Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Modified score statistic based on bias reduction

Relatore Prof. Nicola Sartori
Dipartimento di Scienze Statistiche

Laureando Gabriele Uliano
Matricola 2054150

Anno Accademico 2022/2023

Contents

Introduction

Likelihood inference provides a valuable and general inferential framework for several statistical models, especially parametric ones. However, in some settings maximum likelihood estimators may be affected by a non-negligible presence of bias, thereby requiring research effort towards the mitigation of such issue. As a matter of fact, a branch of statistical literature is specifically devoted to the topic of bias reduction, which constitutes the theoretical background of this thesis.

The main purpose of this work is to investigate the performance of a modified score test statistic, defined within the framework of bias reduction in parametric models. The idea beneath this research is to study whether such statistic provides a valuable alternative to the currently-used tests, typically Wald-type ones. In such case, the modified score statistic could be used not only in place of standard likelihood-based tests, but also of Wald-type statistics based on bias reduction.

The thesis is organized as follows: in the first chapter, we provide a brief overview on the statistical literature on likelihood inference and bias reduction. In the second chapter, we focus our attention on the modified score test statistic, discussing some computational details and showing numerical examples of the corresponding implementation. In the third chapter, we assess the performance of the modified score statistic by means of simulation studies, considering different possible settings.

Chapter 1

Likelihood inference and bias reduction

1.1 Model specification

In the following, we provide a general overview on likelihood inference and, more importantly, on the theory of bias reduction. As a starting point, we choose to first discuss the problem of model specification for two reasons. In the first place, it can be regarded as an important and often delicate phase of statistical inference. Secondly, it allows us to introduce the relevant notation used throughout this work.

Let us consider the data $y = (y_1, \ldots, y_n)$, where n is the sample size. Assuming that y is generated by an underlying unknown distribution $p^0(y)$, specifying a statistical model means defining a family of probability distributions $\mathcal{F}$ such that, provided a correct specification, $p^0(y)$ belongs to $\mathcal{F}$. Here, $p^0(y)$ corresponds to either a joint density function in the continuous case or a joint probability mass function in the case of discrete data.

As described in more detail in Pace & Salvan (1997, Section 1.3.1), a statistical model can be specified following one of three levels of specification: parametric, semi-parametric and nonparametric. Throughout this work, we exclusively focus on parametric statistical models, namely

$$\mathcal{F} = \{p(y;\theta), y \in \mathcal{Y}, \theta \in \Theta \subseteq \mathbb{R}^p\}, \tag{1.1}$$

where $\mathcal{Y}$ is the sample space, Θ is the parameter space and θ is a p-dimensional real parameter. For the sake of simplicity, we generally assume that the model is correctly specified, which means that $p^0(y) = p(y;\theta_0)$ for a $\theta_0 \in \Theta$.

Following the principle of repeated sampling, we denote by Y the random variable whose realization corresponds to the observed data y. Given a generic function of the data $g(y)$, possibly depending on the parameter θ, we indicate with $E_\theta[g(Y)]$ the expectation of $g(Y)$ with respect to the density $p(y;\theta)$ in (1.1), assuming θ as the true value of the parameter. Analogously, the same notation holds for the variance, denoted by $\text{var}_\theta[g(Y)]$.

1.2 Likelihood inference

With respect to the parametric statistical model (1.1), the likelihood function can be defined as

$$L(\theta) = c(y)p(y;\theta), \tag{1.2}$$

where $c(y)$ is a positive constant of proportionality. For algebraic manageability, the log-likelihood function $\ell(\theta) = \log L(\theta)$ is often used in place of (1.2). The likelihood function is the core of likelihood inference, the influential and widespread inferential framework addressed in this section. Nonetheless, here we focus only on a few key concepts required for the main topic of this work. A thorough and comprehensive volume on likelihood inference is Pace & Salvan (1997), which provides the main source for this section.

Relevant likelihood-related quantities include the score function

$$U(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) \tag{1.3}$$

and the observed information

$$j(\theta) = -\frac{\partial}{\partial\theta^\top}U(\theta). \tag{1.4}$$

Respectively, (1.3) and (1.4) correspond to the gradient and to the negative Hessian matrix (or simply the second derivative in the case of a scalar parameter) of the log-likelihood. Furthermore, the quantity

$$i(\theta) = E_\theta[j(\theta)] = E_\theta[j(\theta;Y)]$$

is referred to as the expected information. It is worth mentioning that, when using the principle of repeated sampling, we may not explicit the dependence of the involved functions on Y in order to preserve compactness of notation. Notable properties that

hold under regular problems (see for example Cox & Hinkley, 1974, Section 4.8) include

$$E_\theta[U(\theta; Y)] = 0$$

and

$$\text{var}_\theta[U(\theta; Y)] = E_\theta[U(\theta; Y)U(\theta; Y)^\top] = i(\theta).$$

Among the several quantities related to the likelihood, a central role in likelihood inference is played by the maximum likelihood estimate, defined as the value $\hat{\theta} \in \Theta$ such that

$$L(\hat{\theta}) \geq L(\theta), \quad \theta \in \Theta,$$

where strict inequality holds when the likelihood has a unique global maximum. Provided its existence, in several applications the maximum likelihood estimate is obtained by solving the score equation

$$U(\theta) = 0, \tag{1.5}$$

which corresponds to the first-order conditions. Then, the solution $\hat{\theta}$ is such that $j(\hat{\theta})$ is positive definite.

Under suitable regularity conditions it can be shown (Cox & Hinkley, 1974, Sections 9.1 and 9.2) that the maximum likelihood estimator $\hat{\theta} = \hat{\theta}(Y)$ satisfies desirable frequentist properties, namely asymptotic unbiasedness, asymptotic efficiency and consistency. Furthermore, it holds that

$$\hat{\theta} \dot{\sim} N(\theta, i(\theta)^{-1}), \tag{1.6}$$

for large n and assuming θ as the true value of the parameter. Here and in the following the symbol "$\dot{\sim}$" reads as "is approximately distributed as". Another useful result is given by the equivariance, which allows to obtain the maximum likelihood estimate for a reparameterized model by simply transforming $\hat{\theta}$ through the desired reparameterization.

In several practical settings with p-dimensional θ, for $p > 1$, it is often of interest to focus only on a subset of components of the parameter. That is, partitioning $\theta = (\psi, \lambda)$, we address ψ as a p_0-dimensional parameter of interest, whereas λ is a $(p - p_0)$-dimensional nuisance parameter. Here and in the following, if v and u are p_1 and p_2-dimensional vectors respectively, we denote as (v, u) the $(p_1 + p_2)$-dimensional vector obtained by concatenation of v and u, therefore we avoid the use of transposition symbols to avoid excessive notational clutter.

Among the possible pseudo-likelihoods used to tackle the presence of nuisance parameters (see for example Pace & Salvan, 1997, Section 4), a standard implicit choice

is the profile likelihood $L_P(\psi) = L(\psi, \hat{\lambda}_\psi)$, where $\hat{\lambda}_\psi$ is the constrained maximum likelihood estimate of λ, obtained by maximizing the likelihood function in λ with fixed ψ. It may be worth noting that $\hat{\lambda}_{\hat{\psi}} = \hat{\lambda}$, where we wrote the maximum likelihood estimate $\hat{\theta}$ as $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$.

In the presence of interest and nuisance parameters, let us write the score function as

$$U(\theta) = (U_\psi(\theta), U_\lambda(\theta))$$

and the observed information matrix as

$$j(\theta) = \begin{bmatrix} j_{\psi\psi}(\theta) & j_{\psi\lambda}(\theta) \\ j_{\lambda\psi}(\theta) & j_{\lambda\lambda}(\theta) \end{bmatrix},$$

with the corresponding inverse indicated as

$$j(\theta)^{-1} = \begin{bmatrix} j^{\psi\psi}(\theta) & j^{\psi\lambda}(\theta) \\ j^{\lambda\psi}(\theta) & j^{\lambda\lambda}(\theta) \end{bmatrix}.$$

An analogous notation will be used for the expected information matrix $i(\theta)$, with respect to the subscripts and superscripts referring to the components of θ.

Then, the constrained estimate $\hat{\lambda}_\psi$ can be generally obtained by solving in λ

$$U_\lambda(\psi, \lambda) = 0, \tag{1.7}$$

keeping ψ fixed. However, due to the computational efforts required to solve (1.7), it is worth mentioning that a first-order approximation is available through a linear expansion of $U_\lambda(\psi, \hat{\lambda}_\psi)$ around $\hat{\theta}$, namely

$$U_\lambda(\psi, \hat{\lambda}_\psi) \doteq U_\lambda(\hat{\psi}, \hat{\lambda}) + \frac{\partial}{\partial \psi^\top} U_\lambda(\hat{\psi}, \hat{\lambda})(\psi - \hat{\psi}) + \frac{\partial}{\partial \lambda^\top} U_\lambda(\hat{\psi}, \hat{\lambda})(\hat{\lambda}_\psi - \hat{\lambda}).$$

Observing that $U_\lambda(\psi, \hat{\lambda}_\psi) = U_\lambda(\hat{\psi}, \hat{\lambda}) = 0$, we can therefore write

$$\hat{\lambda}_\psi \doteq \hat{\lambda} + j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})^{-1} j_{\lambda\psi}(\hat{\psi}, \hat{\lambda})(\hat{\psi} - \psi), \tag{1.8}$$

which corresponds to updating the respective component of $\hat{\theta}$ by a weighted difference between $\hat{\psi}$ and ψ. The derivation of (1.8) is also shown in Cox & Hinkley (1974, page 308), following from the asymptotic normality (1.6) and from the properties of multivariate Gaussian distributions.

1.3 Approximate pivots

In order to make inferences on θ we can use approximate pivots, available from the aforementioned likelihood-related quantities. Assuming fixed θ, the Wald statistic can be defined as

$$W_e(\theta) = (\hat{\theta} - \theta)^\top j(\hat{\theta})(\hat{\theta} - \theta), \tag{1.9}$$

where $j(\theta)$, $i(\theta)$ or $i(\hat{\theta})$ can be interchangeably used in place of $j(\hat{\theta})$. A second quantity of interest is the score statistic, which is defined for fixed θ as

$$W_u(\theta) = U(\theta)^\top i(\theta)^{-1} U(\theta), \tag{1.10}$$

where the maximum likelihood estimate is not required for its computation. Thirdly, it is important to mention the log-likelihood ratio statistic

$$W(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\}. \tag{1.11}$$

The quantities (1.9), (1.10) and (1.11) are closely linked by asymptotic equivalence to the first order (Azzalini, 1996, Section 4.2.2). Furthermore, they asymptotically follow a χ^2_p distribution if θ is the true value of the parameter, which allows to use them as approximate pivotal quantities. Moreover, if computed in a hypothesized parameter value $\theta_0 \in \Theta$, they can be used to test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ with respect to an approximate significance level.

Although being asymptotically equivalent, in general the quantity (1.11) is preferred since it allows to use directly the shape of the likelihood, thereby providing qualitatively superior confidence and acceptance regions when the quadratic approximation of $\ell(\theta)$ around $\hat{\theta}$ is not suitable (Pace & Salvan, 1997, page 92). Moreover, a further difference among the aforementioned approximate pivots is that (1.10) and (1.11) are parameterization-invariant, while in contrast (1.9) is not.

Also in the case of profile likelihood, approximate pivots are available for large sample inference. That is, the profile Wald statistic is given by

$$W_{Pe}(\psi) = (\hat{\psi} - \psi)^\top j^{\psi\psi}(\hat{\theta})^{-1}(\hat{\psi} - \psi), \tag{1.12}$$

while the profile score statistic is

$$W_{Pu}(\psi) = U_\psi(\psi, \hat{\lambda}_\psi)^\top i^{\psi\psi}(\psi, \hat{\lambda}_\psi) U_\psi(\psi, \hat{\lambda}_\psi). \tag{1.13}$$

Thirdly, the profile likelihood ratio statistic is defined as

$$W_P(\psi) = 2\{\ell_P(\hat{\psi}) - \ell_P(\psi)\}, \tag{1.14}$$

where $\ell_P(\psi) = \log L_P(\psi)$. Assuming $\theta = (\psi, \lambda)$ as the true value of the parameter, the quantities (1.12), (1.13) and (1.14) approximately follow a $\chi^2_{p_0}$ distribution for large n (see for example Pace & Salvan, 1997, Section 4.6). It is worth observing that, while (1.9) does not require $\hat{\lambda}_\psi$, both (1.13) and (1.14) do, which may require more computational effort when constructing confidence regions. In such situations, the approximation (1.8) may help reducing the computing time, which can prove useful in models when p is not negligible and/or in simulation studies.

When $p_0 = 1$, it is possible to use the corresponding one-sided, or signed, approximate pivots, namely the signed profile Wald statistic

$$r_{Pe}(\psi) = (\hat{\psi} - \psi)/\sqrt{j^{\psi\psi}(\hat{\theta})}, \tag{1.15}$$

the signed profile score statistic

$$r_{Pu}(\psi) = U_\psi(\psi, \hat{\lambda}_\psi)\sqrt{i^{\psi\psi}(\psi, \hat{\lambda}_\psi)} \tag{1.16}$$

and the signed root likelihood ratio statistic

$$r_P(\psi) = \text{sign}(\hat{\psi} - \psi)\sqrt{W_P(\psi)}. \tag{1.17}$$

The quantities (1.15), (1.16) and (1.17) asymptotically follow a standard Gaussian distribution, assuming (ψ, λ) as the true value of the parameter. Furthermore, they can be used to test for one-sided alternative hypotheses and, besides, they can prove useful when computing confidence intervals.

1.4 Issues with maximum likelihood estimation

As briefly described in the previous sections, likelihood inference provides a valuable framework which not only allows to obtain "automatic" parameter estimates, but also gives useful quantities for approximate hypothesis testing and for the construction of approximate confidence regions. That being said, however, maximum likelihood estimation is affected by issues which, in realistic settings with finite n, may deteriorate its overall performance.

In the first place, the maximum likelihood estimator is generally biased. Despite not being an issue with diverging n, being the bias generally of order $O(n^{-1})$, in finite datasets such bias may not be negligible. Some examples of such situations can be found, for instance, in Cordeiro & McCullagh (1991).

A second related problem involved in maximum likelihood estimation is frequently encountered when the parameter size p is large, especially if compared to n. Without taking into account problems where p is larger than n, suppose that $p < n$ and that they both diverge with $p/n \to \kappa$, for an arbitrary constant $\kappa \in [0, 1/2)$. Then, as notably shown in Sur & Candès (2019) with respect to logistic regression models, maximum likelihood inference not only fails to yield reliable estimates, but also the standard asymptotic distribution does not hold anymore. This result is also important since it shows that, even for a relatively small p/n ratio, a large number of parameters may likewise affect the properties of maximum likelihood inference.

A third issue involved in maximum likelihood estimation is given by the phenomenon known as complete or quasi-complete data separation that may occur in models for discrete data. For instance, given a binary response vector and p covariates (possibly including a unit term associated with an intercept), this problem occurs when there exists a p-dimensional hyperplane which separates the response classes. An artificially generated toy example of such situation, with $p = 2$ continuous covariates, is illustrated in Figure 1.1.

As shown in Albert & Anderson (1984), given a logistic regression model for separated data, the maximum likelihood estimate does not exist. Analogously, this holds also with respect to quasi-separated data. In such cases, standard software typically yields estimates with meaningless standard errors and the Iteratively Reweighted Least Squares algorithm (Green, 1984) does not converge. A thorough discussion on the existence of the maximum likelihood estimates in logistic regression model is provided by Candès & Sur (2020), where the authors investigate the relationship between the limiting p/n ratio κ and the underlying signal strength.

In general, manually checking for data separation or quasi-separation becomes an increasingly difficult task if $p > 2$. For this reason, there are linear programming algorithms which allow to solve such a problem. An implementation of such routines is provided in the `R` package `detectseparation` (Kosmidis et al., 2022).

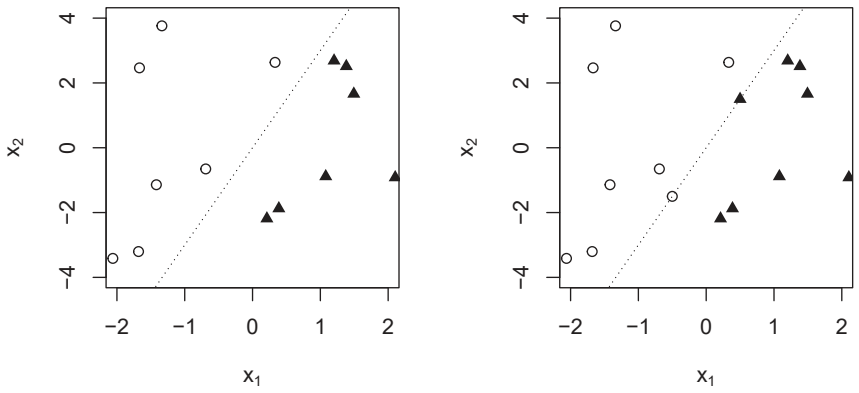

FIGURE 1.1: Illustration of complete data separation (left) and quasi-complete data separation (right), with respect to a binary classification problem with two continuous covariates, denoted by x_1 and x_2. The point shapes denote the class, while the dotted line is the true data-separating hyperplane.

1.5 Bias reduction in parametric models

In this section, we provide a brief literature review on the topic of bias reduction. Indeed, our work is contextualized within this particular framework, therefore here we aim at illustrating its key concepts.

Given a regular parametric statistical model in the form (1.1) and given the maximum likelihood estimator $\hat{\theta}$, the corresponding bias $B(\theta) = E_\theta(\hat{\theta} - \theta)$ can be expressed as (see for example Kosmidis, 2014)

$$B(\theta) = \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + \frac{b_3(\theta)}{n^3} + O(n^{-4}), \tag{1.18}$$

where $b_1(\theta), b_2(\theta)$ and $b_3(\theta)$ are $O(1)$ functions as n diverges. Although a natural way to define an unbiased estimator is $\tilde{\theta} = \hat{\theta} - B(\theta)$, in several cases this is unfeasible since $B(\theta)$ generally depends on θ and, moreover, the exact expression of (1.18) may not be available in closed form.

Following the schematic structure as in Kosmidis (2014), bias reduction can be achieved by means of two approaches: explicit or implicit bias reduction. Both strategies generally succeed in removing the first-order term in the expansion (1.18), namely $b_1(\theta)/n$, yielding in some cases second-order efficient estimators (Firth, 1993). Nonetheless, in both cases there are advantages and drawbacks that need to be carefully addressed.

1.5.1 Explicit bias reduction

The methods entailed in the class of explicit bias reduction are based on the definition of an estimator

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

following thereby a corrective rather than preventive approach with respect to $\hat{\theta}$ (Firth, 1993). This correction can be achieved through computationally intensive methods such as Jackknife (Quenouille, 1949) and bootstrap (Efron, 1979), which do not require any analytical calculation. A further class of explicit bias reduction is given by asymptotic techniques, whose aim is to directly obtain the expression of $b_1(\theta)/n$. A notable example of this approach is provided in the work of Cordeiro & McCullagh (1991), where the authors derive the explicit formulae of the first-order bias with respect to generalized linear models. Further landmark studies and their respective roles on this approach are concisely summarized in the review of Kosmidis (2014).

Of particular interest for our discussion is the general expression of the first-order asymptotic bias, shown in matrix notation in Kosmidis & Firth (2010), which is

$$\frac{b_1(\theta)}{n} = -i(\theta)^{-1}A^{\star}(\theta), \tag{1.19}$$

where the t-th component of $A(\theta)$ is given by

$$A_t^{\star}(\theta) = \frac{1}{2}\text{trace}\left\{i(\theta)^{-1}[P_t(\theta) + Q_t(\theta)]\right\}, \quad t = 1, \ldots, p, \tag{1.20}$$

with $P_t(\theta) = E_\theta[U(\theta)U(\theta)^\top U_t(\theta)]$ and $P_t(\theta) = -E_\theta[j(\theta)U_t(\theta)]$. Note that we use the subscript t in order to refer to each component of the involved vectors, with $t = 1, \ldots, p$. The main issue with explicit bias reduction is that it relies on the existence of the maximum likelihood estimate, which, as seen in the previous section, is not always guaranteed. Thus, in the case of models for binary response with linearly separated datasets, explicit bias reduction is not feasible.

1.5.2 Implicit bias reduction

A class of bias reduction methods which allows to deal with non-existent maximum likelihood estimates is given by implicit bias reduction. Originating from the work in Firth (1993), such class of methods is based on the definition of adjusted score equations,

which can be expressed in the form

$$\tilde{U}(\theta) = U(\theta) + A(\theta) = 0, \tag{1.21}$$

where $A(\theta)$ is a p-dimensional adjusting vector. In both Firth (1993) and Kosmidis & Firth (2010), there are two different expressions for $A(\theta)$, one based on the observed information matrix and the other using the expected information. More specifically, the latter defines $A(\theta) = A^{\star}(\theta)$ as in (1.20), therefore the adjusting vector does not depend on the data. Throughout this work, we only refer to this version, addressing it as "mean bias reduction" as in Kosmidis et al. (2020).

Moreover, Firth (1993) shows that, for the canonical parameter θ of an exponential family, the adjusting vector corresponds to the derivative of the Jeffreys prior in logarithmic scale, namely

$$A^{\star}(\theta) = \frac{1}{2}\frac{\partial}{\partial\theta}\log|i(\theta)|, \tag{1.22}$$

where the operator $|\cdot|$ indicates the determinant. A useful incidental property following from (1.22) is the possibility of defining a corresponding penalized log-likelihood function

$$\tilde{\ell}(\theta) = \ell(\theta) + \frac{1}{2}\log|i(\theta)|, \tag{1.23}$$

which, from a Bayesian perspective, is the logarithm of the unscaled posterior density resulting from using the Jeffreys prior. As a result, we can regard (1.21) as the first-order conditions associated to (1.23). Therefore, maximizing (1.23) equivalently yields the bias-reduced estimate $\tilde{\theta}$. It may be worth noting that, outside of canonical exponential family models, a corresponding penalized likelihood function does not generally exist, therefore $\tilde{\theta}$ is only the solution of the estimating equation (1.21).

An important issue of mean bias reduction consists on its dependence on the working parameterization. That is, given a smooth one-to-one function $\phi(\cdot)$ and the mean bias-reduced estimator $\tilde{\theta}$, it is in general not true that $\tilde{\phi} = \phi(\tilde{\theta})$ is the mean bias-reduced estimator corresponding to the reparameterized model. Nevertheless, as shown for instance in Kosmidis et al. (2020), there is exact invariance with respect to the class of affine transformations, which includes for instance parameter contrasts.

A further implicit method for bias reduction is introduced in Kenne Pagui et al. (2017) under the name of "median bias reduction". By means of such technique, it is possible to obtain bias-reduced estimates by solving (1.21), defining another adjusting vector $A(\theta) = A^{\dagger}(\theta)$. The resulting estimators share some degree of mean bias reduction, however the emphasis is placed in drawing the probability of underestimating the true

value of the parameter closer to 1/2 than $\hat{\theta}$ does, reducing as a matter of fact the median bias. Moreover, in Kenne Pagui et al. (2017) it is shown that the median bias-reduced estimators have an invariance property with respect to nonlinear and componentwise reparameterizations.

With respect to generalized linear models indexed by the parameter $\theta = (\beta, \phi)$, where β is a vector of regression coefficients and ϕ denotes a dispersion parameter, the work of Kosmidis et al. (2020) proposes a mixed adjustment approach. More specifically, this method exploits the Fisher orthogonality of β and ϕ, defining the corresponding adjusted score equation (1.21) in such a way that $A(\theta) = (A^{\star}_{\beta}(\theta), A^{\dagger}_{\phi}(\theta))$, where $A^{\star}_{\beta}(\theta)$ is the mean bias-reducing adjustment for β and $A^{\dagger}_{\phi}(\theta)$ is the median bias-reducing adjustment for ϕ. As a result, the estimators $\tilde{\beta}$ and $\tilde{\phi}$ respectively share the invariance properties of mean and median bias reduction, which is desirable in practical settings.

This approach is currently implemented as the default fitting routine in the `R` package `brglm2` (Kosmidis, 2023), however the mixed adjustment collapses to a mean bias reduction in the case of generalized linear models for Poisson and binomial data, where the dispersion parameter is fixed as $\phi = 1$.

As far as our work is concerned, we especially focus on binomial generalized linear models. For this reason and to favor more simplicity of discussion, throughout this thesis we restrict our attention towards mean bias reduction, also when illustrating examples of other parametric models.

1.5.3 Quasi-Fisher scoring

A practical aspect involved in bias reduction is to solve the adjusted score equations (1.21). In general, as with the usual score equations (1.5), it is not possible to find a solution in closed form, especially when dealing with more complex models. As a result, it may be necessary to rely on numerical methods to obtain a solution $\tilde{\theta}$.

The maximum likelihood estimate can be numerically computed by means of the well-known Newton-Raphson algorithm (see Cox & Hinkley, 1974, page 308)

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + j(\hat{\theta}^{(k)})^{-1} U(\hat{\theta}^{(k)}),$$

where the superscript denotes the iteration step $k \geq 0$, increasing until a proper convergence criterion is satisfied. In several cases, such as in generalized linear models, the expected information matrix $i(\theta)$ substitutes $j(\theta)$. Such alternative is typically addressed as the Fisher scoring algorithm.

Following from the work Kosmidis & Firth (2010), the bias reduced estimate $\tilde{\theta}$ can be numerically obtained in a similar fashion through the quasi-Fisher scoring step

$$\tilde{\theta}^{(k+1)} = \tilde{\theta}^{(k)} + i(\tilde{\theta}^{(k)})^{-1}\tilde{U}(\tilde{\theta}^{(k)}), \tag{1.24}$$

the difference being the adjusted score function $\tilde{U}(\theta)$. In particular, the term "quasi" derives from using the expected information instead of the actual negative Jacobian $\tilde{\jmath}(\theta) = -\partial\tilde{U}(\theta)/\partial\theta^{\top}$ or the corresponding expectation (Kosmidis & Firth, 2010). A similar strategy is used in the case of mixed adjustments, as described by Kosmidis et al. (2020), by simultaneously applying quasi-Fisher scoring steps with respect to the regression coefficients β and the dispersion parameter ϕ.

Chapter 2

Modified score statistic

2.1 Available approximate pivots

In the previous chapter, we have briefly addressed the methods of bias reduction in parametric models with respect to point estimation. As a matter of fact, focusing on implicit bias reduction, we have only discussed the problem of finding the solution of (1.21), which yields the bias-reduced estimate $\tilde{\theta}$. Nonetheless, from an inferential point of view we are especially interested in characterizing the uncertainty underlying the parameter estimates, which amounts to define suitable procedures for hypothesis testing and for constructing confidence regions.

In this respect, as in maximum likelihood estimation, finding an exact pivotal quantity or test statistic in the case of implicit bias reduction methods can be unfeasible, especially when the estimator $\tilde{\theta} = \tilde{\theta}(Y)$ cannot be expressed in closed form. Indeed, in typical applications we rely on numerical methods such as the quasi-Fisher scoring iterations (1.24). For this reason, in this section we address some available approximate pivotal quantities and test statistics which provide automatic inferential procedures, in a similar fashion as the approximate pivots for likelihood inference.

In the first place, in both cases of mean and median bias reduction, the asymptotic distribution of $\tilde{\theta}$ is given by (see for example Kosmidis et al., 2020, Section 3.1)

$$\tilde{\theta} \dot{\sim} N(\theta, i(\theta)^{-1}), \tag{2.1}$$

for diverging n and provided that θ is the true parameter value. From such result, it is possible to construct the Wald-type statistic

$$\tilde{W}_e(\theta) = (\tilde{\theta} - \theta)^\top i(\tilde{\theta})(\tilde{\theta} - \theta), \tag{2.2}$$

where $i(\tilde{\theta})$ could be replaced with $i(\theta)$. In this work, we refer to (2.2) as the modified Wald statistic since it can be considered as a modification of (1.9). Assuming that θ is the true value of the model parameter, the statistic (2.2) asymptotically follows a χ^2_p distribution, as a result of the property (2.1).

In the second place, under exponential family models in canonical parameterization and in the case of mean bias reduction, it is possible to define a test statistic based on the penalized log-likelihood $\tilde{\ell}(\theta)$, given by (1.23), in the form

$$\tilde{W}(\theta) = 2\{\tilde{\ell}(\tilde{\theta}) - \tilde{\ell}(\theta)\}. \tag{2.3}$$

Since the quantity (2.3) resembles (1.11), in our work we refer to it as the modified likelihood ratio statistic. Besides, due to the fact that (1.23) is a penalized log-likelihood with a $O(1)$ penalization term, it holds that (2.3) asymptotically follows a χ^2_p distribution (see for example Sartori, 2006, Section 2), assuming θ as the true parameter value.

Although the approximate pivots (2.2) and (2.3) allow the construction of approximate confidence regions and hypothesis testing, in practice it can be useful to focus only on a p_0-dimensional parameter of interest ψ, where θ can be partitioned as $\theta = (\psi, \lambda)$ and λ is a $(p - p_0)$-dimensional nuisance parameter. For this reason, we should consider the profile versions of both (2.2) and (2.3), in a similar way as with profile likelihood inference.

It is easy to define the modified profile Wald test statistic as

$$\tilde{W}_{Pe}(\psi) = (\tilde{\psi} - \psi)^\top i^{\psi\psi}(\tilde{\theta})^{-1}(\tilde{\psi} - \psi), \tag{2.4}$$

which has a limiting $\chi^2_{p_0}$ distribution under θ. Moreover, when $\tilde{\ell}(\theta)$ exists, there exists a modified profile likelihood ratio test, which can be expressed as

$$\tilde{W}_P(\psi) = 2\{\tilde{\ell}_P(\tilde{\psi}) - \tilde{\ell}_P(\psi)\}, \tag{2.5}$$

where $\tilde{\ell}_P(\psi) = \tilde{\ell}(\psi, \tilde{\lambda}_\psi)$ and $\tilde{\lambda}_\psi$ is the constrained estimate obtained by solving the λ-component of (1.21), keeping ψ fixed. For instance, such test statistic was studied with respect to logistic regression models by Heinze & Schemper (2002), where it is stated that it follows a $\chi^2_{p_0}$ limiting distribution as n diverges and assuming θ as the true parameter value.

2.2 Modified score statistic

2.2.1 Definition and motivation

Considering the case of implicit bias reduction and with respect to the adjusted score $\tilde{U}(\theta)$ as in (1.21), it is easy to note that

$$E_\theta[\tilde{U}(\theta)] = E_\theta[U(\theta)] + A(\theta) = A(\theta),$$

where $A(\theta) = O(1)$ for diverging n in both cases of mean and median bias reduction. Furthermore, the corresponding variance is given by

$$\mathrm{var}_\theta[\tilde{U}(\theta)] = \mathrm{var}_\theta[U(\theta) + A(\theta)] = i(\theta),$$

which follows from the fact that $A(\theta)$ is a non-stochastic quantity. Along with the asymptotic normality of $U(\theta)$, such properties allow us to derive the approximate pivot

$$\tilde{W}_u(\theta) = \tilde{U}(\theta)^\top i(\theta)^{-1} \tilde{U}(\theta), \tag{2.6}$$

which has a limiting χ^2_p distribution for large n if θ is the true value of the parameter. The latter property is satisfied since $A(\theta)$ is a $O(1)$ term and it is dominated by $i(\theta)$ and $U(\theta)$, which are respectively $O(n)$ and $O_p(\sqrt{n})$ as n diverges (see for example Pace & Salvan, 1997, Section 3.4). In this work, we refer to (2.6) as the modified score statistic since it can be regarded as a modification of (1.10).

With respect to the partition of the model parameter $\theta = (\psi, \lambda)$, let us write in block notation the modified score function $\tilde{U}(\psi, \lambda)$ as

$$\tilde{U}(\psi, \lambda) = (\tilde{U}_\psi(\psi, \lambda), \tilde{U}_\lambda(\psi, \lambda)),$$

and let us define the corresponding negative Jacobian matrix as

$$\tilde{\jmath}(\psi, \lambda) = -\frac{\partial \tilde{U}(\theta)}{\partial \theta^\top} = \begin{bmatrix} \tilde{\jmath}_{\psi\psi}(\psi, \lambda) & \tilde{\jmath}_{\psi\lambda}(\psi, \lambda) \\ \tilde{\jmath}_{\lambda\psi}(\psi, \lambda) & \tilde{\jmath}_{\lambda\lambda}(\psi, \lambda) \end{bmatrix}.$$

This thesis originates from the discussion in Kosmidis et al. (2020, page 58), where the test statistic

$$\tilde{W}_{Pu}(\psi) = \tilde{U}_\psi(\psi, \tilde{\lambda}_\psi)^\top i^{\psi\psi}(\psi, \tilde{\lambda}_\psi) \tilde{U}_\psi(\psi, \tilde{\lambda}_\psi) \tag{2.7}$$

is proposed as an alternative to Wald-type inference within the framework of mean and

median bias reduction. Furthermore, in Kosmidis et al. (2020, page 58) it is stated that (2.7) asymptotically follows a $\chi^2_{p_0}$ distribution, assuming $\theta = (\psi, \lambda)$ as the true parameter, which follows from the asymptotic normality of the score statistic and from the fact that the adjustment is of order $O(1)$. Such result holds in both cases of mean and median bias reduction. The quantity (2.7) is denoted in our work by the name of modified profile score statistic because it can be regarded as the profile version of (2.6).

In the case of a one-dimensional parameter of interest, namely when $p_0 = 1$, it can be useful to use the signed version of (2.7) given by

$$\tilde{r}_{Pu}(\psi) = \tilde{U}_\psi(\psi, \tilde{\lambda}_\psi)\sqrt{i^{\psi\psi}(\psi, \tilde{\lambda}_\psi)},$$

which can be easily handled when computing confidence intervals for ψ and has asymptotic standard Gaussian distribution if θ is the true value of the parameter.

It is important to highlight the core reasons which motivate a more in-depth study of both (2.6) and (2.7). In the first place, Wald-type inference provided by (2.2) can yield unsatisfactory results in cases where the (profile) penalized likelihood (1.23), if available, is highly asymmetric (see Heinze & Schemper, 2002). As a matter of fact, analogously as the usual Wald test (1.9), only symmetric confidence intervals can be obtained in such a way, however this is typically appropriate when the log-likelihood (or its penalized version) can be well approximated by the corresponding quadratic approximation around the global maximum. On the contrary, the modified score statistic may provide non-elliptical confidence regions that could better reflect the behaviour of the penalized log-likelihood.

A second motivation for investigating the performance of the modified (profile) score statistic is given by the absence of (1.23) in more general settings. Despite its nice properties addressed in the literature, a prominent example being Heinze & Schemper (2002) with respect to logistic regression, outside of canonical exponential families and mean bias reduction the equivalence (1.22) does not hold in general. On the contrary, the approximate pivots (2.6) and (2.7) can be defined with respect to more general models, in the same way as Wald-type inference.

In the case of mean and median bias reduction for generalized linear models, the inferential procedures are typically carried out by means of Wald tests (2.2) and (2.4). As an example, the default procedures for computing p-values and confidence intervals for the model parameters in the `brglm2` package are based on Wald-type inference. Therefore, in our work we also provide an implementation of alternative inferential procedures based on (2.6) and (2.7), which can be used with respect to `brglmFit` objects

with known dispersion parameter. The `R` code of such implementation is reported in the Appendix.

2.2.2 Computational considerations

An important issue, which has not yet been addressed in this work, is how to compute the constrained bias-reduced estimate $\tilde{\lambda}_\psi$. It is worth noting that such quantity is required for both modified profile likelihood ratio test (2.5) and modified profile score statistic (2.7), in the same way as the constrained maximum likelihood estimate of the nuisance parameter λ is needed to compute (1.13) and (1.14).

To begin with, the quantity $\tilde{\lambda}_\psi$ is the solution of

$$\tilde{U}_\lambda(\psi, \lambda) = U_\lambda(\psi, \lambda) + A_\lambda(\psi, \lambda) = 0, \tag{2.8}$$

keeping ψ fixed, which amounts to solving the λ-related component of the adjusted equations (1.21).

Although such approach may remind that of solving the usual λ-related score equation (1.7) to obtain $\hat{\lambda}_\psi$, there is however a distinction to be made. On the one hand, solving (1.7) amounts to finding the maximum likelihood estimate of λ in the sub-model with ψ kept fixed. This result follows from the fact that equations (1.7) and (1.5) coincide. On the other hand, in the case of mean or median bias reduction such equivalence is no longer true. Such discrepancy can be explained by the fact that the adjusted score equations (1.21) depend on the adjustment term $A(\theta)$, which is model-dependent in both mean and median bias reduction. As a consequence, considering the sub-model with fixed ψ, the bias-reduced estimate of the nuisance parameter is a quantity denoted by $\tilde{\lambda}^\star_\psi$ which generally differs from $\tilde{\lambda}_\psi$.

In general, it is unfeasible to solve (2.8) analytically, therefore numerical procedures are required. Keeping ψ fixed, a first-order Taylor expansion of $\tilde{U}_\lambda(\psi, \tilde{\lambda}_\psi)$ around the point λ yields

$$\tilde{U}_\lambda(\psi, \tilde{\lambda}_\psi) \doteq \tilde{U}_\lambda(\psi, \lambda) + \frac{\partial}{\partial \lambda^\top} \tilde{U}_\lambda(\psi, \lambda)(\tilde{\lambda}_\psi - \lambda),$$

from which it follows that

$$\tilde{\jmath}_{\lambda\lambda}(\psi, \lambda)(\tilde{\lambda}_\psi - \lambda) \doteq \tilde{U}_\lambda(\psi, \lambda),$$

where $\tilde{\jmath}_{\lambda\lambda}(\psi, \lambda) = -\partial \tilde{U}_\lambda(\psi, \lambda)/\partial \lambda^\top$. Therefore, we can write

$$\tilde{\lambda}_\psi \doteq \lambda + \tilde{\jmath}_{\lambda\lambda}(\psi, \lambda) \tilde{U}_\lambda(\psi, \lambda),$$

from which we can derive the Newton-Raphson step

$$\tilde{\lambda}_{\psi}^{(k+1)} = \tilde{\lambda}_{\psi}^{(k)} + \tilde{j}_{\lambda\lambda}(\psi, \tilde{\lambda}_{\psi}^{(k)})^{-1}\tilde{U}_{\lambda}(\psi, \tilde{\lambda}_{\psi}^{(k)}), \quad k = 0, 1, \ldots, \tag{2.9}$$

where $\tilde{\lambda}_{\psi}^{(k)}$ denotes the current constrained estimate for λ. A difficulty which may arise is that the term $\partial A_{\lambda}(\psi, \lambda)/\partial\lambda^{\top}$ can be algebraically tedious. For this reason, we have tried two possible solutions. Firstly, we can rely on the numerical derivation of the negative Jacobian matrix $\tilde{j}_{\lambda\lambda}(\psi, \lambda)$ at each iteration of (2.9). A second possibility, inspired by the quasi-Fisher scheme of (1.24), is to use as negative Jacobian the λ-related block of expected information matrix, namely $i_{\lambda\lambda}(\psi, \lambda)$, thereby omitting the term $\partial A_{\lambda}(\psi, \lambda)/\partial\lambda^{\top}$. Throughout our simulation studies, the latter approach has proven numerically more stable in the case of a relatively high number of nuisance parameters. An efficient implementation of (2.9) in `R` is available by means of the routine `nleqslv`, contained in the homonymous package (Hasselman, 2023), which also allows to use the quasi-Fisher counterpart by imposing $-i_{\lambda\lambda}(\psi, \lambda)$ as Jacobian matrix.

An issue that needs to be addressed is that solving (2.8) may entail expensive computational effort. As a matter of fact, considering for instance the implementation of confidence regions, it is required to compute the constrained solution $\tilde{\lambda}_{\psi}$ several times with respect to a suitable grid of ψ values. Besides, such an approach may be unfeasible when the nuisance parameter has a non-negligible size, for example when dealing with generalized linear models, when profiling for each regression parameter.

For this reason, in our work we also take into consideration an approximate solution to (2.8), given by

$$\tilde{\lambda}_{\psi} \doteq \tilde{\lambda} + \tilde{j}_{\lambda\lambda}(\tilde{\theta})^{-1}\tilde{j}_{\lambda\psi}(\tilde{\theta})(\tilde{\psi} - \psi). \tag{2.10}$$

Such result reminds that of (1.8). Indeed, it follows from a first-order Taylor expansion of $\tilde{U}_{\lambda}(\psi, \tilde{\lambda}_{\psi})$ around $\tilde{\theta} = (\tilde{\psi}, \tilde{\lambda})$, namely

$$\tilde{U}_{\lambda}(\psi, \tilde{\lambda}_{\psi}) \doteq U_{\lambda}(\tilde{\psi}, \tilde{\lambda}) + \frac{\partial}{\partial\psi^{\top}}\tilde{U}_{\lambda}(\tilde{\psi}, \tilde{\lambda})(\psi - \tilde{\psi}) + \frac{\partial}{\partial\lambda^{\top}}\tilde{U}_{\lambda}(\tilde{\psi}, \tilde{\lambda})(\tilde{\lambda}_{\psi} - \tilde{\lambda}).$$

The approximation (2.10) allows a fast implementation of the inferential procedures based not only on (2.7), but also on (2.5), which analogously needs the computation of the constrained estimate $\tilde{\lambda}_{\psi}$.

A further issue that has been encountered throughout our numerical trials concerns the initialization of algorithm (2.9), namely $\tilde{\lambda}_{\psi}^{(0)}$. As a matter of fact, a generic initial value for the nuisance parameter can result in a very slow convergence towards the constrained estimate $\tilde{\lambda}_{\psi}$, if not even numerical errors. Besides, such a problem may

occur even in cases where ψ is not too far away from $\tilde{\psi}$, for instance when computing confidence intervals. For this reason, it is important to rely on a proper initialization strategy that takes into account the value of ψ requested by the user.

We propose two possible solutions, which in our trials have proven the most numerically stable and effective. As a first option, we suggest as initial value $\tilde{\lambda}_{\psi}^{(0)}$ the linear approximation provided by (2.10) since it can be readily obtained through a reasonable computational effort.

A second possibility is provided by the initialization $\tilde{\lambda}_{\psi}^{(0)} = \tilde{\lambda}_{\psi}^{\star}$, namely by using the bias-reduced estimate of the sub-model with fixed ψ. If ψ is a subset of coefficients of a generalized linear model, it is possible to obtain $\tilde{\lambda}_{\psi}^{\star}$ by fitting the model with an offset term, where the latter is a linear combination of ψ and the associated columns of the model matrix. In our numerical trials, obtaining $\tilde{\lambda}_{\psi}^{\star}$ was easier than $\tilde{\lambda}_{\psi}$ since the fitting routine in `brglmFit` uses automatic and stable initial points.

Throughout our numerical trials, such approach proved fairly more stable than the first one in the case of logistic regression models with relatively high number of covariates. Nonetheless, this numerical stability is achieved at the cost of more computing time, considering that $\tilde{\lambda}_{\psi}^{\star}$ is obtained through numerical procedures such as (1.24). For this reason, our proposed strategy is to use $\tilde{\lambda}_{\psi}^{\star}$ as initial point in case the first initialization approach failed.

2.3 Examples

In this section, we illustrate some numerical examples with respect to simple parametric models. In particular, we mainly consider cases in which $p = 2$, allowing us to investigate the behaviour of the modified profile score statistic (2.6), the constrained estimates $\tilde{\lambda}_{\psi}$ and also the corresponding approximations, as described in the previous sections. Furthermore, we also consider different sample sizes, namely $n \in \{20, 50, 100, 200\}$, in order to illustrate the effect of increasing the amount of information about the model parameters. At the end of this section, we also provide a numerical illustration which involves a real data set, with respect to the logistic regression model, considering that the latter is of particular interest for bias reduction.

2.3.1 Canonical gamma model

Let us consider n independent observations $y_1, \ldots, y_n$ sampled from a gamma distribution under canonical parameterization $\theta = (\alpha, \lambda)$, where $\alpha, \lambda > 0$. Then, the corresponding log-likelihood is

$$\ell(\theta) = -n \log \Gamma(\alpha) + n\alpha \log \lambda + (\alpha - 1) \sum_{i=1}^{n} \log y_i - \lambda \sum_{i=1}^{n} y_i,$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$ is the gamma function. From this, we derive the score function

$$U(\theta) = \begin{bmatrix} n \log \lambda - n\Psi(\alpha) + \sum_{i=1}^{n} \log y_i \\ n\alpha/\lambda - \sum_{i=1}^{n} y_i \end{bmatrix}$$

where $\Psi(\alpha) = \partial \log \Gamma(\alpha)/\partial\alpha$ denotes the digamma function. The observed information matrix is given by

$$j(\theta) = \begin{bmatrix} n\Psi^{(1)}(\alpha) & -n/\lambda \\ -n/\lambda & n\alpha/\lambda^2 \end{bmatrix}$$

where $\Psi^{(k)}(\alpha) = \partial^k \Psi(\alpha)/\partial\alpha^k$ denotes the polygamma function, for $k \in \{1, 2, \ldots\}$. We observe that $j(\theta)$ is equal to the expected information $i(\theta) = E_\theta\,[(j(\theta; Y)]$ due to the canonical parameterization.

The maximum likelihood estimate $\hat{\theta}$ can be obtained by numerically solving (1.5), whereas the mean bias-reduced estimate $\tilde{\theta}$ can be computed by solving (1.21). For the latter, we need the expression for the adjusting vector $A(\theta)$, which can be conveniently derived through equality (1.22). In this regard, the Jeffreys prior in logarithmic scale is given by

$$\frac{1}{2} \log|i(\theta)| = \frac{1}{2} \log \left\{ \frac{n^2}{\lambda^2} [\alpha\Psi^{(1)}(\alpha) - 1] \right\} = \log n - \log \lambda + \frac{1}{2} \log[\alpha\Psi^{(1)}(\alpha) - 1],$$

from which we can express

$$A(\theta) = \frac{\partial}{\partial\theta} \left\{ \frac{1}{2} \log \left| i(\theta) \right| \right\} = \begin{bmatrix} \frac{1}{2} \frac{\Psi^{(1)}(\alpha) + \alpha\Psi^{(2)}(\alpha)}{\alpha\Psi^{(1)}(\alpha) - 1} \\ -1/\lambda \end{bmatrix}.$$

By simulating n independent gamma observations, with $n = 20, 50, 100, 200$ and $\theta = (5, 2)$, we compute the maximum likelihood estimate $\hat{\theta} = (\hat{\alpha}, \hat{\lambda})$ and the mean bias-reduced estimates $\tilde{\theta} = (\tilde{\alpha}, \tilde{\lambda})$, shown in Table 2.1. In this case, even if it is only a single sample, it is noteworthy that mean bias reduction yields fairly more accurate results than maximum likelihood estimation. Besides, as n increases, both methods lead to

TABLE 2.1: Maximum likelihood and mean bias-reduced estimates of the canonical gamma model with respect to different sample sizes.

	$n = 20$	$n = 50$	$n = 100$	$n = 200$
$\hat{\alpha}$	8.5979	6.3699	6.6308	5.4012
$\hat{\lambda}$	3.9633	2.6586	2.8236	2.1416
$\tilde{\alpha}$	7.3396	6.0005	6.4384	5.3234
$\tilde{\lambda}$	3.3602	2.4961	2.7374	2.1088

estimates that are closer to the true parameter value, as expected from the asymptotic theory.

In Figure 2.1, we illustrate the 95% confidence regions obtained with the modified score statistic (2.6). As expected, the increasing sample size shrinks the area of the confidence regions, thereby locating with more certainty plausible values for the model parameters. Furthermore, we also observe that with large sample size all the modified statistics yield almost juxtaposing confidence regions, as expected from their asymptotic equivalence.

The confidence intervals for each parameter can be obtained by profiling both α and λ through (2.7), for which the constrained bias-reduced estimates $\tilde{\lambda}_\alpha$ and $\tilde{\alpha}_\lambda$ are needed respectively.

In the first place, we need to solve

$$\tilde{U}_\lambda(\alpha, \lambda) = n\alpha/\lambda - \sum_{i=1}^{n} y_i - 1/\lambda = 0,$$

whose solution can be easily expressed in closed form as

$$\tilde{\lambda}_\alpha = \frac{n\alpha - 1}{\sum_{i=1}^{n} y_i}.$$

Such constrained estimate is a linear function in α, hence the corresponding linear approximation given by (2.10) yields the same result. In Figure 2.2, we illustrate the modified score function $\tilde{W}_{Pu}(\alpha)$, along with $\tilde{W}_{Pe}(\alpha)$ and $\tilde{W}_P(\alpha)$, where we highlight the resulting approximate 95% confidence interval. We can observe that, in accordance with the global modified score regions, increasing n concentrates the modified profile score statistic around $\tilde{\alpha}$, thereby defining narrower confidence intervals for α. Furthermore, all the modified profile statistics yield almost analogous results considering a large n.

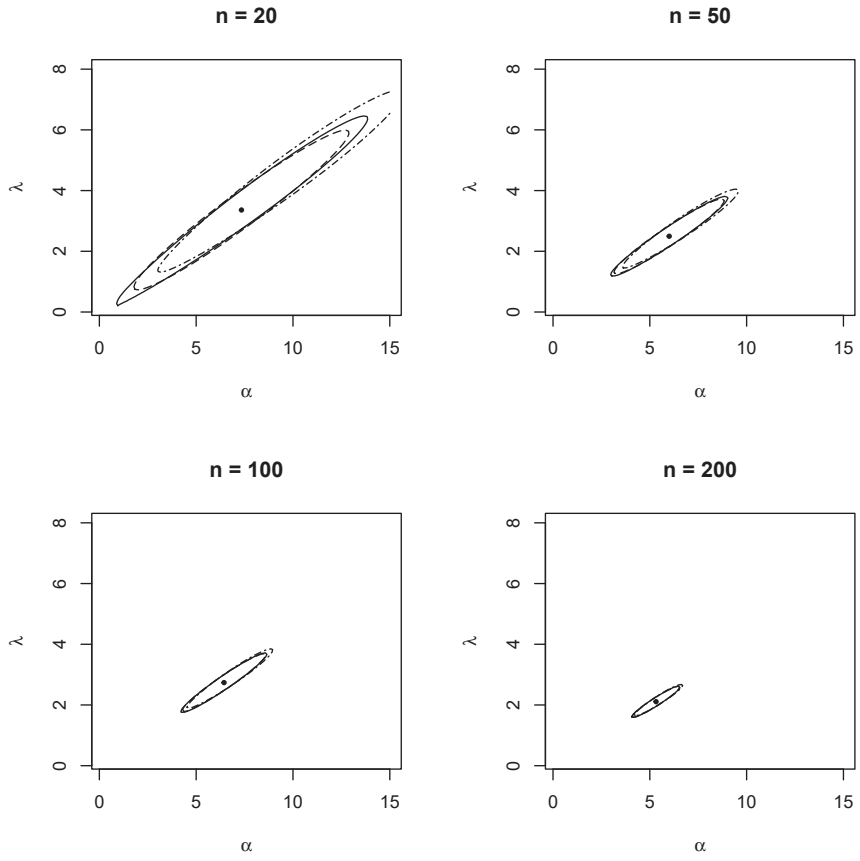

FIGURE 2.1: Approximate 95% confidence regions of θ obtained through $\tilde{W}_u(\theta)$ (solid line), $\tilde{W}_e(\theta)$ (long-dashed line) and $\tilde{W}(\theta)$ (dot-dashed line) for the canonical gamma model, with respect to different sample sizes. The dot corresponds to the mean bias-reduced estimates.

In the second place, we compute $\tilde{\alpha}_\lambda$ by solving

$$\tilde{U}_\alpha(\alpha, \lambda) = n \log \lambda - n\Psi(\alpha) + \sum_{i=1}^{n} \log y_i + \frac{1}{2} \frac{\Psi^{(1)}(\alpha) + \alpha\Psi^{(2)}(\alpha)}{\alpha\Psi^{(1)}(\alpha) - 1} = 0,$$

which is a nonlinear equation in α that needs to be solved numerically. In Figure 2.3, we show the constrained bias-reduced estimate $\tilde{\alpha}_\lambda$ as a function of λ and the corresponding linear approximation. Quite unexpectedly, there is no appreciable difference between the exact and approximate solutions, although $\tilde{\alpha}_\lambda$ is nonlinear in λ. In Figure 2.4, we illustrate the modified profile score statistic $\tilde{W}_{Pu}(\lambda)$, computed with both the exact and approximate solutions $\tilde{\alpha}_\lambda$. A slight difference can be noticed between the two curves, considering the case of $n = 20$ and small values of λ, however such difference becomes almost negligible for higher sample sizes.

Numerical results with respect to the confidence intervals are summarized in Table

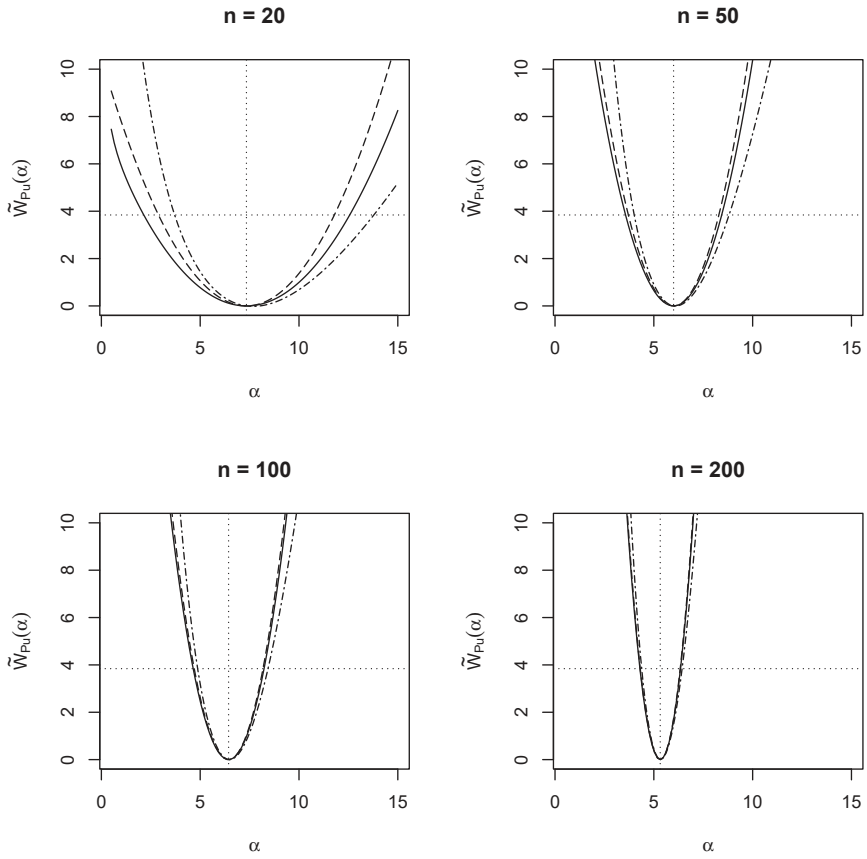

FIGURE 2.2: Modified profile score statistic for α (solid line) in the canonical gamma model, along with $\tilde{W}_{Pe}(\alpha)$ (long-dashed line) and $\tilde{W}_P(\alpha)$ (dot-dashed line), with respect to different sample sizes. The horizontal dotted line corresponds to the 0.95-quantile of a χ^2_1 distribution, while the vertical dotted line is the mean bias-reduced estimate $\tilde{\alpha}$.

2.2. We can notice that, at least for the simulated sample trajectory, all the confidence intervals include the true parameter value.

TABLE 2.2: Confidence intervals obtained through the modified profile score statistic in the canonical Gamma model, for each parameter and with respect to different sample sizes.

n	α	λ
20	(2.1459, 12.5653)	(0.8746, 5.8431)
50	(3.5693, 8.4338)	(1.4386, 3.5534)
100	(4.6456, 8.2316)	(1.9439, 3.5309)
200	(4.2961, 6.3508)	(1.6817, 2.5358)

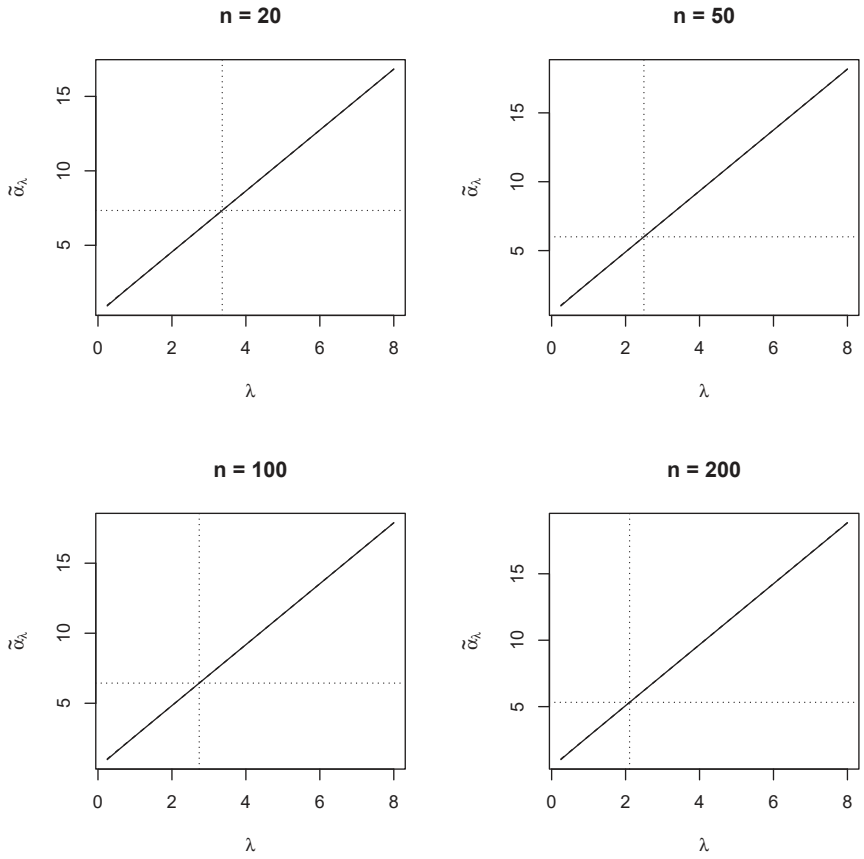

FIGURE 2.3: Constrained estimates of α as functions of λ for the canonical gamma model, with respect to different sample sizes. The solid line corresponds to the exact solution, whereas the dashed line shows the linear approximation, although in this example they are graphically indistinguishable. The dotted horizontal and vertical lines correspond to the global solution $\tilde{\theta}$.

2.3.2 Canonical inverse Gaussian model

Let us consider n independent observations $y_1, \ldots, y_n$ as realizations of an inverse Gaussian distribution in canonical parameterization, namely with joint density

$$p(y;\theta) = \prod_{i=1}^{n} \sqrt{\frac{\lambda}{2\pi y_i^3}} \exp\left\{\sqrt{\lambda\phi} - \frac{\lambda}{2y_i} - \frac{\phi y_i}{2}\right\},$$

where $y_i > 0$ for all $i = 1, \ldots, n$ and $\theta = (\lambda, \phi)$ is such that $\lambda > 0$ and $\phi \geq 0$. Discarding additive constants independent of θ, we express the log-likelihood of such model as

$$\ell(\theta) = \frac{n}{2}\log\lambda + n\sqrt{\lambda\phi} - \frac{\lambda}{2}\sum_{i=1}^{n}\frac{1}{y_i} - \frac{\phi}{2}\sum_{i=1}^{n} y_i.$$

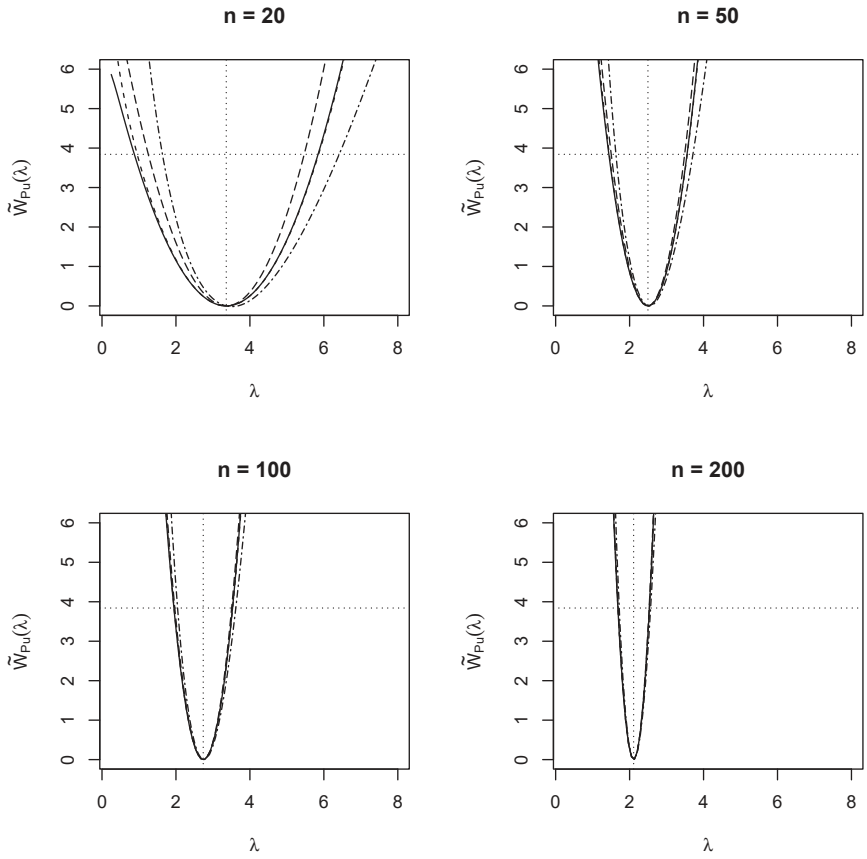

FIGURE 2.4: Modified profile score statistic (solid line) for λ in the canonical gamma model, along with the corresponding approximation (dashed line), $\tilde{W}_{Pe}(\lambda)$ (long-dashed line) and $\tilde{W}_P(\lambda)$, with respect to different sample sizes. The horizontal dotted line corresponds to the 0.95-quantile of a χ^2_1 distribution, while the vertical dotted line is the mean bias-reduced estimate $\tilde{\lambda}$.

The specified model is an exponential family with minimal sufficient statistic given by $(-1/(2\sum_{i=1}^n y_i), -\sum_{i=1}^n y_i/2)$ and canonical parameter θ. Nonetheless, such exponential family is full but not regular, since the natural parameter space is not open in $\mathbb{R}^2$, admitting indeed $\phi = 0$ (see for more detail Pace & Salvan, 1997, pages 175 and 186). For this reason, throughout the following computations we assume that $\phi > 0$.

The score function can be expressed as

$$U(\theta) = \begin{bmatrix} \frac{n}{2\lambda} + \frac{n}{2}\sqrt{\frac{\phi}{\lambda}} - \frac{1}{2}\sum_{i=1}^n \frac{1}{y_i} \\ \frac{n}{2}\sqrt{\frac{\lambda}{\phi}} - \frac{1}{2}\sum_{i=1}^n y_i \end{bmatrix}$$

and the corresponding observed information matrix

$$j(\theta) = \begin{bmatrix} \frac{n}{2\lambda^2} + \frac{n}{4}\sqrt{\phi}\lambda^{-3/2} & -\frac{n}{4}(\phi\lambda)^{-1/2} \\ -\frac{n}{4}(\phi\lambda)^{-1/2} & \frac{n}{4}\sqrt{\lambda}\phi^{-3/2} \end{bmatrix}.$$

Due to the canonical parameterization $j(\theta) = i(\theta)$, then equality (1.22) still holds, which allows us to easily obtain the adjustment vector for mean bias reduction

$$A(\theta) = \frac{\partial}{\partial\theta}\left\{\frac{1}{2}\log\left|i(\theta)\right|\right\} = -\frac{3}{4}\begin{bmatrix} 1/\lambda \\ 1/\phi \end{bmatrix}.$$

After generating $n \in \{20, 50, 100, 200\}$ pseudo-observations from an inverse Gaussian distribution with $\theta = (3, 2)$, we obtain the maximum likelihood and the bias-reduced estimates $\hat{\theta}$ and $\tilde{\theta}$ by solving (1.5) and (1.21). Defining $s_1 = \sum_{i=1}^{n} 1/y_i$ and $s_2 = \sum_{i=1}^{n} y_i$, we can express $\hat{\theta}$ in closed form as

$$\hat{\phi} = \frac{n^3}{s_1 s_2^2 - n^2 s_2},$$

and

$$\hat{\lambda} = \frac{n s_2}{s_1 s_2 - n^2}.$$

Note that the positivity of $\hat{\lambda}$ and $\hat{\phi}$ follows from the fact that $s_1 s_2 > n^2$ since $s_1 s_2 = n^2\bar{y}/\bar{y}^a$ and $\bar{y} > \bar{y}^a$, using Jensen's inequality for the latter, where $\bar{y}$ and $\bar{y}^a$ are the sample arithmetic and harmonic mean respectively.

Also in the case of mean bias reduction, it is possible to obtain $\tilde{\theta}$ analytically, however this requires fairly more algebraic effort. In the first place, we can obtain the constrained estimate $\tilde{\lambda}_\phi$ as

$$\tilde{\lambda}_\phi = \left[\frac{n\sqrt{\phi} + \sqrt{n^2\phi + 4s_1(n - 3/2)}}{2s_1}\right]^2,$$

Secondly, by substituting $\tilde{\lambda}_\phi$ into the ϕ-related modified score equation, we obtain a quadratic equation whose solutions are given by

$$\phi_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

where $a = 4s_1 s_2(s_1 s_2 - 1)$, $b = 6(2s_1 s_2 - n^2)/n^2 - 4s_1(n - 3/2)$ and $c = 9s_1^2/n^2$. Thus, the modified score equation admits two solutions. However, our numerical trials suggest that we use

$$\tilde{\phi} = \frac{-b + \sqrt{b^2 - 4ac}}{2a},$$

in order to obtain reasonable estimates for θ. The existence of two solutions could be explained from the fact that the penalized log-likelihood is

$$\tilde{\ell}(\theta) = \frac{n}{2}\log\lambda + n\sqrt{\lambda\phi} - \frac{\lambda}{2}\sum_{i=1}^{n}\frac{1}{y_i} - \frac{\phi}{2}\sum_{i=1}^{n} y_i - \frac{3}{4}\log\lambda - \frac{3}{4}\log\phi, \tag{2.11}$$

which is not globally concave, since $\tilde{\ell}(\theta)$ diverges to infinity as $\phi \to 0^+$. As far as our numerical trials are concerned, the inspection of the negative Hessian matrix of (2.11)

$$\tilde{j}(\theta) = \begin{bmatrix} \frac{n}{2\lambda^2} + \frac{n}{4}\sqrt{\phi}\lambda^{-3/2} - \frac{3}{4\lambda^2} & -\frac{n}{4}(\phi\lambda)^{-1/2} \\ -\frac{n}{4}(\phi\lambda)^{-1/2} & \frac{n}{4}\sqrt{\lambda}\phi^{-3/2} - \frac{3}{4\phi^2} \end{bmatrix}$$

reveals that the point $(\tilde{\lambda}, \tilde{\phi})$ corresponds to a local maximum of (2.11) because $\tilde{j}(\tilde{\theta})$ is positive definite. Instead, computing $\tilde{j}(\theta)$ in the other solution leads to an indefinite matrix, therefore solving the modified score equation also locates a saddlepoint of (2.11).

In Table 2.3, we show the maximum likelihood and the mean bias-reduced estimates with respect to the simulated data, with $n \in \{20, 50, 100, 200\}$. As in the previous example, the mean bias-reduced estimates are generally closer to the true parameter value than the maximum likelihood ones.

TABLE 2.3: Maximum likelihood and mean bias-reduced estimates of the canonical inverse Gaussian model with respect to different sample sizes.

	$n = 20$	$n = 50$	$n = 100$	$n = 200$
$\hat{\lambda}$	3.3175	3.5664	3.6483	3.4560
$\hat{\phi}$	2.5799	2.4069	2.1302	1.9493
$\tilde{\lambda}$	2.8118	3.3512	3.5385	3.4041
$\tilde{\phi}$	2.0523	2.2121	2.0431	1.9088

In Figure 2.5, we illustrate the 95% confidence regions determined through the modified score statistic $\tilde{W}_u(\lambda, \phi)$ and we can observe that, as n increases, the confidence regions shrink around the mean bias-reduced estimate. We also note that, in the case of $n = 20$, the confidence region could not be fully determined near the boundary of the parameter space.

The constrained estimate $\tilde{\phi}_\lambda$ can be easily expressed in closed form as

$$\tilde{\phi}_\lambda = \left[\frac{n\sqrt{\lambda} + \sqrt{n^2\lambda - 6s_2}}{2s_2}\right]^2,$$

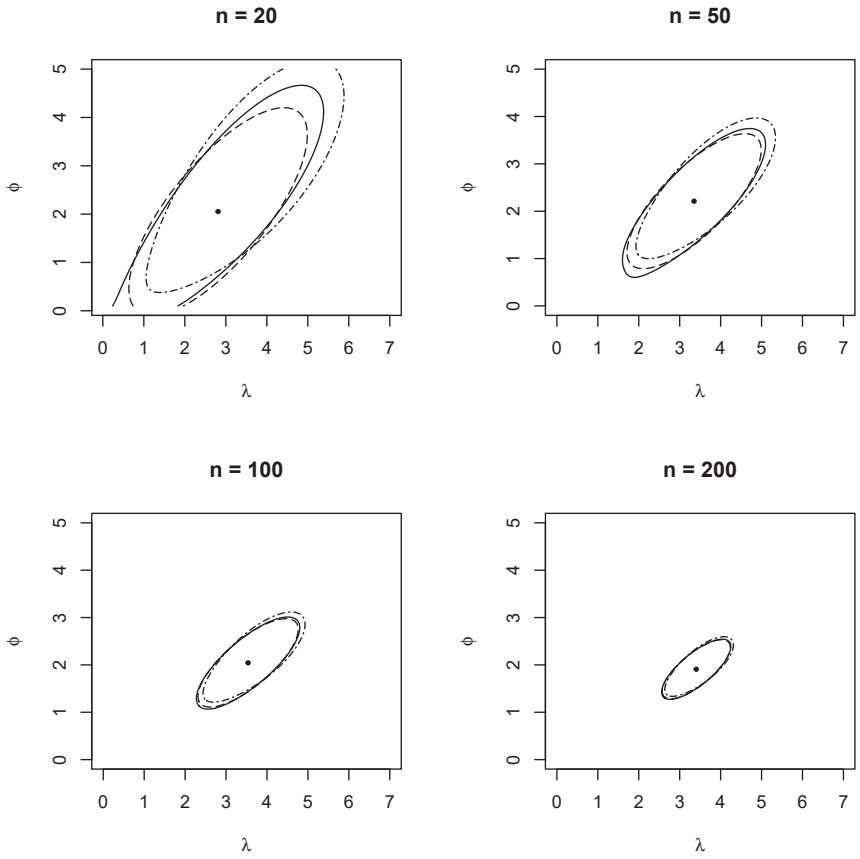

FIGURE 2.5: Approximate 95% confidence regions of θ obtained through $\tilde{W}_u(\theta)$ (solid line), $\tilde{W}_e(\theta)$ (long-dashed line) and $\tilde{W}(\theta)$ (dot-dashed line) for the canonical inverse Gaussian model, with respect to different sample sizes. The dot corresponds to the mean bias-reduced estimates.

which is a nonlinear function of λ. Nevertheless, we note that such estimate is available provided that $\lambda > 6s_2$ due to the root term. Such condition can be problematic when computing confidence intervals for λ, considering that we may need to obtain the modified score statistic $\tilde{W}_{Pu}(\lambda)$ for sufficiently small λ.

The availability of a closed-form expression for $\tilde{\phi}_\lambda$ makes the corresponding linear approximation useless from a computational perspective. However, we are also interested in studying the suitability of such approximation, especially as n varies. For this reason, we provide an illustration of both versions of the constrained estimates of ϕ in Figure 2.6, where we can observe an almost complete overlap between the two curves in each panel.

The modified score function for λ is illustrated in Figure 2.7, where we also highlight the resulting approximated 95% confidence interval. We can notice that the confidence intervals become narrower as n increases, in accordance with the global confidence regions.

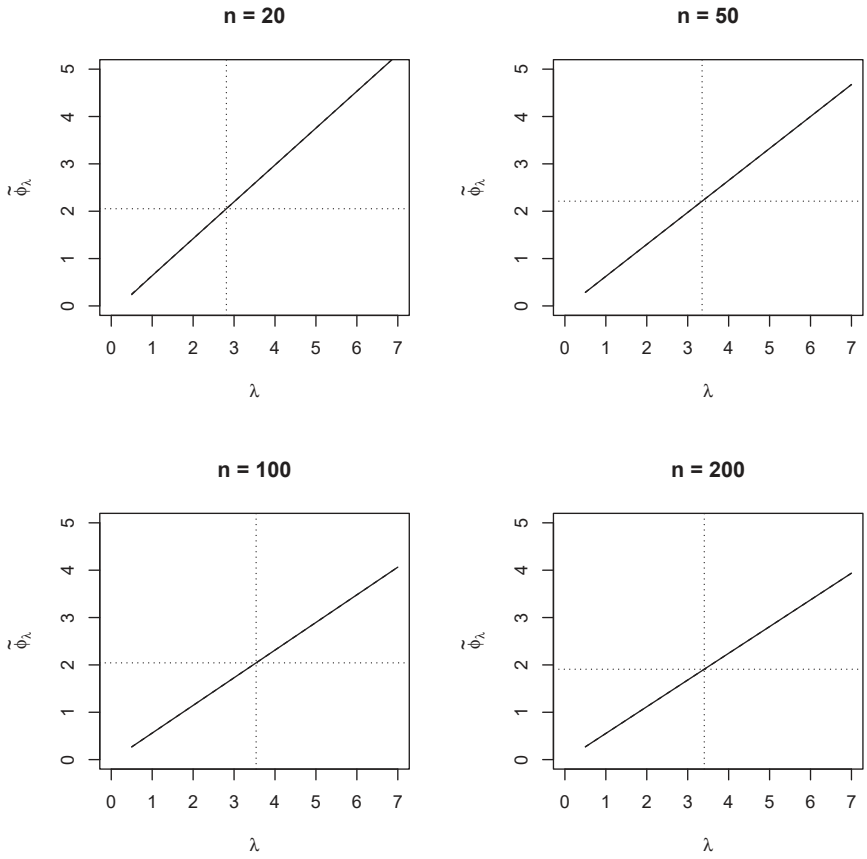

FIGURE 2.6: Constrained estimates of ϕ as functions of λ in the canonical inverse Gaussian model, with respect to different sample sizes. The solid line corresponds to the exact solution, whereas the dashed line shows the linear approximation, although in this example they are graphically indistinguishable. The dotted horizontal and vertical lines correspond to the global solution $\tilde{\theta}$.

Considering ϕ as the parameter of interest, the expression for the constrained estimate $\tilde{\lambda}_\phi$ is already available from obtaining $\tilde{\theta}$ and corresponds to a nonlinear function of ϕ. Unlike $\tilde{\phi}_\lambda$, in this case we do not require further constraints with respect to the parameter space, since for a reasonable problem we assume that $n \geq 2$. The constrained estimate and the corresponding linear approximation are displayed in Figure 2.8, where we can notice a slight difference between the two curves in each panel, especially when ϕ is sufficiently close to 0.

Such discrepancy becomes more appreciable when we consider the modified score statistic $\tilde{W}_{Pu}(\phi)$, as represented in Figure 2.9. From such illustration, we can address two relevant phenomena. In the first place, the modified score statistic $\tilde{W}_{Pu}(\phi)$ is shaped in such a way that we cannot explicitly obtain a lower bound for the confidence interval in the case $n = 20$. This also holds when considering the linear approximation for $\tilde{\lambda}_\phi$ since both curves reach a local maximum when ϕ is small enough.

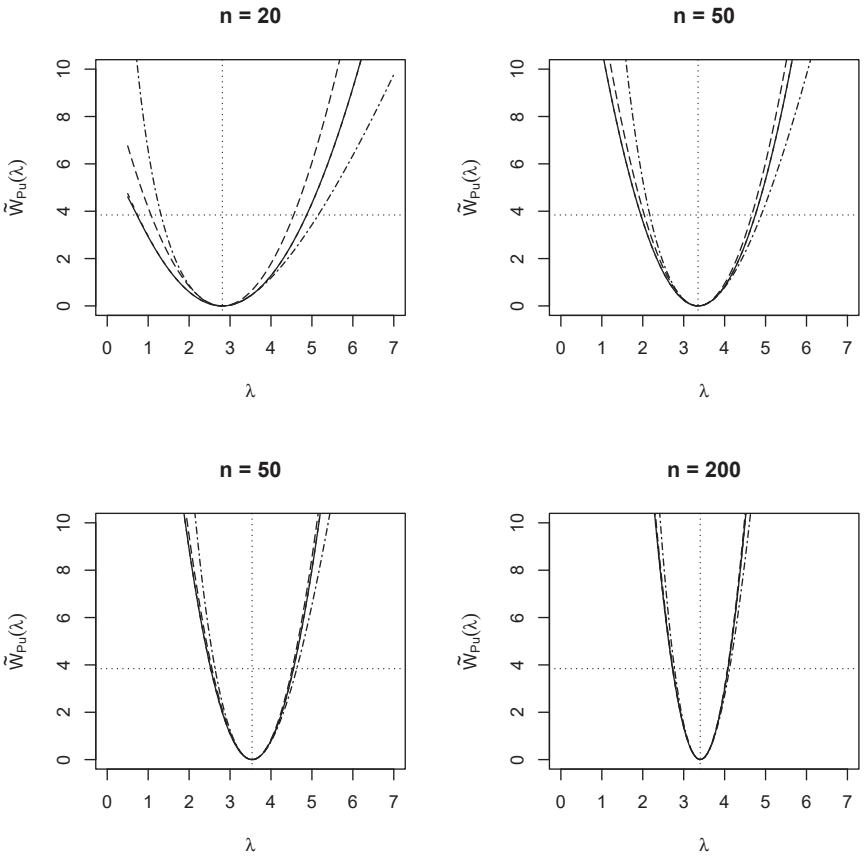

FIGURE 2.7: Modified profile score statistic (solid line) and corresponding approximation (dashed line) for λ in the canonical inverse Gaussian model, with respect to different sample sizes. In this case, the two curves are graphically indistinguishable. We also display $\tilde{W}_{Pe}(\lambda)$ (long-dashed line) and $\tilde{W}_P(\lambda)$ (dot-dashed line). The horizontal dotted line corresponds to the 0.95-quantile of a χ^2_1 distribution, while the vertical dotted line is the mean bias-reduced estimate $\tilde{\lambda}$.

In the second place, we see that the performance of the linear approximation of $\tilde{\lambda}_\phi$ degrades as ϕ gets closer to 0. Nonetheless, we also note that such discrepancy becomes irrelevant when n is relatively large, following from the fact that the modified score statistic becomes more concentrated around a neighborhood of $\tilde{\phi}$ where the approximation seems to be sufficiently accurate.

In Table 2.4, we numerically show the confidence intervals, with respect to the simulated data. Also in this case, even though it is only one simulated sample, all of them include the true value of the model parameters.

2.3.3 Gamma ratio

Let us consider n independent pairs $(y_{11}, y_{12}), \ldots, (y_{i1}, y_{i2}), \ldots (y_{n1}, y_{n2})$ as realizations of the random vector (Y_{i1}, Y_{i2}) such that $Y_{i1} \sim \text{Exp}(\psi\lambda)$, $Y_{i2} \sim \text{Exp}(\lambda)$, where the symbol

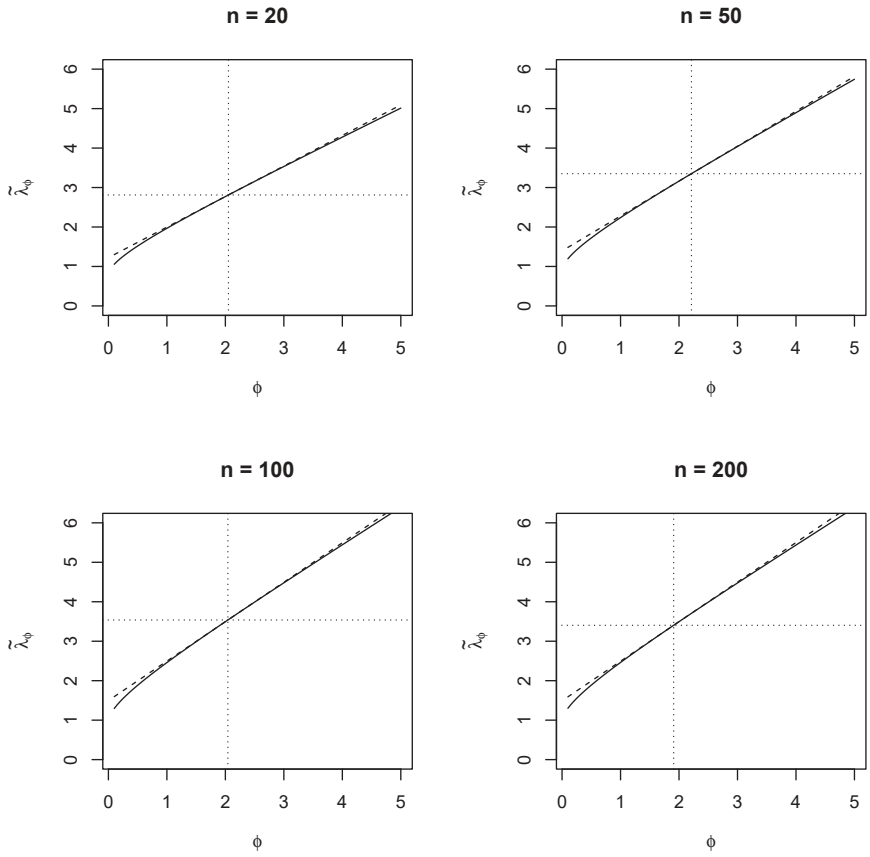

FIGURE 2.8: Constrained estimates of λ as functions of ϕ in the canonical inverse Gaussian model, with respect to different sample sizes. The solid line corresponds to the exact solution, whereas the dashed line shows the linear approximation. The dotted horizontal and vertical lines correspond to the global solution $\tilde{\theta}$.

$\text{Exp}(\omega)$ denotes an exponential distribution parameterized with rate $\omega > 0$, and Y_{i1} is independent of Y_{i2} for $i = 1, \ldots, n$. Then, we can define $Y_1 = \sum_{i=1}^n Y_{i1}$ and analogously $Y_2 = \sum_{i=1}^n Y_{i2}$, where for the well-known properties of the exponential distribution it follows that $Y_1 \sim \text{Ga}(n, \psi\lambda)$ and $Y_2 \sim \text{Ga}(n, \lambda)$. Such a model is parameterized by $\theta = (\psi, \lambda)$, with $\psi, \lambda > 0$.

This example is a simplified version of a more general problem, addressed for instance in Severini (1998, Example 4) and in the thesis of Emireni (2004, Section 3.2), in which the main issue is to carry out inference in the presence of several nuisance parameters. Dealing with such a problem is outside of the scope of the present example, where we aim at showing an application of the techniques of our interest in a relatively simple model.

Given that the parameter of interest is typically ψ, namely the ratio $E_\theta(Y_2)/E_\theta(Y_1)$, we refer to this model as "gamma ratio" for mere simplicity. Nevertheless, we carry out

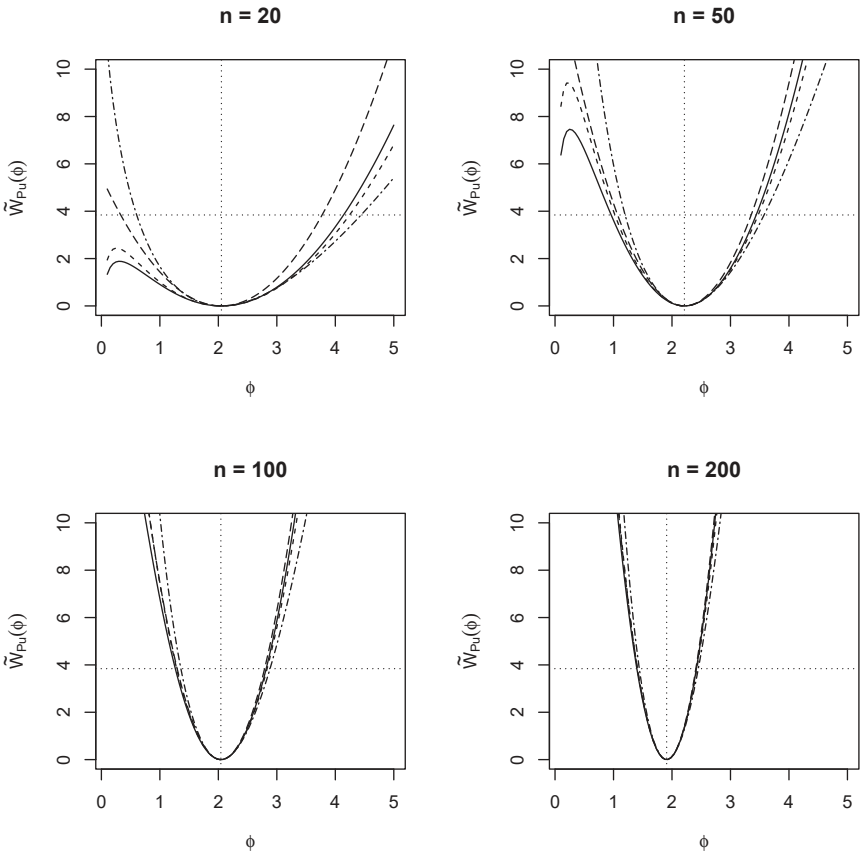

FIGURE 2.9: Modified profile score statistic (solid line) and corresponding approximation (dashed line) for ϕ in the canonical inverse Gaussian model, along with $\tilde{W}_{Pe}(\phi)$ (long-dashed line) and $\tilde{W}_P(\phi)$ (dot-dashed line), with respect to different sample sizes. The horizontal dotted line highlights the 0.95-quantile of a χ^2_1 distribution, while the vertical dotted line is the mean bias-reduced estimate $\tilde{\phi}$.

inference through the modified profile score with respect to both parameters, as shown in the previous examples.

Using the independence between Y_1 and Y_2, it is easy to express their joint density in the form

$$p(y_1, y_2; \theta) = \frac{\psi^n \lambda^{2n}}{\Gamma(n)^2} (y_1 y_2)^{(n-1)} \exp\left\{-(\psi\lambda y_1 + \lambda y_2)\right\},$$

from which we can obtain the log-likelihood, discarding additive constants, as

$$\ell(\theta) = n \log \psi + 2n \log \lambda - \psi\lambda y_1 - \lambda y_2.$$

From the previous expression or, equivalently, from the joint density, we note that the gamma ratio model is an exponential family with minimal sufficient statistic (y_1, y_2) and canonical parameter $\omega = \omega(\theta) = (-\psi\lambda, -\lambda)$. For this reason, θ is not the canonical parameter, hence in the case of mean bias reduction equality (1.22) does not hold and

TABLE 2.4: Confidence intervals obtained through the modified profile score statistic in the canonical inverse Gaussian model, for each parameter and with respect to different sample sizes. The symbol † indicates that it was not possible to explicitly determine the confidence limit, therefore 0 was assigned.

n	λ	ϕ
20	(0.7246, 4.8715)	($0^{\dagger}$, 4.1411)
50	(1.9523, 4.7496)	(0.9601, 3.4395)
100	(2.5272, 4.5498)	(1.264, 2.8183)
200	(2.7267, 4.0815)	(1.399, 2.4177)

we cannot base our inferential procedures on a penalized log-likelihood as in (1.23).

The score function can be expressed as

$$U(\theta) = \begin{bmatrix} U_{\psi}(\theta) \\ U_{\lambda}(\theta) \end{bmatrix} = \begin{bmatrix} n/\psi - \lambda y_1 \\ 2n/\lambda - \psi y_1 - y_2 \end{bmatrix},$$

from which we obtain the observed information matrix

$$j(\theta) = \begin{bmatrix} \frac{n}{\psi^2} & y_1 \\ y_1 & \frac{2n}{\lambda^2} \end{bmatrix}$$

and the expected information matrix

$$i(\theta) = E_{\theta}[j(\theta; Y_1, Y_2)] = \begin{bmatrix} \frac{n}{\psi^2} & \frac{n}{\psi\lambda} \\ \frac{n}{\psi\lambda} & \frac{2n}{\lambda^2} \end{bmatrix},$$

whose inverse can be easily expressed as

$$i(\theta)^{-1} = \begin{bmatrix} \frac{2\psi^2}{n} & -\frac{\psi\lambda}{n} \\ -\frac{\psi\lambda}{n} & \frac{\lambda^2}{n} \end{bmatrix}.$$

It is easy to solve the score equation $U(\theta) = 0$, leading to the solution $\hat{\theta} = (\hat{\psi}, \hat{\lambda}) = (y_2/y_1, n/y_2)$. Note that, unlike canonical exponential families, $j(\theta) \neq i(\theta)$ and therefore it is not guaranteed that the likelihood is globally concave. However, the log-likelihood is concave in $\hat{\theta}$, which follows from the fact that $|j(\hat{\theta})| = y_1 > 0$ and $n/\psi^2 > 0$, which suffices to prove that $j(\hat{\theta})$ is definite positive and that $\hat{\theta}$ is the maximum likelihood estimate of θ.

In order to obtain the adjusting vector for mean bias reduction, we need to explicitly use (1.20), namely

$$A(\theta) = \begin{bmatrix} A_\psi(\theta) \\ A_\lambda(\theta) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \text{trace}\,\{i(\theta)^{-1}[P_\psi(\theta) + Q_\psi(\theta)]\} \\ \text{trace}\,\{i(\theta)^{-1}[P_\lambda(\theta) + Q_\lambda(\theta)]\} \end{bmatrix},$$

where $P_\psi = E_\theta[U(\theta)U(\theta)^\top U_\psi(\theta)]$, $P_\lambda = E_\theta[U(\theta)U(\theta)^\top U_\lambda(\theta)]$, $Q_\psi(\theta) = E_\theta[-j(\theta)U_\psi(\theta)]$ and $Q_\lambda(\theta) = E_\theta[-j(\theta)U_\lambda(\theta)]$. Tedious but straightforward algebra, along with the properties $E[Y_1^j] = (n+j-1)!/[(n-1)!\psi^j\lambda^j]$ and $E[Y_2^j] = (n+j-1)!/[(n-1)!\lambda^j]$ for $j \in \{1,2,3\}$, allows us to obtain

$$\begin{aligned} P_\psi(\theta) &= E_\theta \begin{bmatrix} (\frac{n}{\psi} - \lambda Y_1)^3 & (\frac{n}{\psi} - \lambda Y_1)^2(\frac{2n}{\lambda} - \psi Y_1 - Y_2) \\ (\frac{n}{\psi} - \lambda Y_1)^2(\frac{2n}{\lambda} - \psi Y_1 - Y_2) & (\frac{n}{\psi} - \lambda Y_1)(\frac{2n}{\lambda} - \psi Y_1 - Y_2)^2 \end{bmatrix} \\ &= \begin{bmatrix} -\frac{2n}{\psi^3} & -\frac{2n}{\psi^2\lambda} \\ -\frac{2n}{\psi^2\lambda} & -\frac{2n}{\psi\lambda^2} \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} P_\lambda(\theta) &= E_\theta \begin{bmatrix} (\frac{n}{\psi} - \lambda Y_1)^2(\frac{2n}{\lambda} - \psi Y_1 - Y_2) & (\frac{n}{\psi} - \lambda Y_1)(\frac{2n}{\lambda} - \psi Y_1 - Y_2)^2 \\ (\frac{n}{\psi} - \lambda Y_1)(\frac{2n}{\lambda} - \psi Y_1 - Y_2)^2 & (\frac{2n}{\lambda} - \psi Y_1 - Y_2)^3 \end{bmatrix} \\ &= \begin{bmatrix} -\frac{2n}{\psi^2\lambda} & -\frac{2n}{\psi\lambda^2} \\ -\frac{2n}{\psi\lambda^2} & -\frac{4n}{\lambda^3} \end{bmatrix}. \end{aligned}$$

Analogously, the remaining matrices $Q_\psi(\theta)$ and $Q_\lambda(\theta)$ can be expressed as

$$\begin{aligned} Q_\psi(\theta) &= -E_\theta \begin{bmatrix} \frac{n}{\psi^2}(\frac{n}{\psi} - \lambda Y_1) & Y_1(\frac{n}{\psi} - \lambda Y_1) \\ Y_1(\frac{n}{\psi} - \lambda Y_1) & \frac{2n}{\lambda^2}(\frac{n}{\psi} - \lambda Y_1) \end{bmatrix} \\ &= \begin{bmatrix} 0 & \frac{n}{\psi^2\lambda} \\ \frac{n}{\psi^2\lambda} & 0 \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} Q_\lambda(\theta) &= -E_\theta \begin{bmatrix} \frac{n}{\psi^2}(\frac{2n}{\lambda} - \psi Y_1 - Y_2) & Y_1(\frac{2n}{\lambda} - \psi Y_1 - Y_2) \\ Y_1(\frac{2n}{\lambda} - \psi Y_1 - Y_2) & \frac{2n}{\lambda^2}(\frac{2n}{\lambda} - \psi Y_1 - Y_2) \end{bmatrix} \\ &= \begin{bmatrix} 0 & \frac{n}{\psi\lambda^2} \\ \frac{n}{\psi\lambda^2} & 0 \end{bmatrix}. \end{aligned}$$

Such quantities allow us to express

$$\begin{aligned} A_\psi(\theta) &= \frac{1}{2}\text{trace}\left\{ \begin{bmatrix} \frac{2\psi^2}{n} & -\frac{\psi\lambda}{n} \\ -\frac{\psi\lambda}{n} & \frac{\lambda^2}{n} \end{bmatrix} \begin{bmatrix} -\frac{2n}{\psi^3} & -\frac{n}{\psi^2\lambda} \\ -\frac{n}{\psi^2\lambda} & -\frac{2n}{\psi\lambda^2} \end{bmatrix} \right\} \\ &= \frac{1}{2}\left(-\frac{4\psi^2}{n}\frac{n}{\psi^3} + \frac{\psi\lambda}{n}\frac{n}{\psi^2\lambda} + \frac{\psi\lambda}{n}\frac{n}{\psi^2\lambda} - \frac{\lambda^2}{n}\frac{2n}{\psi\lambda^2} \right) \\ &= -\frac{2}{\psi} \end{aligned}$$

and analogously

$$\begin{aligned} A_\lambda(\theta) &= \frac{1}{2}\text{trace}\left\{ \begin{bmatrix} \frac{2\psi^2}{n} & -\frac{\psi\lambda}{n} \\ -\frac{\psi\lambda}{n} & \frac{\lambda^2}{n} \end{bmatrix} \begin{bmatrix} -\frac{2n}{\psi^2\lambda} & -\frac{n}{\psi\lambda^2} \\ -\frac{n}{\psi\lambda^2} & -\frac{4n}{\lambda^3} \end{bmatrix} \right\} \\ &= \frac{1}{2}\left(-\frac{4\psi^2}{n}\frac{n}{\psi^2\lambda} + \frac{\psi\lambda}{n}\frac{n}{\psi\lambda^2} + \frac{\psi\lambda}{n}\frac{n}{\psi\lambda^2} - \frac{\lambda^2}{n}\frac{4n}{\lambda^3} \right) \\ &= -\frac{3}{\lambda}. \end{aligned}$$

The mean bias reduced estimate $\tilde{\theta} = (\tilde{\psi}, \tilde{\lambda})$ can be obtained by solving the modified score equation

$$\tilde{U}(\theta) = \begin{bmatrix} \tilde{U}_\psi(\theta) \\ \tilde{U}_\lambda(\theta) \end{bmatrix} = \begin{bmatrix} \frac{n}{\psi} - \lambda y_1 - \frac{2}{\psi} \\ \frac{2n}{\lambda} - \psi y_1 - y_2 - \frac{3}{\lambda} \end{bmatrix},$$

whose solution can be analytically expressed as

$$\tilde{\theta} = \begin{bmatrix} \tilde{\psi} \\ \tilde{\lambda} \end{bmatrix} = \begin{bmatrix} \frac{n-2}{n-1}\frac{y_2}{y_1} \\ \frac{n-1}{y_2} \end{bmatrix}.$$

We simulate the pseudo-observation (y_1, y_2) following the gamma ratio model with $\theta = (3, 7)$. Furthermore, we also consider different $n \in \{20, 50, 100, 200\}$, which can be considered either as the sample size or as an index that quantifies the amount of information available from the data. With respect to the generated data, we show in Table 2.5 the maximum likelihood and the mean bias reduced estimates of the model parameters. We see that the estimates for ψ improve as n increases and also that $\tilde{\psi}$ performs slightly better than $\hat{\psi}$, nonetheless this is evidently not the case if we consider the estimates for λ with respect to $n < 200$.

Focusing our attention on the modified score statistic $\tilde{W}_u(\theta)$, we are able to obtain the resulting 95% confidence region of the model parameters, as illustrated in Figure 2.10, where we see that increasing n concentrates the modified score statistic around the mean bias-reduced estimate. Unlike the previous examples, in this case the non-elliptical

TABLE 2.5: Maximum likelihood and mean bias-reduced estimates of the gamma ratio model with respect to different sample sizes.

	$n = 20$	$n = 50$	$n = 100$	$n = 200$
$\hat{\psi}$	4.4043	3.9461	2.6627	2.8981
$\hat{\lambda}$	5.5755	6.9425	6.7210	7.9614
$\tilde{\psi}$	4.1725	3.8656	2.6358	2.8835
$\tilde{\lambda}$	5.2967	6.8037	6.6538	7.9216

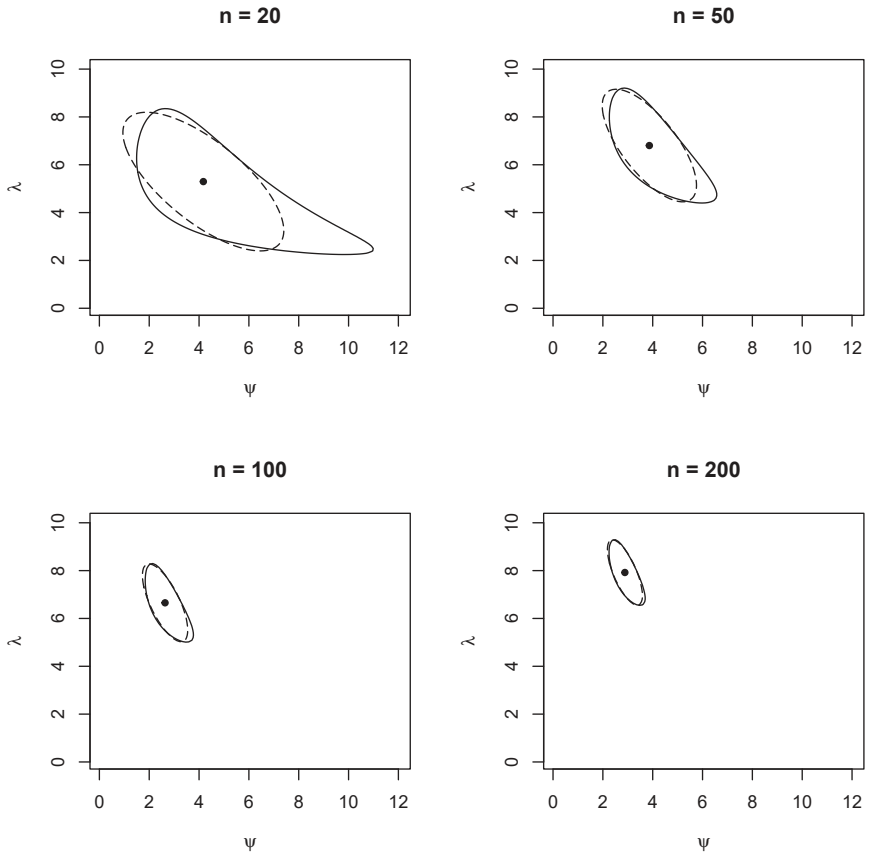

FIGURE 2.10: Approximate 95% confidence regions of θ obtained through $\tilde{W}_u(\theta)$ (solid line) and $\tilde{W}_e(\theta)$ (long-dashed line) for the gamma ratio model, with respect to different sample sizes. The dot corresponds to the mean bias-reduced estimates.

shape of the confidence regions is much more appreciable, however as n increases such feature seems to vanish, resembling Wald-type elliptical confidence regions. This is a quite expected outcome since $\tilde{W}_u(\theta)$ is asymptotically equivalent to $\tilde{W}_e(\theta)$, therefore both approximate pivots should lead to analogous results as n diverges.

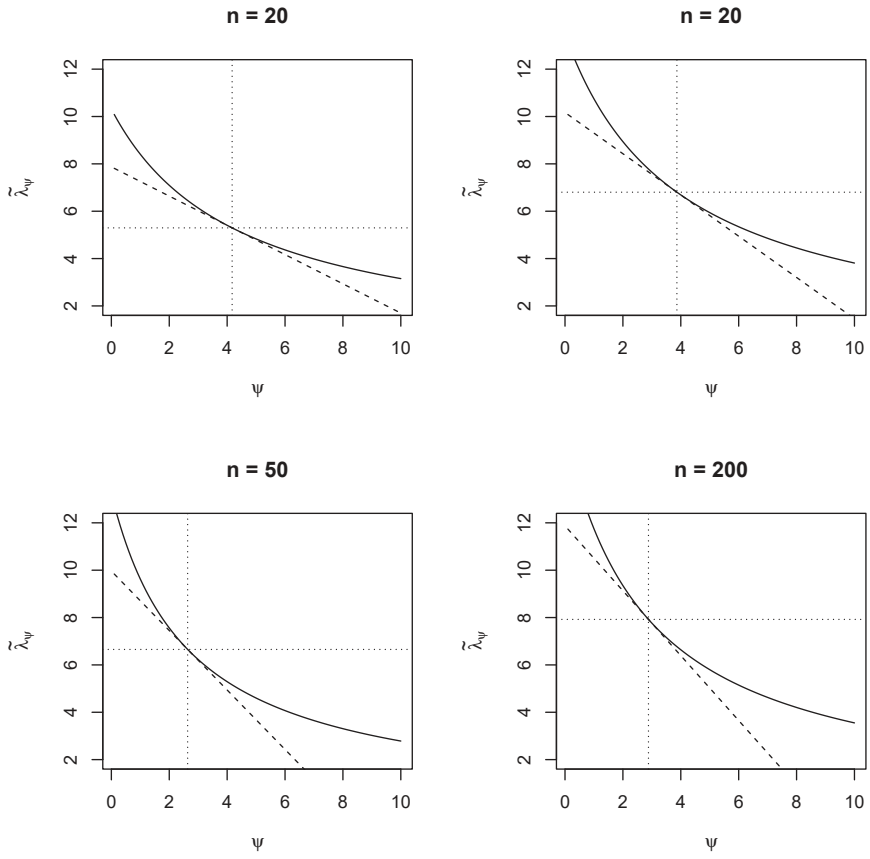

FIGURE 2.11: Constrained estimates of λ as functions of ψ in the gamma ratio model, with respect to different sample sizes. The solid line corresponds to the exact solution, whereas the dashed line shows the linear approximation. The dotted horizontal and vertical lines correspond to the global solution $\tilde{\theta}$.

The constrained estimate of λ given ψ is readily available as

$$\tilde{\lambda}_\psi = \frac{2n-3}{\psi y_1 + y_2},$$

which is a nonlinear function of ψ, as shown in Figure 2.11. Despite the availability of a closed-form expression for $\tilde{\lambda}_\psi$, we also consider the corresponding linear approximation. In this case, there is a noticeable discrepancy between the two curves in each panel, which suggests inadequacy of the approximation for values of ψ that are too far from the bias-reduced estimate $\tilde{\psi}$.

The effect of such discrepancy becomes more apparent when considering the modified profile score statistic for ψ, which is illustrated in Figure 2.12. Quite clearly, the shape of $\tilde{W}_{Pu}(\psi)$ computed through the linear approximation of $\tilde{\lambda}_\psi$ fails to yield a reliable result, with respect to the exact version of the statistic, considering even a relatively large sample size such as $n = 100$.

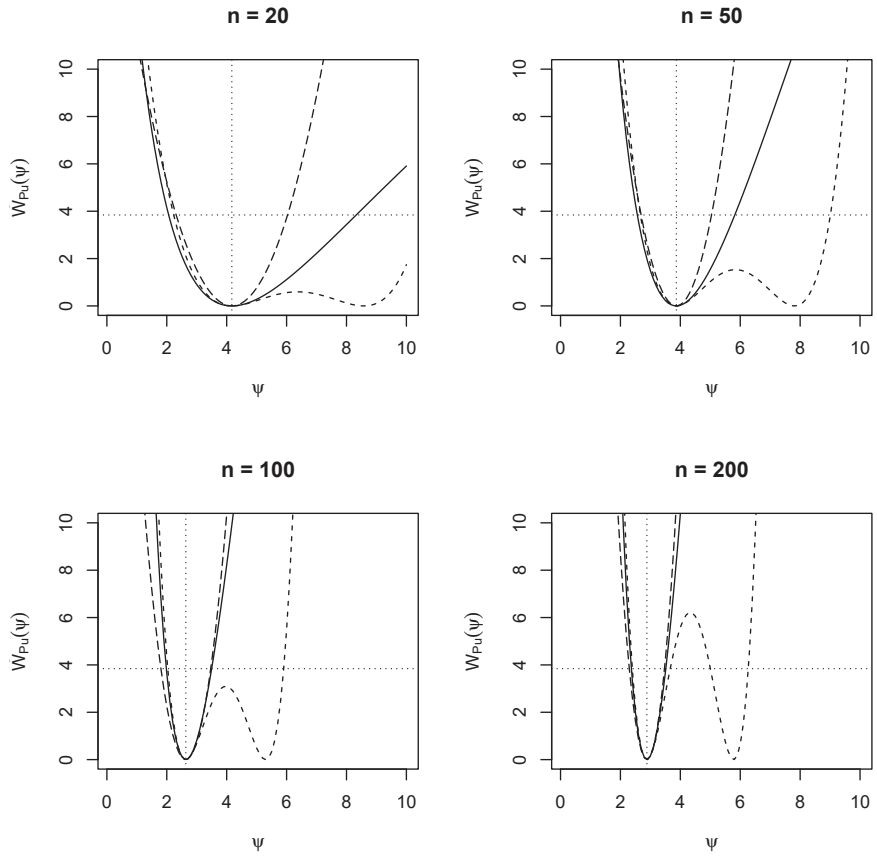

FIGURE 2.12: Modified profile score statistic (solid line) and corresponding approximation (dashed line) for ψ in the gamma ratio model, along with $\tilde{W}_{Pe}(\psi)$ (long-dashed line), with respect to different sample sizes. The horizontal dotted line highlights the 0.95-quantile of a χ^2_1 distribution, while the vertical dotted line is the mean bias-reduced estimate $\tilde{\psi}$.

In accordance with our observations regarding the global confidence regions, the confidence intervals resulting from $\tilde{W}_{Pu}(\psi)$ are clearly asymmetrical for small to modest sample size, namely $n = 20$ and $n = 50$, whereas for larger n such asymmetry becomes less apparent.

Considering λ as the parameter of interest, we obtain quite similar results with respect to the constrained estimate $\tilde{\psi}_\lambda = (n-2)/(\lambda y_1)$, shown in Figure 2.13. Analogously, the shape of $\tilde{W}_{Pu}(\lambda)$, illustrated in Figure 2.14 appreciably differs from the corresponding approximated version. Therefore, once again the linear approximation of the constrained estimate yields unsatisfactory results, in this case at least for $n = 50$.

In contrast to the confidence intervals of ψ, those of λ do not show appreciable asymmetry, even considering a relatively small sample size.

In Table 2.6, we show the numerical results as regards the approximate 95% confidence intervals of both ψ and λ, taking into consideration different sample sizes. As in

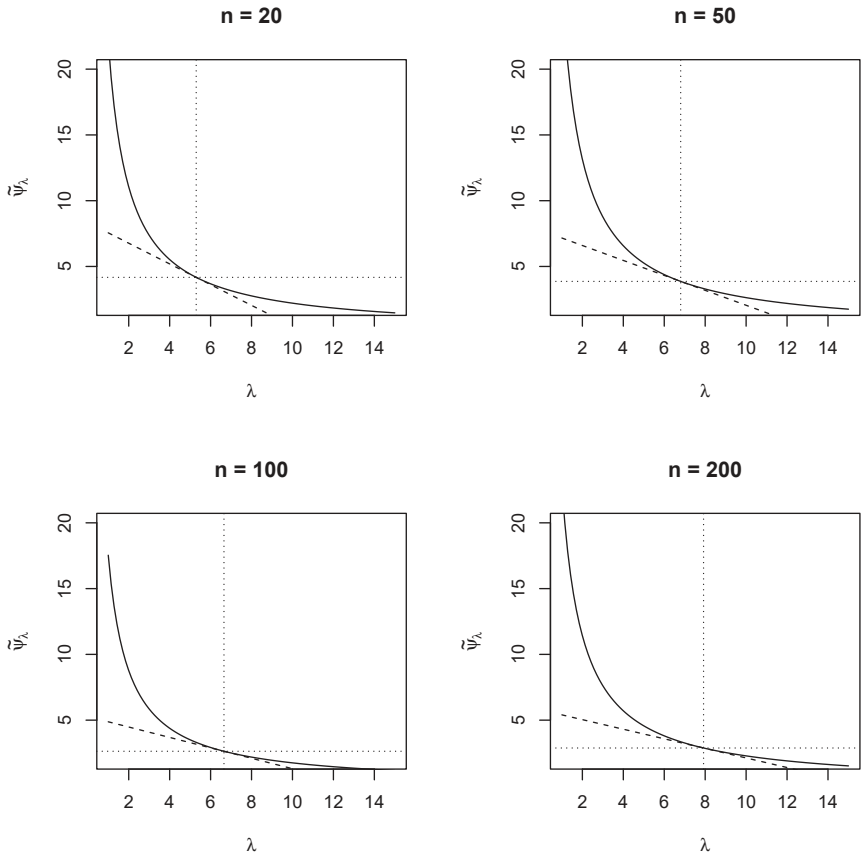

FIGURE 2.13: Constrained estimates of ψ as functions of λ in the gamma ratio model, with respect to different sample sizes. The solid line corresponds to the exact solution, whereas the dashed line shows the linear approximation. The dotted horizontal and vertical lines correspond to the global solution $\tilde{\theta}$.

the previous examples, also in this simulated case all the confidence intervals include the true value of the parameters.

TABLE 2.6: Confidence intervals obtained through the modified profile score statistic in the gamma ratio model, for each parameter and with respect to different sample sizes.

n	ψ	λ
20	(2.0629, 8.3249)	(2.8532, 7.7402)
50	(2.5637, 5.8185)	(4.8793, 8.728)
100	(1.9851, 3.4983)	(5.3365, 7.9711)
200	(2.3652, 3.5152)	(6.8182, 9.0249)

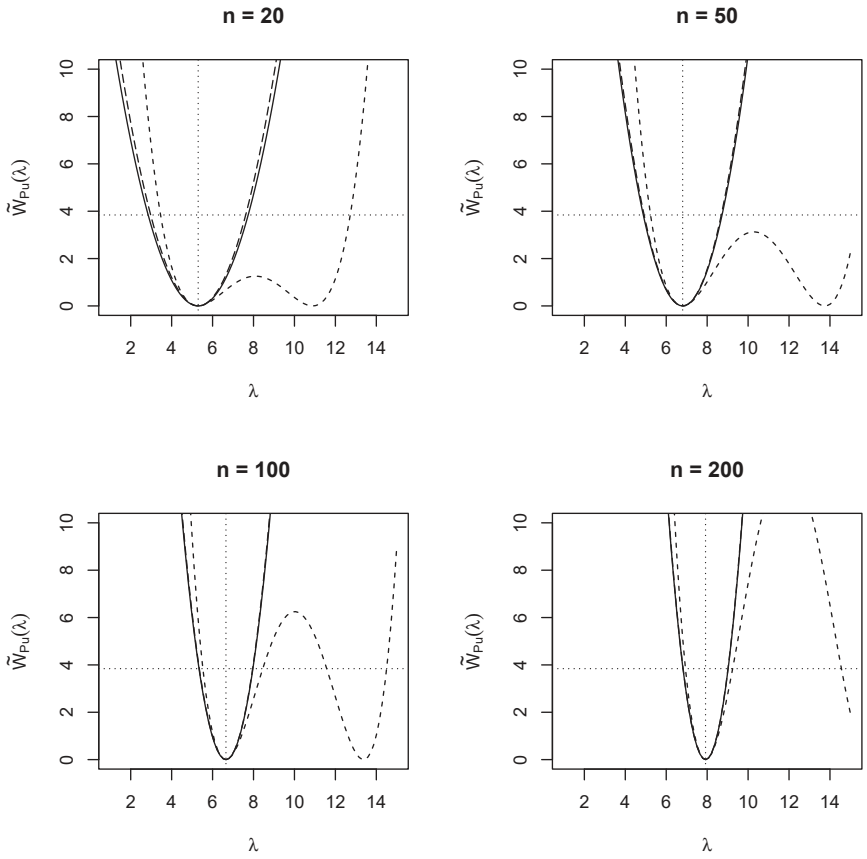

FIGURE 2.14: Modified profile score statistic (solid line) and corresponding approximation (dashed line) for λ in the gamma ratio model, along with $\tilde{W}_{Pe}(\lambda)$ (long-dashed line), with respect to different sample sizes. The horizontal dotted line highlights the 0.95-quantile of a χ^2_1 distribution, while the vertical dotted line is the mean bias-reduced estimate $\tilde{\lambda}$.

2.3.4 Logistic regression model

In this section, we illustrate two different examples related to the logistic regression model. In the first case, we consider a simulated n-dimensional binary response with respect to a $(n \times 2)$ fixed design matrix X, whose columns are obtained by simulating from two independent standard Gaussian distributions. Such a case consists of a logistic regression model with two covariates and no intercept. As a second example, we illustrate a case study involving a true data set, which is of our particular interest mainly due to the presence of quasi-complete data separation.

Let us consider n observations $y_1, \ldots, y_n$ and let $y = (y_1, \ldots, y_n)$. With respect to a fixed $(p \times n)$ design matrix denoted by X, let us assume that y is generated by a random

vector $Y = (Y_1, \ldots, Y_n)$ whose joint density is

$$p(y; \beta) = \prod_{i=1}^{n} \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i},$$

where $\pi_i = E_\beta[Y_i] = g^{-1}(x_i\beta)$, $g(\cdot)$ is a smooth one-to-one link function, x_i is the i-th row of X and $\beta = (\beta_1, \ldots, \beta_p)$ is a vector of regression parameters in $\mathbb{R}^p$. As far as our examples are concerned, we assume that $m_i = 1$ for all $i = 1, \ldots, n$ and that $g(\pi) = \log[\pi/(1-\pi)]$, with $\pi \in (0, 1)$, is the canonical logit link.

The log-likelihood of the model can be easily written as

$$\ell(\beta) = \sum_{i=1}^{n} \{y_i \log[\pi_i/(1 - \pi_i)] + \log(1 - \pi_i)\},$$

from which we derive the score function, expressed in matrix notation as

$$U(\beta) = X^\top (y - \mu),$$

where $\mu = (\pi_1, \ldots, \pi_n)$, and the observed information matrix

$$j(\theta) = X^\top W X,$$

where $W = \text{diag}(w_1, \ldots, w_n)$, with $w_i = \pi_i(1 - \pi_i)$ for $i = 1, \ldots, n$. Note that such model is once again a canonical exponential family, where the canonical parameter is given by β. For this reason, the mean bias-reduced estimate $\tilde{\beta}$ of β can be obtained by maximizing the penalized log-likelihood

$$\tilde{\ell}(\beta) = \ell(\beta) + \frac{1}{2} \log|i(\beta)|.$$

Such interpretation of the mean bias-reduced estimate is quite useful because $\log|i(\beta)|$ is strictly concave, provided that X has full rank. For this reason, it follows that $\tilde{\ell}(\beta)$ has a unique global maximum $\tilde{\beta}$, as shown for example in Firth (1993, Section 3.3) and in Kosmidis & Firth (2020, Section 2). Among other useful quantities with respect to exponential dispersion families, Kosmidis et al. (2020) provides the expression of the adjusting vector for mean bias reduction, which in our case corresponds to

$$A(\beta) = X^\top \xi,$$

where $\xi = (\xi_1, \ldots, \xi_n)$, $\xi_i = h_i(1 - 2\pi_i)/2$ and h_i is the i-th diagonal element of the "hat" matrix

$$H = X(X^\top W X)^{-1} X^\top W.$$

Therefore, it follows that the modified score statistic for β can be written as

$$\tilde{U}(\beta) = X^\top (y - \mu + \xi),$$

and the modified score equation $\tilde{U}(\beta) = 0$ is a nonlinear equation in β with unique solution given by $\tilde{\beta}$. We note that $\tilde{\beta}$ generally needs to be found numerically, for instance by means of the quasi-Fisher algorithm (1.24), with an exception for special cases in which $\tilde{\beta}$ has explicit solution (see Firth, 1993, page 31).

Considering the first example of a logistic regression with two covariates and no intercept and denoting as x_{ij} the generic element of X, with $i = 1, \ldots, n$ and $j = 1, 2$, we can express the score function as

$$U(\theta) = \begin{bmatrix} U_{\beta_1}(\beta) \\ U_{\beta_2}(\beta) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i x_{i1} - \sum_{i=1}^n x_{i1}\pi_i \\ \sum_{i=1}^n y_i x_{i2} - \sum_{i=1}^n x_{i2}\pi_i \end{bmatrix}$$

and the observed information matrix as

$$j(\beta) = i(\beta) = \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 \pi_i(1-\pi_i) & \sum_{i=1}^n x_{i1}x_{i2}\pi_i(1-\pi_i) \\ \sum_{i=1}^n x_{i1}x_{i2}\pi_i(1-\pi_i) & \sum_{i=1}^n x_{i2}^2 \pi_i(1-\pi_i) \end{bmatrix}.$$

The adjusting vector for mean bias reduction can be obtained by using (1.22) due to the canonical parameterization, namely

$$A(\beta) = \frac{1}{2}|i(\beta)|^{-1} \begin{bmatrix} a_1(\beta) \\ a_2(\beta) \end{bmatrix}$$

where, defining for compactness of notation $\nu_i = \pi_i(1-\pi_i)(1-2\pi_i)$ and $\zeta_i = \pi_i(1-\pi_i)$, we can write

$$\begin{aligned} a_j(\beta) &= \frac{\partial}{\partial \beta_j}|i(\beta)| \\ &= \sum_{i=1}^n x_{i1}^2 x_{ij}\nu_i \sum_{i=1}^n x_{i2}^2 \zeta_i + \sum_{i=1}^n x_{i2}^2 x_{ij}\nu_i \sum_{i=1}^n x_{i1}^2 \zeta_i \\ &\quad - 2\sum_{i=1}^n x_{i1}x_{i2}x_{ij}\nu_i \sum_{i=1}^n x_{i1}x_{i2}\zeta_i \end{aligned}$$

for $j = 1, 2$.

To provide an illustration of this case, we first simulate a (200×2) design matrix X from a bivariate standard Gaussian distribution with independent components. Then, we consider the corresponding first n rows, with $n = 20, 50, 100, 200$ and a parameter $\beta = (\beta_1, \beta_2) = (-0.3, 0.4)$ from which we simulate the binary response y for each n.

From the simulated data, we obtain the maximum likelihood and mean bias-reduced estimates provided in Table 2.7. In particular, we note that the mean bias-reduced esti-

TABLE 2.7: Maximum likelihood and mean bias-reduced estimates of the biparametric logistic regression with respect to different sample sizes.

	$n = 20$	$n = 50$	$n = 100$	$n = 200$
$\hat{\beta}_1$	-0.4534	-0.3387	-0.3349	-0.4178
$\hat{\beta}_2$	0.7775	0.9037	0.5895	0.5123
$\tilde{\beta}_1$	-0.3877	-0.3206	-0.3238	-0.4097
$\tilde{\beta}_2$	0.6613	0.8391	0.5676	0.5031

mates are shrunk towards 0, which for the simulated sample trajectory results in slightly more accurate estimates with respect to the true parameter value. Such shrinkage property is a well-known aspect of mean bias reduction and of the penalization based on the Jeffreys prior for binomial logistic regression (Firth, 1993, Section 3.3).

By means of the aforementioned relevant quantities, we are able to obtain the modified score statistic $\tilde{W}_u(\beta)$ and to construct confidence regions, an illustration of which is provided in Figure 2.15 with respect to a 0.95 approximate confidence level. We can observe that the confidence regions shrink around $\tilde{\beta}$ and that they seem more elliptically-shaped as n increases, in accordance with the results of the previous examples and with the asymptotic equivalence to the modified Wald statistic $\tilde{W}_e(\beta)$ (and $\tilde{W}(\beta)$).

Let us focus our attention on the regression parameter β_2, while in contrast considering β_1 as a nuisance parameter. Then, denoting the constrained bias-reduced estimate of β_1 given β_2 as $\tilde{\beta}_1(\beta_2)$ to avoid an excessive clutter due to subscripts, we obtain such estimate by numerically solving the β_1-related adjusted equation through the Newton-Raphson iterations (2.9). We provide an illustration of $\tilde{\beta}_1(\beta_2)$ in Figure 2.16, along with the respective linear approximation.

Clearly, $\tilde{\beta}_1(\beta_2)$ is a nonlinear function of β_2, resulting in a discrepancy between its exact and approximated version. However, as shown in Figure 2.17, such difference does not seem to dramatically reflect on the approximation of the modified score statistic $\tilde{W}_{Pu}(\beta_2)$, even when considering $n = 20$. Moreover, in accordance with the global confidence regions, we can notice that, for instance considering $n = 20$, $\tilde{W}_{Pu}(\beta_2)$ follows

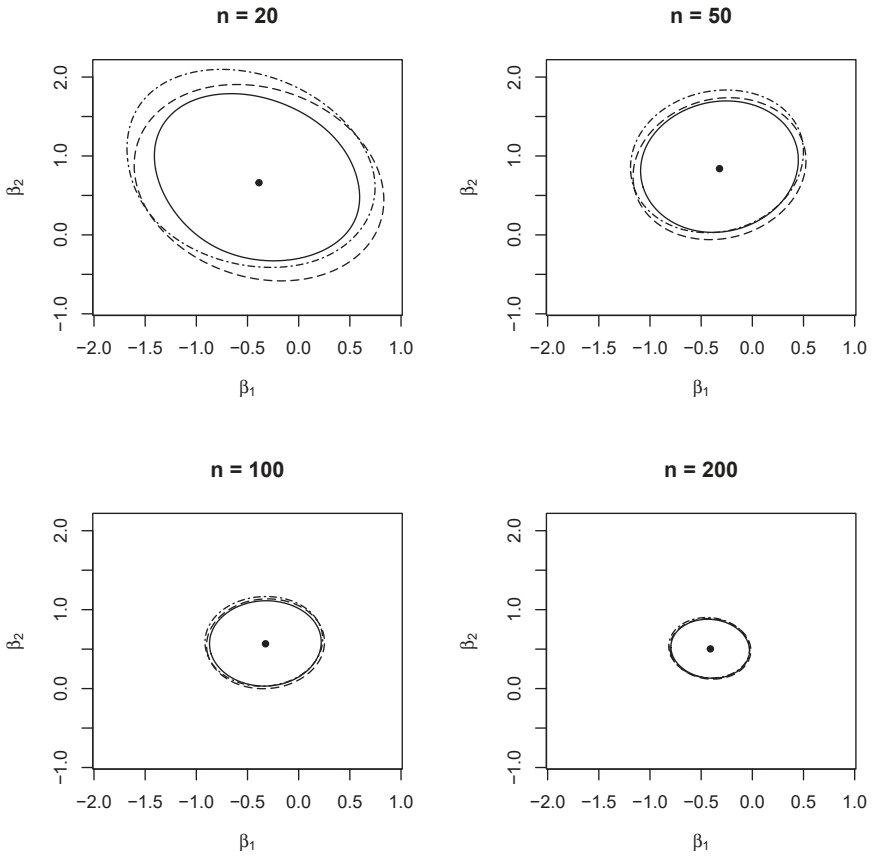

FIGURE 2.15: Approximate 95% confidence regions of β obtained through $\tilde{W}_u(\beta)$ (solid line), $\tilde{W}_e(\beta)$ (long-dashed line) and $\tilde{W}(\beta)$ (dot-dashed line) for the biparametric logistic regression model, with respect to different sample sizes. The dot corresponds to the mean bias-reduced estimates.

a slightly asymmetric shape around $\tilde{\beta}_2$, which may be regarded as a desirable property for confidence intervals.

In order to avoid redundancy, we do not show the converse case, namely the situation of β_1 as parameter of interest. Indeed, in this example we may regard the roles of the two regression parameters as equivalent, given the structure of the model matrix.

Nonetheless, for completeness of discussion we show in Table 2.8 the approximate 95% confidence intervals obtained through the modified profile score statistic for β_1 and β_2. Also in this case, the confidence intervals include the true value of the parameters, but we can also notice that the effect of the covariates may be regarded as non-significant, especially for small or modest n.

Let us now carry out our analysis on the `endometrial` data set, available for instance in the `brglm2` package. Such data were provided by Dr E. Asseryanis from the Medical University of Vienna and were first analyzed in the work of Heinze & Schemper (2002), to which we refer for a more thorough description.

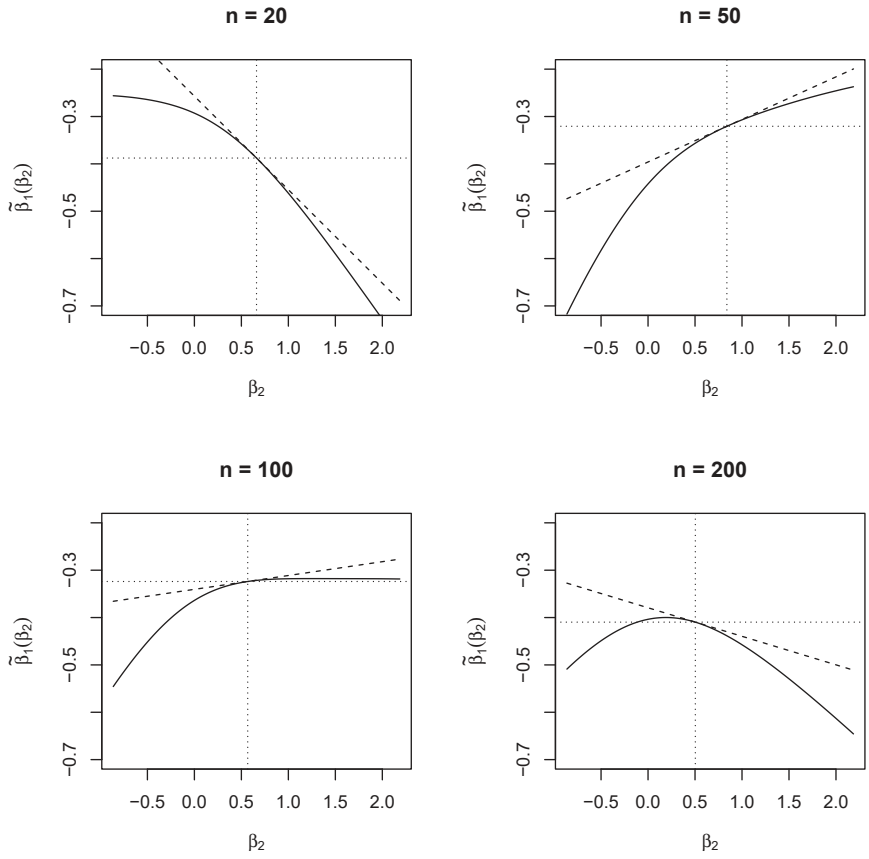

FIGURE 2.16: Constrained estimates of β_1 as functions of β_2 in the biparametric logistic regression model, with respect to different sample sizes. The solid line corresponds to the exact solution, whereas the dashed line shows the linear approximation. The dotted horizontal and vertical lines correspond to the global solution $\tilde{\beta}$.

The data set consists of $n = 79$ statistical units, corresponding to subjects diagnosed with endometrial cancer. For each statistical unit, four variables are measured. In the first place, each subject is classified with respect to histology (HG), which is a binary variable such that 0 corresponds to grade 0-II and 1 denotes grade III-IV of endometrial cancer. A second variable is given by neovascularization (NV), which is dichotomous and coded as 0 and 1 to respectively denote absence and presence of neovascularization for each subject. A third variable measures the pulsatility index of arteria uterina (PI), which is a continuous variable ranging from 0 to 49 with mean 17.38 and standard deviation 9.93. The fourth and last variable is given by the endometrium height (EH), which is continuous, ranging from 0.27 to 3.61 and with mean 1.66 and standard deviation 0.66.

As in Heinze & Schemper (2002), we carry out our analysis considering HG as the response variable and NV, PI, EH as risk factors. In particular, we specify the logistic

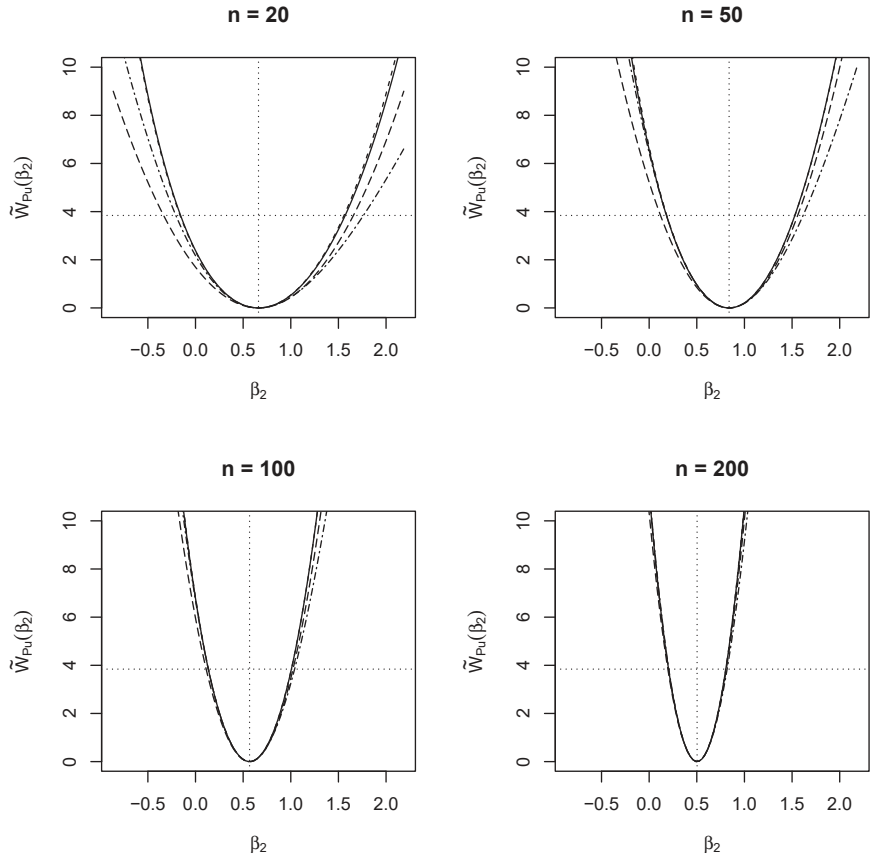

FIGURE 2.17: Modified profile score statistic (solid line) and corresponding approximation (dashed line) for β_2 in the biparametric logistic regression model, along with $\tilde{W}_{Pe}(\beta_2)$ (long-dashed line) and $\tilde{W}_P(\beta_2)$ (dot-dashed line), with respect to different sample sizes. The horizontal dotted line corresponds to the 0.95-quantile of a χ^2_1 distribution, while the vertical dotted line is the mean bias-reduced estimate $\tilde{\beta}_2$.

regression model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_{\text{NV}} x_{i1} + \beta_{\text{PI}} x_{i2} + \beta_{\text{EH}} x_{i3}, \quad i = 1, \ldots, 79, \tag{2.12}$$

with respect to the parameter vector $\beta = (\beta_0, \beta_{\text{NV}}, \beta_{\text{PI}}, \beta_{\text{EH}})$ and to the model matrix X, whose rows are denoted by $x_i = (1, x_{i1}, x_{i2}, x_{i3})$.

To briefly illustrate the relationship between the response variable and the aforementioned risk factors, we show in Figure 2.18 that HG = 1 seems to be associated with a visibly lower endometrium height and to a slightly lower pulsatility index. We place more emphasis on the results shown in Table 2.9, where we note that no statistical unit such that HG = 0 is associated with NV = 1, which causes quasi-complete linear separation in the data. As illustrated in Agresti (2015, section 5.7.1), the maximum likelihood estimate of β_{NV} is infinite and the corresponding confidence interval based

TABLE 2.8: Confidence intervals obtained through the modified profile score statistic in the biparametric logistic regression model, for each parameter and with respect to different sample sizes.

n	β_1	β_2
20	(-1.2174, 0.4143)	(-0.1568, 1.5711)
50	(-0.9444, 0.3013)	(0.1813, 1.5293)
100	(-0.7649, 0.1148)	(0.1345, 1.0065)
200	(-0.7196, -0.1019)	(0.2032, 0.8047)

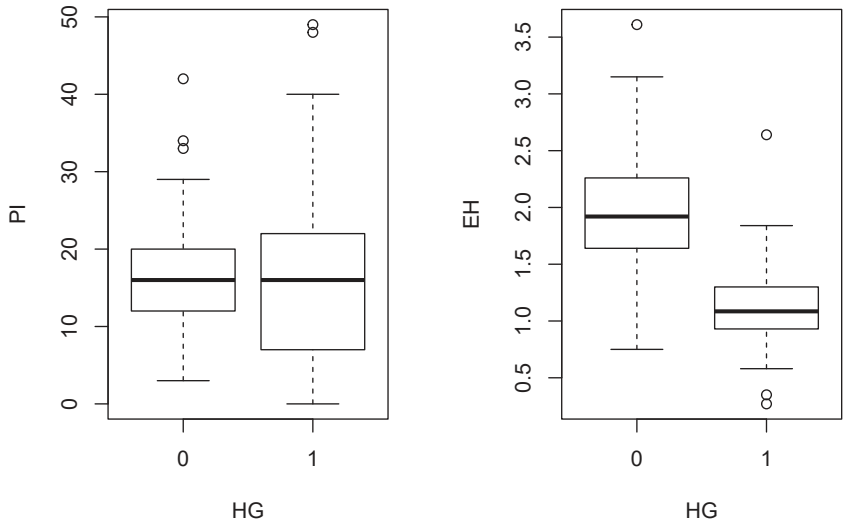

FIGURE 2.18: Boxplots of PI (left) and EH (right), illustrated conditionally on HG, with respect to the `endometrial` data set.

on the profile likelihood ratio test (1.14) has no upper bound. Moreover, Wald-type inference is unfeasible due to the non-finiteness of the maximum likelihood estimate.

TABLE 2.9: Absolute frequency contingency table of HG and NV in the `endometrial` data set.

	NV = 0	NV = 1
HG = 0	49	0
HG = 1	17	13

Quite the contrary, applying mean bias-reduced estimation to such problem allows us to obtain finite and therefore interpretable estimates, along with Wald-type confidence

intervals based on (2.4), as shown in Table 2.10.

TABLE 2.10: Summary of the mean bias-reduced parameter estimates for the `endometrial` data and approximate 95% Wald-type confidence intervals.

	Estimate	Standard error	Confidence interval
β_0	3.7746	1.4887	(0.8568, 6.6923)
β_{NV}	2.9293	1.5508	(-0.1102, 5.9687)
β_{PI}	-0.0348	0.0396	(-0.1123, 0.0428)
β_{EH}	-2.6042	0.7760	(-4.1251, -1.0832)

Besides, as displayed in Figure 2.19 the profile score for β_{NV} follows a horizontal asymptote that prevents from finding a finite solution to the corresponding equation. In contrast, we can see that mean bias reduction results in a correction to the profile score function that allows to find a finite estimate.

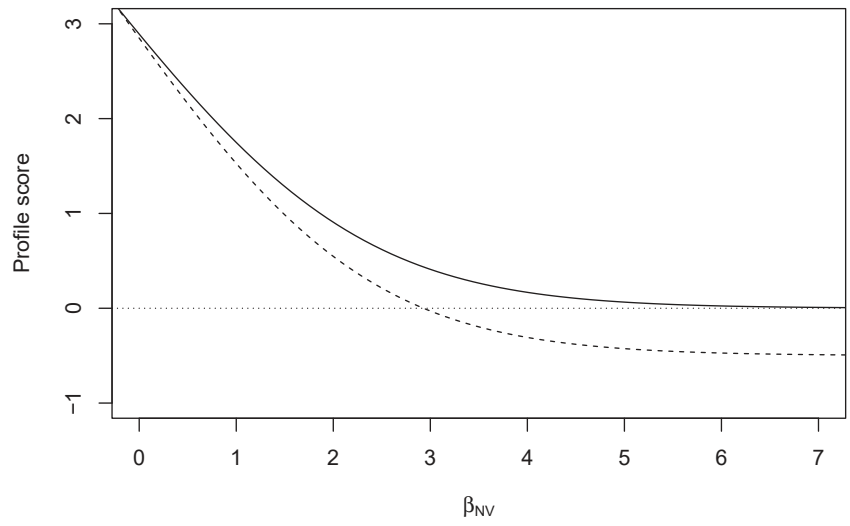

FIGURE 2.19: Profile score function (solid line) and modified profile score function (dashed line) for β_{NV}, with respect to the `endometrial` data.

Nonetheless, if we base our inferential analysis on Wald-type inference (2.4) and provided that we use a 0.05 or a lower significance level, we may regard the effect of NV as non-significant, since the corresponding confidence interval includes 0. In the work of Heinze & Schemper (2002), this particular problem is addressed by using the modified profile likelihood ratio statistic (2.5), which allows to obtain confidence regions that take into account the particular shape of the penalized likelihood (1.23). Indeed, the

penalized profile log-likelihood for β_{NV}, displayed in the top-right panel in Figure 2.20, has a clearly skewed shape that may not be well characterized by the quadratic shape of $\tilde{W}_{Pe}(\beta_{\text{NV}})$. This observation seems to hold considering the other model coefficients as well, as illustrated in Figure 2.20.

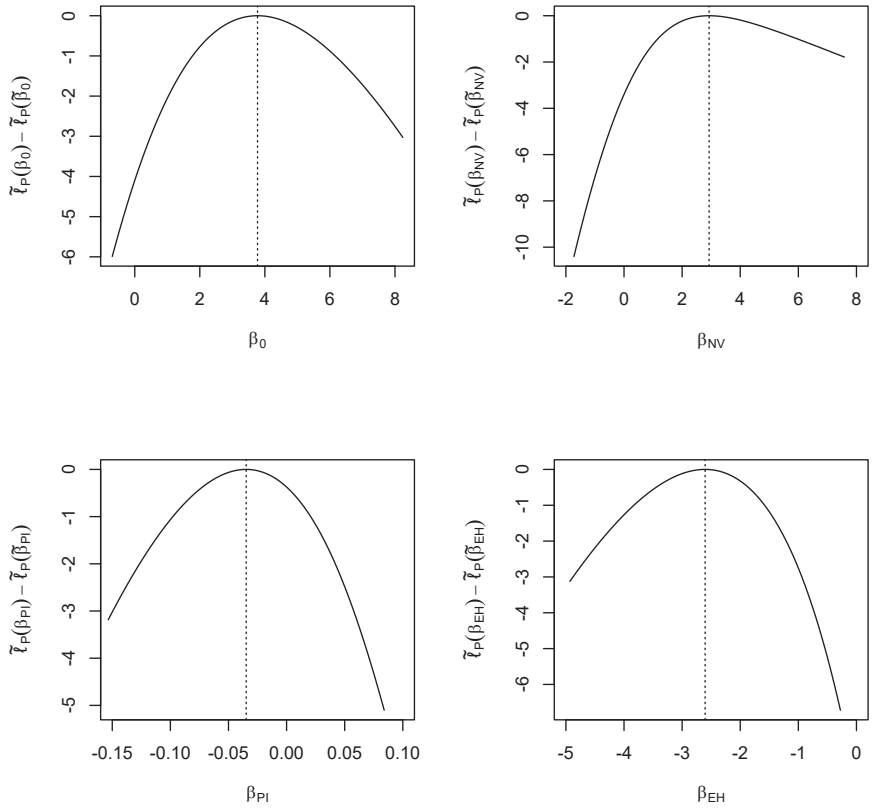

FIGURE 2.20: Relative penalized profile log-likelihood for each regression coefficient, with respect to the `endometrial` data. The vertical dotted line in each panel indicates the mean bias-reduced estimate of the corresponding parameter.

Focusing on β_{NV}, we decide to compare the shapes of the profile modified Wald statistic $\tilde{W}_{Pe}(\beta_{\text{NV}})$, the modified profile log-likelihood ratio $\tilde{W}_P(\beta_{\text{NV}})$ and the modified profile score statistic $\tilde{W}_{Pu}(\beta_{\text{NV}})$, as displayed in Figure 2.21. As expected, we can see that $\tilde{W}_{Pe}(\beta_{\text{NV}})$ is symmetric around $\tilde{\beta}_{\text{NV}}$, while in contrast $\tilde{W}_P(\beta_{\text{NV}})$ follows a more complex shape and defines wider and asymmetric confidence regions for β_{NV}. In particular, $\tilde{W}_P(\beta_{NV})$ allows to obtain confidence intervals that include a relatively large upper bound, which can be explained by the fact that NV is the data-separating covariate.

Quite unexpectedly, the shape of $\tilde{W}_{Pu}(\beta_{\text{NV}})$ seems to be almost symmetric around $\tilde{\beta}_{\text{NV}}$ and defines narrower confidence regions than the other approximate pivots. Moreover, we notice that $\tilde{W}_{Pu}(\beta_{\text{NV}})$ closely follows $\tilde{W}_P(\beta_{\text{NV}})$ as regards the definition of lower

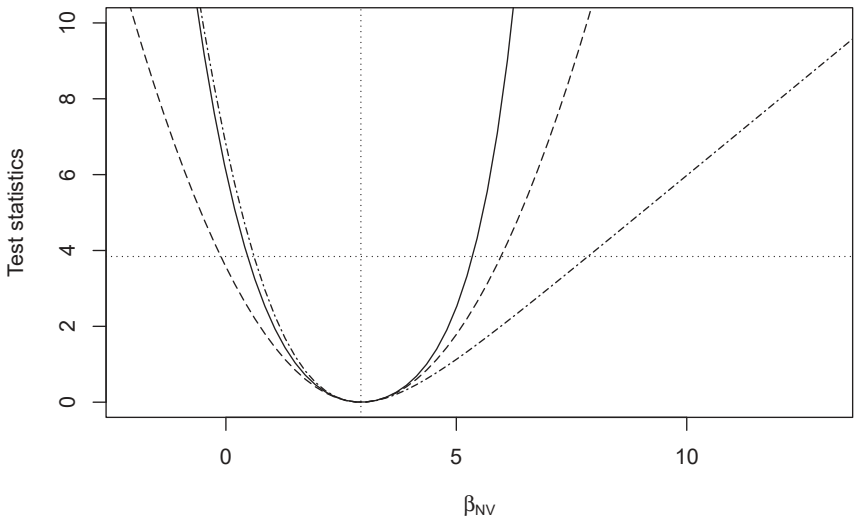

FIGURE 2.21: Shapes of the approximate test statistics based on mean bias reduction for β_{NV} in the logistic regression model, with respect to the `endometrial` data. The solid line indicates $\tilde{W}_{Pu}(\beta_{\text{NV}})$, the long-dashed line is $\tilde{W}_{Pe}(\beta_{\text{NV}})$ and the dot-dashed line corresponds to $\tilde{W}_{P}(\beta_{\text{NV}})$. The dotted vertical line indicates the mean bias-reduced estimate of β_{NV}, while the dotted horizontal line corresponds to the 0.95-quantile of a χ^2_1 distribution.

bounds for β_{NV}, nonetheless the two statistics show a very different behaviour as soon as $\beta_{\text{NV}} > \tilde{\beta}_{\text{NV}}$.

A further point of interest is to investigate the suitability of the linear approximation (2.10) for the estimation of nuisance parameters. Since profiling each regression coefficients requires estimating a three-dimensional nuisance parameter, we cannot show a direct comparison between exact and approximate constrained estimates. Nonetheless, we are more interested in the effect of plugging the approximate solution in the modified score statistic, a representation of which is displayed in Figure 2.22.

The linear approximation seems to be inadequate for β_{EH} and, most notably, for the intercept β_0, where the resulting statistic follows an irregular shape compared to its exact version. Such phenomenon also occurred in the gamma ratio model, as displayed in Figure 2.12 and 2.14, and highlights the fact that the approximate version $\tilde{r}_{Pu}(\cdot)$ is not guaranteed to be a monotonic function of the parameter of interest. Perhaps, such undesirable behaviour occurs when the constrained bias-reduced estimate is affected by a too prominent curvature, for which a linear approximation is unsuitable.

In Table 2.11, we report the confidence intervals obtained through the modified profile score test and the corresponding approximation, with respect to a nominal confidence

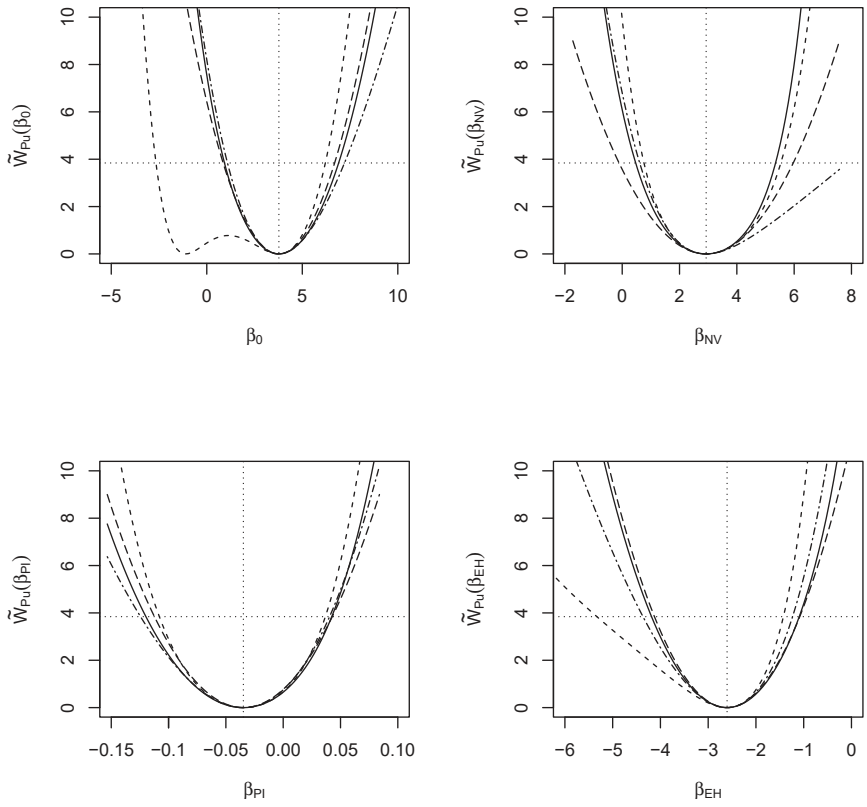

FIGURE 2.22: Modified profile score statistic (solid line) and corresponding approximation (dashed line) for each regression parameter, along with $\tilde{W}_{Pe}$ (long-dashed line) and $\tilde{W}_P$ (dot-dashed line) with respect to the `endometrial` data. The horizontal dotted line corresponds to the 0.95-quantile of a χ^2_1 distribution, while the vertical dotted line is the mean bias-reduced estimate for each parameter.

level of 0.95. Such results point out an important issue of the approximated modified profile score, considering that the lower bound of the approximated confidence interval for β_0 is quite lower than 0. From a hypothesis testing perspective, this means that

TABLE 2.11: Approximate 95% confidence intervals of the model parameters based on the modified profile score statistic and the corresponding approximation, with respect to the `endometrial` data.

	Exact	Approximate
β_0	(0.9254, 6.9029)	(-2.6587, 6.2074)
β_{NV}	(0.4700, 5.3512)	(0.7587, 5.5351)
β_{PI}	(-0.1198, 0.0424)	(-0.1081, 0.0368)
β_{EH}	(-4.1861, -1.0928)	(-5.3299, -1.4246)

using the linear approximation of the constrained estimate could lead us to dramatically different inferential conclusions than by using the exact modified profile score. Of course, we are typically not interested in the intercept term, nonetheless in other data sets such an issue may reflect on possibly important coefficients.

Although we considered confidence intervals constructed through the modified profile score statistic, we can analogously test for nullity of the model parameters and of sets of parameters, which allows for example to compare nested models. In this respect, we notice from the confidence intervals in Table 2.8 that the modified profile score test rejects the null hypotheses $H_0 : \beta_{\text{NV}} = 0$ against $H_1 : \beta_{\text{NV}} \neq 0$ in the same way as the modified log-likelihood ratio test would, considering at least a nominal significance level of 0.05.

Due to the good properties of the modified likelihood ratio statistic addressed in Heinze & Schemper (2002), the similar outcome with respect to the aforementioned hypothesis testing problem could be regarded as desirable. Nonetheless, this does not suffice to regard inference based on the modified (profile) score statistic as a competitive alternative to Wald-type inference. In order to assess this, we need proper simulation studies from which we can further analyze the statistical properties beneath our proposed test statistic.

Chapter 3

Simulation studies

3.1 Structure of our simulation studies

In this chapter, we illustrate some simulation studies taking into consideration the models introduced in the previous chapter. In this respect, we particularly focus our attention on logistic regression models for two reasons. In the first place, such models are extensively used in practical settings, therefore simulation studies in this respect may be particularly useful. The second is that logistic regression models allow us to address potentially interesting situations such as linear separation of the data and an increasing number of covariates, the latter implying an increasing parameter dimension.

Our aim is to investigate the performance of the proposed modified (profile) score statistic, especially in comparison with Wald-type inference. A second objective is to assess whether the approximate version of the modified profile score statistic yields accurate results, so that it could be a competitive alternative to both its exact version and to the modified Wald test.

To better explain the logic beneath our simulation studies, let us consider a parametric model specified as (1.1) and a known parameter value θ_0. Then, we simulate a sufficiently large number B of datasets and from each one we compute the modified score statistic $\tilde{W}_u^{(b)}(\theta_0)$, where $b = 1, \ldots, B$. We note that such statistics are computed to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, therefore the empirical distribution of $\tilde{W}_u^{(1)}(\theta_0), \ldots, \tilde{W}_u^{(B)}(\theta_0)$ is the simulated null distribution of the modified score statistic. From this result, we can compare the simulated null distribution with the corresponding asymptotic χ^2_p distribution, for instance through quantile-quantile plots. For simplicity, from now on we denote as θ the true value of the parameter, instead of θ_0.

Another point of interest is to assess whether the simulated significance level

$$\alpha^{(B)} = \frac{1}{B}\sum_{b=1}^{B} I(\tilde{W}_u^{(b)}(\theta) > \chi^2_{p;1-\alpha}),$$

where $I(\cdot)$ is the indicator function and $\chi^2_{p;1-\alpha}$ is the $(1-\alpha)$-quantile of a χ^2_p distribution, is sufficiently close to the nominal significance level α. Moreover, by evaluating $1-\alpha^{(B)}$ we obtain the simulated coverage of the corresponding confidence region without explicitly computing the latter. Indeed, if θ is the true value of the model parameter, there is a one-to-one correspondence between the acceptance and confidence regions. This result allows to carry out simulation studies with more reasonable computational effort.

Clearly, the same holds if we consider other approximate pivots instead of the modified score statistic and, analogously, focusing on a p_0-dimensional parameter of interest for inference. With respect to the latter case, it is of interest to also study the attained coverage of the modified profile score obtained by plugging in the approximation (2.10).

The following simulation studies are carried out using `R`, version 4.2.3 (R Core Team, 2023).

3.2 Simple biparametric models

In this section, we show the results from our simulation studies with respect to the canonical gamma (Section 2.3.1), canonical inverse Gaussian (Section 2.3.2) and gamma ratio (Section 2.3.3) models. In our work, such models play the role of introductory toy examples that illustrate the key aspects of the modified (profile) score statistic. Nonetheless, for completeness of discussion we report the simulation studies related also to these models.

3.2.1 Simulation from the canonical gamma model

Considering the gamma model with canonical parameter $\theta = (\alpha, \lambda) = (5, 2)$, we simulate $B = 10000$ data sets and obtain the simulated coverage of the modified statistics based on bias reduction, along with their profile versions for α and λ, considering different sample sizes $n \in \{20, 50, 100, 200\}$. We also take into account the standard likelihood-based approximate pivots for a more complete comparison.

The results are reported in Table 3.1, considering a fixed nominal significance level of 0.05. Here and in the following tables, we denote the approximate pivots omitting the

argument, considering that the latter shall be clear within the structure of the tables. Similarly, we indicate as $\tilde{W}^*_{Pu}$ the approximate version of the modified profile score statistic.

TABLE 3.1: Simulated coverage of the approximate 95% confidence regions based on the modified and standard statistics for the canonical gamma model, with respect to different sample sizes.

Parameter	Statistic	$n = 20$	$n = 50$	$n = 100$	$n = 200$
Global	$\tilde{W}_u$	0.9212	0.9338	0.9428	0.9467
	$\tilde{W}_e$	0.8981	0.9243	0.9397	0.9447
	$\tilde{W}$	0.9452	0.9462	0.9476	0.9511
	W_u	0.9509	0.9493	0.9490	0.9506
	W_e	0.9577	0.9522	0.9512	0.9516
	W	0.9425	0.9429	0.9465	0.9506
α	$\tilde{W}_{Pu}$	0.9375	0.9399	0.9473	0.9463
	$\tilde{W}_{Pe}$	0.9032	0.9247	0.9417	0.9427
	$\tilde{W}_P$	0.9452	0.9438	0.9475	0.9477
	W_{Pu}	0.9612	0.9506	0.9512	0.9486
	W_{Pe}	0.9613	0.9506	0.9512	0.9486
	W_P	0.9394	0.9407	0.9472	0.9485
λ	$\tilde{W}_{Pu}$	0.9401	0.9420	0.9491	0.9462
	$\tilde{W}^*_{Pu}$	0.9424	0.9431	0.9494	0.9466
	$\tilde{W}_{Pe}$	0.9035	0.9290	0.9404	0.9424
	$\tilde{W}_P$	0.9476	0.9453	0.9503	0.9496
	W_{Pu}	0.9619	0.9511	0.9523	0.9510
	W_{Pe}	0.9619	0.9511	0.9523	0.9510
	W_P	0.9390	0.9435	0.9487	0.9495

The results in Table 3.1 highlight that inference based on bias reduction may not always provide a better outcome. In this case, the standard global likelihood-based statistics provide overall better coverage than the modified counterparts, although with the exception of $\tilde{W}(\theta)$, which yields satisfying performance for all n.

The modified profile score and likelihood ratio statistics seem to provide overall competitive results, with attained coverage sufficiently close to the nominal 0.95 coverage. The same holds for the $\tilde{W}^*_{Pu}$ as well. Conversely, the modified profile Wald statistic seems to provide unreliable inference with respect to lower sample sizes.

In accordance with their asymptotic properties, increasing the sample size enhances the performance of both modified and standard likelihood-based statistics.

In simulation studies with lower α (not reported here), $\tilde{W}$ and $\tilde{W}_P$ were less accurate, while the results of $\tilde{W}_u$ and $\tilde{W}_{Pu}$ were comparable to those in Table 3.1.

3.2.2 Simulation from the canonical inverse Gaussian model

With respect to the canonical inverse Gaussian model, we set the true value of the parameter as $\theta = (\lambda, \phi) = (3, 2)$ and simulate $B = 10000$ data sets, considering the sample sizes $n \in \{20, 50, 100, 200\}$. Then, computing for each replication the modified and standard likelihood-based statistics in the true parameter value, along with their profile versions for both λ and ϕ, we obtain the results shown in Table 3.2.

TABLE 3.2: Simulated coverage of the approximate 95% confidence regions based on the modified and standard statistics for the canonical inverse Gaussian model, with respect to different sample sizes.

Parameter	Statistic	$n = 20$	$n = 50$	$n = 100$	$n = 200$
Global	$\tilde{W}_u$	0.9130	0.9343	0.9448	0.9493
	$\tilde{W}_e$	0.8724	0.9219	0.9388	0.9470
	$\tilde{W}$	0.9487	0.9469	0.9549	0.9522
	W_u	0.9486	0.9488	0.9553	0.9533
	W_e	0.9540	0.9511	0.9564	0.9539
	W	0.9446	0.9468	0.9539	0.9520
α	$\tilde{W}_{Pu}$	0.9381	0.9441	0.9485	0.9529
	$\tilde{W}^*_{Pu}$	0.9386	0.9441	0.9485	0.9529
	$\tilde{W}_{Pe}$	0.8941	0.9285	0.9416	0.9495
	$\tilde{W}_P$	0.9473	0.9479	0.9498	0.9534
	W_{Pu}	0.9621	0.9538	0.9538	0.9548
	W_{Pe}	0.9621	0.9538	0.9538	0.9548
	W_P	0.9412	0.9442	0.9486	0.9479
λ	$\tilde{W}_{Pu}$	0.9362	0.9448	0.9492	0.9518
	$\tilde{W}^*_{Pu}$	0.9461	0.9483	0.9512	0.9515
	$\tilde{W}_{Pe}$	0.8848	0.9271	0.9408	0.9479
	$\tilde{W}_P$	0.9481	0.9470	0.9522	0.9507
	W_{Pu}	0.9641	0.9541	0.9558	0.9523
	W_{Pe}	0.9615	0.9532	0.9552	0.9520
	W_P	0.9413	0.9421	0.9519	0.9502

As in the case of the canonical gamma model, the standard global likelihood-based statistics provide overall better results than the modified counterparts, in terms of attained coverage.

Similarly as in the previous example, the modified profile score and likelihood ratio statistics provide acceptable results, even in cases of relatively small sample size. As regards the approximation $\tilde{W}^*_{Pu}$, it provides acceptable performance, in comparison to both the exact statistic and to the expected coverage. In contrast, the modified profile

Wald statistic seems to be unreliable from an inferential point of view, considering its lack of attained coverage for relatively small n.

As the sample size increases, all the involved statistics improve in terms of simulated coverage. Such result may be of interest since the inverse Gaussian model in canonical parameterization lacks the regularity conditions related to its parameter space, and besides the corresponding modified score equation admits multiple solutions.

3.2.3 Simulation from the gamma ratio model

Focusing our attention on the gamma ratio model, we generate $B = 10000$ data sets by setting the true parameter value as $\theta = (\psi, \lambda) = (3, 7)$. Computing the modified and standard likelihood-based statistics in θ and considering $n \in \{20, 50, 100, 200\}$, we obtain the results illustrated in Table 3.3.

TABLE 3.3: Simulated coverage of the approximate 95% confidence regions based on the modified and standard statistics for the gamma ratio model, with respect to different sample sizes.

Parameter	Statistic	$n = 20$	$n = 50$	$n = 100$	$n = 200$
Global	$\tilde{W}_u$	0.9214	0.9394	0.9443	0.9463
	$\tilde{W}_e$	0.8601	0.9138	0.9298	0.9404
	W_u	0.9483	0.9493	0.9513	0.9514
	W_e	0.9227	0.9378	0.9419	0.9479
	W	0.9473	0.9510	0.9504	0.9508
α	$\tilde{W}_{Pu}$	0.9644	0.9569	0.9550	0.9539
	$\tilde{W}^*_{Pu}$	0.9766	0.9761	0.9754	0.9633
	$\tilde{W}_{Pe}$	0.9113	0.9292	0.9432	0.9445
	W_{Pu}	0.9511	0.9527	0.9524	0.9520
	W_{Pe}	0.9358	0.9396	0.9487	0.9480
	W_P	0.9456	0.9504	0.9513	0.9510
λ	$\tilde{W}_{Pu}$	0.9408	0.9482	0.9497	0.9489
	$\tilde{W}^*_{Pu}$	0.9682	0.9696	0.9599	0.9561
	$\tilde{W}_{Pe}$	0.9283	0.9438	0.9477	0.9477
	W_{Pu}	0.9510	0.9515	0.9504	0.9509
	W_{Pe}	0.9510	0.9515	0.9504	0.9509
	W_P	0.9452	0.9504	0.9498	0.9492

It is clear that, as in the previous examples, the modified statistics for θ seem to provide worse performance than the standard likelihood-based statistics. Due to the fact that θ is not the canonical parameter, in this case the modified likelihood ratio statistic is not available for comparison.

Considering the profile cases, the modified profile score statistic provides acceptable performance in terms of attained coverage. The corresponding approximation, however, shows an overall conservative behaviour, providing fairly too large coverage than expected. Such outcome seems to agree with the results in Section 2.3.3, where the exact constrained estimates show a curvature that cannot be suitably approximated.

In contrast, modified Wald-type inference seems to provide unreliable inference due to lack of attained coverage, analogously as in the previous simulation studies. As n increases, its performance seems to improve, as for the other involved statistics.

3.3 Biparametric logistic regression

Let us consider the biparametric logistic regression model without intercept, introduced in Section 2.3.4. In contrast to the other biparametric models addressed in the previous section, in this case we develop a more detailed simulation study, involving the modified likelihood ratio test (2.3) and its profile version (2.5), as well as all the other likelihood-based approximate pivots.

We simulate $B = 10000$ data sets from $\beta = (\beta_1, \beta_2) = (-0.4, 0.3)$ and we compute the approximate statistics in the true parameter value. The model matrix is kept fixed for all the simulated data sets. As regards the corresponding profile versions, we only consider β_2 as the parameter of interest since both β_1 and β_2 have a similar scale and are associated with covariates with almost same scale and position. Furthermore, we also take into account an increasing sample size $n \in \{20, 50, 100, 200\}$ in order to better characterize the asymptotic properties involved.

We mention that, in the case of $n = 20$, one of the simulated data sets is affected by linear separation, which prevents us from obtaining the standard likelihood-based tests, except for the global score statistic (1.10). Excluding the latter, for the standard likelihood-based statistics we compute the simulated coverage discarding the linearly separated data set.

In Table 3.4, we illustrate the results obtained in this simulation study, considering also different nominal significance levels $\alpha \in \{0.01, 0.05, 0.10\}$. Considering the cases of $n = 20$ and $n = 50$, we can observe that the modified score statistic performs well since the attained coverage is close to each nominal confidence level. In this respect, we also notice a slightly better performance than the modified likelihood ratio test.

In contrast, Wald-type inference with both (2.2) and (1.9) provides too broad confidence intervals, resulting in a higher simulated coverage than expected. From a hypothesis testing perspective, the standard and modified Wald tests tend to reject the

TABLE 3.4: Simulated coverage of the standard and modified likelihood-based statistics for the biparametric logistic regression model, with respect to different nominal significance levels α and different sample sizes n. The values denoted by the symbol $*$ are obtained excluding the linear separated data set.

n	Statistic	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
20	$\tilde{W}_u$	0.9913	0.9504	0.8956
	$\tilde{W}_e$	1.0000	0.9983	0.9818
	$\tilde{W}$	0.9942	0.9611	0.9129
	W_u	0.9943	0.9553	0.9024
	W_e	1.0000*	0.9956*	0.9635*
	W	0.9857*	0.9354*	0.8760*
50	$\tilde{W}_u$	0.9902	0.9516	0.9019
	$\tilde{W}_e$	0.9978	0.9778	0.9385
	$\tilde{W}$	0.9924	0.9555	0.9090
	W_u	0.9917	0.9543	0.9046
	W_e	0.9970	0.9685	0.9227
	W	0.9884	0.9467	0.8957
100	$\tilde{W}_u$	0.9907	0.9511	0.8989
	$\tilde{W}_e$	0.9952	0.9663	0.9196
	$\tilde{W}$	0.9909	0.9543	0.9017
	W_u	0.9909	0.9544	0.8990
	W_e	0.9936	0.9607	0.9073
	W	0.9892	0.9497	0.8941
200	$\tilde{W}_u$	0.9896	0.9477	0.8970
	$\tilde{W}_e$	0.9925	0.9565	0.9073
	$\tilde{W}$	0.9902	0.9486	0.9001
	W_u	0.9900	0.9479	0.8986
	W_e	0.9914	0.9524	0.9038
	W	0.9890	0.9470	0.8947

null hypothesis less often than desired, which may result in too conservative acceptance regions. In this respect, we also note that the modified Wald statistic yields even worse results than the unmodified counterpart.

An analogous analysis is reported in Table 3.5 as regards the profile version of each statistic. Also in this case, the modified profile score yields satisfactory results as regards the attained coverage, especially in comparison to the modified Wald test and in the case of $n = 20$ and $n = 50$. Unlike the global case, we notice that the modified profile likelihood ratio test provides a coverage that is slightly closer to the nominal confidence levels than that of our proposed statistic. Besides, as theoretically expected from the asymptotic results, increasing n in both global and profile cases improves the overall

TABLE 3.5: Simulated coverage of the standard and modified profile likelihood-based statistics for the biparametric logistic regression model, with respect to different nominal significance levels α and different sample sizes n. The values denoted by the symbol $*$ are obtained excluding the linear separated data set.

n	Statistic	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
20	$\tilde{W}_{Pu}$	0.9893	0.9444	0.8934
	$\tilde{W}_{Pe}$	0.9995	0.9843	0.9487
	$\tilde{W}_{P}$	0.9929	0.9529	0.9060
	W_{Pu}	0.9899*	0.9435*	0.8902*
	W_{Pe}	0.9990*	0.9752*	0.9213*
	W_{P}	0.9844*	0.9315*	0.8741*
50	$\tilde{W}_{Pu}$	0.9902	0.9483	0.8970
	$\tilde{W}_{Pe}$	0.9960	0.9651	0.9205
	$\tilde{W}_{P}$	0.9906	0.9524	0.9032
	W_{Pu}	0.9897	0.9479	0.8953
	W_{Pe}	0.9944	0.9568	0.9048
	W_{P}	0.9869	0.9437	0.8910
100	$\tilde{W}_{Pu}$	0.9914	0.9486	0.8973
	$\tilde{W}_{Pe}$	0.9940	0.9588	0.9099
	$\tilde{W}_{P}$	0.9918	0.9507	0.9017
	W_{Pu}	0.9914	0.9483	0.8973
	W_{Pe}	0.9932	0.9520	0.9016
	W_{P}	0.9906	0.9471	0.8951
200	$\tilde{W}_{Pu}$	0.9899	0.9483	0.8973
	$\tilde{W}_{Pe}$	0.9923	0.9526	0.9038
	$\tilde{W}_{P}$	0.9901	0.9494	0.8981
	W_{Pu}	0.9899	0.9481	0.8973
	W_{Pe}	0.9911	0.9503	0.8986
	W_{P}	0.9888	0.9463	0.8940

performance of all statistics, thereby reducing differences in this respect.

A more in-depth analysis with respect to the modified score statistic requires studying the corresponding simulated distribution in more detail, especially assessing its correspondence to the asymptotic null distribution. In order to address such a problem, we compare the simulated and asymptotic null distribution by means of quantile-quantile plots.

For the global modified score statistic, we obtain the quantile-quantile plots displayed in Figure 3.1, with respect to the theoretical χ^2_2 distribution. Taking into consideration the varying sample size as well, we observe that there is no substantial discrepancy between the simulated and asymptotic null distribution of the modified score test, even

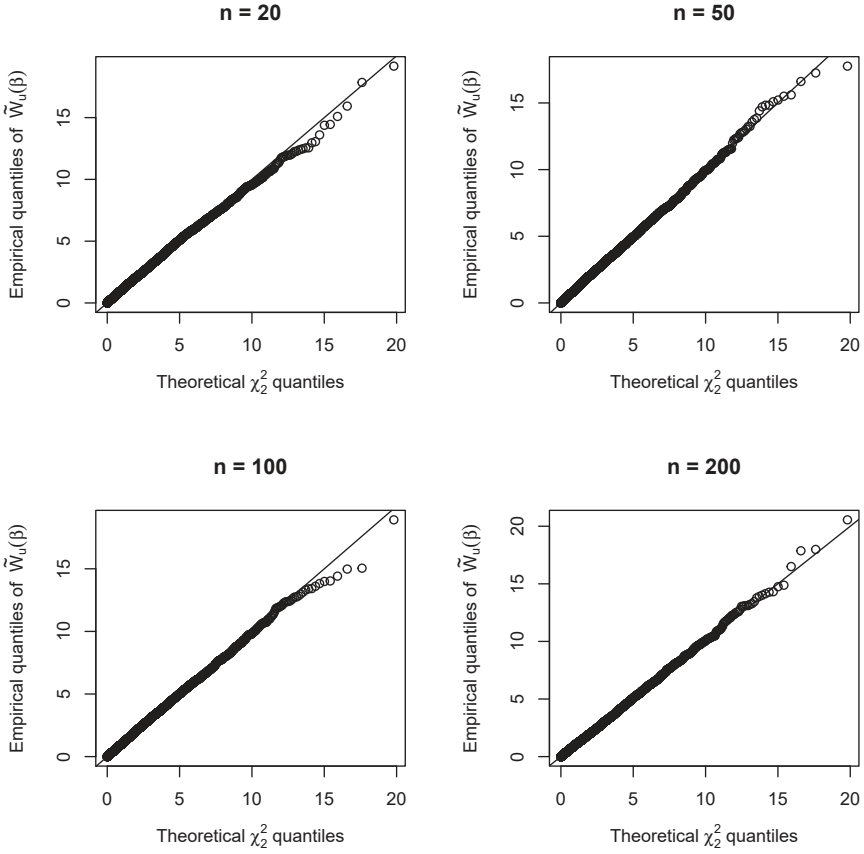

FIGURE 3.1: Biparametric logistic regression model. Quantile-quantile plots comparing the simulated null distribution of the modified score statistic for β to the theoretical asymptotic χ^2_2 null distribution for increasing sample size.

in the case of $n = 20$.

In contrast, the simulated distribution of the modified Wald statistic shows a lighter right tail than the χ^2_2, as illustrated in Figure 3.2. This result explains the over-coverage shown by the modified Wald statistic. We also note that the discrepancy between the simulated and expected distribution vanishes as the sample size increases, in accordance with the asymptotic theory and with the results in Table 3.4.

As regards the modified profile score statistic for β_2, comparing the corresponding simulated distribution to its asymptotic null χ^2_1 distribution reveals a substantial correspondence between the former and the latter. As illustrated in Figure 3.3, this holds also for smaller sample sizes.

In contrast, the modified profile Wald statistic shows a worse performance in this respect, as clear from Figure 3.4. Especially for smaller sample sizes, the simulated distribution of $\tilde{W}_{Pe}(\beta_2)$ is affected by a lighter right tail than expected, which may result in too wide confidence intervals or, analogously, in a too conservative bilateral

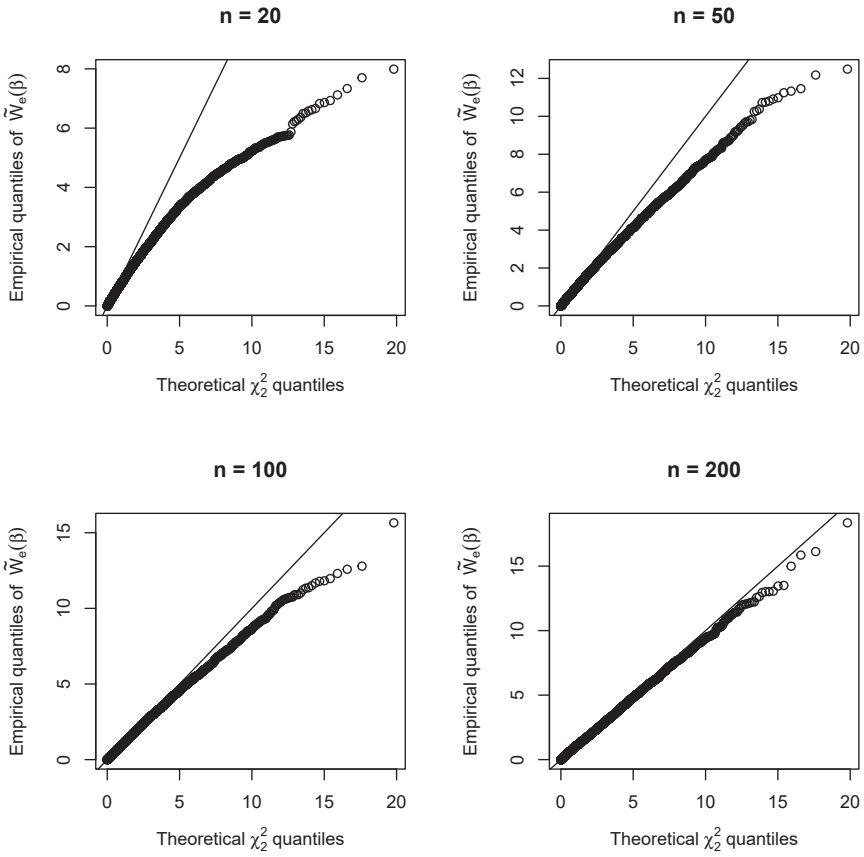

FIGURE 3.2: Biparametric logistic regression model. Quantile-quantile plots comparing the simulated null distribution of the modified Wald statistic for β to the theoretical asymptotic χ^2_2 null distribution for increasing sample size.

hypothesis testing.

In order to assess the reliability of using the linear approximation instead of the actual nuisance parameter estimate, we report in Table 3.6 a comparison between $\tilde{W}_{Pu}$ and the corresponding approximation denoted by $\tilde{W}^*_{Pu}$, in terms of simulated coverage. In this case, we observe a substantially satisfactory performance since the simulated coverage is quite close to the nominal levels. Besides, we see that the results almost coincide for $n = 200$. Hence, in this case the linear approximation seems to provide a reliable alternative for hypothesis testing and the construction of confidence intervals.

3.4 Simulation from endometrial cancer data

The desirable properties of the modified (profile) score statistic in the case of a biparametric logistic regression with no intercept have been assessed in the previous section.

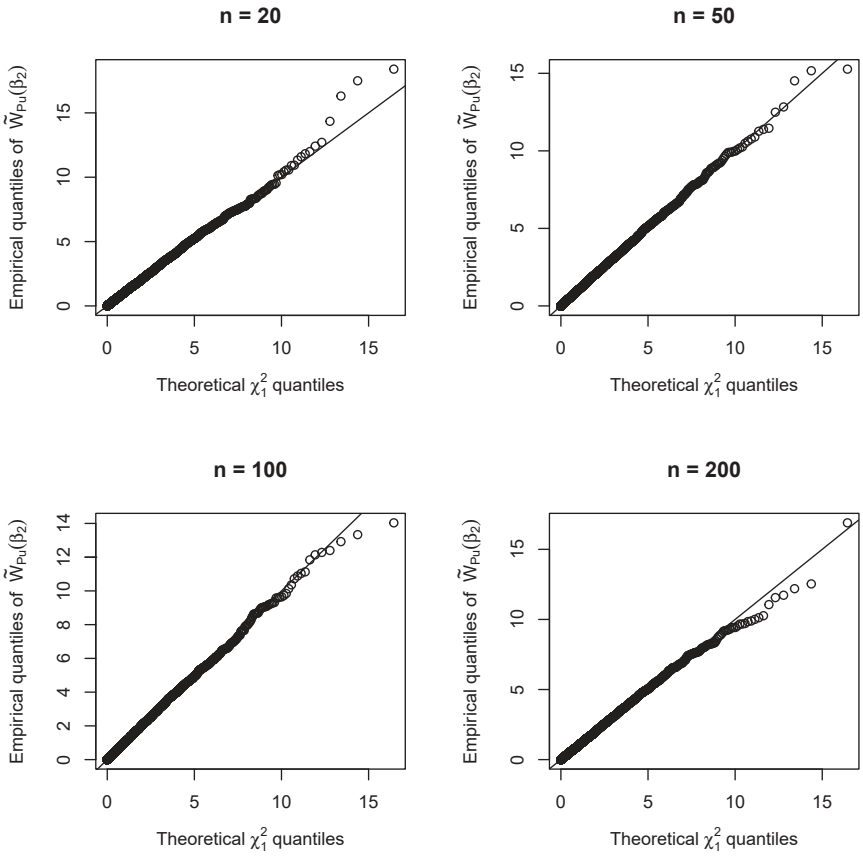

FIGURE 3.3: Biparametric logistic regression model. Quantile-quantile plots comparing the simulated null distribution of the modified profile score statistic for β_2 to the theoretical asymptotic χ^2_1 null distribution for increasing sample size.

Nevertheless, such setting is both artificial and relatively simple, which may result unrepresentative with respect to real data problems. For this reason, we focus on the `endometrial` data set and on the corresponding model (2.12) illustrated in Section 2.3.4, for which $n = 79$ and $p = 4$. Besides, considering such a problem for a simulation study may be of particular interest due to the presence of quasi-complete separation.

In this case, we simulate $B = 10000$ data sets from model (2.12) setting $\beta = \tilde{\beta} = (3.7746, 2.9293, -0.0348, -2.6042)$ and keeping the model matrix fixed. This approach amounts to generating B parametric bootstrap replicates of the model response from the mean bias-reduced estimate $\tilde{\beta}$.

Our simulations lead to 6013 linearly separated data sets, as assessed by using the `detectseparation` package. For this reason, unlike the previous simulation study, we focus exclusively on the modified approximate tests based on bias reduction, thereby excluding the standard likelihood-based approximate pivots.

In Table 3.7, we report the simulated coverage with respect to the modified statistics

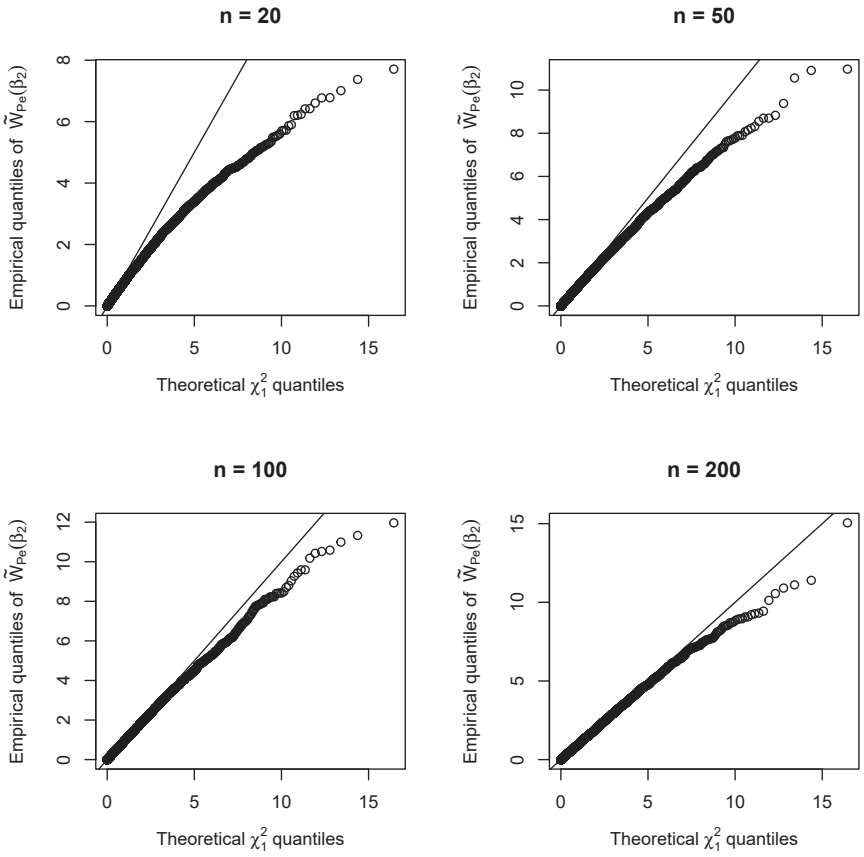

FIGURE 3.4: Biparametric logistic regression model. Quantile-quantile plots comparing the simulated null distribution of the modified profile Wald statistic for β_2 to the theoretical asymptotic χ^2_1 null distribution for increasing sample size.

and to their profile counterparts, considering also each regression parameter. Furthermore, we also include in the comparison the approximate version of the modified profile score statistic, denoted by $\tilde{W}^*_{Pu}$. For the sake of completeness, we also report the results related to the intercept term β_0, nonetheless we are usually more interested in the other regression coefficients.

Unlike the previous simulation studies, in this case the modified score statistic seems to underperform in terms of simulated coverage, since the attained values are quite far from the nominal ones. For global confidence regions, the modified likelihood ratio test seems the most adequate, while Wald-type inference yields still better results than the modified score.

As regards the intercept term, the corresponding profile modified score yields acceptable performance, nonetheless its approximation fails to provide reliable confidence regions. This outcome seems to be in agreement to the issue illustrated in Figure 2.22, where the inadequacy of the linear approximation becomes quite evident considering β_0.

TABLE 3.6: Simulated coverage of the modified profile score statistic and the corresponding approximation for β_2 in the biparametric logistic regression model, with respect to different nominal significance levels α and different sample sizes n.

n	Statistic	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
20	$\tilde{W}_{Pu}$	0.9893	0.9444	0.8934
	$\tilde{W}^*_{Pu}$	0.9887	0.9492	0.9036
50	$\tilde{W}_{Pu}$	0.9902	0.9483	0.8970
	$\tilde{W}^*_{Pu}$	0.9908	0.9491	0.8973
100	$\tilde{W}_{Pu}$	0.9914	0.9486	0.8973
	$\tilde{W}^*_{Pu}$	0.9915	0.9489	0.8975
200	$\tilde{W}_{Pu}$	0.9899	0.9483	0.8973
	$\tilde{W}^*_{Pu}$	0.9900	0.9484	0.8973

For the coefficient β_{NV}, the modified profile score statistic provides a lower empirical coverage than the nominal one, thereby yielding too narrow confidence intervals. Such phenomenon seems in accordance with Figure 2.21. Quite unexpectedly, in this case the approximation $\tilde{W}^*_{Pu}$ seems to perform better in this respect. In contrast, both Wald-type inference and the modified likelihood ratio test yield an overall better performance.

Considering β_{PI}, the modified profile score statistic shows an acceptable performance in terms of simulated coverage, however we note that the corresponding approximation provides a simulated coverage that is greater than the nominal one. Also in this case, the modified Wald and likelihood ratio tests equally show good results. The same considerations hold for β_{EH} as well.

To better investigate the simulated distribution of the modified score statistic, we consider the quantile-quantile plots displayed in Figure 3.5. It is clear that the empirical distribution of $\tilde{W}_u(\beta)$ is far from its asymptotic null distribution. As a matter of fact, the former shows a heavier right tail than the χ^2_4 distribution, which explains why the simulated coverage is so low. In contrast, the modified Wald statistic seems more adequate with respect to the χ^2_4 distribution, although there is still a deviation for larger values. As far as the modified likelihood ratio statistic is concerned, its asymptotic null distribution seems to be the most accurate with respect to its simulated distribution.

Considering the modified profile score statistic for each parameter, we display in Figure 3.6 the corresponding quantile-quantile plots. It becomes evident that the simulated distribution of $\tilde{W}_{Pu}(\beta_{\text{NV}})$ yields a much heavier right tail than desired, which results in under-covering confidence intervals or, equivalently, in hypothesis testing with a too large significance level if the χ^2_1 distribution is employed. Such an issue seems to occur

TABLE 3.7: Simulated coverage of the modified statistics and of the corresponding profile versions in the simulation study from the `endometrial` data, with respect to different nominal significance levels α.

Parameter	Statistic	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Global	$\tilde{W}_u$	0.9594	0.8821	0.8348
	$\tilde{W}_e$	0.9790	0.9338	0.8774
	$\tilde{W}$	0.9908	0.9537	0.9105
β_0	$\tilde{W}_{Pu}$	0.9888	0.9546	0.9101
	$\tilde{W}^*_{Pu}$	0.9565	0.9125	0.8757
	$\tilde{W}_{Pe}$	0.9914	0.9533	0.9053
	$\tilde{W}_P$	0.9918	0.9561	0.9058
β_{NV}	$\tilde{W}_{Pu}$	0.9692	0.9218	0.8763
	$\tilde{W}^*_{Pu}$	0.9841	0.9484	0.9033
	$\tilde{W}_{Pe}$	0.9850	0.9518	0.9076
	$\tilde{W}_P$	0.9923	0.9653	0.9337
β_{PI}	$\tilde{W}_{Pu}$	0.9919	0.9573	0.9138
	$\tilde{W}^*_{Pu}$	0.9960	0.9741	0.9454
	$\tilde{W}_{Pe}$	0.9942	0.9542	0.9025
	$\tilde{W}_P$	0.9912	0.9505	0.9050
β_{EH}	$\tilde{W}_{Pu}$	0.9871	0.9517	0.9079
	$\tilde{W}^*_{Pu}$	0.9928	0.9719	0.9374
	$\tilde{W}_{Pe}$	0.9880	0.9496	0.9007
	$\tilde{W}_P$	0.9921	0.9557	0.9060

also with respect to β_0 and β_{EH}. However, this affects larger quantiles than the commonly used critical thresholds, resulting in a still acceptable performance considering a nominal approximate significance level $\alpha \in \{0.01, 0.05, 0.10\}$.

Clearly, inference on the coefficient β_{NV} is quite problematic when using our proposed statistic. Perhaps, such phenomenon is related to the fact that NV is the data-separating covariate. For this reason, let us focus our attention on such parameter, addressing in more detail the distributions of the modified Wald and likelihood ratio statistics as well, as shown in Figure 3.7. Such illustration highlights the fact that also $\tilde{W}_{Pe}(\beta_{\text{NV}})$ and $\tilde{W}_P(\beta_{\text{NV}})$ seem to be affected by the same issue, namely a clear deviation from the expected asymptotic distribution, although not in the same magnitude as $\tilde{W}_{Pu}(\beta_{\text{NV}})$.

To investigate this further, an inspection of the simulated density of the modified signed statistics $\tilde{r}_{Pu}(\beta_{\text{NV}})$, $\tilde{r}_{Pe}(\beta_{\text{NV}})$ and $\tilde{r}_P(\beta_{\text{NV}})$ reveals an appreciable discrepancy from the asymptotic standard Gaussian distribution. Indeed, from the histograms reported in Figure 3.8 the simulated distributions resemble a mixture of two bell-shaped

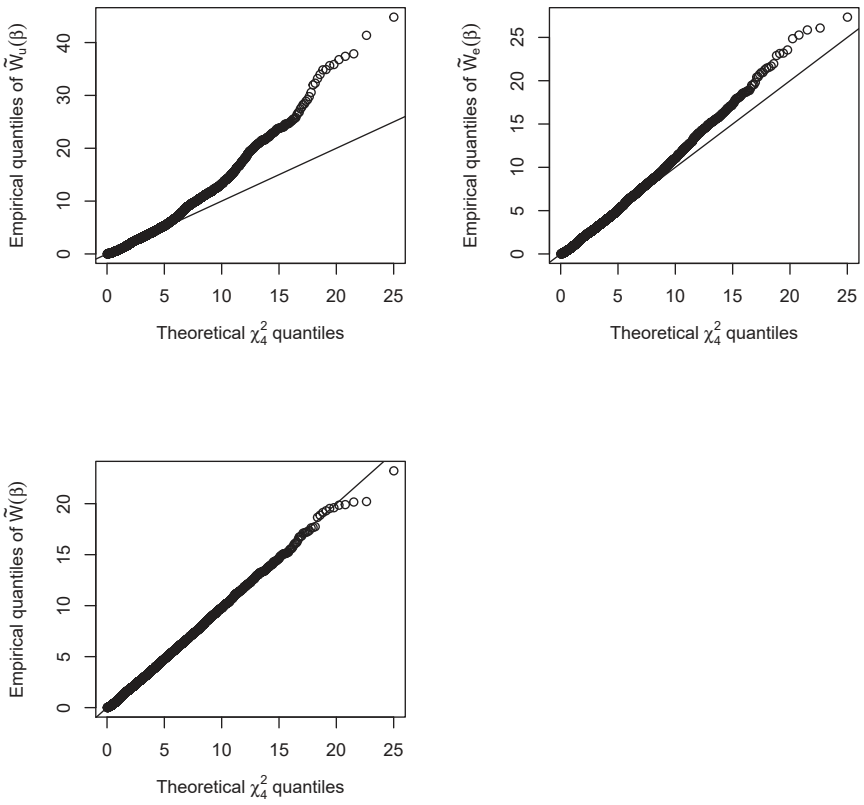

FIGURE 3.5: Simulation study based on the `endometrial` data set. Quantile-quantile plots comparing the simulated null distribution of the modified score (top-left), Wald (top-right) and likelihood ratio (bottom-left) statistics for β to the theoretical asymptotic χ^2_4 null distribution.

distributions. Indeed, conditionally on data separation, the two conditional simulated distributions clearly differ in location and scale, as illustrated in the three left-hand side panels in Figure 3.8.

To get a better understanding of the problem, a possibility is to increase the amount of information contained in the data with respect to the model parameters, which can be easily done by increasing the sample size. We achieve such a task by conducting a second simulation study from the `endometrial` data set, however this time we duplicate the rows of the model matrix. As a result, we simulate from the same model, obtaining $B = 10000$ data sets of size $n = 158$.

Nonetheless, such a sample size does not protect from the presence of complete or quasi-complete separation, indeed in our simulation 3634 data sets are still affected by such an issue. In spite of this, it is of interest to assess possible improvements of our proposed statistic stemming from an increased sample size.

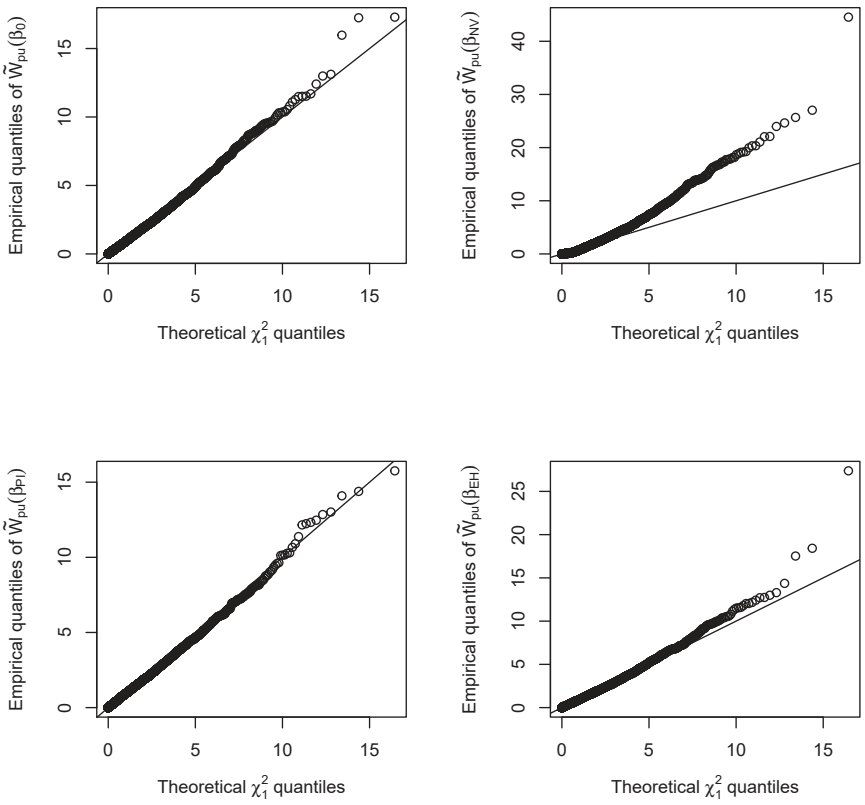

FIGURE 3.6: Simulation study based on the `endometrial` data set. Quantile-quantile plots comparing the simulated null distribution of the modified profile score statistic for each component of β to the theoretical asymptotic χ^2_1 null distribution.

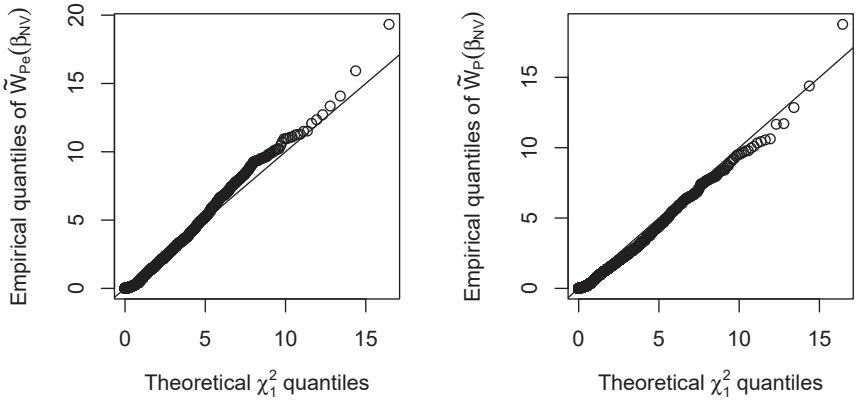

FIGURE 3.7: Simulation study based on the `endometrial` data set. Quantile-quantile plots comparing the simulated null distribution of the modified profile Wald (left) and likelihood ratio (right) statistics for β_{NV} to the theoretical asymptotic χ^2_1 null distribution.

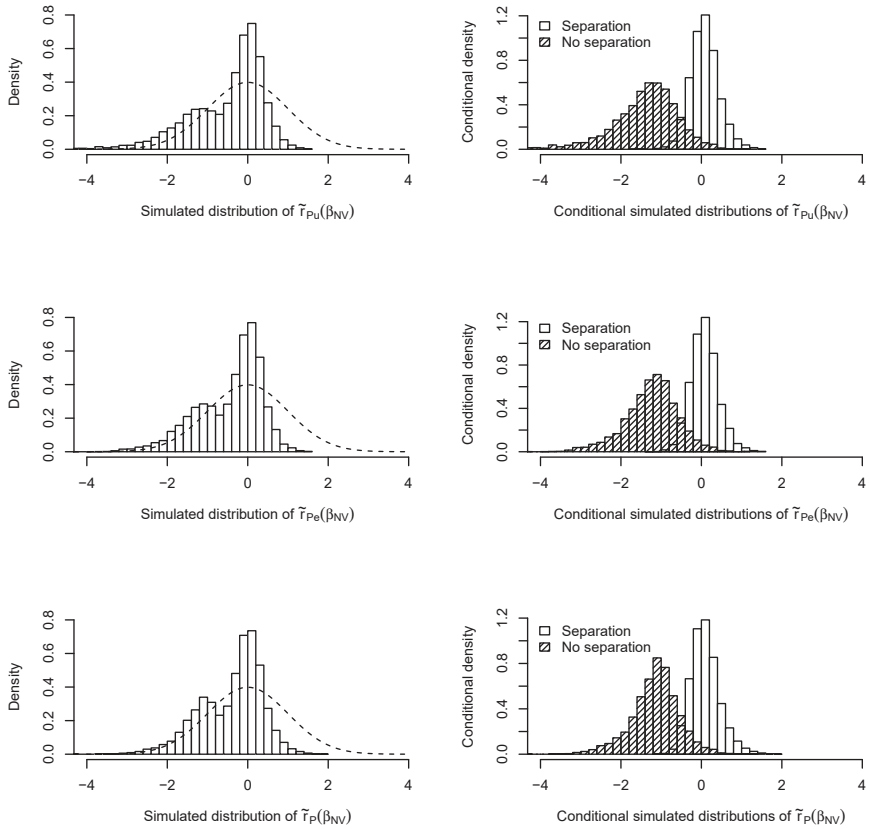

FIGURE 3.8: Histograms of the simulated distribution of the signed modified statistics (left-hand side) and of their simulated distribution conditionally on data separation, with respect to the simulation study based on the `endometrial` data set. The dashed curve in the left-hand side panels corresponds to the standard normal density.

The results obtained from this simulation study are reported in Table 3.8. In the first place, we can see a slight improvement of the modified score statistic for β since the simulated coverage is slightly greater than that shown in Table 3.7. However, its overall performance is still unsatisfying, considering that Wald-type inference provides a fairly better protection against a too large type-I error probability.

Secondly, increasing the sample size seems to only slightly improve the attained coverage of the modified score test in the most critical case, namely that of β_{NV} as parameter of interest. In this respect, the modified profile likelihood ratio statistic shows no appreciable change in terms of simulated coverage, whereas Wald-type inference provides slightly better results.

As regards the other regression coefficients, increasing the sample size results in an overall improvement of our proposed statistic and of the corresponding approximation. Nonetheless, considering the latter and β_0 the attained coverage is still rather unsatisfying, which underlines a difficulty in providing a reliable approximation of the modified

TABLE 3.8: Simulated coverage of the modified statistics and of the corresponding profile versions in the second simulation study from the `endometrial` data, with respect to different nominal significance levels α.

Parameter	Statistic	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Global	$\tilde{W}_u$	0.9700	0.9197	0.8653
	$\tilde{W}_e$	0.9829	0.9334	0.8841
	$\tilde{W}$	0.9923	0.9563	0.9078
β_0	$\tilde{W}_{Pu}$	0.9908	0.9528	0.9070
	$\tilde{W}^*_{Pu}$	0.9685	0.9310	0.8954
	$\tilde{W}_{Pe}$	0.9910	0.9510	0.9020
	$\tilde{W}_P$	0.9912	0.9524	0.9039
β_{NV}	$\tilde{W}_{Pu}$	0.9752	0.9346	0.8925
	$\tilde{W}^*_{Pu}$	0.9859	0.9503	0.9094
	$\tilde{W}_{Pe}$	0.9860	0.9498	0.9096
	$\tilde{W}_P$	0.9936	0.9658	0.9305
β_{PI}	$\tilde{W}_{Pu}$	0.9898	0.9534	0.9058
	$\tilde{W}^*_{Pu}$	0.9947	0.9659	0.9250
	$\tilde{W}_{Pe}$	0.9890	0.9477	0.8957
	$\tilde{W}_P$	0.9887	0.9476	0.8978
β_{EH}	$\tilde{W}_{Pu}$	0.9897	0.9519	0.9032
	$\tilde{W}^*_{Pu}$	0.9923	0.9617	0.9178
	$\tilde{W}_{Pe}$	0.9897	0.9498	0.8989
	$\tilde{W}_P$	0.9928	0.9518	0.9041

profile score in this respect.

From the simulation studies from the `endometrial` data, we can draw possibly more insightful conclusions than with those based on the biparametric logistic regression. In the first place, in the case of a high probability of data separation it seems that the theoretical assumptions related to the usual asymptotic distributions do not hold, according to the results in Figure 3.8. In this case, such a problem does not seem to dramatically affect the attained coverage of the modified Wald and log-likelihood ratio statistics, provided that the usual nominal approximate levels are used. Nevertheless, problems may arise when testing for unidirectional alternative hypotheses. We limit ourselves to pointing out such an issue, nonetheless further theoretical investigation may be required.

In the second place, the modified score statistic does not seem to be reliable in the presence of data separation, for which the other alternatives provide a qualitatively superior inference. Regardless, the profile version of the modified score statistic seems

to be sufficiently reliable in terms of attained coverage, although with the exception of β_{NV}. In this respect, a higher sample size does not seem to substantially protect from unsuitable coverage. Perhaps, such a problem persists due to the fact that β_{NV} is associated with the data-separating covariate.

Thirdly, approximating the modified profile score function does not seem to provide reliable results in this case, considering the intercept term β_0. This problem seems to persist even with a double sample size, although in a slightly lower magnitude. Perhaps, this issue may analogously derive from the high probability of data separation. In this respect, the suitability of such approximation needs to be carefully considered in the following simulation studies.

3.5 Logistic regression with increasing number of covariates

In order to provide a more thorough characterization of our proposed statistic, in this section we focus our attention on the dimensionality of the parameter space. Referring to logistic regression models for both ease of discussion and for the existence of the penalized log-likelihood (1.23), our aim is to study possible interactions between the performance of the modified (profile) score statistic and an increasing dimension p of the parameter, keeping the sample size n fixed.

We simulate $B = 10000$ data sets consisting of $n = 200$ observations and a p-dimensional regression parameter $\beta = (\beta_1, \ldots, \beta_p)$, with $p \in \{5, 10, 20, 30\}$. More specifically, we first generate the p components of β from a standard Gaussian distribution. Then, for each p we simulate $p - 1$ covariates from independent standard normal distributions, while keeping an intercept term corresponding to β_1. Such model matrix is kept fixed for each p.

In our simulations, complete or quasi-complete separation occurred only in the case of $p = 30$, resulting in a total of 242 separated data sets. Considering a higher p resulted in a much greater amount of separated data sets, therefore we limit ourselves to $p \leq 30$.

Considering nominal approximate significance levels $\alpha \in \{0.01, 0.05, 0.10\}$, we report in Table 3.9 the simulated coverage of the modified statistics. The structure of our simulation study allows us to focus on a single parameter of interest, namely β_2, in order to compare the corresponding modified profile statistics.

Clearly, $\tilde{W}_u(\beta)$ seems to lose attained coverage as p increases, which entails misleading inference on the model parameters. In contrast, the modified Wald and likelihood ratio

TABLE 3.9: Simulated coverage of the modified statistics and of the corresponding profile versions for β_2 in the logistic regression model, with respect to different nominal significance levels α and parameter dimensions p.

p	Parameter	Statistic	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5	Global	$\tilde{W}_u$	0.9863	0.9439	0.8928
		$\tilde{W}_e$	0.9923	0.9591	0.9169
		$\tilde{W}$	0.9891	0.9505	0.9000
	β_2	$\tilde{W}_{Pu}$	0.9896	0.9477	0.8973
		$\tilde{W}^*_{Pu}$	0.9902	0.9486	0.8987
		$\tilde{W}_{Pe}$	0.9930	0.9557	0.9068
		$\tilde{W}_P$	0.9906	0.9501	0.8996
10	Global	$\tilde{W}_u$	0.9787	0.9303	0.8729
		$\tilde{W}_e$	0.9921	0.9590	0.9157
		$\tilde{W}$	0.9910	0.9532	0.9061
	β_2	$\tilde{W}_{Pu}$	0.9902	0.9494	0.8971
		$\tilde{W}^*_{Pu}$	0.9920	0.9535	0.9026
		$\tilde{W}_{Pe}$	0.9932	0.9553	0.9030
		$\tilde{W}_P$	0.9906	0.9477	0.8972
20	Global	$\tilde{W}_u$	0.9608	0.8883	0.8163
		$\tilde{W}_e$	0.9897	0.9613	0.9248
		$\tilde{W}$	0.9920	0.9551	0.9073
	β_2	$\tilde{W}_{Pu}$	0.9909	0.9515	0.9046
		$\tilde{W}^*_{Pu}$	0.9930	0.9575	0.9089
		$\tilde{W}_{Pe}$	0.9925	0.9551	0.9081
		$\tilde{W}_P$	0.9897	0.9511	0.9019
30	Global	$\tilde{W}_u$	0.9102	0.8047	0.7216
		$\tilde{W}_e$	0.9775	0.9302	0.8832
		$\tilde{W}$	0.9924	0.9610	0.9182
	β_2	$\tilde{W}_{Pu}$	0.9942	0.9618	0.9169
		$\tilde{W}^*_{Pu}$	0.9958	0.9702	0.9282
		$\tilde{W}_{Pe}$	0.9933	0.9560	0.9053
		$\tilde{W}_P$	0.9913	0.9528	0.9059

statistics show a satisfying behaviour with respect to the corresponding asymptotic null distributions.

As regards $\tilde{W}_{Pu}(\beta_2)$, the corresponding attained coverage seems to be sufficiently close to the nominal ones and comparable to the other modified profile statistics, if not for slight over-coverage in the case $p = 30$. Furthermore, the corresponding approximation denoted by $\tilde{W}^*_{Pu}$ provides an overall acceptable result in terms of simulated coverage, although when $p = 30$ it shows a more conservative behaviour than $\tilde{W}_{Pu}$.

A further comparison can be made considering also the standard likelihood-based statistics, whose simulated coverage is reported in Table 3.10.

TABLE 3.10: Simulated coverage of the standard likelihood-based statistics and their profile versions for β_2 in the logistic regression model, with respect to different nominal significance levels α and parameter dimensions p. The values denoted "$*$" were computed considering the 9758 non-separated data sets.

p	Parameter	Statistic	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5	Global	W_u	0.9888	0.9488	0.8993
		W_e	0.9924	0.9585	0.9130
		W	0.9870	0.9449	0.8887
	β_2	W_{Pu}	0.9898	0.9476	0.8946
		W_{Pe}	0.9909	0.9500	0.8972
		W_P	0.9894	0.9450	0.8937
10	Global	W_u	0.9885	0.9511	0.9026
		W_e	0.9957	0.9711	0.9339
		W	0.9846	0.9371	0.8805
	β_2	W_{Pu}	0.9882	0.9375	0.8824
		W_{Pe}	0.9908	0.9412	0.8863
		W_P	0.9860	0.9322	0.8791
20	Global	W_u	0.9867	0.9466	0.9009
		W_e	0.9985	0.9870	0.9703
		W	0.9677	0.8911	0.8126
	β_2	W_{Pu}	0.9803	0.9170	0.8568
		W_{Pe}	0.9864	0.9248	0.8620
		W_P	0.9756	0.9093	0.8515
30	Global	W_u	0.9835^*	0.9412^*	0.8974^*
		W_e	0.9995^*	0.9965^*	0.9918^*
		W	0.9009^*	0.7593^*	0.6494^*
	β_2	W_{Pu}	0.9554^*	0.8535^*	0.7752^*
		W_{Pe}	0.9813^*	0.8854^*	0.7972^*
		W_P	0.9309^*	0.8311^*	0.7606^*

We notice that the score statistic for β yields considerable performance in terms of attained coverage, even as p becomes larger. Nonetheless, all the standard likelihood-based profile statistics for β_2 clearly fail to provide sufficient coverage with respect to each nominal α, especially as p becomes larger. In contrast, a more reliable profile inference is available by means of the modified profile statistics based on bias reduction, including the approximation $\tilde{W}^*_{Pu}$.

In order to get a further understanding on the simulated distribution of the modified score statistic, we provide a comparison with respect to the asymptotic null χ^2_p distribution for different p, as shown in Figure 3.9.

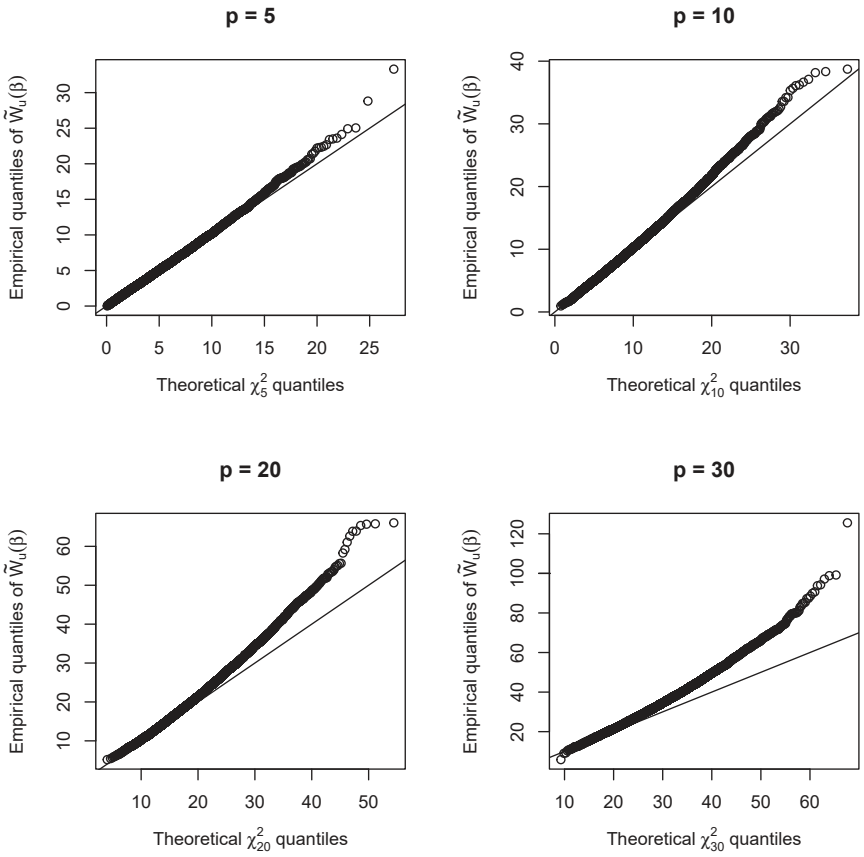

FIGURE 3.9: Logistic regression model with varying parameter dimension p. Quantile-quantile plots comparing the simulated distribution of $\tilde{W}_u(\beta)$ to the corresponding asymptotic χ^2_p null distribution.

It is clear that, as p becomes larger, the empirical simulated distribution of $\tilde{W}_u(\beta)$ becomes affected by a too heavy right tail if compared to that of a χ^2_p distribution, which results in insufficient attained coverage. In contrast, even considering $p = 30$, Wald-type inference and the modified likelihood ratio statistic provide more cohesion between the simulated and asymptotic null distributions, as illustrated in Figure 3.10.

Focusing our attention on the modified profile score statistic for β_2, there seems to be no dramatic departure from its asymptotic χ^2_1 distribution. Only considering $p = 30$, the right tail of the corresponding simulated distribution is appreciably lighter than the expected one, which may result in too conservative confidence intervals and hypothesis testing for β_2.

Again, considering the highest dimension $p = 30$, we also report the quantile-quantile plots of the modified profile Wald and likelihood ratio statistics in Figure 3.12. Such illustration highlights that $\tilde{W}_{Pe}(\beta_2)$ provides an overall worse cohesion to the χ^2_1 distribution than our proposed statistic does, especially considering the right tail of the

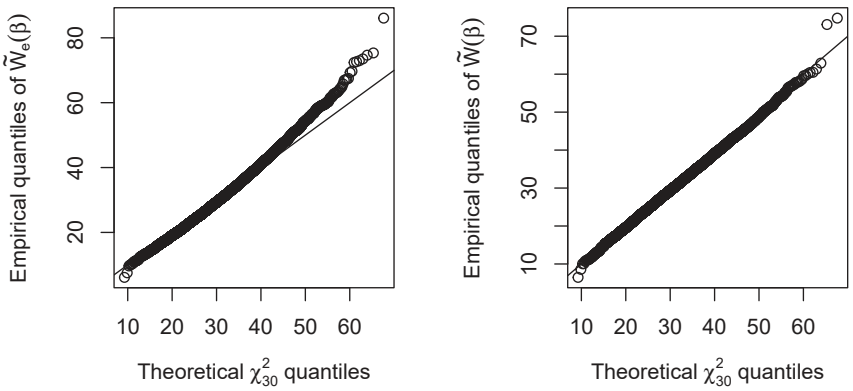

FIGURE 3.10: Logistic regression model with $p = 30$. Quantile-quantile plots comparing the simulated distribution of $\tilde{W}_e(\beta)$ (left) and $\tilde{W}(\beta)$ (right) to the corresponding asymptotic χ^2_{30} null distribution.

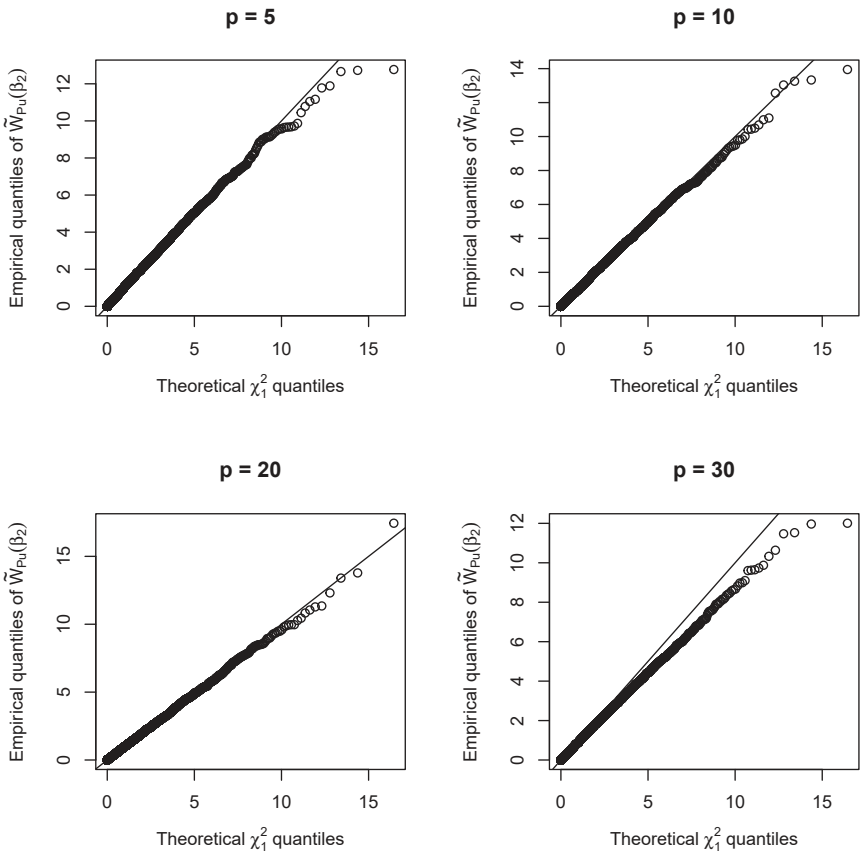

FIGURE 3.11: Logistic regression model with varying parameter dimension p. Quantile-quantile plots comparing the simulated distribution of $\tilde{W}_{Pu}(\beta_2)$ to the corresponding asymptotic χ^2_1 null distribution.

simulated distribution. Nonetheless, this does not seem to affect the corresponding performance as regards the usual nominal confidence levels, for which Wald-type inference provides a better simulated coverage. Furthermore, focusing on $\tilde{W}_P(\beta_2)$, it is evident that such statistic yields the best overall performance.

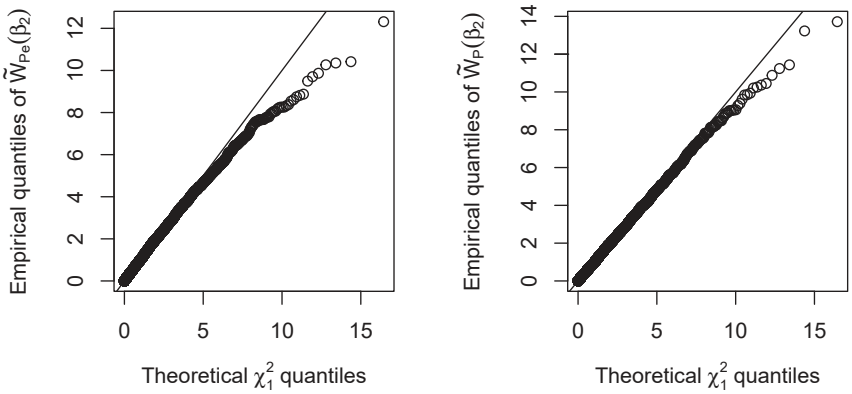

FIGURE 3.12: Logistic regression model with $p = 30$. Quantile-quantile plots comparing the simulated distribution of $\tilde{W}_{Pe}(\beta_2)$ (left) and $\tilde{W}_P(\beta_2)$ (right) to the corresponding asymptotic χ^2_1 null distribution.

From this simulation study, it is clear that a higher dimensionality of the parameter dramatically affects the performance of the modified score statistic, while in contrast its profile counterpart seems to be more stable in this respect. Furthermore, the previous simulation study based on the `endometrial` data provides a similar outcome, namely showing the inadequacy of the global modified score statistic, although in the latter example the dimensionality is relatively low.

The issues of the modified score statistic do not seem to be related to the overall parameter dimension p. Therefore, an important question that stems from these results is whether the dimension p_0 of the parameter of interest affects the overall performance of the modified profile score statistic. To investigate this further, let us consider the case $p = 30$ and an increasing dimension of the parameter of interest. Then, for each simulated data set we compute the modified profile score statistic with respect to the first p_0 components of the regression parameter β, considering $p_0 \in \{2, 5, 8, 11, 14, 17, 20, 23, 26, 29\}$. Given a nominal significance level $\alpha = 0.05$, we compute the attained coverage for each p_0 for both $\tilde{W}_{Pu}$ and its approximation $\tilde{W}^*_{Pu}$, and display the outcome in Figure 3.13 as functions of p_0. For a broader comparison, we also consider $\tilde{W}_{Pe}$ and $\tilde{W}_P$.

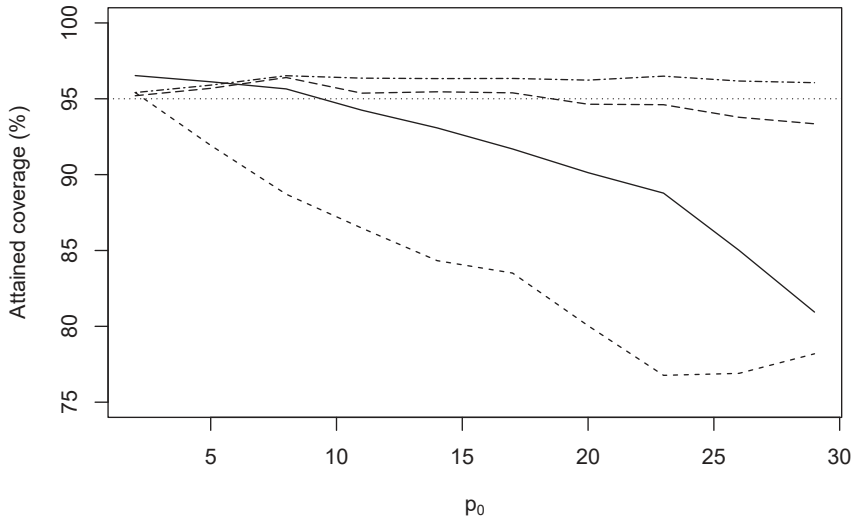

FIGURE 3.13: Percentage of attained coverage of $\tilde{W}_{Pu}$ (solid line), $\tilde{W}^*_{Pu}$ (dashed line), $\tilde{W}_{Pe}$ (long-dashed line) and $\tilde{W}_P$ (dot-dashed line), with respect to the logistic regression model with $p = 30$ and a nominal significance level $\alpha = 0.05$. The horizontal dotted line shows the desired 95% coverage.

Quite evidently, the performance of the proposed statistic degrades as we consider higher-dimensional parameters of interest, in accordance with the previous results. This suggests that the modified (profile) score statistic should be used with respect to relatively low-dimensional parameters, whereas for higher p_0 the other modified statistics are more suitable. Moreover, from Figure 3.13 we notice that the approximate modified profile score statistic is affected by a faster and more accentuated degradation in terms of attained coverage. Regardless, for a relatively small value of p_0 it still provides a qualitatively satisfying inference.

This simulation study clearly illustrates that the modified profile score statistic and its approximation can yield acceptable performance even in cases of a relatively high parameter dimension. Nonetheless, the corresponding inferential procedures should be carried out considering low dimensional parameters of interest. In practical settings this could be an unimportant limitation, provided that for instance confidence intervals or single-parameter significance tests are of interest. Conversely, if the practitioner is interested in testing for higher-dimensional parameters, additional care needs to be taken when using the modified score and profile score statistics.

3.6 Simulation from infertility data

In the previous section, we considered a simulation study based on an artificially generated problem, which could provide us with a better understanding on the modified score statistic, especially with respect to the related issues in the presence of relatively high-dimensional parameters of interest. Nonetheless, analogously as with the simulations based on the biparametric logistic regression, such an artificial setting may be unrepresentative with respect to real data problems.

For this reason, let us consider a further simulation study based on a real data set. In this respect, we take into consideration the `infert` data, available for instance from the `datasets` (R Core Team, 2023) package in `R`. Such data were collected in the work of Trichopoulos et al. (1976), whose objective was to study the relationship between secondary sterility and abortion.

As described in Kosmidis et al. (2020, Section 5.4), from which our study is inspired, the `infert` data stems from a retrospective and matched case-control study involving 83 patients affected by secondary sterility. For each patient, two healthy controls are matched based on corresponding age, education and parity, with the exception of a patient for which only one control is available. As a result, the data set consists of a total of $n = 248$ rows corresponding to either cases or matched controls.

In the first place, for each statistical unit a dichotomous variable denoted by `case` is available and coded as 0 and 1 to indicate controls and cases respectively. In the second place, `stratum` is a qualitative ordinal variable coded with integers ranging from 1 to 83, labeling each matched set of cases and controls. Thirdly, for each subject the variables `spontaneous` and `induced` respectively measure the presence of spontaneous and induced abortions. Both variables are coded in such a way that 0 corresponds to absence of abortions, 1 indicates one abortion and 2 denotes that two or more abortions have occurred. Further information such as age and education is available for each subject, nonetheless our focus is exclusively on the aforementioned variables. A brief summary of the data is reported in Table 3.11, considering the absolute frequencies of each combination of `case`, `spontaneous` and `induced`.

Based on Kosmidis et al. (2020, Formula (12)), a suitable statistical model for such data is

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \lambda_i + \beta_1 x_{ij} + \beta_2 x'_{ij} + \beta_3 z_{ij} + \beta_3 z'_{ij}, \tag{3.1}$$

where π_{ij} corresponds to the probability of secondary infertility for the j-th subject that belongs to i-th matched set, for $i = 1, \ldots, 83$ and $j = 1, \ldots, n_i$, with n_i being the total subjects in the i-th stratum. Model (3.1) corresponds to a logistic regression model,

TABLE 3.11: Absolute frequencies with respect to the variables `case`, `spontaneous` and `induced` in the `infert` data.

		spontaneous		
case	induced	0	1	2
0	0	60	25	11
	1	33	11	1
	2	20	4	0
1	0	7	22	18
	1	12	5	6
	2	9	4	0

where λ_i are strata-specific intercepts that can be regarded as 83 nuisance parameters. In contrast, $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$ denotes the vector of regression parameters associated with the model covariates, therefore it can be regarded as our parameter of interest. As regards the covariates, x_{ij} and x'_{ij} are dummy variables corresponding to `spontaneous` being 1 and 2 respectively, while analogously z_{ij} and z'_{ij} indicate `induced` being 1 and 2 respectively. For this reason, the design matrix corresponding to model (3.1) consists of $n = 248$ rows and $p = 87$ columns.

Following the terminology and definitions of Pace & Salvan (1997, page 124), in this case the nuisance parameter $\lambda = (\lambda_1, \ldots, \lambda_{83})$ corresponds to an incidental parameter, considering that its dimension depends on the sample size. In contrast, the parameter β can be regarded as a structural parameter since it characterizes the common structure of the data.

In this problem, we also note that the sample size within each stratum is at most 3, while in contrast the number of strata is 83. In such settings, suitable modifications of the profile likelihood can provide a qualitatively superior inference than profile likelihood, considering that the size of the incidental parameter is $O(n)$ and causes inconsistency of the maximum likelihood estimator of β (for a more thorough discussion, see Sartori, 2003).

An additional and well-known approach used to deal with incidental parameters is given by the conditional likelihood, obtained by conditioning on a partially sufficient statistic for λ (see for example Pace & Salvan, 1997, pages 128 and 133). Nonetheless, as mentioned in Kosmidis et al. (2020, Section 5.4), such a method lacks the generality of both mean and median bias reduction, considering that the existence of a sufficient statistic for the strata-specific parameters is not guaranteed, for instance considering probit regression instead of model (3.1). Furthermore, inference based on mean and

median bias reduction allows to obtain estimates for the nuisance parameters as well, in contrast with conditional likelihood.

Through the application of mean bias reduction, we obtain the estimate for β reported in Table 3.12, along with the corresponding standard errors and approximately 95% Wald-type confidence intervals. From such results, we can see that the components of the mean bias-reduced estimate $\tilde{\beta}$ are positive and statistically significant, with respect to a component-wise hypothesis testing against nullity and considering a nominal approximate significance level of 0.05.

TABLE 3.12: Summary of the mean bias-reduced parameter estimates for the `infert` data and approximate 95% Wald-type confidence intervals.

	Estimate	Standard error	Confidence interval
β_1	2.0550	0.4721	(1.1297, 2.9804)
β_2	3.9538	0.7077	(2.5669, 5.3408)
β_3	1.3050	0.4742	(0.3756, 2.2345)
β_4	2.7145	0.7438	(1.2567, 4.1723)

Considering the profile modified score statistic and its approximation for each regression parameter, we obtain the approximately 95% confidence intervals reported in Table 3.13. From such results, we observe that the approximate modified profile score statistic yields a comparable outcome with respect to its exact counterpart. Furthermore, from a hypothesis testing perspective both statistics lead to analogous results as with Wald-type inference, at least considering testing against nullity of the model coefficients.

TABLE 3.13: Approximate 95% confidence intervals of the model parameters based on the modified profile score statistic and the corresponding approximation, with respect to the `infert` data.

	Exact	Approximate
β_1	(1.3015, 2.9385)	(1.2115, 2.7761)
β_2	(2.8368, 5.2472)	(2.8393, 5.1487)
β_3	(0.5280, 2.2038)	(0.4176, 2.0331)
β_4	(1.4947, 4.0783)	(1.2625, 3.8542)

In this case, it could also be of interest to profile the parameters (β_1, β_2) and (β_3, β_4), which amounts to respectively testing for nullity of the effects of `spontaneous` and `induced` on the probability of second sterility. Furthermore, in this case it could be of

interest to assess the significance of the overall effect of both spontaneous and induced abortions, which amounts to testing for $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$.

In Table 3.14, we report the results obtained with the modified profile score statistic, along with the corresponding approximation denoted by W^*_{Pu} and with the modified Wald and likelihood ratio statistics. Clearly, in this case all the employed statistics lead to an analogous result, namely that of considering the main effects of induced and/or spontaneous abortions as statistically significant, at least with respect to an approximate significance level of 0.05.

TABLE 3.14: Hypothesis testing for nullity of the effects of `spontaneous`, `induced` and their overall effect β, with respect to the `infert` data

	Degrees of freedom	$\tilde{W}_{Pu}$	$\tilde{W}^*_{Pu}$	$\tilde{W}_{Pe}$	$\tilde{W}_P$
β	4	69.0699	45.5388	34.6281	56.8201
(β_1, β_2)	2	69.1558	48.2024	34.6103	56.6829
(β_3, β_4)	2	22.5631	10.5476	14.0661	19.0974

To better investigate the inferential reliability of the modified score and profile score statistics, we conduct a simulation study by generating $B = 3000$ data sets from model (3.1), setting $(\lambda, \beta) = (\tilde{\lambda}, \tilde{\beta})$, where $\tilde{\lambda}$ is the mean bias-reduced estimate for the strata-specific parameters in λ. The model matrix is kept fixed as that of the original data. Due to the computational effort entailed by the relatively large number of covariates, in this case we could consider an appreciably lower number of simulated replications.

Unfortunately, all of our simulated data sets are affected by complete or quasi-complete data separation, which prevents us from considering standard likelihood-based inference for further comparisons in this simulation study. As assessed by using the `detectseparation` package, in 2982 cases separation occurred exclusively due to covariates associated with λ.

After computing all the modified profile statistics for β, (β_1, β_2), (β_3, β_4) and for each scalar component of β in the true parameter value, we obtain the simulated coverage reported in Table 3.15, considering different nominal approximate significance levels.

From such results, we observe that the modified profile score statistic provides slightly low but not unsatisfying attained coverage when profiling β. Conversely, in accordance with the previous simulation study, the approximate modified profile score statistic shows noticeable inadequacy in terms of coverage. As regards the modified Wald statistic, we notice a clear over-coverage with respect to each nominal confidence level.

TABLE 3.15: Simulated coverage of the modified statistics and of the corresponding profile versions in the logistic regression model, with respect to the simulation study based on the `infert` data and to different nominal significance levels.

Parameter	Statistic	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
β	$\tilde{W}_{Pu}$	0.9850	0.9373	0.8840
	$\tilde{W}^*_{Pu}$	0.9657	0.9010	0.8427
	$\tilde{W}_{Pe}$	0.9983	0.9783	0.9487
	$\tilde{W}_P$	0.9937	0.9583	0.9210
(β_1, β_2)	$\tilde{W}_{Pu}$	0.9870	0.9510	0.9097
	$\tilde{W}^*_{Pu}$	0.9783	0.9363	0.8880
	$\tilde{W}_{Pe}$	0.9960	0.9730	0.9410
	$\tilde{W}_P$	0.9927	0.9647	0.9267
(β_3, β_4)	$\tilde{W}_{Pu}$	0.9850	0.9400	0.8883
	$\tilde{W}^*_{Pu}$	0.9720	0.9193	0.8697
	$\tilde{W}_{Pe}$	0.9967	0.9693	0.9283
	$\tilde{W}_P$	0.9910	0.9570	0.9050
β_1	$\tilde{W}_{Pu}$	0.9937	0.9530	0.9033
	$\tilde{W}^*_{Pu}$	0.9873	0.9517	0.9073
	$\tilde{W}_{Pe}$	0.9950	0.9703	0.9247
	$\tilde{W}_P$	0.9953	0.9620	0.9123
β_2	$\tilde{W}_{Pu}$	0.9910	0.9633	0.9270
	$\tilde{W}^*_{Pu}$	0.9843	0.9487	0.9137
	$\tilde{W}_{Pe}$	0.9953	0.9693	0.9310
	$\tilde{W}_P$	0.9950	0.9683	0.9287
β_3	$\tilde{W}_{Pu}$	0.9883	0.9413	0.8863
	$\tilde{W}^*_{Pu}$	0.9810	0.9390	0.8843
	$\tilde{W}_{Pe}$	0.9960	0.9603	0.9117
	$\tilde{W}_P$	0.9920	0.9497	0.8963
β_4	$\tilde{W}_{Pu}$	0.9880	0.9520	0.9017
	$\tilde{W}^*_{Pu}$	0.9780	0.9400	0.8947
	$\tilde{W}_{Pe}$	0.9947	0.9637	0.9250
	$\tilde{W}_P$	0.9920	0.9580	0.9120

Furthermore, while $\tilde{W}_{Pu}$ provides acceptable results when profiling the overall effect of `spontaneous` and `induced`, the corresponding approximation $\tilde{W}^*_{Pu}$ fails to yield sufficient coverage with respect to each nominal confidence level. The converse case is given by Wald-type inference, where we observe that the corresponding confidence regions result in a too large coverage, or equivalently to conservative hypothesis testing.

Considering scalar parameters of interest, the modified profile score statistic and the corresponding approximation provide a simulated coverage that is quite close to each

nominal one. In contrast, also in this case Wald-type inference seems to show a too conservative behaviour.

To further investigate the simulated distribution of the modified profile score statistic for β and profiling the effects of `spontaneous` and `induced`, we illustrate the quantile-quantile plots in Figure 3.14, comparing the simulated and asymptotic null distributions. Clearly, the modified score function shows considerable cohesion with respect to its

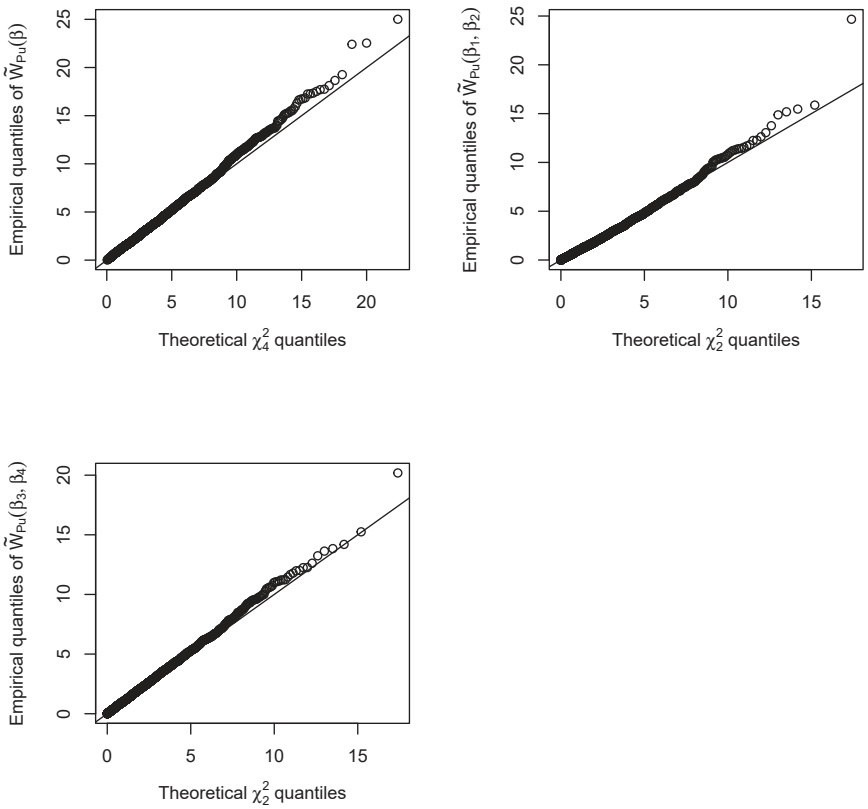

FIGURE 3.14: Simulation study based on the `infert` data. Quantile-quantile plots comparing the simulated distribution of $\tilde{W}_{Pu}$ for β, (β_1, β_2) and (β_3, β_4) to the corresponding asymptotic null distributions.

asymptotic distribution, although the simulated distribution of $\tilde{W}_{Pu}(\beta)$ has a slightly heavier tail than expected, resulting in under-coverage as shown in Table 3.15.

On the contrary, the simulated distribution of $\tilde{W}_{Pe}$ when profiling β and the main effects of `spontaneous` and `induced`, as displayed in Figure 3.15, shows a more evident departure from the corresponding asymptotic null distributions, namely with lighter right tails than expected. Such discrepancy entails the over-coverage of Wald-type inference, consistently to the results in Table 3.15.

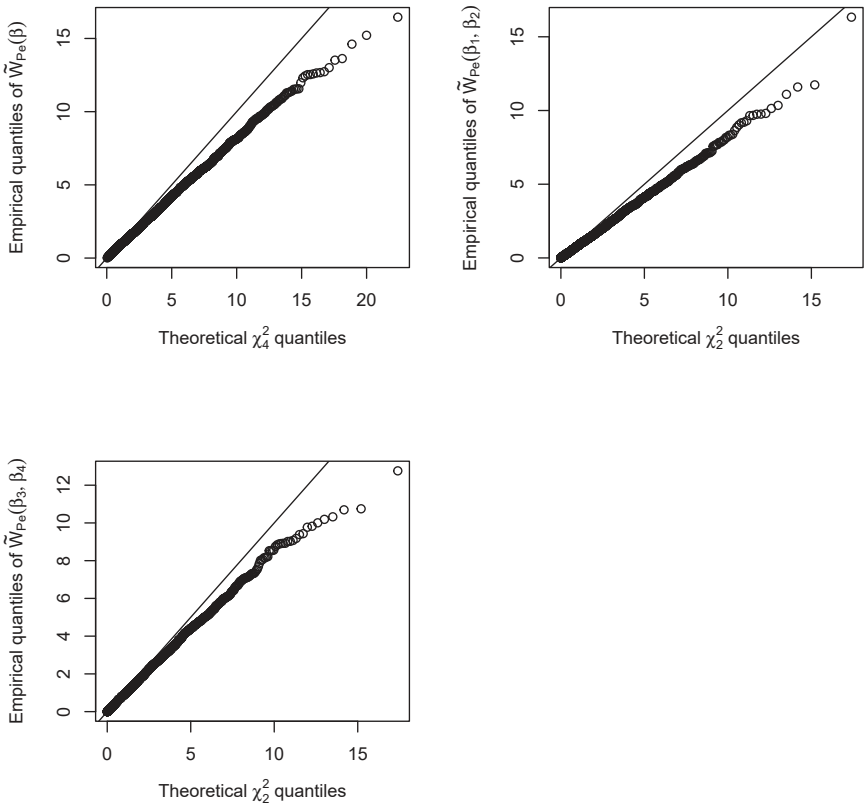

FIGURE 3.15: Simulation study based on the `infert` data. Quantile-quantile plots comparing the simulated distribution of $\tilde{W}_{Pe}$ for β, (β_1, β_2) and (β_3, β_4) to the corresponding asymptotic null distributions.

Focusing on scalar parameters of interest, we consider the simulated distribution of $\tilde{W}_{Pu}$ for each component of the structural parameter β, as illustrated in Figure 3.16. In accordance with the simulated coverage shown in Table 3.15, there seems to be no substantial departure from the χ^2_1 distribution, entailing an overall reliable inference with respect to scalar parameters of interest.

Conversely, addressing in more detail the simulated distribution of the modified Wald statistic as in Figure 3.17, we notice a comparably worse performance in terms of cohesion to the χ^2_1 distribution. Clearly, the presence of a lighter right tail than expected may result in a too conservative inference on the model coefficients, in accordance with the over-covering confidence intervals as reported in Table 3.15.

This simulation study allows us to regard the modified profile score statistic as reliable from an inferential perspective, even considering a real data setting with a relatively high-dimensional nuisance parameter.

In contrast to the simulation study based on the `endometrial` data, in this case the

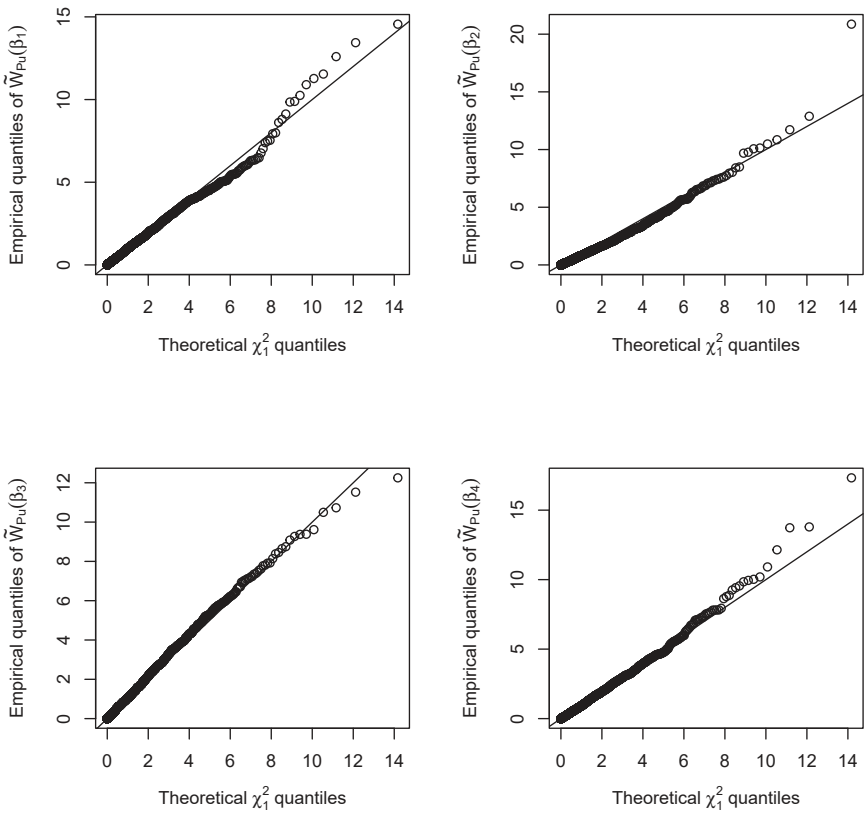

FIGURE 3.16: Simulation study based on the `infert` data. Quantile-quantile plots comparing the simulated distribution of $\tilde{W}_{Pu}$ for each scalar component of β to the corresponding asymptotic null distribution.

presence of exclusively complete or quasi-complete separated data sets in our simulation does not seem to affect the inferential quality provided by the modified profile score statistic. Perhaps, the lack of such a problem could be explained by the fact that, in this case, the data-separating covariates are mostly associated with the incidental parameter λ and not with β, on which we focused our attention.

As regards the approximate modified profile score statistic, in this case it has proven unsuitable in the presence of non-scalar parameters of interest, with particular regard to β. Such result seems to agree with that of the previous simulation study, namely the effect of the increasing dimension of the parameter of interest, as was shown in Figure 3.13. In contrast to the simulation study based on `endometrial`, where the inadequacy of W^*_{Pu} was evident even considering scalar parameters of interest, in this case such a problem does not seem to occur.

For what concerns the modified profile likelihood ratio statistic, as expected from the previous simulation studies, it clearly provides the most accurate inference on the

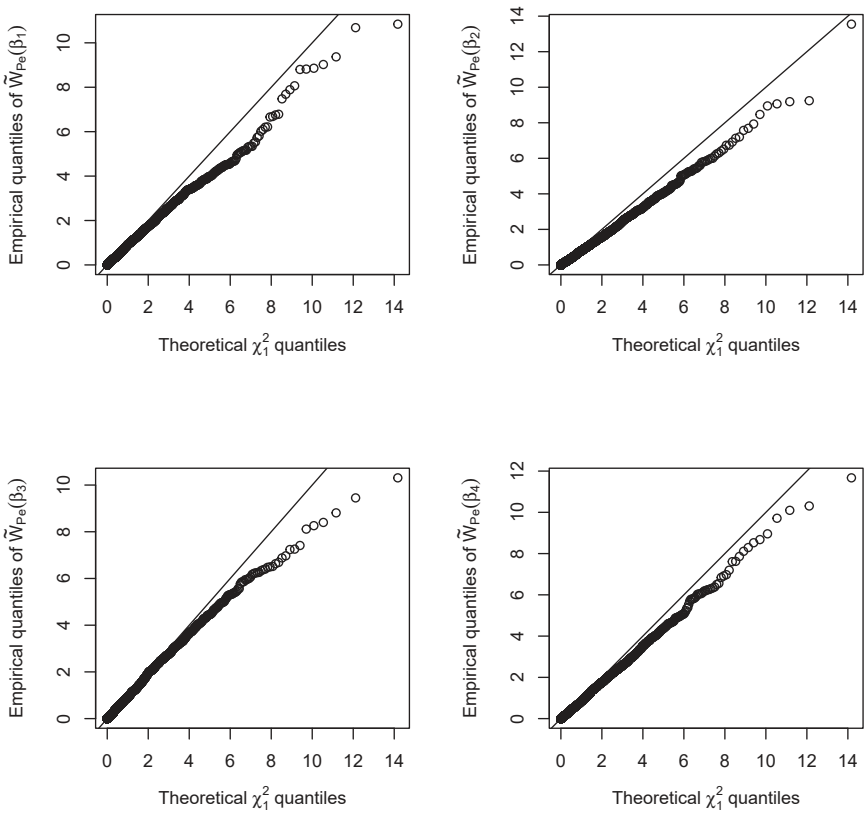

FIGURE 3.17: Simulation study based on the `infert` data. Quantile-quantile plots comparing the simulated distribution of $\tilde{W}_{Pe}$ for each scalar component of β to the corresponding asymptotic null distribution.

model parameters, with respect to the corresponding simulated coverage in Table 3.15. A considerably accurate performance is provided in all cases, also when profiling β, where in contrast the modified profile score statistic underperforms. Nonetheless, in more general settings we should also consider that the existence of a penalized likelihood as (1.23) is no longer guaranteed, considering for instance non-canonical link functions, such as the probit.

Conclusion

Inference within the framework of bias reduction can provide a valuable alternative to standard likelihood inference, as also addressed in this work. For this reason, further research towards suitable approximate pivotal quantities based on bias reduction can be of particular interest.

The foregoing simulation studies highlight that the modified score statistic, along with its profile version, can provide a valuable alternative to Wald-type inference within the framework of bias reduction. Although the modified profile likelihood ratio statistic can yield more reliable inference, such results are of interest for models in which the penalized log-likelihood as (1.23) does not exist.

More specifically, the previous results suggest that the proposed statistic should be preferably used in the presence of relatively low-dimensional parameters of interest, for instance when constructing confidence intervals. Conversely, inference on higher-dimensional parameters should be carried out with caution due to the potentially high type-I error probability.

Furthermore, in the presence of binary data affected by complete or quasi-complete separation, our results suggest that the modified profile score statistic should be used when profiling parameters that are not involved in the data separation. On the contrary, inference involving such parameters needs to be carried out with additional care when using the proposed statistic.

As far as the approximate modified profile score statistic is concerned, our simulations illustrate that it can provide a reliable and fast alternative to the corresponding exact version, provided that we consider low-dimensional parameters of interest. This may not be too strict of a limitation, since the computational advantage of the approximation proves useful when constructing confidence intervals.

Nonetheless, such approximation should be used with caution in the presence of constrained bias-reduced estimates with too prominent curvature or in the case of data separation. In the latter situation, even inference on parameters that are not associated with the data-separating hyperplane could be misleading.

Surely, our work has limitations that needs to be addressed. To begin with, throughout our simulation studies we mainly considered binary logistic regression models. This is motivated by the fact that such models are notorious for the non-existence of maximum likelihood estimate and, moreover, we could provide a comparison including the modified likelihood ratio statistic.

Nonetheless, additional simulation studies could be required when considering non-canonical link functions and binomial regression models with non-unit weights. Furthermore, other studies could be of interest with respect to regression models with Poisson response. Our implementation, reported in the Appendix, could be useful in this respect, considering that it allows to carry out the inferential procedures addressed in our work with respect to most of the generalized linear models with unit dispersion parameter.

In contrast, a further implementation effort is needed with respect to generalized linear models with unknown dispersion parameter, an example being gamma regression. Of course, further simulation studies could prove useful to address such problems as well.

Another limitation of our work is that we exclusively considered the case of mean bias reduction, due to both ease of discussion and to the corresponding invariance properties, which can be of more interest in practical settings. Nevertheless, it is clear that further simulation studies considering mean bias reduction and the mixed adjustment approach could provide a further understanding on the modified score statistic.

Further research could focus on studying the performance of

$$\tilde{W}_u^{\dagger}(\theta) = \tilde{U}(\theta)^{\top} i(\tilde{\theta})^{-1} \tilde{U}(\theta),$$

namely the modified score statistic with $i(\theta)$ replaced with $i(\tilde{\theta})$. Considering the partition of the model parameter $\theta = (\psi, \lambda)$, the corresponding profile version for ψ is

$$\tilde{W}_{Pu}^{\dagger}(\psi) = \tilde{U}_{\psi}(\psi, \tilde{\lambda}_{\psi})^{\top} i^{\psi\psi}(\tilde{\theta}) \tilde{U}_{\psi}(\psi, \tilde{\lambda}_{\psi}).$$

The replacement of the information matrix computed in θ could have noticeable effects on the performance of such statistics. The implementation provided in the Appendix also allows to use such a replacement, which could be useful for simulation studies in this respect.

Further investigation is required to study the theoretical asymptotic distribution of $\tilde{r}_P(\psi)$, $\tilde{r}_{Pe}(\psi)$ and $\tilde{r}_{Pu}(\psi)$, when ψ is involved in data separation. As a matter of fact, the results in Figure 3.8 point out a potential issue in the distribution of the signed

modified statistics based on bias reduction. Namely, the empirical distribution of such statistics resembles a mixture of two different distributions instead of the asymptotic standard Gaussian one. Perhaps, this phenomenon could be explained by the fact that the standard asymptotic results hold no longer in case of data separation, since the maximum likelihood estimator is located in the boundary of the parameter space with non-negligible probability.

Appendix

```
### Libraries
require(brglm2)

###================================================
### 1. Test statistic
###================================================
### Modified score
modified.score.fun = function(object, beta){
    y = object$y
    x = model.matrix(object)
    nobs = nrow(x)
    nvars = ncol(x)
    linkfun = object$family$linkfun
    linkinv = object$family$linkinv
    mu.eta = object$family$mu.eta
    d2mu.deta = object$family$d2mu.deta
    variance = object$family$variance

    etas = drop(x %*%  beta)
    mus = linkinv(etas)
    d1mus = mu.eta(etas)
    d2mus = d2mu.deta(etas)
    varmus <- variance(mus)
    weights = object$prior.weights
    working_weights <- weights * d1mus^2 / varmus

    score_components <- weights * d1mus  * (y - mus) / varmus * x
```

```
    ## Score function
    s_beta = .colSums(score_components, nobs, nvars, TRUE)

    ## Mean bias reduction adjustment:
    ## firstly obtain the "hat" values. As in the original code from
    ## ‘brglmFit‘, we use the QR decomposition to compute them
    wx <- sqrt(working_weights) * x
    qr_decomposition <- qr(wx)
    Qmat <- qr.Q(qr_decomposition)
    hatvalues = .rowSums(Qmat * Qmat, nobs, nvars, TRUE)

    ## Adjustment (mean bias reduction)
    A_beta = .colSums(0.5 * hatvalues * d2mus/d1mus * x,
    nobs, nvars, TRUE)

    ## Modified score
    return(s_beta + A_beta)
}

### Expected information matrix
i.fun = function(object, beta, inverse = FALSE){
    y = object$y
    x = model.matrix(object)
    nobs = nrow(x)
    nvars = ncol(x)
    linkfun = object$family$linkfun
    linkinv = object$family$linkinv
    mu.eta = object$family$mu.eta
    d2mu.deta = object$family$d2mu.deta
    variance = object$family$variance

    etas = drop(x %*%  beta)
    mus = linkinv(etas)
    d1mus = mu.eta(etas)
    varmus <- variance(mus)
    weights = object$prior.weights
```

```
    working_weights <- weights * d1mus^2 / varmus

    ## As in the original code from brglmFit, use QR decomposition
    wx <- sqrt(working_weights) * x
    qr_decomposition <- qr(wx)
    R_matrix <- qr.R(qr_decomposition)
    if (inverse) {
        ## return(dispersion * tcrossprod(solve(R_matrix)))
        return(chol2inv(R_matrix))
    }else{
        return(crossprod(R_matrix))
    }
}

### Jacobian matrix of the modified score function
### (we derive it numerically)
jtilde.fun = function(object, beta){
    -numDeriv::jacobian(function(x) modified.score.fun(object, x),
    x = beta)
}

### Global test
Wu.fun = function(object, beta, type = "variable"){
    s_beta = modified.score.fun(object, beta)
    i_betabeta = switch(type,
    variable = i.fun(object, beta, TRUE),
    local = vcov(object))
    return(drop(crossprod(s_beta, i_betabeta) %*% s_beta))
}

### Profile test
### 1. Approximate constrained estimate
nuisance_approximate = function(object, parm, beta, jtilde = NULL,
jtilde.inv = NULL){
    ## jtilde.inv refers to j_{\lambda \lambda}^{-1} computed in
    ## the bias-reduced estimate
```

```
    if(is.null(jtilde)) jtilde = jtilde.fun(object, object$coefficients)
    if(is.null(jtilde.inv)) jtilde.inv = solve(jtilde[-parm,-parm])
    interest.br = object$coefficients[parm]
    nuisance.br = object$coefficients[-parm]
    drop(nuisance.br + jtilde.inv %*% jtilde[-parm,parm] %*%
    (interest.br - beta))
}

### 2. Exact constrained estimate
nuisance_exact = function(object, parm, beta, init = NULL,
quasifisher = TRUE, maxit = 500){
    if(is.null(init)){
        init = nuisance_approximate(object, parm, beta)
    }

    if(!quasifisher){
        out = nleqslv::nleqslv(init,
        function(x){
            beta.all = object$coefficients
            beta.all[-parm] = x # nuisance
            beta.all[parm] = beta # interest
            modified.score.fun(object,
            beta.all)[-parm]
        },
        method = "Newton",
        global = "none",
        control = list(maxit = maxit,
        allowSingular = TRUE))
    }

    else{
        out = nleqslv::nleqslv(init,
        function(x){
            beta.all = object$coefficients
            beta.all[-parm] = x # nuisance
            beta.all[parm] = beta # interest
```

```
            modified.score.fun(object,
            beta.all)[-parm]
        },
        jac = function(x){
            beta.all = object$coefficients
            beta.all[-parm] = x # nuisance
            beta.all[parm] = beta # interest
            -i.fun(object,
            beta.all)[-parm, -parm]
        },
        method = "Newton",
        global = "none",
        control = list(maxit = maxit,
        allowSingular = TRUE))
    }
    if(!(out$termcd %in% c(1,2))){
        print(out$message)
        return(NA)
    }
    out$x
}

### 3. Constrained estimate computed by refitting the model.
###    It does not yield the exact constrained solution,
###    but it may provide a more stable starting point than
###    the linear approximation
nuisance_refit = function(object, parm, beta){
    x = model.matrix(object)
    y = object$y
    beta.est = object$coefficients
    offset = drop(as.matrix(x[,parm]) %*% beta)
    nuisance = glm(y ~ x[,-parm] - 1, method = object$method,
    type = object$type, family = object$family,
    offset = offset,
    control = list(maxit = 10000))$coefficients
    names(nuisance) = names(beta.est[-parm])
```

```
    return(nuisance)
}

### Modified profile score test statistic
Wpu.fun = function(object, parm, beta, approximate = FALSE,
init = NULL, quasifisher = TRUE,
maxit = 500, jtilde = NULL, jtilde.inv = NULL,
type = "variable"){
    ## type = "variable" means that the expected
    ##        information is compute in the parameter value
    ## type = "local" computes the expected information in
    ##        the bias-reduced estimate, but unbounded
    ##        confidence intervals may occur

    if(is.character(parm)){
        parm = which(names(object$coefficients) == parm)
    }
    if(is.logical(parm)){
        parm = which(parm)
    }
    if(is.null(jtilde)) jtilde = jtilde.fun(object,
    object$coefficients)
    if(is.null(jtilde.inv)) jtilde.inv = solve(jtilde[-parm,-parm])
    if(approximate){
        nuisance = nuisance_approximate(object,
        parm,
        beta,
        jtilde,
        jtilde.inv)
    }
    else{
        if(is.null(init)){
            init = nuisance_approximate(object,
            parm,
            beta,
            jtilde,
```

```
            jtilde.inv)
        }
        if(any(init == "refit")){
            init = nuisance_refit(object, parm, beta)
        }
        nuisance = nuisance_exact(object, parm, beta,
        init, quasifisher, maxit)
        if(any(is.na(nuisance)) & any(init != "refit")){
            ## try a better but slower initial point
            init = nuisance_refit(object, parm, beta)
            nuisance = nuisance_exact(object, parm, beta,
            init, quasifisher, maxit)
        }
    }
    if(any(is.na(nuisance))) stop(
    "A numerical error occurred. Please try a better inital point"
    ) ## in case of errors
    beta.all = object$coefficients
    beta.all[parm] = beta
    beta.all[-parm] = nuisance

    s_interest = modified.score.fun(object, beta.all)[parm]
    i_interest = switch(type,
    variable = i.fun(object,
    beta.all,
    TRUE)[parm, parm],
    local = vcov(object)[parm, parm])
    return(drop(crossprod(s_interest, i_interest) %*% s_interest))
}

###=============================================================
### 2. Confidence intervals
###=============================================================
### Modified profile score test statistic (one parameter)
rpu.fun = function(object, parm, beta, approximate = FALSE,
init = NULL, quasifisher = TRUE, maxit = 500,
```

```
jtilde = NULL, jtilde.inv = NULL,
type = "variable"){

    if(is.character(parm)){
        parm = which(names(object$coefficients) == parm)
    }

    if(is.logical(parm)){
        parm = which(parm)
    }

    if(approximate){
        nuisance = nuisance_approximate(object,
        parm,
        beta,
        jtilde,
        jtilde.inv)
    }
    else{
        if(is.null(init)){
            init = nuisance_approximate(object,
            parm,
            beta,
            jtilde,
            jtilde.inv)
        }
        if(any(init == "refit")){
            init = nuisance_refit(object, parm, beta)
        }
        nuisance = nuisance_exact(object, parm, beta,
        init, quasifisher, maxit)
        if(any(is.na(nuisance)) & any(init != "refit")){
            ## try a better but slower initial point
            init = nuisance_refit(object, parm, beta)
            nuisance = nuisance_exact(object, parm, beta,
            init, quasifisher, maxit)
```

```
        }
    }

    if(any(is.na(nuisance))) stop(
    "A numerical error occurred. Please try a better initalization"
    ) ## in case of errors
    beta.all = object$coefficients
    beta.all[parm] = beta
    beta.all[-parm] = nuisance

    s_interest = modified.score.fun(object, beta.all)[parm]
    i_interest = switch(type,
    variable = i.fun(object,
    beta.all,
    TRUE)[parm, parm],
    local = vcov(object)[parm, parm])
    return(sqrt(i_interest)*s_interest)
}

### Confint
confint.brglmFit = function(object, parm, level = 0.95,
approximate = FALSE, init = NULL,
quasifisher = TRUE, maxit = 500,
type = "variable", trace = TRUE){
    if(missing(parm)){
        parm = seq.int(length(object$coefficients))
    }
    if(is.character(parm)){
        parm = which(names(object$coefficients) == parm)
    }
    if(is.logical(parm)){
        parm = which(parm)
    }
    alpha = 1-level
    out = matrix(nrow = length(parm), ncol = 2)
    colnames(out) = paste(round(c(alpha/2, 1-alpha/2) * 100,
```

```
digits = 1), "%")
rownames(out) = names(object$coefficients[parm])
rownumber = 1
jtilde = jtilde.fun(object, object$coefficients)
if(trace) cat("Profiling ...\n")
for(i in parm){
    mle = object$coefficients[i]
    std.err = sqrt(summary(object)$cov.unscaled[i,i])
    jtilde.inv = solve(jtilde[-i, -i])

    ## initialize search space considering 3 standard errors
    search = c(3, 3)
    grid.ok = rep(FALSE, 2)

    ## check if the grid contains the required quantiles
    while(any(!grid.ok)){
        grid = seq(mle - search[1] * std.err,
        mle + search[2] * std.err,
        length = 100)
        rval.limits = c(rpu.fun(object, i, grid[1], approximate,
        init, quasifisher, maxit, jtilde,
        jtilde.inv,
        type),
        rpu.fun(object, i, grid[100], approximate,
        init, quasifisher, maxit, jtilde,
        jtilde.inv,
        type))
        grid.ok = rval.limits^2 > qchisq(1-alpha, df = 1)
        ## extend search space
        search[!grid.ok] = search[!grid.ok] + 1
        if(any(search > 15)){
            break
        }
    }
    rval = numeric(100)
    rval[1] = rval.limits[1]
```

```
rval[100] = rval.limits[2]
rval[2:99] = sapply(grid[2:99],
function(x){
    rpu.fun(object, i, x,
    approximate, init,
    quasifisher, maxit, jtilde,
    jtilde.inv, type)
})
if(approximate){
    ## linear approximation may yield non-monotonic
    ## signed modified profile
    ## score, therefore we cannot use interpolation
    ## through splines
    out[rownumber,] = c(-Inf, Inf)
    if(grid.ok[1]){
        out[rownumber, 1] = uniroot(function(x){
            Wpu.fun(object, i, x,
            approximate, init,
            quasifisher, maxit,
            jtilde, jtilde.inv,
            type) - qchisq(1-alpha,
            df = 1)
        },
        interval = c(grid[1], mle))$root

    }
    if(grid.ok[2]){
        out[rownumber, 2] = uniroot(function(x){
            Wpu.fun(object, i, x,
            approximate, init,
            quasifisher, maxit,
            jtilde, jtilde.inv,
            type) - qchisq(1-alpha,
            df = 1)
        },
        interval = c(mle, grid[100]))$root
```

```
            }
        }
        else{
            mod = pspline::sm.spline(rval, grid)
            out[rownumber,] = c(-Inf, Inf)
            out[rownumber, grid.ok] = predict(mod,
            qnorm(c(1-alpha/2,
            alpha/2)[grid.ok]))
        }
        if(trace){
            cat("    Profiled ", rownumber, " out of ",
            length(parm), "\n")
        }
        rownumber = rownumber + 1
    }
    return(out)
}
```

Bibliography

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley.

Albert, A. & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1–10.

Azzalini, A. (1996). *Statistical inference: based on the likelihood*. Chapman & Hall, 1st ed.

Candès, E. J. & Sur, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics* **48**, 27 – 42.

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

Cox, D. & Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall/CRC, 1st ed.

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

Emireni, G. (2004). *Verosimiglianza profilo generalizzata*. Tesi di Laurea, Facoltà di Scienze Statistiche, Università degli Studi di Padova.

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

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)* **46**, 149–192.

Hasselman, B. (2023). *nleqslv: Solve Systems of Nonlinear Equations*. R package version 3.3.4.

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

Kenne Pagui, E. C., Salvan, A. & Sartori, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika* **104**.

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

Kosmidis, I. (2023). *brglm2: Bias Reduction in Generalized Linear Models*. R package version 0.9.

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

Kosmidis, I. & Firth, D. (2020). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika* **108**, 71–82.

Kosmidis, I., Kenne Pagui, E. C. & Sartori, N. (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing* **30**, 43–59.

Kosmidis, I., Schumacher, D. & Schwendinger, F. (2022). *detectseparation: Detect and Check for Separation and Infinite Maximum Likelihood Estimates*. R package version 0.3.

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

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Methodological)* **11**, 68–84.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Sartori, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *Journal of Statistical Planning and Inference* **136**, 4259–4275.

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

Sur, P. & Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences* **116**, 14516–14525.

Trichopoulos, D., Handanos, N., Danezis, J., Kalandidi, A. & Kalapothaki, V. (1976). Induced abortion and secondary infertility. *BJOG: An International Journal of Obstetrics & Gynaecology* **83**, 645–650.