$.

When, rather, we conjecture that incompleteness of the data originates from a nonignorable process, Monte Carlo simulation comes to our aid for approximating the unconditional expected value $I_{\lambda_i\lambda_i}(\hat{\varphi}_\psi; \hat{\varphi})$, whose exact formulation remains undefined. The expression of $l_{\widetilde{M}^*}(\psi)$ in the probit setting may be obtained by double substitution of $\Phi(\lambda_i + \beta x_{it})$ and $\phi(\lambda_i + \beta x_{it})$ for π_{it} and f_{it}, respectively, in equations (3.18)–(3.21).

The optimization methods employed for the various functions under forthcoming comparison correspond to those of the logistic case. Also, all the comments and justifications made on this point in the previous section apply here as well. It is certainly worthwhile mentioning that, when the link in model (3.15) is non-canonical, the computational instabilities driven by the problematic estimation of γ in case of MNAR assumption appear to be more pronounced and execution times of numerical routines are sensitively longer. Naturally, even in this framework, exclusion of the non-informative clusters by the dataset must take place prior to the fitting phase.

The basic structure of the two experiments now performed considering a probit link

TABLE 3.8: Inference on $\beta = 1/1.6$ in the probit regression for MCAR longitudinal data. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\beta)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\beta)$	0.495	0.419	0.459	0.675	0.421	0.645	0.677
		$l_{\widetilde{M}}(\beta)$	0.086	0.076	0.198	0.216	0.135	1.075	0.977
		$l_{\widetilde{M}^*}(\beta)$	0.092	0.077	0.213	0.232	0.140	1.001	0.964
	6	$l_P(\beta)$	0.284	0.259	0.241	0.373	0.261	0.760	0.686
		$l_{\widetilde{M}}(\beta)$	0.053	0.046	0.152	0.161	0.103	0.989	0.957
		$l_{\widetilde{M}^*}(\beta)$	0.050	0.042	0.154	0.162	0.102	0.968	0.951
	10	$l_P(\beta)$	0.150	0.144	0.136	0.203	0.147	0.865	0.766
		$l_{\widetilde{M}}(\beta)$	0.026	0.022	0.107	0.110	0.070	0.986	0.948
		$l_{\widetilde{M}^*}(\beta)$	0.024	0.019	0.107	0.109	0.070	0.985	0.948
100	4	$l_P(\beta)$	0.445	0.398	0.300	0.536	0.398	0.678	0.464
		$l_{\widetilde{M}}(\beta)$	0.071	0.063	0.142	0.159	0.102	1.057	0.954
		$l_{\widetilde{M}^*}(\beta)$	0.075	0.064	0.151	0.169	0.107	0.995	0.941
	6	$l_P(\beta)$	0.264	0.251	0.165	0.312	0.251	0.776	0.485
		$l_{\widetilde{M}}(\beta)$	0.046	0.040	0.108	0.117	0.075	0.992	0.944
		$l_{\widetilde{M}^*}(\beta)$	0.041	0.035	0.109	0.116	0.074	0.974	0.942
	10	$l_P(\beta)$	0.140	0.137	0.098	0.171	0.137	0.855	0.620
		$l_{\widetilde{M}}(\beta)$	0.018	0.017	0.077	0.079	0.052	0.974	0.950
		$l_{\widetilde{M}^*}(\beta)$	0.016	0.014	0.077	0.078	0.052	0.971	0.948
250	4	$l_P(\beta)$	0.398	0.384	0.171	0.433	0.384	0.720	0.129
		$l_{\widetilde{M}}(\beta)$	0.054	0.050	0.087	0.102	0.068	1.068	0.943
		$l_{\widetilde{M}^*}(\beta)$	0.056	0.051	0.092	0.108	0.070	1.006	0.931
	6	$l_P(\beta)$	0.252	0.247	0.105	0.273	0.247	0.796	0.172
		$l_{\widetilde{M}}(\beta)$	0.038	0.037	0.069	0.079	0.054	1.001	0.932
		$l_{\widetilde{M}^*}(\beta)$	0.034	0.033	0.070	0.078	0.052	0.983	0.932
	10	$l_P(\beta)$	0.136	0.134	0.061	0.149	0.134	0.858	0.295
		$l_{\widetilde{M}}(\beta)$	0.016	0.015	0.048	0.051	0.033	0.976	0.942
		$l_{\widetilde{M}^*}(\beta)$	0.014	0.013	0.048	0.050	0.033	0.975	0.942

between the response variable and the predictor is held unchanged. In the first, missing observations are chosen according to an MCAR mechanism with $\gamma_1 = -0.5$ and $\gamma_2 = 0.3$; in the second, the true missingness generation process is MNAR with $\gamma_1 = -1$, $\gamma_2 = 0.3$ and $\gamma_3 = 2$. The unique covariate is again simulated from the $N(0, 1)$ distribution and the N incidental parameters are consequently set equal to $\lambda_i = \sum_{t=1}^{T} x_{it}/T + u_i$, where $u_i \sim N(0, 1)$ $(i = 1, \dots, N)$. Exploiting the well-known relation between the logistic and normal distributions (Amemiya, 1981) in order to obtain data and quantity of informative groups comparable to the logistic setting, the complete fictitious samples are generated by fixing $\beta = 1/1.6$ under the MCAR scenario and $\beta = 2/1.6$ under the MNAR one.

Tables 3.8 and 3.9 summarize in the customary manner results based on $S = 2000$ simulations of the study regarding MCAR data. Relative behaviours of the three MCAR log-likelihoods illustrated by Table 3.8 do not differentiate from those viewed in Table 3.5 for the logit link. In more detail, the defective performance of $l_P(\beta)$ is greatly corrected by the adjustment proposed by Severini, from any relevant inferential perspective and for all possible couples (T, N) with $T = 4, 6, 10$ and $N = 50, 100, 250$. Moreover, even in this case, accuracies achieved by $l_{\widetilde{M}}(\beta)$ and $l_{\widetilde{M}^*}(\beta)$ are basically indistinguishable, thanks to the validity of the MCAR hypothesis.

Table 3.9: Inference on $\beta = 1/1.6$ in the probit regression for MCAR longitudinal data. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	0.414	0.342	0.483	0.636	0.355	0.493	0.507
		$l_{\widetilde{M}^*}(\psi)$	0.051	0.035	0.239	0.244	0.143	0.880	0.960
	6	$l_P(\psi)$	0.218	0.194	0.259	0.338	0.207	0.601	0.517
		$l_{\widetilde{M}^*}(\psi)$	0.038	0.029	0.164	0.169	0.107	0.918	0.940
	10	$l_P(\psi)$	0.117	0.114	0.150	0.190	0.130	0.759	0.717
		$l_{\widetilde{M}^*}(\psi)$	0.022	0.018	0.110	0.113	0.070	0.969	0.945
100	4	$l_P(\psi)$	0.364	0.335	0.341	0.499	0.342	0.508	0.377
		$l_{\widetilde{M}^*}(\psi)$	0.043	0.031	0.176	0.181	0.108	0.849	0.940
	6	$l_P(\psi)$	0.205	0.192	0.187	0.278	0.196	0.572	0.370
		$l_{\widetilde{M}^*}(\psi)$	0.030	0.025	0.118	0.122	0.081	0.904	0.936
	10	$l_P(\psi)$	0.119	0.116	0.107	0.160	0.119	0.767	0.610
		$l_{\widetilde{M}^*}(\psi)$	0.012	0.008	0.080	0.081	0.053	0.950	0.939
250	4	$l_P(\psi)$	0.335	0.329	0.205	0.393	0.329	0.509	0.150
		$l_{\widetilde{M}^*}(\psi)$	0.026	0.020	0.118	0.120	0.073	0.783	0.921
	6	$l_P(\psi)$	0.204	0.201	0.125	0.239	0.201	0.584	0.207
		$l_{\widetilde{M}^*}(\psi)$	0.020	0.018	0.081	0.083	0.054	0.860	0.927
	10	$l_P(\psi)$	0.125	0.125	0.069	0.143	0.125	0.783	0.358
		$l_{\widetilde{M}^*}(\psi)$	0.007	0.006	0.050	0.051	0.033	0.954	0.943

Like in the preceding part, inferences on β drawn using the MNAR methods for the same samples are displayed by Table 3.9. However, in this probit regression setting the empirical properties of $l_{\widetilde{M}^*}(\psi)$ are in line with the theory and much more favourable than those of the corresponding unmodified function. This is well reflected by all bias and Wald coverage indicators. In addition, the MNAR likelihoods seem to supply better point estimation but less trustworthy confidence intervals compared to their MCAR counterparts. The former in fact still tend to underestimate the variance of their maximizers. Altogether, we can say that with the probit link $l_{\widetilde{M}^*}(\psi)$ succeeds in detecting the underlying ignorable missingness process, which plainly represents a reduced form of the full MNAR model presupposed by that Monte Carlo MPL.

TABLE 3.10: Inference on $\beta = 2/1.6$ in the probit regression for MNAR longitudinal data. Figures based on a simulation study with 4000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	10	$l_P(\psi)$	0.377	0.313	0.392	0.543	0.323	0.335	0.307
		$l_{\widetilde{M}}(\beta)$	-0.084	-0.102	0.174	0.193	0.145	1.097	0.920
		$l_{\widetilde{M}^*}(\psi)$	-0.054	-0.083	0.229	0.235	0.164	0.908	0.925
	20	$l_P(\psi)$	0.183	0.174	0.193	0.266	0.187	0.445	0.404
		$l_{\widetilde{M}}(\beta)$	-0.072	-0.080	0.130	0.148	0.109	0.985	0.886
		$l_{\widetilde{M}^*}(\psi)$	0.015	0.005	0.147	0.148	0.094	0.927	0.947
	30	$l_P(\psi)$	0.104	0.101	0.129	0.166	0.110	0.613	0.618
		$l_{\widetilde{M}}(\beta)$	-0.082	-0.087	0.102	0.131	0.099	0.985	0.840
		$l_{\widetilde{M}^*}(\psi)$	0.035	0.031	0.106	0.111	0.073	0.955	0.939
100	10	$l_P(\psi)$	0.387	0.359	0.286	0.481	0.360	0.158	0.084
		$l_{\widetilde{M}}(\beta)$	-0.093	-0.100	0.122	0.154	0.116	1.109	0.887
		$l_{\widetilde{M}^*}(\psi)$	0.025	-0.019	0.235	0.237	0.118	0.630	0.928
	20	$l_P(\psi)$	0.184	0.183	0.139	0.230	0.185	0.268	0.194
		$l_{\widetilde{M}}(\beta)$	-0.077	-0.080	0.091	0.119	0.089	1.001	0.837
		$l_{\widetilde{M}^*}(\psi)$	0.039	0.033	0.107	0.114	0.070	0.884	0.955
	30	$l_P(\psi)$	0.105	0.105	0.097	0.143	0.108	0.556	0.500
		$l_{\widetilde{M}}(\beta)$	-0.082	-0.084	0.071	0.108	0.086	1.009	0.763
		$l_{\widetilde{M}^*}(\psi)$	0.045	0.042	0.075	0.088	0.058	0.930	0.905
250	10	$l_P(\psi)$	0.358	0.346	0.219	0.420	0.346	0.043	0.017
		$l_{\widetilde{M}}(\beta)$	-0.103	-0.106	0.077	0.128	0.107	1.110	0.768
		$l_{\widetilde{M}^*}(\psi)$	0.032	-0.019	0.194	0.197	0.095	0.471	0.890
	20	$l_P(\psi)$	0.199	0.198	0.111	0.228	0.198	0.053	0.014
		$l_{\widetilde{M}}(\beta)$	-0.082	-0.085	0.057	0.100	0.085	1.010	0.679
		$l_{\widetilde{M}^*}(\psi)$	0.041	0.030	0.090	0.099	0.051	0.648	0.932
	30	$l_P(\psi)$	0.120	0.120	0.076	0.142	0.122	0.203	0.104
		$l_{\widetilde{M}}(\beta)$	-0.089	-0.090	0.044	0.099	0.090	1.015	0.474
		$l_{\widetilde{M}^*}(\psi)$	0.032	0.029	0.051	0.060	0.038	0.865	0.903

Evidence resulting from the last simulation experiment is presented in Table 3.10, which is referred to incomplete datasets with MNAR units having dimensions varying in $N = 50, 100, 250$ and $T = 10, 20, 30$. Notice that the total number of iterations run is raised to $S = 4000$ in order to compensate for the convergence difficulties encountered throughout the estimation of the probit regression with nonignorable missing values. Unlike what emerged by Table 3.7 for the second series of simulations in the MNAR logistic framework, here $l_{\widetilde{M}^*}(\psi)$ always appears to have higher inferential precision than the analytical MPL which ignores the missingness model. Specifically, taking the true nonignorable missing-data mechanism into account via Monte Carlo simulation is practically translated into improved bias and coverage properties of the MNAR MPL.

3.4.4 Logistic regression with missing covariates

In this section, attention is turned to clustered binary observations with incomplete covariate data. Missing covariates are almost ubiquitous in biostatistics and especially present a continual challenge in matched case-control studies (Cho Paik, 2004). A comparative review of methods for inference in GLMs with possibly unobserved predictors is provided by Ibrahim *et al.* (2005). Hereafter, we consider in particular the approach proposed by Lipsitz *et al.* (1998) to handle incomplete covariate information in general logistic regressions with many nuisance parameters. Concentrating on the habitual grouped structure of the data, let us assume the multiple logistic regression model for independent binary observations y_{it}

$$\begin{aligned} Y_{it} \sim Bern(\pi_{it}), \quad \pi_{it} = \pi_{it}(\theta) = \text{logit}^{-1}(\lambda_i + \beta_1 x_{it} + \beta_2 z_{it}), \\ i = 1, \dots, N, \quad t = 1, \dots, T, \end{aligned} \tag{3.28}$$

where the global parameter $\theta = (\psi, \lambda)$ has components $\psi = \beta = (\beta_1, \beta_2) \in \mathrm{IR}^2$ and $\lambda = (\lambda_1, \dots, \lambda_N) \in \mathrm{IR}^N$. It is well established that in this framework one may eliminate from $L(\theta)$ the cluster-specific intercepts by conditioning on suitable sufficient statistics (see, e.g., Bellio and Sartori, 2003). The resulting function is called conditional likelihood and enjoys standard first-order inferential properties (Andersen, 1970), not being affected by the Neyman & Scott problems. Lipsitz *et al.* (1998) modified this function in order to account also for the presence of MAR regressors. Since the MPL is a popular approximation to the original conditional likelihood (Barndorff-Nielsen, 1983), we can start from the intuition of Lipsitz *et al.* (1998) to derive a new version of $l_{\widetilde{M}}(\psi)$ aimed at dealing with missing covariates in logistic regressions.

By way of illustration, suppose that the response y_{it} and the covariate x_{it} are entirely recorded, whereas some values of z_{it} are missing ($i = 1, \dots, N$, $t = 1, \dots, T$). As usual, generalization of the next steps to circumstances with more than one complete predictor and/or more than one incomplete comes naturally. In this setting, redefine the missingness indicator M_{it}, so that $M_{it} = 0$ if z_{it} is observed and $M_{it} = 1$ if z_{it} is missing. Under the hypothesis of MAR covariate data, the conditional distribution of such random variable may be formulated as

$$M_{it}|Y_{it} = y_{it} \sim Bern(\zeta_{it}), \qquad i = 1, \dots, N, \quad t = 1, \dots, T, \tag{3.29}$$

with

$$\zeta_{it} = \zeta_{it}(\gamma) = \text{logit}^{-1}(\gamma_1 + \gamma_2 x_{it} + \gamma_3 y_{it}). \tag{3.30}$$

If one is willing to base inference only on cases with complete predictor information, the reference distribution of the dependent variable is

$$Y_{it}|M_{it} = 0 \sim Bern(\pi^c_{it}), \qquad i = 1, \dots, N, \quad t = 1, \dots, T, \tag{3.31}$$

where the conditional probability of success $\pi^c_{it} = P(Y_{it} = 1|M_{it} = 0)$ can be obtained by straightforward application of Bayes' rule. Specifically, it is simple to prove that the following equality holds:

$$\begin{aligned}\pi^c_{it} = \pi^c_{it}(\theta, \gamma) &= \frac{P(M_{it} = 0|Y_{it} = 1)P(Y_{it} = 1)}{\sum_{y_{it}=0}^{1} P(M_{it} = 0|Y_{it} = y_{it})P(Y_{it} = y_{it})} \\ &= \text{logit}^{-1}\big(\lambda_i + \beta_1 x_{it} + \beta_2 z_{it} + \delta_{it}\big),\end{aligned}$$

where $\delta_{it} = \delta_{it}(\gamma) = \log\{(1 - \zeta^1_{it})/(1 - \zeta^0_{it})\}$, with $\zeta^0_{it} = \text{logit}^{-1}(\gamma_1 + \gamma_2 x_{it})$ and $\zeta^1_{it} = \text{logit}^{-1}(\gamma_1 + \gamma_2 x_{it} + \gamma_3)$. At this point, instead of removing the incidental parameters by conditioning, we may compute the MPL on the grounds of model (3.31) in order to make accurate inference on ψ. The presence of the offset δ_{it} in the logistic regression permits to take the probability of a complete unit having totally observed data into consideration, avoiding so the bias otherwise implied by the exclusion of incomplete cases from the analysis (Lipsitz *et al.*, 1998). Evidently, in practice δ_{it} is unknown and needs to be estimated. One obvious consistent estimate is $\hat{\delta}_{it} = \delta_{it}(\hat{\gamma})$, where $\hat{\gamma}$ results from the ML fit of the logistic regression specified by formulae (3.29) and (3.30). Henceforward, the conditional probability of success obtained upon plug-in of $\hat{\gamma}$ will be indicated by $\pi^{\hat{c}}_{it} = \pi^{\hat{c}}_{it}(\theta) = \pi^c_{it}(\theta, \hat{\gamma})$. Such a substitution entails that the asymptotic variance of the estimator for ψ shall depend upon the distribution of $\hat{\gamma}$.

To facilitate the present exposition, refer to the partition $z = (z_{it}) = (z^{obs}, z^{mis})$, where $z_{it} \in z^{obs}$ if $m_{it} = 0$ and $z_{it} \in z^{mis}$ if $m_{it} = 1$ $(i = 1, \dots, N,\ t = 1, \dots, T)$. Provided the independence of groups in the sample, the expression of the ith contribution to the log-likelihood function for the conditional model (3.31) with $\pi^{c}_{it} = \pi^{\hat{c}}_{it}$ clearly is

$$l^{i,\hat{c}}(\theta) = \sum_{t\colon z_{it} \in z^{obs}} \big\{ y_{it} \log \pi^{\hat{c}}_{it} + (1 - y_{it}) \log(1 - \pi^{\hat{c}}_{it}) \big\}, \qquad i = 1, \dots, N. \tag{3.32}$$

The partial score resulting from differentiating with regard to λ_i the right-hand side of the last equality is then

$$\begin{aligned} l^{\hat{c}}_{\lambda_i}(\theta) &= \sum_{t\colon z_{it} \in z^{obs}} \big\{ y_{it} - \pi^{\hat{c}}_{it}(\theta) \big\} \\ &= \sum_{t\colon z_{it} \in z^{obs}} \big\{ y_{it} - \text{logit}^{-1}\big(\lambda_i + \beta_1 x_{it} + \beta_2 z_{it} + \hat{\delta}_{it}\big) \big\}, \end{aligned}$$

and numerical solution for λ_i of the ith cluster-related likelihood equation $l^{\hat{c}}_{\lambda_i}(\theta) = 0$ gives the constrained ML estimate $\hat{\lambda}^{\hat{c}}_{i\psi}$ $(i = 1, \dots, N)$. As always, by replacement of the overall parameter with $\hat{\theta}^{\hat{c}}_{\psi} = (\psi, \hat{\lambda}^{\hat{c}}_{\psi})$ in (3.32) one obtains the ith additive component of the profile log-likelihood $l^{\hat{c}}_P(\psi) = \sum_{i=1}^N l^{i,\hat{c}}_P(\psi) = \sum_{i=1}^N l^{i,\hat{c}}(\hat{\theta}^{\hat{c}}_{\psi})$. Maximization of the latter yields to the ML estimate $\hat{\psi}^{\hat{c}}$, so that $\hat{\theta}^{\hat{c}} = (\hat{\psi}^{\hat{c}}, \hat{\lambda}_{\hat{\psi}^{\hat{c}}})$. In order to calculate the modification term of Severini $\widetilde{M}^{\hat{c}}(\psi)$, we derive the expression for $|j^{\hat{c}}_{\lambda\lambda}(\theta)| = \prod_{i=1}^N j^{\hat{c}}_{\lambda_i\lambda_i}(\theta)$, where

$$j^{\hat{c}}_{\lambda_i\lambda_i}(\theta) = \sum_{t\colon z_{it} \in z^{obs}} \text{logit}^{-1}\big(\lambda_i + \beta_1 x_{it} + \beta_2 z_{it} + \hat{\delta}_{it}\big) \big\{ 1 - \text{logit}^{-1}\big(\lambda_i + \beta_1 x_{it} + \beta_2 z_{it} + \hat{\delta}_{it}\big) \big\}.$$

Contrary to most of the situations discussed earlier in the chapter, now it is not necessary to approximate the expectation in $l_{\widetilde{M}^{\hat{c}}}(\psi)$ via the Monte Carlo method. Indeed, in the current case computing such expected value with respect to the conditional distribution of Y_{it} given $M_{it} = 0$ is correct, as we found the manner to model it by accounting also for the missingness mechanism. Particularly, one may easily show that

$$I^{\hat{c}}_{\lambda_i\lambda_i}(\hat{\theta}_{\psi}; \hat{\theta}) = \sum_{t\colon z_{it} \in z^{obs}} \big[1 - \text{logit}^{-1}\big(\hat{\lambda}_i + \hat{\beta}_1 x_{it} + \hat{\beta}_2 z_{it} + \hat{\delta}_{it}\big) \big], \qquad i = 1, \dots, N,$$

and hence, as typically occurs in the presence of a logit link, this part of Severini's adjustment does not play a role in the estimation of ψ. The appropriate version of the MPL in this logistic regression with one MAR predictor can be then formulated as $l_{\widetilde{M}^{\hat{c}}}(\psi) = l^{\hat{c}}_P(\psi) + \widetilde{M}^{\hat{c}}(\psi)$, where $\widetilde{M}^{\hat{c}}(\psi) = \frac{1}{2} \log |j^{\hat{c}}_{\lambda\lambda}(\hat{\theta}^{\hat{c}}_{\psi})|$.

In the subsequent part, results of simulation studies performed to compare the inferential accuracy of several methods will be reported. Precisely, the competitors shall be the profile and modified profile log-likelihoods $l_P(\psi)$ and $l_{\widetilde{M}}(\psi)$, relating to the unconditional model (3.28) under the MCAR covariate assumption, and their homologous in the MAR setting computed on the basis of the conditional distribution (3.31), i.e. $l_P^{\hat{c}}(\psi)$ and $l_{\widetilde{M}^{\hat{c}}}(\psi)$.

Simulation studies

All the presented experiments are carried out on $S = 2000$ samples, simulated with single group size equal to $T = 4, 6, 10$ and number of clusters equal to $N = 50, 100, 250$. For each pair (T, N), complete covariates x_{it} and z_{it} are independently drawn from the standard normal distribution and binary responses y_{it} are generated under model (3.28) with $\beta_1 = -1$, $\beta_2 = 2$ and $\lambda_i = \sum_{t=1}^{T} x_{it}/T + u_i$, where $u_i \sim N(0, 1)$ $(i = 1, \ldots, N)$. The simulation setups can be characterized by the specified probability of missing values in z_{it}. Specifically, in the first two studies we consider the MAR structure hypothesized in (3.30), while in the other two the true missingness process is nonignorable, with $\zeta_{it} = \text{logit}^{-1}(\gamma_1 + \gamma_2 x_{it} + \gamma_3 y_{it} + \gamma_4 z_{it})$.

Results in Tables 3.11 and 3.12 refer to inference on β_1 and β_2, respectively, under the first MAR scenario, with true probability of unobserved z_{it} fixed at $\zeta_{it} = \text{logit}^{-1}(-1 - 0.5x_{it} + 0.5y_{it})$ in order to get datasets with a proportion of missing values ranging between 30% and 35%. In agreement with the simulation-based evidence shown by Lipsitz *et al.* (1998), the former table is definitely the most interesting from the viewpoint of comparing the procedures which do not acknowledge the missing-data problem to those which do. Indeed, no such relevant differences in the estimation accuracy of the coefficient associated to the incomplete regressor are recorded in Table 3.12. Conversely, Table 3.11 not only illustrates the well-known inferential enhancements determined by adjusting the profile likelihood in models with incidental parameters, but also reflects the disparity in supposition about the generating process of missingness. Quite peculiarly, at the same time the validity of such assumption seems to refine the precision of the MPL on one side and to further deteriorate the quality of ordinary ML inference on the other. In greater detail, the worse performance of $l_P^{\hat{c}}(\psi)$ with regard to $l_P(\psi)$ is principally due to the larger empirical bias of its estimator, but some plausible justification for this finding at the moment cannot be provided. The opposite comment applies instead to $l_{\widetilde{M}^{\hat{c}}}(\psi)$ and $l_{\widetilde{M}}(\psi)$: in this case, accounting for the MAR predictor sensibly results in more adequate point and interval estimation of β_1 for almost all the sample sizes in question.

Table 3.11: Inference on $\beta_1 = -1$ in the logistic regression for stratified data with MAR covariate generated with missingness probability $\zeta_{it} = \text{logit}^{-1}(-1 - 0.5x_{it} + 0.5y_{it})$. Figures based on a simulation study with 2000 trials.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	-0.916	-0.715	1.195	1.506	0.740	0.546	0.758
		$l_{\widetilde{M}}(\psi)$	0.101	0.130	0.707	0.714	0.248	0.519	0.939
		$l_P^{\hat{c}}(\psi)$	-0.968	-0.761	1.196	1.538	0.775	0.546	0.725
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.051	0.077	0.667	0.669	0.231	0.550	0.958
	6	$l_P(\psi)$	-0.581	-0.502	0.557	0.805	0.507	0.723	0.736
		$l_{\widetilde{M}}(\psi)$	0.091	0.108	0.241	0.257	0.176	1.104	0.948
		$l_P^{\hat{c}}(\psi)$	-0.630	-0.547	0.556	0.840	0.549	0.724	0.696
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.042	0.052	0.240	0.244	0.165	1.107	0.967
	10	$l_P(\psi)$	-0.277	-0.254	0.299	0.408	0.268	0.836	0.818
		$l_{\widetilde{M}}(\psi)$	0.028	0.039	0.205	0.207	0.141	1.014	0.947
		$l_P^{\hat{c}}(\psi)$	-0.330	-0.305	0.298	0.445	0.309	0.839	0.748
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.024	-0.014	0.204	0.206	0.135	1.018	0.958
100	4	$l_P(\psi)$	-0.785	-0.695	0.667	1.030	0.701	0.689	0.626
		$l_{\widetilde{M}}(\psi)$	0.193	0.191	0.218	0.291	0.210	1.170	0.886
		$l_P^{\hat{c}}(\psi)$	-0.837	-0.742	0.667	1.070	0.743	0.689	0.587
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.141	0.138	0.218	0.260	0.178	1.170	0.926
	6	$l_P(\psi)$	-0.477	-0.444	0.351	0.592	0.444	0.757	0.590
		$l_{\widetilde{M}}(\psi)$	0.089	0.098	0.178	0.199	0.141	1.059	0.924
		$l_P^{\hat{c}}(\psi)$	-0.527	-0.496	0.349	0.632	0.496	0.760	0.525
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.039	0.045	0.177	0.181	0.127	1.066	0.958
	10	$l_P(\psi)$	-0.236	-0.228	0.196	0.307	0.229	0.842	0.710
		$l_{\widetilde{M}}(\psi)$	0.043	0.045	0.139	0.145	0.099	1.004	0.932
		$l_P^{\hat{c}}(\psi)$	-0.287	-0.280	0.194	0.347	0.280	0.848	0.603
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.008	-0.006	0.138	0.138	0.091	1.014	0.955
250	4	$l_P(\psi)$	-0.693	-0.663	0.362	0.782	0.663	0.717	0.280
		$l_{\widetilde{M}}(\psi)$	0.217	0.218	0.129	0.253	0.218	1.155	0.702
		$l_P^{\hat{c}}(\psi)$	-0.744	-0.714	0.361	0.827	0.714	0.719	0.215
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.167	0.168	0.129	0.211	0.170	1.161	0.826
	6	$l_P(\psi)$	-0.433	-0.424	0.210	0.481	0.424	0.780	0.277
		$l_{\widetilde{M}}(\psi)$	0.102	0.103	0.112	0.152	0.111	1.062	0.864
		$l_P^{\hat{c}}(\psi)$	-0.484	-0.476	0.210	0.527	0.476	0.782	0.185
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.052	0.052	0.111	0.123	0.083	1.066	0.924
	10	$l_P(\psi)$	-0.230	-0.226	0.127	0.263	0.226	0.845	0.446
		$l_{\widetilde{M}}(\psi)$	0.052	0.055	0.090	0.104	0.073	1.006	0.899
		$l_P^{\hat{c}}(\psi)$	-0.282	-0.279	0.127	0.309	0.279	0.849	0.266
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.000	0.001	0.090	0.090	0.059	1.013	0.956

TABLE 3.12: Inference on $\beta_2 = 2$ in the logistic regression for stratified data with MAR covariate generated with missingness probability $\zeta_{it} = \text{logit}^{-1}(-1 - 0.5x_{it} + 0.5y_{it})$. Figures based on a simulation study with 2000 trials.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	2.088	1.641	2.815	3.505	1.641	0.351	0.515
		$l_{\widetilde{M}}(\psi)$	-0.052	-0.158	2.317	2.317	0.265	0.188	0.964
		$l_P^{\hat{c}}(\psi)$	2.080	1.640	2.600	3.329	1.640	0.376	0.515
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.062	-0.159	2.067	2.067	0.265	0.210	0.965
	6	$l_P(\psi)$	1.342	1.170	0.979	1.661	1.170	0.624	0.440
		$l_{\widetilde{M}}(\psi)$	-0.079	-0.089	0.297	0.308	0.215	1.155	0.955
		$l_P^{\hat{c}}(\psi)$	1.342	1.170	0.979	1.661	1.170	0.624	0.440
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.079	-0.089	0.297	0.308	0.215	1.155	0.955
	10	$l_P(\psi)$	0.673	0.618	0.436	0.802	0.618	0.798	0.562
		$l_{\widetilde{M}}(\psi)$	0.032	0.007	0.263	0.265	0.175	1.036	0.966
		$l_P^{\hat{c}}(\psi)$	0.673	0.618	0.436	0.802	0.618	0.798	0.563
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.032	0.007	0.263	0.265	0.175	1.036	0.966
100	4	$l_P(\psi)$	1.766	1.599	1.036	2.048	1.599	0.617	0.197
		$l_{\widetilde{M}}(\psi)$	-0.247	-0.254	0.208	0.323	0.259	1.373	0.897
		$l_P^{\hat{c}}(\psi)$	1.766	1.599	1.038	2.048	1.599	0.617	0.197
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.248	-0.255	0.208	0.323	0.259	1.373	0.896
	6	$l_P(\psi)$	1.133	1.088	0.532	1.251	1.088	0.735	0.178
		$l_{\widetilde{M}}(\psi)$	-0.092	-0.091	0.203	0.223	0.150	1.181	0.942
		$l_P^{\hat{c}}(\psi)$	1.133	1.087	0.532	1.251	1.087	0.735	0.178
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.093	-0.091	0.203	0.223	0.150	1.181	0.942
	10	$l_P(\psi)$	0.606	0.588	0.290	0.672	0.588	0.806	0.286
		$l_{\widetilde{M}}(\psi)$	0.017	0.012	0.186	0.187	0.129	1.007	0.962
		$l_P^{\hat{c}}(\psi)$	0.606	0.589	0.290	0.672	0.589	0.806	0.286
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.018	0.012	0.186	0.187	0.129	1.007	0.962
250	4	$l_P(\psi)$	1.555	1.502	0.519	1.639	1.502	0.708	0.005
		$l_{\widetilde{M}}(\psi)$	-0.312	-0.313	0.124	0.335	0.313	1.382	0.559
		$l_P^{\hat{c}}(\psi)$	1.555	1.502	0.519	1.639	1.502	0.708	0.005
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.312	-0.313	0.124	0.335	0.313	1.382	0.558
	6	$l_P(\psi)$	1.034	1.026	0.314	1.081	1.026	0.752	0.007
		$l_{\widetilde{M}}(\psi)$	-0.087	-0.087	0.132	0.158	0.107	1.149	0.922
		$l_P^{\hat{c}}(\psi)$	1.034	1.026	0.314	1.081	1.026	0.752	0.007
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.087	-0.087	0.132	0.158	0.107	1.149	0.921
	10	$l_P(\psi)$	0.600	0.596	0.177	0.626	0.596	0.823	0.014
		$l_{\widetilde{M}}(\psi)$	0.002	0.001	0.111	0.111	0.075	1.042	0.953
		$l_P^{\hat{c}}(\psi)$	0.600	0.596	0.177	0.626	0.596	0.823	0.013
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.002	0.001	0.111	0.111	0.075	1.042	0.953

In the second setup with MAR covariate, the probability of one missing datum is $\zeta_{it} = \text{logit}^{-1}(-0.5 + x_{it} + y_{it})$, so that about 50% of the values z_{it} are deleted in the final simulated samples. Results of the study conducted with such experimental design can be found in Tables 3.13 and 3.14. Again, the latter basically outlines that inference on β_2 is unaffected by the incompleteness of the corresponding regressor. Rather, a more careful analysis of Table 3.13 appears worthwhile. Relative patterns in the behaviour of the two profile and the two modified profile log-likelihoods change when the amount of unrecorded observations in the data grows. Now, $l_P^{\hat{c}}(\psi)$ outperforms $l_P(\psi)$ in terms of empirical bias and coverages properties for any values of T and N. By contrast, the superiority of $l_{\widetilde{M}^{\hat{c}}}(\psi)$ on $l_{\widetilde{M}}(\psi)$ which incorrectly postulates an MCAR missing-covariate process remains unquestionable only when $T = 10$. In our view, the reason of this trend reversal is unfortunately not obvious.

The missingness mechanism considered in the third experiment is nonignorable, with probability of not observing z_{it} set equal to $\zeta_{it} = \text{logit}^{-1}(-1 - 0.5x_{it} + 0.5y_{it} + 0.5z_{it})$. The latter is chosen in such a way as to obtain in the artificial samples a percentage of missing data varying in 30-35%, as was in the first MAR framework. Notice that, against this background, $l_P^{\hat{c}}(\psi)$ and $l_{\widetilde{M}^{\hat{c}}}(\psi)$ now underspecify the true model of missingness, neglecting the dependence of ζ_{it} on the possibly unobserved predictor. Table 3.15 and 3.16 display the outcomes of this study. Usual comments are pertinent to the second table associated with inference on β_2. As for the estimation of the other parameter of interest β_1, the differing tendencies identified in the first simulation study are recognisable also in Table 3.15. Furthermore, the global performance of $l_P^{\hat{c}}(\psi)$ and $l_{\widetilde{M}^{\hat{c}}}(\psi)$ does not look particularly altered by the misspecification of the missing-data process.

The last scenario examined relates to the case of an MNAR covariate whose missingness is described by the probability $\zeta_{it} = \text{logit}^{-1}(-0.5 + x_{it} + y_{it} + 2z_{it})$. Such a definition delivers datasets where z_{it} results unrecorded around 50% of the times. The various aspects pertaining to inference on β_1 and β_2 in this framework are detailed by Tables 3.17 and 3.18. Unsurprisingly, the general accuracy achieved by the four methods is the least satisfactory among the several situations discussed, especially in terms of empirical bias of the estimators. Similarly to the second MAR setting, properties of $l_P^{\hat{c}}(\psi)$ in Table 3.17 are surely more valuable than those of its direct competitor, even if now the former does not take the right missing-data mechanism into consideration. Yet $l_{\widetilde{M}}(\psi)$, derived under the hypothesis of MCAR predictor, proves to be the most reliable inferential tool for any couple (T, N). The performance of $l_{\widetilde{M}^{\hat{c}}}(\psi)$ is remarkably poorer than in the previous circumstance of nonignorable missingness. Aside from the larger

TABLE 3.13: Inference on $\beta_1 = -1$ in the logistic regression for stratified data with MAR covariate generated with missingness probability $\zeta_{it} = \text{logit}^{-1}(-0.5 + x_{it} + y_{it})$. Figures based on a simulation study with 2000 trials.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	-5.012	-1.539	30.978	31.373	1.557	0.177	0.767
		$l_{\widetilde{M}}(\psi)$	-3.268	-0.039	31.065	31.228	0.288	0.015	0.927
		$l_P^{\hat{c}}(\psi)$	-4.288	-1.317	17.465	17.979	1.347	0.310	0.849
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-2.575	0.180	17.742	17.923	0.339	0.026	0.890
	6	$l_P(\psi)$	-1.518	-1.136	2.118	2.606	1.137	0.395	0.576
		$l_{\widetilde{M}}(\psi)$	-0.203	-0.056	1.872	1.882	0.209	0.196	0.975
		$l_P^{\hat{c}}(\psi)$	-1.300	-0.920	2.102	2.471	0.925	0.386	0.704
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.009	0.156	1.866	1.865	0.246	0.196	0.934
	10	$l_P(\psi)$	-0.800	-0.761	0.459	0.923	0.761	0.763	0.382
		$l_{\widetilde{M}}(\psi)$	-0.197	-0.194	0.241	0.311	0.210	1.050	0.926
		$l_P^{\hat{c}}(\psi)$	-0.591	-0.553	0.451	0.744	0.553	0.775	0.640
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.012	0.020	0.235	0.236	0.159	1.072	0.965
100	4	$l_P(\psi)$	-1.497	-1.269	1.207	1.923	1.269	0.603	0.504
		$l_{\widetilde{M}}(\psi)$	0.036	0.063	0.560	0.561	0.179	0.591	0.978
		$l_P^{\hat{c}}(\psi)$	-1.287	-1.068	1.202	1.760	1.069	0.606	0.642
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.240	0.275	0.616	0.661	0.287	0.535	0.883
	6	$l_P(\psi)$	-1.099	-1.039	0.582	1.244	1.039	0.712	0.244
		$l_{\widetilde{M}}(\psi)$	-0.042	-0.042	0.205	0.209	0.142	1.164	0.983
		$l_P^{\hat{c}}(\psi)$	-0.892	-0.833	0.576	1.062	0.833	0.719	0.444
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.164	0.162	0.202	0.260	0.187	1.176	0.908
	10	$l_P(\psi)$	-0.738	-0.717	0.312	0.801	0.717	0.783	0.146
		$l_{\widetilde{M}}(\psi)$	-0.175	-0.171	0.174	0.247	0.177	1.044	0.877
		$l_P^{\hat{c}}(\psi)$	-0.529	-0.506	0.306	0.611	0.506	0.797	0.439
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.034	0.038	0.170	0.173	0.118	1.069	0.953
250	4	$l_P(\psi)$	-1.294	-1.216	0.587	1.420	1.216	0.688	0.107
		$l_{\widetilde{M}}(\psi)$	0.088	0.087	0.153	0.176	0.122	1.267	0.960
		$l_P^{\hat{c}}(\psi)$	-1.087	-1.011	0.583	1.233	1.011	0.693	0.246
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.293	0.291	0.152	0.330	0.291	1.271	0.699
	6	$l_P(\psi)$	-1.011	-0.991	0.354	1.071	0.991	0.729	0.029
		$l_{\widetilde{M}}(\psi)$	-0.073	-0.075	0.142	0.159	0.108	1.126	0.966
		$l_P^{\hat{c}}(\psi)$	-0.800	-0.774	0.351	0.874	0.774	0.735	0.157
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.137	0.140	0.140	0.196	0.148	1.135	0.877
	10	$l_P(\psi)$	-0.718	-0.716	0.197	0.744	0.716	0.800	0.006
		$l_{\widetilde{M}}(\psi)$	-0.156	-0.156	0.111	0.191	0.157	1.060	0.776
		$l_P^{\hat{c}}(\psi)$	-0.507	-0.506	0.194	0.543	0.506	0.813	0.126
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.054	0.056	0.109	0.122	0.083	1.080	0.940

Table 3.14: Inference on $\beta_2 = 2$ in the logistic regression for stratified data with MAR covariate generated with missingness probability $\zeta_{it} = \text{logit}^{-1}(-0.5 + x_{it} + y_{it})$. Figures based on a simulation study with 2000 trials.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	8.029	2.210	51.457	52.067	2.210	0.179	0.761
		$l_{\widetilde{M}}(\psi)$	5.167	-0.322	51.669	51.913	0.412	0.009	0.861
		$l_P^{\hat{c}}(\psi)$	7.917	2.216	52.119	52.704	2.216	0.148	0.763
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	5.102	-0.318	52.324	52.559	0.413	0.009	0.862
	6	$l_P(\psi)$	2.142	1.592	3.653	4.234	1.592	0.320	0.488
		$l_{\widetilde{M}}(\psi)$	-0.040	-0.262	3.373	3.372	0.288	0.111	0.911
		$l_P^{\hat{c}}(\psi)$	2.139	1.592	3.416	4.029	1.592	0.324	0.487
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.032	-0.263	3.152	3.151	0.290	0.119	0.909
	10	$l_P(\psi)$	0.957	0.880	0.638	1.150	0.880	0.736	0.507
		$l_{\widetilde{M}}(\psi)$	-0.035	-0.062	0.296	0.298	0.197	1.069	0.956
		$l_P^{\hat{c}}(\psi)$	0.957	0.879	0.638	1.150	0.879	0.736	0.506
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.036	-0.063	0.297	0.299	0.197	1.069	0.956
100	4	$l_P(\psi)$	2.092	1.740	1.634	2.654	1.740	0.539	0.342
		$l_{\widetilde{M}}(\psi)$	-0.351	-0.395	0.851	0.920	0.396	0.381	0.813
		$l_P^{\hat{c}}(\psi)$	2.095	1.748	1.612	2.643	1.748	0.549	0.337
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.341	-0.392	0.873	0.937	0.395	0.372	0.814
	6	$l_P(\psi)$	1.454	1.343	0.821	1.669	1.343	0.665	0.242
		$l_{\widetilde{M}}(\psi)$	-0.245	-0.249	0.228	0.335	0.264	1.215	0.862
		$l_P^{\hat{c}}(\psi)$	1.453	1.342	0.821	1.669	1.342	0.665	0.242
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.245	-0.250	0.228	0.334	0.264	1.216	0.862
	10	$l_P(\psi)$	0.870	0.825	0.410	0.962	0.825	0.767	0.209
		$l_{\widetilde{M}}(\psi)$	-0.057	-0.069	0.197	0.205	0.141	1.104	0.951
		$l_P^{\hat{c}}(\psi)$	0.870	0.825	0.410	0.962	0.825	0.767	0.208
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.057	-0.070	0.197	0.205	0.140	1.104	0.952
250	4	$l_P(\psi)$	1.823	1.701	0.780	1.983	1.701	0.662	0.034
		$l_{\widetilde{M}}(\psi)$	-0.471	-0.478	0.139	0.491	0.478	1.436	0.307
		$l_P^{\hat{c}}(\psi)$	1.825	1.703	0.780	1.984	1.703	0.663	0.033
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.470	-0.476	0.139	0.490	0.476	1.435	0.308
	6	$l_P(\psi)$	1.337	1.296	0.462	1.415	1.296	0.717	0.014
		$l_{\widetilde{M}}(\psi)$	-0.217	-0.223	0.142	0.260	0.224	1.254	0.796
		$l_P^{\hat{c}}(\psi)$	1.337	1.296	0.462	1.415	1.296	0.717	0.014
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.217	-0.223	0.142	0.260	0.224	1.253	0.796
	10	$l_P(\psi)$	0.837	0.828	0.243	0.872	0.828	0.790	0.009
		$l_{\widetilde{M}}(\psi)$	-0.087	-0.088	0.115	0.144	0.103	1.139	0.908
		$l_P^{\hat{c}}(\psi)$	0.838	0.828	0.243	0.873	0.828	0.790	0.009
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.086	-0.088	0.115	0.144	0.103	1.139	0.908

TABLE 3.15: Inference on $\beta_1 = -1$ in the logistic regression for stratified data with MNAR covariate generated with missingness probability $\zeta_{it} = \text{logit}^{-1}(-1 - 0.5x_{it} + 0.5y_{it} + 0.5z_{it})$. Figures based on a simulation study with 2000 trials.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	-0.928	-0.729	1.283	1.583	0.748	0.518	0.784
		$l_{\widetilde{M}}(\psi)$	0.049	0.090	0.850	0.852	0.236	0.451	0.941
		$l_P^{\hat{c}}(\psi)$	-1.014	-0.816	1.260	1.617	0.826	0.527	0.744
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.033	0.004	0.785	0.786	0.225	0.490	0.965
	6	$l_P(\psi)$	-0.560	-0.479	0.562	0.793	0.489	0.699	0.734
		$l_{\widetilde{M}}(\psi)$	0.066	0.081	0.255	0.264	0.180	1.046	0.940
		$l_P^{\hat{c}}(\psi)$	-0.652	-0.567	0.560	0.859	0.570	0.701	0.656
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.027	-0.012	0.254	0.255	0.168	1.054	0.967
	10	$l_P(\psi)$	-0.280	-0.256	0.297	0.408	0.274	0.835	0.809
		$l_{\widetilde{M}}(\psi)$	0.021	0.025	0.205	0.206	0.138	1.009	0.949
		$l_P^{\hat{c}}(\psi)$	-0.372	-0.353	0.296	0.475	0.354	0.837	0.702
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.072	-0.067	0.204	0.216	0.139	1.013	0.963
100	4	$l_P(\psi)$	-0.789	-0.706	0.693	1.050	0.707	0.663	0.619
		$l_{\widetilde{M}}(\psi)$	0.165	0.173	0.258	0.307	0.202	1.002	0.901
		$l_P^{\hat{c}}(\psi)$	-0.878	-0.799	0.692	1.117	0.799	0.664	0.549
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.078	0.083	0.227	0.240	0.163	1.142	0.954
	6	$l_P(\psi)$	-0.464	-0.433	0.345	0.578	0.433	0.758	0.598
		$l_{\widetilde{M}}(\psi)$	0.080	0.087	0.178	0.195	0.137	1.056	0.927
		$l_P^{\hat{c}}(\psi)$	-0.549	-0.513	0.343	0.648	0.513	0.763	0.472
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.006	0.003	0.176	0.177	0.118	1.066	0.967
	10	$l_P(\psi)$	-0.233	-0.224	0.193	0.302	0.225	0.849	0.714
		$l_{\widetilde{M}}(\psi)$	0.039	0.045	0.138	0.143	0.100	1.009	0.940
		$l_P^{\hat{c}}(\psi)$	-0.320	-0.309	0.192	0.373	0.309	0.855	0.528
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.048	-0.042	0.137	0.145	0.094	1.018	0.949
250	4	$l_P(\psi)$	-0.683	-0.654	0.355	0.770	0.654	0.736	0.284
		$l_{\widetilde{M}}(\psi)$	0.200	0.204	0.130	0.238	0.204	1.174	0.767
		$l_P^{\hat{c}}(\psi)$	-0.774	-0.749	0.353	0.851	0.749	0.739	0.178
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.109	0.114	0.129	0.169	0.126	1.183	0.919
	6	$l_P(\psi)$	-0.428	-0.416	0.208	0.476	0.416	0.789	0.286
		$l_{\widetilde{M}}(\psi)$	0.088	0.093	0.113	0.143	0.106	1.069	0.885
		$l_P^{\hat{c}}(\psi)$	-0.515	-0.505	0.207	0.555	0.505	0.792	0.139
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.001	0.006	0.112	0.112	0.075	1.073	0.965
	10	$l_P(\psi)$	-0.221	-0.216	0.126	0.255	0.216	0.856	0.478
		$l_{\widetilde{M}}(\psi)$	0.050	0.051	0.091	0.104	0.074	1.007	0.913
		$l_P^{\hat{c}}(\psi)$	-0.311	-0.307	0.126	0.336	0.307	0.860	0.202
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.040	-0.040	0.091	0.099	0.065	1.013	0.939

Table 3.16: Inference on $\beta_2 = 2$ in the logistic regression for stratified data with MNAR covariate generated with missingness probability $\zeta_{it} = \text{logit}^{-1}(-1 - 0.5x_{it} + 0.5y_{it} + 0.5z_{it})$. Figures based on a simulation study with 2000 trials.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	1.983	1.544	2.148	2.923	1.544	0.463	0.570
		$l_{\widetilde{M}}(\psi)$	-0.040	-0.124	1.303	1.304	0.268	0.357	0.961
		$l_P^{\hat{c}}(\psi)$	1.980	1.543	2.133	2.910	1.543	0.466	0.569
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.049	-0.124	1.231	1.232	0.269	0.378	0.962
	6	$l_P(\psi)$	1.212	1.070	0.899	1.509	1.070	0.663	0.521
		$l_{\widetilde{M}}(\psi)$	-0.078	-0.097	0.314	0.324	0.222	1.132	0.949
		$l_P^{\hat{c}}(\psi)$	1.212	1.069	0.899	1.509	1.069	0.663	0.522
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.078	-0.098	0.314	0.324	0.223	1.132	0.949
	10	$l_P(\psi)$	0.596	0.565	0.429	0.734	0.565	0.806	0.625
		$l_{\widetilde{M}}(\psi)$	-0.010	-0.027	0.268	0.268	0.181	1.023	0.953
		$l_P^{\hat{c}}(\psi)$	0.596	0.565	0.429	0.734	0.565	0.806	0.625
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.010	-0.026	0.268	0.268	0.181	1.023	0.953
100	4	$l_P(\psi)$	1.656	1.487	1.040	1.956	1.487	0.613	0.262
		$l_{\widetilde{M}}(\psi)$	-0.253	-0.267	0.334	0.419	0.275	0.884	0.881
		$l_P^{\hat{c}}(\psi)$	1.656	1.487	1.040	1.955	1.487	0.613	0.262
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.259	-0.267	0.220	0.340	0.274	1.340	0.881
	6	$l_P(\psi)$	1.029	0.970	0.528	1.157	0.970	0.731	0.250
		$l_{\widetilde{M}}(\psi)$	-0.121	-0.125	0.211	0.243	0.170	1.157	0.927
		$l_P^{\hat{c}}(\psi)$	1.029	0.970	0.528	1.156	0.970	0.730	0.251
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.121	-0.125	0.211	0.243	0.170	1.157	0.926
	10	$l_P(\psi)$	0.533	0.504	0.288	0.606	0.504	0.807	0.404
		$l_{\widetilde{M}}(\psi)$	-0.024	-0.038	0.188	0.189	0.130	1.002	0.944
		$l_P^{\hat{c}}(\psi)$	0.533	0.504	0.288	0.606	0.504	0.807	0.403
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.024	-0.038	0.188	0.189	0.130	1.002	0.945
250	4	$l_P(\psi)$	1.456	1.402	0.532	1.550	1.402	0.688	0.018
		$l_{\widetilde{M}}(\psi)$	-0.314	-0.318	0.133	0.341	0.318	1.329	0.578
		$l_P^{\hat{c}}(\psi)$	1.456	1.402	0.532	1.550	1.402	0.688	0.018
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.314	-0.318	0.133	0.341	0.318	1.329	0.578
	6	$l_P(\psi)$	0.944	0.923	0.313	0.994	0.923	0.753	0.022
		$l_{\widetilde{M}}(\psi)$	-0.107	-0.107	0.138	0.175	0.127	1.125	0.897
		$l_P^{\hat{c}}(\psi)$	0.944	0.923	0.313	0.995	0.923	0.753	0.022
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.107	-0.107	0.138	0.175	0.126	1.125	0.897
	10	$l_P(\psi)$	0.502	0.496	0.175	0.532	0.496	0.830	0.072
		$l_{\widetilde{M}}(\psi)$	-0.044	-0.048	0.115	0.123	0.084	1.025	0.931
		$l_P^{\hat{c}}(\psi)$	0.503	0.496	0.175	0.532	0.496	0.830	0.072
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.044	-0.048	0.115	0.123	0.084	1.025	0.932

Table 3.17: Inference on $\beta_1 = -1$ in the logistic regression for stratified data with MNAR covariate generated with missingness probability $\zeta_{it} = \text{logit}^{-1}(-0.5 + x_{it} + y_{it} + 2z_{it})$. Figures based on a simulation study with 2000 trials.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	-11.264	-1.330	129.903	130.358	1.344	0.075	0.842
		$l_{\widetilde{M}}(\psi)$	-9.711	-0.161	129.982	130.312	0.362	0.004	0.936
		$l_P^{\hat{c}}(\psi)$	-8.080	-0.872	81.790	82.167	0.999	0.108	0.924
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-6.503	0.299	81.868	82.105	0.457	0.007	0.860
	6	$l_P(\psi)$	-1.167	-0.896	1.667	2.035	0.896	0.377	0.634
		$l_{\widetilde{M}}(\psi)$	-0.185	-0.104	1.376	1.388	0.230	0.266	0.971
		$l_P^{\hat{c}}(\psi)$	-0.753	-0.487	1.557	1.729	0.555	0.409	0.850
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.229	0.294	1.232	1.252	0.332	0.296	0.862
	10	$l_P(\psi)$	-0.660	-0.618	0.440	0.793	0.618	0.740	0.501
		$l_{\widetilde{M}}(\psi)$	-0.164	-0.156	0.249	0.298	0.196	0.989	0.941
		$l_P^{\hat{c}}(\psi)$	-0.209	-0.162	0.419	0.469	0.273	0.775	0.897
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.285	0.297	0.234	0.369	0.303	1.048	0.736
100	4	$l_P(\psi)$	-1.352	-1.099	1.630	2.118	1.099	0.442	0.575
		$l_{\widetilde{M}}(\psi)$	-0.054	0.017	1.255	1.256	0.194	0.271	0.968
		$l_P^{\hat{c}}(\psi)$	-0.943	-0.693	1.646	1.897	0.720	0.429	0.808
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.348	0.429	1.313	1.358	0.433	0.259	0.747
	6	$l_P(\psi)$	-0.957	-0.904	0.515	1.087	0.904	0.741	0.294
		$l_{\widetilde{M}}(\psi)$	-0.060	-0.057	0.201	0.209	0.141	1.169	0.984
		$l_P^{\hat{c}}(\psi)$	-0.561	-0.505	0.503	0.753	0.506	0.757	0.712
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.332	0.330	0.196	0.385	0.331	1.196	0.709
	10	$l_P(\psi)$	-0.608	-0.592	0.284	0.671	0.592	0.798	0.252
		$l_{\widetilde{M}}(\psi)$	-0.155	-0.150	0.176	0.234	0.165	1.013	0.891
		$l_P^{\hat{c}}(\psi)$	-0.200	-0.188	0.272	0.338	0.221	0.830	0.851
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.252	0.257	0.167	0.302	0.258	1.061	0.676
250	4	$l_P(\psi)$	-1.164	-1.096	0.590	1.305	1.096	0.674	0.187
		$l_{\widetilde{M}}(\psi)$	0.033	0.035	0.179	0.182	0.124	1.171	0.974
		$l_P^{\hat{c}}(\psi)$	-0.780	-0.717	0.580	0.972	0.717	0.685	0.527
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.413	0.410	0.174	0.448	0.410	1.204	0.480
	6	$l_P(\psi)$	-0.861	-0.841	0.308	0.914	0.841	0.786	0.063
		$l_{\widetilde{M}}(\psi)$	-0.080	-0.078	0.140	0.161	0.111	1.151	0.967
		$l_P^{\hat{c}}(\psi)$	-0.453	-0.432	0.301	0.544	0.432	0.803	0.552
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.326	0.328	0.136	0.353	0.328	1.183	0.462
	10	$l_P(\psi)$	-0.598	-0.589	0.181	0.625	0.589	0.840	0.024
		$l_{\widetilde{M}}(\psi)$	-0.147	-0.142	0.112	0.185	0.143	1.068	0.803
		$l_P^{\hat{c}}(\psi)$	-0.197	-0.189	0.175	0.264	0.193	0.870	0.748
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	0.253	0.256	0.108	0.275	0.256	1.111	0.424

Table 3.18: Inference on $\beta_2 = 2$ in the logistic regression for stratified data with MNAR covariate generated with missingness probability $\zeta_{it} = \text{logit}^{-1}(-0.5 + x_{it} + y_{it} + 2z_{it})$. Figures based on a simulation study with 2000 trials.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	17.943	1.490	254.958	255.525	1.543	0.069	0.887
		$l_{\widetilde{M}}(\psi)$	15.705	-0.233	255.094	255.513	0.534	0.003	0.879
		$l_P^{\hat{c}}(\psi)$	12.350	1.505	124.928	125.506	1.544	0.121	0.889
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	10.021	-0.221	125.052	125.422	0.528	0.006	0.883
	6	$l_P(\psi)$	1.315	0.927	2.087	2.466	0.943	0.430	0.785
		$l_{\widetilde{M}}(\psi)$	-0.240	-0.316	1.475	1.494	0.394	0.333	0.877
		$l_P^{\hat{c}}(\psi)$	1.306	0.924	1.962	2.356	0.937	0.476	0.786
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.253	-0.320	1.274	1.298	0.395	0.387	0.877
	10	$l_P(\psi)$	0.473	0.396	0.608	0.770	0.442	0.752	0.830
		$l_{\widetilde{M}}(\psi)$	-0.245	-0.265	0.346	0.424	0.317	1.006	0.855
		$l_P^{\hat{c}}(\psi)$	0.473	0.396	0.609	0.771	0.441	0.752	0.830
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.246	-0.267	0.346	0.425	0.317	1.006	0.855
100	4	$l_P(\psi)$	1.585	1.180	2.047	2.588	1.192	0.460	0.680
		$l_{\widetilde{M}}(\psi)$	-0.345	-0.451	1.550	1.588	0.463	0.277	0.815
		$l_P^{\hat{c}}(\psi)$	1.591	1.185	2.065	2.606	1.197	0.448	0.677
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.334	-0.450	1.612	1.646	0.462	0.267	0.816
	6	$l_P(\psi)$	0.839	0.748	0.752	1.127	0.753	0.705	0.666
		$l_{\widetilde{M}}(\psi)$	-0.440	-0.450	0.290	0.527	0.452	1.117	0.694
		$l_P^{\hat{c}}(\psi)$	0.838	0.749	0.752	1.125	0.752	0.705	0.666
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.443	-0.452	0.289	0.529	0.455	1.118	0.692
	10	$l_P(\psi)$	0.395	0.368	0.391	0.555	0.385	0.800	0.762
		$l_{\widetilde{M}}(\psi)$	-0.278	-0.285	0.232	0.362	0.293	1.041	0.758
		$l_P^{\hat{c}}(\psi)$	0.394	0.367	0.390	0.555	0.384	0.800	0.763
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.278	-0.285	0.232	0.362	0.293	1.042	0.758
250	4	$l_P(\psi)$	1.257	1.134	0.795	1.487	1.134	0.665	0.371
		$l_{\widetilde{M}}(\psi)$	-0.543	-0.544	0.212	0.583	0.544	1.222	0.426
		$l_P^{\hat{c}}(\psi)$	1.258	1.135	0.795	1.488	1.135	0.665	0.372
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.544	-0.545	0.212	0.583	0.545	1.223	0.425
	6	$l_P(\psi)$	0.777	0.741	0.445	0.895	0.741	0.742	0.392
		$l_{\widetilde{M}}(\psi)$	-0.380	-0.387	0.187	0.423	0.387	1.129	0.542
		$l_P^{\hat{c}}(\psi)$	0.777	0.741	0.445	0.895	0.741	0.742	0.393
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.379	-0.386	0.187	0.423	0.386	1.129	0.544
	10	$l_P(\psi)$	0.372	0.360	0.248	0.447	0.360	0.805	0.541
		$l_{\widetilde{M}}(\psi)$	-0.294	-0.295	0.149	0.329	0.295	1.037	0.514
		$l_P^{\hat{c}}(\psi)$	0.372	0.360	0.248	0.447	0.361	0.806	0.542
		$l_{\widetilde{M}^{\hat{c}}}(\psi)$	-0.294	-0.295	0.149	0.329	0.295	1.039	0.514

amount of incomplete data, the MPL assuming an MAR process might suffer from the greater dependence of ζ_{it} on the unobserved regressor: the coefficient γ_3 relating to z_{it} is larger than before in absolute value, so the impact of its omission from the missingness model is likely to be more adverse.

3.5 Survival model for censored data

3.5.1 Introduction

Time-to-event data subject to censoring are routinely collected in a wide variety of applied contexts: health science, engineering and biomedicine are just some of the examples. Subdivision in groups of such commonly named survival or failure times is very frequent for reasons related with stratified sampling, confounding factors or adjustments due to violation of model assumptions (Cortese and Sartori, 2016). Common clustering variables in these settings range from geographical areas and individuals to measuring methods and operating conditions. Nevertheless, as often occurs with grouped observations, the primary concern of the study is not the inter-cluster variability.

In survival analysis, the random effects approach outlined in Section 3.1 is put into practice by the renowned frailty models. Although parsimonious, these formulations are founded upon rather improbable presumptions and may lead to results which are sensitive to the supposed distribution of the involved group-specific random variables. However, when the amount of clusters in the sample is high relative to the within-group size, fixed effects specifications are as always hampered by the incidental parameters problem.

Under this special scenario, inferential solutions to the latter usual issue need also to deal with the presence of censored observations. In fact, application of the MPL has been experimented only to a limited extent because its computation is particularly far from straightforward in regression frameworks with general censoring scheme. The technique proposed by Pierce and Bellio (2006) to overcome such complications in fully parametric settings relies on Monte Carlo simulations like ours, but benefits from the complete definition of the censoring model. Later, Pierce and Bellio (2015) considered also higher-order asymptotics for semiparametric Cox regressions. In that case, the likelihood-based adjustment pertaining to effects of fitting nuisance parameters and equivalent to the MPL was obtained either by implementation of a parametric bootstrap employing a reference censoring model or by simulation. Instead, elimination of cluster-related parameters in parametric survival models for highly grouped censored data was achieved via Severini's frequentist integrated likelihood in the work of Cortese and Sartori (2016).

Importantly, the authors managed to prove the inferential superiority of their approach to random effects models with seriously misspecified frailty distribution.

This part of the dissertation is devoted to illustrating how to approximate the MPL through the expedient presented in Section 3.2 within the context of survival analysis. We introduce below the general setup, which may be viewed as an extension of the regression scenarios on which Cortese and Sartori (2016) focused on.

3.5.2 Notation and setup

Let independent clustered failure times $\widetilde{Y}_{it}$ follow a Weibull distribution with probability density function

$$p_{\widetilde{Y}_{it}}(\tilde{y}_{it};\psi;\lambda_i,x_{it},z_{it}) = \eta_{it}\xi\big(\eta_{it}\tilde{y}_{it}\big)^{\xi-1}\exp\big\{-\big(\eta_{it}\tilde{y}_{it}\big)^{\xi}\big\}, \qquad i=1,\dots,N,\ t=1,\dots,T, \tag{3.33}$$

for $\tilde{y}_{it} \geq 0$ and where $\eta_{it} = \exp\big\{-(\lambda_i + \beta_1 x_{it} + \beta_2 z_{it})\big\} > 0$ are the scale parameters. The interest is on estimating the common shape parameter $\xi > 0$ and the regression coefficients in $\beta = (\beta_1, \beta_2) \in \mathbb{R}^2$ while treating the vector of group-related intercepts $\lambda = (\lambda_1, \dots, \lambda_N) \in \mathbb{R}^N$ as a nuisance component. Therefore we shall have $\theta = (\psi, \lambda)$, with $\psi = (\xi, \beta) \in \mathbb{R}^+ \times \mathbb{R}^2$. Note that, as usual, the presence of whatever number of covariates in the study is not a paramount modeling aspect from the standpoint of the methodology aimed at deriving the MPL. On the contrary, application of the integrated likelihood in regressions similar to the present calls for extra computational effort, as borne out by Section 8.3 of Cortese and Sartori (2016).

Provided these premises, $\widetilde{Y}_{it}$ has survival function of the form $S_{\widetilde{Y}_{it}}(\tilde{y}_{it};\psi;\lambda_i,x_{it},z_{it}) = P_\theta(\widetilde{Y}_{it} > \tilde{y}_{it}) = \exp\big\{-\big(\eta_{it}\tilde{y}_{it}\big)^{\xi}\big\}$ and hazard function equal to

$$\begin{aligned} h_{\widetilde{Y}_{it}}(\tilde{y}_{it};\psi;\lambda_i,x_{it},z_{it}) = \frac{p_{\widetilde{Y}_{it}}(\tilde{y}_{it};\psi;\lambda_i,x_{it},z_{it})}{S_{\widetilde{Y}_{it}}(\tilde{y}_{it};\psi;\lambda_i,x_{it},z_{it})} &= h_0(\tilde{y}_{it};\xi)\eta_{it}^{\xi} \\ &= h_{0i}(\tilde{y}_{it};\xi,\lambda_i)e^{-\xi(\beta_1 x_{it} + \beta_2 z_{it})}, \end{aligned}$$

where $h_0(\tilde{y}_{it};\xi) = \xi\tilde{y}_{it}^{\xi-1}$ is the baseline hazard parametrically modeled and shared by all clusters, whereas $h_{0i}(\tilde{y}_{it};\xi,\lambda_i) = h_0(\tilde{y}_{it};\xi)e^{-\xi\lambda_i}$ can be seen as the equivalent for the ith group $(i = 1, \dots, N)$. Thus (3.33) has the advantage of being a proportional hazards model. Moreover, its logarithmic transformation coincides with a so-called accelerated failure time model, largely used in several scientific fields (for more details see Section 6 of Cortese and Sartori, 2016).

Since observations may be right-censored, data actually consist of realizations of the pair $\left(Y_{it}, \Delta_{it}\right)$, where $Y_{it} = \min\left(\widetilde{Y}_{it}, C_{it}\right)$ with C_{it} censoring time and Δ_{it} is the censoring indicator equal to 1 if $\widetilde{Y}_{it} \leq C_{it}$ and equal to 0 otherwise. The random censoring mechanism is only hypothesized to be independent and non-informative, meaning that each C_{it} is unrelated to the other survival or censoring times and its continuous distribution does not depend on θ. In particular, as opposed to what done in Section 4 of Cortese and Sartori (2016), we prefer to avoid the formal specification of a parametric density for C_{it}. On the one hand, such choice relaxes the assumptions of the analysis, but on the other, it prevents Severini's MPL from being exactly calculated. Nonetheless, in the next part our Monte Carlo approach will be shown flexible enough to tackle also this difficulty.

3.5.3 Monte Carlo modified profile likelihood

Consider the observed couple $\left(y_{it}, \delta_{it}\right)$ introduced above. If the censoring times c_{it} are independent realizations of a continuous random variable with generic density $p_{C_{it}}(c_{it}; \varsigma)$ and survival function $S_{C_{it}}(c_{it}; \varsigma) = P_\varsigma(C_{it} > c_{it})$, then those data are drawn from the joint density

$$p_{Y_{it},\Delta_{it}}(y_{it}, \delta_{it}; \theta, \varsigma) = \left\{p_{\widetilde{Y}_{it}}(y_{it}; \theta) S_{C_{it}}(y_{it}; \varsigma)\right\}^{\delta_{it}} \left\{p_{C_{it}}(y_{it}; \varsigma) S_{\widetilde{Y}_{it}}(y_{it}; \theta)\right\}^{1-\delta_{it}}, \qquad (3.34)$$

where, in the interests of conciseness, dependence on covariates is disregarded. Notwithstanding, since the distribution of C_{it} is independent of the parameter θ, the likelihood function based on the whole dataset $\left(y_{it}, \delta_{it}\right)$ $(i = 1, \ldots, N, t = 1, \ldots, T)$ can be formulated by

$$L(\theta) = \prod_{i=1}^{N} \prod_{t=1}^{T} \left\{p_{\widetilde{Y}_{it}}(y_{it}; \theta)\right\}^{\delta_{it}} \left\{S_{\widetilde{Y}_{it}}(y_{it}; \theta)\right\}^{1-\delta_{it}},$$

as pointed out in Example 1.2 of Pace and Salvan (1997). Denoting the number of failures recorded in the ith cluster by $\delta_{i\cdot} = \sum_{t=1}^{T} \delta_{it}$ $(i = 1, \ldots, N)$ and consequently their total number in the sample by $\delta_{\cdot\cdot} = \sum_{i=1}^{N} \delta_{i\cdot}$ allows to write the corresponding log-likelihood as

$$l(\theta) = \xi \sum_{i=1}^{N} \sum_{t=1}^{T} \delta_{it} \log \eta_{it} + \delta_{\cdot\cdot} \log \xi + (\xi - 1) \sum_{i=1}^{N} \sum_{t=1}^{T} \delta_{it} \log y_{it} - \sum_{i=1}^{N} \sum_{t=1}^{T} (\eta_{it} y_{it})^{\xi}, \quad (3.35)$$

where we remark that $\eta_{it} = \eta_{it}(\theta) = \exp\left\{-(\lambda_i + \beta_1 x_{it} + \beta_2 z_{it})\right\}$. Now, differentiation with respect to the ith nuisance component gives the connected element of the partial score

$$l_{\lambda_i}(\theta) = l_{\lambda_i}(\psi, \lambda_i) = -\xi\delta_{i\cdot} + \xi\sum_{t=1}^{T}(\eta_{it}y_{it})^{\xi}, \qquad i = 1, \dots, N. \tag{3.36}$$

By equating (3.36) to 0 and solving analytically for λ_i, one may find the group-specific constrained ML estimate

$$\hat{\lambda}_{i\psi} = \frac{1}{\xi}\left\{ \log\sum_{t=1}^{T} y_{it}^{\xi} e^{-\xi(\beta_1 x_{it} + \beta_2 z_{it})} - \log\delta_{i\cdot} \right\}, \qquad i = 1, \dots, N, \tag{3.37}$$

which in its turn delivers $\hat{\lambda}_{\psi} = (\hat{\lambda}_{1\psi}, \dots, \hat{\lambda}_{N\psi})$ and $\hat{\theta}_{\psi} = (\psi, \hat{\lambda}_{\psi})$. If in (3.35) we substitute each incidental parameter with the last expression, we get the profile log-likelihood function for ψ

$$\begin{aligned} l_P(\psi) = & \sum_{i=1}^{N} \delta_{i\cdot} \left\{ \log\delta_{i\cdot} - \log\sum_{t=1}^{T} y_{it}^{\xi} e^{-\xi(\beta_1 x_{it} + \beta_2 z_{it})} \right\} - \xi\sum_{i=1}^{N}\sum_{t=1}^{T}\delta_{it}(\beta_1 x_{it} + \beta_2 z_{it}) \\ & + \delta_{\cdot\cdot}(\log\xi - 1) + (\xi - 1)\sum_{i=1}^{N}\sum_{t=1}^{T}\delta_{it}\log y_{it}, \end{aligned} \tag{3.38}$$

reaching its maximum at $\hat{\psi} = (\hat{\xi}, \hat{\beta})$. Once the latter is obtained numerically, the full ML estimate of the model clearly is $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$, with $\hat{\lambda} = \hat{\lambda}_{\hat{\psi}}$.

The first quantity to be computed in the ith summand of Severini's modification term follows immediately from the derivative of (3.36) with regard to λ_i. Specifically, after a change of sign, it is possible to express it as

$$j_{\lambda_i,\lambda_i}(\hat{\theta}_{\psi}) = \xi^2\sum_{t=1}^{T}(\tilde{\eta}_{it}y_{it})^{\xi}, \qquad i = 1, \dots, N,$$

where $\tilde{\eta}_{it} = \eta_{it}(\hat{\theta}_{\psi}) = \exp\left\{-(\hat{\lambda}_{i\psi} + \beta_1 x_{it} + \beta_2 z_{it})\right\}$. With right-censored data, explicit calculation of the expected value $I_{\lambda_i\lambda_i}(\hat{\theta}_{\psi}; \hat{\theta})$ should be carried out with reference to the joint probability density function (3.34), comprising also the distribution of the censoring times. Yet, as claimed at the end of Section 3.5.1, we are not willing to constrain $p_{C_{it}}(c_{it}; \varsigma)$ and $S_{C_{it}}(c_{it}; \varsigma)$ to take one specific parametric form. Fortunately, such a restriction is not required to calculate the MPL through the Monte Carlo strategy reported in (3.3), because estimation of such functions can be implemented nonparametrically. Turning to the technicalities of the procedure prescribed in the current situation, the empirical mean (3.3) is given by

$$I^{*}_{\lambda_i\lambda_i}(\hat{\theta}_{\psi}; \hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\left[\left\{ -\xi\delta^{r}_{i\cdot} + \xi\sum_{t=1}^{T}(\tilde{\eta}_{it}y^{r}_{it})^{\xi} \right\}\left\{ -\hat{\xi}\delta^{r}_{i\cdot} + \hat{\xi}\sum_{t=1}^{T}(\hat{\eta}_{it}y^{r}_{it})^{\hat{\xi}} \right\}\right], \tag{3.39}$$

where $\hat{\eta}_{it} = \eta_{it}(\hat{\theta}) = \exp\left\{-(\hat{\lambda}_i + \hat{\beta}_1 x_{it} + \hat{\beta}_2 z_{it})\right\}$, $\delta^r_{i\cdot} = \sum_{t=1}^{T} \delta^r_{it}$ and $(y^r_{it}, \delta^r_{it})$ are the data in the rth Monte Carlo sample $(r = 1, \ldots, R)$ simulated as explained in the sequel. Firstly, failures $\tilde{y}^r_{it}$ are generated from the ML fit of model (3.33). Secondly, new censoring times c^r_{it} are determined by performing the conditional bootstrap described in Algorithm 3.1 of Davison and Hinkley (1997, p. 85). In particular, if the original indicator δ_{it} equals 0 we set $c^r_{it} = c_{it}$, otherwise we draw c^r_{it} from the conditional distribution of $C_{it}|C_{it} > y_{it}$ computed as

$$\widehat{S}_{C_{it}|C_{it}>y_{it}}(c_{it}) = \frac{\widehat{S}_{C_{it}}(c_{it})}{\widehat{S}_{C_{it}}(y_{it})},$$

where $\widehat{S}_{C_{it}}$ is the Kaplan-Meier nonparametric estimator of the survival function of C_{it} (Kaplan and Meier, 1958). Precisely, each c^r_{it} corresponding to $\delta_{it} = 1$ is found as the unique solution c to the equation $\widehat{S}_{C_{it}}(c) = u^r_{it}\widehat{S}_{C_{it}}(y_{it})$, with $u^r_{it} \sim U(0, 1)$. Eventually, the censored survival times are $y^r_{it} = \min(\tilde{y}^r_{it}, c^r_{it})$ and hence the new failure indicators δ^r_{it} are defined accordingly $(i = 1, \ldots, N,\ t = 1, \ldots, T)$.

Complying with the practice adopted during all this chapter, in what follows some simulation results will shed light on the possibility to solve the Neyman & Scott problems using the MPL in the Weibull regression model for clustered time-to-event data with unspecified random censoring scheme. The studies will especially examine on a comparative basis the profile log-likelihood $l_P(\psi)$ in (3.38) and its Monte Carlo adjustment $l_{\widetilde{M}^*}(\psi)$ derived by the approximation (3.39).

3.5.4 Simulation studies

Two experiments of $S = 2000$ simulations are conducted to study inference on ψ in the survival model for censored observations presented in Section 3.5.1. Focusing on the two-index asymptotic setting at issue, the within-group size and the number of clusters in the artificial datasets are $T = 4, 6, 10$ and $N = 50, 100, 250$, respectively. The first binary covariate x_{it} in each ith group $(i = 1, \ldots, N)$ is obtained by imposing $x_{it} = 0$ for $t = 1, \ldots, T/2$ and $x_{it} = 1$ for $t = T/2 + 1, \ldots, T$. The second regressor z_{it}, differently, is drawn from the standard normal distribution. We set the common shape parameter ξ equal to 1.5 and $\beta = (-1, 1)$, while each cluster-related intercept is independently sampled as $\lambda_i \sim N(0.5, 0.5^2)$. Failures $\tilde{y}_{it}$ are then simulated via the Weibull density function (3.33). The censoring times c_{it} can be obtained by random generation from the distribution $Exp(\varsigma)$, where the parameter is chosen in such a way as to control the overall proportion P_c of censored data. In detail, given the quantities above and for a certain P_c, ς is fixed to the value solving the equation

$$\frac{1}{TN}\sum_{i=1}^{N}\sum_{t=1}^{T} P_{\varrho}(\widetilde{Y}_{it} > C_{it}) = \frac{1}{TN}\sum_{i=1}^{N}\sum_{t=1}^{T}\int_0^{+\infty} S_{\widetilde{Y}_{it}}(y;\psi;\lambda_i,x_{it},z_{it})p_{C_{it}}(y;\varsigma)dy = P_c,$$

where $\varrho = (\theta,\varsigma)$ and $p_{C_{it}}(y;\varsigma) = \varsigma e^{-\varsigma y}$. Then, in each one of the S fictitious samples, observations $\big(y_{it},\delta_{it}\big)$ stem from the usual definitions of censored failures and censoring indicators, i.e. $y_{it} = \min(\tilde{y}_{it}, c_{it})$ and $\delta_{it} = 1$ when $\tilde{y}_{it} \le c_{it}$, otherwise $\delta_{it} = 0$ ($i = 1,\dots,N$, $t = 1,\dots,T$).

The first series of simulations considers data with average censoring probability $P_c = 0.2$, the second relates instead to situations with higher proportion of censored observations, namely $P_c = 0.4$. Inferences from the profile likelihood and from the Monte

Table 3.19: Inference on $\xi = 1.5$ in the Weibull regression model for grouped survival data with unspecified censoring scheme and probability of censoring $P_c = 0.2$. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	0.392	0.385	0.145	0.418	0.385	0.858	0.111
		$l_{\widetilde{M}^*}(\psi)$	0.010	0.004	0.112	0.112	0.074	0.979	0.956
	6	$l_P(\psi)$	0.231	0.228	0.102	0.252	0.228	0.884	0.291
		$l_{\widetilde{M}^*}(\psi)$	0.008	0.005	0.087	0.088	0.056	0.964	0.943
	10	$l_P(\psi)$	0.124	0.123	0.066	0.141	0.123	0.976	0.517
		$l_{\widetilde{M}^*}(\psi)$	0.005	0.004	0.060	0.061	0.041	1.029	0.961
100	4	$l_P(\psi)$	0.371	0.369	0.103	0.385	0.369	0.840	0.015
		$l_{\widetilde{M}^*}(\psi)$	-0.006	-0.006	0.079	0.079	0.053	0.966	0.936
	6	$l_P(\psi)$	0.219	0.216	0.070	0.230	0.216	0.903	0.063
		$l_{\widetilde{M}^*}(\psi)$	-0.000	-0.003	0.060	0.060	0.041	0.987	0.947
	10	$l_P(\psi)$	0.119	0.117	0.048	0.128	0.117	0.938	0.259
		$l_{\widetilde{M}^*}(\psi)$	0.001	-0.001	0.044	0.044	0.030	0.989	0.943
250	4	$l_P(\psi)$	0.366	0.366	0.065	0.372	0.366	0.847	0.000
		$l_{\widetilde{M}^*}(\psi)$	-0.009	-0.011	0.050	0.050	0.034	0.972	0.939
	6	$l_P(\psi)$	0.214	0.213	0.045	0.218	0.213	0.890	0.000
		$l_{\widetilde{M}^*}(\psi)$	-0.005	-0.005	0.038	0.039	0.026	0.972	0.934
	10	$l_P(\psi)$	0.116	0.116	0.030	0.120	0.116	0.943	0.018
		$l_{\widetilde{M}^*}(\psi)$	-0.002	-0.002	0.028	0.028	0.019	0.993	0.949

Table 3.20: Inference on $\beta_1 = -1$ in the Weibull regression model for grouped survival data with unspecified censoring scheme and probability of censoring $P_c = 0.2$. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	-0.001	0.001	0.122	0.122	0.082	0.825	0.898
		$l_{\widetilde{M}^*}(\psi)$	0.006	0.007	0.120	0.120	0.082	0.973	0.944
	6	$l_P(\psi)$	-0.003	-0.003	0.097	0.097	0.066	0.867	0.911
		$l_{\widetilde{M}^*}(\psi)$	0.002	0.001	0.096	0.096	0.065	0.969	0.940
	10	$l_P(\psi)$	-0.000	-0.002	0.070	0.070	0.047	0.943	0.933
		$l_{\widetilde{M}^*}(\psi)$	0.002	0.001	0.070	0.070	0.048	1.008	0.950
100	4	$l_P(\psi)$	-0.007	-0.006	0.089	0.089	0.062	0.804	0.885
		$l_{\widetilde{M}^*}(\psi)$	0.001	0.001	0.088	0.088	0.061	0.950	0.940
	6	$l_P(\psi)$	-0.004	-0.004	0.069	0.069	0.044	0.864	0.901
		$l_{\widetilde{M}^*}(\psi)$	0.001	-0.000	0.069	0.069	0.044	0.964	0.935
	10	$l_P(\psi)$	-0.003	-0.002	0.050	0.050	0.034	0.932	0.928
		$l_{\widetilde{M}^*}(\psi)$	0.000	0.001	0.050	0.050	0.034	0.996	0.954
250	4	$l_P(\psi)$	-0.007	-0.007	0.056	0.056	0.037	0.814	0.894
		$l_{\widetilde{M}^*}(\psi)$	0.002	0.002	0.055	0.055	0.037	0.961	0.946
	6	$l_P(\psi)$	-0.003	-0.003	0.042	0.042	0.027	0.893	0.920
		$l_{\widetilde{M}^*}(\psi)$	0.002	0.003	0.042	0.042	0.027	0.997	0.954
	10	$l_P(\psi)$	-0.003	-0.002	0.032	0.032	0.022	0.936	0.926
		$l_{\widetilde{M}^*}(\psi)$	0.000	0.000	0.032	0.032	0.022	1.002	0.945

Carlo MPL on the structural component ψ are investigated as already done in the previous examples. Notice that, before proceeding to maximize the two functions for every simulated dataset, non-informative clusters with only censored failure times need to be discarded from the study. Indeed, (3.37) shows that $\hat{\lambda}_{i\psi}$ is not finite if $\delta_{i\cdot} = 0$ and hence the ith group does not make any contribution to estimating ψ $(i = 1, \ldots, N)$. Numerical optimization of both $l_P(\psi)$ and $l_{\widetilde{M}^*}(\psi)$ is implemented by the `R` function `optim`. Specifically, in the former case we choose the method `L-BFGS-B` (Byrd *et al.*, 1995) which enables to find the solution $\hat{\psi}$ in a bounded set, while in the latter we search for $\hat{\psi}_{\widetilde{M}^*} = \left(\hat{\xi}_{\widetilde{M}^*}, \hat{\beta}_{\widetilde{M}^*}\right)$ by means of the Nelder-Mead algorithm, with no constraints imposed on the parameters but initial value set to the ML estimate.

Results of the first experiment may be seen in Tables 3.19, 3.20 and 3.21 by reference to ξ, β_1 and β_2, respectively. The accuracy of $l_{\widetilde{M}^*}(\psi)$ is extremely good for all

TABLE 3.21: Inference on $\beta_2 = 1$ in the Weibull regression model for grouped survival data with unspecified censoring scheme and probability of censoring $P_c = 0.2$. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	0.006	0.005	0.074	0.074	0.051	0.835	0.905
		$l_{\widetilde{M}^*}(\psi)$	-0.001	-0.002	0.073	0.073	0.050	0.985	0.946
	6	$l_P(\psi)$	0.003	0.003	0.056	0.056	0.037	0.863	0.911
		$l_{\widetilde{M}^*}(\psi)$	-0.002	-0.002	0.055	0.055	0.036	0.963	0.936
	10	$l_P(\psi)$	0.002	0.002	0.041	0.041	0.027	0.931	0.933
		$l_{\widetilde{M}^*}(\psi)$	0.000	0.000	0.041	0.041	0.027	0.997	0.947
100	4	$l_P(\psi)$	0.008	0.008	0.053	0.054	0.037	0.828	0.890
		$l_{\widetilde{M}^*}(\psi)$	0.000	0.000	0.052	0.052	0.037	0.980	0.946
	6	$l_P(\psi)$	0.005	0.004	0.040	0.040	0.027	0.868	0.906
		$l_{\widetilde{M}^*}(\psi)$	-0.001	-0.001	0.040	0.040	0.026	0.971	0.944
	10	$l_P(\psi)$	0.001	0.001	0.028	0.028	0.018	0.937	0.936
		$l_{\widetilde{M}^*}(\psi)$	-0.001	-0.001	0.028	0.028	0.018	1.000	0.952
250	4	$l_P(\psi)$	0.006	0.006	0.033	0.033	0.022	0.834	0.890
		$l_{\widetilde{M}^*}(\psi)$	-0.002	-0.002	0.032	0.032	0.021	0.984	0.946
	6	$l_P(\psi)$	0.004	0.005	0.025	0.025	0.017	0.887	0.917
		$l_{\widetilde{M}^*}(\psi)$	-0.000	0.000	0.025	0.025	0.016	0.990	0.944
	10	$l_P(\psi)$	0.002	0.002	0.018	0.018	0.012	0.949	0.936
		$l_{\widetilde{M}^*}(\psi)$	-0.001	-0.001	0.017	0.018	0.012	1.017	0.952

unknown quantities and diverse dimensions of the data. The presence of many nuisance parameters does not seem to be of great importance to the estimation of the regression coefficients, yet inferential conclusions on ξ drawn via $l_P(\psi)$ are found quite misguided. In particular, Table 3.19 testifies how the Monte Carlo modification is capable not only of greatly reducing the severe empirical bias of the ML estimator but also of correcting the excessively low actual Wald coverages derived by the profile likelihood. In fact, these can also be ascribed to the supplied standard errors of $\hat{\xi}$, prominently downward biased for smaller T, independently of N. Estimated variability of $\hat{\xi}_{\widetilde{M}^*}$ is, conversely, much more trustworthy. Although Tables 3.20 and 3.21 confirm the sufficient adequacy of the profile likelihood to make inference on β, due to better estimation of the standard errors of $\hat{\beta}_{\widetilde{M}^*} = (\hat{\beta}_{1\widetilde{M}^*}, \hat{\beta}_{2\widetilde{M}^*})$ the Monte Carlo MPL is still undoubtedly superior in terms of appropriateness of confidence intervals' coverage for both coefficients.

Performances of the two inferential tools under examination in the second simulation study are summarized by Tables 3.22, 3.23 and 3.24. For what concerns the shape parameter, Table 3.22 proves the convenience of $l_{\widetilde{M}^*}(\psi)$ even in the occasion of more observations subject to censoring. Indeed, also when $P_c = 0.4$ the empirical bias of $\hat{\xi}_{\widetilde{M}^*}$ is systematically lower than that of $\hat{\xi}$, reaching negligible values when N and T increase. In contrast, the imprecise point estimation provided by $l_P(\psi)$ is especially critical when the within-group size is smaller and remains basically constant as N grows, coherently with the existing theoretical knowledge (Sartori, 2003). Furthermore, the empirical coverage probabilities based on the Monte Carlo MPL are all very close to the nominal level, while those based on the profile likelihood are well below it, even for the aforementioned unreliable estimated standard errors of $\hat{\xi}$. Statistical indicators displayed in Tables 3.23 and 3.24 about inference on β let us conclude once again that when Neyman & Scott problems arise the Monte Carlo adjustment is still valuable to further improve the quality of standard ML procedures under the regression scenario.

TABLE 3.22: Inference on $\xi = 1.5$ in the Weibull regression model for grouped survival data with unspecified censoring scheme and probability of censoring $P_c = 0.4$. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	0.462	0.449	0.179	0.495	0.449	0.834	0.122
		$l_{\widetilde{M}^*}(\psi)$	-0.001	-0.010	0.125	0.125	0.084	0.994	0.950
	6	$l_P(\psi)$	0.266	0.258	0.120	0.292	0.258	0.880	0.312
		$l_{\widetilde{M}^*}(\psi)$	-0.002	-0.008	0.097	0.097	0.069	0.985	0.949
	10	$l_P(\psi)$	0.143	0.141	0.082	0.165	0.141	0.910	0.524
		$l_{\widetilde{M}^*}(\psi)$	0.001	-0.001	0.073	0.073	0.049	0.974	0.946
100	4	$l_P(\psi)$	0.445	0.436	0.126	0.462	0.436	0.827	0.009
		$l_{\widetilde{M}^*}(\psi)$	-0.013	-0.017	0.089	0.090	0.060	0.982	0.940
	6	$l_P(\psi)$	0.254	0.251	0.084	0.268	0.251	0.889	0.067
		$l_{\widetilde{M}^*}(\psi)$	-0.009	-0.012	0.068	0.068	0.047	0.993	0.940
	10	$l_P(\psi)$	0.142	0.142	0.057	0.153	0.142	0.915	0.234
		$l_{\widetilde{M}^*}(\psi)$	0.000	0.000	0.051	0.051	0.033	0.981	0.947
250	4	$l_P(\psi)$	0.430	0.427	0.080	0.437	0.427	0.815	0.000
		$l_{\widetilde{M}^*}(\psi)$	-0.021	-0.022	0.056	0.060	0.041	0.969	0.915
	6	$l_P(\psi)$	0.248	0.247	0.053	0.254	0.247	0.888	0.002
		$l_{\widetilde{M}^*}(\psi)$	-0.014	-0.015	0.043	0.045	0.031	0.993	0.938
	10	$l_P(\psi)$	0.135	0.134	0.035	0.139	0.134	0.941	0.018
		$l_{\widetilde{M}^*}(\psi)$	-0.006	-0.006	0.031	0.032	0.022	1.004	0.945

TABLE 3.23: Inference on $\beta_1 = -1$ in the Weibull regression model for grouped survival data with unspecified censoring scheme and probability of censoring $P_c = 0.4$. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	-0.018	-0.018	0.152	0.153	0.104	0.787	0.880
		$l_{\widetilde{M}^*}(\psi)$	-0.001	-0.002	0.149	0.149	0.100	0.970	0.948
	6	$l_P(\psi)$	-0.007	-0.008	0.116	0.116	0.080	0.856	0.905
		$l_{\widetilde{M}^*}(\psi)$	0.003	0.003	0.114	0.114	0.079	0.984	0.946
	10	$l_P(\psi)$	-0.005	-0.006	0.083	0.083	0.057	0.943	0.941
		$l_{\widetilde{M}^*}(\psi)$	0.000	-0.000	0.082	0.082	0.056	1.026	0.955
100	4	$l_P(\psi)$	-0.016	-0.014	0.106	0.107	0.074	0.798	0.882
		$l_{\widetilde{M}^*}(\psi)$	-0.000	0.001	0.104	0.104	0.071	0.983	0.949
	6	$l_P(\psi)$	-0.012	-0.010	0.081	0.081	0.056	0.875	0.912
		$l_{\widetilde{M}^*}(\psi)$	-0.001	-0.001	0.079	0.079	0.053	1.006	0.947
	10	$l_P(\psi)$	-0.007	-0.006	0.059	0.059	0.039	0.939	0.926
		$l_{\widetilde{M}^*}(\psi)$	-0.002	-0.001	0.058	0.058	0.039	1.017	0.947
250	4	$l_P(\psi)$	-0.014	-0.012	0.067	0.069	0.044	0.796	0.868
		$l_{\widetilde{M}^*}(\psi)$	0.004	0.006	0.065	0.065	0.043	0.988	0.945
	6	$l_P(\psi)$	-0.008	-0.008	0.052	0.053	0.036	0.854	0.901
		$l_{\widetilde{M}^*}(\psi)$	0.003	0.003	0.051	0.051	0.035	0.985	0.947
	10	$l_P(\psi)$	-0.004	-0.004	0.038	0.038	0.025	0.931	0.932
		$l_{\widetilde{M}^*}(\psi)$	0.001	0.001	0.037	0.037	0.025	1.010	0.953

Ultimately, a thorough comparison between the outcomes of the two experiments reviewed in this section may be helpful to check whether and how the incidence of censored data in the sample affects the accuracy of the statistical techniques employed for inferences on the parameter of interest in the Weibull survival model. In general, to recognize a clear performance pattern by looking at the various tables is not immediate. Perhaps, with special regard to the quantity whose estimation is the most harmed by the presence of incidental parameters, $l_P(\psi)$ appears to suffer more than $l_{\widetilde{M}^*}(\psi)$ from a high average censoring probability. Indeed, in making inference on ξ via the profile likelihood, only the empirical coverages when $N = 50$ are slightly more adequate with $P_c = 0.4$. To the contrary, conclusions descending from the MPL look less impacted by the percentage of observations subject to censoring.

To end the discussion, it seems worthwhile stressing that such empirical findings are substantially in accordance with those relating to the contrast between the profile likelihood and the integrated likelihood in Cortese and Sartori (2016). Nonetheless, there exist three main motivations to prefer the Monte Carlo MPL approach illustrated here. First, it is far less computationally expensive, as the effort implied by the numerical integration to calculate Severini's integrated likelihood in the regression setting is considerable. Second, its basic procedure easily lends itself to encompass the bootstrap for nonparametric estimation of the censoring mechanism, permitting to protect against misspecification risks. And third, it may be readily generalized to cope with a different distribution of the failure times $\widetilde{Y}_{it}$, such as logNormal or Gamma, whereas the method of Cortese and Sartori (2016) demands to derive ad hoc formulae for finding a suitable reparametrization of the model (Severini, 2007).

TABLE 3.24: Inference on $\beta_2 = 1$ in the Weibull regression model for grouped survival data with unspecified censoring scheme and probability of censoring $P_c = 0.4$. Figures based on a simulation study with 2000 trials and $R = 500$ Monte Carlo replicates to compute $l_{\widetilde{M}^*}(\psi)$.

N	T	Method	B	MB	SD	RMSE	MAE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	0.020	0.018	0.104	0.106	0.071	0.776	0.874
		$l_{\widetilde{M}^*}(\psi)$	0.003	0.001	0.101	0.101	0.067	0.964	0.941
	6	$l_P(\psi)$	0.015	0.013	0.068	0.070	0.047	0.870	0.909
		$l_{\widetilde{M}^*}(\psi)$	0.005	0.003	0.067	0.067	0.046	1.003	0.954
	10	$l_P(\psi)$	0.006	0.005	0.051	0.052	0.036	0.920	0.929
		$l_{\widetilde{M}^*}(\psi)$	0.001	0.001	0.051	0.051	0.035	0.997	0.953
100	4	$l_P(\psi)$	0.017	0.017	0.070	0.072	0.048	0.787	0.867
		$l_{\widetilde{M}^*}(\psi)$	0.001	0.001	0.068	0.068	0.046	0.974	0.943
	6	$l_P(\psi)$	0.009	0.009	0.049	0.050	0.033	0.871	0.909
		$l_{\widetilde{M}^*}(\psi)$	-0.000	-0.001	0.048	0.048	0.032	1.002	0.946
	10	$l_P(\psi)$	0.005	0.004	0.034	0.034	0.022	0.920	0.926
		$l_{\widetilde{M}^*}(\psi)$	-0.000	-0.001	0.034	0.034	0.022	0.995	0.941
250	4	$l_P(\psi)$	0.014	0.014	0.041	0.044	0.030	0.822	0.869
		$l_{\widetilde{M}^*}(\psi)$	-0.001	-0.002	0.040	0.040	0.028	1.017	0.958
	6	$l_P(\psi)$	0.009	0.010	0.031	0.033	0.022	0.858	0.895
		$l_{\widetilde{M}^*}(\psi)$	-0.001	-0.000	0.031	0.031	0.020	0.986	0.943
	10	$l_P(\psi)$	0.006	0.006	0.022	0.022	0.015	0.942	0.931
		$l_{\widetilde{M}^*}(\psi)$	0.000	-0.000	0.021	0.021	0.015	1.021	0.954

3.6 Discussion and further work

The prime objective of the present chapter has been to show how to exploit Monte Carlo simulation for widening the field of application of the MPL (Barndorff-Nielsen, 1983). Severini (1998b) made a first valuable step in this direction, yet his approximation is still not approachable enough to deal with the today's degree of modeling sophistication. A new solution was then needed to fill such a gap in accessibility.

Clustered data are often collected from studies designed with extreme care, like clinical trials. As a consequence, statistical models for grouped observations not only are particularly subject to Neyman & Scott problems for the reasons discussed in Section 3.1, but also are likely to incorporate complex assumptions due to the experimental design. Section 3.2 served the purpose to introduce the Monte Carlo strategy for computing the MPL in those nonstandard situations. The procedure essentially foresees to approximate the expected value implicated in Severini's modification by means of an empirical mean. Such approach is easy, implementable in broad generality and reasonably fast.

In Section 3.3 the suggested methodology was applied to the nonstationary autoregressive model for panel data with fixed effects. Under these hypotheses, analytical calculation of Severini's MPL is practicable but is not a simple task. By contrast, we saw that Monte Carlo simulation may be used in a straightforward manner to estimate the required expectation. Results of simulations reported in Section 3.3.4 empirically confirm the $(T \times N)$-asymptotic properties of the profile and modified profile likelihoods derived by Sartori (2003) for the case of independent observations (see Section 1.4.3). In fact, inferential improvements determined by the adjustment when the group size T is much smaller than the number of clusters N seem to be remarkable even allowing for dependence in the data. Moreover, the findings concerning the unconventional form of the MPL function in this setting are consistent with previous works (Lancaster, 2002; Dhaene and Jochmans, 2014; De Bin *et al.*, 2015).

Issues in inferences on the structural parameter related to the presence of missing values in binary grouped data were addressed by Section 3.4. Specifically, in Section 3.4.3 we considered univariate arbitrary patterns of missingness in the response. In this case, since the density of the dependent variable and the mechanism of missingness is not jointly specified but factorized in two parts, the usual expected value in the function by Severini is not exactly computable. On the opposite its approximation, stemming from a two-step procedure to simulate the Monte Carlo samples, permits to correctly take the missing-data generation process into account with no need to assume a common distribution.

Results of two simulation studies were initially presented for the logistic regression scenario. From the first analysis of MCAR observations, Monte Carlo simulation was found unnecessary to compute the MPL. In particular, the greater inferential precision of the MCAR Monte Carlo MPL relative to the profile likelihood appears equivalent to that of the analytical MPL by Severini which disregards the missing data. Indeed the MCAR hypothesis implies that complete units are a random sample from the original population, hence the expectation taken on the unconditional sampling distribution equals the conditional one based only on the observed data (Kenward and Molenberghs, 1998). Furthermore it turned out that, curiously, the MNAR variant of the MPL is inappropriate to reliably estimate the parameter of interest, although the model of nonrandom missingness supposed is just a generalization of the true one. Additional investigations regarding larger within-group sizes may contribute to clarify this aspect.

The second experiment under the logit framework examined instead the situation of nonignorable missing data. In this case, the MNAR Monte Carlo MPL accounting for the missing values proved to be more accurate than Severini's function provided that T was not too small, for any value of N considered. Justifications for this outcome were already given in Section 3.4.3, nevertheless it might be worth understanding why the quality of fitting for the missingness model seems to depend on the amount of information in the specific group and not on the global sample size.

Analogue simulation studies in the binary regression setting with probit link and possibly missing response showed a quite different inferential behaviour of the MNAR Monte Carlo MPL. Firstly, when used to analyze MCAR data, the latter set an example of robustness against nonignorable incompleteness. Secondly, with a correct specification of the missingness mechanism, its application was more recommendable than that of Severini's function even for the lowest value of T taken into consideration. Given the aforesaid ability of the MNAR MPL to identify the true ignorable process of missingness in datasets with smaller groups, it is likely that the same performance pattern under the MNAR scenario is retained when $T < 10$, as opposed to what observed within the logistic setup.

At the time being, such discrepancies following the change of link function from logit to probit have reasons not apparent to the writer. One rather vague presumption is that the two logistic models in the former framework, the one for the dependent variable and the one for the missingness indicator, come somehow into conflict with each other, causing convergence difficulties perhaps related to identifiability issues during the global fit through the MNAR Monte Carlo MPL. Future studies, preferably involving probit specifications for the missing-data mechanism, might be helpful to elaborate on this matter.

Section 3.4.4 coped with inference in the event of MAR regressors in fixed-effect logistic regressions for clustered observations. Differently from before, in this case the Monte Carlo expedient was not used to compute the MPL, because one analytical version which is able to account for the incomplete covariate information was derived as an approximation to the modified conditional likelihood of Lipsitz *et al.* (1998). The main indication resulting from the simulation experiments is that this approach seems more suitable than the classical MPL to estimate the parameters of interest when the percentage of missing predictors does not exceed 35%. When such percentage grows, the solution is advisable only for larger groups. Note that these observations apply both to the circumstance of correctly specified MAR mechanism and to that of underpecified true MNAR mechanism. Under this last scenario, however, the inferential accuracy might be refined by considering the nonignorability of the missingness process, as done in the conditional likelihood proposed by Cho Paik (2004).

Clustered survival times subject to right-censoring were discussed by Section 3.5. In the context of a Weibull regression model with group-related intercepts, our proposed approximation to Severini's MPL was made necessary by the lack of distributional assumptions on the random censoring mechanism. Indeed, an explicit calculation of the modification term requires full parametric specification of the density for the censoring times, whereas the Monte Carlo strategy allows to estimate it nonparametrically, using a conditional bootstrap (Davison and Hinkley, 1997, Algorithm 3.1). Experimental outcomes examined in Section 3.5.4 substantially corroborated for this other framework the theory pertaining to inference in the standard two-index asymptotic setting, described at the end of Chapter 1. Estimation of the parameter of interest via the Monte Carlo MPL is notably preferable to that via the profile likelihood in every relevant respect, even though inferences on regression coefficients were found less affected by Neyman & Scott problems. In addition, the proportion of censored data in the sample does not appear to have a significant effect on the ensuing precision of the MPL. Note finally that the computational burden demanded by existing alternative statistical procedures (Cortese and Sartori, 2016) is much heavier than that of the one adopted here.

In this area, the potential room for future developments is extremely vast. The generality inherent in the suggested method enables in fact to take advantage of the MPL's properties in numerous models suffering from the incidental parameters problem. Furthermore, several aspects of our study, emerged here or earlier in this chapter, deserve to be further investigated. Among others, some open topics we plan to tackle in the forthcoming work are listed below:

i) Explore the usefulness of the Monte Carlo strategy in the presence of MAR binary response, considering also possible misspecifications of the missingness generation process.

ii) Apply the Monte Carlo strategy to clustered data with continuous incomplete response.

iii) Derive a MPL function for handling MNAR covariates in logistic regressions by approximating the conditional likelihood of Cho Paik (2004).

iv) Analyze real clustered data from a clinical trial involving HIV-infected patients (Carlin and Hodges, 1999;Cohn *et al.*, 1999) adopting a version of the Monte Carlo MPL equivalent to that described in Section 3.5; compare such results with those obtained by means of the integrated likelihood in Cortese and Sartori (2016).

v) Extend the application of the Monte Carlo strategy to semiparametric regression models where the incidental nuisance parameters are expressed as unknown real-valued functions, like those treated by He and Severini (2014) via the integrated likelihood.

Appendix

Recalling that, when $k = 1$, under model (2.1) the Wald z-statistic is $\widehat{T} = (\hat{\theta} - \theta_0)\hat{\nu}_{1,1}^{1/2}$, the analogue of function (2.27) is simply

$$T = T(\theta; \theta_0) = (\theta - \theta_0)\nu_{1,1}^{1/2}. \tag{A.1}$$

Therefore, by the general definition (2.29), one shall derive the first term in the asymptotic bias expansion of $\widehat{T}$ as

$$B_T(\theta; \theta_0) = B(\theta)T'(\theta; \theta_0) + \frac{1}{2}\nu_{1,1}^{-1}T''(\theta; \theta_0), \tag{A.2}$$

where $T'(\theta; \theta_0)$ and $T''(\theta; \theta_0)$ are the first and the second derivative, respectively, of (A.1) with respect to the scalar argument θ. In particular, they take the form

$$\begin{aligned} T'(\theta; \theta_0) &= \nu_{1,1}^{1/2} - \frac{\theta - \theta_0}{2}\frac{\nu'_{1,1}}{\nu_{1,1}^{1/2}}, \\ T''(\theta; \theta_0) &= \frac{\nu'_{1,1}}{\nu_{1,1}^{1/2}} + \frac{\theta - \theta_0}{2}\left(\frac{\nu''_{1,1}}{\nu_{1,1}^{1/2}} - \frac{\nu'^{2}_{1,1}}{2\nu_{1,1}^{3/2}}\right). \end{aligned}$$

By exploiting Bartlett's identities and adopting the power notation, it is not hard to check that double differentiation of the expected information leads to

$$\begin{aligned} \nu'_{1,1} &= -(\nu_3 + \nu_{1,2}), \\ \nu''_{1,1} &= -(\nu_4 + 2\nu_{1,3} + \nu_{2,2} + \nu_{1,1,2}), \end{aligned}$$

and so we have

$$\begin{aligned} T'(\theta; \theta_0) &= \nu_{1,1}^{1/2} - \frac{\theta - \theta_0}{2}\frac{\nu_3 + \nu_{1,2}}{\nu_{1,1}^{1/2}}, \\ T''(\theta; \theta_0) &= -\left\{\frac{\nu_3 + \nu_{1,2}}{\nu_{1,1}^{1/2}} + \frac{\theta - \theta_0}{2}\left(\frac{\nu_4 + 2\nu_{1,3} + \nu_{2,2} + \nu_{1,1,2}}{\nu_{1,1}^{1/2}} + \frac{\nu_3^2 + 2\nu_3\nu_{1,2} + \nu_{1,2}^2}{2\nu_{1,1}^{3/2}}\right)\right\}, \end{aligned}$$

which are both $O\big(n^{1/2}\big)$ like T. From the latter expressions, the first-order bias of the ML estimate in (2.32) and equation (A.2) follows directly

$$B_T(\theta;\theta_0) = \frac{\nu_{1,2}}{2\nu_{1,1}^{3/2}} - \frac{\theta-\theta_0}{8}\left(\frac{3\nu_3^2 + 8\nu_3\nu_{1,2} + 5\nu_{1,2}^2}{\nu_{1,1}^{5/2}} + \frac{2\nu_4 + 4\nu_{1,3} + 2\nu_{2,2} + 2\nu_{1,1,2}}{\nu_{1,1}^{3/2}}\right).$$

Differentiating (A.2) once with regard to θ gives

$$B'_T(\theta;\theta_0) = B'(\theta)T'(\theta;\theta_0) + B(\theta)T''(\theta;\theta_0) + \frac{1}{2}\left\{-\frac{\nu'_{1,1}}{\nu_{1,1}^2}T''(\theta;\theta_0) + \nu_{1,1}^{-1}T'''(\theta;\theta_0)\right\}, \quad \text{(A.3)}$$

and it is straightforward to show that

$$\begin{aligned} B'(\theta) &= \frac{\nu_4 + 3\nu_{1,3} + 2\nu_{2,2} + 2\nu_{1,1,2}}{2\nu_{1,1}^2} + \frac{\nu_3^2 + 3\nu_3\nu_{1,2} + 2\nu_{1,2}^2}{\nu_{1,1}^3}, \qquad \text{(A.4)} \\ T'''(\theta;\theta_0) &= \frac{3}{2}\left(\frac{\nu''_{1,1}}{\nu_{1,1}^{1/2}} - \frac{\nu'^2_{1,1}}{2\nu_{1,1}^{3/2}}\right) + \frac{\theta-\theta_0}{2}\left(\frac{\nu'''_{1,1}}{\nu_{1,1}^{1/2}} - \frac{3\nu'_{1,1}\nu''_{1,1}}{2\nu_{1,1}^{3/2}} + \frac{3\nu'^3_{1,1}}{2\nu_{1,1}^{5/2}}\right). \end{aligned}$$

In order to express $B'_T(\theta_0;\theta_0)$ as reported in (2.36), we need to evaluate the derivatives of $T(\theta;\theta_0)$ at θ_0. Precisely, we obtain:

$$\begin{aligned} T'(\theta_0;\theta_0) &= \mathring{\nu}_{1,1}^{1/2}, \\ T''(\theta_0;\theta_0) &= \frac{\mathring{\nu}'_{1,1}}{\mathring{\nu}_{1,1}^{1/2}} = -\left(\frac{\mathring{\nu}_3 + \mathring{\nu}_{1,2}}{\mathring{\nu}_{1,1}^{1/2}}\right), \\ T'''(\theta_0;\theta_0) &= \frac{3}{2}\left(\frac{\mathring{\nu}''_{1,1}}{\mathring{\nu}_{1,1}^{1/2}} - \frac{\mathring{\nu}'^2_{1,1}}{2\mathring{\nu}_{1,1}^{3/2}}\right) = -\frac{3}{2}\left(\frac{\mathring{\nu}_4 + 2\nu_{1,3} + \mathring{\nu}_{2,2} + \mathring{\nu}_{1,1,2}}{\mathring{\nu}_{1,1}^{1/2}} + \frac{\mathring{\nu}_3^2 + 2\mathring{\nu}_3\mathring{\nu}_{1,2} + \mathring{\nu}_{1,2}^2}{2\mathring{\nu}_{1,1}^{3/2}}\right). \end{aligned}$$

Furthermore, terms $B(\theta_0)$ and $B'(\theta_0)$ are readily available by substitution of θ with θ_0 in formulae (2.32) and (A.4), respectively. Finally, employing such quantities and following definition (A.3), it is easy to see that

$$B'_T(\theta_0;\theta_0) = -\left(\frac{\mathring{\nu}_4 - \mathring{\nu}_{2,2} - \mathring{\nu}_{1,1,2}}{4\mathring{\nu}_{1,1}^{3/2}} + \frac{3\mathring{\nu}_3^2 + 2\mathring{\nu}_3\mathring{\nu}_{1,2} - \mathring{\nu}_{1,2}^2}{8\mathring{\nu}_{1,1}^{5/2}}\right).$$

Bibliography

Agresti, A. and Coull, B. A. (1998) Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician* **52**, 119–126.

Amemiya, T. (1981) Qualitative response models: A survey. *Journal of Economic Literature* **19**, 1483–1536.

Andersen, E. B. (1970) Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)* **32**, 283–301.

Anderson, J. and Richardson, S. (1979) Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics* **21**, 71–78.

Arellano, M. and Bond, S. (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies* **58**, 277–297.

Baker, S. G. (1995) Marginal regression for repeated binary data with outcome subject to non-ignorable non-response. *Biometrics* **51**, 1042–1052.

Baraldi, A. N. and Enders, C. K. (2010) An introduction to modern missing data analyses. *Journal of School Psychology* **48**, 5–37.

Barndorff-Nielsen, O. E. (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–365.

Barndorff-Nielsen, O. E. (1986) Inference on full and partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307–322.

Barndorff-Nielsen, O. E. (1988) *Parametric Statistical Models and Likelihood.* Springer-Verlag, Berlin Heidelberg.

Barndorff-Nielsen, O. E. (1991) Modified signed log likelihood ratio. *Biometrika* **78**, 557–563.

Barndorff-Nielsen, O. E. (1994) Adjusted versions of profile likelihood and directed likelihood, and extended likelihood. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 125–140.

Barndorff-Nielsen, O. E. (1995) Stable and invariant adjusted profile likelihood and directed likelihood for curved exponential models. *Biometrika* **82**, 489–500.

Barndorff-Nielsen, O. E. and Cox, D. R. (1979) Edgeworth and saddle-point approximations with statistical applications. *Journal of the Royal Statistical Society. Series B (Methodological)* **41**, 279–312.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994) *Inference and Asymptotics.* Chapman & Hall, London.

Bartlett, M. S. (1937) Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A (Mathematical and Physical Sciences)* **160**, 268–282.

Bartlett, M. S. (1953) Approximate confidence intervals. II. More than one unknown parameter. *Biometrika* **40**, 306–317.

Bartolucci, F., Bellio, R., Salvan, A. and Sartori, N. (2016) Modified profile likelihood for fixed-effects panel data models. *Econometric Reviews* **35**, 1271–1289.

Bellio, R. and Pierce, D. (2015) `likelihoodAsy`*: Functions for Likelihood Asymptotics.* `http://CRAN.R-project.org/package=likelihoodAsy`.

Bellio, R. and Sartori, N. (2003) Extending conditional likelihood in models for stratified binary data. *Statistical Methods and Applications* **12**, 121–132.

Bellio, R. and Sartori, N. (2015) `panelMPL`*: Modified profile likelihood estimation for fixed-effects panel data models.* `http://ruggerobellio.weebly.com/software.html`.

Beran, R. (1987) Prepivoting to reduce level error of confidence sets. *Biometrika* **74**, 457–468.

Beran, R. (1988) A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association* **83**, 687–697.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001) Interval estimation for a binomial proportion. *Statistical Science* **16**, 101–133.

Bull, S. B., Lewinger, J. P. and Lee, S. S. (2007) Confidence intervals for multinomial logistic regression in sparse data. *Statistics in Medicine* **26**, 903–918.

Bull, S. B., Mak, C. and Greenwood, C. M. (2002) A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics & Data Analysis* **39**, 57–74.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995) A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16**, 1190–1208.

Carlin, B. P. and Hodges, J. S. (1999) Hierarchical proportional hazards regression models for highly stratified data. *Biometrics* **55**, 1162–1170.

Cho Paik, M. (2004) Nonignorable missingness in matched case–control data analyses. *Biometrics* **60**, 306–314.

Cohn, D. L., Fisher, E. J., Peng, G. T., Hodges, J. S., Chesnut, J., Child, C. C., Franchino, B., Gibert, C. L., El-Sadr, W., Hafner, R. *et al.* (1999) A prospective randomized trial of four three-drug regimens in the treatment of disseminated mycobacterium avium complex disease in aids patients: excess mortality associated with high-dose clarithromycin. *Clinical Infectious Diseases* **29**, 125–133.

Cook, R., Tsai, C.-L. and Wei, B. (1986) Bias in nonlinear regression. *Biometrika* **73**, 615–623.

Cordeiro, G. M. and Barroso, L. P. (2007) A third-order bias corrected estimate in generalized linear models. *Test* **16**, 76–89.

Cordeiro, G. M. and Cribari-Neto, F. (2014) *An Introduction to Bartlett Correction and Bias Reduction.* Springer, New York.

Cordeiro, G. M. and Ferrari, S. L. P. (1991) A modified score test statistic having chi-squared distribution to order n^{-1}. *Biometrika* **78**, 573–582.

Cordeiro, G. M. and McCullagh, P. (1991) Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)* **53**, 629–643.

Cortese, G. and Sartori, N. (2016) Integrated likelihoods in parametric survival models for highly clustered censored data. *Lifetime Data Analysis* **22**, 382–404.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics.* Chapman & Hall, London.

Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)* **49**, 1–39.

Cox, D. R. and Snell, E. J. (1968) A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)* **30**, 248–275.

Cruddas, A., Reid, N. and Cox, D. (1989) A time series illustration of approximate conditional likelihood. *Biometrika* **76**, 231–237.

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and their Application*. Cambridge University Press.

De Bin, R., Sartori, N. and Severini, T. (2015) Integrated likelihoods in models with stratum nuisance parameters. *Electronic Journal of Statistics* **9**, 1474–1491.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38.

Dhaene, G. and Jochmans, K. (2014) Likelihood inference in an autoregression with fixed effects. *Econometric Theory* **First View**, 1–38.

Dhaene, G. and Jochmans, K. (2016) Bias-corrected estimation of panel vector autoregressions. *Economics Letters* **145**, 98–103.

DiCiccio, T. J., Martin, M. A., Stern, S. E. and Young, G. A. (1996) Information bias and adjusted profile likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 189–203.

DiCiccio, T. J. and Stern, S. E. (1993) An adjustment to profile likelihood based on observed information. Technical report, Department of Statistics, Stanford University.

DiCiccio, T. J. and Stern, S. E. (1994) Constructing approximately standard normal pivots from signed roots of adjusted likelihood ratio statistics. *Scandinavian Journal of Statistics* **21**, 447–460.

Diggle, P. and Kenward, M. G. (1994) Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **43**, 49–93.

Efron, B. (1975) Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics* **3**, 1189–1242.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **7**, 1–26.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap.* Chapman & Hall, New York.

Firth, D. (1993) Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2008) *Longitudinal Data Analysis.* Chapman & Hall/CRC.

Fitzmaurice, G. M., Laird, N. M. and Lipsitz, S. R. (1994) Analysing incomplete longitudinal binary responses: a likelihood-based approach. *Biometrics* **50**, 601–612.

Fitzmaurice, G. M., Molenberghs, G. and Lipsitz, S. R. (1995) Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 691–704.

Fraser, D. A. S. and Reid, N. (1988) On conditional inference for a real parameter: a differential approach on the sample space. *Biometrika* **38**, 251–274.

Fraser, D. A. S. and Reid, N. (1989) Adjustments to profile likelihoods. *Biometrika* **76**, 477–488.

Gad, A. M. and Ahmed, A. S. (2006) Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm. *Computational Statistics & Data Analysis* **50**, 2702–2714.

Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **41**, 337–348.

Hall, P. (1992) *The Bootstrap and Edgeworth Expansion.* Springer Science+Business Media, New York.

Hall, P. and Martin, M. A. (1988) On bootstrap resampling and iteration. *Biometrika* **75**, 661–671.

Hayakawa, T. and Puri, M. L. (1985) Asymptotic expansions of the distributions of some test statistics. *Annals of the Institute of Statistical Mathematics* **37**, 95–108.

He, H. and Severini, T. (2014) Integrated likelihood inference in semiparametric regression models. METRON - *International Journal of Statistics* **72**, 185–199.

Heinze, G. and Schemper, M. (2002) A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409–2419.

Hsiao, C. (2003) *Analysis of Panel Data.* Second edition. Cambridge University Press.

Hsiao, C. (2007) Panel data analysis—advantages and challenges. *TEST* **16**, 1–22.

Ibrahim, J. G., Chen, M.-H. and Lipsitz, S. R. (1999a) Monte Carlo EM for missing covariates in parametric regression models. *Biometrics* **55**, 591–596.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R. and Herring, A. H. (2005) Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association* **100**, 332–346.

Ibrahim, J. G. and Lipsitz, S. R. (1996) Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics* **52**, 1071–1078.

Ibrahim, J. G., Lipsitz, S. R. and Chen, M.-H. (1999b) Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society. Series B (Methodological)* **61**, 173–190.

Ibrahim, J. G., Lipsitz, S. R. and Horton, N. (2001) Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **50**, 361–373.

Johnson, N. J. (1978) Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association* **73**, 536–544.

Jørgensen, B. (1987) Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)* **49**, 127–162.

Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

Kent, J. (1982) Robust properties of likelihood ratio tests. *Biometrika* **69**, 19–27.

Kenward, M. G. and Molenberghs, G. (1998) Likelihood based frequentist inference when data are missing at random. *Statistical Science* **13**, 236–247.

Kim, D. K. and Taylor, J. M. (1995) The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters. *Journal of the American Statistical Association* **90**, 708–716.

Kosmidis, I. (2007) *Bias Reduction in Exponential Family Nonlinear Models* (unpublished doctoral dissertation). University of Warwick, United Kingdom.

Kosmidis, I. (2014) Bias in parametric estimation: reduction and useful side-effects. *WIREs Computational Statistics* **6**, 185–196.

Kosmidis, I. (2016) `brglm2`*: Estimation and inference for generalized linear models using explicit and implicit methods for bias reduction.* `https://github.com/ikosmidis/brglm2`.

Kosmidis, I. and Firth, D. (2009) Bias reduction in exponential family nonlinear models. *Biometrika* **96**, 793–804.

Kosmidis, I. and Firth, D. (2010) A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics* **4**, 1097–1112.

Kosmidis, I. and Firth, D. (2011) Multinomial logit bias reduction via the Poisson log-linear model. *Biometrika* **98**, 755–759.

Lancaster, T. (2000) The incidental parameter problem since 1948. *Journal of Econometrics* **95**, 391–413.

Lancaster, T. (2002) Orthogonal parameters and panel data. *Review of Economic Studies* **69**, 647–666.

Lawley, D. N. (1956) A general method for approximating to the distribution of the likelihood ratio criteria. *Biometrika* **71**, 233–244.

Lee, S. M. S. and Young, G. A. (2003) Prepivoting by weighted bootstrap iteration. *Biometrika* **90**, 393–410.

Lehmann, E. L. and Romano, J. P. (2006) *Testing Statistical Hypotheses.* Third edition. Springer Science+Business Media, New York.

Lipsitz, S. R., Parzen, M. and Ewell, M. (1998) Inference using conditional logistic regression with missing covariates. *Biometrics* **54**, 295–303.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data.* First edition. Wiley, New York.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data.* Second edition. Wiley, New York.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models.* Chapman & Hall, London.

McCullagh, P. and Tibshirani, R. (1990) A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological)* **52**, 325–344.

Mealli, F. and Rubin, D. B. (2015) Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **102**, 995–1000.

Michiels, B., Molenberghs, G. and Lipsitz, S. R. (1999) Selection models and pattern-mixture models for incomplete data with covariates. *Biometrics* **55**, 978–983.

Nelder, G. and Wedderburn, R. (1972) Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* **135**, 370–384.

Nelder, J. A. and Mead, R. (1965) A simplex method for function minimization. *The Computer Journal* **7**, 308–313.

Neyman, J. and Scott, E. (1948) Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

Pace, L. and Salvan, A. (1997) *Principles of Statistical Inference from a Neo-Fisherian Perspective.* World Scientific Publishing, Singapore.

Pace, L. and Salvan, A. (2006) Adjustments of the profile likelihood from a new perspective. *Journal of Statistical Planning and Inference* **136**, 3554–3564.

Parzen, M., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G. and Troxel, A. (2006) Pseudo-likelihood methods for longitudinal binary data with non-ignorable missing responses and covariates. *Statistics in Medicine* **25**, 2784–2796.

Peers, H. W. and Iqbal, M. (1985) Asymptotic expansions for confidence limits in the presence of nuisance parameters, with applications. *Journal of the Royal Statistical Society. Series B (Methodological)* **47**, 547–554.

Pfanzagl, J. (1973) Asymptotic expansions related to minimum contrast estimators. *The Annals of Statistics* **1**, 993–1026.

Pierce, D. A. and Bellio, R. (2006) Effects of the reference set on frequentist inferences. *Biometrika* **93**, 425–438.

Pierce, D. A. and Bellio, R. (2015) Beyond first-order asymptotics for Cox regression. *Bernoulli* **21**, 401–419.

Pierce, D. A. and Peters, D. (1992) Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)* **54**, 701–737.

Portnoy, S. (1988) Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics* **16**, 356–366.

Reid, N. (1988) Saddlepoint methods and statistical inference. *Statistical Science* **3**, 213–238.

Reid, N. (2003) Asymptotics and the theory of inference. *The Annals of Statistics* **31**, 1695–1731.

Rubin, D. B. (1976) Inference and missing data. *Biometrika* **63**, 581–592.

dos Santos, S. J. P. and Cordeiro, G. M. (1999) Corrected Wald test statistics for one-parameter exponential family models. *Communications in Statistics - Theory and Methods* **28**, 1391–1414.

Sartori, N. (2003) Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90**, 533–549.

Schaefer, R. L. (1983) Bias correction in maximum likelihood logistic regression. *Statistics in Medicine* **2**, 71–78.

Schafer, J. L. and Graham, J. W. (2002) Missing data: our view of the state of the art. *Psychological Methods* **7**, 147–177.

Severini, T. A. (1998a) Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika* **85**, 507–522.

Severini, T. A. (1998b) An approximation to the modified profile likelihood function. *Biometrika* **85**, 403–411.

Severini, T. A. (2000) *Likelihood Methods in Statistics.* Oxford University Press.

Severini, T. A. (2007) Integrated likelihood functions for non-Bayesian inference. *Biometrika* **94**, 529–542.

Shenton, L. R. and Bowman, K. (1963) Higher moments of a maximum-likelihood estimate. *Journal of the Royal Statistical Society. Series B (Methodological)* **25**, 305–317.

Shenton, L. R. and Bowman, K. (1977) *Maximum Likelihood Estimation in Small Samples.* Charles Griffin, London.

Sinha, S. and Maiti, T. (2008) Analysis of matched case–control data in presence of nonignorable missing exposure. *Biometrics* **64**, 106–114.

Sinha, S. K., Troxel, A. B., Lipsitz, S. R., Sinha, D., Fitzmaurice, G. M., Molenberghs, G. and Ibrahim, J. G. (2011) A bivariate pseudolikelihood for incomplete longitudinal binary data with nonignorable nonmonotone missingness. *Biometrics* **67**, 1119–1126.

Skovgaard, I. (1989) A review of higher order likelihood methods. *Bulletin of the International Statistical Institute* **3**, 331–351.

Skovgaard, I. (1996) An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2**, 145–165.

Stafford, J. E. (1992) *Symbolic Computation and the Comparison of Traditional and Robust Test Statistics* (unpublished doctoral dissertation). University of Toronto, Canada.

Stern, S. (1997) A second-order adjustment to the profile likelihood in the case of a multidimensional parameter of interest. *Journal of the Royal Statistical Society. Series B (Methodological)* **59**, 653–665.

Stern, S. E. (2006) Simple and accurate one-sided inference based on a class of M-estimators. *Biometrika* **93**, 973–987.

Sun, J., Loader, C. and McCormick, W. P. (2000) Confidence bands in generalized linear models. *The Annals of Statistics* **28**, 429–460.

Taniguchi, M. (1991) Third-order asymptotic properties of a class of test statistics under a local alternative. *Journal of Multivariate Analysis* **37**, 223–238.

Troxel, A. B., Harrington, D. P. and Lipsitz, S. R. (1998a) Analysis of longitudinal data with non-ignorable non-monotone missing values. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **47**, 425–438.

Troxel, A. B., Lipsitz, S. R. and Harrington, D. P. (1998b) Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. *Biometrika* **85**, 661–672.

Varin, C., Reid, N. and Firth, D. (2011) An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–42.

Warm, T. A. (1989) Weighted likelihood estimation of ability in item response theory. *Psychometrika* **54**, 427–450.

Wei, G. C. and Tanner, M. A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85**, 699–704.

Xu, J. and Gupta, A. K. (2005) Confidence intervals for the mean value of response function in generalized linear models. *Statistica Sinica* **15**, 1081–1096.

Yang, S. and Kim, J. K. (2016) Likelihood-based inference with missing data under missing-at-random. *Scandinavian Journal of Statistics* **43**, 436–454.

Young, G. A. (2003) Better bootstrapping by constrained prepivoting. *METRON* **61**, 227–242.

Young, G. A. and Smith, R. L. (2005) *Essentials of Statistical Inference.* Cambridge University Press.

Zorn, C. (2005) A solution to separation in binary response models. *Political Analysis* **13**, 157–170.

Claudia Di Caterina

CURRICULUM VITAE

Contact Information

University of Padova,
Department of Statistics,
via Cesare Battisti, 241-243,
35121, Padova, Italy.
Phone: +39 049 827 4174
e-mail: dicaterina@stat.unipd.it

Current Positions

Since January 2017;
Postdoctoral Research Fellow.
University of Padova, Department of Statistical Sciences.
Research project title: *Approximate Likelihood Inference with High-dimensional Models*
Supervisor: Prof. Nicola Sartori

Since January 2014; (expected completion: March 2017)
PhD Candidate in Statistical Sciences, admitted to the final exam.
University of Padova, Department of Statistical Sciences.
Thesis title: *Reducing the Impact of Bias in Likelihood Inference for Prominent Model Settings*
Supervisor: Prof. Nicola Sartori
Co-supervisor: Dr. Ioannis Kosmidis

Research interests

- Likelihood and pseudo likelihood methods.
- Likelihood asymptotics.
- Statistical treatment of nuisance parameters.
- Statistical computing.

Education

September 2011 – July 2013
Master degree (*laurea magistrale*) in Statistical Sciences.
University of Padova, Department of Statistical Sciences.
Title of dissertation: *Modified Profile Likelihood in Dynamic Panel Data Models*
Supervisor: Prof. Nicola Sartori
Final mark: 110/110 *cum laude*

September 2008 – July 2011
Bachelor degree (*laurea triennale*) in Statistics, Economics and Finance.
University of Padova, Faculty of Statistical Sciences.
Title of dissertation: *Bootstrap for Time Series*
Supervisor: Prof. Luisa Bisaglia
Final mark: 110/110 *cum laude*

Visiting period

September 2015 – September 2016
University College,
London, United Kingdom.
Supervisor: Dr. Ioannis Kosmidis

Computer skills

- Programming Languages: R, C (basic), Java (basic).
- Scripting Languages: PHP (basic).
- Databases: MySql (basic).
- OS environments: Mac OS X, Windows.
- Packages: LaTeX, MS Office, OpenOffice, Stata, SPSS, Gretl.

Language skills

Italian: native; English: fluent; French: moderate; Spanish: moderate.

Publications

Articles in proceedings

Di Caterina, C. and Kosmidis, I. (2016). Bias corrected z-tests for regression models. *Proceedings of the 31st International Workshop on Statistical Modelling (Dupuy, J.-F. and Josse, J., editors)* **1**, 87–92.

Abstracts

Di Caterina, C. and Sartori, N. (2016). Modified profile likelihood in complex models with many nuisance parameters. *Book of Abstracts of the 22nd International Conference on Computational Statistics (COMPSTAT 2016)*, Oviedo, Spain, August 23–26.

Bellio, R., Di Caterina, C. and Sartori, N. (2013). Monte Carlo modified likelihood for panel data models. *Book of Abstracts of the 6th International Conference of the ERCIM Working Group on Computational and Methodological Statistics*, London, UK, December 14–16.

Working papers

Di Caterina, C., Cortese, G., Bellio, R. and Sartori, N. (2016). Monte Carlo modified profile likelihood for panel data models. *In preparation.*

Di Caterina, C. and Kosmidis, I. (2016). Adjusted z-tests in regression settings. *In preparation.*

Conference presentations

Di Caterina, C. and Sartori, N. (2016). Monte Carlo modified profile likelihood for panel data models (poster). *22nd International Conference on Computational Statistics (COMPSTAT 2016)*, Oviedo, Spain, August 23–26.

Di Caterina, C. and Kosmidis, I. (2016). Adjusted z-tests for regression models (talk). *31st International Workshop on Statistical Modelling*, Rennes, France, July 4–8.

Teaching experience

October 2012 – October 2013
Tutor.
Exercises and short lectures for undergraduate students.
University of Padova, Department of Statistical Sciences.

References

Prof. Nicola Sartori
University of Padova,
Department of Statistical Sciences,
via Cesare Battisti, 241-243,
35121, Padova, Italy.
Phone: +39 049 827 4127
e-mail: sartori@stat.unipd.it

Dr. Ioannis Kosmidis
University College London,
Department of Statistical Science,
Gower Street, London WC1E6BT,
United Kingdom.
Phone: +44 20 7679 1862
e-mail: i.kosmidis@ucl.ac.uk