Advanced Review

Bias in parametric estimation: reduction and useful side-effects

Ioannis Kosmidis*

The bias of an estimator is defined as the difference of its expected value from the parameter to be estimated, where the expectation is with respect to the model. Loosely speaking, small bias reflects the desire that if an experiment is repeated indefinitely then the average of all the resultant estimates will be close to the parameter value that is estimated. The current article is a review of the still-expanding repository of methods that have been developed to reduce bias in the estimation of parametric models. The review provides a unifying framework where all those methods are seen as attempts to approximate the solution of a simple estimating equation. Of particular focus is the maximum likelihood estimator, which despite being asymptotically unbiased under the usual regularity conditions, has finite-sample bias that can result in significant loss of performance of standard inferential procedures. An informal comparison of the methods is made revealing some useful practical side-effects in the estimation of popular models in practice including: (1) shrinkage of the estimators in binomial and multinomial regression models that guarantees finiteness even in cases of data separation where the maximum likelihood estimator is infinite and (2) inferential benefits for models that require the estimation of dispersion or precision parameters.

How to cite this article:
WIREs Comput Stat 2014, 6:185–196. doi: 10.1002/wics.1296

Keywords: jackknife/bootstrap; indirect inference; penalized likelihood; infinite estimates; separation in models with categorical responses

IMPACT OF BIAS IN ESTIMATION

By its definition, bias necessarily depends on how the model is written in terms of its parameters and this dependence makes it not a strong statistical principle in terms of evaluating the performance of estimators; e.g., unbiasedness of the familiar sample variance S^2 as an estimator of σ^2 does not deliver an unbiased estimator of σ itself. Despite this fact, an extensive amount of literature has focused on unbiased estimators (estimators with zero bias) as the basis of refined statistical procedures (e.g., finding minimum variance unbiased estimators). In such work unbiasedness plays the dual role of a condition that (1) allows the restriction of the class of possible estimators in order to obtain something useful (like minimum variance amongst unbiased estimators), and (2) ensures that estimation is performed in an impartial

*Correspondence to: i.kosmidis@ucl.ac.uk

Department of Statistical Science, University College London, London, United Kingdom

Conflict of interest: The author has declared no conflicts of interest for this article.

Supplementary material for this article is available upon request and includes a script that reproduces the analyses of the gasoline yield data using Beta regression models (wireGasoline.R). The script for the wine tasting data case study with cumulative link models is available at the supplementary material of Kosmidis[30] and can be obtained following the associated DOI link in the References section. A change in behavior at the latest version of the ordinal R package at the time of writing (version 2013.9-30) returns nonavailable values for the earlier estimated standard errors, and *Z*-values for the parameters of cumulative link models when at least one infinite estimate is detected. For this reason, the wine data analyses can be reproduced using version 2012.01-19 of the ordinal package which is available for download at http://cran.r-project.org/src/contrib/Archive/ordinal/

way, ruling out estimators that would favour one or more parameter values at the cost of neglecting other possible values. Lehmann and Casella[1] is a thorough review of statistical methods that are optimal once attention is restricted to unbiased estimators.

Another stream of literature has focused in reducing the bias of estimators, as a means to alleviating the sometimes considerable problems that bias can cause in inference. This literature, despite dating back to the early years of statistical science, is resurfacing as increasingly relevant as the complexity of models used in practice increases and pushes traditional estimation methods to their theoretical limits.

The current review focuses on the latter literature, explaining the link between the available methods for bias reduction and their relative merits and disadvantages through the analysis of real data sets.

The following case study demonstrates the direct consequences that the bias in the estimation of a single nuisance (or incidental) parameter can have in inference, even if all parameters of interest are estimated with negligible bias.

Gasoline Yield Data

To demonstrate how bias can in some cases severely affect estimation and inference we follow the gasoline yield data example in Kosmidis and Firth[2] and Grün et al.[3] The gasoline yield data[4] consists of $n=32$ observations on the proportion of crude oil converted to gasoline after distillation and fractionation on 10 distinct experimental settings for the triplet (1) temperature in degrees Fahrenheit at which 10% of crude oil has vaporized, (2) crude oil gravity, and (3) vapor pressure of crude oil. The temperature at which all gasoline has vaporized is also recorded in degrees Fahrenheit for each one of the 32 observations.

The task is to fit a statistical model that links the proportion of crude oil converted to gasoline with the experimental settings and the temperature at which all gasoline has vaporized. For this we assume that the observed proportions of crude oil converted to gasoline $y_1, y_2, \ldots, y_n$ are realizations of independent Beta distributed random variables $Y_1, \ldots, Y_n$, where $\mu_i = E(Y_i)$ and $var(Y_i) = \mu_i(1-\mu_i)/(1+\phi)$. Hence, in this parameterization, ϕ is a precision parameter $(i=1, \ldots, n)$. Then, the mean μ_i of the ith response can be linked to a linear combination of covariates and regression parameters via the logistic link function as

$$\log\frac{\mu_i}{1-\mu_i} = \alpha + \sum_{t=1}^{9}\gamma_t s_{it} + \delta t_i \quad (i=1, \ldots, n). \quad (1)$$

In the above expression, $s_{i1}, \ldots, s_{i9}$ are the values of nine dummy covariates which represent the 10 distinct experimental settings in the data set and t_i is the temperature in degrees Fahrenheit at which all gasoline has vaporized for the ith observation $(i=1, \ldots, n)$.

The parameters $\theta = (\alpha, \gamma_1, \ldots, \gamma_9, \delta, \phi)$ are estimated using maximum likelihood and the estimated standard errors for the estimates are calculated using the square roots of the diagonal elements of the inverse of the Fisher information matrix for model 1. The parameter ϕ is considered here to be a nuisance (or incidental) parameter which is only estimated to complete the specification of the Beta regression model.

Table 1 shows the parameter estimates with the corresponding estimated standard errors and the 95% Wald-type confidence intervals. One immediate observation from the table of coefficients is the very large estimate for the precision parameter ϕ. If this is merely the effect of upward bias then this bias will result in underestimation of the standard errors because for such a model the entries of the Fisher information matrix corresponding to the regression parameters $\alpha, \gamma_1, \ldots, \gamma_9, \delta$ are quantities of the form 'ϕ times a function of θ' (see Refs 2, 3 for expressions on the Fisher information). Hence, if the estimation of ϕ is prone to upward bias, then this can lead to confidence intervals that are shorter than expected at any specified nominal level and/or anti-conservative hypothesis testing procedures, which in turn result in spuriously strong conclusions.

To check whether this is indeed the case a small simulation study has been designed where 50000 samples are simulated from the maximum likelihood fit shown in Table 1. Maximum likelihood is used to fit model 1 on each simulated sample and the bias of the maximum likelihood estimator is estimated using the resultant parameter estimates. The estimated bias for α is 0.010 while the estimated biases for $\gamma_1, \ldots, \gamma_9, \delta$ are all less than 0.005 in absolute value, providing indications that bias on the regression parameters is of no consequence. Nevertheless, the estimated bias for ϕ is 299.779 which indicates a strong upward bias in the estimation of ϕ. To check how the upward bias in the precision parameter can affect the usual Wald-type inferences, we estimate the coverage probability (the probability that the confidence intervals contains the true parameter value) of the individual Wald-Type confidence intervals at levels 90, 95, and 99%. Table 2 shows the results. It is clear that the Wald-type confidence intervals systematically undercover the true parameter value across parameters.

Such behavior is observed even when the precision parameter is linked to covariates through a link

TABLE 1 | Maximum Likelihood Estimates for the Parameters of Model 1 with the Corresponding Estimated Standard Errors and the Wald-Type 95% Confidence Intervals ('estimate' ± 1.96 'estimated standard error').

Parameter	Estimate	Estimated Standard Error	95% Confidence Interval	
α	−6.160	0.182	−6.517	−5.802
γ_1	1.728	0.101	1.529	1.926
γ_2	1.323	0.118	1.092	1.554
γ_3	1.572	0.116	1.345	1.800
γ_4	1.060	0.102	0.859	1.260
γ_5	1.134	0.104	0.931	1.337
γ_6	1.040	0.106	0.832	1.248
γ_7	0.544	0.109	0.330	0.758
γ_8	0.496	0.109	0.282	0.709
γ_9	0.386	0.119	0.153	0.618
δ	0.011	0.000	0.010	0.012
ϕ	440.278	110.026	224.632	655.925

TABLE 2 | Estimated Coverage of Wald-Type Confidence Intervals at Nominal Level 90, 95, and 99%. Estimated Standard Errors Are Calculated Using the Fisher Information at the Maximum Likelihood Estimates.

	Nominal Level		
Parameter	90%	95%	99%
α	80.2	87.2	94.9
δ_1	80.3	87.3	95.2
δ_2	80.2	87.1	95.1
δ_3	80.2	87.1	94.8
δ_4	80.2	87.5	95.2
δ_5	80.5	87.5	95.2
δ_6	80.4	87.4	95.1
δ_7	80.6	87.4	95.1
δ_8	80.2	87.3	95.1
δ_9	80.5	87.3	95.0
ϕ	79.9	87.1	94.9

function, like logarithm (see e.g. Ref 3). More generally, similar consequences of bias in inference are present in all exponential family models that involve the estimation of dispersion (or precision) parameters.

CONSISTENCY, BIAS, AND VARIANCE

Suppose that interest is in the estimation of a p-vector of parameters θ, from data $y^{(n)}$ assumed to be realizations of a random quantity $y^{(n)}$ distributed according to a parametric distribution M_θ, $\theta=(\theta_1, \ldots, \theta_p)^T \in \Theta \subset \Re^p$. The superscript n here is used as an indication of the information in the data and is usually the sample size in the sense that the realization of $Y^{(n)}$ is $y^{(n)}=(y_1, \ldots, y_n)^T$. An estimator of θ is a function $\hat\theta \equiv t\left(Y^{(n)}\right)$ and in the presence of data the estimate would be $t(y^{(n)})$.

An estimator $\hat\theta = t\left(Y^{(n)}\right)$ is consistent if it converges in probability to the unknown parameter θ as $n \to \infty$. Consistency is usually an essential requirement for a good estimator because given that the family of distributions M_θ is large enough, it ensures that as n increases the distribution of $\hat\theta$ becomes concentrated around the parameter θ, essentially providing a practical reassurance that for very large n the estimator recovers θ.

The bias of an estimator is defined as

$$B(\theta) = E_\theta\left(\hat\theta - \theta\right).$$

Loosely speaking, small bias reflects the desire that if an experiment that results in data $y^{(n)}$ is repeated indefinitely, then the long-run average of all the resultant estimates will not be far from θ. Small bias is a much weaker and hence less useful requirement than consistency. Indeed, one may get an inconsistent estimator with zero bias or a consistent estimator that is biased. For example, if $Y^{(n)}=(Y_1, \ldots, Y_n)^T$, with $Y_1, \ldots, Y_n$ mutually independent random variables with $Y_i \sim N(\mu,\sigma^2)$ then $t(Y^{(n)})=Y_1$ is an unbiased but inconsistent estimator of μ. On the other hand, $t\left(Y^{(n)}\right) = \sum_{i=1}^n Y_i + 1/n$ is a consistent estimator for μ but has bias $B(t(Y^{(n)}))=1/n$. So, bias becomes relevant only if it is accompanied by guarantees of consistency, or more generally when the variability of V around θ is small (see Ref 5, § 8.1 for a discussion along this lines).

The bias function also appears directly in the expression for the lowest attainable variance of an estimator. The Cramér-Rao inequality states that the variance of any estimator $\hat{\theta}^{(n)}$ satisfies

$$var\left(\hat{\theta}^{(n)}\right) \geq \left\{1_p + \nabla_\theta B(\theta)\right\}^T \{F(\theta)\}^{-1} \left\{1_p + \nabla_\theta B(\theta)\right\}, \tag{2}$$

where 1_p is the $p \times p$ identity matrix and the inequality $A \geq C$ means that $A - C$ is a positive semidefinite matrix. The matrix $F(\theta)$ is the Fisher (or expected) information matrix which is defined as $F(\theta) = E_\theta\{S(\theta)S(\theta)^T\}$, where $S(\theta) = \nabla_\theta l(\theta)$ and $l(\theta)$ is the log-likelihood function for θ. The Cramér-Rao inequality shows what is the 'lowest' attainable variance for an estimator in terms of the derivatives of its bias and the Fisher information.

Maximum Likelihood Estimation

Denote $f(y;\theta)$ the joint density or probability mass function implied by the family of distributions M_θ. The maximum likelihood estimator $\hat{\theta}$ is the value of θ which maximizes the log-likelihood function $l(\theta;y^{(n)}) = \log f(y^{(n)};\theta)$. Given that the log-likelihood function is sufficiently smooth on θ, $\hat{\theta}$ can be obtained as the solution of the score equations

$$S(\theta) = \nabla_\theta l(\theta) = 0,$$

provided that the observed information matrix $I(\theta) = -\nabla_\theta \nabla_\theta^T l(\theta)$ is positive definite when evaluated at $\hat{\theta}$. An appealing property of the maximum likelihood estimator is its invariance under one-to-one reparameterizations of the model. If $\theta' = g(\theta)$ for some one-to-one function $g : \Re^p \to \Re^p$, then the maximum likelihood estimator of θ' is simply $g\left(\hat{\theta}\right)$. This result states that when obtaining the maximum likelihood estimator of θ, we automatically obtain the maximum likelihood estimator of $g(\theta)$ for any function g that is one-to-one, simply by calculating $g\left(\hat{\theta}\right)$ without the need of maximizing the likelihood on $g(\theta)$.

It can also be shown that the maximum likelihood estimator θ has certain optimality properties if the 'usual regularity conditions' are satisfied. Informally, the usual regularity conditions imply, among others, that (1) M_θ is identifiable (i.e., $M_\theta \neq M_{\theta'}$, for any pair (θ,θ') such that $\theta \neq \theta'$, apart from sets of probability zero), (2) p is finite, (3) that the parameter space Θ does not depend on the sample space (which implies that p does not depend on n and iv) that there exists a sufficient number of log-likelihood derivatives and expectations of those under M_θ. A more technical account of those conditions can be found in McCullagh[6], § 7.1,7.2, or equivalently in Cox and Hinkley[5], §9.1.

If these conditions are satisfied, then $\hat{\theta}$ is consistent and has bias of asymptotic order $O(n^{-1})$, which means that its bias vanishes as $n \to \infty$. Moreover, the maximum likelihood estimator has the property that as $n \to \infty$ its distribution converges to a multivariate Normal distribution with expectation θ and variance-covariance matrix $\{F(\theta)\}^{-1}$. Hence, the variance of the asymptotic distribution of the maximum likelihood estimator is exactly the Cramér-Rao lower bound $\{F(\theta)\}^{-1}$ given in Eq. (2).

Reducing Bias

All the above shows that under the usual regularity conditions as $n \to \infty$, the maximum likelihood estimator $\hat{\theta}$ has optimal properties, a fact that makes it a default choice in applications. However, for finite n these properties may deteriorate, in some cases causing severe problems in inference. Such an effect has been seen in the gasoline yield data case study where the bias of $\hat{\theta}$ affects the performance of tests and the construction of confidence intervals based on the asymptotic normality of $\hat{\theta}$.

Before reviewing the basic methods for reducing bias, it is necessary to emphasize again that bias necessarily depends on the parameterization of the model; if the bias of any estimator $\hat{\theta}$ is reduced resulting to a less biased estimator $\tilde{\theta}$, then it is not necessary that the same will happen for the estimator $g\left(\tilde{\theta}\right)$. In fact, the bias of the $g\left(\tilde{\theta}\right)$ as an estimator of $g(\theta)$ may be considerably inflated. Hence, correction of the bias of the maximum likelihood estimator comes at the cost of destroying its invariance properties under reparameterization. Therefore, all the methods for bias reduction that are described in the current review should be seen with scepticism if invariance is a necessary requirement for the analysis. On the other hand if the parameterization is fixed by the problem or practitioner, one can do much better in terms of estimation and inference by reducing the bias. Furthermore, as it will be seen later, for some models reduction of bias produces useful side-effects which in many cases have motivated its routine use in applications. A thorough discussion on considerations on bias and variance and examples of exactly unbiased estimators that are useless or irrelevant can be found at Cox and Hinkley[5] §8.2 and Lehmann and Casella[1] § 1.1.

BIAS REDUCTION—A SIMPLE RECIPE WITH MANY DIFFERENT IMPLEMENTATIONS

For a general not necessarily the maximum likelihood estimator $\hat{\theta}$ taking values in $\Theta \subset \Re^p$, consider the solution of the equation

$$\hat{\theta} - \tilde{\theta} = B(\theta), \quad (3)$$

with respect to a new estimator $\tilde{\theta}$. Equation (3) is a moment-matching equation which links the properties of the estimation method to the properties of M_θ through $\hat{\theta}$ and $B(\theta)$, respectively. If both the function $B(\theta)$ and θ were known then it is straightforward to show that $\tilde{\theta} = \hat{\theta} - B(\theta)$ has zero bias and hence, smaller mean squared error than $\tilde{\theta}$. If, in addition, the initial estimator $\tilde{\theta}$ has vanishing variance-covariance matrix as $n \to \infty$ then an application of Chebyschev's inequality shows that $\tilde{\theta}$ is consistent, even if $\tilde{\theta}$ is not. Of course, if θ is known then there is no reason for estimation, and furthermore, usually the function $B(\theta)$ cannot be written in closed-form. The importance of Eq. (3) is that, despite of its limited practical value, all known methods to reduce bias can be usefully thought of as attempts to approximate its solution. These methods can be distinguished into *explicit* and *implicit*.

EXPLICIT METHODS

Explicit methods rely on an *one-step* procedure where $B(\theta)$ is estimated and then subtracted from $\hat{\theta}$ resulting in the new estimator $\tilde{\theta}$. The most popular explicit methods for reducing bias are the jackknife, the bootstrap, and methods which use approximations of the bias function through asymptotic expansions of $B(\theta)$.

Jackknife

For many common estimators including the maximum likelihood estimator, the bias function can be expanded in decreasing powers of n as

$$B(\theta) = \frac{b(\theta)}{n} + \frac{b_2(\theta)}{n^2} + \frac{b_3(\theta)}{n^3} + O\left(n^{-4}\right), \quad (4)$$

for an appropriate sequence of functions $b(\theta), b_2(\theta)$, $b_3(\theta), \ldots$, and so on, that are $O(1)$ as $n \to \infty$.. From Eq. (4), the estimator $\hat{\theta}_{(-j)}$ which results from leaving the jth random variable out of the original set of n variables has the same bias expansion as in Eq. (4) but with n replaced with $n-1$. In light of this observation, Quenouille[7] noticed that the estimator

$$\tilde{\theta} = n\hat{\theta} - (n-1)\,\overline{\theta},$$

where $\overline{\theta}$ is the average of the n possible leave-one-out estimators $\hat{\theta}_{(-1)}, \ldots, \hat{\theta}_{(-n)}$, has bias expansion $-b_2(\theta)/n^2 + O(n^{-3})$ which is of smaller asymptotic order than the $O(n^{-1})$ bias of $\hat{\theta}$. This procedure is called jackknife (see Ref 8 for an overview of jackknife). Efron[9] §2.3 shows the basic geometric argument behind the jackknife; the jackknife is estimating the bias based on a linear extrapolation of the expected value of the estimator as a function of $1/n$. The same procedure can be carried out for correcting bias in higher orders essentially replacing the linear extrapolation by quadratic extrapolation and so on. Schucany et al.[10] give an elegant way of deriving such higher order corrections in bias with the jackknife being a prominent special case of their method. The jackknife is an explicit method because the new estimator $\hat{\theta}$ is simply

$$\tilde{\theta} = \hat{\theta} - B^{(\text{jack})},$$

where $B^{(\text{jack})} = (n-1)\left(\overline{\theta} - \hat{\theta}\right)$ is the jackknife estimator of the bias.

Bootstrap

Another class of popular explicit methods for the correction of the bias comes from the bootstrap framework. Bootstrap is a collection of methods that can be used to improve the accuracy of inference and operates under the principle that the '*bootstrap sample* is for the sample, what the sample is for the population'. Then the same procedures that are applied on the sample can equally well be applied on the bootstrap sample giving direct access to estimated sampling distributions of statistics (see Ref 11 for an overview of bootstrap). The two dominant ways to obtain a bootstrap sample are (1) by sampling from the empirical distribution function (hence sampling with replacement from the original sample) giving rise to nonparametric bootstrap methods, and (2) by sampling from the fitted parametric model giving rise to parametric bootstrap methods. In all cases, the bias of an estimator can be estimated by $B^{(boot)} = \overline{\theta}^* - \hat{\theta}$, where $\overline{\theta}^*$ is the average of the estimates based on each of the bootstrap samples. Efron and Tibshirani[12] and Davison and Hinkley[13] are thorough treatments of bootstrap methodology. Under general conditions Hall and Martin[14] showed that, if $\hat{\theta}$ has a bias of $O(n^{-1})$ which can be consistently estimated, then the estimator

$$\tilde{\theta} = \hat{\theta} - B^{(\text{boot})} = 2\hat{\theta} - \overline{\theta}^*$$

has $O(n^{-2})$bias.

The estimates of the bias in the gasoline yield data case study were obtained by simulation from the fitted model in Table 1, and thus are parametric bootstrap estimates of the bias.

Asymptotic Bias Correction

Another widely used class of explicit methods involves the approximation of $B(\theta)$ by $b\left(\hat{\theta}\right)/n$ which is the first-term in the right hand side of Eq. (4) evaluated at $\hat{\theta}$. Cox and Snell,[15] in their investigation of higher order properties of residuals in general parametric models, derive an expression for the first-order bias term $b(\theta)/n$ in Eq. (4), when $\hat{\theta}$ is the maximum likelihood estimator. That expression has sparked a still-active research stream in correcting the bias by using the estimator

$$\tilde{\theta} = \hat{\theta} - b\left(\hat{\theta}\right)/n.$$

Efron[16] showed that $\tilde{\theta}$ has bias of order $o(n^{-1})$ which is of smaller order than the $O(n^{-1})$ bias of the maximum likelihood estimator and that the asymptotic variance of any estimator with $O(n^{-2})$ bias is greater or equal to the asymptotic variance of $\tilde{\theta}$ (second-order efficiency). For the interested reader, Pace and Salvan[17] give a thorough discussion of those properties.

Landmark studies in the literature for asymptotic bias corrections are Cook et al.[18] who investigate correcting the bias in nonlinear regression models with Normal errors and Cordeiro and McCullagh[19] who treat generalized linear models with interesting results on the shrinkage properties of the reduced-bias estimators in binomial regression models and an attractive implementation through one supplementary reweighted least squares iteration. Furthermore, Botter and Cordeiro[20] and Cordeiro and Toyama Udo[21] extend the results in Cordeiro and McCullagh[19] and derive the first-order biases for generalized linear and nonlinear models with dispersion covariates.

The general form of the first-order bias term of the maximum likelihood estimator can be found in matrix form in Kosmidis and Firth[22]. Specifically,

$$\frac{b(\theta)}{n} = -\{F(\theta)\}^{-1} A(\theta),$$

where $A(\theta)$ is a p-dimensional vector with components

$$A_t(\theta) = \frac{1}{2}\mathrm{tr}\left[\{F(\theta)\}^{-1}\left\{P_t(\theta) + Q_t(\theta)\right\}\right] \quad (t = 1, \ldots, p), \tag{5}$$

and where

$$P_t(\theta) = E_\theta\left\{S(\theta)S^T(\theta)S_t(\theta)\right\} \quad (t = 1, \ldots, p),$$

$$Q_t(\theta) = -E_\theta\left\{I(\theta)S_t(\theta)\right\} \quad (t = 1, \ldots, p),$$

are higher order joint null moments of the gradient and the matrix of second derivatives of the log-likelihood.

Breslow and Lin[23] derive the expressions for the asymptotic biases in generalized linear mixed models for various estimation methods and used those to correct for the bias. Higher order corrections have also appeared in the literature[24] where expressions for $b(\theta)/n + b_2(\theta)/n^2$ in Eq. (4) are obtained. The expressions involved for such higher order corrections are too cumbersome requiring enormous effort in derivation and implementation, and there is always the danger that the benefits in estimation from this effort are only marginal, if any, compared to methods that are based on simply removing the first-order bias term.

Advantages and Disadvantages of Explicit Methods

The main advantage of all explicit methods is the simplicity of their application. Once an estimate of bias is available, reduction of bias is simply a matter of an one-step procedure where the estimated bias is subtracted from the estimates. Nevertheless because of their explicit dependence on $\hat{\theta}$, explicit methods directly inherit any of the instabilities of the original estimator. Such cases involve models with categorical responses where there is a positive probability that the maximum likelihood estimator is not finite (see Ref 25 for conditions that characterize when infinite estimates occur in multinomial response models) and have been the subject of study in works like Mehrabi and Matthews,[26] Heinze and Schemper,[27] Bull et al.,[28] Kosmidis and Firth,[29] Kosmidis and Firth,[2] and Kosmidis.[30] In particular, Kosmidis[30] relates to the case study of the proportional odds models, discussed below.

Furthermore, asymptotic bias correction methods have the disadvantage that are only applicable when $b(\theta)/n$ can be obtained in closed-form, which can be a tedious or even impractical task for many models (see, e.g., Ref 3 where the expressions for $b(\theta)/n$ are given for Beta regression models).

IMPLICIT METHODS

Implicit methods approximate $B(\theta)$ at the target estimator $\tilde{\theta}$ and then solve Eq. (3) with respect to $\tilde{\theta}$. Hence, $\tilde{\theta}$ is the solution of an implicit equation.

Indirect Inference

Indirect inference is a class of inferential procedures that appeared in the Econometrics literature in

Gourieroux et al.[31] and can be used for bias reduction. The simplest approach to bias reduction via indirect inference attempts to solve the equation

$$\tilde{\theta} = \hat{\theta} - B\left(\tilde{\theta}\right),$$

by approximating $B(\theta)$ at $\tilde{\theta}$ through parametric bootstrap. Kuk[32] independently produced the same idea for reducing the bias in the estimation of generalized linear models with random effects and Jiang and Turnbull[33] give a comprehensive review of indirect inference from a statistical point of view. Furthermore, Gourieroux et al.[34] and Phillips[35] discuss bias reduction through indirect inference in econometric applications. Pfeffermann and Correa[36] give an alternative approach to bias reduction which is in line with the basic idea of indirect inference.

Bias-Reducing Adjusted Score Equations

For the case where $\hat{\theta}$ is the maximum likelihood estimator and under the usual regularity conditions, Firth[37] and Kosmidis and Firth[29] investigate what penalties need to be added to $S(\theta)$ in order to get an estimator that has asymptotically smaller bias than that of the maximum likelihood estimator. In its simplest form such an approach requires finding $\tilde{\theta}$ by solving the adjusted score equations

$$S\left(\tilde{\theta}\right) + A\left(\tilde{\theta}\right) = 0, \qquad (6)$$

with $A(\theta)$ as given in Eq. (5). Then $\tilde{\theta}$ is an estimator with $o(n^{-1})$ bias. Equation 6 can be rewritten as

$$\left\{F\left(\tilde{\theta}\right)\right\}^{-1} S\left(\tilde{\theta}\right) = \frac{b\left(\tilde{\theta}\right)}{n},$$

which reveals that $\tilde{\theta}$ is another approximate solution to Eq. (3) because $b(\theta)/n$ approximates $B(\theta)$ up to order $O(n^{-2})$ and $\{F(\theta)\}^{-1}S(\theta)$ is the $O(n^{-1/2})$ term in the asymptotic expansion of $\hat{\theta} - \theta$ evaluated at $\theta := \tilde{\theta}$.

An important property of the estimator based on adjusted score functions, which is also shared by the estimator from asymptotic bias correction, is that it has the same asymptotic distribution as the maximum likelihood estimator, namely a Normal distribution centered at the true parameter value with variance-covariance matrix the Cramér-Rao lower bound $\{F(\theta)\}^{-1}$. Hence, the first-order methods that are used for the maximum likelihood estimator, like Wald-type confidence intervals, score tests for model comparison, and so on, are unaltered in their form and apply directly by using the new estimators.

It is noteworthy that in the case of full exponential families (e.g., logistic regression and Poisson log-linear models) the solution of Eq. (6) can be obtained by direct maximization of a penalized likelihood where the penalty is the Jeffreys[38] invariant prior (see Ref 37 for details). It should also be stressed that not all models admit a penalized likelihood interpretation of bias reduction via adjusted scores. Kosmidis and Firth[29] give an easy-to-check necessary and sufficient condition that identifies which univariate generalized linear models admit such penalized likelihood interpretation and provide the form of the resultant penalties when the condition holds. That condition is a restriction on the variance function of the responses in terms of the derivative of the chosen link function.

Advantages and Disadvantages

The main disadvantage of implicit methods is that their application requires the solution of a set of implicit equations which in most of the useful cases requires numerical optimization. This task is even more computationally demanding for indirect inference approaches in general models because of the necessity to approximate the bias function in a p-dimensional space. Furthermore, indirect inference approaches inherit the disadvantages of explicit methods because they explicitly depend on the original estimator.

The approach in Firth[37] and Kosmidis and Firth,[29] on the other hand, does not directly depend on $\hat{\theta}$ and hence has gained considerable attention compared to the other approaches. Another reason for the considerable adaptation of this method are recent advances which simplify application through either iterated first-order bias adjustments (see Refs 3, 22) or iterated maximum likelihood fits on pseudo observations (see Refs 2, 29, 30). Of course, as for the asymptotic bias correction methods the adjusted score equation approach to bias reduction has the disadvantage of being directly applicable only under the same conditions that guarantee the good limiting behavior of the maximum likelihood estimator and only when the score functions, Fisher information and the first-order bias term of the maximum likelihood estimator are available in closed-form.

PROPORTIONAL ODDS MODELS

This example was analyzed in Kosmidis.[30] The data set in Table 3 is from Randall[39] and concerns a factorial experiment for investigating factors that

TABLE 3 | The Wine Tasting Data (Randall[39]).

		Bitterness Scale				
Temperature	Contact	1	2	3	4	5
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

affect the bitterness of white wine. There are two factors in the experiment, 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 nine 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 five ordered categories, 1, 2, 3, 4, 5.

The task of the analysis is to test whether there are departures from the assumption of proportional odds. For performing such a test we use the more general partial proportional odds model of Peterson et al.[40] with

$$\log \frac{\gamma_{rs}}{1-\gamma_{rs}} = \alpha_s - \beta_s w_r - \theta z_r \quad (r=1,\dots,4; s=1,\dots,4), \tag{7}$$

where w_r and z_r are dummy variables representing the factors temperature and contact, respectively, $\alpha_1, \alpha_2, \alpha_3\, \alpha_4, \beta_1, \beta_2, \beta_3, \beta_4, \theta$ are model parameters and γ_{rs} is the cumulative probability for the sth category at the rth combination of levels for temperature and contact. Then we can check for departures from the proportional odds assumption by testing the hypothesis $\beta_1 = \beta_2 = \beta_3 = \beta_4$, effectively comparing Eq. (7) to the proportional odds nested model that is implied by the hypothesis.

Table 4 shows the maximum likelihood estimates for model Eq. (7) and the corresponding estimated standard errors as reported by the clm function of the R package ordinal.[41] It is directly apparent that the absolute value of the estimates and estimated standard errors for the parameters α_4, β_1 and β_4 is very large. Actually, these would diverge to infinity as the stopping criteria of the iterative fitting procedure used become stricter and the number of allowed iterations increases. The estimates for the remaining parameters are all finite and will preserve the value shown in Table 4 even if the number of allowed iterations increases. This is an instance of the problems that practitioners may face when dealing with categorical response models. Using a Wald-type statistic based on the maximum likelihood estimator for testing the hypothesis of proportional odds would be adventurous here because such a statistic explicitly depends on the estimates of β_1, β_2, β_3 and β_4. Of course, given that the likelihood is close to its maximal value at the estimates in Table 3, a likelihood ratio test can be used instead; the likelihood ratio test for this particular example has been carried out in Christensen.[42]

TABLE 4 | The Maximum Likelihood and the Reduced-Bias Estimates for the Parameters of Model (7), the Corresponding Estimated Standard Errors (in parenthesis) and the Values of the Corresponding *Z*-Statistic for the Hypothesis that the Corresponding Parameter is zero. The Maximum Likelihood Estimates and *Z*-statistics are as Reported by the clm R Package Ordinal[41].

	Maximum Likelihood		Adjusted Score Equations	
Parameter	Estimates	*Z*-Statistic	Estimates	*Z*-Statistic
α_1	−1.27 (0.51)	−2.46	−1.19 (0.50)	−2.40
α_2	1.10 (0.44)	2.52	1.06 (0.44)	2.42
α_3	3.77 (0.80)	4.68	3.50 (0.74)	4.73
α_4	28.90 (193125.63)	0.00	5.20 (1.47)	3.52
β_1	25.10 (112072.69)	0.00	2.62 (1.52)	1.72
β_2	2.15 (0.59)	3.65	2.05 (0.58)	3.54
β_3	2.87 (0.82)	3.52	2.65 (0.75)	3.51
β_4	26.55 (193125.63)	0.00	2.96 (1.50)	1.98
θ	1.47 (0.47)	3.13	1.40 (0.46)	3.02

Note here that methods like the bootstrap and jackknife would require special considerations for their application in a well-designed experiment like the above, the question to be answered being what comprises an observation to be resampled or left-out. Even if such considerations were resolved, bootstrap and jackknife would be prone to the problem of infinite estimates. The latter is also true for the estimator based on asymptotic bias corrections and for indirect inference.

Kosmidis[30] derives the adjusted score equations for cumulative link models, and uses them to calculate the reduced-bias estimates shown in the right of Table 4. The reduced-bias estimates based on the adjusted score functions are finite and, through the asymptotic normality of the reduced-bias estimator, they can form the basis of a Wald-test for the hypothesis $\beta_1 = \beta_2 = \beta_3 = \beta_4$. This test has been carried out in Kosmidis[30] and gives a p-value of 0.861, providing no evidence against the hypothesis of proportional odds.

Furthermore, the values of the Z-statistics for α_4, β_1 and β_4 in Table 4 are essentially zero when based on the maximum likelihood estimator. This is typical behavior when the estimates diverge to infinity and it happens because the estimated standard errors diverge much faster than the estimates, irrespective of whether or not there is evidence against the individual hypotheses. This is also true if we were testing individual hypothesis at values other than zero, and can lead to invalid conclusions if the maximum likelihood output is interpreted naively; as shown in Table 4, the Z-statistics based on the reduced-bias estimates are far from being zero.

Such inferential pitfalls with the use of the maximum likelihood estimator are not specific to partial proportional odds models. For most models for categorical and discrete data (binomial-response models like the logistic regression, multinomial response models, Poisson log-linear models, and so on) there is a positive probability of infinite estimates. Bias reduction through adjusted score functions has been found to provide a solution to those problems and the corresponding methodology is quickly gaining in popularity and has found its way to commercial software like Stata and SAS. Open-source solutions include the logistf R package[43] for logistic regressions which is based on the work in Heinze and Schemper,[27] the pmlr R package[44] for multinomial logistic regressions based on the work of Bull et al.[28] and the brglm R package[45,46] which at the time of writing handles all binomial-response models. At the time of writing, the brglm R package is being extended for the next major update which will handle all generalized linear models, including multinomial logistic regression[2] and ordinal response models.[30]

GASOLINE YIELD DATA REVISITED

In this section, the reduced-bias estimates for the parameters of model 1 are calculated using jackknife, bootstrap, asymptotic bias correction, and the approach of bias-reducing adjusted score functions. The full parametric bootstrap estimate of the bias has been obtained in our earlier treatment showing that the bias on the regression parameters is of no consequence. A fully nonparametric bootstrap where the bootstrap samples are produced by sub-sampling with replacement the full response-covariate combinations $(y_i, s_{i1}, \ldots, s_{i9}, t_i)$ $(i = 1, \ldots, n)$ is not advisable here because the 9 dummy variables $s_{i1}, \ldots, s_{i9}$ are representing 10 distinct experimental settings and sub-sampling those will result in singular fits with high probability (see also Ref 13, § 6.3 for a description of such problems in the simpler case of multiple linear regression). An intermediate sub-sampling strategy is to resample residuals and use them with the original model matrix to get samples for the response. This strategy lies between fully nonparametric bootstrap and fully parametric bootstrap (see Ref 13, § 7). Residual resampling works well in multiple linear regression because the response is related linearly to the regression parameters, which is not true for Beta regression. For more complicated models like generalized linear models and Beta regression, an appropriate residual definition has to be chosen. Because Beta responses are restricted in (0,1), the best option is to resample residuals on the scale of the linear predictor and then transform back to the response scale using the inverse of the logistic link, obtaining bootstrap samples for the response (see Ref 13 expression (7.13) for rationale and implementation). In the current case we choose the 'standardized weighted residual 2' of Espinheira et al.[47] because it appears to be the one that is least sensitive to the inherent skewness of the response.

The reduced-bias estimates of ϕ using jackknife, residual-resampling bootstrap (with 9999 bootstrap samples), asymptotic bias correction and bias-reducing adjusted score functions are 165.682, 236.003, 261.206, and 261.038, respectively, all indicating that the maximum likelihood estimator of ϕ is prone to substantial upward bias. The simulations in Kosmidis and Firth[22] illustrate that asymptotic bias correction and the bias-reducing adjusted score functions, correctly inflate the estimated standard errors to the extent that almost the exact coverage of the first-order Wald-type confidence intervals is recovered.

DISCUSSION AND CONCLUSION

As can be seen from the earlier case-studies, reduced-bias estimators can form the basis of asymptotic inferential procedures that have better performance than the corresponding procedures based on the initial estimator. Heinze and Schemper,[27] Bull et al.,[48] Kosmidis,[49] Kosmidis and Firth,[22,29] and Grün et al.[3] all demonstrate that such improved procedures are delivered either by using the penalized likelihood that results from the approximation of Eq. (3), or by replacing the initial estimator with the reduced-bias estimator in Wald-type pivots, as was done in the case-studies of this review.

At the time of writing the current review there is no general answer to which of the methods that have been reviewed here produces better results. All methods deliver estimators that have $o(n^{-1})$ bias which is asymptotically smaller than the $O(n^{-1})$ bias of the maximum likelihood estimator. In models with categorical or discrete responses, the adjusted score equations approach is preferable to the other bias reduction approaches because the resultant estimates appear to be always finite, even in cases where the maximum likelihood estimates are infinite (see, e.g., Refs 27, 48, 29, 2, 30 for generalized linear and nonlinear models with binomial, multinomial, and Poisson responses). This has led researchers to promoting the routine use of the adjusted score equations in such models as an improved alternative to maximum likelihood.

The general use, though, of the adjusted score equations approach is limited by its dependence on a closed-form expression for the first-order bias of the maximum likelihood estimator which may not be readily available or even intractable (e.g., generalized linear mixed effects models).

At this point, we should also stress that improving bias does not always have desirable effects; an improvement in bias can sometimes result in inflation of the mean squared error, through an inflation in the estimator's variance. The use of simple simulation studies, similar to the one in Kosmidis and Firth[22] is recommended for checking whether that is the case. If that is the case then the use of reduced-bias estimates in test statistics and confidence intervals is not recommended.

Furthermore, bias is a parameterization-specific quantity and any attempt to improve it will violate the invariance properties of the maximum likelihood estimator. Hence, bias-reduction methods should be used with care, unless the parameterization is fixed either by the context or by the practitioner.

All the discussion in the current review has focused on the effect that bias can have and the benefits of its reduction in cases where the usual regularity conditions are satisfied. An important research avenue is the reduction of bias under departures from the regularity conditions and especially when the dimension of the parameter space increases with the sample size. Lancaster[50] gives a review of the issues that econometricians and applied statisticians face in such settings. A viable route toward reduction of bias in such cases comes from the use of modified profile likelihood methods (see, Ref 51 for a brief introduction), which have been successfully used for reducing the bias in the estimation of dynamic panel data models in Bartolucci et al.[52] Another route is the appropriate adaptation of indirect inference approaches or of other approximate solutions of Eq. (3) in such settings. The Econometric community is currently active in this direction, with a recent example being Gouriéroux et al.[53] where indirect inference is applied to dynamic panel data models. These early attempts are only indicative that, there is still much to be explored and much work to be done on the topic of bias reduction in parametric estimation.

ACKNOWLEDGMENTS

The author thanks the Editor, the Review Editor, and three anonymous Referees for detailed, helpful comments which have substantially improved the presentation.

REFERENCES

1. Lehmann EL, Casella G. Theory of Point Estimation. 2nd ed. New York: Springer; 1998.
2. Kosmidis I, Firth D. Multinomial logit bias reduction via the poisson log-linear model. *Biometrika* 2011, 98:755–759.
3. Grün B, Kosmidis I, Zeileis A. Extended β regression in R: Shaken, stirred, mixed, and partitioned. *J Statist Softw* 2012, 48:1–25.
4. Prater NH. Estimate gasoline yields from crudes. *Petroleum Refiner* 1956, 35:236–238.
5. Cox DR, Hinkley DV. *Theoretical Statistics*. London: Chapman & Hall Ltd.; 1974.
6. McCullagh P. *Tensor Methods in Statistics*. London: Chapman and Hall; 1987.
7. Quenouille MH. Notes on bias in estimation. *Biometrika* 1956, 43:353–360.
8. Chernick MR. The jackknife: a resampling method with connections to the bootstrap. *Wiley Interdiscip Rev Comput Stat* 2012, 4:224–226.
9. Efron B. *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia, PA: SIAM [Society for Industrial and Applied Mathematics]; 1982.
10. Schucany WR, Gray HL, Owen DB. On bias reduction in estimation. *J Am Stat Assoc* 1971, 66:524–533.
11. Hesterberg T. Bootstrap. *Wiley Interdiscip Rev Comput Stat* 2011, 3:497–526.
12. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York: Chapman & Hall Ltd; 1993.
13. Davison AC, Hinkley DV. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press; 1997.

14. Hall P, Martin MA. Exact convergence rate of bootstrap quantile variance estimator. *Probab Theory Rel Fields* 1988, 80:261–268.
15. Cox DR, Snell EJ. A general definition of residuals (with discussion). *J Roy Statist Soc Ser B Methodol* 1968, 30:248–275.
16. Efron B. Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *Ann Stat* 1975, 3:1189–1217.
17. Pace L, Salvan A. *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. London: World Scientific; 1997.
18. Cook RD, Tsai C-L, Wei BC. Bias in nonlinear regression. *Biometrika* 1986, 73:615–623.
19. Cordeiro GM, McCullagh P. Bias correction in generalized linear models. *J Roy Statist Soc Ser B Methodol* 1991, 53:629–643.
20. Botter DA, Cordeiro G. Improved estimators for generalized linear models with dispersion covariates. *J Stat Comput Simul* 1998, 62:91–104.
21. Cordeiro G, Toyama Udo M. Bias correction in generalized nonlinear models with dispersion covariates. *Commun Stat Theory Methods* 2008, 37:2219–225.
22. Kosmidis I, Firth D. A generic algorithm for reducing bias in parametric estimation. *Electron J Stat* 2010, 4:1097–1112.
23. Breslow NE, Lin X. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* 1995, 82:81–91.
24. Cordeiro G, Barroso L. A third-order bias corrected estimate in generalized linear models. *Test* 2007, 16:76–89.
25. Albert A, Anderson J. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 1984, 71:1–10.
26. Mehrabi Y, Matthews JNS. Likelihood-based methods for bias reduction in limiting dilution assays. *Biometrics* 1995, 51:1543–1549.
27. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med* 2002, 21:2409–2419.
28. Bull SB, Mak C, Greenwood C. A modified score function estimator for multinomial logistic regression in small samples. *Comput Stat Data Anal* 2002, 39: 57–74.
29. Kosmidis I, Firth D. Bias reduction in exponential family nonlinear models. *Biometrika* 2009, 96:793–804.
30. Kosmidis I. Improved estimation in cumulative link models. *J Roy Stat Soc Ser B* 2013b. doi:10.1111/rssb.12025. *ArXiv e-prints*. arXiv:1204.0105v3 [stat.ME].
31. Gourieroux C, Monfort A, Renault E. Indirect inference. *J Appl Econom* 1993, 8S:85–118.
32. Kuk AYC. Asymptotically unbiased estimation in generalized linear models with random effects. *J Roy Statist Soc Ser B Methodol* 1995, 57:395–407.
33. Jiang W, Turnbull B. The indirect method: inference based on intermediate statistics – synthesis and examples. *Statist Sci* 2004, 19:239–263.
34. Gourieroux C, Renault E, Touzi N. Chapter 13: Calibration by simulation for small sample bias correction. In: Mariano R, Schuermann T, Weeks MJ, eds. *Simulation-based Inference in Econometrics: Methods and Applications*. 1st ed. Cambridge: Cambridge University Press; 2000, 328–358.
35. Phillips PCB. Folklore theorems, implicit maps, and indirect inference. *Econometrica* 2012, 80: 425–454.
36. Pfeffermann D, Correa S. Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation. *Biometrika* 2012, 99: 457–472.
37. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993, 80:27–38.
38. Jeffreys H. An invariant form for the prior probability in estimation problems. *Proc Roy Soc Lond* 1946, 186:453–461.
39. Randall J. The analysis of sensory data by generalised linear model. *Biom J* 1989, 7:781–793.
40. Peterson B, Harrell J, Frank E. Partial proportional odds models for ordinal response variables. *Appl Stat* 1990, 39:205–217.
41. Christensen RHB. *Ordinal—regression models for ordinal data*. R package version 2012.01-19; 2012b. Available at: http://cran.r-project.org/src/contrib/Archive/ordinal/. (Accessed September 20, 2013).
42. Christensen RHB. *Ordinal—a tutorial on fitting cumulative link models with the ordinal package*. R package version 2012.01-19; 2012a. Available at: http://cran.r-project.org/src/contrib/Archive/ordinal/. (Accessed September 20, 2013).
43. Heinze G, Ploner M, Dunkler D, Southworth H. *Logistf: Firth's bias reduced logistic regression*. R package version 1.21; 2013.
44. Colby S, Lee S, Lewinger JP, Bull S. *Pmlr: Penalized Multinomial Logistic Regression*. R package version 1.0; 2010.
45. Kosmidis I. *On iterative adjustment of responses for the reduction of bias in binary regression models*. Technical Report 09-36. CRiSM working paper series; 2009.
46. Kosmidis I. *brglm: Bias reduction in binary-response Generalized Linear Models*. R package version 0.5-9; 2013a.
47. Espinheira PL, Ferrari SLP, Cribari-Neto F. On β regression residuals. *J Appl Stat* 2008, 35: 407–419.
48. Bull SB, Lewinger JB, Lee SSF. Confidence intervals for multinomial logistic regression in sparse data. *Stat Med* 2007, 26:903–918.

49. Kosmidis I. *Bias reduction in exponential family non-linear models*. Ph.D. thesis, Department of Statistics, University of Warwick; 2007.
50. Lancaster T. The incidental parameter problem since 1948. *J Econom* 2000, 95:391–413.
51. Reid N. Likelihood inference. *Wiley Interdiscip Rev Comput Stat* 2010, 2:517–525.
52. Bartolucci F, Bellio R, Salvan A, Sartori N. Modified profile likelihood for panel data models. Technical report. Available at SSRN repository; 2012.
53. Gouriéroux C, Phillips PCB, Yu J. Indirect inference for dynamic panel models. *J Econom* 2010, 157: 68–77.