Mean and median bias reduction: A concise review and application to adjacent-categories logit models

Ioannis Kosmidis
ioannis.kosmidis@warwick.ac.uk

Department of Statistics, University of Warwick
Coventry CV4 7AL, UK

January 25, 2022

Abstract

The estimation of categorical response models using bias-reducing adjusted score equations has seen extensive theoretical research and applied use. The resulting estimates have been found to have superior frequentist properties to what maximum likelihood generally delivers and to be finite, even in cases where the maximum likelihood estimates are infinite. We briefly review mean and median bias reduction of maximum likelihood estimates via adjusted score equations in an illustration-driven way, and discuss their particular equivariance properties under parameter transformations. We then apply mean and median bias reduction to adjacent-categories logit models for ordinal responses. We show how ready bias reduction procedures for Poisson log-linear models can be used for mean and median bias reduction in adjacent-categories logit models with proportional odds and mean bias-reduced estimation in models with non-proportional odds. As in binomial logistic regression, the reduced-bias estimates are found to be finite even in cases where the maximum likelihood estimates are infinite. We also use the approximation of the bias of transformations of mean bias-reduced estimators to correct for the mean bias of model-based ordinal superiority measures. All developments are motivated and illustrated using real-data case studies and simulations.

Keywords: infinite estimates; bias reduction; adjusted score equations; data separation

1 Overview

The first part of this chapter provides an example-driven, concise review of the developments in a fast growing body of literature about mean and median bias reduction (BR) in parametric estimation via adjusted score equations; see Firth (1993) for mean BR (mBR) and Kenne Pagui et al. (2017) for median BR (mdBR). Particular focus is placed on how these methods can be used as a remedy for the numerical and inferential consequences of boundary maximum likelihood (ML) estimates in categorical response models, which are illustrated in Section 2. Sections 3 and 4 describe how the mean and median bias of the ML estimator can be reduced in general parametric models through the appropriate adjustment of the gradient of the log-likelihood. Section 5 discusses the validity of inference when the ML estimates are replaced by mBR or mdBR estimates in standard first-order procedures. Section 6 takes a close look at the equivariance properties of mBR and mdBR estimators under transformation of the model parameters. We also present an approximation of the bias of general transformations of mBR estimators, which can be used to correct for the bias of transformations of the model parameters using only the mBR estimates, the second derivatives of the transformation, and the expected information matrix. The bias approximation is used to get mBR estimates of odds ratios from mBR estimates of regression coefficients in logistic regression models.
The second part of this chapter uses the results from the first to develop, for the first time, mBR and mdBR procedures for adjacent-categories logit (ACL) models for ordinal responses (see, for example, Agresti, 2010, Chapter 4 for an introduction). Section 7 reviews the proportional odds (PO) and non-proportional odds (NPO) versions of the ACL models, and their key properties, including their equivalence to baseline-category logit (BCL) models, and discusses how that equivalence can be exploited for ML estimation. A real-data case study is used to illustrate that boundary estimates can also cause numerical and inferential issues for ACL models. Section 8 then details how and when the equivariance properties of mBR and mdBR, and implementations of the latter for BCL models, can be used for mBR and mdBR for the PO and NPO versions of ACL models. Finally, Section 9 details how the mBR estimates can be used for the explicit correction of the estimates of ordinal superiority summaries.

2 Boundary estimates in categorical response models

It is well known that ML estimation of regression models with categorical responses may result in estimates on the boundary of the parameter space. The data patterns that result in boundary estimates in general multinomial logistic regression models (also known as baseline category models models; see Agresti, 2002, Section 7.1) have been studied extensively and are completely characterized. For a range of binomial regression models, Silvapulle (1981) proves that a certain degree of “overlap” on the data is a necessary and sufficient condition for the ML estimates to have finite values. Albert and Anderson (1984) enrich the arguments in Silvapulle (1981) generalizing the results in the case of baseline-category logit (BCL) models for nominal responses. In particular, Albert and Anderson (1984) categorize the possible configurations for the sample points into complete separation, quasi-complete separation, and overlap, and then show that separation is necessary and sufficient for the ML estimate to have at least one infinite-valued component. Geometric representations of (quasi-)complete separation for binomial logistic regression — when the ACL and BCL models reduce to exactly the same form — are given in Albert and Anderson (1984, Figure 1), and for multinomial responses in Lesaffre and Albert (1989, Figure 1).

Example 1: Separation in logistic regression A simple illustration of a completely separated data set is shown in Figure 1. The data consists of 100 realizations of two continuous covariates $x_2$ and $x_3$, and a response $y$ that ends up being 0 whenever $x_2 + 2x_3 > 0$. ML estimation of the logistic regression model with $\log \left\{ \frac{\pi}{1 - \pi} \right\} = \beta_1 + \beta_2 x_2 + \beta_3 x_3$, where $\pi$ is the probability of observing $y = 1$ given $x$, results in the estimated logistic discriminant line in Figure 1, with the log-likelihood attaining its global maximum value of 0, and the fitted value 0 being assigned to all observations with $y = 0$, and 1 to the rest. The detectseparation R package (Kosmidis and Schumacher, 2021) that implements the methods in the unpublished PhD thesis by Konis (2007) can be used to show that the ML estimates of $\beta_1$, $\beta_2$ and $\beta_3$ are $-\infty$, $+\infty$ and $+\infty$ respectively.

While there is no ambiguity in reporting infinite estimates, estimates on the boundary of the parameter space can i) cause numerical instabilities to fitting procedures, ii) lead to misleading output when estimation is based on iterative procedures with a stopping criterion, and more importantly, iii) cause havoc to asymptotic inferential procedures, and especially to the ones that depend on estimates of the standard error of the estimators (for example, Wald tests and related confidence intervals), oftentimes leading to wrong inferences. For example, the ML estimates in Table 1 have been obtained using the glm() function in R (R Core Team, 2021). Despite the fact that the ML estimates for $\beta_1$, $\beta_2$ and $\beta_3$ are in reality infinite, the stopping criteria of the fitting procedure that glm() implements are met for finite values of the parameters, which are returned. The reported estimated standard errors are also finite and substantially larger.
Table 1: ML, mBR and mdBR estimates for the logistic regression model in Example 1. The estimated standard errors (S.E.) are based on the expected information matrix at the estimates. The $z$-statistic is computed as estimate over estimated S.E., and the $p$-value is computed as $2 \min(\Phi(z), 1 - \Phi(z))$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal distribution.

| Parameter | Estimate | Estimated S.E. | $z$-statistic | $p$-value |
|-----------|----------|----------------|-------------|-----------|
| Maximum likelihood | | | | |
| $\beta_1$ | -22.397 | 13879.616 | -0.002 | 0.999 |
| $\beta_2$ | 62.578 | 20968.761 | 0.003 | 0.998 |
| $\beta_3$ | 132.228 | 44964.541 | 0.003 | 0.998 |
| Mean bias reduction | | | | |
| $\beta_1$ | -2.001 | 1.552 | -1.289 | 0.197 |
| $\beta_2$ | 5.266 | 1.997 | 2.637 | 0.008 |
| $\beta_3$ | 11.166 | 3.984 | 2.803 | 0.005 |
| Median bias reduction | | | | |
| $\beta_1$ | -2.583 | 1.984 | -1.302 | 0.193 |
| $\beta_2$ | 6.325 | 2.628 | 2.406 | 0.016 |
| $\beta_3$ | 13.321 | 5.293 | 2.517 | 0.012 |

Figure 1: The data described in Example 1. The dashed line is the line $0 = \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$, where the fitted probabilities are all 0.5.

than the estimates. This results in small, in absolute value, $z$-statistics, and hence no evidence against the individual hypotheses $\beta_2 = 0$ and $\beta_3 = 0$; one would expect at least some evidence against the hypotheses given that the value of the response has been fully determined from the values of $x_2$ and $x_3$.

One way to circumvent the numerical and inferential issues associated with boundary ML estimates is to replace ML with an alternative estimation method that i) has comparable or
sometimes better asymptotic properties than the ML estimator generally does, and ii) tends to result or results in estimates away from the boundary of the parameter space. Popular examples of such alternative estimation methods are the mean bias-reducing adjusted score functions approach in Firth (1993), and the median bias-reducing adjusted score functions approach in Kenne Pagui et al. (2017), which we briefly review in Section 3, Section 4, and Section 6.

3 Mean bias reduction

Let $\ell(\theta)$ be the log-likelihood about a parameter vector $\theta$ with $\theta \in \mathbb{R}^v$. Assuming that the model at hand is appropriate, then under fairly general regularity conditions about the model, the ML estimator $\hat{\theta} = \arg \max \ell(\theta)$ has mean bias $E_\theta(\hat{\theta} - \theta) = O(N^{-1})$, where $N$ is a measure of information about $\theta$, usually — but not necessarily — the sample size.

If $S(\theta) = \nabla \ell(\theta)$, Firth (1993) shows that we can define an alternative estimator $\theta^*$ with mean bias $E_\theta(\theta^* - \theta) = O(N^{-2})$, which is asymptotically smaller than the bias of $\hat{\theta}$, as the solution of

$$S(\theta) + A(\theta) = 0_v,$$

where

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

In the above expression, $P_t(\theta) = E_\theta(S(\theta)S(\theta)^\top S_t(\theta))$ and $Q_t(\theta) = -E_\theta(j(\theta)S_t(\theta))$, and $j(\theta) = -\nabla \nabla^\top \ell(\theta)$ and $i(\theta) = E_\theta(S(\theta)S(\theta)^\top)$ are the observed and expected information matrix about $\theta$, respectively, with all expectations taken with respect to the model.

Mean bias reduction has been found to result in estimates away from the boundary of the parameter space in a range of categorical data models; see, for example, Firth (1993) and Heinze and Schumper (2002) for binomial logistic regression; Mehrabi and Matthews (1995) for the estimation of simple complementary log–log-models; Kosmidis and Firth (2009, Section 6) for row-column association models; Bull et al. (2002), Kosmidis and Firth (2011), and Kosmidis et al. (2020, Section 6) for BCL models; and Kosmidis (2014) for cumulative link models.

If $\theta$ is the canonical parameter of a full exponential family (see Pace and Salvan, 1997, Chapter 5), like in binomial and multinomial logistic regression, then $j(\theta) = i(\theta)$ and $j(\theta)$ does not depend on the stochastic part of the model. Hence, $Q_t(\theta) = 0_{v \times v}$, where $0_{v \times v}$ is a $v \times v$ matrix of zeros, and some algebra (see Firth, 1993, Section 3) gives that the solution of the mean bias-reducing adjusted score equations (1) is equivalent to the maximization of the penalized log-likelihood

$$\ell(\theta) + \frac{1}{2} \log \det \{ i(\theta) \},$$

where the penalty is the logarithm of the Jeffreys prior. Recent work by Kosmidis and Firth (2021) considers the impact of penalized likelihoods like (2) in the estimation of many well-used binomial-response generalized linear models, including logistic, probit, complementary log-log, and cauchit regression. Among other results, Kosmidis and Firth (2021) prove that maximizing the likelihood after penalizing it by arbitrary positive powers of the Jeffreys prior always results in finite estimates, and derive the shrinkage directions implied by the penalty.

4 Median bias reduction

The median bias-reducing adjusted score functions of Kenne Pagui et al. (2017) is another method that has been found to result in finite estimates in extensive simulation studies with logistic regression and BCL models (see Kosmidis et al., 2020, Section 6) and with cumulative link models (Gioia et al., 2021).
The ML estimator generally has median bias $P(\hat{\theta}_t \leq \theta_t) = 1/2 + O(N^{-1/2})$. Kenne Pagui et al. (2017) show that we can define an alternative estimator $\theta^\dagger$ with $P(\hat{\theta}_t^\dagger \leq \theta_t) = 1/2 + O(N^{-3/2})$, which is asymptotically closer to 1/2 than the median bias of $\hat{\theta}$, as the solution of

$$S(\theta) + A(\theta) - i(\theta)F(\theta) = 0_v. \tag{3}$$

In the above expression, $F_t(\theta) = [\tilde{i}(\theta)^{-1}]^\dagger_{tu} \tilde{F}_t(\theta)$, with

$$\tilde{F}_tu(\theta) = \text{trace} \left[ \tilde{i}_u(\theta) \left\{ \frac{1}{3}P_t(\theta) + \frac{1}{2}Q_t(\theta) \right\} \right] \quad (t = 1, \ldots, g),$$

and $\tilde{i}_u(\theta) = [\tilde{i}(\theta)^{-1}]_{uu} [\tilde{i}(\theta)^{-1}]_{uu}^\dagger [\tilde{i}(\theta)^{-1}]_{uu}$ $(u = 1, \ldots, v)$, where $A_u$ and $A_{tu}$ denote the $u$th column and $(t, u)$th element of a matrix $A$.

When $j(\theta) = \tilde{i}(\theta)$, expression (3) simplifies in a similar manner as expression (1) does. In fact, for one-parameter models ($v = 1$) that are exponential families in canonical parameterization, it can be shown that mdBR is formally equivalent to the maximization of $\tilde{i}(\theta) + \log \det \{\tilde{i}(\theta)\}/6$ (see Kenne Pagui et al., 2017, Section 2.1). However, mdBR has no penalized likelihood interpretation for $v > 1$.

## 5 Inference with mean and median bias reduction

According to the results in Firth (1993) and Kenne Pagui et al. (2017), both $\theta^*$ and $\theta^\dagger$ have the same asymptotic distribution as the ML estimator generally does, and hence are asymptotically efficient. Therefore, the distribution of those estimators for finite samples can be approximated by a Normal with mean $\theta$ and variance-covariance matrix $\{i(\theta)\}^{-1}$. The derivation of this result relies on the fact that both the adjustments $A(\theta)$ and $A(\theta) - i(\theta)F(\theta)$ to the score functions for mBR and mdBR in (1) and (3), respectively, are of order $O(1)$ as $N \to \infty$. Hence, the score function $S(\theta)$, which is $O_p(\sqrt{N})$, dominates the adjustments as information increases. The implication is that standard errors for the components of $\theta^*$ and $\theta^\dagger$ can be computed exactly as for the ML estimator, using the square roots of the diagonal elements of $\{i(\theta)\}^{-1}$ of $\{j(\theta)\}^{-1}$ at the estimates. Furthermore, first-order inferences, like standard Wald tests and Wald-type confidence intervals and regions are constructed in a plugin fashion, by replacing the ML estimates with the mBR or mdBR estimates in the usual procedures in standard software.

**Example 2: Separation in logistic regression (continued)** Continuing from Example 1, Table 1 provides the estimates of $\beta_1$, $\beta_2$ and $\beta_3$ from mBR and mdBR. The estimates have been computed using the default arguments of the `brglm_fit()` method of the `brglm2` R package (Kosmidis, 2021). `brglm_fit()` implements a variant of the quasi-Fisher scoring procedure

$$\theta^{(k+1)} = \theta^{(k)} + \{i(\theta^{(k)})\}^{-1}U(\theta^{(k)}), \tag{4}$$

where $U(\theta) := S(\theta) + A(\theta)$ if the intention is to compute the mean BR estimates, and $U(\theta) := S(\theta) + A(\theta) - i(\theta)F(\theta)$ if the intention is to compute the mdBR estimates; see Kosmidis et al. (2020) for details on the quasi-Fisher iterations and the form of the adjusted scores for mBR and mdBR in generalized linear models. Convergence has been rapid and `brglm_fit()` reported no issues for either mBR or mdBR. Furthermore, the estimates and estimated standard errors from mBR are typically closer in absolute value to zero than those from mdBR. Importantly, the z-statistics for $\beta_2$ and $\beta_3$ are all away from zero, and, in contrast to ML, both mBR and mdBR suggest at least some evidence against the individual hypothesis $\beta_2 = 0$ and $\beta_3 = 0$, which agrees with the fact that the value of the response has been fully determined from the values of $x_2$ and $x_3$. 

5
6 Bias reduction and parameter transformation

6.1 Maximum likelihood estimation and general parameter transformations

The ML estimator is equivariant in the sense that the ML estimator of \( g(\theta) \) is exactly \( g(\hat{\theta}) \) for any one-to-one transformation \( g(\cdot) \). Hence, there is no need to maximize the log-likelihood about \( g(\theta) \) if the ML estimator of \( \theta \) has already been computed. In contrast, the mBR and mdBR estimators are equivariant only for specific transformations \( g(\cdot) \).

6.2 Mean bias reduction and linear parameter transformations

The mBR estimator is equivariant under linear transformations for the parameters, in the sense that the mBR estimator of \( C\theta \) for a known matrix \( C \) is exactly \( C\theta^* \). The same is not true for the mdBR estimator.

For example, using Table 1, the mBR estimate of \( \beta_2 - \beta_3 \) in Example 2 is simply 5.266 – 11.166 = -5.9. The mdBR estimate, however, is not 6.325 – 13.321 = 6.996, but rather -7.227, which is obtained by reparameterizing the model in terms of \( \beta_2 - \beta_3 \) and computing the mdBR estimate by solving (3) in the new parameterization.

6.3 Median bias reduction and component-wise parameter transformations

On the other hand, the mdBR estimator of \( (g_1(\theta_1), \ldots, g_e(\theta_e))^\top \) is \( (g_1(\theta_1^*), \ldots, g_e(\theta_e^*))^\top \) for any set of one-to-one functions \( g_1(\cdot), \ldots, g_e(\cdot) \). In other words, the mdBR estimator is equivariant under component-wise transformations. The same is not true for the mBR estimator. For example, the mdBR estimate of the odds-ratio \( \exp(\beta_2) \) in Example 2 is exactly \( \exp(6.325) \), but \( \exp(5.266) \) is not an mBR estimate of \( \exp(\beta_2) \).

6.4 Mean bias reduction and general parameter transformations

Di Caterina and Kosmidis (2019) show that there is a simple way to derive the mean bias of \( h(\theta^*) \) for any three-times differentiable function \( h : C \to D \), with \( C \subset \mathbb{R}^p \) and \( D \subset \mathbb{R} \), where \( \theta^* \) is an mBR estimator of \( \theta \) with \( O(N^{-2}) \) bias. In particular, Di Caterina and Kosmidis (2019) show that the estimator \( h(\theta^*) \) of \( \zeta = h(\theta) \) has mean bias

\[
E(h(\theta^*) - h(\theta)) = \frac{1}{2} \text{trace} \left\{ i(\theta)^{-1} \nabla \nabla^\top h(\theta) \right\} + O(N^{-2}),
\]

where \( \nabla \nabla^\top h(\theta) \) is the hessian of \( h(\cdot) \) at \( \theta \). Note that for linear transformations, \( \nabla \nabla^\top h(\theta) = 0_{r \times r} \), and hence \( E(h(\theta^*) - h(\theta)) = O(N^{-2}) \), which confirms the discussion in Section 6.2 that the mBR estimator is exactly equivariant for linear transformations of the parameters. The first term in the right-hand side of (5) can be evaluated at \( \theta^\ast \) and be used to derive mean BR estimators of \( h(\theta) \), based only on \( \hat{\theta}^\ast \), \( i(\hat{\theta}^\ast) \), and \( \nabla \nabla^\top h(\theta^\ast) \). An obvious mean BR estimator resulting from (5) is \( h(\theta^\ast) - \text{trace} \left\{ i(\theta^\ast)^{-1} \nabla \nabla^\top h(\theta^\ast) \right\} / 2 \).

For example, consider the special case of estimation of the odds-ratio \( \exp(\beta_j) \) in Example 1, which was estimated using the equivariance properties of mdBR in Section 6.3. Expression (5) gives that the odds-ratio at the mBR estimator has

\[
E(\exp(\beta_j^*)) = \exp(\beta_j) \left[ 1 + \frac{1}{2} v_{jj}(\theta) \right] + O(N^{-2}),
\]

where \( v_{jj}(\theta) = [i(\theta)^{-1}]_{jj} \). Hence, two mean BR estimators of \( \zeta_j = \exp(\beta_j) \) with \( O(N^{-2}) \) bias are

\[
\zeta_j^* = \exp(\beta_j^*) \left[ 1 - \frac{1}{2} v_{jj}(\theta^*) \right] \quad \text{and} \quad \zeta_j^{**} = \frac{\exp(\beta_j^*)}{1 + v_{jj}(\theta^*)/2}.
\]
arising from subtracting an estimate of the bias at \( \theta := \theta^* \) from \( \exp(\beta_j^* \right) \), and dividing \( \exp(\beta_j^* \right) \) by the correction factor \( 1 + v_{jj}(\theta^*)/2 \) from the right-hand side of (6), respectively. The estimator \( \zeta_j^* \) for the odds-ratio \( \zeta \) has the advantage of being always positive, while \( \zeta_j^* \) takes negative values if \( v_{jj}(\theta^*) > 2 \). For example, to the accuracy reported in Table 1, \( \zeta_2^* = \exp(5.266)(1 - 1.997^2/2) = 192.48 \), which is clearly nonsensical as an odds-ratio estimate. In contrast, \( \zeta_2^* = \exp(5.266)/(1 + 1.997^2/2) = 64.66 \). The approximation \( \exp\{v_{jj}(\theta)/2\} \approx 1 + v_{jj}(\theta)/2 \) for small \( v_{jj}(\theta) \) can be used to show that the mean BR estimator \( \zeta_j^{**} \) closely relates to the mean BR estimator \( \zeta_j^{***} = \exp(\beta_j^* - v_{jj}(\theta^*)/2) \) derived in Lyles et al. (2012).

The discussion in Section 5 implies that estimated standard errors for mBR estimators of transformed parameters constructed on the basis of (6) can be computed using the delta method, as for the ML estimator.

7 Adjacent-categories logit models

7.1 Proportional and non-proportional odds models

We now turn our attention in applying mBR and mdBR from Section 3 and Section 4 to ACL models.

Adjacent-categories logit models (see, for example, Agresti, 2010, Chapter 4 for an introduction) are a prominent family of regression models for ordinal responses, where the local odds ratios of consecutive categories of an ordinal response variable are linked with linear combinations of parameters and explanatory variables. Suppose that we observe realizations of \( n \) independent random vectors of frequencies \( Y_1, \ldots, Y_n \), where \( Y_i = (Y_{i1}, \ldots, Y_{ik})^\top \) has a \( k \)-category multinomial distribution with ordered categories \( 1 < 2 < \ldots < k \). total \( m_i = \sum_{j=1}^k Y_{ij} \) and probability vector \( (\pi_1(x_i), \ldots, \pi_k(x_i))^\top \) with \( \sum_{j=1}^k \pi_j(x_i) = 1 \), where \( x_i = (x_{i1}, \ldots, x_{ip})^\top \) is a \( p \)-vector of covariate values. An ACL model has

\[
\log \frac{\pi_j(x)}{\pi_{j+1}(x)} = \eta_j(x) \quad (j = 1, \ldots, k - 1),
\]

where \( \eta_j(x) \) is typically a linear combination of unknown model parameters and a covariate vector \( x \).

The specification of \( \eta_j(x) \) results in ACL models with particular properties. The PO version of the ACL model has

\[
\eta_j(x) = \alpha_j + \beta^\top x,
\]

and \( p + k - 1 \) scalar model parameters \( \theta = (\alpha_1, \ldots, \alpha_{k-1}, \beta_1, \ldots, \beta_p)^\top \). Straightforward algebra starting from (7) gives that

\[
\frac{\pi_j(x_2)}{\pi_{j+1}(x_2)} = \exp(\beta^\top (x_2 - x_1)) \cdot \frac{\pi_j(x_1)}{\pi_{j+1}(x_1)} \quad \text{for any } x_1, x_2 \in \mathbb{R}^p \text{ and } j \in \{1, \ldots, k - 1\}.
\]

As a result, adjacent-categories odds are indeed proportional with a constant of proportionality that does not depend on the category. The NPO version of the ACL model has

\[
\eta_j(x) = \alpha_j + \beta_j^\top x,
\]

with \( (k - 1)(p + 1) \) scalar model parameters \( \theta = (\alpha_1, \ldots, \alpha_{k-1}, \beta_1^\top, \ldots, \beta_{k-1}^\top)^\top \), where \( \beta_j = (\beta_{j1}, \ldots, \beta_{jp})^\top \). Figure 2 shows the adjacent-categories log-odds for two distinct categories under the PO and the NPO versions of the ACL model, for \( x \in \mathbb{R} \). Note that under the PO version of the model the log-odds for distinct categories are parallel lines, which in turn implies (9) for any pair of categories. On the other hand, under the NPO version of the model the log-odds are not parallel lines, so (9) is not generally satisfied.
Figure 2: The adjacent-categories log-odds for categories \( j \) and \( k \), \( j \neq k \), under the proportional odds (left) and the non-proportional odds (right) versions of the model, for \( x \in \mathbb{R} \). The probability for category \( j \) at covariate value \( x \) is denoted \( \pi_j(x) \).

The most general version of the ACL model is the partial proportional odds model with

\[
\eta_j(x) = \alpha_j + \xi_j^\top x^{(np)} + \rho^\top x^{(p)},
\]

where \( x^{(np)} \) and \( x^{(p)} \) are sub-vectors of \( x \) with distinct components characterizing the PO and NPO effects, respectively. All subsequent derivations, results, and discussions can be written in terms of the more general partial proportional odds version, and then PO and NPO can be presented as special cases. Nevertheless, we focus on the PO and NPO versions separately, to keep the notation concise, and because some of the following results are specific to PO and not to NPO.

Expressions (8) and (10) immediately imply that the ACL model provides valid category probabilities across the parameter space and regardless of whether the local odds \( \log \frac{\pi_j(x)}{\pi_{j+1}(x)} \) are modelled as proportional or non-proportional. This is in contrast to other popular ordinal-response regression models, like cumulative-logit models (McCullagh, 1980), whose NPO versions (Peterson and Harrell, 1990) may provide invalid category probabilities in subsets of the parameter space and covariate space, and, hence, result in hard-to-circumvent issues with estimation, inference, and prediction.

7.2 Equivalence with baseline-category logit models

Writing \( \log \frac{\pi_j(x)}{\pi_k(x)} = \sum_{l=j}^{k-1} \log \frac{\pi_l(x)}{\pi_{l+1}(x)} \), it is simple to show that both the PO and NPO versions of the ACL model for ordinal responses can be written as BCL models for nominal responses (see Agresti, 2010, Section 4.1) where the \( k \) category is used as reference.

In particular, the NPO version of the ACL model in (10) is equivalent to a BCL model with

\[
\log \frac{\pi_j(x)}{\pi_k(x)} = \gamma_j + \delta_j^\top x \quad (j = 1, \ldots, k-1),
\]

where \( \gamma_j = \sum_{l=j}^{k-1} \alpha_l \) and \( \delta_j = \sum_{l=j}^{k-1} \beta_l \). The PO version of the ACL model in (8) is equivalent
to a BCL model with
\[ \log \frac{\pi_j(x)}{\pi_k(x)} = \gamma_j + (k - j)\xi^T x \quad (j = 1, \ldots, k - 1), \] (12)

where \( \gamma_j = \sum_{l=j}^{k-1} \alpha_l \) and \( \beta = \zeta \).

### 7.3 Maximum likelihood estimation

A consequence of the equivalence between the BCL and ACL models is that we can estimate the
latter using the ML estimates for the former. The equivariance of the maximum ML estimator
under one-to-one transformations of the model parameters guarantees that after computing the
ML estimates for the parameters of BCL model (11), the model parameters of the NPO version
of the ACL can be estimated as \( \hat{\alpha}_j = \gamma_j - \hat{\gamma}_j + 1 \) and \( \hat{\beta}_j = \hat{\delta}_j - \hat{\delta}_j + 1 \) (\( j = 1, \ldots, k - 1 \)) with \( \hat{\gamma}_k = 0 \) and \( \hat{\beta}_k = 0 \), where 0 is a \( p \)-vector of zeros. Correspondingly, once the ML estimates for the
parameters of BCL model (12) have been obtained, the model parameters of the PO version of
the ACL model can be estimated as \( \hat{\alpha}_j = \gamma_j - \hat{\gamma}_j + 1 \) and \( \hat{\beta}_j = \hat{\zeta} \) (\( j = 1, \ldots, k - 1 \)).

So, ML estimation of ACL models can be performed using ready ML implementations for
fitting the BCL models (11) and (12), like the \texttt{multinom()} function of the \texttt{nnet} R package
(Venables and Ripley, 2002) that exploits the equivalence of BCL models with neural networks,
and the \texttt{brmultinom()} function of the \texttt{brglm2} R package (Kosmidis, 2021) that exploits the
equivalence of BCL models with Poisson log-linear models.

### 7.4 Exponential families

The BCL model is a full exponential family distribution with natural parameters \( \gamma_j \) and \( \delta_j \) for
the NPO version (11) of the ACL model, and \( \gamma_j \) and \( \zeta \) for the PO version (12) of the ACL
model (\( j = 1, \ldots, k - 1 \)). Hence, another consequence of the equivalence of ACL models to
BCL models is that both the PO and NPO versions of the ACL model are full exponential
families. Specifically, the sufficient statistics in the NPO parameterization are \( \sum_{l=1}^{j} \sum_{i=1}^{n} y_{il} \) for \( \alpha_j \), \( \sum_{k=1}^{j} \sum_{i=1}^{n} y_{il} x_i \) for \( \beta_j \), and \( \sum_{l=1}^{j} \sum_{i=1}^{n} y_{il} \) for \( \alpha_j \) and \( \sum_{k=1}^{j-1} \sum_{i=1}^{n} (k - j) y_{ij} x_i \) for \( \beta \) in the
PO parameterization (\( j = 1, \ldots, k - 1 \)) (see, also, Agresti, 2010, Section 4.1).

### 7.5 Infinite maximum likelihood estimates

As is the case for their equivalent BCL models, depending on the data configuration, the ML
estimates of ACL models can have infinite components, resulting in issues for both iterative
estimation procedures and for first-order inference about the parameters. In fact, infinite ML
estimates for the PO and NPO versions of the ACL model result if, and only if, separation
occurs for the equivalent BCL models. Example 3 below uses a real data set to illustrate the
consequences that separation can have in the estimation of, and inference from, ACL models.

**Example 3: Infinite ML estimates in ACL models** The data set in Table 2 comes from
Randall (1989) and concerns an experiment for investigating factors that affect the bitterness
of white wine. There are two factors in the experiment, namely temperature at the time of
crushing the grapes (with two levels, “cold” and “warm”) and contact of the juice with the skin
(with two levels “Yes” and “No”). For each combination of factors two bottles were rated on
their bitterness by a panel of 9 judges. The responses of the judges on the bitterness of the wine
were taken on a continuous scale in the interval from 0 (“None”) to 100 (“Intense”) and then
they were grouped correspondingly into 5 ordered categories, labelled as “1”, “2”, “3”, “4”, and
“5”.

Figure 3 shows the empirical adjacent logits \( \log\left\{(y_{ij} + 1/2)/(y_{ij+1} + 1/2)\right\} \) (\( j = 1, \ldots, 4 \))
for the bitterness rating for all combinations of temperature and contact. Note that 1/2 has
Table 2: The wine tasting data (Randall, 1989).

| Temperature | Contact | Bitterness rating |
|-------------|---------|-------------------|
| Cold        | No      | 4 9 5 0 0         |
| Cold        | Yes     | 1 7 8 2 0         |
| Warm        | No      | 0 5 8 3 2         |
| Warm        | Yes     | 0 1 5 7 5         |

Figure 3: The empirical adjacent logits \( \log \left\{ \left( \frac{y_{ij} + 1}{2} \right) / \left( \frac{y_{ij+1} + 1}{2} \right) \right\} \) \( j = 1, \ldots, 4 \) for the bitterness rating for all combinations of levels for temperature and contact.

been added to all frequencies as a means of getting estimates of the adjacent-categories logits with second-order mean bias (see, for example, Haldane, 1955), avoiding infinite estimates in the process.

There seems to be evidence that the adjacent logits for the combinations of temperature and contact are parallel (see, also, Figure 2), or in other words, the adjacent odds ratios across temperature and/or contact levels do not depend on the rating. The latter hypothesis can be formally tested by estimating the NPO version of the ACL model

\[
\log \frac{\pi_j(t, c)}{\pi_{j+1}(t, c)} = \alpha_j + \beta_{1j} t + \beta_{2j} c \quad (j = 1, \ldots, 4),
\]

where \( t \) is 1 if temperature is warm and 0 otherwise, \( c \) is 1 if contact is yes and 0 otherwise, and \( \pi_j(t, c) \) is the probability of a bitterness rating \( j \) at \( t \) and \( c \). The hypotheses of parallel adjacent logits can then be written in terms of the model parameters as \( \beta_{11} = \ldots = \beta_{14} = \beta_1 \) and \( \beta_{21} = \ldots = \beta_{24} = \beta_2 \), and tested using the value of the Wald statistic

\[
W = \hat{\theta}^T C^T \left\{ C i(\hat{\theta})^{-1} C^T \right\}^{-1} C \hat{\theta},
\]

where \( \hat{\theta} \) is the ML estimate of \( \theta = (\alpha_1, \ldots, \alpha_4, \beta_{11}, \ldots, \beta_{14}, \beta_{21}, \ldots, \beta_{24})^T \) for model (13), and \( i(\theta) \) is the expected information matrix at \( \theta \). The contrast matrix \( C \) we use in (14) has the form

\[
C = \begin{bmatrix}
0_{3 \times 4} & C_1 \\
0_{3 \times 4} & 0_{3 \times 4} & C_1
\end{bmatrix}
\]

with \( C_1 = \begin{bmatrix}
1 & 0 & 0 & -1 \\
0 & 1 & 0 & -1 \\
0 & 0 & 1 & -1
\end{bmatrix} \).
Table 3: Top: ML estimates and estimated standard errors (in parenthesis) from fitting the ACL model in (13) on the data in Table 2. The estimates are obtained using the `vglm()` function of the VGAM R package (Yee, 2021) version 1.1-5 with default converge criteria (`epsilon = 10^{-7}` in `vglm.control()`). Bottom: ML estimates and estimated standard errors using stricter convergence criteria (`epsilon = 10^{-9}` in `vglm.control()`). The estimated standard errors are computed as the square roots of the diagonal of the inverse of the expected information matrix at the ML estimates. The column $\ell(\hat{\theta})$ gives the maximized log-likelihood for each fit.

| epsilon | $\ell(\hat{\theta})$ | rating (j) | $\hat{\alpha}_j$ | $\hat{\beta}_{1j}$ | $\hat{\beta}_{2j}$ |
|---------|---------------------|-----------|------------------|------------------|------------------|
| $10^{-7}$ | -15.29 | 1 | -0.83 (0.59) | -20.26 (10047.96) | -1.10 (1.21) |
|         |         | 2 | 0.67 (0.52) | -1.21 (0.66) | -0.87 (0.64) |
|         |         | 3 | 3.08 (1.05) | -1.98 (0.92) | -1.54 (0.83) |
|         |         | 4 | 20.22 (10732.18) | -19.89 (10732.18) | 0.04 (1.08) |
| $10^{-9}$ | -15.29 | 1 | -0.83 (0.59) | -25.26 (122409.18) | -1.10 (1.21) |
|         |         | 2 | 0.67 (0.52) | -1.21 (0.66) | -0.87 (0.64) |
|         |         | 3 | 3.08 (1.05) | -1.98 (0.92) | -1.54 (0.83) |
|         |         | 4 | 25.22 (130748.2) | -24.89 (130748.24) | 0.04 (1.08) |

where $0_{a \times b}$ is an $a \times b$ matrix of zeros. General results about the limiting distribution of the ML estimator under mild regularity conditions (see, for example, McCullagh, 2018, Section 7.1 and Section 7.2 and Cox and Hinkley, 1974, Section 9.1) can be used to show that the Wald statistic has asymptotically a $\chi^2_6$ distribution.

Table 3 shows the ML estimates of the ACL model in (13), as computed using the `vglm()` function of the VGAM R package (Yee, 2021). No warnings or errors were returned when fitting the model. As has been the case in the logistic regression model of Example 1, the estimates and estimated standard errors for $\alpha_4$, $\beta_{11}$ and $\beta_{14}$ are atypically large in absolute value. It is also clear that these estimates and estimated standard errors increase in absolute value as the convergence criteria get stricter, while the maximized log-likelihood value remains the same to the displayed accuracy.

These issues are not due to the implementation of the `vglm()` function; instead they are consequences of quasi-complete separation for this particular combination of data and model (13). The ML estimates $\hat{\alpha}_4$, $\hat{\beta}_{11}$ and $\hat{\beta}_{14}$ in Table 3 are formally $\infty$, $-\infty$ and $-\infty$, the corresponding estimated standard errors are all $\infty$, and the likelihood surface has an asymptote at $-15.29$ as $\alpha_4$, $\beta_{11}$ and $\beta_{14}$ diverge to $\infty$, $-\infty$ and $-\infty$, respectively, along a ray in the parameter space.

Note here that the estimated standard errors appear to diverge faster than the ML estimates do as the convergence criteria get stricter. As a result, the typically reported $Z$-statistics for individual hypothesis tests about the parameters will tend to be spuriously small in absolute value regardless of the strength of the evidence against the hypotheses. Hence, the naive use of the computer output for inference about the parameters of ACL models is likely to lead to invalid conclusions when data separation occurs. More importantly, having estimates on the boundary of the parameter space violates the assumptions required for the asymptotic $\chi^2$ distribution of (14). Consequently, it is hard to justify the performance and validity of the Wald statistic in that case.
8 Mean and median bias reduction for ACL models

A consequence of the ACL models being full exponential family distributions (see Section 7.4) is that mean BR can be implemented by maximizing the penalized likelihood in (2). Nevertheless, as for ML, mean BR estimates for ACL models can be conveniently computed through a ready implementation for mean BR in BCL models coupled with the equivariance of the mean BR estimator under linear transformations (see Section 6.2).

Kosmidis and Firth (2011) prove that the equivalence of BCL models and Poisson log-linear models (see, also, Palmgren, 1981 and Baker, 1994 for authoritative descriptions of that equivalence) extends to the mBR estimates, and describe a simple algorithm for mBR estimation of BCL models, each iteration of which consists of the following steps:

P1 Rescale the Poisson means to match the observed multinomial totals.

P2 Add half a leverage based on the rescaled means to the observed multinomial frequencies.

P3 Estimate, using ML, the equivalent Poisson log-linear model to the adjusted frequencies.

Iteration stops when the differences between successive estimates or, alternatively, the mean BR adjusted scores in (1) are smaller than a pre-determined, small positive constant. An alternative criterion can be based on the change of the mean BR penalized likelihood (2) between successive iterations. mBR estimates for ACL models can then be computed as follows

S1 Compute mBR estimates of the parameters $\gamma_j$ and $\delta_j$ of the BCL model in (11) for the NPO version (or $\gamma_j$ and $\zeta$ of the BCL model in (12) for the PO version) ($j = 1, \ldots, q$) by iterating steps P1, P2, and P3.

S2 Calculate the mBR estimates for the NPO version of the ACL model as $\alpha_j^* = \gamma_j^* - \gamma_j^* + 1$ and $\beta_j^* = \delta_j^* - \delta_j^* + 1$ (or $\alpha_j^* = \gamma_j^* - \gamma_j^* + 1$ and $\beta^* = \zeta^*$ for the PO version) ($j = 1, \ldots, q$), with $\gamma_k^* = 0$ and $\beta_k^* = 0$.

Implementation of mdBR for ACL models is not as direct as that of mBR. A maximum penalized likelihood interpretation of mdBR does not exist for general ACL models, like it does for mBR. Also, since contrasts of parameters are not component-wise transformations, algorithms for mdBR for BCL models (see Kosmidis et al., 2020, Section 6 for extensions of the results in Kosmidis and Firth, 2011) can only be used to get mdBR estimates $\beta^\dagger$ of $\beta$ in the PO version of the ACL model. In other words, the estimates $\gamma_j^\dagger - \gamma_{j+1}^\dagger$ and $\beta_j^\dagger = \delta_j^\dagger - \delta_{j+1}^\dagger$ ($j = 1, \ldots, q$) are not mdBR estimates, unless $k = 2$. Hence, for general ACL models, computing the mdBR estimates $\theta^\dagger$ must rely on implementing and solving the mdBR adjusted score equations (3). That can certainly be done (using, for example, the quasi-Fisher scoring iteration (4)), with the only effort being in deriving $P_t(\theta)$ using the expressions for mBR in BCL models in Kosmidis (2007, Appendix B.5).

Example 4: Infinite ML estimates in ACL models (continued) Table 4 gives the mBR estimates from fitting the ACL model in (13) on the data in Table 2. The mBR estimates are computed using the bracl() function of the brglm2 R package, which implements mBR through the corresponding Poisson log-linear model, as detailed earlier. No convergence issues have been reported; the absolute values of the components of the adjusted score functions in (1) at the mBR estimates are all less than $10^{-6}$, and all estimates and estimated standard errors remain unchanged to the reported accuracy as the convergence criteria get stricter.

The Wald statistic (14) when $\hat{\theta}$ is replaced by $\theta^\dagger$ has value 1.067, which is small compared to the value of the 95% quantile of a $\chi^2_6$ distribution (12.592), providing no evidence against the simpler PO model with $\beta_{11} = \ldots = \beta_{14} = \beta_1$ and $\beta_{21} = \ldots = \beta_{24} = \beta_2$. 

12
Table 4: Mean BR estimates and estimated standard errors (in parenthesis) from fitting the ACL model in (13) on the data in Table 2. The estimates are obtained using the `bracl()` function of the `brglm2` R package (Kosmidis, 2021) version 0.7.2 with default convergence criteria. The estimated standard errors are computed as the square roots of the diagonal of \( i(θ^*)^{-1} \).

| rating (j) | \( α_j^* \) | \( β_{1j}^* \) | \( β_{2j}^* \) |
|------------|----------------|----------------|----------------|
| 1          | -0.76 (0.59)   | -1.65 (1.60)   | -0.82 (1.08)   |
| 2          | 0.62 (0.52)    | -1.12 (0.66)   | -0.80 (0.64)   |
| 3          | 2.73 (0.99)    | -1.75 (0.87)   | -1.38 (0.81)   |
| 4          | 1.53 (1.83)    | -1.26 (1.68)   | 0.07 (1.03)    |

Comparing the mBR estimates in Table 4 to the ML ones in Table 3, we notice that the mBR estimates are shrunken towards zero relative to ML ones. As a result, the fitted multinomial probabilities at the mBR estimates are closer to \( (1/5, 1/5, 1/5, 1/5, 1/5)^\top \) than ones at the ML estimates. In other words, mBR shrinks the model towards equi-probability across observations. This is a generalization of the shrinkage effect of mBR we observed in Example 2 and that Kosmidis and Firth (2021) study theoretically in the special case of logistic regression \((k = 2)\).

It is interesting to note that the shrinkage direction of mBR in cumulative logit models for global cumulative odds (Kosmidis, 2014) is rather different; the fitted multinomial probabilities at the mBR estimates for the PO version of the cumulative logit model would be closer to \( (1/2, 0, 0, 0, 1/2)^\top \) than the ones at the ML estimates. In other words, mBR shrinks the cumulative logit model towards a logistic regression model for the end categories.

The ML, mdBR, and mBR estimates for \( β_1 \) for the PO version of the ACL model are \(-1.69, -1.61, \) and \(-1.56, \) respectively, with corresponding estimated standard errors \(0.41, 0.39, \) and \(0.38. \) The respective estimates for \( β_2 \) are \(-0.96, -0.92, \) and \(-0.90, \) respectively, with corresponding estimated standard errors \(0.32, 0.31, \) and \(0.31. \) The shrinkage towards equi-probability that mBR delivers is also apparent in the estimates for the PO version of the ACL model. As is the case in logistic regression, mdBR also tends to shrink estimates towards zero, but that shrinkage effect is less strong than from mBR.

9 Mean bias reduction of ordinal superiority summaries

The mean BR estimates for ACL models can be used to get improved estimates of other model summaries by using the bias of transformations of the mean RB estimator in expression (5).

A prominent example of such a summary are the ordinal, model-based superiority measures for comparing distributions of two groups, adjusted for covariates that are introduced in Agresti and Kateri (2017). In ordinal-response models with a latent variable interpretation, such as cumulative-link models (McCullagh, 1980), ordinal superiority measures can be defined directly on the latent scale, which results in exact (for probit, log-log, and complementary log-log link) or approximate expressions (for logit link) that are functions of only the coefficient of the indicator variable characterizing the two groups being compared. This fact has been exploited in Gioia et al. (2021), who used the equivariance properties of the mdBR estimator (see Section 6.3) to directly transform the mdBR estimates of the group indicator parameter to deliver mdBR estimates of ordinal superiority measures.

In more general models for ordinal responses that may also lack a latent variable interpretation (like ACL models), ordinal superiority measures are instead defined in terms of category probabilities that necessarily depend on all model parameters. Suppose that the covariate vector
is \((w^\top, z)^\top\), where \(z\) is a group indicator variable taking value 0 for group 1 and value 0 for group 2, and denote by \(\pi_j(w, 1)\) and \(\pi_j(w, 0)\) \((j = 1, \ldots, k)\) the model-based probabilities of category \(j\) at covariate values \(w\), for group 1 and group 2, respectively. The dependence of the probabilities on the model parameters has been suppressed here for notational convenience.

Agresti and Kateri (2017) propose comparing the distribution of the ordinal response at group 1 to that at group 2, at covariate values \(w\), through the ordinal superiority measure

\[
\Delta(w; \theta) = \sum_{r > s} \pi_r(w, 1)\pi_s(w, 0) - \sum_{s > r} \pi_r(w, 1)\pi_s(w, 0) .
\]

(15)

If the two distributions are identical then \(\Delta(w; \theta) = 0\). Positive values of \(\Delta(w; \theta)\) indicate that for covariates \(w\), it is more likely to observe higher response categories in group 1 than in group 2, and vice versa for negative values. A related ordinal superiority measure is

\[
\gamma(w; \theta) = 2\Delta(w; \theta) - 1 ,
\]

(16)

which takes values between 0 and 1, and is interpreted as the probability that the response category in group 1 is higher than the response category in group 2, while adjusting for covariates \(w\) (see Klotz, 1966, for details). In practice, the covariate setting \(w\) can be taken to be a representative value from a sample of covariate values \(w_1, \ldots, w_n\), e.g. \(\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i/n\). Alternatively, if the sample of covariate values is representative of the population of interest then summary ordinal superiority measures can be defined as

\[
\hat{\Delta}(\theta) = \frac{1}{n} \sum \Delta(w_i; \theta) \quad \text{and} \quad \hat{\gamma}(\theta) = \frac{1}{n} \sum \gamma(w_i; \theta) .
\]

(17)

Agresti and Kateri (2017) propose estimating the ordinal superiority measures by replacing \(\theta\) in expressions (15), (16), and (17) by the ML estimator \(\hat{\theta}\), and use the delta method to construct inferences about those measures. Note here that because of the specific equivariance properties of the mBR and mdBR estimator (see Section 6), replacing \(\theta\) by the mBR estimator \(\theta^\star\) or a mdBR estimator \(\theta^\ddagger\) does not, in general, result in mBR or mdBR estimators of the measures. In fact, despite it being the case that the resulting estimators will be consistent under the same conditions that their ML counterparts are, they may end up having much worse finite-sample mean and/or median bias properties than the ML version does.

mdBR estimators of (15), (16), and (17) are not easy to construct. In contrast, an easy-to-compute mBR estimator of \(\Delta(w; \theta)\) and of the other ordinary superiority measures can be derived using expression (5). In particular, an mBR estimator of \(\Delta(w; \theta)\) is

\[
\Delta^*(w; \theta^\star) = \Delta(w; \theta^\star) - B^*(w; \theta^\star) ,
\]

where

\[
B^*(w; \theta) = \frac{1}{2} \text{trace} \left\{ i(\theta)^{-1} \nabla \nabla^\top \Delta(w; \theta) \right\} ,
\]

is the first term in the right-hand side of expression (5). Computing \(\Delta^*(w; \theta^\star)\) requires only the mBR estimator \(\theta^\star\) that can be obtained using the procedures in Section 8, the corresponding estimated category probabilities at \((w^\top, 1)\) and \((w^\top, 0)\), the matrix \(i(\theta^\star)^{-1}\), and the hessian \(\nabla \nabla^\top \Delta(w; \theta^\star)\). All these quantities, except \(\nabla \nabla^\top \Delta(w; \theta^\star)\), are readily available or can be readily computed once the model has been estimated using mBR, as is done, for example, in Section 3 for ACL models and in Kosmidis (2014) for cumulative link models. For specific ordinal-response models, the hessian \(\nabla \nabla^\top \Delta(w; \theta^\star)\) can be analytically obtained with some algebraic effort. For example, if \(\pi_j(w, z)\) is based on cumulative link models one can work with the expressions for the derivatives of \(\Delta(w; \theta)\) in Agresti and Kateri (2017, Web appendix A). Alternatively, a very accurate approximation of \(\nabla \nabla^\top \Delta(w; \theta^\star)\) can be obtained for general models using a ready implementation of \(\Delta(w; \theta)\) and numerical differentiation routines, like the ones provided in the numDeriv
R package (Gilbert and Varadhan, 2019). This is the route that the ordinal_superiority() method of the brglm2 R package takes.

Due to the equivariance properties of mBR estimation in Section 6.2 under linear transformations, mBR estimators of \(\gamma(w; \theta), \Delta(w; \theta)\), and \(\gamma^*(\theta)\) are readily obtained by replacing \(\Delta(w; \theta)\) by \(\Delta^*(w; \theta^*)\) in expressions (16) and (17). Wald-type inferences about the mBR estimators of the ordinal superiority measures can be constructed as proposed in Agresti and Kateri (2017, Section 5), using the mBR estimates of the ordinal superiority measures along with estimated standard errors obtained using the delta method, based on \(i(\theta^*)\), and numerical gradients.

**Example 5: mBR for ordinal superiority measures from ACL model** In order to assess the finite sample properties of the mBR estimator of ordinal superiority scores in ACL models, we consider the example in Christensen (2019, Section 4.3), where the bitterness ratings “2”, “3”, and “4” in Table 2 are merged into a single rating “2-4”. Like the PO version of the cumulative logit model (see Christensen, 2019, Section 4.8), the ML estimates for \(\alpha_{2-4}\) and \(\beta_1\) for the PO version of the ACL model (13) are \(+\infty\) and \(−\infty\) respectively. The mBR estimates of \(\theta = (\alpha_1, \alpha_{2-4}, \beta_1, \beta_2)^\top\), on the other hand, take the finite values \(\theta^* = (-1.247, 5.331, -3.291, -1.181)^\top\). If \(\gamma(w, \theta)\) is the ordinal superiority measure for temperature setting \(w\) (\(w = 0\) for cold and \(w = 1\) for warm), and \(z\) indicates contact \((z = 1)\) or not \((z = 0)\) of the juice with the skin, then \(\gamma(0, \theta^*) = 0.594\) and \(\gamma(1, \theta^*) = 0.575\), indicating that there is almost 60% chance of higher bitterness ratings when there is contact of the juice with the skin.

We simulate 10,000 samples from the PO version of the ACL model at \(\bar{\theta}\), and we compute \(\gamma(w, \bar{\theta})\) and \(\gamma^*(w, \theta^*)\) for each sample. The simulation-based estimates of the finite-sample relative biases of \(\gamma(w, \bar{\theta})\) are 0.84% and 1.56% for \(w = 0\) and \(w = 1\), respectively. As expected, the mBR version \(\gamma^*(w, \theta^*)\) is found to have smaller finite-sample relative biases at 0.13% and −0.02% for \(w = 0\) and \(w = 1\), respectively. The corresponding percentages of underestimation are 48.48% and 44.69% for \(\gamma(w, \bar{\theta})\), and 52.12% and 51.14% for \(\gamma^*(w, \theta^*)\). Hence, in this case, mBR also results in improvements in median bias. Finally, both estimators appear to perform satisfactorily in terms of Wald-type inferences based on them. The coverage probability of the nominally 95% Wald-type confidence intervals based on \(\gamma(w, \bar{\theta})\) are 94.8% (\(w = 0\)) and 94.6% (\(w = 1\)), and 94.7% (\(w = 0\)) and 95.1% (\(w = 1\)) for those based on \(\gamma^*(w, \theta^*)\).

### 10 Supplementary material

The supplementary material consists of three scripts that replicate all the numerical results and graphics reported in the paper, and is available at https://ikosmidis.com/files/bracl_supplementary_v0.2.zip. The results are exactly reproducible in R version 4.1.2, and with the following packages: VGAM version 1.1-5 (Yee, 2021), tibble 3.1.6 (Müller and Wickham, 2021), dplyr 1.0.7 (Wickham et al., 2021), ggplot2 3.3.5 (Wickham, 2016), colorspace 2.0-2 (Zeileis et al., 2020), and ordinal 2019.12-10 (Christensen, 2019), brglm2 0.8.2 (Kosmidis, 2021), enrichwith 0.3.1 (Kosmidis, 2020), and detectseparation 0.2 (Kosmidis and Schumacher, 2021).

### 11 Acknowledgements

The author greatly appreciates the constructive discussions with Alan Agresti and Anestis Touloumis during the Challenges for Categorical Data Analysis 2018 Workshop in Aachen University on the equivalence between ACL and BCL models, which informed this work. Ioannis Kosmidis has been partially supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.
References
Agresti, A. (2002). *Categorical Data Analysis* (2nd ed ed.). New York: Wiley-Interscience.
Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). Hoboken, NJ: Wiley.
Agresti, A. and M. Kateri (2017). Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics* 73(1), 214–219.
Albert, A. and J. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71(1), 1–10.
Baker, S. G. (1994). The multinomial-Poisson transformation. *Journal of the Royal Statistical Society: Series D (The Statistician)* 43(4), 495.
Bull, S. B., C. Mak, and C. M. T. Greenwood (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis* 39, 57–74.
Christensen, R. H. B. (2019). ordinal—regression models for ordinal data. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.
Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. London: Chapman & Hall Ltd.
Di Caterina, C. and I. Kosmidis (2019). Location-adjusted Wald statistics for scalar parameters. *Computational Statistics & Data Analysis* 138, 126–142.
Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80(1), 27–38.
Gilbert, P. and R. Varadhan (2019). numDeriv: Accurate Numerical Derivatives. R package version 2016.8-1.1.
Gioia, V., E. C. Kenne Pagui, and A. Salvan (2021). Median bias reduction in cumulative link models. *Communications in Statistics*, 1–17.
Haldane, J. (1955). The estimation of the logarithm of a ratio of frequencies. *Annals of Human Genetics* 20, 309–311.
Heinze, G. and M. Schemper (2002, August). A solution to the problem of separation in logistic regression. *Statistics in Medicine* 21(16), 2409–2419.
Kenne Pagui, E. C., A. Salvan, and N. Sartori (2017). Median bias reduction of maximum likelihood estimates. *Biometrika* 104(4), 923–938.
Klotz, J. H. (1966). The Wilcoxon, Ties, and the Computer. *Journal of the American Statistical Association* 61(315), 772–787.
Konis, K. (2007). *Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models*. Ph. D. thesis, University of Oxford.
Kosmidis, I. (2007). Bias reduction in exponential family nonlinear models. Ph. D. thesis, University of Warwick.
Kosmidis, I. (2014). Improved estimation in cumulative link models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1), 169–196.
Kosmidis, I. (2020). *enrichwith: Methods to enrich list-like R objects with extra components*. R package version 0.3.1.

Kosmidis, I. (2021). *brglm2: Bias Reduction in Generalized Linear Models*. R package version 0.7.2.

Kosmidis, I. and D. Firth (2009). Bias reduction in exponential family nonlinear models. *Biometrika* 96(4), 793–804.

Kosmidis, I. and D. Firth (2011). Multinomial logit bias reduction via the poisson log-linear model. *Biometrika* 98(3), 755–759.

Kosmidis, I. and D. Firth (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika* 108(1), 71–82.

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

Kosmidis, I. and D. Schumacher (2021). *detectseparation: Detect and Check for Separation and Infinite Maximum Likelihood Estimates*. R package version 0.2.

Lesaffre, E. and A. Albert (1989). Partial separation in logistic discrimination. *Journal of the Royal Statistical Society. Series B (Methodological)* 51(1), 109–116.

Lyles, R. H., Y. Guo, and S. Greenland (2012). Reducing bias and mean squared error associated with regression-based odds ratio estimators. *Journal of Statistical Planning and Inference* 142(12), 3235–3241.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)* 42, 109–142.

McCullagh, P. (2018). *Tensor methods in statistics* (2nd ed.). Mineola, NY: Dover Publications.

Mehrabi, Y. and J. N. S. Matthews (1995). Likelihood-based methods for bias reduction in limiting dilution assays. *Biometrics* 51, 1543–1549.

Müller, K. and H. Wickham (2021). *tibble: Simple Data Frames*. R package version 3.1.6.

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

Palmgren, J. (1981). The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables. *Biometrika* 68(2), 563.

Peterson, B. and J. Harrell, Frank E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics* 39, 205–217.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Randall, J. H. (1989). The analysis of sensory data by generalised linear model. *Biometrical Journal* 7, 781–793.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)* 43(3), 310–313.
Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. ISBN 0-387-95457-0.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., R. François, L. Henry, and K. Müller (2021). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.7.

Yee, T. W. (2021). *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.1-5.

Zeileis, A., J. C. Fisher, K. Hornik, R. Ihaka, C. D. McWhite, P. Murrell, R. Stauffer, and C. O. Wilke (2020). *colorspace: A toolbox for manipulating and assessing colors and palettes*. *Journal of Statistical Software* 96(1), 1–49.