Bias reduction as a remedy to the consequences of infinite estimates in Poisson and Tobit regression

Susanne Köll\textsuperscript{a}, Ioannis Kosmidis\textsuperscript{b}, Christian Kleiber\textsuperscript{c}, Achim Zeileis\textsuperscript{a,∗}

\textsuperscript{a}Faculty of Economics and Statistics, Universität Innsbruck, Austria
\textsuperscript{b}Department of Statistics, University of Warwick & The Alan Turing Institute, United Kingdom
\textsuperscript{c}Faculty of Business and Economics, Universität Basel, Switzerland

Abstract

Data separation is a well-studied phenomenon that can cause problems in the estimation and inference from binary response models. Complete or quasi-complete separation occurs when there is a combination of regressors in the model whose value can perfectly predict one or both outcomes. In such cases, and such cases only, the maximum likelihood estimates and the corresponding standard errors are infinite. It is less widely known that the same can happen in further microeconometric models. One of the few works in the area is Santos Silva and Tenreyro (2010) who note that the existence of a point mass at zero implies that $F(0; \mu_i, \phi) = F(0; \mu_i, \phi)$, where $F(\cdot; \mu_i, \phi)$ is the density or probability mass function corresponding to $F(\cdot; \mu_i, \phi)$.

The simplest but arguably often-encountered occurrence of data separation in practice is when there is a regressor $x_{i,k} \in \{0, 1\}$ such that $y_i = 0$ for all $i \in \{1, \ldots, n\}$ with $x_{i,k} = 1$. Assuming that $y_1, \ldots, y_n$ are independent conditionally on $x_1, \ldots, x_n$, the log-likelihood $\ell(\beta, \phi)$ for the model defined by (1) and (2) can be decomposed as

$$
\ell(\beta, \phi) = \sum_{x_{i,k}=0} \log f(y_i; h(x^{\top}_{i,-k} \beta_{-k}), \phi) + \sum_{x_{i,k}=1} \log F(0; h(x^{\top}_{i,-k} \beta_{-k} + \beta_k), \phi),
$$

where $a_{-k}$ indicates the sub-vector formed from a vector $a$ after omitting its $k$-th component.

Term (3) is exactly the log-likelihood without the $k$-th regressor and based only on the observations with $x_{i,k} = 0$. Under the extra assumption that $F(0; \mu_i, \phi)$ is monotonically decreasing with $\mu_i$ (which is true, for example, in Poisson and Tobit regression models), $\beta_k$ will diverge to $-\infty$ during maximization, so that (4) achieves its maximum value of 0. Then, the maximization of term (3) with respect to $\beta_{-k}$ yields the maximum likelihood (ML) estimate of $\beta_{-k}$. So, the ML estimate of $\beta_{-k}$ will be the same as the ML estimate obtained by maximizing the log-likelihood without the $k$-th regressor over the subset of observations with $x_{i,k} = 0$.

1. Sources of separation in regression models

Suppose that the non-negative random variable $y_i$ has a distribution with a point mass at zero. Suppose that the distribution function of $y_i$ is $F(\cdot; \mu_i, \phi)$ ($i = 1, \ldots, n$), where the scalar parameter $\mu_i$ is a centrality measure (e.g., the mean), and the parameter $\phi$ represents higher-order characteristics of the distribution (e.g., dispersion).

A regression model can be formulated as

$$
y_i \sim F(\cdot; \mu_i, \phi), \quad \mu_i = h(x_i^{\top} \beta) \quad (i = 1, \ldots, n),
$$

where $x_i$ is a vector of regressors with $\text{dim}(x_i) = p$, which is observed along with $y_i$, and $h(\cdot)$ is a monotonically increasing function that links $\mu_i$ to $x_i$ and a parameter vector $\beta$. The model specification in (1) and (2) covers a range of models, including models for binary, multinomial, ordinal, and count models, models for limited dependent variables such as the Tobit model and its extensions, and zero-inflated and two-part hurdle models.

The existence of a point mass at zero implies that $f(0; \mu_i, \phi) = F(0; \mu_i, \phi)$, where $f(\cdot; \mu_i, \phi)$ is the density or probability mass function corresponding to $F(\cdot; \mu_i, \phi)$.

Keywords: Bias reduction, Data separation, Shrinkage

\textit{JEL:} C13, C24, C25, C52

\textsuperscript{∗}Corresponding author

Email addresses: Susanne.Berger@Student.uibk.ac.at (Susanne Köll), Ioannis.Kosmidis@Warwick.ac.uk (Ioannis Kosmidis), Christian.Kleiber@unibas.ch (Christian Kleiber), Achim.Zeileis@uibk.ac.at (Achim Zeileis)

\textsuperscript{1}Note that the discussion here extends to the case where the support of the response is bounded below or above. If the lower boundary is a constant $b \neq 0$, we can use $y_i - b$. Similarly, if the upper boundary is $b$, we can use $5 - y_i$. 
As Santos Silva and Tenreyro (2010) show for Poisson regression, the same situation can occur more generally, when separation occurs for a certain linear combination of regressors. Our discussion above extends their considerations beyond log-link models and Poisson regression.

2. Estimating regression models with separated data

Albert and Anderson (1984) showed that infinite estimates in multinomial logistic regression occur if and only if there is data separation. Since then, the consequences of infinite estimates to estimation and inference have been well-studied for binomial and multinomial responses.

A popular remedy in the statistics literature is to replace the ML estimator with shrinkage estimators that are guaranteed to take finite values (see, for example, Gelman et al. (2008) for using shrinkage priors in the estimation of binary regression models). The probably most-used estimator of this kind comes from the solution of the bias-reducing adjusted score equations in Firth (1993) (see, for example, Heine and Schenper (2002) and Zorn (2005) for accessible detailed accounts), which guarantee estimators with smaller asymptotic bias than what the ML estimator typically has (Firth, 1993; Kosmidis and Firth, 2009).

In contrast, the majority of methods that have been put forward in the econometrics literature are typically based on omitting the regressors that are responsible for the infinite estimates. Such practice can be problematic as we discuss in the following sections.

2.1. Omitting regressors affected by separation

Santos Silva and Tenreyro (2010) show that the regressors responsible for separation in Poisson models can be easily identified by running a least squares regression on the non-boundary observations and checking for perfect collinearity among the regressors. The same strategy is also applicable for Tobit regression models.

Having identified the collinear regressors associated with separation, Santos Silva and Tenreyro (2010) propose to simply omit those and re-estimate the model using the full data set with all n observations. The same strategy is also adopted in Cameron and Trivedi (2013, Chapter 6.2), who suggest to drop the separating regressor from the binary model part of a count data hurdle model.

However, this strategy only leads to consistent estimates if the omitted regressors are, in fact, not relevant, or were constructed to specifically indicate a zero response (e.g., in the artificial data set used in the illustrations of Santos Silva and Tenreyro (2011). In contrast, when a highly informative regressor is omitted, separation will be replaced by a systematic misspecification of the model (Heine and Schenper (2002) Zorn (2005)). In that situation, consistent estimates can be obtained by not only omitting the regressor but also the observations responsible for separation, i.e., considering only the first term in the likelihood and dropping (4).

2.2. Bias reduction

Kosmidis and Firth (2020) have formally shown that, in logistic regression models with full-rank model matrix, the bias-reduced (BR) estimators coming from the adjusted score equations in Firth (1993) (i) have always finite value and (ii) shrink towards zero in the direction of maximizing the Fisher information about the parameters. There are also strong empirical findings that the finiteness of the BR estimator extends beyond logit models.

A desirable feature of the bias-reducing adjustments to the score functions is that they are asymptotically dominated by the score functions. As a result, inference that relies on the BR estimates (Wald tests, information criteria, etc.) can be performed as usual by simply using the BR estimates in place of the ML estimates. This makes BR estimation a rather attractive alternative approach for dealing with separation, without omitting regressors.

While bias reduction is a well-established remedy for data separation in binary regression models, it is less well known that it is effective also in more general settings such as generalized nonlinear models (Kosmidis and Firth, 2009), and, as illustrated here, the models in Section 1.

3. Illustration

Similarly to Santos Silva and Tenreyro (2011), we consider models with intercept $x_{i,1} = 1$ and regressors $x_{i,2}$ and $x_{i,3}$ ($i = 1, \ldots, n$). The values for $x_{i,2}$ are generated from a uniform distribution as $x_{i,2} \sim U(-1, 1)$. The values for $x_{i,3}$ are, then, generated from Bernoulli distributions as $x_{i,3} \sim B(\pi)$ if $x_{i,2} > 0$ and $x_{i,3} \sim B(1 - \pi)$ otherwise, in order to allow for correlation between the two regressors.

The responses for the Poisson model are generated from $h(x_i^\top \beta) = \exp(x_i^\top \beta)$ and the Poisson distribution for $F$ (with known dispersion $\phi = 1$). The responses for the Tobit model are generated from a latent normal distribution $N(x_i^\top \beta, \phi)$ with variance $\phi = 2$ and subsequent censoring by setting all negative responses to 0.

For illustration purposes, we generate a single artificial data set involving $n = 100$ regressor values with $\pi = 0.25$, and Poisson and Tobit responses using $\beta_1 = 1, \beta_2 = 1$ and $\beta_3 = -10$. In both cases, separation occurs due to the extreme value for the coefficient of $x_{i,3}$. In the Appendix, we carry out a thorough simulation study with 10,000 data sets for a range of combinations of $n$ and $\pi$ and $\beta_2 = -3$ so that separation occurs with smaller probability.

We estimate the models from the artificial data using ML and BR estimation using all $n = 100$ observations, and ML estimation of the reduced model after omitting $x_{i,3}$ either by using just the subset of the data set with $x_{i,2} = 0$ (ML/sub), or all $n = 100$ observations as proposed by Santos Silva and Tenreyro (2010) (ML/SST).

The bias-reducing adjusted score equations for the Poisson regression are $\sum_{i=1}^n(y_i + h_i/2 - \mu_i)x_i = 0_p$, where $0_p$ is a p-vector of zeros and $h_i = x_i^\top(X^\topWX)^{-1}x_i\mu_i$ with $W = \text{diag}\{\mu_1, \ldots, \mu_n\}$ (Firth, 1992). It is solved with the
Table 1: Comparison of different approaches when dealing with separation in a Poisson model. N is the number of observations used.

|        | ML    | BR    | ML/sub | ML/SST |
|--------|-------|-------|--------|--------|
| (Intercept) | 0.951 | 0.958 | 0.951 | 0.350  |
|         | (0.100) | (0.099) | (0.100) | (0.096) |
| x2     | 1.011 | 1.006 | 1.111 | 1.662  |
|         | (0.158) | (0.157) | (0.158) | (0.144) |
| x3     | -20.907 | -5.174 |         |        |
|         | (2242.463) | (1.416) |         |        |
| Log-likelihood | -107.364 | -107.809 | -107.364 | -109.028 |
| N      | 100   | 100   | 55     | 100    |

Table 2: Comparison of different approaches when dealing with separation in a Tobit model. N is the number of observations used.

|        | ML    | BR    | ML/sub | ML/SST |
|--------|-------|-------|--------|--------|
| (Intercept) | 1.135 | 1.142 | 1.135 | -0.125 |
|         | (0.208) | (0.210) | (0.208) | (0.251) |
| x2     | 0.719 | 0.705 | 0.719 | 2.074  |
|         | (0.364) | (0.359) | (0.364) | (0.404) |
| x3     | -11.238 | -4.218 |         |        |
|         | (60452.270) | (0.891) |         |        |
| (Variance) | 1.912 | 1.970 | 1.912 | 3.440  |
|         | (0.422) | (0.434) | (0.422) | (0.795) |
| Log-likelihood | -87.633 | -88.101 | -87.633 | -118.935 |
| N      | 100   | 100   | 55     | 100    |

The brglm_fit method from the R package brglm2 [Kosmidis 2020]. For the Tobit model we derived the adjusted score equations along with an implementation in the R package brtobit [Köll et al. 2021]. The derivations are tedious but not complicated and are provided in the Appendix. Tables 1 and 2 show the results from estimating the Poisson and Tobit models, respectively, with the four different strategies. The following remarks can be made:

- Standard ML estimation using all observations leads to a large estimate of $\beta_3$ with even larger standard error. As a result, a standard Wald test results in no evidence against the hypothesis that $x_3$ should not be in the model, despite the fact that $\beta_3 = -10$ when generating the data makes $x_3$ perhaps the most influential regressor.\(^2\)

- The ML/sub strategy, i.e., estimating the model without $x_2$ only for the 0 observations with $x_{1,2} = 0$, yields exactly the same estimates as ML because it optimizes the term \((\hat{y} - \hat{\mu})^2\) after setting \(\hat{\beta}_3\) to zero.

- Compared to ML and ML/sub, BR has the advantage of returning a finite estimate and standard error for $\beta_3$. Hence a Wald test can be directly used to examine the evidence against $\beta_3 = 0$. The other parameter estimates and the log-likelihood are close to ML. Similarly to binary response models, bias reduction here slightly shrinks the parameter estimates of $\beta_2$ and $\beta_3$ towards zero.

- Finally, the estimates from ML/SST, where regressor $x_3$ is omitted and all observations are used, appear to be far from the values we used to generate the data. This is due to the fact that $x_3$ is not only highly informative but also correlated with $x_2$.

Moreover, the simulation experiments in the Appendix provide evidence that the BR estimates are always finite, and result in Wald-type intervals with better coverage.

References

Albert, A., Anderson, J.A., 1984. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71, 1–10. doi:10.1093/biomet/71.1.1

Amemiya, T., 1973. Regression analysis when the dependent variable is truncated normal. Econometrica 41, 997–1016. doi:10.2307/1914031

Cameron, A.C., Trivedi, P.K., 2013. Regression Analysis of Count Data. 2nd ed., Cambridge University Press, New York.

Firth, D., 1992. Bias reduction of maximum likelihood estimates. Biometrika 79, 527–532. doi:10.1093/biomet/79.1.527

Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.S., 2008. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics 2, 1360–1383. doi:10.1214/08-aoas141

Gourieroux, C., 2000. Econometrics of Qualitative Dependent Variables. Cambridge University Press, Cambridge.

Heinze, G., Schepers, M., 2002. A solution to the problem of separation in logistic regression. Statistics in Medicine 21, 2409–2419. doi:10.1002/sim.1047

Köll, S., Kosmidis, I., Kleiber, C., Zeileis, A., 2021. brtobit: Bias-Reduced Tobit Regression Models. URL: https://R-Forge.R-project.org/projects/topmodels/ R package version 0.1-1/r1146.

Kosmidis, I., 2020. brglm2: Bias Reduction in Generalized Linear Models. URL: https://CRAN.R-project.org/package=brglm2 R package version 0.6-2.

Kosmidis, I., Firth, D., 2009. Bias reduction in exponential family parametric estimation. Electronic Journal of Statistics 3, 498–514. doi:10.1214/08-ajoas105

Kosmidis, I., Firth, D., 2010. A general algorithm for reducing bias in parametric estimation. Electronic Journal of Statistics 4, 1097–1112. doi:10.1214/10-ejs579

Kosmidis, I., Firth, D., 2020. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika doi:10.1093/biomet/asaa055

Kosmidis, I., Firth, D., 2010. A generic algorithm for reducing bias in parametric estimation. Electronic Journal of Statistics 4, 1097–1112. doi:10.1214/10-ejs579

Kosmidis, I., Firth, D., 2020. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika doi:10.1093/biomet/asaa055

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. URL: https://www.R-project.org/

Santos Silva, J.M.C., Tenreyro, S., 2010. On the existence of the maximum likelihood estimates in Poisson regression. Economics Letters 107, 310–312. doi:10.1016/j.econlets.2010.02.020

Santos Silva, J.M.C., Tenreyro, S., 2011. Poisson: Some convergence issues. Stata Journal 11, 207–212. doi:10.1177/153687661101100209

Zorn, C., 2005. A solution to separation in binary response models. Political Analysis 13, 157–170. doi:10.1093/pan/mp1009
Appendix A. Bias-reducing adjusted score functions for Tobit regression

The Tobit model is one of the classic models of microeconometrics. Fundamental results were obtained by Amemiya [1973]. A detailed account of basic properties is available in, e.g., Gourieroux (2000). Here we provide the building blocks for bias-reduced estimation of the Tobit model.

Denote by $\ell(\theta)$ the log-likelihood function for a Tobit regression model with full-rank, $n \times p$ model matrix $X$ with rows the $p$-vectors $x_1, \ldots, x_n$, and a $(p+1)$-vector of parameters $\theta = (\beta^\top, \phi)^\top$ with regression parameters $\beta$ and variance $\phi$. Then, $\ell(\theta) = \sum_{i=1}^n [(1-d_i) \log(1-F_i) - d_i \log(\phi)/2 - d_i (y_i - \eta_i)^2/(2\phi^2)],$ where $d_i = 1$ if $y_i > 0$ and $d_i = 0$ if $y_i \leq 0$, $\eta_i = x_i^\top \beta$, and $F_i$ is the standard normal distribution function at $\eta_i/\sqrt{\phi}$. The score vector is

$$s(\theta) = \nabla \ell(\theta) = \left[ s_\beta(\theta) \quad s_\phi(\theta) \right] = \left[ \sum_{i=1}^n \left\{ \frac{(d_i-1) \lambda_i}{\sqrt{\phi} \phi} + \frac{d_i (y_i - \eta_i)}{\phi} \right\} x_i \right],$$

where $\lambda_i = f_i/(1-F_i)$ and $f_i$ is the density function of the standard normal distribution at $\eta_i/\sqrt{\phi}$.

The observed information matrix, $j(\theta) = -\nabla^2 \ell(\theta)$, has the form

$$j(\theta) = \begin{bmatrix} j_{\beta\beta}(\theta) & j_{\beta\phi}(\theta) \\ j_{\phi\beta}(\theta) & j_{\phi\phi}(\theta) \end{bmatrix},$$

where, setting $\nu_i = f_i/(1-F_i)^2$,

$$j_{\beta\beta}(\theta) = \sum_{i=1}^n \left[ \frac{\nu_i (d_i - 1)}{\phi} \left\{ \frac{f_i}{\sqrt{\phi}} - \frac{(1-F_i) \eta_i}{\phi} \right\} - \frac{d_i}{\phi} \right] x_i x_i^\top,$$

$$j_{\beta\phi}(\theta) = \sum_{i=1}^n \left[ \frac{\nu_i (d_i - 1)}{2\phi^{3/2}} \left\{ \frac{(1-F_i) \eta_i^2}{\phi} - 1 + F_i - \eta_i f_i \right\} - \frac{d_i (y_i - \eta_i)}{\phi^2} \right] x_i,$$

$$j_{\phi\beta}(\theta) = j_{\beta\phi}(\theta)^\top,$$

$$j_{\phi\phi}(\theta) = \sum_{i=1}^n \left[ \frac{\nu_i (1-d_i)}{4\phi^{3/2}} \left\{ \frac{(1-F_i) \eta_i^2}{\phi} - 3(1-F_i) \eta_i - \frac{\eta_i^2 f_i}{\phi} \right\} + \frac{d_i}{2\phi^2} - \frac{d_i (y_i - \eta_i)^2}{\phi^2} \right].$$

As shown in Kosmidis and Firth (2010), a BR estimator for $\theta$ results as the solution of the adjusted score equations $s(\theta) + A(\theta) = 0_{p+1}$, where the vector $A(\theta)$ has $t$-th component $A_t(\theta) = \text{tr}\{[i(\theta)]^{-1} [P_t(\theta) + Q_t(\theta)]\}/2$ ($t = 1, \ldots, p+1$). In the above expression, $Q_t(\theta) = -E(j(\theta)s_t(\theta))$ and $P_t(\theta) = E(s(\theta)s_t^\top(\theta)s_t(\theta))$, where $i(\theta) = E(j(\theta))$ is the expected information matrix. The R package brtobit implements $i(\theta)$, $Q_t(\theta)$, and $P_t(\theta)$, and solves the bias-reducing adjusted score equations for general Tobit regressions using the quasi Fisher-scoring scheme proposed in Kosmidis and Firth (2010).

The matrices $i(\theta)$, $Q_t(\theta)$ and $P_t(\theta)$ have the same block structure as $j(\theta)$ and, directly by their definition, closed-form expressions for their blocks result by taking expectations of the appropriate products of blocks of $s(\theta)$ and $j(\theta)$. By direct inspection of the expressions for $s(\theta)$ and $j(\theta)$, the required expectations result by noting that $E(d_i^{m/2}) = F_i, E((1-d_i)^m) = 1 - F_i, E(d_i^m(1-d_i)^j) = 0, E(d_i^m(1-d_i)^j(y_i - \eta_i)^6) = 0,$ and by computing $E(d_i^m(y_i - \eta_i)^l)$ $(k,l,m = 1, \ldots, 6)$. For the latter expression, note that $E(d_i^m(y_i - \eta_i)^l) = F_i E((y_i - \eta_i)^l \mid y_i > 0)$, and that some algebra gives

$$E(y_i - \eta_i \mid y_i > 0) = \sqrt{\phi} \xi_i,$$

$$E((y_i - \eta_i)^2 \mid y_i > 0) = \phi - \sqrt{\phi} \eta_i \xi_i,$$

$$E((y_i - \eta_i)^3 \mid y_i > 0) = \sqrt{\phi} \xi_i (\eta_i^2 + 2\phi),$$

$$E((y_i - \eta_i)^4 \mid y_i > 0) = 3\phi^2 - \eta_i^4 \sqrt{\phi} \xi_i - 3\phi^{3/2} \eta_i \xi_i,$$

$$E((y_i - \eta_i)^5 \mid y_i > 0) = \sqrt{\phi} \eta_i^4 \xi_i + 4\phi^{3/2} \xi_i (\eta_i^2 + 2\phi),$$

$$E((y_i - \eta_i)^6 \mid y_i > 0) = -\eta_i \sqrt{\phi} \xi_i (\eta_i^4 + 5\eta_i^2 \phi + 15\phi^2) + 15\phi^3,$$

where $\xi_i = f_i/F_i$. The expected information,

$$i(\theta) = \begin{bmatrix} E(j_{\beta\beta}(\theta)) & E(j_{\beta\phi}(\theta)) \\ E(j_{\phi\beta}(\theta)) & E(j_{\phi\phi}(\theta)) \end{bmatrix},$$

is a matrix of the same form as $j(\theta)$.
has elements

\[
\begin{align*}
E(\beta_\beta(\theta)) &= -\frac{1}{\phi} \sum_{i=1}^{n} \left\{ \frac{\eta_i f_i}{\phi} - \lambda_i f_i - F_i \right\} x_i x_i^T, \\
E(\beta_\phi(\theta)) &= \frac{1}{2\phi^{3/2}} \sum_{i=1}^{n} f_i \left\{ \frac{\eta_i^2}{\phi} + 1 - \lambda_i \frac{\eta_i}{\sqrt{\phi}} \right\} x_i^T, \\
E(\phi(\theta)) &= E(\phi(\theta))^T, \\
E(\phi(\theta)) &= -\frac{1}{4\phi^2} \sum_{i=1}^{n} \left\{ f_i \eta_i^2 + f_i \frac{\eta_i}{\sqrt{\phi}} - \lambda_i f_i \frac{\eta_i^2}{\phi} - 2F_i \right\}.
\end{align*}
\]

Furthermore, for \( t \in \{1, \ldots, p\} \),

\[
Q_t(\theta) = - \left[ \begin{array}{c}
E(\beta_\beta s_{\beta_t}) \\
E(\beta_\phi s_{\beta_t}) \\
E(\phi_\beta s_{\beta_t}) \\
E(\phi_\phi s_{\beta_t}) \\
E(s_{\beta s_{\beta_t}}) \\
E(s_{\beta s_{\phi}}) \\
E(s_{\phi s_{\beta_t}}) \\
E(s_{\phi s_{\phi}})
\end{array} \right] \quad \text{and} \quad P_t(\theta) = \left[ \begin{array}{c}
E(\beta_\beta s_{\beta_t}) \\
E(\beta_\phi s_{\beta_t}) \\
E(\phi_\beta s_{\beta_t}) \\
E(\phi_\phi s_{\beta_t}) \\
E(s_{\beta s_{\beta_t}}) \\
E(s_{\beta s_{\phi}}) \\
E(s_{\phi s_{\beta_t}}) \\
E(s_{\phi s_{\phi}})
\end{array} \right],
\]

and for \( t = p + 1 \),

\[
Q_{p+1}(\theta) = - \left[ \begin{array}{c}
E(\beta_\beta s_{\phi}) \\
E(\beta_\phi s_{\phi}) \\
E(\phi_\beta s_{\phi}) \\
E(\phi_\phi s_{\phi}) \\
E(s_{\beta s_{\beta_t}}) \\
E(s_{\beta s_{\phi}}) \\
E(s_{\phi s_{\beta_t}}) \\
E(s_{\phi s_{\phi}})
\end{array} \right] \quad \text{and} \quad P_{p+1}(\theta) = \left[ \begin{array}{c}
E(\beta_\beta s_{\beta_t}) \\
E(\beta_\phi s_{\beta_t}) \\
E(\phi_\beta s_{\beta_t}) \\
E(\phi_\phi s_{\beta_t}) \\
E(s_{\beta s_{\beta_t}}) \\
E(s_{\beta s_{\phi}}) \\
E(s_{\phi s_{\beta_t}}) \\
E(s_{\phi s_{\phi}})
\end{array} \right],
\]

where

\[
\begin{align*}
E(\beta_\beta s_{\beta_t}) &= \sum_{i=1}^{n} \left[ \frac{-f_i}{\phi^{3/2}} \left( \lambda_i^2 - \lambda_i \frac{\eta_i}{\sqrt{\phi}} + 1 \right) \right] x_i x_i^T x_{i,t}, \\
E(\beta_\phi s_{\beta_t}) &= \sum_{i=1}^{n} \left[ \frac{1}{2\phi^2} \lambda_i f_i \left\{ \frac{\eta_i^2}{\phi} + 1 + \lambda_i \frac{\eta_i}{\sqrt{\phi}} \right\} + \frac{1}{\phi^2} \left\{ F_i - \frac{\eta_i f_i}{\sqrt{\phi}} \right\} \right] x_i^T x_{i,t}, \\
E(\phi_\beta s_{\beta_t}) &= \sum_{i=1}^{n} \left[ \frac{1}{\phi^{5/2}} \lambda_i \frac{f_i \eta_i}{4\sqrt{\phi}} \left\{ \frac{\eta_i^2}{\phi} - 3 - \lambda_i \frac{\eta_i}{\sqrt{\phi}} \right\} + \frac{f_i \eta_i^2}{\phi} + 3f_i \frac{\eta_i}{2} \right] x_{i,t}, \\
E(\phi_\phi s_{\beta_t}) &= \sum_{i=1}^{n} \left[ \frac{-f_i^2}{2\phi^{7/2}} \left\{ \lambda_i - \frac{\eta_i}{\sqrt{\phi}} \right\} - \frac{\eta_i f_i}{2\phi^{7/2}} \right] x_i x_i^T x_{i,t}, \\
E(\beta_\phi s_{\phi}) &= \sum_{i=1}^{n} \left[ \frac{\eta_i f_i}{4\sqrt{\phi}} \left\{ \frac{\eta_i^2}{\phi} - 1 - \lambda_i \frac{\eta_i}{\sqrt{\phi}} \right\} + \frac{f_i \eta_i}{2\phi^{3/2}} \left\{ 1 + \frac{\eta_i^2}{\phi} \right\} \right] x_i^T x_{i,t}, \\
E(\phi_\phi s_{\phi}) &= \sum_{i=1}^{n} \left[ \lambda_i \frac{\eta_i^2}{8\phi^2} \left\{ -\lambda_i f_i + 3f_i + \lambda_i \frac{f_i \eta_i}{\sqrt{\phi}} \right\} + \frac{F_i}{\phi^3} - \frac{3\eta_i f_i}{4\phi^{7/2}} - \frac{f_i^3}{2\phi^{9/2}} \right] x_i^T x_{i,t}, \\
E(s_{\beta s_{\beta_t}}) &= \sum_{i=1}^{n} \left[ \frac{-\lambda_i^2 f_i}{\phi^{3/2}} + \frac{f_i}{\phi^{7/2}} \left\{ \eta_i^2 + 2\phi \right\} \right] x_i x_i^T x_{i,t}, \\
E(s_{\beta s_{\phi}}) &= \sum_{i=1}^{n} \left[ \frac{\eta_i f_i}{2\phi^{7/2}} \left\{ \lambda_i^2 - 2 - \frac{\eta_i^2}{\phi} \right\} + \frac{F_i}{\phi^3} \right] x_i x_i^T x_{i,t}, \\
E(s_{\beta s_{\phi}}) &= \sum_{i=1}^{n} \left[ \frac{f_i \eta_i^2}{2\phi^{7/2}} \left\{ \lambda_i^2 - 2 \right\} + \frac{F_i}{\phi^2} - \frac{f_i \eta_i^2}{2\phi^{7/2}} \right] x_i x_i^T x_{i,t}, \\
E(s_{\phi s_{\phi}}) &= \sum_{i=1}^{n} \left[ \frac{-f_i \eta_i^2}{2\phi^{7/2}} \left\{ -\lambda_i^2 + 1 \right\} + \frac{f_i}{4\phi^{5/2}} \left\{ 5 + \frac{\eta_i^2}{\phi^2} \right\} \right] x_i x_i^T x_{i,t}, \\
E(s_{\phi s_{\phi}}) &= \sum_{i=1}^{n} \left[ \frac{-f_i \eta_i^2}{8\phi^{9/2}} \left\{ \lambda_i^2 - 2 - \frac{\eta_i^2}{\phi} \right\} + \frac{F_i}{\phi^3} - \frac{9f_i \eta_i}{8\phi^{7/2}} \right].
\end{align*}
\]
Appendix B. Simulation

The aim of the simulation experiment is to compare the performance of the BR and ML estimator in count and limited dependent variable models with varying probabilities of infinite ML estimates. The comparison here is in terms of bias, variance, and empirical coverage of nominally 95% Wald-type confidence intervals based on the asymptotic normality of the estimators. Our results were obtained using R 4.0.3 (R Core Team [2020]). Random variables were generated using the default methods for the relevant distributions, which in turn rely on uniform random numbers obtained by the Mersenne Twister, currently R’s default generator.

The same data generating process as in Section 3 of the main paper is considered, with the coefficient of the binary regressor $x_2$ set to the less extreme value $\beta_3 = -3$. The amount of correlation between $x_2$ and $x_3$ varies with $\pi \in \{0, 1/8, 1/4, 3/8, 1/2\}$ so that increasing the value of $\pi$ leads to decreasing the probability of infinite estimates. The sample sizes we consider are $n \in \{25, 50, 100, 200, 400\}$. For each combination of $\pi$ and $n$, 10,000 independent samples are simulated, and the parameters of the Poisson and Tobit regression models in Section 3 are estimated using maximum likelihood and bias reduction. The estimates are then used to compute simulation-based estimates of the bias, variance, and coverage probability for $\beta_3$.

For the ML estimator, the bias, variance, and coverage probabilities are computed conditionally on the finiteness of the ML estimates. We classify an ML estimate as infinite if the corresponding estimated standard error exceeds 20. In effect, we are assuming that if the standard error exceeds 20, the Fisher scoring iteration for ML stopped while moving along an asymptote on the log-likelihood surface, hence, at a point where the inverse negative hessian has at least one massive diagonal element. The heuristic value 20 is conservative even for $n = 25$. This has been verified through a pilot simulation study to estimate the variance of the reduced-bias estimator, which has the same asymptotic distribution as the ML estimator. No convergence issues were encountered and the maximum estimated standard error of the reduced-bias estimators across simulation settings, parameters, and sample sizes was 8.3 for Tobit and 5.5 for Poisson regression.

For BR estimation, the estimates appear to be always finite. So, we estimate biases, variances and coverage probabilities both conditionally on the finiteness of the ML estimates and unconditionally. We note here that a direct comparison of conditional and unconditional summaries is not formally valid, but gets more and more informative as the probability of infinite estimates decreases.

Figures B.1, B.2, B.3, and B.4 show the estimated probability that the ML and BR estimate of $\beta_3$ are infinite, the estimated bias, the estimated variance, and the estimated coverage probability of 95% Wald-type confidence intervals, respectively, for the Poisson model. Figures B.5, B.6, B.7, and B.8 show the corresponding results for the Tobit model.

The results for Poisson and Tobit regression lead to similar insights:

- Bias reduction via adjusted score functions always yields finite estimates.
- The BR estimator has bias close to zero even for small sample sizes.
- Wald-type confidence intervals based on BR estimates have good coverage properties.
- The variances of the BR and ML estimator get closer to each other and closer to zero as $n$ increases. This is exactly what the theory suggests because the score functions asymptotically dominate the bias-reducing adjustments.
Figure B.1: Probability of infinite estimates for $\beta_3$ (Poisson).

Figure B.2: Bias of estimates for $\beta_3$ (Poisson).

Figure B.3: Variance of estimates for $\beta_3$ (Poisson).

Figure B.4: Coverage of 95% Wald-type confidence intervals for $\beta_3$ (Poisson).
Figure B.5: Probability of infinite estimates for $\beta_3$ (Tobit).

Figure B.6: Bias of estimates for $\beta_3$ (Tobit).

Figure B.7: Variance of estimates for $\beta_3$ (Tobit).

Figure B.8: Coverage of 95% Wald-type confidence intervals for $\beta_3$ (Tobit).