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*.*

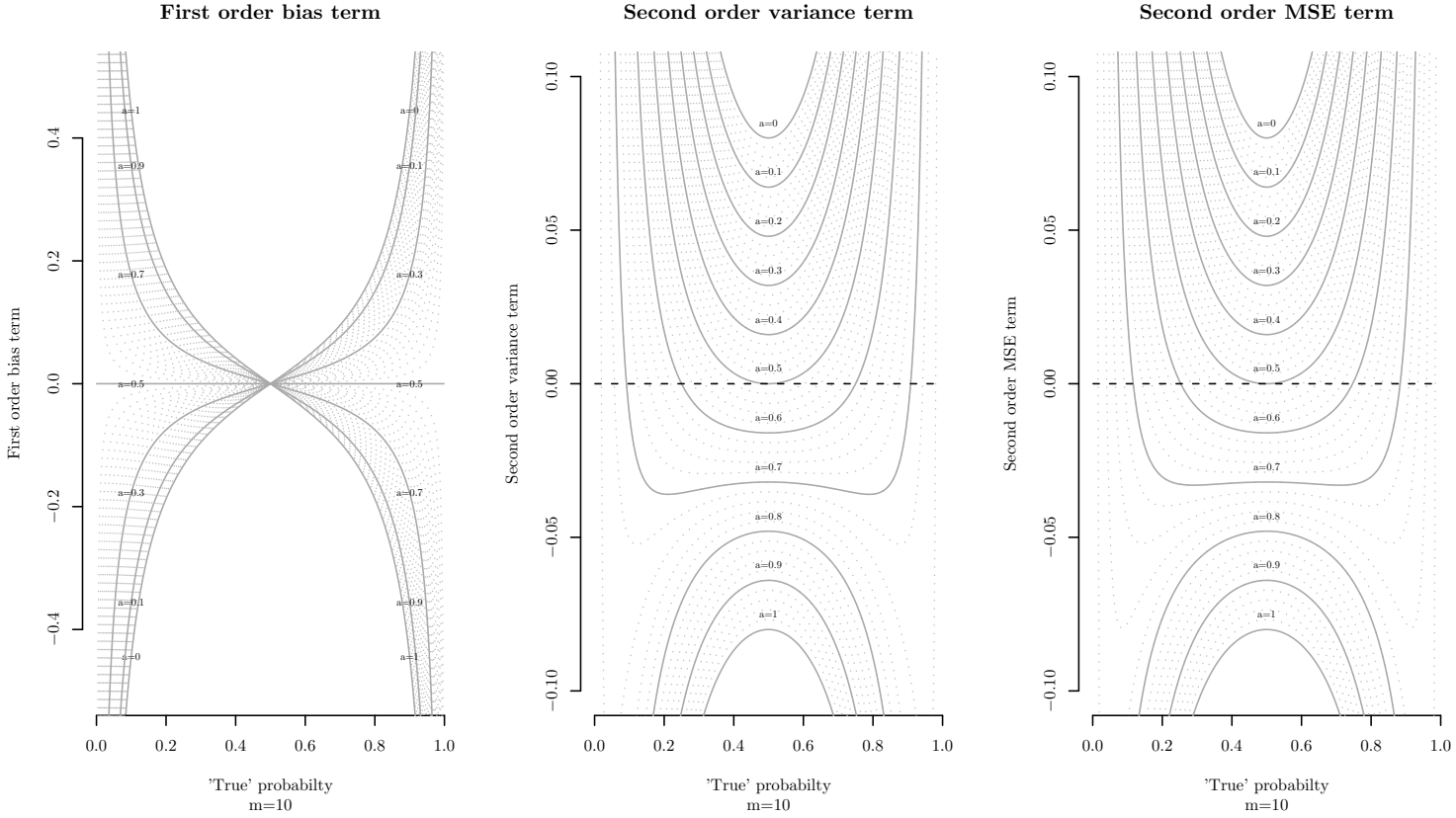
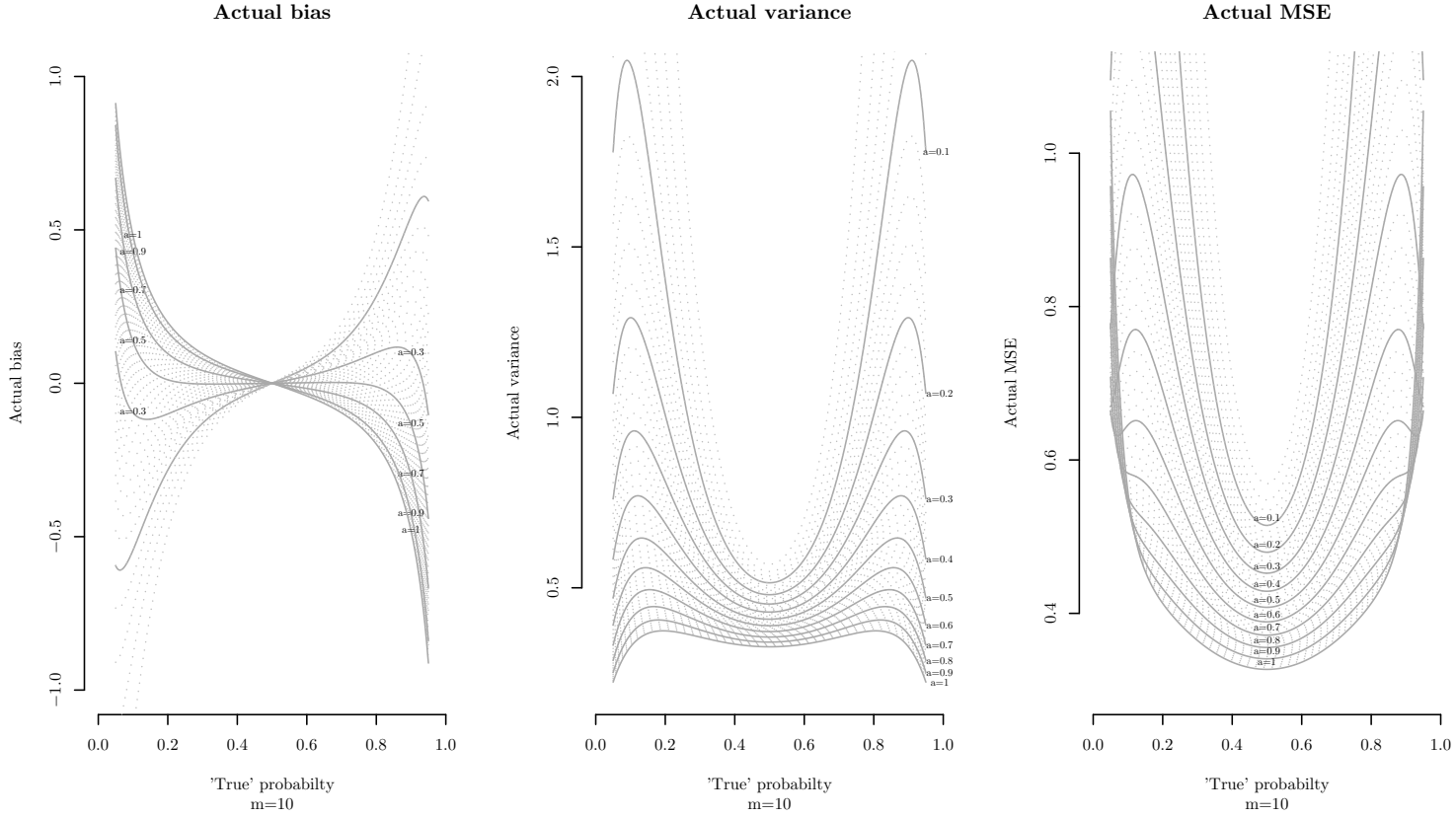


Figure 4.2: Actual bias, actual MSE and actual variance of $\tilde{\beta}^{(a)}$, for a grid of values of $a \in (0, 1]$ against the true probability of success. The dotted curves represent values of a between the reported ones and with step 0.02.



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.

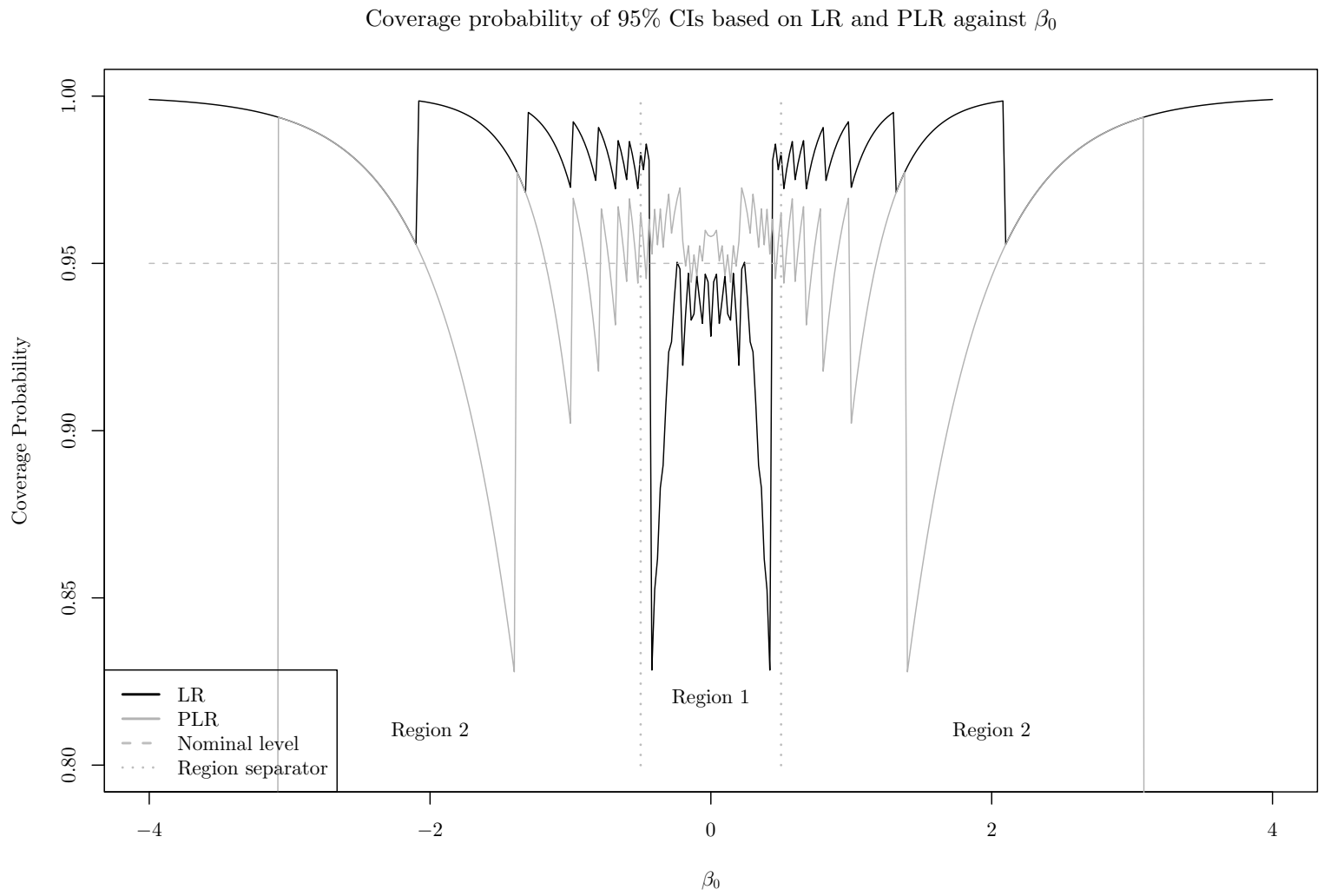


Figure 4.4: Coverage probability of the 95 *per cent confidence interval defined as the union of the intervals* $C_{LR}(y, 0.05)$ *and* $C_{PLR}(y, 0.05)$*, for a fine grid of values of the true parameter* β_0*.*