Finite Sample Bias Reduction For Misspecified Models With Extensions to High Dimensional Data

Andrew T. Humbert

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Brian Leroux, Co-Chair

Noah Simon, Co-Chair

Marco Carone

Program Authorized to Offer Degree:

Biostatistics – Public Health

University of Washington

Abstract

Finite Sample Bias Reduction for Misspecified Models
With Extensions to High Dimensional Data

Andrew T. Humbert

Chairs of the Supervisory Committee:

Professor Brian Leroux
Biostatistics

Professor Noah Simon
Biostatistics

While maximum likelihood estimates (MLEs) from generalized linear models have desirable asymptotic properties, for finite samples, these estimates can be biased for finite samples. Current bias reduction methods account for either misspecification of the model or separation in the data; however current methods cannot address both simultaneously. We provide a detailed characterization of the finite sample bias for log-linear models to understand how model components contribute to the bias of MLEs. This dissertation proposes a new robust bias reduction method that effectively reduces finite sample bias in the presence of misspecification and separation and does not result in a loss of asymptotic performance. These results are demonstrated analytically as well as empirically through simulations. This method is extended to clustered data using modified generalized estimating equations. We explore the effects finite sample bias and bias reduction methods have on transformed estimates as well as developing a post-transformation bias reduction process. The effects of finite sample bias and bias reduction methods are also explored for meta-analysis. Lastly, we discuss scenarios where these bias reduction methods may or may not be effective in high dimensional data settings with a focus on sparse models.

Contents

List of Figures

List of Tables

Acknowledgments

I gratefully acknowledge the support provide by the University of Washington Department of Biostatistics.

I would like to thank the members of my advisory committee: Marco Carone, Susanne May, Chris Delaney, and Brandon Guthrie for their questions, comments, and suggestions.

In particular, I would like to thank my co-advisors Brian Leroux, who has stuck with me since the start of my journey, and Noah Simon, who helped me push through to the end. I would also like to thank my fellow classmates who provided support and inspiration both inside and outside of the classroom.

Finally, I would like to express my appreciation for my family, whose unwaivering support started long before my dissertation.

Chapter 1

Introduction

1.1 Motivating Scenario

Binary outcomes are frequently observed in medical research. Often, the goal is to assess the association between a binary outcome and a predictor of interest while adjusting for other variables, such as confounders. The association can be quantified in many different ways. Risk difference (RD), relative risk (RR), and odds ratio (OR) are three commonly used measurements for the association. The strengths and weaknesses of each measure have been previously discussed [63]. We focus on settings where a measure of association on a multiplicative scale is preferred (RR or OR).

While the OR is a useful quantity in case control studies, where RR and RD cannot be estimated, the OR is not easy to interpret. Schwarz et al critique the reporting of the OR from a study looking at the association between race and referral for cardiac catherization [64]. The study found the OR comparing the rates of referrals for cardiac catherization between black and white patients was 0.6. The media misinterpreted the OR as a RR, claiming black patients were 40% less likely to receive a referral. The actual RR was 0.93 corresponding to a 7% lower rate of referrals. A similar misinterpretation was found from a review that found authors' affiliation to the tobacco industry was associated with an 88 times higher odds of publishing a conclusion that passive smoking is

not harmful compared to authors without an affiliation with the tobacco industry [4]. The OR was misinterpreted as a RR in a news report which concluded "authors with affiliations to the tobacco industry are 88 times more likely to conclude that passive smoking is not harmful than if the review was written by authors with no connection to the tobacco industry." [79] The true RR was 7 which is notably smaller than the OR of 88 [3]. Because of the potential for misinterpretation of the OR, the RR are preferred when possible.

For rare outcomes, the RR can be reasonably approximated by the OR; however, in settings where the event is common, the OR can misrepresent the RR away from the null (as seen in the examples above). Zhang and Yu propose a method to approximate RR from an OR even when the outcome is common [83]. However, Zhang and Yu's method tends to be biased away from the null and can lead to anti-conservative confidence intervals [53]. Thus, estimating the RR directly is preferred [41, 46].

While risk differences can be calculated using least squares regression, a more general family of models is needed to estimate RR. We begin discussing simple linear regression and its limitations before discussing the broader family of generalized linear models (GLMs), which can be used to estimate RR, and its challenges, most notably the presence of finite sample bias of parameter estimates.

1.2 Generalized Linear Models

Ordinary Linear Regression

Linear regression is a statistical method to assess the association between an outcome and a predictor(s). Least squares (LS) regression is a commonly used estimation technique for linear regression. Ordinary LS regression assumes the outcomes, Y, can be modeled by a linear combination of covariates, represented by the design matrix X, and parameter values, β such that $\mathrm{E}[Y|X] = X\beta$. Estimation of β allows for inference on the association between the outcome and predictor(s) and can also be used to make predictions on future outcomes. LS estimates minimize the sum

of squared residuals between predicted and observed values of the outcome. These estimates are asymptotically normal under mild regularity conditions. When the errors are homoskedastic and uncorrelated, least squares estimates have minimum variance among unbiased estimates. If the homoskedasticity assumption is not met, weighted least squares can be used to obtain unbiased estimates with minimum variance.

However, in many scenarios, LS regression may not be appropriate. More flexible models are needed when the variance is a function of the mean or when the linear combination of predictors is associated with a function of the mean outcome, rather than the mean itself. This is particularly relevant for binary data.

While risk differences may be calculated using ordinary linear regression, an alternative mean model is needed to calculate the relative risk. Furthermore, with binary data, the variance can be written as a function of the mean. Ordinary linear regression does not have the flexibility to account for this association. GLMs are a more general family of models that can allow us to estimate RR. We provide a brief overview of GLMs before discussing how they can be used to estimate RR.

Generalized Linear Models

Generalized linear models are a flexible family of linear regression models that includes both logistic and Poisson regression [55]. GLMs assume the outcome, Y, follows a density from the exponential family with the form $f(Y;\theta,\alpha) = \exp((Y\theta - b(\theta)/\alpha + c(Y,\alpha))$ for known functions $b(\cdot)$ and $c(\cdot)$ and an unknown canonical parameter θ and unknown nuisance parameter, α. In this form, $\mathrm{E}[Y] = b'(\theta)$ and $\mathrm{Var}(Y) = b''(\theta)/\alpha$. For n independent outcomes, y, the likelihood can be written as $f_n(y;\theta,\alpha) = \prod_{i=1}^{n} f(y_i;\theta_i,\alpha)$. The likelihood can then be maximized with respect to $\theta = \{\theta_1,...,\theta_n\}$ to obtain the maximum likelihood estimate (MLE), $\hat{\theta}$. In practice, the MLE is found by solving the score $U_n(\hat{\theta}) = \frac{\partial log f_n(y;\hat{\theta},\alpha)}{\partial\theta} = 0$

Instead of estimating the natural parameter, θ, it is often of interest to estimate parameters that quantify the association between predictors and the outcome. With GLMs, different scales of the

Table 1.1: Distribution, link, and score functions for commonly used GLM

Distribution	Link function	$U(\beta)$
Gaussian	identity	$X^t(y - X\beta)$
Poisson	log	$X^t(y - \exp(X\beta))$
Binomial	logit	$X^t(y - \frac{\exp(X\beta)}{1+\exp(X\beta)})$
Binomial	log	$X^t \frac{\exp(X\beta)}{1-\exp(X\beta)}(y - \exp(X\beta))$

mean can be modeled by using a link function, g, on the mean. The mean model now has the form $g(\mathrm{E}[y_i|X_i]) = X_i\beta$ for an n-length vector of outcomes, y, an $n \times p$ design matrix X and a p-length vector of parameters, β. The score can now be written in terms of β rather than θ. The score then has the form

$$U_n(\beta) = D^T V^{-1}(y - g^{-1}(X\beta))$$

where $D^T = \frac{\partial g^{-1}(X\beta)}{\partial(\beta)}$ and V is a diagonal matrix where $V_{ii} = \mathrm{Var}(y_i)$, dependent on its probability distribution.

Solving $U_n(\hat{\beta}) = 0$ gives the MLE, $\hat{\beta}$, which maximizes the likelihood with respect to β. In general, no closed-form expression for $\hat{\beta}$ exists and iterative algorithms, such as the Newton-Raphson method or iterated weighted least squares are used to calculate the MLE.

Table 1.1 shows the score function for some commonly used distributions and link functions. When an identity link is used and a Gaussian distribution is specified, $\hat{\beta}$ is identical to the least squares estimate. Thus, ordinary LS can be viewed as a special case of a GLM. We also note that a binomial distribution and log link function can be used the estimate the log-RR which can be exponentiated to obtain an estimate of the RR.

1.2.1 Properties of the MLE

In addition to being used in estimation, the score is also useful in understanding the properties of the MLE. For ease of notation consider the score from a single-parameter model. The bias for

multi-parameter models is provided in Chapter 2. The Taylor expansion about $U(\beta)$ is

$$U(\hat{\beta}) = U(\beta) + \dot{U}(\beta)(\hat{\beta} - \beta) + r_n$$

where $\dot{U}(\beta) = \frac{\partial U(\beta)}{\partial \beta}$ and r_n is the higher order terms of the Taylor expansion. Taking the expected value and rearranging terms gives

$$\sqrt{n}(\hat{\beta} - \beta + B_n) \rightarrow N(0, I^{-1}(\beta))$$

where B_n is the bias of $\hat{\beta}$ and is $O(n^{-1})$ and $I(\beta) = -\mathrm{E}[\dot{U}(\beta)]$. Asymptotically, $\hat{\beta}$ is consistent, normally distributed, and achieves the Cramer-Rao lower bound. However, finite sample bias may be present. In settings where the score is linear, such as least squares regression, $B_n = 0$ and $\hat{\beta}$ is unbiased for finite samples. In settings where the score has curvature (as is the case for Poisson and logistic regression), $B_n \neq 0$ and the MLE will have finite sample bias.

In this dissertation, we engage with the first order approximation for B_n. Using a higher order Taylor expansion on the score, the first order bias approximation, $\frac{b_1}{n}$, can be written as

$$\frac{b_1}{n} = I^{-1}(\beta) I^{-1}(\beta) (J(\beta) + \frac{K(\beta)}{2}) \tag{1.1}$$

where $I(\beta) = -\mathrm{E}[\dot{U}(\beta)]$, $J(\beta) = \mathrm{E}[U(\beta)\dot{U}(\beta)]$ and $K = \mathrm{E}[\ddot{U}(\beta)]$ [17].

This expression will be used in the dissertation to provide an in depth characterization of the bias for correctly specified log-linear models in Chapter 2.

Bias Reduction Methods

Many methods are available to reduce the finite sample bias of the MLE. Bias reduction methods may be split into corrective and preventative methods. Corrective methods apply a post-hoc

modification to the MLE while preventative methods modify the score equation to directly calculate bias-reduced estimates. We denote the bias reduced estimate by $\tilde{\beta}$. Both corrective and preventative methods can be used to remove the first order bias such that $\mathrm{E}[\tilde{\beta} - \beta] = o(n^{-1})$.

Corrective Methods

A straightforward bias reduction estimate can be obtained by subtracting a bias approximation from the MLE. This method is referred to as a corrective method as it modifies the MLE. When using a theoretical approximation for the bias, the working distribution needs to be correctly specified to produce reliable bias reduction. The jackknife is a corrective method that does not rely on a theoretical expression for the bias approximation, as it estimates the bias non-parametrically. Both methods rely on an finite MLE. In many settings, the MLE may not be finite. This may be the result of separation in the data, in which the MLE may be infinity, or due to issues in model fitting algorithms that fail to converge to the MLE. Both of these issues frequently occur when trying to estimate the RR with a log-Binomial model.

Preventative Methods

Firth proposed a preventative method that does not rely on an estimable MLE [25]. By shifting the score function, a modified score function can be solved that calculates bias reduced estimates directly. The modified score can be solved using methods similar to those for solving the standard GLM score. Even when separation occurs in the data, estimates may still be calculated using the modified score [35]. Furthermore, estimates from the modified score function are asymptotically equivalent to the standard MLE, preserving asymptotic normality and efficiency.

1.3 Misspecified Models

Previous discussion assumed the model had been correctly specified; however, in practice the true model is rarely known and a working distribution from the exponential family is used instead.

Misspecification comes in many forms including misspecification of the mean model as well as misspecification of the working distribution. While the estimation procedure is the same under a misspecified distribution as a correctly specified distribution, the estimate is not an MLE as the true likelihood is not being maximized. Instead, estimates calculated using a misspecified distribution are referred to as quasi-maximum likelihood estimates (QMLEs). Under misspecification, when mild regularity conditions are met, the QMLE is consistent for the true parameter; however, robust estimation methods are needed to ensure the estimated variance of $\hat{\beta}$ is consistent [78]. A more detailed discussion of misspecified models will be provided in Chapter 3.

A misspecified model may be used even when the distribution of the outcome is known. This frequently occurs when estimating relative risks for binary outcomes as desired in our motivating scenario. While the natural model to use is a log-binomial model (correctly using a Bernoulli distribution and a log link function), many fitting algorithms have problems converging to the MLE. For this reason, Poisson regression is often used to estimate log-RR. Poisson regression uses a Poisson working distribution and a log link function; however, the Poisson model overestimates the variance of $\hat{\beta}$. The use of robust standard errors provides consistent estimates for the variance and asymptotically valid inference.

As in the case of a correctly specified model, while estimates from a misspecified model may be asymptotically consistent, finite sample bias may still be present. Furthermore, while the bias of the MLE and QMLE are of the same order, they are not identical. This is problematic when using bias reduction methods that rely on the expression in equation (1.1) as the approximation relies on a correctly specified distribution. The effect of misspecification on finite sample bias has not been evaluated.

1.3.1 Bias Reduction for Misspecified Models

Methods relying on equation (1.1) may not provide reliable bias reduction when the model is misspecified. While methods such as jackknife may still provide reliable bias reduction, no preventative bias reduction is available under misspecification. In Chapter 3, we develop a robust,

preventative bias reduction method that can be used in the presence of separation even when the working model is misspecified.

Clustered Data

Clustered data frequently occurs in application, when observations may not be independent. Repeated measurements on individuals or data collected by sites (such as hospitals) are common scenarios where clustering occurs. Because the assumption of independent outcomes is violated, standard inference techniques for GLM are no longer appropriate. Generalized linear mixed models (GLMMs) and generalized estimating equations (GEEs) are two models commonly used for clustered data. GLMMs are used to estimate conditional effects and GEEs are used to estimate marginal effects. Firth's method has been extended to GEEs for a correctly specified model by treating the estimating functions as if they were likelihood score functions [58]. However, this method assumes the working covariance matrix is independent of the parameters, which is often not the case. We propose an alternative bias reduction method, extending the robust bias reduction from Chapter 3 to account for clustering of the data.

1.4 Indirect Consequences of Finite Sample Bias

In addition to having a direct impact on estimates, bias can also have an indirect impact in settings where transformed estimates or pooled estimates are the primary focus. We provide an overview of two such settings below.

Bias of Transformed Estimates

In some scenarios, a transformed estimate is of primary interest. The delta-method can be used to obtain asymptotically consistent estimates; however, finite sample bias of the transformed estimate may still be present. In particular, we note that bias reduction of the estimate does not protect against bias in the transformed estimate. In Chapter 4, we discuss the impact of estimate bias and

bias reduction on transformed estimates and propose a post-transformation bias reduction method.

Bias in Meta-Analysis

In Chapter 4, we also explore the effect finite sample bias and bias reduction methods have in meta-analysis. Estimates from individual studies are often pooled in meta-analysis, allowing for more precise estimates, but not necessarily reducing the bias. The impact finite sample bias in individual studies has on the pooled estimates has not previously been explored.

1.5 Bias Reduction in High Dimensional Data Settings

A growing amount of research in the biomedical world uses high dimensional data, which pose a unique challenge since the number of features is larger than the sample size. Penalized regression is commonly used to shrink parameter estimates and prevent over-fitting, leading to better performance of the estimates. The lasso, ridge regression, and elastic net are three commonly used examples of penalized regression [70, 86]. These methods utilize a loss function (often a negative log likelihood) and a penalty term which induces shrinkage of parameter estimates. In particular, we focus on the lasso estimate

$$\hat{\beta}^L = \arg\min_{\beta}(l(\beta) + \lambda||\beta||_1) \tag{1.2}$$

where $l(\beta)$ is a loss function and λ is a tuning parameter which controls how much sparsity is induced. The solution to this minimization problem satisfies the Karush-Kuhn-Tucker condition, $U(\hat{\beta}) = \lambda\hat{k}$ where $\hat{k} = sign(\hat{\beta})$.

While the lasso estimate can lead to improved prediction, $\hat{\beta}^L$ is not unbiased and, in general, valid inference cannot be made from this estimate. The bias of $\hat{\beta}^L$ can be split into two components, the sparsity-induced bias which is a result of the ℓ_1-penalty, and the curvature-induced bias, which is a result of the curvature of the loss function. Methods for high dimensional inference, such as

desparsification and low-dimensional projection, utilize a 1-step correction to the lasso estimate to reduced the penalty-induced bias and allow for valid inference [73, 82]. While the curvature-induced bias does not impact asymptotic inference, it may be problematic in finite samples.

As an alternative to penalized regression, sub-models can also be used to make inference. We discuss sub-models that use a screening method to first select a subset of features. The subset of features is then used to make inference using a standard regression technique. In this setting, there is no sparsity-induced bias; however, curvature-induced bias is still present.

No known work addresses the curvature-induced bias in either of these settings. In Chapter 5, we explore the impact of extending bias-reduction methods from Chapter 3 to the high dimensional setting to reduce the curvature-induced bias.

1.6 Research Aims

We now provide a brief overview of the aims of the research for this dissertation.

In Chapter 2, we provide an in-depth characterization of the bias for correctly specified log-linear models and assess the impact the distribution of the outcome, the joint distribution of predictors, parameter values, and sample size all have on bias. In Chapter 3, we explore problems with current bias reduction methods when the distribution assumption for the model is incorrect. We also develop a robust bias reduction method that is able to provide reliable bias reduction even when the distribution assumption is not met. This method is extended to clustered data using modified GEEs. Chapter 4 assesses the impact of bias and bias reduction methods when using the delta-method or pooling estimates in meta-analysis. In Chapter 5, we explore extending bias reduction methods to high dimensional data through high dimensional inference methods as well as special cases of sub-models. Final conclusions and areas for future work are discussed in Chapter 6.

Chapter 2

Characterization of Bias for Correctly Specified GLMs

2.1 Introduction

Generalized linear models (GLMs) are a flexible family of regression models widely used in data analysis to assess the association between an outcome and covariates. GLMs consist of two components: 1) a working mean model and 2) a working distribution. In this chapter, we analyze factors that contribute to bias when both the mean model and working distribution are correctly specified. We provide a brief overview of GLMs before discussing how the bias of parameter estimates can be estimated. Through different scenarios, the bias is characterized analytically and empirically.

2.2 Background

2.2.1 Generalized Linear Models

GLMs are a generalization of ordinary linear regression formulated by Nelder and Wedderburn [55]. The flexibility of GLMs is a result of two components: a working distribution that can be

non-Gaussian and a mean model which allows for the mean of the response to be a non-linear function of the linear model. We discuss each of these components in more detail.

The first component of a GLM is a working probability distribution of the outcome that is a member of the exponential family. Let Y be an outcome with a corresponding parameter, θ, and a nuisance parameter, a, such that the working probability distribution of Y is

$$p(Y;\theta) = \exp((Y\theta - h(\theta))/a + c(Y,a))$$

for functions $h(\theta)$ and $c(Y,a)$. In this chapter, we assume $p(Y;\theta)$ is the true probability distribution. However, in many settings the true probability distribution is unknown or misspecified. We discuss settings where the working and true probability distribution are not equal in Chapter 3.

The second characteristic of a GLM is the mean model. The mean model relates the linear combination of predictors to the mean of the response through a (potentially) non-linear link function, $g()$. For a given outcome y, with $\mu = E[y|x]$, a $p \times 1$ vector of features, x, and a $p \times 1$ parameter of interest, β^*, the mean model is

$$g(\mu) = x^T\beta^*.$$

The working distribution and mean model can be used together to derive maximum likelihood estimates. This process is formulated below.

Since $p(Y;\theta)$ has the form of an exponential family $\mathrm{E}[Y|\theta,\alpha] = \mu = h'(\theta)$ and $\mathrm{Var}(Y) = h''(\theta)/a$. Note that $h'(\theta) = \mu = g^{-1}(x\beta^*)$. Then the probability distribution can be written in terms of $X\beta^*$ rather than θ. The resulting likelihood for n independent observations, $L(y;X\beta) = \prod_{i=1}^{n} p(y_i;X_i\beta)$, can be maximized with respect to β, where $y = \{y_1,...y_n\}$ and X_i is the i^{th} row of the $n \times p$ design matrix, X. The nuisance parameter is omitted from the notation as it does not

affect the estimate of the parameter of interest. The corresponding maximum likelihood estimate

$$\hat{\beta} = \arg\max_{\beta} L(y; X\beta)$$

is consistent for β^*. In general, no closed form expression exists to calculate $\hat{\beta}$. Instead, iterative fitting algorithms, such as the Newton-Raphson method or iteratively re-weighted least squares, are used to solve the score equation

$$U(\beta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log(L(y_i; X_i\beta))}{\partial \beta} = 0$$

for which the MLE is the solution.

While the MLE is asymptotically consistent, normally distributed and achieves the Cramer-Rao lower bound, it may still be biased in finite sample settings. Understanding the finite sample bias of $\hat{\beta}$ is important since many frequently used regression models, such as logistic regression or Poisson regression, are in the GLM family.

2.2.2 Bias of Maximum Likelihood Estimates for GLMs

The existence of the bias of $\hat{\beta}$ has been well documented, with parametric and non-parametric methods available to approximate the bias [16, 17, 21, 51, 71]. However, little work has been done to characterize the bias to understand what components of the model contribute to bias. Nemes et al. explored finite sample bias for logistic regression but primarily explored the effect of sample size and only explored two scenarios: one using a continuous predictor variable and the other a discrete predictor variable. Other work looks into optimal sampling techniques but is limited to stratified data using logistic regression [20]. None of the current literature attempts to extensively characterize bias or provide insight into what model components contribute to bias.

In this section, detailed characterization is provided, both theoretical and empirical, of the bias in numerous scenarios as well as intuition into what types of factors contribute to bias to provide

a broader understanding of bias of maximum likelihood estimates for GLMs. In particular we explore how the distribution of outcomes, the joint distribution of predictors, parameter values, and sample size all contribute to bias.

First Order Bias Approximation

To better understand the characteristics of bias, we first need an approximation for finite sample bias. Cox and Snell provide an expression for the first order bias for a correctly specified model [17]. We reiterate how the bias approximation is derived below, in the special case of a single parameter (for ease of notation) before providing the more general, multi-parameter expression (see Cox and Snell [17] for the original derivation). In this dissertation, we derive a generalized form of this approximation that allows for model misspecification (see Chapter 3). In the bias approximation below, we assume a correctly specified model.

First, we take a Taylor expansion of $U(\hat{\beta})$ centered at β^*:

$$0 = U(\beta^*) + \dot{U}(\beta^*)(\hat{\beta} - \beta^*) + \frac{\ddot{U}(\beta^*)(\hat{\beta} - \beta^*)^2}{2} + o(n^{-1}) \tag{2.1}$$

where $\dot{U}(\beta) = \frac{\partial U(\beta)}{\partial \beta}$ and $\ddot{U}(\beta) = \frac{\partial^2 U(\beta)}{\partial \beta^2}$.

Rearranging the terms in equation (2.1) and taking the expectation gives

$$\begin{aligned} \mathrm{E}[\dot{U}(\beta^*)]\,\mathrm{E}[\hat{\beta} - \beta^*] = &\mathrm{Cov}[U(\beta^*), (\hat{\beta} - \beta^*)] + \frac{1}{2}\mathrm{E}[\ddot{U}(\beta^*)]\,\mathrm{E}[(\hat{\beta} - \beta^*)^2] \\ &+ \frac{1}{2}\mathrm{Cov}[\ddot{U}_n(\beta^*), (\hat{\beta} - \beta^*)^2] + o(n^{-1}). \end{aligned}$$

A lower order Taylor expansion is used to substitute $I^{-1}\dot{U}_n(\beta^*) + o(n^{-1/2})$ for $\hat{\beta} - \beta^*$ on the right side of the equation. After working through the expectations and rearranging terms, the bias is

$$\begin{aligned} \mathrm{E}[\hat{\beta} - \beta^*] &= I^{-1} I^{-1} (J + \frac{K}{2}) + o(n^{-1}) \\ &= b(\beta^*) + o(n^{-1}) \end{aligned} \tag{2.2}$$

where $I = \mathrm{E}[\dot{U}(\beta^*)]$, $J = \mathrm{E}[U(\beta^*)\dot{U}(\beta^*)]$, $K = \mathrm{E}[\ddot{U}(\beta^*)]$.

This expansion can readily be extended to multi-parameter settings as well to derive the multivariate expression for the first order bias:

$$b(\beta_j^*) = \sum_r \sum_s \sum_t I^{rj} I^{st} (J_{t,rs} + \frac{K_{rst}}{2}) \tag{2.3}$$

where $U_r(\beta^*)$ is the r^{th} element of the score vector, $I_{rs} = \mathrm{E}[-\frac{\partial(U_r(\beta^*))}{\partial \beta_s^*}]$, $J_{t,rs} = \mathrm{E}[\frac{\partial(U_r(\beta^*))}{\partial \beta_s^*} U_t(\beta^*)]$, $K_{rst} = \mathrm{E}[-\frac{\partial^2(U_r(\beta^*))}{\partial \beta_s^* \partial \beta_t^*}]$ and I^{rs} denotes the $(r,s)^{th}$ element of I^{-1} [16, 17]. We view this expression as an alternative to tensor notation used by McCullagh [51].

2.3 Characterization of Bias

The general form of bias provides little insight into what factors contribute to bias. By exploring specific scenarios we are not only able to characterize the bias in these scenarios, but also provide intuition into what factors influence bias in more general settings. While sample size is one component that clearly impacts bias, we also want to explore the effects the true parameter value, the probability distribution of the outcome, and the distribution of the predictors all have on bias.

To explore these components, we choose to provide a characterization of bias for models using a log-link function. The reasons for using a log-link function are threefold.

First, the log-link function with a binomial likelihood is the basis for calculating relative risks for binary outcomes. While logistic regression is another option for assessing the relative association between a predictor and outcome through odds ratios (OR), OR can be difficult to interpret

and are often misinterpreted as RR [4, 64]. For rare events, the OR is a reasonable approximation for RR but for common events, the OR provides an approximation further from the null than the true RR. For this reason, directly estimating RR with a log-link is preferred when appropriate [41, 46]. Furthermore, some work has already been done in characterizing the bias for logistic regression [55]. Risk differences (RD) are also used to measure the association between a predictor and outcome. However, RD are calculated using a linear link function and therefore do not have finite sample bias.

Second, a log-link function is commonly used with different probability distributions in practice, such as the Poisson distribution, allowing us to assess the impact the distribution of outcomes has on bias for models commonly used.

Third, the inverse of the log link function is $g^{-1}(X\beta) = \exp(X\beta)$ and the derivative is $\frac{\partial g^{-1}(X\beta)}{\partial \beta} = X\exp(X\beta)$. This property allows for a simplification in the calculation of the derivatives needed to approximate bias. This simplification allows for bias characteristics to be explicitly seen.

Using three data scenarios, we explore how different model components influence the bias of the MLE. In addition to reporting the bias, we also explore other ways of measuring bias such as the percent bias and the bias relative to the standard error (SE).

Scenario 1: Intercept Only (Exploration of Distribution Effect and Methods for Classifying Bias)

To gain a basic understanding of how the probability distribution of the outcome, sample size, and parameter values impact bias, we consider a simple, intercept only model.

For the $n \times 1$ vector of independent outcomes Y, let $\mu = \mathrm{E}[Y]$ and $\log(\mu) = \beta_0$. While this model has limited utility in practice, it allows for a simple, closed form expression of the first order bias that makes the dependency of bias on the distribution of the outcome, parameter value, and sample size clear. Using the bias approximation in equation (2.2), the first order bias can be calculated for a variety of different probability distributions of the outcome.

The bias is approximated as $\frac{1}{2n} - \frac{1}{2n\exp(\beta_0)}$ for Bernoulli outcomes, $-\frac{1}{2n\exp(\beta_0)}$ for Poisson out-

comes, and $-\frac{1}{2n}$ for exponential outcomes (the respective inverse Fisher's information in these settings). From the expressions of bias, the dependency on sample size and the distribution of the outcome is clear. The value of β_0 also generally impacts the bias though interestingly, the bias for exponential outcomes does not depend on the parameter value for the log link function.

Table 2.1 contains the first order bias approximations for a wider variety of commonly used distributions and link functions, though we only explore scenarios using the log link in this chapter. Appendix A.3 contains details on the derivations of the first order bias for the scenarios presented in the table.

Table 2.1: Bias for intercept only GLM for commonly used distributions and link functions

Distribution	Identity	Log	Logit	Inverse
Binomial	0	$\frac{\exp(\beta_0)-1}{2n\exp(\beta_0)}$	$\frac{1-\exp(2\beta_0)}{2n\exp(\beta_0)}$	–
Poisson	0	$\frac{1}{2n\exp(\beta_0)}$	–	–
Exponential	0	$-\frac{1}{2n}$	–	$\frac{\beta_0}{n}$

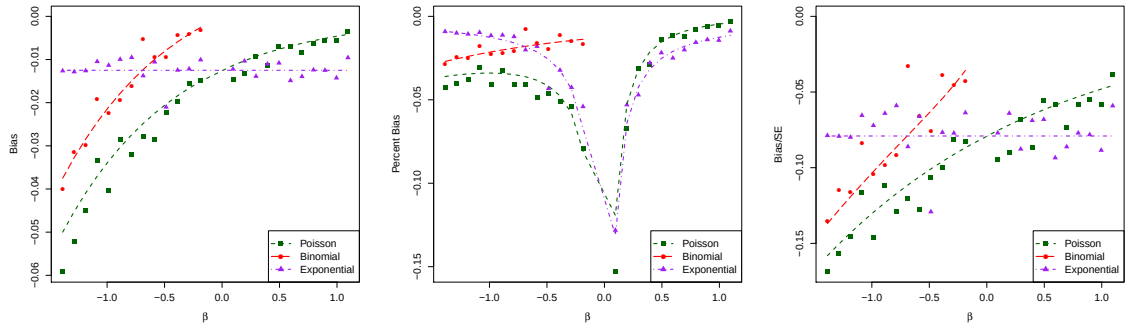
Simulations

Two scenarios are used to illustrate the behavior and accuracy of the first order bias approximation. The first scenario evaluates the impact β_0 has on bias for three probability distributions of outcomes, binomial, Poisson, and exponential, for a fixed sample size. The second scenario evaluates the impact sample size has on the first order bias approximation for binomial, Poisson, and exponential probability distributions. Each scenario used 5,000 replications.

Figure 2.1 (a) shows the theoretical first order bias approximation and simulated bias of β_0 for a range of values from $\log(.2)$ to $\log(3)$ ($\log(.2)$ to $\log(.9)$ for binomial outcomes) for a sample size of 30. As expected, for an exponential distribution, the bias of $\hat{\beta}_0$ does not depend on the parameter value. For both Poisson and binomial distributions, the magnitude of bias is lower for larger parameter values as seen in the first order approximation. The bias from the Poisson distribution is consistently more negative (larger magnitude) relative to the binomial distribution while the

bias of the exponential distribution relative to the Poisson and binomial distributions depends on the parameter value. The theoretical bias approximation accurately portrayed the simulated bias across all parameter values and distributions.

Other ways to quantify bias may also be used. Figure 2.1 (b) displays the percent bias and figure 2.1 (c) displays the bias/SE for theoretical first order approximations and simulated values. The magnitude of percent bias mirrors the trend of bias for a binomial distribution, with larger parameter values lead to a lower magnitude of percent bias. For both the Poisson and exponential distributions, the magnitude of percent bias increases as β_0 approaches 0. However, this increase is attributed to the decreasing denominator when calculating the percent bias rather than an increase in the bias. The bias/SE ratio follows the same trend as the bias. These figures indicate the relationship between β_0 and bias depends on the distribution of outcomes as well as the way in which bias is quantified.

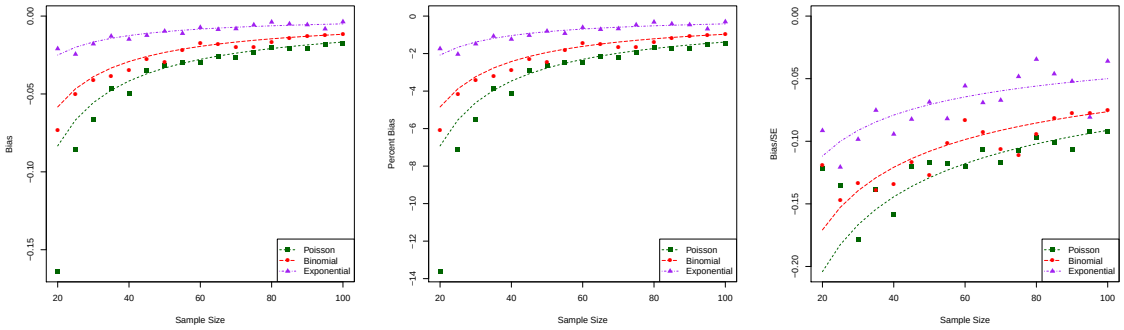


(a) (b) (c)

Figure 2.1: Bias (a), percent bias (b), and bias/SE (c) with respect to β for a single parameter model with $n = 30$ for Poisson, binomial, and exponential outcomes. Points represent simulated values and lines represent theoretical approximations.

Figure 2.2 shows how bias (a), percent bias (b), and bias/SE (c) all depend on the sample size with a fixed parameter ($\beta_0 = \log(.3)$). All three values decrease in magnitude as the sample size increases; however, bias/SE decreases at a smaller rate ($n^{-1/2}$) than bias and percent bias (n^{-1}) as expected since the SE is of order $n^{-1/2}$.

While the effect of sample size is similar regardless of the probability distribution, the sample



(a) (b) (c)

Figure 2.2: Bias (a), percent bias (b), and bias/SE (c) with respect to n for a single parameter model with $\beta = \log(0.3)$ for Poisson, binomial, and exponential outcomes. Points represent simulated values and lines represent theoretical approximations.

size needed to ensure the bias is below a given threshold varies greatly depending on the distribution of outcomes. For example, to ensure the magnitude of bias is less than 0.02, a sample size of 30 is sufficient for exponential outcomes; however, a sample size around 100 is needed for binomial outcomes. An even larger sample size is needed for a Poisson outcome.

In summary, the simulated and first order approximation of bias match up closely, indicating the effect of the higher order terms of the bias are minimal even for small sample sizes. In addition to sample size, the distribution and parameter value also influences the magnitude of bias. Furthermore, how bias is classified may also impact the perceived bias. While the effects of sample size are similar across distributions, the effect of the parameter on bias depends on the probability distribution. The method for measuring bias also affects the impact the parameter and distribution have on bias. A decrease in one measurement of bias does not necessarily equate to a decrease in all measures of bias as demonstrated for absolute bias and percent bias, as seen in figure 2.1. While the preferred method of measuring bias may differ based on the scenario, understanding the distinctions and the expected trends between them is important. In particular, caution may be needed in interpreting percent bias for parameter values particularly close to 0 to ensure the bias is meaningful and not an artifact of a small denominator.

2.3.1 Scenario 2: Single Predictor (Exploring the Effect of Predictor Distribution on Bias)

Now consider the setting of analyzing the association between a predictor of interest and the outcome. Let Y be an $n \times 1$ vector of independent outcomes with $\mathrm{E}[Y] = \mu$ and X be an $n \times 1$ vector of independent observations for the predictor of interest. Define $\log(\mu_i) = \beta_0 + \beta_1 X_i$ for the i^{th} observation. Then, using equation (2.3), the bias approximation for our parameter of interest, β_1, can be written as

$$b(\beta_1) = \frac{-\gamma_w}{2\sum w_i s_w} \tag{2.4}$$

where γ_w and s_w^2 are the sample weighted skewness and weighted variance of x with weights $w_i = \mu_i^2 V_i^{-1}(\mu_i)$ where μ_i is defined above and $V_i(\mu_i)$ is the variance of y_i which may depend on the mean and probability distribution (details in Appendix A.4). From equation (2.4), the relationship between weighted skewness, weighted variance, and bias is clear; larger weighted skewness and smaller weighted variance lead to a larger bias of $\hat{\beta}_1$.

In the special case where Y has an exponential probability distribution, $w_i = 1$ and the bias approximation further simplifies to

$$b(\beta_1) = \frac{\gamma}{2ns} \tag{2.5}$$

where γ is the skewness and s^2 is the variance of x and the bias does not depend on parameter values (as was previously seen in scenario 1).

The weights for Poisson distributed outcomes are $w_i = \exp(\beta_0 + \beta_1 x_i)$ while the weights for binomial distributed outcomes are $w_i = \frac{\exp(\beta_0+\beta_1 x_i)}{1-\exp(\beta_0+\beta_1 x_i)}$. While these can be used to calculate the first order bias, they do not allow for further simplification of equation (2.4) as was the case for outcomes with an exponential probability distribution.

In general, features with higher skewness or lower sampling variance will have a higher bias.

However, it is important to note that a feature with a higher skewness does not necessarily have a higher weighted skewness. Even if x has no skewness, the weighted skewness may still be nonzero, leading to bias. Similarly, the weighted skewness may be 0 even if x is skewed. This holds for the variance and weighted variance as well. While the distribution of the feature is a good indicator of the presence of bias (or lack thereof), it is not definitive. We provide an example, with a binary predictor, of how a non-skewed predictor can have a non-zero weighted skewness later in this section.

The relationship between the first order bias and the parameter values and skewness of the predictor for Poisson outcomes is displayed in figure 2.3. While the skewness of the predictor is displayed, in this setting higher skewness of the predictor leads to higher weighted skewness. This figure contains contour plots for 4 different types of predictors; high skewness/low variance (panel (a)), high skewness/high variance (panel (b)), no skewness/low variance (panel (c)), and no skewness/high variance (panel (d)). Note the scales are different between the plots, so caution is needed when comparing across scenarios. From the contour plots, the high skewness/low variance predictor has high bias for small β_1 values . In the high skewness/high variance and no skewness/low variance predictors, values of β_1 near -0.3 leads to the largest bias. The no skewness/high variance predictor has the highest bias when β_1 was near 0. These trends were consistent regardless of the values of β_0; however, larger β_0 led to a smaller bias of β_1 across all scenarios. The varying bias across parameter values demonstrates how the weighted skewness and variance can vary even if the predictors are fixed. From these contour plots, the high skewness/low variance predictor is most vulnerable to large bias, while the no skewness/high variance predictor has low bias relative to other predictors.

Binary Predictor

The approximations above can further be simplified in the special case of a binary predictor. Using the mean model, $\log(\mu_i) = \beta_0 + \beta_1 X_i$, let $X = 0$ for n_0 cases (group 0) and $X = 1$ for n_1 cases (group 1) where $n_0 + n_1 = n$ is the total sample size. In this scenario, the parameters may be defined by

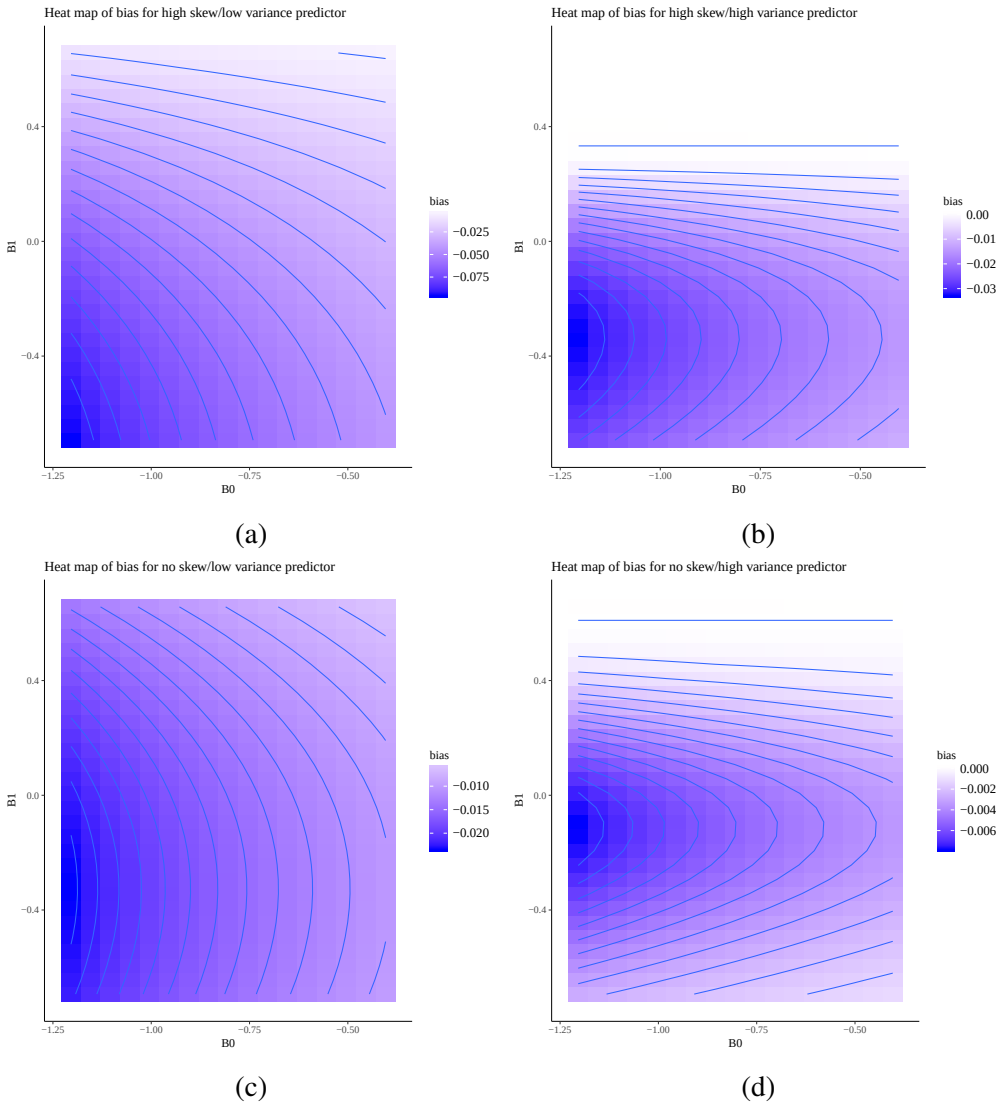


Figure 2.3: Contour plot for $b(\beta_1)$ for varying values of β_0 and β_1 for $n = 40$ from a 2-parameter log-Poisson model. Panel (a) depicts a high skewness/low variance predictor, panel (b) a high skewness/high variance predictor, panel (c) is no skewness/low variance predictor, and (d) is no skewness/high variance predictor.

$\beta_0 = \log(\mu_0)$ (the log mean for group 0) and $\beta_1 = \log(\mu_1) - \log(\mu_0)$ (the difference in log of the means between group 1 and group 0). Intuitively, the bias of β_1 should then be the difference in the bias of $\log(\mu_0)$ and $\log(\mu_1)$, that is $b(\beta_1) = b(\log(\mu_1)) - b(\log(\mu_0))$.

Analytical assessment of the bias approximation supports this intuition. See Appendix A.5 for detailed derivations. For exponential data,

$$b(\beta_1) = \frac{1}{2n_1} - \frac{1}{2n_0}$$

for binomial data,

$$b(\beta_1) = (\frac{1}{2n_1} - \frac{1}{2n_1 \exp(\beta_0 + \beta_1)}) - (\frac{1}{2n_0} - \frac{1}{2n_0 \exp(\beta_0)})$$

and for Poisson data,

$$b(\beta_1) = \frac{1}{2n_1 \exp(\beta_0 + \beta_1)} - \frac{1}{2n_0 \exp(\beta_0)},$$

all of which are the respective differences in biases of the log of the means for groups 1 and 0.

For constant weights, as in the case of exponential data, the weighted skewness and the bias are 0 when the sample sizes between the groups are equal. However, for balanced groups, the bias for Poisson and binomial data is only 0 when $\beta_1 = 0$. When $\beta_1 \neq 0$ the weighted skewness are non-zero for Poisson and binomial outcomes despite the predictor having no skewness. However, situations also exist where x is skewed but the weighted skewness (and thus bias) is 0. In order to minimize the weighted skewness for Poisson data, the group sample sizes must be proportional to their mean parameters, such that $n_1\mu_1 = n_0\mu_0$. For binomial data the group sample sizes must be proportional to their respective odds, $n_1 \frac{\mu_1}{1-\mu_1} = n_0 \frac{\mu_0}{1-\mu_0}$.

Contour plots are used to explore how bias is impacted by the group means (through parameter values) for settings where $n_0 = n_1$, $n_0 = 2n_1$, and $2n_0 = n_1$ all with a total sample size $n = 60$ for Poisson outcomes. Across all three settings, lower values of β_0 lead to a higher magnitude of bias.

The effect β_1 has on bias greatly depends on the sampling scheme. When $n_0 = n_1$ (a non-skewed predictor), bias is minimized when $\beta_1 = 0$, corresponding to the setting where both groups had the same mean. When $n_0 = 2n_1$, the bias is minimized when $\beta_1 = \log(2)$ and when $2n_0 = n_1$ the bias is minimized for $\beta_1 = \log(.5)$ as anticipated.

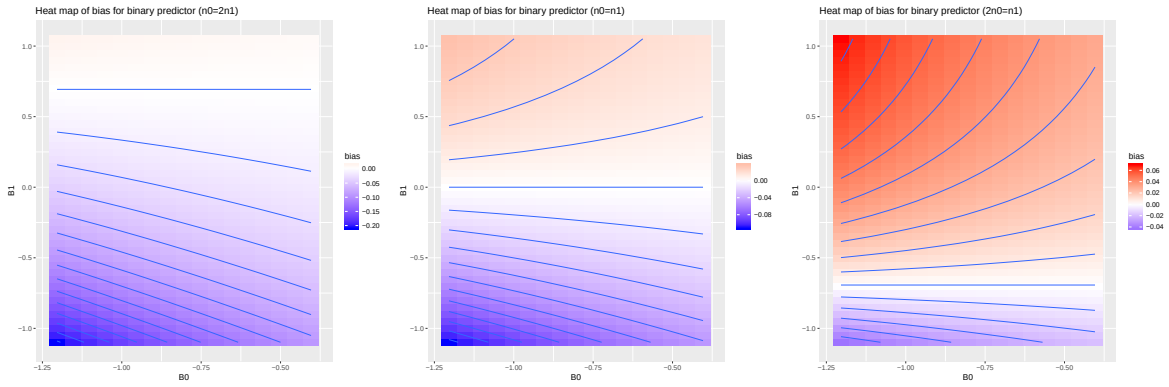


Figure 2.4: Contour plot for $b(\beta_1)$ for varying values of β_0 and β_1 for $n = 40$ from a 2-parameter log-Poisson model with a binary predictor of interest. Panel (a) depicts the setting $n_0 = 2n_1$, panel (b) $n_0 = n_1$, and panel (c) $n_0 = \frac{n_1}{2}$.

Simulations

Continuous Predictor

We now explore the accuracy of the first order approximation empirically for a continuous predictor. In these scenarios, x is generated from a gamma distribution centered at 0 which enables us to easily control both the skewness and variance of the predictor. Four different distributions of X were used, representing the scenarios used in the contour plots: high skewness/low variance, high skewness/high variance, low skewness/low variance, and low skewness/high variance. The variance of the predictor is 1 in the low variance scenario and 9 in the high variance scenario. The skewness of the predictor was 2 in the high skewness scenarios and $2/3$ in the low skewness scenarios. Parameter values of $\beta_0 = \log(.5)$ and $\beta_1 = \log(.9)$ were chosen to represent a scenario with moderate bias based off the contour plots (approximately the center of the plots).

Each of these scenarios is explored for exponential, Poisson, and binomial outcomes for sample

sizes ranging from 30 to 100. Both bias and bias/SE were reported (percent bias was not reported as it is a scaled version of bias for fixed parameter values).

Results are shown in figure 2.5. Panels (a)-(c) demonstrate the relationship between the skewness and variance of the predictor and the bias of β_1. The high skewness/low variance predictor leads to the greatest bias across all distributions while the low skewness/high variance predictor has the smallest bias. The bias of high skewness/high variance and low skewness/low variance predictors is similar for Binomial and Poisson outcomes and equivalent for exponential outcomes in this setting .

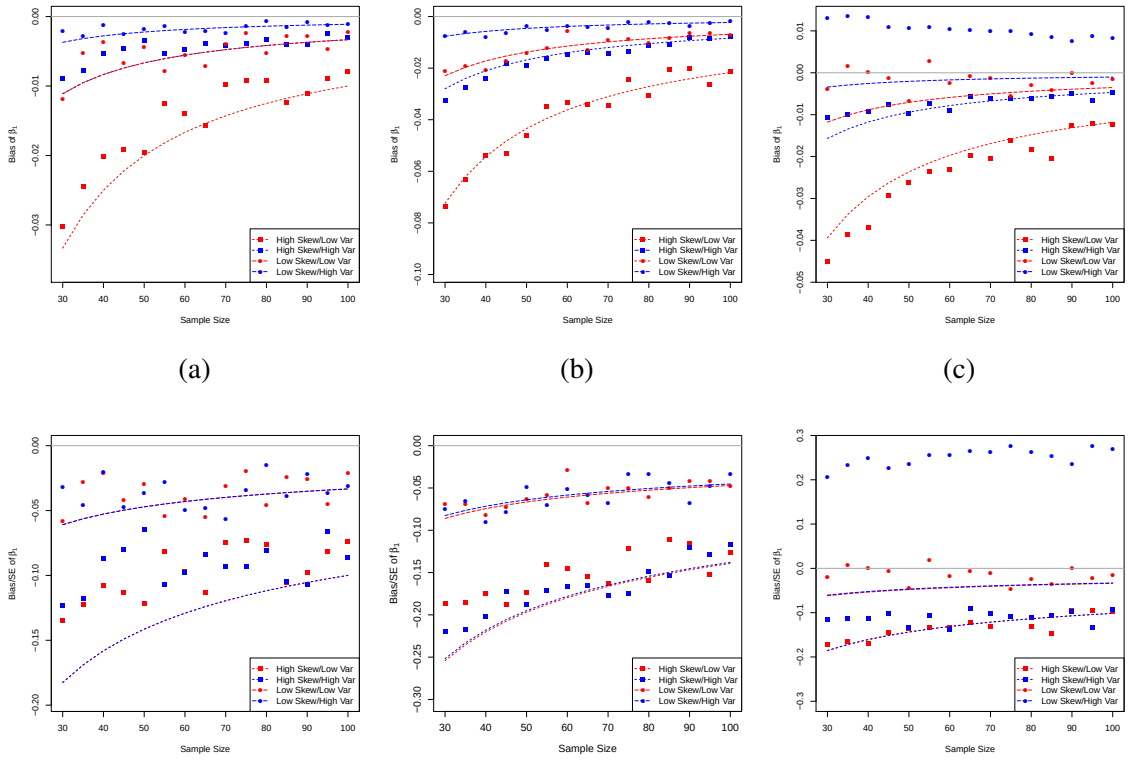


Figure 2.5: Bias (a)-(c) and bias/SE (d)-(f) for exponential, Poisson, and binomial outcomes respectively from a 2-parameter model with log-link function and $\beta_0 = \log(0.7)$ and $\beta_1 = \log(0.9)$ for varying sample sizes and distributions of the predictor of interest.

Across all distributions, greater sample sizes are needed to achieve the same level of bias when the weighted skewness is high and the weighted variance is low than when the weighted skewness is low and the weighted variance high. A sample size of approximately 70 is needed for a high

skewness/high variance predictor to achieve the same bias from a low skewness/low variance predictor with a sample size of 30. For a low skewness/low variance, a sample three times as large is needed to obtain the same bias as a low skewness/high variance predictor. This indicates the distribution of the predictor is a key component in the presence of bias.

In general, the bias approximations are accurate, with the exception of the bias for the low skewness/low variance predictor with a Binomial outcome. However, the fitting algorithms have poor rates of convergence ($> 25\%$ using standard R software) making the simulated bias unreliable for this situation. Skewness is more influential than the variance of the predictor for the bias/SE in all the models as demonstrated in figure 2.5 panels (d)-(f).

The bias/SE ratio is constant for a given skewness as the SE cancels with the weighted variance in the expression for the bias approximation. In general the approximations for bias/SE for low skewness predictors are accurate as well; however, the approximation tends to exaggerate the bias/SE for skewed predictors, due to the underestimation of the SE when using the expected Fisher's information in these settings. The simulated bias/SE for Bernoulli outcomes is inaccurate as well for the low skewness/low variance predictor due to issues with convergence.

Binary Predictor

Simulations are used to assess the accuracy of the theoretical approximations. Based on the contour plots, values of $\beta_0 = \log(.3)$ and $\beta_1 = 2$ are used to represent moderate levels of bias. We explore settings where the group sample sizes are balanced ($n_0 = n_1$), exploring sample sizes in each group from 15 to 50 as well as imbalanced sample sizes where $n_1 = 20$ but n_0 is allowed to vary from 10 to 80 (in both settings the total sample size ranges from 30 to 100).

Figure 2.6 (a) shows the bias for balanced sampling for exponential, Poisson, and binomial distributions. The first order bias for Poisson and binomial distributions are equivalent (an artifact of the equal sample sizes between the two groups) and decreases as the sample size increases. For exponential outcomes, the first order bias is 0 as expected. Figure 2.6 (b) shows the bias/SE. Due to the lower standard error from a binomial outcome than Poisson outcomes, the bias/SE is higher

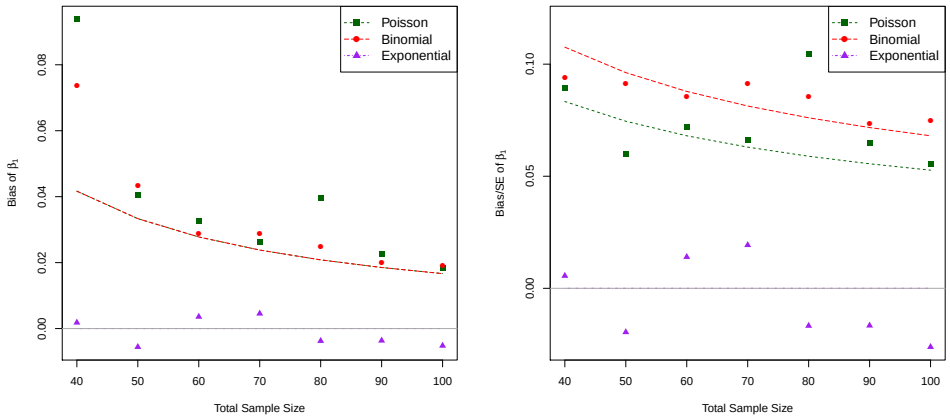


Figure 2.6: Bias (a) and bias/SE (b) for a 2-parameter model with log-link function with a binary predictor of interest and a balanced sample ($n_0 = n_1 = \frac{n}{2}$) where $\beta_0 = \log(0.3)$ and $\beta_1 = \log(2)$ for exponential, Poisson, and binomial outcomes. Points represent the simulated values and lines represent the theoretical bias approximation.

for binomial outcomes. Both bias/SE decrease as the sample size increases. For exponential distributions, bias/SE is 0 for all sample sizes. Figure 2.7 (a) shows the bias for unbalanced sampling. For exponential outcomes, the first order bias is 0 when $n_0 = 20$ (which corresponds to balanced sampling) and increases in magnitude as n_0 increases. The first order bias for Poisson and binomial outcomes is monotonic decreasing; however, the magnitude of bias is decreasing as n_0 approaches the optimal sample size, as the weighted skewness is reduced, after which, further increases of n_0 increase the magnitude of bias but with the opposite sign. The bias from Poisson and binomial outcomes is more heavily affected by n_0 than the bias from exponential outcomes. The trends for bias/SE are similar (panel (b)).

2.3.2 Scenario 3: Multiple Predictors (Exploration of Effects of Correlated and Uncorrelated Predictors on Bias)

Now let X be an $n \times p$ design matrix such that $p > 2$ and β^* be a p-length vector. Then the mean model for an n-length vector of independent outcomes Y is $\log(\mathrm{E}[Y]) = X\beta^*$. Succinctly writing expressions for bias as was done in the previous scenarios is no longer feasible. Rather

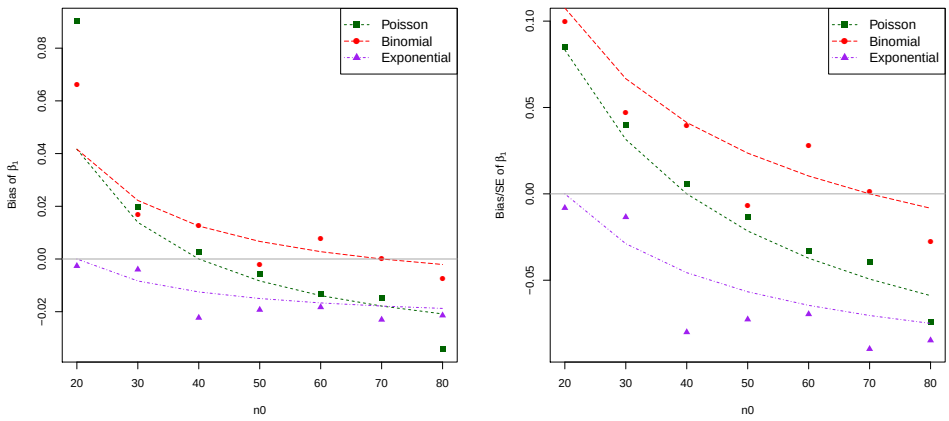


Figure 2.7: Bias (a) and bias/SE (b) for a 2-parameter model with log-link function with a binary predictor of interest and an unbalanced sample ($n_1 = 20$, $n_0 = n - 20$) where $\beta_0 = \log(0.3)$ and $\beta_1 = \log(2)$ for exponential, Poisson, and binomial outcomes. Points represent the simulated values and lines represent the theoretical bias approximation.

than analyzing the bias as a single quantity, we can explore the individual components I, J, and K to gain insight into what factors contribute to bias.

The component I is the most straightforward as it is the standard Fisher's information. Instead of discussing I directly, we consider I^{-1} which, in addition to being a component of bias, is the covariance of $\hat{\beta}$. Two components that are known to contribute to variance are sample size and correlation among features. In general, larger sample sizes and smaller correlations lead to smaller variance estimates, which may in turn lead to a smaller bias of parameter estimates.

The components J and K can be viewed together as a weighted three-dimensional sample covariance array. When using a log link function, $J_{r,st} + \frac{K_{r,s,t}}{2} = \frac{1}{2}\sum X_{i,r}X_{i,s}X_{i,t}w_i$ for weights w_i where i denotes the observations and r, s, and t denote the features. This expression includes the weighted, third non-central moment of individual features ($\sum X_{i,r}^3 w_i$) on the main diagonal. This is a measure of weighted skewness analogous to the finding in the previous scenario. On the off diagonal, the expression includes the three-way weighted, non-central sample covariance between three features ($\sum X_{i,r}X_{i,s}X_{i,t}w_i$ for $r \neq s \neq t$) and the two-way weighted expression of non-central sample covariance between a feature and the square of another feature ($\sum X_{i,r}^2 X_{i,s}w_i$ for $r \neq s$).

In the special case where the off-diagonals are all 0, the bias for each parameter is the same as the bias from a single feature model for each feature. However, it is important to note that even when features are uncorrelated, the off-diagonals may still be non-zero. In general, $\sum X_{i,r}X_{i,s}w_i = 0$ does not imply $\sum X_{i,r}^2 X_{i,s}w_i = 0$. If the square of a feature is correlated with other features, additional bias may be present even when features are uncorrelated.

In general, understanding the cumulative effects these moments have on the bias is difficult; however, estimates may be prone to higher bias when these moments are large.

Simulations

We simulated data for two cases using an $n \times 3$ design matrix $X = (x_0, x_1, x_2)$. The first feature, x_0 is a column of 1's representing the intercept. In the first case, the other two features, x_1 and x_2 are correlated with each other but not with x_0, each with a marginal centered gamma probability distribution with shape and scale parameters both 1. These features were generated using a multivariate normal copula with correlation coefficients ranging from -0.9 to 0.9. In the second case, the second feature, x_1 is normally distributed with mean 1 and variance 0.5. The third feature, x_2, is normally distributed with mean αx_1^2 and variance 0.5 for values of α ranging from -0.5 to 0.5.

Using these design matrices we generate independent outcomes, Y, such that $\log(\mathrm{E}[Y|X]) = X\beta$ for $\beta = (\log(.5), \log(.9), \log(.9))$. In these cases we assume that the first feature is the predictor of interest which is the focus of assessing the bias while the second feature is an adjustment variable; however, similar trends hold when x_3 is the predictor of interest.

Figure 2.8 shows the results from the first case when the two features are positively skewed. The magnitude of the bias increases when the features are negatively correlated. However, the magnitude of bias is slightly smaller when the features are positively correlated. Similar trends are seen for the bias/SE; however, the difference between uncorrelated and positively correlated features is greater, likely due to SE estimates being larger when strong correlation is present. These trends held across Poisson, binomial, and exponential distributions of the outcomes. In the case of Poisson and exponential probability distributions, the theoretical first order bias approximation

provides a good estimate of the true bias. This was not true for the binomial probability distribution, possibly due to non-convergence of fitting algorithms (non-convergence rate $> 30\%$). When the features were independent, the bias for a given feature was similar to the bias of the feature when it was the only predictor in the model. This demonstrates that the bias for a given feature does not substantially change when independent features are added to the model.

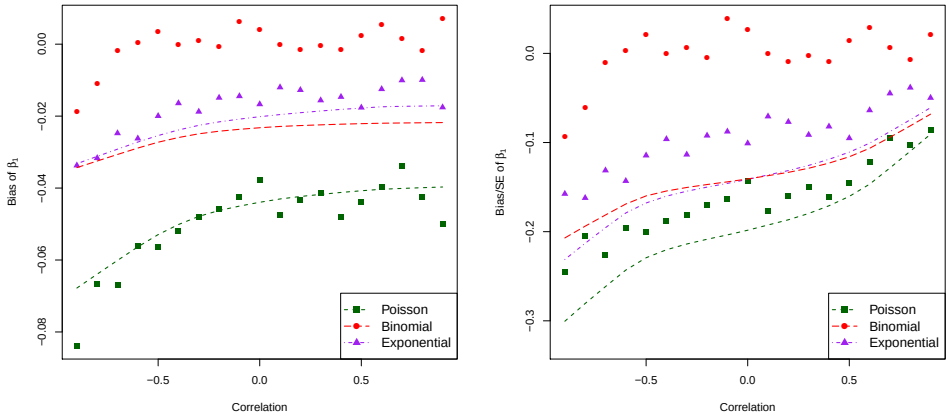


Figure 2.8: Bias (a) and bias/SE (b) for a 3-parameter model with 2 high skewness/low variance features with a correlation ranging from -0.9 to 0.9. for exponential, Poisson, and binomial outcomes. Points represent the simulated values and lines represent the theoretical bias approximation.

The results for the second case are found in figure 2.9. Regardless of the distribution, more extreme values of α led to larger bias. In this scenario, positive and negative α values had a similar effect on bias, as seen by the symmetry in the graphs. Poisson data was the most susceptible and exponential data was the least susceptible to bias for extreme α values. Similar results held for bias/SE. In general, the theoretical first order bias approximations were close to the simulated values with the exception of binomial outcomes which had a high frequency of non-convergence in the fitting algorithm (as in the first case).

In the first scenario, positive correlation was protective while negative correlation was detrimental. In the second scenario, both positive and negative correlation led to a larger bias. This exemplifies the complex relationship between the bias and the joint distribution of the features.

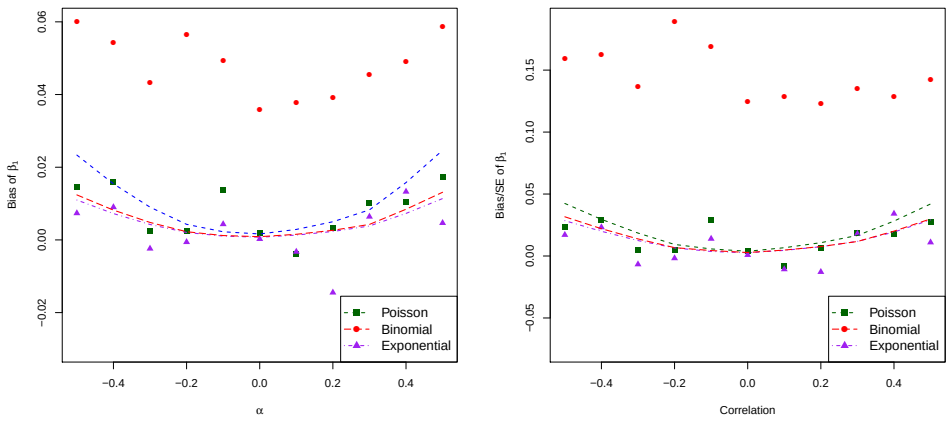


Figure 2.9: Bias (a) and bias/SE (b) for a model with an intercept, a normally distributed predictor, x_2, with mean 0 and a normally distributed predictor, x_3, with mean αx_2^2 for α ranging from -0.5 to 0.5.

2.4 Discussion

These results demonstrate the distribution of outcomes, the joint distribution of predictors, parameter values, and sample size all contribute to bias. In addition to the probability distribution directly impacting bias, the probability distribution also impacts how other model components contribute to bias. For Poisson and Binomial likelihoods, the bias is impacted by the true parameter values; however, for an exponential likelihood, the bias is not impacted by the parameter values for a log link function. Furthermore, higher skewness and lower variance of the predictor (or their weighted counterparts) directly translate to larger bias. In many cases, weighted skewness and weighted variance are correlated with the skewness and variance of the predictor. As seen in the case for a binary predictor, this is not always the case. Similarly, in general, an increase in sample size led to a reduction in bias; however, settings exist where an increase in the total sample size alone is not sufficient to reduce bias.

The relationship between predictors in a model also affects bias. When predictors are independent, the effect of additional features is minimal. In cases where predictors are dependent, the relationship is less clear. In one scenario, positive correlation between predictors was found to

decrease bias while negative correlation increased bias. However, in another scenario, bias was higher for both negative and positive correlations. Thus, caution is needed when using correlated predictors as the effect on bias is not immediately clear. Further characterization of the bias for a wider range of models is needed to confirm if the trends found between the distribution of the predictor and bias with a log-link function are generalizable to other models.

While bias is not explored for other link functions or likelihoods, the trends are anticipated to be similar. In particular, any model that satisfies $I_{jk} = x_j x_k w(X\beta)$ and $J_{jkl} + \frac{K_{jkl}}{2} = x_j x_k x_l w(x\beta)$ for some weights $w(x\beta)$ will produce identical findings to the results above with the exception of different weights being used. The logit and inverse link are two common links that could be explored in addition to the Gaussian likelihood. Further simulations in these settings could confirm these trends are generalizable beyond the settings explored here.

This characterization has assumed the likelihood is correctly specified. However, in many settings the working distribution may not be the same as the true likelihood of the outcomes. Quasi-likelihood methods can be used to obtain asymptotically consistent estimates of the parameters. Finite sample bias is still present for these methods; however, the bias approximation from a Taylor expansion relies on the correctly specified likelihood and is therefore not valid in this setting. Further work to characterize the bias for quasi-MLE is needed.

These findings stress the importance of bias reduction methods for small sample studies. Even for moderate sample sizes, the bias may be noticeable, particularly if the skewness of the predictor is high or the variance of the predictor is low. While many bias reduction methods exist, a better understanding of the reduction of bias they provide in application for small sample studies is needed, particularly when the assumed model may be misspecified [21, 25, 71].

In the next chapter, we discuss bias reduction methods and their limitations, with a particular focus on when the model is misspecified and propose an improved bias reduction method that is robust to model misspecification.

Chapter 3

Robust Bias Correction for Misspecified Models

3.1 Introduction

In the previous chapter we explored the impact the distribution of the outcome and predictors has on bias for a correctly specified model. In practice, the assumption that the model is correctly specified is rarely true. The true mean model and probability distribution of the outcomes are often unknown. In this chapter we discuss bias reduction methods for estimates from misspecified models.

In Section 3.3, we present the concept of misspecified models in a GLM framework and properties of estimates derived from these models as well as discuss currently available bias reduction methods in the context of both correctly specified and misspecified models. In particular we draw attention to the scenario where separation occurs. Separation occurs when the outcomes can be separated by a linear combination of the predictors. This can lead to perfect prediction of the outcome and results in the likelihood being maximized at infinity [2]. While preventative bias reduction methods can give finite solutions under separation, these methods are only applicable to correctly specified models [34]. No known bias reduction method exists when separation and

misspecification occur simultaneously. In Section 3.4 we develop a preventative robust bias reduction method for which the solution is a bias reduced estimate. This method uses a modified score function that utilizes a robust first order bias approximation. This method provides accurate bias reduction in the presence of separation and misspecification. Simulations are used to assess our new method against other bias reduction methods empirically in Section 3.5. Lastly, in Section 3.6, we explore extensions to the GEE framework in the presence of clustered data.

3.2 Motivation

Before engaging with the theory of misspecified models, we use a motivating example to demonstrate the challenges of bias reduction when the model is misspecified. Details for calculating the bias reduced estimates presented in this section will be discussed in more detail later in the chapter.

Motivating Example: Estimating Log-risk

Suppose we want to estimate the log-risk of a disease in a population. Let Y be a sample of independent binary outcomes with a mean model only dependent on a single parameter, β_0 such that $\log(\mathrm{E}[Y]) = \beta_0$. Then the score function for a correctly specified, log-binomial model is

$$U_n = \frac{1}{n}\sum_{i}^{n} \frac{y_i - \exp(\beta)}{1 - \exp(\beta)}$$

with a corresponding point estimate, $\hat{\beta} = \log(\bar{y})$ where $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$. Using the results from Chapter 2, the parametric first order bias approximation is $\frac{1}{2n} - \frac{1}{2n(\exp(\beta_0)}$. This approximation, when used with a bias reduction method, can be used to derive the naive bias reduction (NBR) estimate

$$\tilde{\beta} = \log\left(\frac{\bar{y} + \frac{1}{2n}}{\frac{2n+1}{2n}}\right). \tag{3.1}$$

While a closed form expression exists for an intercept only model, in more complex settings, an iterative algorithm is needed for parameter estimates. In some settings, fitting algorithms may

not converge to parameter estimates when using the correctly specified binomial probability distribution. Instead a misspecified Poisson working distribution is frequently used to ensure fitting algorithms converge to estimates.

Suppose we instead use a Poisson regression to estimate β_0. Then the corresponding score function is

$$U_n(\beta) = \frac{1}{n}\sum y_i - \exp(\beta) \tag{3.2}$$

while the point estimate remains the same, $\hat{\beta} = \log(\bar{y})$. Despite having identical point estimates, the parametric bias approximation is $-\frac{1}{2n}\frac{1}{\exp(\beta_0)}$ which differs from that of the correctly specified model. Furthermore, the NBR parameter estimate is

$$\tilde{\beta} = \log(\bar{y} - \frac{1}{2n}). \tag{3.3}$$

This is referred to as a misspecified NBR estimate as it incorrectly assumes the model is correctly specified. In this setting, we see the misspecified NBR estimate is too large and can even lead to an increase in the magnitude of bias compared to the standard MLE when $\exp(\beta_0) > \frac{1}{2}$.

While parameter estimates are asymptotically consistent under misspecification, this demonstrates that accurate parametric bias reduction methods rely on correct model assumptions to provide accurate bias reduction.

Non-parametric bias reduction methods exist (such as jackknife and bootstrap methods) that do not rely on a correctly specified model; however, these methods rely on a finite MLE which may not always exist. Before discussing these bias reduction methods in more depth, we outline what a misspecified model is as well as properties of estimates derived from misspecified models.

3.3 Background

In this section, the estimation procedure and properties of estimates from misspecified GLM are defined before discussing current bias reduction methods.

3.3.1 Misspecification of Generalized Linear Models

For correctly specified models, a likelihood is obtained using a mean model and probability distribution of the outcomes. This likelihood is maximized to obtain the maximum likelihood estimate (MLE) which is asymptotically consistent, efficient, and normally distributed. However, in many settings, the model may be misspecified. This misspecification comes in two forms.

The first is a misspecification of the mean model. This occurs when the working linear combination of covariates does not align with the true deterministic relationship between (possibly unknown) covariates and the outcome.

The second type of misspecification is of the working distribution. In some settings, the true distribution of the outcomes may be known (such as the case of binary outcomes). But in many settings the distribution of outcomes is unknown, either due to over-dispersion, where the form of the distribution is known but the variance is off by a multiplicative factor or because the form of the distribution is unknown. Furthermore, even when the true distribution is known, issues with fitting algorithms converging to parameter estimates may lead to the intentional misspecification of the working distribution (such as the case of Poisson regression when estimating the relative risk from binomial data.)

In misspecified settings the true mean model and probability distribution are replaced by a working mean model and working probability distribution. Together these form a quasi-likelihood. This quasi-likelihood can be maximized with respect to the parameters to obtain quasi-maximum likelihood estimates (QMLE). In many settings the QMLE is consistent for the target parameter that minimizes the Kullback-Leibler distance, which can be thought of as minimizing the ignorance of the working model relative to the unknown true model [1]. The QMLE is also asymptotically

consistent and normally distributed under some regularity conditions [78]. Standard error estimates based on the working model are no longer consistent; however, consistent standard error estimates can be obtained using robust standard error estimation techniques [78].

The theory of estimating equations can be used to provide a more formal discussion of the asymptotic properties of QMLE. We first discuss general results from estimating equations before demonstrating how these results can be applied to (quasi-) maximum likelihood estimation in the GLM framework.

3.3.2 Estimating Equations

Estimating equations provide a general framework for calculating parameter estimates from a variety of models. The general concept is to solve a set of equations to obtain parameter estimates. Estimating equations have the form

$$\mathbf{U}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{U}_i(\beta) = 0, \tag{3.4}$$

which is the sum of n independent estimating equations. As a closed form solution often does not exist, iterative methods are frequently used to solve the system of equations to obtain the parameter estimate

$$\hat{\beta} : \mathbf{U}_n(\hat{\beta}) = \mathbf{0}. \tag{3.5}$$

Under suitable conditions, $\hat{\beta}$ is asymptotically consistent and normally distributed for the population parameter, β^0 which is the solution to the equation when setting the expected value of the estimating equation equal to 0:

$$\mathrm{E}[\mathbf{U}_n(\beta^0)] = 0.$$

Huber identified conditions necessary for consistency and asymptotic normality of $\hat{\beta}$ when

the observations are independent and identically distributed. Inagaki generalized these findings to settings in which observations are independent but not necessarily identically distributed [39]. Juan and Yennrich provide a more general set of assumptions that ensures the existence, consistency, and asymptotic normality of $\hat{\theta}$ [81]. The following are assumptions listed by Juan and Yennrich.

Assumptions:

A1 $\mathbf{U}_n(\beta^0) \to 0$ with probability one

A2 These exists a neighborhood of β^0 such that, with probability one, all $\mathbf{U}_n(\beta)$ are continuously differentiable and $\dot{U}_n(\beta) = \frac{\partial \mathbf{U}_n(\beta)}{\partial \beta}$ converges uniformly to a nonsingular, non-stochastic limit at β^0.

A3 $\mathrm{E}[\mathbf{U}_n(\beta)]$ has a unique zero at β^0

A4 $\sqrt{n}\mathbf{U}_n(\boldsymbol{\beta^0}) \to N(0,V)$ where $\mathbf{V} = \mathrm{E}[\mathbf{U}(\beta^0)^T\mathbf{U}(\boldsymbol{\beta^0})]$

where $\mathbf{U}_n(\beta)^T$ denotes the transpose of $\mathbf{U}_n(\beta)$.

With these assumptions, Juan and Yennrich derived the following theorems proving the consistency and asymptotic normality of $\hat{\beta}$.

Theorem 1 (Consistency of $\hat{\beta}$)**.** *If assumptions A1-A3 are satisfied then there exist estimates* $\hat{\beta}$ *satisfying* $\mathbf{U}_n(\hat{\beta}) = 0$ *such that* $\hat{\beta} \to \beta^0$

Theorem 2 (Asymptotic Normality of $\hat{\beta}$)**.** *If* $\hat{\beta} \to_p \beta^0$ *and assumptions A1-A4 are satisfied, then* $\sqrt{n}(\hat{\beta} - \beta) \to N(0, \mathbf{V})$ *where* $\mathbf{V} = \mathrm{E}[\dot{\mathbf{U}}(\beta_0)]^{-1}\mathrm{E}[\mathbf{U}(\beta^0)\mathbf{U}(\beta^0)^T]\mathrm{E}[\dot{\mathbf{U}}(\beta_0)]^{-1}$

We note that no assumptions about the mean model or probability distribution being correctly specified were needed to obtain the asymptotic properties, making this theory appealing for misspecified models. We note that the target of inference for $\hat{\beta}$ is the population-level solution to the estimating equations.

GLM as Estimating Equations

We now break down the process of (quasi-)maximum likelihood estimation to understand how it fits into the framework of estimating equations. Suppose for a given vector of independent outcomes, Y and design matrix X with parameters β we have a (quasi-)likelihood $L(y;X\beta)$. The (Q)MLE is defined as

$$\hat{\beta} = \arg\max_{\beta} \prod_{i=1}^{n} L(y_i; X_i\beta). \tag{3.6}$$

This is equivalent to maximizing the (quasi-)log likelihood $l_n(\beta) = \log(L_n(y, X\beta))$ which is the solution to $\frac{\partial l_n(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{1}{n} \frac{\partial l(y_i;, X_i\beta)}{\partial \beta} = 0$. Thus solving an estimating equation where

$$U(\beta) = \frac{\partial l_n(\beta)}{\partial \beta} = 0 \tag{3.7}$$

is equivalent to (quasi-)maximum likelihood estimation.

Using the results from estimating equation theory, the QMLE is asymptotically consistent for the population parameter, β^* that is the solution to the expected score equation:

$$\mathrm{E}[U_n(\hat{\beta}^*)] = 0. \tag{3.8}$$

Furthermore, the QMLE is asymptotically normal and robust standard error estimation can be used to calculate consistent standard error estimates. Furthermore, in the case of a correctly specified GLM, the resulting estimating equation has optimality properties as well [30].

While the QMLE has desirable asymptotic properties, finite sample bias is still present. Ideally, bias reduction methods can be extended from MLE to QMLE. While this is straightforward for some of the bias reduction methods, other methods are not suitable to be used when misspecification is present. In the next section we discuss bias reduction methods in more detail.

3.3.3 Current Bias Reduction Methods

Bias reduction methods come in two forms: corrective and preventative. Corrective bias reduction involves a post-estimation correction to the MLE. Preventative bias reduction involves solving a modified score equation for which the solution is a biased reduced estimate and no further correction is needed. While neither of these methods directly rely on model assumptions, they do rely on an accurate estimate of the bias. The bias can be estimated non-parametrically using re-sampling methods or a parametric first order approximation can be used, as was derived in Chapter 2. We first discuss the distinction between corrective and preventative bias reduction methods in more depth before discussing methods of bias estimation.

Corrective Bias Reduction

The implementation of corrective bias reduction is a straightforward two-step process. The first step is to calculate $\hat{\beta}$. The second step is to subtract an estimate of the first order bias, $b(\hat{\beta})$, giving a bias reduced estimate $\tilde{\beta} = \hat{\beta} - b(\hat{\beta})$. The bias can be estimated parametrically, using the first order approximation derived in Chapter 2, as well as non-parametrically using a re-sampling method such as the jackknife or bootstrap. However, this method relies on a finite MLE as an initial estimate (as well as a finite MLE in all subsamples when using a non-parametric method) which may not exist when separation occurs in the data.

Preventative Bias Reduction

In contrast, the preventative bias reduction method does not rely on the MLE to be finite. Instead, a modified score function is solved to directly calculate a bias reduced estimate. Firth demonstrated that for a correctly specified model with a parametric approximation of the bias, the solution to the modified score equation $U^*(\beta) = U(\beta) - Ib(\beta) = 0$, has no first order bias [25]. We refer to this method as naive bias reduction (NBR) as the bias approximation used assumes the model is correctly specified. The solution to $U^*(\beta) = 0$ is finite even in the presence of separation. Of

note, this modified score function has connections with Bayesian inference. Under a canonical link function, the modification is Jeffrey's invariant prior and can be used to reduce the bias of the posterior mode [25]. However, this method uses a parametric expression for $b(\beta)$. While a non-parametric estimate could be used in theory, since the score equation is solved iteratively, the method would need to be applied for every iteration which makes the implementation impractical. This method effectively avoids issues that occur due to separation but has the restriction of using a parametric expression for bias.

Methods for Estimating Bias

While the frameworks for corrective and preventative methods do not rely on correct model specification, some of the bias estimates used in these methods do. Analyzing parametric and non-parametric methods for bias estimation allows us to better understand what types of bias reduction methods are robust to misspecification and which ones are susceptible to mis-correction of the bias.

Parametric Bias Estimation

In Chapter 2, we derived the following parametric expression for the first order bias approximation. We use the bias approximation for a single parameter to simply notation. The first order bias approximation is

$$b(\beta) = I^{-1}I^{-1}(J + \frac{1}{2}K)$$

where $I = \mathrm{E}[\dot{U}(\beta^*)]$, $K = \mathrm{E}[\ddot{U}(\beta^*)]$, and $J = \mathrm{E}[\dot{U}(\beta^*)U(\beta^*)]$. The NBR method explicitly uses the working model to derive model based values I_m, J_m and K_m which take the expectation under the working model. If the model is correctly specified $I_m = I$, $J_m = J$, and $K_m = K$ and the working model leads to the exact first order bias approximation.

If the working mean model is misspecified then in general $I_m \neq I$, $J_m \neq J$, and $K_m \neq K$ and the first order bias approximation based on the working model will be misspecified as well. Fur-

thermore, while not obvious from the expression for bias above, in the derivation in Chapter 2, it was assumed that $\mathrm{E}[\dot{U}(\beta)] = \mathrm{E}[U(\beta)^2]$. However, this equality does not hold under misspecification. Thus in the presence of misspecification, methods that rely on the parametric first order approximation are not valid.

Non-parametric Bias Estimation

Non-parametric methods such as jackknife and bootstrap rely on resampling to estimate the bias. As these methods do not rely on parametric assumptions to estimate the bias, they are valid under model misspecification. However, this method is susceptible to separation in the data, particularly for binary outcomes. Even if the MLE is finite, the subsample MLE may not be. Because of this, non-parametric bias estimation is vulnerable to separation in the data (or sub-samples in the data).

Bias Reduction for Misspecified Models with Separation

Table 3.1 has a summary of valid bias reduction methods under different model assumptions. In conclusion, non-parametric methods can provide reliable bias reduction under model misspecification but not in the presence of separation. In contrast, parametric bias approximations can be used even in the presence of separation (when used in tandem with the preventative bias reduction method) but relies on a correctly specified parametric model [35]. These limitations to current bias reduction methods are problematic as no known method exists for misspecified models in the presence of separation. We develop a bias approximation that is robust to model misspecification that can be used in a preventative bias reduction framework to create a robust bias reduction method. This robust bias reduction method can be used for misspecified models and in the presence of separation in the data.

Table 3.1: Summary of validity of bias reduction methods under different model scenarios

	No Separation		Separation	
	Correctly Specified	Misspecified	Correctly Specified	Misspecified
Parametric Corrective	Yes	No	No	No
Non-parametric Corrective	Yes	Yes	No	No
Parametric Preventative	Yes	No	Yes	No
Robust Bias Reduction	**Yes**	**Yes**	**Yes**	**Yes**

3.4 Robust Bias Reduction

There are clear shortcomings in current bias reduction methods, including the inability to handle separation in the data or misspecification of the working model. In particular, no bias reduction method is available when the model is misspecified and separation occurs in the data. For this reason, a more robust preventative bias reduction method is needed to help address these shortcomings. A preventative method that uses a robust first order bias approximation would fill this void in the literature, providing a method to reduce bias across settings with misspecification and separation in the data.

In the next section we develop a more general form of the first order bias approximation presented in Chapter 2. This robust first order bias approximation does not depend on a correctly specified model. We then demonstrate that using this robust bias approximation in a preventative bias reduction framework allows for accurate bias reduction even under model misspecification.

3.4.1 Robust Bias Approximation

We begin by developing a robust bias approximation using a Taylor expansion but not assuming the model is correctly specified. This approximation consists of components that rely on the specified model as well as components that rely on the true distribution of the data. In settings where the working model is identical to the true model, the approximation simplifies to that found in Cox and Snell and used by Firth [17, 25]. We will demonstrate that in settings where the true model is known but a misspecified working model is used, the bias approximation can be explicitly derived.

Lastly, when the true distribution is unknown, we can estimate the first order bias approximation empirically.

In order to obtain a finite approximation of the first order bias, assumptions on the estimating equation and parameter estimates are needed. This research focuses on finite sample modification to estimating equations that are well-behaved asymptotically (i.e. assumptions A1-A4 are met and parameter estimates from the unmodified estimating equations are asymptotically consistent and normally distributed). Further assumptions are needed on the derivatives of the estimating equations to ensure the first order bias approximation is well-defined.

Assumptions:

B1 For some δ-neighborhood around β^*, each component of $\ddot{\mathbf{U}}_n(\beta)$ is Lipschitz. That is, for all β such that $|\beta - \beta^*| < \delta$, each component of $\ddot{\mathbf{U}}_n(\beta)$ satisfies

$$|\ddot{U}_n(\beta) - \ddot{U}_n(\beta^*)| \leqslant R|\beta - \beta^*| \text{ for some constant } R \geqslant 0.$$

B2 $\mathrm{E}[\dot{\mathbf{U}}_n(\beta^*)]$ exists and is nonsingular, and $\mathrm{E}[\mathbf{U}_n(\beta^*)\dot{\mathbf{U}}_n(\beta^*)]$ exists and each component is finite

B3 $\hat{\beta} - \beta^* = O_p(n^{-1/2})$

B4 $\mathrm{E}[\mathbf{U}_n(\beta^*)] = 0$.

Assumption B1 and B3 are used to demonstrate the higher order remainder of the Taylor expansion is sufficiently small; B2 ensures the bias approximation is finite; B4 defines the target parameter for which estimates are consistent.

We also define the following notation

$$
\begin{aligned}
I &= \mathrm{E}[U_n(\beta^*)] \\
V &= I^{-1}\,\mathrm{E}[U_n(\beta^*)^T U_n(\beta^*)]I^{-1} \\
J_{s,rt} &= \mathrm{E}[U_r(\beta^*)\dot{U}_{st}(\beta^*)] \\
K_{rst} &= \mathrm{E}[\ddot{U}_{rst}(\beta^*)]
\end{aligned}
$$

where $U_r(\beta^*)$ is the r^{th} component of $U(\beta^*)$, $\dot{U}_{rs}(\beta) = \frac{\partial U_r}{\partial \beta_s}$, and $\ddot{U}_{rst}(\beta) = \frac{\partial \dot{U}_{rs}}{\partial \beta_t}$. We note that I and V are $p \times p$ matrices and J and K are $p \times p \times p$ arrays. Theorem 3 provides a general form of the first-order bias approximation.

Theorem 3 (Robust Bias Approximation)**.** *Let $U_n(\beta) = 0$ be an estimating equation that satisfies A1-A3, with the solution $\hat{\beta}$ such that $U_n(\hat{\beta}) = 0$. Under assumptions B1-B3,*

$$
b(\beta_j) = \sum_r \sum_s \sum_t I^{jr}(I^{st}J_{s,rt} + \frac{1}{2}V_{st}K_{rst})
$$

is the first order approximation of the bias of $\hat{\beta}_j$, that is $\mathrm{E}[\hat{\beta}_j - \beta_j^*] = b(\beta_j) + o(n^{-1})$.

For ease of notation, we demonstrate that Theorem 3 holds for a single-parameter model. A more detailed derivation for single and multi-parameter models is available in Appendices A.1 and A.2 respectively. While not shown here, the extension to the multi-parameter setting can be seen using tensor notation as used by McCullagh [51]. The Taylor expansion has the form

$$
U_n(\hat{\beta}) = U_n(\beta^*) + \dot{U}_n(\beta^*)(\hat{\beta} - \beta^*) + \frac{1}{2}\ddot{U}_n(\beta^\dagger)(\hat{\beta} - \beta^*)^2, \tag{3.9}
$$

where $\beta^\dagger$ is an intermediate point between β^* and $\hat{\beta}$. Then by assumption B1 and B3

$$|\ddot{U}_n(\beta^\dagger)(\hat{\beta}-\beta^*)^2 - \ddot{U}_n(\beta^*)(\hat{\beta}-\beta^*)^2| \leqslant R|\beta^\dagger-\beta^*|(\hat{\beta}-\beta^*)^2$$
$$\leqslant R|\hat{\beta}-\beta|^3 = O_p(n^{3/2})$$

.

Using this result, equation (3.9) can be written as

$$U_n(\hat{\beta}) = U_n(\beta^*) + \dot{U}_n(\beta^*)(\hat{\beta}-\beta^*) + \frac{1}{2}\ddot{U}_n(\beta^*)(\hat{\beta}-\beta^*)^2 + o_p(n^{-1}) \tag{3.10}$$

.

Taking the expectation and rearranging terms gives the expression

$$\mathrm{E}[U_n'(\beta^*)]\,\mathrm{E}(\hat{\beta}-\beta^*) = \mathrm{Cov}(U_n'(\beta^*),\hat{\beta}-\beta^*) + \frac{1}{2}\mathrm{E}[U_n''(\beta^*)]\,\mathrm{E}[(\hat{\beta}-\beta^*)^2]$$
$$+ \frac{1}{2}\mathrm{Cov}(U_n''(\beta^*),(\hat{\beta}-\beta^*)^2) + o(n^{-1})$$

On the right side of the equation the substitution $\hat{\beta}-\beta^* = I^{-1}U_n(\beta^*) + O_p(n^{-1})$ is used. Furthermore $\mathrm{Cov}(U_n''(\beta^*), I^{-1}U_n(\beta^*)^2 I^{-1}) = O(n^{-2})$. Then the bias is

$$\mathrm{E}[\hat{\beta}-\beta] = I^{-1}(I^{-1}J + \frac{1}{2}VK) + o(n^{-1})$$

for $V = I^{-1}\,\mathrm{E}[U_n(\beta^*)U_n(\beta^*)]I^{-1}$. The first order approximation to the bias can be written as

$$b(\beta) = I^{-1}(I^{-1}J + \frac{1}{2}VK) \tag{3.11}$$

which is the single-parameter expression given in Theorem 3.

This can be seen as a generalization of the bias found by Cox and Snell [17]. In the special case of a correctly specified working distribution the bias approximation can be explicitly calculated, where $\mathrm{E}[U_n(\beta^*)U_n(\beta^*)] = I$ and $V = I^{-1}$. The first order bias approximation is $I^{-1}I^{-1}(J + \frac{K}{2})$ which is the expression derived by Cox and Snell.

When the true mean model and probability distribution are known (but not necessarily equal to the working mean model and distribution) the expectations and bias can be worked out explicitly. In this setting we refer to the bias approximation as the corrected-bias approximation (CBA) as the bias approximation derived from a misspecified model is being 'corrected' for the true mean model and distribution.

However, in general, when the mean model or probability distribution are misspecified, the expectations may not calculable and an empirical estimate of the first order bias is needed. This can be done by empirically estimating each component of the bias. Empirical estimates for each of the components are provided below:

$$\begin{aligned}
\hat{I} &= \frac{1}{n}\sum U_i(\beta) \\
\hat{J} &= \frac{1}{n^2}\sum U_i(\beta)\dot{U}_i(\beta) \\
\hat{K} &= \frac{1}{n}\sum \ddot{U}_i(\beta) \\
\hat{V} &= \frac{1}{n^2}\sum \hat{I}^{-1}U_i(\beta)^2\hat{I}^{-1}.
\end{aligned}$$

which can be used to estimate the bias empirically

$$\hat{b}(\beta) = \hat{I}^{-1}(\hat{I}^{-1}\hat{J} + \frac{1}{2}\hat{V}\hat{K}).$$

By the law of large numbers, these empirical estimates will be consistent for their respective true

components ($\hat{I} - I = o(1)$, $n(\hat{J} - J) = o(1)$, $\hat{K} - K = o(1)$, and $n(\hat{V}) - V = o(1)$). Then

$$\hat{b}(\beta) - b(\beta) = o(n^{-1}). \tag{3.12}$$

We refer to this empirical estimation of the first order bias approximation as the robust bias approximation (RBA) as it does not rely on correct model assumptions

In the special case of a canonical link, $J = 0$, and $\hat{I} = I$ and $\hat{K} = K$ since the empirical estimates do not depend on Y. In this setting, V is the only quantity that needs to be estimated empirically.

The use of empirical estimates for I, J, and K has previously been suggested. Firth argued that using empirical values can lead to an improvement in efficiency of the standard error of the parameter estimates. However, we have demonstrated the importance of these empirical estimates, as well as using an empirical estimate for V, to ensure an accurate approximation of the bias under model misspecification when the true model is unknown.

Both the robust bias and corrected bias approximations provide an accurate first order bias approximation without relying on a correctly specified working model (though CBA does require correct model assumptions). These approximations can be used to develop preventative robust and corrected bias reduction methods.

3.4.2 Robust Preventative Bias Reduction

We present the corrected bias reduction (CBR) method using the CBA. The robust bias reduction (RBR) method, which use the RBA, has the same form with an additional remainder term from the empirical estimate. Later, we explicitly show this additional remainder does not affect the first order bias reduction during the derivations of the bias of CBR estimates.

The CBR (and RBR) method is a direct application of the CBA (and RBA) within Firth's preventative bias reduction framework. The modified estimating equation can be written as $U^*(\beta) = U(\beta) - Ib(\beta)$ where $U(\beta)$ is a standard score, $I = -\mathrm{E}[\dot{U}(\beta^*)]$ and $b(\beta)$ has the form of the bias approximation established in equation (3.11). The bias reduced estimate is the solution to the

modified estimating equation

$$\tilde{\beta} : U^*(\tilde{\beta}) = 0. \tag{3.13}$$

This estimate has the same asymptotic properties as the MLE but reduced finite sample bias. We first demonstrate consistency and asymptotic normality of $\tilde{\beta}$ before demonstrating it effectively reduces the bias.

In order to demonstrate consistency and asymptotic normality, we demonstrate that, under an additional assumption on $Ib(\beta)$, the criteria for Theorems 1 and 2 are satisfied.

Assumption

C1 Each component of $Ib(\beta)$ is continuously differentiable and $\frac{\partial Ib(\beta)}{\partial \beta} = O(n^{-1})$ element-wise.

This assumption ensures the derivative from the modification to the estimating equation is of the same order as the modification itself.

Theorem 4 (Consistency of $\tilde{\beta}$)**.** *Let $U_n(\beta)$ be an estimating equation that satisfies A1 and A3. Define $U_n^* = U_n(\beta) - Ib(\beta)$ and $\tilde{\beta}$ such that $U_n^*(\tilde{\beta}) = 0$. If assumption A1.1 is met, then $\tilde{\beta}$ is a consistent estimator for β^*.*

Proof. To prove this theorem we need to demonstrate that the assumptions A1 and A2 for Theorem 1 are met.

We first need to show $U_n^*(\beta^*) \to 0$ with probability one. We note that each component of $-Ib(\beta^*)$ is $O(n^{-1})$, therefore $Ib(\beta^*) \to 0$ with probability one. Then by assumption A1 on $U_n(\beta)$, $U_n^* = U_n(\beta^*) - Ib(\beta^*) \to 0$ with probability one and $U^*(\beta)$ satisfies assumption A1.

We now need to show assumption A2 is met. From assumption A2, $U_n(\beta)$ is continuously differentiable. Under assumption A1.1 $Ib(\beta)$ is differentiable as well. Since $U_n^*(\beta) = U_n(\beta^*) - Ib(\beta^*)$ is the sum of continuously differentiable functions, $\mathbf{U}_n^*(\beta)$ is continuously differentiable as well.

Secondly, since $\dot{U}_n(\beta)$ converges uniformly to I and $\frac{\partial Ib(\beta)}{\partial \beta}$ converges in probability to 0, then $\dot{U}_n^*(\beta)$ converges uniformly to I which is non-stochastic and nonsingular. Therefore, assumption A2 is also met for $U^*(\beta)$. Then by theorem 1, $\tilde{\beta}$ is consistent for β^*.

□

Theorem 5 (Asymptotic Normality of $\tilde{\beta}$)**.** *Let $U_n(\beta)$ be an estimating equation that satisfies assumptions A1-A4 and $\tilde{\beta}$ be the estimator in equation* (3.13)*. If assumption A1.1 is met, then $\sqrt{n}(\tilde{\beta}-\beta^*) \to N(0,V)$ where $V = I^{-1}\mathrm{E}[U(\beta^*)U(\beta^*)^T]I^{-1}$ for $I = \mathrm{E}[\dot{U}(\beta^*)]$.*

Proof. To prove this theorem we need to further demonstrate assumption A4 is met for $U_n^*(\beta)$. Since $\sqrt{n}U_n(\beta^*) \to N(0,V)$ and $\sqrt{n}Ib(\beta^*) \to 0$ in probability, then by Slutsky's theorem, $U_n^*(\beta^*) \to N(0,V)$ and assumption A4 is met for $U_n^*(\beta)$.

We have demonstrated that the estimate $\tilde{\beta}$ is asymptotically consistent and normally distributed. Furthermore, we have shown there is no loss of asymptotic efficiency as the limiting distribution is the same as estimates from an unmodified estimating equation. Since $\tilde{\beta}$ has the same asymptotic variance as $\hat{\beta}$, robust forms of the standard error estimates for $\hat{\beta}$ are still consistent for the variance of $\tilde{\beta}$ as well.

□

We now show the finite sample bias of $\tilde{\beta}$ is $o(n^{-1})$

3.4.3 Bias of $\tilde{\beta}$

A Taylor expansion about $U_n^*(\tilde{\beta}) = 0$ can be used to derive the bias of $\tilde{\beta}$. The following additional assumptions are needed to ensure the higher order remainder of the modification is sufficiently small.

D1 Each component of $\frac{\partial^2 Ib(\beta)}{\partial \beta^2}$ is a Lipschitz function

D2 $\frac{\partial Ib(\beta)}{\partial \beta}$ and $\frac{\partial^2 Ib(\beta)}{\partial \beta^2}$ are $O(n^{-1})$

Note that since $\ddot{U}_n^*$ is the sum of two Lipschitz functions, $\ddot{U}_n^*$ is also Lipschitz. Therefore

$$|\ddot{U}_n^*(\beta^\dagger)(\tilde{\beta}-\beta^*)^2 - \ddot{U}_n(\beta^*)(\tilde{\beta}-\beta^*)^2| = o(n^{-1}). \tag{3.14}$$

where $\beta^\dagger$ is an intermediate point between $\tilde{\beta}$ and β^*, as was shown in the previous derivation of the bias approximation.

Then, using a Taylor expansion and equation (3.14),

$$\begin{aligned} 0 &= U_n^*(\beta^*) + \dot{U}_n^*(\beta^*)(\tilde{\beta}-\beta^*) + \frac{1}{2}\ddot{U}_n^*(\beta^\dagger)(\tilde{\beta}-\beta^*)^2 \\ &= U_n^*(\beta^*) + \dot{U}_n^*(\beta^*)(\tilde{\beta}-\beta^*) + \frac{1}{2}\ddot{U}_n^*(\beta^*)(\tilde{\beta}-\beta^*)^2 + o(n^{-1}) \end{aligned}$$

following the same steps going from equation (3.9) to equation (3.10).

Rewriting $U_n^*(\beta^*)$ (and its derivatives) using the standard estimating function $U_n(\beta^*)$ and the modification $Ib(\beta^*)$

$$0 = U_n(\beta^*) + \dot{U}_n^*(\beta^*)(\tilde{\beta}-\beta^*) - Ib(\beta^*) + \frac{1}{2}U_n''(\beta^*)(\tilde{\beta}-\beta^*)^2 + r_n + o_p(n^{-1})$$

where $r_n = \frac{\partial Ib(\beta)}{\partial \beta}(\tilde{\beta}-\beta^*) + \frac{1}{2}\frac{\partial^2 Ib(\beta)}{\partial \beta \partial \beta}(\tilde{\beta}-\beta^*) = o(n^{-1})$ by assumption D2.

Then the expected value is

$$\begin{aligned} \mathrm{E}[\tilde{\beta}-\beta^*] &= I\left(I^{-1}(I^{-1}J + \frac{1}{2}VK)\right) - Ib(\beta^*) + o(n^{-1}) \\ &= Ib(\beta^*) - Ib(\beta^*) + o(n^{-1}) \\ &= o(n^{-1}). \end{aligned} \tag{3.15}$$

Thus the bias of $\tilde{\beta}$ is of order $o(n^{-1})$.

When the first order bias approximation is unknown, the empirical estimate for bias, $\hat{b}(\beta)$ in

equation (3.12) can be used in place of $b(\beta)$. The modified estimating function using $\hat{\beta}$ is

$$\begin{aligned} U_n^*(\beta) &= U_n(\beta) - I\hat{b}(\beta) \\ &= U_n(\beta) - Ib(\beta) + o(n^{-1}) \end{aligned}$$

and, using the result in equation (3.15), the bias of $\tilde{\beta}$ is

$$\mathrm{E}[\tilde{\beta} - \beta^*] = o(n^{-1}).$$

3.4.4 Motivating Example Revisited

We now revisit the motivating example at the start of this chapter. Since the true distribution of the outcomes is known and the mean model is correctly specified, we can apply the CBR method to a Poisson regression model used to estimate relative risks. Then based off the working Poisson distribution and true binomial distribution, $V = \frac{1-\exp(\beta_0)}{\exp(\beta_0)}$ in equation (3.11). As the canonical link is used $I = \exp(\beta_0)$, $J = 0$, and $K = -\frac{1}{2n}\exp(\beta_0)$ which are equivalent to the values from the NBR method.

Then the first order bias approximation is

$$\frac{1}{2n}\left(\frac{\exp(\beta_0) - 1}{\exp(\beta_0)}\right).$$

This is identical to the bias approximation from a correctly specified model. Furthermore, the parameter estimate derived from the CBR method is identical to the estimate from a correctly specified NBR method. While equivalence between the methods is not generally true, it illustrates a setting where despite using a misspecified model, we obtain an estimate as though it were from a correctly specified model. This is particularly useful as we can still recover the correctly specified, bias reduced estimate even if such a model cannot be directly fit due to issues with fitting algorithms.

In general, no closed form expression is available for bias reduced estimates. In the next section, we describe an iterative fitting algorithm that can be used to solve the modified estimating equations.

3.4.5 Implementation of the Robust Bias Reduction Method

The robust bias reduction method can easily be implemented using a quasi Newton-Raphson method proposed by Kosmidis and Firth [42]. For a modified score function, $U^*(\beta)$, the quasi Newton-Raphson iteration is

$$\beta^{j+1} = \beta^j + I_j^{-1} U_n^*(\beta^j) \tag{3.16}$$

where $I_j = \mathrm{E}[U_n(\beta^j)]$, the information from the unmodified score.

3.5 Simulations

In this section, we compare the performance of the CBR, RBR, NBR, and jackknife bias reduction techniques along with estimates from a standard GLM using log-binomial data. We report the percent bias and bias/SE for the predictor of interest in each of the scenarios. We also record the rates of non-convergence in the fitting algorithms to identify the practicality of using the models in application. We explore three different scenarios. Estimates from the following methods are reported:

- **Standard-Poisson**: A log-Poisson GLM with no bias reduction
- **NBR-Poisson**: A naive bias reduction method for log-Poisson GLM
- **NBR-Binomial**: A naive bias reduction method for log-binomial GLM

- **CBR-Poisson**: A corrected bias reduction method for log-Poisson GLM with a binomial correction
- **RBR-Poisson**: A robust bias reduction method for log-Poisson GLM
- **Jacknife**: A jackknife bias reduction for log-Poisson GLM.

Each scenario uses 1,000 replications. Details for each scenario are provided below.

3.5.1 Data Scenarios

Scenario 1: Skewed Predictor

For the first scenario, we consider the association between a continuous predictor, X, and a binary outcome Y such that

$$\mathrm{E}[Y|X] = \exp(\beta_0 + \beta_1 X) \tag{3.17}$$

for $\beta_0 = \log(.5)$ and $\beta_1 = \log(.9)$. In this setting, X is a skewed predictor reflecting a scenario with a moderate amount of bias based on the characterizations in Chapter 2.

Scenario 2: Binary Predictor

The second scenario estimates the association between a balanced binary predictor, X, and a binary outcome, Y with the mean model

$$\mathrm{E}[Y|X] = \exp(\beta_0 + \beta_1 X) \tag{3.18}$$

for parameters $\beta_0 = \log(.3)$ and $\beta_1 = \log(2)$. Since x has an equal number of 0's and 1's parameter estimates and bias for log-Poisson GLM and log-Binomial GLM are equivalent. In this case NBR-Poisson and NBR-Binomial estimates are equivalent.

Scenario 3: Misspecified Mean Model

In the final scenario we assume X is a discrete predictor with values of 0, 1, 2, 3, and 4 occurring at the same rates. The mean model for the binary outcome Y is

$$\mathrm{E}[Y|X] = \exp(\log(.5) + \log(.9)X^2) \tag{3.19}$$

However, the working mean model $\exp(\beta_0 + \beta_1 X)$ is used to assess the linear relationship between the predictor, X and the log risk of an event. In this case the mean model is misspecified and the true population-level working model parameters are $\beta_0^* = -0.664$ and $\beta_1^* = -0.203$.

3.5.2 Results

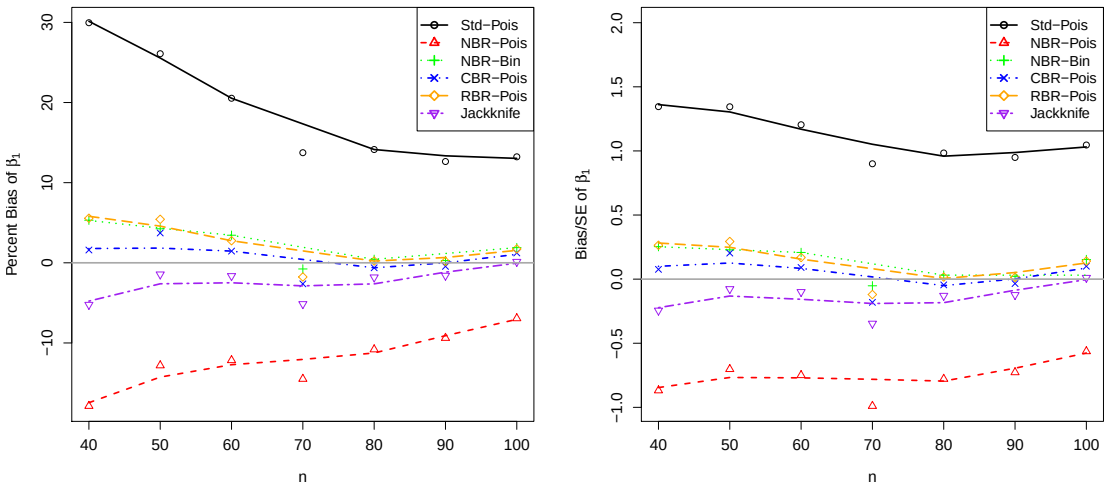
Table 3.2 shows the worst case scenario for rates of fitting algorithms not converging for each method which occurred for the smallest sample size we explored, 40. The NBR-binomial had the worst rates of non-convergence, particularly for a continuous predictor. Both CBR-Poisson and RBR-Poisson methods had periodic issues with non-convergence but these occurred infrequently. Even at a sample size of 100, NBR-binomial methods still had non-convergence rates of approximately 3%, while no other methods had occurrences of non-convergence. These non-convergence rates are not prohibitively large; however, smaller sample sizes are likely to lead to an increase in non-convergence for both NBR-binomial and CBR-binomial methods as the estimated probability is more likely to fall outside $(0,1)$ during the fitting process.

Table 3.2: Percentage of data scenarios where the model did not converge to parameter estimates with a sample size of 40

Method	Scenario 1	Scenario 2	Scenario 3
GLM-Poisson	0	0	0
NBR-Poisson	0	0	0
NBR-Binomial	5.0	0	4.7
CBR-Poisson	< 0.1	0	< 0.1
RRB-Poisson	0	0.1	0
Jackknife	0	0	0

Scenario 1: Skewed Predictor

From figure 3.1, the magnitude of bias for standard Poisson estimates and NBR-Poisson estimates are similar, though of opposite sign, as the NBR-Poisson is over-correcting the bias. The jackknife had poor performance for small sample sizes but effectively reduced the bias for larger sample sizes. The CBR-Poisson and RBR-Poisson performed similarly to the NBR-Binomial model indicating the robust bias reduction methods had similar performance as a correctly specified bias reduction method. However, the fitting algorithms for the robust and corrected bias reduction methods had better performance than the NBR-Binomial estimates.



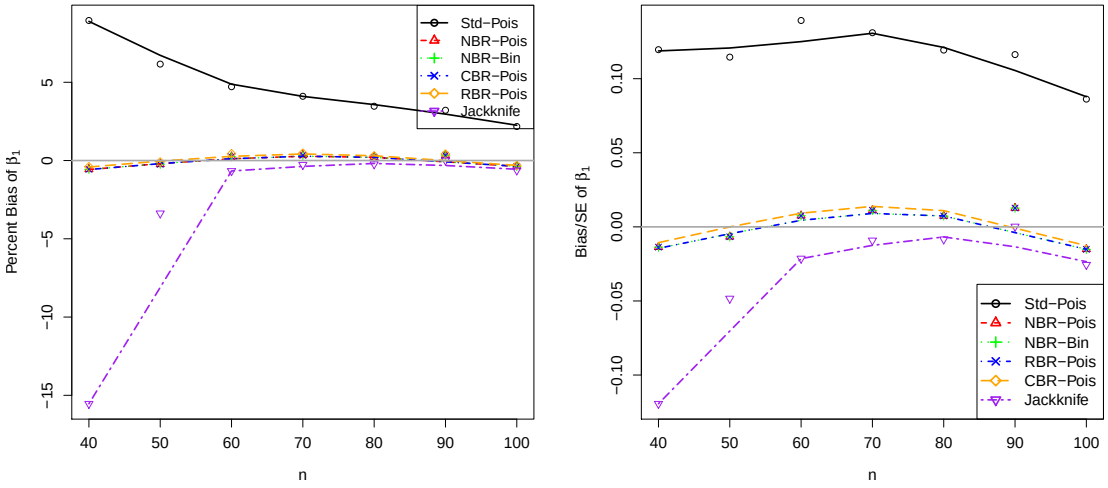
(a) Bias (b) Bias/SE

Figure 3.1: Percent bias (a) and bias/SE (b) of standard Poisson (Std-Pois), NBR Poisson (NBR-Pois), NBR Binomial (NBR-Bin), CBR Poisson (CBR-Pois), RBR Poisson (RBR-Pois), and jackknife methods for a continuous predictor of interest.

Similar trends were seen when considering the bias/SE ratio. However, we see the rates of convergence to 0 appear to be slower, particularly for the standard Poisson estimates as was hypothesized in Chapter 2.

Scenario 2: Binary Predictor

Results are shown in figure 3.2. As expected, the bias of the NBR-Poisson, NBR-Binomial, and CBR-Poisson have identical performance and nearly eliminate the bias compared to the standard Poisson estimates. For small sample sizes, the jackknife method has the worst performance of the bias reduction methods, likely due to separation of the data in subsamples. As the sample size increases, the performance of the jackknife improves. For moderate sample sizes, all bias reduction methods had similar performance. Despite not exploiting knowledge of the true model, the RBR method provides a reduction in bias comparable to bias reduction methods that take advantage of this knowledge. This held even for sample sizes as small as 40



(a) Bias

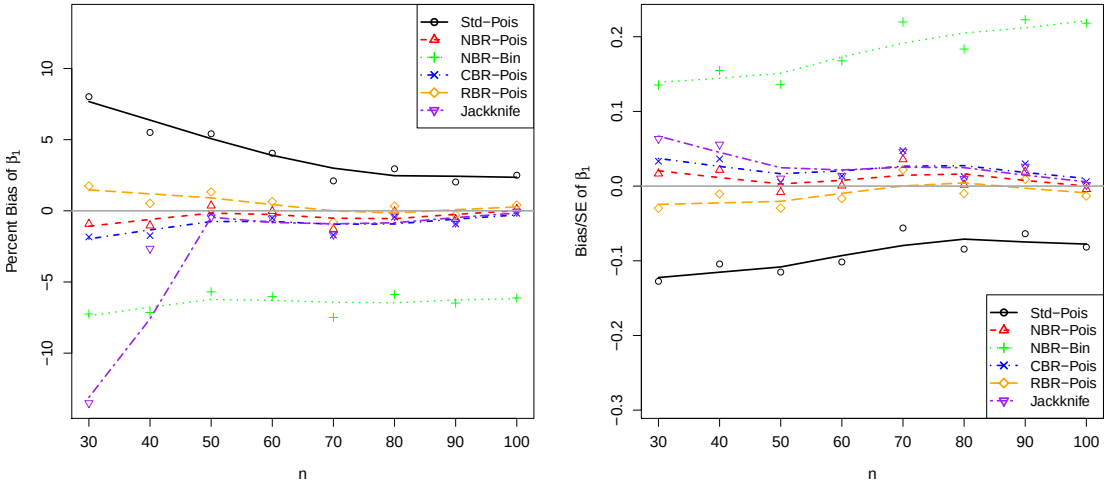
(b) Bias/SE

Figure 3.2: Percent bias (a) and bias/SE (b) of standard Poisson (Std-Pois), NBR Poisson (NBR-Pois), NBR Binomial (NBR-Bin), CBR Poisson (CBR-Pois), RBR Poisson (RBR-Pois), and jackknife methods for a binary predictor of interest.

As in scenario 1, the bias/SE has a similar trend to bias, but seems to converge at a slower rate. It also appears to be increasing in magnitude for large sample sizes; however, this can be attributed to sampling variability as the bias is negligible at the larger sample sizes as seen in the graph for bias.

3.5.3 Scenario 3: Misspecified Mean Model

Figure 3.3 displays the results for Scenario 3. In this scenario, the NBR-binomial model had poor performance, attributed to the frequent rates at which fitting algorithms did not converge. The jackknife also failed to provide bias reduction for small sample sizes (similar to Scenario 2). All other bias reduction methods effectively reduced the bias and performed better than standard estimates. When looking at the bias/SE ratio, similar trends held with the exception that jackknife had improved performance for small samples due to the larger SE in jackknife estimates.



(a) Bias (b) Bias/SE

Figure 3.3: Percent bias (a) and bias/SE (b) of standard Poisson (Std-Pois), NBR Poisson (NBR-Pois), NBR Binomial (NBR-Bin), CBR Poisson (CBR-Pois), RBR Poisson (RBR-Pois), and jackknife methods for a misspecified mean model.

3.5.4 Summary

As was demonstrated through examples and simulations, using a naive bias reduction when the assumptions are not met can lead to problems in bias reduction, even leading to an increase in the magnitude of bias in some settings. While the NBR is preferred when the assumptions are valid, problems with fitting algorithms may prevent this model from being used. The CBR method allows

for reliable bias reduction and can greatly improve the frequency that fitting algorithms converge when the true distribution of the outcomes is known. In more general settings, the RBR bias reduction method and the jackknife can both provide reliable bias reduction when the distribution of outcomes is unknown and the sample size is reasonably large. However, the jackknife had poor performance in small sample sizes. In contrast, the RBR method effectively reduced bias even with sample sizes as low as 40 and performed comparably to the NBR-binomial method but with a lower frequency of non-convergence. Both the CBR and RBR had fewer issues of non-convergence, allowing for accurate bias reduction in a wider variety of settings with little downside.

While the RBR method has been demonstrated to effectively reduce bias, further work is needed to assess the impact this bias reduction may have on inference, such as coverage of a nominal 95% confidence interval. Further work is also needed to assess the performance of these models for very small sample sizes (< 30) as the RBR method may have worse performance due to its reliance on empirical estimation. If performance is poor in small samples, further modification may be needed, similar to improvements in robust standard error estimation for small sample sizes.

3.6 Extensions of Bias Reduction to Correlated Data

The previous work assumes the outcomes being measured are independent. However, in many medical settings our outcomes are correlated. In these settings, GLMs break down as they are unable to account for this correlation, leading to inconsistent standard errors.

Generalized linear mixed models (GLMM) and generalized estimating equations (GEEs) are two common approaches that account for correlation. GLMM accounts for the correlation by adding a random effect based on clustering which allows for parameter estimates to be interpreted conditionally. GEE allows for the specification of a working correlation matrix that leads to unconditional parameter estimates by assuming the marginal variance is a function of the marginal mean of the outcome.

In this section we focus on extensions of the bias reduction methods for estimating equations

that can be easily extended to the GEE framework. In Section 3.6.1 we provide a brief overview of GEEs and derive the general form of a modified GEE to reduce bias. In Section 3.6.2 we engage with the unique challenges of bias reduction that arise from GEE models. In addition we propose a robust bias reduction method that assumes an independent working covariance but uses a robust estimation of the bias approximation that accounts for clustering. Lastly, this method is assessed empirically in Section 3.6.3.

3.6.1 Generalized Estimating Equations

GEE can be seen as an extension of estimating equations that accounts for the correlation of observations within clusters and assumes the clusters are independent. Each cluster, $i = 1, ..., m$ has a sample size n_i, with a corresponding $n_i \times 1$ vector of outcomes, y_i and an $n_i \times p$ matrix of covariates X_i. Then

$$U_n^{GEE}(\beta) = \sum_i^m D_i^T W_i^{-1}(y_i - g^{-1}(X_i\beta)) \tag{3.20}$$

where $D_i = \frac{\partial g^{-1}(X_i\beta)}{\partial \beta}$ and $W_i = \text{Cov}(y_i, y_i)$. In the case of an independence working correlation, the off-diagonals of W_i are 0 and the estimating equations are equivalent to those from a GLM. Similar to estimating equations for uncorrelated data, the target of inference is the zero of the expected value of U_n^{GEE}:

$$\beta^* : \text{E}[U_n^{GEE}(\beta^*)] = 0. \tag{3.21}$$

The parameter estimate, $\hat{\beta}_{GEE}$ is the solution to the estimating equation $U_n^{GEE}(\beta) = 0$. The estimate $\hat{\beta}_{GEE}$ is asymptotically consistent for β^*_{GEE} even when the working covariance is misspecified. However, finite sample bias is still present. Similarly as for estimating equations, a Taylor expansion can be used to derive a first order bias approximation for GEE estimates. The

Taylor expansion about $U_n^{GEE}(\beta)$ (in the single parameter setting) is

$$U_n^{GEE}(\beta^*) + (\hat{\beta} - \beta^*)\dot{U}_n^{GEE}(\beta^*) + \frac{1}{2}(\hat{\beta} - \beta^*)^2\ddot{U}_n^{GEE}(\beta^*) + o_p(n^{-1}) = 0. \tag{3.22}$$

Working through these expectations as done in equation (3.10), the bias approximation becomes

$$b_{GEE}(\beta_j) = \sum_r \sum_s \sum_t I_{GEE}^{rj} V_{st}^{GEE}(J_{t,rs}^{GEE} + \frac{1}{2}K_{rts}^{GEE}) \tag{3.23}$$

where $I^{GEE} = \mathrm{E}[-\dot{U}_n^{GEE}(\beta^*)]$, $J_{r,st}^{GEE} = \mathrm{E}[\dot{U}_{st}^{GEE}(\beta^*)U_r^{GEE}(\beta^*)]$, $K_{rst}^{GEE} = \mathrm{E}[\ddot{U}_{rst}^{GEE}(\beta^*)]$ and I_{GEE}^{rs} denotes the $(r,s)^{th}$ component of $(I^{GEE})^{-1}$.

This bias approximation has a similar form to the one previously used for independent data; however, the use of the GEE sub/super script makes clear the generalized estimating equations used for the bias approximation in equation (3.23) account for clustering in the data. This bias approximation can be used to create a set of modified GEE:

$$U_n^{GEE*}(\beta) = U_n^{GEE}(\beta) - I_{GEE}(\beta)b_{GEE}(\beta) \tag{3.24}$$

.

The bias reduced estimate $\tilde{\beta}^{GEE}$ is the zero of this modified GEE.

Similar to the CBR method discussed in the setting of independent outcomes, a similar method can be used in the GEE framework when the true form of the covariance matrix is known. In this case, the knowledge of the true covariance matrix can be used to 'correct' the bias approximation even when the working covariance is misspecified. We refer to this method as the cluster-corrected bias reduction method (cl-CBR).

In many cases the true covariance matrix is unknown and a robust empirical method is needed

to estimate the components of the first order bias, which we refer to as the cluster-robust bias reduction method (cl-RBR).

In the next section, we argue that using a simplified cl-RBR with an independent working covariance structure is a straightforward extension of previous bias reduction methods that can be used in the presence of clustering.

3.6.2 GEE Bias Reduction in Application

While the expression for the modified GEE appears straightforward, deriving the form of the first order bias approximation can be difficult in practice due to the correlation between outcomes. Paul et al tried to work around this problem by using a decoupling method which allows the covariance matrix to be treated as constant relative to the parameters of interest [58]. However, this assumes the working covariance matrix is constant with respect to β. This seems overly limiting as the covariance matrix frequently depends on β through the mean-variance relationship. We propose an alternative method of simplifying the working covariance matrix which allows for the covariance matrix to be a function of β. The most straightforward way to ensure this is to assume a working independence matrix, which simplifies the estimating equations to those from a GLM. However, the modified GEE is not the same as a modified estimating equation. The modified GEE has the form

$$U^*_{GEE}(\beta) = U_n(\beta) - Ib_{GEE-I}(\beta) \tag{3.25}$$

where

$$b_{GEE-I}(\beta_s) = \sum_r \sum_t \sum_u I^{rs}_{GEE-I} V^{tu}_{GEE} (J^{GEE-I}_{t,ru} + \frac{1}{2} K^{GEE-I}_{rtu}). \tag{3.26}$$

We note that when using an independent working covariance matrix, $I_{GEE-I} = I$, $J_{GEE-I} = J$, and $K_{GEE-I} = K$ where I, J, and K are the corresponding expressions from the GLM framework for

independent data. However, $V_{GEE} \neq V$ and needs to be calculated using a method that accounts for clustering.

If the outcomes were known to be independent, V_{GEE} would be a diagonal matrix and the modified estimating equations would be equal to those used in the RBR and CBR framework. For correlated outcomes, a robust standard error estimate, $\hat{V}_{GEE}$, will provide accurate bias reduction without needing to specify the specific correlation of our data. Similar to the RBR method, a robust bias reduction method allowing for clustering can also be used to achieve a reduction in bias even when the working covariance matrix is misspecified. While restricting to an independence working distribution may seem limiting, and open to a similar critique that this is an oversimplification for the sake of easing derivations, the use of an independence working distribution is not as problematic as it may seem.

In many settings an independent working correlation is reasonably efficient [52]. While specification of other working correlations may lead to an improvement in efficiency if correct, they may also lead to a decrease in efficiency if incorrect compared to an independent working correlation. This is not necessarily true if we have time-varying covariates, in which case a large reduction in efficiency may occur [26].

Mancl and Leroux identified other distributions of covariates that may also lead to large inefficiency when using an independence working correlation, particularly when covariates are poorly balanced between clusters [49]. Mancl and Leroux also found when covariates are constant within a cluster, vary within clusters but are mean-balanced between clusters, or the correlation within clusters is small, high efficiency is obtained with an independence working correlation.

In addition, Pepe and Anderson demonstrated that using an independence working correlation guarantees consistency of parameter estimates, while the use of other working correlations requires additional restrictions if we have time-dependent covariates [59]. Thus if we are unsure if the condition is met, an independence working correlation is reasonable as it guarantees consistency and will have high efficiency in many settings regardless of the true working distribution. The use of an independent working correlation also leads to a simplification in the computation of bias

reduced parameters.

While methods exist for selecting an appropriate working covariance matrix that would likely lead to an increase in efficiency, they are generally used for model selection and require the fitting of multiple models [15,36,57,75]. While settings exist where this is feasible, in many research settings we may prefer to specify a working correlation a priori. In these settings, without knowledge of the true correlation, an independent working correlation is reasonable.

In the next section, we assess the performance of an cl-RBR GEE that is based on an independence working covariance.

3.6.3 Simulations

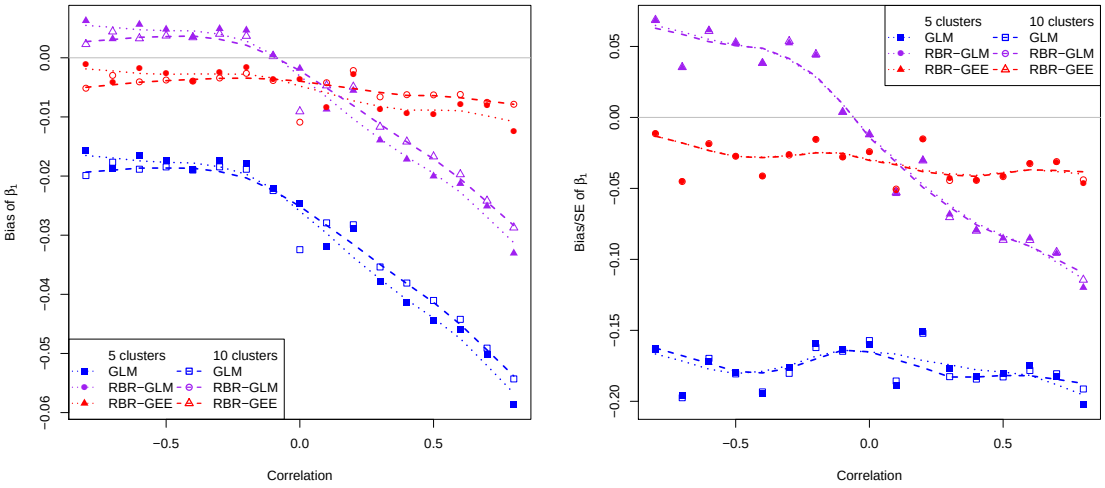
Scenarios

Let Y consist of 50 Bernoulli observations with mean $\mathrm{E}[Y|x] = \log(0.5) + \log(0.9)x$ where x has a centered gamma distribution with shape and scale equal to 1. We consider two clustering structures for y. The first setting consists of 5 clusters each with sample size 10 while the second consists of 10 clusters each with sample size 5. In both of these settings, the within cluster correlation is exchangeable and clusters are independent of each other. This is meant to assess the impact the number of clusters has on estimates assuming the total sample size is fixed.

We fit standard Poisson, RBR-Poisson, and cl-RBR-Poisson models, all of which assume an independence working correlation. The RBR-Poisson method uses a robust bias approximation that fully relies on the independence assumption but allows for misspecification of the marginal distribution while the cl-RBR-Poisson method allows for misspecification of the working correlation and marginal distribution of the outcomes. Bias and bias/SE ratio are reported for all models for both the intercept and slope parameters.

Results

Figure 3.4 shows the bias and bias/SE ratio for the intercept, β_0. Both the RBR and the cl-RBR estimates have reduced bias relative to standard GLM estimates. For negative correlations the RBR and cl-RBR estimates have a similar magnitude of bias but with opposite signs. For positive correlations the cl-RBR has the smallest bias, particularly for large correlations. Notably, the cl-RBR estimate has relatively constant bias relative to the correlation while the bias of standard and RBR estimates depends on the correlation. A similar trend is seen with the bias/SE ratio. Both the standard and cl-RBR estimates are relatively constant with respect to the correlation; however, cl-RBR have a much smaller bias/SE ratio. Estimates were minimally affected by the choice of cluster size in this scenario.



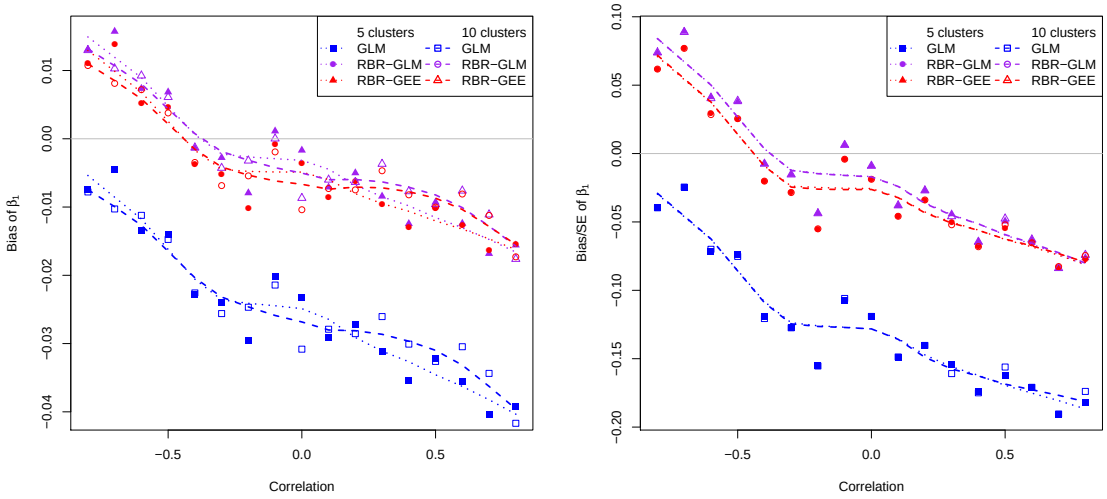
(a) Bias

(b) Bias/SE

Figure 3.4: Bias (a) and bias/SE ratio (b) for the intercept parameter, β_0, for clustered data with an independent working covariance matrix.

The results in figure 3.5 demonstrate that RBR methods accounting for clusters perform similarly to those that do not under an independent working covariance when considering β_1. Both of these methods had improved performance for positive correlations and moderate correlations relative to standard estimates. For extreme negative correlations, standard estimates performed

similarly but had bias of the opposite sign. The impact of the number of clusters, holding the total sample size constant, was limited, with similar performance between 10 clusters each with sample size 5, or 5 clusters each with sample size 10.



(a) Bias

(b) Bias/SE

Figure 3.5: Bias (a) and bias/SE ratio (b) for the slope parameter, β_1, for clustered data with an independent working covariance matrix.

3.6.4 Summary

These simulations demonstrate that bias reduction methods that rely on independent outcomes may not provide accurate bias reduction. However, the RBR method can be extended to account for clustering in the data through GEE. A modified GEE using a independence working distribution but using the cl-RBR method that allows for clustering provides a straightforward solution to bias reduction in the presence of clustering. While this method theoretically extends to other working correlations, in application, these methods require a complex derivation of the first order bias approximation. Further work in developing these methods could improve the performance of cl-RBR methods, particularly when the form of the correlation is known.

Additional work is also needed to more thoroughly investigate the performance of cl-RBR methods in a wider range of scenarios, including different forms of the true correlation structure

and the effect the cluster size and number of clusters has on bias and bias reduction.

Extending bias reduction methods to generalized linear mixed models is one direction for future exploration.

3.7 Discussion

In this chapter we have developed a robust bias approximation that can be used in the preventative bias reduction framework to provide accurate bias reduction. This fills an important gap in the bias reduction literature as this method can be used under misspecified models and separation of the data, which no previous methods were able to do. Furthermore, in misspecified settings where the true model is known, the true model can be exploited to derive corrected bias reduction methods which may have slightly better performance in some settings.

We demonstrated empirically that the RBR method effectively reduces bias in correctly and misspecified models and performs well with sample sizes as low as 30. We extended the RBR method to GEE which allows for clustering in the data. This extension seems promising based on an initial empirical setting; however, future work is needed to fully assess the performance in clustered settings.

In the next chapter, we explore consequences of bias and bias reduction for transformed parameters and meta-analyses.

Chapter 4

Indirect Consequences of Bias Reduction

4.1 Introduction

In the previous chapter the effect of bias reduction on estimates from a single study has been evaluated. However, in many settings, coefficient estimates from a given model may not be of primary interest. We explore two such scenarios.

In Section 4.2, we explore the setting where the true parameter of interest is a transformation of the coefficient parameter. In this setting the transformed estimate is consistent for the transformed parameter, but finite sample bias may still be present. We discuss how bias reduction of the coefficient estimates may not reduce the bias of transformed estimates. In addition, we present a corrective bias-reduction method for transformations of both standard and bias reduced coefficient estimates. We further demonstrate that transforming bias reduced estimates simplifies the bias reduction of transformed estimates.

In Section 4.3, we consider pooling estimates from multiple studies through meta-analysis, which leads to a gain in precision. However, bias may still be present in meta-analysis which can lead to an increase in type I error as precision increases. We explore the impact bias and bias reduction methods have on the bias of pooled estimates, demonstrating that in some scenarios, bias reduction can lead to an increase in the pooled estimate bias.

4.2 Bias Reduction and the Delta Method

In some settings the model parameter, β^*, is not of primary interest. Instead, we are interested in a transformation of the coefficient. One particular setting where this occurs is estimating relative risks. The coefficient, β, represents the log-RR. The estimate, $\hat{\beta}$ can be obtained from GLM. However, for ease of interpretation, the RR is preferred and can be obtained by exponentiating β. In this setting, $\exp(\beta)$ is the parameter of interest and $\exp(\hat{\beta})$ is consistent for $\exp(\beta)$.

Now suppose we are interested in a general transformation of a parameter, $\omega^* = h(\beta)$, with corresponding estimate, $\hat{\omega} = h(\hat{\beta})$ such that $\dot{h}(\beta^*) = \frac{\partial h(\beta^*)}{\partial \beta^*}$ exists and is non-zero.

Under mild regularity conditions, if

$$\sqrt{n}(\hat{\beta} - \beta^*) \to N(0, \Sigma), \tag{4.1}$$

then by the delta-method

$$\sqrt{n}(\hat{\omega} - \omega^*) \to N(0, \dot{h}(\beta^*)\Sigma\dot{h}(\beta^*)). \tag{4.2}$$

Under the conditions discussed in Chapter 4, the bias reduced estimate $\tilde{\beta}$ has the same asymptotic distribution as $\hat{\beta}$, that is

$$\sqrt{n}(\tilde{\beta} - \beta^*) \to N(0, \Sigma),$$

the asymptotic distribution is identical if a bias reduced estimate is used instead. Then $\tilde{\omega} = h(\tilde{\beta})$ can be used in place of $\hat{\omega}$ and equation (4.2) still holds for $\tilde{\omega}$. Asymptotically, the two estimates are equivalent; however, the finite sample biases of $\hat{\omega}$ and $\tilde{\omega}$ are not equivalent. Furthermore, the presence of bias in the coefficient estimate does not imply bias in the transformed estimate. In fact, $\hat{\omega}$ may be unbiased even when $\hat{\beta}$ is biased. For example, when estimating the population mean, $\bar{y}$

is unbiased for ω^* even though $\log(\bar{y})$ is biased for β^*.

We calculate the respective biases of $\hat{\omega}$ and $\tilde{\omega}$ below.

Bias of $h(\hat{\beta})$

We now derive the bias of $h(\hat{\beta})$. From the Taylor approximation

$$\hat{\omega} = \omega + \dot{h}(\beta)(\hat{\beta} - \beta^*) + \frac{1}{2}\ddot{h}(\beta^*)(\hat{\beta} - \beta^*)^2 + O(n^{-3/2})$$

the bias can be expressed as

$$\mathrm{E}[\hat{\omega} - \omega] = \dot{h}(\beta^*)\,\mathrm{E}[\hat{\beta} - \beta^*] + \frac{1}{2}\ddot{h}(\beta^*)\,\mathrm{E}[(\hat{\beta} - \beta^*)^2] + O(n^{-3/2}). \tag{4.3}$$

For convenience, we write $\beta^* = \mathrm{E}[\hat{\beta}] + b(\beta^*) + o(n^{-1})$, where $b(\beta^*)$ is the first order bias approximation. Then

$$\mathrm{E}[\hat{\beta} - \beta^*] = \mathrm{E}[\hat{\beta} - \mathrm{E}[\hat{\beta}] + b(\beta^*) + o(n^{-1})] = b(\beta^*) + o(n^{-1}) \tag{4.4}$$

and

$$\begin{aligned}\mathrm{E}[(\hat{\beta} - \beta^*)^2] &= \mathrm{E}[(\hat{\beta} - \mathrm{E}[\hat{\beta}])^2] + \mathrm{E}[b(\beta^*)^2] - 2\,\mathrm{E}[\hat{\beta} b(\beta^*)] - 2\,\mathrm{E}[\hat{\beta}] b(\beta^*)] \\ &= \mathrm{Var}(\hat{\beta}) + o(n^{-1}).\end{aligned} \tag{4.5}$$

Then using equations (4.4) and (4.5) with equation (4.3), the bias can be rewritten as

$$\mathrm{E}[\hat{\omega} - \omega] = \dot{h}(\beta^*) b(\beta^*) + \frac{1}{2}\ddot{h}(\beta^*)\,\mathrm{Var}(\hat{\beta}) + o(n^{-1}) \tag{4.6}$$

and the first order bias approximation is

$$b(\hat{\omega}) = \dot{h}(\beta^*) b(\beta^*) + \frac{1}{2}\ddot{h}(\beta^*)\,\mathrm{Var}(\hat{\beta}). \tag{4.7}$$

Therefore, when using a transformation of $\hat{\beta}$, the first order bias of $\hat{\omega}$ depends on the transformation, the bias of $\hat{\beta}$ and the variance of $\hat{\beta}$.

Bias of $h(\tilde{\beta})$

Similar steps can be used to derive the first order approximation of $h(\tilde{\beta})$:

$$b(\tilde{\omega}) = \frac{1}{2}\ddot{h}(\beta^*)\,\mathrm{Var}(\tilde{\beta}). \tag{4.8}$$

When using a transformation of $\tilde{\beta}$, $b(\beta^*) = 0$ and the first order bias only depends on the transformation and the variance of $\tilde{\beta}$.

From these expressions it is tempting to say $\tilde{\omega}$ has the smaller first order bias as $\tilde{\beta}$ has no first order bias, however, this is not generally true. In some settings $\dot{h}(\beta^*)b(\beta^*) = -\ddot{h}(\beta^*)\,\mathrm{Var}(\hat{\beta})$ and $\hat{b}(\omega) = 0$ but $b(\tilde{\omega}) = \ddot{h}(\beta^*)\,\mathrm{Var}(\tilde{\beta}) \neq 0$. We illustrate this phenomenon by considering an exponential transformation on an intercept only, log-linear model.

Exponential transformation

Suppose we want to estimate the population mean, ω^*. We note in this setting, the sample mean, $\bar{y}$ is an unbiased estimate of ω. We can also estimate the population mean using the coefficient estimate from a log-linear model, $\hat{\beta} = log(\bar{y})$. Then $\exp(\hat{\beta})$ is an estimate of ω. The first order bias approximation of $\exp(\hat{\beta})$, is

$$\begin{aligned} b(\hat{\omega}) =& \exp(\hat{\beta})b(\hat{\beta}) + \frac{1}{2}\exp(\hat{\beta})\,\mathrm{Var}(\hat{\beta}) \\ =& -\frac{1}{2n(\exp(\hat{\beta}))}\exp(\hat{\beta})\sigma^2 + \frac{1}{2}\exp(\hat{\beta})\frac{1}{n(\exp(\hat{\beta}))} \\ =& 0 \end{aligned}$$

as expected since $\exp(\hat{\beta}) = \bar{y}$. If we instead use the bias reduced estimate, $\exp(\tilde{\beta})$, the first order bias approximation is

$$b(\tilde{\omega}) = \frac{1}{2}\exp(\tilde{\beta})\frac{1}{n(\exp(\tilde{\beta}))}.$$

In this case, reducing the bias of $\hat{\beta}$ led to an increase in bias of the estimate of ω^*. However, it is important to note that this is not true in general, but rather a special case.

In most settings, both $b(\hat{\omega})$ and $b(\tilde{\omega})$ will be non-zero and it may not be obvious which estimate leads to a lower bias of the transformed estimate. In general, a bias reduction method on the transformed estimates is needed to reduce the first order bias. A one-step bias correction can be used to eliminate the first order bias for transformations of either $\hat{\beta}$ or $\tilde{\beta}$.

In the next section we outline a general method for reducing the bias of transformed estimates. We then argue that using bias-reduced parameter estimates is preferred as it simplifies the bias reduction of the transformed estimate.

4.2.1 Bias Reduction of Transformed Estimates

A simple corrective bias reduction approach can be used on transformed estimates. The bias reduced transformed estimate using $\hat{\beta}$ is

$$\hat{\omega}^{BR} = \hat{\omega} - b(\hat{\omega}). \tag{4.9}$$

Similarly, the bias reduced transformed estimate using $\tilde{\beta}$ is

$$\tilde{\omega}^{BR} = \tilde{\omega} - b(\tilde{\omega}). \tag{4.10}$$

While both $\hat{\beta}$ and $\tilde{\beta}$ can be used to obtain bias reduced estimates of ω^*, using $\tilde{\beta}$ has an inherent advantage. The primary advantage of transforming $\tilde{\beta}$ is that the bias reduction step for the

transformed predictor is simpler and can be implemented with only regression output.

We note that although $\hat{\omega}^{BR}$ and $\tilde{\omega}^{BR}$ both require a first order approximation of $\hat{\beta}$, there are advantages to using $\tilde{\omega}^{BR}$. While estimating $\hat{\beta}$ does not require a first order bias approximation, a first order bias approximation of $\hat{\beta}$ is needed to calculate $\hat{\omega}^{BR}$. In this case, while $\hat{\omega}^{BR}$ is approximately unbiased for ω^*, $\hat{\beta}$ is not unbiased for β^*. Furthermore, in many settings, calculating the first order bias approximation requires the complete data. In these scenarios, $\hat{\omega}^{BR}$ cannot be calculated using only regression estimates.

In contrast, estimation methods for $\tilde{\beta}$ require a first order bias approximation; however, no first order bias approximation $b(\beta^*)$ is needed when calculating $\tilde{\omega}^{BR}$. In this case, both $\tilde{\beta}$ and $\tilde{\omega}^{BR}$ are bias reduced estimates for β^* and ω^* respectively. Furthermore, calculating $\tilde{\omega}^{BR}$ does not require the complete data and can be calculated using standard regression output.

We also note that for a linear transformation, $\hat{b}(\omega) \neq 0$ but $\tilde{b}(\omega) = 0$. Thus, for linear transformations, using a bias reduced model estimate ensures there is no first order bias of the transformed estimate and no correction is needed.

4.2.2 Simulations

Simulations are used to assess the impact of finite sample bias and bias reduction methods on a transformed predictor.

Scenarios

Let Y be an n-length vector of independent outcomes and X be an n-length vector of binary predictors such that

$$\mathrm{E}[Y|X] = \exp(\beta_0 + X\beta_1). \tag{4.11}$$

Standard Poisson MLE estimates, $\hat{\beta}$, and RBR estimates, $\tilde{\beta}$, are used to calculate the transformed estimates, $\hat{\omega} = \exp(\hat{\beta})$ and $\tilde{\omega} = \exp(\tilde{\beta})$, respectively. Additionally, we fit the bias reduced

estimate, $\tilde{\omega}^{BR}$. We simulate data for two sets of parameter values. In each scenario, 5,000 repetitions were used.

Scenario 1

In this scenario, $\beta_0 = \log(.7)$ and $\beta_1 = \log(3/7)$. These parameter values correspond to probabilities of events of 70% in the group with $X = 0$ and 30% in the group with $X = 1$. In this scenario, the bias of $\hat{\beta}$ is negative resulting in $b(\hat{\omega})$ being smaller than $b(\tilde{\omega})$.

Scenario 2

In this scenario, $\beta_0 = \log(.3)$ and $\beta_1 = \log(7/3)$ and the probability of a success in the two groups are switched compared to scenario 1. This changes the sign of the bias and consequently, $b(\hat{\omega})$ is larger than $b(\tilde{\omega})$. This scenario is meant to contrast scenario 1 to demonstrate how predicting the effects the bias of model estimates has on transformed estimates is difficult.

Results

Figure 4.1 displays the results for scenario 1. In this scenario, the transformed RBR estimate has the largest bias. The transformed standard estimate and corrected transformed RBR estimate have similar performance.

Figure 4.2 shows the results for scenario 2. In this scenario we see the bias from the standard model estimate is highest while the corrected transformed RBR estimate has the best performance. In particular, we see the bias of the transformed RBR estimate and the corrected transformed RBR estimate are similar between the two graphs. Thus the bias of transformed RBR estimates is robust to parameterization. However, the bias of the transformed standard estimate is not robust to parameterization as can be seen by the different bias between the two scenarios.

From these simulations, the impact of the choice of estimate of β has on bias is clear, with bias-reduced estimates of β leading to higher bias in some settings but lower bias in other settings. However, discerning which coefficient estimate leads to a lower transformed bias may not

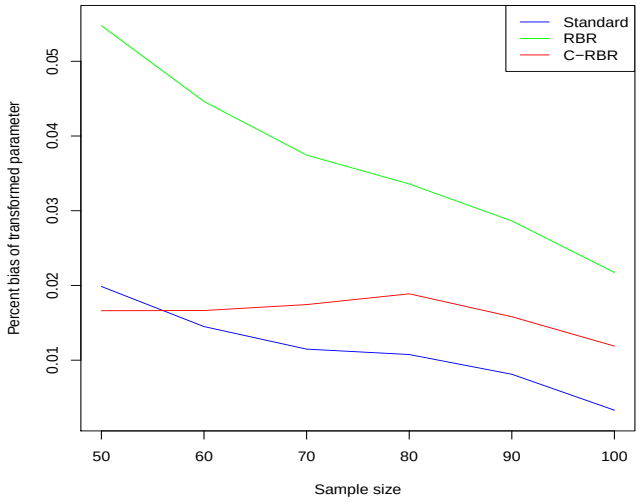


Figure 4.1: Percent bias for $\exp(\beta)$ for standard and RBR estimates of of $\hat{\beta}$ as well as a one-step corrected estimate of $\exp(\beta)$ (C-RBR) for a parameter with negative bias

be straightforward in practice. Therefore, choosing the coefficient estimation method with the aim of reducing the bias of the transformed estimate is not feasible. Instead, a bias reduction method can be used to reliably reduce the bias of the transformed estimate.

4.2.3 Discussion

The use of a one-step correction on the transformed bias reduced estimate effectively reduces the bias of the transformed estimate. Though a similar correction can be used on transformed standard coefficient estimates, using bias reduced estimates is preferred. Using bias reduced estimates can simplify the post-transformation correction process. Furthermore, a bias approximation is needed to ensure a reduced bias of the transformed estimate. This bias approximation can be used when estimating β^* or during the post-transformation step. However, by applying it when estimating β^*, bias reduction occurs in both model estimates and transformed estimates without adding any substantial complexity to the overall estimation procedure.

Future work is needed to provide a more in depth empirical analysis of bias in transformed estimates; however, this work demonstrates that reducing the bias of model estimates may not

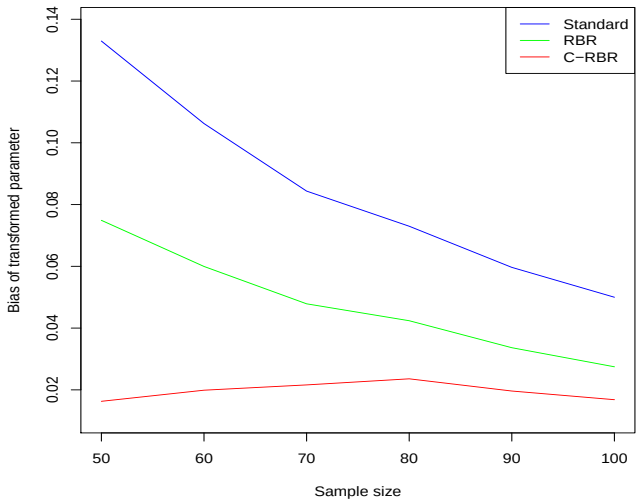


Figure 4.2: Percent bias for $\exp(\beta)$ for standard and RBR estimates of of $\hat{\beta}$ as well as a one-step corrected estimate of $\exp(\beta)$ (C-RBR) for a parameter with positive bias

reduce the bias of transformed estimates.

One other area of future work is to derive a preventative bias reduction method for transformed estimates. In this scenario, a modified score function based on the desired transformation can be solved. The concept is straightforward as the bias approximation used in the RBR method would need to be replaced by an expression that may introduce bias into the coefficient estimate but would reduce the bias of the transformed estimate. However, the expression needed to replace the bias approximation is not immediately clear and requires further investigation.

4.3 Finite Sample Bias and Bias Reduction in Meta-Analysis

4.3.1 Introduction

The previous two chapters discussed the bias and bias reduction methods for individual studies. In this section we explore the impact finite sample bias and bias reduction have in meta-analysis, where multiple study estimates are pooled together. In particular, we show pooling estimates can inflate the bias/SE ratio leading to a higher type II error. Furthermore, using bias reduction methods on individual studies is not always effective at reducing the bias of the pooled estimate.

4.3.2 Background

Systematic reviews can synthesize evidence from a variety of studies in order to make more informed clinical or public health decisions. Meta-analysis provides a formal statistical framework to combine estimates from multiple studies into one pooled estimate. Fixed effect and random effects models are two frameworks used to pool study estimates.

Fixed effects models assume the treatment effect is homogeneous across all studies being pooled. In general, this assumes the study populations, exposure, and other factors are the same in all studies. In this setting, the only variation between study estimates is the random sampling variability. In these models, inference can be made on a common effect, as the effect is assumed to be the same in all studies. Lin and Zheng demonstrated that under certain assumptions, pooled estimates have the same efficiency as if the raw data were analyzed including study as a fixed effect [47].

In contrast, random effects models do not assume a homogeneous effect. Random effects models allow for heterogeneity in the treatment effect between studies, which can occur due to differences in study populations, exposures, or health access as well as many other factors. In this setting, there is variability in the treatment effect as well as sampling variability. In many scenarios, studies may not be homogeneous and random effects models are necessary. However, results from

random-effects models are frequently misinterpreted as a common (homogeneous) effect rather than an average treatment effect across heterogeneous populations.

When the appropriate assumptions are met for fixed or random effects estimates, these pooled estimates lead to improved efficiency and inference. However, bias can be present in the pooled estimate, leading to more precise estimates (lower standard error) but not necessarily more accurate estimates (lower bias). This bias can arise from a variety of causes. We break the discussion of bias into three general categories: selection bias, design bias, and estimation bias.

Selection bias is the bias that occurs due to the inclusion or exclusion of specific studies. This bias occurs when the estimates from selected studies do not reflect the true underlying distribution of estimated effects. Selection bias has many different causes.

Publication bias is one cause of selection bias discussed extensively in the literature [6, 61, 69]. Publication bias occurs when only select studies (generally those with positive findings) get published while others (generally those without significant findings) do not. As meta-analyses are based on combining results of published studies, if studies without significant results are not published, pooled estimates will be overly optimistic.

Other less discussed causes of selection bias are language bias, database bias, and citation bias which all can impact whether study results are identified to be included in the meta-analysis [22]. Furthermore, even if there is no bias present in the available studies, the inclusion criteria used by investigators can lead to bias.

Funnel plots can be a useful tool to help identify selection bias. A funnel plot is a scatter plot of study estimates against sample sizes. Larger studies generally have greater precision and make up the 'tip' of the funnel, while smaller studies with less precision make up the 'base'. In an ideal setting, the pattern in the scatter plot is funnel shaped. Deviations from this pattern may indicate selection bias is present. Frequently this occurs with a gap in the funnel corresponding to small studies that did not have significant findings. More objective statistical methods are also available to test for selection bias [5, 24, 48]. The performance of these methods under various conditions have been assessed under different scenarios [48, 60].

The second type of bias is a result of low quality design of the individual studies. Poor quality study design (such as improper blinding of investigators) can contribute to bias in individual studies and overly optimistic results. In particular, it has been shown that studies with smaller sample sizes tend to have lower quality of study design, which can lead to larger estimated effects [23, 66]. Funnel plots, and other methods to detect selection bias, may still be effective at detecting this bias as systematic patterns may still arise, particularly the lack of small sample studies with small estimated effects.

The third type of bias is estimation bias and is the inherent bias attributed to the method used for estimation. This bias can be introduced during the process of pooling estimates or be a result of the estimation bias in individual study estimates.

The method for pooling estimates can introduce bias even when study estimates are unbiased. Inverse-variance weighting is one common way to pool study estimates which can introduce bias. Because the true variance of estimates is unknown, estimated variances are commonly used for weights. However, in many settings, the estimated variance is correlated with the parameter estimate which can lead to a bias in the pooled estimate even when study estimates are unbiased [31, 33]. When the weights are not correlated to parameter estimates, no bias is introduced through the pooling process. Unit weights, and in many scenarios sample size weights, are not correlated with the parameter estimate.

Regardless of the weights used, pooled estimates may be biased if study estimates are biased. Biased study estimates can be the result of finite sample modeling bias as discussed in Chapter 2. Many maximum likelihood estimates for generalized linear models (GLMs) have finite sample bias. This is problematic as many treatment effects for binary outcomes, such as relative risks (RR) and odds ratios (OR), use GLMs. Logistic regression and Poisson regression are two of the more commonly used GLMs and estimate log-OR and log-RR respectively.

In their discussion, Nemes et al acknowledge that finite sample bias could be problematic in meta-analysis; however, no known work has been done to assess the impact finite sample modeling bias may have in meta-analysis [56]. If the bias is not reduced through pooling estimates, but the

standard error is, the resulting pooled estimate will have a large bias/SE ratio, potentially leading to high type I error rates.

Furthermore, there is evidence that meta-analyses with small studies are common. Turner *et al* provide a review of nearly 15,000 meta-analyses in Cochrane reviews to study the effects underpowered studies have on meta-analysis. In 70% of the meta-analyses, all studies included were underpowered to detect a RR of 0.7 for a binary predictor and only 17% included two or more adequately powered studies. While the concept of power is not directly related to bias, the frequency of underpowered studies indicates that small sample studies (which have higher finite sample bias) are common in meta-analysis. These small studies are susceptible to larger finite sample bias.

In Chapter 3 we described different bias reduction approaches for correctly specified and misspecified models. These methods can effectively remove the first order bias of individual study estimates. However, the effect bias reduction has on pooled estimates has not been assessed. Intuitively, the expectation is the pooled bias will also decrease; however, as we will demonstrate later, this is not always the case.

In this section, we assess the impact of finite sample modeling bias on fixed-effects meta-analysis using three different weighting schemes for pooling estimates: inverse variance weighting, sample size weighting, and unit weighting. In Section 4.3.3 we consider the effects bias and bias reduction of individual studies have on the pooled estimates. In particular we demonstrate that pooled estimates using inverse-variance weights are particularly vulnerable to bias. Bias reduction on individual studies is effective at reducing the bias of estimates using sample size and unit weights, as well as special cases of inverse-variance weighting. However, reducing the bias of individual studies does not always reduce the bias of pooled estimates using inverse-variance weights and in some scenarios can lead to an increase in bias. Section 4.3.4 contains results empirically assessing the impact of bias and bias reduction on pooled estimates through simulations.

4.3.3 Estimates from Meta-Analysis

In this section we explicitly discuss pooling estimates for the fixed-effects models. Although the general concepts extend to random-effects models as well, further work is needed to explicitly engage with the impact finite sample bias and bias reduction method have when there is not a common treatment effect.

Pooled Estimates

The process of synthesizing study estimates into a pooled estimate is commonly done using a weighted average of individual study estimates. Suppose for k studies, the estimated parameter for the i^{th} study is $\hat{\beta}_i$ with a given weight W_i. The general form of the weighted average is

$$\hat{\beta}^W = \frac{\sum_{i=1}^{k} W_i \hat{\beta}_i}{\sum_{i=1}^{k} W_i}. \tag{4.12}$$

Different methods are used to assign weights to studies. We discuss three types of weights: 1) inverse-variance weights, 2) sample size weights, and 3) unit weights. Estimates using these weights are presented below.

The first method we discuss is inverse-variance weights as described by Hedges and Vevea [34]. This method explicitly gives studies with greater precision (lower variance) more weight. As the true variance of $\hat{\beta}_i$ is rarely known, the variance estimate, V_i, is often used for the weight. Then

$$W_i = \frac{1}{V_i}$$

and equation (4.12) becomes

$$\hat{\beta}^{IV} = \frac{\sum_{i=1}^{k} \frac{\hat{\beta}_i}{V_i}}{\sum_{i=1}^{k} \frac{1}{V_i}}.$$

The estimated variance of $\hat{\beta}^{IV}$ is

$$\widehat{\text{Var}}(\hat{\beta}^{IV}) = \frac{1}{\sum_{i=1}^{k} \frac{1}{V_i}}.$$

Hunter and Schmidt use sample size weights as an alternative to inverse variance weighting [38]. This method has philosophical similarities to inverse-variance weights. Both methods give greater weights to studies that are considered to have 'better' estimates. Inverse-variance weights give more weight to studies with greater precision (smaller standard error) and sample size weighting gives more weight to studies with larger sample sizes (which is often an indicator of greater precision). Formally, the weights are

$$W_i = n_i$$

where n_i is the sample size of study i. The pooled estimate from equation (4.12) then becomes

$$\hat{\beta}^{SS} = \frac{\sum_{i=1}^{k} n_i \hat{\beta}_i}{\sum_{i=1}^{k} n_i}$$

and the estimated variance of $\hat{\beta}^{SS}$ is

$$\widehat{\text{Var}}(\hat{\beta}^{SS}) = \frac{\sum_{i=1}^{k} n_i (\hat{\beta}_i - \hat{\beta}^{SS})^2}{k \sum_{i=1}^{k} n_i}.$$

In the setting where $V_i = V/n_i$ for some common variance V across all studies, inverse-variance weights and sample size weights will give identical pooled estimates.

An alternative to the two weighting methods above, which both give weights based on a measure of precision, unit weighting gives a common weight to all studies. When unit weights are used, the unweighted average of study estimates is the pooled estimate:

$$\hat{\beta}^U = \frac{\sum_{i=1}^{k} \hat{\beta}_i}{k}. \tag{4.13}$$

The estimated variance for $\hat{\beta}^U$ is

$$\widehat{\text{Var}}(\hat{\beta}^U) = \frac{\sum_{i=1}^{k} (\hat{\beta}_i - \hat{\beta}^{SS})^2}{k^2} \tag{4.14}$$

If study sample sizes are equal, sample size and unit weights are equivalent.

These different methods of weights are compared extensively in the literature [8,50,62]. Additional work has also analyzed the effects of combining multiple weighting schemes [7]. However, current literature does not address the impact weights have on the perpetuation of finite sample estimation bias in pooled estimates.

We note that other methods for estimating the variance of the pooled estimates are available and may improve inference [38,43]. However, we do not engage with these methods as our focus is on the bias of the pooled estimate.

Finite Sample Modeling Bias

Before deriving the bias of the pooled estimates we briefly review finite sample modeling bias of individual studies.

MLEs are asymptotically consistent and normally distributed, making them ideal for large sample inference. However, for finite sample sizes, MLEs may be biased. In GLM settings, this occurs when a non-linear link function is used. This is problematic as logit and log link functions are frequently used to estimate log-OR and log-RR for binary outcomes. The bias can be expressed as $bias = \frac{B}{n} + o(n^{-1})$. Cox and Snell provide an expression for the first order bias, $\frac{b}{n}$, which is a function of the sample size, n, and some constant, b dependent on the model. This bias was discussed in detail in Chapter 2. To discuss the bias of pooled estimates, we use an alternative representation of bias. Let $b_i = \hat{\beta}_i - \beta^*$ denote the deviation of $\hat{\beta}_i$ from the true parameter β^*.

Then $\mathrm{E}[b_i] = \mathrm{E}[\hat{\beta}_i - \beta^*]$ is the bias. This representation of bias is useful when calculating the bias of the pooled estimates.

Pooled Estimate Bias

We now provide the bias of the pooled estimates for the three different weighting methods above. These calculations assume a common treatment effect, β^*. This assumption means the first order bias for each study can be represented as a function of the same constant, B, and the sample size of the study. That is $\mathrm{E}[b_i] = \frac{B}{n_i} + o(n_i^{-1})$ for all i. Complete details of the derivations of the pooled estimate biases are available in the Appendix A.6. The bias of the pooled estimates for inverse-variance, sample size, and unit weights are as follows

$$\mathrm{E}[\hat{\beta}^{IV} - \beta^*] = \mathrm{E}\left[\frac{\sum_{i=1}^k \frac{b_i}{V_i}}{\sum_{i=1}^k \frac{1}{V_i}}\right].$$

$$\begin{aligned}\mathrm{E}[\hat{\beta}^{SS} - \beta^*] &= \mathrm{E}\left[\frac{\sum_{i=1}^k n_i b_i}{\sum_{i=1}^k n_i}\right] \\ &\approx \frac{kB}{\sum_{i=1}^k n_i}.\end{aligned}$$

$$\begin{aligned}\mathrm{E}[\hat{\beta}^{U} - \beta^*] &= \mathrm{E}\left[\frac{\sum_{i=1}^k b_i}{k}\right] \\ &\approx \frac{B\sum_{i=1}^k \frac{1}{n_i}}{k}.\end{aligned}$$

The approximations for the sample size weights and unit weights are a result of the last expression being the first order approximation of the bias. The expression of bias for inverse-variance weights cannot be further simplified due to the potential correlation between parameter and variance estimates.

From these expressions of bias, we note that the bias using sample size or unit weights are

both scaled versions of B. Unit weights are particularly susceptible to a small sample bias as small studies are given the same weight as large studies. When using inverse-variance weights, the pooled estimate bias is no longer proportional to B, but rather depends on $\mathrm{E}[\frac{b_i}{V_i}]$, the expected value of the weighted deviations. If V_i only depends on the sample size, then inverse-variance weights are a scaled version of sample size weights and the bias will be identical. However, in general $\mathrm{E}[\frac{b_i}{V_i}] \neq 0$ even when $\mathrm{E}[b_i] = 0$ [33].

The presence of bias indicates that pooling estimates may not reduce the bias regardless of the weights used. In some cases, pooling estimates can even introduce bias when inverse-variance weights are used. This can be particularly problematic as pooling estimates improves the precision. In all three weighting methods, the variance of the pooled estimate decreases as more study estimates are pooled. However, if the bias is unchanged, this leads to a high bias/SE ratio which may be problematic in hypothesis testing. This can be easily demonstrated in settings where the studies have the same sample size and variance estimates, V (and thus all three weighting methods are identical). In this case the bias of the pooled estimate is equivalent to the bias of the individual studies. However, the standard error of the pooled estimate is $\sqrt{\frac{V}{k}}$ where V is the common variance of the individual studies. Then the bias/SE ratio is

$$\frac{\sqrt{(k)}B}{V}. \tag{4.15}$$

In this special case, the bias/SE is proportional to the square root of the number of samples pooled. While this exact relationship is only true in special settings, in many settings, pooling more studies can increase the bias/SE ratio. We explore this in more realistic settings empirically in Section 4.3.4.

Impact of Bias Reduction of Study Estimates

With the finite sample modeling bias of pooled estimates quantified, we now discuss the impact bias reduction methods on individual studies may have on the pooled estimates. Let $\tilde{\beta}_i$ be the bias

reduced estimate for study i such that $\tilde{\beta}_i = \hat{\beta} - \frac{B}{n_i}$. Then $\mathrm{E}[\tilde{\beta}_i] = \beta^* + \mathrm{E}[b_i] - \frac{B}{n_i} = \beta^* + o(n_i^{-1})$.

Using this estimate in place of $\hat{\beta}$ in the pooled estimates gives the following expressions for the first order bias for inverse-variance weights

$$\mathrm{E}[\tilde{\beta}^{IV} - \beta^*] = \mathrm{E}\left[\frac{\sum_{i=1}^{k} \frac{n_i b_i - B}{n_i V_i}}{\sum_{i=1}^{k} \frac{1}{V_i}}\right]. \tag{4.16}$$

In general, since V_i is a function of $\tilde{\beta}$ (and thus b_i), $\mathrm{E}[\frac{b_i}{V_i}] \neq B\mathrm{E}[\frac{1}{V_i}]$, and the first order bias is not reduced. Furthermore, in some settings $\mathrm{E}[\frac{b_i}{V_i}]$ and $\mathrm{E}[\frac{B}{V_i}]$ are of opposite sign. In this setting, bias reduction on individual studies leads to an increase in the bias of the pooled estimate.

As sample size and unit weights are not functions of the parameter estimate, the weighted averages have $\mathrm{E}[b_i] - \frac{B}{n_i} = 0$ in the numerator and the first order bias is 0 when bias reduced estimates are pooled. This is also true for the inverse variance weight when $V_i = \frac{V}{n_i}$ for some constant V. In these settings, bias reduction methods on individual studies can provide improvement in the coverage probability of confidence intervals built around the pooled estimate.

In the next section we assess the impact finite sample modeling bias and bias reduction of study estimates has on the bias, bias/SE ratio, and coverage of pooled estimates through simulations.

4.3.4 Results

We begin by exploring three scenarios to better understand the impact finite sample modeling bias and bias reduction methods can have on pooled estimates. For all simulation scenarios, 5,000 replications were used.

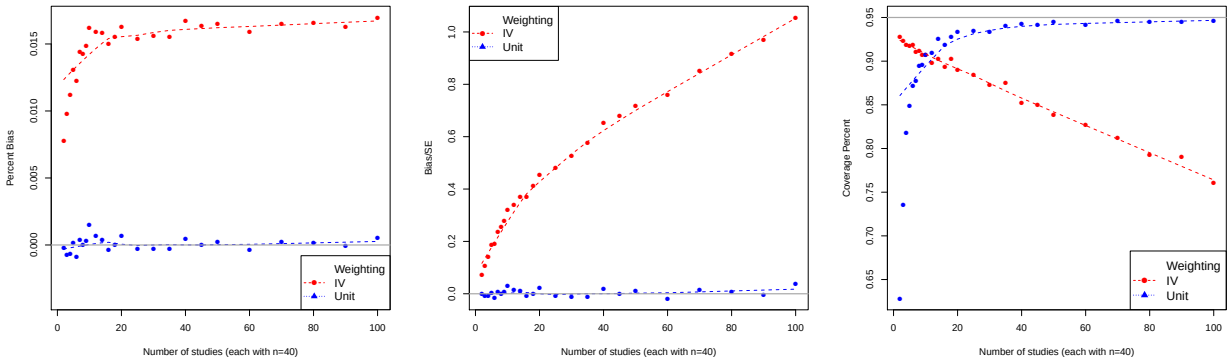
Scenario 1

The first scenario illustrates the phenomenon of inverse-variance weights introducing bias to the pooled estimate despite individual study estimates being unbiased. In this scenario, let

$$\mathrm{E}[y|x] = \beta_0 + \beta_1 x$$

where y a vector of independent binomial outcomes, x is a balanced binary predictor, $\beta_0 = .3$, and $\beta_1 = .3$. Since the outcomes are binomial, parameter estimates and variance estimates are correlated. In this scenario, the RD is being estimated using a linear link function, which means the finite sampling bias for individual studies is 0. The number of studies pooled ranged from 2 to 100 and each study had a sample size of 40. We present the bias, bias/SE, and coverage (using a nominal 95% confidence interval) of pooled estimates using inverse-variance and unit weights (which are equivalent to sample size weights in this scenario). As no finite sample modeling bias is presented, no bias reduction estimates were used.

From figure 4.3, the bias is 0 when unit weights are used as expected. However, bias is present when inverse-variance weights are used. This bias increases slightly as more studies are pooled but is relatively constant when at least 25 studies are pooled. This trend is mirrored when considering the bias/SE ratio; however, the bias/SE ratio for inverse-variance weights continues to increase as the number of studies pooled increases. Despite the bias when using inverse variance weights, the coverage for a nominal 95% confidence interval is closer to the nominal value than unit weights when the number of studies pooled is small due to the poor standard error estimates when using unit weights. Under unit weights, the true coverage approaches the nominal 95% value when a large number of studies are pooled, as standard error estimates improve. Coverage for inverse-variance weights gets worse as more studies are pooled due to lower standard errors without a reduction in bias.



(a) Bias (b) Bias/SE (c) Coverage

Figure 4.3: Bias, bias/SE and coverage probability for inverse-variance and sample size weights of estimated risk differences for studies with a sample size of 40

Scenario 2

In the second scenario, the mean model is

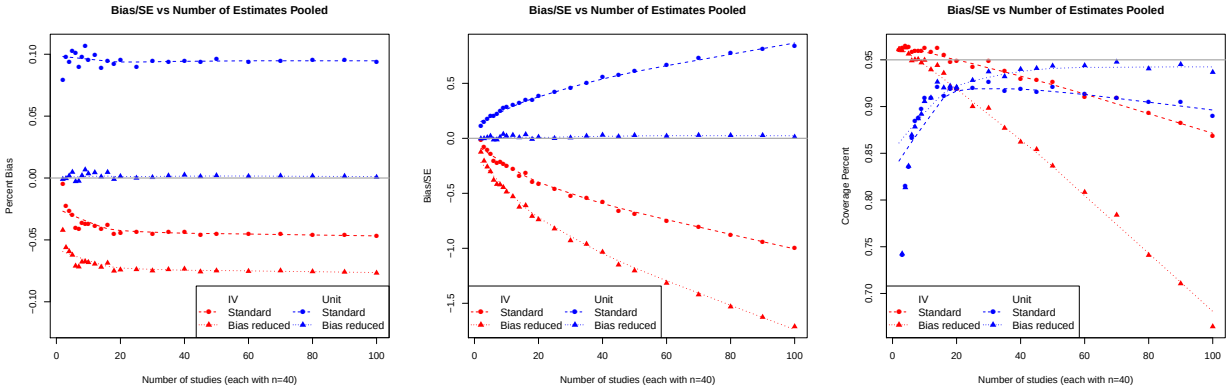
$$\text{logit}(\text{E}[y|x]) = \beta_0 + \beta_1 x$$

where y is a vector of independent binomial outcomes, x is a balanced binary predictor, $\beta_0 = \log(.3)$, and $\beta_1 = \log(2)$. The log-OR is the target of inference in these scenarios and is estimated using the logit link function. As this is a non-linear link function, finite sample modeling bias is present.

In this scenario, we explore how the number of studies pooled affects the bias. The number of studies pooled ranges from 2 to 100, each with a sample size of 40. In this scenario, finite sample modeling bias is present and standard MLE estimates and bias reduced estimates were used to calculate the pooled estimate to assess the impact bias reduction has on the pooled estimate. We present the bias, bias/SE, and coverage (using a nominal 95% confidence interval) of pooled estimates using inverse-variance and unit weights (which are equivalent to sample size weights in this scenario).

Figure 4.4 shows the results from this scenario. Bias is present for both weighting methods

when standard estimates are used. However, the bias when using inverse-variance weights has the opposite sign of the bias from unit weights, indicating the pooled estimate has bias of the opposite sign as individual studies. When using bias-reduced estimates, the pooled estimate effectively has the bias reduced when unit weights are used. When using inverse-variance weights, the bias increases when bias-reduced estimates are used. The bias/SE has similar trends but continues to increase as the number of studies pooled increases with the exception of unit weighting with bias-reduced estimates.



(a) Bias (b) Bias/SE (c) Coverage

Figure 4.4: Bias, bias/SE and coverage probability for inverse-variance and sample size weights of estimated log odds ratios for studies with a sample size of 40

The number of studies pooled also has an effect on coverage. When using inverse-variance weights the coverage gets worse as more studies are pooled regardless of the study estimates used. However, bias reduced estimates lead to a faster drop in coverage compared to standard estimates.

The coverage is poor for unit weighting when the number of studies pooled is small due to the variance estimate underestimating the variance. As the number of studies pooled increases, the estimated variance approaches the true variance. When bias reduced estimates are pooled, the improvement in variance estimation leads to a coverage probability close to the nominal 95% value. When standard estimates are pooled, an increase in the number of studies temporarily leads to improved coverage as the variance estimate is more accurate. However, after approximately 30 studies, the coverage begins to decrease as more studies are pooled. While the variance estimate

is more accurate (and decreasing as the number of studies pooled increases), the bias has a more substantial impact, leading to worse coverage.

Scenario 3

In the third scenario, we also use the mean model

$$\text{logit}(\text{E}[y|x]) = \beta_0 + \beta_1 x$$

In this scenario, we explore the impact one study can have on the pooled estimates. The first case consists of 20 studies each with a sample size of 500; the second case is similar to case 1 with addition of one study with sample size 50; the third case has 20 studies each with sample size 50; the fourth case has 20 studies with sample size 50 and one with sample size 500. These cases are used to assess the impact one small or one large study has on the pooled estimates. The bias, bias/SE, and coverage (using a nominal 95% confidence interval) are presented for pooled estimates using inverse-variance, sample size, and unit weights. As finite sample bias is present, pooled estimates for all weights were calculated using standard estimates and bias reduced estimates.

The complete results are found in table 4.1. As expected, the meta-analysis with 20 studies of sample size 500 had lower bias than studies with sample size of 50, demonstrating pooling studies with larger sample sizes lowers the pooled bias. However, pooled estimates using unit weights were more susceptible to bias if even one study had a small sample size. By adding one study with a sample size of 50 to 20 studies with sample size 500, the bias nearly doubled (0.44% to 0.84%). While the bias was still relatively small, estimates using inverse-variance and sample size weights were more resistant to a single small study, as more weight was given to the larger studies.

A similar phenomenon was seen when adding one large study ($n = 500$) to twenty smaller studies ($n = 50$). Estimates using inverse-variance and, in particular, sample size weights had a notable reduction in bias. However, estimates using unit weights only had a minor improvement in bias. Similar to above, this can be attributed to the extra weight inverse-variance and sample

Table 4.1: Bias, Bias/SE, and coverage for inverse-variance, sample size, and unit weights using standard estimates or bias reduced estimates for four different cases: 1) 20 studies each with n=1,000 2) 20 studies with n=1,000, 1 study with n=40 3) 20 studies with n=40 4) 20 studies with n=40, 1 study with n=1000.

		Inverse-variance		Unit		Sample Size	
		Standard	Bias Reduced	Standard	Bias Reduced	Standard	Bias Reduced
Bias	Case 1	-0.48	-0.95	0.44	-0.05	0.44	-0.05
	Case 2	-0.46	-0.95	0.84	0.05	0.52	-0.05
	Case 3	-4.86	-8.65	7.14	-0.03	7.14	-0.03
	Case 4	-3.37	-6.11	6.83	0.23	4.81	0.04
Bias/SE	Case 1	-0.11	-0.21	0.10	-0.01	0.10	-0.01
	Case 2	-0.10	-0.21	0.14	0.01	0.11	-0.00
	Case 3	-0.35	-0.65	0.32	-0.00	0.32	-0.00
	Case 4	-0.30	-0.56	0.34	0.02	0.32	0.00
Coverage	Case 1	95.2	90.8	92.7	93.9	92.7	93.9
	Case 2	94.5	94.1	94.6	94.3	92.9	92.8
	Case 3	95.4	94.9	93.0	93.1	93.0	93.1
	Case 4	95.0	91.9	91.8	92.7	91.8	92.7

size weights assign to larger, less biased studies. In all cases, using bias reduced estimates nearly eliminated the bias when using unit and sample size weights, but introduced additional bias in estimates using inverse-variance weights.

Summary

These results support the theoretical findings previously discussed. The bias of individual studies is not a good indicator of the bias for pooled estimates using inverse-variance weights. When inverse-variance weights are used, the lack of bias in individual studies does not guarantee an unbiased pooled estimate. Furthermore, the sign of the bias can also differ between individual study estimates and the bias of the pooled estimate. This also leads to inconsistent results when using bias reduced estimates of individual studies. In some settings the bias reduced estimates lead to a minimal reduction in bias and in other settings can lead to an increase in bias.

For sample size and unit weights, the bias of individual studies is a good indicator of the bias of the pooled estimate. Pooled estimates using unit weights are more vulnerable to small sample bias as they give equal weights to studies with small sample sizes, which have higher finite sample

modeling bias. When individual study estimates are unbiased, pooled estimates using sample size or unit weights are also unbiased. Therefore, bias reduction methods on individual studies effectively reduce the bias of pooled estimates when sample size and unit weights are used.

Similar trends were seen for bias/SE, although bias/SE is more dependent on the number of studies pooled. The bias/SE can be seen to increase as the number of studies pooled increases even in settings where the bias remains constant. Bias reduction methods that effectively reduced the bias were also effective at reducing bias/SE. Similarly, when bias reduction methods were ineffective at reducing bias, they were also ineffective at reducing bias/SE.

The coverage was poor for sample size and unit weights when only a few studies were pooled and can be attributed to poor standard error estimation. As the number of studies pooled increased, standard error estimates were more reliable, and coverage for unbiased estimates approached the nominal 95% level.

Data Example: Meta analysis

The next scenarios are motivated by two meta-analyses analyzing the effects of closed-incision negative-pressure therapy (ciNPT) systems on surgical site infections (SSI), one based on randomized clinical trials (RCT) and the other on observational data [65]. The meta-analysis of the RCTs pooled estimates from 11 studies with sample sizes ranging from 44 to 441. The aggregate rate of SSI was 19.8% in the control group and 8.5% in the ciNPT group. In the meta-analysis based on the observational studies, 19 study estimates were pooled with sample sizes ranging from 47 to 3,745 (though the study with 3,745 participants was heavily imbalanced with3 3,508 controls and 237 ciNPT participants). The aggregate rate of SSI was 5.6% in the control group and 2.9% in the ciNPT. Aggregate event rates were used as the 'true' event rate and to calculate the true log-OR for the RCT meta analysis. For the observational study meta analysis we used 4 times the aggregate rates (21.8% and 8.7% in the control and ciNPT groups respectively), as the given aggregate rates led to a high frequency of studies with 0 events, which leads to infinite estimates without additional modification.

Table 4.2: Bias, Bias/SE, and coverage based on two meta-analysis for RCT and observational studies in Singh et al. for inverse-variance, sample size, and unit weights using standard estimates or bias reduced estimates.

		RCT			Observational		
Weights	**Study Estimates**	**Bias**	**Bias/SE**	**Coverage**	**Bias**	**Bias/SE**	**Coverage**
Inverse-variance	Standard	-3.4	-0.22	94.9	-2.8	-0.27	94.3
	Bias Reduced	-7.0	-0.47	94.1	-5.7	-0.56	91.7
Sample	Standard	5.4	0.31	90.6	3.4	0.25	79.7
	Bias Reduced	-0.7	-0.04	89.6	-0.38	-0.03	77.7
Unit	Standard	6.4	0.30	100.0	7.6	0.44	100.0
	Bias Reduced	-1.8	-0.10	90.5	-2.1	-0.14	91.9

To replicate these studies, the sample sizes from each study and aggregate rates from each meta-analysis were used to randomly generated SSI events. Individual studies were then analyzed using standard regression methods and bias reduction methods for OR. Pooled estimates and standard errors were then calculated using inverse-variance, sample size, and unit weights. Bias, Bias/SE and coverage probabilities were calculated and compared across weighting methods.

In both meta-analyses, noticeable bias was present across all studies when using standard study estimates. When using bias-reduced estimates, the bias for sample size and unit weights decreased, while magnitude of bias increased when inverse-variance weights were used (as seen in table 4.2). Inverse variance weights led to the lowest bias when standard estimates were used while sample size weights led to the smallest bias when bias reduced estimates were used. Inverse-variance weights using standard study estimates had the best coverage across all models, indicating the bias had minimal impact on the coverage probability in these settings.

Additionally we compared the estimates using inverse-variance weighting to the estimate from the single largest study and the estimate from a theoretical, single study with sample size equal to the aggregate sample size across all pooled studies.

Table 4.3 shows the results comparing the bias from a pooled estimate using inverse-variance weights, a single theoretical study with sample size equal to the aggregate sample size of pooled studies, and the single largest study in the meta-analysis. Across all metrics, the single, theoretical study performed had the lowest bias of all estimates. Interestingly, the pooled estimates had the

Table 4.3: Bias, bias percent, and bias/SE for the pooled estimate using inverse-variance weights, the estimate from a theoretical, single study with the same aggregate sample as the meta-analysis, and the estimate from the largest study in the meta-analysis.

Study	Method	Absolute Bias	Percent Bias	Bias/SE
Singh *et al*: RCT	Pooled Estimates	-0.032	-3.3	-0.22
	Single Study	0.004	0.4	0.02
	Largest Sample	0.019	1.9	0.06

worst performance, with the largest magnitude of bias and bias/SE across all studies. In particular, the bias/SE was over three times as large as the other two estimates.

4.3.5 Discussion

The presence of finite sample modeling bias is problematic in meta-analyses. Pooled estimates using sample size weights were the most resistant to finite sample modeling bias. The presence of one or two studies with large samples effectively controlled the bias of the pooled estimates. Pooled estimates using unit weights were susceptible to bias if any of the studies had a small sample. This distinction is intuitive as large sample studies (which tend to have smaller bias) are given more weight than small sample studies (which tend to have higher bias) when sample size weights are used. When unit weights are used, both studies are given the same weight and finite sample bias remains prominent.

The inclusion of a large sample study also provided some protection against finite sample bias for inverse-variance weights. However, inverse-variance weights can lead to bias in the pooled estimate even when study estimates are unbiased if standard error and parameter estimates are correlated. Despite this phenomenon, the use of inverse-variance weights led to the best coverage when a small number of study estimates were pooled. This is likely because inverse-variance weights calculate the standard error based on study estimates of the standard error, while sample size and unit weights rely on the sample variance of study estimates, which relies on a larger number of studies for accurate standard error estimation.

While bias reduction methods are effective on individual studies, reducing the bias of individual

studies only led to a reduction in bias when sample size or unit weights were used. Bias reduction of individual studies was ineffective at reducing the pooled bias when inverse-variance weights were used and can lead to an increase in bias.

In some settings, the study estimates all share a common variance. In this scenario, one possible solution to reduce the bias pooled estimates using inverse variance weights is to use an estimate of the common variance in the weights. The first step is to calculate the pooled estimate using inverse-variance weights from study effects and standard error estimates. The pooled estimate can then be used to estimate common variance for all study estimates. Pooled estimates can be recalculated using study estimates for the effects but use the common variance estimate (weighted accordingly by sample size) as the weights. A similar suggestion has proven effective when pooling correlation estimates [38]. However, further work is needed to assess the performance of such a correction.

Regardless of the weights used, bias-reduced study estimates are needed to reduce the bias of pooled estimates. In practice, meta-analyses rarely have access to bias reduced estimates, as bias reduction methods are not widely used. This is problematic as, in most scenarios, bias reduced estimates cannot be calculated by the summary statistics frequently provided from studies; complete data from the analysis is needed. Because of this, bias reduction of individual study estimates is not a feasible solution to pooled sample bias in practice.

Non-parametric re-sampling methods may be one method for post-estimation correction of pooled estimates, using sub-sampling of the studies selected for analysis and is an area for future research.

Another limitation of our findings is that we assumed an ideal meta-analysis scenario with homogeneous populations across studies. While not explored in this section, the interaction between finite sample bias and other types of bias in meta-analysis could further contribute to the bias of pooled estimates. In more realistic settings, studies may differ on key elements such as design or target population. Furthermore, publication bias can also influence the study estimates that are available. These shortcomings further complicate the finite sample bias present in meta-analysis. Further research needs to be done to understand how finite sample bias interacts with these other

types of bias.

Chapter 5

Bias Reduction Methods in High Dimensional Data

In the previous chapters we characterized the bias of maximum likelihood estimates and developed a robust bias reduction method for low dimensional models. This chapter extends and evaluates bias reduction methods in high dimensional data settings. In particular, we show that the bias reduction methods from Chapter 3 are ineffective for sparsity debiased lasso estimates but are effective for some screen and refit models.

5.1 Introduction

With the increase in the scale of data collection, the presence of high dimensional data (where the number of features, p, is large relative to the sample size, n) is becoming more common in all areas of research. Estimates from classical regression methods, such as (quasi-)maximum likelihood estimation with generalized linear models, are not well defined in the high dimensional setting. This has led to the development of new estimation techniques that can be used for high dimensional data.

In this chapter we discuss regularized regression estimates, with a focus on the lasso. Regularization techniques introduce bias into parameter estimates by shrinking parameter estimates. In

many settings, these estimates result in better prediction, but inference using these estimates may no longer be valid. We discuss the lasso in more detail in the next section.

While high dimensional settings are usually defined as those wherein the number of features is greater than the sample size, $p > n$, classical regression methods may have a high type II error rate (particularly if there is a high correlation among features) and be subject to overfitting even when $p < n$ [32, 44, 67, 80]. While the methods in this chapter are presented for high dimensional settings, they also extend to low-dimensional settings where regularization or feature selection is used.

In Section 5.2, we provide a brief background on modeling with high dimensional data before engaging with the bias of lasso estimates. In particular, we distinguish between sparsity-induced bias, which results from regularization using the ℓ_1 penalty (and was not present in the low-dimensional results from Chapters 2 and 3), and curvature-induced bias which results from the curvature of the loss function (which has the same form of bias as GLM estimates discussed in previous chapters.) In Section 5.3, we discuss current methods to make inference with high dimensional data. These methods can be divided into two families: sub-models and sparsity debiasing methods. Sub-models use a low-dimensional subset of features to make inference. Sparsity debiasing methods use a post-estimation correction of lasso estimates to make valid inference. We describe methods for reducing the curvature-induced bias for a special case of sub-models and lasso estimates in Section 5.4. Section 5.5 contains simulations empirically assessing the proposed bias reduction methods from Section 5.4. Final conclusions and areas for future work are discussed in Section 5.6.

5.2 Background

We begin by discussing regression models in high dimensional data settings.

5.2.1 Regression Models

Let (y, x^0) be a set of random observations where y is a scalar and x^0 is a (possibly unknown) vector of features. Consider the true model

$$\mathrm{E}\left[y|x^0\right] = f\left(x^0\right) \tag{5.1}$$

for a (possibly unknown) function of the covariates $f(x^0)$. Rather than trying to directly estimate $f(x^0)$, we are often interested in the association between the outcome, y, and a given set of features, x (which may not be x^0). Minimizing a loss function is one technique to assess this association. In the case of squared-error loss, the minimizer of the loss function is precisely the mean of the outcome conditional on the features, x. This conditional mean is often modeled as a linear combination of the features. Other types of loss functions frequently used are those with the form of GLMs, which commonly link a linear combination of features to a (possibly non-linear) function of the mean outcome. When minimizing a loss function, the target parameter of interest is the population parameter that minimizes the loss function, which we formally define below.

For the random observation (y, x), where y is a univariate outcome and x is a $p \times 1$ vector of features, the target of inference is the p-length vector of parameters

$$\beta^* = \arg\min_{\beta} \mathrm{E}[l(y, x; \beta)] \tag{5.2}$$

for a given loss function $l(y, x; \beta)$. For example, the loss function for squared error loss is $l(y, x; \beta) = (y - x\beta)^2$.

Now consider the n independent and identically distributed realizations (y_i, x_i) for $i = 1, ..., n$. For estimation, we minimize an empirical loss function:

$$l_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, x_i; \beta). \tag{5.3}$$

In the low-dimensional setting, classical inference can be made using the estimate that minimizes $l_n(\beta)$:

$$\hat{\beta} = \arg\min_{\beta} l_n(\beta). \tag{5.4}$$

Under mild regularity conditions, $\hat{\beta}$ is consistent for β^*. This estimate is identical to the (quasi-) maximum likelihood estimates derived from a generalized linear model when the loss function is derived from a probability distribution in the exponential family.

However, in high dimensional settings, sensible estimates cannot be obtained by solving equation (5.4) and new methods are required to estimate β^*. In the last two decades, a growing body of research focuses on estimating β^* in the high dimensional setting. Regularization (or penalized regression) is one common method for obtaining estimates in high dimensional settings by adding a penalty to the loss function. Estimates have the form

$$\hat{\beta}^p = \arg\min_{\beta} l_n(\beta) + \lambda P(\beta) \tag{5.5}$$

for some tuning parameter, λ, and some penalty $P(\beta)$. The penalty is used to "shrink" parameter estimates which can help prevent over-fitting. Bridge regression is a family of models with $P(\beta) = ||\beta||_\gamma$ for $\gamma > 0$ [27]. The lasso and ridge regression are special cases of bridge regression when $\gamma = 1$ and $\gamma = 2$ respectively [37, 70]. The elastic net is a regularization method that uses a linear combination of ℓ_1 and ℓ_2 penalties [86].

The inclusion of the penalty term in all these methods introduces bias to the parameter estimates but can improve overall performance, particularly in settings where prediction is the focus. The tuning parameter λ controls how much shrinkage is induced; larger values induce greater shrinkage and $\lambda = 0$ induces no shrinkage, fitting the classical regression model in equation (5.4). When using the lasso estimate (discussed below), we refer to the bias induced by the penalty as sparsity-induced bias. In contrast the bias attributed to the curvature of the loss function (which was the focus of Chapters 2 and 3 in low-dimensional settings), is referred to as curvature-induced bias.

We discuss these biases in more depth later in this section.

In this chapter we focus on lasso estimates

$$\hat{\beta}^L = \arg\min_{\beta}(l_n(\beta) + \lambda||\beta||_1) \tag{5.6}$$

where $l_n(\beta)$ is a convex loss function and $||\beta||_1$ is the ℓ_1 norm. In addition to minimizing equation (5.6), $\hat{\beta}^L$ also satisfies the Karush-Kuhn-Tucker (KKT) condition

$$-\dot{l}_n(\hat{\beta}^L) + \lambda\hat{k} = 0 \tag{5.7}$$

for $||\hat{k}||_\infty \leqslant 1$ where $\hat{k}_j = sign(\hat{\beta}_j)$ for $\hat{\beta}_j \neq 0$ and $\dot{l_n}(\beta) = \frac{\partial l_n(\beta)}{\partial \beta}$.

One strength of the lasso over many other regularization methods is that it induces sparsity in the model, allowing for feature selection in a similar spirit to best subset regression. As an added benefit, the lasso is less computationally intensive than best subset regression and can therefore be used in high dimensional settings when best subset regression may not be feasible [28, 29].

5.2.2 Bias of Lasso Estimates

The focus of much of the high dimensional data literature is on consistency of feature selection, prediction, and oracle properties [11, 14, 18, 74, 77, 84]. However, more recent developments have also focused on making inference in high dimensional settings [40, 73, 73, 82]. This is particularly challenging due to the presence of bias that prevents a non-degenerate limiting distribution from being obtained.

While the bias in lasso estimates has been well documented, no known work has attempted to distinguish the model components that induce bias. By using a Taylor expansion, the bias of lasso estimates can be attributed to two components: the bias attributed to sparsity and the bias attributed to the curvature of the loss function. We note this is only relevant for non-least squares problems, as least squares problems have no curvature-induced bias.

Taking a Taylor expansion of $\dot{l}$ gives equation

$$\dot{l}_n(\beta^*) + \ddot{l}_n(\beta^*)(\hat{\beta}^L - \beta^*) + r_n + \lambda\hat{k} = 0. \tag{5.8}$$

Rearranging terms gives the expression

$$\ddot{l}_n(\beta^*)(\hat{\beta}^L - \beta^*) = -\lambda\hat{k} - \dot{l}_n(\beta^*) - r_n. \tag{5.9}$$

From this expression, $\lambda\hat{k}$ is the bias induced by regularization, $\dot{l}_n(\beta^*)$ is the term that gives asymptotic normality, and $r_n = o(n^{-1/2})$ (under the conditions discussed in Van de Geer [73]) is the higher order bias caused by the curvature of the loss function. In the case of squared error loss, $\dddot{l}_n(\beta) = 0$, and $r_n = 0$ (*i.e.* when the derivative of the loss function is linear, there is no higher order, curvature-induced bias.) The presence of bias is not problematic in prediction settings but leads to difficulty in making inference on parameters. Of particular concern for inference is the sparsity-induced bias which makes the limiting distribution intractable.

In the next section we discuss methods for making inference in high dimensional settings.

5.3 Inference with High Dimensional Data

Recent methods have been developed to make valid inference in high dimensional data settings. These methods can be be split into two general camps: methods making inference on sub-model parameters and methods making inference on full-model parameters. We provide a brief overview of both camps below before discussing each in more depth. A more detailed comparison of these methods can be found in Dezeure *et al* [19].

Sub-model parameter inference is a two step approach that first involves a screening process to select a subset of features. In the second step, the selected subset of features is used with standard estimation techniques. Sample splitting can be used to avoid over-optimism from using the same data to select features and fit a model with the selected features [76]. Meinshausen, Meier and

Buhlmann proposed a multi-sample splitting procedure to reduce the sensitivity of the process with respect to the choice of sample split [54].

Full-model parameter inference can be made using a one-step estimator with lasso estimates as the initial value. Zhang and Zhang, Van de Geer et al., and Javanmard and Montanri proposed similar methods that reduce the bias induced by regularization [40, 73, 82]. The corresponding estimates have a tractable limiting distribution and valid inference can be made asymptotically. Bootstrap methods have also been proposed as a tool for making inference in high dimensional data settings; however, we do not consider these methods in this dissertation [12, 13].

Before engaging with methods to reduce the curvature-induced bias for sub-model and full model estimates, we discuss sub-model and full-model inference in turn.

5.3.1 Sub-Models

Rather than making inference on parameters associated with the full set of features in equation (5.4), inference can be made on parameters associated with a (suitably small) subset of the features using a sub-model.

We will discuss two challenges that arise when using sub-models. The first is identifying the target of inference. Care must be taken here as parameters in the sub-model are not necessarily identical to a subset of parameters in the full model. The second challenge is making valid inference when a data adaptive approach is used to select the features used in the sub-model. We first identify the target parameter from a sub-model before discussing ways to make inference on this parameter.

Sub-Model Parameters

Consider the special case of the mean model in equation (5.1)

$$\mathrm{E}[y|x] = g^{-1}(x\beta^*). \tag{5.10}$$

Let $\mathscr{M} \subset \{1,...,p\}$ be an index set of size m. Let x_M be an m-length vector of features such that $x_M = \{x_j; j \in \mathscr{M}\}$ and $\beta^{M*} = \{\beta_j^*; j \in \mathscr{M}\}$. Similarly $x_C = \{x_j; j \notin \mathscr{M}\}$ and $\beta^{C*} = \{\beta_j; j \notin \mathscr{M}\}$ are $(p-m)$-length vectors.

Instead of using a model with the full set of features, x, suppose a sub-model is used to make inference on the subset of features, x_M. As a sub-model makes inference on different parameters than the full model, we need to determine what the target parameter is for a sub-model. As motivation, we first discuss sub-model parameters from a least squares regression model before considering a more general family of models.

Consider the mean model from equation (5.10) where g is the identity function

$$\mathrm{E}[y|x] = x\beta^*.$$

Then for a squared error loss function on the full set of parameters, the target of inference is

$$\beta^* = \arg\min_{\beta} \mathrm{E}[(y - x^T\beta)^2].$$

However, for a sub-model on the features x_M, the target of inference is

$$\beta_M^* = \arg\min_{\beta_M} \mathrm{E}[(y - x_M^T\beta_M)^2].$$

From these expression it is clear that β^* is a function of the full set of features x, while β_M^* is a function of the subset of features x_M. Note that in general, the parameters from the sub-model are not equal to the subset of parameters from the full model corresponding to x_m.

Additionally, we have the following alternative representation of the sub-model parameter

$$0 = \mathrm{E}[x_M(y - x_M^T\beta_M^*)]. \tag{5.11}$$

This representation is particularly useful for explicitly writing the relationship between full-

model and sub-model parameters. The following lemma relates the parameters from a sub-model, β_M^* to the parameters from the full model, β^*, under a squared error loss function.

Lemma 1. *Let x be a p-length vector of features, and y be a univariate outcome with mean* $\mathrm{E}[y|x] = x\beta^*$ *with* β^{M*} *denoting the subset of* β^* *corresponding to a subset of features,* x_M*. Then* β_M^**, the target of inference for the subset of features* x_M*, can be written as*

$$\beta_M^* = \beta^{M*} + \Sigma_{M,M}^{-1}\Sigma_{M,C}\beta^{C*}$$

where $\Sigma_{M,M} = \mathrm{E}[x_M x_M^T]$ *is the non-centered covariance of* x_M *and* $\Sigma_{M,C} = \mathrm{E}[x_M x_C^T]$ *is the non-centered covariance of* x_M *and* x_C*.*

See proof in Appendix A.7.

From Lemma 1, β_M^* is only equal to β^{M*} in the special case where $\Sigma_{M,C}\beta^{C*} = 0$.

A similar relationship between the sub-parameters and full model parameters exists for a more general family of loss functions. Although the parameter β_M^* cannot be worked out explicitly in terms of β^*, a similar dependency between β_M^* and β^* can be seen.

Consider the mean model in equation (5.10) for a general link function. Then for a loss function with the form of a generalized linear model, the sub-model parameters can be expressed using a similar representation to equation (5.11)

$$\begin{aligned} 0 &= \mathrm{E}[D_M^T V_M^{-1}(y - g^{-1}(x_M\beta_M^*))] \\ &= \mathrm{E}[D_M^T V_M^{-1}(g^{-1}(x_M\beta^{M*} + x_C\beta^{C*}) - g^{-1}(x_M\beta_M^*))]. \end{aligned} \tag{5.12}$$

where $D = \frac{\partial g^{-1}(x_M\beta_M^*)}{\partial\beta}$ and V_M is a matrix with elements $v(x_M\beta_M^*)$ on the main diagonal and 0 on the off diagonal for some function $v()$ determined by the working model.

Thus, we have that if $x_C\beta^{C*} = 0$, then $\beta_M^* = \beta^{M*}$; however, in general, $\beta_M^* \neq \beta^{M*}$. As we noted, this is precisely what was demonstrated in the least squares case.

So far we have discussed sub-models with no restrictions on the parameters and demonstrated that only under special conditions is the target of inference equivalent between full and sub-models. We now explore the setting of a sparse model, where many of the parameters are 0.

Sparse Model Parameters

Now suppose x consists of active features x_A and inactive features x_C such that for $x_j \in x_A$, $\beta_j^* \neq 0$ and for $x_j \in x_C$, $\beta_j^* = 0$ for $j = 1, ..., p$. Let β^{A*} denote the subset of non-zero parameters and β^{C*} denote the subset of parameters that are zero. In this setting we may be interested in estimating parameters for only the (unknown) active features. For the set of active features, x_A, the target parameter of inference is

$$\beta_A^* = \arg\min_{\beta} \mathrm{E}[l(y, x_A; \beta_A)]. \tag{5.13}$$

.

The following lemma relates the target parameters of interest from the full model (equation (5.2)) to the target parameters of the sub-model on only the active features (equation (5.13)).

Lemma 2. *Let y be a univariate outcome and x be a p-length vector of features with parameters satisfying $\beta^* = \arg\min_\beta \mathrm{E}[\dot{l}(y, x; \beta)]$ for a strictly convex loss function $l(y, x; \beta)$. Suppose that $A \equiv \{j, \beta_j^* \neq 0\}$ is a fixed index set for the active features. Let $x_A \equiv \{x_j; j \in A\}$ denote the set of active features with a corresponding subset of parameters $\beta^{A*} \equiv \{\beta_j; j \in A\}$. Suppose the sub-model on X_A has parameters as defined in equation* (5.13)*. Then $\beta_A^* = \beta^{A*}$.*

Proof. From equation (5.12), since $\beta^{C*} = 0$, then $x_C \beta^{C*} = 0$. Therefore $\beta_A^* = \beta^{A*}$. □

Thus in the case of a sparse model, the full and sub-models have the same target of inference for the active features. With a better understanding of the target parameter of interest in sub-models, we can now discuss the challenges and methods used to make inference on these parameters.

Sub-Model Inference

In the previous section, we discussed the implications of using a sub-model on the target of inference, in particular how the target of inference is defined by the features included in the sub-model. In this section, we contrast the implications of selecting X_M *a priori* versus using a data adaptive approach on inference.

In this discussion, we consider n independent and identically distributed realizations (y_i, x_i) for $i = 1, ..., n$. Let X denote the design matrix with the i^{th} row equal to x_i. The matrix X_M denotes a design matrix with only the features indexed by M

A priori X_M

When the set of features X_M is determined *a priori*, the target parameter of inference is well defined and does not depend on the data. Assuming X_M is suitably low-dimensional, classical inference can be made. The standard estimate

$$\hat{\beta}_M = \arg\min_{\beta} l_n(\beta_M).$$

is consistent for β_M^* and asymptotically normal. Standard confidence intervals and p-values can be used to make inference on the parameters.

However, many settings, the desired subset of features is unknown *a priori* and an alternative method is needed to select X_M.

Data Adaptive X_M

A data adaptive approach can also used to select X_M. This can be done using the lasso estimate in equation (5.6) by defining $\hat{M} = \{j; \hat{\beta}_j^L \neq 0\}$ or through another screening method such as stepwise regression. We use the notation $X_{\hat{M}}$ to make it clear that there is randomness to the features selected and differentiate from the prior setting, where the features selected are fixed *a priori*.

For the set of features, $x_{\hat{M}}$,

$$\beta^*_{\hat{M}} = \arg\min_{\beta_{\hat{M}}} \mathrm{E}[l(y, x_{\hat{M}}; \beta_{\hat{M}})]. \tag{5.14}$$

This expression makes clear the dependency of the target parameter of inference, $\beta^*_{\hat{M}}$ on the estimated set of active features, $x_{\hat{M}}$.

While it is tempting to make classical inference using the estimate

$$\hat{\beta}_{\hat{M}} = \arg\min_{\beta_{\hat{M}}} l_n(\beta_{\hat{M}}),$$

two problems arise from not having $\hat{M}$ specified *a priori*. The first is that the target parameter of inference is dependent on the features selected and not well-defined *a priori*. The second problem is that estimates may be overly optimistic as the same data is used to select features as well as make inference on those features.

Selective-inference literature attempts to address these issues to allow for valid inference [68]. Data-splitting and using a truncated normal distribution for parameter estimates are two methods to make valid inference under rigid assumptions. More recent work has demonstrated that data-splitting may not be necessary under further assumptions, notably when the selected features, $\hat{M}$, are deterministic and non-data dependent with high probability [45, 85].

Even when selective-inference methods are valid, making inference on a parameter that is not well defined may not be preferable. In the next section, and the rest of this chapter, we focus on parameter estimation using sparse models, where the sub-model has the same target of inference as the full model for the active features.

SCARF Inference

Now consider the special case of a sparse model. Let $A = \{j; \beta^*_j \neq 0\}$ and $A^c = \{j; \beta^*_j = 0\}$ be the index sets for the active and inactive features respectively such that $x_j \in X_A$ when $j \in A$ and $x_j \in X_{A^c}$

when $j = A^c$ for $j = 1, ..., p$. Let β^{A*} denote the non-zero parameters and β^{C*} denote parameters that are zero.

If X_A is known, then classical inference can be made on the parameters β_A using estimates

$$\hat{\beta}_A = \arg\min_{\beta}(l_n(\beta_A)). \tag{5.15}$$

In general, X_A is unknown and a data adaptive approach is used to estimate the set of active features before using a standard regression techniques to make inference on the features selected. In this chapter we use the lasso to screen for active features and a GLM for parameter estimation. We refer to this method as the screen and refit (SCARF) method. The two stage estimation procedure for SCARF estimates is formalized below.

Stage 1: Use the lasso to estimate the set of active features.

For the lasso estimate defined in equation (5.6), let $\hat{A} = \{j; \hat{\beta}_j^L \neq 0\}$ and $\hat{A}^c = \{j; \hat{\beta}_j^L = 0\}$ index the estimated set of active and non-active features.

Stage 2: Use the estimated active set to fit a standard regression model

Assuming $X_{\hat{A}}$ is suitably low-dimensional, the standard regression estimate

$$\hat{\beta}_{\hat{A}} = \arg\min_{\beta_{\hat{A}}} l_n(\beta_{\hat{A}}) \tag{5.16}$$

can be used for inference.

At a glance, this method should have the same problems as the general family of sub-models: using the data twice and having an undefined target parameter. However, due to the sparsity assumption of the true model and Lemma 1, the target of inference is constant as long as the active features are included in the estimated active set. Furthermore, when the estimated active set of features converges to a deterministic set that includes the active features with probability one, there is no randomness to the feature selection. Thus selective inference methods are not needed. We now discuss these concepts in more detail.

We assume $\hat{A}$ converges in probability to a deterministic set A^* which contains the indexes of

the true active features ($A \subseteq A^*$). We denote the parameter in equation (5.14) as $\beta^*_{A^*}$ when $\hat{A} = A^*$ and $\beta^*_{A^c}$ when $\hat{A} \neq A^*$. Note that $\beta^*_{A^*}$ is a well defined parameter as it corresponds to a fixed set of features. However, $\beta^*_{A^c}$ is not well defined as it corresponds to a random set the features.

The parameter estimates in equation (5.16) can be written as

$$\hat{\beta}_{\hat{A}} = \begin{cases} \hat{\beta}_{A^*} : X_{\hat{A}} = X_{A^*} \\ \hat{\beta}_{A^c} : X_{\hat{A}} \neq X_{A^*} \end{cases}$$

Then

$$\sqrt{n}(\hat{\beta}_{\hat{A}} - \beta^*_{\hat{A}}) = \sqrt{n}(\hat{\beta}_{A^*} - \beta^*_{A^*})\mathbb{I}(X_{\hat{A}} = X_{A^*}) + \sqrt{n}(\hat{\beta}_{A^c} - \beta^*_{A^c})\mathbb{I}(X_{\hat{A}} \neq X_{A^*})$$

where $\mathbb{I}$ is an indicator, $\sqrt{n}(\hat{\beta}_{A^*} - \beta^*_{A^*}) \sim N(0, V_{A^*})$ for some covariance matrix, V_{A^*}, and $\sqrt{n}(\hat{\beta}_{A^c} - \beta^*_{A^c}) \sim D$ for some distribution D that is not well defined. Note that $\mathbb{I}(X_{\hat{A}} = X_{A^*}) \to 1$ in probability (and thus $\mathbb{I}(X_{\hat{A}} \neq X_{A^*}) \to 0$). Then by Slutsky's Theorem,

$$\sqrt{n}(\hat{\beta}_{\hat{A}} - \beta^*_{A^*}) \to N(0, V_{A^*}).$$

Thus, when we have a sparse model and the feature selection is consistent, valid inference can be made on the original, full model target parameters using SCARF estimates.

5.3.2 Full-Model Parameters

In this section we discuss methods to make inference on parameters from the full model. The idea is to remove remove the sparsity component of the bias from the lasso estimate. Zhang and Zhang proposed a low dimension projection estimate (LDPE) for linear models, that uses a relaxed form of the projection of x_j to the column space of X_{-j} by using lasso residuals [82]. Van de Geer et al. provide a similar correction by directly inverting the KKT conditions using a relaxed form of the inverse of the estimated covariance matrix for the more general family of loss functions based on generalized linear models [73]. In some cases, this de-biased estimate can achieve asymptotic efficiency [72]. Javanmard and Montanari proposed a related correction assuming a Gaussian

model with known covariance [40]. These methods are similar in that they involve a one-step correction on initial lasso estimates to obtain sparsity-debiased (SDB) estimates that can be used to make valid inference. SDB methods can also be used under model misspecification [9]. We provide a heuristic overview of the method proposed in detail by Van de Geer et al. [73]

Sparsity Debiasing

Suppose that $\hat{\Theta}$ is a relaxed inverse of $\ddot{l}_n(\beta^*)$ such that $(\hat{\Theta}\ddot{l}_n(\beta^*) - I)(\hat{\beta}^L - \beta^*) = o_p(n^{-1/2})$. Note that equation (5.7) gives $\dot{l}_n(\hat{\beta}^L) = \lambda\hat{k}$. Then equation (5.9) can be rewritten as

$$(\hat{\beta}^L - \beta^*) = \hat{\Theta}\dot{l}_n(\hat{\beta}^L) + \hat{\Theta}\dot{l}_n(\beta^*) + \hat{\Theta}r_n + (\hat{\Theta}\ddot{l}_n(\beta^*) - I)(\hat{\beta}^L - \beta^*). \tag{5.17}$$

Then the bias is

$$\mathrm{E}[(\hat{\beta}^L - \beta^*)] = \mathrm{E}[\hat{\Theta}\dot{l}_n(\hat{\beta}^L)] + \mathrm{E}[o_p(n^{-1/2})].$$

This leads to the sparsity-debiased estimate

$$\hat{\beta}^{SDB} = \hat{\beta}^L - \hat{\Theta}\lambda\hat{k}. \tag{5.18}$$

When the necessary conditions are met (see Theorem 3.1 from Van de Geer et al. [73])

$$\sqrt{n}(\hat{\beta}_j^{SDB} - \beta_j^*)/\hat{\sigma}_j = C_j + \delta_j$$

where $C_j \to N(0,1)$, $\delta_j = o_p(1)$, and $\hat{\sigma}_j^2$ is a consistent estimate for the variance [73].

In practice, a lasso for nodewise regression can be used on a weighted design matrix $X_{GLM} = WX$ to estimate $\hat{\Theta}$ for standard GLMs with weights, W, that satisfy $X_{GLM}^T X_{GLM}^T = \ddot{l}_n(y, X\beta)$. Note that for least squares regression, $X_{GLM} = X$.

In a sparse linear model, it has further been shown that $\hat{\beta}^{SDB}$ achieves the semi-parametric

efficiency bound [73]. While these methods were originally developed for a correctly specified model, they have extended to potentially misspecified linear models as well [10].

5.3.3 Curvature-Induced Bias

While methods have been developed to make valid asymptotic inference, by either removing the sparsity-induced bias or refitting models that do not induce sparsity, curvature-induced bias may still be present. The curvature-induced bias is asymptotically negligible and therefore does not prohibit valid asymptotic inference; however, the bias may be concerning for finite sample estimates. While the curvature-induced bias has been explored for low dimensional data in previous literature and Chapters 2 and 3 of this dissertation, the curvature-induced bias has not been explored in the context of high dimensional data. No current methods address the curvature-induced bias in high dimensional data. In the next section, we propose methods that combine elements of bias reduction from Chapter 3 with sparsity-debiasing methods in high dimensional data settings. The aim of this research is to produce estimates that can be used for asymptotically valid inference with smaller finite sample bias in the high dimensional setting. In the next section, we propose three methods for bias reduction that address the first order bias caused by the curvature of the loss-function.

5.4 Finite Sample Curvature-Bias Reduction

In this section we describe heuristically how curvature-induced bias reduction methods, previously discussed in Chapter 3, can be used to reduce the finite sample bias when used in tandem with SCARF or sparsity-debiasing methods. While rigorous conditions are not given, we provide an outline of how curvature-induced bias reduction methods can be implemented and discuss settings where reducing the curvature-induced bias can lead to an improvement in the overall finite sample bias. In Section 5.5, we assess curvature-induced bias reduction methods empirically through simulations.

We will introduce and discuss three methods for reducing the curvature-induced bias in high

dimensional data. Here we provide a brief, high level overview of the three methods. In what follows we give full details of how these methods work.

Method 1: C-BR SCARF This method uses the two stage SCARF approach. The first stage consists of a screening method to select a set of features. The second stage is a direct application of the preventative robust bias reduction method discussed in Chapter 3 on the selected features.

Method 2: Preventative C-BR/S-DB lasso This method directly modifies the loss function used in a lasso model, similar in nature to the preventative robust bias reduction method discussed in Chapter 3. The resulting estimates will have the curvature-induced bias reduced. A post-estimation S-DB method can then be used to reduce the sparsity-induced bias.

Method 3: Corrective C-BR/S-DB lasso This method uses a two-step, post-estimation correction to lasso estimates. The first step uses a S-DB method to reduce the sparsity-induced bias. The second step uses a one-step corrective robust bias reduction method from Chapter 3 on the S-DB estimate to reduce the curvature-induced bias.

We split the discussion of these methods into two sections. In Section 5.4.1 we explicitly discuss the curvature-induced bias for SCARF estimates and discuss Method 1 in more depth. Section 5.4.2 engages with the curvature induced bias for lasso estimates and provides a detailed discussion of Methods 2 and 3.

5.4.1 Curvature-Bias Reduction of SCARF Estimates

At first glance, reducing the curvature-induced bias should be straightforward. The general concept is to use the same feature selection in stage 1. In stage 2, a modified estimating equation of the form discussed in Chapter 3 is used, for which the solution is a bias-reduced estimate. However, the modified score function discussed in Chapter 3 assumes the set of features is fixed and may not be valid when using data adaptive feature selection. We first discuss the challenges in approximating the bias approximation for data adaptive approaches before discussing a naive application of the bias reduction method discussed in Chapter 3 and provide intuition for settings in which the naive application of the curvature-induced bias reduction method may still effectively reduce the

curvature-induced bias.

Bias of SCARF estimates

Before deriving the bias, we introduce the following notation:

$$\dot{l}_r(\beta) = \frac{\partial l_n(\beta)}{\partial \beta_r}$$
$$\ddot{l}_{rs}(\beta) = \frac{\partial^2 l_n(\beta)}{\partial \beta_r \partial \beta_s}$$
$$\dddot{l}_{rst}(\beta) = \frac{\partial^3 l_n(\beta)}{\partial \beta_r \partial \beta_s \partial \beta_t}.$$

Note that the estimate in equation (5.16) is equivalent to the solution of the estimating equation

$$0 = \dot{l}_n(\beta_{A*})\mathbb{I}(X_{\hat{A}} = X_{A*}) + \dot{l}_n(\beta_{A^c})\mathbb{I}(X_{\hat{A}} \neq X_{A*}). \tag{5.19}$$

Using a Taylor expansion as done in Chapter 3, the bias can be written as

$$\mathrm{E}[\hat{\beta}_{\hat{A}} - \beta^*_{\hat{A}}] = b_{A*}(\hat{\beta}_{A*}) + b_{\hat{A}^c}(\hat{\beta}_{\hat{A}^c})$$

where

$$b_{A*}(\hat{\beta}^j_{A*}) = \sum_{r=1}^{p}\sum_{s=1}^{p}\sum_{t=1}^{p} I^{rj}_{A*}(I^{st}_{A*}J^{A*}_{s,rt} + \frac{1}{2}R^{A*}_{st}K^{A*}_{rst})$$

for

$$I^{A*}_{rs} = \mathrm{E}[\ddot{l}_{rs}(\beta)\mathbb{I}(X_{\hat{A}} = X_{A*})]$$
$$R^{A*}_{rs} = \sum_{t=1}^{p}\sum_{u=1}^{p} I^{rt}_{A*}\,\mathrm{E}[\dot{l}_t(\beta)\dot{l}_u(\beta)\mathbb{I}(X_{\hat{A}} = X_{A*})]I^{us}_{A*}$$
$$J^{A*}_{s,rt} = \mathrm{E}[\dot{l}_s(\beta)\ddot{l}_{rt}(\beta)\mathbb{I}(X_{\hat{A}} = X_{A*})]$$
$$K^{A*}_{rst} = \mathrm{E}[\dddot{l}_{rst}(\beta)\mathbb{I}(X_{\hat{A}} = X_{A*})]$$

where I^{st}_{A*} denotes the (s,t) element of $(I^{A^*})^{-1}$.

A similar expression for the first order bias approximation can be written for $b_{\hat{A}^c}(\hat{\beta}_{\hat{A}^c})$ by using the indicator $\mathbb{I}(X_{\hat{A}} \neq X_{A*})$ instead of $\mathbb{I}(X_{\hat{A}} = X_{A*})$ in the above expressions.

In the case of *a priori* feature selection, $P(X_{\hat{A}} = X_{A*}) = 1$. In this setting, we omit the subscript and use the notation I, J, K, and R for the elements of the bias approximation. The bias approximation is then equivalent to the expression derived in Chapter 3. We denote this bias as $b(\hat{\beta}_{A*})$ and omit the subscript on b as well.

In general, $b_{A*}(\hat{\beta}_{A*}) \neq b(\hat{\beta}_{A*})$. If the parameter estimate was independent of the selected features, the indicator could be moved outside the expectation and we would have a weighted average of bias approximations of the form in Chapter 3. However, the indicator and parameter estimate are not independent. A feature is more likely to be selected in data realizations where the parameter estimate is large. Rather than trying to approximate $b_{A*}(\hat{\beta}_{A*})$ directly, we show that $nb_{A*}(\hat{\beta}_{A*}) \to B(\hat{\beta}_{A*})$ where $\frac{1}{n}B(\hat{\beta}_{A*}) = b(\hat{\beta}_{A*})$ without assuming the indicator and estimate are independent.

We demonstrate that $I_{A*} \to I$. Similar steps can also be used to prove the convergence of J_{A*}, K_{A*}, and R_{A*}.

We assume that $\mathrm{E}\,|\dot{l}_{rs}(\beta^*)| < \infty$ for all r and s and that $\mathbb{I}(X_{\hat{A}} = X_{A*}) \to_p 1$. Note that

$$\ddot{l}_{rs}(\beta^*)\mathbb{I}(X_{\hat{A}} = X_{A*}) - \ddot{l}_{rs}(\beta^*) \to_p 0$$

and

$$|\dot{l}_{rs}(\beta^*)\mathbb{I}(X_{\hat{A}} = X_{A*})| \leqslant |I_{rs}(\beta^*)|.$$

Then by the Dominated Convergence Theorem,

$$I^{A^*}_{rs} - I_{rs} \to 0.$$

Similarly,

$$n(J^{A^*}_{s,rt} - J_{s,rt}) \to 0$$
$$K^{A^*}_{rst} - K_{rst} \to 0$$
$$n(R^{A^*}_{rs} - R_{rs}) \to 0.$$

for all r, s, t. Then by continuous mapping theorem, $I^{-1}_{A^*} \to I^{-1}$ and

$$nb_{A*}(\hat{\beta}_{A*}) \to B(\hat{\beta}_{A*}). \tag{5.20}$$

Then $b_{A*}(\hat{\beta}_{A*}) - b(\beta_{A*}) = o(n^{-1})$ and $b_{A^c}(\hat{\beta}_{A^c}) - 0 = o(n^{-1})$.

In the next section, we demonstrate how this result can be used to obtain bias-reduced SCARF estimates.

Method 1: C-BR SCARF

In this section, we develop the curvature-bias reduction SCARF (C-BR SCARF) method using a modified estimating equations during the refitting stage. Under certain conditions, a naive application (in the sense it does not account for a data adaptive approach to feature selection) of the robust bias reduction method to stage 2 of the SCARF estimation process can lead to a reduction in the curvature-induced bias. Recall from Chapter 3 that the robust bias reduction method solves the modified score function, $\dot{l}^*_n(\beta) = \dot{l}_n(\beta) - Ib(\beta)$.

This modified score function can be used in place of $\dot{l}_n(\beta)$ in equation (5.19):

$$\begin{aligned} 0 &= \dot{l}^*_n(\beta_{\hat{A}})\mathbb{I}(X_{\hat{A}} = X_{A*}) + \dot{l}^*_n(\beta_{\hat{A}})\mathbb{I}(X_{\hat{A}} \neq X_{A*}) \\ &= [\dot{l}_n(\beta_{\hat{A}}) - Ib(\hat{\beta}_{\hat{A}})]\mathbb{I}(X_{\hat{A}} = X_{A*}) + [\dot{l}_n(y; X_{\hat{A}}\hat{\beta}_{\hat{A}}) - Ib(\hat{\beta}_{\hat{A}^c})]\mathbb{I}(X_{\hat{A}} \neq X_{A*}). \end{aligned}$$

.

Then the bias can be written as

$$\mathrm{E}[\hat{\beta}_{\hat{A}} - \beta^*_{\hat{A}}] = b_{A*}(\hat{\beta}_{A*}) - b(\hat{\beta}_{A*})\mathbb{I}(X_{\hat{A}} = X_{A*}) + b_{\hat{A}^c}(\hat{\beta}_{\hat{A}^c}) - b(\hat{\beta}_{\hat{A}^c})\mathbb{I}(X_{\hat{A}} \neq X_{A*}).$$

Recall that $B(\hat{\beta}_{A*})$ is a constant such that $\frac{1}{n}B(\hat{\beta}_{A*}) = b(\hat{\beta}_{A*})$.

Then

$$nb_{A*}(\hat{\beta}_{A*}) - B(\hat{\beta}_{A*})\mathbb{I}(X_{\hat{A}} = X_{A*}) \to_p B(\hat{\beta}_{A*}) - B(\hat{\beta}_{A*}) = 0.$$

Therefore

$$b_{A*}(\hat{\beta}_{A*}) - b(\hat{\beta}_{A*})\mathbb{I}(X_{\hat{A}} = X_{A*}) = o_p(n^{-1}).$$

Similarly,

$$b_{\hat{A}^c}(\hat{\beta}_{\hat{A}^c}) - b(\hat{\beta}_{\hat{A}^c})\mathbb{I}(X_{\hat{A}} \neq X_{A*}) = o_p(n^{-1}).$$

Furthermore,

$$\begin{aligned}\mathrm{E}[\hat{\beta}_{\hat{A}} - \beta^*_{\hat{A}}] &= b_{A*}(\hat{\beta}_{A*}) - b(\hat{\beta}_{A*})\mathbb{I}(X_{\hat{A}} = X_{A*}) + b_{\hat{A}^c}(\hat{\beta}_{\hat{A}^c}) - b(\hat{\beta}_{\hat{A}^c})\mathbb{I}(X_{\hat{A}} \neq X_{A*}) \\ &= o(n^{-1}).\end{aligned}$$

Asymptotically, the robust bias reduction method removes the first order bias under the conditions needed for SCARF estimates. While no formal criteria is given, if the estimate set of features converges to the deterministic set at a fast enough rate, then the curvature-induced bias of the modified estimate will be of reduced. One way to view this assumption is that the sample size needed to ensure high probability of correct feature selection is less than the sample size needed before the curvature-induced bias is negligible. This is explored empirically through simulations in Section 5.5.

5.4.2 Curvature-Bias Reduction for Lasso Estimates

Before discussing curvature-bias reduction methods, we revisit the expression for the bias of lasso estimates in equation (5.17). We note that sparsity debiasing methods address $\hat{\Theta}\dot{l}_n(\hat{\beta}^L)$ and $\hat{\Theta}r_n$ is the curvature-induced bias, which is non-zero when $\dot{l}_n(\beta)$ is non-linear. The last term, $(\hat{\Theta}\ddot{l}_n(\beta^*) - I)(\hat{\beta}^L - \beta^*)$, is a result of using an estimated inverse of $\ddot{l}_n(\beta)$. While these remainders do not impede asymptotic inference, further reducing the magnitude of the remainder terms may improve finite sample performance. In this section we propose extending the bias reduction methods from Chapter 3 to reduce the magnitude of $\hat{\Theta}r_n$. We refer to these methods as curvature-bias reduction (C-BR) methods to distinguish from the S-DB methods previously discussed. While necessary conditions for these methods are not discussed in detail, we provide insight into the implementation and effectiveness of curvature induced bias reduction methods. As this is an extension of the work by Van de Geer et al., all assumptions needed for sparsity-bias reduced estimation are needed.

We first examine in more detail the bias C-BR methods are targeting before discussing the methods for reducing this bias.

Suppose $\dddot{l}_{rst}(\beta)$ is Lipschitz, $|\mathrm{E}[\dddot{l}_{rst}(\beta^*)]| < K$ for some constant K for all r, s, and t, and that $\sum_{i=1}^n |x_i(\hat{\beta}^L - \beta^*)|^2 = o_p(n^{-1/2})$. Then a higher order Taylor expansion (as was done in Chapter 3) can be used to write equation (5.9) as

$$-\ddot{l}_n(\beta^*)(\hat{\beta}^L - \beta^*) = \lambda\hat{k} + \dot{l}_n(\beta^*) + W(\beta^\dagger) \tag{5.21}$$

where the r^{th} element of $W(\beta^\dagger)$ is $W_r(\beta^\dagger) = \sum_{t=1}^p \sum_{u=1}^p \left[\ddot{l}_{rtu}(\beta^\dagger)(\hat{\beta}_t^L - \beta_t^*)(\hat{\beta}_u^L - \beta_u^*)\right]$ and $\beta^\dagger$ is an intermediate point between $\hat{\beta}^L$ and β^*.

Under the assumption that $\dddot{l}_{rst}(\beta)$ is Lipschitz and explicitly working out the chain rule in the

derivatives,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}|\dddot{l}_{i,rst}(\beta^{\dagger})x_{i,r}(\hat{\beta}_r^L-\beta_r^*)x_{i,s}(\hat{\beta}_s^L-\beta_s^*)-\dddot{l}_{i,rst}(\beta^*)x_{i,r}(\hat{\beta}_r^L-\beta_r^*)x_{i,s}(\hat{\beta}_s^L-\beta_s^*)| \\ &\leqslant \frac{1}{n}\sum_{i=1}^{n}|x_{i,t}(\beta_t^{\dagger}-\beta_t^*)||x_{i,r}(\hat{\beta}_r^L-\beta_r^*)x_s(\hat{\beta}_s^L-\beta_s^*)| \\ &\leqslant \frac{1}{n}\sum_{i=1}^{n}|x_{i,t}(\hat{\beta}_t^L-\beta_t^*)x_{i,r}(\hat{\beta}_r^L-\beta_r^*)x_{i,s}(\hat{\beta}_s^L-\beta_s^*)| \\ &= o_p(n^{-3/4}).\end{aligned}$$

Then equation (5.21) can be rewritten as

$$-\ddot{l}_n(\beta^*)(\hat{\beta}^L-\beta^*)=\lambda\hat{k}+\dot{l}_n(\beta^*)+W(\beta^*)+o_p(n^{-3/4}). \tag{5.22}$$

Taking the expectation, as done in Chapter 3, gives the expression

$$\mathrm{E}[-\ddot{l}_n(\beta^*)]\,\mathrm{E}[\hat{\beta}-\beta^*]=\mathrm{E}[\lambda\hat{k}]+C(\beta^*)+\mathrm{E}[W(\beta^*)]+o(n^{-3/4}) \tag{5.23}$$

where the r^{th} element of $C(\beta^*)$ is $C_r(\beta^*)=\sum_{t=1}^{p}\sum_{u=1}^{p}\mathrm{Cov}[\ddot{l}_{ru}(\beta^*),(\hat{\beta}_t-\beta_t^*)]$.

A substitution for $\hat{\beta}-\beta^*$ in $W(\beta^*)$ and $C(\beta^*)$ can be found by rearranging equation (5.8):

$$\hat{\beta}_j-\beta_j^*=\Theta_j\dot{l}_n(\beta^*)+\Theta_j\lambda\hat{k}+o_p(n^{-1/2}).$$

Then equation (5.23) becomes

$$\mathrm{E}[-\ddot{l}_n(\beta^*)]\,\mathrm{E}[\hat{\beta}-\beta^*]=\lambda\hat{k}+B1+B2+o(n^{-3/4}). \tag{5.24}$$

where the r^{th} elements of $B1$ and $B2$ are

$$B1_r = \sum_{t=1}^{p}\sum_{u=1}^{p}\left[\Theta_{tu}\text{Cov}(\ddot{l}_{ru}(\beta^*), \dot{l}_t(\beta^*)) + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\Theta_{ti}\dot{l}_i(\beta^*)\dot{l}_j(\beta^*)\Theta_{ju}\right)\text{E}[\dddot{l}_{rtu}(\beta^*)]\right]$$

$$B2_r = \sum_{t=1}^{p}\sum_{u=1}^{p}\left[\Theta_{tu}\text{Cov}(\ddot{l}_{ru}(\beta^*), \lambda\hat{k}_t) + \sum_{i=1}^{p}\sum_{j=1}^{p}\left(\Theta_{ti}(\lambda k_i(\lambda k_j + \dot{l}_j(\beta^*)))\Theta_{ju}\right)\text{E}[\dddot{l}_{rtu}(\beta^*)]\right]$$

for $\Theta = \text{E}[-\ddot{l}_n(\beta^*)]^{-1}$. Then the bias is

$$\text{E}[\hat{\beta}_j - \beta_j^*] = \Theta_j(\lambda\hat{k} + B1 + B2 + o(n^{-3/4}))$$

The term $\Theta\lambda\hat{k}$ is the bias targeted by sparsity-bias reduction method, $\Theta B1$ is the bias we wish to target with curvature-bias reduction methods, and $\Theta B2$ is the bias from the combination of sparsity and curvature that neither method addresses. While we provide a method for addressing $\Theta B1$, the impact this has on the overall bias is unclear. The bias term, $\Theta B2$, and the remainder after using the sparsity debiasing method on $\Theta\lambda\hat{K}$ may limit the impact of curvature-bias reduction methods unless these bias terms are also accounted for. We assess the impact reducing $\Theta B1$ has on the overall bias empirically in Section 5.5.

We note that $B1$ has the form of the first order bias approximation discussed in Chapter 3 and the the first order bias approximation in equation (5.20) without the indicators:

$$\Theta B1 = b(\beta_j^*) = \sum_{r=1}^{p}\sum_{s=1}^{p}\sum_{t=1}^{p}\Theta_{rj}(\Theta_{st}J_{s,rt} + R_{st}K_{rst}). \tag{5.25}$$

In the next section, we demonstrate how the curvature-bias reduction methods can be used to reduce $\Theta B2$ when paired with a sparsity debiasing method to reduce $\Theta B1$. In Section 5.5 we empirically assess the performance of using sparsity bias reduction methods on sparsity bias reduced estimates.

5.4.3 Bias Reduction of Debiased Estimates

Two natural methods arise for reducing the curvature-induced bias of debiased lasso estimates. The first method involves modifying the loss function used in the lasso to reduce the curvature induced bias. This method is similar to the preventative bias reduction method discussed in Chapter 3. The second method is a corrective bias reduction method. This method is a post-estimation correction of the sparsity debiased estimate that reduces the curvature induced bias. Both of these methods are discussed in turn.

Method 2: Preventative C-BR/S-DB Lasso

The idea behind preventative curvature-induced bias reduction and sparsity de-biased (C-BR/S-DB lasso) estimation is to use a modified loss function in equation (5.6). Resulting lasso estimates would have the curvature-induced bias reduced. A sparsity de-biasing method can then be used to reduce the sparsity induced bias of this estimate to allow for valid inference.

Instead of directly replacing the loss function, the bias reduction methods from Chapter 3 can be used to modify the derivative of the loss function, $\dot{l}_n(\beta)$ in equation (5.7). The derivative of the loss function can be modified as was done in Chapter 3. Let $\dot{l}^*_n(\beta) = \dot{l}_n(\beta) - E[-\dot{l}_n(\beta)]b(\beta)$. Then replacing $\dot{l}_n(\beta)$ with $\dot{l}^*_n(\beta)$ in equation (5.7) gives

$$\dot{l}^*_n(\tilde{\beta}^L) + \lambda\hat{k} = 0$$

Using a Taylor expansion about $\dot{l}^*_n(\tilde{\beta}^L)$ as done in equation (5.8), the bias of $\tilde{\beta}^L$ is

$$\begin{aligned}\mathrm{E}[-\ddot{l}_n(\beta^*)]\,\mathrm{E}[\tilde{\beta}^L - \beta^*] =& \lambda\hat{k} + B1 - E[-\dot{l}_n(\beta^*)]b(\beta^*) + B2 + o(n^{-3/4}) \\ =& \lambda\hat{k} + B2 + o(n^{-3/4})\end{aligned}$$

Using a modified loss function can effectively eliminate the $B1$ term, reducing the curvature-induced bias. A sparsity debiasing method can now be used to reduce the bias $\lambda\hat{k}$ to allow for valid inference with this estimates. One notable difficulty with using this method is that $\dot{l}^*(\beta)$ may not be convex, making optimization computationally difficult. Instead, we propose an alternative post-estimation correction to achieve a similar reduction in the curvature-induced bias that is simple to implement.

Method 3: Corrective C-BR/S-DB Lasso

Rather than trying to solve a (potentially) non-convex function, the corrective C-BR/S-DB method uses a 2-step post-estimation correction on the original lasso estimates to reduce curvature and sparsity-induced bias of the estimates. The first step in this process is to use a sparsity debiasing method on the lasso estimates to reduce $\lambda\hat{k}$, which allows for valid inference. The second step is to apply a curvature-induced bias reduction method, which targets the B1 term of the bias, to the sparsity debiased estimate to reduce the curvature-induced bias. The bias approximations from Chapter 3 can be used in the bias reduction method. This process is formalized in the steps below:

1. Obtain an initial estimate, $\hat{\beta}^L$ from the lasso
2. Use a one-step correction to reduce the sparsity-induced bias $\hat{\beta}^L$ (we use the form of sparsity debiasing as present in van de Geer [73]):

$$\hat{\beta}^{SDB} = \hat{\beta}^L - \hat{\Theta}\dot{l}_n(\hat{\beta}^L)$$

3. Use a one-step correction to reduce the curvature-induced bias of $\hat{\beta}^{SDB}$ for finite samples and obtain the C-BR/S-DB estimate

$$\tilde{\beta}^{SDB} = \hat{\beta}^{SDB} - b(\hat{\beta}^{SDB}).$$

We now demonstrate that the corrective C-BR/S-DB estimate, $\tilde{\beta}^{SDB}$, has reduced both the

sparsity-induced and curvature-induced bias. We assume that $b(\beta)$ is continuous on the interval $[\beta^*, \hat{\beta}^{SDB}]$ (assuming $\beta^* < \hat{\beta}^{SDB}$) with out loss of generality) and differentiable on $(\beta^*, \hat{\beta}^{SDB})$.

Then by the mean value theorem,

$$b(\hat{\beta}^{SDB}) - b(\beta^*) = \dot{b}(\beta^\dagger)(\hat{\beta}^{SDB} - \beta^*) = o(n^{-3/2})$$

since $\hat{\beta}^{SDB} - \beta^* = o(n^{-1/2})$ and $b(\hat{\beta}^{SDB}) = O(n^{-1})$.

Then the bias of $\tilde{\beta}^c$ can be written as

$$\begin{aligned}
\mathrm{E}[\tilde{\beta}^{SDB} - \beta^*] &= \mathrm{E}[\hat{\beta}^L - \beta^*] - \hat{\Theta}\dot{l}_n(\hat{\beta}^L) - b(\hat{\beta}^{SDB}) \\
&= (\Theta\lambda\hat{k} - \hat{\Theta}\dot{l}_n(\hat{\beta}^L)) + (\Theta B1 - b(\hat{\beta})) + \Theta B2 + o(n^{3/4}) \\
&= o(n^{-1/2}) + o(n^{-3/2}) + O(n^{-1}) + o(n^{-3/4}) \\
&= o(n^{-1/2})
\end{aligned}$$

While we have successfully reduced the curvature induced bias, the order of magnitude of the overall bias has not changed. Thus the practical impact of reducing the curvature-induced bias may be negligible unless further correction is applied to the other remainder terms. Through simulations we assess the impact of reducing the curvature-induced bias.

5.5 Simulations

5.5.1 Method

We compare estimates from five estimation techniques: C-BR-SCARF, C-BR/S-DB, standard lasso, SCARF, and S-DB estimates empirically using logistic regression models. Let $\mathbf{X} = \{X_1, X_2, ...X_p\}$ be a multivariate design matrix with columns having a marginal gamma distribution generated using Gaussian copulas with exchangeable covariance. A gamma distribution was chosen since the skew of $\mathbf{X}$ was found to influence the curvature-induced bias in Chapter 2 and the skew can eas-

ily be controlled with the shape parameter. All covariates were standardized to have mean 0 and variance 1. Assuming a sparse model, for the vector of parameters $\beta = (\beta_1, \beta_2, ...\beta_p)$, $\beta_j = 1$ for $i \leqslant 5$ and $\beta_j = 0$ for $i > 5$. The outcomes Y, are independent Bernoulli events such that $logit(\mathrm{E}[y|X]) = \mathbf{X}\beta$.

In the simulations we explored how the correlation, skew, and number of features impacts the performance of the different models. We consider low, moderate, and high levels of skew corresponding to shape parameters of 1, 5, and 25 for the gamma distribution respectively and correlations of 0 (uncorrelated), 0.25, and 0.5. In these settings we assume the number of active features is fixed at 5, but allow the number of inactive features to vary such that the total number of features in the model was 25, 50, 100, or 300.

The lasso estimates were obtained using 10-fold cross validation to choose the tuning parameter, λ. In some scenarios, separation in the data occurred leading to at least one parameter estimate diverging to $\pm\infty$ for SCARF estimates although finite (but not always sensible) estimates are returned using the standard GLM function in R. We present the bias including estimates when separation occurs to demonstrate the limitations of SCARF estimation. We also present the bias when data realizations with separation are removed from SCARF estimates to compare bias across methods, conditional on a well-defined, finite estimate existing.

The average absolute bias is calculated for each scenario. Selection frequency is reported for both the active features and the non-active features for the SCARF methods. We also engage with other ways to quantify bias, such as the bias of only the active features and the bias of SCARF estimates conditional on a feature being selected. Each scenario consisted of 250 data realizations.

5.5.2 Results

Separation can occur in SCARF models, leading to large parameter estimates during the refitting step when using standard R software to fit a GLM. We illustrate the effect this has bias in one scenario. However, in all other scenarios, data realizations with separation were excluded for the SCARF model but included for all other models.

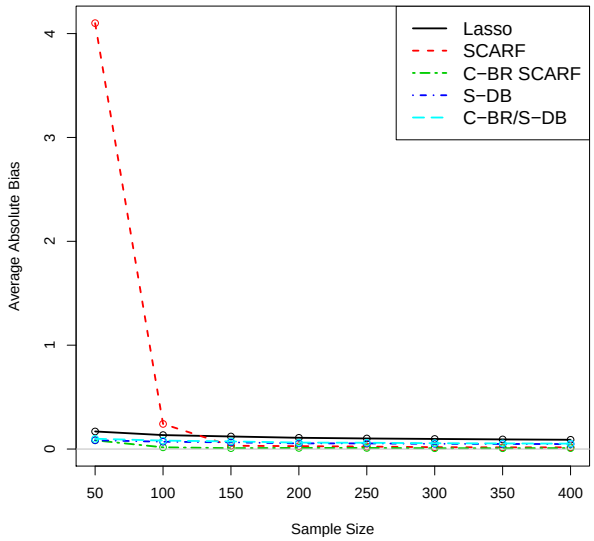
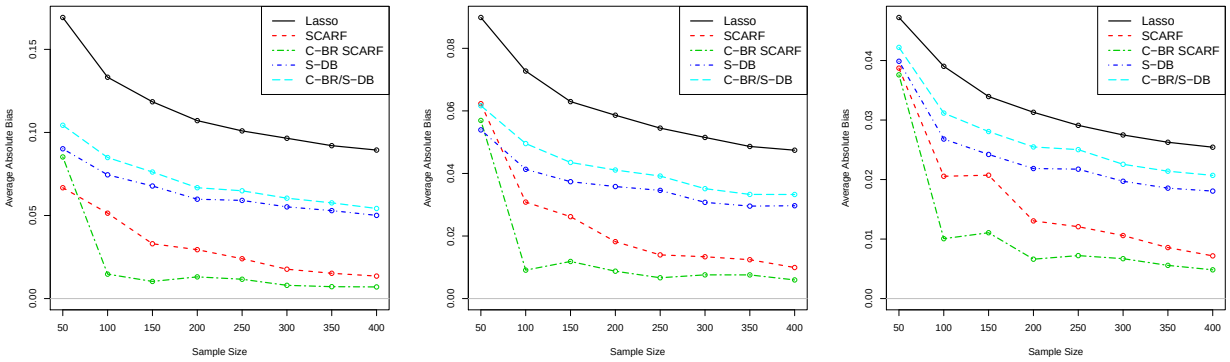


Figure 5.1: Average absolute bias of the active features for lasso, SCARF, BR-SCARF, Debias, and BR-Debias estimates for uncorrelated features, including SCARF estimates when separation occurred

Figure 5.1 presents a graph with SCARF estimates using standard R software to fit GLM for all realizations. The effect on bias is clear and makes comparing the bias of other estimates difficult. This figure demonstrates that regularization and bias reduction methods are effective even when separation in the data occurs. This trend occurred in all simulation settings; however, moving forward, we present results where estimates with separation in the data are removed from the bias of SCARF estimates, which were the only estimates impacted by separation.

Figure 5.2 displays the results for uncorrelated features across a range of sample sizes and number of parameters. We see that the trend is similar regardless of the number of features in the model (recall in each scenario there were only 5 active features.) In these scenarios we see that lasso estimates have the largest bias. For small sample sizes, SCARF and C-BR SCARF estimates have large bias due to poor feature selection. However, even at low sample sizes the bias is comparable to that of S-BR and CS-BR estimates. For sample sizes of at least 100, the BR-SCARF estimates have the lowest bias and the SCARF estimates have the second lowest bias



(a) 25 non-active features (b) 45 non-active features (c) 95 non-active features

Figure 5.2: Average absolute bias of all features for lasso, SCARF C-BR SCARF, S-DB, and C-BR/S-DB estimates for uncorrelated, moderately skewed features (the SCARF bias omits realizations where separation occurred)

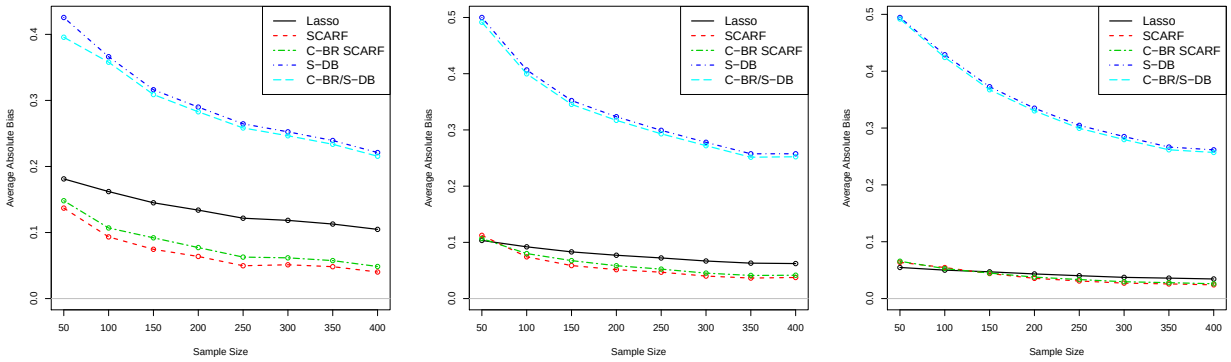
and outperform the other three methods. Both S-DB and C-BR/S-DB estimates had less bias than standard lasso estimates. However, C-BR/S-DB estimates had a larger bias than S-DB estimates indicating the bias reduction process introduced additional finite sample bias.

For correlated data (figure 5.3), the S-DB estimate had the highest bias when considering all features and was comparable to the C-BR/S-DB estimate. However, when only considering the bias of the active features (figure 5.4) we see the S-DB and C-BR/S-DB estimates have lower bias than lasso estimates. This indicates that while S-DB estimates effectively reduce the bias in active features, in small samples, they introduce non-negligible bias in non-active features.

In this setting, the SCARF estimates also had slightly lower bias than the C-BR SCARF estimates when considering either all features or only the active features, though both had lower bias than lasso estimates. The poorer performance of SCARF estimates for correlated data can be attributed to worse selection percentage as seen in table 5.1. The selection percentages are lower for active features and higher for non-active features when correlation is present in the features.

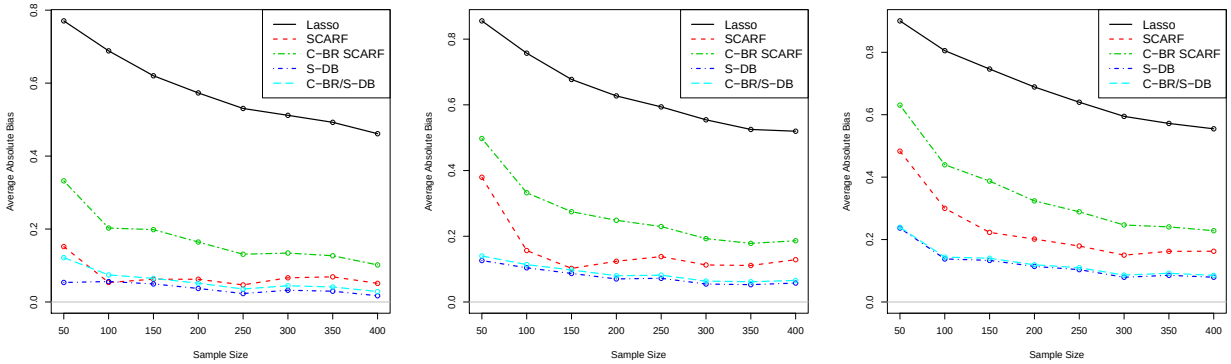
This reinforces the concept that the performance of SCARF methods (and C-BR SCARF methods by extension) rely on the correct features being selected.

While the increase in bias when using a C-BR method is counterintuitive, this phenomena can be explained. Figure 5.5 shows the average bias of the active features for two data scenarios. In



(a) 25 non-active features (b) 45 non-active features (c) 95 non-active features

Figure 5.3: Average absolute bias of all features for lasso, SCARF, C-BR SCARF, S-DB, and C-BR/S-DB estimates for moderately correlated, moderately skewed features



(a) 25 non-active features (b) 45 non-active features (c) 95 non-active features

Figure 5.4: Average absolute bias of the active features for lasso, SCARF, C-BR SCARF, S-DB, and C-BR/S-DB estimates for correlated features

the first scenario consisting of uncorrelated features, the bias for SCARF estimates is (generally) positive while the bias for BR-Debias estimates is negative. Recalling the expression of bias from Chapter 2, in this scenario the bias approximation is positive. This means the C-BR methods will pull estimates in the negative direction. When the bias is positive, the negative correction reduces the absolute bias. However, when the bias is negative, the negative correction leads to an increase in the absolute bias.

This also explains why C-BR SCARF estimates may have higher bias when correlation between features exists. When correlation is high between features (as seen in the second scenario),

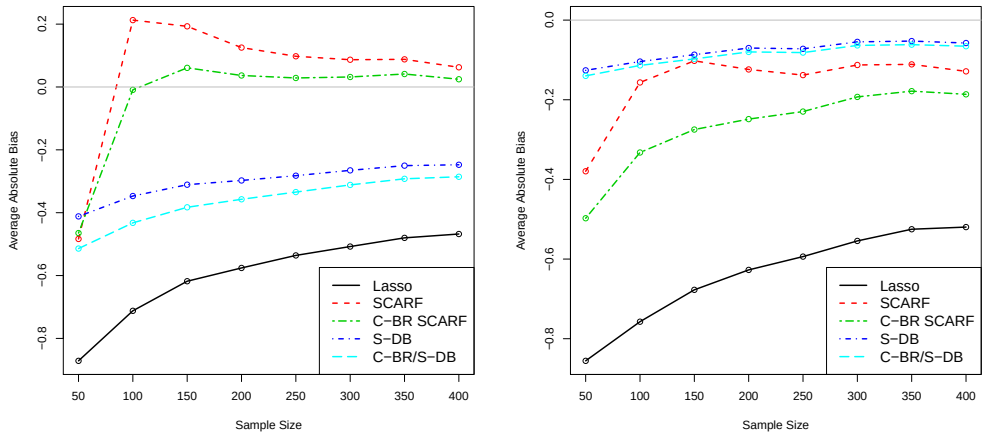
	Active Features				Non-Active Features			
	p=25		p=100		p=25		p=100	
n	$\rho = 0$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$
50	48.3	58.6	29.5	36.0	7.3	14.8	2.3	5.5
100	93.0	83.5	82.3	66.6	11.0	19.3	5.1	8.3
150	99.1	93.8	96.8	83.9	9.9	19.9	5.0	9.1
200	99.7	97.8	99.5	93.1	10.0	20.2	4.4	9.0
250	100	99.3	99.8	96.1	9.4	18.9	4.6	9.7
300	100	99.7	100	98.2	9.3	20.1	4.3	9.6
350	100	100	100	99.7	8.5	19.5	4.3	10.1
400	100	100	100	99.8	6.6	18.4	3.6	9.7

Table 5.1: Selection percentage of SCARF methods for scenarios with correlated and uncorrelated, moderately skewed features dichotomized by active and non-active features.

selection rates of the active features is worse, meaning a bias approximation relying on the correct set of features may no longer be accurate. The result is that the bias approximation is not in the same direction as the true bias. In the scenarios with high correlation, SCARF estimates have negative bias, while the bias approximation is positive. This means C-BR methods will lower parameter estimates further, increasing the absolute bias, further reinforcing that correlation may be problematic for SCARF and C-BR SCARF estimates. C-BR/S-DB and S-DB estimates effectively reduce the bias of the active features at the cost of finite sample bias in the non-active features which may lead to an overall increase in the ℓ_1-norm of the bias for a large number of non-active features.

5.6 Discussion

This work demonstrates that methods to reduce the curvature-induced bias may be viable when using a method that involves screening for active features and then refitting a standard regression model based on the selected active features. The bias of SCARF estimates can be reduced by replacing the loss function in the refitting step with a modified loss function. This method relies on selection consistency, and in particular, for the features selected through the screening process to converge to a deterministic active set at a rate faster than the curvature-induced bias becomes neg-



(a) $\rho = 0$ (b) $\rho = 0.5$

Figure 5.5: Average absolute bias the active features for lasso, SCARF, C-BR SCARF, S-DB, and C-BR/S-DB estimates for uncorrelated (a) and correlated (b) features, (SCARF estimates when separation occurred are excluded)

ligible. For uncorrelated features, we found this was a reasonable assumption, with high selection percentages of the active features even with sample sizes of 100 for 100 features. For correlated data, the selection percentage of the active features was lower and the selection percentage of the non-active features was higher. This led to worse performance for both SCARF and C-BR SCARF estimates. Notably, in some instances, this led to higher bias of C-BR SCARF estimates than unmodified SCARF estimates. However, in small samples, SCARF estimates were susceptible to separation in the data while C-BR SCARF estimates were able to provide reasonable estimates even when separation occurred.

Using a C-BR method directly on lasso estimates did not lead to a reduction in bias. Even when paired with a S-DB method, the bias attributed to the curvature of the loss function is of lower magnitude than other bias terms attributed to sparsity. While the curvature-induced bias can be reduced, the impact on the overall bias is minimal. The sparsity (even with S-DB) can result in an overall bias with the opposite sign of the curvature-induced bias. This can lead to an increase in bias when curvature-induced bias reduction methods are used. The increase in bias was confirmed through simulations, with C-BR/S-DB estimates having higher bias than S-DB estimates, though

the difference was small. This indicates that additional methods to further reduce the sparsity-induced bias are needed before C-BR methods can have an noticeable impact.

While valid inference can be made in high dimensional settings, much work is still needed to further reduce the bias in finite sample settings. Further simulations are also needed to better identify settings where reducing the curvature-induced bias for SCARF estimates may or may not be effective.

Chapter 6

Summary

6.1 Characterization of Bias

In Chapter 2, we provided a detailed characterization for the bias of a correctly specified GLM with a log-link function. From this characterization, insight was gained into model components that contribute to bias. The distribution of the outcomes, link function, sample size, and joint distribution of the predictor are all characteristics that influence bias. In particularly, predictors that are highly skewed and have low variance are more susceptible to bias. We demonstrated these trends both analytically and empirically. While the explicit characterization was done for a log-link function, these trends may apply to other link functions as well.

Further characterization of bias is still need for a wider family of models. This characterization could provide further insights into how generalizable the results for log-linear models are to the entire GLM family.

6.2 Robust Bias Reduction Method

In Chapter 3, we demonstrated that previous bias reduction methods may not be valid when separation occurs in the data and the model is misspecified. This is particularly problematic when estimating relative risks for binary outcomes. We developed a robust bias reduction method that uses a

general form of the first order bias approximation within a preventative bias reduction framework. An empirical estimate of the first order bias approximation can be used to calculate bias reduced estimates without knowledge of the true model. However, when the true model is known, this information can be used to explicitly derive the first order approximation even when a misspecified working model is used.

More extensive simulations are needed to asses the performance of robust bias reduction methods for independent data to understand settings where the robust bias reduction methods may not be effective.

6.2.1 Clustered Data

These methods can readily be extend to clustered data using the GEE framework. When an independence working distribution is used, GEE and and GLM estimates are identical. However, using a robust bias approximation that accounts for clustering allows for improved bias reduction when clustering is present in the data.

A more comprehensive simulation study is needed to understand how different correlation structures and the number and sample size of clusters impacts the bias and bias reduction methods. Additional methods also needed for clustered data. This includes adapting the robust bias reduction method to allow for non-independent working covariances as well as extensions of bias reduction methods to generalized linear mixed models.

6.3 Indirect Consequences of Finite Sample Bias

In Chapter 5, consequences of bias and bias reduction were discussed in the context of the delta-method and meta-analysis.

6.3.1 Transformed Estimate

The delta-method is used to derive the asymptotic distribution of a transformed predictor. We demonstrated that in some settings, the bias of a model estimate can increase the bias of the transformed estimate. However in other settings, bias of a model estimate can reduce the bias of the transformed estimate. We proposed a one-step correction to effectively reduce the bias of the transformed estimate. This method is valid for standard and bias reduced model estimates, though bias reduced estimates simplify the bias reduction process for the transformed estimate.

The development of a preventative bias reduction method for transformed estimates is one area for future work. This method would modify the estimating equations such that the resulting estimate is biased, but the transformed estimate is bias. While this concept is straightforward, identifying the appropriate modification is more challenging.

6.3.2 Meta-Analysis

For meta-analysis, we demonstrated that pooling estimates may decrease the standard error but not necessarily the bias, leading to an inflated type I error in some scenarios. Further more, the choice of weights greatly impacts the effect of bias and the effectiveness of bias reduction methods. In particular, reducing the bias of individual studies can effectively reduce the bias for sample size and unit weights but can lead to an increase in bias when inverse-variance weights are used. However, bias reduced estimates are generally not available for meta-analyses and an alternative bias reduction method is needed.

The development of a bias reduction method that does not rely on reducing the bias of individual studies is needed, particular for inverse-variance weights. Using the jackknife or bootstrap on study estimates may be one method to provide reliable bias reduction. A comprehensive simulation study could provide insight into whether these re-sampling methods provide reliable bias reduction

Additional research also needs to be done to understand the interaction between finite sample

estimation and other types of bias (such as publication bias) in meta-analyses.

6.4 Bias Reduction in High Dimensional Data

In Chapter 5, we extended bias reduction methods to high dimensional data. A direct application of the robust bias reduction method from Chapter 3, after using a screening method to select active features, proved effective when the true active variables were selected with high probability. However, modifications to sparsity de-biased lasso estimates were ineffective at further reducing the bias.

More comprehensive simulations studies are needed to assess curvature-bias reduction methods in the screen and refit method. Future work is also needed to further reduce the sparsity-induced bias to reduce the finite sample bias of lasso estimates.

Appendix A

A.1 Derivation of First Order Bias Approximation (Single Parameter)

We first walk through the derivation of the general form of the first order bias approximation for a single parameter. In Appendix A.2 we discuss the multi-parameter setting.

Let $U(\beta)$ be a set of estimating equations satisfying assumptions A1-A4 and B1-B2 with parameter estimate $\hat{\beta}$ such that $U(\hat{\beta}) = 0$. Define $\dot{U}(\beta) = \frac{\partial U(\beta)}{\partial \beta}$ and $\ddot{U}(\beta) = \frac{\partial \dot{U}(\beta)}{\partial \beta}$. From assumptions A3 and A4, we note that $\mathrm{E}[U(\beta^*)] = 0$ and that $\hat{\beta} - \beta^* = o(n^{-1/2})$.

Then, using a Taylor expansion

$$U(\hat{\beta}) = U(\beta^*) + \dot{U}(\beta^*)(\hat{\beta} - \beta^*) + \ddot{U}(\beta^\dagger)(\hat{\beta} - \beta^*)^2$$

for an intermediate point, $\beta^\dagger$, between $\hat{\beta}$ and β^*.

Now, since $\ddot{U}(\beta)$ is Lipschitz (assumption B1),

$$|\ddot{U}(\beta^\dagger) - \ddot{U}(\beta^*)||(\hat{\beta} - \beta^*)^2| \leqslant |(\hat{\beta} - \beta^*)^3| = O(n^{-3/2}).$$

Then

$$\begin{aligned}0 =& U(\beta^*) + \dot{U}(\beta^*)(\hat{\beta} - \beta^*) + \ddot{U}(\beta^*)(\hat{\beta} - \beta^*)^2 + (\ddot{U}(\beta^\dagger) - \ddot{U}(\beta^*))(\hat{\beta} - \beta^*) \\ =& U(\beta^*) + \dot{U}(\beta^*)(\hat{\beta} - \beta^*) + \ddot{U}(\beta^*)(\hat{\beta} - \beta^*)^2 + O(n^{-3/2}). \end{aligned} \tag{A.1}$$

Rearranging terms and taking the expected value gives

$$\begin{aligned}\mathrm{E}[\dot{U}(\beta^*)]\,\mathrm{E}[(\hat{\beta} - \beta^*)] =& \mathrm{Cov}[\dot{U}(\beta^*), (\hat{\beta} - \beta^*)] + \mathrm{E}[\ddot{U}(\beta^*)]\,\mathrm{E}[(\hat{\beta} - \beta^*)^2] \\ &+ \mathrm{Cov}[\ddot{U}(\beta^*), (\hat{\beta} - \beta^*)^2] + O(n^{-3/2}). \end{aligned} \tag{A.2}$$

Using a first order Taylor approximation, substituting $\hat{\beta} - \beta^* = I^{-1}(U(\beta^*) + O(n^{-1/2}))$ into the right side of equation A.2 gives

$$\begin{aligned}I\,\mathrm{E}[(\hat{\beta} - \beta^*)] =& I^{-1}\,\mathrm{Cov}[\dot{U}(\beta^*), U(\beta^*) + O(n^{-1/2})] + \mathrm{E}[\ddot{U}(\beta^*)]I^{-1}\,\mathrm{E}[(U(\beta^*) + O(n^{-1/2}))^2]I^{-1} \\ &+ I^{-1}\,\mathrm{Cov}[\ddot{U}(\beta^*), \left(U(\beta^*) + O(n^{-1/2})\right)^2]I^{-1} + O(n^{-3/2}). \end{aligned}$$

We now note that

$$\begin{aligned}\mathrm{Cov}[\dot{U}(\beta^*), U(\beta^*) + O(n^{-1/2})] &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}[\dot{U}_i(\beta^*)U_i(\beta^*)] + \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}[\dot{U}_i(\beta^*)O(n^{-1/2})] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}[\dot{U}_i(\beta^*)U_i(\beta^*)] + O(n^{-3/2}) \\ &= \mathrm{E}[\dot{U}(\beta^*), U(\beta^*)] + O(n^{-3/2}),\end{aligned}$$

$$\begin{aligned}\mathrm{E}[(U(\beta^*)+O(n^{-1/2}))^2] &= \mathrm{E}[\frac{1}{n^2}\sum_{i=1}^{n}U_i(\beta^*)^2]+\mathrm{E}[\frac{1}{n^2}\sum_{i=1}^{n}\Big(2U_i(\beta^*)O(n^{-1/2})+O(n^{-1})\Big)]\\ &= \mathrm{E}[\frac{1}{n^2}\sum_{i=1}^{n}U_i(\beta^*)^2]+O(n^{-3/2})\\ &= \mathrm{E}[U(\beta^*)^2]+O(n^{-3/2})\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}[\ddot{U}(\beta^*),(U(\beta^*))^2] &= \frac{1}{n^3}\sum_{i=1}^{n}\mathrm{E}[\ddot{U}_i(\beta^*)(U_i(\beta)+O(n^{-1/2}))^2]\\ &= O(n^{-3/2}).\end{aligned}$$

Then equation A.3 can be written as

$$\begin{aligned}I\,\mathrm{E}[(\hat{\beta}-\beta^*)] &= I^{-1}\,\mathrm{E}[\dot{U}(\beta^*),U(\beta^*)]+\mathrm{E}[(U(\beta^*)^2]I^{-1}\,\mathrm{E}[(U(\beta^*)^2]I^{-1}+O(n^{-3/2})\\ &= I^{-1}J+\frac{1}{2}VK+O(n^{-3/2})\end{aligned}$$

for $J=\mathrm{E}[\dot{U}(\beta^*)U(\beta^*)]$, $V=I^{-1}\,\mathrm{E}[(U(\beta^*)^2]I^{-1}$, and $K=\mathrm{E}[\ddot{U}(\beta^*)]$.

Finally, the bias can be written as

$$\mathrm{E}[(\hat{\beta}-\beta^*)] = I^{-1}(I^{-1}J+\frac{1}{2}VK)+O(n^{-3/2}).$$

In the special case of a correctly specified model, $\mathrm{E}[\ddot{U}(\beta^*)]=I$, $V=I^{-1}$, and

$$\mathrm{E}[(\hat{\beta}-\beta^*)] = I^{-1}(I^{-1}(J+\frac{1}{2}K)+O(n^{-3/2}).$$

A.2 Derivation of First Order Bias Approximation (Multi- Parameter)

We now repeat the derivation for p parameters, omitting some of the intermediary steps.

Let $U(\beta)$ be a set of estimating equations satisfying assumptions A1-A4 and let $U_r(\beta)$ denote the r^{th} element of the estimating equation. Define the $p \times p$ matrix $\dot{U}_{rs}(\beta) = \frac{\partial U(\beta_r)}{\partial \beta_s}$ and the $p \times p \times p$ array $\ddot{U}_{rst}(\beta) = \frac{\partial \dot{U}_{rs}(\beta)}{\partial \beta_t}$. From assumptions A3 and A4, we note that $\mathrm{E}[U(\beta^*)] = 0$ and that $\hat{\beta} - \beta^* = o(n^{-1/2})$. This derivation follows the same steps as that for a single parameter.

Then, using a Taylor expansion

$$U(\hat{\beta}) = U(\beta^*) + \dot{U}(\beta^*)(\hat{\beta} - \beta^*) + W(\beta^\dagger)$$

where the r^{th} element of $W(\beta)$ is

$$W_r(\beta) = \sum_{i=1}^{p} \sum_{j=1}^{p} [\ddot{U}_{rst}(\beta)(\hat{\beta}_i - \beta_i^*)(\hat{\beta}_j - \beta_j^*)]$$

for an intermediate point, $\beta^\dagger$, between $\hat{\beta}$ and β^*.

Now, since $\ddot{U}(\beta)$ is Lipschitz (assumption B1),

$$|\ddot{U}_{rst}(\beta^\dagger) - \ddot{U}_{rst}(\beta^*)||(\hat{\beta}_r - \beta_r^*)||(\hat{\beta}_s - \beta_s^*)| \leqslant |(\hat{\beta}_r - \beta_r^*)||(\hat{\beta}_s - \beta_s^*)||(\hat{\beta}_t - \beta_t^*)| = O(n^{-3/2}).$$

Then

$$0 = U(\beta^*) + \dot{U}(\beta^*)(\hat{\beta} - \beta^*) + W(\beta^*) + W(\beta^\dagger) - W(\beta^*)$$

$$U(\beta^*) + \dot{U}(\beta^*)(\hat{\beta} - \beta^*) + W(\beta^*) + O(n^{-3/2})$$

.

Taking the expectation and rearranging terms gives

$$\mathrm{E}[\dot{U}(\beta^*)]\,\mathrm{E}[\hat{\beta}-\beta^*] = C(\beta^*) + \mathrm{E}[W(\beta^*)] + O(n^{-3/2})$$

where the r^{th} element of C is $C_r(\beta^*) = \sum_{s=1}^{p}\sum_{t=1}^{p} Cov(\dot{U}_{rt}, U_s + O(n^{-1/2}))$

We note that, without explicitly working out the summations as was done the single parameter case,

$$\mathrm{Cov}(\dot{U}_{rs}(\beta^*), U_t(\beta^*) + O(n^{-1/2}) = \mathrm{E}[\dot{U}_{rs}(\beta^*)U_t(\beta^*)] + O(n^{-3/2})$$

and

$$\mathrm{E}[W_{rst}(\beta^*)] = \mathrm{E}[\ddot{U}_{rst}(\beta^*)]V_{st} + O(n^{-3/2})$$

where V_{st} is the (s,t) element of $V = I^{-1}\,\mathrm{E}[U(\beta^*)U(\beta^*)^T]I^{-1}$.

Then the bias is

$$\mathrm{E}[\hat{\beta}_j - \beta_j^*] = \sum_{r=1}^{p}\sum_{s=1}^{p}\sum_{t=1}^{p} I^{rj}(I^{st}J_{s,rt} + V_{st}K_{rst}) + O(n^{-3/2})$$

As previously shown, when the model is correctly specified, $\mathrm{E}[\ddot{U}(\beta)^*] = I^{-1}$, $V = I-1$, and

$$\mathrm{E}[\hat{\beta}_j - \beta_j^*] = \sum_{r=1}^{p}\sum_{s=1}^{p}\sum_{t=1}^{p} I^{rj}I^{st}(J_{s,rt} + \frac{1}{2}K_{rst})$$

A.3 First Order Bias for Common GLM Probability Distributions and Link Functions

Below we work through examples for calculating the bias for a single parameter (intercept-only) model for common GLM. We break the components down into preliminary equations ((($L(\beta_0)$, $U(\beta_0)$, $\dot{U}(\beta_0)$, and $\ddot{U}(\beta_0)$, the components (I, K and K), and the bias ($b(\beta_0)$).

A.3.1 Binomial Models

Binomial: Identity Link

1. Preliminaries

$$
\begin{aligned}
L(\beta_0) &= \beta_0^y(1-\beta_0)^{1-y} \\
U(\beta_0) &= \frac{y}{\beta_0} - \frac{1-y}{1-\beta_0} \\
\dot{U}(\beta_0) &= \frac{-y}{\beta_0^2} - \frac{1-y}{(1-\beta_0)^2} \\
\ddot{U}(\beta_0) &= \frac{2y}{\beta_3} - \frac{2(1-y)}{(1-\beta_0)^3}
\end{aligned}
$$

2. Components

$$
\begin{aligned}
I(\beta_0) &= \frac{1}{\beta_0(1-\beta_0)} \\
J(\beta_0) &= \frac{-1}{\beta_0^2} + \frac{1}{(1-\beta_0)^2} \\
K(\beta_0) &= \frac{2}{\beta_0^2} - \frac{2}{(1-\beta_0)^2}
\end{aligned}
$$

3. Bias

$$b(\beta_0) = (\beta_0(1-\beta_0))^2(\frac{-1}{\beta_0^2} + \frac{1}{(1-\beta_0)^2} + \frac{1}{2}(\frac{2}{\beta_0^2} - \frac{2}{(1-\beta_0)^2}))$$

$$= 0$$

Binomial: Log Link

1. Preliminaries

$$
\begin{aligned}
L(\beta_0) &= exp(y\beta_0)(1-exp(\beta_0))^{1-y} \\
U(\beta_0) &= y + (1-y)\frac{exp(\beta_0)}{1-exp(\beta_0)} \\
\dot{U}(\beta_0) &= \frac{(y-1)exp(\beta_0)}{(1-exp(\beta_0))^2} \\
\ddot{U}(\beta_0) &= \frac{(y-1)exp(\beta_0)(1+exp(\beta_0))}{(1-exp(\beta_0))^3}
\end{aligned}
\tag{A.3}
$$

2. Components

$$
\begin{aligned}
I(\beta_0) &= \frac{exp(\beta_0)}{1-exp(\beta_0)} \\
J(\beta_0) &= \frac{exp(2\beta_0)}{(1-exp(\beta_0))^2} \\
K(\beta_0) &= -\frac{exp(\beta_0)(1+exp(\beta_0))}{(1-exp(\beta_0))^2}
\end{aligned}
$$

3. Bias

$$
\begin{aligned}
b(\beta_0) &= \left(\frac{1-exp(\beta_0)}{exp(\beta_0)}\right)^2 \left(\frac{exp(2\beta_0)}{(1-exp(\beta_0))^2} + \frac{1}{2}\frac{-exp(\beta_0)(1+exp(\beta_0))}{(1-exp(\beta_0))^2}\right) \\
&= \frac{exp(\beta_0)-1}{2nexp(\beta_0)}
\end{aligned}
$$

Binomial: Logit Link

1. Preliminaries

$$L(\beta_0) = (\frac{exp(\beta_0)}{1+exp(\beta_0)})^y(\frac{1}{1+exp(\beta_0)})^{1-y}$$
$$U(\beta_0) = y - \frac{exp(\beta_0)}{1+exp(\beta_0)}$$
$$\dot{U}(\beta_0) = -\frac{exp(\beta_0)}{(1+exp(\beta_0))^2}$$
$$\ddot{U}(\beta_0) = -\frac{exp(\beta_0)(1-exp(\beta_0))}{(1+exp(\beta_0))^3}$$

2. Components

$$I(\beta_0) = \frac{exp(\beta_0)}{(1+exp(\beta_0))^2}$$
$$J(\beta_0) = 0$$
$$K(\beta_0) = -\frac{exp(\beta_0)(1-exp(\beta_0))}{(1+exp(\beta_0))^3}$$

3. Bias

$$b(\beta_0) = (\frac{(1+exp(\beta_0))^2}{exp(\beta_0)})^2(0+\frac{1}{2}(\frac{exp(\beta_0)(1-exp(\beta_0))}{(1+exp(\beta_0))^3})$$
$$= -\frac{1}{2}\frac{(1-exp(2\beta_0))}{exp(\beta_0)}$$

A.3.2 Poisson Models

Poisson: Identity Link

1. Preliminaries

$$L(\beta_0) = \beta_0^y \frac{exp(-\beta_0)}{y!}$$
$$U(\beta_0) = \frac{y}{\beta_0} - n$$
$$\dot{U}(\beta_0) = \frac{y}{\beta_0^2}$$
$$\ddot{U}(\beta_0) = -\frac{2y}{\beta_0^3}$$

2. Components

$$I(\beta_0) = \frac{1}{\beta_0}$$
$$J(\beta_0) = \frac{1}{\beta_0^2}$$
$$K(\beta_0) = -\frac{2}{\beta_0^2}$$

3. Bias

$$b(\beta_0) = (exp(\beta_0))^2 \left(\frac{1}{\beta_0^2} + \frac{1}{2}\frac{-2}{\beta_0^2}\right)$$
$$= 0$$

Poisson: Log Link

1. Preliminaries

$$L(\beta_0) = exp(y\beta_0)$$

$$U(\beta_0) = \bar{y} - exp(\beta_0)$$

$$\dot{U}(\beta_0) = -exp(\beta_0)$$

$$\ddot{U}(\beta_0) = -exp(\beta_0)$$

2. Components

$$I(\beta_0) = exp(\beta_0)$$

$$J(\beta_0) = 0$$

$$K(\beta_0) = -exp(\beta_0)$$

3. Bias

$$\begin{aligned} b(\beta_0) &= \frac{1}{exp(\beta_0)} \frac{1}{exp(\beta_0)} (0 + \frac{1}{2}(-exp(\beta_0))) \\ &= -\frac{1}{2exp(\beta_0)} \end{aligned}$$

A.3.3 Exponential Models

Exponential: Identity Link

1. Preliminaries

$$L(\beta_0) = \frac{exp(\frac{-y}{\beta_0})}{\beta_0}$$
$$U(\beta_0) = \frac{y}{\beta_0^2} - \frac{1}{\beta_0}$$
$$\dot{U}(\beta_0) = \frac{-2y}{\beta_0^3} + \frac{1}{\beta_0^2}$$
$$\ddot{U}(\beta_0) = \frac{6y}{\beta_0^4} - \frac{2}{\beta_0^3}$$

2. Components

$$I(\beta_0) = \frac{1}{\beta_0^2}$$
$$J(\beta_0) = \frac{2}{\beta_0^3}$$
$$K(\beta_0) = -\frac{4}{\beta_0^3}$$

3. Bias

$$b(\beta_0) = \left(\beta_0^2\right)^2 \left(-2\beta_0^3 + \frac{1}{2}4\beta_0^3\right)$$
$$= 0$$

Exponential: Log Link

1. Preliminaries

$$L(\beta_0) = exp(\beta_0)exp(-yexp(-\beta_0))$$

$$U(\beta_0) = -n + yexp(\beta_0)$$

$$\dot{U}(\beta_0) = -yexp(\beta_0)$$

$$\ddot{U}(\beta_0) = yexp(\beta_0)$$

2. Components

$$I(\beta_0) = n$$

$$J(\beta_0) = -n$$

$$K(\beta_0) = n$$

3. Bias

$$\begin{aligned} b(\beta_0) &= n^2\left(-n + \frac{1}{2}n\right) \\ &= -\frac{1}{2n} \end{aligned}$$

Exponential: Inverse Link

1. Preliminaries

$$
\begin{aligned}
L(\beta_0) &= \beta_0 exp(-y\beta_0) \\
U(\beta_0) &= \frac{1}{\beta_0} - y \\
\dot{U}(\beta_0) &= -\frac{1}{\beta_0^2} \\
\ddot{U}(\beta_0) &= \frac{2}{\beta_0^3}
\end{aligned}
$$

2. Components

$$
\begin{aligned}
I(\beta_0) &= \frac{1}{\beta_0^2} \\
J(\beta_0) &= 0 \\
K(\beta_0) &= \frac{2}{\beta_0^3}
\end{aligned}
$$

3. Bias

$$
\begin{aligned}
b(\beta_0) &= \left(\beta_0^2\right)^2 2\left(0 + \frac{1}{2}\frac{2}{\beta_0^3}\right) \\
&= \beta_0
\end{aligned}
$$

A.4 Bias for 2 parameter model with log link function

Let Y be an $n \times 1$ vector of independent outcomes with expected value $\mathrm{E}[Y] = \mu$. Suppose] for a log-link function and an $n \times 1$ vector of features x, $\mu = exp(\beta_0 + \beta_1 x)$. Let ω_x be a function of weights such that $I = \begin{bmatrix} \sum \omega_x & \sum x\omega_x \\ \sum x\omega_x & \sum x^2\omega_x \end{bmatrix}$.

Taking the inverse gives

$I^{-1} = \frac{1}{\sum \omega_x \sum x^2\omega_x - (\sum x\omega_x)^2} \begin{bmatrix} \sum x^2\omega_x & -\sum x\omega_x \\ -\sum \omega_x & \sum \omega_x \end{bmatrix}$.

We also note that

$J_{1ij} + \frac{K_{1ij}}{2} = -\frac{1}{2} \begin{bmatrix} \sum \omega_x & \sum x\omega_x \\ \sum x\omega_x & \sum x^2\omega_x \end{bmatrix}$

and $J_{2ij} + \frac{K_{2ij}}{2} = -\frac{1}{2} \begin{bmatrix} \sum x\omega_x & \sum x^2\omega_x \\ \sum x^2\omega_x & \sum x^3\omega_x \end{bmatrix}$.

Then using the multivariate expression for bias in equation 2.3, the first order approximation for the bias for the intercept can be written as

$$\begin{aligned} b(\beta_0) &= I^{11}I^{11}K_{111} + I^{11}I^{12}K_{112} + I^{11}I^{21}K_{121} + I^{11}I^{22}K_{122} + I^{12}I^{11}K_{211} + I^{12}I^{12}K_{212} + I^{12}I^{21}K_{221} + I^{12}I^{22}K_{222} \\ &= I^{11}I^{11}K_{111} + 3I^{11}I^{12}K_{112} + I^{11}I^{22}K_{122} + 2I^{12}I^{12}K_{212} + I^{12}I^{22}K_{222} \\ &= \frac{-(\sum x^2\omega_x)^2 \sum \omega_x + 3\sum x^2\omega_x (\sum x\omega_x)^2 - \sum \omega_x (\sum x^2\omega_x)^2 - 2(\sum x\omega_x)^2 \sum x^2\omega_x + \sum \omega_x \sum x\omega_x \sum x^3\omega_x}{2[\sum \omega_x \sum x^2\omega_x - (\sum x\omega_x)^2]^2} \\ &= \frac{-2(\sum x^2\omega_x)^2 \sum \omega_x + \sum x^2\omega_x (\sum x\omega_x)^2 + \sum \omega_x \sum x\omega_x \sum x^3\omega_x}{2[\sum \omega_x \sum x^2\omega_x - (\sum x\omega_x)^2]^2} \end{aligned}$$

and the first order approximation of the bias for the slope can be written as

$$\begin{aligned} b(\beta_1) &= I^{21}I^{11}K_{111} + I^{21}I^{12}K_{112} + I^{21}I^{21}K_{121} + I^{21}I^{22}K_{122} + I^{22}I^{11}K_{211} + I^{22}I^{12}K_{212} + I^{22}I^{21}K_{221} + I^{22}I^{22}K_{222} \\ &= I^{21}I^{11}K_{111} + 2I^{21}I^{12}K_{112} + I^{21}I^{22}K_{122} + I^{22}I^{11}K_{211} + 2I^{22}I^{12}K_{212} + I^{22}I^{22}K_{222} \\ &= \frac{3\sum \omega_x \sum x\omega_x \sum x^2\omega_x - 2(\sum x\omega_x)^3 - (\sum \omega_x)^2 \sum x^3\omega_x}{2[\sum \omega_x \sum x^2\omega_x - (\sum x\omega_x)^2]^2}. \end{aligned}$$

Let $\bar{x}_\omega$ denote the weighted average of x. Then the weighted sample variance of x can be written as

$$\begin{aligned} s_w = \frac{\sum w_i(x_i - \bar{x}_\omega)^2}{\sum w_i} &= \frac{\sum w_i x_i^2 - 2\sum w_i x_i \bar{x}_\omega + \sum w_i \bar{x}_\omega^2}{\sum w_i} \\ &= \frac{\sum w_i x_i^2 - 2\frac{(\sum w_i x_i)^2}{\sum w_i} + \frac{(\sum w_i x_i)^2}{\sum w_i}}{\sum w_i} \\ &= \frac{\sum w_i x_i^2 - \frac{(\sum w_i x_i)^2}{\sum w_i}}{\sum w_i}. \end{aligned}$$

Then

$$(\sum w_i)^2 s_w = \sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2. \tag{A.4}$$

This expression can be seen on the denominator of the expressions for bias

The weighted skew can also be written as

$$\begin{aligned} \gamma_w &= \frac{\sum w_i(x_i - \tilde{x})^3}{s_w^{3/2}\sum w_i} \\ &= \frac{\sum w_i x_i^3 - 3\sum w_i x_i^2 \tilde{x} + 3\sum w_i x_i \tilde{x}^2 - \sum w_i \tilde{x}^3}{s_w^{3/2}\sum w_i} \\ &= \frac{\sum w_i x_i^3 - 3\sum w_i x_i^2 \frac{\sum w_i x_i}{\sum w_i} + 3\sum w_i x_i (\frac{\sum w_i x_i}{\sum w_i})^2 - \sum w_i (\frac{\sum w_i x_i}{\sum w_i})^3}{s_w^{3/2}\sum w_i} \\ &= \frac{\sum w_i x_i^3 - 3\sum w_i x_i^2 \frac{\sum w_i x_i}{\sum w_i} + 2\sum w_i x_i (\frac{\sum w_i x_i}{\sum w_i})^2}{s_w^{3/2}\sum w_i} \\ &= \frac{\sum w_i x_i^3 - 3\sum w_i x_i^2 \frac{\sum w_i x_i}{\sum w_i} + 2\frac{(\sum w_i x_i)^3}{(\sum w_i)^2}}{s_w^{3/2}\sum w_i} \end{aligned}$$

. Then

$$s_w^{3/2}(\sum w_i)^3 \gamma_w = (\sum w_i)^2 \sum w_i x_i^3 - 3(\sum w_i)\sum w_i x_i^2 \sum w_i x_i + 2(\sum w_i x_i)^3 \tag{A.5}$$

which is in the numerator of the expression for the bias approximation for β_1. Now using equations A.4 and A.5, the bias for β_1 can be written as

$$\begin{aligned} b(\beta_1) &= \frac{3\sum\omega_i\sum x\omega_i\sum x^2\omega_i - 2(\sum x\omega_i)^3 - (\sum\omega_i)^2\sum x^3\omega_i}{2[\sum\omega_i\sum x^2\omega_i - (\sum x\omega_i)^2]^2} \\ &= \frac{-s_w^{3/2}(\sum w_i)^3\gamma_w}{((\sum w_i)^2 s_w)^2} \\ &= \frac{-\gamma_w}{\sum w_i s_w^{1/2}} \end{aligned}$$

which is the ratio of the sample weighted skew and sample weighted variance.

A.5 Bias for binary predictor with log link function

We now consider the special case of a binary predictor. Then $x = x^2 = x^3$ and $\sum x\omega_x = \sum x^2\omega_x \sum x^3\omega_x = n_1\omega_1$. Furthermore $\sum\omega_x = n_0\omega_0 + n_1\omega_1$ where ω_0 is the weight when $x = 0$ and ω_1 is the weight when $x = 1$. Then

$$\begin{aligned} b(\beta_0) &= \frac{-2(\sum x\omega_x)^2\sum\omega_x(\sum x\omega_x)^3 + \sum\omega_x(\sum x\omega_x)^2}{2[\sum\omega_x\sum x\omega_x - (\sum x\omega_x)^2]^2} \\ &= \frac{-(\sum x\omega_x)^2\sum\omega_x + (\sum x\omega_x)^3}{2[\sum\omega_x\sum x\omega_x - (\sum x\omega_x)^2]^2} \\ &= \frac{-(\sum x\omega_x)[(\sum x\omega_x)\sum\omega_x - (\sum x\omega_x)^2]}{2[\sum\omega_x\sum x\omega_x - (\sum x\omega_x)^2]^2} \\ &= \frac{-(\sum x\omega_x)}{2[\sum\omega_x\sum x\omega_x - (\sum x\omega_x)^2]} \\ &= \frac{-n_1\omega_1}{2[n_1\omega_1^2 + n_1 n_0\omega_1\omega_0 - (n_1\omega_1)^2]} \\ &= \frac{-1}{2n_0\omega_0} \end{aligned}$$

and

$$
\begin{aligned}
b(\beta_1) &= \frac{3\sum \omega_x(\sum x\omega_x)^2 - 2(\sum x\omega_x)^3 - (\sum \omega_x)^2 \sum x\omega_x}{2[\sum \omega_x \sum x\omega_x - (\sum x\omega_x)^2]^2} \\
&= \frac{3[n_1^3\omega_1^3 + n_0 n_1^2 \omega_0 \omega_1^2] - 2(n_1^3\omega_1^3) - [n_0^2 n_1 \omega_0^2 \omega_1 + 2n_0 n_1^2 \omega_0 \omega_1^2 + n_1^3 \omega_1^3]}{2[n_0 n_1 \omega_0 \omega_1]^2} \\
&= \frac{n_0 n_1^2 \omega_0 \omega_1^2 - n_0^2 n_1 \omega_0^2 \omega_1}{2[n_0 n_1 \omega_0 \omega_1]^2} \\
&= \frac{1}{2n_0\omega_0} - \frac{1}{2n_1\omega_1}.
\end{aligned}
$$

A.6 Derivations of Bias of Pooled Estimates for Different Weights

Let $\hat{\beta}_i = \beta + b_i$ and $E[b_i] = \frac{B}{n_i}$ for $i = 1, ..., k$.

Inverse-Weights

$$
\begin{aligned}
E\left[\hat{\beta}^{IV}\right] &= E\left[\frac{\sum_{i=1}^{n}\frac{\hat{\beta}_i}{V_i}}{\sum_{i=1}^{n}\frac{1}{V_i}} - \beta^*\right] \\
&\quad E\left[\frac{\sum_{i=1}^{n}\frac{\beta^*+b_i}{V_i}}{\sum_{i=1}^{n}\frac{1}{V_i}} - \beta^*\right] \\
&\quad E\left[\frac{\sum_{i=1}^{n}\frac{b_i}{V_i}}{\sum_{i=1}^{n}\frac{1}{V_i}} + \frac{\sum_{i=1}^{n}\frac{\beta^*}{V_i}}{\sum_{i=1}^{n}\frac{1}{V_i}} - \beta^*\right] \\
&= E\left[\frac{\sum_{i=1}^{n}\frac{b_i}{V_i}}{\sum_{i=1}^{n}\frac{1}{V_i}}\right]
\end{aligned}
$$

Sample Size Weights

$$
\begin{aligned}
E\left[\hat{\beta}^{SS} - \beta^*\right] &= E\left[\frac{\sum_{i=1}^{n}\hat{\beta}_i n_i}{\sum_{i=1}^{n} n_i} - \beta^*\right] \\
&= E\left[\frac{\sum_{i=1}^{n}\beta^* + b_i n_i}{\sum_{i=1}^{n} n_i} - \beta^*\right] \\
&= E\left[\frac{\sum_{i=1}^{n} b_i n_i}{\sum_{i=1}^{n} n_i} + \frac{\sum_{i=1}^{n}\beta^* n_i}{\sum_{i=1}^{n} n_i} - \beta^*\right] \\
&= \frac{\sum_{i=1}^{n} E\left[b_i\right] n_i}{\sum_{i=1}^{n} n_i} \\
&= \frac{\sum_{i=1}^{n}\frac{B}{n_i} n_i}{\sum_{i=1}^{n} n_i} \\
&= \frac{kB}{\sum_{i=1}^{n} n_i}
\end{aligned}
$$

Unit Weights

$$
\begin{aligned}
E\left[\hat{\beta}^U - \beta^*\right] &= E\left[\frac{\sum_{i=1}^n \hat{\beta}_i}{k} - \beta^*\right] \\
&= E\left[\frac{\sum_{i=1}^n \beta^* + b_i}{k} - \beta^*\right] \\
&= E\left[\frac{\sum_{i=1}^n b_i}{k} + \frac{\sum_{i=1}^n \beta^*}{k} - \beta^*\right] \\
&= \frac{\sum_{i=1}^n E\left[b_i\right]}{k} \\
&= \frac{\sum_{i=1}^n \frac{B}{n_i}}{k} \\
&= B\frac{\sum_{i=1}^n \frac{1}{n_i}}{k}
\end{aligned}
$$

A.7 Proof of Lemma 1

From equation 5.11,

$$
\begin{aligned}
0 &= E[x_M(y - x_M^T\beta_M^*)] \\
&= E[E[x_M(y - x_M^T\beta_M^*)|x]] \\
&= E[x_M(x\beta^* - x_M^T\beta_M^*)] \\
&= E[x_M(x_M^T\beta^{M*} + x_C^T\beta^{C*} - x_M^T\beta_M^*)] \\
&= \Sigma_{M,M}\beta^{M*} + \Sigma_{M,C} - \Sigma_{M,M}\beta_M^*)
\end{aligned}
$$

where $\Sigma_{M,M} = E[x_M x_M^T]$ and $\Sigma_{M,C} = E[x_M x_C^T]$. Then

$$
\Sigma_{M,M}\beta_M^*) = \Sigma_{M,M}\beta^{M*} + \Sigma_{M,C}
$$

and

$$\beta_M^*) = \beta^{M*} + \Sigma_{M,M}^{-1}\Sigma_{M,C}.$$

Bibliography

[1] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In E Parzen, K Tanabe, and G Kitagawa, editors, *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, New York, NY, 1998.

[2] Adelin Albert and John A Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[3] Douglas Altman, Jonathon Deeks, and David Sackett. Odds ratios should be avoided when events are common. *BMJ*, 317(7168):1318, 1998.

[4] Deborah Barnes and Lisa Bero. Why review articles on the health effects of passive smoking reach different conclusions. *JAMA*, 279(19):1566–1570, 05 1998.

[5] Colin Begg and Madhuchhanda Mazumdar. Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50(4):1088–1101, 1994.

[6] Colin B Begg and Jesse A Berlin. Publication bias and dissemination of clinical research. *Journal of the National Cancer Institute*, 81(2):107–115, 1989.

[7] Michael Brannick, Sean Potter, Bryan Benitez, and Scott Morris. Bias and precision of alternate estimators in meta-analysis: benefits of blending schmidt-hunter and hedges approaches. *Organizational Research Methods*, 22(2):490–514, 2019.

[8] Michael Brannick, Liu-Qin Yang, and Guy Cafri. Comparison of weights for meta-analysis of r and d under realistic conditions. *Organizational Research Methods*, 14(4):587–607, 2011.

[9] Peter Bühlmann, Sara van de Geer, et al. High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1):1449–1473, 2015.

[10] Peter Bühlmann, Sara van de Geer, et al. High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1):1449–1473, 2015.

[11] Florentina Bunea, Alexandre Tsybakov, Marten Wegkamp, et al. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[12] Arindam Chatterjee and Soumendra Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

[13] Arindam Chatterjee and Soumendra Lahiri. Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, 41(3):1232–1259, 2013.

[14] Sourav Chatterjee and Jafar Jafarov. Prediction error of cross-validated lasso. *arXiv preprint arXiv:1502.06291*, 2015.

[15] Jien Chen and Nicole Lazar. Selection of working correlation structure in generalized estimating equations via empirical likelihood. *Journal of Computational and Graphical Statistics*, 21(1):18–41, 2012.

[16] Gauss Cordeiro and Peter McCullagh. Bias correction in generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):629–643, 1991.

[17] David Cox and Joyce Snell. A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):248–275, 1968.

[18] Arnak Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 2017.

[19] Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical Science*, 30(4):533–558, 2015.

[20] Jason Dietrich. The effects of sampling strategies on the small sample properties of the logit estimator. *Journal of Applied Statistics*, 32(6):543–554, 2005.

[21] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[22] Matthias Egger and George Davey Smith. Meta-analysis: bias in location and selection of studies. *BMJl*, 316(7124):61–66, 1998.

[23] Matthias Egger, George Davey Smith, and Andrew N Phillips. Meta-analysis: principles and procedures. *BMJ*, 315(7121):1533–1537, 1997.

[24] Matthias Egger, George Davey Smith, Martin Schneider, and Christoph Minder. Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109):629–634, 1997.

[25] David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

[26] Garrett Fitzmaurice. A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, 51:309–317, 04 1995.

[27] Ildiko E Frank and Jerome Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[28] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[29] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[30] Vidyadhar Godambe. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211, 1960.

[31] Elizabeth Hamman, Paula Pappalardo, James Bence, Scott Peacor, and Craig Osenberg. Bias in meta-analyses using hedges d. *Ecosphere*, 9(9), 2018.

[32] Frank Harrell Jr, Kerry Lee, and Daniel Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996.

[33] Larry Hedges. Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, 92(2):490 – 499, 1982.

[34] Larry Hedges and Jack Vevea. Fixed-and random-effects models in meta-analysis. *Psychological Methods*, 3(4):486, 1998.

[35] Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419, 8 2002.

[36] Lin-Yee Hin and You-Gan Wang. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine*, 28(4):642–658, 2009.

[37] Arthur Hoerl and Robert Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[38] John Hunter and Frank Schmidt. *Methods of meta-analysis: Correcting error and bias in research findings*. Sage, 2004.

[39] Nobuo Inagaki. Asymptotic relations between the likelihood estimating function and the maximum likelihood estimator. *Annals of the Institute of Statistical Mathematics*, 25(1):26, 1973.

[40] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

[41] Kenneth Katz. The (relative) risks of using odds ratios. *Archives of Dermatology*, 142(6):761–764, 2006.

[42] Ioannis Kosmidis and David Firth. A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics*, 4:1097–1112, 2010.

[43] Ioannis Kosmidis, Annamaria Guolo, and Cristiano Varin. Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression. *Biometrika*, 104(2):489–496, 2017.

[44] Matthew Ryan Lavery, Parul Acharya, Stephen A Sivo, and Lihua Xu. Number of predictors and multicollinearity: What are their effects on error and bias in regression? *Communications in Statistics-Simulation and Computation*, 48(1):27–38, 2019.

[45] Hannes Leeb, Benedikt M Pötscher, Karl Ewald, et al. On various confidence intervals post-model-selection. *Statistical Science*, 30(2):216–227, 2015.

[46] Akiva Liberman. How much more likely? the implications of odds ratios for probabilities. *American Journal of Evaluation*, 26(2):253–266, 2005.

[47] DY Lin and D Zeng. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 34(1):60–66, 2010.

[48] Petra Macaskill, Stephen D. Walter, and Les Irwig. A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, 20(4):641–654, 2001.

[49] Lloyd Mancl and Brian Leroux. Efficiency of regression estimates for clustered data. *Biometrics*, 52(2):500–511, 1996.

[50] Fulgencio Marín-Martínez and Julio Sánchez-Meca. Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, 70(1):56–73, 2010.

[51] Peter McCullagh. *Tensor Methods in Statistics*. Chapman and Hall, London, 1987.

[52] Barry McDonald. Estimating logistic regression parameters for bivariate binary data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2):391–397, 1993.

[53] Louise-Anne McNutt, Chuntao Wu, Xiaonan Xue, and Jean Paul Hafner. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology*, 157(10):940–943, 2003.

[54] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

[55] John Nelder and Robert Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[56] Szilard Nemes, Junmei Miao Jonasson, Anna Genell, and Gunnar Steineck. Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology*, 9(1):56, 2009.

[57] Wei Pan. On the robust variance estimator in generalised estimating equations. *Biometrika*, 88(3):901–906, 2001.

[58] Sudhir Paul and Xuemao Zhang. Small sample gee estimation of regression parameters for longitudinal data. *Statistics in Medicine*, 33(22):3869–3881, 2014.

[59] Margaret Sullivan Pepe and Garnet Anderson. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation*, 23(4):939–951, 1994.

[60] Jaime Peters, Alex Sutton, David Jones, Keith Abrams, and Lesley Rushton. Comparison of two methods to detect publication bias in meta-analysis. *JAMA*, 295(6):676–680, 02 2006.

[61] Hannah Rothstein, Alexander Sutton, and Michael Borenstein. Publication bias in meta-analysis. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, pages 1–7, 2005.

[62] Julio Sánchez-Meca and Fulgencio Marin-Martinez. Weighting by inverse variance or by sample size in meta-analysis: A simulation study. *Educational and Psychological Measurement*, 58(2):211–220, 1998.

[63] Edna Schechtman. Odds ratio, relative risk, absolute risk reduction, and the number needed to treatwhich of these should we use? *Value in Health*, 5(5):431–436, 2002.

[64] Lisa Schwartz, Steven Woloshin, and H. Gilbert Welch. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *New England Journal of Medicine*, 341(4):279–283, 1999.

[65] Devinder Singh, Allen Gabriel, Javad Parvizi, Michael Gardner, and Ralph Jr D'Agostino. Meta-analysis of comparative trials evaluating a single-use closed-incision negative-pressure therapy system. *Plastic and Reconstructive Surgery*, 143:41S–46S, Jan 2019.

[66] Jonathan AC Sterne, David Gavaghan, and Matthias Egger. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, 53(11):1119–1129, 2000.

[67] Jyothi Subramanian and Richard Simon. Overfitting in prediction models–is it a problem only in high dimensions? *Contemporary Clinical Trials*, 36(2):636–641, 2013.

[68] Jonathan Taylor and Robert Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

[69] Alison Thornton and Peter Lee. Publication bias in meta-analysis: its causes and consequences. *Journal of Clinical Epidemiology*, 53(2):207–216, 2000.

[70] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[71] John Tukey. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29:614, 1958.

[72] Sara van de Geer. On the efficiency of the de-biased lasso. *arXiv preprint arXiv:1708.07986*, 2017.

[73] Sara van de Geer, Peter Bühlmann, Yaacov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[74] Sara A van de Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[75] You-Gan Wang and Liya Fu. Selection of working correlation structure in generalized estimating equations. *Statistics in Medicine*, 36(14):2206–2219, 2017.

[76] Larry Wasserman and Kathryn Roeder. High-dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, 10 2009.

[77] Fengrong Wei and Jian Huang. Consistent group selection in high-dimensional linear regression. *Bernoulli*, 16(4):1369, 2010.

[78] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[79] Jacqui Wise. Links to tobacco industry influences review conclusions. *BMJ*, 316(7144):1553, 1998.

[80] Wonsuk Yoo, Robert Mayberry, Sejong Bae, Karan Singh, Qinghua Peter He, and James Lillard Jr. A study of effects of multicollinearity in the multivariable analysis. *International Journal of Applied Science and Technology*, 4(5):9, 2014.

[81] Ke-Hai Yuan and Robert Jennrich. Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, 65(2):245–260, May 1998.

[82] Cun-Hui Zhang and Stephanie Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[83] Jun Zhang and Kai F. Yu. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA*, 280(19):1690–1691, 11 1998.

[84] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006.

[85] Sen Zhao, Ali Shojaie, and Daniela Witten. In defense of the indefensible: A very naive approach to high-dimensional inference. *arXiv preprint arXiv:1705.05543*, 2017.

[86] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.