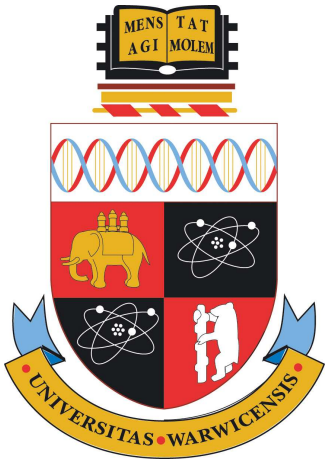


Bias Reduction in Exponential Family Nonlinear Models

by

Ioannis Kosmidis

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

August 2007

THE UNIVERSITY OF
WARWICK

To Stella

Πειθαρχία, νά ἡ ἀνώτατη ἀρετή. Ἔτσι μονάχα σοζυγιάζεται ἡ δύναμη μὲ τὴν ἐπιθυμία καὶ καρπίζει ἡ προσπάθεια τοῦ ἀνθρώπου.

N. KAZANTZAKHΣ,
'AΣKHTIKH'

Discipline is the highest of all virtues. Only so may strength and desire be counterbalanced and the endeavors of man bear fruit.

N. KAZANTZAKIS,
'THE SAVIORS OF GOD:
SPIRITUAL EXERCISES'
Translated by Kimon Friar

Contents

Acknowledgements

It would be difficult to fully express my deep gratitude to my supervisor Professor David Firth. His guidance, help, inspiration, great intuition and willingness to try to explain things simply and clearly, helped to make the beginning of my journey in the statistical science enjoyable and interesting. Furthermore, I would like to thank Professor David Firth for his continuous encouragement during the time of writing and for proof reading the final draft of the current thesis, providing valuable advice on how to improve the presentation and refine the language.

I am really grateful to all my friends and generally to all the members of the department of Statistics of the University of Warwick for providing a stimulating and friendly environment during my PhD studies. I am also indebted to the friends and members of the Department of Statistics of the Athens University of Economics and Business, where I was first taught statistics and I was provided with a background that played a vital role in my graduate studies.

I would also like to thank my flatmates Konstantinos and Konstantinos, for their understanding and support. I am particularly grateful to Konstantinos Kritsis for reading certain parts of the thesis and for providing considerable help on linguistic matters.

I wish I had the words to thank my parents, my brother and my sister. Without their love, support, encouragement and understanding, I would not have made it this far.

I am also grateful to the University of Warwick and to the EPSRC (Engineering and Physical Sciences Research Council) for the financial support they provided during my PhD studies.

Computing environment and typeset

For the computational requirements of the thesis, the *R language* (R Development Core Team, 2007) was used and all the figures were created using R's *PostScript* device. The thesis is typeset using LaTeX under the *TeX live* distribution (`www.tug.org/texlive`). The aid provided by the excellent LaTeX editor *Kile* (`kile.sourceforge.net`) is beyond description. Many thanks to all the people involved in the development of the above software.

Declaration

I hereby declare that the contents of the current thesis are based upon my own research in accordance with the regulations of the University of Warwick. The results, figures, tables and generally any included material is original, except otherwise indicated by reference to other authors or organizations. The current thesis has not been submitted for examination at any other university than the University of Warwick.

Abstract

The modified-score functions approach to bias reduction (Firth, 1993) is continually gaining in popularity (e.g. Mehrabi & Matthews, 1995; Pettitt et al., 1998; Heinze & Schemper, 2002; Bull et al., 2002; Zorn, 2005; Sartori, 2006; Bull et al., 2007), because of the superior properties of the bias-reduced estimator over the traditional maximum likelihood estimator, particularly in models for categorical responses. Most of the activity is noted for logistic regression, where the bias-reduction method neatly corresponds to penalization of the likelihood by Jeffreys prior and the bias-reduced estimates are always finite and beneficially shrink towards the origin.

The recent applied and methodological interest in the bias-reduction method motivates the current thesis and the aim is to explore the nature and widen the applicability of the method, identifying cases where bias reduction is beneficial. Particularly, the current thesis focuses on the following three targets:

i) To explore the nature of the bias-reducing modifications to the efficient scores and to obtain results that facilitate the application and the theoretical assessment of the bias-reduction method.

ii) To establish theoretically that the bias-reduction method should be considered as an improvement over traditional ML for logistic regressions.

iii) To deviate from the flat exponential family and explore the effect of bias reduction in some commonly used curved models for categorical responses.

Notation

Unless otherwise stated, the following notational conventions are used throughout the current thesis. For the reader's convenience, in addition to their statement here, they are also described in their first occurrence in each chapter.

$\Re$	The set of real numbers
$\Re^p$	The p-dimensional Euclidean space
$\xrightarrow{p}$	Converges in probability
$\xrightarrow{d}$	Converges in distribution
$\|x\|$, $x \in \Re$	absolute value of x
$\|\|x\|\|$	the norm of x in the domain of x
$\mathrm{E}(X)$, $\mathrm{Var}(X)$, $\mathrm{Cov}(X)$	expected value, variance, covariance,
$\mathrm{Cum}_r(X)$	r-th order cumulant ($r = 1, \ldots, n$)
A^T	the transpose of a matrix A
A^{-1}	the inverse of a square matrix A
$\det A$	the determinant of a square matrix A
$\mathrm{trace}\, A$	the trace of a square matrix A
$\mathrm{diag}(a)$	the diagonal matrix with diagonal elements the components of some vector a

$\mathrm{diag}\{a_s; s = 1, \ldots, p\}$	$\mathrm{diag}\, a$, $a = (a_1, a_2, \ldots, a_p)$
1_p	the $p \times p$ identity matrix
J_p	A $p \times p$ matrix of ones
L_p	A $p \times 1$ vector of ones
0_p	A $p \times 1$ vector of zeros
$A \otimes B$	the Kronecker product of matrix A with matrix B
$\nabla_x f(x)$, $f : \Re^p \to \Re$	the gradient of f with respect to x, ie. i.e., $\nabla_x f(x) = (\partial f(x)/\partial x_1, \partial f(x)/\partial x_2, \ldots, \partial f(x)/\partial x_p)$

Abbreviations

The following abbreviations are used in the main text. In addition to their statement here, for the readers convenience, they are re-introduced in each chapter.

BR	**b**ias-**r**educed
BC	**b**ias-**c**orrected
GLM	**g**eneralized **l**inear **m**odel
IWLS	**i**terative re-**w**eighted **l**east **s**quares
IGLS	**i**terative **g**eneralized **l**east **s**quares
LR	**l**ikelihood **r**atio
ML	**m**aximum **l**ikelihood
MSE	**m**ean **s**quared **e**rror
PLR	**p**enalized-**l**ikelihood **r**atio
MPL	**m**aximum **p**enalized **l**ikelihood

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Towards the removal of bias: A brief review

Bias in estimation is a common concern of practitioners and researchers in statistics. Its magnitude plays an important role in estimation and if large it can result in potentially misleading inferences. In this perspective, the maximum likelihood (ML) estimator has asymptotically desirable behaviour. Given the regularity of the statistical problem (see for example, Cox & Hinkley, 1974, §9.1, for an account on regularity conditions for ML), it can be shown that the ML estimator is asymptotically unbiased with a leading term in its bias expansion of order $\mathcal{O}(n^{-1})$. For example, in the case of quantal response models, Sowden (1972) studied the bias of estimators based on the ML, the minimum chi-squared and the modified minimum chi-squared methods. In this study, an elegant expression that connects the first-order bias terms of the resultant estimators is derived and it is illustrated that the first-order bias term of the ML estimator is the smallest among the three alternatives. Sowden (1972) only considered the case where the probabilities are linked with a linear combination of the model parameters through the inverse of the standard normal distribution function, but it is correctly suggested that the same result may extend to different link functions. However, the first-order bias term of the ML estimator could be large for small or even moderate sample sizes. There has been much work on the ways which could be used for reducing the bias of the ML estimator and in view of a part of the substantial literature that is related to this task, we can distinguish two classes of methods that from now on are referred to as "bias correction" and "bias reduction".

1.1.1 Bias correction

The word "correction" refers to the fact that all the bias-correction methods are based on the following two step calculation:

i) Obtain the first-order bias term of the ML estimator.

ii) Subtract it from the ML estimates.

This way, the ML estimates and the first-order bias of the ML estimator evaluated at the ML estimates are the building blocks of the bias-corrected estimates. The expected value of the ML estimator $\hat{\beta}$ for the parameters β of a parametric model, can be generally expressed as

$$\mathrm{E}\left(\hat{\beta}\right) = \beta_0 + \frac{b_1(\beta_0)}{n} + \frac{b_2(\beta_0)}{n^2} + \frac{b_3(\beta_0)}{n^3} + \dots ,$$

where n is some measure of the units of information — usually the sample size, β_0 is the true but unknown parameter value and b_t $(t = 1, 2, \dots)$ are $\mathcal{O}(1)$ functions of β, which can be explicitly obtained once the model is specified. So, the simple rearrangement

$$\mathrm{E}\left(\hat{\beta}\right) - \frac{b_1(\beta_0)}{n} = \beta_0 + \frac{b_2(\beta_0)}{n^2} + \frac{b_3(\beta_0)}{n^3} + \dots ,$$

corrects the bias of $\hat{\beta}$ up to order $\mathcal{O}(n^{-1})$. This is the line of argument behind the corrective methods. Cox & Snell (1968, §3) derived the expression for $b_1(\beta_0)/n$ for a very general family of models. Anderson & Richardson (1979) and Schaefer (1983), based upon the results in Cox & Snell (1968), calculated the first-order bias terms for logistic regressions and obtained the bias-corrected estimates. Both studies conclude that bias-correction is desirable because for such models the bias of the ML estimator is large for small and moderate sample sizes. They also note that, in that particular case, the mean squared error (MSE) beneficially shrinks along with the bias of the estimator. Cordeiro & McCullagh (1991) extended these results and treated bias-correction for the class of generalized linear models (GLMs, Nelder & Wedderburn, 1972). They showed that bias-correction can be achieved by means of a supplementary re-weighted least squares iteration and gave several interesting results on the behaviour of the bias in binomial-response models. These results demonstrate the beneficial — in terms of bias and MSE — shrinkage of the bias-corrected estimates towards the origin of the scale imposed by the link function.

However, the bias-corrected estimates depend upon the finiteness (existence, in the terminology in Albert & Anderson, 1984) of the ML estimates. By definition, the bias-corrected estimates are undefined when the ML estimates are infinite. This is the case for many categorical-response models and is associated with the configuration of zero observations for the response (Albert & Anderson, 1984; Santner & Duffy, 1986; Lesaffre & Albert, 1989, study and classify such configurations for logistic regression models). Further, for small sample sizes the bias-correction method tends to correct beyond the true parameter value. This is illustrated through the empirical studies in Bull et al. (1997) where they compare bias correction with a bias-reduction method for logistic regressions.

1.1.2 Bias reduction

The main difference between bias correction and bias-reduction methods is that the latter do not directly depend on the ML estimates. Instead, new estimators are derived which are known to have smaller or even zero first-order term in the asymptotic expansion of their bias. In this sense the nature of these methods is bias-preventive rather than bias-corrective. A popular estimator of this kind is the Haldane estimator (Haldane, 1956). If

y is the observed number of successes in a binomial trial with total m and probability of success π, the ML estimator of the log-odds $\beta = \log(\pi/(1-\pi))$ is

$$\hat{\beta} = \log \frac{y}{m-y} .$$

Haldane (1956) showed that a simple replacement of y with $y^* = y + 1/2$ and m with $m^* = m + 1$ in the above expression results in an estimator for β which is free of the first-order bias term (see also Cox & Snell, 1989, §2.1.6).

1.1.2.1 Jackknife estimators

Quenouille (1956) was the first to develop a bias-reduction method that is applicable to general families of distributions. This is the jackknife procedure which aims the removal of bias terms up to a specified order. An important reference point for this method is Farewell (1978), who shows that the jackknife estimator can be improved by the reflection of any special structure of the data (for example, fixed totals for contingency tables, fixed sample size ratio in two sample problems, etc.) directly in the jackknife calculations. However, if the ML estimator is not in closed form, jackknifing can become expensive because the ML estimates have to be obtained iteratively for each of all the possible subsets of the sample according to the partitioning scheme considered. Further, several considerations have to be made in cases where the ML estimates for a subset of the sample are infinite. In the case of logistic regression, Bull et al. (1997, §3.2) deal with this problem but in a rather ad-hoc way. As seen therein, these methods reduce the first-order bias term but they do not eliminate it.

1.1.2.2 Estimators based on modified score functions

For a very special scalar-parameter item response theory model, Warm (1989) derives an alternative bias-reduction method that is free from the defects of bias-correction and jackknifing. Based on a conjecture which is proved later in the same paper, Warm (1989) gives the form for a modified score function which results in an estimator with bias of order $o(n^{-1})$ and notices the possible extensions to more general families.

Starting from a rather different point than Warm (1989) and based upon formal asymptotic arguments for regular families, Firth (1993) developed a general method for removing the first-order term in the asymptotic expansion of the bias of the ML estimator. The efficient score functions are appropriately modified so that the roots of the resultant modified score equations result in first-order unbiased estimators. He showed that for exponential families in canonical parameterization, the method reduces to penalization of the likelihood function by the Jeffreys invariant prior (Jeffreys, 1946). The application of the method in generalized linear models (GLMs) with canonical link is studied in Firth (1992a,b), and emphasis is given on the properties of the resultant estimator in some special but important cases such as binomial logistic regression models, Poisson log-linear models and Gamma reciprocal-linear models. For these models, it is demonstrated how the bias-reduction method can be implemented by appending appropriate flattening quantities to the responses, at each step of the iterative re-weighted least squares (IWLS) fitting procedure.

However, the main advantage of the method is not the facility of obtaining estimates, but the properties that the bias-reduced estimator can have for specific models.

Heinze & Schemper (2002) and Zorn (2005) studied the behaviour of the bias-reduced estimator in the canonical case of binary-response logistic regression. Based on empirical findings they note that the bias-reduced estimates are always finite, even in cases where the ML estimates are infinite. They also indicate that the bias-reduced estimates shrink towards the origin of the logistic scale. Bull et al. (2002), through simulation studies, extended the conclusions in Heinze & Schemper (2002) to multinomial responses. They also performed an empirical comparison of the bias-reduced estimator with estimators based on jackknifing and on bias correction. This comparison concluded in favour of the estimator based on the modified scores, both in terms of bias and MSE. Similar conclusions can be found in Mehrabi & Matthews (1995) where the modified score functions are used for estimation in a scalar-parameter complementary log-log model with non-linear predictor. Heinze & Schemper (2002) and Bull et al. (2007) went further and illustrated that confidence intervals based on the ratio of penalized likelihoods (with penalization by Jeffreys prior) seem to outperform Wald-type confidence intervals and intervals based on the ordinary likelihood ratio, in terms of coverage.

1.2 Outline

The recent applied and methodological interest (Mehrabi & Matthews, 1995; Heinze & Schemper, 2002; Bull et al., 2002, 2007; Zorn, 2005) motivates the in-depth study of the bias-reduction method (Firth, 1993) and the derivation of general expressions for the modified score functions in the case of exponential family models. Furthermore, the beneficial improvement in estimation for logistic regressions (Heinze & Schemper, 2002; Bull et al., 2002, 2007; Zorn, 2005) and in complementary log-log models (Mehrabi & Matthews, 1995), points towards the direction of an extensive study of the behaviour of the BR estimator in categorical response models.

The current thesis is organized in the following way. In Chapter 2, we set up some of the notation that is used throughout the thesis and we give a brief description of the class of exponential family non-linear models. It is a very wide class of parametric models including as special cases both univariate and multivariate generalized linear models, as well as more general nonlinear regressions in which the variance has a specified relationship with the mean.

In Chapter 3 we review the bias-reduction method and we explore, from a theoretical point of view, several aspects of the bias-reducing modifications to the efficient score functions. Furthermore, we derive explicit expressions for the modified scores for exponential family non-linear models. These expressions facilitate the application and study of the bias-reduction method for more general models than the ones already considered in the literature. The main theoretical results in this chapter are derived using index notation and the Einstein summation convention. For a complete treatment of such notation and, generally, tensor methods in statistics, see McCullagh (1987) and Pace & Salvan (1997, Chapter 9). Also, Appendix A is recommended for a short — but sufficient for the contents of the thesis — account of index notation and tensors.

Chapter 4 is the starting point of our treatise on models for categorical responses. We present a systematic and theoretical treatment on logistic regression, both for binomial and multinomial responses. We formally prove the finiteness and shrinkage properties of the bias-reduced estimator, filling the theoretical gaps in Heinze & Schemper (2002), Zorn (2005) and Bull et al. (2002), and we exploit the beneficial impact of the shrinkage effect on the variance and MSE of the bias-reduced estimator. Additionally, the easy implementation of the bias-reduction method is shown. Summing up, we conclude that the bias-reduction method should be regarded as an overall improvement over traditional ML.

In Chapter 5 we deviate from the case of exponential families in canonical parameterization. The bias-reduction method is applied and evaluated

i) in binomial-response GLMs with non-canonical link functions and

ii) in two commonly-used item response theory models.

For obtaining the bias-reduced estimates, we propose a fitting algorithm that uses already implemented software via pseudo-data representations. Furthermore, for every case, the properties of the bias-reduced estimator are explored, with particular emphasis on shrinkage.

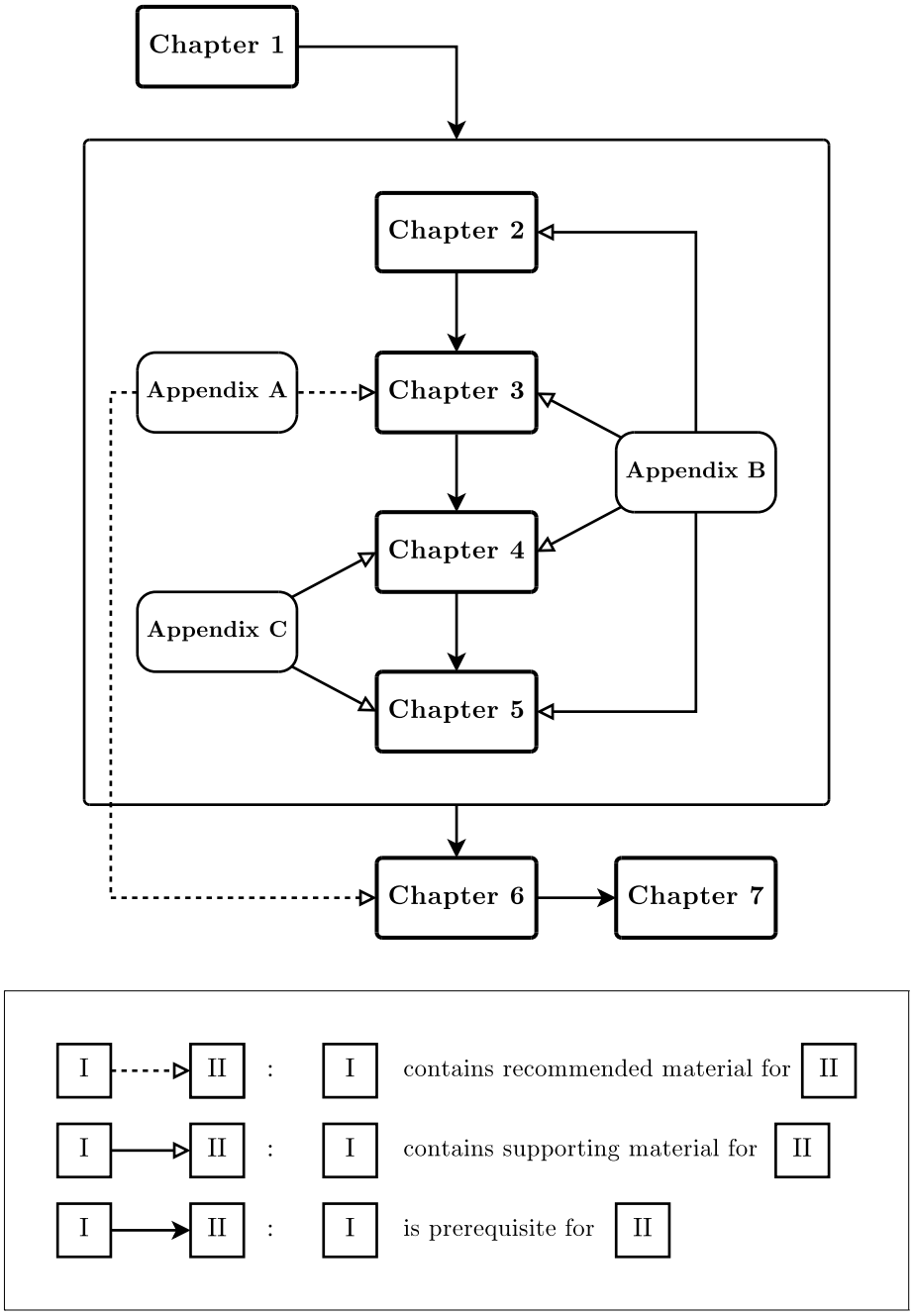
Chapter 6 presents results that are intended to be used for further research in the area. The expansions therein are interesting in their own right and can be used to derive alternative modifications to the score function, which in turn could result in classes of estimators with certain improved properties.

A summary of the main results is given in Chapter 7, and we indicate some related open topics for further work in the area.

Appendix B includes proofs of some theorems and the algebraic derivations of several results presented in the main text. Lastly, Appendix C contains four long tables of estimates that are referred to in the main text.

For the reader's convenience, Figure 1.1 presents a schematic representation of the organization of the thesis contents.

Figure 1.1: Schematic representation of the organization of the contents of the thesis.



Chapter 2

An outline of exponential family non-linear models

2.1 The components of an exponential family non-linear model

In this chapter we introduce the class of exponential family non-linear models with known dispersion. It is a very wide class of parametric models including as special cases both univariate and multivariate generalized linear models (GLMs), as well as more general non-linear regressions in which the variance has a specified relationship with the mean. Classical multivariate examples in this class are multinomial logistic regression models and non-linear extensions of them. Our main aims are i) to review some standard results for exponential family non-linear models and ii) to introduce some of the notation used throughout the thesis. For a detailed technical treatment and study of the geometry of these models the reader is referred to Wei (1997). Also, McCullagh & Nelder (1989) and Fahrmeir & Tutz (2001) are the standard statistical textbooks for GLMs for univariate and multivariate responses, respectively.

2.1.1 Description of the model

Consider a q-dimensional random variable Y in a sample space $\mathcal{Y}$. We say that Y has a distribution function $F(y|\theta,\lambda)$ from the exponential family of distributions $\mathcal{F}$ if and only if its corresponding density or probability mass function has the general form,

$$f(y|\theta,\lambda) = \exp\left\{\frac{y^T\theta - b(\theta)}{\lambda} + c(y,\lambda)\right\}, \tag{2.1}$$

where $c(y,\lambda) \geq 0$ and measurable and $b(\theta)$ is a scalar-valued function of the q-vector θ. The requirement of measurability for the function $c(y,\lambda)$ is necessary in order to have a well defined density on $\mathcal{Y}$.

The parametric family $\mathcal{F}$ indexed by θ is subject to the following conditions:

i) $\theta \in \Theta$ and Θ is a subset of $\Re^q$ with q finite,

ii) $\lambda \in \Lambda$ and Λ is an open subset of $\Re^+$.

iii) The probability distributions defined by any two different values of θ are distinct elements of $\mathcal{F}$,

iv) The inequality

$$0 < \int_{\mathcal{Y}} \exp\left\{\frac{y^T\theta}{\lambda} + c(y,\lambda)\right\} dy < \infty\,,$$

holds. Given the defining property of the density f of having unit integral over $\mathcal{Y}$ with respect to y, the above inequality ensures the existence and finiteness of $b(\theta)$. In the case of vectors of discrete random variables, the integral corresponds to summation of the integrand values over the sample space $\mathcal{Y}$.

For given λ, which is the case we are considering in the thesis, Jørgensen (1987) calls (2.1) the defining form of the densities in a *linear exponential family*. Then, θ is called the canonical parameter and Θ is the canonical parameter space. Furthermore, for given λ, it is easily proved that y is sufficient for θ.

Under the previous assumptions and definitions, the cumulant generator $b(\theta)$ of the family is analytic in the interior of Θ ($\operatorname{int}\Theta$) and all moments of Y exist. Emphasis is given to the first three derivatives of $b(\theta)$ from which we obtain the identities

$$\mathrm{E}(Y;\theta) = \mu_*(\theta) = \nabla_\theta b(\theta)\,, \tag{2.2}$$

$$\mathrm{Cov}(Y;\theta) = \Sigma_*(\theta) = \lambda\mathcal{D}^2\left(b(\theta);\theta\right)\,, \tag{2.3}$$

$$\mathrm{Cum}_3(Y;\theta) = K_*(\theta) = \lambda^2\mathcal{D}^2\left(\mu_*(\theta);\theta\right)\,, \tag{2.4}$$

where $\mathrm{Cum}_3(Y;\theta)$ is the $q^2 \times q$ matrix of third order cumulants of Y, as a function of θ, and $\mathcal{D}^2\left(b(\theta);\theta\right)$ stands for the $q \times q$ Hessian of $b(\theta)$ with respect to θ . Generally, in what follows, if a and b are p and q dimensional vectors, respectively, $\mathcal{D}\left(a;b\right)$ is the $p \times q$ matrix of first derivatives of a with respect to b and $\mathcal{D}^2\left(a;b\right)$ is a $pq \times q$ blocked Hessian matrix of the second derivatives of a with respect to b, having blocks the $q \times q$ matrices $\mathcal{D}^2\left(a_i;b\right)$, $i = 1,\ldots,p$ (see Appendix B, Section B.1 for analytic description).

The covariance matrix $\Sigma_*(\theta)$ is assumed to be positive definite in $\operatorname{int}\Theta$. Furthermore

$$\mu_* : \operatorname{int}\Theta \longrightarrow \mathrm{M} \equiv \mu_*(\operatorname{int}\Theta)$$

is injective (one-to-one), and hence invertible. If we denote by $\theta_*(\mu)$ the inverse of μ_* and substitute into (2.3), the variance-covariance matrix of Y can be written as a function of the expectation μ and the dispersion parameter λ,

$$\mathrm{Cov}(Y;\mu) = \lambda v(\mu) = \lambda\,\mathcal{D}^2\left(b(\theta);\theta\right)\big|_{\theta=\theta_*(\mu)}\,,$$

where $v(\mu)$ is called the variance function.

An exponential family non-linear model (sometimes referred to as a generalized non-linear model) consists of three components:

i) Random component : The random variable Y has density or probability mass function $f(y|\theta, \lambda)$ of the form (2.1).

ii) Systematic component : The systematic part of the model is a function η of the p-vector of parameters β, and $\eta(\beta)$ takes values in an open subset of $\Re^q$. The parameter vector β belongs to the parameter space B which is an open subset of $\Re^p$. Furthermore, the predictor function $\eta(\beta)$ is assumed to be at least three times continuously differentiable with respect to β.

iii) Linking structure : The expectation μ of Y is linked with the systematic part $\eta(\beta)$ through an assumed vector-valued function $g : \mathrm{M} \to \Re^{\mathrm{q}}$,

$$g(\mu) = \eta(\beta)\,. \tag{2.5}$$

The link function g is assumed to be monotonic and differentiable up to third order.

By the structure of a generalized non-linear model, the canonical parameter θ is related to the predictor $\eta(\beta)$ through the function $u = (g \circ \mu_*)^{-1} = \theta_* \circ h$, with h the inverse of the function g. So,

$$\theta = u(\eta(\beta)) = \theta_*(h(\eta(\beta)))\,. \tag{2.6}$$

Hence, by the monotonicity of the link function and the fact that the mapping μ_* is injective, the predictor η indexes the family of distributions $\mathcal{F}$, and the probability distributions defined by any two different values of η are distinct elements of $\mathcal{F}$.

Note that if we set $\eta(\beta) = Z\beta$ in (2.5), where the $q \times p$ matrix $Z = Z'(x)$ is an appropriate function of a known covariate p'-vector x not depending on β, (2.5) corresponds to a multivariate GLM (see Fahrmeir & Tutz, 2001, for a thorough study). Furthermore, if we drop the dimension of the response to $q = 1$, we obtain a univariate GLM.

2.1.2 Canonical link functions

Consider the case of a link function g such that (2.5) takes the form

$$\theta = g(\mu) = \eta(\beta)\,. \tag{2.7}$$

Such link function will be called canonical. This definition of canonical links is parallel to the corresponding definition for GLMs, in the sense that the canonical parameter θ is equated to the predictor η. However, the familiar property that there exists a sufficient statistic having the same dimension as β is no longer generally valid, because curvature is introduced to the family of distributions by the non-linearity of the predictor η. On the other hand, in the case of a *linear* predictor $\eta(\beta) = Z\beta$, there is always a sufficient statistic T such that $\dim T = \dim \beta = p$ and the model corresponds to a flat exponential family in canonical parameterization.

Furthermore, we can always represent a GLM with general link function as a generalized non-linear model with canonical link. So, if a GLM has the form

$$g(\mu) = \eta(\beta) = Z\beta\,,$$

with Z a $q \times p$ matrix not depending on β, then the equivalent canonically-linked generalized non-linear model will have the form

$$\theta = \tilde{\eta}(\beta) = \theta_*(h(Z\beta))\,,$$

where the functions θ_* and h are as defined earlier.

2.2 Notational conventions

Because of the multivariate nature of the responses in the generic setup of generalized non-linear models, sequences of matrices or multidimensional arrays appear as quantities of interest under repeated sampling. This suggests the introduction of a notational framework that enables us to write multidimensional arrays as matrices with certain blocking structure, in a consistent way.

Consider a sequence of 3-way arrays $\{E_r; r = 1, \ldots, k\}$. Each array E_r is a 3-way arrangement of scalars e_{rstu} with $s = 1, \ldots, l$, $t = 1, \ldots, m$ and $u = 1, \ldots, n$. Note that the scalar components of E_r are denoted by lower case letters. Generally, in what follows the scalar components of an array will be denoted by the corresponding lower case letters. The array E_r can be represented as a $lm \times n$ blocked-matrix having the form

$$E_r = \begin{bmatrix} E_{r1} \\ E_{r2} \\ \vdots \\ E_{rl} \end{bmatrix},$$

with E_{rs} a $m \times n$ matrix. Writing E_{rst} we denote the t-th row of the matrix E_{rs} as a row vector, i.e., having dimension $1 \times n$.

Similarly, consider a sequence of 4-way arrays $\{E_r; r = 1, \ldots, k\}$. Such array E_r is a 4-way arrangement of scalars e_{rstuv} with $s = 1, \ldots, l$, $t = 1, \ldots, m$, $u = 1, \ldots, n$ and $v = 1, \ldots, q$. The array E_r can be represented as a $ln \times mq$ blocked-matrix having the form

$$E_r = \begin{bmatrix} E_{r11} & E_{r12} & \cdots & E_{r1m} \\ E_{r21} & E_{r22} & \cdots & E_{r2m} \\ \vdots & \vdots & \ddots & \vdots \\ E_{rl1} & E_{rn2} & \cdots & E_{rlm} \end{bmatrix},$$

with E_{rst} a $n \times q$ matrix. Writing E_{rstu} we denote the u-th row of the matrix E_{rst} as a row vector, i.e. having dimension $1 \times q$.

Similar conventions are used for 2-dimensional arrays or matrices. So, if we consider the sequence of $l \times m$ matrices $\{E_r; r = 1, \ldots, k\}$ then each E_r is an arrangement of scalars e_{rst} with $s = 1, \ldots, l$, $t = 1, \ldots, m$. In this case, E_{rs} denotes the s-th row of E_r, as a row vector, i.e. having dimension $1 \times m$.

Under these notational conventions, we should stress the difference in the assignment of dimensions between a q-vector μ_r that has dimension $q \times 1$ and the s-th row of a $p \times q$ matrix E_r denoted as E_{rs} having dimension $1 \times q$.

In what follows, the blocking-structure and dimensions of any quantity will be clearly stated or will be clear directly from the context.

Lastly, unless otherwise stated, all the algebraic structures that are functions of the parameters β are denoted simply by the corresponding letter. For example, consider a $p \times p$ matrix $Q_t(\beta)$. This is denoted just as Q_t.

2.3 Likelihood related quantities

2.3.1 Log-likelihood function

Consider realizations $y_1, y_2, \ldots, y_n$ of independent random variables $Y_1, Y_2, \ldots, Y_n$ from the exponential family of distributions (2.1). In the light of these observations, the log-likelihood function for the parameters β of an exponential family non-linear model with known dispersion parameter is the sum of n independent contributions,

$$l \equiv l(\beta|\{y_r\}, \lambda) \equiv \log \prod_r f(y_r|\beta, \lambda_r) = \sum_r \left\{ \frac{y_r^T \theta_r - b(\theta_r)}{\lambda_r} + c(y_r; \lambda_r) \right\} , \tag{2.8}$$

where θ_r is related to β as in (2.6), i.e. $\theta_r = \theta_r(\beta) = u(\eta_r(\beta))$, and λ_r is known for every $r = 1, \ldots, n$. Further, the term $c(y_r; \lambda_r)$ does not depend on β and thus it can be omitted from the summands of the log-likelihood for obtaining the maximum likelihood (ML) estimator.

The usual regularity conditions, in the spirit of the ones in Cramér (1946, §33.2) (see also Cox & Hinkley, 1974, §9.1) apply to this case as follows:

i) The admissible parameter space B is an open subset of $\Re^p$.

ii) The parameter space B does not depend on the sample space $\mathcal{Y}$.

iii) $h(\eta_r(\beta)) \in \mathrm{M} = \mu_*(\operatorname{int} \Theta)$, $r = 1, \ldots, n$, for all $\beta \in \mathrm{B}$.

iv) Two different values of β correspond to two different members of the family of distributions $\mathcal{F}$.

v) The log-likelihood function is almost surely three times continuously differentiable with respect to β in some neighbourhood of the true value β_0. Further, for $\epsilon > 0$ and for all β in the ball $|\beta - \beta_0| < \epsilon$,

$$n^{-1} \left| \frac{\partial^3 l(\beta)}{\partial \beta_s \partial \beta_t d\beta_u} \right| \leq a_{stu} \quad (s, t, u = 1, \ldots, p) ,$$

with a_{stu} finite and $\mathrm{E}(a_{stu}) < \infty$.

Condition v) ensures the continuity of the first three derivatives of the log-likelihood. Specifically, condition v) plays a crucial role in establishing the asymptotic normality of the ML estimator. Condition ii) ensures that the processes of integration over the sample space $\mathcal{Y}$ and differentiation with respect to β can be interchanged. Conditions iii) and

iv) are necessary for a well-defined generalized non-linear model. Especially condition iv) ensures that given a fixed parameterization β for the model, inferences based on some specific value of β cannot be the same for any other values. For the models considered in the thesis i), ii), iii) and v) are satisfied. In the case of linear predictors, the requirement of identifiability (condition iv)) can be easily satisfied, for example by using an appropriate set of linear constraints on β with direct impact on the rank of the design matrix (see, for example, McCullagh & Nelder, 1989, §3.5). However, for any given model with non-linear predictor, the identifiability requirement should be carefully examined and verified.

2.3.2 Score functions

The score vector for the parameters β of an exponential family non-linear model with known dispersion has the form

$$U = \sum_r Z_r^T D_r \Sigma_r^{-1} (y_r - \mu_r)\,,$$

where $Z_r = \mathcal{D}(\eta_r; \beta)$, $D_r^T = \mathcal{D}(\mu_r; \eta_r)$, and $\mu_r = \mathrm{E}(Y_r; \beta)$, $\Sigma_r = \mathrm{Cov}(Y_r; \beta)$ are as defined in (2.2) and (2.3), respectively. For the analytic derivation of the vector of score functions see (B.2) in Appendix B. Re-expressing the above equation in terms of the matrix of 'working weights' $W_r = D_r \Sigma_r^{-1} D_r^T$, we can obtain the alternative equivalent expression

$$U = \sum_r Z_r^T W_r \mathcal{D}(\eta_r; \mu_r)(y_r - \mu_r)\,. \tag{2.9}$$

The working (or quadratic) weight matrix W_r for generalized non-linear models is defined exactly as for the case of GLMs (see, for example, Fahrmeir & Tutz (2001, Appendix A.1) for the multivariate case or McCullagh & Nelder (1989, Section 2.5.1) for univariate GLMs) and plays an important role in ML fitting procedures, such as iterative generalized least squares.

From (2.9), it is clear that the score functions have zero expectation, since the responses appear in the equations only through their centered form $Y_r - \mu_r$. This result is usually referred to as the unbiasedness of the score function and it is standard in ML theory under the aforementioned usual regularity conditions.

If the log-likelihood function is concave on β, the ML estimator $\hat{\beta}$ is the solution of the likelihood equations

$$U(\hat{\beta}) = 0\,.$$

2.3.3 Information measures

By the usual differentiation rules and the independence of $Y_1, Y_2, \ldots, Y_n$, the observed and the Fisher information matrices are sums of n independent contributions. For the

observed information on β we have

$$I = \sum_r Z_r^T W_r Z_r - \sum_r \sum_{s=1}^{q} \lambda_r^{-1} Z_r^T V_{rs} Z_r (y_{rs} - \mu_{rs}) \\ - \sum_r \sum_{s,u=1}^{q} \mathcal{D}^2 (\eta_{ru}; \beta) \, k_{rsu} (y_{rs} - \mu_{rs}) \,, \tag{2.10}$$

where $V_{rs} = \mathcal{D}^2 (\theta_{rs}; \eta_r)$ and k_{rsu} is the (s,u)-th element of the matrix $\Sigma_r^{-1} D_r^T$.

Under the validity of the Bartlett identities (see Appendix B, Section B.1), the Fisher (or expected) information F on β can be obtained as the expectation of the expression in (2.10). Because the two last terms in the right hand side of (2.10) have zero expectation, we have that

$$F = \mathrm{E}(I) = \sum_r Z_r^T W_r Z_r \,. \tag{2.11}$$

The derivations of the above formulae are given analytically in Section B.1 of Appendix B.

For canonically-linked models, $V_r = \mathcal{D}^2 (\theta_r; \eta_r) = 0$ and $D_r = \lambda_r^{-1} \Sigma_r$. So, (2.9), (2.10) and (2.11) are considerably simplified, taking the forms

$$U = \sum_r \lambda_r^{-1} Z_r^T (y_r - \mu_r) \,, \tag{2.12}$$
$$F = \sum_r \lambda_r^{-2} Z_r^T \Sigma_r Z_r \,,$$
$$I = F - \sum_r \sum_{s=1}^{q} \mathcal{D}^2 (\eta_{rs}; \beta) \, (y_{rs} - \mu_{rs}) \,.$$

For *linear* predictors $\eta(\beta)$, the score functions have the same form as in (2.9), the only difference being that Z_r is no longer a function of β rather than some appropriate function of a p'-vector of covariates x. The same applies to the Fisher information given in (2.11). The significant change is noted in the formulae for the observed information. Except for the difference in the nature of Z_r's, the last term in the right hand side of (2.10) vanishes because $\mathcal{D}^2 (\eta_r; \beta)$ is zero for linear predictors. In the case of GLMs with canonical link, the same simplifications apply to the expressions in (2.12). Nevertheless, such models are exponential families in canonical parameterization, and thus the Hessian of the log-likelihood with respect to β does not depend on Y_r's. Hence, the Fisher and the observed information coincide.

2.4 Fitting exponential family non-linear models

The most standard fitting procedure used for maximizing the log-likelihood for exponential family non-linear models is Fisher scoring; see for example McCullagh & Nelder (1989, §2.5) for a description for univariate GLMs and Fahrmeir & Tutz (2001, §3.4.1) for multivariate GLMs.

In the univariate case, Fisher scoring can be viewed as an iterative re-weighted least squares (IWLS) procedure; see for example Agresti (2002, §4.6.3) for a thorough description of the connection. However, in the multivariate case, the generalization of the same method corresponds to iterative *generalized* least squares (IGLS) because the working weight matrix is block diagonal rather than diagonal as in the univariate case.

2.4.1 Newton-Raphson and Fisher scoring

Consider a first-order Taylor approximation to the likelihood equations $U(\hat{\beta}) = 0$,

$$0 = U(\hat{\beta}) \approx U(\beta_0) - I(\beta_0)(\hat{\beta} - \beta_0)\,,$$

with $I(\beta_0)$ the observed information matrix at the true unknown parameter value β_0. Re-expressing in terms of $\hat{\beta}$ we get

$$\hat{\beta} \approx \beta_0 + \{I(\beta_0)\}^{-1} U(\beta_0)\,,$$

which suggests the Newton-Raphson iteration,

$$\beta_{(c+1)} = \beta_{(c)} + \{I^{-1}U\}_{(c)}\,, \tag{2.13}$$

with $\beta_{(c)}$ the parameter value at the c-th iteration, and for $\{I^{-1}U\}_{(c)}$, the subscript (c) denotes evaluation at $\beta_{(c)}$.

If we replace the observed information I with the expected information F in the right hand side of (2.13), we obtain the Fisher scoring iteration

$$\beta_{(c+1)} = \beta_{(c)} + \{F^{-1}U\}_{(c)}\,. \tag{2.14}$$

Given good starting values and the concavity of the log-likelihood function, ML estimates can be obtained using either iteration (2.13) or iteration (2.14), until an appropriate stopping criterion is satisfied; for example, the change to the value of the log-likelihood function between successive iterations is sufficiently small.

Comparing the two alternative fitting procedures which are defined by iterations (2.13) and (2.14), Fisher scoring produces an estimate of the asymptotic variance-covariance matrix of the ML estimator as its byproduct, namely the inverse of the Fisher information evaluated at the last iteration. Also, the Fisher information is necessarily non-negative definite and so iteration (2.14) can never cause a decrease to the log-likelihood function. Further, there is a neat equivalence between Fisher scoring and the generalized least squares method for ordinary linear regressions. However, despite the conveniences, Fisher scoring does not guarantee that the convergence to the ML estimates is at quadratic rate as the Newton-Raphson fitting procedure does.

There is a vast repository of alternative optimization methods as well as modifications of the Newton-Raphson and Fisher scoring methods, each with its own advantages (see for example Monahan, 2001). The choice of fitting procedure is highly model dependent. In the case of exponential family non-linear models the expected information has a much simpler form than the observed (compare (2.11) with (2.10)) and thus Fisher scoring is

preferable to Newton-Raphson. However, there might be non-linear models where even the expected information is hard or expensive to evaluate. In this case, quasi-Newton methods provide a good alternative approach, where the Hessian for a Newton-Rapshon iteration does not need to be specified and an approximation is used instead. A well-used example is the variable metric algorithm or 'BFGS' method (see Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) which is well-implemented through the `optim` function in *R language* (R Development Core Team, 2007).

2.4.2 Fisher scoring and iterative generalized least squares

To illustrate the equivalence between Fisher scoring and IGLS, consider the 'working observation' vector

$$\zeta_r = Z_r\beta + (D_r^T)^{-1}(y_r - \mu_r)\,, \tag{2.15}$$

with scalar components $\zeta_{r1}, \zeta_{r2}, \ldots, \zeta_{rq}$. In the case of multivariate GLMs, where Z_r does not depend on β, ζ_r is the locally linearized form of the link function g evaluated at y_r. Note that the Fisher information for an exponential family non-linear model can be written as

$$F = \sum_r Z_r^T W_r Z_r = Z^T W Z\,,$$

where $Z^T = (Z_1^T, \ldots, Z_n^T)$ is a blocked matrix of dimension $p \times nq$ and W is the $nq \times nq$ block-diagonal matrix, with diagonal blocks the matrices of working weights W_r, $r = 1, \ldots, n$. Thus, by (2.15), the equivalent IGLS iteration to (2.14) is

$$\{Z^T W Z\}_{(c)}\beta_{(c+1)} = \{Z^T W \zeta\}_{(c)} \tag{2.16}$$

so that

$$\beta_{(c+1)} = \left\{(Z^T W Z)^{-1} Z^T W \zeta\right\}_{(c)}\,, \tag{2.17}$$

where $\zeta = (\zeta_{r1}, \ldots, \zeta_{rq}, \ldots, \zeta_{n1}, \ldots, \zeta_{nq})^T$. Ignoring the subscript (c), equations (2.16) are the normal equations for fitting a multivariate response linear model with response vector ζ, model matrix Z and normal errors with zero expectation and variance covariance matrix W^{-1}, all evaluated at the c-th iteration.

2.4.3 Hat matrix

For the special case of multivariate GLMs, where Z_r does not depend on the parameters, iteration of the above scheme regresses $\zeta_{(c)}$ on Z using the weight matrix $W_{(c)}$, to obtain a new estimate $\beta_{(c+1)}$ through (2.17). Then this estimate is used to calculate a new linear predictor value $\eta_{(c+1)} = Z\beta_{(c+1)} = H_{(c)}\beta_{(c)}$ and hence new working observations $\zeta_{(c+1)}$ using (2.15). The cycle is continued until some appropriate convergence criterion is satisfied.

The matrix

$$H = Z\left(Z^T W Z\right)^{-1} Z^T W$$

is the projection or 'hat' matrix that has the well-known leverage interpretation in linear models. For multivariate GLMs, the n^2 blocks of H can be still used as generalized

influence measures (Fahrmeir & Tutz, 2001, §4.2.2), but its key characteristic is that it projects the current working observation to the next value of the linear predictor in an IGLS iteration. In the case of univariate generalized non-linear models, Wei (1997, §6.5) proposes an alternative generalized leverage measure which is based on the instantaneous rate of change of the predictions for the response relative to the observed response values and seems to have a more natural interpretation than H. However, the performance of alternative influence measures is out of the scope of this thesis. Despite its questionable leverage interpretation for exponential family non-linear models, the hat matrix H is going to play an important role in the results of later chapters. The reasons for this are i) algebraic convenience and ii) the fact that H is readily available from the quantities required for the implementation of the IGLS fitting procedure.

Chapter 3

A family of modifications to the efficient score functions

3.1 Introduction

The most common estimation method in the frequentist school is maximum likelihood (ML). The reasons for this are mainly the neat asymptotic properties of the ML estimator (asymptotic normality, asymptotic sufficiency, unbiasedness and efficiency) and further the easy implementation of fitting procedures. However, there are cases where the ML estimator can have appreciable bias, especially when the sample size is small, resulting in potentially misleading inferences. Cordeiro & McCullagh (1991) derived explicit expressions for the asymptotic bias of the ML estimator in the case of univariate generalized linear models (GLMs) and showed how the first-order ($\mathcal{O}(n^{-1})$) term in the asymptotic expansion of the bias of the ML estimator could be eliminated by a supplementary weighted regression. However, such bias correction depends upon the existence of the ML estimates and so it does not apply, for example, to situations where the ML estimates are found to be infinite-valued. Motivated partly by this, Firth (1993) developed a fairly general method for removing the first-order ($\mathcal{O}(n^{-1})$) term in the asymptotic expansion of the bias of the ML estimator. Specifically, for regular problems the efficient score functions are appropriately modified such that the roots of the resultant modified score equations result in first-order unbiased estimators. Firth (1993) studied the case of canonically linked GLMs, emphasizing the properties of the bias-reduced (BR) estimator in some special but important cases such as binomial logistic regression models, Poisson log-linear models and Gamma reciprocal-linear models (Firth, 1992a,b). Heinze & Schemper (2002) through empirical studies verified the properties of the BR estimator for binomial logistic regression models and illustrated the superiority of profile penalized-likelihood based confidence intervals over Wald-type intervals. Similar studies have been carried out by Mehrabi & Matthews (1995) who applied the bias-reduction method to a binomial-response complementary log-log model, and by Bull et al. (2002) who evaluated it on multinomial logistic

regression. In all these cases, they conclude on the superior properties of the BR estimator relative to the ML estimator.

In this chapter we give a brief description of the bias-reduction method as given in Firth (1993). The general results in the latter paper are quoted in index notation and using Einstein summation convention; for a complete treatment of such notation and, generally, tensor methods in statistics see McCullagh (1987), Pace & Salvan (1997, Chapter 9) or for a short — but sufficient for the contents of the thesis — account, Appendix A. For our purposes, there will be an interchange between index notation under the conventions of Appendix A and usual matrix notation under the notational conventions of Section 2.2. Whenever such an interchange takes place, it will be clearly stated. We re-express the results in Firth (1993) in usual matrix notation and for the case of a scalar-parameter statistical problem, we study the relation of the modified scores with statistical curvature (Efron, 1975). Further, we derive a necessary and sufficient condition on the existence of a penalized likelihood corresponding to the modified scores and we comment on the consistency and asymptotic normality of the BR estimator.

Lastly, motivated by the recent applied and methodological research interest in the penalized likelihood approach to bias reduction (e.g., Heinze & Schemper, 2002; Bull et al., 2002, 2007; Zorn, 2005), we derive explicit formulae for the modified score vector for the wide class of multivariate-response exponential family non-linear models with known dispersion; as mentioned in Chapter 2, this class includes as special cases both univariate and multivariate GLMs, as well as more general non-linear regressions in which the variance has a specified relationship with the mean. The formulae derived involve quantities that are readily available from the output of standard computing packages and they can be used directly for the implementation of the modified scores approach to bias reduction. In addition, they can be used theoretically in order to gain insight into the nature of the modifications — for example, whether their effect is a shrinkage effect or something potentially less advantageous — in any specific application. This will be the topic of later chapters.

3.2 Family of modifications. Removal of the first-order asymptotic bias term from the ML estimator

3.2.1 General family of modifications

Consider a model with parameters the components of the vector $\beta = (\beta^1, \beta^2, \ldots, \beta^p)$. Assume that, in the light of n realizations of independent random variables, we formulate the log-likelihood function $l(\beta)$. Under regularity conditions in the spirit of the ones in Cox & Hinkley (1974, §9.1), Firth (1993) developed a general family of modifications to the efficient scores. Therein, it is shown that location of the roots of the modified scores results in an estimator with second-order ($o(n^{-1})$) bias.

Using index notation and under the Einstein summation convention (see Appendix A), assume we modify the r-th component of the efficient score vector according to

$$U_r^* = U_r + A_r \,,$$

where A_r is $\mathcal{O}_p(1)$ as $n \to \infty$ and is allowed to depend on the data, and U_r is the ordinary score. Note that in this case, the letter r indexes parameters ($r \in \{1, 2, \ldots, p\}$) in contrast to its use for indexing observations under the usual matrix notation in Chapter 2.

Firth (1993) proposed and studied two alternatives for A_r that result in removal of the first-order term from the asymptotic expansion of the bias of the ML estimator; one based on the expected information,

$$A_r = A_r^{(E)} = -\kappa_{r,s} b_1^s \tag{3.1}$$

and one on the observed

$$A_r = A_r^{(O)} = n^{-1} U_{rs} b_1^s \,, \tag{3.2}$$

where

$$n^{-1} b_1^r = -n^{-1} \kappa^{r,s} \kappa^{t,u} (\kappa_{s,t,u} + \kappa_{s,tu})/2 \tag{3.3}$$

is the first-order term in the asymptotic expansion of the bias of the ML estimator. In the above formulae the quantities named U refer to log-likelihood derivatives. Thus,

$$U_r = \frac{\partial l(\beta)}{\partial \beta^r} \,; \quad U_{rs} = \frac{\partial^2 l(\beta)}{\partial \beta^r \partial \beta^s} \,; \quad U_{rst} = \frac{\partial^3 l(\beta)}{\partial \beta^r \partial \beta^s d\beta^t} \,;$$

and so on. Also, the quantities named κ refer to the *null* cumulants of log-likelihood derivatives, *per observation*

$$\kappa_{r,s} = n^{-1} \mathrm{E}\,(U_r U_s) \;; \quad \kappa_{r,st} = n^{-1} \mathrm{E}\,(U_r U_{st}) \;; \quad \kappa_{r,s,t} = n^{-1} \mathrm{E}\,(U_r U_s U_t) \;;$$

and so on. The word 'null' is used to indicate that both the operations of differentiation and expectation take place for the same value of the parameter β. Also, $\kappa^{r,s}$ in (3.3) is the matrix-inverse of $\kappa_{r,s}$ and all κ's, as defined above, are $\mathcal{O}(1)$ (see Subsection A.4.2 in Appendix A for a note on the definitions and the calculus of $\mathcal{O}$, $\mathcal{O}_p$, o, o_p). The same expression for the first-order bias term (3.3) in the single parameter case, is derived in Cox & Hinkley (1974, §9.2 (vii)).

Under the notational rules of Section 2.2 and letting F and I be the Fisher and observed information on β, we can express the above results in matrix notation. From (3.1) and (3.2), removal of the first-order bias term occurs if either

$$A_t \equiv A_t^{(E)} = \frac{1}{2} \operatorname{trace} \left\{ F^{-1} (P_t + Q_t) \right\} \quad (t = 1, \ldots, p) \tag{3.4}$$

or

$$A_t \equiv A_t^{(O)} = I_t F^{-1} A^{(E)} \quad (t = 1, \ldots, p) \,, \tag{3.5}$$

based on the expected or observed information, respectively. In the above expressions, $A^{(E)} = (A_1^{(E)}, \ldots, A_p^{(E)})^T$, $P_t = \mathrm{E}(UU^T U_t)$ stands for the t-th block of the $p^2 \times p$ matrix of the third order cumulants of the scores, and $Q_t = \mathrm{E}(-IU_t)$ for the t-th block of the $p^2 \times p$ blocked matrix of the covariance of the first and second log-likelihood derivatives. Note that by the symmetry of UU^T and I, P_t and Q_t are symmetric matrices for every $t = 1, \ldots, p$. Also, the $1 \times p$ vector I_t denotes the t-th row of I. All the expectations are

taken with respect to the model and at β. From (3.3), it is immediate that the vector of first-order biases can be expressed in terms of $A^{(E)}$ as

$$n^{-1}b_1 = -F^{-1}A^{(E)}\,. \tag{3.6}$$

We should mention that according to the derivation of the above modifications (see Section 6.6 for a detailed derivation), the modifications studied in Firth (1993) are just a special case of a more general family of modifications. Using index notation, if we let

$$\mu_{r,s} = \mathrm{E}\left(U_r U_s\right)\;; \quad \mu_{r,st} = \mathrm{E}\left(U_r U_{st}\right)\;; \quad \mu_{r,s,t} = \mathrm{E}\left(U_r U_s U_t\right)\;,$$

be the joint null moments of log-likelihood derivatives, the generic modifications are defined as

$$A_r = e^s_r \mu^{t,u}\left(\mu_{s,tu} + \mu_{s,t,u}\right)/2 + \bar{R}_r\,, \tag{3.7}$$

where either $e^t_r = \left(\mu_{r,s} + R_{rs}\right)\mu^{s,t}$ or $e^t_r = \left(-U_{rs} + R_{rs}\right)\mu^{s,t}$ or $e^t_r = \left(U_r U_s + R_{rs}\right)\mu^{s,t}$ and R_{rs} and $\bar{R}_r$ are any quantities that depend on the data and the parameters and have expectations of order at most $\mathcal{O}(n^{1/2})$ and at most $\mathcal{O}(n^{-1/2})$, respectively. Any modification that results from the above generic definition results in an estimator with $o(n^{-1})$ bias.

We consider only the case where $\bar{R}_r = 0$. In matrix notation, one can write the modifications (3.7), as a weighted sum of the modifications based on the expected information, defined in (3.4). This allows the modified scores to be written in the general form

$$U^*_t = U_t + \sum_{u=1}^{p} e_{tu} A^{(E)}_u \quad (t = 1, \ldots, p)\;, \tag{3.8}$$

where e_{tu} is defined as either

$$e_{tu} \equiv e^{(E)}_{tu} = \left[RF^{-1} + 1_p\right]_{tu}$$

or

$$e_{tu} \equiv e^{(O)}_{tu} = \left[(I + R)F^{-1}\right]_{tu}$$

or

$$e_{tu} \equiv e^{(S)}_{tu} = \left[(UU^T + R)F^{-1}\right]_{tu}\;,$$

with 1_p the $p \times p$ identity matrix. In order to obtain the modifications based on the Fisher information (3.4), we merely choose $e^{(E)}_{tu}$ in (3.8) and set $R = 0$, so that $e_{tu} = 1$ for $t \neq u$ and $e_{tt} = 1$. Also, for obtaining the modifications based on the observed information (3.5), we just chose $e^{(O)}_{tu}$ in (3.8) and set $R = 0$. Lastly, note that by the weak law of large numbers both $n^{-1}I$ and $n^{-1}UU^T$ converge to $n^{-1}F$ and that RF^{-1} converges in probability to zero as $n \to \infty$, so that every possibility for the modified scores is asymptotically equivalent to $U^*_t = U_t + A^{(E)}_t$.

3.2.2 Special case: exponential family in canonical parameterization

Now, consider the case where β is the canonical parameter of an exponential family model. For such flat exponential family the observed information I and the Fisher information F on β coincide, because the second derivatives of the log-likelihood with respect to β do not depend on the responses. So, the cumulant matrices $Q_t = \mathrm{E}(-IU_t)$ are zero. Further, since I_t does not depend on the data

$$\frac{\partial}{\partial\beta^t}F = \frac{\partial}{\partial\beta^t}\mathrm{E}(UU^T) = \mathrm{E}(-I_t^T U^T) + \mathrm{E}(-UI_t) + \mathrm{E}(UU^T U_t) = \mathrm{E}(UU^T U_t) = P_t\,.$$

Thus, for $R_t = 0$,

$$\begin{aligned} A_t = A_t^{(E)} = A_t^{(O)} &= \frac{1}{2}\,\mathrm{trace}\left\{F^{-1}\frac{\partial F}{\partial\beta^t}\right\} \\ &= \frac{\partial}{\partial\beta^t}\left\{\frac{1}{2}\log\det F\right\} \quad (t = 1,\ldots,p)\ , \end{aligned}$$

where $\det F$ denotes the determinant of F. Hence, the equations $U_t^* = 0$ locate a stationary point of a penalized log-likelihood function of the form

$$l^*(\beta) = l(\beta) + \frac{1}{2}\log\det F(\beta)\,,$$

which corresponds to a penalized likelihood of the form

$$L^*(\beta) = L(\beta)\,(\det F(\beta))^{1/2}\ . \tag{3.9}$$

Thus, as noted in Firth (1993), the bias-reduction method in the case of flat exponential family models reduces to penalization of the likelihood function by Jeffreys invariant prior (Jeffreys, 1946). From a Bayesian point of view, obtaining the maximum penalized likelihood estimates is equivalent to obtaining the posterior mode using Jeffreys invariant prior.

3.2.3 Existence of penalized likelihoods for general exponential families

A natural question that arises in this context is about the existence of penalized likelihoods that correspond to the modified scores in the case of general exponential families. By 'existence' we mean that there is a unique —up to a multiplicative constant not depending on the parameters— function of the parameters that is the integral of the modified scores. As in the case of existence of quasi-likelihoods in McCullagh & Nelder (1989, § 9.3.2), a necessary and sufficient condition for the existence of a penalized likelihood is that the derivative matrix of $U^*(\beta)$ is symmetric. Using index notation, this condition translates to

$$\frac{\partial U_r^*(\beta)}{\partial\beta^s} = \frac{\partial U_s^*(\beta)}{\partial\beta^r}\,,$$

where the indices r and s take values in $\{1, 2, \ldots, p\}$. Applying the theorem in Skovgaard (1986) for the differentiation of cumulants of log-likelihood derivatives (see Theorem A.4.1 in Appendix A) we have that

$$\frac{\partial}{\partial \beta^r}\mu_{s,t} = \mu_{rs,t} + \mu_{s,rt} + \mu_{r,s,t}\,,$$

so that the modified scores based on the expected information can be re-expressed as

$$\begin{aligned} U_r^* &= U_r + \frac{1}{2}\mu^{s,t}(\mu_{r,st} + \mu_{r,s,t}) \\ &= U_r + \frac{1}{2}\mu^{s,t}\frac{\partial}{\partial \beta^r}\mu_{s,t} + \frac{1}{2}\mu^{s,t}(\mu_{r,st} - \mu_{rs,t} - \mu_{rt,s}) \\ &= U_r + \frac{1}{2}\mu^{s,t}\frac{\partial}{\partial \beta^r}\mu_{s,t} + \frac{1}{2}\mu^{s,t}(\mu_{r,st} - 2\mu_{rs,t})\,. \end{aligned} \tag{3.10}$$

The second term in the right hand side of the latter expression is the derivative of the logarithm of the Jeffreys prior and U_r are the derivatives of a proper log-likelihood. So, the existence of penalized log-likelihoods depends solely on the symmetry of the derivatives, T_{rs}, of the third term. These derivatives are

$$\begin{aligned} T_{rs} &= \frac{\partial}{\partial \beta^s}\left\{\mu^{t,u}(\mu_{r,tu} - 2\mu_{rt,u})\right\} \\ &= \mu^{t,u}\frac{\partial}{\partial \beta^s}\left(\mu_{r,tu} - 2\mu_{rt,u}\right) - \mu^{t,w}\mu^{u,v}\left(\mu_{r,tu} - 2\mu_{rt,u}\right)\frac{\partial}{\partial \beta^s}\mu_{w,v} \\ &= \mu^{t,u}\left(\mu_{rs,tu} + \mu_{r,stu} + \mu_{r,s,tu} - 2\mu_{rst,u} - 2\mu_{rt,su} - 2\mu_{rt,s,u}\right) \\ &\quad - \mu^{t,w}\mu^{u,v}\left(\mu_{r,tu} - 2\mu_{rt,u}\right)\left(\mu_{sw,v} + \mu_{w,sv} + \mu_{s,w,v}\right)\,. \end{aligned}$$

Hence,

$$\begin{aligned} T_{rs} = {} & \mu^{t,u}\left(\mu_{rs,tu} + \mu_{r,stu} + \mu_{r,s,tu} - 2\mu_{rst,u} - 2\mu_{rt,su} - 2\mu_{rt,s,u}\right) \\ & + \mu^{t,w}\mu^{u,v}\left(\mu_{r,tu}\mu_{swv} - 2\mu_{rt,u}\mu_{swv} + \mu_{r,tu}\mu_{s,wv} - 2\mu_{rt,u}\mu_{s,wv}\right)\,. \end{aligned} \tag{3.11}$$

This last expression results from the use of the third order Bartlett identity (see (A.5) in Appendix B) which ensures that

$$\mu_{swv} + \mu_{s,wv} = -\mu_{sw,v} - \mu_{w,sv} - \mu_{s,w,v}\,.$$

In (3.11) the terms $\mu^{t,u}\left(\mu_{rs,tu} + \mu_{r,s,tu} - 2\mu_{rst,u} - 2\mu_{rt,su}\right)$ and $\mu^{t,w}\mu^{u,v}\mu_{r,tu}\mu_{s,wv}$ are symmetric under interchanges of the indices r and s. The remainder is

$$\mu^{t,u}\left(\mu_{r,stu} - 2\mu_{rt,s,u}\right) + \mu^{t,w}\mu^{u,v}\left(\mu_{r,tu}\mu_{swv} - 2\mu_{rt,u}\mu_{swv} - 2\mu_{rt,u}\mu_{s,wv}\right)\,, \tag{3.12}$$

which is not guaranteed to be symmetric. Thus the integral of the modified scores is not path independent, and so for general exponential families the existence of a pseudo-likelihood corresponding to the modified scores is not guaranteed. Specifically, a pseudo-likelihood corresponding to the modified scores exists if and only if (3.12) is invariant under interchanges of r and s.

3.3 The bias-reduction method and statistical curvature

Given that the integral of the modified scores based on the Fisher information exists, expression (3.10) corresponds to a penalized likelihood of the form

$$\text{“likelihood”} \times \text{“Jeffreys prior”} \times \text{“extra penalty”} .$$

Also, by (3.9), in the case of flat exponential families, the extra penalty term disappears, which suggests that the bias-reduction method depends on the curvature of the family. The following example illustrates this argument.

Using standard notation, assume that $\mathcal{F}_\beta$ is indexed by a scalar parameter β. Further, assume we formulate the log-likelihood $l(\beta)$ and obtain the score functions $U(\beta)$ for β. The modified score function using expected information is

$$U^*(\beta) = U(\beta) + \frac{\mu_{1,1,1}(\beta) + \mu_{1,2}(\beta)}{2\mu_{1,1}(\beta)} ,$$

where

$$\mu_r(\beta) = \mathrm{E}\left(\frac{\partial^r l}{\partial \beta^r}\right) ; \, \mu_{r,s}(\beta) = \mathrm{E}\left(\frac{\partial^r l}{\partial \beta^r}\frac{\partial^s l}{\partial \beta^s}\right) \quad (r, s = 1, 2, \ldots) ,$$

and so on. By (3.10), $U^*(\beta)$ can be written in the form

$$U^*(\beta) = U(\beta) + \frac{1}{2}\frac{d}{d\beta}\log \mu_{1,1}(\beta) - \frac{1}{2}\frac{\mu_{1,2}(\beta)}{\mu_{1,1}(\beta)} .$$

On the other hand, according to Efron (1975), the statistical curvature of $\mathcal{F}_\beta$ is,

$$\gamma_\beta = \left(\frac{\mu_{2,2}(\beta)}{\mu_{1,1}(\beta)^2} - 1 - \frac{\mu_{1,2}(\beta)^2}{\mu_{1,1}(\beta)^3}\right)^{1/2} .$$

Since $\mu_{1,1}(\beta) > 0$, the dependence of U^* to the statistical curvature is revealed through the expression

$$\begin{aligned} U^*(\beta) = U(\beta) &+ \frac{1}{2}\frac{d}{d\beta}\log \mu_{1,1}(\beta) \\ &- \frac{1}{2}\operatorname{sgn}\left(\mu_{1,2}(\beta)\right)\left(\frac{\mu_{2,2}(\beta)}{\mu_{1,1}(\beta)} - \mu_{1,1}(\beta)(1+\gamma_\beta^2)\right)^{1/2} , \end{aligned} \tag{3.13}$$

where $\operatorname{sgn}(x)$ is -1 if $x < 0$, 0 if $x = 0$ and 1 if $x > 0$. The roots of (3.13) are stationary points of

$$l^*(\beta) = l(\beta) + \frac{1}{2}\log F(\beta) + \frac{1}{2}\psi(\beta) ,$$

where $F = \mu_{1,1}$ is the expected information on β and the function ψ is understood as the solution of the differential equation

$$\frac{d\psi}{d\beta} = -\operatorname{sgn}\left(\mu_{1,2}(\beta)\right)\left(\frac{\mu_{2,2}(\beta)}{\mu_{1,1}(\beta)} - \mu_{1,1}(\beta)(1+\gamma_\beta^2)\right)^{1/2} .$$

If $\mathcal{F}_\beta$ is a flat subset of a wider exponential family $\mathcal{F}$, then $\gamma_\beta = 0$ and $\mu_{2,2} = \mu_{1,1}^2$ so that

$$\psi(\beta) = -\int_a^\beta \operatorname{sgn}\left(\mu_{1,2}(t)\right) \left(\frac{\mu_{2,2}(t)}{\mu_{1,1}(t)} - \mu_{1,1}(t)(1+\gamma_t^2)\right)^{1/2} dt = c\,,$$

with c a real constant and with a an arbitrary, fixed element of the parameter space. Thus, in this case, the penalized log-likelihood takes the form

$$l^*(\beta) = l(\beta) + \frac{1}{2}\log\det F(\beta)\,, \tag{3.14}$$

which, as expected, coincides with the result which Firth (1993) proved for flat exponential families.

3.4 Parameterization invariance for penalized likelihoods

In the previous section we discussed the existence of the penalized likelihood corresponding to the modified scores. Moving a bit further and assuming the existence of a penalized likelihood, its invariance is not generally guaranteed for one-to-one transformation of the parameters, as the invariance of the ordinary likelihood is. The reason is that the purpose of the modified scores is bias reduction and bias itself is a parameterization-wise defined quantity. More formally, consider the case of modifications based on the expected information. Using index notation, the modified scores for β are

$$U_r^* = U_r + \mu^{s,t}\left(\mu_{r,st} + \mu_{r,s,t}\right)/2\,. \tag{3.15}$$

Consider an injective and smooth re-parameterization $\gamma = \phi(\beta)$ and consider the modified scores on γ

$$\bar{U}_r^* = \bar{U}_r + \bar{\mu}^{s,t}\left(\bar{\mu}_{r,st} + \bar{\mu}_{r,s,t}\right)/2\,,$$

with all the barred quantities referring to γ parameterization. Also, let $\beta_s^r = \partial\beta^r/\partial\gamma^s$, $\beta_{st}^r = \partial^2\beta^r/\partial\gamma^s\partial\gamma^t$ and $\gamma_s^r = \partial\gamma^r/\partial\beta^s$. By these definitions $\beta_s^r\gamma_t^s = \delta_t^r$, with δ_t^r the Kronecker delta function with value 1 for $r = t$ and 0 else.

By Section A.3 in Appendix A, the ordinary score functions U_r transform as covariant tensors since $\bar{U}_r = \beta_r^s U_s$. Also, by Example A.3.1 in Appendix A, the Fisher information transforms as a covariant tensor and its matrix-inverse as a contravariant one. So, $\bar{\mu}_{r,s} = \beta_r^t\beta_s^u\mu_{t,u}$ for the Fisher information and $\bar{\mu}^{r,s} = \gamma_t^r\gamma_u^s\mu^{t,u}$ for its matrix-inverse. Directly by their definition, for the joint null moments $\mu_{r,s,t}$ and $\mu_{r,st}$ we have that

$$\bar{\mu}_{r,s,t} = \beta_r^u\beta_s^v\beta_t^w\mu_{u,v,w}$$

and

$$\bar{\mu}_{r,st} = \beta_r^u\beta_s^v\beta_t^w\mu_{u,vw} + \beta_r^u\beta_{st}^v\mu_{u,v}\,,$$

and so $\mu_{r,s,t}$ transforms as a covariant tensor and $\mu_{r,st}$ is not a tensor on account of the presence of the second derivatives with respect to γ in the above formulae. Substituting

in (3.15) we have

$$\begin{aligned}\bar{U}_r^* &= \beta_r^{r_1} U_{r_1} + \frac{1}{2}\gamma_{s_1}^{s}\gamma_{t_1}^{t}\mu^{s_1,t_1}\left(\beta_r^{r_2}\beta_s^{s_2}\beta_t^{t_2}\mu_{r_2,s_2t_2} + \beta_r^{r_2}\beta_{st}^{s_2}\mu_{r_2,s_2} + \beta_r^{r_2}\beta_s^{s_2}\beta_t^{t_2}\mu_{r_2,s_2,t_2}\right)\\ &= \beta_r^{r_1} U_{r_1} + \frac{1}{2}\beta_r^{r_1}\mu^{s_1,t_1}\left(\mu_{r_1,s_1t_1} + \mu_{r_1,s_1,t_1}\right) + \frac{1}{2}\gamma_{s_1}^{s}\gamma_{t_1}^{t}\beta_r^{r_2}\beta_{st}^{s_2}\mu^{s_1,t_1}\mu_{r_2,s_2}\\ &= \beta_r^{r_1} U_{r_1}^* + \frac{1}{2}\gamma_{s_1}^{s}\gamma_{t_1}^{t}\beta_r^{r_2}\beta_{st}^{s_2}\mu^{s_1,t_1}\mu_{r_2,s_2}\,,\end{aligned}$$

and thus U_r^* does not transform as a tensor because the second summand of the last expression depends on β_{st}^r. Hence, the value of the corresponding penalized log-likelihood $l^*(\beta)$ is not generally parameterization invariant because $\bar{l}^*(\gamma) = \bar{l}^*(\phi(\beta)) \neq l^*(\phi^{-1}(\gamma)) = l^*(\beta)$. However notice that for affine transformations $\gamma^r = a^r + a_s^r\beta^s$, with a^r and a_s^r some constants, we have that $\beta_{st}^r = 0$ and so $\bar{U}_r^* = \beta_r^{r_1} U_{r_1}^*$. Thus $\bar{l}^*(\gamma) = l^*(\phi^{-1}(\gamma)) = l^*(\beta)$ and the penalized log-likelihood is invariant under affine transformations. This is a direct consequence of the tensorial properties of the bias of an estimator under the group of linear parameter transformations. Given that R_{rs} and $\bar{R}_r$ are tensors, the above discussion extends for penalized likelihoods corresponding to the more general family of modifications defined in (3.7).

One might argue that from a philosophical point of view, the corresponding penalized log-likelihoods are not appropriate for statistical inference, because a change in parameterization might alter the inferences made. However, for statistical applications and after the choice of a specific parameterization has been made, the modified score functions can be used to obtain first-order unbiased counterparts to the ML estimator and in several situations, as we will see in later chapters, estimators that are superior to the ML estimator in various other ways.

3.5 Consistency and asymptotic normality of the bias-reduced estimator

The consistency and the asymptotic normality of the BR estimator are direct consequences of the general results in Section 6.3 and in Section 6.5 by using any bias-reducing modification in their derivations. If $\tilde{\beta}$ is the BR estimator for exponential family non-linear models with known dispersion and β_0 is the true but unknown value of the parameter vector, then under the regularity conditions of Subsection 2.3.1, $\tilde{\beta} - \beta_0 = o_p(1)$ and specifically $(\tilde{\beta} - \beta_0)$ has asymptotically a p-dimensional normal distribution with zero mean and variance-covariance matrix the inverse of the Fisher information. The proof for the consistency of $\tilde{\beta}$ as given in Section 6.3 depends on the consistency of the ML estimator and, in turn, the consistency of the ML estimator relies heavily on the assumption of it taking always finite values in a compact subset of the parameter space B. However, there might be cases of exponential family models, like logistic regression, where this assumption is not generally valid. In these cases, we could treat $\tilde{\beta}$ as a "Z-estimator" and possibly proceed according to the consistency proofs in van der Vaart (1998, § 5.2).

3.6 Modified scores for exponential family non-linear models: Multivariate Responses

3.6.1 Multivariate response generalized non-linear models

3.6.1.1 General links

We consider the usual setting of Chapter 2, that is realizations $y_1, y_2, \ldots, y_n$ of independent q-dimensional random vectors $Y_1, Y_2, \ldots, Y_n$ from an exponential family of distributions. Also, for their expectations, consider an exponential family non-linear model with known dispersion in its most general form (2.5). Then, the sum of the cumulant matrices P_t and Q_t is given by

$$\begin{aligned} P_t + Q_t = & \sum_r \sum_{s=1}^{q} Z_r^T ([D_r \Sigma_r^{-1}]_s \otimes 1_q) \mathcal{D}^2 (\mu_r; \eta_r) Z_r z_{rst} \\ & + \sum_r \sum_{s=1}^{q} (W_{rs} \otimes 1_q) \mathcal{D}^2 (\eta_r; \beta) z_{rst} , \end{aligned}$$

where W_{rs} is the s-th row of the $q \times q$ matrix $W_r = D_r \Sigma_r^{-1} D_r^T$ as a $1 \times q$ vector and $[D_r \Sigma_r^{-1}]_s$ is the s-th row of $D_r \Sigma_r^{-1}$ as a $1 \times q$ vector.

Hence, the modified scores (3.8) for exponential family non-linear models with known dispersion are written as

$$\begin{aligned} U_t^* = U_t & + \frac{1}{2} \sum_r \sum_{s=1}^{q} \operatorname{trace} \left\{ H_r W_r^{-1} ([D_r \Sigma_r^{-1}]_s \otimes 1_q) \mathcal{D}^2 (\mu_r; \eta_r) \right\} \sum_{u=1}^{p} e_{tu} z_{rsu} \qquad (3.16) \\ & + \frac{1}{2} \sum_r \sum_{s=1}^{q} \operatorname{trace} \left\{ F^{-1} (W_{rs} \otimes 1_q) \mathcal{D}^2 (\eta_r; \beta) \right\} \sum_{u=1}^{p} e_{tu} z_{rsu} \quad (t = 1, \ldots, p) , \end{aligned}$$

with $H_r = Z_r F^{-1} Z_r^T W_r$ (see Subsection 2.4.3 for a description of H_r) and e_{tu} as in (3.8). For analytic derivations of the above results see Section B.2 in Appendix B.

The usefulness of expression (3.16) lies in the fact that it involves quantities which are easily obtained once a specific exponential family non-linear model is selected and which usually are readily available in the output of standard computing packages. The modified scores could be expressed in terms of the derivatives of the logarithm of Jeffreys invariant prior, following (3.10). However, while this would reveal the dependence on Jeffreys prior, it would involve more complicated terms than (3.16).

3.6.1.2 Canonical Links

In the case of canonically linked models (2.7), $D_r = \lambda_r^{-1} \Sigma_r$ and so $W_r = \lambda_r^2 \Sigma_r$. Further, by (2.4)

$$([D_r \Sigma_r^{-1}]_s \otimes 1_q) \mathcal{D}^2 (\mu_r; \eta_r) = \lambda_r^{-1} \mathcal{D}^2 (\mu_{rs}; \eta_r) = \lambda_r^{-3} K_{rs} ,$$

where K_{rs} denotes the s-th block of rows of K_r, $s = 1, \ldots, q$, with K_r the blocked $q^2 \times q$ matrix of third-order cumulants of the random vector Y_r. Thus, the expression (3.16) is

considerably simplified and the modified scores take the form

$$U_t^* = U_t + \frac{1}{2}\sum_r \sum_{s=1}^{q} \lambda_r^{-1} \operatorname{trace}\left\{H_r \Sigma_r^{-1} K_{rs}\right\} \sum_{u=1}^{p} e_{tu} z_{rsu} \tag{3.17}$$
$$+ \frac{1}{2}\sum_r \sum_{s=1}^{q} \lambda_r^{-2} \operatorname{trace}\left\{F^{-1}(\Sigma_{rs} \otimes 1_p)\mathcal{D}^2\left(\eta_r;\beta\right)\right\} \sum_{u=1}^{p} e_{tu} z_{rsu} \quad (t = 1, \ldots, p) \ .$$

3.6.2 Multivariate-response generalized linear models

3.6.2.1 General links

Generalized linear models have the form (2.5), with $\eta_r(\beta) = Z(x_r)$, where the $q \times p$ design matrix $Z(x_r)$ $(r = 1, \ldots, n)$ is a function of the covariate vector x_r and does not depend on β. Thus, in (3.16), $\mathcal{D}^2\left(\eta_r;\beta\right) = 0$ and the modified scores reduce to

$$U_t^* = U_t + \frac{1}{2}\sum_r \sum_{s=1}^{q} \operatorname{trace}\left\{H_r W_r^{-1}([D_r \Sigma_r^{-1}]_s \otimes 1_q)\mathcal{D}^2\left(\mu_r;\eta_r\right)\right\} \sum_{u=1}^{p} e_{tu} z_{rsu} \ , \tag{3.18}$$

for $t = 1, \ldots, p$. Again, all the quantities, except $Z_r \equiv Z_r(x)$, in (3.18) are functions of β.

3.6.2.2 Canonical Links

By the same arguments as in the derivation of (3.17), for canonical link functions the modified scores have the form

$$U_t^* = U_t + \frac{1}{2}\sum_r \sum_{s=1}^{q} \lambda_r^{-1} \operatorname{trace}\left\{H_r \Sigma_r^{-1} K_{rs}\right\} \sum_{u=1}^{p} e_{tu} z_{rsu} \quad (t = 1, \ldots, p) \ . \tag{3.19}$$

3.7 Modified scores for exponential family non-linear models: Univariate Responses

3.7.1 Univariate-response generalized non-linear models

3.7.1.1 General links

Suppose now that the response variable is scalar. All of the above applies, just by dropping the dimension of the response to $q = 1$. For notational simplicity, in the univariate case, write $\kappa_{2,r} = \operatorname{Var}(Y_r)$ and $\kappa_{3,r} = \operatorname{Cum}_3(Y_r)$ for the variance and the third cumulant of Y_r, respectively. The modified scores for an exponential family non-linear model with known dispersion are written in the form

$$U_t^* = U_t + \frac{1}{2}\sum_r h_r \frac{d_r'}{d_r} \sum_{s=1}^{p} e_{ts} z_{rs} \tag{3.20}$$
$$+ \frac{1}{2}\sum_r w_r \operatorname{trace}\left\{F^{-1}\mathcal{D}^2\left(\eta_r;\beta\right)\right\} \sum_{s=1}^{p} e_{ts} z_{rs} \quad (t = 1, \ldots, p) \ .$$

In (3.20), $z_{rs} = \partial\eta_r/\partial\beta_s$, $d_r = \partial\mu_r/\partial\eta_r$, $d'_r = \partial^2\mu_r/\partial\eta_r^2$, $w_r = d_r^2/\kappa_{2,r}$ and $\mathcal{D}^2(\eta_r;\beta)$ is the $p \times p$ Hessian matrix of η_r with respect to β. The quantity h_r is the r-th diagonal element of the projection matrix

$$H = ZF^{-1}Z^TW\,,$$

where, if $\eta = (\eta_1, \ldots, \eta_n)^T$, Z is the $n \times p$ Jacobian of η with respect to β and $W = \text{diag}\{w_r;\, r = 1, 2, \ldots, n\}$.

3.7.1.2 Canonical links

For canonical link functions in the non-linear case, we have $d_r = \lambda_r^{-1}\kappa_{2,r}$ and $d'_r = \lambda_r^{-2}\kappa_{3,r}$. Thus

$$\begin{aligned} U_t^* = U_t &+ \frac{1}{2}\sum_r \lambda_r^{-1} h_r \frac{\kappa_{3,r}}{\kappa_{2,r}} \sum_{s=1}^{p} e_{ts} z_{rs} \\ &+ \frac{1}{2}\sum_r \lambda_r^{-2}\kappa_{2,r}\,\text{trace}\left\{F^{-1}\mathcal{D}^2(\eta_r;\beta)\right\} \sum_{s=1}^{p} e_{ts} z_{rs} \quad (t = 1, \ldots, p)\,, \end{aligned} \tag{3.21}$$

3.7.2 Univariate-response generalized linear models

3.7.2.1 General links

For univariate GLMs, the third summand in the expression (3.20) disappears and $z_{rs} = x_{rs}$ does not depend on β, leaving us with the following form for the modified scores for general link functions

$$U_t^* = U_t + \frac{1}{2}\sum_r h_r \frac{d'_r}{d_r} \sum_{s=1}^{p} e_{ts} x_{rs} \quad (t = 1, \ldots, p)\,. \tag{3.22}$$

Note that the term d'_r/d_r depends solely on the inverse of the link function.

3.7.2.2 Canonical links

Canonically-linked GLMs are flat exponential families and further simplification is possible since the Fisher and the observed information coincide. Hence, in this case,

$$U_t^* = U_t + \frac{1}{2}\sum_r \lambda_r^{-1} h_r \frac{\kappa_{3,r}}{\kappa_{2,r}} \sum_{s=1}^{p} e_{ts} x_{rs} \quad (t = 1, \ldots, p)\,, \tag{3.23}$$

where e_{ts} simplifies to the (t, s)-th element of $1_p + RF^{-1}$ for $e_{ts} \equiv e_{ts}^{(O)} = e_{ts}^{(E)}$. Further, if R is a matrix of zeros we have that

$$U_t^* = U_t + \frac{1}{2}\sum_r \lambda_r^{-1} h_r \frac{\kappa_{3,r}}{\kappa_{2,r}} x_{rt}\,, \tag{3.24}$$

which is the same elegant result given in Firth (1992a,b).

3.7.3 Relation to Cordeiro & McCullagh (1991) and pseudo-responses

In the case of univariate GLMs, the above expressions are directly connected to the results in Cordeiro & McCullagh (1991) (see also McCullagh & Nelder, 1989, § 15.2). For (3.22) we have that

$$\begin{aligned} U_t^* &= U_t + \frac{1}{2}\sum_r h_r \frac{d'_r}{d_r}\sum_{s=1}^p e_{ts}x_{rs} \qquad (3.25)\\ &= U_t - \sum_r h_r \frac{\xi_r}{S_{rr}}\sum_{s=1}^p e_{ts}x_{rs} \quad (t=1,\ldots,p)\ , \end{aligned}$$

with

$$\xi_r = -\frac{1}{2}\frac{d'_r}{d_r}S_{rr}\,,$$

as defined in Cordeiro & McCullagh (1991), and S_{rr} the r-th diagonal element of $S = X(X^TWX)^{-1}X^T$. The $n \times n$ matrix S is the asymptotic variance-covariance matrix of the ML estimator of $\eta = (\eta_1, \ldots, \eta_n)^T$. Noting that $h_r = S_{rr}w_r$, (3.25) is written as

$$U_t^* = U_t - \sum_r w_r\xi_r \sum_{s=1}^p e_{ts}x_{rs} \quad (t=1,\ldots,p)\ ,$$

so that the vector of modified scores is given by

$$U^* = U - X^{*T}W\xi\,,$$

where X^* is the $n \times p$ matrix with (r,t)-th component $x^*_{rt} = \sum_{s=1}^p e_{ts}x_{rs}$. Letting $D = \mathrm{diag}\{d_r; r = 1,\ldots,n\}$, the ordinary scores in this case have the usual form $U = X^TWD^{-1}(y-\mu)$, with $y = (y_1,\ldots,y_n)^T$ and $\mu = (\mu_1,\ldots,\mu_n)^T$. So, the vector of modified scores can be written in the form

$$U^* = X^TWD^{-1}(y-\mu) - X^{*T}W\xi\,,$$

This re-expression is a direct consequence of the initial definition of the modifications (see (3.1) and (3.2)) and the fact that, as is shown in Cordeiro & McCullagh (1991), the vector of the first-order biases of the ML estimator (see (3.6) for its expression in terms of $A^{(E)}$) can be written as

$$n^{-1}b_1 = (X^TWX)^{-1}X^TW\xi\,,$$

which is, also, the basis for the supplementary re-weighted least squares approach to bias correction therein.

Moving a bit further, in the case of modifications based on the expected information where $e_{ts} = 1$ if $t = s$ and 0 else, we have that $X^* = X$. Thus,

$$U^* = X^TWD^{-1}(y - D\xi - \mu)\,,$$

which has components

$$U_t^* = \sum_r \frac{d_r}{\kappa_{2,r}} \left(y_r - d_r\xi_r - \mu_r\right) x_{rt} \tag{3.26}$$
$$= \sum_r \frac{d_r}{\kappa_{2,r}} \left(y_r + \frac{1}{2} h_r \frac{d_r'}{w_r} - \mu_r\right) x_{rt} \quad (t = 1, \ldots, p) \; .$$

This latter expression reveals an important feature of the modified scores in this setting that will be used extensively in later chapters. If $h_r d_r'/(2w_r)$ or equivalently $-d_r\xi_r$ $(r = 1, \ldots, n)$ were known constants then the bias-reduction method would be formally equivalent to maximum likelihood when the pseudo-responses

$$y_r^* = y_r + \frac{1}{2} h_r \frac{d_r'}{w_r} \quad (r = 1, \ldots, n) \; ,$$

(or equivalently $y_r^* = y_r - d_r\xi_r$) are used instead of y_r. In this way the implementation of fitting procedures for obtaining the BR estimates is greatly facilitated, since we can just replace y_r by y_r^* in the IWLS step. Note that generally $h_r d_r'/w_r$ depends on the parameters and so the value of y_r^* will be updated according to the current estimates at each step of the cycle. For canonical link functions, $d_r = \lambda_r^{-1}\kappa_{2,r}$, $w_r = \lambda_r^{-2}\kappa_{2,r}$ and $d_r' = \lambda_r^{-2}\kappa_{3,r}$ and so the pseudo-responses reduce to

$$y_r^* = y_r + \frac{1}{2} h_r \frac{\kappa_{3,r}}{\kappa_{2,r}} \quad (r = 1, \ldots, n) \; ,$$

as is, also, directly apparent by (3.24). In Table 3.2 we derive the form of the pseudo-responses for some commonly used GLMs. The modified IWLS step can more conveniently be described in terms of modified working observations: replacing y_r with y_r^* in (2.15) we have

$$\zeta_r^* = \eta_r + \frac{y_r^* - \mu_r}{d_r} \tag{3.27}$$
$$= \eta_r + \frac{y_r - \mu_r}{d_r} - \xi_r = \zeta_r - \xi_r \; .$$

Thus, if we modify the working observation ζ_r by adding $-\xi_r = h_r d_r'/(2d_r w_r)$ to it, the iteration of the usual IWLS scheme returns the BR estimates.

An issue that could arise when using the pseudo-responses in already implemented fitting software relates to the sign of $h_r d_r'/w_r$. If the response has a restricted range, which is the most usual case (for example, positive real for gamma and inverse Gaussian and non-negative integer for binomial and Poisson responses), the pseudo-responses could violate the restriction. Since $h_r \in [0, 1]$ and the working weight w_r is necessarily non-negative as the ratio of a square over a variance, the sign of $h_r d_r'/w_r$ is the sign of d_r' and thus it directly depends on the link function. Many implementations of fitting procedures — correctly — refuse to proceed under violations of the response range. For example, in order to fit a binomial-response GLM through the `glm` procedure in *R language* (R

Development Core Team, 2007) the response must be in $[0, 1]$ and outside this range an error message is returned. So, special care is needed when using the pseudo-responses within readily available procedures. In special cases this could be dealt by simple algebraic manipulation of the pseudo-data representation. For example, as is done in Chapter 5, for binomial response models we can 'trade' quantities between the pseudo-responses and the binomial totals so that the positivity of both is ensured. However this raises other issues, that are discussed in Chapter 5.

These elegant results do not extend in any obvious way in the case of *multivariate* responses, on account of the fact that W is no longer diagonal but is block diagonal and because the vector of first-order biases can be expressed, at best, as a function of traces of products of matrices with no other apparent simplification. Also, for univariate generalized *non-linear* models, expressions of this elegance cannot in general be obtained because of the existence of the third term in the right hand side of (3.20). This term corresponds to the extra term in the expression for the first-order bias that incorporates the non-linearity of the predictor.

3.7.4 Existence of penalized likelihoods for univariate GLMs

In the case of general exponential families, we have derived a general necessary and sufficient condition for the existence of a penalized likelihood corresponding to the modified scores based on the expected information (see Subsection 3.2.3). The derivation was based on the re-expression of the modified scores as ordinary scores plus derivative of the logarithm of Jeffreys prior plus an extra term. This extra term is zero for canonical families and so the modified scores correspond to penalization of the likelihood by Jeffreys prior.

Here we present a more specialized but not less important result which asserts that within the class of univariate GLMs, a penalized likelihood corresponding to the modified scores based on the expected information exists if and only if the canonical link is used.

Theorem 3.7.1: *Within the class of univariate generalized linear models, a penalized likelihood corresponding to the modified scores based on the expected information exists if and only if the link is canonical.*

Proof. Noting that $d_r'/d_r x_{rt} = \partial \log d_r/\partial\beta_t$ and by (3.22), the modified scores based on the expected information can be written as

$$U_t^* = U_t + \frac{1}{2}\,\mathrm{trace}\{HE_t\} \quad (t = 1, \ldots, p)\ , \tag{3.28}$$

where $E_t = \mathrm{diag}\{\partial \log d_r/\partial\beta_t; r = 1, \ldots, n\}$.

The vector of modified scores corresponds to a pseudo-likelihood expression if and only if

$$\frac{\partial U_s^*(\beta)}{\partial \beta_t} = \frac{\partial U_t^*(\beta)}{\partial \beta_s}\,,$$

for every $s, t \in \{1, 2, \ldots, p\}$, which by (3.28) reduces to the condition

$$\frac{\partial}{\partial \beta_t}\,\mathrm{trace}\{HE_s\} = \frac{\partial}{\partial \beta_s}\,\mathrm{trace}\{HE_t\}\,.$$

So, a necessary and sufficient condition for U_t^* $(t = 1, \ldots, p)$ to be integrable over the parameter space is that $\partial \text{trace}\{HE_t\}/\partial\beta_s$ is invariant under interchanges of the subscripts s and t. We have

$$\frac{\partial}{\partial\beta_s}\text{trace}\{HE_t\} = \text{trace}\left\{\left(\frac{\partial}{\partial\beta_s}H\right)E_t\right\} + \text{trace}\left\{H\left(\frac{\partial}{\partial\beta_s}E_t\right)\right\}.$$

Note that $\partial E_t/\partial\beta_s = \text{diag}\{\partial^2 \log d_r/\partial\beta_s\partial\beta_t; r = 1, \ldots, n\}$. Hence the second term in the above expression is invariant under interchanges of s and t. Thus, the condition reduces to involve only the first term. Now,

$$\begin{aligned}\frac{\partial}{\partial\beta_s}H &= \frac{\partial}{\partial\beta_s}\left(X(X^TWX)^{-1}X^TW\right) \\ &= -X(X^TWX)^{-1}X^TW_sX(X^TWX)^{-1}X^TW + X(X^TWX)^{-1}X^TW_s\,,\end{aligned} \tag{3.29}$$

where $W_s = \text{diag}\{\partial w_r/\partial\beta_s; r = 1, \ldots, n\}$. But

$$\begin{aligned}\frac{\partial}{\partial\beta_s}w_r &= \frac{\partial}{\partial\beta_s}\frac{d_r^2}{\kappa_{2,r}} \\ &= 2\frac{d_r}{\kappa_{2,r}}\frac{\partial}{\partial\beta_s}d_r - \frac{d_r^2}{\kappa_{2,r}^2}\frac{\partial}{\partial\beta_s}\kappa_{2,r} \\ &= w_r(2\frac{\partial}{\partial\beta_s}\log d_r - \frac{\partial}{\partial\beta_s}\log\kappa_{2,r})\,,\end{aligned}$$

so that $W_s = W(2E_s - \Lambda_s)$, where $\Lambda_s = \text{diag}\{\partial\log\kappa_{2,r}/\partial\beta_s; r = 1, \ldots, n\}$. Substituting in (3.29) we have

$$\begin{aligned}\frac{\partial}{\partial\beta_s}H &= -X(X^TWX)^{-1}X^TW(2E_s - \Lambda_s)X(X^TWX)^{-1}X^TW \\ &\quad + X(X^TWX)^{-1}X^TW(2E_s - \Lambda_s) \\ &= H(2E_s - \Lambda_s)(1_n - H)\,,\end{aligned}$$

where 1_n is the $n \times n$ identity matrix. Thus,

$$\begin{aligned}\text{trace}\left\{\left(\frac{\partial}{\partial\beta_s}H\right)E_t\right\} &= \text{trace}\left\{H(2E_s - \Lambda_s)(1_n - H)E_t\right\} \\ =&2\,\text{trace}\{HE_sE_t\} - 2\,\text{trace}\{HE_sHE_t\} - \text{trace}\{H\Lambda_sE_t\} + \text{trace}\{H\Lambda_sHE_t\}\,.\end{aligned} \tag{3.30}$$

Note that the first two terms of the latter expression are invariant under interchanges of s and t, because E_t is diagonal and because of the properties of the trace function. For the remaining terms note that $\partial\log d_r/\partial\beta_t = x_{rt}\partial\log d_r/\partial\eta_r$, so that $E_t = \tilde{E}\tilde{X}_t$, with $\tilde{E} = \text{diag}\{\partial\log d_r/\partial\eta_r; r = 1, \ldots, n\}$ and $\tilde{X}_t = \text{diag}\{x_{rt}; r = 1, \ldots, n\}$. By the same argument $\Lambda_s = \tilde{\Lambda}\tilde{X}_s$, with obvious notational correspondences. So, for the third term of (3.30) we have

$$\text{trace}\{H\Lambda_sE_t\} = \text{trace}\{H\tilde{\Lambda}\tilde{X}_s\tilde{E}\tilde{X}_t\} = \text{trace}\{H\tilde{\Lambda}\tilde{X}_t\tilde{E}\tilde{X}_s\}\,.$$

The last equality is satisfied because $\tilde{\Lambda}$, $\tilde{X}_s$, $\tilde{E}$ are by definition diagonal and so matrix multiplication is commutative for them. So, the condition is reduced to the invariance of $\text{trace}\{H\Lambda_s H E_t\}$. The projection matrix H can be written as $H = SW$, where S is as defined in the previous section. Thus, using diagonality and the properties of trace we have

$$\text{trace}\{H\Lambda_s H E_t\} = \text{trace}\{SW\tilde{\Lambda}\tilde{X}_s SW\tilde{E}\tilde{X}_t\} = \text{trace}\{\tilde{X}_t S\tilde{X}_s W\tilde{\Lambda}S\tilde{E}W\}\,. \tag{3.31}$$

Changing the position of s and t we have

$$\text{trace}\{H\Lambda_t H E_s\} = \text{trace}\{SW\tilde{\Lambda}\tilde{X}_t SW\tilde{E}\tilde{X}_s\} = \text{trace}\{\tilde{X}_t S\tilde{X}_s W\tilde{E}S\tilde{\Lambda}W\}\,. \tag{3.32}$$

Hence (3.31) and (3.32) are equal if and only if $\tilde{\Lambda}S\tilde{E}$ is symmetric or alternatively, by the symmetry of S, if and only if

$$\frac{\partial}{\partial\eta_r}\log d_r = c_r\frac{\partial}{\partial\eta_r}\log\kappa_{2,r} \quad (r = 1,\ldots,n)\ ,$$

with $\{c_r\}$ a sequence of real constants. This equality is valid if and only if the link function is the canonical one, since then and only then $d_r = \lambda_r^{-1}\kappa_{2,r}$. This completes the proof. □

3.8 General remarks

We have derived explicit, general formulae for the modified scores that produce first-order unbiased estimators, starting from the wide class of multivariate-response exponential family non-linear models and narrowing down to the simple case of canonically-linked GLMs. It should be noted that further simplification of the formulae is possible for other special cases, for example generalized bilinear models, by exploiting the specific structure of the various quantities involved.

Statistical properties of the penalized-likelihood estimator, on the other hand, must be examined case by case. For example, Heinze & Schemper (2002) illustrated that in binomial logistic regression, ordinary ML can beneficially be replaced by the penalized version because of the clear shrinkage interpretation, ensuring at the same time finiteness of the BR estimates even in separated cases where the ML estimate is infinite-valued. In contrast, in some other situations the reduction of bias in this way may not be beneficial. An example is the estimation of the mean of a normal population with known coefficient of variation, where reduction in bias is accompanied by inflation of variance (Firth, 1993).

Further, we note that the ML and the BR estimator agree in their first-order asymptotic variance-covariance matrix in regular problems, this being the inverse of the Fisher information.

Finally, the only restriction we imposed in the general family of modifications (3.7) was that $\bar{R}_r = 0$. This enabled the derivation of generic formulae for the modified scores for the whole class of exponential family non-linear models, maintaining at the same time a certain level of elegance in the expressions.

In later chapters, while we give the general form of modifications (with $\bar{R}_r = 0$), we focus only on modifications based on the expected information. This is done because i) they

have the simplest form in contrast to the other members of the family and ii) as already seen for the case of univariate GLMs, they allow the most elegant theoretical derivations and enable the ready implementation of fitting procedures in practice. More complicated modifications can be obtained using modifications based either on the score function or the observed information and/or controlling both R_{rs} and $\bar{R}_r$ under the restriction of having expectations of order at most $\mathcal{O}(n^{1/2})$ and $\mathcal{O}(n^{-1/2})$, respectively. However, all modifications are asymptotically equivalent and as already discussed, in terms of bias they all result in first-order unbiased estimators. The choice among them should be application dependent and possibly be made towards further improvement of other asymptotic properties of the resultant estimators. This is the subject of further work and will not be pursued in the current thesis.

Table 3.1: Characteristics of commonly used exponential families with known dispersion.

	Normal	Binomial	Poisson	Gamma	Inverse Gaussian
Range of y	$(-\infty, +\infty)$	$0, 1, \ldots, m$	$0, 1, \ldots$	$[0, +\infty)$	$(0, +\infty)$
Parameters	$\mu \in \Re$, $\sigma^2 > 0$	$\pi \in (0,1)$, $m = 1, 2, \ldots$	$\mu > 0$	$\mu > 0$, $\nu > 0$	$\lambda > 0$, $\mu > 0$
Density or pmf	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{(y-\mu)^2}{2\sigma^2}\right\}$	$\binom{m}{y}\pi^y(1-\pi)^{m-y}$	$\frac{\exp\{-\mu\}\,\mu^y}{y!}$	$\frac{\left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1}\exp\left\{-\frac{\nu}{\mu}y\right\}}{\Gamma(\nu)}$	$\frac{1}{\sqrt{2\pi\lambda y^3}}\exp\left\{-\frac{(y-\mu)^2}{2\lambda\mu^2 y}\right\}$
λ	σ^2	1	1	$1/\nu$	λ
$b(\theta)$	$\theta^2/2$	$m\log(1+e^\theta)$	$\exp(\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$\mu = \mu_*(\theta) = \mathrm{E}(Y;\theta)$	μ	$me^\theta/(1+e^\theta)$	$\exp(\theta)$	$-1/\theta$	$(-2\theta)^{-1/2}$
$\kappa_2 = \mathrm{Var}(Y)$	σ^2	$m\pi(1-\pi)$	μ	μ^2/ν	$\lambda\mu^3$
$\kappa_3 = \mathrm{Cum}(Y)$	0	$m\pi(1-\pi)(1-2\pi)$	μ	$2\mu^3/\nu^2$	$3\lambda^2\mu^5$
$\theta = \theta_*(\mu)$	μ	$\log(\pi/(1-\pi))$	$\log\mu$	$-1/\mu$	$-1/(2\mu^2)$

Table 3.2: Derivation of pseudo-responses for several commonly used GLMs (see Section 3.7).

Distribution	Link function $\eta = g(\mu)$ (canonical link: *)	Inverse link function $\mu = h(\eta)$	$d = \partial\mu/\partial\eta$	$d' = \partial^2\mu/\partial\eta^2$	Quadratic weight $w = d^2/\kappa_2$	Pseudo-responses $y^* = y + hd'/(2w)$
Normal (μ, σ^2)	$\eta = \mu$ *	$\mu = \eta$	1	0	$1/\sigma^2$	$y^* = y$
Binomial (m, π)	$\eta = \log\frac{\pi}{1-\pi}$ *	$\pi = \frac{e^\eta}{1+e^\eta}$	$m\pi(1-\pi)$	$m\pi(1-\pi)(1-2\pi)$	$m\pi(1-\pi)$	$y^* = y + h(1/2 - \pi)$
	$\eta = \Phi^{-1}(\pi)$	$\pi = \Phi(\eta)$	$m\phi(\eta)$	$-m\eta\phi(\eta)$	$m\frac{(\phi(\eta))^2}{\pi(1-\pi)}$	$y^* = y - h\pi(1-\pi)\frac{\eta}{2\phi(\eta)}$
	$\eta = \log(-\log(1-\pi))$	$\pi = 1 - e^{-e^\eta}$	$m(1-\pi)e^\eta$	$m(1-\pi)e^\eta(1-e^\eta)$	$m\frac{e^{2\eta}(1-\pi)}{\pi}$	$y^* = y + h\pi\frac{(1-e^\eta)}{2e^\eta}$
	$\eta = -\log(-\log(\pi))$	$\pi = e^{-e^{-\eta}}$	$m\pi e^{-\eta}$	$m\pi e^{-\eta}(e^{-\eta}-1)$	$m\frac{e^{-2\eta}\pi}{1-\pi}$	$y^* = y + h(1-\pi)\frac{e^{-\eta}-1}{2e^{-\eta}}$
Poisson (μ)	$\eta = \log\mu$ *	$\mu = e^\eta$	μ	μ	μ	$y^* = y + \frac{h}{2}$
Gamma (μ, ν)	$\eta = -\frac{1}{\mu}$ *	$\mu = -\frac{1}{\eta}$	μ^2	$2\mu^3$	$\nu\mu^2$	$y^* = y + h\frac{\mu}{\nu}$
	$\eta = \log\mu$	$\mu = e^\eta$	μ	μ	ν	$y^* = y + h\frac{\mu}{2\nu}$
	$\eta = \mu$	$\mu = \eta$	1	0	$\frac{\nu}{\mu^2}$	$y^* = y$
Inverse Gaussian (λ, μ)	$\eta = -\frac{1}{2\mu^2}$ *	$\mu = (-2\eta)^{-1/2}$	μ^3	$3\mu^5$	$\frac{\mu^3}{\lambda}$	$y^* = y + h\frac{3\lambda\mu^2}{2}$

Chapter 4

Bias reduction and logistic regression

4.1 Introduction

Bias correction in logistic regression has attracted the attention of many authors, for example Anderson & Richardson (1979), Schaefer (1983), Copas (1988), Cordeiro & McCullagh (1991) and other references therein.

Recently, Heinze & Schemper (2002) and Zorn (2005) investigated the bias-reduction method, described in the previous chapter, for estimation in binomial-response logistic regression. By extensive empirical studies, they illustrated the superior properties of the resultant estimator relative to the maximum likelihood (ML) estimator. Specifically, they emphasized the finiteness of the bias-reduced (BR) estimates even in cases of complete or quasi-complete separation (see Albert & Anderson, 1984, for definitions) and their shrinkage properties. Corresponding empirical studies in Bull et al. (2002) extended these remarks to the case of multinomial-response logistic regression. They compared the bias-reduction method with other bias correction methods and accorded it a preferred position amongst them, again in terms of the properties of the resultant estimator. However, Bull et al. (2002) do not give any generalization for the elegant form of the modified scores in the binomial case. Instead, due to the involvement of the multinomial variance-covariance matrix in the calculations, they keep a pessimistic attitude towards this direction and they proceed to the empirical studies keeping the unnecessary, for logistic regressions, redundancy in the expressions from the general definition of the bias-reducing modifications.

In this chapter we proceed to a systematic, theoretical treatment of the bias-reduction method for logistic regression by filling the theoretical gaps in the aforementioned work. The chapter is organized in two parts.

The first part deals with binomial-response logistic regression. We briefly review the work in Firth (1992a,b) giving explicit expressions for the modified scores and the iterative re-weighted least squares (IWLS) variant for obtaining the BR estimates. The core of this part consists of new material presenting the theorems and corresponding proofs that

formally attribute to the BR estimator the properties that have been conjectured by the results of the empirical studies in Heinze & Schemper (2002) and Zorn (2005). In this way we round off any previous work for such models by formally concluding that the maximum penalized likelihood is an improvement to the traditional ML approach.

The second part deals with the extension of the results to the multinomial case. More specifically, the simple and elegant form of the modified score equations is derived and discussed, focusing mainly on the way that the results in Firth (1992a,b) generalize in the multinomial setting. It is also shown how the corresponding Poisson log-linear model can be used to derive the BR estimator. An iterative generalized least squares (IGLS) algorithm for the BR estimates is proposed and we illustrate how it proceeds by applying appropriate 'flattening' modifications to the response at each IGLS iteration.

4.2 Binomial-response logistic regression

4.2.1 Modified score functions

Consider realizations $y_1, y_2, \ldots, y_n$ of n independent binomial random variables Y_1, Y_2, ..., Y_n with probabilities of *success* $\pi_1, \pi_2, \ldots, \pi_n$ and binomial totals $m_1, m_2, \ldots, m_n$, respectively. Furthermore, consider a logistic regression model of the form

$$\log \frac{\pi_r}{1-\pi_r} = \eta_r = \sum_{t=1}^{p} \beta_t x_{rt} \quad (r = 1, \ldots, n) \;, \tag{4.1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional parameter vector and x_{rt} is the (r,t)-th element of a $n \times p$ design matrix X, assumed to be of full rank; if an intercept parameter is to be included in the model we can just set the first column of X to be a column of ones. This is a generalized linear model (GLM) with canonical link, and by a substitution of $\kappa_{2,r} = \pi_r(1-\pi_r)$ and $\kappa_{3,r} = \pi_r(1-\pi_r)(1-2\pi_r)$ in (3.23), the modified scores for the parameters β are given by the expression

$$U_t^* = U_t + \frac{1}{2} \sum_{r}^{n} h_r (1 - 2\pi_r) \sum_{s=1}^{p} e_{ts} x_{rs} \quad (t = 1, \ldots, p) \;, \tag{4.2}$$

where $U_t = \sum_r (y_r - m_r \pi_r) x_{rt}$ are the ordinary score functions. In this case $e_{ts}^{(E)} = e_{ts}^{(O)} = [1_p + RF^{-1}]_{ts}$ in (3.8), and if we set $R = 0$ we get the elegant form of the modifications in Firth (1992a,b),

$$\begin{aligned} U_t^* &= \sum_r (y_r - m_r \pi_r) x_{rt} + \frac{1}{2} \sum_r h_r (1 - 2\pi_r) x_{rt} \\ &= \sum_r \left(y_r + \frac{1}{2} h_r - (m_r + h_r) \pi_r \right) x_{rt} \quad (t = 1, \ldots, p) \;. \end{aligned} \tag{4.3}$$

As already mentioned in Subsection 3.2.2, for flat exponential families, the solution of the modified scores equation based on the expected information locates a stationary point

of the penalized log-likelihood

$$l^*(\beta) = l(\beta) + \frac{1}{2}\log\det F(\beta)\,,$$

with $F = X^T W X$ the Fisher information on β, and $W = \mathrm{diag}\{m_r\pi_r(1-\pi_r); r = 1,\ldots,n\}$.

An alternative bias-reducing expression for the modified scores is obtained for $e_{tu} = e_{tu}^{(S)}$. To illustrate the extra complexity of the alternative formulae, for $R = 0$ and for every $t = 1,\ldots,p$ we have

$$\begin{aligned}
U_t^{(S)} &= \sum_r (y_r - m_r\pi_r)x_{rt} + \sum_r h_r(1/2 - \pi_r)\sum_{s=1}^{p}\left[X^T(y-\mu)(y-\mu)^T X(X^T W X)^{-1}\right]_{ts} x_{rs} \\
&= \sum_r (y_r - m_r\pi_r)x_{rt} + \sum_r h_r(1/2 - \pi_r)\left[X(X^T W X)^{-1}X^T(y-\mu)(y-\mu)^T X\right]_{rt} \\
&= \sum_r (y_r - m_r\pi_r)x_{rt} + \sum_r h_r(1/2 - \pi_r)\sum_{s,u=1}^{n} h_{rs}\frac{(y_s-\mu_s)(y_u-\mu_u)}{m_s\pi_s(1-\pi_s)}x_{ut}\,,
\end{aligned}$$

with h_{rs} the (r,s)-th element of H. Apparently the above expression is unwieldy compared to (4.3) mainly because it involves all the elements of the projection matrix H. Given that both expressions result in first-order unbiased estimators and that $U_t^{(S)}$ does not correspond to a penalized log-likelihood in closed form as U_t^* does, we emphasize in what follows the case of modifications based on the expected information. As a final comment on the form of $U_t^{(S)}$, note that by taking the expectation of the second summand of the right hand side above, expression (4.3) is recovered.

4.2.2 IWLS procedure for obtaining the bias-reduced estimates

As in Firth (1992a,b), if we treat h_r $(r = 1,\ldots,n)$ in (4.3) as if they were known constants, then the BR estimator is formally equivalent to the use of ML after making the following adjustments to the response frequencies and totals:

$$\begin{array}{ll} \text{Pseudo-successes} & y_r^* = y_r + \frac{1}{2}h_r \\ \text{Pseudo-totals} & m_r^* = m_r + h_r \end{array} \qquad (r = 1,\ldots,n)\,. \tag{4.4}$$

The above pseudo-data representation can be viewed as a generalization of the flattening modifications used in Clogg et al. (1991) for the re-calibration of the industry and occupation codes on 1970 census public-use samples to the 1980 standard. In our context, the Clogg et al. (1991) proposal is to use

$$\begin{array}{ll} \text{Pseudo-successes} & y_r^* = y_r + a\frac{p}{\sum_i m_i} \\ \text{Pseudo-totals} & m_r^* = m_r + \frac{p}{\sum_i m_i} \end{array} \qquad (r = 1,\ldots,n)\,, \tag{4.5}$$

where $a \in (0,1)$. Therein a was chosen as the observed proportion of successes, namely $a = \sum_r y_r / \sum_r m_r$. The stated aim in Clogg et al. (1991) was not bias reduction but

rather an applicable way of eliminating the possibility of infinite ML estimates in the large application they considered. They made this choice of flattening modification based on standard Bayesian arguments which relate to the behaviour of Jeffreys prior among every possible logistic regression model and design. More specifically, this choice yields to the same average prior variance for any design and model when the prior is in the conjugate form (see also the last section in Rubin & Schenker, 1987, for a thorough discussion). However, in the same year, Cordeiro & McCullagh (1991) showed that the vector of first-order biases of the estimators of the parameters β in a logistic regression model can be approximated by $p\beta/\sum_r m_r$. In this way Cordeiro & McCullagh (1991) attribute to the vector of first-order biases approximate collinearity with the parameter vector. In terms of this approximation, Clogg et al. (1991) append to the responses an appropriate fraction a of the first-order relative bias. In the case of (4.4) we append to the responses half a leverage. If we depart from the aggregated case we considered, and consider that the responses are just "1"-"0" Bernoulli trials, then $m_r = 1$ for every $r = 1, \ldots, n$ and $\sum_r m_r = n$. In this way the balanced choice $h_r = p/n$ makes both pseudo-data representations equivalent for $a = 1/2$. In this sense the bias-reducing pseudo-data representation is more general than (4.5). As will be shown shortly, (4.4) is equally easy to apply in practice and by the point of origin of its derivation has the advantage a clearer interpretation in terms of second-order asymptotics.

However, the pseudo-data representation (4.4) has the disadvantage of incorrectly inflating the binomial totals and for this reason could lead to underestimation of the asymptotic standard errors, if care is not taken. On the other hand such a pseudo-data representation ensures positive pseudo-responses y_r^*. In order not to have to re-adjust to the correct binomial totals once the bias-reduced estimates have been obtained, we could define the alternative pseudo-data representation

$$y_r^* = y_r + \frac{1}{2}h_r - h_r\pi_r\,, \tag{4.6}$$

(derived also in Table 3.2) which however does not necessarily respect the non-negativity of the binomial responses. The choice of a pseudo-data representation to be used in practice is merely a matter of whether already implemented software is used for obtaining the BR estimates. So, for example, as in the discussion at the end of Chapter 3, if we intend to use the `glm` procedure in the *R language* (R Development Core Team, 2007), we should use the first representation, since `glm` —correctly— will return an error message if the pseudo-responses become negative. However in this case, after the final iteration we have to re-adjust each working weight w_r by dividing it by $m_r + h_r$, with h_r evaluated at the estimates, and multiplying it by m_r. In this way we recover the correct estimated standard errors (see Chapter 5, for more details).

Both pseudo-data representations in (4.4) and (4.6) can be used to derive a modified IWLS procedure (see Section 3.8 for a general description) for the bias-reduced estimates. If the estimates at the c-th iteration have value $\beta_{(c)}$, then an updated value can be obtained via the modified Fisher scoring iteration

$$\beta_{(c+1)} = \beta_{(c)} + \left(X^T W_{(c)} X\right)^{-1} U^*_{(c)}\,.$$

or, in terms of modified IWLS iteration,

$$\beta_{(c+1)} = (X^T W_{(c)} X)^{-1} X^T W_{(c)} \zeta^*_{(c)} \,. \tag{4.7}$$

From (3.27), ζ^* is the n-vector of the modified working variates with elements

$$\begin{aligned} \zeta^*_r &= \log \frac{\pi_r}{1-\pi_r} + \frac{y_r/m_r - \pi_r}{\pi_r(1-\pi_r)} + \frac{h_r(1/2-\pi_r)}{m_r \pi_r (1-\pi_r)} \\ &= \zeta_r - \xi_r \quad (r = 1, \ldots, n) \,, \end{aligned}$$

where $\xi_r = h_r(\pi_r - 1/2)/\{m_r \pi_r (1-\pi_r)\}$ and ζ_r are the working variates for maximum likelihood IWLS. Again, since both h_r and π_r generally depend on the parameters, the value of the modified frequencies y^*_r is updated with the estimates at each cycle of this scheme. Thus, the replacement of the responses y_r with y^*_r results in the additive adjustment of ζ_r by $-\xi_r$.

As starting values for the above scheme we can use the ML estimates of β after adding $1/2$ to the initial frequencies. By this simple device we eliminate the possibility of infinite ML estimates.

4.2.3 Properties of the bias-reduced estimator

Data separation in logistic regression has been extensively studied in Albert & Anderson (1984) and Lesaffre & Albert (1989) (see also Section B.4 in Appendix B for some of the results therein expressed in our notation). With separated datasets, the ML estimate has at least one infinite-valued component, which usually implies that some fitted probabilities are exactly zero or one and causes fitting procedures to fail to converge. Recently, Heinze & Schemper (2002) illustrated, with two extensive empirical studies, that the bias-reduction method provides a solution to the problem of separation in binary logistic regression, and they conjectured that it guarantees that the BR estimates are finite for general regressions. Further, they note that the BR estimates are typically smaller in absolute value than the corresponding ML estimates. This is natural, since the asymptotic bias in this case increases with the distance of the true parameter values from the origin, with the ML estimator being exactly unbiased when all of the true log-odds are zero (see, for example, Copas, 1988). Also, Zorn (2005) gives an excellent review and examples on separation in binomial-response logistic regression and focuses on the finiteness of the BR estimates in such cases. However, no formal theoretical account on the aforementioned properties of the BR estimator has appeared.

The remainder of this section is devoted to the formal statement and proof of the finiteness and shrinkage properties of the BR estimator in the case of binomial-response logistic regression. Specifically, it is shown that the estimates shrink towards zero, with respect to a metric based on the Fisher information. Further, we comment on the direct but beneficial impact of shrinkage upon the variance and, consequently, on the mean squared error (MSE) of the estimator.

4.2.3.1 A motivating example

Before continuing to the theoretical results, we give an illustration of the finiteness and shrinkage properties of the BR estimator, by considering the case of a $2\times2\times2$ contingency table with one binomial response and two cross-classified factors C_1 and C_2, with two levels each, as explanatory variables. The response counts are sampled independently at each combination of the levels of C_1 and C_2 (covariate settings) from binomial distributions with totals m_1, m_2, m_3, m_4. The model to be fitted is

$$\log\frac{\pi_r}{1-\pi_r} = \alpha + \beta x_{r1} + \gamma x_{r2} \quad (r = 1, \ldots, 4) \ ,$$

where x_{r1} is equal to 1 if $C_1 = \text{II}$ and 0 otherwise and x_{r2} is 1 if $C_2 = \text{B}$ and 0 otherwise; see Table 4.1.

In this simple case, it is possible to identify every data configuration that causes separation simply by looking at the likelihood equations

$$\begin{aligned} T_\alpha &= m_1\pi_1 + m_2\pi_2 + m_3\pi_3 + m_4\pi_4 \,, \\ T_\beta &= m_3\pi_3 + m_4\pi_4 \,, \\ T_\gamma &= m_2\pi_2 + m_4\pi_4 \,, \end{aligned} \tag{4.8}$$

where $T_\alpha = \sum_{r=1}^{4} y_r$, $T_\beta = y_3 + y_4$, $T_\gamma = y_2 + y_4$ are the sufficient statistics for α, β and γ, respectively, and π_r is the probability of success for the r-th covariate setting ($r = 1, 2, 3, 4$). Infinite maximum likelihood estimates correspond to fitted probabilities 0 or 1 and so, by (4.8), they occur if and only if at least one of the following conditions holds:

$$\begin{aligned} T_\alpha = 0 \quad &\text{or} \quad m_1 + m_2 + m_3 + m_4 \,, \\ T_\beta = 0 \quad &\text{or} \quad m_3 + m_4 \,, \\ T_\gamma = 0 \quad &\text{or} \quad m_2 + m_4 \,, \\ T_\alpha - T_\beta = 0 \quad &\text{or} \quad m_1 + m_2 \,, \\ T_\alpha - T_\gamma = 0 \quad &\text{or} \quad m_1 + m_3 \,, \\ T_\beta - T_\gamma = m_3 \quad &\text{or} \quad -m_2 \,, \\ T_\alpha - T_\beta - T_\gamma = m_1 \quad &\text{or} \quad -m_4 \,. \end{aligned}$$

Using the above conditions, Table 4.2 is constructed, in which all the possible separated data configurations are shown. They are characterized as completely or quasi-completely separated according to definitions in Albert & Anderson (1984) (Definition B.4.1 and Definition B.4.2 in Appendix B, here). Also, by the theorems therein (see Theorem B.4.3 in Appendix B) any data configuration that is not recorded in the table is an "overlapping" configuration and results in unique and finite ML estimates. A data set that has the tenth configuration of the first row of the quasi-separated part of Table 4.2 was used in Clogg et al. (1991, Table 6) to illustrate the problematic behaviour of ML for these cases. The sampling scheme in their example is retrospective (column totals fixed, row totals random)

Table 4.1: A two-way layout with a binomial response and totals m_1, m_2, m_3, m_4 for each combination of the categories of the cross-classified factors C_1 and C_2

Covariate Setting	Covariates		Response		Totals
	C_1	C_2	Success	Failure	
1	I	A	y_1	$m_1 - y_1$	m_1
2		B	y_2	$m_2 - y_2$	m_2
3	II	A	y_3	$m_3 - y_3$	m_3
4		B	y_4	$m_4 - y_4$	m_4

while we use a prospective sampling scheme (row totals fixed, column totals random). However, as discussed in McCullagh & Nelder (1989, § 4.3.3) the logistic model applies with the same β and γ but with different constant α, so that Table 4.2 covers separated configurations under either sampling scheme.

Here, we consider the severe case with $m_1 = m_2 = m_3 = m_4 = 2$, where 50 (62.5%) of the 81 possible data configurations are separated. For these cases, the vector of ML estimates involves infinite components; as in all logistic regressions the bias and the variance of the ML estimators are infinite. In Table C.1 in Appendix C we present the ML estimates, the bias-corrected (BC) estimates and the BR estimates for every possible data set in this setting. The BC estimates (Cordeiro & McCullagh, 1991) are the ML estimates after subtracting from them the first-order bias terms that are given by (3.3), and so their value is undefined when the ML estimates are infinite. In contrast, the BR estimator is finite in all 81 cases. The shrinkage effect from bias reduction is directly noted by comparing the finite ML estimates and the BR estimates in Table C.1. Further, note that the shrinkage of the BC estimates is stronger. This agrees with the empirical results in Heinze & Schemper (2002) and Bull et al. (2002), where it is illustrated that the BC estimates correct the bias of the ML estimator beyond the true value and that such overcorrection is dangerous because is accompanied by small estimated variance.

In Table 4.3 we calculate the expected value, bias and variance of the BR estimator for several vectors of true parameter values. In the last case of the table the true parameter vector has the extreme for this setting value $(2, 0.4, 2.1)$, implying probabilities $\pi_1 = 0.88$, $\pi_2 = 0.98$, $\pi_3 = 0.92$ and $\pi_4 = 0.99$ at the four covariate settings. Despite the high probability of separation (0.99), the BR estimates are finite and hence we can explicitly calculate expectations and variances. However, we have to be cautious with penalized-likelihood based inferences on data generated for this setting with $m = 2$. Despite the fact that the BR estimates exist in contrast to the ML estimates, they have considerable bias and very small variance so that concerns about the coverage properties of classical

Table 4.2: All possible separated data configurations for a two-way layout and a binomial response (see Table 4.1). The notions of quasi-complete and complete separation are defined in Definition B.4.1 and Definition B.4.2 in Appendix B

Separation Type	Data configuration (x: positive count, 0: zero count)
Complete	$\begin{bmatrix}0 & x\\0 & x\\0 & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}0 & x\\x & 0\\0 & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}0 & x\\0 & x\\0 & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}0 & x\\x & 0\\0 & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}0 & x\\0 & x\\x & 0\\x & 0\end{bmatrix}$ $\begin{bmatrix}0 & x\\x & 0\\x & 0\\0 & x\end{bmatrix}$ $\begin{bmatrix}0 & x\\0 & x\\x & 0\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & 0\\0 & x\\0 & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & 0\\0 & x\\0 & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & 0\\0 & x\\x & 0\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & 0\\0 & x\\x & 0\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & 0\\x & 0\\0 & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & 0\\x & 0\\0 & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & 0\\x & 0\\x & 0\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & 0\\x & 0\\x & 0\\x & 0\end{bmatrix}$
Quasi-Complete	$\begin{bmatrix}0 & x\\x & x\\0 & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}0 & x\\0 & x\\x & 0\\x & x\end{bmatrix}$ $\begin{bmatrix}0 & x\\0 & x\\x & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}0 & x\\0 & x\\x & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}0 & x\\0 & x\\0 & x\\x & x\end{bmatrix}$ $\begin{bmatrix}0 & x\\0 & x\\0 & x\\x & x\end{bmatrix}$ $\begin{bmatrix}0 & x\\x & x\\x & 0\\x & 0\end{bmatrix}$ $\begin{bmatrix}0 & x\\x & x\\x & 0\\x & 0\end{bmatrix}$ $\begin{bmatrix}0 & x\\x & x\\0 & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}0 & x\\x & x\\0 & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & x\\0 & x\\0 & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}0 & x\\x & 0\\0 & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}0 & x\\x & 0\\x & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & x\\0 & x\\0 & x\\0 & x\end{bmatrix}$
	$\begin{bmatrix}x & x\\0 & x\\x & 0\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & x\\0 & x\\x & 0\\x & x\end{bmatrix}$ $\begin{bmatrix}x & x\\0 & x\\x & 0\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & x\\0 & x\\x & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & x\\0 & x\\x & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & x\\0 & x\\x & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & x\\x & x\\0 & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & x\\x & x\\0 & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & x\\x & x\\x & 0\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & x\\x & x\\x & 0\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & 0\\0 & x\\0 & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & 0\\0 & x\\0 & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & 0\\0 & x\\x & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & 0\\0 & x\\x & x\\x & 0\end{bmatrix}$
	$\begin{bmatrix}x & 0\\x & x\\x & 0\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & 0\\x & x\\x & 0\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & 0\\x & x\\0 & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & 0\\x & 0\\0 & x\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & 0\\x & 0\\x & 0\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & 0\\x & 0\\x & 0\\x & 0\end{bmatrix}$ $\begin{bmatrix}x & 0\\x & 0\\x & x\\0 & x\end{bmatrix}$ $\begin{bmatrix}x & 0\\x & 0\\x & x\\x & 0\end{bmatrix}$

Table 4.3: Expectations, biases and variances for the bias-reduced estimator $(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})$ to three decimal places for several different settings of the true parameter vector $(\alpha_0, \beta_0, \gamma_0)$.

$(\alpha_0, \beta_0, \gamma_0)$			Expected values			Biases			Variances			Probability of separation
α_0	β_0	γ_0	$\mathrm{E}(\tilde{\alpha})$	$\mathrm{E}(\tilde{\beta})$	$\mathrm{E}(\tilde{\gamma})$	$\mathrm{E}(\tilde{\alpha}-\alpha_0)$	$\mathrm{E}(\tilde{\beta}-\beta_0)$	$\mathrm{E}(\tilde{\gamma}-\gamma_0)$	$\mathrm{Var}(\tilde{\alpha})$	$\mathrm{Var}(\tilde{\beta})$	$\mathrm{Var}(\tilde{\gamma})$	
−0.5	−0.5	−0.5	−0.472	−0.423	−0.423	0.028	0.077	0.077	1.389	1.9	1.9	0.667
−0.5	−0.5	0	−0.474	−0.452	0	0.026	0.048	0	1.432	1.949	1.988	0.58
−0.5	−0.5	0.5	−0.475	−0.471	0.471	0.025	0.029	−0.029	1.459	1.982	1.982	0.526
−0.5	0	−0.5	−0.474	0	−0.452	0.026	0	0.048	1.432	1.988	1.949	0.58
−0.5	0	0	−0.483	0	0	0.017	0	0	1.48	2.009	2.009	0.497
−0.5	0	0.5	−0.486	0	0.486	0.014	0	−0.014	1.499	2.018	1.988	0.464
−0.5	0.5	0	−0.486	0.486	0	0.014	−0.014	0	1.499	1.988	2.018	0.464
0	−0.5	−0.5	−0.004	−0.471	−0.471	−0.004	0.029	0.029	1.47	1.982	1.982	0.526
0	−0.5	0.5	0	−0.486	0.486	0	0.014	−0.014	1.498	1.988	1.988	0.464
0	0	−0.5	0	0	−0.486	0	0	0.014	1.498	2.018	1.988	0.464
0	0	0	0	0	0	0	0	0	1.514	2.018	2.018	0.43
0	0	0.5	0	0	0.486	0	0	−0.014	1.498	2.018	1.988	0.464
0	0.5	−0.5	0	0.486	−0.486	0	−0.014	0.014	1.498	1.988	1.988	0.464
0	0.5	0	0	0.486	0	0	−0.014	0	1.498	1.988	2.018	0.464
0	0.5	0.5	0.004	0.471	0.471	0.004	−0.029	−0.029	1.47	1.982	1.982	0.526
0.5	−0.5	−0.5	0.486	−0.486	−0.486	−0.014	0.014	0.014	1.499	1.988	1.988	0.464
0.5	−0.5	0	0.486	−0.486	0	−0.014	0.014	0	1.499	1.988	2.018	0.464
0.5	−0.5	0.5	0.475	−0.471	0.471	−0.025	0.029	−0.029	1.459	1.982	1.982	0.526
0.5	0	−0.5	0.486	0	−0.486	−0.014	0	0.014	1.499	2.018	1.988	0.464
0.5	0	0	0.483	0	0	−0.017	0	0	1.48	2.009	2.009	0.497
0.5	0	0.5	0.474	0	0.452	−0.026	0	−0.048	1.432	1.988	1.949	0.58
0.5	0.5	−0.5	0.475	0.471	−0.471	−0.025	−0.029	0.029	1.459	1.982	1.982	0.526
0.5	0.5	0.5	0.472	0.422	0.422	−0.028	−0.078	−0.078	1.389	1.9	1.9	0.667
1.5	−1.5	−1.5	1.309	−1.309	−1.309	−0.191	0.191	0.191	1.324	1.723	1.723	0.662
2	0.4	2.1	1.4	0.112	0.454	−0.6	−0.288	−1.646	0.62	0.764	0.681	0.99

confidence intervals may arise.

Also, it should be noted that for general designs, the BR estimates are not strictly smaller than their ML counterparts, which suggests that they do not shrink towards the origin according to the Euclidean distance in the parameter space. Obvious candidates of distances that can be used to verify shrinkage are the ones that depend directly to the form of the penalized likelihood, that is distances depending on the Jeffreys invariant prior (see Subsection 4.2.3.3 below).

4.2.3.2 Finiteness

Consider estimation of β for model (4.1) by maximization of a penalized log-likelihood of the form

$$l^{(a)}(\beta) = l(\beta) + a \log \det F(\beta)\,, \tag{4.9}$$

where a is a fixed positive constant. The case $a = 1/2$ corresponds to penalization of the likelihood by the Jeffreys invariant prior.

Theorem 4.2.1: *If any component of $\eta = X\beta$ is infinite-valued, the penalized log-likelihood $l^{(a)}(\beta)$ has value $-\infty$.*

Proof. It is sufficient to prove the argument when exactly one component of η is infinite-valued. Without loss of generality, suppose that η_1 is infinite-valued. For the corresponding binomial probability $\pi_1 = \exp(\eta_1)/\{1 + \exp(\eta_1)\}$ we either have $\pi_1 = 1$ ($\eta_1 = +\infty$) or $\pi_1 = 0$ ($\eta_1 = -\infty$). We can re-parameterize the model by defining new parameters $\gamma = \gamma(\beta) = Q\beta$, where Q is a $p \times p$ non-singular matrix. The new design matrix is defined as $G = XQ^{-1}$. By the spectral decomposition theorem and the symmetry of the Fisher information, we can find a matrix Q — which possibly depends on β — that has the following two properties:

i) $G^T W G = \mathrm{diag}\{i_1, \ldots, i_p\}$, with $W = \mathrm{diag}\{w_r; r = 1, \ldots, n\}$, $w_r = m_r \pi_r (1 - \pi_r)$. That is the information on γ is a diagonal matrix with diagonal elements the eigenvalues of the information on β.

ii) $g_{1t} = \begin{cases} 1, & \text{for } t = 1 \\ 0, & \text{otherwise} \end{cases} \qquad (t = 1, \ldots, p)$.

Hence, the new parameterization is constructed in order to have $\gamma_1 = \eta_1$ as its first parameter and in such a way that all of its parameters are mutually orthogonal. Because η_1 is infinite-valued, η_r is necessarily infinite-valued for all r such that g_{r1} is non-zero ($\eta_r = \sum_{t=1}^{p} \gamma_t g_{rt}$). Collect all these r in a set $C \subset \{1, \ldots, n\}$. Then, for $r \in C$ the binomial variances $w_r = m_r \pi_r (1 - \pi_r)$ are zero. Hence

$$i_1 = \sum_{r=1}^{n} g_{r1}^2 w_r = \sum_{r \in C} g_{r1}^2 w_r = 0\,.$$

By the orthogonality of Q, $(\det Q)^2 = 1$ and so

$$\det F = \det\{X^T W X\} = \det\{G^T W G\} = \prod_{t=1}^{p} i_t = 0$$

and the result follows by noting that the binomial log-likelihood $l(\beta)$ is bounded above by zero. Note that the requirement of mutual orthogonality of the parameters, can be relaxed to orthogonality of γ_1 to $\gamma_2, \ldots, \gamma_p$. In this case the Fisher information on β is again singular because if we keep the second requirement for G valid, the first row and column of the Fisher information are zero. □

The above theorem enables us to state the following corollary, the main result towards the finiteness of the BR estimates.

Corollary 4.2.1: Finiteness of the bias-reduced estimates. *The vector $\hat{\beta}^{(a)}$ that maximizes $l^{(a)}(\beta)$ has all of its elements finite, for positive a.*

Proof. If any component of β is infinite, at least one component of the corresponding $\eta(\beta)$ is infinite-valued and so, by Theorem 4.2.1, $l^{(a)}(\beta)$ has value $-\infty$. Hence, there exists $\hat{\beta}^{(a)}$, with finite components, such that

$$\hat{\beta}^{(a)} = \arg\max_{\beta} l^{(a)}(\beta) \tag{4.10}$$

because $l^{(a)}(\beta)$ can always take finite values — for example, by the choice $\beta = 0$, which corresponds to binomial probabilities $\pi_r = 1/2$ for every $r = 1, \ldots, n$ and is the point where the determinant of the Fisher information attains its global maximum (see Theorem 4.2.2 below). □

For $a = 1/2$, the above corollary refers to the finiteness of the BR estimates for binary logistic regression models.

4.2.3.3 Shrinkage towards the origin

The shrinkage of the BR estimates towards the origin is a direct consequence of the penalization of the likelihood function by Jeffreys invariant prior. This is shown through the two following theorems that describe the functional behaviour of $\log\det F(\beta)$.

Theorem 4.2.2: *Let β be the p-dimensional parameter vector of a binary logistic regression model. If $F(\beta)$ is the Fisher information on β, the function $\det F(\beta)$ is globally maximized at $\beta = 0$.*

Proof. The design matrix X is by assumption of full rank p and so we can always orthogonalize it. In practice this can be achieved by applying the Gram–Schmidt procedure to its columns and thus expressing it in the form $X = QR$ with Q a $n \times p$ matrix with orthonormal columns ($Q^TQ = I_p$) and R a $p \times p$ non-singular matrix. In this way we can write

$$\det\left\{X^TW(\beta)X\right\} = \frac{\det\left\{Q^TW(\beta)Q\right\}}{(\det R)^2}\,.$$

Note that R does not depend on β and thus, since $(\det R)^2$ is positive, $\det\{X^TW(\beta)X\}$ and $\det\{Q^TW(\beta)Q\}$ have stationary points of the same kind for the same values of β. Further, the eigenvalues of W are its diagonal elements $w_r = m_r\pi_r(1-\pi_r)$. Denote the ordered set of w_r's as $\{w_{(r)}; r = 1, 2, \ldots, n\}$ with $w_{(1)} \leq w_{(2)} \leq \ldots \leq w_{(n)}$.

Lemma B.3.1 in Appendix B shows that

$$\prod_{t=1}^{p} \lambda_t \leq \det\left\{G^T A G\right\} \leq \prod_{t=1}^{p} \lambda_{n-p+t}\,,$$

for every positive definite $n \times n$ matrix A with eigenvalues $\lambda_1 \leq \ldots \leq \lambda_n$ and any $n \times p$ matrix G satisfying $G^T G = 1_p$, with 1_p the $p \times p$ identity matrix. Thus

$$\prod_{t=1}^{p} w_{(t)}(\beta) \leq \det\left\{Q^T W(\beta) Q\right\} \leq \prod_{t=1}^{p} w_{(n-p+t)}(\beta)\,. \tag{4.11}$$

Note that, for every $r = 1, \ldots, n$, $0 < w_r(\beta) \leq 1/4$, with the upper bound achieved when $\pi_r = 1/2$. Hence

$$\prod_{t=1}^{p} w_{(t)}(\beta) \leq \frac{1}{4^p} \qquad \text{and} \qquad \prod_{t=1}^{p} w_{(n-p+t)}(\beta) \leq \frac{1}{4^p}\,.$$

By the form of the logistic regression model (4.1), the probability π_r for a subject r is $1/2$ only if all the components of β, associated with π_r, are zero. Thus, at $\beta = 0$, inequalities (4.11) become

$$\frac{1}{4^p} \leq \det\left\{Q^T W(0) Q\right\} \leq \frac{1}{4^p}$$

and hence $\det\{Q^T W(0) Q\} = 1/4^p$, which is the maximum value that $\det\{Q^T W(\beta) Q\}$ can take. Thus, $\det\{X^T W(\beta) X\}$ is globally maximized at $\beta = 0$. □

Theorem 4.2.3: *Let β be the p-dimensional parameter vector of a binary logistic regression model. Further, let π be the n-dimensional vector of binomial probabilities. If $F(\beta)$ is the Fisher information on β, let $\bar{F}(\pi(\beta)) = F(\beta)$. Then, the function $f(\pi) = \det \bar{F}(\pi)$ is log-concave.*

Proof. Let $\bar{W}(\pi(\beta)) = W(\beta)$ and denote $\bar{w}_r(\pi_r)$ the diagonal elements of $\bar{W}(\pi)$. Then $\bar{F}(\pi) = X^T \bar{W}(\pi) X$. For $\theta \in (0,1)$, $\tilde{\theta} = 1 - \theta$ and any pair of n-vectors of probabilities π and ϕ,

$$\bar{w}_r(\theta\pi_r + \tilde{\theta}\phi_r) \geq \theta \bar{w}_r(\pi_r) + \tilde{\theta}\bar{w}_r(\phi_r) \quad (r = 1, \ldots, n)$$

because $\bar{w}_r(\pi_r) = m_r \pi_r (1 - \pi_r)$ and thus concave. Hence, by Lemma B.3.3 in Appendix B,

$$\det\left\{X^T \bar{W}(\theta\pi + \tilde{\theta}\phi) X\right\} \geq \det\left\{\theta X^T \bar{W}(\pi) X + \tilde{\theta} X^T \bar{W}(\phi) X\right\}$$

and so, by the monotonicity of the logarithm function and using Lemma B.3.4,

$$\log\det\left\{X^T \bar{W}(\theta\pi + \tilde{\theta}\phi) X\right\} \geq \theta \log\det\left\{X^T \bar{W}(\pi) X\right\} + \tilde{\theta} \log\det\left\{X^T \bar{W}(\phi) X\right\}\,,$$

which completes the proof. □

Once again, consider estimation by maximization of a penalized log-likelihood as in (4.9) but now for non-negative a, and let $a_1 > a_2 \geq 0$. Further, let $\hat{\beta}^{(a_1)}$ and $\hat{\beta}^{(a_2)}$ be the maximizers of $l^{(a_1)}$ and $l^{(a_2)}$, respectively and $\pi^{(a_1)}$ and $\pi^{(a_2)}$ the corresponding fitted n-vectors of probabilities. Then, by the concavity of $\log\det F'(\pi)$, the vector $\pi^{(a_1)}$ is closer to $(1/2,\ldots,1/2)^T$ than is $\pi^{(a_2)}$, in the sense that $\pi^{(a_1)}$ lies within the hull of that convex contour of $\log\det F'(\pi)$ containing $\pi^{(a_2)}$. With the specific values $a_1 = 1/2$ and $a_2 = 0$ the last result refers to penalization of the likelihood by Jeffreys invariant prior, and to un-penalized likelihood, respectively.

Specifically, the above conclusion can be written as follows. Since $a_1 > a_2 \geq 0$, by Lemma B.3.5 with $f(x)$ replaced by $l^{(a_2)}(\beta)$ and $g(x)$ replaced by $(a_1 - a_2)\log\det F(\beta)$, we obtain that

$$\log\det F'(\pi^{(a_1)}) \geq \log\det F'(\pi^{(a_2)})\,. \tag{4.12}$$

Now let $\beta, \gamma \in \mathrm{B}$. If, for example, we define

$$d\,(\beta,\gamma) = \left[\log\det F\,(\beta) - \log\det F\,(\gamma)\right]^2 = \left(\log\det\left\{F(\beta)F^{-1}(\gamma)\right\}\right)^2\,,$$

then, since $\log\det F(\beta) = \log\det F'(\pi(\beta))$,

$$d\,(\beta,\gamma) = \left[\log\det F'\,(\pi(\beta)) - \log\det F'\,(\pi(\gamma))\right]^2\,. \tag{4.13}$$

Hence, by (4.12) and the concavity of $\log\det F'(\pi)$, we have that $d(\hat{\beta}^{(a_1)},0) \leq d(\hat{\beta}^{(a_2)},0)$. Thus the BR estimates shrink towards the origin, relative to the ML estimates, with respect to this metric based on the Fisher information.

It is important to mention here that since the BR estimates are typically smaller in absolute value than the ML estimates, the asymptotic variance of the BR estimator is, correspondingly, typically smaller than that of the ML estimator, whenever the latter exists. The same is true for the estimated first-order variances. Further, since the BR estimator has bias of order $\mathcal{O}(n^{-2})$ and smaller asymptotic variance than the ML estimator, it also has smaller asymptotic MSE. These remarks summarize the importance of the shrinkage effect in this setting. The following example illustrates the shrinkage in the variance and hence in the MSE of the estimator in the simple case of the estimation of the log-odds of success for a single binomial trial.

Example 4.2.1: *Estimation of the log-odds of success for a single binomial trial.* Consider a modified score function of the form $U^*(\beta) = U(\beta) + A(\beta)$, where β is a scalar parameter, U_β is the ordinary score function and A_r is a $\mathcal{O}(1)$ modification. Further denote the true but unknown parameter value as β_0 and let $A \equiv A(\beta_0)$. For a flat exponential family indexed by the parameter β, and m units of information, the MSE of the resultant estimator $\tilde{\beta}$ can be expressed as (see (6.7) in Chapter 6)

$$\mathrm{E}\left((\tilde{\beta}-\beta_0)^2\right) = \frac{1}{\mu_{1,1}} \ddot{+} \frac{\mu_4 + 3\mu_3 A + \mu_{1,1}(2\dot{A} + A^2)}{\mu_{1,1}^3} + \frac{11\mu_3^2}{4\mu_{1,1}^4} \ddot{+} \mathcal{O}(m^{-3})\,, \tag{4.14}$$

where $\dot{A}$ is the first derivative of A with respect to β evaluated at β_0, $\ddot{+}$ denotes a drop of the asymptotic order by $\mathcal{O}(m^{-1})$ and

$$\mu_r \equiv \mu_r(\beta_0) = \mathrm{E}\left(\frac{\partial^r l}{\partial\beta^r};\ \beta_0\right);\ \mu_{r,s} \equiv \mu_{r,s}(\beta_0) = \mathrm{E}\left(\frac{\partial^r l}{\partial\beta^r}\frac{\partial^s l}{\partial\beta^s};\ \beta_0\right) \quad (r,s = 1,2,\ldots)\,.$$

Also, the corresponding expression for the variance of $\tilde{\beta}^{(a)}$ (see (6.10) in Chapter 6) is

$$\mathrm{Var}\left(\tilde{\beta}\right) = \frac{1}{\mu_{1,1}} \ddot{+} \frac{\mu_4 + 2\mu_3 A + 2\mu_{1,1}\dot{A}}{\mu_{1,1}^3} + \frac{10\mu_3^2}{4\mu_{1,1}^4} \ddot{+} \mathcal{O}(m^{-3})\,. \tag{4.15}$$

As an illustrative example of the effect of the penalized likelihood to the MSE and the variance of the resultant estimators in binary logistic regression, we consider the simplest case of a single realization y of a binomial random variable Y with index m and probability of success π. We are interested on the estimation of the log-odds $\beta = \log(\pi/(1-\pi))$ by a penalized likelihood of the form (4.9) with $a \geq 0$. For $a = 1/2$, the penalized likelihood refers to the bias-reduction method and for $a = 0$ to ML. In this context, $A(\beta) = a(1 - 2\pi(\beta))$, $\dot{A}(\beta) = -2a\pi(\beta)(1-\pi(\beta))$ and $U(\beta) = y - m\pi(\beta)$. Thus, the resultant modified score equation is $y + a - (m+a)\pi(\beta) = 0$ and so the resultant estimator has the familiar form $\tilde{\beta}^{(a)} = \log((Y+a)/(m-Y+a))$ (cf., Cox & Snell, 1989, §2.1.6). Also, in this setting

$$\mu_{1,1} = m\pi(1-\pi)\,; \quad \mu_3 = -m\pi(1-\pi)(1-2\pi)\,; \quad \mu_4 = -m\pi(1-\pi)(1-6\pi(1-\pi))\,,$$

where $\pi \equiv \pi(\beta_0)$. A simple substitution to (4.14) and to (4.15) gives

$$\mathrm{E}\left((\tilde{\beta}^{(a)} - \beta_0)^2\right) = \frac{1}{m\pi(1-\pi)} \ddot{+} \frac{7 - 4a(3-a) - (20 - 16a(2-a))\pi(1-\pi)}{4m^2\pi^2(1-\pi)^2} \ddot{+} \mathcal{O}(m^{-3})\,, \tag{4.16}$$

for the MSE and

$$\mathrm{Var}\left(\tilde{\beta}^{(a)}\right) = \frac{1}{m\pi(1-\pi)} \ddot{+} \frac{3 - 4a + 8(a-1)\pi(1-\pi)}{2m^2\pi^2(1-\pi)^2} \ddot{+} \mathcal{O}(m^{-3})\,, \tag{4.17}$$

for the variance. Similar expressions are derived in Gart et al. (1985), who study the performance of several estimators for the log-odds of success. From (4.16) and (4.17) up to the $\mathcal{O}(m^{-1})$ term, the candidate estimators have the same asymptotic MSE and variance. Their difference is in the $\mathcal{O}(m^{-1})$ order term, which depends on a. For $m = 10$, Figure 4.1 shows how the first-order bias, second-order variance and second-order MSE terms behave for various values of $\alpha \in [0,1]$ as the true probability of success (and consequently β_0) varies. Also, Figure 4.2 includes the corresponding graphs for the actual bias, variance and MSE. Note that the curves for $a = 0$ have been excluded from the latter figure since $\tilde{\beta}^{(0)}$ has infinite bias, variance and MSE.

The choice $a = 0.5$ is asymptotically optimal in terms of bias since $\tilde{\beta}^{(1/2)}$ has zero first-order bias term and consequently its bias vanishes with $\mathcal{O}(m^{-2})$ rate, for every value of the true probability, as m increases. However, in the present context, discussions on the optimal choice of a based on MSE-related criteria should be avoided since such choice depends on the true parameter value. For example, note the behaviour of $(a \in (0.7, 0.8))$-estimators, where the second-order MSE term (Figure 4.1) increases rapidly for extreme true probabilities. Also, while for moderate probabilities the actual MSE of the $(a > 1/2)$-estimators is smaller than that of $(a < 1/2)$-estimators (see Figure 4.2), it increases rapidly as the true probability gets close to zero or close to one (or, equivalently when β_0 takes

Figure 4.1: First order bias term, second order MSE term and second order variance term of $\tilde{\beta}^{(a)}$*, for a grid of values of* $a \in [0, 1]$ *against the true probability of success. The dotted curves represent values of a between the reported ones and with step* 0.02*.*

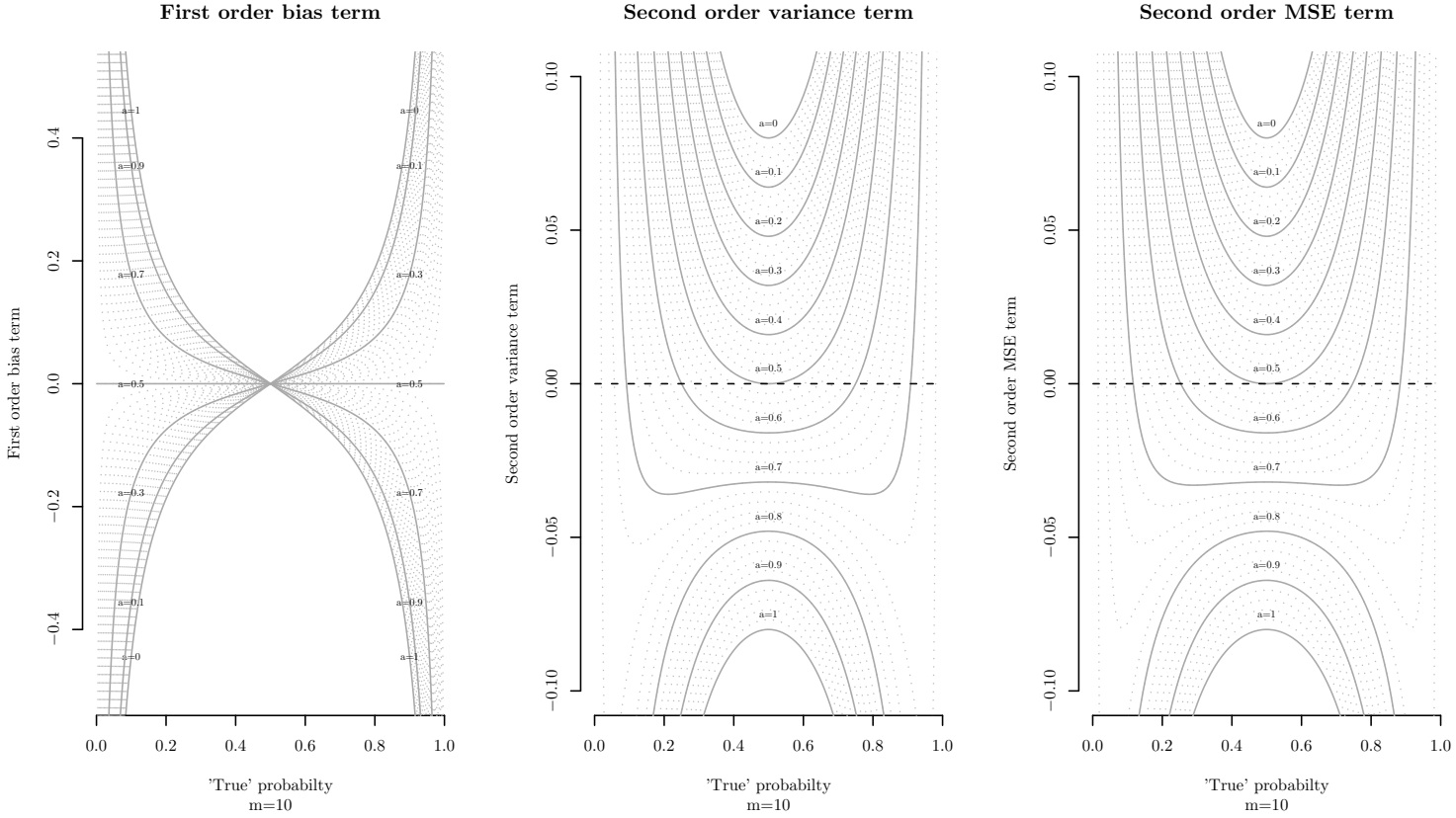
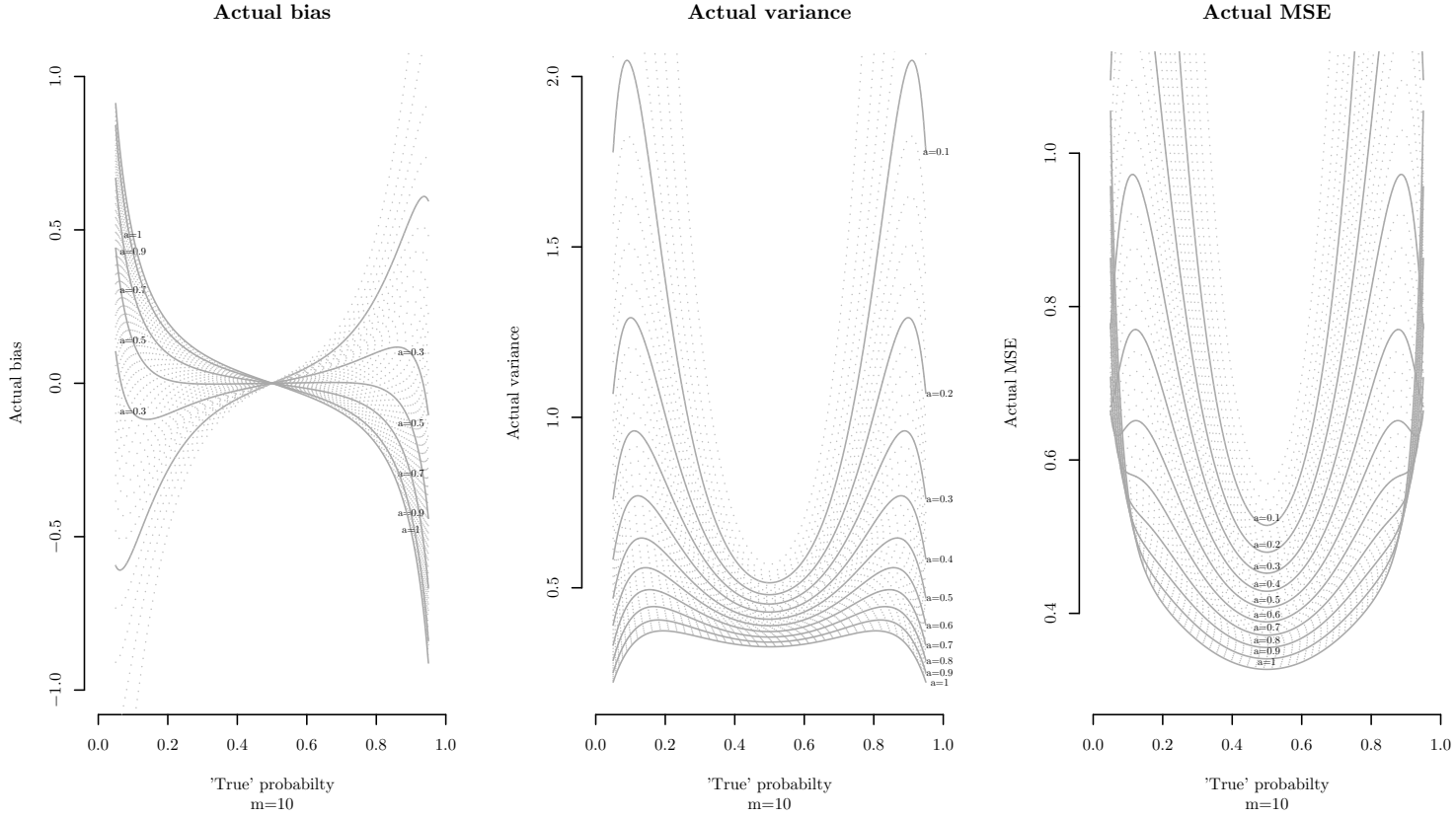


Figure 4.2: Actual bias, actual MSE and actual variance of $\tilde{\beta}^{(a)}$, for a grid of values of $a \in (0, 1]$ against the true probability of success. The dotted curves represent values of a between the reported ones and with step 0.02.



extreme positive or negative values). This is caused by the large bias that such estimators have for very small or very large probabilities. Hence, generic claims like "Values of $c > 1/2$ correspond to priors stronger than Jeffreys, further reducing MSE at the cost of introducing negative bias on the log-odds-ratio scale" (Bull et al., 2007, § 2.1), where c is a in our setup, should be more carefully examined in terms of generality. One thing to note is that $a = 0.5$ results in estimators that have the least positive second order MSE and variance terms, revealing the beneficial impact of the shrinkage effect in terms of variance and MSE when compared with $(a < 1/2)$-estimators. To conclude, the choice $a = 1/2$ results in estimators with the smallest first-order bias and can be characterized as balanced in terms of variance and MSE.

4.3 Generalization to multinomial responses

4.3.1 Baseline category representation of logistic regression

Consider a multinomial response Y with k categories labelled as $1, 2, \ldots, k$, and corresponding category probabilities $\pi_1, \pi_2 \ldots, \pi_k$. In multinomial logistic regression the log-odds for category s versus category b of the response is represented as follows:

$$\log \frac{\pi_s}{\pi_b} = (\beta_s - \beta_b)^T x \quad (s, b = 1, \ldots, k) \ , \tag{4.18}$$

with x a vector of p covariate values (with its first element set to one if a constant is to be included in the linear predictor) and $\beta_s \in \Re^p$ (see, for example, Cox & Snell, 1989, §5.3, for a thorough description). If, for identifiability reasons, we set $\beta_h = 0$, where h is the label of a reference category, the baseline category representation of the model (Agresti, 2002, §7.1) results:

$$\log \frac{\pi_s}{\pi_h} = \eta_s = \beta_s^T x \quad (s = 1, 2, \ldots, h-1, h+1, \ldots, k) \ . \tag{4.19}$$

Likelihood-based inferences using this model are invariant to the choice of the reference category because the only thing affected is the parameterization used. Thus, without loss of generality in what follows, we set as reference the k-th category.

4.3.2 Modified scores

Let $q = k - 1$ and $\gamma^T = (\beta_1^T, \ldots, \beta_q^T)$ be the vector of the pq model parameters. Assume that we have observed n pairs (y_r, x_r) with $y_r = (y_{r1}, \ldots, y_{rq})^T$ the vector of observed frequencies which are realizations of a multinomially distributed random vector Y_r with index m_r, and x_r a $p \times 1$ vector of known covariate values. The observed frequency for the k-th category is $y_{rk} = m_r - \sum_{s=1}^{q} y_{rs}$. Also, denote by $\pi_r = (\pi_{r1}, \ldots, \pi_{rq})^T$ the vector of the corresponding category probabilities. By definition, the probability of the k-th category is $\pi_{rk} = 1 - \sum_{s=1}^{q} \pi_{rs}$. The multinomial log-likelihood can be written as

$$l(\gamma; X) = \sum_r \sum_{s=1}^{q} y_{rs} \log \frac{\pi_{rs}}{\pi_{rk}} + \sum_r m_r \log \pi_{rk} \, ,$$

where the log-odds $\log(\pi_{rs}/\pi_{rk})$ is modelled according to (4.19). In what follows, the matrix X with rows x_r^T is assumed to be of full rank and if an intercept parameter is present in the model the first element of x_r is set to one for every $r = 1, 2, \ldots, n$. Writing $Z_r = 1_q \otimes x_r^T$ for the $q \times pq$ model matrix, we can express (4.19) as

$$\log \frac{\pi_{rs}}{\pi_{rk}} = \eta_{rs} = \sum_{t=1}^{pq} \gamma_t z_{rst} \quad (r = 1, \ldots, n \ ; \ s = 1, \ldots, q) \,, \tag{4.20}$$

where z_{rst} is the (s,t)-th element of Z_r and 1_q is the $q \times q$ identity matrix.

Model (4.20) is a multivariate GLM with canonical link and hence by (2.12), the score vector is

$$U(\gamma) = \sum_r U_r(\gamma) = \sum_r Z_r^T (y_r - m_r \pi_r) \,. \tag{4.21}$$

Also, for the current case, the Fisher information for γ takes the form

$$F(\gamma) = Z^T W Z = \sum_r Z_r W_r Z_r^T \,,$$

with W being a $nq \times nq$ block-diagonal matrix with non-zero blocks the $q \times q$ incomplete covariance matrices $W_r = \mathrm{Var}(Y_r) = m_r \,\mathrm{diag}\,(\pi_r) - m_r \pi_r \pi_r^T$ and $Z^T = (Z_1^T, ..., Z_n^T)$.

Furthermore, by (3.19), the modified scores based on the expected information are

$$\begin{aligned} U_t^*(\gamma) &= U_t(\gamma) + \frac{1}{2} \sum_r \sum_{s=1}^{q} \mathrm{trace}\left\{ H_r W_r^{-1} K_{rs} \right\} z_{rst} \qquad (4.22) \\ &= \sum_r \sum_{s=1}^{q} \left(y_{rs} - m_r \pi_{rs} + \frac{1}{2} \,\mathrm{trace}\left\{ H_r W_r^{-1} K_{rs} \right\} \right) z_{rst} \quad (t = 1, \ldots, pq) \,, \end{aligned}$$

where K_{rs} is a $q \times q$ symmetric matrix with (u,v)-th element the third order cumulants of Y_r, these being

$$\kappa_{rsuv} = \mathrm{Cum}_3(Y_{rs}, Y_{ru}, Y_{rv}) = \begin{cases} m_r \pi_{rs}(1 - \pi_{rs})(1 - 2\pi_{rs}) & s = t = u \\ -m_r \pi_{rs} \pi_{ru}(1 - \pi_{rs}) & s = t \neq u \\ 2 m_r \pi_{rs} \pi_{rt} \pi_{ru} & s, t, u \quad \text{distinct} \,, \end{cases}$$

with $r = 1, \ldots, n$ and $s, u, v = 1, \ldots, q$ (see, for example, McCullagh & Nelder, 1989, p. 167, for the analytic form of higher order cumulants of the multinomial distribution). Also,

$$W_r^{-1} = \frac{1}{m_r} \left(\frac{1}{\pi_{rk}} L_q + \mathrm{diag} \left\{ \frac{1}{\pi_{rs}}; s = 1, \ldots, q \right\} \right) \quad (r = 1, \ldots, n) \,,$$

where L_q is a $q \times q$ matrix of ones. The matrix H_r denotes the r-th diagonal block of the $nq \times nq$ matrix $H = Z \left(Z^T W Z \right)^{-1} Z^T W$ consisting of n^2 blocks, each of dimension $q \times q$. As already mentioned in Subsection 2.4.3, the matrix H is an asymmetric form of the 'hat matrix' as is defined in the framework of multivariate GLMs (see, for example,

Fahrmeir & Tutz, 2001, §4.2.2, for definition and properties). Despite the fact that we are not going to consider the case of more general bias-reducing modifications for the reasons mentioned in Section 3.8, note that, in the first equation of (4.22), a simple replacement of z_{rst} with $z^*_{rst} = \sum_{u=1}^{p} e_{tu} z_{rsu}$ can be used to deduce modified scores based on the score vector or possibly more generic modifications by controlling the matrix R. The scalars e_{tu} are as defined in (3.8).

After some algebra (see Section B.5, Appendix B) the modified score functions are found to take the form

$$U^*_t(\gamma) = \sum_r \sum_{s=1}^{q} \left[y_{rs} + \frac{1}{2} h_{rss} - \left(m_r + \frac{1}{2} \operatorname{trace} H_r \right) \pi_{rs} - \frac{1}{2} \sum_{u=1}^{q} \pi_{ru} h_{rus} \right] z_{rst} \,, \qquad (4.23)$$

for $t = 1, \ldots, pq$, where h_{rsu} is the (s, u)-th element of H_r.

When $q = 1$, (4.19) reduces to the binary logistic regression model with y_r and π_r representing the number of successes observed and the probability of success for the r-th subject, respectively. Also, the matrices H_r reduce to the scalars h_r, which are the diagonal elements of the hat matrix in the univariate case. Thus, in the binary case, (4.23) reduces to

$$U^*_t(\gamma) = \sum_r \left(y_r + \frac{1}{2} h_r - (m_r + h_r)\pi_r \right) z_{rt} \quad (t = 1, \ldots, p) \,,$$

confirming the form of the modified scores in (4.3).

4.3.3 The 'Poisson trick' and bias reduction

At this point we note that an alternative version of (4.23) can be obtained by making use of the equivalence between multinomial logit models and Poisson log-linear models (Palmgren, 1981).

The equivalent log-linear model to (4.19) is

$$\begin{aligned} &\log \mu_{rs} = \tilde{\eta}_{rs} = \phi_r + \eta_{rs} \,, \\ &\log \mu_{rk} = \tilde{\eta}_{rk} = \phi_r \quad (r = 1, \ldots, n \,;\; s = 1, \ldots, q) \,, \end{aligned}$$

where $\mu_{rs} = \tau_r \pi_{rs}$ are the expectations of the independent Poisson random variables Y_{rs}, $\tau_r = \sum_{s=1}^{k} \mu_{rs}$, η_{rs} as in (4.20), and ϕ_r nuisance parameters. According to the above model

$$\tau_r = \left(1 + \sum_{s=1}^{q} e^{\eta_{rs}} \right) \exp \phi_r$$

and so,

$$\phi_r = \log(\tau_r) - \log \left(1 + \sum_{s=1}^{q} e^{\eta_{rs}} \right) .$$

Hence applying the transformation $(\gamma, \phi) \to \delta = (\gamma, \tau)$, we obtain the equivalent log-linked non-linear model,

$$\log \mu_{rs} = \tilde{\eta}_{rs} = \log \tau_r + \eta_{rs} - \log \left(1 + \sum_{u=1}^{q} e^{\eta_{ru}} \right) \quad (s = 1, \ldots, q) \; ; \tag{4.24}$$

$$\log \mu_{rk} = \tilde{\eta}_{rk} = \log \tau_r - \log \left(1 + \sum_{u=1}^{q} e^{\eta_{ru}} \right) \; ,$$

where τ_r are nuisance parameters.

Palmgren (1981), using the parameterization on (γ, τ), decomposed the Poisson log-likelihood as the sum of a marginal, Poisson log-likelihood for τ and a conditional log-likelihood given the observed totals, and proved the equivalence of the Fisher information matrix on γ for the two alternative models, when the parameter space is restricted by equating the nuisances τ_r with the multinomial totals. We proceed through the same route, taking advantage of the orthogonality of τ and γ.

By (3.23) the modified scores in the case of a univariate canonically-linked non-linear model and using penalties based on the expected information are

$$\begin{aligned} \tilde{U}_t^* = \tilde{U}_t &+ \frac{1}{2} \sum_r \sum_{s=1}^{k} \tilde{h}_{rss} \frac{\mathrm{Var}(Y_{rs})}{\mathrm{Cum}_3(Y_{rs})} z_{rst}^* \\ &+ \frac{1}{2} \sum_r \sum_{s=1}^{k} \mathrm{Var}(Y_{rs}) \, \mathrm{trace} \left\{ \tilde{F}^{-1} \mathcal{D}^2 \left(\tilde{\eta}_{rs}; \delta \right) \right\} z_{rst}^* \quad (t = 1, \ldots, n + pq) \; , \end{aligned} \tag{4.25}$$

where $\mathcal{D}^2 \left(\tilde{\eta}_{rs}; \delta \right)$ is the Hessian of $\tilde{\eta}_{rs}$ with respect to δ, and $\tilde{U}$ and $\tilde{F}$ are the scores and the Fisher information on δ, respectively. Furthermore, $\tilde{h}_{rss}$ is the s-th diagonal element of the $k \times k$, r-th diagonal block $\tilde{H}_r$ of the asymmetric hat matrix $\tilde{H} = Z^* \tilde{F}^{-1} Z^{*T} \tilde{W}$ for model (4.24) (see, Section B.6 in Appendix B for the identities connecting the elements of $\tilde{H}_r$ with the elements of H_r) . Here, $Z^{*T} = (Z_1^{*T}, \ldots, Z_n^{*T})$ where $Z_r^*(\delta) = \mathcal{D} \left(\tilde{\eta}_r; \delta \right)$ is the $k \times (n + pq)$ Jacobian of $\tilde{\eta}_{rs}$ with respect to the parameters δ and has (s, t)-th element z_{rst}^*. Also, because of the independence of the Poisson variates, $\tilde{W}$ is a diagonal matrix with diagonal elements $\mathrm{Var}(Y_{rs}) = \mathrm{Cum}_3(Y_{rs}) = \mu_{rs}$ for $s = 1, \ldots, k$ and $r = 1, \ldots, n$. By exploiting the structure of Z_r^* (see Section B.6 in Appendix B) and noting that $\mathcal{D}^2 \left(\tilde{\eta}_{rs}; \delta \right)$ is the same for every $s = 1, 2, \ldots, k$, the third summand in the right hand side of (4.25) is zero and the modified scores for γ are found to take the elegant form

$$\tilde{U}_t^* = \sum_r \sum_{s=1}^{q} \left[y_{rs} + \frac{1}{2} \tilde{h}_{rss} - \left(\tau_r + \frac{1}{2} \, \mathrm{trace} \, \tilde{H}_r \right) \pi_{rs} \right] z_{rst} \, ,$$

for $t = 1, \ldots, pq$. On the parameter space restricted by $\tau_r = m_r$, the scores and the Fisher information on γ are equal to their counterparts for the multinomial logit model. Hence, in the restricted parameter space, the modified scores for γ corresponding to model (4.24) are given by

$$\tilde{U}_t^* = \sum_r \sum_{s=1}^{q} \left[y_{rs} + \frac{1}{2} \tilde{h}_{rss} - \left(m_r + \frac{1}{2} \, \mathrm{trace} \, \tilde{H}_r \right) \pi_{rs} \right] z_{rst} \, , \tag{4.26}$$

for $t = 1, \ldots, pq$.

The key results for the equivalence of (4.26) with the modified scores (4.23) for the multinomial logistic regression model are given in the following theorem and corollary.

Theorem 4.3.1: *Let H_r be the $q \times q$, r-th block of the asymmetric hat matrix H for the multinomial logistic regression model with parameters γ, and $\tilde{H}_r$ the $k \times k$, r-th block of the asymmetric hat matrix $\tilde{H}$ for the equivalent Poisson log-linear model in (γ, ϕ) parameterization. If we restrict the parameter space by $\tau_r = \sum_{s=1}^{k} \mu_{rs} = m_r$, for $r = 1, 2, \ldots, n$, we have*

$$\tilde{h}_{rss} = \pi_{rs} + h_{rss} - \sum_{u=1}^{q} \pi_{ru} h_{rus} \quad (s = 1, \ldots, q) \; ;$$

$$\tilde{h}_{rkk} = \pi_{rk} + \sum_{s,u=1}^{q} \pi_{ru} h_{rus} \, .$$

The proof is in Section B.6, Appendix B.

Corollary 4.3.1: *Using the same notation and conditions as in Theorem 4.3.1,*

$$\operatorname{trace} \tilde{H}_r = \operatorname{trace} H_r + 1 \quad (r = 1, \ldots, n) \; . \tag{4.27}$$

Proof. If we consider the sum $\sum_{s=1}^{k} \tilde{h}_{rss}$ and replace $\tilde{h}_{rss}$ by Theorem 4.3.1, the result follows by the fact that $\sum_{s=1}^{k} \pi_{rs} = 1$. □

Obviously, in the light of these results,

$$\tilde{U}_t^* = U_t^* \quad (t = 1, \ldots, pq) \; ,$$

and so either approach can be used for obtaining the BR estimator of γ. In ML, the likelihood equations for the nuisances are $\hat{\tau}_r = m_r$, $r = 1, \ldots, n$ and so the parameter space is restricted automatically. In contrast, for maximum penalized likelihood, it is necessary to restrict the parameter space by $\{\tau_r = m_r\}$; only $\hat{\gamma}$ should be affected by the bias-reducing modification, not $\hat{\tau}$. Despite the fact that by the orthogonality of γ and τ, both restricted and unrestricted maximum penalized likelihood result in the same estimates for γ, the reason for the restriction is that without it, the modified score equations would result in an estimator for τ, of the form

$$\tilde{\tau}_r = m_r + \frac{1}{2} \operatorname{trace} \tilde{H}_r \quad (r = 1, \ldots, n) \; ; \tag{4.28}$$

the multinomial totals in the fitted model would then be incorrect.

4.3.4 Iterative adjustments of the response

The forms of (4.23) and (4.26) suggest two alternative pseudo-response representations:

i)

$$y^*_{rs} = y_{rs} + \frac{1}{2}h_{rss} - \frac{1}{2}\,\mathrm{trace}\, H_r\pi_{rs} - \frac{1}{2}\sum_{u=1}^{q}\pi_{ru}h_{rus}\,, \tag{4.29}$$

$$y^*_{rk} = y_{rk} - \frac{1}{2}\,\mathrm{trace}\, H_r\pi_{rk} + \frac{1}{2}\sum_{s,u=1}^{q}\pi_{ru}h_{rus} \quad (r = 1,\ldots,n\ ;\ s = 1,\ldots,q)\,,$$

ii)

$$\tilde{y}^*_{rs} = y_{rs} + \frac{1}{2}\tilde{h}_{rss} - \frac{1}{2}\,\mathrm{trace}\,\tilde{H}_r\pi_{rs} \quad (r = 1,\ldots,n\ ;\ s = 1,\ldots,k)\,.$$

Note that both pseudo-data representations above are constructed in order to have the multinomial pseudo-totals equal to the totals observed or fixed by design. In this way and by the same arguments as in the binomial case, we avoid any possible systematic underestimation of standard errors by some artificial inflation of the multinomial totals.

If the two last terms in the right of the above expressions were known constants, the BR estimator would then be formally equivalent to the use of ML after adjusting the response y_r to y^*_r. However, in general, both H_r and $\tilde{H}_r$ depend on γ, exceptions to this being very special cases such as saturated models. The utility of the above definitions of pseudo-observations is that they directly suggest simple, iterative computational procedures for obtaining the BR estimates (Section 4.3.7 below).

4.3.5 Saturated models and Haldane correction

Consider a saturated model of the form (4.19). The model then has nq parameters and the hat matrix H is the identity. Hence the modified score equations in this case take the form

$$0 = \sum_r\sum_{s=1}^{q}\left(y_{rs} + \frac{1}{2} - \left(m_r + \frac{k}{2}\right)\pi_{rs}\right)z_{rst} \quad (t = 1,\ldots,nq)\ .$$

Thus the maximum penalized likelihood method is equivalent to the addition of 1/2 to each frequency and then the application of ML using the modified responses. So, in this case, the maximum penalized likelihood is equivalent to the Haldane correction (Haldane, 1956) introduced for avoiding singularities in the estimation of log-odds in sparse arrays and producing the well-known bias-reducing "empirical logistic transform". Parameter estimates in this case are obtained by solving with respect to γ the equations

$$\eta_{rs}(\gamma) = \sum_{t=1}^{nq}\gamma_t z_{rst} = \log\frac{y_{rs} + 1/2}{y_{rk} + 1/2} \quad (r = 1,\ldots,n;\ s = 1,\ldots,q)\,.$$

4.3.6 Properties of the bias-reduced estimator

The finiteness properties of the BR estimator for binomial-response logistic regression generalize directly to the case of multinomial response model. In particular, the finiteness of the BR estimator can be proved by direct use of the results in Albert & Anderson (1984), Santner & Duffy (1986) and Lesaffre & Albert (1989).

4.3.6.1 Finiteness

Theorem B.4.4 in Appendix B (Lesaffre & Albert, 1989) shows the behaviour of the inverse of the Fisher information in an iterative fitting procedure when complete or quasi-complete separation of the sample points occurs. Specifically, if $F_{(c)}$ is the Fisher information on γ evaluated at the c-th iteration, complete or quasi-complete separation occurs if and only if at least one diagonal element of $F_{(c)}^{-1}$ diverges as c grows. Thus trace $\left(F_{(c)}^{-1}\right)$ diverges as the number of iterations tends to infinity so that at least one eigenvalue of $F_{(c)}^{-1}$ diverges. Hence, $\det\left(F_{(c)}\right) \to 0$ as c tends to infinity.

Now, consider estimation by maximization of a penalized log-likelihood function of the form

$$l^{(a)}(\gamma) = l(\gamma) + a \log\det F(\gamma)\,,$$

where a is a fixed positive constant, and denote by $\tilde{\gamma}$ the resultant estimator. Below we show that $\tilde{\gamma}$ takes finite values even when either complete or quasi-complete separation occurs.

Theorem 4.3.2: *In the case of complete separation of the data points, the estimator that results from the maximization of $l^{(a)}$ takes finite values.*

Proof. Let Γ^C be the set of all γ's satisfying (B.11) in Definition B.4.1. Then, as in Albert & Anderson (1984), Γ^C is the interior of a convex cone. The generic element of Γ^C can be denoted as $k\alpha$, with $\alpha \in \Gamma^C$ and $k > 0$. By Theorem B.4.1 the ML estimate $\hat{\gamma}$ falls on the boundary of the parameter space; this is proved in Albert & Anderson (1984) by showing that if we move along any ray $k\alpha$ in Γ^C and let k increase towards infinity the likelihood attains its maximum value of 1. In addition, the strict concavity of the log-likelihood guarantees that this maximum value is attained only when γ is of the form $\lim_{k\to\infty} k\alpha$, for every $\alpha \in \Gamma^C$. Further, by our previous discussion $\det(F(k\alpha)) \to 0$ as $k \to \infty$ and hence the value of the penalized likelihood diverges towards $-\infty$. Consequently, since there is always a choice of γ, for example $\gamma = 0$, such that $l^{(a)}(\gamma)$ is finite, the maximum penalized likelihood estimator $\tilde{\gamma}$ does not have the form $\lim_{k\to\infty} k\alpha$ with $\alpha \in \Gamma^C$. Further, by the strict concavity argument above, $0 = l(\hat{\gamma}) > l(\tilde{\gamma})$, giving

$$\det F(\tilde{\gamma}) > \det F(\hat{\gamma}) = 0\,.$$

Hence, there always exists $\tilde{\gamma} = \arg\max_{\gamma\in\Gamma} l^{(\alpha)}(\gamma)$ with finite components. □

Theorem 4.3.3: *In the case of quasi-complete separation of the data points, the estimator that results from the maximization of $l^{(a)}$ takes finite values.*

Proof. We use the same line of argument as in the proof of Theorem 4.3.2 but with some modifications. First, we replace the set Γ^C by Γ^Q which is the set of all vectors γ satisfying (B.12) in Definition B.4.2 of quasi-complete separation. Santner & Duffy (1986), correcting technical details in the proofs in Albert & Anderson (1984), show that if Γ^C is empty and $\Gamma^Q \neq \{0\}$ then Γ^Q is a convex cone. Further, they define $\gamma(k, \alpha^*, \alpha) = \alpha^* + k\alpha$ with $\alpha^* \in \Re^{pq}$ and $\alpha \in \Gamma^Q \setminus \{0\}$. By this construction any vector in $\Re^{pq}$ can be described as

the sum of some arbitrary vector $\alpha^* \in \Re^{pq}$ and a vector in the convex cone Γ^Q, excluding the zero vector. In Albert & Anderson (1984), it is proved that for fixed α^* and α, $l(\gamma(k, \alpha^*, \alpha))$ is a strictly increasing function of k with some upper asymptote $l_u < 0$ and so the ML estimates do not exist. Hence, if we use l_u in the place of the value 0 for the log-likelihood in the case of complete separation, the same arguments as in the proof of Theorem 4.3.2 can be used for the finiteness of $\tilde{\gamma}$ in quasi-separated cases. □

4.3.6.2 Shrinkage

As in the binomial case, a complete proof of shrinkage should consist of two parts; a part showing that Jeffreys prior has a maximum at $\gamma = 0$ and a part for the log-concavity of Jeffreys prior with respect to the category probabilities. Then the same discussion as in the binomial response case applies.

The first part has already been covered by Poirier (1994) who proves analytically that the Jeffreys prior for multinomial response logistic regression models has a local mode at $\gamma = 0$.

However, the second part is much more complicated to put formally in the multinomial setting on account of the fact that W is no longer diagonal but is block diagonal. Despite the fact that we have not encountered empirical evidence contradicting shrinkage (see also the empirical results in Bull et al., 2002) with respect to the metric (4.13), a formal proof remains to be formulated and it is the subject of further work.

4.3.7 IGLS procedure for obtaining the bias-reduced estimates

4.3.7.1 Iterative algorithm

For general models, we propose a modification of the IGLS algorithm for ML estimation. By (2.15), the r-th working variate vector for ML has the form

$$\zeta_r = Z_r\gamma + W_r^{-1}(y_r - m_r\pi_r) \quad (r = 1, \ldots, n) ,$$

and so its components have the form

$$\begin{aligned}\zeta_{rs} &= \log\frac{\pi_{rs}}{\pi_{rk}} + \sum_{u=1}^{q}\frac{y_{ru} - m_r\pi_{ru}}{m_r\pi_{rk}} + \frac{y_{rs} - m_r\pi_{rs}}{m_r\pi_{rs}} \\ &= \log\frac{\pi_{rs}}{\pi_{rk}} + \frac{y_{rs}\pi_{rk} - y_{rk}\pi_{rs}}{m_r\pi_{rs}\pi_{rk}} ,\end{aligned}$$

for $r = 1, \ldots, n$ and $s = 1, \ldots, q$. If we replace the observed responses with the pseudo-responses in (4.29), we obtain the modified working variate that can be used for obtaining the BR estimates as follows.

Assume that the current estimates are $\gamma_{(c)}$. The updated estimate $\gamma_{(c+1)}$ is obtained from the following three steps:

i) Calculate

$$H_{(c)} = Z\left(Z^T W_{(c)} Z\right)^{-1} Z^T W_{(c)} \quad ,$$

with $H_{(c)} = H\left(\gamma_{(c)}\right)$.

ii) At $\gamma_{(c)}$ evaluate the current value of the modified working variates

$$\zeta^*_{rs} = \log \frac{\pi_{rs}}{\pi_{rk}} + \frac{y^*_{rs}\pi_{rk} - y^*_{rk}\pi_{rs}}{m_r \pi_{rs}\pi_{rk}},$$

for $r = 1, \ldots, n$ and $s = 1, \ldots, q$, where y^*_{rs} are as in (4.29).

iii) The updated estimate is then

$$\gamma_{(c+1)} = \left(Z^T W_{(c)} Z\right)^{-1} Z^T W_{(c)} \zeta^*_{(c)},$$

with $\zeta^* = (\zeta^*_{11}, \ldots, \zeta^*_{1q}, \ldots, \zeta^*_{n1}, \ldots, \zeta^*_{nq})^T$.

The iteration of the above scheme until, for example, the changes to the estimates are sufficiently small, returns the BR estimates. By construction, the iteration is exactly the same as the IGLS iteration for ML but with the observed counts y_{rs} replaced by y^*_{rs} in the working variate formulae. In this sense this is a modification of the standard IGLS that is often used for ML estimation. More specifically, if we replace y^*_{rs} and y^*_{rk} as in (4.29) and use identity (B.21) in Appendix B, the modified working variates can be written in the elegant form

$$\zeta^*_{rs} = \zeta_{rs} - \xi_{rs},$$

where $\xi_{rs} = -h_{rss}/(2m_r\pi_{rs}) + \sum_{u=1}^{q} h_{rsu}/(2m_r\pi_{rk})$, for $r = 1, \ldots, n$ and $s = 1, \ldots, q$. So the bias-reduction method can be implemented simply by subtracting ξ_{rs} from the working variates of the standard IGLS procedure for ML.

Note that if we drop the dimension of the response to $q = 1$, everything reduces to the results of the previous section for binary response models.

4.3.7.2 Nature of the fitting procedure

As starting values $\gamma^{(0)}$ for the parameters we can use the ML estimates after adding 1/2 to the initial frequencies. The correction to the initial frequencies is made in order to ensure the finiteness of the starting values even in cases of complete or quasi-complete separation. Also, this procedure will generally converge with linear rate, in contrast to the standard IGLS which converges with quadratic rate. In terms of the equivalent Fisher scoring procedure, the reason is that only the first term $F(\gamma) = Z^T W(\gamma) Z$ of the Jacobian of the modified score vector is used. However, in all of the various examples in which we have applied the procedure with the above starting values, satisfactory convergence is achieved after a very small number of iterations, and the difference in run-time from the standard IGLS for ML is small.

Further, note that since the Fisher information $F(\gamma)$ is positive definite, the above iteration will always deliver an increase in the penalized log-likelihood.

4.3.7.3 Estimated standard errors

By the general results of Section 3.5, the variance of the asymptotic distribution of the BR estimator agrees with the variance of the asymptotic distribution of the ML estimator,

both being the inverse of the Fisher information evaluated at the true parameter value γ_0. Thus, estimated standard errors for the BR estimates can be obtained as a byproduct of the suggested procedure by using the square roots of the diagonal elements of $\left(Z^T W(\gamma) Z\right)^{-1}$ evaluated at the final iteration.

4.4 On the coverage of confidence intervals based on the penalized likelihood

Heinze & Schemper (2002) and later Bull et al. (2007) illustrated through empirical work that confidence intervals for the BR estimates based on the ratio of the profiles of the penalized likelihood (Heinze-Bull intervals, for short) have better coverage properties than both the usual Wald-type intervals and the ordinary likelihood-ratio intervals. However, we object that such confidence intervals could exhibit low or even zero coverage for hypothesis testing on extreme parameter values. This is a direct consequence of the shape of the penalized likelihood, which does not allow confidence intervals with extreme left or right endpoints.

The same behaviour appears for symmetric confidence intervals for the log-odds in a contingency table. For example, for the log odds-ratio β of a 2×2 contingency table with counts y_{11}, y_{12}, y_{21} and y_{22}, Gart (1966) proposes a $100(1-\alpha)$ per cent confidence interval of the form

$$\log \frac{(y_{11}+1/2)(y_{22}+1/2)}{(y_{12}+1/2)(y_{21}+1/2)} \pm \Phi^{-1}(\alpha/2) \sqrt{\sum_{r=1}^{2}\sum_{s=1}^{2} \frac{1}{y_{rs}+1/2}}\,, \tag{4.30}$$

where $\Phi^{-1}(\alpha/2)$ is the $\alpha/2$ quantile of the normal distribution. So, the counts are modified by appending $1/2$ to them (the same effect as the bias-reduction method would have for such an estimation problem), and then a Woolf interval (Woolf, 1955) is constructed based on the modified counts. Agresti (1999) illustrates that the coverage of intervals of the form (4.30) deteriorates as the true parameter value increases, because "for any such interval with given n_1 and n_2, there exists $\theta_{L0} < \theta_{U0}$ such that, for all $\theta < \theta_{L0}$ and $\theta > \theta_{U0}$, the actual coverage probability equals zero" (Agresti, 1999, §2, p. 599), where, in our notation, n_1 and n_2 are $m_1 = y_{11} + y_{12}$ and $m_2 = y_{21} + y_{22}$, respectively, θ is $\exp(\beta)$ and, for any given n_1 and n_2 , i) θ_{L0}, ii) θ_{U0} are some i) lower and ii) upper finite bounds for the values of the i) lower and ii) upper end-points of the Gart interval (actually, Agresti, 1999, deals with confidence intervals for the odds ratio and considers the Gart interval by exponentiating its endpoints, but (4.30) for the log odds-ratio has the same behaviour).

The same argument, as the quoted above, applies for Heinze-Bull confidence intervals. As a non-trivial illustration of our objection, we consider a variant of the simple example in Copas (1988, §2.1). Assume that binomial observations y_1, y_2 ,y_3 ,y_4 ,y_5, each with totals m, are made independently at each one of five design points $x_r = cr - c$ $(r = 1, \ldots, 5)$, where c is some real constant. The model to be fitted is

$$\log \frac{\pi_r}{1-\pi_r} = \beta x_r \quad (r = 1, \ldots, 5)\ .$$

This is a non-trivial example in the sense that the bias-reduction method iteratively inflates the observed counts by quantities that depend on the parameter value (half a leverage).

We set $c = 2$ and perform a complete enumeration of the 1024 possible samples that could arise for $m = 3$. We consider confidence intervals for β based on the ordinary likelihood ratio (LR) statistic $W(\beta) = 2l(\hat{\beta}) - 2l(\beta)$ and on the penalized-likelihood ratio (PLR) statistic $W^*(\beta) = 2l^*(\tilde{\beta}) - 2l^*(\beta)$, where $\hat{\beta}$ and $\tilde{\beta}$ are the ML and BR estimates for β. The latter statistic is the one that was used by Heinze & Schemper (2002) and Bull et al. (2002). Imitating the construction of the ordinary likelihood ratio interval, the endpoints of the $100(1-\alpha)$ per cent Heinze-Bull interval are obtained by the solution of the inequality $W^*(\beta) < \chi_{1-\alpha}$ where $\chi_{1-\alpha}$ is the $1-\alpha$ quantile of a chi-squared distribution with 1 degree of freedom.

For the vector of observed responses $y = (y_1, \ldots, y_5)^T$, denote by $C_{\mathrm{LR}}(y, \alpha)$ and by $C_{\mathrm{PLR}}(y, \alpha)$ the $100(1-\alpha)$ per cent LR and Heinze-Bull (or PLR) confidence intervals for β. Based upon the complete enumeration, we calculate the corresponding coverage probabilities $\mathrm{E}[I(\beta_0 \in C_{\mathrm{LR}}(Y, 0.05))]$ and $\mathrm{E}[I(\beta_0 \in C_{\mathrm{PLR}}(Y, 0.05))]$ on a fine grid of values for the true parameter β_0, where $I(B)$ takes value 1 if condition B is satisfied and 0 else. Figure 4.3 shows the coverage probabilities plotted against β_0.

First, note the familiar oscillating effect of the coverage that is caused by the discrete nature of the responses (see, for example Brown et al., 2001, where oscillation is studied for intervals for a binomial proportion).

In Region 1 (see Figure 4.3) the PLR based interval outperforms the LR interval in terms of coverage. More explicitly, within that region the mean coverage for LR is 0.926 to three decimals, in contrast to 0.956 for the PLR. This illustrates the favourable behaviour of PLR intervals for moderate parameter values, having coverage very close to the nominal within Region 1 and avoiding the undesirable drop of coverage (long spikes) that the LR interval illustrates for $|\beta_0| \simeq 0.4$. However, outside Region 1 the PLR confidence interval starts to misbehave by illustrating severe oscillation in its coverage with long spikes below the nominal level. Eventually, the coverage drops to zero for $|\beta_0| \gtrsim 3.1$. In contrast the LR confidence interval tends to have coverage 1 as $|\beta| \to \infty$ because the expected length of the interval tends to ∞ as $|\beta| \to \infty$.

Increasing the absolute value of the c constant, we can construct much more severe examples where the loss of coverage occurs arbitrarily close to $\beta_0 = 0$. For example, for $c = 3$ (the plots are not shown here) the loss of coverage occurs for $|\beta_0| \gtrsim 2$. However, since c controls just the scale of the covariate values (and thus the scale of the estimate), one might argue that decreasing the absolute value of c, we can achieve the drop of coverage to take place after arbitrarily large absolute values of the true parameter. However, the loss of coverage will always take place, eventually. This is an undesirable property of such intervals and is mentioned neither in Heinze & Schemper (2002) nor in Bull et al. (2007).

A conservative workaround could be the definition of an interval having the form

$$C_{\mathrm{LR}}(y, \alpha) \cup C_{\mathrm{PLR}}(y, \alpha)\,.$$

A $100(1-\alpha)$ per cent interval of this form has coverage probability

$$\mathrm{E}\left[I\left(\{\beta_0 \in C_{\mathrm{LR}}(Y, \alpha)\} \cup \{\beta_0 \in C_{\mathrm{PLR}}(Y, \alpha)\}\right)\right]\,.$$

In Figure 4.4 we give the corresponding to Figure 4.3 ($c = 2$, $m = 3$, $\alpha = 0.05$) plot for such an interval. Despite its global conservativeness, this interval inherits the desirable properties of $C_{\mathrm{PLR}}(y, \alpha)$ in Region 1 (mean coverage for Region 1 is 0.969) and at the same time avoids the irregular oscillation and the complete loss of coverage in Region 2. The extension of our proposal in problems where the target parameter β has dimension $p > 1$ is direct: for every component β_t ($t = 1, \ldots, p$) of the parameter vector, replace $l(\beta)$ and $l^*(\beta)$ in the statistics $W(\beta)$ and $W^*(\beta)$ by $l_{\mathrm{P}}(\beta_t)$ and $l^*_{\mathrm{P}}(\beta_t)$, which are the profile likelihood and profile penalized likelihood, respectively.

4.5 General remarks and further work

By the finiteness and shrinkage properties of the BR estimator and the fact that the BR estimates can be easily obtained via a modified IGLS procedure, we conclude that the application of the bias-reduction method is rather attractive and should be regarded as an improvement over traditional ML in logistic regression models. All the theoretical results that have been presented can be supported by the extensive empirical studies in Heinze & Schemper (2002) and Bull et al. (2002).

In addition, as illustrated in the previous section, PLR based intervals (Heinze & Schemper, 2002; Bull et al., 2007) could misbehave with complete loss of coverage. In this direction, we have proposed an alternative interval which, despite being conservative, it avoids the loss of coverage for large parameter values and its coverage probability illustrates smaller oscillation across the parameter space.

A formal framework for measuring the goodness of fit is still lacking, and further work is required in this direction.

Figure 4.3: Coverage probability of 95 *per cent confidence intervals based on the likelihood ratio (LR) and the penalized-likelihood ratio (PLR), for a fine grid of values of the true parameter* β_0.

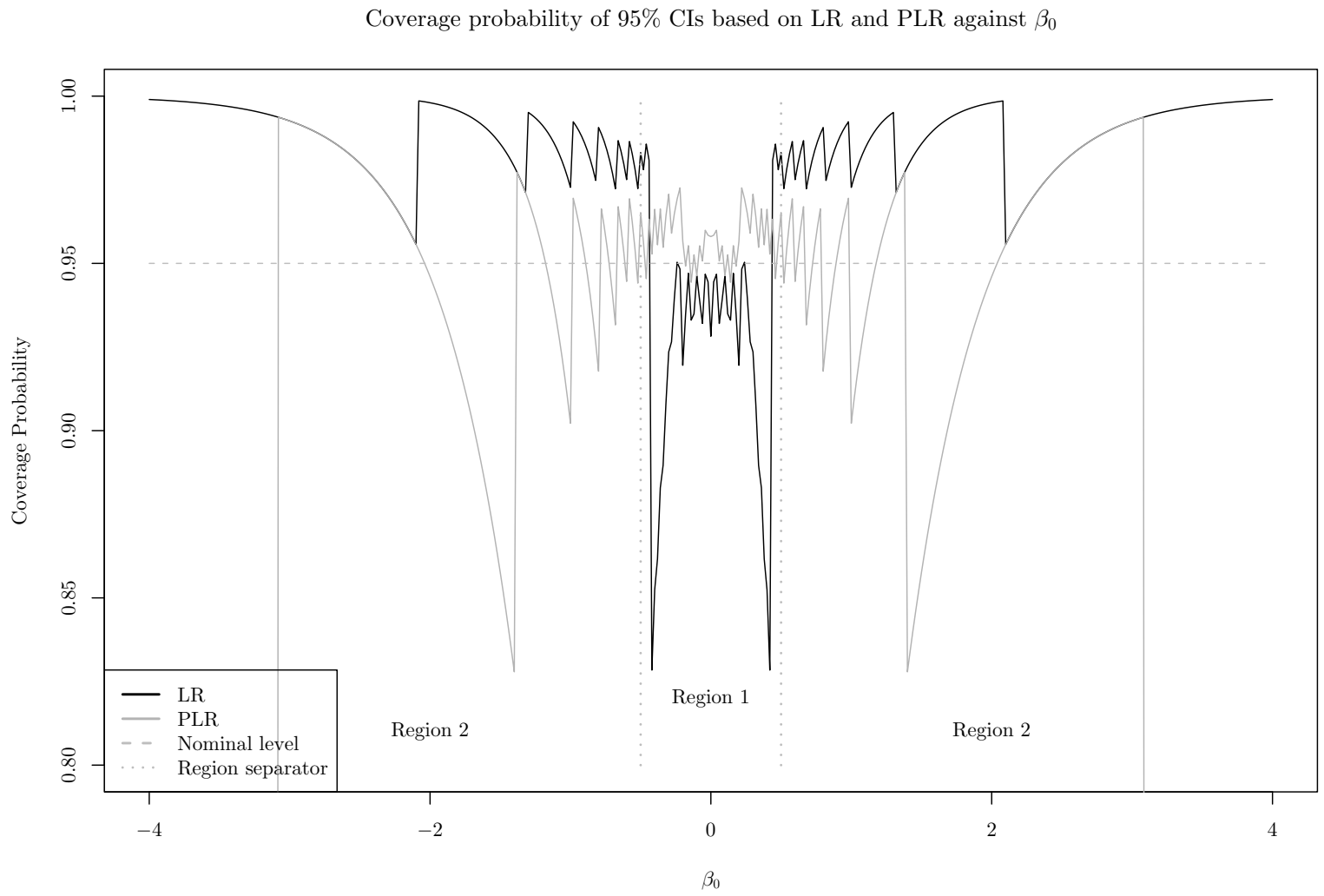
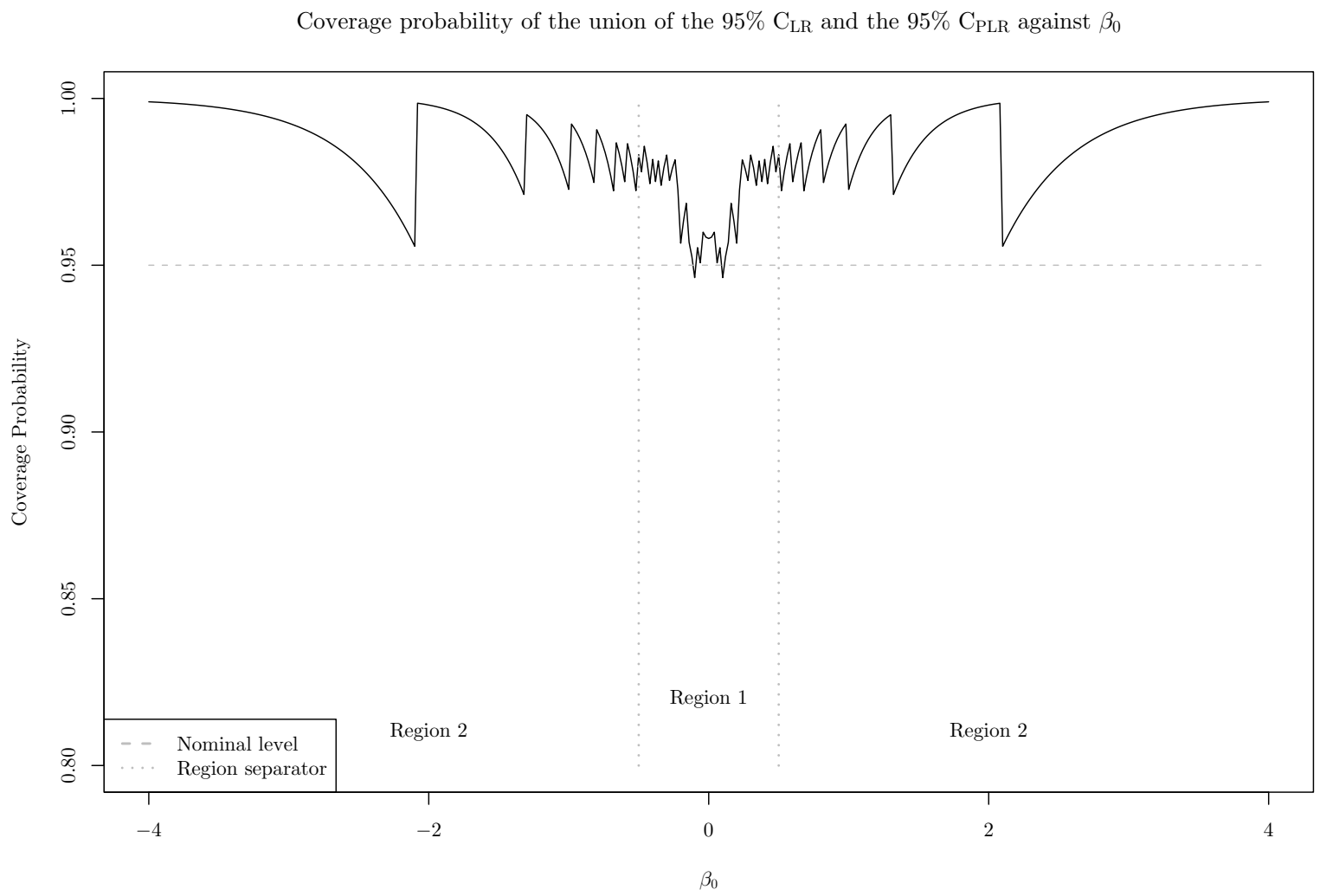


Figure 4.4: Coverage probability of the 95 *per cent confidence interval defined as the union of the intervals* $C_{LR}(y, 0.05)$ *and* $C_{PLR}(y, 0.05)$*, for a fine grid of values of the true parameter* β_0*.*



Chapter 5

Development for some curved models

5.1 Introduction

In the previous chapter we considered the effect of the bias-reduction method when applied to logistic regression, both for binomial and multinomial responses. These models are flat exponential families in canonical parameterization and both the application and the further derivation of theoretical results were significantly facilitated by the re-expression of the problem in terms of a penalized likelihood function, where the penalty is the Jeffreys invariant prior. However, as already seen in Chapter 3, when we deviate from flat exponential families, a penalized likelihood corresponding to the modified scores can have an intractable form or does not even exist. In most cases, this makes the derivation of elegant theoretical results difficult.

Continuing our treatise on models for categorical responses, the bias-reduction method is applied to the cases of binomial-response probit, complementary log-log and log-log models, deviating in this way from the canonical (logistic) link. Specifically, in the case of the complementary log-log link, some work has already been done by Mehrabi & Matthews (1995), who only consider a non-linear predictor complementary log-log model, with a single parameter and an offset, for modelling limiting dilution assays. We extend their derivation to more general regression settings, but with linear predictors. We also consider the "2-parameter logistic" (2PL) model (Birnbaum, 1968). Apart from its methodological importance in item response theory, our interest in the 2PL model model stems from the special form of the predictor which, despite being a non-linear function of the parameters, is connected with the extensively studied case of logistic regression. All of the aforementioned models can result in infinite maximum likelihood (ML) estimates with positive probability and it is illustrated that the bias-reduced (BR) estimates are always finite. Also, in each case we discuss aspects of the nature of shrinkage, based mainly on empirical results.

5.2 Binomial response models with non-canonical links

This section concerns the behaviour of the BR estimator under changes of the link structure in binary-response generalized linear models (GLMs). The alternatives considered are the commonly-used probit, complementary log-log (c-log-log) and log-log links. Applying the results of Chapter 3, we derive explicit expressions for the modified score functions. Their form enables the construction of pseudo-data representations by 'trading' quantities between the responses and the binomial totals. Such a pseudo-data representation is used to derive a general fitting algorithm for obtaining the BR estimates, which can be used with already implemented ML fitting procedures and is an alternative to the general modified iterative re-weighted least squares (IWLS) procedure that was suggested in Section 3.7. We conclude with empirical studies, that mainly demonstrate the extension of the finiteness and shrinkage properties of the BR estimator for these curved families.

5.2.1 Modified score functions

We consider the same setting as in Section 4.2, namely realizations $y_1, \ldots, y_n$ of n independent binomial random variables $Y_1, \ldots, Y_n$ with probabilities of *success* $\pi_1, \ldots, \pi_n$ and binomial totals $m_1, \ldots, m_n$, respectively. We denote the p-dimensional parameter vector as $\beta = (\beta_1, \ldots, \beta_p)^T$, and x_{rt} denotes the (r,t)-th element of a $n \times p$ design matrix X, assumed to be of full rank; if an intercept parameter is to be included in the models we can just set the first column of X to be a column of ones. For linear predictors $\eta_r = \sum_{t=1}^{p} \beta_t x_{rt}$ $(r = 1, \ldots, n)$, the form of the probit, c-log-log and log-log models is given in (5.1), where $\Phi(.)$ denotes the cumulative distribution function of the standard normal distribution and $\Phi^{-1}(.)$ its inverse.

$$\begin{array}{ll} \text{Link} & \text{Model Specification} \\ \text{probit} & \Phi^{-1}(\pi_r) = \eta_r \quad (r = 1, \ldots, n) \\ \text{c-log-log} & \log(-\log(1-\pi_r)) = \eta_r \quad (r = 1, \ldots, n) \\ \text{log-log} & -\log(-\log(\pi_r)) = \eta_r \quad (r = 1, \ldots, n) \end{array} \ . \tag{5.1}$$

For these models, the modified score functions based on the expected information can be obtained by a simple substitution of the corresponding second derivatives $d'_r = \partial^2 \mu_r / \partial \eta_r^2$ and of the corresponding working weights w_r (both given explicitly in Table 3.2) into the expression

$$U_t^* = \sum_r \frac{d_r}{\kappa_{2,r}} \left(y_r + \frac{1}{2} h_r \frac{d'_r}{w_r} - m_r \pi_r \right) x_{rt}^* \quad (t = 1, \ldots, p) \ ,$$

where $\kappa_{2,r} = m_r \pi_r (1 - \pi_r)$ are the binomial variances. This expression has been derived in Section 3.7 (see (3.26)). By Table 3.2, the probit model has $d'_r = -m_r \eta_r \phi(\eta_r)$ and $w_r = m_r [\phi(\eta_r)]^2 / [\pi_r (1 - \pi_r)]$, and so the modified score functions take the form

$$U_t^* = \sum_r \frac{\phi(\eta_r)}{\pi_r (1 - \pi_r)} \left(y_r - \frac{1}{2} h_r \frac{\pi_r (1 - \pi_r) \eta_r}{\phi(\eta_r)} - m_r \pi_r \right) x_{rt} \quad (t = 1, \ldots, p) \ , \tag{5.2}$$

where $\phi(.)$ denotes the density function of the standard normal distribution. For the c-log-log model, $d'_r = m_r(1-\pi_r)e^{\eta}_r(1-e^{\eta}_r)$ and $w_r = m_r e^{2\eta_r}(1-\pi_r)/\pi_r$. Hence,

$$U^*_t = \sum_r \frac{e^{\eta_r}}{\pi_r}\left(y_r + \frac{1}{2}h_r\frac{\pi_r(1-e^{\eta_r})}{e^{\eta_r}} - m_r\pi_r\right)x_{rt} \quad (t=1,\ldots,p)\ . \tag{5.3}$$

For the log-log model $d'_r = m_r\pi_r e^{-\eta_r}(e^{-\eta_r}-1)$ and $w_r = m_r e^{-2\eta_r}\pi_r/(1-\pi_r)$, and so

$$U^*_t = \sum_r \frac{e^{-\eta_r}}{1-\pi_r}\left(y_r + \frac{1}{2}h_r\frac{(1-\pi_r)(1-e^{-\eta_r})}{e^{-\eta_r}} - m_r\pi_r\right)x_{rt} \quad (t=1,\ldots,p)\ . \tag{5.4}$$

We should mention that in contrast to the case of logistic regression, Theorem 3.7.1 shows that there does not exist a penalized likelihood corresponding to either (5.2), (5.3), or (5.4). Hence, the location of their roots does not correspond to a maximization problem. The BR estimator in this cases should be viewed merely as a first-order unbiased "Z-estimator" (van der Vaart, 1998, § 5.2). This hinders the development of formal results on its finiteness and shrinkage properties; the method of proof that was used for logistic regressions in Chapter 4 cannot be applied here.

5.2.2 Obtaining the bias-reduced estimates via IWLS

The BR estimates can be obtained by the modified IWLS procedure described in Section 3.8. As is shown therein, the algorithm can be implemented simply by modifying the usual working observations ζ_r for ML to $\zeta_r - \xi_r$, with $\xi_r = -S_{rr}d'_r/(2d_r)$, where $S_{rr} = x_r^T(X^TWX)^{-1}x_r$ is the asymptotic variance of the ML estimator of η_r and W is the diagonal matrix that has diagonal elements the working weights w_r, $r=1,\ldots,n$. By the results in Table 3.2, the form of ξ_r for the logistic regression model (see Subsection 4.2.2) and the models considered here is given in Table 5.1. The same table is also presented in McCullagh & Nelder (1989, § 15.2.2) and in Cordeiro & McCullagh (1991, Table 1).

Table 5.1: Adjustments ξ_r for the modified IWLS in the case of logit, probit, c-log-log and log-log links.

Model	ξ_r
logit	$S_{rr}(\pi_r - 1/2)$
probit	$S_{rr}\eta_r/2$
c-log-log	$S_{rr}(e^{\eta_r}-1)/2$
log-log	$S_{rr}(1-e^{-\eta_r})/2$

5.2.3 *Refinement of the pseudo-data representation: Obtaining the bias-reduced estimates using already implemented software*

An alternative procedure for obtaining the BR estimates arises directly from the form of the modified score functions in (5.2), (5.3) and (5.4). The differences from the corresponding ordinary score functions

$$U_t = \sum_r \frac{d_r}{m_r \pi_r (1-\pi_r)} (y_r - m_r \pi_r) x_{rt} \quad (t = 1, \ldots, q), \tag{5.5}$$

are the extra terms that depend on h_r. These terms are additive to the responses and thus we can directly 'trade' parts of them between y_r and the totals m_r. In this way we can obtain a pseudo-data representation (pseudo-responses and pseudo-totals) that has certain desirable properties. Table 3.2 gives the obvious forms for such pseudo-data; the extra term is added to the response and the totals are left as observed or fixed by design. However, in that form the pseudo-responses can take negative values or become even greater than the binomial totals, thus violating the range of the actual responses. A natural requirement of the pseudo-data would be to be able to imitate the nature of the actual responses and totals. More formally we would like to find pseudo-responses y_r^* and pseudo-totals m_r^* which satisfy the condition $0 \leq y_r^* \leq m_r^*$ but which, at the same time, respect the form of the modified score functions. There is not a general solution to this task and each model has to be treated separately. The only general rule, which is imposed directly by the form of the modified scores, is that every quantity that is transferred from the responses to the totals is divided by $-\pi_r$ before attributing it to the pseudo-totals. This rule can be used alongside with the obvious calculation of adding and subtracting the same quantity to the responses.

For the probit model, and temporarily omitting the subject index r, the crude pseudo-responses in Table 3.2 are re-expressed as

	Pseudo-responses y^*	$\overset{\div(-\pi)}{\longrightarrow}$	Pseudo-totals m^*
P1.	$y - \frac{1}{2} h \frac{\pi(1-\pi)\eta}{\phi(\eta)}$		m
P2.	$y - \frac{1}{2} h \frac{\pi\eta}{\phi(\eta)}$		$m - \frac{1}{2} h \frac{\pi\eta}{\phi(\eta)}$
P3.	$y - \frac{1}{2} h \frac{\pi\eta I(\eta<0)}{\phi(\eta)}$		$m + \frac{1}{2} h \frac{\eta[I(\eta\geq 0)-\pi]}{\phi(\eta)}$,

where $I(B) = 1$ if the condition B is satisfied, and 0 else. For the c-log-log model we have

	Pseudo-responses y^*	$\overset{\div(-\pi)}{\longrightarrow}$	Pseudo-totals m^*
P4.	$y + \frac{1}{2} h \frac{\pi(1-e^\eta)}{e^\eta}$		m
P5.	$y + \frac{1}{2} h \frac{\pi}{e^\eta}$		$m + \frac{1}{2} h$

and for the log-log model,

	Pseudo-responses y^*	$\stackrel{\div(-\pi)}{\longrightarrow}$	Pseudo-totals m^*
P6.	$y + \frac{1}{2}\, h\, \frac{(1-\pi)(e^{-\eta}-1)}{e^{-\eta}}$		m
P7.	$y + \frac{1}{2}\, h(1-\pi) - \frac{1}{2}\, h\, \frac{(1-\pi)}{e^{-\eta}}$		m
P8.	$y + \frac{1}{2}\, h(1 - \frac{(1-\pi)}{e^{-\eta}})$		$m + \frac{1}{2}\, h$.

Note that the pseudo-data representations derived in this way for each of the three cases (P1, P2, P3 for the probit, P4, P5 for the c-log-log and P6, P7, P8 for the log-log) are equivalent in the sense that if we replace the actual responses y_r with y_r^* and the actual totals m_r with m_r^* in the expression (5.5) for the ordinary score functions, we obtain the corresponding modified scores. Also, note that by substituting η_r according to the corresponding model specification in (5.1), the derived pseudo-data representations can be expressed in the general form

$$\begin{array}{ll} \text{Pseudo-responses} & y_r^* = y_r + \dfrac{1}{2} h_r a_{\mathrm{R}}(\pi_r) \\ \text{Pseudo-totals} & m_r^* = m_r + \dfrac{1}{2} h_r a_{\mathrm{T}}(\pi_r) \end{array} \qquad (r = 1, \ldots, n)\ . \tag{5.6}$$

According to Section 4.2, in the case of logistic regression $a_{\mathrm{R}}(\pi_r) = 1$ and $a_{\mathrm{T}}(\pi_r) = 2$. For the other link functions, the form of the adjustment functions $a_{\mathrm{R}}(\pi)$ and $a_{\mathrm{T}}(\pi)$ corresponding to P3, P5, and P8 is given in Table 5.2.

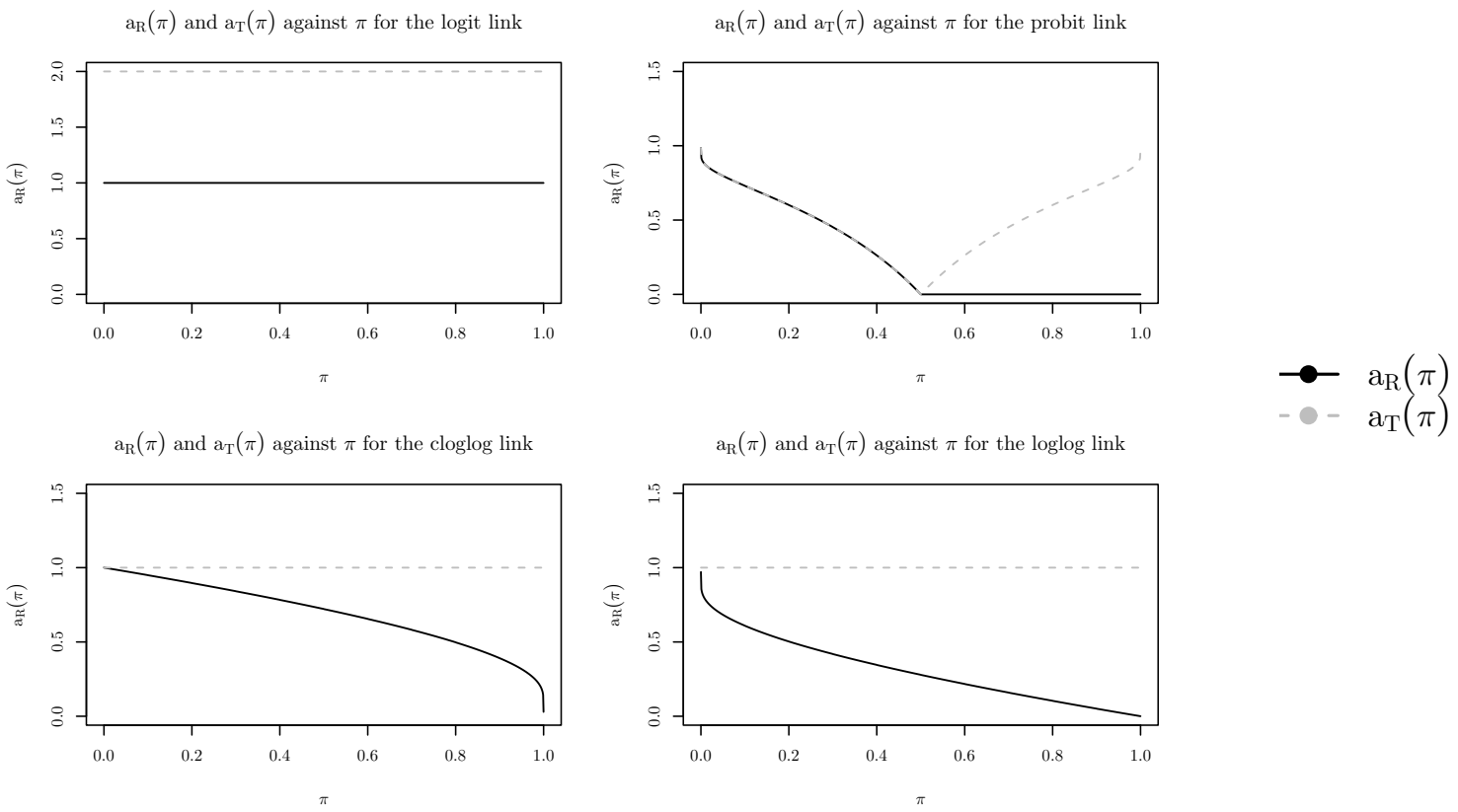
Table 5.2: Adjustment functions $a_R(\pi)$ and $a_T(\pi)$ for the logit, probit, c-log-log and log-log links in binary response GLMs

Link	$a_{\mathrm{R}}(\pi)$	$a_{\mathrm{T}}(\pi)$
logit	1	2
probit	$-\frac{\pi\Phi^{-1}(\pi) I(\pi<1/2)}{\phi(\Phi^{-1}(\pi))}$	$\frac{\Phi^{-1}(\pi)[I(\pi\geq 1/2)-\pi]}{\phi(\Phi^{-1}(\pi))}$
c-log-log	$-\frac{\pi}{\log(1-\pi)}$	1
log-log	$1 + \frac{1-\pi}{\log \pi}$	1

Since the leverages h_r take values in $[0, 1]$, the requirement $0 \leq y_r^* \leq m_r^*$ is translated in terms of $a_{\mathrm{R}}(.)$ and $a_{\mathrm{T}}(.)$ to $0 \leq a_{\mathrm{R}}(\pi_r) \leq a_{\mathrm{T}}(\pi_r)$. All of the adjustments in Table 5.2 satisfy the latter inequality and so they respect the range of the actual responses and totals (see Figure 5.1 for a visual justification). Thus, they can be directly used for obtaining the BR estimates using already implemented software, without the danger of error messages.

The general algorithm is given in Figure 5.2. The algorithm is given in a fairly general form and can be used for any choice of link function given that the derived pseudo-data

Figure 5.1: Adjustment functions $a_R(\pi)$ and $a_T(\pi)$ for the logit, probit, c-log-log and log-log links against $\pi \in (0, 1)$.



representation satisfies $0 \leq a_{\mathrm{R}}(\pi) \leq a_{\mathrm{T}}(\pi)$. In the initialization part (part A), we adjust the actual responses and the actual totals by appending to them 1/2 and 1, respectively, so that we eliminate the possibility of infinite ML estimates. Note that in the main iteration (part B, step ii)) we re-adjust the reported working weights $w^*_{r,(j)}$ to $w_{r,(j)}$ so that they agree with the actual totals. In this way it is ensured that the correct leverages are used.

As already mentioned in Section 3.5, the variance-covariance matrix of the asymptotic distribution of the BR estimator is the inverse of the Fisher information. However, extra care is needed for the estimation of the standard errors. In order to obtain the correct estimated standard errors, we need to repeat the adjustment of the working weights (step B.ii)) after convergence. The estimated standard errors are then the square roots of the diagonal elements of $(X^T W X)^{-1}$ where W is the diagonal matrix of the re-adjusted weights. In this way we avoid the underestimation of the standard errors by the artificial inflation of the binomial totals. Also, by the asymptotic normality of the BR estimator (see Section 3.5), the BR estimates could be accompanied by the familiar Wald-type asymptotic confidence intervals.

5.2.4 Empirical studies

5.2.4.1 Finiteness and shrinkage towards the origin

In order to assess the properties of the BR estimator in the case of binomial-response GLMs with non-canonical links, we proceed to a complete enumeration study based on the same design as the study conducted in Subsection 4.2.3.1. For the illustration of the finiteness of the BR estimates, we consider again two cross-classified two-level factors C_1 and C_2, and independent realizations of binomial random variables at each combination of levels (covariate settings) of C_1, C_2 with totals m_1, m_2, m_3, m_4, respectively (see Table 4.1). We set $m_1 = m_2 = m_3 = m_4 = 2$ and consider models with linear predictors of the form

$$\eta_r = \alpha + \beta x_{r1} + \gamma x_{r2} \quad (r = 1, \ldots, 4) \ , \tag{5.7}$$

where x_{r1} is equal to 1 if $C_1 = \mathrm{II}$ and 0 else and x_{r2} is 1 if $C_2 = \mathrm{B}$ and 0 else. For every possible data configuration with the above row totals, Table C.2 , Table C.3 and Table C.4 in Appendix C give the ML estimates, the bias-corrected (BC) estimates (Cordeiro & McCullagh, 1991) and the BR estimates for the probit, c-log-log and log-log links, respectively. The finiteness of the BR estimates is apparent since, in contrast to BC and and ML, they exist in all cases. In Subsection 4.2.3.1, Table 4.2 was constructed by formal arguments relating to the values of the sufficient statistics for the parameters of the logistic regression model. Those arguments do not apply for models with non-canonical link because such models are curved families and consequently the sufficient statistic has different dimension than the natural parameter (see Cox & Hinkley, 1974, Example 2.20). However, it is worth noting that for the three link functions we considered, infinite ML estimates occur for and only for the data configurations in Table 4.2. We have as yet not developed formal arguments explaining this behaviour, but according to further empirical evidence (not reported here) it is also noted for more general designs. A candidate starting point towards the formalization is that the theorems in Albert & Anderson (1984) and Santner & Duffy

Figure 5.2: Algorithm for obtaining the bias-reduced estimates for binomial-response models, using pseudo-data representations along with already implemented ML fitting procedures.

SCOPE:

- Binary response GLMs with link function $g(.)$ and parameter vector $\beta = (\beta_1, \ldots, \beta_n)$.

REQUIRES:

- An already implemented ML fitting procedure for such models.

INPUT:

- Observed response vector $y = (y_1, \ldots, y_n)^T$,
- binomial totals $m = (m_1, \ldots, m_n)^T$,
- the $n \times p$ design matrix X,
- other options for the implemented ML fitting procedure,
- tolerance $\epsilon > 0$ for the stopping criterion.

OUTPUT:
The bias-reduced estimates for β using modifications based on the Fisher information.

A. **INITIALIZATION: (0-th iteration)**

i) `Set` $j = 0$.

ii) `Set` $y^*_{r,(0)} = y_r + 1/2$ and $m^*_{r,(0)} = m_r + 1$, $r = 1, \ldots, n$. [a]

iii) `Fit` $y^*_{(0)} \sim X$ using totals $m^*_{(0)}$ and link $g(.)$.

B. **MAIN ITERATION ($(j+1)$-th iteration) :**

i) From the previous iteration `get`

- pseudo-totals $m^*_{r,(j)}$, $r = 1, \ldots, n$,
- fitted probabilities $\pi_{r,(j)}$, $r = 1, \ldots, n$,
- modified working weights $w^*_{r,(j)}$, $r = 1, \ldots, n$,
- estimated parameters $\beta_{t,(j)}$, $t = 1, \ldots, p$.

ii) `Set` $w_{r,(j)} = w^*_{r,(j)} m_r / m^*_{r,(j)}$, $r = 1, \ldots, n$.

iii) `Set` $H_{(j)} = X(X^T W_{(j)} X)^{-1} X^T W_{(j)}$. [b]

iv) `Set`

- $y^*_{r,(j+1)} = y_r + h_{r,(j)} a_{\mathrm{R}}(\pi_{r,(j)})/2$,
- $m^*_{r,(j+1)} = m_r + h_{r,(j)} a_{\mathrm{T}}(\pi_{r,(j)})/2$, $r = 1, \ldots, n$. [c]

v) `Fit` $y^*_{(j+1)} \sim X$ using totals $m^*_{(j+1)}$, link $g(.)$ and starting values $\beta_{(j)}$.

vi) `Set` $j = j + 1$.

vii) `Repeat` until either
a) $\sum_{t=1}^{p} |\beta_{t,(j+1)} - \beta_{t,(j)}| < \epsilon$, $\epsilon > 0$,
or alternatively
b) $\sum_{t=1}^{p} |U^*_t(\beta_{(j+1)})| < \epsilon$, $\epsilon > 0$

---

[a] For a quantity Q, $Q_{(c)}$ denotes the value of Q evaluated at the c-th iteration.
[b] $W_{(j)} = \mathrm{diag}\{w_{r,(j)}, r = 1, \ldots, n\}$.
[c] $h_{r,(j)}$ is the r-th diagonal element of $H_{(j)}$.

(1986) could be extended to cover more general link functions than the logistic since, given their monotonicity, they result in similar simple discrimination rules: "for $r = 1, \ldots, n$, the r-th observation is assigned to the group of successes if $\hat{\eta}_r > g(0.5)$, where $g(.)$ is the link function and $\hat{\eta}_r$ is the linear predictor η_r evaluated at the ML estimates for β and γ".

For the assessment of the nature of the apparent shrinkage of the BR estimates in Table C.2, C.3 and C.4, we proceed to a larger complete enumeration study on the same contingency table, increasing the row totals to $m_1 = m_2 = m_3 = m_4 = 4$. We consider only the non-separated datasets. Our main tools are the plots of the fitted probabilities $\hat{\pi}_{\mathrm{ML}}$ that correspond to the ML estimates, against the fitted probabilities $\hat{\pi}_{\mathrm{BR}}$ that correspond to the BR estimates. For any configuration of counts there are four fitted probabilities, one for each covariate setting in Table 4.1. However, for each link function, the four plots that result from the complete enumeration are identical. The reason is that for any given configuration of counts of successes we also consider the configurations resulting from the 4! possible permutations of these counts on the covariate settings. Thus, in Figure 5.3 we present 4 and not 16 plots. If there were no apparent shrinkage effect then we would expect the points on the plots to lie on a 45^o line. Shrinkage is identified if there is a single point of the form (c, c), $c \in (0, 1)$, for which the points $(\hat{\pi}_{\mathrm{BR}}, \hat{\pi}_{\mathrm{ML}})$ fall above the 45^o line if $\hat{\pi}_{\mathrm{BR}} > c$ and below it for $\hat{\pi}_{\mathrm{BR}} < c$.

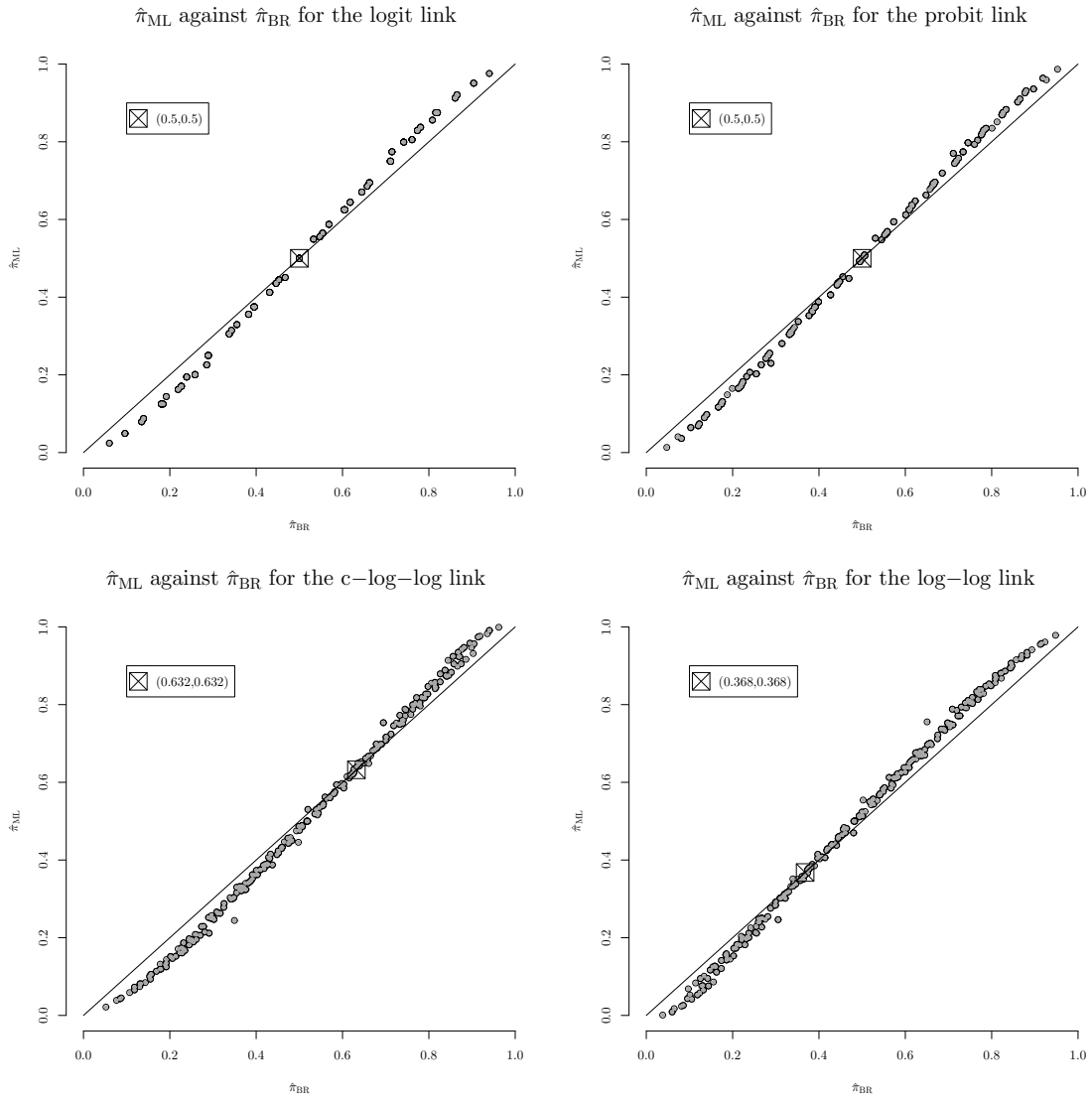
Table 5.3: The probability towards which $\hat{\pi}_{BR}$ shrinks for the logit, probit, c-log-log and log-log links.

Link	$P(Z_l < 0)$
logit	$\exp(0)/(1 + exp(0)) = 0.5$
probit	$\Phi(0) = 0.5$
c-log-log	$1 - \exp(-\exp(0)) \simeq 0.632$
log-log	$\exp(-\exp(0)) \simeq 0.368$

As is demonstrated in Figure 5.3, $\hat{\pi}_{\mathrm{BR}}$ shrinks towards $P(Z_l < 0)$ where Z_l is some latent random variable following the distribution imposed by the link function (or, equivalently, the distribution function of Z_l is the inverse link function; see Table 5.3). Consequently, the BR estimates shrink towards the origin on the scale of the link function. For the logit link this has already been formally proved in Section 4.2. For the remaining links, this agrees with the results for binary-response GLMs in Cordeiro & McCullagh (1991), where it is shown that the first-order bias vector is approximately collinear with the parameter vector (see Section 4.2 for a more analytic description of this result).

The case of log-log linked models will not be pursued any further, because the estimators for its parameters have identical bias, variance and mean squared error (MSE) with the estimators of the parameters of the c-log-log model. This is a direct consequence of the fact that c-log-log$(\pi) = -$log-log$(1 - \pi)$, or equivalently that, on the logistic scale, the

Figure 5.3: Demonstration of shrinkage of the fitted probabilities for the logit, probit, c-log-log and log-log links.



log-log link function is the reflection of the c-log-log link function on the 45^o line. Hence, from now on, any comments on the behaviour of the estimators for the c-log-log link apply for the log-log link, as well.

5.2.4.2 A Monte Carlo simulation study: Bias reduction versus bias correction

As it has been demonstrated by the previous complete enumeration studies, the advantage of the BR estimator over the BC estimator of Cordeiro & McCullagh (1991) is that the latter is undefined when the ML estimates are infinite. Here, a Monte Carlo simulation study allows the deeper comparison of the two estimators in terms of estimated bias and estimated variance. We only consider the probit and the c-log-log links. The case of logit links has been theoretically covered in Chapter 4, and extensive empirical results are already available in Heinze & Schemper (2002) and Bull et al. (2002).

Working on the two-way layout with a binomial response (table (Table 4.1)), we consider models with linear predictors of the form (5.7) and we investigate the BR and BC estimators under several different settings for the true parameter vector $(\alpha_0, \beta_0, \gamma_0)$. Among the wide range of true parameter settings we have considered, we present and comment on the results for the settings $(0, -0.5, 0.5)$, $(0, -2, 1.7)$ and $(-1, 0, 1.3)$, which from now on they will be referred to as 'A', 'B' and 'C', respectively. These were found to be good representatives for the illustration of the general behaviour of the estimators. Setting A is moderate for both the c-log-log and the probit link and assumes same probability for the first and fourth covariate setting (see Table 5.4). The parameter setting B has more extreme effects on the scale of each link function, implying very large and very small probabilities for the second and third covariate settings, respectively. The last parameter setting describes a moderate situation where the second and the fourth covariate settings are favoured in terms of probability.

Table 5.4: Implied probabilities by the probit and c-log-log links, for the parameter settings A, B and C

Link	Parameter setting	Implied probabilities			
		π_1	π_2	π_3	π_4
probit	A	0.5	0.691	0.309	0.5
	B	0.5	0.955	0.023	0.382
	C	0.159	0.618	0.159	0.618
c-log-log	A	0.632	0.808	0.455	0.632
	B	0.632	0.996	0.127	0.523
	C	0.308	0.741	0.308	0.741

Under each parameter setting, we simulate 20000 samples for the probit model and 20000 samples for c-log-log model. The whole set of simulations is performed once for row totals $m = m_1 = m_2 = m_3 = m_4 = 5$ and once for row totals $m = m_1 = m_2 = m_3 = m_4 = 25$, so that we can have an indication on the effect of the sample size. For each simulated sample, the ML, BR and BC estimates are obtained, and after the removal of

separated cases, we calculate the estimated bias, estimated variance and estimated MSE corresponding to each estimator. For the BR estimates, we use algorithm 5.2 with the stopping criterion B.vii).a) that monitors the change on the estimated values between successive iterations. The tolerance we set is $\epsilon = 10^{-10}$. In this way, the error from the numerical approximation of the roots of the modified score function is negligible compared to the bias and the variance of the BR estimator. For all of the simulated samples the algorithm converged after a few iterations (on average around 11 iterations for the probit link and around 14 for the c-log-log link), and the absolute value of the modified score function for β evaluated at the estimated value was at most of order 10^{-11}. Despite the fact that the removal of separated samples favours the ML and BC estimators, it is necessary in order to obtain finite estimated biases, variances and MSEs for them. So the following results and conclusions are conditional on the finiteness of the ML estimates.

Table 5.5 and Table 5.6 present the results of the empirical study. The following remarks can be made.

Remark 1. First, note that in most of the cases for $m = 5$ and parameter settings B and C, the estimated biases of the ML estimator appear to be considerably smaller than those of the BR and BC estimators. This is just a side-effect of the exclusion of the separated datasets from the calculations, and is also related to the — necessary for bias correction/reduction — shrinkage of the estimates towards the origin. The inclusion of only un-separated datasets corresponds to averaging over the part of the sample space in where the ML estimates are finite. Thus, we systematically exclude the long right (or left, depending on the direction of the true effects) tail of the distribution of the ML estimator. Further, a comparison of the corresponding estimated biases for $m = 25$, demonstrates that the phenomenon becomes less apparent as m increases. The reason is that as m increases, the probability of separation decreases, but at the same time, un-separated samples that allow for stronger estimated effects on the scale of each link function, are included in the study. Note that, when we include the separated samples in the study (see bracketed quantities in Table 5.5 and Table 5.6), the estimated biases for the BR estimator are smaller than the estimated biases based only on the un-separated cases. The reason is that we take into account values from the tails of the distribution of the BR estimator that are still finite but away from zero.

Remark 2. For the majority of the cases in Table 5.5 and Table 5.6, the BC estimator has smaller estimated variance than the BR estimator. However, a parallel comparison of the corresponding estimated biases and MSEs, reveals that the BC estimator has larger bias than the BR estimator, and in fact so large that in most cases the favourable picture for its estimated variance is reversed for the estimated MSE. In contrast to the BC estimator, the BR estimator seems to behave better in terms of estimated MSE, preserving a balance between estimated bias and estimated variance. The same behaviour is also noted for logistic regressions in both Heinze & Schemper (2002) and Bull et al. (2002).

Remark 3. For the c-log-log link (Table 5.6) and the parameter setting C, both the BR and BC estimators seem to correct the bias of the ML estimator beyond the true value. Especially for $m = 5$, the overcorrection is more severe for the BC estimator which, on parameter γ (true value 1.3) and to three decimal places, has estimated bias -0.211 , in

contrast to -0.140 of the BR estimator. The corresponding estimated variances are 0.288 and 0.338, respectively. The estimated bias and variance of the ML estimator for γ are 0.094 and 0.530, respectively. The same situation, but in smaller scales, is noted for the moderate parameter setting A, both for the probit and the c-log-log link. In more extreme settings such overcorrection when combined with the systematically smaller estimated variance of the BC estimator could raise serious concerns about its overall performance, since the estimator illustrates small variance but around the wrong value.

Remark 4. All of the above conclusions are conditional on the existence of the ML estimates. Thus, the tails of the distribution of the BR estimator were systematically excluded, not allowing it to take large finite values. In this way the shrinkage properties of the BR estimator seriously affected the results, and in several cases the BR estimator had artificially large estimated biases and small estimated variances. The bracketed quantities in Table 5.5 and Table 5.6 illustrate how this situation is improved when we include all the datasets in the study so that the estimated biases, variances and MSEs refer to the full distribution of the BR estimator. Histograms and kernel density estimates allow the visualization of the effect of conditioning. As an example, we consider the BR estimates of β for the probit link (Table 5.5) and under the parameter setting B. This setting has $\beta_0 = -2$, which is the largest assumed effect we considered in the study, and so the effect of conditioning is most apparent.

Figure 5.4 shows the histogams for the BR estimator of β. For $m = 5$ and when only un-separated samples are included in the study, the true value β_0 is only covered by the tail of the distribution and so the estimated bias is large. When all samples are included, -2 is much closer to the mean of the distribution. For $m = 25$ (the two plots in the bottom), separation is rarer (3636 separated samples out of 20000 simulated) and both histograms look similar, with the one based on all samples having slightly longer left tail on account of the inclusion of extreme negative estimated effects. The histograms for the c-log-log link (the corresponding plots are not shown here) illustrate the same behaviour but with longer left tails because the asymmetry of the c-log-log link allows it to accommodate larger effects without resulting in separation. This discussion is supported by the corresponding results in Table 5.5 and Table 5.6.

In conclusion, even if we exclude the part of the sample space that the ML estimator takes infinite values, the BR estimator has overall smaller estimated bias than the BC estimator. Conditional on the existence of the ML estimates, for small sample sizes and extreme parameter settings, the BC estimator has the tendency to overcorrect the bias beyond the true value, and at the same time it has small variance. This could be dangerous since the estimator gets less dispersed but around the wrong value. On the other hand, as the Monte Carlo study suggests, the BR estimator illustrates smaller overcorrection and despite being conservative on its variance, it preserves a lower MSE than the BC estimator.

As the sample size increases, the differences between the ML, the BC and the BR estimator diminish, and application of the bias-reduction method bears only the small extra cost in implementation and computation.

Table 5.5: Probit link. Estimated bias, estimated variance and estimated MSE to three decimal places, excluding the separated samples.

$(\alpha_0, \beta_0, \gamma_0)$	Row totals	Separated samples	Method	Estimated bias			Estimated variance			Estimated MSE		
				α	β	γ	α	β	γ	α	β	γ
A. $(0, -0.5, 0.5)$	$m = 5$	842	ML	0.011	−0.047	0.034	0.306	0.395	0.392	0.306	0.397	0.393
			BC	0.009	0.047	−0.058	0.210	0.264	0.261	0.210	0.266	0.265
			BR	0.009	0.032	−0.044	0.223	0.284	0.281	0.224	0.285	0.283
				[0.006]	[−0.008]	[−0.001]	[0.248]	[0.343]	[0.341]	[0.248]	[0.343]	[0.341]
	$m = 25$	0	ML	−0.001	−0.014	0.014	0.051	0.070	0.070	0.051	0.071	0.070
			BC	−0.001	0.000	0.001	0.048	0.066	0.066	0.048	0.066	0.066
			BR	−0.001	0.000	0.001	0.048	0.066	0.066	0.048	0.066	0.066
				[−0.001]	[0.000]	[0.001]	[0.048]	[0.066]	[0.066]	[0.048]	[0.066]	[0.066]
B. $(0, -2, 1.7)$	$m = 5$	14620	ML	−0.017	0.562	−0.536	0.297	0.210	0.227	0.297	0.526	0.515
			BC	−0.004	0.867	−0.793	0.189	0.131	0.143	0.189	0.882	0.772
			BR	−0.005	0.811	−0.746	0.206	0.145	0.156	0.206	0.802	0.713
				[0.004]	[0.271]	[−0.267]	[0.288]	[0.299]	[0.299]	[0.288]	[0.373]	[0.371]
	$m = 25$	3636	ML	−0.004	0.004	−0.013	0.060	0.098	0.096	0.060	0.099	0.096
			BC	−0.001	0.150	−0.149	0.056	0.072	0.070	0.056	0.094	0.092
			BR	−0.001	0.133	−0.133	0.056	0.076	0.075	0.056	0.094	0.092
				[0.001]	[0.008]	[−0.008]	[0.057]	[0.139]	[0.138]	[0.057]	[0.140]	[0.138]
C. $(-1, 0, 1.3)$	$m = 5$	4511	ML	0.044	−0.002	−0.037	0.249	0.414	0.308	0.251	0.414	0.310
			BC	0.228	−0.002	−0.274	0.143	0.254	0.188	0.196	0.254	0.263
			BR	0.196	−0.002	−0.232	0.160	0.277	0.210	0.199	0.277	0.264
				[0.011]	[−0.006]	[−0.009]	[0.312]	[0.395]	[0.382]	[0.312]	[0.395]	[0.382]
	$m = 25$	5	ML	−0.036	0.002	0.041	0.074	0.083	0.088	0.075	0.083	0.090
			BC	0.001	0.002	−0.004	0.065	0.077	0.079	0.065	0.077	0.079
			BR	0.000	0.002	−0.002	0.066	0.077	0.079	0.066	0.077	0.079
				[−0.001]	[0.002]	[−0.002]	[0.066]	[0.077]	[0.080]	[0.066]	[0.077]	[0.080]

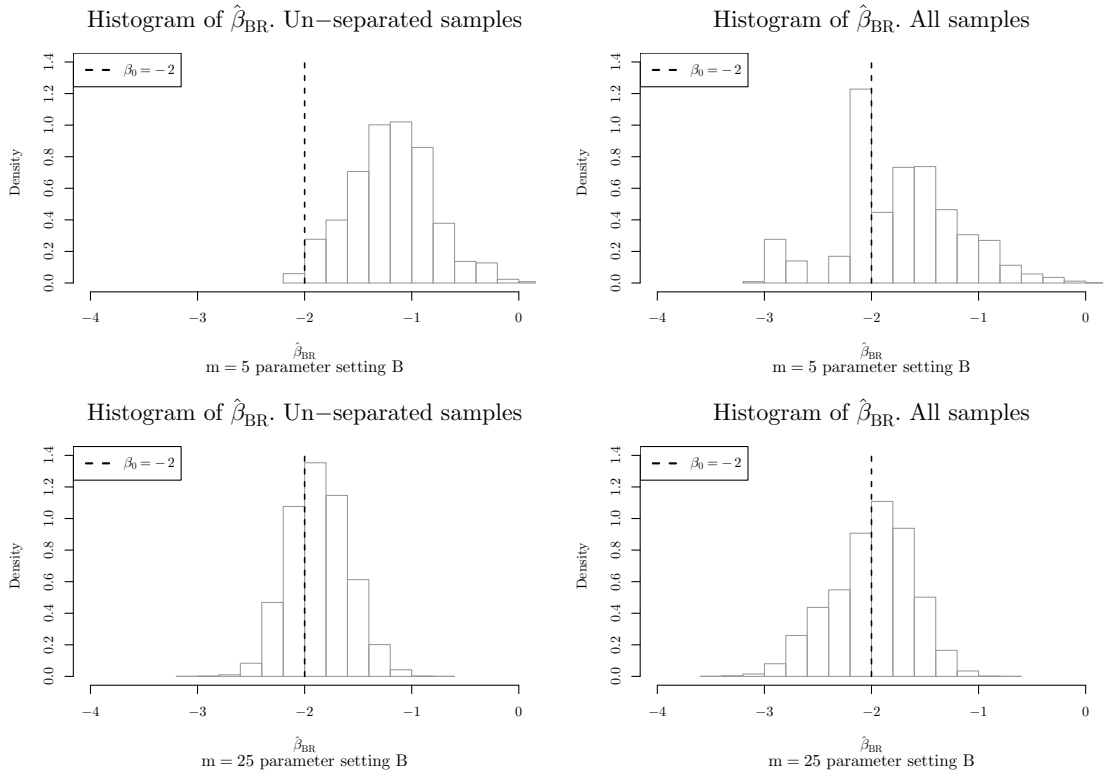
The bracketed ([.]) quantities refer to the corresponding estimated quantities for the BR estimator, when all the samples are included.

Table 5.6: C-log-log link. Estimated bias, estimated variance and estimated MSE to three decimal places, excluding the separated samples.

$(\alpha_0, \beta_0, \gamma_0)$	Row totals	Separated samples	Method	Estimated bias			Estimated variance			Estimated MSE		
				α	β	γ	α	β	γ	α	β	γ
A. $(0, -0.5, 0.5)$	$m = 5$	1728	ML	-0.022	-0.069	0.069	0.330	0.488	0.485	0.331	0.493	0.489
			BC	-0.043	0.075	-0.076	0.211	0.264	0.263	0.213	0.270	0.269
			BR	-0.034	0.038	-0.039	0.230	0.309	0.307	0.232	0.311	0.309
				[-0.003]	[-0.006]	[0.004]	[0.272]	[0.373]	[0.368]	[0.272]	[0.373]	[0.368]
	$m = 25$	0	ML	0.001	-0.018	0.018	0.056	0.078	0.078	0.056	0.079	0.079
			BC	0.000	-0.001	0.002	0.053	0.073	0.073	0.053	0.073	0.073
			BR	0.000	-0.002	0.002	0.053	0.073	0.073	0.053	0.073	0.073
				[0.000]	[-0.002]	[0.002]	[0.053]	[0.073]	[0.073]	[0.053]	[0.073]	[0.073]
B. $(0, -2, 1.7)$	$m = 5$	11199	ML	0.010	0.321	-0.274	0.243	0.272	0.269	0.243	0.375	0.345
			BC	-0.077	0.970	-0.886	0.163	0.192	0.176	0.168	1.132	0.960
			BR	-0.041	0.740	-0.663	0.172	0.182	0.166	0.174	0.730	0.605
				[-0.064]	[0.411]	[-0.434]	[0.311]	[0.385]	[0.348]	[0.315]	[0.554]	[0.536]
	$m = 25$	633	ML	0.025	-0.197	0.191	0.066	0.268	0.269	0.067	0.307	0.306
			BC	0.001	0.096	-0.096	0.063	0.174	0.179	0.063	0.183	0.188
			BR	-0.006	0.089	-0.090	0.063	0.154	0.156	0.063	0.162	0.164
				[-0.006]	[0.034]	[-0.035]	[0.063]	[0.257]	[0.257]	[0.064]	[0.258]	[0.258]
C. $(-1, 0, 1.3)$	$m = 5$	2383	ML	-0.084	-0.001	0.094	0.500	0.566	0.530	0.507	0.566	0.538
			BC	0.114	-0.002	-0.211	0.286	0.287	0.288	0.300	0.287	0.332
			BR	0.078	-0.001	-0.140	0.327	0.351	0.338	0.333	0.351	0.358
				[-0.004]	[-0.004]	[0.001]	[0.459]	[0.424]	[0.501]	[0.459]	[0.424]	[0.501]
	$m = 25$	0	ML	-0.027	0.005	0.035	0.094	0.092	0.106	0.095	0.092	0.107
			BC	0.003	0.005	-0.005	0.087	0.085	0.098	0.087	0.085	0.098
			BR	0.002	0.005	-0.004	0.088	0.085	0.098	0.088	0.085	0.098
				[0.002]	[0.005]	[-0.004]	[0.088]	[0.085]	[0.098]	[0.088]	[0.085]	[0.098]

The bracketed ([.]) quantities refer to the corresponding estimated quantities for the BR estimator, when all the samples are included.

Figure 5.4: Histograms of the values of the BR estimator of β ($\hat{\beta}_{BR}$), under the parameter setting B, when only the un-separated samples are included in the study and when all the samples are included.



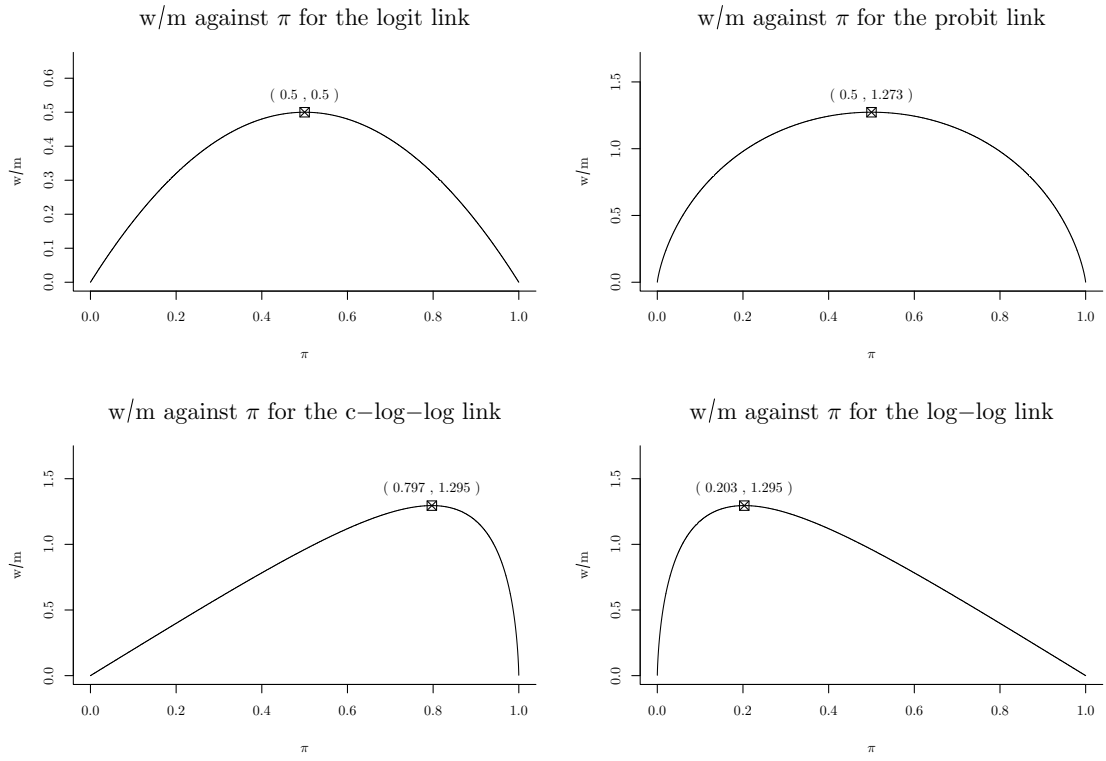
5.2.5 Do the fitted probabilities always shrink towards the point where the Jeffreys prior is maximized?

The general answer is no. The point where the Jeffreys prior for binary response GLMs is maximized is exactly the same point where the working weights are simultaneously maximized. This is directly proved if we use the working weights that correspond to any given model and follow exactly the same steps as in the proof of Theorem 4.2.2.

In Figure 5.5 we plot the working weights divided by the binomial totals for the logistic, probit, c-log-log and log-log links and indicate their maxima. For the logistic and probit link the working weight is maximized for $\pi = 0.5$ which is the same point towards which $\hat{\pi}_{\text{BR}}$ shrinks. However, for the c-log-log and log-log links the maxima are attained at different probabilities to the ones indicated in Table 5.3.

Given that the Jeffreys prior is concave, if we used it to penalize the likelihood and obtained the maximum penalized likelihood (MPL) estimates, the fitted probabilities would

Figure 5.5: Average working weight w/m (see Table 3.2) against the probability of success π for the logit, probit, c-log-log and log-log links.



shrink towards the points indicated in Figure 5.5. Moreover, as with the BR estimates, the MPL estimates are always finite and this can be shown by the steps in the proof of Theorem 4.2.1, if the logistic working weights are replaced with the working weights that correspond to each link function.

Because of the finiteness and shrinkage properties of the MPL estimator, it appears that the latter could also yield an improvement over ML. However, the leading term in the expansion of the bias of the MPL estimator is of order $\mathcal{O}(m^{-1})$ and thus, asymptotically, it cannot outperform the BR estimator in terms of first-order bias. Despite the fact that the penalization of the likelihood by the Jeffreys prior in curved families is not supported by any bias-related asymptotic results, the finiteness and shrinkage properties of the resultant estimator motivate its further comparison with the BR estimator. This is intended for further work and will not be pursued here.

5.2.6 Discussion and further work

It has been shown that the implementation of the bias-reduction method for binomial-response models can be easily performed by the use of available ML estimation software through pseudo-data representations. The estimation bears only a small extra cost in computation when compared to ML. Further, it has been demonstrated that the finiteness and shrinkage properties of the BR estimator extend beyond the case of canonical-links.

Certainly, further work is needed on the formalization of the finiteness and shrinkage properties of the bias-reduced estimator. However, in addition to the presented results in this section, we have not encountered empirical results that contradict these properties and it seems that the superiority of the BR estimator over the ML and BC estimators extends beyond canonical links.

As already mentioned in Subsection 5.2.3, since the BR estimator has asymptotically a normal distribution (see Section 3.5), the usual Wald-type intervals could be used for statements on the uncertainty of the obtained estimates. However, such intervals would illustrate poor coverage properties and in many cases the coverage probabilities would be far below the nominal level. For logistic regressions (see Section 4.4), confidence intervals based on both the likelihood ratio and penalized-likelihood ratio were constructed, further improving the suggestions of Heinze & Schemper (2002) and Bull et al. (2007). However, as Theorem 3.7.1 shows, there is no penalized likelihood corresponding to non-canonical models within the class of GLMs. Thus, for the BR estimates, there is no argument motivating the use of confidence intervals based on penalization of the likelihood by Jeffreys prior. In this perspective, it could be argued that the MPL estimator of Subsection 5.2.5 has an advantage over the BR estimator, since, as is done in the case of logistic regression, its definition motivates the use of such confidence intervals. Lastly, as in the case of logistic regression, a formal framework for measuring the goodness of fit is still lacking. All of the aforementioned issues could form subjects for future work in the area.

5.3 Non-linear Rasch models

In this section, we illustrate how the properties of the bias-reduction method for logistic regression extend to the case of the non-linear predictor "2-parameter logistic" (2PL) model (Birnbaum, 1968), well known in the item response theory literature. It is a generalized non-linear model in canonical parameterization (see Chapter 2). Similarly to logistic regression, this model has infinite ML estimates with positive probability and the ML estimator exhibits considerable bias under deviations from the origin of the logistic scale. Apart from its methodological importance in item response theory, our interest on 2PL models stems from the special form of the predictor which despite being a non-linear function of the parameters, is connected with the well-studied — in Chapter 4 — case of linear predictors.

5.3.1 The 1PL and 2PL models, and partial linearity

The well-known in item response theory Rasch (or 1PL) model has the simple, linear on the logistic scale, form

$$\log \frac{\pi_{rs}}{1-\pi_{rs}} = \eta_{rs} = \alpha_r + \gamma_s \,, \tag{5.8}$$

with $r = 1, \ldots, N$ and $s = 1, \ldots, n$. Here π_{rs} can be thought as the probability that the person s *succeeds* in item r of an achievement test and α_r, γ_s are unknown fixed parameters. In the framework of item response theory, parameter γ_s is interpreted as a measure of the ability of person s when taking an achievement test and α_r (or $-\alpha_r$) corresponds to a measure of the ease (or difficulty) of the item r of the test.

A much-used extension is the 2PL model that has the form

$$\log \frac{\pi_{rs}}{1-\pi_{rs}} = \tilde{\eta}_{rs} = \alpha_r + \beta_r \gamma_s \,, \tag{5.9}$$

where α_r, β_r and γ_s are unknown fixed parameters. The parameter β_r is usually referred to as the "discrimination parameter" of item r. The larger the discrimination or slope parameter, the steeper is the item response function (IRF), which is the curve that maps γ_s to π_{rs}. In contrast to 1PL, the log-odds η_{rs} is assumed to be a non-linear function of the parameters. Specifically, η_{rs} is a 'partially-linear' combination of parameters; if we fix either β_r's or γ_s's the result is a representation of the 1PL or a simple logistic regression model, respectively.

These models can be approached from three different perspectives: i) parameters of interest are only person-specific, so that all the item parameters are nuisances; ii) parameters of interest are item-specific, and so the person parameters are nuisances (see Hoijtink & Boomsma, 1995; Molenaar, 1995, respectively, for elegant reviews on these two perspectives for the 1PL model); iii) all parameters are of interest. Further, depending on the perspective, they can be fitted using either ML for all the parameters (joint ML) or conditional ML or marginal ML. Tuerlinckx et al. (2005) gives a review of these methods. Here we only consider joint ML.

Both models as given above are overparameterized and we have to pose certain restrictions on the parameter space in order to be able to identify the inferences made based on the fitted model. For example, in the case of the 2PL model the log-odds of success is unchanged if we multiply the discrimination parameter by a constant c and divide the ability parameter by c. Various constraints can be used for achieving identifiability in the 2PL models, such as, fixing the abilities of two persons, forcing the abilities to have zero mean and variance one, etc. Actually, there is an infinity of such choices and all are equivalent in terms of the fitted probabilities. However, the choice of an appropriate set of constraints is always discussed in item response theory literature (see, for example, the aforementioned studies in Hoijtink & Boomsma, 1995; Molenaar, 1995). [1] For the sake of generality, the

[1] An alternative way of fitting and inference on the fitted parameters is via generalized inverses and quasi-standard errors (see Firth, 2003; Firth & De Menezes, 2004), where no restrictions need to be imposed on the parameters. These methods are implemented in the `gnm` library in *R language* (R Development Core Team, 2007).

modified score functions are given for all the parameters involved in the model. Then, any restrictions on the parameter space are directly passed as constraints to the optimization problem of finding the roots of the system of the modified score equations.

5.3.2 The earlier work of Warm (1989)

Four years earlier than Firth (1993), Warm (1989) working on a item response theory model with a single person parameter derived the bias-reduction method using modifications based on the expected information. The motivation for the derivation was the reduction of the bias of the estimator. The model considered in Warm (1989) refers only to a single person and has the form

$$\pi_r = c_r + \frac{1 - c_r}{1 + \exp(-1.7a_r(\theta - b_r))} \quad (r = 1, \ldots, N) \ ,$$

where π_r is the probability that the person succeeds on item r, a_r, b_r and c_r are some item specific constants that are known from the context and θ is the unknown person parameter to be estimated. The starting point of Warm (1989) is a conjecture of the form the modified scores should have, and he succeeds in proving that a first-order unbiased estimator of θ can be obtained by locating the roots of

$$U^*(\theta) = U(\theta) + \frac{\mathrm{E}\left[U(\theta)^3\right] - \mathrm{E}\left[U(\theta)I(\theta)\right]}{F(\theta)} \, ,$$

where $U(\theta)$, $F(\theta)$, $I(\theta)$ are the ordinary score, the Fisher information and the observed information on θ, respectively (see Warm (1989) for their explicit forms). This is exactly the modified score function that is derived in Section 3.3. Moreover Warm (1989) recognises that the BR estimates are always finite even in cases where the ML estimates are infinite. The term "weighted likelihood estimation" is used and it is noted that when $c_r = 0$ the modified scores correspond to penalization (or "weighting" in Warm (1989) terminology) of the ordinary likelihood by $\sqrt{F(\theta)}$, the Jeffreys prior. This latter case when extended to more than one person is exactly the application of the bias-reduction method to the 1PL model that we consider in the following subsection.

5.3.3 Bias reduction for the 1PL and 2PL models

Consider realizations y_{rs} of independent binomial random variables Y_{rs} with totals m_{rs}, $r = 1, \ldots, N$, $s = 1, \ldots, n$. Let $\delta = (\alpha^T, \gamma^T)^T$ be the $(N + n)$-vector of parameters of the 1PL model (5.8). This is a logistic regression model and directly from the results in Section 4.2 the modified scores take the form

$$U_t^{(1PL)} = \sum_{r=1}^{n} \left(y_{rs} + \frac{1}{2} h_{rss} - (m_{rs} + h_{rss})\pi_{rs} \right) z_{rst} \quad (t = 1, \ldots, N + n) \ . \tag{5.10}$$

In the above expression h_{rss} is the s-th diagonal element of the $n \times n$ projection matrix $H_r = Z_r F^{-1} Z_r^T \Sigma_r$ where $F = \sum_r Z_r^T \Sigma_r Z_r$ is the Fisher information on δ, Z_r is the

$n \times (N+n)$ matrix with elements $z_{rst} = \partial \eta_{rs}/\partial \delta_t$ which do not depend on δ, and Σ_r is a diagonal matrix with s-th diagonal element $\kappa_{2,rs} = \mathrm{Var}(Y_{rs}) = m_{rs}\pi_{rs}(1-\pi_{rs})$.

Now let $\tilde{\delta} = (\alpha^T, \beta^T, \gamma^T)^T$ be the $(2N+n)$-vector of parameters of the 2PL model (5.9). By (3.21) and using the modifications based on the Fisher information, the modified score functions have the form

$$U_t^* = U_t + \frac{1}{2}\sum_{r=1}^{N}\sum_{s=1}^{n} \tilde{h}_{rss}\frac{\kappa_{3,rs}}{\kappa_{2,rs}}\tilde{z}_{rst} \qquad (5.11)$$
$$+ \frac{1}{2}\sum_{r=1}^{N}\sum_{s=1}^{n} \kappa_{2,rs}\,\mathrm{trace}\left\{\tilde{F}^{-1}\mathcal{D}^2\left(\tilde{\eta}_{rs};\tilde{\delta}\right)\right\}\tilde{z}_{rst} \quad (t = 1,\ldots,2N+n)\ .$$

Here, $\tilde{h}_{rss}$ has the same interpretation as does h_{rss} in the 1PL model, with the difference that Z_r is replaced by $\tilde{Z}_r$, and $\kappa_{3,rs} = m_{rs}\pi_{rs}(1-\pi_{rs})(1-2\pi_{rs})$ is the third cumulant of Y_{rs}, $(r = 1,\ldots,N\ ;\ s = 1,\ldots,n)$. The matrix $\tilde{Z}_r$ is the $n \times (2N+n)$ matrix of first derivatives of $\tilde{\eta}_r = (\tilde{\eta}_{r1},\ldots,\tilde{\eta}_{rn})$ with respect to $\tilde{\delta}$ and has components $\tilde{z}_{rst} = \partial\tilde{\eta}_{rs}/\partial\tilde{\delta}_t$ which in this case depend on $\tilde{\delta}$. Further, $\tilde{F} = \sum_r \tilde{Z}_r^T\Sigma_r\tilde{Z}_r$ is the Fisher information on $\tilde{\delta}$ and $\mathcal{D}^2\left(\tilde{\eta}_{rs};\tilde{\delta}\right)$ is the $(2N+n)\times(2N+n)$ matrix of second derivatives of $\tilde{\eta}_{rs}$ with respect to the parameter vector $\tilde{\delta}$. By the results in Chapter 2, the ordinary scores U_t for the 2PL model have the form

$$U_t = \sum_{r=1}^{N}\sum_{s=1}^{n}(y_{rs} - m_{rs}\pi_{rs})\,\tilde{z}_{rst} \quad (t = 1,\ldots,p)$$

and so (5.11) could be written in a more condensed from as

$$U_t^* = \sum_{r=1}^{N}\sum_{s=1}^{n}\left(y_{rs} + \frac{1}{2}\tilde{h}_{rss}\frac{\kappa_{3,rs}}{\kappa_{2,rs}} + \frac{1}{2}\kappa_{2,rs}\,\mathrm{trace}\left\{\tilde{F}^{-1}\mathcal{D}^2\left(\tilde{\eta}_{rs};\tilde{\delta}\right)\right\} - m_{rs}\pi_{rs}\right)\tilde{z}_{rst}\,. \quad (5.12)$$

At first glance the above expression might seem unwieldy, but the special form of the predictor of the 2PL model (5.9) suggests that there is much structure to be exploited in the quantities involved. First, note that $\kappa_{3,rs}/\kappa_{2,rs} = 1-2\pi_{rs}$. In addition, by the form of the predictor $\tilde{\eta}_{rs}$ and the definition of $\mathcal{D}^2\left(\tilde{\eta}_{rs};\tilde{\delta}\right)$ (see Theorem B.1.2 in Appendix B) we have that

$$\mathcal{D}^2\left(\tilde{\eta}_{rs};\tilde{\delta}\right) = \begin{bmatrix} D_{rs}^{\alpha,\alpha} & D_{rs}^{\alpha,\beta} & D_{rs}^{\alpha,\gamma} \\ (D_{rs}^{\alpha,\beta})^T & D_{rs}^{\beta,\beta} & D_{rs}^{\beta,\gamma} \\ (D_{rs}^{\alpha,\gamma})^T & (D_{rs}^{\beta,\gamma})^T & D_{rs}^{\gamma,\gamma} \end{bmatrix} = \begin{bmatrix} 0_{N\times N} & 0_{N\times N} & 0_{N\times n} \\ 0_{N\times N} & 0_{N\times N} & D_{rs}^{\beta,\gamma} \\ 0_{n\times N} & (D_{rs}^{\beta,\gamma})^T & 0_{n\times n} \end{bmatrix},$$

with $r = 1,\ldots,N$ and $s = 1,\ldots,n$, where $0_{n\times N}$ is the $n \times N$ matrix of zeros and $D_{rs}^{\beta,\gamma}$ denotes the $N \times n$ matrix with elements $\partial^2\tilde{\eta}_{rs}/\partial\beta_u\partial\gamma_w$ $(u = 1,\ldots,N\ ;\ w = 1,\ldots,n)$. The direct calculation of these second derivatives gives that $D_{r,s}^{\beta,\gamma}$ is a matrix of zeros with only its (r,s)-th element equal to 1. Thus, by the symmetry of $\tilde{F}^{-1}$,

$$\mathrm{trace}\left\{\tilde{F}^{-1}\mathcal{D}^2\left(\tilde{\eta}_{rs};\tilde{\delta}\right)\right\} = 2c_{rs}\,,$$

where c_{rs} is the (r,s)-th component of the $N \times n$ sub-matrix of $\tilde{F}^{-1}$ with rows referring to β and columns referring to γ. In other words, c_{rs} is the asymptotic covariance of the ML (or BR) estimators of β_r and γ_s.

Substitution of the above results into (5.12) gives the modified scores for the 2PL model in the elegant form

$$U_t^{(2PL)} = \sum_{r=1}^{N}\sum_{s=1}^{n}\left(y_{rs} + \frac{1}{2}\tilde{h}_{rss} - (m_{rs} + \tilde{h}_{rss})\pi_{rs} + c_{rs}\kappa_{2,rs}\right)\tilde{z}_{rst}\,, \tag{5.13}$$

for $t = 1, \ldots, 2N + n$.

5.3.4 Comparison of $U_t^{(1PL)}$ and $U_t^{(2PL)}$

The term $c_{rs}\kappa_{2,rs}$ in $U_t^{(2PL)}$ reflects the aforementioned partial linearity of the predictor; fixing either β's or γ's, the 2PL model reduces to a simple logistic regression model and the extra term disappears. For example, if we fix β_r's, $\tilde{z}_{rst}$ does not depend on the parameters and $c_{rs}\kappa_{2,rs} = 0$, retrieving the form $U_t^{(1PL)}$.

5.3.5 Obtaining the bias-reduced estimates

The form of the modified scores (5.13) for the 2PL model suggests a pseudo-data representation for the responses and the totals, of the form

$$\begin{array}{ll} \text{Pseudo-successes} & y_{rs}^* = y_{rs} + \dfrac{1}{2}\tilde{h}_{rss} + m_{rs}c_{rs}\pi_{rs}(1-\pi_{rs})\,, \\ \text{Pseudo-totals} & m_{rs}^* = m_{rs} + \tilde{h}_{rss}\,, \end{array} \tag{5.14}$$

with $r = 1, \ldots, n$ and $s = 1, \ldots, N$. If we replace the actual responses and totals with their pseudo counterparts in the expression for the usual IWLS working variates for ML estimation and iterate, the BR estimates are obtained at convergence.

However, as in the case of binary response GLMs there is an alternative way of obtaining the BR estimates. Note that the representation (5.14) has the same undesirable behaviour as the crude pseudo-responses presented in Table 3.2 for GLMs; the value of the pseudo-response might violate the range of the original response and it is not necessarily smaller than the value of the pseudo-totals. As in the previous section for GLMs, if we re-express (5.14) so that $0 < y_{rs}^* < m_{rs}^*$, then already-implemented software for fitting 2PL models via joint ML could be used in order to obtain the BR estimates. A simple re-expression of (5.14) gives

$$\begin{array}{ll} \text{Pseudo-successes} & y_{rs}^* = y_{rs} + \dfrac{1}{2}\tilde{h}_{rss} + m_{rs}c_{rs}\pi_{rs}I(c_{rs} \geq 0)\,, \\ \text{Pseudo-totals} & m_{rs}^* = m_{rs} + \tilde{h}_{rss} + m_{rs}c_{rs}(\pi_{rs} - I(c_{rs} < 0))\,, \end{array} \tag{5.15}$$

Now, algorithm 5.2 could be used for the implementation of the bias-reduction method with the following minor adjustments. We replace the subscript r with the subscripts rs and the pseudo-data specification in step B.iv) is replaced by (5.15), where the pseudo-specifications are modified by the addition of the subscript $(j+1)$ to the quantities of

the left hand sides and (j) to the quantities of the right. In order to obtain the correct estimated standard errors, we should adjust the resultant modified working weights (binomial variances here) after convergence, precisely in the same way as is the description of algorithm (5.2) in the previous section.

An issue that could arise in fitting these models relates to the choice of starting values for the iterative process. Even in the case of ML estimation their choice is crucial because the log-likelihood surface exhibits flat regions away from the maximum. Good starting values can be provided through the `residSVD` function of the `gnm` library in the *R language* (R Development Core Team, 2007) which decomposes appropriately (in an item-person way, here) the residuals of a simpler model using a singular value decomposition. We recommend these starting values to be used for fitting the model in the initialization part of algorithm 5.2 (step A.iii)).

5.3.6 Finiteness of the bias-reduced estimator

The ML estimates for both 1PL and 2PL models can take infinite values, for example, when the persons are perfectly separated with respect to a specific item. In practice, this causes fitting procedures to fail to converge and the standard ML asymptotic theory cannot be applied, since the ML estimate is located on the boundary of the parameter space. As in the case of logistic regression (Lesaffre & Albert, 1989, §4), fitted probabilities very close to 0 or 1 and/or very large estimated standard errors during the fitting procedure reflect well such situations. In contrast, from our experience with various simulated datasets the BR estimates are finite even when the ML estimates are infinite. This is also illustrated in the results of the following empirical study.

5.3.7 Issues and considerations

In the derivation of the modified scores we considered the case where instead of a single Bernoulli trial we allow more than one independent trial at each person-item combination.

The more usual asymptotic framework that is used in item response theory for these models uses the persons as the units of information (Molenaar, 1995, §3.2). In that case, the joint ML estimator is inconsistent since the dimension of the parameter space increases with the number of persons that are included in the study.

In order to avoid the inconsistency of the ML estimator we assume that both n and N are fixed so that the information on the parameters grows with the totals m_{rs} of each person-item combination. This is a restrictive assumption for these models but it will enable us to perform a small empirical study for the comparison of the BR and the ML estimator under a common basis. The form of the modified scores is not affected by these considerations. However, outside this restrictive assumption, the BR estimator is inconsistent since there is a $\mathcal{O}(1)$ term in the asymptotic expansion of its bias that persists as the number of persons increases. This situation is not considered here and is the subject of further work.

5.3.8 A small empirical study

Consider a 5-person (A,B,C,D,E) and 3-item (1,2,3) design, where each person replies 20 times on each item. For identifiability reasons we set the abilities γ_A and γ_E to -2 and 2, respectively. The true probabilities of positive reply for each person-item combination are fixed as in Table 5.7 (see Table 5.8 for the corresponding parameter values).

Table 5.7: True probabilities for the simulation study for the comparison of the BR and ML estimators on 2PL models.

	Item 1	Item 2	Item 3
Person A	0.083	0.182	0.354
Person B	0.231	0.378	0.550
Person C	0.500	0.622	0.731
Person D	0.769	0.818	0.858
Person E	0.917	0.924	0.931

Under these probabilities, we simulate 2500 samples and for each sample we obtain the ML and the BR estimates. In Figure 5.6, the estimated IRFs based on the BR and the ML estimates are shown. The IRFs based on the BR estimates exhibit less variation than the IRFs based on the ML estimates, *shrinking* towards the true curve. Further, note the simulated cases of (quasi-) perfect separation of groups of persons (dark-grey curves). The ML estimates are infinite and consequently the IRFs are steepest. In contrast, the corresponding IRFs based on the BR estimates are slightly adjusted, thus avoiding any singularities. As an example consider the dark-grey and dotted curves that correspond to a separated case. The fitted probabilities for persons A and B on the first item are zero and one. More specifically, based on ML, the fitted probabilities on item 1 to four decimal places are (0.0000 0.1283, 0.4470, 0.7747, 1.0000) and based on BR, a clear shrinkage towards 0.5 is noted, with fitted probabilities (0.0017, 0.1760, 0.4615, 0.7424, 0.9877) for persons A, B, C, D, E, respectively.

In Table 5.8 we calculate the estimated biases and the estimated MSEs for the ML and BR estimators, based on a larger simulation (10^4 samples) where the separated configurations (162 in total) have been removed. This removal, favourable for the ML estimator, takes place in order to enable us to report non-infinite values for the corresponding estimated measures of dispersion. As expected, the estimated bias of the BR estimator is smaller. The same behaviour is noted for the estimated MSE and consequently, the BR estimator has, also, smaller estimated variance.

Figure 5.6: Estimated IRFs for the 3 items from a simulation study of size 2500 and true probabilities as in Table 5.7.

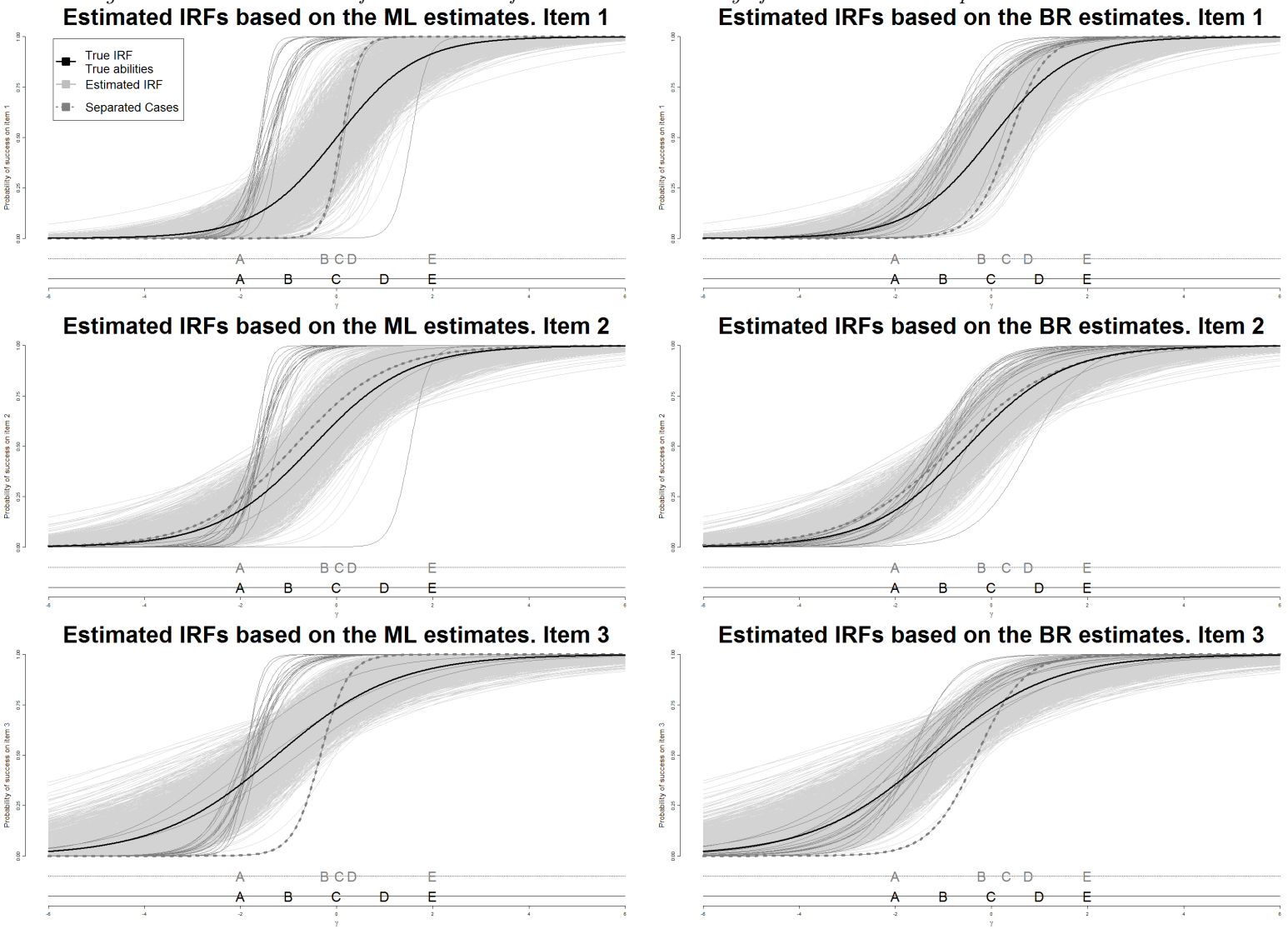


Table 5.8: Estimated bias and MSE to three decimal places, based on 10^4 simulated samples (162 separated samples were removed).

		α_1	α_2	α_3	β_1	β_2	β_3
True value		0	0.5	1	1.2	1	0.8
Est. Bias	ML	0.038	0.073	0.09	0.136	0.082	0.063
	BR	−0.009	−0.007	−0.003	0.003	−0.006	−0.001
Est. MSE	ML	0.359	0.245	0.201	0.16	0.099	0.074
	BR	0.188	0.137	0.121	0.068	0.055	0.047

		γ_B	γ_C	γ_D
True value		−1	0	1
Est. Bias	ML	−0.014	−0.022	−0.021
	BR	−0.004	0.008	0.01
Est. MSE	ML	0.193	0.19	0.314
	BR	0.174	0.163	0.254

5.3.9 Discussion and further work

We have considered the bias-reduction method for a different asymptotic framework than the one usually used in item response theory. However, according to the small empirical study above, the BR estimator outperforms the ML estimator in terms of estimated bias and estimated MSE. Furthermore, the BR estimator takes always finite values, even in cases where the persons are completely separated for a specific item (the ML estimates are infinite). In addition, the estimated probabilities shrink towards 0.5 when compared to the fitted probabilities based on the ML estimates. However, statements on the shrinkage of the estimates should be avoided in the case of 2PL models because the parameters are related to the log-odds in a non-linear way. The only thing to be noted is that the BR estimates are adjusted so that the estimated predictors η_{rs} $(r = 1, \ldots, n\ ;\ s = 1, \ldots, N)$ shrink towards zero.

In addition to the topics outlined in Subsection 5.2.6 for future work in binomial-response GLMs, there are more issues to deal with in item response theory models. The main topic is dealing with the inconsistency of the estimators when the information on the parameters grows with the number of persons. A different bias-reduction method has to be developed, that is robust to the increase of the dimension of the parameter space, as the number of persons increases. The target is the removal of the persistent $\mathcal{O}(1)$ term in the bias expansion of the ML estimator, and the parallel construction of consistent

estimators.

On the other hand, for conditional ML and marginal ML (Hoijtink & Boomsma, 1995; Molenaar, 1995), the consistency of the resultant estimators is ensured because both methods assume that information on the parameters grows with the number of persons (or items) and the parameters to be estimated are only item-specific (or person-specific), thus assuming fixed number of items (or persons). Hence, another possible direction for further research is the application of bias reduction to conditional and marginal ML estimation.

Chapter 6

Further topics: Additively modified scores

6.1 Introduction

Maximum likelihood (ML) is the dominant method of estimation in the frequentist school, mainly because of the neat asymptotic properties of the ML estimator (asymptotic normality, asymptotic sufficiency, unbiasedness and efficiency) and further the easy implementation of fitting procedures. Given the log-concavity of the likelihood function $L(\beta)$ on some parameterization β, estimation is performed by the solution of the efficient score equations $\nabla \log L(\beta) = 0$. However there are cases where the ML estimator has undesirable properties. Firth (1993), motivated partly from the reduction of the bias of the ML estimator for logistic regressions in small samples, developed a class of modifications to the efficient scores that result in first-order ($\mathcal{O}(n^{-1})$) unbiased estimators. As already described in Chapter 3, the core of his work lies on an additive modification to the original scores, that depends on the parameters and possibly on the data. By requiring first-order unbiasedness of the solution of the modified score equations, the appropriate form of the modifications is obtained from standard asymptotic expansions. By requiring a different asymptotic property from the solution of the additively modified scores, we can obtain estimators, for example, with smaller variance, smaller mean squared error, etc.

Despite the fact that this thesis deals with the aforementioned bias-reduction method, the results in this chapter refer generally to the asymptotic properties of an estimator resulted from the additive modification of the score functions with a $\mathcal{O}_p(1)$ quantity under repeated sampling. The asymptotic expressions that are given are interesting on their own right because they are derived in full generality and cover any situation where the usual regularity conditions are satisfied. In this way we are presented with several alternative directions for further work in the area.

For our purposes, the use of index notation and Einstein summation convention is necessary, and thus this chapter is recommended to be accompanied by Appendix A.

6.2 Additively modified score functions

Hereafter, we assume that all the necessary regularity conditions for likelihood inference (see Cox & Hinkley, 1974, Section 9.1) hold. Under the notation of Subsection A.4.1 in Appendix A, assume we modify the ordinary scores U_r according to

$$U_r^* = U_r + A_r \,,$$

where A_r is at least three times differentiable with respect to β and is allowed to depend on the data. Also, A_r is $\mathcal{O}_p(1)$ as $n \to \infty$. Let $\tilde{\beta}$ be the resultant estimator by the solution of the modified score equations $U_r^*(\tilde{\beta}) = 0$. In contrast, the ML estimator $\hat{\beta}$ is the solution of the ordinary score equations $U_r(\hat{\beta}) = 0$.

6.3 Consistency of $\tilde{\beta}$

The consistency of $\tilde{\beta}$ is a direct consequence of the consistency of the ML estimator and the fact that the adjustments to the scores have order $\mathcal{O}_p(1)$. Under the regularity conditions and the extra assumption that the ML estimate always exists and is unique within the parameter space B, it can be proved that $\hat{\beta}$ is a consistent estimator of β and that $\sqrt{n}(\hat{\beta} - \beta_0)$ is asymptotically normally distributed with zero mean and variance-covariance matrix the inverse of the Fisher information per observation evaluated at β_0, with β_0 the true but unknown parameter vector. So,

$$\hat{\beta}^r - \beta_0^r = \mathcal{O}_p(n^{-1/2}) \,. \tag{6.1}$$

Hence, if the difference of $\hat{\beta}$ from $\tilde{\beta}$ is smaller in order than $\mathcal{O}_p(n^{-1/2})$, then the bias-reduced estimator is also a consistent estimator of β_0.

Let $\epsilon^r = \tilde{\beta}^r - \hat{\beta}^r$. Since, by definition, $U_r^* = U_r + \mathcal{O}_p(1)$, we have that the modified scores, as defined in (6.2), are of order $\mathcal{O}_p(n^{1/2})$ and their derivatives with respect to the parameters are $\mathcal{O}_p(n)$, as their ordinary likelihood counterparts. A first-order Taylor expansion of the modified scores vector at $\tilde{\beta}$ around $\hat{\beta}$ gives

$$0 = U_r^*(\tilde{\beta}) \approx U_r^*(\hat{\beta}) + \epsilon^s U_{rs}^*(\hat{\beta}) \,,$$

with $U_{rs}^* = \partial U_r^* / \partial \beta^s$. Since $\hat{\beta}$ is the solution of the likelihood equations we have that $U_r(\hat{\beta}) = 0$ and thus $U_r^*(\hat{\beta}) = A_r(\hat{\beta})$. Re-expressing in terms of ϵ^r we get

$$\epsilon^r \approx A_s(\hat{\beta}) I^{*rs}(\hat{\beta}) \,,$$

where I^{*rs} is understood as the matrix inverse of $-U_{rs}^*$. By the above relation, we have that $\epsilon^r = \mathcal{O}_p(n^{-1})$ and combining with (6.1) we get

$$\tilde{\beta}^r - \beta_0^r = \mathcal{O}_p(n^{-1/2}) \,.$$

An application of Theorem A.4.2 with $a = 1/2$ and $t = 1/2$ gives $\tilde{\beta}^r - \beta_0^r = o_p(1)$ and the consistency of $\tilde{\beta}$ is derived.

The above derivation relies heavily on the assumption of existence and finiteness of the ML estimator in a compact subset of the parameter space B. However, there are cases, like logistic regression, where this assumption is not valid. In these cases, we could treat $\tilde{\beta}$ as a general "Z-estimator" and proceed according to the proofs in van der Vaart (1998, Section 5.2).

6.4 Expansion of $\tilde{\beta} - \beta_0$

Let $\delta = \tilde{\beta} - \beta_0$ and consider the expansion of U_r^* evaluated at $\tilde{\beta}$. By the consistency of $\tilde{\beta}$ we have

$$0 = U_r^*(\tilde{\beta}) = U_r^* + \delta^s U_{rs}^* + \frac{1}{2}\delta^s\delta^t U_{rst}^* + \frac{1}{6}\delta^s\delta^t\delta^u U_{rstu}^* + \mathcal{O}_p(n^{-1})\,,$$

where, hereafter, U_{rs}^*, U_{rst}^*, U_{rstu}^*, ... denote partial derivatives of U_r^*, evaluated at β_0. For the sake of compactness of the following expressions, let $\delta^{R_a} = \delta^{r_1}\ldots\delta^{r_a}$, $a = 1, 2, \ldots$. Also, A_{rs}, A_{rst}, A_{rstu} denote higher order derivatives of the $\mathcal{O}_p(1)$ modifications A_r and thus are also $\mathcal{O}_p(1)$. The above expansion is written as

$$0 = U_r + A_r + \delta^s U_{rs} + \delta^s A_{rs} + \frac{1}{2}\delta^{st} U_{rst} + \frac{1}{6}\delta^{stu} U_{rstu} + \mathcal{O}_p(n^{-1})\,.$$

The terms $\delta^{st}A_{rst}$ and $\delta^{stu}A_{rstu}$ are of order $\mathcal{O}_p(n^{-1})$ and $\mathcal{O}_p(n^{-3/2})$, respectively, and so they are incorporated in the $\mathcal{O}_p(n^{-1})$ remainder in the expansion above.

By the second Bartlett identity in (A.5) (see Appendix A), $\mu_{r,s} = -\mu_{rs}$ and so, for the matrix inverse of the Fisher information we have $\mu^{r,s} = -\mu^{rs}$. Re-expressing in terms of the centered log-likelihood derivatives $H_{R_a} = U_{R_a} - \mu_{R_a}$ and solving with respect to δ^r, we have that

$$\delta^r = U^r \mathbin{\dot{+}} A^r + H_s^r\delta^s + \frac{1}{2}\mu_{st}^r\delta^{st} \mathbin{\dot{+}} A_s^r\delta^s + \frac{1}{2}H_{st}^r\delta^{st} + \frac{1}{6}\mu_{stu}^r\delta^{stu} \mathbin{\dot{+}} \mathcal{O}_p(n^{-2})\,, \tag{6.2}$$

where $\mathbin{\dot{+}}$ denotes a drop in asymptotic order of $O_p(n^{-1/2})$ for the subsequent terms, and U^r, A^r, A_s^r, $H_{s_1\ldots s_a}^r$, $\mu_{s_1\ldots s_a}^r$ are the outcomes of the contraction of U_r, A_r, A_{rs}, $H_{rs_1\ldots s_a}$, $\mu_{rs_1\ldots s_a}$, respectively, with $\mu^{r,s}$. That is

$$\begin{aligned}
U^r &= \mu^{r,s}U_s = \mathcal{O}_p(n^{-1/2})\,,\\
A^r &= \mu^{r,s}A_s = \mathcal{O}_p(n^{-1})\,,\\
A_{s_1}^r &= \mu^{r,s}A_{ss_1} = \mathcal{O}_p(n^{-1})\,,\\
H_{s_1\ldots s_a}^r &= \mu^{r,s}H_{ss_1\ldots s_a} = \mathcal{O}_p(n^{-1/2})\,,\\
\mu_{s_1\ldots s_a}^r &= \mu^{r,s}\mu_{ss_1\ldots s_a} = \mathcal{O}_p(1)\,.
\end{aligned}$$

By iteratively substituting $\delta^r = U^r \mathbin{\dot{+}} A^r + H_s^r\delta^s + \frac{1}{2}\mu_{st}^r\delta^{st} \mathbin{\dot{+}} \mathcal{O}_p(n^{-3/2})$ in (6.2) (see Pace & Salvan, 1997, § 9.3.2, for a detailed description of the "iterative substitution method"

for stochastic Taylor expansions), we have

$$\begin{aligned}\delta^r = U^r \dot{+} A^r + H^r_s U^s + \frac{1}{2}\mu^r_{st}U^{st} \\ \dot{+} H^r_s A^s + \mu^r_{st}U^s A^t + A^r_s U^s + H^r_s H^s_t U^t + \mu^r_{st}H^s_u U^{tu} \\ + \frac{1}{2}H^r_s \mu^s_{tu}U^{tu} + \frac{1}{2}\mu^r_{st}\mu^s_{uv}U^{tuv} + \frac{1}{2}H^r_{st}U^{st} + \frac{1}{6}\mu^r_{stu}U^{stu} \dot{+} \mathcal{O}_p(n^{-2})\,,\end{aligned} \tag{6.3}$$

where $U^{r_1\dots r_a} = U^{r_1}\dots U^{r_a}$.

Here, we should mention that if we remove the terms depending on the modifications A_r and their derivatives A_{rs} from (6.3) and let $\delta = \hat{\beta} - \beta_0$, the resultant asymptotic expansion is exactly that of $\hat{\beta} - \beta_0$. This is a direct consequence of the fact that the modifications A_r come into the modified scores additively. The same would be true if the modified scores were dependant to the modifications through any affine transformation. However, this elegant correspondence would not be valid for more general non-additive relationships between the efficient scores and the modifications.

6.5 Asymptotic normality of $\tilde{\beta}$

The asymptotic normality of $\tilde{\beta}$ is directly apparent by (6.3). Specifically,

$$\sqrt{n}\delta^r = \sqrt{n}\mu^{r,s}U_s \dot{+} \mathcal{O}_p(n^{-1/2})$$

and using Theorem A.4.2 with $a = 1/2$ and $t = 1/2$ we have

$$\sqrt{n}\delta^r = \sqrt{n}\mu^{r,s}U_s + o_p(1)\,.$$

Thus, the application of the central limit theorem on the standardized score vector and of the Slutzky lemma (see van der Vaart, 1998, Section 2.1) gives that $\sqrt{n}(\tilde{\beta}^r - \beta^r)$ is asymptotically distributed according to a normal distribution with zero mean and variance-covariance matrix the inverse of the Fisher information per observation $\kappa^{r,s} = n\mu^{r,s}$.

6.6 Asymptotic bias of $\tilde{\beta}$

The asymptotic bias of $\tilde{\beta}$ is obtained by taking expectations in both sides of (6.3) and by using rule (A.9) for the asymptotic orders of the expectation of each term. After the exploitation of the contractions, an elementary algebraic manipulation gives

$$\begin{aligned}\mathrm{E}(\delta^r) &= \mu^{r,s}\mathrm{E}(A_s) + \frac{1}{2}\mu^{r,s}\mu^{t,u}\left(2\mu_{st,u} + \mu_{stu}\right) \ddot{+} \mathcal{O}(n^{-3/2}) \\ &= \mu^{r,s}\mathrm{E}(A_s) - \frac{1}{2}\mu^{r,s}\mu^{t,u}\left(\mu_{s,tu} + \mu_{s,t,u}\right) \ddot{+} \mathcal{O}(n^{-3/2})\,,\end{aligned} \tag{6.4}$$

where $\ddot{+}$ denotes a drop in asymptotic order for $\mathcal{O}_p(n^{-1})$ for the subsequent terms. This expression for the asymptotic bias is the basis for the results in Firth (1993); note that when

$$\mu^{r,s}\mathrm{E}(A_s) - \frac{1}{2}\mu^{r,s}\mu^{t,u}\left(\mu_{s,tu} + \mu_{s,t,u}\right) = \mathcal{O}(n^{-3/2})\,,$$

the resultant estimator has bias of order smaller than $\mathcal{O}(n^{-1})$ and hence a class of such estimators is obtained from the location of the roots of the modified scores using one of the following modifications

$$
\begin{aligned}
A_r \equiv A_r^{(E)} &= \frac{1}{2}\left(\mu_{r,s} + R_{rs}\right)\mu^{s,t}\mu^{u,v}\left(\mu_{t,uv} + \mu_{t,u,v}\right) + \bar{R}_r\,, \\
A_r \equiv A_r^{(O)} &= \frac{1}{2}\left(-U_{rs} + R_{rs}\right)\mu^{s,t}\mu^{u,v}\left(\mu_{t,uv} + \mu_{t,u,v}\right) + \bar{R}_r\,, \\
A_r \equiv A_r^{(S)} &= \frac{1}{2}\left(U_r U_s + R_{rs}\right)\mu^{s,t}\mu^{u,v}\left(\mu_{t,uv} + \mu_{t,u,v}\right) + \bar{R}_r\,,
\end{aligned}
$$

with R_{rs} and $\bar{R}_r$ any quantities that depend on the data and the parameters and have expectations of order at most $\mathcal{O}(n^{1/2})$ and at most $\mathcal{O}(n^{-1/2})$, respectively. Substituting any of the above modifications back to (6.3) and taking expectations in both sides, we conclude that the estimator which results from the solution of the modified score equations $U_r^* = 0$, has bias of order $o(n^{-1})$.

6.7 Asymptotic mean-squared error of $\tilde{\beta}$

Directly from its definition, the variance of an estimator $\tilde{\beta}$ of a scalar parameter β can be expressed as

$$
\mathrm{Var}(\tilde{\beta}) = \mathrm{E}\left\{(\tilde{\beta} - \beta_0)^2\right\} - \left[\mathrm{E}(\tilde{\beta} - \beta_0)\right]^2\,,
$$

where $\mathrm{E}\{(\tilde{\beta} - \beta_0)^2\}$ is the mean squared error (MSE) of $\tilde{\beta}$ and $\mathrm{E}(\tilde{\beta} - \beta_0)$ is its bias. An analogous expression can be obtained in the case of a multi-dimensional target parameter β_0. Denoting V^{rs} the variance-covariance matrix of $\tilde{\beta}$ we have

$$
\begin{aligned}
V^{rs} &= \mathrm{E}\left\{(\tilde{\beta}^r - \mathrm{E}(\tilde{\beta}^r))(\tilde{\beta}^s - \mathrm{E}(\tilde{\beta}^s))\right\} \\
&= \mathrm{E}\left\{(\tilde{\beta}^r - \mathrm{E}(\tilde{\beta}^r) + \beta_0^r - \beta_0^r)(\tilde{\beta}^s - \mathrm{E}(\tilde{\beta}^s) + \beta_0^s - \beta_0^s)\right\} \\
&= \mathrm{E}\left(\delta^r\delta^s\right) - \mathrm{E}(\delta^r)\mathrm{E}(\delta^s)\,,
\end{aligned}
\tag{6.5}
$$

with $\delta^r = \tilde{\beta}^r - \beta_0^r$. The matrix $\mathrm{E}\left(\delta^r\delta^s\right)$ is the generalization of the MSE in the case of more than one parameters. In the previous section we obtained an asymptotic expression for $\mathrm{E}(\delta^r)$ and in order to obtain the asymptotic expression of V^{rs} we first need to consider the asymptotic expression for $\mathrm{E}\left(\delta^r\delta^s\right)$ and, as we do in the next section, substitute in the identity (6.5).

By (6.3) and the symmetry of $\delta^r\delta^s$, some algebraic manipulation gives

$$
\begin{aligned}
\delta^r\delta^s = U^{rs} &\dot{+} 2H_t^r U^{st} + 2U^r A^s + \mu_{tu}^r U^{stu} \\
&\dot{+} 2H_t^r U^s A^t + 2H_t^r A^s U^t + 2\mu_{tu}^r U^{st} A^u + \mu_{tu}^r A^s U^{tu} + 2A_t^r U^{st} + 2H_t^r H_u^t U^{su} \\
&+ H_t^r H_u^s U^{tu} + 2\mu_{tu}^r H_v^t U^{suv} + \mu_{tu}^r H_v^s U^{tuv} + H_t^r \mu_{uv}^t U^{suv} + \mu_{tu}^r \mu_{vw}^t U^{suvw} \\
&+ \mu_{tu}^r \mu_{vw}^s U^{tuvw}/4 + H_{tu}^r U^{stu} + \mu_{tuv}^r U^{stuv}/3 + A^r A^s \dot{+} \mathcal{O}_p(n^{-5/2})\,.
\end{aligned}
$$

After exploiting the contractions and taking expectations in both sides above (see Example A.4.1 in Appendix A) we have

$$\begin{aligned}\mathrm{E}(\delta^r\delta^s) = \mu^{r,s} \ddot{+} \mu^{r,v}\mu^{s,w}\mu^{t,y}\{&2\mu_{vt,w,y} + 2\mu_{vt,wy} + (\mu_{vt,wy} - \mu_{vt}\mu_{wy}) + \mu_{vty,w} \\ &+ 2\mu_{vwt,y} + \mu_{vwty} \\ &+ 4\mu_{w,y}\mathrm{E}(U_vA_t) + 2\mu_{w,y}\mathrm{E}(U_tA_v) + \mu_{w,y}\mathrm{E}(A_vA_t) \\ &+ 2\mathrm{E}(A_yU_{vt}U_w) + 2\mathrm{E}(A_wU_{vt}U_y) + 2\mathrm{E}(A_{vt}U_wU_y)\} \\ + \mu^{r,v}\mu^{s,w}\mu^{t,y}\mu^{u,z}\{&2\mu_{vtu}\mathrm{E}(U_wU_yA_z) + \mu_{vtu}\mathrm{E}(U_yU_zA_w) \\ &+ \mu_{vt,z}(2\mu_{yu,w} + \mu_{wu,y}) + \mu_{wyu}(4\mu_{vt,z} + \mu_{v,tz}) \\ &+ (\mu_{vwt} + \mu_{w,vt})(2\mu_{yu,z} + \mu_{yuz}) + 3\mu_{vtu}\mu_{wyz}/2 \\ &+ \mu_{vt,y}\mu_{wu,z} + \mu_{vty}\mu_{wu,z} + \mu_{vty}\mu_{wuz}/4\} \ddot{+} \mathcal{O}(n^{-3})\end{aligned} \tag{6.6}$$

For the special case where the modifications A_r do not depend on the data, expression (6.6) is considerably simplified, taking the form

$$\begin{aligned}\mathrm{E}(\delta^r\delta^s) = \mu^{r,s} \ddot{+} \mu^{r,v}\mu^{s,w}\mu^{t,y}\{&2\mu_{vt,w,y} + 2\mu_{vt,wy} + (\mu_{vt,wy} - \mu_{vt}\mu_{wy}) + \mu_{vty,w} \\ &+ 2\mu_{vwt,y} + \mu_{vwty} \\ &+ \mu_{w,y}A_vA_t + 2A_y\mu_{vt,w} + 2A_w\mu_{vt,y} + 2A_{vt}\mu_{w,y} \\ &+ 2A_y\mu_{vwt} + A_w\mu_{vty}\} \\ + \mu^{r,v}\mu^{s,w}\mu^{t,y}\mu^{u,z}\{&\mu_{vt,z}(2\mu_{yu,w} + \mu_{wu,y}) + \mu_{wyu}(4\mu_{vt,z} + \mu_{v,tz}) \\ &+ (\mu_{vwt} + \mu_{w,vt})(2\mu_{yu,z} + \mu_{yuz}) + 3\mu_{vtu}\mu_{wyz}/2 \\ &+ \mu_{vt,y}\mu_{wu,z} + \mu_{vty}\mu_{wu,z} + \mu_{vty}\mu_{wuz}/4\} \ddot{+} \mathcal{O}(n^{-3})\end{aligned}$$

Both for the above expression and for (6.6), if we remove the terms depending on the modifications A_r and the derivatives of the modifications A_{rs}, and let $\delta^r = \hat{\beta} - \beta_0$, we obtain the corresponding expressions for the ML estimator.

In the case of a scalar target parameter β_0 and letting $U_k = \partial^k l(\beta)/\partial\beta^k$, $\mu_k = \mathrm{E}(U_k)$, $\mu_{k,m} = \mathrm{E}(U_kU_m)$ and so on, (6.6) is written in the form

$$\begin{aligned}\mathrm{E}(\delta^2) = F^{-1} \ddot{+} F^{-3}\{&2\mu_{1,1,2} + 2\mu_{2,2} + (\mu_{2,2} - F^2) + 3\mu_{3,1} + \mu_4 \\ &+ 6F\mathrm{E}(U_1A) + F\mathrm{E}(A^2) + 4\mathrm{E}(AU_1U_2) + 2\mathrm{E}(\dot{A}l_1^2)\} \\ + F^{-4}\{&3\mu_3\mathrm{E}(U_1^2A) + 6\mu_{2,1}^2 + 9\mu_3\mu_{2,1} + 11\mu_3^2/4\} \ddot{+} \mathcal{O}(n^{-3}),\end{aligned} \tag{6.7}$$

where $F = \mu_{1,1}$ is the Fisher information and $\dot{A}$ the derivative of the modification evaluated at β_0. If A does not depend on the data the corresponding expression is

$$\begin{aligned}\mathrm{E}(\delta^2) = F^{-1} \ddot{+} F^{-3}\{&2\mu_{1,1,2} + 2\mu_{2,2} + (\mu_{2,2} - F^2) + 3\mu_{3,1} + \mu_4 \\ &+ FA^2 + 4A\mu_{1,2} + 2F\dot{A} + 3\mu_3A\} \\ + F^{-4}\{&6\mu_{2,1}^2 + 9\mu_3\mu_{2,1} + 11\mu_3^2/4\} \ddot{+} \mathcal{O}(n^{-3}).\end{aligned}$$

All of the above expressions are considerably simplified in the case of exponential families in canonical parameterization, because $U_{R_a} = \mu_{R_a}$ for $a > 1$ and so $\mu_{R_{a_1},\ldots,R_{a_k},s} = 0$

for $a_1, \ldots, a_k > 1$. For example, in the case of a single parameter exponential family in canonical parameterization, $\mu_{1,1,2} = -\mu_{2,2}^2 = -F^2$ and $\mu_{3,1} = \mu_{1,2} = 0$, so that the above expression reduces to

$$\mathrm{E}(\delta^2) = F^{-1} \ddot{+} F^{-3}\{\mu_4 + FA^2 + 2F\dot{A} + 3\mu_3 A\} + 11F^{-4}\mu_3^2/4 \ddot{+} \mathcal{O}(n^{-3})\,, \tag{6.8}$$

Note that we can substitute the bias-reducing modifications A_r in the derived expressions in order to check the effect of bias-reduction to the mean squared error of the resultant estimator. This has been done in Section 4.2 for the estimation of the log-odds from the realization of a single binomial random variable.

Also, these expressions can be directly used to produce MSE-reducing modifications by working on a similar way as in the previous section. For example, in the case of a single parameter exponential family in canonical parameterization, the second-order term in (6.8) is eliminated if

$$F^{-3}\{\mu_4 + FA^2 + 2F\dot{A} + 3\mu_3 A\} + 11F^{-4}\mu_3^2/4 = \mathcal{O}(n^{-3})$$

or alternatively if we could find A that satisfies the differential equation

$$\dot{A} = -\frac{2F(\mu_4 + 3\mu_3 A) + 2F^2A^2 + 11\mu_3}{4F^2} \ddot{+} \mathcal{O}(n^{-1})\,.$$

6.8 Asymptotic variance of $\tilde{\beta}$

The asymptotic expression for the variance of $\tilde{\beta}$ results from the direct substitution of (6.6) and (6.4) into (6.5). After some rearrangement we have

$$\begin{aligned} V^{rs} = \mu^{r,s} \ddot{+} \mu^{r,v}\mu^{s,w}\mu^{t,y}\{&2\mu_{vt,w,y} + 2\mu_{vt,wy} + (\mu_{vt,wy} - \mu_{vt}\mu_{wy}) + \mu_{vty,w} \\ &+ 2\mu_{vwt,y} + \mu_{vwty} \\ &+ 4\mu_{w,y}\mathrm{E}(U_vA_t) + 2\mu_{w,y}\mathrm{E}(U_tA_v) + \mu_{w,y}\mathrm{E}(A_vA_t) \\ &+ 2\mathrm{E}(A_yU_{vt}U_w) + 2\mathrm{E}(A_wU_{vt}U_y) + 2\mathrm{E}(A_{vt}U_wU_y) \\ &- (2\mu_{wt,y} + \mu_{wty})\mathrm{E}(A_v) - \mu_{w,y}\mathrm{E}(A_vA_t)\} \\ + \mu^{r,v}\mu^{s,w}\mu^{t,y}\mu^{u,z}\{&2\mu_{vtu}\mathrm{E}(U_wU_yA_z) + \mu_{vtu}\mathrm{E}(U_yU_zA_w) \\ &+ \mu_{vt,z}(2\mu_{yu,w} + \mu_{wu,y}) + \mu_{wyu}(4\mu_{vt,z} + \mu_{v,tz}) \\ &+ (\mu_{vwt} + \mu_{w,vt})(2\mu_{yu,z} + \mu_{yuz}) + 3\mu_{vtu}\mu_{wyz}/2\} \ddot{+} \mathcal{O}(n^{-3})\,. \end{aligned} \tag{6.9}$$

When the modifications A_r do not depend on the data, expression (6.9) takes the form

$$\begin{aligned} V^{rs} = \mu^{r,s} \ddot{+} \mu^{r,v}\mu^{s,w}\mu^{t,y}\{&2\mu_{vt,w,y} + 2\mu_{vt,wy} + (\mu_{vt,wy} - \mu_{vt}\mu_{wy}) + \mu_{vty,w} \\ &+ 2\mu_{vwt,y} + \mu_{vwty} \\ &+ \mu_{w,y}A_vA_t + 2A_y\mu_{vt,w} + 2A_w\mu_{vt,y} + 2A_{vt}\mu_{w,y} \\ &+ 2A_y\mu_{vwt} + A_w\mu_{vty} \\ &- (2\mu_{wt,y} + \mu_{wty})A_v - \mu_{w,y}A_vA_t\} \\ \ldots \end{aligned}$$

$$\ldots + \mu^{r,v}\mu^{s,w}\mu^{t,y}\mu^{u,z}\{\mu_{vt,z}(2\mu_{yu,w} + \mu_{wu,y}) + \mu_{wyu}(4\mu_{vt,z} + \mu_{v,tz}) + (\mu_{vwt} + \mu_{w,vt})(2\mu_{yu,z} + \mu_{yuz}) + 3\mu_{vtu}\mu_{wyz}/2 \ddot{+} \mathcal{O}(n^{-3})\,.$$

For both expressions above, removal of the terms depending on the modifications results in the asymptotic expression for the variance-covariance matrix of $\hat{\beta}$, up to and including the $\mathcal{O}(n^{-2})$ terms. This expression is given in Peers & Iqbal (1985).

In the case of a scalar parameter, the variance of $\tilde{\beta}$ is given by the asymptotic expression

$$\begin{aligned}\mathrm{Var}(\tilde{\beta}) = F^{-1} \ddot{+} F^{-3}\{&2\mu_{1,1,2} + 2\mu_{2,2} + (\mu_{2,2} - F^2) + 3\mu_{3,1} + \mu_4 \\ &+ 6F\mathrm{E}(U_1A) + 4\mathrm{E}(AU_1U_2) + 2\mathrm{E}(\dot{A}U_1^2) - 2\mu_{2,1}\mathrm{E}(A) - \mu_3\mathrm{E}(A)\} \\ &F^{-4}\{4\mu_{2,1}^2 + 8\mu_3\mu_{2,1} + 5\mu_3^2/2 + 3\mu_3\mathrm{E}(U_1^2A)\} \ddot{+} \mathcal{O}(n^{-3})\,,\end{aligned} \tag{6.10}$$

and in the case of modifications that do not depend on the data

$$\begin{aligned}\mathrm{Var}(\tilde{\beta}) = F^{-1} \ddot{+} F^{-3}\{&2\mu_{1,1,2} + 2\mu_{2,2} + (\mu_{2,2} - F^2) + 3\mu_{3,1} + \mu_4 \\ &+ 4A\mu_{1,2} + 2F\dot{A} - 2\mu_{2,1}A + 2\mu_3A\} \\ &F^{-4}\{4\mu_{2,1}^2 + 8\mu_3\mu_{2,1} + 5\mu_3^2/2\} \ddot{+} \mathcal{O}(n^{-3})\,.\end{aligned}$$

All of the above asymptotic expressions can be used in the same manner as the expressions for the MSE in the previous section, for producing estimators with smaller variance.

6.9 General remarks

In this short chapter, we have given the asymptotic expressions for the bias, the variance and the MSE of an estimator resulted from the solution of the modified score equations

$$U_r^* = U_r + A_r = 0\,,$$

where A_r is an arbitrary function of the parameters and possibly the data, and is $\mathcal{O}(1)$ as $n \to \infty$. The expressions have been given in full generality and they provide a baseline for future work towards the construction of estimators that have improved properties relative to the traditional ML estimator. The most interesting direction seems the use of the MSE expressions, because MSE is a measure that incorporates the trade-off between bias and variance.

Chapter 7

Final remarks

7.1 Summary of the thesis

The modified-score functions approach to bias reduction (Firth, 1993) is continually gaining in popularity (e.g. Mehrabi & Matthews, 1995; Pettitt et al., 1998; Heinze & Schemper, 2002; Bull et al., 2002; Zorn, 2005; Sartori, 2006; Bull et al., 2007), because of the superior properties of the bias-reduced (BR) estimator over the traditional maximum likelihood (ML) estimator, particularly in models for categorical responses. Most of the activity has been noted for logistic regressions where, as the empirical studies in Heinze & Schemper (2002) and Bull et al. (2002) illustrated, bias reduction has a clear shrinkage interpretation and the BR estimates are always finite. Furthermore, the implementation of the bias-reduction method is greatly facilitated by the fact that logistic regressions are flat exponential families and as shown in Firth (1993), the bias-reduction method neatly corresponds to the penalization of the ordinary likelihood by Jeffreys invariant prior.

The current thesis has been mainly motivated by the recent applied and methodological interest in the bias-reduction method and we aimed to widen the applicability of the method, identifying cases where bias reduction is beneficial. Our target has been threefold:

i) To explore the nature of the bias-reducing modifications to the efficient scores and to obtain results that facilitate the application and the theoretical assessment of the bias-reduction method.

ii) To establish theoretically that the bias-reduction method should be considered as an improvement over traditional ML for logistic regressions.

iii) To deviate from the flat exponential family and explore the effect of bias reduction to some commonly used curved families for categorical responses.

For target i), we have dedicated Chapter 3. We have given the form of the general family of modifications that result in modified score equations, which in turn result in estimators with bias of order $o(n^{-1})$. This family includes as special members the proposals in

Firth (1993), namely the simplest case of modifications based on the expected information and the slightly more elaborate modifications based on the observed information. The class of modifications that is given in Chapter 3 has the basic property that every member of it can be written as a shifted weighted sum of the modifications based on the expected information, shifted by a quantity that has expectation of order at most $\mathcal{O}(n^{-1/2})$. Next, we have shown that, in contrast to flat exponential families, the existence of a penalized likelihood corresponding to the modified scores for general families is not generally guaranteed, even in the simplest case of modifications based on the expected information. In this direction, we have derived a necessary and sufficient condition that matches the level of generality of the applicability of the bias-reduction method (all regular problems). The validity of this condition guarantees the existence of a penalized likelihood, and vice versa. It has also been shown that under the consistency of the ML estimator, the consistency and asymptotic normality of the BR estimator is guaranteed. In this way, standard techniques based on the same asymptotic properties for the ML estimator, could be used also for the BR estimator.

Despite the theoretical importance of the above results, the core of Chapter 3 is the derivation of explicit formulae for the modified score vector for the wide class of exponential family non-linear models with known dispersion (reviewed in Chapter 2), which includes as special cases both univariate and multivariate GLMs, as well as more general non-linear regressions in which the variance has a specified relationship with the mean. The formulae that have been derived involve quantities that are readily available when the model to be used has been specified and usually result from the output of standard computing packages. These expressions can be used directly for the implementation of the bias-reduction method or even theoretically in order to gain insight into the nature of the modifications in any specific application. We have focused on univariate GLMs, where despite the fact that penalized likelihoods exist only for canonical links, we have shown that the implementation of the bias-reduction method can be achieved by a modified iterative re-weighted least squares procedure (IWLS) with the simple subtraction of the Cordeiro & McCullagh (1991) ξ-quantities from the usual ML working variates. In this way we have generalized the modified IWLS procedure in Firth (1992a,b) for canonically linked-models and showed the neat connections with Cordeiro & McCullagh (1991).

Chapter 4 achieves target ii). For logistic regressions, we have theoretically shown that the BR estimates are always finite, even in separated cases where the ML estimates are infinite. Moreover, we have formally verified the shrinkage properties of the BR estimator and have shown that shrinkage takes place according to a metric based on the Fisher information rather than the usual Euclidean distance. We have demonstrated that the Heinze & Schemper (2002) and Bull et al. (2007) profile penalized-likelihood ratio (PLR) confidence interval can have poor coverage properties, particularly for extreme parameter values. In view of this, for assessing the uncertainty on the BR estimates, we propose a conservative alternative to the PLR confidence interval, which preserves the good coverage properties of the PLR interval for moderate parameter values, and approaches coverage 1 as the true parameter value tends to ∞. Furthermore, as far as multinomial responses are concerned, we have derived the modified iterative generalized least squares (IGLS)

procedure that produces the BR estimates by simply subtracting appropriate quantities from the ML working variates at each iteration.

In this way, previous published work on logistic regression has been rounded off and we concluded that the bias-reduction method should be considered as an improvement over traditional ML.

For target iii), in Chapter 5 we have considered

a) the probit, the complementary log-log and the log-log models for binary responses and,

b) the 1-parameter and 2-parameter logistic models (Birnbaum, 1968) from item response theory.

The above models belong to the class of exponential family non-linear models with known dispersion and the application of the bias-reduction method has been greatly facilitated by the results of Chapter 3.

For binomial-response GLMs, apart from the modified IWLS procedure described in Chapter 3, we have derived an alternative algorithm for the implementation of bias reduction. The algorithm depends on pseudo-data representations and on already implemented ML fitting procedures, so that it allows the quick implementation of the method. Furthermore, we have pursued extensive empirical studies, mainly demonstrating that the finiteness and shrinkage properties of the BR estimator extend beyond canonical-links. In addition, a comparison of the BR, the ML and the bias-corrected (BC, Cordeiro & McCullagh, 1991) estimators has been performed, illustrating the superiority of the BR estimator over the ML and BC estimators.

In the case of the two item response theory models, the 1PL model consists of parallel logistic regressions and the results of Chapter 4 apply directly. For the non-linear predictor 2PL model, the modified score functions have been explicitly given and their neat correspondence with the modified scores for the 1PL model has been shown. Apart from the standard modification to the usual IWLS, we have, also, shown how the alternative fitting algorithm for binomial-response GLMs could be used for obtaining the BR estimates. This was again achieved by a pseudo-data representation. Under an alternative asymptotic framework we have conducted a small empirical study that illustrates the finiteness and shrinkage properties of the BR estimator and thus its superiority relative to the ML estimator.

Although the above two studies showed that the neat properties of the BR estimator extend beyond the logistic regression case, our results raised more new questions than the ones answered and, certainly, further work is required in the area. In what follows we list some of the open topics.

7.2 Further work on bias reduction

The following list could serve as an agenda for future work in the area. Most of the items have been already mentioned in the concluding remarks and discussion parts of each chapter. Nevertheless, we mention them here among others, in a more structured form.

i) *Develop a formal model-comparison framework for logistic regressions that is based upon the penalized likelihood.*

ii) *Develop methods for the construction of confidence intervals for the bias-reduced estimates in curved binary-response GLMs, where a penalized likelihood corresponding to the modified scores does not exist.*

iii) *Consider the penalization of the likelihood by Jeffreys prior for curved models and compare the resultant maximum penalized likelihood (MPL) estimators with the BR estimators.*
The proposed confidence interval for the MPL estimates in logistic regressions should perform equally well in curved cases.

iv) *Develop formal proofs for the finiteness and shrinkage properties of the BR estimates for general binomial-response GLMs.*

v) *Research on how to reduce the bias in problems where the dimension of the parameter space increases with the sample size.*
Such models are for example, the 1PL and 2PL models and binary matched pair models (Cox & Snell, 1989, §2.4). The ML estimator for such models is inconsistent and so is the BR estimator that has been studied in the current thesis. For such models, a different bias-reduction method has to be developed, that is robust to the increase of the dimension of the parameter space, as the number of information units increases. The target is the removal of the persistent $\mathcal{O}(1)$ term in the bias expansion of the ML estimator, and the parallel construction of consistent estimators.

vi) *Extend the bias-reduction method to cover the estimation of the dispersion parameter in exponential family non-linear models.*

vii) *Use the results in Chapter 6 to construct estimators with other improved properties.*
The area of additively modified scores seems fruitful and particularly, MSE-reduction is most attractive because MSE is a measure that incorporates the trade-off between bias and variance.

viii) *Explore the properties of the resultant estimators when more elaborate modifications than the modifications based on the expected information are used.*
In the current thesis we thoroughly considered the case of modifications based on the expected information, mainly because they have the simplest form among the members of the family of possible bias-reducing modifications. Thus, they allowed the development of neat theoretical and applied results. An 'inter-comparison' of the properties of the estimators resulting from other modifications is still needed.

ix) *Apply the modified-scores approach to bias reduction to random/mixed effects models.*
The application of the bias-reduction method in the literature and in the current thesis has been restricted to fixed effects models. The reason is that the modifications for the efficient scores involve higher order cumulants of the log-likelihood

derivatives. When random effects are included in the model, the complex form of the log-likelihood — due to the integrals involved — is inherited by the cumulants of its derivatives, making the application of the method difficult.

The topic here involves research on how to apply the bias-reduction method in random/mixed effects models and thereupon the study of the properties of the BR estimators, determining whether the neat properties noted in fixed effects models for categorical data generalize for random/mixed effects models.

A good reference point seems to be the work in Breslow & Lin (1995).

x) *Explore more models in order to identify further cases where the use of the modified scores is beneficial.*
The results in Chapter 3 cover many models that are used in statistical applications and allow the easy implementation and study of the impact of bias reduction. Interest lies in finding other models where the modification of the efficient scores is desirable.

xi) *Development of statistical software for the public use of the results of the current thesis.*
The implementation of the results in the current thesis is intended to be released in *CRAN* (`cran.r-project.org`), as a contributed package for the *R language* (R Development Core Team, 2007).

Appendix A

Index notation and tensors

A.1 Introduction

The term "index notation" is used to describe a set of rules and conventions that facilitate the notation and description of multi-dimensional algebraic structures. McCullagh (1984) recognised the importance of index notation for theoretical work in statistics and introduced a variant that allowed him to express in an elegant fashion the, otherwise messy, identities giving the generalized cumulants of multi-dimensional random variables in terms of the ordinary cumulants. In the same paper he gives a generic identity connecting the joint cumulant generating function of any polynomial transformation with the cumulants of the original variables. Later, in McCullagh (1987), a complete treatment of index notation is given and it is applied to various contexts of statistics. Pace & Salvan (1997) use index notation and derive several likelihood related expansions for statistical problems with multi-dimensional target parameter. Both McCullagh (1987) and Pace & Salvan (1997) insist on the great value of index notation for the study of transformation rules for statistics under the re-parameterization of a statistical model. Towards this direction, they focus on tensors. Tensors are arrays that transform in a special way under specific groups of transformations and constitute the 'raw materials' for the construction of invariants, namely quantities that their value is unaffected from the choice of coordinate system within the group of transformations.

In this appendix we give an elementary review of index notation and tensors, focusing mainly on results related to the contents of the thesis. For a detailed treatment the reader is referred to the aforementioned material.

A.2 Index notation and Einstein summation convention

A.2.1 Some examples of index notation

Consider a parameter vector $\beta = (\beta^1, \beta^2, \ldots, \beta^p)$ in $\Re^p$. In index notation this is denoted just as β^r and the range of r is understood from the context. Similarly, the score vector U having p-components is denoted as U_r. Also, under the same convention, U_{rs} is used to

dentote the $p \times p$ Hessian matrix of second derivatives of the log-likelihood with respect to the parameters and I^{rs} denotes minus the matrix-inverse of U_{rs}, namely the observed information. The same notational conventions can be used for 3-way arrays, for example, U_{rst}, κ_{rst}, ν_t^{rs}, for 4-way arrays, for example ν_{tu}^{rs}, and so on. All r,s,t and u above take values in the index set $\{1, 2, \ldots, p\}$. The position of the index (upper or lower or both) plays an important role in the understanding of the nature of a quantity. Many times, it is merely a prior convention for the quantities involved in some algebraic manipulation. However, being consistent with this convention can reveal important properties of the outcome and the way this behaves under certain groups of transformations. An upper index, such as r in β^r is called a *contravariant index* and a lower one a *covariant index*. Without loss of generality, all multiply-indexed quantities are considered symmetric under index permutations within their contravariant group of indices and within their covariant group of indices.

A.2.2 Einstein summation convention

The power of index notation is noted after the introduction of the summation convention. Whenever the same index appears once in the contravariant group of indices of a quantity a and once in the covariant group of another quantity b, summation is implied over that index.

$$a^r b_r := \sum_{r=1}^{p} a^r b_r \,.$$

Example A.2.1: The quadratic form $U^T F^{-1} U$, involving the score vector U and the Fisher information F, can be written as $\mu^{r,s} U_r U_s$ where $\mu_{r,s} = \mathrm{E}(U_r U_s)$ is the Fisher information and $\mu^{r,s}$ its matrix-inverse. The correspondence in this case might seem trivial and a waste of subscripts and superscripts. But moving to higher-order relationships, expressions in usual matrix notation can get unsightly and hard to understand. Consider the cubic form $\beta^r \beta^s \beta^t U_{rst}$, where U_{rst} are the third partial derivatives of the log-likelihood with respect to β^r, β^s and β^t. This quantity appears in stochastic expansions of the score vector. Under the notational rules of Section 2.2

$$\beta^r \beta^s \beta^t U_{rst} := \sum_{r=1}^{q} \beta^r \beta^T S_r \beta \,,$$

with $S_r = \mathcal{D}^2 (U_r; \beta)$, the $p \times p$ Hessian of the r-th component of the score vector with respect to β. Extra care should be taken in the above summation. Any re-arrangement of quantities should not alter the position of the quantities in $\beta^T S_r \beta$ because then the fundamental definition of matrix multiplication is violated. This is not the case for index notation using the summation convention. The expressions $\beta^r \beta^s \beta^t U_{rst}$, $\beta^r \beta^s U_{rst} \beta^t$ and any other re-arrangement of β^r, β^s, β^t and U_{rst} refer to the same quantity. Imagining how higher degree homogeneous polynomials would look in matrix notation, the superiority of index notation for theoretical work becomes apparent.

A.2.3 Free indices, dummy indices and transformations

Example A.2.2: Consider two vectors of random variables $Y = [Y^r]$, $X = [X^r]$ taking values in $\Re^q$ and having variance-covariance matrices $\Sigma_X = \text{Var}(X)$ and $\Sigma_Y = \text{Var}(Y)$, respectively. Also, let $\Sigma_{XY} = [Cov(X_r, Y_s)]$. Consider linear transformations $g(X) = AX$ and $h(Y) = BY$, with A and B two $k \times q$ matrices of known constants. It is known that, the transformed random variable $Z = g(X) + h(Y)$ has variance-covariance matrix

$$\Sigma = A\Sigma_X A^T + A\Sigma_{XY} B^T + B\Sigma_{XY}^T A^T + B\Sigma_Y B^T \,. \tag{A.1}$$

In index notation the above equality is expressed as

$$\sigma^{rs} = a_t^r a_u^s \sigma_X^{tu} + a_t^r b_u^s \sigma_{XY}^{tu} + b_t^r a_u^s \sigma_{XY}^{tu} + b_t^r b_u^s \sigma_Y^{tu} \,, \tag{A.2}$$

with $A = [a_t^r]$, $B = [b_t^r]$ and obvious correspondences between 'Σ' matrices in (A.1) and 'σ' quantities in (A.2). Again, any re-arrangement of quantities within any summand of (A.2) has no effect on the result. This example illustrates that, under index notation, the effect of the variance and the covariance operators to linear transformations of multi-dimensional random variables is the same as in the univariate case.

There is a useful distinction between the indices appearing in the summands of expressions like (A.2). *Free indices* are the ones appearing only once in each summand and *dummy indices* are the ones used in the application of the summation convention. For example, in any summand of (A.2) r, s are the free indices and t, u are the dummy ones. This distinction provides two devices for easy detection of algebraic mistakes:

i) Once the letters for the free indices for one side of an equation have been chosen, the letters used for the free indices in the other side cannot be different.

ii) For sums of products of arrays, any dummy index can appear only two times in each summand, once in the covariant part of some quantity and once in the contravariant part of another.

Rule ii) implies that if we change the letter for a dummy index, we have to change the letter for its other occurrence.

A.2.4 Differentiation

In Theorem B.1.1 and Theorem B.1.2 and using matrix notation we give the general expressions for the Jacobian and Hessian matrices of composite vector valued functions. For real-valued composite functions and in index notation, the expressions in Theorem B.1.1 and Theorem B.1.2 are special cases of an elegant expression for the higher order derivatives of composite functions.

Let $S \subset \Re^m$, and assume that $f : S \to \Re^n$ is smooth at a point b in the interior of S. Let $T \subset \Re^n$ such that $f(x) \in T$, for every $x \in S$, and assume that $g : T \to \Re$ is smooth at a point $c = f(b)$ in the interior of T. Then the composite function $h : S \to \Re$ defined by

$$h(x) = (g \circ f)(x) = g(f(x)) \,,$$

is smooth at b. Indicating the generic arguments of f and g by $b = [b^r] \in S$ and $c = [c^s] \in T$, respectively, the real-valued composite function $h(x) = g(f(x))$ will have partial derivatives

$$h_r = \frac{\partial h(b)}{\partial b^r} = g_s f^s_r \,,$$

where $g_s = \partial g(c)/\partial c^s\big|_{c=f(b)}$ and $f^s_r = \partial f(b)^s/\partial b^r$. Furthermore, the chain rule for Hessian matrices in Theorem B.1.2 is given by

$$h_{r_1 r_2} = \frac{\partial^2 h(b)}{\partial b^{r_1} \partial b^{r_2}} = g_{s_1 s_2} f^{s_1}_{r_1} f^{s_2}_{r_2} + g_{s_1} f^{s_1}_{r_1 r_2} \,,$$

where $g_{s_1 s_2} = \partial^2 g(c)/\partial c^{s_1} \partial c^{s_2}\big|_{c=f(b)}$ and $f^{s_1}_{r_1 r_2} = \partial^2 f(b)^{s_1}/\partial b^{r_1} \partial b^{r_2}$. Note that, the above two equations describe the generic elements of the Jacobian and Hessian matrices as defined in Theorem B.1.1 and Theorem B.1.2 with no special reference to the dimension of the algebraic structures involved and with validity of the commutative property for the multiplication of arrays.

As we move to higher order derivatives, the relevant expressions become increasingly unsightly in matrix notation. Index notation not only keeps the formulae elegant but provides the means for the generic expression of derivatives of arbitrary degree. For this purpose, the concept of a *multiindex* is introduced (see, also Pace & Salvan, 1997, §9.1). Arrays such as $h_{r_1 r_2 \ldots r_a}$ can be denoted in a more compact form as h_{R_a}, where $R_a = \{r_1, r_2, \ldots, r_a\}$ is a finite set of a indices and is called a *multiindex* of order a. Using multiindices, the generic a-th order derivative of $h(b)$ with respect to b is given by the expression

$$h_{R_a} = \sum_{t=1}^{a} g_{S_t} F^{S_t}_{R_a} \,, \tag{A.3}$$

with

$$F^{S_t}_{R_a} = \sum_{R_a/t} f^{s_1}_{R_{a_1}} \ldots f^{s_t}_{R_{a_t}} \,,$$

where $t \leq a$ and $\sum_{R_a/t}$ denotes summation over all possible partitions of R_a into t non-empty subsets $R_{a_1}, R_{a_2}, \ldots R_{a_t}$. Also,

$$h_{R_a} = \frac{\partial^a h(b)}{\partial b^{r_1} b^{r_2} \ldots b^{r_a}} \,, \qquad g_{S_t} = \left. \frac{\partial^t g(c)}{\partial c^{s_1} \ldots c^{s_t}} \right|_{c=f(b)} \,, \qquad f^s_{R_a} = \frac{\partial f(b)^s}{\partial b^{r_1} \ldots b^{r_a}} \,.$$

For example, by (A.3), the generic element of the array of the third order derivatives of h with respect to b is

$$\begin{aligned} h_{R_3} &= \sum_{t=1}^{3} g_{S_t} F^{S_t}_{R_3} = \sum_{t=1}^{3} g_{S_t} \sum_{R_3/t} f^{s_1}_{R_{q_1}} \ldots f^{s_t}_{R_{q_t}} \\ &= g_{s_1} f^{s_1}_{r_1 r_2 r_3} + g_{s_1 s_2} \left\{ f^{s_1}_{r_1 r_2} f^{s_2}_{r_3} + f^{s_1}_{r_1 r_3} f^{s_2}_{r_2} + f^{s_1}_{r_2 r_3} f^{s_2}_{r_1} \right\} + g_{s_1 s_2 s_3} f^{s_1}_{r_1} f^{s_2}_{r_2} f^{s_3}_{r_3} \,. \end{aligned}$$

A.3 Tensors

A.3.1 Definition

Consider an array $\omega_{T_b}^{S_a} = \omega_{t_1 \dots t_b}^{s_1 \dots s_a}$ that is a function of β. Such array is called a tensor of contravariant degree a and covariant degree b, or more concisely, a (a, b) tensor, if under a specified injective and smooth re-parameterization $\gamma = g(\beta)$, it transforms to $\tilde{\omega}_{V_b}^{U_a}$ according to the rule

$$\tilde{\omega}_{V_b}^{U_a} = \gamma_{s_1}^{u_1} \dots \gamma_{s_a}^{u_a} \omega_{T_b}^{S_a} \beta_{v_1}^{t_1} \dots \beta_{v_b}^{t_b} ,$$

with $\omega_{T_b}^{S_a}$ evaluated at $\beta = h(\gamma)$, h the inverse of the transformation g and $\beta_v^t = \partial \beta^t / \partial \gamma^v$, $\gamma_s^u = \partial \gamma^u / \partial \beta^s$. Also, $\beta_u^t \gamma_s^u = \delta_s^t$, with δ_s^t the Kronecker delta function, which takes value 1 for $t = s$ and 0 else.

Example A.3.1: Consider the Fisher information $\mu_{r,s} = \mathrm{E}(U_r U_s; \beta)$ on β. The comma in the covariant group of indices is just a convention and its utility will become apparent later. If we re-parameterize to $\gamma = g(\beta)$, where g is a smooth injective map, the Fisher information on γ is obtained by the relationship

$$\tilde{\mu}_{r,s} = \beta_t^r \beta_u^s \mu_{r,s} .$$

So $\mu_{r,s}$ is a $(0, 2)$ tensor or a covariant tensor of degree 2 under the group of smooth injections. Note that for the inverse of the Fisher information on γ we get

$$\tilde{\mu}^{r,s} = \gamma_t^r \gamma_u^s \mu^{t,u} , \tag{A.4}$$

with $\mu^{r,s}$ the matrix-inverse of the Fisher information on β. Thus, $\mu^{r,s}$ is a $(2, 0)$ tensor or a contravariant tensor of degree 2.

When we attribute the tensorial property to a quantity we have to state the group of transformations we are referring to. For example while the covariance σ^{rs} in (A.2) is a $(2, 0)$ tensor under general affine transformations $c^r + c_s^r Z^s$, with c^r and c_s^r constants, it is not a tensor under more general non-linear transformations.

A.3.2 Direct Kronecker products and contraction

A useful property of index notation is the facility of the transition from two vector spaces to the product space constructed by them and vice-versa.

Example A.3.2: Consider the setting of Example A.2.2 and assume a sequence of n independent copies of the q-dimensional random variable Z. This sequence can be thought as an element of the product space $\Re^n \times \Re^p$. The covariance matrix of the sequence is $\delta^{ij} \sigma^{rs}$, with δ^{ij} the Kronecker delta function, $i, j \in \{1, \dots, n\}$ and $r, s \in \{1, \dots, q\}$. Note that δ^{ij} is a tensor under the action of the symmetric group of permutations $\mathcal{G}_1$ in $\Re^n$ and, as already mentioned, σ^{rs} is a tensor under the group of general affine transformations $\mathcal{G}_2$ in $\Re^q$. The Kronecker product $\delta^{ij} \sigma^{rs}$ inherits the tensorial properties of both δ^{ij} under $\mathcal{G}_1$ and σ^{rs} under $\mathcal{G}_2$ and is a tensor under the direct product group of transformations $\mathcal{G}_1 \times \mathcal{G}_2$ acting on $\Re^n \times \Re^p$. This is necessary because the joint distribution of the n independent copies of Z is not affected by permutations of them.

In matrix notation, $\delta^{ij}\sigma^{rs}$ can be denoted as $1_n \otimes \Sigma$, where 1_n is the $n \times n$ identity matrix and the symbol $\otimes$ is used to denote the Kronecker product operator. However, $1_n \otimes \Sigma$ is generally a different matrix than $\Sigma \otimes 1_n$. No such distinction has to be made in index notation because the product of arrays is a commutative operation.

Another way of combining two existing tensors to construct a new one is to sum over pairs of indices, a process known as *contraction*. For example, if ω_{rst} and ω^{rs} are both tensors under some group of transformations $\mathcal{G}$, then $\tilde{\omega}^u_{rs} = \omega_{rst}\omega^{tu}$ is a $(1,2)$ tensor under the same group $\mathcal{G}$. Contraction is a basic operation for the construction of scalar invariants. For example, if A_{S_a} is a $(0,a)$ tensor and C^{S_a} is a $(a,0)$ tensor, then the scalar obtained by the contraction $A_{S_a}C^{S_a}$ is invariant, since under transformation we have

$$\begin{aligned}\tilde{A}_{S_a}\tilde{C}^{S_a} &= A_{T_a}\beta^{t_1}_{s_1}\dots\beta^{t_a}_{s_a}C^{U_a}\gamma^{s_1}_{u_1}\dots\gamma^{s_a}_{u_a}\\ &= A_{T_a}C^{U_a}\delta^{t_1}_{u_1}\dots\delta^{t_a}_{u_a}\\ &= A_{T_a}C^{T_a}\,.\end{aligned}$$

A.4 Likelihood quantities

A.4.1 Null moments and null cumulants of log-likelihood derivatives

In stochastic expansions of likelihood related quantities, the joint null moments and cumulants of log-likelihood derivatives play an important role. The word null refers to the fact that the operations of differention with respect to β and averaging over the sample space take place at the same value for β. We define these quantities similarly to Pace & Salvan (1997, Chapter 9) and introduce some notational conventions which allow the effective utilization of the compactness of index notation and, at the same time, keep the transparency on the interpretation of expressions.

Consider a statistical model with parameters the components of the p-vector β and log-likelihood function $l(\beta)$. We denote the log-likelihood derivatives by

$$U_{R_a} = U_{r_1r_2\dots r_a} = \frac{\partial^a l(\beta)}{\partial\beta^{r_1}\beta^{r_2}\dots\beta^{r_a}}\,.$$

For the expectations of products of log-likelihood derivatives we have

$$\begin{aligned}\mu_{R_a} &= \mathrm{E}(U_{R_a};\beta)\,,\\ \mu_{R_a,S_b} &= \mathrm{E}(U_{R_a}U_{S_b};\beta)\,,\\ \mu_{R_a,S_b,T_c} &= \mathrm{E}(U_{R_a}U_{S_b}U_{T_c};\beta)\,,\end{aligned}$$

and so on. The joint null moments of log-likelihood derivatives, as defined above, are symmetric under permutations of multiindices and under permutations of indices within each multiindex. However, they are not generally invariant under exchange of indices between the comma-separated groups, except in the case where the cardinality of two or more comma-separated multiindices is one so that their indices can be interchanged. Note that when the target parameter β is scalar, index notation gets clumsy and tedious. In such cases we use an alternative notation, denoting the log-likelihood derivatives as $U_k = \partial^k l(\beta)/\partial\beta^k$ and $\mu_k = \mathrm{E}(U_k)$, $\mu_{k,m} = \mathrm{E}(U_kU_m)$, etc.

Under the validity of the regularity conditions, the first four Bartlett identities can be written as

$$\begin{aligned} \mu_r &= 0\,, \\ \mu_{rs} + \mu_{r,s} &= 0\,, \\ \mu_{rst} + \mu_{r,st}[3] + \mu_{r,s,t} &= 0\,, \\ \mu_{rstu} + \mu_{rs,tu}[3] + \mu_{r,stu}[4] + \mu_{r,s,tu}[6] + \mu_{r,s,t,u} &= 0\,, \end{aligned} \tag{A.5}$$

where $[k]$ indicates the sum over the k possible permutations of indices by exchanging them between comma-separated groups and keeping the number of groups constant. So, for example,

$$\mu_{r,s,tu}[6] = \mu_{r,s,tu} + \mu_{r,t,su} + \mu_{r,u,st} + \mu_{s,t,ru} + \mu_{s,u,rt} + \mu_{t,u,rs}\,.$$

It is usually preferable to work with the null cumulants of the log-likelihood derivatives rather that the moments. In order to avoid the introduction of further notational rules for separating between groups of indices for the null cumulants of log-likelihood derivatives, the letter κ is reserved for their notation. The null cumulants can be expressed in terms of moments as follows,

$$\begin{aligned} n\kappa_{R_a} &= \mathrm{E}(U_{R_a};\beta) = \mu_{R_a}\,, \\ n\kappa_{R_a,S_b} &= \mathrm{Cov}(U_{R_a},U_{S_b};\beta) = \mu_{R_a,S_b} - \mu_{R_a}\mu_{S_b}\,, \\ n\kappa_{R_a,S_b,T_c} &= \mathrm{Cum}(U_{R_a},U_{S_b},U_{T_c};\beta) = \mu_{R_a,S_b,T_c} - \mu_{R_a}\mu_{S_b,T_c}\{3\} + 2\mu_{R_a}\mu_{S_b}\mu_{T_c}\,, \\ n\kappa_{R_a,S_b,T_c,Q_d} &= \mathrm{Cum}(U_{R_a},U_{S_b},U_{T_c},U_{Q_D};\beta) = \mu_{R_a,S_b,T_c,Q_d} - \mu_{R_a}\mu_{S_b,T_c,Q_d}\{4\} \\ &\quad - \mu_{R_a,S_b}\mu_{T_c,Q_d}\{3\} + 2\mu_{R_a}\mu_{S_b}\mu_{T_c,Q_d}\{6\} - 6\mu_{R_a}\mu_{S_b}\mu_{T_c}\mu_{Q_d}\,, \end{aligned} \tag{A.6}$$

and so on [1], where the notation highlights the fact that the joint cumulants of log-likelihood derivatives are of order $\mathcal{O}(n)$ under random sampling of size n and $\{k\}$ indicates summation over the k possible interchanges of multiindices between the quantities involved in the expression preceding it. For example,

$$\mu_{R_a}\mu_{S_b,T_c,Q_d}\{4\} = \mu_{R_a}\mu_{S_b,T_c,Q_d} + \mu_{S_b}\mu_{R_a,T_c,Q_d} + \mu_{T_c}\mu_{R_a,S_b,Q_d} + \mu_{Q_d}\mu_{R_a,S_b,T_c}\,.$$

Skovgaard (1986) proved a useful and elegant theorem on the differentiation of null cumulants of log-likelihood derivatives.

Theorem A.4.1: Differentiation of cumulants of log-likelihood derivatives (Skovgaard, 1986)
Let

$$n\kappa_{R_a,S_b,T_c,\ldots} = \mathrm{Cum}(U_{R_a},U_{S_b},U_{T_c},\ldots;\beta) = \mathrm{Cum}(D_1,D_2,\ldots,D_m;\beta)\,.$$

Then,

$$n\frac{\partial\kappa_{R_a,S_b,T_c,\ldots}}{\partial\beta^v} = \sum_i \mathrm{Cum}(D_1,D_2,\ldots,\frac{\partial D_i}{\partial\beta^v},\ldots,D_m;\beta) + \mathrm{Cum}(D_1,D_2,\ldots,D_m,U_v;\beta)\,.$$

□

[1] see the *exlog relations* (Barndorff-Nielsen & Cox, 1989, Section 5.4) which are the generic formulae expressing cumulants in terms of moments and vice-versa.

The relation described in the above theorem is valid for null moments, too. That is, if we replace κ with μ and $\mathrm{Cum}(D_1, D_2, \ldots, D_m; \beta)$ with $\mathrm{E}(D_1 D_2 \ldots D_m; \beta)$ wherever they appear, the identity remains valid. This fact can be used to obtain the m-th Bartlett-type identity by differentiating both sides of the $(m-1)$-th identity. This enables the replacement of μ with κ in (A.5) and therefore we obtain the identities

$$\begin{aligned}
\kappa_r &= 0\,, \\
\kappa_{rs} + \kappa_{r,s} &= 0\,, \\
\kappa_{rst} + \kappa_{r,st}[3] + \kappa_{r,s,t} &= 0\,, \\
\kappa_{rstu} + \kappa_{rs,tu}[3] + \kappa_{r,stu}[4] + \kappa_{r,s,tu}[6] + \kappa_{r,s,t,u} &= 0\,,
\end{aligned}$$

for the null cumulants *per observation*.

In order to be able to separate the stochastic from the deterministic part of the log-likelihood derivatives, we define the centered random variables

$$H_{R_a} = U_{R_a} - \mu_{R_a}\,.$$

These quantities can be used to facilitate the assignment of asymptotic orders to the terms of stochastic expansions of likelihood related quantities. The set of joint null cumulants of the triangular sequence of random variables $U_r, U_{rs}, U_{rst}, \ldots$ is the same as the set of cumulants of the triangular sequence of their centered counterparts. From this set of correspondences are excluded the expectations of U_{R_a} because $\kappa_{R_a} = n^{-1}\mathrm{E}(U_{R_a})$, which is zero only for $a = 1$ and the corresponding expectations of H_{R_a} are zero for every $a \geq 1$. All these are direct consequences of the shift invariance of the cumulants of second order and above, and of the unbiasedness of the score vector.

Therefore, using (A.6), the expressions giving the null cumulants in terms of the null moments of H_{R_a} (or null centered moments of U_{R_a}) are

$$\begin{aligned}
\kappa_{R_a,S_b} &= n^{-1}\nu_{R_a,S_b}\,, \\
\kappa_{R_a,S_b,T_c} &= n^{-1}\nu_{R_a,S_b,T_c}\,, \\
\kappa_{R_a,S_b,T_c,Q_d} &= n^{-1}\nu_{R_a,S_b,T_c,Q_d} - n^{-1}\nu_{R_a,S_b}\nu_{T_c,Q_d}\{3\}
\end{aligned} \tag{A.7}$$

and so on, where $\nu_{R_a,S_b} = \mathrm{E}(H_{R_a} H_{S_b}; \beta)$, $\nu_{R_a,S_b,T_c} = \mathrm{E}(H_{R_a} H_{S_b} H_{T_c}; \beta)$ etc.

A.4.2 Stochastic order and Landau symbols

Stochastic Taylor expansions are widely used in statistics for the derivation of asymptotic expressions for complicated likelihood related quantities by omitting *small* order terms, namely terms that become negligible under repeated sampling. Thus, the recognition of the order of the terms involved in a stochastic expansion has to be done in a systematic way. The stochastic order symbols $\mathcal{O}_p(.)$ and $o_p(.)$ are the most commonly used symbols for describing the asymptotic order of random quantities and are defined as follows:

Definition A.4.1: *Consider a sequence of scalar random variables X_n, $n = 1, 2, \ldots$. We write $X_n = o_p(a_n)$ if for every $\epsilon > 0$ and for every $\delta > 0$ there exists an integer $N(\delta, \epsilon)$ such that*

$$\textit{if} \quad n \geq N(\delta, \epsilon) \quad \textit{then} \quad P\left(\tfrac{|X_n|}{|a_n|} < \delta\right) \geq 1 - \epsilon\,,$$

where a_n is a sequence of constants.

Definition A.4.2: *Consider a sequence of scalar random variables X_n, $n = 1, 2, \ldots$. We write $X_n = \mathcal{O}_p(a_n)$ if for every $\epsilon > 0$ there exists $K(\epsilon) > 0$ and an integer $N(\epsilon)$ such that*

$$\text{if} \quad n \geq N(\epsilon) \quad \text{then} \quad P\left(\frac{|X_n|}{|a_n|} \leq K(\epsilon)\right) \geq 1 - \epsilon\,,$$

a_n is a sequence of constants.

If $\{X_n\}$ is a sequence of random vectors in $\Re^p$, then $X_n = o_p(a_n)$ if $||X_n|| = o_p(a_n)$ and $X_n = \mathcal{O}_p(a_n)$ if $||X_n|| = \mathcal{O}_p(a_n)$, where $||.||$ denotes some norm in $\Re^p$. Further, if the sequence of random vectors $\{X_n\}$ converges in distribution to a random vector X ($X_n \xrightarrow{d} X$) then $X_n = O_p(1)$ (X_n is bounded in probability) as n grows towards infinity. The converse is not generally true. Also, $X_n = o_p(1)$ if and only if X_n converges in probability to zero ($X_n \xrightarrow{p} 0$) as $n \to \infty$ (for definitions and examples on the various types of stochastic convergence the reader is referred to van der Vaart, 1998, Chapter 2). These symbols first appeared in Mann & Wald (1943) among several other symbols denoting different kinds of stochastic relationships, and they are generalizations of their deterministic counterparts $o(.)$ and $\mathcal{O}(.)$ (usually referred to as the Landau symbols).

Definition A.4.3: *Consider a sequence of real numbers b_n, $n = 1, 2, \ldots$. We write $b_n = o(a_n)$ if for every $\epsilon > 0$ there exists positive integer $N(\epsilon)$ such that*

$$\text{if} \quad n \geq N(\epsilon) \quad \text{then} \quad |b_n| < \epsilon |a_n|$$

or, alternatively, if $\lim_{n\to\infty} |b_n|/|a_n| = 0$

Definition A.4.4: *Consider a sequence of real numbers b_n, $n = 1, 2, \ldots$. We write $b_n = \mathcal{O}(a_n)$ if there exists $\epsilon > 0$ and positive integer $N(\epsilon)$ such that*

$$\text{if} \quad n \geq N(\epsilon) \quad \text{then} \quad |b_n| < \epsilon |a_n|$$

or, alternatively, $\limsup_{n\to\infty} |b_n|/|a_n| < \infty$

By the above definitions, for any real constant c, $\mathcal{O}_p(a_n)$, $o_p(a_n)$, $\mathcal{O}(a_n)$, $o(a_n)$ are equivalent to $ca_n\mathcal{O}_p(1)$, $ca_no_p(1)$, $ca_n\mathcal{O}(1)$, $ca_no(1)$, respectively. Also, while $\mathcal{O}_p(n^c) = \mathcal{O}_p(n^{c+1})$, $\mathcal{O}_p(n^{c+1}) \neq \mathcal{O}_p(c^t)$ and the same is true when $\mathcal{O}_p$ is replaced with either $\mathcal{O}$ or o or o_p. So, in expressions like $X_n = o_p(a_n)$ the use of the equality symbol is a slight abusage of notation. However, its use is convenient and it denotes the assignment of the property in the right hand side to the quantities of the left hand side. To completely clarify this asymmetry of the definition of these symbols, $X_n = \mathcal{O}_p(a_n)$ should be understood as "X_n is at most of order a_n in probability" and $X_n = o_p(a_n)$ as "X_n is of order smaller than a_n in probability". The same interpretation should be used for their deterministic counterparts omitting the expression "in probability".

A complete treatment of stochastic order symbols and illustrative examples of their use is given in Bishop et al. (1975, Section 14.4). Some of the properties of stochastic and deterministic order symbols are given below. They are used without comment throughout

the thesis. If a, b are real numbers and $k = \max\{a, b\}$ then

Products	Sums
$o(n^a)o(n^b) = o(n^{a+b})$	$o(n^a) + o(n^b) = o(n^k)$
$\mathcal{O}(n^a)\mathcal{O}(n^b) = \mathcal{O}(n^{a+b})$	$\mathcal{O}(n^a) + \mathcal{O}(n^b) = \mathcal{O}(n^k)$
$\mathcal{O}(n^a)o(n^b) = o(n^{a+b})$	$\mathcal{O}(n^a) + o(n^b) = \mathcal{O}(n^k)$
$o_p(n^a)o_p(n^b) = o_p(n^{a+b})$	$o_p(n^a) + o_p(n^b) = o_p(n^k)$
$\mathcal{O}_p(n^a)\mathcal{O}_p(n^b) = \mathcal{O}_p(n^{a+b})$	$\mathcal{O}_p(n^a) + \mathcal{O}_p(n^b) = \mathcal{O}_p(n^k)$
$\mathcal{O}_p(n^a)o_p(n^b) = o_p(n^{a+b})$	$\mathcal{O}_p(n^a) + o_p(n^b) = \mathcal{O}_p(n^k)$
$o_p(n^a)o(n^b) = o_p(n^{a+b})$	Compositions
$\mathcal{O}_p(n^a)o(n^b) = o_p(n^{a+b})$	$\mathcal{O}_p(\mathcal{O}(n^a)) = \mathcal{O}_p(n^a)$
$\mathcal{O}(n^a)o_p(n^b) = o_p(n^{a+b})$	$o(\mathcal{O}_p(n^a)) = o_p(n^a)$
$\mathcal{O}(n^a)\mathcal{O}_p(n^b) = \mathcal{O}_p(n^{a+b})$	$o_p(\mathcal{O}_p(n^a)) = o_p(n^a)$

The compositions above represent the effect of a linear function $f(x)$ on a quantity of known stochastic or deterministic order. For example, the equivalence $o(\mathcal{O}_p(n^a)) = o_p(n^a)$ is interpreted as if $f(x) = o(x)$ and $X_n = \mathcal{O}_p(n^a)$, then $f(X_n) = o_p(n^a)$. Further, we mention and prove a very useful result that describes when a O_p quantity is o_p and gives the appropriate order for the convergence in probability. This is a generalization of the result of exercise 6 in Bishop et al. (1975, Section 14.4.6) and is extensively used in stochastic expansions to formally justify the omission of lower order terms, ensuring that under repeated sampling they converge in probability to zero faster than the included terms.

Theorem A.4.2: A connection between $\mathcal{O}_p$ and o_p.
If $X_n = \mathcal{O}_p(n^{-a})$ with $a > 0$ then $X_n = o_p(n^{-a+t})$, for every $t > 0$.

Proof. By Definition A.4.2 we have that for every $\epsilon > 0$ there exists some constant $K(\epsilon) > 0$ and positive integer $N(\epsilon)$ such that if $n \geq N(\epsilon)$ then $P\left(n^a|X_n| \leq K(\epsilon)\right) \geq 1 - \epsilon$ or, equivalently, if $n \geq N(\epsilon)$ then $P\left(n^{a-t}|X_n| \leq n^{-t}K(\epsilon)\right) \geq 1 - \epsilon$, for $t > 0$.

Fix $\epsilon > 0$. For every $n \geq N(\epsilon)$ there exists $\delta > n^{-t}K(\epsilon)$ with

$$1 - \epsilon \leq P\left(n^{a-t}|X_n| \leq n^{-t}K(\epsilon)\right) \leq P\left(n^{a-t}|X_n| < \delta\right) . \tag{A.8}$$

Let $\Delta_n = \{\delta : \delta > n^{-t}K(\epsilon)|t > 0, \epsilon > 0\}$ and so $\Delta_n \subset \Re+$, with $\Re^+$ the set of all positive real numbers. Also, let $\bar{\Delta}_n = \Re^+ - \Delta_n$ and for a subset A of $\Re^+$ define the the length of A to be $\text{len}\, A = \sup A - \inf A$. Hence, $\text{len}\, \bar{\Delta}_n = n^{-t}K(\epsilon)$.

Relation (A.8) is satisfied for any δ in $\Delta_{N(\epsilon)}$ but not necessarily for δ in $\bar{\Delta}_{N(\epsilon)}$. However, note that $\Delta_\infty \equiv \Re^+$ and $\bar{\Delta}_\infty \equiv \emptyset$, with $\emptyset$ the empty set. Additionally, $\text{len}\, \bar{\Delta}_n$ is a strictly decreasing function of n. Hence, the validity of $(A.8)$ can be justified for any choice of $\delta \in \Re^+$ as long as $n \geq N^*(\delta, \epsilon)$, with $N^*(\delta, \epsilon)$ a positive integer sufficiently larger that $N(\epsilon)$.

Thus for every $\delta \in \Re^+$ there exists $N^*(\delta, \epsilon) > N(\epsilon)$ such that if $n \geq N^*(\delta, \epsilon)$ then $P\left(n^{a-t}|X_n| < \delta\right) \geq 1-\epsilon$, for $t > 0$. By the arbitrariness of choice of ϵ and definition A.4.1, we conclude that $X_n = o_p(n^{-a+t})$ for $t > 0$. □

A.4.3 Asymptotic order of null moments and cumulants of log-likelihood derivatives

Under random sampling of size n, the log-likelihood derivatives are all sums of n independent contributions and therefore

$$U_{R_a} = \begin{cases} \mathcal{O}_p(n^{1/2}) & \text{for } a = 1 \\ \mathcal{O}_p(n) & \text{for } a > 1 \end{cases} ,$$

since U_r has mean zero. Analogously, by an application of the central limit theorem, the centered random variables H_{R_a} are distributed according to a normal distribution with zero mean and $\mathcal{O}(n^{-1})$ variance-covariance matrix for large n. Thus, they are of order $\mathcal{O}_p(n^{1/2})$ for every $a \geq 1$. Also, all the joint null cumulants of log-likelihood derivatives are of order $\mathcal{O}(n)$.

For the identification of the asymptotic order of central moments of log-likelihood derivatives, Pace & Salvan (1997, Section 9.22) use the *exlog relations* (Barndorff-Nielsen & Cox, 1989, Section 5.4) in order to derive the following rule for the expectations of products of centered log-likelihood derivatives.

$$\nu_{R_{a_1},S_{a_2},\ldots,T_{a_d}} = \begin{cases} \mathcal{O}(n^{d/2}) & \text{if } d \text{ is even} \\ \mathcal{O}(n^{(d-1)/2}) & \text{if } d \text{ is odd} \end{cases} . \tag{A.9}$$

Example A.4.1: To illustrate the use of the above rule consider the expectation of products of centered log-likelihood derivatives $\nu_{rs,tu,v,w} = \mathrm{E}(H_{rs}H_{tu}U_vU_w)$.

By (A.7),

$$n\kappa_{rs,tu,v,w} = \nu_{rs,tu,v,w} - \nu_{rs,tu}\nu_{v,w}\{3\} ,$$

where $\{3\}$ indicates summation over the three possible permutations of comma separated groups of indices for the summand preceding it. Since $H_r = U_r$ we have that $\nu_{r_1,r_2,\ldots,r_d} = \mu_{r_1,r_2,\ldots,r_d}$. Hence

$$\begin{aligned} \nu_{rs,tu,v,w} &= \nu_{rs,tu}\nu_{v,w}\{3\} + \mathcal{O}(n) \\ &= \nu_{rs,tu}\nu_{v,w} + \nu_{rs,v}\nu_{tu,w} + \nu_{rs,w}\nu_{tu,v} + \mathcal{O}(n) \\ &= \mu_{v,w}(\mu_{rs,tu} - \mu_{rs}\mu_{tu}) + \mu_{rs,v}\mu_{tu,w} + \mu_{rs,w}\mu_{tu,v} + \mathcal{O}(n) , \end{aligned}$$

where the $\mathcal{O}(n)$ term $n\kappa_{rs,tu,v,w}$ is omitted from the expressions and only $\mathcal{O}(n^2)$ terms are involved. By the second Bartlett relation $\mu_{r,s} = -\mu_{rs} = n\kappa_{r,s}$ (the total Fisher information) and by the fact that $\nu_{R_a,S_b} = n\kappa_{R_a,S_b}$ for every $a, b \geq 1$, the above equation can be re-expressed as

$$\nu_{rs,tu,v,w} = n^2\kappa_{rs,tu}\kappa_{v,w} + n^2\kappa_{rs,v}\kappa_{tu,w} + n^2\kappa_{rs,w}\kappa_{tu,v} + \mathcal{O}(n) .$$

This is an evaluation of $\mathrm{E}(H_{rs}H_{tu}U_vU_w)$ in terms of cumulants, omitting any $\mathcal{O}(n)$ terms.

Appendix B

Some complementary results and algebraic derivations

This appendix contains some complementary material to support proofs of theorems and derivation of results in Chapter 4 and Chapter 5. Despite the fact that in several cases they constitute the corner stones for the results in the main text, they are stated within an appendix since their inclusion might be distracting to the reader.

Each section depends on the material in the main text, that the section refers to. Thus, the current appendix is standalone neither in notation nor in material, but every section has to be read in conjunction with the corresponding part of the main text.

B.1 Score functions and information measures for exponential family non-linear models

B.1.1 Some tools on the differentiation of matrices

In order to derive the score functions and the information measures for exponential family non-linear models we use two main results on the differentiation of vectors and matrices.

Theorem B.1.1: Chain Rule for Jacobian matrices (Magnus & Neudecker, 1999, §5.12). *Let $S \subset \Re^m$, and assume that $f : S \to \Re^n$ is differentiable at an interior point b of S. Let $T \subset \Re^n$ such that $f(x) \in T$, for every $x \in S$, and assume that $g : T \to \Re^p$ is differentiable at an interior point $c = f(b)$ of T. Then the composite function $h : S \to \Re^p$ defined by*

$$h(x) = (g \circ f)(x) = g(f(x)) \,,$$

is differentiable at b and the $p \times m$ Jacobian of h with respect to b is defined by

$$\mathcal{D}\left(h(b); b\right) = \mathcal{D}\left(g(c); c\right) \mathcal{D}\left(c; b\right) \,.$$

□

Theorem B.1.2: Chain Rule for Hessian matrices (Magnus & Neudecker, 1999, §6.10). *Let $S \subset \Re^m$, and assume that $f : S \to \Re^n$ is twice differentiable at an interior point b of S. Let $T \subset \Re^n$ such that $f(x) \in T$, for every $x \in S$, and assume that $g : T \to \Re^p$ is twice differentiable at an interior point $c = f(b)$ of T. Then the composite function $h : S \to \Re^p$ defined by*

$$h(x) = (g \circ f)(x) = g(f(x)) \,,$$

is twice differentiable at b and the $pm \times m$ Hessian of h with respect to b is defined by

$$\mathcal{D}^2 \left(h(b); b\right) = \left(1_p \otimes \mathcal{D}\left(c; b\right)\right)^T \mathcal{D}^2 \left(g(c); c\right) \mathcal{D}\left(c; b\right) + \left(\mathcal{D}\left(g(c); c\right) \otimes 1_m\right) \mathcal{D}^2 \left(c; b\right) ,$$

where, for example, the Hessian of a function $\alpha : \Re^n \to \Re^p$ with respect to its argument $x \in \Re^n$ is the $pn \times n$ matrix

$$\mathcal{D}^2 \left(\alpha; x\right) = \begin{bmatrix} \mathcal{D}^2 \left(\alpha_1; x\right) \\ \mathcal{D}^2 \left(\alpha_2; x\right) \\ \vdots \\ \mathcal{D}^2 \left(\alpha_p; x\right) \end{bmatrix} ,$$

with

$$\mathcal{D}^2 \left(\alpha_i; x\right) = \begin{bmatrix} \partial^2\alpha_i/\partial x_1^2 & \partial^2\alpha_i/\partial x_1 \partial x_2 & \dots & \partial^2\alpha_i/\partial x_1 \partial x_n \\ \partial^2\alpha_i/\partial x_1 \partial x_2 & \partial^2\alpha_i/\partial x_2^2 & \dots & \partial^2\alpha_i/\partial x_2 \partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2\alpha_i/\partial x_n \partial x_1 & \partial^2\alpha_i/\partial x_n \partial x_2 & \dots & \partial^2\alpha_i/\partial x_n^2 \end{bmatrix} ,$$

and 1_p and 1_n are the $p \times p$ and $n \times n$ identity matrices, respectively. □

B.1.2 Score functions and information measures

From the form of the log-likelihood function for exponential family non-linear models with known dispersion (2.8), the log-likelihood contribution of observation y_r is

$$l_r \equiv l_r(\beta; y_r, \lambda_r) = \frac{y_r^T \theta_r - b(\theta_r)}{\lambda_r} \,, \tag{B.1}$$

with λ_r known and fixed.

For the r-th contribution, the transformation from the parameter space B to the real line (where (B.1) takes values), could be represented visually by the diagrams below:

$$\mathrm{B} \longrightarrow \mathrm{H} \longrightarrow \Theta \longrightarrow \Re$$

with $\mathrm{B} \subset \Re^{\mathrm{p}}$, $\mathrm{H} \subset \Re^{\mathrm{q}}$ and $\Theta \subset \Re^q$.

We move progressively obtaining the Jacobian and the Hessian matrices for these transformations, making use of the relationships

$$\begin{aligned} \mathrm{E}(Y; \theta) &= \mu_*(\theta) = \nabla_\theta b(\theta) \,, \\ \mathrm{Cov}(Y; \theta) &= \Sigma_*(\theta) = \lambda \mathcal{D}^2 \left(b(\theta); \theta\right) \,, \end{aligned}$$

and Theorem B.1.1 and Theorem B.1.2.

We have:

- The $q \times p$ Jacobian of the transformation B $\longrightarrow$ H is

$$\mathcal{D}\left(\eta_r;\beta\right) = Z_r\,,$$

and the corresponding $qp \times p$ Hessian is

$$\mathcal{D}^2\left(\eta_r;\beta\right) = \left[\begin{array}{c} \mathcal{D}^2\left(\eta_{r1};\beta\right) \\ \mathcal{D}^2\left(\eta_{r2};\beta\right) \\ \vdots \\ \mathcal{D}^2\left(\eta_{rq};\beta\right) \end{array}\right].$$

- The $q \times q$ Jacobian of the transformation H $\longrightarrow \Theta$ is

$$\mathcal{D}\left(\theta_r;\eta_r\right) = \mathcal{D}\left(\theta_r;\mu_r\right)\mathcal{D}\left(\mu_r;\eta_r\right) = \lambda_r \Sigma_r^{-1} D_r^T\,,$$

and the $q^2 \times q$ Hessian is

$$\mathcal{D}^2\left(\theta_r;\eta_r\right) = V_r = \left[\begin{array}{c} V_{r1} \\ V_{r2} \\ \vdots \\ V_{rq} \end{array}\right],$$

with $V_{rs} = \mathcal{D}^2\left(\theta_{rs};\eta_r\right)$.

- The $q \times p$ Jacobian of the transformation B $\longrightarrow \Theta$ is

$$\mathcal{D}\left(\theta_r;\beta\right) = \mathcal{D}\left(\theta_r;\eta_r\right)\mathcal{D}\left(\eta_r;\beta\right) = \lambda_r \Sigma_r^{-1} D_r^T Z_r\,,$$

and the $qp \times p$ Hessian is

$$\begin{aligned}\mathcal{D}^2\left(\theta_r;\beta\right) &= \left(1_q \otimes \mathcal{D}\left(\eta_r;\beta\right)\right)^T \mathcal{D}^2\left(\theta_r;\eta_r\right)\mathcal{D}\left(\eta_r;\beta\right) + \left(\mathcal{D}\left(\theta_r;\eta_r\right)\otimes 1_p\right)\mathcal{D}^2\left(\eta_r;\beta\right) \\ &= \left(1_q \otimes Z_r\right)^T V_r Z_r + \lambda_r\left[\left(\Sigma_r^{-1} D_r^T\right)\otimes 1_p\right]\mathcal{D}^2\left(\eta_r;\beta\right)\,.\end{aligned}$$

- The $1 \times q$ gradient of the transformation $\Theta \longrightarrow \Re$ is

$$\mathcal{D}\left(l_r;\theta_r\right) = \lambda_r^{-1}\left(y_r - \mu_r\right)^T\,,$$

and the $q \times q$ Hessian is

$$\mathcal{D}^2\left(l_r;\theta_r\right) = -\lambda_r^{-1}\mathcal{D}^2\left(b(\theta_r);\theta_r\right) = -\lambda_r^{-2}\Sigma_r\,.$$

Notice that, in accordance with the notational rules in Section 2.2, all the matrices appearing above that are functions of the parameter vector β are denoted just by their corresponding letter. So, $Z_r \equiv Z_r(\beta)$, $\Sigma_r \equiv \Sigma_r(\beta)$, $D_r \equiv D_r(\beta)$, $V_r \equiv V_r(\beta)$ and $\mu_r \equiv \mu_r(\beta)$.

Since the log-likelihood itself is a sum of n independent contributions, the vector of score functions, the observed information matrix and the Fisher information matrix are all sums of n independent contributions.

Score functions

Using the above results the p-vector of score functions $U \equiv U(\beta)$ is the sum of n independent contributions and has the form

$$U = \sum_{r=1}^{n} \left(\mathcal{D}\left(l_r;\beta\right)\right)^T = \sum_{r=1}^{n} \left(\mathcal{D}\left(l_r;\theta_r\right)\mathcal{D}\left(\theta_r;\beta\right)\right)^T = \sum_{r=1}^{n} Z_r^T D_r \Sigma_r^{-1}(y_r - \mu_r)\,. \tag{B.2}$$

So, in the case of exponential family non-linear models with known dispersion, the t-th component of the score vector has the generic form

$$U_t = \sum_{r=1}^{n}\sum_{s=1}^{q} c_{rts}(y_{rs} - \mu_{rs}) \quad (t = 1, \ldots, p)\ ,$$

where c_{rts} is the (t,s)-th element of the $p \times q$ matrix $C_r = Z_r^T D_r \Sigma_r^{-1}$ and y_{rs}, μ_{rs} denote the s-th components of y_r and μ_r respectively.

Observed information

The observed information $I \equiv I(\beta)$ is defined as minus the Hessian of the log-likelihood with respect to the parameters β. Thus, for the $p \times p$ observed information matrix I on β we have

$$\begin{aligned} I &= -\sum_{r=1}^{n} \mathcal{D}^2\left(l_r;\beta\right) \\ &= -\sum_{r=1}^{n} \left[\mathcal{D}\left(\theta_r;\beta\right)^T \mathcal{D}^2\left(l_r;\theta_r\right)\mathcal{D}\left(\theta_r;\beta\right) + \left(\mathcal{D}\left(l_r;\theta_r\right)\otimes 1_p\right)\mathcal{D}^2\left(\theta_r;\beta\right)\right]\ , \end{aligned}$$

so that

$$\begin{aligned} I = \sum_{r=1}^{n} Z_r^T W_r Z_r &- \sum_{r=1}^{n}\sum_{s=1}^{q} \lambda_r^{-1} Z_r^T V_{rs} Z_r (y_{rs} - \mu_{rs}) \\ &- \sum_{r=1}^{n}\sum_{s,u=1}^{q} (y_{rs} - \mu_{rs}) k_{rsu} \mathcal{D}^2\left(\eta_{ru};\beta\right)\ , \end{aligned} \tag{B.3}$$

with $W_r = D_r \Sigma_r D_r^T$ and k_{rsu} the (s,u)-the element of the matrix $\Sigma_r^{-1} D_r^T$.

Expected information

The expected or Fisher information $F \equiv F(\beta)$ is defined as the variance-covariance matrix of the score vector. Under the independence of the random variables $Y_1, Y_2, \ldots, Y_n$, for

the $p \times p$ expected information matrix F on β we have

$$\begin{aligned} F &= \mathrm{E}\left(UU^T\right) \\ &= \mathrm{E}\left(\sum_{r=1}^{n} Z_r^T D_r \Sigma_r^{-1}(Y_r - \mu_r) \sum_{i=1}^{n} (Y_i - \mu_i)^T \Sigma_i^{-1} D_i^T Z_i\right) \\ &= \sum_{r=1}^{n} Z_r^T D_r \Sigma_r^{-1} \mathrm{E}\left[(Y_r - \mu_r)(Y_r - \mu_r)^T\right] \Sigma_r^{-1} D_r^T Z_r \\ &= \sum_{r=1}^{n} Z_r^T W_r Z_r \,. \end{aligned} \tag{B.4}$$

By (B.2), (B.3), (B.4) we can verify that the well-known Bartlett relations (Bartlett, 1953, Section 2) are satisfied for the class of exponential family non-linear models with known dispersion under the regularity conditions. So,

$$\begin{aligned} \mathrm{E}(U) &\equiv \mathrm{E}\left[\mathcal{D}\left(l(\beta);\beta\right)\right] = 0\,, \\ \mathrm{E}(I) &\equiv -\mathrm{E}\left[\mathcal{D}^2\left(l(\beta);\beta\right)\right] = \mathrm{E}\left[\mathcal{D}\left(l(\beta);\beta\right)^T \mathcal{D}\left(l(\beta);\beta\right)\right] \equiv F\,, \end{aligned}$$

because the expectation of the two last summands in the right hand side of (B.3) is zero.

B.2 Modified scores for exponential family non-linear models

B.2.1 Introduction

In Section 3.2 we give the general form of the modified scores that result in first-order unbiased estimators in matrix notation:

$$U_t^* = U_t + \frac{1}{2}\sum_{u=1}^{p} e_{tu}\,\mathrm{trace}\left\{F^{-1}(P_u + Q_u)\right\} \quad (t = 1, \ldots, p)\ , \tag{B.5}$$

where $P_u = \mathrm{E}(UU^T U_u)$ stands for the u-th block of the $p^2 \times p$ matrix of the third order cumulant of the scores, and $Q_u = \mathrm{E}(-IU_u)$ for the u-th block of the $p^2 \times p$ blocked matrix of the covariance of the first and second derivatives of the log-likelihood with respect to the parameters. Also, the alternatives for e_{tu} are either

$$e_{tu} \equiv e_{tu}^{(E)} = \left[RF^{-1} + 1_p\right]_{tu}$$

or

$$e_{tu} \equiv e_{tu}^{(O)} = \left[(I + R)F^{-1}\right]_{tu}$$

or

$$e_{tu} \equiv e_{tu}^{(S)} = \left[(UU^T + R)F^{-1}\right]_{tu}\ ,$$

with 1_p the $p \times p$ identity matrix.

This section is devoted to the detailed derivation of the form of explicit formulae for the modified scores in the case of exponential family non-linear models with known dispersion and it is intended to accompany Section 3.6 of the main text.

B.2.2 Derivation of the modified scores for exponential family non-linear models

As already mentioned in Section 3.2 any choice of e_{tu} among the alternatives in (B.5) results in first-order unbiased estimators and should depend upon the model under consideration, the chosen implementation procedure and generally to user-related choices for any specific application. For this reason they are going to remain unchanged in the explicit formulae derived here. Further, the form of the scores U, the Fisher information F and the observed information I on the model parameters β is derived in Subsection B.1.2 and we only need to derive explicit formulae for the cumulant matrices P_t and Q_t, $t = 1, \ldots, p$. Under the notational rules of Section 3.2, from (B.2) and the independence of the sequence of the random variables $\{Y_r\}$ we have

$$P_t = \mathrm{E}(UU^T U_t) = \sum_r \sum_{s=1}^{q} c_{rts} C_r \mathrm{E}\left\{(Y_{rs} - \mu_{rs})(Y_r - \mu_r)(Y_r - \mu_r)^T\right\} C_r^T \,.$$

In the above expression, c_{rts} is the (t, s)-th element of the $p \times q$ matrix $C_r = Z_r^T D_r \Sigma_r^{-1}$, where Σ_r is the $q \times q$ variance-covariance matrix of Y_r, D_r is the $q \times q$ matrix $\mathcal{D}\left(\mu_r; \eta_r\right)^T$ and Z_r is the $q \times p$ matrix $\mathcal{D}\left(\eta_r; \beta\right)$. Substituting, we have that

$$P_t = \mathrm{E}(UU^T U_t) = \sum_r \sum_{s=1}^{q} [Z_r^T D_r \Sigma_r^{-1}]_{ts} Z_r^T D_r \Sigma_r^{-1} K_{rs} \Sigma_r^{-1} D_r^T Z_r \,,$$

where K_{rs} denotes the s-th block of rows of K_r, $s = 1, \ldots, q$, with K_r the blocked $q^2 \times q$ matrix of cumulants of order three and degree three of the random vector Y_r. So,

$$P_t = \mathrm{E}(UU^T U_t) = \sum_r \sum_{s,v=1}^{q} [D_r \Sigma_r^{-1}]_{vs} Z_r^T D_r \Sigma_r^{-1} K_{rs} \Sigma_r^{-1} D_r^T Z_r z_{rvt} \,, \tag{B.6}$$

Further, from (B.2) and (B.3) we have

$$\begin{aligned}
Q_t = -\mathrm{E}(IU_t) &= \sum_r \lambda_r^{-1} \sum_{s,u=1}^{q} c_{rts} \mathrm{E}\left\{(Y_{ru} - \mu_{ru})(Y_{rs} - \mu_{rs})\right\} Z_r^T V_{ru} Z_r \\
&\quad + \sum_r \sum_{s,u,w=1}^{q} c_{rtw} \mathrm{E}\left[(Y_{rw} - \mu_{rw})(Y_{rs} - \mu_{rs})\right] k_{rsu} \mathcal{D}^2\left(\eta_{ru}; \beta\right) \\
&= \sum_r \lambda_r^{-1} \sum_{s,u=1}^{q} c_{rts} \sigma_{rsu} Z_r^T V_{ru} Z_r \\
&\quad + \sum_r \sum_{s,u,w=1}^{q} c_{rtw} \sigma_{rws} k_{rsu} \mathcal{D}^2\left(\eta_{ru}; \beta\right) \,,
\end{aligned}$$

where $V_{rs} = \mathcal{D}^2\left(\theta_{rs}; \eta_r\right)$ is the s-th $q \times q$ block of the $q^2 \times q$ blocked matrix $V_r = \mathcal{D}^2\left(\theta_r; \eta_r\right)$, k_{rsu} is the (s, u)-the element of the matrix $\Sigma_r^{-1} D_r^T$ and σ_{rsu} is the (s, u)-th component of

Σ_r. So,

$$
\begin{aligned}
Q_t = & \sum_r \lambda_r^{-1} \sum_{s,u=1}^{q} [D_r \Sigma_r^{-1}]_{us} Z_r^T (\Sigma_{rs} \otimes 1_q) V_r Z_r z_{rut} \\
& + \sum_r \sum_{s,u,w=1}^{q} [D_r \Sigma_r^{-1}]_{us} [D_r^T]_{sw} \mathcal{D}^2 (\eta_{rw}; \beta) z_{rut} \\
= & \sum_r \lambda_r^{-1} \sum_{s,v=1}^{q} [D_r \Sigma_r^{-1}]_{vs} Z_r^T (\Sigma_{rs} \otimes 1_q) V_r Z_r z_{rvt} \\
& + \sum_r \sum_{s,v=1}^{q} [D_r \Sigma_r^{-1}]_{vs} ([D_r^T]_s \otimes 1_q) \mathcal{D}^2 (\eta_r; \beta) z_{rvt} ,
\end{aligned} \tag{B.7}
$$

By (B.6) and (B.7) and the definitions of Section B.1, we have that

$$
\begin{aligned}
P_t + Q_t = & \sum_r \sum_{s,v=1}^{q} [D_r \Sigma_r^{-1}]_{vs} Z_r^T \left\{ D_r \Sigma_r^{-1} K_{rs} \Sigma_r^{-1} D_r^T + \lambda_r^{-1} (\Sigma_{rs} \otimes 1_q) V_r \right\} Z_r z_{rvt} \\
& + \sum_r \sum_{s,v=1}^{q} [D_r \Sigma_r^{-1}]_{vs} ([D_r^T]_s \otimes 1_q) \mathcal{D}^2 (\eta_r; \beta) z_{rvt} ,
\end{aligned}
$$

Now, $\mathcal{D}^2 (\mu_{rs}; \theta_r) = \lambda_r^{-2} K_{rs}$, and by Theorem B.1.2, for the first summand on the right hand side of the above expression we have that

$$
\begin{aligned}
& D_r \Sigma_r^{-1} K_{rs} \Sigma_r^{-1} D_r^T + \lambda_r^{-1} (\Sigma_{rs} \otimes 1_q) V_r \\
= & \mathcal{D} (\mu_r; \eta_r)^T \mathcal{D} (\theta_r; \mu_r)^T \mathcal{D}^2 (\mu_{rs}; \theta_r) \mathcal{D} (\theta_r; \mu_r) \mathcal{D} (\mu_r; \eta_r) + (\mathcal{D} (\mu_{rs}; \theta_r) \otimes 1_q) \mathcal{D}^2 (\theta_r; \eta_r) \\
= & \mathcal{D} (\theta_r; \eta_r)^T \mathcal{D}^2 (\mu_{rs}; \theta_r) \mathcal{D} (\theta_r; \eta_r) + (\mathcal{D} (\mu_{rs}; \theta_r) \otimes 1_q) \mathcal{D}^2 (\theta_r; \eta_r) = \mathcal{D}^2 (\mu_r; \eta_r) ,
\end{aligned}
$$

which is the matrix of second derivatives of the inverse link function with respect to the elements of the predictors vector η_r, and thus it depends only on the linking structure of the model.

So, given that $[D_r^T]_s = \mathcal{D} (\mu_{rs}; \eta_r)$ a representation for $P_t + Q_t$ is

$$
\begin{aligned}
P_t + Q_t = & \sum_r \sum_{s,v=1}^{q} [D_r \Sigma_r^{-1}]_{vs} \left\{ Z_r^T \mathcal{D}^2 (\mu_{rs}; \eta_r) Z_r \right. \\
& \left. + (\mathcal{D} (\mu_{rs}; \eta_r) \otimes 1_q) \mathcal{D}^2 (\eta_r; \beta) \right\} z_{rvt}
\end{aligned} \tag{B.8}
$$

or

$$
\begin{aligned}
P_t + Q_t = & \sum_r \sum_{s=1}^{q} Z_r^T ([D_r \Sigma_r^{-1}]_s \otimes 1_q) \mathcal{D}^2 (\mu_r; \eta_r) Z_r z_{rst} \\
& + \sum_r \sum_{s=1}^{q} (W_{rs} \otimes 1_q) \mathcal{D}^2 (\eta_r; \beta) z_{rst} ,
\end{aligned} \tag{B.9}
$$

where W_{rs} is the s-th row of the $q \times q$ matrix $W_r = D_r \Sigma_r^{-1} D_r^T$ as a $1 \times q$ vector and $[D_r \Sigma_r^{-1}]_s$ is the s-th row of $D_r \Sigma_r^{-1}$ as a $1 \times q$ vector. The above expression can be furher simplified by noting that

$$\begin{aligned}
&Z_r^T \mathcal{D}^2 \left(\mu_{rs}; \eta_r\right) Z_r + \left(\mathcal{D}\left(\mu_{rs}; \eta_r\right) \otimes 1_q\right) \\
=&\mathcal{D}\left(\eta_r; \beta\right)^T \mathcal{D}^2 \left(\mu_{rs}; \eta_r\right) \mathcal{D}\left(\eta_r; \beta\right) + \left(\mathcal{D}\left(\mu_{rs}; \eta_r\right) \otimes 1_q\right) \\
=&\mathcal{D}^2 \left(\mu_{rs}; \beta\right) .
\end{aligned}$$

Hence another representation of the sum of cumulants would be

$$P_t + Q_t = \sum_r \sum_{s,v=1}^{q} [D_r \Sigma_r^{-1}]_{vs} \mathcal{D}^2 \left(\mu_{rs}; \beta\right) z_{rvt} . \tag{B.10}$$

Despite the apparent simplicity of the above expression when compared to (B.8) or (B.9), the latter expressions are preferred because they explicitly illustrate how the sum $P_t + Q_t$ can be decomposed to a part that depends on the linking structure through $\mathcal{D}^2 \left(\mu_r; \eta_r\right)$ and a part that depends on the, generally, non-linear structure of the predictor with respect to the model parameters through $\mathcal{D}^2 \left(\eta_r; \beta\right)$.

Substituting (B.9) in (B.5) we have that

$$\begin{aligned}
U_t^* &= U_t + \frac{1}{2} \sum_{u=1}^{p} e_{tu} \operatorname{trace} \left\{ F^{-1} (P_u + Q_u) \right\} \\
&= U_t + \frac{1}{2} \sum_{u=1}^{p} e_{tu} \operatorname{trace} \left\{ F^{-1} \sum_r \sum_{s=1}^{q} Z_r^T ([D_r \Sigma_r^{-1}]_s \otimes 1_q) \mathcal{D}^2 \left(\mu_r; \eta_r\right) Z_r z_{rsu} \right\} \\
&\quad + \frac{1}{2} \sum_{u=1}^{p} e_{tu} \operatorname{trace} \left\{ F^{-1} \sum_r \sum_{s=1}^{q} (W_{rs} \otimes 1_q) \mathcal{D}^2 \left(\eta_r; \beta\right) z_{rsu} \right\} \\
&= U_t + \frac{1}{2} \sum_r \sum_{s=1}^{q} \operatorname{trace} \left\{ F^{-1} Z_r^T ([D_r \Sigma_r^{-1}]_s \otimes 1_q) \mathcal{D}^2 \left(\mu_r; \eta_r\right) Z_r \right\} \sum_{u=1}^{p} e_{tu} z_{rsu} \\
&\quad + \frac{1}{2} \sum_r \sum_{s=1}^{q} \operatorname{trace} \left\{ F^{-1} (W_{rs} \otimes 1_q) \mathcal{D}^2 \left(\eta_r; \beta\right) \right\} \sum_{u=1}^{p} e_{tu} z_{rsu} \\
&= U_t + \frac{1}{2} \sum_r \sum_{s=1}^{q} \operatorname{trace} \left\{ Z_r F^{-1} Z_r^T W_r W_r^{-1} ([D_r \Sigma_r^{-1}]_s \otimes 1_q) \mathcal{D}^2 \left(\mu_r; \eta_r\right) \right\} \sum_{u=1}^{p} e_{tu} z_{rsu} \\
&\quad + \frac{1}{2} \sum_r \sum_{s=1}^{q} \operatorname{trace} \left\{ F^{-1} (W_{rs} \otimes 1_q) \mathcal{D}^2 \left(\eta_r; \beta\right) \right\} \sum_{u=1}^{p} e_{tu} z_{rsu} \\
&= U_t + \frac{1}{2} \sum_r \sum_{s=1}^{q} \operatorname{trace} \left\{ H_r W_r^{-1} ([D_r \Sigma_r^{-1}]_s \otimes 1_q) \mathcal{D}^2 \left(\mu_r; \eta_r\right) \right\} \sum_{u=1}^{p} e_{tu} z_{rsu} \\
&\quad + \frac{1}{2} \sum_r \sum_{s=1}^{q} \operatorname{trace} \left\{ F^{-1} (W_{rs} \otimes 1_q) \mathcal{D}^2 \left(\eta_r; \beta\right) \right\} \sum_{u=1}^{p} e_{tu} z_{rsu} ,
\end{aligned}$$

where H_r is as defined in Section 2.4.

B.3 Some lemmas

Lemma B.3.1: Magnus & Neudecker (1999, Ch. 11, theorem 15). *Let A be a positive definite $n \times n$ matrix with eigenvalues $\lambda_1 \leq \lambda_1 \leq \ldots \leq \lambda_n$. Then*

$$\min_{G^T G = I_k} \det G^T A G = \prod_{i=1}^{k} \lambda_i \,,$$

$$\max_{G^T G = I_k} \det G^T A G = \prod_{i=1}^{k} \lambda_{n-k+i} \,.$$

Lemma B.3.2: Magnus & Neudecker (1999, Ch. 11, theorem 9). *Let $\lambda_k(C)$ be the k-th eigenvalue of a square matrix C. For any symmetric $n \times n$ matrix A and positive semidefinite matrix B,*

$$\lambda_r (A + B) \geq \lambda_r (A) \,, \qquad r = 1, 2, \ldots, n \,.$$

If B is positive definite, then the inequality is strict.

Lemma B.3.3: *If A and B are both diagonal $n \times n$ matrices with non-negative diagonal elements $\{a_r\}$ and $\{b_r\}$, respectively , and $a_r \geq b_r$, for every $r = 1, \ldots, n$, then, if X is a $n \times p$ matrix,* $\det\{X^T A X\} > \det\{X^T B X\}$,.

Proof. Since $A \geq B$, elementwise, $A = B + C$, where C is a diagonal matrix with non-negative entries. Further, $X^T A X$, $X^T B X$ and $X^T C X$ are positive semidefinite, by the non-negativity of the diagonal elements of A, B and C, respectively. Hence, by Lemma B.3.2,

$$\lambda_t \left(X^T A X\right) \geq \lambda_t \left(X^T B X\right) \,, \qquad t = 1, 2, \ldots, p \,.$$

Since the determinant of a matrix is the product of its eigenvalues the result follows. □

Lemma B.3.4: Magnus & Neudecker (1999, Ch. 11, theorem 25). *For any two positive semidefinite matrices A and B of the same order, and $0 < \theta < 1$, we have*

$$\det A^{\theta} \det B^{(1-\theta)} \leq \det \{\theta A + (1 - \theta) B\} \,.$$

Thus, the function $f(A) = \log \det A$ is concave in the space of all positive semidefinite matrices.

Lemma B.3.5: *Consider continuous functions f and g defined on some subset of $\Re^p$ and taking values in $\Re$. Further, let $x_m = \arg\max f(x)$ and $y_m = \arg\max\{f(x) + g(x)\}$. Then, $g(y_m) \geq g(x_m)$.*

Proof. Since $y_m = \arg\max\{f(x) + g(x)\}$ we have that

$$f(y_m) + g(y_m) \geq f(x_m) + g(x_m) \,.$$

But by the definition of x_m, $f(x_m) \geq f(y_m)$ so that necessarily $g(y_m) \geq g(x_m)$. □

B.4 Definition of separation for logistic regression

In this section we give the definition of separation and the theorems on the finiteness (existence, in the terminology in Albert & Anderson, 1984) of the maximum likelihood estimates found in Albert & Anderson (1984) and Lesaffre & Albert (1989). The reader is also referred to Santner & Duffy (1986) for corrections on the proofs of some theorems in Albert & Anderson (1984).

The setting is the same as in Section 4.3.7.3. Consider n independent realizations of the categorical random variable G taking values $1, 2, \ldots, k$ and a p-dimensional covariate setting x_r that corresponds to each realization of G, $r = 1, \ldots, n$. The baseline category representation of logistic regression for multinomial responses can be written as

$$\log \frac{\pi_{rs}}{\pi_{rk}} = \eta_s = \beta_s^T x_r \quad (s = 1, 2, \ldots, q)\,,$$

where $\pi_{rs} = P(G = s|x_r)$, $q = k - 1$ and x_r^T is the r-th row of the $n \times p$ design matrix X, assumed to be of full rank; if an intercept parameter is to be included in the model the the first column of X is a column of ones.

Here, $\gamma^T = (\beta_1^T, \ldots, \beta_q^T)^T$ and the parameter space Γ is assumed to be an open subset of $\Re^{pq}$. Let $E = \{1, 2, \ldots, n\}$ and E_s the set of row identifiers of X for observations with category label s, such that $\bigcup_{s=1}^{k} E_s = E$. Writing $r(s)$ we denote all those row identifiers that belong to E_s. Also, let $C = \{1, 2, \ldots, k\}$ and $\beta_k = 0$.

Definition B.4.1: Complete separation (Albert & Anderson, 1984, Section 3.2). *We say that there is complete separation of the sample points, if there exists vector $\gamma \in \Gamma$ such that for every $s \in C$ and for every $r(s)$ and $t \in C \setminus \{s\}$,*

$$(\beta_s - \beta_t)^T x_r > 0\,. \tag{B.11}$$

Definition B.4.2: Quasi-complete separation (Albert & Anderson, 1984, Section 3.3). *We say that there is quasi-complete separation of the sample points, if there exists vector $\gamma \in \Gamma$ such that for every $s \in C$ and for every $r(s)$ and $t \in C \setminus \{s\}$,*

$$(\beta_s - \beta_t)^T x_r \geq 0\,, \tag{B.12}$$

with equality satisfied for at least one triplet (r, s, t).

Definition B.4.3: Overlap (Albert & Anderson, 1984, Section 3.4). *We say that the sample points are overlapping if neither complete nor quasi-complete separation occur. That is if for every vector $\gamma \in \Gamma$ there exists a triplet (r, s, t) with $s \in C$ and $t \in C \setminus \{s\}$ such that*

$$(\beta_s - \beta_t)^T x_r < 0\,.$$

Theorem B.4.1: (Albert & Anderson, 1984, Theorem 1).
If there is complete separation of the sample points, the maximum likelihood estimator $\hat{\gamma}$ does not exist and

$$\max_{\gamma \in \Gamma} L(\gamma; X) = 1\,,$$

where $L(\gamma; X)$ is the likelihood function for γ.

Theorem B.4.2: (Albert & Anderson, 1984, Theorem 2). *If there is quasi-complete separation of the sample points, the maximum likelihood estimator $\hat{\gamma}$ does not exist and*

$$\max_{\gamma \in \Gamma} L(\gamma; X) < 1 \,.$$

Theorem B.4.3: (Albert & Anderson, 1984, Theorem 3). *If the sample points are overlapping, the maximum likelihood estimate $\hat{\gamma}$ exists and is unique.*

Theorem B.4.4: (Lesaffre & Albert, 1989, Theorem 2). *Let $F_{(c)}$ be the Fisher information on γ evaluated at the c-th iteration of the fitting procedure, $c = 1, 2, \ldots$.*

i. *If* $\operatorname{rank} X < p$ *then, for every c,* $\det(F_{(c)}) = 0$.

ii. *If* $\operatorname{rank} X = p$ *then, for every finite c,* $\det(F_{(c)}) > 0$. *There is (quasi-) complete separation of the sample points if and only if there exists a diagonal element of $F_{(c)}^{-1}$ which diverges as we maximize the likelihood $L(\gamma; X)$.*

B.5 Derivation of the modified scores for multinomial logistic regression models

By (4.22), the modified scores based on the expected information for a multinomial logistic regression model are

$$U_t^*(\gamma) = \sum_r \sum_{s=1}^{q} \left(y_{rs} - m_r \pi_{rs} + \frac{1}{2} \operatorname{trace} \left\{ H_r W_r^{-1} K_{rs} \right\} \right) z_{rst} \quad (t = 1, \ldots, pq) \,, \quad \text{(B.13)}$$

where, $W_r = m_r \operatorname{diag}(\pi_r) - m_r \pi_r \pi_r^T$, K_{rs} is a $q \times q$ symmetric matrix with (u, v)-th element the third order cumulants of Y_r and H_r is the r-th diagonal block of the $nq \times nq$ asymmetric hat matrix H. The (u, v)-th element of K_{rs} is given by

$$\kappa_{rsuv} = \operatorname{Cum}(Y_{rs}, Y_{ru}, Y_{rv}) = \begin{cases} m_r \pi_{rs}(1 - \pi_{rs})(1 - 2\pi_{rs}) & s = t = u \\ -m_r \pi_{rs} \pi_{ru}(1 - \pi_{rs}) & s = t \neq u \\ 2 m_r \pi_{rs} \pi_{rt} \pi_{ru} & s, t, u \quad \text{distinct} \end{cases} . \quad \text{(B.14)}$$

Also, W_r^{-1} has (s, u)-th element

$$\rho_{rsu} = \begin{cases} m_r^{-1}(\pi_{rs}^{-1} + \pi_{rk}^{-1}) & s = u \\ m_r^{-1} \pi_{rk}^{-1} & s \neq u \end{cases} , \quad \text{(B.15)}$$

with $\pi_{rk} = 1 - \sum_{s=1}^{q} \pi_{rs}$.

We can further simplify (B.13) by exploiting the structure of $\operatorname{trace}\{H_r W_r^{-1} K_{rs}\}$. We have

$$\operatorname{trace}\{H_r W_r^{-1} K_{rs}\} = \sum_{u,v=1}^{q} h_{ruv} b_{rsvu} \,, \quad \text{(B.16)}$$

where b_{rsvu} is the (v, u)-th element of $B_{rs} = W_r^{-1} K_{rs}$ and h_{ruv} is the (u, v)-th element of H_r. By using (B.14), (B.15) and letting $S = \{1, 2, \ldots, q\}$ we have,

- for $s = u$:

$$\begin{aligned} b_{rsvs} &= \sum_{w \in S} \rho_{rvw} \kappa_{rsws} \\ &= \rho_{rvs} \kappa_{rsss} + \sum_{w \in S \setminus \{s\}} \rho_{vw} \kappa_{rsws} \\ &= m_r \pi_{rs} (1 - \pi_{rs})(1 - 2\pi_{rs}) \rho_{rvs} - m_r \pi_{rs} (1 - 2\pi_{rs}) \sum_{w \in S \setminus \{s\}} \pi_{rw} \rho_{rvw} \,, \end{aligned}$$

  - for $v = s = u$,

$$\begin{aligned} b_{rsss} &= \pi_{rs}(1 - \pi_{rs})(1 - 2\pi_{rs})(\pi_{rs}^{-1} + \pi_{rk}^{-1}) - \pi_{rs}(1 - 2\pi_{rs}) \sum_{w \in S \setminus \{s\}} \pi_{rw}/\pi_{rk} \\ &= (1 - \pi_{rs})(1 - 2\pi_{rs}) + \pi_{rs}(1 - \pi_{rs})(1 - 2\pi_{rk})/\pi_{rk} - \pi_{rs}(1 - 2\pi_{rs})(1 - \pi_{rk})/\pi_{rk} \\ &= 1 - 2\pi_{rs} \,, \end{aligned}$$

  - for $v \neq s = u$,

$$\begin{aligned} b_{rsvs} &= \pi_{rs}(1 - \pi_{rs})(1 - 2\pi_{rs})/\pi_{rk} - \pi_{rs}(1 - 2\pi_{rs}) \sum_{w \in S \setminus \{s,v\}} \pi_{rw} \\ &\qquad - \pi_{rs}(1 - 2\pi_{rs}) - \pi_{rs}\pi_{rv}(1 - 2\pi_{rs})/\pi_{rk} \\ &= \pi_{rs}(1 - 2\pi_{rs})/\pi_{rk} - \pi_{rs}(1 - 2\pi_{rs})(1 - \pi_{rk})/\pi_{rk} - \pi_{rs}(1 - 2\pi_{rs}) \\ &= 0 \,, \end{aligned}$$

- for $s \neq u$:

$$\begin{aligned} b_{rsvu} &= \sum_{w \in S} \rho_{rvw} \kappa_{rswu} \\ &= \rho_{rvs} \kappa_{rssu} + \rho_{rvu} \kappa_{rsuu} + \sum_{w \in S \setminus \{s,u\}} \rho_{rvw} \kappa_{rswu} \\ &= -m_r \pi_{rs} \pi_{ru} (1 - 2\pi_{rs}) \rho_{rvs} - m_r \pi_{rs} \pi_{ru} (1 - 2\pi_{ru}) \rho_{rvu} \\ &\qquad + 2 m_r \pi_{rs} \pi_{ru} \sum_{w \in S \setminus \{s,u\}} \pi_{rw} \rho_{rvw} \,, \end{aligned}$$

  - for $v = s$ but $v \neq u$,

$$\begin{aligned} b_{rssu} &= -\pi_{ru}(1 - 2\pi_{rs}) - \pi_{rs}\pi_{ru}(1 - 2\pi_{rs})/\pi_{rk} - \pi_{rs}\pi_{ru}(1 - 2\pi_{ru})/\pi_{rk} \\ &\qquad + 2\pi_{rs}\pi_{ru} \sum_{w \in S \setminus \{s,u\}} \pi_{rw}/\pi_{rk} \\ &= -\pi_{ru}(1 - 2\pi_{rs}) - 2\pi_{rs}\pi_{ru}/\pi_{rk} + 2\pi_{rs}\pi_{ru}(1 - \pi_{rk})/\pi_{rk} \\ &= -\pi_{ru} \,, \end{aligned}$$

- for $v = u$ but $v \neq s$,

$$\begin{aligned} b_{rsuu} &= -\pi_{rs}\pi_{ru}(1-2\pi_{rs})/\pi_{rk} - \pi_{rs}(1-2\pi_{ru}) - \pi_{rs}\pi_{ru}(1-2\pi_{ru})/\pi_{rk} \\ &\quad + 2\pi_{rs}\pi_{ru} \sum_{w \in S\setminus\{s,u\}} \pi_{rw}/\pi_{rk} \\ &= -\pi_{rs} \end{aligned}$$

because this expression can be obtained from the expression of b_{rssu} by interchanging u and s.

- for $v \neq s$, $v \neq u$ and $s \neq u$,

$$\begin{aligned} b_{rsvu} &= -\pi_{rs}\pi_{ru}(1-2\pi_{rs})/\pi_{rk} - \pi_{rs}\pi_{ru}(1-2\pi_{ru})/\pi_{rk} + 2\pi_{rs}\pi_{ru} + 2\pi_{rs}\pi_{ru}\pi_{rv}/\pi_{rk} \\ &\quad 2\pi_{rs}\pi_{ru} \sum_{w \in S\setminus\{s,u,v\}} \pi_{rw} \\ &= -2\pi_{rs}\pi_{ru}/\pi_{rk} + 2\pi_{rs}\pi_{ru} + 2\pi_{ru}\pi_{rs}(1-\pi_{rk})/\pi_{rk} \\ &= 0\,. \end{aligned}$$

Thus, for $s \in S$

$$B_{rs} = \begin{bmatrix} -\pi_{rs} & 0 & \dots & 0 & \dots & 0 \\ 0 & -\pi_{rs} & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ -\pi_{r1} & -\pi_{r2} & \dots & 1-2\pi_{rs} & \dots & -\pi_{rq} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & -\pi_{rs} \end{bmatrix},$$

where $1 - 2\pi_{rs}$ is the s-th diagonal element of B_{rs}. So, by (B.16) we have

$$\begin{aligned} \operatorname{trace}\{H_r W_r^{-1} K_{rs}\} &= (1-2\pi_{rs})h_{rss} + \sum_{v \in S\setminus\{s\}} h_{rsv} b_{rsvs} + \sum_{u \in S\setminus\{s\}} h_{rus} b_{rssu} \\ &\quad + \sum_{u \in S\setminus\{s\}} h_{ruu} b_{rsuu} + \sum_{\substack{u,v=1 \\ u\neq v, v\neq s, u\neq s}}^{q} h_{ruv} b_{rsvu} \\ &= h_{rss} - \sum_{u \in S} \pi_{ru} h_{rus} - \pi_{rs} \operatorname{trace} H_r\,. \end{aligned}$$

Substituting in (B.13) we obtain the modified scores expressed in terms of the elements of H_r,

$$U_t^*(\gamma) = \sum_r \sum_{s=1}^{q} \left[y_{rs} + \frac{1}{2} h_{rss} - \left(m_r + \frac{1}{2} \operatorname{trace} H_r \right) \pi_{rs} - \frac{1}{2} \sum_{u=1}^{q} \pi_{ru} h_{rus} \right] z_{rst}\,,$$

for $t = 1, \dots, pq$.

B.6 Proof of theorem 4.3.1

Note that $\tilde{H}_r$ is invariant under the choice of either (γ, ϕ) or (γ, τ) parameterization, since the transformation $(\gamma, \phi) \to (\gamma, \tau)$ is invertible. For our purposes, it is more convenient to work in the (γ, τ) parameterization. Thus

$$\tilde{H}_r = Z_r^* \tilde{F}^{-1} Z_r^{*T} \tilde{W}_r \,, \tag{B.17}$$

with $\tilde{W}_r = \text{diag}\,\{\mu_{rs}; s = 1, \ldots, k\}$. Now,

$$Z_r^* = [\mathcal{D}\,(\tilde{\eta}_r; \gamma)\,|\mathcal{D}\,(\tilde{\eta}_r; \tau)] = [D_{1,r}|D_{2,r}]\,,$$

where, $\tilde{\eta}_r = (\tilde{\eta}_{r1}, \ldots, \tilde{\eta}_{rk})$, $D_{1,r}^T = [Z_r^T|0_p] - [L_q^T \otimes (Z_r^T \pi_r)|Z_r^T \pi_r]$ is a $pq \times k$ matrix with 0_p and L_q a $p \times 1$ vector of zeros and a $q \times 1$ vector of ones, respectively, and $\pi_r = (\mu_{r1}/\tau_r, \ldots, \mu_{rq}/\tau_r)^T$. Further, $D_{2,r}$ is a $k \times n$ matrix of zeros with the elements of its r-th column equal to τ_r^{-1}. Also,

$$\tilde{F} = \left[\begin{array}{c|c} \tilde{F}_\gamma & \\ \hline & \tilde{F}_\tau \end{array}\right],$$

where $\tilde{F}_\gamma$ is the Fisher information on γ and $\tilde{F}_\tau = \text{diag}\left\{m_r/\tau_r^2; r = 1, \ldots, n\right\}$. Palmgren (1981) showed that if we restrict the parameter space by $\tau_r = m_r$, $\tilde{F}_\gamma = F_\gamma$, where F_γ is the Fisher information on γ for the corresponding multinomial logistic regression model. So, on the restricted parameter space, we have

$$\tilde{H}_r = D_{1,r} F_\gamma^{-1} D_{1,r}^T \tilde{W}_r + D_{2,r}\ \tilde{F}_\tau^{-1}\Big|_{\tau_r = m_r} D_{2,r}^T \tilde{W}_r \,. \tag{B.18}$$

For the second summand of the above equation we have that

$$D_{2,r}\ \tilde{F}_\tau^{-1}\Big|_{\tau_r = m_r} D_{2,r}^T \tilde{W}_r = \begin{bmatrix} \pi_{r1} & \pi_{r2} & \ldots & \pi_{rk} \\ \vdots & \vdots & \vdots & \vdots \\ \pi_{r1} & \pi_{r2} & \ldots & \pi_{rk} \end{bmatrix}. \tag{B.19}$$

For the first summand we get

$$D_{1,r} F_\gamma^{-1} D_{1,r}^T \tilde{W}_r = A_{1,r} - A_{2,r} - A_{2,r}^T + A_{3,r} \tag{B.20}$$

where

$$A_{1,r} = \left[\begin{array}{c|c} Z_r F_\gamma^{-1} Z_r^T & 0_q \\ \hline 0_q^T & 0 \end{array}\right] = \left[\begin{array}{c|c} H_r W_r^{-1} & 0_q \\ \hline 0_q^T & 0 \end{array}\right],$$

$$A_{2,r} = \left[\begin{array}{c} L_k^T \otimes \left(Z_r F_\gamma^{-1} Z_r^T \pi_r\right) \\ \hline 0_k^T \end{array}\right] = \left[\begin{array}{c} L_k^T \otimes \left(H_r W_r^{-1} \pi_r\right) \\ \hline 0_k^T \end{array}\right] \text{ and}$$

$$A_{3,r} = (\pi_r^T Z_r F_\gamma^{-1} Z_r^T \pi_r) J_k = (\pi_r^T H_r W_r^{-1} \pi_r) J_k$$

where L_k is a $k \times 1$ vector of ones, J_k is a $k \times k$ matrix of ones and $H_r = Z_r F_\gamma^{-1} Z_r^T W_r$ is the $q \times q$, r-th diagonal block of the asymmetric hat matrix H for the corresponding multinomial logistic regression model. Substituting (B.19) and (B.20) in (B.18) and after straightforward but tedious calculation, we get the following identities

$$\begin{aligned}
\tilde{h}_{rsu} &= \pi_{ru} + h_{rsu} - \frac{\pi_{ru}}{\pi_{rk}} \sum_{v=1}^{q} h_{ruv} + \frac{\pi_{ru}}{\pi_{rk}} \sum_{v,w=1}^{q} h_{rvw} \pi_{rv} \,, \\
\tilde{h}_{rku} &= \pi_{ru} - \frac{\pi_{ru}}{\pi_{rk}} \sum_{v=1}^{q} h_{ruv} + \frac{\pi_{ru}}{\pi_{rk}} \sum_{v,w=1}^{q} h_{rvw} \pi_{rv} \,, \\
\tilde{h}_{rsk} &= \pi_{rk} - \sum_{u=1}^{q} h_{rsu} + \sum_{u,v=1}^{q} h_{ruv} \pi_{ru} \quad \text{and} \\
\tilde{h}_{rkk} &= \pi_{rk} + \sum_{s,u=1}^{q} h_{rsu} \pi_{rs} \,,
\end{aligned}$$

for $s, u \in \{1, \ldots, q\}$. Hence, the diagonal elements of $\tilde{H}_r$ have the form

$$\begin{aligned}
\tilde{h}_{rss} &= \pi_{rs} + h_{rss} - \frac{\pi_{rs}}{\pi_{rk}} \sum_{u=1}^{q} h_{rsu} + \frac{\pi_{rs}}{\pi_{rk}} \sum_{u,v=1}^{q} h_{ruv} \pi_{ru} \,, \\
\tilde{h}_{rkk} &= \pi_{rk} + \sum_{s,u=1}^{q} h_{rsu} \pi_{rs} \,,
\end{aligned}$$

for $s = 1, \ldots, q$. The proof is completed after showing that

$$\sum_{u=1}^{q} h_{rus} \pi_{ru} = \frac{\pi_{rs}}{\pi_{rk}} \sum_{u=1}^{q} h_{rsu} - \frac{\pi_{rs}}{\pi_{rk}} \sum_{u,v=1}^{q} h_{ruv} \pi_{ru} \,. \tag{B.21}$$

Note that by exploiting the structure of H_r and Z_r,

$$h_{rsu} = x_r^T F_{su}^- x_r \pi_{ru} - \sum_{v=1}^{q} x_r^T F_{sv}^- x_r \pi_{ru} \pi_{rv} \,,$$

where F_{su}^- is the (s,u)-th, $p \times p$ partition of F_γ^{-1} $(s, u \in \{1, \ldots, q\})$. So, substituting in the left hand side of (B.21) and noting that $F_{st}^- = F_{ts}^-$, we have that

$$\sum_{u=1}^{q} h_{rus} \pi_{ru} = \pi_{rs} \sum_{u=1}^{q} x_r^T F_{su}^- x_r \pi_{ru} - \pi_{rs} \sum_{u,v=1}^{q} x_r^T F_{vu}^- x_r \pi_{ru} \pi_{rv} \,.$$

The same substitution in the right hand side of (B.21) gives the same result and so (B.21) is valid.

Appendix C

Results of complete enumeration studies for binary response GLMs

In this appendix we present the results of the complete enumeration studies pursued in Chapter 4 and Chapter 5. Therein, these results are used to demonstrate that the bias-reduced (BR) estimates are finite even in cases of complete of quasi-complete separation of the data points, in which cases the maximum likelihood (ML) estimates are infinite. The BR estimates in Table C.1, Table C.2, Table C.3 and Table C.4 are obtained using algorithm 5.2 with the `glm` function in the *R language* (R Development Core Team, 2007). We use the criterion based on the sum of absolute changes on the estimates between successive iterations (step B.vii).a) in the algorithm) with tolerance $\epsilon > 10^{-10}$. In all cases convergence was rapid and it required just a few additional iterations from the starting values that are set in the initialization step of algorithm 5.2. For the ML estimates the `glm` function was used with the option

```
control = glm.control(epsilon = 1e-12, maxit = 100, trace = FALSE) .
```

Table C.1: Logistic link. Maximum likelihood estimates, bias-corrected estimates and bias-reduced estimates for (α, β, γ) to three decimal places, for every possible data configuration in Table 4.1 with $m_1 = m_2 = m_3 = m_4 = 2$.

Success Counts				Maximum Likelihood estimates			Bias-corrected estimates			Bias-reduced estimates		
y_1	y_2	y_3	y_4	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\alpha}_c$	$\hat{\beta}_c$	$\hat{\gamma}_c$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\gamma}$
0	0	0	0	$-\infty$	0	0	–	0	0	−1.846	0	0
0	0	0	1	$-\infty$	∞	∞	–	–	–	−2.869	1.363	1.363
0	0	0	2	$-\infty$	∞	∞	–	–	–	−4.504	3.003	3.003
0	0	1	0	$-\infty$	∞	$-\infty$	–	–	–	−1.505	1.363	−1.363
0	0	1	1	$-\infty$	∞	0	–	–	0	−1.967	1.967	0
0	0	1	2	$-\infty$	∞	∞	–	–	–	−2.869	3.011	1.363
0	0	2	0	$-\infty$	∞	$-\infty$	–	–	–	−1.501	3.003	−3.003
0	0	2	1	$-\infty$	∞	$-\infty$	–	–	–	−1.505	3.011	−1.363
0	0	2	2	$-\infty$	∞	0	–	–	0	−1.846	3.692	0
0	1	0	0	$-\infty$	$-\infty$	∞	–	–	–	−1.505	−1.363	1.363
0	1	0	1	$-\infty$	0	∞	–	0	–	−1.967	0	1.967
0	1	0	2	$-\infty$	∞	∞	–	–	–	−2.869	1.363	3.011
0	1	1	0	−1.099	0	0	−0.599	0	0	−0.762	0	0
0	1	1	1	−1.781	1.187	1.187	−0.944	0.63	0.63	−1.207	0.804	0.804
0	1	1	2	$-\infty$	∞	∞	–	–	–	−1.967	1.967	1.967
0	1	2	0	−0.594	1.187	−1.187	−0.315	0.63	−0.63	−0.402	0.804	−0.804
0	1	2	1	−1.099	2.197	0	−0.599	1.197	0	−0.762	1.524	0
0	1	2	2	$-\infty$	∞	∞	–	–	–	−1.505	3.011	1.363
0	2	0	0	$-\infty$	$-\infty$	∞	–	–	–	−1.501	−3.003	3.003
0	2	0	1	$-\infty$	$-\infty$	∞	–	–	–	−1.505	−1.363	3.011
0	2	0	2	$-\infty$	0	∞	–	0	–	−1.846	0	3.692
0	2	1	0	−0.594	−1.187	1.187	−0.315	−0.63	0.63	−0.402	−0.804	0.804
0	2	1	1	−1.099	0	2.197	−0.599	0	1.197	−0.762	0	1.524
0	2	1	2	$-\infty$	∞	∞	–	–	–	−1.505	1.363	3.011
0	2	2	0	0	0	0	0	0	0	0	0	0
0	2	2	1	−0.594	1.187	1.187	−0.315	0.63	0.63	−0.402	0.804	0.804
0	2	2	2	$-\infty$	∞	∞	–	–	–	−1.501	3.003	3.003
1	0	0	0	0	$-\infty$	$-\infty$	0	–	–	−0.142	−1.363	−1.363
1	0	0	1	−1.099	0	0	−0.599	0	0	−0.762	0	0
1	0	0	2	−1.781	1.187	1.187	−0.944	0.63	0.63	−1.207	0.804	0.804
1	0	1	0	0	0	$-\infty$	0	0	–	0	0	−1.967
1	0	1	1	−0.594	1.187	−1.187	−0.315	0.63	−0.63	−0.402	0.804	−0.804
1	0	1	2	−1.099	2.197	0	−0.599	1.197	0	−0.762	1.524	0
1	0	2	0	0	∞	$-\infty$	0	–	–	0.142	1.363	−3.011
1	0	2	1	0	∞	$-\infty$	0	–	–	0	1.967	−1.967
1	0	2	2	0	∞	$-\infty$	0	–	–	−0.142	3.011	−1.363
1	1	0	0	0	$-\infty$	0	0	–	0	0	−1.967	0
1	1	0	1	−0.594	−1.187	1.187	−0.315	−0.63	0.63	−0.402	−0.804	0.804

continued on next page

Success Counts				Maximum Likelihood estimates			Bias-corrected estimates			Bias-reduced estimates		
y_1	y_2	y_3	y_4	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\alpha}_c$	$\hat{\beta}_c$	$\hat{\gamma}_c$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\gamma}$
1	1	0	2	-1.099	0	2.197	-0.599	0	1.197	-0.762	0	1.524
1	1	1	0	0.594	-1.187	-1.187	0.315	-0.63	-0.63	0.402	-0.804	-0.804
1	1	1	1	0	0	0	0	0	0	0	0	0
1	1	1	2	-0.594	1.187	1.187	-0.315	0.63	0.63	-0.402	0.804	0.804
1	1	2	0	1.099	0	-2.197	0.599	0	-1.197	0.762	0	-1.524
1	1	2	1	0.594	1.187	-1.187	0.315	0.63	-0.63	0.402	0.804	-0.804
1	1	2	2	0	∞	0	0	–	0	0	1.967	0
1	2	0	0	0	$-\infty$	∞	0	–	–	0.142	-3.011	1.363
1	2	0	1	0	$-\infty$	∞	0	–	–	0	-1.967	1.967
1	2	0	2	0	$-\infty$	∞	0	–	–	-0.142	-1.363	3.011
1	2	1	0	1.099	-2.197	0	0.599	-1.197	0	0.762	-1.524	0
1	2	1	1	0.594	-1.187	1.187	0.315	-0.63	0.63	0.402	-0.804	0.804
1	2	1	2	0	0	∞	0	0	–	0	0	1.967
1	2	2	0	1.781	-1.187	-1.187	0.944	-0.63	-0.63	1.207	-0.804	-0.804
1	2	2	1	1.099	0	0	0.599	0	0	0.762	0	0
1	2	2	2	0	∞	∞	0	–	–	0.142	1.363	1.363
2	0	0	0	∞	$-\infty$	$-\infty$	–	–	–	1.501	-3.003	-3.003
2	0	0	1	0.594	-1.187	-1.187	0.315	-0.63	-0.63	0.402	-0.804	-0.804
2	0	0	2	0	0	0	0	0	0	0	0	0
2	0	1	0	∞	$-\infty$	$-\infty$	–	–	–	1.506	-1.363	-3.011
2	0	1	1	1.099	0	-2.197	0.599	0	-1.197	0.762	0	-1.524
2	0	1	2	0.594	1.187	-1.187	0.315	0.63	-0.63	0.402	0.804	-0.804
2	0	2	0	∞	0	$-\infty$	–	0	–	1.846	0	-3.692
2	0	2	1	∞	∞	$-\infty$	–	–	–	1.505	1.363	-3.011
2	0	2	2	∞	∞	$-\infty$	–	–	–	1.501	3.003	-3.003
2	1	0	0	∞	$-\infty$	$-\infty$	–	–	–	1.506	-3.011	-1.363
2	1	0	1	1.099	-2.197	0	0.599	-1.197	0	0.762	-1.524	0
2	1	0	2	0.594	-1.187	1.187	0.315	-0.63	0.63	0.402	-0.804	0.804
2	1	1	0	∞	$-\infty$	$-\infty$	–	–	–	1.967	-1.967	-1.967
2	1	1	1	1.781	-1.187	-1.187	0.944	-0.63	-0.63	1.207	-0.804	-0.804
2	1	1	2	1.099	0	0	0.599	0	0	0.762	0	0
2	1	2	0	∞	$-\infty$	$-\infty$	–	–	–	2.869	-1.363	-3.011
2	1	2	1	∞	0	$-\infty$	–	0	–	1.967	0	-1.967
2	1	2	2	∞	∞	$-\infty$	–	–	–	1.505	1.363	-1.363
2	2	0	0	∞	$-\infty$	0	–	–	0	1.846	-3.692	0
2	2	0	1	∞	$-\infty$	∞	–	–	–	1.505	-3.011	1.363
2	2	0	2	∞	$-\infty$	∞	–	–	–	1.501	-3.003	3.003
2	2	1	0	∞	$-\infty$	$-\infty$	–	–	–	2.869	-3.011	-1.363
2	2	1	1	∞	$-\infty$	0	–	–	0	1.967	-1.967	0
2	2	1	2	∞	$-\infty$	∞	–	–	–	1.505	-1.363	1.363
2	2	2	0	∞	$-\infty$	$-\infty$	–	–	–	4.504	-3.003	-3.003
2	2	2	1	∞	$-\infty$	$-\infty$	–	–	–	2.869	-1.363	-1.363
2	2	2	2	∞	0	0	–	0	0	1.846	0	0

Table C.2: Probit link. Maximum likelihood estimates, bias-corrected estimates and bias-reduced estimates for (α, β, γ) to three decimal places, for every possible data configuration in Table 4.1 with $m_1 = m_2 = m_3 = m_4 = 2$.

Success Counts				Maximum Likelihood estimates			Bias-corrected estimates			Bias-reduced estimates		
y_1	y_2	y_3	y_4	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\alpha}_c$	$\hat{\beta}_c$	$\hat{\gamma}_c$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\gamma}$
0	0	0	0	$-\infty$	0	0	–	0	0	−1.189	0	0
0	0	0	1	$-\infty$	∞	∞	–	–	–	−1.688	0.765	0.765
0	0	0	2	$-\infty$	∞	∞	–	–	–	−2.817	1.878	1.878
0	0	1	0	$-\infty$	∞	$-\infty$	–	–	–	−0.922	0.765	−0.765
0	0	1	1	$-\infty$	∞	0	–	–	0	−1.278	1.278	0
0	0	1	2	$-\infty$	∞	∞	–	–	–	−1.688	1.845	0.765
0	0	2	0	$-\infty$	∞	$-\infty$	–	–	–	−0.939	1.878	−1.878
0	0	2	1	$-\infty$	∞	$-\infty$	–	–	–	−0.922	1.845	−0.765
0	0	2	2	$-\infty$	∞	0	–	–	0	−1.189	2.378	0
0	1	0	0	$-\infty$	$-\infty$	∞	–	–	–	−0.922	−0.765	0.765
0	1	0	1	$-\infty$	0	∞	–	0	–	−1.278	0	1.278
0	1	0	2	$-\infty$	∞	∞	–	–	–	−1.688	0.765	1.845
0	1	1	0	−0.674	0	0	−0.44	0	0	−0.504	0	0
0	1	1	1	−1.136	0.757	0.757	−0.696	0.464	0.464	−0.806	0.538	0.538
0	1	1	2	$-\infty$	∞	∞	–	–	–	−1.232	1.232	1.232
0	1	2	0	−0.325	0.649	−0.649	−0.207	0.414	−0.414	−0.248	0.497	−0.497
0	1	2	1	−0.674	1.349	0	−0.44	0.879	0	−0.504	1.008	0
0	1	2	2	$-\infty$	∞	∞	–	–	–	−0.922	1.845	0.765
0	2	0	0	$-\infty$	$-\infty$	∞	–	–	–	−0.939	−1.878	1.878
0	2	0	1	$-\infty$	$-\infty$	∞	–	–	–	−0.922	−0.765	1.845
0	2	0	2	$-\infty$	0	∞	–	0	–	−1.189	0	2.378
0	2	1	0	−0.325	−0.649	0.649	−0.207	−0.414	0.414	−0.248	−0.497	0.497
0	2	1	1	−0.674	0	1.349	−0.44	0	0.879	−0.504	0	1.008
0	2	1	2	$-\infty$	∞	∞	–	–	–	−0.922	0.765	1.845
0	2	2	0	0	0	0	0	0	0	0	0	0
0	2	2	1	−0.325	0.649	0.649	−0.207	0.414	0.414	−0.248	0.497	0.497
0	2	2	2	$-\infty$	∞	∞	–	–	–	−0.939	1.878	1.878
1	0	0	0	0	$-\infty$	$-\infty$	0	–	–	−0.157	−0.765	−0.765
1	0	0	1	−0.674	0	0	−0.44	0	0	−0.504	0	0
1	0	0	2	−0.974	0.649	0.649	−0.621	0.414	0.414	−0.745	0.497	0.497
1	0	1	0	0	0	$-\infty$	0	0	–	0	0	−1.232
1	0	1	1	−0.379	0.757	−0.757	−0.232	0.464	−0.464	−0.269	0.538	−0.538
1	0	1	2	−0.674	1.349	0	−0.44	0.879	0	−0.504	1.008	0
1	0	2	0	0	∞	$-\infty$	0	–	–	0.157	0.765	−1.845
1	0	2	1	0	∞	$-\infty$	0	–	–	0	1.278	−1.278
1	0	2	2	0	∞	$-\infty$	0	–	–	−0.157	1.845	−0.765
1	1	0	0	0	$-\infty$	0	0	–	0	0	−1.232	0
1	1	0	1	−0.379	−0.757	0.757	−0.232	−0.464	0.464	−0.269	−0.538	0.538

continued on next page

Success Counts				Maximum Likelihood estimates			Bias-corrected estimates			Bias-reduced estimates		
y_1	y_2	y_3	y_4	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\alpha}_c$	$\hat{\beta}_c$	$\hat{\gamma}_c$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\gamma}$
0	0	0	0	$-\infty$	0	0	–	0	0	−1.846	0	0
0	0	0	1	$-\infty$	∞	∞	–	–	–	−2.852	1.266	1.266
0	0	0	2	$-\infty$	∞	∞	–	–	–	−3.869	2.304	2.304
0	0	1	0	$-\infty$	∞	$-\infty$	–	–	–	−1.586	1.266	−1.266
0	0	1	1	$-\infty$	∞	0	–	–	0	−2.043	1.78	0
0	0	1	2	$-\infty$	∞	∞	–	–	–	−2.709	2.479	0.985
0	0	2	0	$-\infty$	∞	$-\infty$	–	–	–	−1.565	2.304	−2.304
0	0	2	1	$-\infty$	∞	$-\infty$	–	–	–	−1.723	2.479	−0.985
0	0	2	2	$-\infty$	∞	0	–	–	0	−2.063	2.85	0
0	1	0	0	$-\infty$	$-\infty$	∞	–	–	–	−1.586	−1.266	1.266
0	1	0	1	$-\infty$	0	∞	–	0	–	−2.043	0	1.78
0	1	0	2	$-\infty$	∞	∞	–	–	–	−2.709	0.985	2.479
0	1	1	0	−1.246	0	0	−0.708	0	0	−0.879	0	0
0	1	1	1	−1.673	0.82	0.82	−1.013	0.495	0.495	−1.198	0.589	0.589
0	1	1	2	$-\infty$	∞	∞	–	–	–	−1.693	1.297	1.297
0	1	2	0	−0.875	1.316	−1.316	−0.514	0.759	−0.759	−0.627	0.859	−0.859
0	1	2	1	−1.08	1.709	−0.449	−0.693	1.032	−0.238	−0.815	1.193	−0.233
0	1	2	2	$-\infty$	∞	∞	–	–	–	−1.155	1.716	0.505
0	2	0	0	$-\infty$	$-\infty$	∞	–	–	–	−1.565	−2.304	2.304
0	2	0	1	$-\infty$	$-\infty$	∞	–	–	–	−1.723	−0.985	2.479
0	2	0	2	$-\infty$	0	∞	–	0	–	−2.063	0	2.85
0	2	1	0	−0.875	−1.316	1.316	−0.514	−0.759	0.759	−0.627	−0.859	0.859
0	2	1	1	−1.08	−0.449	1.709	−0.693	−0.238	1.032	−0.815	−0.233	1.193
0	2	1	2	$-\infty$	∞	∞	–	–	–	−1.155	0.505	1.716
0	2	2	0	−0.367	0	0	−0.247	0	0	−0.276	0	0
0	2	2	1	−0.524	0.494	0.494	−0.351	0.322	0.322	−0.423	0.399	0.399
0	2	2	2	$-\infty$	∞	∞	–	–	–	−0.732	1.048	1.048
1	0	0	0	−0.367	$-\infty$	$-\infty$	−0.367	–	–	−0.32	−1.266	−1.266
1	0	0	1	−1.246	0	0	−0.708	0	0	−0.879	0	0
1	0	0	2	−2.191	1.316	1.316	−1.273	0.759	0.759	−1.487	0.859	0.859
1	0	1	0	−0.367	0	$-\infty$	−0.367	0	–	−0.263	0	−1.78
1	0	1	1	−0.852	0.82	−0.82	−0.518	0.495	−0.495	−0.609	0.589	−0.589
1	0	1	2	−1.529	1.709	0.449	−0.93	1.032	0.238	−1.047	1.193	0.233
1	0	2	0	−0.367	∞	$-\infty$	−0.367	–	–	−0.229	0.985	−2.479
1	0	2	1	−0.367	∞	$-\infty$	−0.367	–	–	−0.396	1.297	−1.297
1	0	2	2	−0.367	∞	$-\infty$	−0.367	–	–	−0.65	1.716	−0.505
1	1	0	0	−0.367	$-\infty$	0	−0.367	–	0	−0.263	−1.78	0
1	1	0	1	−0.852	−0.82	0.82	−0.518	−0.495	0.495	−0.609	−0.589	0.589

Table C.3: Complementary log-log link. Maximum likelihood estimates, bias-corrected estimates and bias-reduced estimates for (α, β, γ) to three decimal places, for every possible data configuration in Table 4.1 with $m_1 = m_2 = m_3 = m_4 = 2$.

Success Counts				Maximum Likelihood estimates			Bias-corrected estimates			Bias-reduced estimates		
y_1	y_2	y_3	y_4	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\alpha}_c$	$\hat{\beta}_c$	$\hat{\gamma}_c$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\gamma}$
0	0	0	0	$-\infty$	0	0	–	0	0	−1.846	0	0
0	0	0	1	$-\infty$	∞	∞	–	–	–	−2.852	1.266	1.266
0	0	0	2	$-\infty$	∞	∞	–	–	–	−3.869	2.304	2.304
0	0	1	0	$-\infty$	∞	$-\infty$	–	–	–	−1.586	1.266	−1.266
0	0	1	1	$-\infty$	∞	0	–	–	0	−2.043	1.78	0
0	0	1	2	$-\infty$	∞	∞	–	–	–	−2.709	2.479	0.985
0	0	2	0	$-\infty$	∞	$-\infty$	–	–	–	−1.565	2.304	−2.304
0	0	2	1	$-\infty$	∞	$-\infty$	–	–	–	−1.723	2.479	−0.985
0	0	2	2	$-\infty$	∞	0	–	–	0	−2.063	2.85	0
0	1	0	0	$-\infty$	$-\infty$	∞	–	–	–	−1.586	−1.266	1.266
0	1	0	1	$-\infty$	0	∞	–	0	–	−2.043	0	1.78
0	1	0	2	$-\infty$	∞	∞	–	–	–	−2.709	0.985	2.479
0	1	1	0	−1.246	0	0	−0.708	0	0	−0.879	0	0
0	1	1	1	−1.673	0.82	0.82	−1.013	0.495	0.495	−1.198	0.589	0.589
0	1	1	2	$-\infty$	∞	∞	–	–	–	−1.693	1.297	1.297
0	1	2	0	−0.875	1.316	−1.316	−0.514	0.759	−0.759	−0.627	0.859	−0.859
0	1	2	1	−1.08	1.709	−0.449	−0.693	1.032	−0.238	−0.815	1.193	−0.233
0	1	2	2	$-\infty$	∞	∞	–	–	–	−1.155	1.716	0.505
0	2	0	0	$-\infty$	$-\infty$	∞	–	–	–	−1.565	−2.304	2.304
0	2	0	1	$-\infty$	$-\infty$	∞	–	–	–	−1.723	−0.985	2.479
0	2	0	2	$-\infty$	0	∞	–	0	–	−2.063	0	2.85
0	2	1	0	−0.875	−1.316	1.316	−0.514	−0.759	0.759	−0.627	−0.859	0.859
0	2	1	1	−1.08	−0.449	1.709	−0.693	−0.238	1.032	−0.815	−0.233	1.193
0	2	1	2	$-\infty$	∞	∞	–	–	–	−1.155	0.505	1.716
0	2	2	0	−0.367	0	0	−0.247	0	0	−0.276	0	0
0	2	2	1	−0.524	0.494	0.494	−0.351	0.322	0.322	−0.423	0.399	0.399
0	2	2	2	$-\infty$	∞	∞	–	–	–	−0.732	1.048	1.048
1	0	0	0	−0.367	$-\infty$	$-\infty$	−0.367	–	–	−0.32	−1.266	−1.266
1	0	0	1	−1.246	0	0	−0.708	0	0	−0.879	0	0
1	0	0	2	−2.191	1.316	1.316	−1.273	0.759	0.759	−1.487	0.859	0.859
1	0	1	0	−0.367	0	$-\infty$	−0.367	0	–	−0.263	0	−1.78
1	0	1	1	−0.852	0.82	−0.82	−0.518	0.495	−0.495	−0.609	0.589	−0.589
1	0	1	2	−1.529	1.709	0.449	−0.93	1.032	0.238	−1.047	1.193	0.233
1	0	2	0	−0.367	∞	$-\infty$	−0.367	–	–	−0.229	0.985	−2.479
1	0	2	1	−0.367	∞	$-\infty$	−0.367	–	–	−0.396	1.297	−1.297
1	0	2	2	−0.367	∞	$-\infty$	−0.367	–	–	−0.65	1.716	−0.505
1	1	0	0	−0.367	$-\infty$	0	−0.367	–	0	−0.263	−1.78	0
1	1	0	1	−0.852	−0.82	0.82	−0.518	−0.495	0.495	−0.609	−0.589	0.589

continued on next page

Success Counts				Maximum Likelihood estimates			Bias-corrected estimates			Bias-reduced estimates		
y_1	y_2	y_3	y_4	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\alpha}_c$	$\hat{\beta}_c$	$\hat{\gamma}_c$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\gamma}$
1	1	0	2	−1.529	0.449	1.709	−0.93	0.238	1.032	−1.047	0.233	1.193
1	1	1	0	−0.032	−0.82	−0.82	−0.024	−0.495	−0.495	−0.02	−0.589	−0.589
1	1	1	1	−0.367	0	0	−0.247	0	0	−0.276	0	0
1	1	1	2	−0.936	0.941	0.941	−0.541	0.49	0.49	−0.634	0.615	0.615
1	1	2	0	0.18	0.449	−1.709	0.102	0.238	−1.032	0.146	0.233	−1.193
1	1	2	1	0.005	0.941	−0.941	−0.05	0.49	−0.49	−0.019	0.615	−0.615
1	1	2	2	−0.367	∞	0	−0.367	−	0	−0.278	1.119	0
1	2	0	0	−0.367	$-\infty$	∞	−0.367	−	−	−0.229	−2.479	0.985
1	2	0	1	−0.367	$-\infty$	∞	−0.367	−	−	−0.396	−1.297	1.297
1	2	0	2	−0.367	$-\infty$	∞	−0.367	−	−	−0.65	−0.505	1.716
1	2	1	0	0.18	−1.709	0.449	0.102	−1.032	0.238	0.146	−1.193	0.233
1	2	1	1	0.005	−0.941	0.941	−0.05	−0.49	0.49	−0.019	−0.615	0.615
1	2	1	2	−0.367	0	∞	−0.367	0	−	−0.278	0	1.119
1	2	2	0	0.464	−0.494	−0.494	0.292	−0.322	−0.322	0.375	−0.399	−0.399
1	2	2	1	0.327	0	0	0.214	0	0	0.244	0	0
1	2	2	2	−0.367	∞	∞	−0.367	−	−	0.007	0.563	0.563
2	0	0	0	∞	$-\infty$	$-\infty$	−	−	−	0.738	−2.304	−2.304
2	0	0	1	0.44	−1.316	−1.316	0.244	−0.759	−0.759	0.232	−0.859	−0.859
2	0	0	2	−0.367	0	0	−0.247	0	0	−0.276	0	0
2	0	1	0	∞	$-\infty$	$-\infty$	−	−	−	0.756	−0.985	−2.479
2	0	1	1	0.629	−0.449	−1.709	0.34	−0.238	−1.032	0.378	−0.233	−1.193
2	0	1	2	−0.03	0.494	−0.494	−0.029	0.322	−0.322	−0.024	0.399	−0.399
2	0	2	0	∞	0	$-\infty$	−	0	−	0.787	0	−2.85
2	0	2	1	∞	∞	$-\infty$	−	−	−	0.561	0.505	−1.716
2	0	2	2	∞	∞	$-\infty$	−	−	−	0.316	1.048	−1.048
2	1	0	0	∞	$-\infty$	$-\infty$	−	−	−	0.756	−2.479	−0.985
2	1	0	1	0.629	−1.709	−0.449	0.34	−1.032	−0.238	0.378	−1.193	−0.233
2	1	0	2	−0.03	−0.494	0.494	−0.029	−0.322	0.322	−0.024	−0.399	0.399
2	1	1	0	∞	$-\infty$	$-\infty$	−	−	−	0.9	−1.297	−1.297
2	1	1	1	0.946	−0.941	−0.941	0.44	−0.49	−0.49	0.595	−0.615	−0.615
2	1	1	2	0.327	0	0	0.214	0	0	0.244	0	0
2	1	2	0	∞	$-\infty$	$-\infty$	−	−	−	1.066	−0.505	−1.716
2	1	2	1	∞	0	$-\infty$	−	0	−	0.841	0	−1.119
2	1	2	2	∞	∞	$-\infty$	−	−	−	0.569	0.563	−0.563
2	2	0	0	∞	$-\infty$	0	−	−	0	0.787	−2.85	0
2	2	0	1	∞	$-\infty$	∞	−	−	−	0.561	−1.716	0.505
2	2	0	2	∞	$-\infty$	∞	−	−	−	0.316	−1.048	1.048
2	2	1	0	∞	$-\infty$	$-\infty$	−	−	−	1.066	−1.716	−0.505
2	2	1	1	∞	$-\infty$	0	−	−	0	0.841	−1.119	0
2	2	1	2	∞	$-\infty$	∞	−	−	−	0.569	−0.563	0.563
2	2	2	0	∞	$-\infty$	$-\infty$	−	−	−	1.364	−1.048	−1.048
2	2	2	1	∞	$-\infty$	$-\infty$	−	−	−	1.132	−0.563	−0.563
2	2	2	2	∞	0	0	−	0	0	0.848	0	0

Table C.4: Log-log link. Maximum likelihood estimates, bias-corrected estimates and bias-reduced estimates for (α, β, γ) to three decimal places, for every possible data configuration in Table 4.1 with $m_1 = m_2 = m_3 = m_4 = 2$.

Success Counts				Maximum Likelihood estimates			Bias-corrected estimates			Bias-reduced estimates		
y_1	y_2	y_3	y_4	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\alpha}_c$	$\hat{\beta}_c$	$\hat{\gamma}_c$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\gamma}$
0	0	0	0	$-\infty$	0	0	–	0	0	−0.848	0	0
0	0	0	1	$-\infty$	∞	∞	–	–	–	−1.132	0.563	0.563
0	0	0	2	$-\infty$	∞	∞	–	–	–	−1.365	1.049	1.049
0	0	1	0	$-\infty$	∞	$-\infty$	–	–	–	−0.569	0.563	−0.563
0	0	1	1	$-\infty$	∞	0	–	–	0	−0.841	1.119	0
0	0	1	2	$-\infty$	∞	∞	–	–	–	−1.066	1.716	0.505
0	0	2	0	$-\infty$	∞	$-\infty$	–	–	–	−0.316	1.049	−1.049
0	0	2	1	$-\infty$	∞	$-\infty$	–	–	–	−0.561	1.716	−0.505
0	0	2	2	$-\infty$	∞	0	–	–	0	−0.787	2.85	0
0	1	0	0	$-\infty$	$-\infty$	∞	–	–	–	−0.569	−0.563	0.563
0	1	0	1	$-\infty$	0	∞	–	0	–	−0.841	0	1.119
0	1	0	2	$-\infty$	∞	∞	–	–	–	−1.066	0.505	1.716
0	1	1	0	−0.327	0	0	−0.214	0	0	−0.244	0	0
0	1	1	1	−0.946	0.941	0.941	−0.44	0.49	0.49	−0.595	0.615	0.615
0	1	1	2	$-\infty$	∞	∞	–	–	–	−0.9	1.297	1.297
0	1	2	0	0.03	0.494	−0.494	0.029	0.322	−0.322	0.024	0.399	−0.399
0	1	2	1	−0.629	1.709	0.449	−0.34	1.032	0.238	−0.378	1.193	0.233
0	1	2	2	$-\infty$	∞	∞	–	–	–	−0.756	2.479	0.985
0	2	0	0	$-\infty$	$-\infty$	∞	–	–	–	−0.316	−1.049	1.049
0	2	0	1	$-\infty$	$-\infty$	∞	–	–	–	−0.561	−0.505	1.716
0	2	0	2	$-\infty$	0	∞	–	0	–	−0.787	0	2.85
0	2	1	0	0.03	−0.494	0.494	0.029	−0.322	0.322	0.024	−0.399	0.399
0	2	1	1	−0.629	0.449	1.709	−0.34	0.238	1.032	−0.378	0.233	1.193
0	2	1	2	$-\infty$	∞	∞	–	–	–	−0.756	0.985	2.479
0	2	2	0	0.367	0	0	0.247	0	0	0.276	0	0
0	2	2	1	−0.44	1.316	1.316	−0.244	0.759	0.759	−0.232	0.859	0.859
0	2	2	2	$-\infty$	∞	∞	–	–	–	−0.738	2.304	2.304
1	0	0	0	0.367	$-\infty$	$-\infty$	0.367	–	–	−0.007	−0.563	−0.563
1	0	0	1	−0.327	0	0	−0.214	0	0	−0.244	0	0
1	0	0	2	−0.464	0.494	0.494	−0.292	0.322	0.322	−0.375	0.399	0.399
1	0	1	0	0.367	0	$-\infty$	0.367	0	–	0.278	0	−1.119
1	0	1	1	−0.005	0.941	−0.941	0.05	0.49	−0.49	0.019	0.615	−0.615
1	0	1	2	−0.18	1.709	−0.449	−0.102	1.032	−0.238	−0.146	1.193	−0.233
1	0	2	0	0.367	∞	$-\infty$	0.367	–	–	0.65	0.505	−1.716
1	0	2	1	0.367	∞	$-\infty$	0.367	–	–	0.396	1.297	−1.297
1	0	2	2	0.367	∞	$-\infty$	0.367	–	–	0.229	2.479	−0.985
1	1	0	0	0.367	$-\infty$	0	0.367	–	0	0.278	−1.119	0
1	1	0	1	−0.005	−0.941	0.941	0.05	−0.49	0.49	0.019	−0.615	0.615

continued on next page

Success Counts				Maximum Likelihood estimates			Bias-corrected estimates			Bias-reduced estimates		
y_1	y_2	y_3	y_4	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\alpha}_c$	$\hat{\beta}_c$	$\hat{\gamma}_c$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\gamma}$
1	1	0	2	−0.18	−0.449	1.709	−0.102	−0.238	1.032	−0.146	−0.233	1.193
1	1	1	0	0.936	−0.941	−0.941	0.541	−0.49	−0.49	0.634	−0.615	−0.615
1	1	1	1	0.367	0	0	0.247	0	0	0.276	0	0
1	1	1	2	0.032	0.82	0.82	0.024	0.495	0.495	0.02	0.589	0.589
1	1	2	0	1.529	−0.449	−1.709	0.93	−0.238	−1.032	1.047	−0.233	−1.193
1	1	2	1	0.852	0.82	−0.82	0.518	0.495	−0.495	0.609	0.589	−0.589
1	1	2	2	0.367	∞	0	0.367	−	0	0.263	1.78	0
1	2	0	0	0.367	−∞	∞	0.367	−	−	0.65	−1.716	0.505
1	2	0	1	0.367	−∞	∞	0.367	−	−	0.396	−1.297	1.297
1	2	0	2	0.367	−∞	∞	0.367	−	−	0.229	−0.985	2.479
1	2	1	0	1.529	−1.709	−0.449	0.93	−1.032	−0.238	1.047	−1.193	−0.233
1	2	1	1	0.852	−0.82	0.82	0.518	−0.495	0.495	0.609	−0.589	0.589
1	2	1	2	0.367	0	∞	0.367	0	−	0.263	0	1.78
1	2	2	0	2.191	−1.316	−1.316	1.273	−0.759	−0.759	1.487	−0.859	−0.859
1	2	2	1	1.246	0	0	0.708	0	0	0.879	0	0
1	2	2	2	0.367	∞	∞	0.367	−	−	0.32	1.266	1.266
2	0	0	0	∞	−∞	−∞	−	−	−	0.732	−1.049	−1.049
2	0	0	1	0.524	−0.494	−0.494	0.351	−0.322	−0.322	0.423	−0.399	−0.399
2	0	0	2	0.367	0	0	0.247	0	0	0.276	0	0
2	0	1	0	∞	−∞	−∞	−	−	−	1.155	−0.505	−1.716
2	0	1	1	1.08	0.449	−1.709	0.693	0.238	−1.032	0.815	0.233	−1.193
2	0	1	2	0.875	1.316	−1.316	0.514	0.759	−0.759	0.627	0.859	−0.859
2	0	2	0	∞	0	−∞	−	0	−	2.063	0	−2.85
2	0	2	1	∞	∞	−∞	−	−	−	1.723	0.985	−2.479
2	0	2	2	∞	∞	−∞	−	−	−	1.565	2.304	−2.304
2	1	0	0	∞	−∞	−∞	−	−	−	1.155	−1.716	−0.505
2	1	0	1	1.08	−1.709	0.449	0.693	−1.032	0.238	0.815	−1.193	0.233
2	1	0	2	0.875	−1.316	1.316	0.514	−0.759	0.759	0.627	−0.859	0.859
2	1	1	0	∞	−∞	−∞	−	−	−	1.693	−1.297	−1.297
2	1	1	1	1.673	−0.82	−0.82	1.013	−0.495	−0.495	1.198	−0.589	−0.589
2	1	1	2	1.246	0	0	0.708	0	0	0.879	0	0
2	1	2	0	∞	−∞	−∞	−	−	−	2.709	−0.985	−2.479
2	1	2	1	∞	0	−∞	−	0	−	2.043	0	−1.78
2	1	2	2	∞	∞	−∞	−	−	−	1.586	1.266	−1.266
2	2	0	0	∞	−∞	0	−	−	0	2.063	−2.85	0
2	2	0	1	∞	−∞	∞	−	−	−	1.723	−2.479	0.985
2	2	0	2	∞	−∞	∞	−	−	−	1.565	−2.304	2.304
2	2	1	0	∞	−∞	−∞	−	−	−	2.709	−2.479	−0.985
2	2	1	1	∞	−∞	0	−	−	0	2.043	−1.78	0
2	2	1	2	∞	−∞	∞	−	−	−	1.586	−1.266	1.266
2	2	2	0	∞	−∞	−∞	−	−	−	3.869	−2.304	−2.304
2	2	2	1	∞	−∞	−∞	−	−	−	2.852	−1.266	−1.266
2	2	2	2	∞	0	0	−	0	0	1.846	0	0

Bibliography

Agresti, A. (1999). On logit confidence intervals for the odds ratio with small samples. *Biometrics* **55**, 597–602.

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.

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

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

Barndorff-Nielsen, O. E. & Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. London: Chapman & Hall Ltd.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*, Eds. F. Lord & M. Novick. Reading, MA: Addison-Wesley.

Bishop, Y. M., Fienberg, S. E. & Holland, P. W. (1975). *Discrete Multivariate Analysis : Theory and Practice*. Cambridge, Massachutetts: M.I.T. Press.

Breslow, N. E. & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91.

Brown, L. D., Cai, T. T. & DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statistical Science* **16**, 101–117.

Broyden, C. (1970). The convergence of double-rank minimization algorithms. 1. General considerations. *Journal of the Institute of Mathematics and Its Applications* **6**, 76–90.

Bull, S. B., Greenwood, C. M. T. & Hauck, W. W. (1997). Jackknife bias reduction for polychotomous logistic regression (Corr: 97V16 p2928). *Statistics in Medicine* **16**, 545–560.

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

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

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B. & Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association* **86**, 68–78.

Copas, J. B. (1988). Binary regression models for contaminated data (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological* **50**, 225–265.

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

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

Cox, D. R. & Snell, E. J. (1989). *Analysis of Binary Data*. London: Chapman and Hall, 2nd edn.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princenton: Princenton University Press.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *The Annals of Statistics* **3**, 1189–1217.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer.

Farewell, V. T. (1978). Jackknife estimation with structured data. *Biometrika* **65**, 444–447.

Firth, D. (1992a). Bias reduction, the Jeffreys prior and GLIM. In *Advances in GLIM and statistical modelling: Proceedings of the GLIM 92 conference, Munich*, Eds. L. Fahrmeir, B. Francis, R. Gilchrist & G. Tutz, pp. 91–100. New York: Springer.

Firth, D. (1992b). Generalized linear models and Jeffreys priors: An iterative generalized least-squares approach. In *Computational Statistics I*, Eds. Y. Dodge & J. Whittaker. Heidelberg: Physica-Verlag.

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

Firth, D. (2003). Overcoming the reference category problem in the presentation of statistical models. *Sociological Methodology* **33**, 1–18.

Firth, D. & De Menezes, R. X. (2004). Quasi-variances. *Biometrika* **91**, 65–80.

Fletcher, R. (1970). A new approach to variable metric algorithms. *Computer Journal* **13**, 317–322.

Gart, J. J. (1966). Alternative analyses of contingency tables. *Journal of the Royal Statistical Society, Series B: Methodological* **28**, 164–179.

Gart, J. J., Pettigrew, H. M. & Thomas, D. G. (1985). The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. *Biometrika* **72**, 179–190.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation* **24**, 23–26.

Haldane, J. (1956). The estimation of the logarithm of a ratio of frequencies. *Annals of Human Genetics* **20**, 309–311.

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

Hoijtink, H. & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In *Rasch Models: Foundations, Recent Developments and Applications*, Eds. G. H. Fisher & I. W. Molenaar, chap. 4, pp. 54–68. Springer.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London* **186**, 453–461.

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

Lesaffre, E. & Albert, A. (1989). Partial separation in logistic discrimination. *Journal of the Royal Statistical Society, Series B: Methodological* **51**, 109–116.

Magnus, J. R. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester: Wiley.

Mann, H. B. & Wald, A. (1943). On stochastic limit and order relationships. *The Annals of Mathematical Statistics* **14**, 217–226.

McCullagh, P. (1984). Tensor notation and cumulants of polynomials. *Biometrika* **71**, 461–476.

McCullagh, P. (1987). *Tensor Methods in Statistics*. London: Chapman and Hall.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. London: Chapman and Hall, 2nd edn.

Mehrabi, Y. & Matthews, J. N. S. (1995). Likelihood-based methods for bias reduction in limiting dilution assays. *Biometrics* **51**, 1543–1549.

Molenaar, I. W. (1995). Estimation of item parameters. In *Rasch Models: Foundations, Recent Developments and Applications*, Eds. G. H. Fisher & I. W. Molenaar, chap. 3, pp. 39–51. Springer.

Monahan, J. (2001). *Numerical Methods of Statistics*. Cambridge: Cambridge University Press.

Nelder, J. A. & Wedderburn, W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A: General* **135**, 370–384.

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

Palmgren, J. (1981). The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables. *Biometrika* **68**, 563–566.

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

Pettitt, A. N., Kelly, J. M. & Gao, J. T. (1998). Bias correction for censored data with exponential lifetimes. *Statistica Sinica* **8**, 941–964.

Poirier, D. (1994). Jeffrey's prior for logit models. *Journal of Econometrics* **63**, 327–339.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika* **43**, 353–360.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rubin, D. B. & Schenker, N. (1987). Logit-based interval estimation for binomial data using the Jeffreys prior. *Sociological Methodology* **17**, 131–144.

Santner, T. J. & Duffy, D. E. (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **73**, 755–758.

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.

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

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation* **24**, 647–656.

Skovgaard, I. (1986). A note on the differentiation of cumulants of log-likelihood derivatives. *International Statistical Review* **54**, 29–32.

Sowden, R. R. (1972). On the first-order bias of parameter estimates in a quantal response model under alternative estimation procedures. *Biometrika* **59**, 573–579.

Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., den Noorgate, W. V., Meulders, M. & Boeck, P. D. (2005). Estimation and software. In *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, Eds. P. D. Boeck & M. Wilson, chap. 12. Springer.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika* **54**, 427–450.

Wei, B. (1997). *Exponential Family Nonlinear Models*. New York: Springer-Verlag Inc.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics* **19**, 251–253.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis* **13**, 157–170.