The Likelihood Pivot: Performing Inference with Confidence.

James W. Harmon

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Peter Hoff, Chair

Patrick Heagerty

Jon Wakefield

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

The Likelihood Pivot:
Performing Inference with Confidence.

James W. Harmon

Chair of the Supervisory Committee:
Professor Peter Hoff
Department of Statistics

Maximum likelihood estimation is a popular statistical method. To account for possible model misspecification, the sandwich estimate of variance can be used to generate asymptotically correct confidence intervals. Several ad hoc attempts have been made to correct for small sample bias in the sandwich, but none of these are a generally accepted correction. After analyzing the original proof of the sandwich estimate of variance, we derive a pivot-based method that uses fewer approximating assumptions than the sandwich method and consequently performs better than the sandwich and its adjustments at small sample sizes. The pivot-based confidence intervals prove better than the sandwich alternatives in multiple simulation studies in both fully parametric and semiparametric scenarios. Further, asymptotic efficiency is proven which shows that the pivot-based confidence intervals perform as well as sandwich-based confidence intervals for large sample sizes. All of these results are shown for both one-dimensional parameters of interest and then for multi-dimensional parameters of interest. This pivot is also explored in a Bayesian setting. We show how to use the pivot in a pseudo-Bayesian analysis. Analyzing this pivot and a monotonic transformation of this pivot show that not all pivots provide equal pseudo-Bayesian confidence interval coverage. From this work a principle is derived that further justifies the use of our pivot. A theoretical result is proven which can be used to determine which transformations

of our pivot will also give good confidence interval coverage. A simulation study is performed that shows our pseudo-Bayesian inference provides better confidence interval coverage than standard Bayesian inference, when the likelihood has been misspecified. Finally, a dataset obtained from the Centers for Disease Control is used as a population. We estimate confidence interval coverage generated by our pivot and the sandwich alternatives on subsamples of the dataset for several different models and show that our pivot still performs favorably with real world populations.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to University of Washington, where he has had the opportunity to work with Peter Hoff and the rest of the Department of Statistics.

DEDICATION

to my dear wife, Veronica

Chapter 1

INTRODUCTION AND LITERATURE REVIEW

Maximum likelihood estimation is a widely used technique with many benefits. Under certain regularity conditions, the estimators are asymptotically normal, and the asymptotic variance has an easily computable form. The estimators also have the principle of invariance. If $\hat{\theta}$ is the maximum likelihood estimator of θ, then the maximum likelihood estimator of a one-to-one function, $g(\theta)$, of the parameter of interest is $g(\hat{\theta})$. Maximum likelihoods estimators are also functions of sufficient statistics. Basic maximum likelihood theory can be found in any first-year graduate text on statistical theory such as Casella and Berger [1990] or Arnold [1990].

Though the method of maximum likelihood is a powerful technique for statistical inference, much of the power is lost if the data-generating model is incorrectly specified. If a statistician selects the wrong family of distributions to model the data, then the standard Fisher information-based variance estimates will likely be biased, even asymptotically. Consequently, any inference performed with the incorrect model will be adversely affected. For example true coverage rates of confidence intervals will not match nominal coverage rates even as the sample size increases.

This begs the question of why any analyst would want to use maximum likelihood or any parametric methods at all. The first of two answers is that the parameter in a misspecified model might still be the appropriate parameter to estimate. An example of this might be the measure of linear association between a covariate and the response variable in a regression model. Even if the model has been misspecified, this association might still be of interest to a researcher. Plus, given certain regularity conditions, the maximum likelihood estimator converges to a unique pseudo-true parameter. This pseudo-true parameter has

a nice interpretation: it minimizes the Kullback-Leibler distance between the true data-generating model and assumed family of models [Akaike, 1973]. In other words, the pseudo-true parameter yields the distribution from the misspecified model that is "closest" to the true data-generating distribution.

The second answer is that parametric methods increase inferential efficiency. As one moves through the spectrum of parametric, semi-parametric, and non-parametric methods, assumptions about the data-generating model decrease. The trade-off for fewer assumptions is that the efficiency of one's inference decreases as well. Using fully empirical methods clearly requires the fewest assumptions of the true data-generating model, but the cost in efficiency might be too high for a given application. Once the decision to use parametric or semi-parametric methods has been made, the question then boils down to one of how to perform inference about the pseudo-true parameter.

The standard method for creating confidence intervals using misspecified models was introduced in Huber [1967] and expanded in White [1982]. It is commonly referred to as the "sandwich" estimator of variance[1], so named because the matrix version of this estimator, as we will see in Section 3.1, has the form $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ where the outer matrices are the "bread" and the inner matrix is the "peanut butter". The main benefit of the sandwich is that, given a standard set of regularity conditions, it provides asymptotically correct variance estimates for maximum likelihood parameter estimates. This yields asymptotically correct confidence intervals for pseudo-true parameters. Additionally, it is applicable across a wide range of statistical techniques and is not limited to fully parametric, maximum likelihood methods [Chen and Fan, 2005], [Huber and Ronchetti, 2009].

While the sandwich does provide asymptotically correct inference, the small sample confidence interval coverage is typically quite poor. This has been attributed to the increased variability in the sandwich estimator over other, less robust variance estimators [Kauermann and Carroll, 2001]. Specifically, Efron [1986] in his discussion of Wu [1986] noticed greater

[1]To the best of the author's knowledge, the first published use of the term "sandwich" was made by Lin and Wei [1989].

variability in the sandwich estimator than in the standard variance estimator in his simulation results. Breslow [1990] also finds greater variability in the sandwich estimator in his simulations of overdispersed Poisson regression. Firth [1992] and McCullagh [1992] in their review of Liang et al. [1992] both question the small sample efficiency of the sandwich. And Diggle et al. [2009] notes on p.80 that best results are obtaineds when there are "many experimental units". The result of greater variability in typical small sample scenarios is that nominal 95% confidence intervals may yield much less than 95% coverage of the pseudo-true parameter. Greater variability and consequently poor confidence interval coverage is the result both of multiple approximations being used in the derivation of sandwich-based confidence intervals and of multiple quantities being estimated.

Several authors have attempted to adjust sandwich-based confidence intervals for small sample sizes [Hinkley, 1977], [Horn et al., 1975], [Efron, 1982], [Kauermann and Carroll, 2000], [Fay and Graubard, 2001]. As remarked upon in Hardin [2003], all of these adjustments to the sandwich are "ad hoc solutions". All of them attempt to inflate confidence interval sizes, but none of them do so in a way that directly addresses the structural biases and variability inherent in the sandwich. The methods proposed by Kauermann and Carroll [2000] and Fay and Graubard [2001] do not even attempt to adjust the sandwich at all. Instead, they modify the standard normal quantiles used in confidence interval construction, or they replace asymptotic normal and chi-squared distributions with Student-t and F distributions. To date, there is no universally-agreed-upon adjustment or alternative to the sandwich for small sample sizes that corrects for the sandwich's structural problems.

This paper introduces a pivot-based method of inference for use when model misspecification has been assumed. Pivots are a natural choice for inference. They are functions of data and parameters, the distribution of which is independent of said parameters. The first pivot any student of statistics learns is $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$, and it is the basis for building confidence intervals and constructing hypothesis tests. Pivots can also be used in more complicated scenarios. For example Parzen et al. [1994] introduces a Monte Carlo method using pivots to generate confidence intervals for the parameter of interest. Though for practical concerns, they

restrict their application to single-dimension parameters and situations where the pivotal quantity and the parameter are in a one-to-one relationship.

This pivot method was developed as a result of analyzing the proof of the sandwich estimate of variance. That proof involves several approximating assumptions in order to provide an asymptotically consistent variance estimate. Our idea was to eliminate as many approximating assumptions as possible in order to make more accurate inference. Our work allows us to eliminate the plug-in assumptions in the original sandwich proof, with the end result being a pivot-based method of inference. Our pivot provides better inference in small sample scenarios than is possible with the sandwich or its current adjustments. Additionally, our pivot is asymptotically as efficient as the sandwich, which means our pivot performs as well as the sandwich at large sample sizes.

Our pivot is not just restricted to use in frequentist inference. While a proper Bayesian analysis is not possible with pivots [Monahan and Boos, 1992], a pseudo-Bayesian analysis is possible. Our work with pivots led us to devise a method of pseudo-Bayesian analysis using our pivot. When one begins to use pivots in Bayesian analysis, another issue that arises is the question of which pivot to use. Based on our pseudo-Bayesian procedure, we devised a principle of choosing which pivot to use in this pseudo-Bayesian analysis. As a result, our pivot is a natural choice according to the principle.

The rest of the dissertation is organized in the following manner. In Chapter 2, we derive our pivot-based method. By avoiding plug-in assumptions commonly used with sandwich-based inference, we preserve any mean-variance relationships in the model and improve confidence interval coverage in smaller sample sizes. We also give a proof in the one-parameter case that no asymptotic efficiency has been lost when using the pivot versus the sandwich. In that same section we follow up with several simulation examples that show our method provides more accurate confidence interval coverage in small sample sizes. In Chapter 3, we discuss the pivot in the mutiparameter case. We extend our proof of asymptotic equivalence to multiple parameters. We also perform a simulation exercise to demonstrate our pivot's improved inference on a simple linear regression model. Chapter 4 extends our work on

the pivot into a Bayesian context. We investigate the use of a pseudo-likelihood to perform pseudo-Bayesian inference with pivots. We also propose a principle that one can use in order to select which pivots to use in a pseudo-Bayesian analysis. This principle shows that our pivot makes an ideal choice for this pseudo-Bayesian analysis. Chapter 5 takes data from a real-world dataset and illustrates how our pivot can be used to derive inference on pseudo-true parameters in the presence of more realistic model misspecification. We use our pivot in several scenarios: a multiple regression model, a simple linear model involving survey weights, and a semiparametric copula model. Finally, in Chapter 6 we summarize the main results, mention a few of the issues encountered in our work, and discuss possible extensions.

Chapter 2

ONE-PARAMETER THEORY AND EXAMPLES

2.1 Pivot Method

In order to systematically improve the sandwich, we will first review the assumptions used in the derivation of the sandwich estimate of variance. In the original argument we find four approximating assumptions. Our approach to inference differs from the original and eliminates two of these assumptions, resulting in better confidence interval coverage at small sample sizes.

For simplicity, we describe the method for a one-parameter model under misspecification. Let $y_1, \ldots, y_n$ represent data values. Let the family of distributions assumed by the analyst be described by the density $f(y;\theta)$ with indexing parameter $\theta \in \Theta$, a compact subset of R^p. Let the true distribution of the data be desribed by density $g(y)$, with $g(y)$ not necessarily in $\{f(y;\theta) : \theta \in \Theta\}$. Let $\prod_{i=1}^n f(y_i;\theta)$ be the pseudo-likelihood of independent and identically distributed data. We use the prefix "pseudo-" to highlight the possibility that the true data-generating distribution is not in our assumed model. Let $\sum_{i=1}^n \log(f(y_i;\theta))$ be the pseudo-log-likelihood. Let $\sum_{i=1}^n l_\theta(y_i;\theta)$ be the derivative of the pseudo-log-likelihood with respect to θ and let $\sum_{i=1}^n l_{\theta\theta}(y_i;\theta)$ be the second derivative. Let $\hat{\theta}$ be the maximum misspecified likelihood estimator. Let θ^* be the Kullback-Leibler minimizing parameter, or the pseudo-true parameter. Then by a stochastic Mean Value Theorem from [Jennrich, 1969] there exists $\bar{\theta}$, a weighted average of $\hat{\theta}$ and θ^*, that satisfies the following equation

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_\theta(y_i,\theta^*) = \frac{1}{\sqrt{n}}\left[\sum_{i=1}^{n} l_{\theta\theta}(y_i,\bar{\theta})\right](\hat{\theta}-\theta^*) \tag{2.1}$$

The standard procedure is to solve for $\sqrt{n}(\hat{\theta}-\theta^*)$ and compute the variance of its asymptotic

distribution. The argument begins by applying the Central Limit Theorem to the left-hand side of line (2.1) to get the following approximate distribution

$$\frac{1}{\sqrt{n}} \sum l_\theta(y_i, \theta^*) \dot{\sim} N(0, B(\theta^*)). \tag{2.2}$$

Now we replace the left-hand side of line (2.2) with the right-hand side of line (2.1).

$$\frac{1}{\sqrt{n}} \left[\sum_{i=1}^{n} l_{\theta\theta}(y_i, \bar{\theta}) \right] (\hat{\theta} - \theta^*) \dot{\sim} N(0, B(\theta^*)). \tag{2.3}$$

If we let $A_n(\bar{\theta}) = \frac{1}{n} \sum l_{\theta\theta}(y_i, \bar{\theta})$ then the result is

$$A_n(\bar{\theta})\sqrt{n}(\hat{\theta} - \theta^*) \dot{\sim} N(0, B(\theta^*)). \tag{2.4}$$

Dividing both sides by the square root of $B(\theta^*)$ gives us

$$B^{-1/2}(\theta^*)A_n(\bar{\theta})\sqrt{n}(\hat{\theta} - \theta^*) \dot{\sim} N(0, 1) \tag{2.5}$$

Since $B(\theta^*) = E\{[l_\theta(y, \theta^*)]^2\}$ and that expectation depends on the true data-generating distribution, we will approximate $B(\theta^*)$ with $B_n(\theta^*) = \frac{1}{n} \sum_{i=1}^{n} [l_\theta(y_i, \theta^*)]^2$ which gives us

$$B_n^{-1/2}(\theta^*)A_n(\bar{\theta})\sqrt{n}(\hat{\theta} - \theta^*) \dot{\sim} N(0, 1). \tag{2.6}$$

Since θ^* and hence $\bar{\theta}$ are unknown, we will approximate them with $\hat{\theta}$.

$$B_n^{-1/2}(\hat{\theta})A_n(\hat{\theta})\sqrt{n}(\hat{\theta} - \theta^*) \dot{\sim} N(0, 1). \tag{2.7}$$

We can write this more explicitly as

$$\sqrt{n}(\hat{\theta} - \theta^*) \dot{\sim} N(0, A_n(\hat{\theta})^{-1} B_n(\hat{\theta}) A_n(\hat{\theta})^{-1}) \tag{2.8}$$

where the plug-in form of the sandwich estimate of variance is $A_n(\hat{\theta})^{-1}B_n(\hat{\theta})A_n(\hat{\theta})^{-1}$.

Aggregating all the approximating assumptions made in the above argument, we have the following four assumptions:

1. $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_\theta(y_i, \theta^*) \dot{\sim} N(0, B(\theta^*))$

2. $\frac{1}{n}\sum_{i=1}^{n}[l_\theta(y_i, \theta^*)]^2 = B_n(\theta^*) \approx E\{[l_\theta(y, \theta^*)]^2\} = B(\theta^*)$

3. $\hat{\theta} \approx \theta^*$

4. $\hat{\theta} \approx \bar{\theta}$

We take a moment here to describe some of the ad hoc methods mentioned in the introduction. The simplest adjustment is to multiply the standard sandwich estimator by $n/(n-p)$ where p is the number of parameters in the model and n is the size of the dataset [Hinkley, 1977]. This is analogous to using a degrees of freedom correction to yield unbiased estimates of variance. We call this adjustment "hc2" while we reserve "hc1" for an earlier but more complicated adjustment. Another early attempt to correct for the bias of the sandwich's variance estimate was introduced by Horn et al. [1975]. This correction divided the i^{th} diagonal element in the inner, "peanut butter" matrix by $1 - h_{ii}$, where h_{ii} is the i^{th} diagonal element from the hat matrix. If no heteroscedasticity is present, then the expected value of the residuals differ from the assumed σ^2 by just this factor [MacKinnon and White, 1985]. We call this adjustment "hc1". The jackknife estimator was used by Efron [1982], which is equivalent to dividing the i^{th} diagonal element in the inner, "peanut butter" matrix by $(1 - h_{ii})^2$. We call this adjustment "hc3". In Kauermann and Carroll [2000] an adjusted normal quantile value is computed, instead of adjusting the sandwich directly. Here the confidence interval width is increased by changing quantiles, not by computing a larger standard error. We call this adjustment "hc4". Fay and Graubard [2001] compute the degrees of freedom needed for approximately correct $t-$ and $F-$distributions, instead of using asymptotically correct normal and chi-squared distributions. Similar to hc4, confidence interval

widths are increased by changing the distributions of the quantiles used, not by changing the standard errors. We call this adjustment "hc5".

None of these ad hoc methods is a universally agreed upon adjustment to the sandwich. The hc1 method is universally applicable, but when the sandwich provides much less than 95% coverage then hc1 is only a small improvement. The hc2 method comes closest to being a theoretically justified alternative, but is only justified when heteroscedasticity is absent. Homoscedasticity is not always the case. The hc3 method, being equivalent to the jackknife estimator of variance, is also widely applicable. However, being a very general-purpose estimator means that a more specific method might perform better. The hc4 method uses the extra variance in the sandwich to estimate the correct quantile to use. Unfortunately, this requires one to estimate the variance of the sandwich estimate of variance. Finally, the hc5 method estimates approximately correct $t-$ and $F-$distributions, but assumes extra approximations in the process.

Our goal is to eliminate as many of the approximating assumptions as possible in order to reduce variability, yet still maintain the ability to perform inference about θ^*. We want an alternative method to the sandwich in small sample sizes that is also solidly justified by statistical theory and widely applicable. We remind the reader that our goal is not to find the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta^*)$, as this is only a tool for performing inference about θ^* and is not absolutely necessary. We start with the pseudo-likelihood in line (2.2), however, since we are operating under the assumption of model misspecification, we have no knowledge of the exact distribution of this pseudo-likelihood. The best we can do is to keep the assumption of approximate normality by appealing to the Central Limit Theorem using assumption (1.) above. This may be a crude approximation, but it is the only distributional result we have. Model misspecification also means we cannot compute the asymptotic variance $B(\theta)$, so we estimate it with its empirical approximation, $B_n(\theta)$, which is assumption (2.) from above. With these assumptions alone, we can still perform inference using pivots.

We will now derive our pivot and its approximate, asymptotically correct distribution in

the following lines. We start in the same place as the sandwich argument

$$\frac{1}{\sqrt{n}} \sum l_\theta(y_i, \theta^*) \dot{\sim} N(0, B(\theta^*)). \tag{2.9}$$

At this point we divide both sides of the distributional equation by $B^{-1/2}(\theta^*)$ to get

$$B^{-1/2}(\theta^*) \frac{1}{\sqrt{n}} \sum l_\theta(y_i, \theta^*) \dot{\sim} N(0, 1). \tag{2.10}$$

At this point we replace $B^{-1/2}(\theta^*)$ with its empirical approximation

$$B_n^{-1/2}(\theta^*) \frac{1}{\sqrt{n}} \sum l_\theta(y_i, \theta^*) \dot{\sim} N(0, 1). \tag{2.11}$$

And now we explicitly rewrite $B_n^{-1/2}(\theta^*)$ in terms of the derivative of the log-likelihood

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} [l_\theta(y_i, \theta^*)]^2 \right\}^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_\theta(y_i, \theta^*) \dot{\sim} N(0, 1). \tag{2.12}$$

This is the pivot we will use for inference. In this argument we were able to eliminate two of the four approximating assumptions that were present in the sandwich argument. The assumptions made in deriving our pivot are

1. $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_\theta(y_i, \theta^*) \dot{\sim} N(0, B(\theta^*))$

2. $\frac{1}{n} \sum_{i=1}^{n} [l_\theta(y_i, \theta^*)]^2 = B_n(\theta^*) \approx E\{[l_\theta(y, \theta^*)]^2\} = B(\theta^*)$

Note that we had to keep assumption (1.), the distributional result. We also kept assumption (2.), the empirical approximation. In other words, we need assumptions to counter the fact that we don't know the true data-generating model. What we were able to eliminate were the plug-in assumptions (3.) and (4.) from the sandwich argument.

In order to generate a $100(1-\alpha)\%$ one-sided confidence interval with the pivot in (2.12), we first choose α. We next compute the standard normal quantile $z_{1-\alpha}$. Then, we perform

a linear search to find $\theta_{1-\alpha}$ that satisfies

$$\left\{\frac{1}{n}\sum_{i=1}^{n}[l_\theta(y_i,\theta_{1-\alpha})]^2\right\}^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}l_\theta(y_i,\theta_{1-\alpha})=z_{1-\alpha}. \tag{2.13}$$

If we want a two-sided confidence interval, we perform the same linear search method for the $z_{\alpha/2}$ and $z_{1-\alpha/2}$ quantiles to generate $\theta_{\alpha/2}$ and $\theta_{1-\alpha/2}$, the boundaries of our two-sided confidence interval.

2.2 One-Parameter Asymptotics

We will show that the confidence interval procedure based on pivot (2.12) is asymptotically as efficient as the confidence interval based on the plug-in sandwich estimator. This is important because if the pivot were only better than the sandwich at small sample sizes, but lost efficiency with respect to the sandwich at larger sample sizes, we would have no practical way of determining which method to use in order to perform the best inference we could. The standard way of proving asymptotic relative efficiency is to compute the asymptotic variance of each estimator and then compare those variances. For the likelihood pivot, there is no explicit estimator computed. Thus it is impossible to compute an asymptotic variance of an estimator which does not exist, much less compare that variance to the sandwich. As an alternative approach, we compare the confidence interval boundaries generated by the two procedures. Specifically we will show

$$\sqrt{n}|\theta_{1,n}-\theta_{2,n}|\rightarrow_{a.s.}0$$

where $\theta_{1,n}$ and $\theta_{2,n}$ are defined by the equalities

$$\sqrt{n}B_n(\theta_{1,n})^{-1/2}A_n(\bar{\theta})(\hat{\theta}-\theta_{1,n})=z_{1-\alpha} \tag{2.14}$$

$$\sqrt{n}B_n(\hat{\theta})^{-1/2}A_n(\hat{\theta})(\hat{\theta}-\theta_{2,n})=z_{1-\alpha}, \tag{2.15}$$

which are the boundaries of the one-sided confidence interval using the pivot, line (2.14), and the sandwich, line (2.15).

The reason for this approach is intuitive. If two estimators have asymptotically normal distributions, then relative efficiency of the estimators is equivalent to comparing $\sqrt{n}$ times the confidence interval width. If the widths are the same, then the asymptotic variances are the same, and the estimators are equally efficient. If the widths are different, then the shorter confidence interval is more efficient. Even more specific to our problem, we are comparing confidence interval boundaries directly. If those boundaries converge on the $\sqrt{n}$-scale, then the confidence intervals are asymptotically equal to each other, which also results in equal efficiency.

Expressions that we will use for each of $\theta_{1,n}$ and $\theta_{2,n}$ are

$$\theta_{1,n} = \hat{\theta}_n - \frac{A_n((1-b_n)\hat{\theta}_n + b_n\theta_{1,n})^{-1}B_n(\theta_{1,n})^{1/2}z_{1-\alpha}}{\sqrt{n}} \tag{2.16}$$

$$\theta_{2,n} = \hat{\theta}_n - \frac{A_n(\hat{\theta})^{-1}B_n(\hat{\theta})^{1/2}z_{1-\alpha}}{\sqrt{n}} \tag{2.17}$$

where b_n is the value between 0 and 1 that satisfies the Lemma 3 equality in [Jennrich, 1969]. Notice that $\theta_{1,n}$ is not solved explicity in terms of other quantities. This is not a problem. We will be considering asymptotic behavior, and we prove in the appendix that $\theta_{1,n} \to_{a.s.} \theta^*$. Now consider the scaled difference

$$\begin{aligned}
\sqrt{n}|\theta_{1,n} - \theta_{2,n}| &= \sqrt{n}\left|\hat{\theta}_n - \frac{A_n((1-b_n)\hat{\theta}_n + b_n\theta_{1,n})^{-1}B_n(\theta_{1,n})^{1/2}z_{1-\alpha}}{\sqrt{n}}\right. \\
&\quad \left. - \left(\hat{\theta}_n - \frac{A_n(\hat{\theta})^{-1}B_n(\hat{\theta})^{1/2}z_{1-\alpha}}{\sqrt{n}}\right)\right| && (2.18)\\
&= \left|A_n(\hat{\theta})^{-1}B_n(\hat{\theta})^{1/2}z_{1-\alpha} - A_n((1-b_n)\hat{\theta}_n + b_n\theta_{1,n})^{-1}B_n(\theta_{1,n})^{1/2}z_{1-\alpha}\right| && (2.19)\\
&\to_{a.s.} z_{1-\alpha}\left|A(\theta^*)^{-1}B(\theta^*)^{1/2} - A(\theta^*)^{-1}B(\theta^*)^{1/2}\right| && (2.20)\\
&= 0. && (2.21)
\end{aligned}$$

Line (2.18) is the result of replacing the left-hand side with the equivalent expressions in (2.16) and (2.17). Line (2.19) is the result of algebra. Line (2.20) is the consequence of several convergence results, the first being $\theta_{1,n} \rightarrow_{a.s.} \theta^*$. By the strong law of large numbers and Lemma 3.1 from White [1981] we have $A_n(\hat{\theta}) \rightarrow_{a.s.} A(\theta^*)$, $A_n(\bar{\theta}) \rightarrow_{a.s.} A(\theta^*)$, $B_n(\hat{\theta}) \rightarrow_{a.s.} B(\theta^*)$, and $B_n(\theta_{1,n}) \rightarrow_{a.s.} B(\theta^*)$; and by Mann-Wald we have line (2.20) above. Since the $\sqrt{n}$-distance between confidence interval boundaries converges to zero, the confidence intervals generated by the pivot and the sandwich are asymptotically equivalent. Hence, the two procedures for generating confidence intervals are asymptotically the same.

2.3 One-Parameter Simulation Examples

In the next four subsections, we will take a look at several examples of the pivot. The first is an example with a discrete data. The second example assumes continuous data. The third example is a regression example. The final example is a semiparametric copula example.

2.3.1 Poisson Likelihood

In this example, we assume a working model

$$y_i \overset{i.i.d.}{\sim} Pois(\theta) \qquad i = 1, \ldots, n \tag{2.22}$$

and we wish to perform inference on the parameter of interest θ. A typical misspecification would be an overdispersed Poisson. We can generate such a distribution using the Negative Binomial distribution, since the variance of a Negative Binomial random variable can differ from its mean. The Negative Binomial is what we used as our true data-generating distribution.

In this model we have the following pieces of Equations (2.1) and (2.12)

$$l_\theta(\theta^*) = \frac{1}{\theta^*}\sum_{i=1}^{n} y_i - n \tag{2.23}$$

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.24}$$

$$l_{\theta\theta}(\bar{\theta}) = \frac{-1}{(\bar{\theta})^2}\sum_{i=1}^{n} y_i \tag{2.25}$$

We then plug those into the pivot expresion to get

$$\frac{1}{\sqrt{n}} B_n(\theta^*)^{-1/2} l_\theta(y_1, \ldots, y_n; \theta^*) = \frac{1}{\sqrt{n}} \frac{\frac{1}{\theta^*}\sum_{i=1}^{n} y_i - n}{\sqrt{\frac{1}{(\theta^*)^2}\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^*)^2}} \tag{2.26}$$

$$= \sqrt{n} \frac{\frac{1}{\theta^*}\frac{1}{n}\sum_{i=1}^{n} y_i - 1}{\frac{1}{|\theta^*|}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^*)^2}} \tag{2.27}$$

$$= \sqrt{n} \frac{\frac{1}{n}\sum_{i=1}^{n} y_i - \theta^*}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^*)^2}} \tag{2.28}$$

and we use the linear search method on this pivot to find our confidence interval.

The Sandwich Variance Estimator is calculated as

$$A_n(\hat{\theta})^{-1}B_n(\hat{\theta})A_n(\hat{\theta})^{-1} = \left[\frac{-1}{n}\sum_{i=1}^{n}\frac{y_i}{\hat{\theta}^2}\right]^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\hat{\theta}}-1\right)^2\left[\frac{-1}{n}\sum_{i=1}^{n}\frac{y_i}{\hat{\theta}^2}\right]^{-1} \tag{2.29}$$

$$= \left[\frac{-1}{n}\sum_{i=1}^{n}\frac{y_i}{\hat{\theta}^2}\right]^{-2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\hat{\theta}}-1\right)^2 \tag{2.30}$$

$$= \left[\frac{-\hat{\theta}}{\hat{\theta}^2}\right]^{-2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\hat{\theta}}-1\right)^2 \tag{2.31}$$

$$= \left[\frac{-1}{\hat{\theta}}\right]^{-2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\hat{\theta}}-1\right)^2 \tag{2.32}$$

$$= \hat{\theta}^2\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\hat{\theta}}-1\right)^2 \tag{2.33}$$

$$= \hat{\theta}^2\frac{1}{n\hat{\theta}^2}\sum_{i=1}^{n}\left(y_i-\hat{\theta}\right)^2 \tag{2.34}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{\theta}\right)^2 \tag{2.35}$$

From the above, we see that the Sandwich Variance Estimator estimates the variance of $\hat{\theta}$ as $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\theta})^2$, while the pivot variance factor is $\frac{1}{n}\sum_{i=1}^{n}(y_i-\theta^*)^2$. The only difference in this example is the substitution of θ^* in the pivot quantity with $\hat{\theta}$ in the Sandwich Variance Estimator.

For the simulation study, I used a negative binomial with mean 3 and variance 3.9 as the true data-generating model. I then generated $10,000$ samples of size ranging from 5 to 30. For each sample, I constructed a 95% confidence interval using the Sandwich estimator (95% asymptotically correct) and by conducting a linear search on the parameter in question to find the parameter values that caused the pivot to equal the 2.5% and 97.5% quantiles of the standard normal distribution. The results can be seen in Figure 2.1

The pivot-based confidence intervals performs well at all sample sizes. Coverage hovers around 95% throughout all sample sizes tested. Standard maximum likelihood theory undercovers the true parameter, a reflection of the fact that we are assuming the mean and

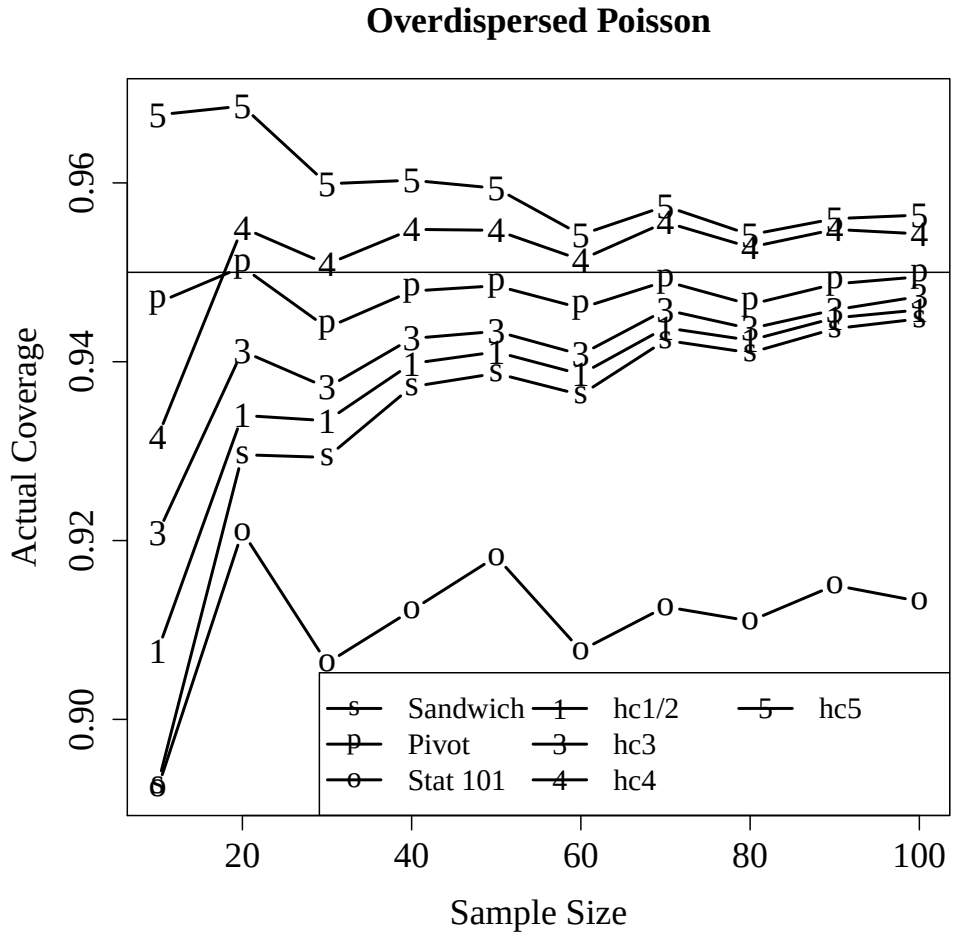

Figure 2.1: Results from simulation study using sandwich-based confidence intervals and confidence intervals generated by the proposed pseudo-likelihood pivot quantity when applied to an assumed Poisson data-generating model.

variance are equal when the variance is actually larger. The sandwich estimate of variance performs as one would expect. It undercovers the true mean in all sample sizes tested, but coverage improves as the sample size increases. The ad hoc methods tend to mirror this as well. The ad hoc method associated with [Fay and Graubard, 2001] tends to overcover the true parameter. This method consistently provided overcoverage in all simulation studies performed for this dissertation. The method uses an adjusted sandwich and the Student t-distribution quantiles instead of the normally distributed quantiles. This two-pronged approach might be over-correcting the sandwich's undercoverage, resulting in the overcoverage we see in the simulation results.

2.3.2 Exponential Likelihood

In this example we assume a working model

$$y_i \overset{i.i.d.}{\sim} Exp(\theta) \qquad i = 1, \ldots, n \tag{2.36}$$

and we wish to perform inference on θ. A typical misspecification is one in which the mean-variance relationship in the Exponential distribution doesn't hold, which can be modeled with the more general Gamma distribution.

In this model, we have the following derivatives of the log-likelihood and MMLE

$$l_\theta(y; \theta^*) = \frac{1}{(\theta^*)^2} \sum_{i=1}^{n} y_i - \frac{n}{\theta^*} \tag{2.37}$$

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2.38}$$

$$l_{\theta\theta}(y; \bar{\theta}) = \frac{-2}{\bar{\theta}^3} \sum_{i=1}^{n} y_i + \frac{n}{\bar{\theta}^2} \tag{2.39}$$

Thus the pivot we will use in practice is

$$\frac{1}{\sqrt{n}} B_n(\theta^*)^{-1/2} l_\theta(y_1, \ldots, y_n; \bar{\theta}) = \frac{1}{\sqrt{n}} \frac{\frac{1}{(\theta^*)^2} \sum_{i=1}^n y_i - \frac{n}{\theta^*}}{\sqrt{\frac{1}{(\theta^*)^4} \cdot \frac{1}{n} \sum_{i=1}^n (y_i - \theta^*)^2}} \tag{2.40}$$

$$= \sqrt{n} \frac{\frac{1}{(\theta^*)^2} \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{\theta^*}}{\frac{1}{(\theta^*)^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \theta^*)^2}} \tag{2.41}$$

$$= \sqrt{n} \frac{\frac{1}{n} \sum_{i=1}^n y_i - \theta^*}{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \theta^*)^2}} \tag{2.42}$$

The sandwich estimator of variance we will use is

$$A_n(\hat{\theta})^{-1} B_n(\hat{\theta}) A_n(\hat{\theta})^{-1} = \left[\frac{-2}{\hat{\theta}^3} \frac{1}{n} \sum_{i=1}^n y_i + \frac{1}{\hat{\theta}^2} \right]^{-1} \left[\frac{1}{n} \sum_{i=1}^n \left(\frac{y_i}{(\hat{\theta})^2} - \frac{1}{\hat{\theta}} \right)^2 \right] \left[\frac{-2}{\hat{\theta}^3} \frac{1}{n} \sum_{i=1}^n y_i + \frac{1}{\hat{\theta}^2} \right]^{-1} \tag{2.43}$$

$$= \left[\frac{-2}{\hat{\theta}^3} \frac{1}{n} \sum_{i=1}^n y_i + \frac{1}{\hat{\theta}^2} \right]^{-2} \left[\frac{1}{n} \sum_{i=1}^n \left(\frac{y_i}{(\hat{\theta})^2} - \frac{1}{\hat{\theta}} \right)^2 \right] \tag{2.44}$$

$$= \left[\frac{-2\hat{\theta} + \hat{\theta}}{\hat{\theta}^3} \right]^{-2} \frac{1}{(\hat{\theta}^4} \left[\frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{\theta} \right)^2 \right] \tag{2.45}$$

$$= \left[\frac{\hat{\theta}^3}{-\hat{\theta}} \right]^2 \frac{1}{(\hat{\theta})^4} \left[\frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{\theta} \right)^2 \right] \tag{2.46}$$

$$= \frac{\hat{\theta}^4}{(\hat{\theta})^4} \left[\frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{\theta} \right)^2 \right] \tag{2.47}$$

$$= \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{\theta} \right)^2 \tag{2.48}$$

which is the same variance estimate up to $n/(n-1)$ as one might use when calculating the sample variance in an introductory statistics course without any model assumptions.

Comparing the two estimates of variance, we see that the sandwich estimator of variance is $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{\theta})^2$, while the pivot variance factor is $\frac{1}{n} \sum_{i=1}^n (y_i - \theta^*)^2$. Again, just as in the Poisson example, the only difference is the replacement of $\hat{\theta}$ by θ^*.

I simulated this example by assuming Exponentially distributed data when the true data-generating mechanism is Gamma with mean 3 and variance 6. I simulated sample sizes from 5 to 30 and drew random samples $10,000$ times each. For each sample I calculated confidence intervals using both the Sandwich Variance Estimator and the pivot quantity. The results can be seen in Figure 2.2.

The pivot performs better than the sandwich in all sample sizes tested, but overall it undercovers the true parameter a little, though pivot-based coverage rates were between 92% and 95% for all sample sizes tested. The sandwich performs as expected: undercoverage of the true parameter in small sample sizes, with coverage improving as sample size increases. The standard maximum likelihood estimator undercovers so dramatically, it is not even shown on the graph. Coverage rates using the standard maximum likelihood method were less than 60% at the smallest n and fell throughout the sample sizes tested. This is due to the fact that we used the same estimate of the mean for the standard deviation in this model. The ad hoc methods one and two were equivalent. The first three ad hoc methods tend to outperform the sandwich, but perform worse when compared to the pivot method. Ad hoc methods four and five actually do pretty well, being the only methods that provided coverage rates that touched the 95% line, but show some slight overcoverage in sample sizes greater than $n = 30$.

2.3.3 Regression Through the Origin

Since linear regression is one of the most powerful and widely used methods in statistics, we will now take a look at a single-parameter-of-interest regression through the origin example. We assume as our working model the two-parameter regression

$$y_i \sim \theta x_i + \epsilon_i \tag{2.49}$$

$$\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2) \qquad i = 1, \ldots, n \tag{2.50}$$

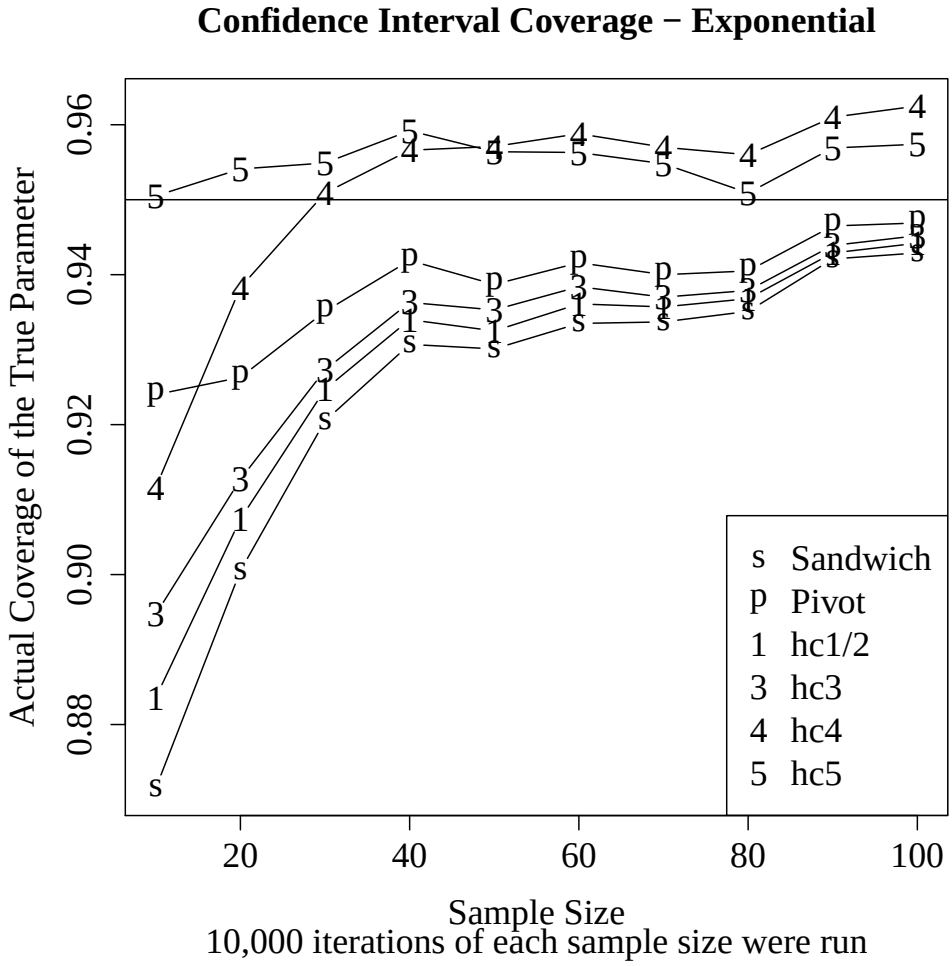

Figure 2.2: Results from simulation study of sandwich-based confidence intervals and confidence intervals generated by our pivot when applied to an assumed Exponentially distributed data-generating model.

where the parameter of interest is θ and σ^2 is a nuisance parameter. In this example σ^2 ends up cancelling in the pivot computation, so we do not need to do anything special to account for it. We let our model misspecification be one of heteroscedastic errors. Specifically the true errors are

$$\epsilon_i \sim N(0, 1+ \mid x_i \mid) \qquad i = 1, \ldots, n \tag{2.51}$$

so that the errors depend directly on the covariate. We then consider the coverage rate of the pseudo-true parameter by sandwich-based and pivot-based confidence intervals.

Under this model, the relevant sandwich and pseudo-likelihood pivot quantities are

$$\sum_{i=1}^{n} l(y_i, \theta) = -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta x_i)^2 \tag{2.52}$$

$$\sum_{i=1}^{n} l_\theta(y_i, \theta) = \frac{-1}{2\sigma^2} \sum_{i=1}^{n} -2x_i(y_i - \theta x_i) = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i(y_i - \theta x_i) \tag{2.53}$$

$$\sum_{i=1}^{n} l_\theta(y_i, \theta)^2 = \frac{1}{\sigma^4} \sum_{i=1}^{n} x_i(y_i - \theta x_i)^2 \tag{2.54}$$

$$\sum_{i=1}^{n} l_{\theta\theta}(y_i, \theta) = \frac{-1}{\sigma^2} \sum_{i=1}^{n} x_i^2 \tag{2.55}$$

respectively. Plugging these into (2.12), gives us the pivot for this example

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} [l_\theta(y_i, \theta^*)]^2 \right\}^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_\theta(y_i, \theta^*) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i y_i - \theta^* x_i^2)}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i y_i - \theta^* x_i^2)^2}}. \tag{2.56}$$

The sandwich estimate of variance is given by

$$A_n(\hat{\theta})^{-1} B_n(\hat{\theta}) A_n(\hat{\theta})^{-1} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i(y_i - \hat{\theta} x_i) \right) \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right)^{-1} \tag{2.57}$$

We simulated data for samples sizes ranging from 10 to 100. For each sample size, we generated $10,000$ datasets. The covariate was generated by random pulls from a standard

normal distribution, while the response is generated according to the above specified model. For each dataset, we computed 95% confidence intervals and determined whether or not the true value of the parameter was covered by the intervals. For comparison, we computed confidence intervals with the pivot, the sandwich, the five ad hoc methods mentioned in the introduction, and standard Wald confidence intervals assuming homoscedastic errors. The results are displayed graphically in Fig. 2.3.

The results of the simulation are what we expected. The standard regression confidence intervals undercover the pseudo-true parameter throughout the range of sample sizes. Since the assumed model does not account for the variability in errors, we expect confidence intervals to be too small and coverage of the true parameter to be less than nominal. The sandwich-based confidence intervals yield about 82% coverage at the smallest sample size simulated, and increase coverage as sample size increases. Since the sandwich has known bias, this is not surprising. Three of the five ad hoc methods hc1, hc2, and hc3 are simply scaled-up versions of the sandwich. As expected they yield coverage similar to but slightly better than the sandwich. The hc4 method uses the sandwich variance but with a slightly wider quantile spread (e.g. the 98% quantile instead of 97.5% quantile). As such it also covers the true parameter with similar to but slightly better coverage than the regular sandwich. The hc5 method yields similar but slightly higher coverage than the pivot. Not only does it use an adjusted sandwich variance estimate, but it also uses Student-t quantiles instead of standard normal. This combination pushes coverage much higher than unadjusted sandwich coverage. The pivot-based intervals perform the best out of all options tested and yield coverage of the pseudo-true parameter close to 95% throughout the range of sample sizes tested.

To test the sensitivity of our method to the distribution of the covariate, we ran the same simulation as before, but now using an exponential distribution with mean three to generate the covariate values. Compared to the standard normal, the exponential distribution with mean three is skewed and only generates non-negative values. Additionally, larger values will be expected which will allow for larger variance in the response, resulting in possible

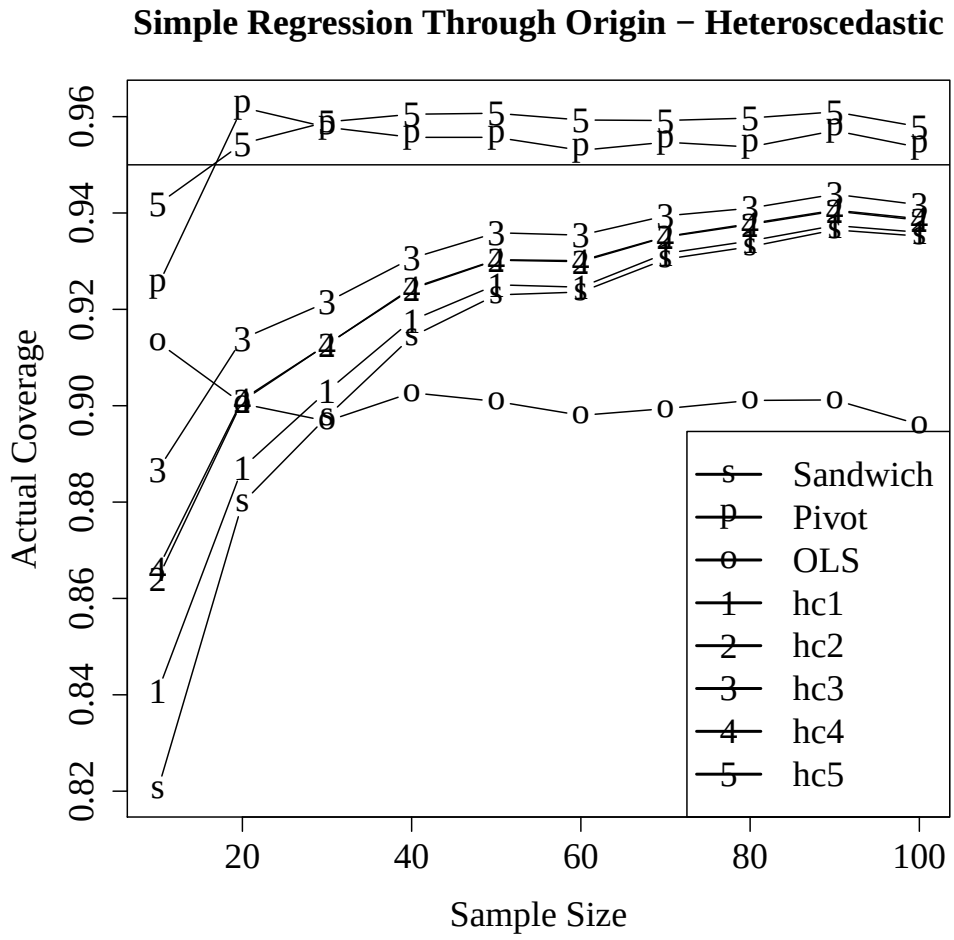

Figure 2.3: Graph showing actual confidence interval coverage of the true slope parameter when using various methods to compute confidence intervals. For each sample size, 10,000 random datasets were generated, maximum misspecified likelihood estimates were calculated, and confidence intervals were computed. The covariate here has a standard normal distribution.

influential points. Results from this simulation can be seen in Figure 2.4.

The results are what we expect. The regular maximum likelihood confidence intervals undercover the true parameter at worse rates than before. This is due to the change in covariate distribution. The sandwich and alternatives hc1-hc4 perform as expected with low coverage at small n and increasing coverage as n increases. In this simulation, the pivot and hc5 perform the best of all the methods tested, and they yield very similar coverage rates at sample sizes $n = 20$ and higher.

2.3.4 Semiparametric Copula

Semiparametric copula models allow for a blend of empirical and parametric analysis. Perhaps an analyst is interested in the association relationship between two random variables, but there is insufficient data to model both that relationship and the marginal distributions parametrically. In a case such as this, the marginal distributions of the random variables are replaced with empirical estimates, but the dependency relationship is modelled parametrically with a copula family. At that point, one can use maximum likelihood theory to estimate the copula parameter. An estimator of the variance of the maximum likelihood estimator in these scenarios was derived in Bhuchongkul [1964] and Ruymgaart [1974]. A plug-in variance estimator was suggested in Genest et al. [1995]. A modified sandwich estimator for specific use in misspecified models was introduced in Chen and Fan [2005].

Copula modelling is a natural place to use semiparametric estimation. A copula is a mathematical function of two or more random variables that represents the joint distribution of those random variables. The copula itself is invariant to monotonic transformations of the marginal data. For this reason, copulas are usually defined on the unit square since every random variable can be mapped monotonically to $[0, 1]$ through its cumulative distribution function. In the bivariate case, we have data $x_{1,1}, \ldots x_{1,n}$ and $x_{2,1}, \ldots x_{2,n}$ from random variables X_1 and X_2 with marginal distribution functions $F_{X_1}(x)$ and $F_{X_2}(x)$. Let $U_1 = F_{X_1}(X_1)$, and $U_2 = F_{X_2}(X_2)$. Then the copula is defined as $C(u_1, u_2) = P(U_1 \leq u_1, U_2 \leq u_2)$. We write the copula density as $c(u_1, u_2) = \partial^2 C(u_1, u_2)/\partial u_1 \partial u_2$. A complete treatment of

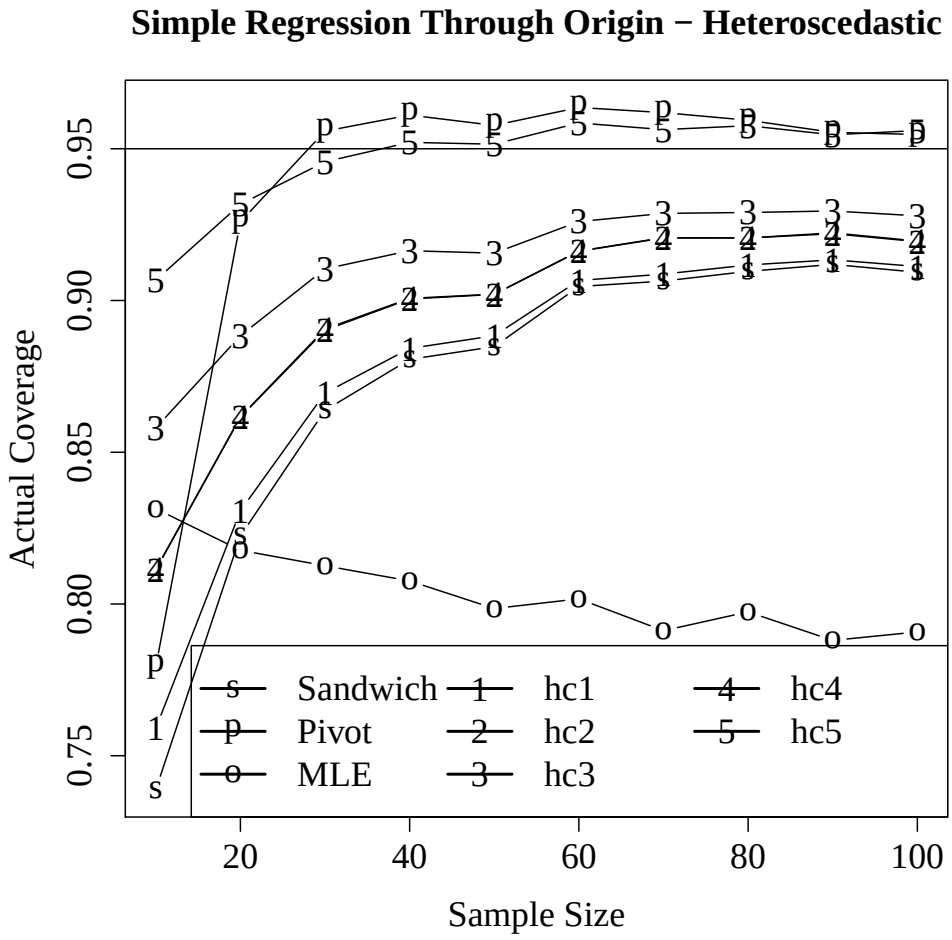

Figure 2.4: Graph showing actual confidence interval coverage of the true slope parameter when using various methods to compute confidence intervals. For each sample size, 10,000 random datasets were generated, maximum misspecified likelihood estimates were calculated, and confidence intervals were computed. The covariate here has an exponential distribution with mean three.

basic copula theory can be found in Nelsen [2006] or Joe [1997].

Under the semiparametric copula scenario we are uninterested in the marginal distributions. If we did know the marginal distributions, we could compute $u_{1,1}, \ldots, u_{1,n}$ and $u_{2,1}, \ldots, u_{2,n}$ from $x_{1,1}, \ldots x_{1,n}$ and $x_{2,1}, \ldots x_{2,n}$ as realizations of U_1 and U_2 which were defined in the previous paragraph. However, we can instead use the empirical rank approximations

$$\hat{u}_{jt} = \frac{1}{n+1} \sum_{s=1}^{n} I(x_{js} \leq x_{jt}), \qquad j = 1, 2 \tag{2.58}$$

where $n + 1$ is used to avoid dealing with possible infinities. Since this is a monotonic transformation of the original data, the underlying copula model will be unaffected. And we can use the $\hat{u}_{jt}$ instead of the actual data in our computations.

In this example we use a one-parameter copula family to model the dependence relationship between two random variables. The working model we will use is a Gaussian copula model with density

$$c_G(u_1, u_2, \theta) = \frac{1}{\sqrt{1-\theta^2}} \exp\left(\frac{-\theta^2(\Phi^{-1}(u_1)^2 + \Phi^{-1}(u_2)^2) + 2\theta\Phi^{-1}(u_1)\Phi^{-1}(u_2)}{2(1-\theta^2)} \right) \tag{2.59}$$

We will set the true copula model as a Farlie-Gumbel-Morgenstern copula with density

$$c_{FGM}(u_1, u_2, \alpha) = 1 + \alpha - 2\alpha(u_1 + u_2) + 4\alpha u_1 u_2. \tag{2.60}$$

In both densities, u_1 and u_2 are distributed as Uniform(0,1) random variables. $\Phi(\cdot)$ is the Gaussian CDF function. Formulas for both the Gaussian and Farlie-Gumbel Morgenstern copula densities can be found in section 5.1 of Joe [1997]. The relevant derivatives of the

working log-likelihood are

$$\sum_{i=1}^{n} l_\theta(u_{1,i}, u_{2,i}, \theta) = \frac{\theta}{1-\theta^2} - \frac{\theta}{(1-\theta^2)^2}\sum_{i=1}^{n}[\Phi^{-1}(u_{1,i})^2 + \Phi^{-1}(u_{2,i})^2] + \frac{1+\theta^2}{(1-\theta^2)^2}\sum_{i=1}^{n}\Phi^{-1}(u_{1,i})\Phi^{-1}(u_{2,i}) \tag{2.61}$$

$$\sum_{i=1}^{n} l_{\theta,1}(u_{1,i}, u_{2,i}, \theta) = \sum_{i=1}^{n}\frac{(1+\theta^2)\Phi^{-1}(u_{2,i}) - 2\theta\Phi^{-1}(u_{1,i})}{(1-\theta^2)^2\phi(\Phi^{-1}(u_{1,i}))} \tag{2.62}$$

$$\sum_{i=1}^{n} l_{\theta,\theta}(u_{1,i}, u_{2,i}, \theta) = \frac{1+\theta^2}{(1-\theta^2)^2} - \frac{(1+3\theta^2)(1-\theta^2)}{(1-\theta^2)^4}\sum_{i=1}^{n}\left[\Phi^{-1}(u_{1,i})^2 + \Phi^{-1}(u_{2,i})^2\right] + \frac{2\theta(3+\theta^2)(1-\theta^2)}{(1-\theta^2)^4}\sum_{i=1}^{n}\Phi^{-1}(u_{1,i})\Phi^{-1}(u_{2,i}) \tag{2.63}$$

where $l_{\theta,1}$ is $\partial^2 l/\partial\theta\partial u_1$. Also, $\phi(\cdot)$ is the standard normal density function. Note that $l_{\theta,2}$ is not listed but is analogous to $l_{\theta,1}$.

The next steps are to construct the sandwich and pivot. The "bread" of the sandwich is defined in the usual way as

$$A(\theta) = -\mathrm{E}\left[l_{\theta\theta}(u_1, u_2, \theta)\right] \tag{2.64}$$

with the expectation taken under the truth. The "peanut butter", however, is defined as

$$B(\theta) = \mathrm{Var}\left[l_\theta(u_1, u_2, \theta) + W_1(u_1, \theta) + W_2(u_2, \theta)\right] \tag{2.65}$$

with the variance also being taken under the true distribution. The W_1 and W_2 terms are included to account for the extra variability that arises from using the empirical distribution, $\hat{u}_{jt}$, $j = 1, 2$; $t = 1, \ldots, n$, on the marginals instead of the actual data, u_{jt}, $j = 1, 2$; $t = 1, \ldots, n$. We define the function W_j for $j = 1, 2$ as

$$W_j(u_{j,t}, \theta) = \mathrm{E}\left[l_{\theta,j}(u_{1,s}, u_{2,s}, \theta)\{I(u_{j,t} \leq u_{j,s}) - u_{j,s}\} \mid u_{j,t}\right] \tag{2.66}$$

where $l_{\theta,j} = \frac{\partial^2 l}{\partial\theta\partial u_j}$ and $I(\cdot)$ is an indicator function.

Using the standard empirical approximation for $A(\theta)$ and the empirical rank approximations to $u_{1,t}$ and $u_{2,t}$, we can now define $A_n(\theta)$ as

$$A_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} l_{\theta,\theta}(\hat{u}_{1,t}, \hat{u}_{2,t}, \theta) \tag{2.67}$$

Defining $B_n(\theta)$ is a little more complicated. We start with $l_\theta(\hat{u}_{1,t}, \hat{u}_{2,t}, \theta)$ where $t = 1, \ldots, n$ and then add in estimates of $W_1(\hat{u}_{1,t}, \theta) + W_2(\hat{u}_{2,t}, \theta)$. We use

$$\hat{W}_1(\hat{u}_{1,t}, \theta) + \hat{W}_2(\hat{u}_{2,t}, \theta) = \frac{1}{n}\sum_{j=t}^{n} l_{\theta,1}\left(\frac{j}{n+1}, \frac{S_j}{n+1}, \theta\right) + \frac{1}{n}\sum_{S_j \geq S_t}^{n} l_{\theta,2}\left(\frac{j}{n+1}, \frac{S_j}{n+1}, \theta\right) \tag{2.68}$$

where S_j is the rank of $u_{2,j}$ among the u_2 data when $u_{1,j}$ is the j^{th} order statistic among the u_1 data. For any given dataset, we can then compute $B_n(\theta)$ as the sample variance of the n values $l_\theta(\hat{u}_{1,t}, \hat{u}_{2,t}, \theta) + \hat{W}_1(\hat{u}_{1,t}, \theta) + \hat{W}_2(\hat{u}_{2,t}, \theta),\ t = 1, \ldots, n$. The sandwich estimate is then computed as $A_n(\hat{\theta})^{-1} B_n(\hat{\theta}) A_n(\hat{\theta})^{-1}$. The pivot and its approximate distribution are

$$\sqrt{n} B_n(\theta)^{-1/2} \frac{1}{n}\sum_{t=1}^{n} l_\theta(\theta, \hat{u}_{1,t}, \hat{u}_{2,t}) \dot{\sim} N(0, 1). \tag{2.69}$$

For the simulations, our true populations followed a Farlie-Gumbel-Morgenstern distribution with parameter values ranging from -1 to 1 in increments of 0.25. We used sample sizes from 10 to 100 in increments of 10. For each combination of sample size and parameter value, we simulated $10,000$ datasets of that size from a Farlie-Gumbel-Morgenstern copula distribution. We then computed the maximum misspecified likelihood estimate of the parameter using the Gaussian copula model. Confidence intervals were computed using both the sandwich estimate of variance and the pivot. Coverage results can be seen in Figure 2.5.

Before describing the results, let us pause for a minute to commoent on our choice of copula truth and working model. We could have chosen any number of true and assumed

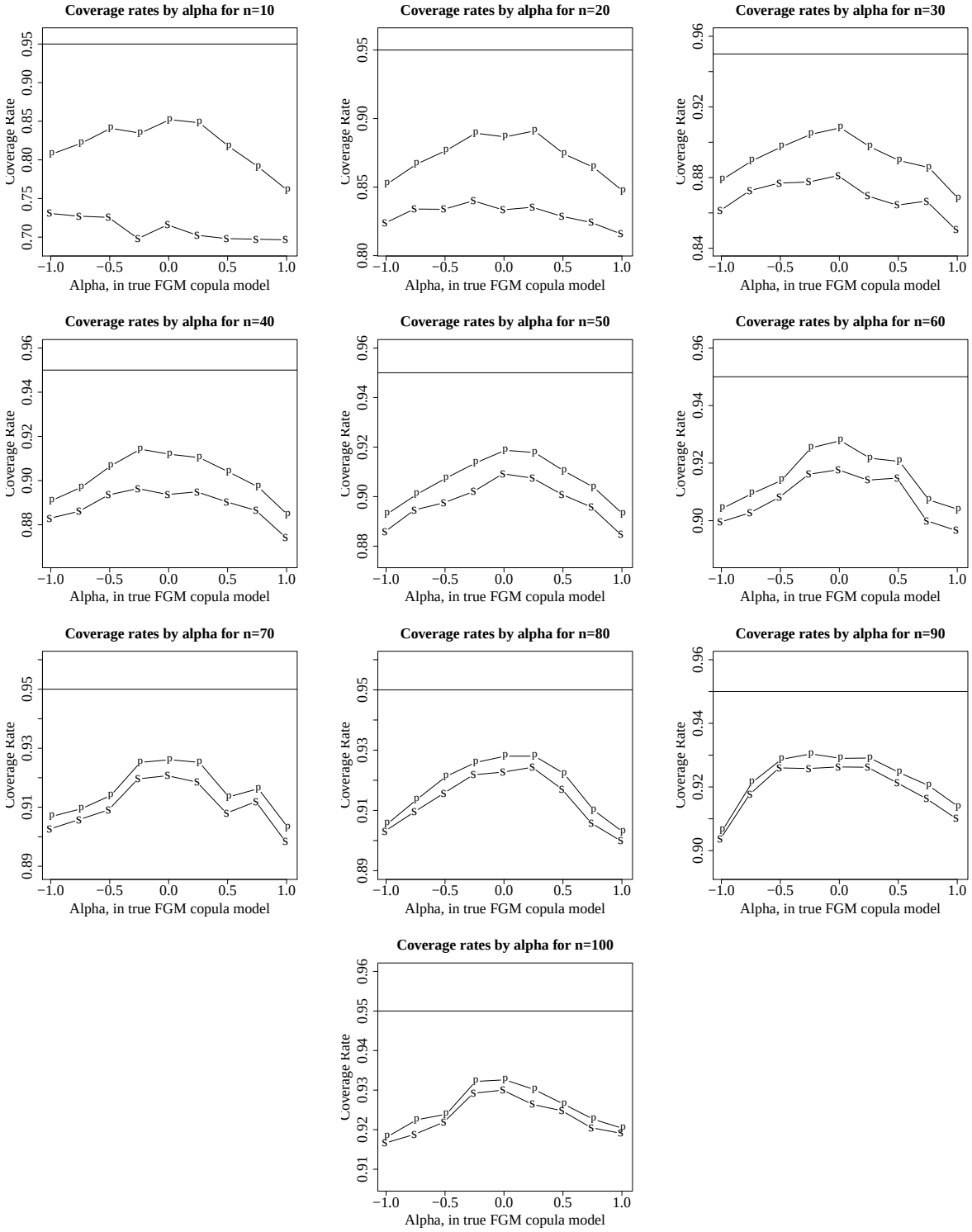

Figure 2.5: Coverage results for the sandwich and pivot in a semiparametric copula simulation. Each graph represents a different sample size. Black lines represent pivot coverage of the true parameter while red lines represent sandwich coverage of the same parameter. In all cases the true copula family was a Farlie-Gumbel-Morgenstern copula, while the assumed copula was Gaussian.

copula models, but there are good reasons for the choice we made. First, the Gaussian copula is a commonly used copula model. If the univariate distributions are normal, the joint distribution will be a multivariate normal distribution. In this two variable case, the parameter θ is the usual correlation coefficient. Second, our choice of the Farlie-Gumbel-Morgenstern copula was made so we could calculate the pseudo true parameter exactly. If the true parameter in a Farlie-Gumbel-Morgenstern distribution is $\alpha_0 \in [-1, 1]$, then the pseudo-true parameter using the Gaussian copula as the working family is $\theta^* = \alpha_0/\pi$.

The results are in line with what we have seen previously. The pivot outperforms the sandwich at small sample sizes. Coverage under each method approaches 95% as sample size increases. The coverage difference between the sandwich and the pivot decreases as sample size increases. One final note is that coverage at any given sample size seems better near $\theta = 0$ than at the extreme values of θ. This is because at $\theta = 0$, we are at independence. At independence, all copula models are the same and there is no model misspecification.

Chapter 3

MULTI-PARAMETER THEORY AND EXAMPLES

3.1 Multi-Parameter Pivot

In the multiparameter case, confidence intervals are replaced by confidence regions. This can be visualized by imagining the difference between an interval on a line versus an ellipse in the plane or a roughly egg-shaped region in three dimensional space. We will review the basic sandwich estimate of variance argument in multiple dimensions and, describe how our pivot is derived, and then explain how to generate confidence regions using the pivot.

The multivariate sandwich starts with a Central Limit Theorem approximation for the distribution of the gradient of the log-likelihood.

$$\sqrt{n}\nabla_{\boldsymbol{\theta}}\left[\frac{1}{n}\sum_{i=1}^{n} l(y_i,\boldsymbol{\theta}^*)\right] \dot{\sim} N(\mathbf{0}_p, \mathbf{B}(\boldsymbol{\theta}^*)) \tag{3.1}$$

where $\nabla_{\boldsymbol{\theta}}$ represents the gradient with respect to the p-dimensional parameter vector $\boldsymbol{\theta}$. In what follows, we will let $l_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}[\frac{1}{n}\sum_{i=1}^{n} l(y_i,\boldsymbol{\theta})]$ for ease of notation. Using an analogue to the Mean Value Theorem for multidimensional functions, we also have that there is a matrix $\tilde{\mathbf{A}}_n$ that makes the following statement true

$$-\sqrt{n}\nabla_{\boldsymbol{\theta}}\left[\frac{1}{n}\sum_{i=1}^{n} l(y_i,\boldsymbol{\theta}^*)\right] = \sqrt{n}\tilde{\mathbf{A}}_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*). \tag{3.2}$$

Combining lines (3.1) and (3.2), we have

$$\sqrt{n}\tilde{\mathbf{A}}_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \dot{\sim} N(\mathbf{0}_p, \mathbf{B}(\boldsymbol{\theta}^*)). \tag{3.3}$$

From this statement, we can construct the following quadratic form and know its approximate distribution

$$n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T \tilde{\mathbf{A}}_n \mathbf{B}(\boldsymbol{\theta}^*)^{-1} \tilde{\mathbf{A}}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \dot{\sim} \chi^2_p. \tag{3.4}$$

Since we do not know the exact distribution, we cannot calculate the matrix $\mathbf{B}(\boldsymbol{\theta})$, however we can use the empirical approximation $\mathbf{B}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum l_{\boldsymbol{\theta}}(\boldsymbol{\theta}) l_{\boldsymbol{\theta}}(\boldsymbol{\theta})^T$. This gives us

$$n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T \tilde{\mathbf{A}}_n \mathbf{B}_n(\boldsymbol{\theta}^*)^{-1} \tilde{\mathbf{A}}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \dot{\sim} \chi^2_p. \tag{3.5}$$

Since we do not know $\boldsymbol{\theta}^*$, we can replace it in the $\mathbf{B}$ matrix with $\hat{\boldsymbol{\theta}}_n$. This gives us

$$n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T \tilde{\mathbf{A}}_n \mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1} \tilde{\mathbf{A}}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \dot{\sim} \chi^2_p. \tag{3.6}$$

Finally, since $\tilde{\mathbf{A}}_n$ also depends on $\boldsymbol{\theta}^*$ we can approximate it with $\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n} \sum l_{\boldsymbol{\theta}\boldsymbol{\theta}}(y_i, \hat{\boldsymbol{\theta}}_n)$ and get the full plug-in sandwich-based pivot

$$n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T \mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) \mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1} \mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \dot{\sim} \chi^2_p. \tag{3.7}$$

A note of explanation for this last approximation will be made in the beginning of the next section. The full list of approximating assumptions made in the construction of this sandwich are

1. $\frac{1}{\sqrt{n}} \sum_{i=1}^n l_{\boldsymbol{\theta}}(y_i, \boldsymbol{\theta}^*) \dot{\sim} N(\mathbf{0}_p, \mathbf{B}(\boldsymbol{\theta}^*))$

2. $\mathrm{E}[l_{\boldsymbol{\theta}}(\boldsymbol{\theta}) l_{\boldsymbol{\theta}}(\boldsymbol{\theta})^T] = \mathbf{B}(\boldsymbol{\theta}) \approx \mathbf{B}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum l_{\boldsymbol{\theta}}(\boldsymbol{\theta}) l_{\boldsymbol{\theta}}(\boldsymbol{\theta})^T$

3. $\hat{\boldsymbol{\theta}}_n \approx \boldsymbol{\theta}^*$

4. $\tilde{\mathbf{A}}_n \approx \mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n} \sum l_{\boldsymbol{\theta}\boldsymbol{\theta}}(y_i, \hat{\boldsymbol{\theta}}_n)$

As in the one-dimensional case, there are four approximating assumptions, and they are very similiar in nature to the ond-dimensional assumptions.

To construct our pivot, we will also start with the Central Limit Theorem approximation

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i, \boldsymbol{\theta}^*) \dot{\sim} N(\mathbf{0}_p, \mathbf{B}(\boldsymbol{\theta}^*)). \tag{3.8}$$

As in the one-dimensional case, we will also remind ourselves that what we truly want is inference on $\boldsymbol{\theta}^*$. From this point, we can create the following pivot with approximate distribution

$$n \left[\frac{1}{n} \sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i, \boldsymbol{\theta}^*) \right] \mathbf{B}(\boldsymbol{\theta}^*)^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i, \boldsymbol{\theta}^*) \right] \dot{\sim} \chi^2_p. \tag{3.9}$$

Since we don't know the true data-generating distribution, we need to approximate $\mathbf{B}(\boldsymbol{\theta}^*)$. We can do this with the same empirical approximation that the sandwich uses $\mathbf{B}(\boldsymbol{\theta}) \approx \mathbf{B}_n(\boldsymbol{\theta})$, to get

$$n \left[\frac{1}{n} \sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i, \boldsymbol{\theta}^*) \right] \mathbf{B}_n(\boldsymbol{\theta}^*)^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i, \boldsymbol{\theta}^*) \right] \dot{\sim} \chi^2_p. \tag{3.10}$$

To conclude this result, we only require the following approximating assumptions

1. $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i, \boldsymbol{\theta}^*) \dot{\sim} N(\mathbf{0}_p, \mathbf{B}(\boldsymbol{\theta}^*))$

2. $\mathrm{E}[l_{\boldsymbol{\theta}}(\boldsymbol{\theta}) l_{\boldsymbol{\theta}}(\boldsymbol{\theta})^T] = \mathbf{B}(\boldsymbol{\theta}) \approx \mathbf{B}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum l_{\boldsymbol{\theta}}(\boldsymbol{\theta}) l_{\boldsymbol{\theta}}(\boldsymbol{\theta})^T$

As in the one-dimensional case, we require a distributional assumption and an empirical approximation because we do not know the true data-generating distribution. However, we do not need the plug-in assumptions used in the sandwich argument.

In order to generate confidence regions, we first specify a confidence level $100(1-\alpha)\%$.

Then we can perform a search throughout the parameter space to find values $\boldsymbol{\theta}$ that satisfy

$$n\left[\frac{1}{n}\sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i,\boldsymbol{\theta})\right]\mathbf{B}_n(\boldsymbol{\theta})^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i,\boldsymbol{\theta})\right] \leq \chi^2_{p,1-\alpha}. \tag{3.11}$$

Since $\hat{\boldsymbol{\theta}}_n$ is a solution to $\sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i,\boldsymbol{\theta}) = 0$, we can center our search on it, and move out from there.

3.2 Multi-Parameter Asymptotics

Lehmann [1999] defines multivariate efficiency in terms of the difference of the asymptotic covariance matrices of two estimators. Since we're not computing asymptotic covariance matrices in our pivot method, we have to do something different. We will prove a result in the multiple parameter case that is similar to the result for the one parameter case, namely that the boundaries of the confidence regions defined by our pivot and the sandwich converge almost surely to each other pointwise as n increases.

For this analytical work we will frequently use the fact that we can find $\overline{\mathbf{A}_n(\boldsymbol{\theta})}$ such that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i,\boldsymbol{\theta}^*) = (\sqrt{n})\overline{\mathbf{A}_n(\boldsymbol{\theta})}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \tag{3.12}$$

where $\mathbf{A}_n(\boldsymbol{\theta}) = (\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^T)[\frac{1}{n}\sum_{i=1}^{n} l(y_i,\boldsymbol{\theta})]$, the matrix of second order partial derivatives of the pseudo-log-likelihood, and $\overline{\mathbf{A}_n(\boldsymbol{\theta})}$ is a weighted average of $\mathbf{A}_n(\boldsymbol{\theta})$ evaluated at multiple points $\boldsymbol{\theta}$. This generates the pivot

$$n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^T\overline{\mathbf{A}_n(\boldsymbol{\theta})}\mathbf{B}_n(\boldsymbol{\theta})^{-1}\overline{\mathbf{A}_n(\boldsymbol{\theta})}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \leq \chi^2_{p,1-\alpha} \tag{3.13}$$

which is equivalent to line (3.11) and will be useful in our analytic arguments.

A few words need to be said about this substitution. Primarily, there is not an exact mean value theorem for vector-valued functions. $\overline{\mathbf{A}_n(\boldsymbol{\theta})}$ is a convex combination of matrices as described in [Furi and Martelli, 1991] as opposed to $\mathbf{A}_n(\boldsymbol{\theta})$ evaluated at one value. That is,

there are $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ and $\omega_1, \ldots, \omega_k \geq 0$ with $\sum_{i=1}^k \omega_i = 1$ such that $\overline{\mathbf{A}_n(\boldsymbol{\theta})} = \sum_{i=1}^k \omega_i \mathbf{A}_n(\boldsymbol{\theta}_i)$. The $\boldsymbol{\theta}_i$ are then chosen as points in p-space where each coordinate is taken from either $\hat{\boldsymbol{\theta}}_n$ or a solution to line (3.13), $\boldsymbol{\theta}$. Thus, $k \leq 2^p$. When we talk about points that satisfy the pivotal quantity, we are referring to solutions $\boldsymbol{\theta}$ and our k points are derived from $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$.

Let us define, for a given confidence level $1-\alpha$, the pivot equality and sandwich equality respectively as

$$n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^T \overline{\mathbf{A}_n(\boldsymbol{\theta})} \mathbf{B}_n(\boldsymbol{\theta})^{-1} \overline{\mathbf{A}_n(\boldsymbol{\theta})} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \chi^2_{p,1-\alpha} \tag{3.14}$$

$$n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^T \mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) \mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1} \mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \chi^2_{p,1-\alpha} \tag{3.15}$$

for any given sample size n.

To describe points in the solution sets above, we start by noticing that $\hat{\boldsymbol{\theta}}_n$ makes both the sandwich quadratic form and the quadratic form for our pivot equal to zero. In other words, $\hat{\boldsymbol{\theta}}_n$ will always be a point in the confidence regions generated by each method. We can then describe points satisfying the sandwich and pivot equalities as $\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,PV}$, and $\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,SW}$ respectively. If we have case that $\tilde{\boldsymbol{\theta}}_{n,PV}/\|\tilde{\boldsymbol{\theta}}_{n,PV}\| = \tilde{\boldsymbol{\theta}}_{n,SW}/\|\tilde{\boldsymbol{\theta}}_{n,SW}\|$, then we can say that the two points are "in the same direction from $\hat{\boldsymbol{\theta}}_n$". We say that a sequence of solutions $\{\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,PV}\}_n$ is "in the same direction" if $\tilde{\boldsymbol{\theta}}_{n,PV}/\|\tilde{\boldsymbol{\theta}}_{n,PV}\|$ is constant for all n. We define "in the same direction" similarly for sequences $\{\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,SW}\}_n$.

What we will show is that the difference in solutions to the pivot and sandwich equalities that are in the same direction from $\hat{\boldsymbol{\theta}}_n$ converges almost surely to zero on the $\sqrt{n}$ scale. This is equivalent to being asymptotically equally efficient for the same reason as in the one-dimensional case. If the boundaries of the pivot and sandwich confidence regions are converging asymptotically, then inference performed using one method will be asymptotically equivalent to inference performed using the other method.

Theorem 1. *For any sequence of solutions, $\{\breve{\boldsymbol{\theta}}_{n,PV}\}_n$, to the pivot equality in the same direction from $\hat{\boldsymbol{\theta}}_n$ and a corresponding sequence of solutions, $\{\breve{\boldsymbol{\theta}}_{n,SW}\}_n$, to the sandwich equality in the same direction from $\hat{\boldsymbol{\theta}}_n$ as $\{\breve{\boldsymbol{\theta}}_{n,PV}\}_n$, $\sqrt{n}\|\breve{\boldsymbol{\theta}}_{n,PV} - \breve{\boldsymbol{\theta}}_{n,SW}\| \to_{a.s.} 0$ as $n \to \infty$.*

Proof. Consider solutions $\breve{\boldsymbol{\theta}}_{n,PV}$ and $\breve{\boldsymbol{\theta}}_{n,SW}$ to the pivot equality and sandwich equality, respectively and both in the same direction from $\hat{\boldsymbol{\theta}}_n$. First, rewrite them as

$$\breve{\boldsymbol{\theta}}_{n,PV} = \hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,PV} \tag{3.16}$$

$$\breve{\boldsymbol{\theta}}_{n,SW} = \hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,SW} \tag{3.17}$$

where we emphasize again that $\tilde{\boldsymbol{\theta}}_{n,PV}$ and $\tilde{\boldsymbol{\theta}}_{n,SW}$ are vectors in the same direction, but with possibly different magnitudes.

We have that $\|\tilde{\boldsymbol{\theta}}_{n,PV}\|$ is $O(1/\sqrt{n})$ with probability one by Lemma 10. This also tells us that with probability one as $n \to \infty$, $\breve{\boldsymbol{\theta}}_{n,PV}$ will become arbitrarily close to $\hat{\boldsymbol{\theta}}_n$ and specifically, there is a constant A such that for sufficiently large n, $\|\tilde{\boldsymbol{\theta}}_{n,PV}\| \leq A/\sqrt{n}$. We also have that $\|\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) - \overline{\mathbf{A}_n(\boldsymbol{\theta})}\mathbf{B}_n(\breve{\boldsymbol{\theta}}_{n,PV})^{-1}\overline{\mathbf{A}_n(\boldsymbol{\theta})}\| \to_{a.s.} 0$ by the continuity of the matrix functions $\mathbf{A}_n(\boldsymbol{\theta})$ and $\mathbf{B}_n(\boldsymbol{\theta})^{-1}$. Pick $\zeta > 0$ small enough so that $\zeta < \chi^2_{p,1-\alpha}$. With probability one there is an M such that for all $n > M$, $\|\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) - \overline{\mathbf{A}_n(\boldsymbol{\theta})}\mathbf{B}_n(\breve{\boldsymbol{\theta}}_{n,PV})^{-1}\overline{\mathbf{A}_n(\boldsymbol{\theta})}\| < \frac{\zeta}{p^2A^2}$.

Now take a solution, $\breve{\boldsymbol{\theta}}_{n,PV} = \hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,PV}$, to the pivot equality and plug it into the sandwich equality

$$(\hat{\boldsymbol{\theta}}_n - \breve{\boldsymbol{\theta}}_{n,PV})^T\mathbf{A}_n(\hat{\boldsymbol{\theta}})\mathbf{B}_n(\hat{\boldsymbol{\theta}})^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_n - \breve{\boldsymbol{\theta}}_{n,PV}) = \tag{3.18}$$

$$(\tilde{\boldsymbol{\theta}}_{n,PV})^T\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)(\tilde{\boldsymbol{\theta}}_{n,PV}) = \tag{3.19}$$

$$(\tilde{\boldsymbol{\theta}}_{n,PV})^T[\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) - \tag{3.20}$$

$$\overline{\mathbf{A}_n(\boldsymbol{\theta})}\mathbf{B}_n(\breve{\boldsymbol{\theta}}_{n,PV})^{-1}\overline{\mathbf{A}_n(\boldsymbol{\theta})} + \overline{\mathbf{A}_n(\boldsymbol{\theta})}\mathbf{B}_n(\breve{\boldsymbol{\theta}}_{n,PV})^{-1}\overline{\mathbf{A}_n(\boldsymbol{\theta})}](\tilde{\boldsymbol{\theta}}_{n,PV}) = \tag{3.21}$$

$$(\tilde{\boldsymbol{\theta}}_{n,PV})^T\overline{\mathbf{A}_n(\boldsymbol{\theta})}\mathbf{B}_n(\breve{\boldsymbol{\theta}}_{n,PV})^{-1}\overline{\mathbf{A}_n(\boldsymbol{\theta})}(\tilde{\boldsymbol{\theta}}_{n,PV}) + \tag{3.22}$$

$$(\tilde{\boldsymbol{\theta}}_{n,PV})^T[\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) - \overline{\mathbf{A}_n(\boldsymbol{\theta})}\mathbf{B}_n(\breve{\boldsymbol{\theta}}_{n,PV})^{-1}\overline{\mathbf{A}_n(\boldsymbol{\theta})}](\tilde{\boldsymbol{\theta}}_{n,PV}) = \tag{3.23}$$

$$\frac{\chi^2_{p,1-\alpha}}{n} + (\tilde{\boldsymbol{\theta}}_{n,PV})^T[\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) - \overline{\mathbf{A}_n(\boldsymbol{\theta})}\mathbf{B}_n(\breve{\boldsymbol{\theta}}_{n,PV})^{-1}\overline{\mathbf{A}_n(\boldsymbol{\theta})}](\tilde{\boldsymbol{\theta}}_{n,PV}). \tag{3.24}$$

By the last line in the previous paragraph we have that the second term in line (3.24) can

be bounded as follows

$$
\begin{aligned}
-\frac{A^2}{n}p^2\frac{\zeta}{p^2A^2} \leq (\tilde{\boldsymbol{\theta}}_{n,PV})^T \Big[&\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) \\
&-\overline{\mathbf{A}_n(\boldsymbol{\theta})}\mathbf{B}_n(\breve{\boldsymbol{\theta}}_{n,PV})^{-1}\overline{\mathbf{A}_n(\boldsymbol{\theta})}\Big] (\tilde{\boldsymbol{\theta}}_{n,PV}) \leq \frac{A^2}{n}p^2\frac{\zeta}{p^2A^2} \\
-\frac{\zeta}{n} \leq (\tilde{\boldsymbol{\theta}}_{n,PV})^T[&\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) - \overline{\mathbf{A}_n(\boldsymbol{\theta})}\mathbf{B}_n(\breve{\boldsymbol{\theta}}_{n,PV})^{-1}\overline{\mathbf{A}_n(\boldsymbol{\theta})}](\tilde{\boldsymbol{\theta}}_{n,PV}) \leq \frac{\zeta}{n}
\end{aligned}
\tag{3.25}
$$

and plugging line (3.25) into line (3.24), performing a little algebra, and recalling that line (3.24) is equal to line (3.19) we have

$$
\frac{\chi^2_{p,1-\alpha}-\zeta}{n} \leq (\tilde{\boldsymbol{\theta}}_{n,PV})^T\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}})^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)(\tilde{\boldsymbol{\theta}}_{n,PV}) \leq \frac{\chi^2_{p,1-\alpha}+\zeta}{n}. \tag{3.26}
$$

Taking each side of line (3.26) separately and multiplying by the appropriate scaling factor, we see

$$
\frac{\chi^2_{p,1-\alpha}}{\chi^2_{p,1-\alpha}+\zeta}(\tilde{\boldsymbol{\theta}}_{n,PV})^T\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)(\tilde{\boldsymbol{\theta}}_{n,PV}) < \frac{\chi^2_{p,1-\alpha}}{n} \tag{3.27}
$$

$$
\frac{\chi^2_{p,1-\alpha}}{\chi^2_{p,1-\alpha}-\zeta}(\tilde{\boldsymbol{\theta}}_{n,PV})^T\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)(\tilde{\boldsymbol{\theta}}_{n,PV}) > \frac{\chi^2_{p,1-\alpha}}{n}. \tag{3.28}
$$

Now we can rewrite $\frac{\chi^2_{p,1-\alpha}}{\chi^2_{p,1-\alpha}+\zeta}$ as $1-\frac{\zeta}{\chi^2_{p,1-\alpha}+\zeta} = 1-\xi_+$. Similarly, $\frac{\chi^2_{p,1-\alpha}}{\chi^2_{p,1-\alpha}-\zeta} = 1+\frac{\zeta}{\chi^2_{p,1-\alpha}-\zeta} = 1+\xi_-$. Note in particular that ζ and ξ_- are in one-to-one correspondence, and as $\zeta \to 0$, $\xi_- \to 0$. This gives us

$$
(1-\xi_+)(\tilde{\boldsymbol{\theta}}_{n,PV})^T\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)(\tilde{\boldsymbol{\theta}}_{n,PV}) < \frac{\chi^2_{p,1-\alpha}}{n} \tag{3.29}
$$

$$
(1+\xi_-)(\tilde{\boldsymbol{\theta}}_{n,PV})^T\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)(\tilde{\boldsymbol{\theta}}_{n,PV}) > \frac{\chi^2_{p,1-\alpha}}{n}. \tag{3.30}
$$

Since $\xi_- > \xi_+$ and $1 - \xi_- < 1 - \xi_+$ we can rewrite the first inequality to get

$$(1-\xi_-)(\tilde{\boldsymbol{\theta}}_{n,PV})^T \mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)(\tilde{\boldsymbol{\theta}}_{n,PV}) < \frac{\chi^2_{p,1-\alpha}}{n} \tag{3.31}$$

$$(1+\xi_-)(\tilde{\boldsymbol{\theta}}_{n,PV})^T \mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)(\tilde{\boldsymbol{\theta}}_{n,PV}) > \frac{\chi^2_{p,1-\alpha}}{n}. \tag{3.32}$$

We have that $\hat{\boldsymbol{\theta}}_n - (\sqrt{1-\xi_-})\tilde{\boldsymbol{\theta}}_{n,PV}$ satisfies the plug-in sandwich quadratic form inequality but $\hat{\boldsymbol{\theta}}_n - (\sqrt{1+\xi_-})\tilde{\boldsymbol{\theta}}_{n,PV}$ does not. By the continuity of the plug-in sandwich quadratic form, the solution to the sandwich equality, i.e. $\breve{\boldsymbol{\theta}}_{n,SW} = \hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,SW}$, must lie on the line segment between these two points.

Once we have selected ζ and consequently ξ_- as described above, $1 - \sqrt{1-\xi_-} > \sqrt{1+\xi_-} - 1$. This means that $\hat{\boldsymbol{\theta}}_n - (\sqrt{1-\xi_-})\tilde{\boldsymbol{\theta}}_{n,PV}$ is the furthest point from $\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,PV}$ in the interval $(\hat{\boldsymbol{\theta}}_n - (\sqrt{1-\xi_-})\tilde{\boldsymbol{\theta}}_{n,PV},\ \hat{\boldsymbol{\theta}}_n - (\sqrt{1+\xi_-})\tilde{\boldsymbol{\theta}}_{n,PV})$. Putting this all together, we have that for all $1 > \xi_- > 0$ with probability one there is an M such that for all $n > M$

$$\begin{aligned}
\sqrt{n}\|(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,PV}) - (\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,SW})\| &= \sqrt{n}\|\tilde{\boldsymbol{\theta}}_{n,SW} - \tilde{\boldsymbol{\theta}}_{n,PV}\| && (3.33)\\
&\leq \|(\sqrt{1-\xi_-})\tilde{\boldsymbol{\theta}}_{n,PV} - \tilde{\boldsymbol{\theta}}_{n,PV}\| && (3.34)\\
&= \sqrt{n}\|\tilde{\boldsymbol{\theta}}_{n,PV}\| \mid 1 - \sqrt{1-\xi_-} \mid && (3.35)\\
&\leq \sqrt{n}\sqrt{\frac{\chi^2_{p,1-\alpha}}{n\lambda_{lo}}}(\mid 1 - \sqrt{1-\xi_-} \mid) && (3.36)\\
&= \sqrt{\frac{\chi^2_{p,1-\alpha}}{\lambda_{lo}}}(\mid 1 - \sqrt{1-\xi_-} \mid) && (3.37)
\end{aligned}$$

but ζ was chosen to be arbitrarily small, which gives us that ξ_- is also arbitrarily small and hence the right-hand side in line (3.37) can be made arbitrarily small. Hence we have that $\sqrt{n}\|(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,PV}) - (\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,SW})\| \rightarrow_{a.s.} 0$ as $n \rightarrow \infty$, as desired. □

3.3 Simple Linear Regression - Simulation Example

In this example we use as our working model a simple linear regression

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i \tag{3.38}$$

$$\epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2) \tag{3.39}$$

where the parameter of interest is $\boldsymbol{\theta} = (\theta_0, \theta_1)^T$. As in the regression through the origin example, we assume σ^2 known, but since it cancels in the pivot we could also assume it is an unknown nuisance parameter.

Our model misspecification will be heteroscedastic errors. Specifically, the truth will be the following model

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i \tag{3.40}$$

$$\epsilon_i \overset{indep}{\sim} N(0, 1 + |x_i|) \tag{3.41}$$

with the errors depending on the covariate. Under this model the relevant term of the pseudo-log-likelihood and necessary partial derivatives are given below:

$$\sum_{i=1}^{n} l(y_i, \boldsymbol{\theta}) = -n \log(\sqrt{2\pi}\sigma) + \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2 \tag{3.42}$$

$$\sum_{i=1}^{n} l_{\theta_0}(y_i, \boldsymbol{\theta}) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i) \tag{3.43}$$

$$\sum_{i=1}^{n} l_{\theta_1}(y_i, \boldsymbol{\theta}) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i (y_i - \theta_0 - \theta_1 x_i). \tag{3.44}$$

We can write the factors in the pivot inequality (3.11) as

$$l_{\boldsymbol{\theta}} = \begin{pmatrix} \frac{1}{2n\sigma^2}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i) \\ \frac{1}{2n\sigma^2}\sum_{i=1}^{n} x_i(y_i - \theta_0 - \theta_1 x_i) \end{pmatrix} \tag{3.45}$$

$$\mathbf{B}_n(y, \boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{4n\sigma^4}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2 & \frac{1}{4n\sigma^4}\sum_{i=1}^{n} x_i(y_i - \theta_0 - \theta_1 x_i)^2 \\ \frac{1}{4n\sigma^4}\sum_{i=1}^{n} x_i(y_i - \theta_0 - \theta_1 x_i)^2 & \frac{1}{4n\sigma^4}\sum_{i=1}^{n} x_i^2(y_i - \theta_0 - \theta_1 x_i)^2 \end{pmatrix} \tag{3.46}$$

and then the pivot is exactly the one as in inequality (3.11).

Our simulation was performed similarly to the one-parameter case. We simulated data for samples sizes ranging from 10 to 100. For each sample size, we simulated $10,000$ datasets. The covariates were generated from a standard normal distribution, and the response was generated from the true model in this section. For each dataset, we computed whether or not the true value of the parameter was covered by the specified confidence regions. The results are shown in Figure 3.1. We only show results for the sandwich, the pivot, and the hc1, hc2, and hc3 sandwich corrections. The hc4 and hc5 methods were designed for scalar, linear combinations of parameters and are not adaptable to a multivariate setting.

The results in this multi-parameter simulation mirror the results from the one-parameter simulation. The standard intervals from simple linear regression undercover the parameter vector consistently throughout all sample sizes tested. The sandwich-based confidence intervals undercover the pseudo-true parameter, with coverage improving as sample size increases. The ad hoc methods all provide slightly better coverage, as they all increase the sandwich variance estimate. The pivot-based confidence intervals provide comparable coverage rates to the hc1, hc2, and hc3 methods, providing the best coverage rates in this simulation in all but the very smallest sample sizes, though the results are not significantly different.

To compare the sensitivity of our results to the distribution of the independent variable, we now look at the same simulation, but with covariates generated from an exponential distribution with mean parameter equal to three. In contrast to the standard normal, the exponential is a skewed distribution and only generates non-negative numbers. This will

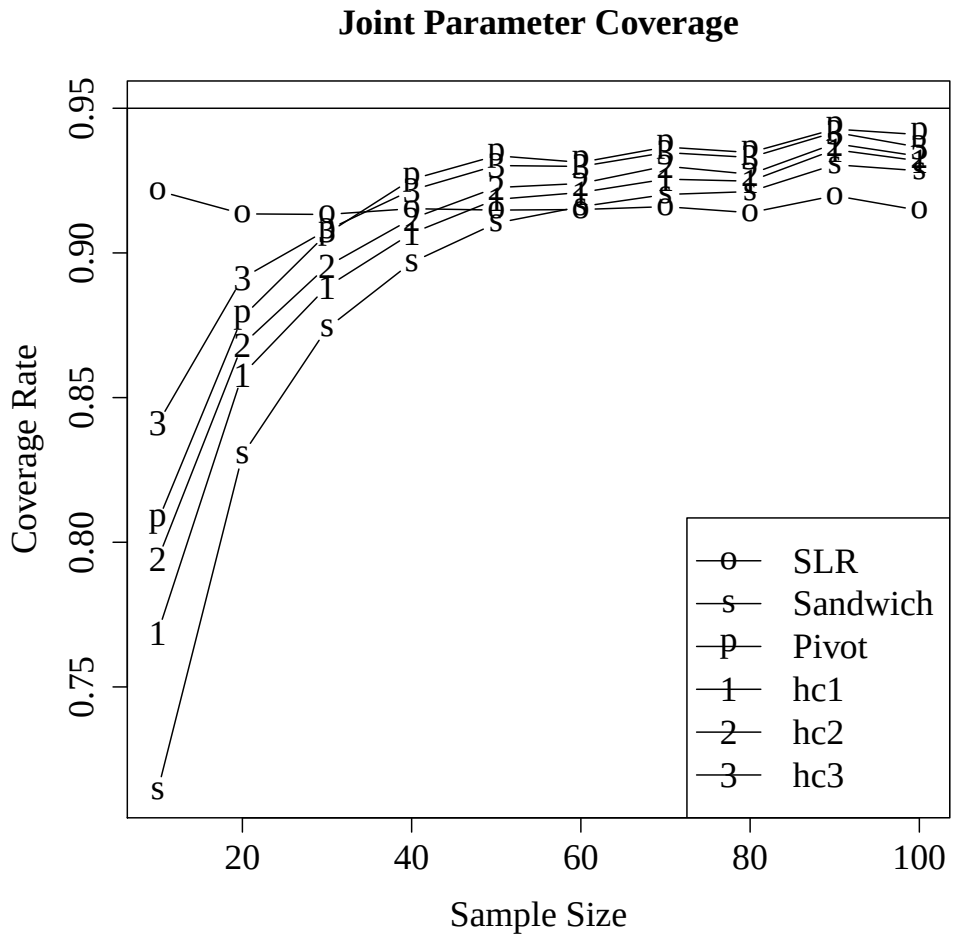

Figure 3.1: Graph showing confidence interval coverage of the true parameter vector using various methods. The covariate here has a standard normal distribution.

give us a sense of how the pivot method compares to the sandwich when the situation is not as "nice" as in the previous example. The results of this second simulation can be seen in Figure 3.2.

We see from these results that standard maximum likelihood confidence intervals now perform the worst of all methods. Coverage rates begin at less than 70% for $n = 10$ and drop from there. The sandwich and its alternatives perform as expected, with low initial coverage rates with respect to n and increasing coverage rates as n increases. The pivot again performs the best of all. Coverage is a little high in the beginning, but drops close to 95% quickly.

3.4 Multi-parameter to One Parameter

We may wish to construct confidence intervals for one parameter at a time in a multi-parameter setting. Alternatively, we may wish to eliminate nuisance parameters in a model with multiple parameters. The pivot described in previous sections can be used to generate a multivariate confidence region, but does not generate individual confidence intervals very easily. In the next two subsections we will describe a couple of ways in which inference on individual parameters might be done. The first subsection details a plug-in method, which is the easier of the two suggestions. Since we are trying to steer away from plug-in methods in general, a profile likelihood method is described in the second section. Each method can be used to generate one-dimensional confidence intervals for the elements of a multiple dimension parameter of interest.

3.4.1 Plug-in

One option to reduce the number of parameters down to one is to use the plug-in principle. We do not know the values of the pseudo-true parameters in our model. However, maximum likelihood theory tells us that the maximum misspecified likelihood estimate is a consistent estimator of the pseudo-true value. The plug-in principle then says that in lieu of using the truth, we can use that estimate as an approximately correct value. This is the principle that

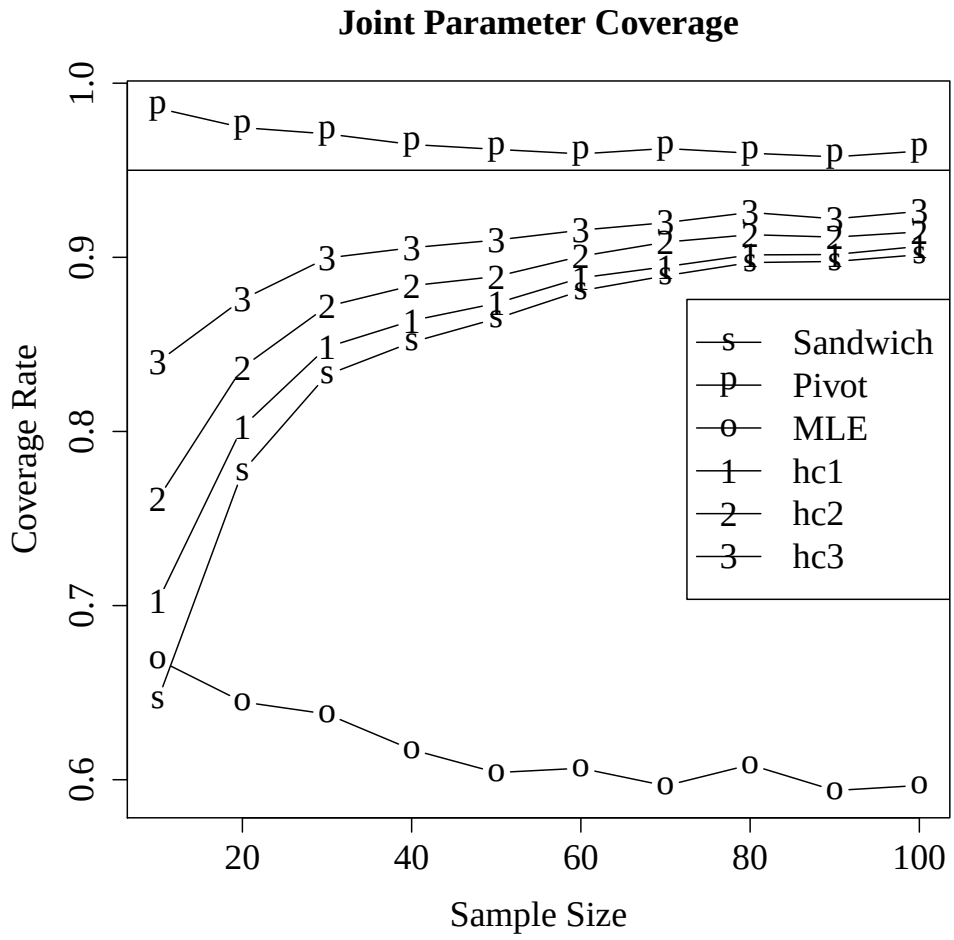

Figure 3.2: Graph showing confidence interval coverage of the true parameter vector using various methods. The covariate here is exponentially distributed.

allows us to use $\hat{\boldsymbol{\theta}}_n \approx \boldsymbol{\theta}^*$ in the sandwich argument. Here we will use that approximation for all but one component of $\boldsymbol{\theta}^*$ in order to generate inference on the remaining component. The next two lines illustrate how that would work in the case of simple linear regression.

$$\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0^* - \hat{\theta}_1 x_i)^2\right)^{-1/2} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(y_i - \theta_0^* - \hat{\theta}_1 x_i) \dot{\sim} N(0,1) \tag{3.47}$$

$$\left(\frac{1}{n}\sum_{i=1}^{n}[x_i(y_i - \hat{\theta}_0 - \theta_1^* x_i)]^2\right)^{-1/2} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i(y_i - \hat{\theta}_0 - \theta_1^* x_i) \dot{\sim} N(0,1) \tag{3.48}$$

We use the first line to find confidence intervals for θ_0 and the second line to find confidence intervals for θ_1.

To check and see if this is reasonable, we perform a simulation study. We simulate covariate data X from a standard normal distribution and response data Y from the model stated above, given constant values of $\boldsymbol{\theta} = (\theta_0, \theta_1)$. Sample sizes ranged from 10 to 100. Each sample size had $10,000$ simulated data sets, from which I estimated $\hat{\boldsymbol{\theta}}$. We then compute marginal confidence intervals using the sandwich estimate of variance, our pivot using plug-in values as in lines (3.47) and (3.48), and standard linear regression theory which does not account for possible model misspecification. The results are shown in Figures 3.3 and 3.4.

The coverage of the intercept parameter θ_0 are shown in Figure 3.3. The sandwich method performs worst while the plug-in pivot perform close to the stated 95% coverage. The standard regression confidence intervals perform nearly as badly as the sandwich based intervals. The ad hoc methods tend to perform better than the sandwich method but not as well as the pivot plug-in method. The hc5 method overcovers the true parameter, as is typical in the simulation studies shown in this dissertation.

The results for the simulated coverage of the covariate coefficient further justify use of our pivot. The pivot using a plug-in method outperforms the sandwich method in terms of coverage of the true parameter, and for sample sizes larger or equal to 20 it shows almost 95% coverage. Standard regression confidence intervals only cover approximately 88% of the time. The alternative ad hoc methods tend to provide better coverage than the unaltered

sandwich method, but not as good coverage as the plug-in pivot method. Again, the hc5 method tends to overcover the true parameter.

3.4.2 Profile Likelihood

I also used the profile likelihood method to compute confidence intervals for θ_0 and θ_1 individually. Profile likelihood is another way to deal with nuisance parameters. Here, neither of θ_0 or θ_1 is inherently a nuisance parameter, but each can become a nuisance when trying to perform inference on the other. To use the general profile likelihood method, you start by maximizing the likelihood with respect to the nuisance parameters alone. The result is a set of estimates for the nuisance parameters that are functions of both the data and the parameters of interest. You then plug these estimates back into the original likelihood in the place of the nuisance parameters. The end result is a profile likelihood that is expressed in terms of data and parameters of interest only. Profile likelihood estimates for the parameters of interest are then computed by maximizing this profile likelihood with respect to the parameters of interest. One of the benefits of the profile likelihood method is that it preserves relationships between nuisance parameters and the parameters of interest. This is a key difference between using profile likelihood and using the standard plug-in method used in the previous subsection.

As an example of using the profile likelihood method, I will describe how to use this method on the covariate coefficient, knowing that it would be performed similarly for the intercept term. First we assume that θ_1 is known. The pseudo-likelihood is then given by

$$l(y; x, \theta_1) = \frac{-1}{2n\sigma^2} \sum (y_i - \theta_0 - \theta_1 x_i)^2 - n \log(\sigma) - \frac{n}{2} \log(2\pi) \tag{3.49}$$

Solving for $\hat{\theta}_0$ assuming θ_1 we take the derivative with respect to θ_0, set it equal to zero,

and solve:

$$\frac{\partial}{\partial\theta_0}l(y;x,\theta_1) = \frac{1}{n\sigma^2}\sum(y_i-\theta_0-\theta_1x_i) \tag{3.50}$$

$$0 = \frac{1}{n\sigma^2}\sum(y_i-\theta_0-\theta_1x_i) \tag{3.51}$$

$$= \sum(y_i-\theta_0-\theta_1x_i) \tag{3.52}$$

$$= n\bar{y}-n\theta_0-n\theta_1\bar{x} \tag{3.53}$$

$$\hat{\theta}_0 = \bar{y}-\theta_1\bar{x}. \tag{3.54}$$

We plug this into the original pseudo-likelihood to profile out θ_0. This results in a pseudo-"likelihood" that is in terms of the data and θ_1 only (ignoring σ for the moment).

$$l(y;x,\theta_1) = \frac{-1}{2n\sigma^2}\sum(y_i-\bar{y}-\theta_1(x_i-\bar{x}))^2 - n\log(\sigma) - \frac{n}{2}\log(2\pi) \tag{3.55}$$

We now take the derivative with respect to θ_1.

$$\frac{\partial}{\partial\theta_1}l(y;x,\theta_1) = \frac{1}{n\sigma^2}\sum(x_i-\bar{x})(y_i-\bar{y}-\theta_1(x_i-\bar{x})). \tag{3.56}$$

This is the function we would set to zero to compute the MMLE. However, we will use this function now to construct our pivot. We compute $B_n(\theta_1)$ as

$$B_n(\theta_1) = \frac{1}{n\sigma^4}\sum(x_i-\bar{x})^2(y_i-\bar{y}-\theta_1(x_i-\bar{x}))^2. \tag{3.57}$$

Our pivot can now be written as

$$\frac{\sqrt{n}}{\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2(y_i-\bar{y}-\theta_1(x_i-\bar{x}))^2}}\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y}-\theta_1(x_i-\bar{x})) \tag{3.58}$$

which has an approximate $N(0,1)$ distribution.

To see how this method compares to both the plug-in, the usual sandwich method, and

the ad hoc methods, I ran the same simulation described in the plug-in section but added in confidence intervals computed using the profile likelihood. The results can be seen in Figures 3.3 and 3.4.

With the intercept parameter, we see that the profile likelihood does a slightly better job of producing 95% confidence intervals than the plug-in method did. In fact, it performs better than the plug-in for most sample sizes tested. Regarding coverage of the true value of the coefficient on the covariate, we see that the profile likelihood method parallels the performance of the plug-in method. This is not terribly surprising, since profile likelihood is a form of plug-in itself.

3.4.3 Non-Normal Covariate

As in previous sections to test the sensitivity of our method to different covariate distributions, we ran the same simulation but with the independent variable distributed as an exponential random variable with mean parameter three. All other aspects of the simulation were kept the same. The results of this modified simulation can be seen in Figures 3.5 and 3.6.

Regarding coverage of the intercept parameter, all methods tend to do pretty well, with coverage rates between 90% and 95% for all methods but hc5 and all sample sizes $n = 20$ and higher. The hc5 method shows high coverage, but that is typical in these examples. The profile likelihood method compares favorably to the sandwich and its alternatives. The only odd result is the poor performance of the plug-in method, since it appears to show coverage that does not converge to 95%. It is possible, given the nature of the skewed covariate distribution, that we have large covariate points that are influental points in the example. This would happen because of the model misspecification tying the variance of the response to the size of the covariate. Excess variability in $\hat{\theta}_1$ could be causing the plug-in to perform poorly, compared to other methods.

Turning now to the slope parameter, we see a more standard set of results. The regular maximum likelihood coverage rates are low throughout the simulation, with a maximum of

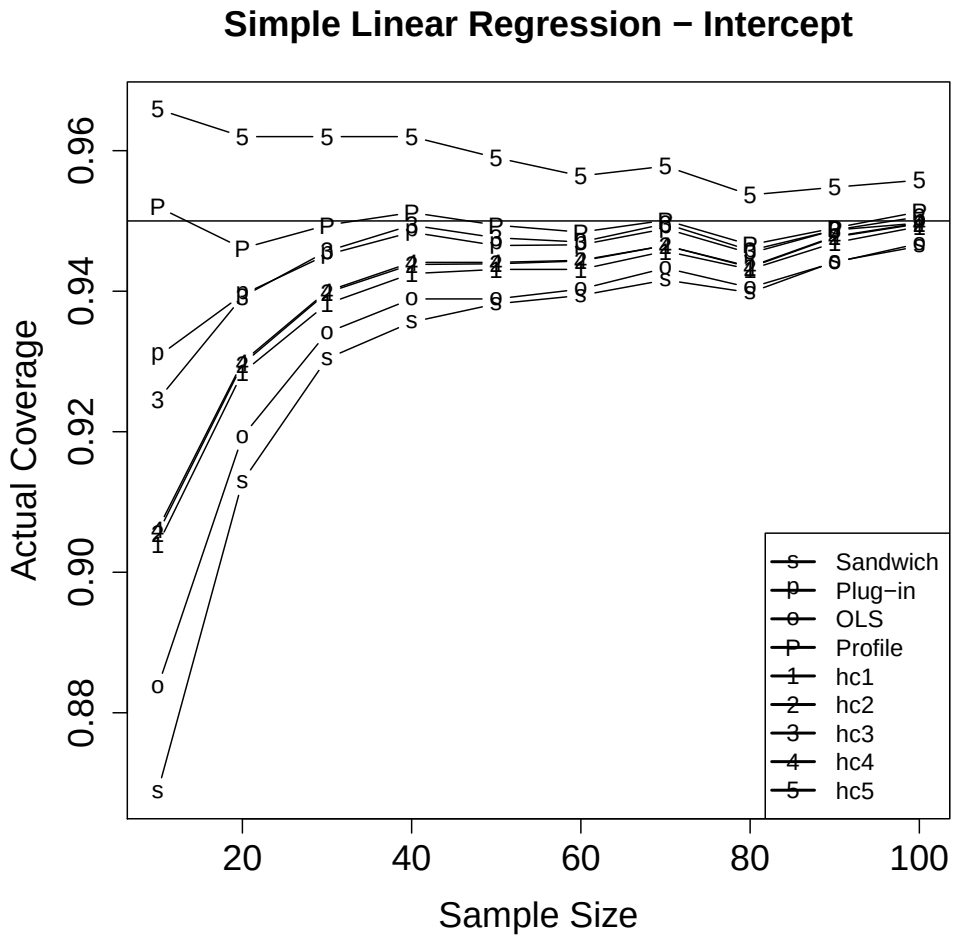

Figure 3.3: Results from a simulation study to compare confidence interval coverage using our pivot with both the plug-in and the profile likelihood methods, the sandwich estimator of variance, and the standard linear regression estimate of variance that does not account for model misspecificiation. The assumed model is simple linear regression with one covariate. The covariate has a standard normal distribution. The model misspecification is heteroscedastic errors. Shown here are the simulated coverage of θ_0, the intercept parameter.

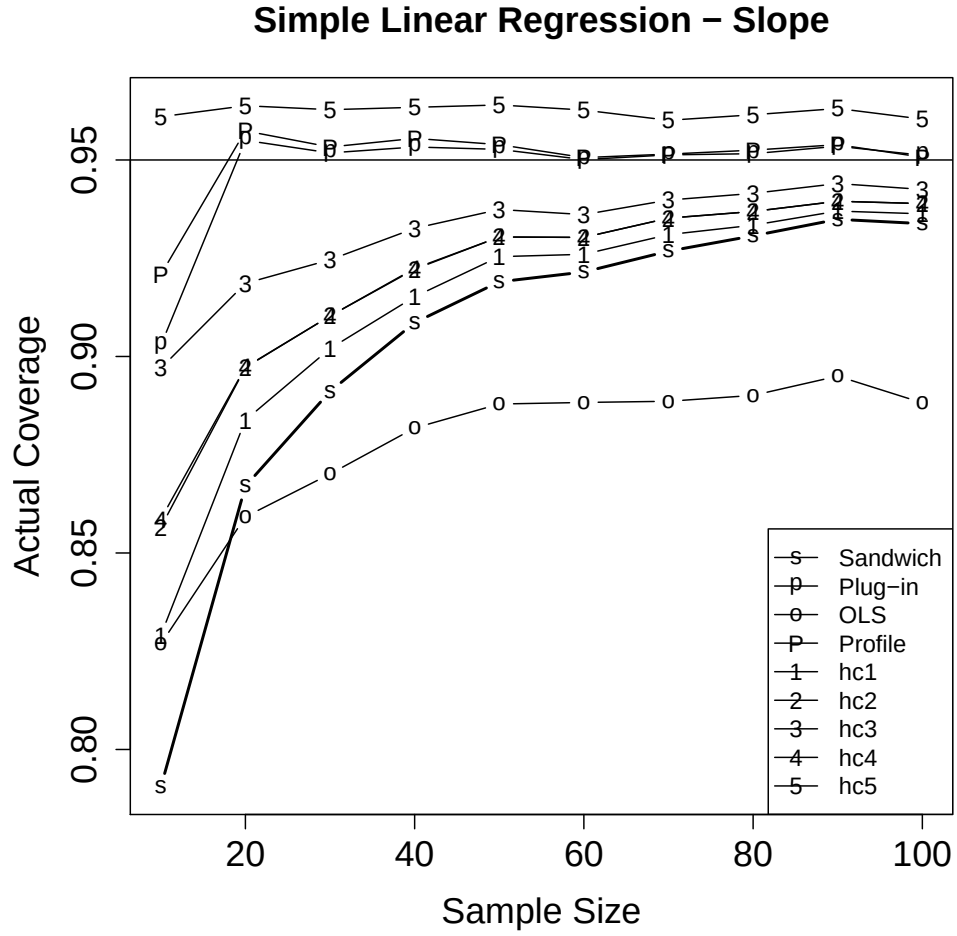

Figure 3.4: Results from a simulation study to compare confidence interval coverage using our pivot with both the plug-in and the profile likelihood methods, the sandwich estimator of variance, and the standard linear regression estimate of variance that does not account for model misspecificiation. The assumed model is simple linear regression with one covariate. The covariate has a standard normal distribution. The model misspecification is heteroscedastic errors. Shown here are the simulated coverage of θ_1, the covariate coefficient parameter.

around 88%. The sandwich and hc1-4 alternatives start with low coverage rates at $n = 10$, with coverage increasing as n increases. The hc5 method shows higher than nominal coverage throught the simulation, as it does in other examples we have seen. The pivot method using the profile likelihood performs favorably to the sandwich and its alternatives at sample size $n = 30$ and higher. The plug-in is again showing odd behavior with high coverage in the smallest sample sizes, and lower than nominal coverage in higher sample sizes.

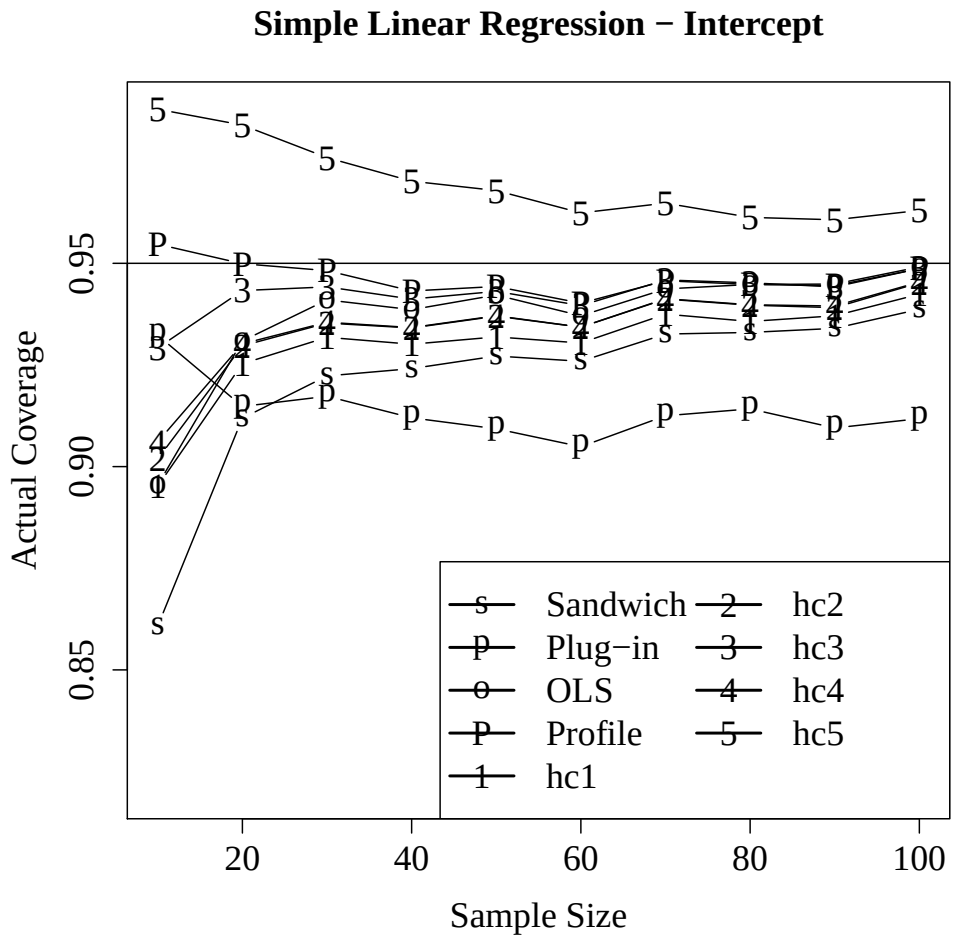

Figure 3.5: Results from a simulation study to compare confidence interval coverage using our pivot with both the plug-in and the profile likelihood methods, the sandwich estimator of variance, and the standard linear regression estimate of variance that does not account for model misspecificiation. The assumed model is simple linear regression with one covariate. The covariate has an exponential distribution with mean three. The model misspecification is heteroscedastic errors. Shown here are the simulated coverage of θ_0, the intercept parameter.

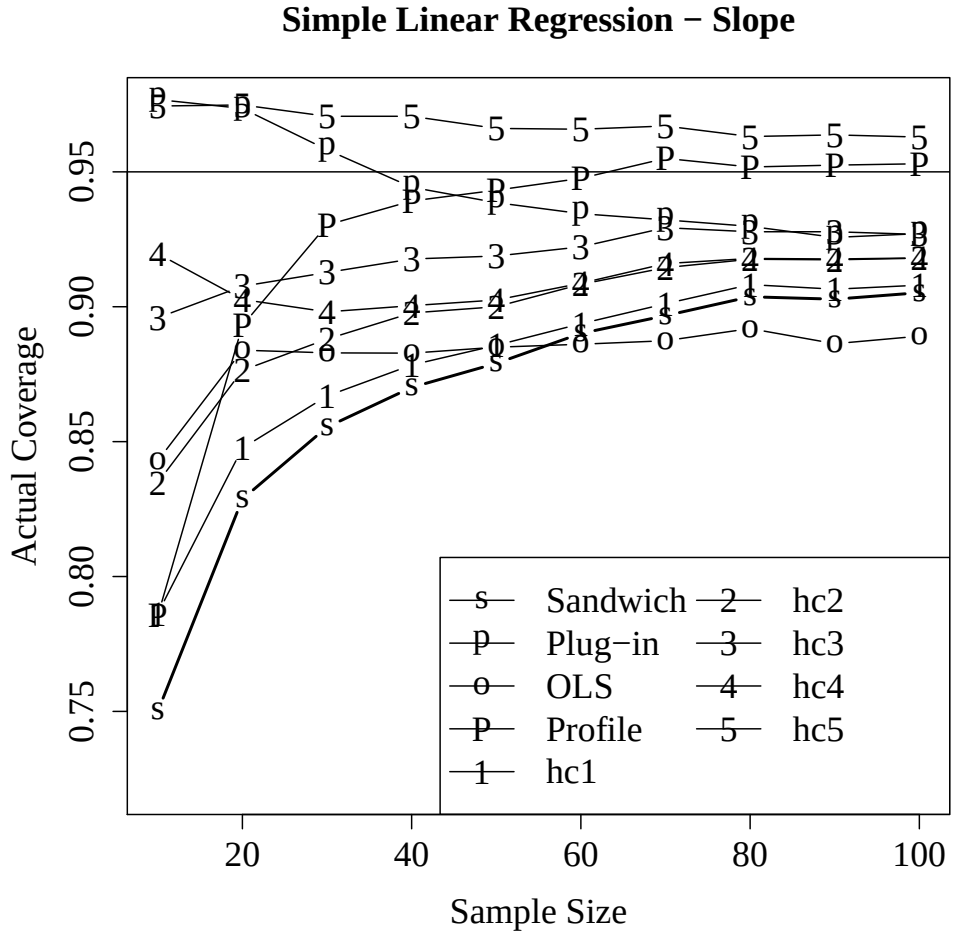

Figure 3.6: Results from a simulation study to compare confidence interval coverage using our pivot with both the plug-in and the profile likelihood methods, the sandwich estimator of variance, and the standard linear regression estimate of variance that does not account for model misspecificiation. The assumed model is simple linear regression with one covariate. The covariate has an exponential distribution with mean three. The model misspecification is heteroscedastic errors. Shown here are the simulated coverage of θ_1, the covariate coefficient parameter.

Chapter 4

PSEUDO-BAYESIAN INFERENCE USING PIVOTS

To this point, we have used the following pivot and approximating distribution for a parameter θ to generate frequentist inference about that parameter:

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} l_\theta(y_i,\theta)^2\right]^{-1/2}\frac{1}{n}\sum_{i=1}^{n} l_\theta(y_i,\theta) \dot{\sim} N(0,1). \tag{4.1}$$

We would now like to use this pivot in a Bayesian context. There are a couple of reasons for doing so. First, Bayesian analysis has a natural method for dealing with nuisance parameters. One simply integrates over their values. When using Monte Carlo methods to approximate the posterior distribution of a parameter set, integration is as easy as summing over the various values of the nuisance parameters. Second, Bayesian analysis allows the analyst to incorporate prior beliefs into the current data analysis.

There are two issues we encountered while investigating pivots in a Bayesian context. First, we cannot use pivots in a proper Bayesian analysis as the probability calculus is violated, a fact noted in Monahan and Boos [1992] and explained in Section 4.1. Also in that section we explore a pseudo-Bayesian alternative using pivots and the densities associated with their approximate distributions. The second issue we face, once we have a pseudo-Bayesian procedure, is the question of which pivot to use. Any monotonic transformation of a pivot results in another pivot. Do all pivots provide equivalent inference? If not, is there a natural choice of pivot to use in pseudo-Bayesian inference? How might we choose a pivot to use for pseudo-Bayesian inference? In Section 4.2 and 4.3 we answer these questions: not all pivots provide equal, pseudo-Bayesian inference; there is a principle that we can use to choose a pivot for use in pseudo-Bayesian inference; and our pivot is a natural choice. In Section

4.4 we illustrate how pseudo-Bayesian inference compares to proper Bayesian inference when the likelihood of the data has been misspecified.

4.1 Pseudo-Bayesian Inference

In traditional Bayesian analysis we select a likelihood, $p(y \mid \theta)$, for our data as a function of the parameter of interest θ. We also select a prior distribution for θ which is described by some prior density $\pi(\theta)$. The posterior distribution $\pi(\theta \mid y)$ can then be computed via the relationship

$$\pi(\theta \mid y) = \frac{p(y \mid \theta)\pi(\theta)}{p(y)} \tag{4.2}$$

$$\propto p(y \mid \theta)\pi(\theta) \tag{4.3}$$

where $p(y)$ is the unconditional distribution of the data, and the proportionality in the second line is with respect to θ.

With pivots, the computations break down. Say we have a pivot $t = t(y;\theta)$ which has a distribution described by density $p(t)$, which does not depend on the parameter θ. Then $p(t)\pi(\theta) \propto \pi(\theta)$ which means we have not updated our beliefs. Suppose we instead use $p(t(y;\theta))$, and plug the actual pivot function into the density $p(\cdot)$. This is no longer necessarily a proper likelihood. When we compute $p(y)$ as

$$p(y) = \int p(t(y;\theta))\pi(\theta)d\theta \tag{4.4}$$

then $p(y)$ is not a density of any recognizable quantity, which means that the posterior computed in line (4.2) is not a proper posterior distribution. Performing a proper Bayesian analysis in this manner is impossible.

Pseudo-Bayesian analysis, on the other hand, is based on the concept of a pseudo-likelihood. Instead of selecting a proper likelihood for our data, we select a function $p^*(y \mid \theta)$ that takes the place of the actual likelihood and which has properties similar to an actual

likelihood. For a given prior distribution, $\pi(\theta)$, we can compute the pseudo-posterior distribution, $\pi^*(\theta \mid y)$, as

$$\pi^*(\theta \mid y) \propto p^*(y \mid \theta)\pi(\theta). \tag{4.5}$$

Monahan and Boos [1992] gives a condition under which a pseudo-likelihood and its corresponding pseudo-posterior may be considered a valid tool for inference. To describe the condition we need to define a couple of terms first. Suppose we have selected a pseudo-likelihood and a prior, and we have computed a pseudo-posterior as in line (4.5). Suppose also that the true data generating distribution is given by some density $f(y \mid \theta)$. Then Monahan and Boos [1992] defines a "posterior coverage set function" of level α for our given prior and pseudo-likelihood as a function $R_\alpha(y)$ such that $\Pr(\theta \in R_\alpha(y)) = \alpha$ for all y, under the measure generated by the pseudo-posterior. Second, a pseudo-posterior is defined as "valid by coverage" for the true model if and only if for every posterior coverage set function $R_\alpha(y)$, $\Pr(\theta \in R_\alpha(y)) = \alpha$ under the joint measure, $\pi(\theta)f(y \mid \theta)$, defined using the true model. Finally, a pseudo-likelihood is a "coverage proper Bayesian likelihood" if and only if for all absolutely continuous priors $\pi(\theta)$, the pseudo-posterior defined in line (4.5) is valid by coverage.

The coverage proper Bayesian likelihood condition, while laudatory, is difficult to verify in reality. The true data-generating model is unknown, so it is impossible to describe $f(y \mid \theta)$. Even if one did know the true model, this condition has to hold for all absolutely continuous prior distributions $\pi(\theta)$. Unless there is a theoretical result proving this for a given pseudo-likelihood, there are quite a few priors to check. For our purposes, we want a pseudo-Bayesian procedure to provide confidence intervals that cover the true parameter at the stated confidence percentage.

Pseudo-Bayesian analysis has been studied by many authors. Efron [1993] proposes a pseudo-likelihood based on confidence intervals that aids in elimination of nuisance parameters. Bertolino and Racugno [1992] use marginal and profile likelihoods as their pseudo-

likelihoods to perform inference on correlation coefficients. Two years later, Bertolino and Racugno [1994] looked at marginal likelihoods as a way to use a Bayesian approach for ANOVA and χ^2 tests. Greco et al. [2008] investigate the behavior of quasi- and empirical likelihood functions in Bayesian analysis.

Our proposal is to use the approximate distribution of our pivot as a pseudo-likelihood. Specifically, recall our pivotal quantity and its approximate distribution in (4.1). We can write our pseudo-likelihood as

$$p^*(t(y_1, \ldots, y_n; \theta)) \propto e^{-\frac{1}{2}t(y_1,\ldots,y_n;\theta)^2} \tag{4.6}$$

where we are abbreviating the notation for the pivot to

$$t(y_1, \ldots, y_n; \theta) = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} l_\theta(y_i, \theta)^2\right]^{-1/2} \frac{1}{n}\sum_{i=1}^{n} l_\theta(y_i, \theta). \tag{4.7}$$

As long as our pseudo-likelihod behaves like a likelihood, we can perform inference in a pseudo-Bayesian fashion. We select a prior distribution, $\pi(\theta)$, for our parameter of interest. Then we can compute our pseudo-posterior distribution as

$$\pi^*(\theta \mid y_1, \ldots, y_n) \propto \pi(\theta) \times e^{-\frac{1}{2}t(y_1,\ldots,y_n;\theta)^2}. \tag{4.8}$$

If the right hand side has a finite integral with respect to θ, then it can function as a proper density. We show this integrability now. For any value θ,

$$0 < e^{-\frac{1}{2}t(y_1,\ldots,y_n;\theta)^2} \leq e^0 = 1. \tag{4.9}$$

Since we know that $\pi(\theta)$ is a proper density, we have

$$0 < \int e^{-\frac{1}{2}t(y_1,\ldots,y_n;\theta)^2}\pi(\theta)d\theta \tag{4.10}$$

$$\leq \int 1 \times \pi(\theta)d\theta \tag{4.11}$$

$$= 1 \tag{4.12}$$

and hence the right-hand side of line (4.8) is integrable as long as $\pi(\theta)$ is a proper density. We can now estimate quantiles of the pseudo-posterior by using Monte Carlo techniques and the Metropolis-Hastings algorithm, and then use appropriate quantiles for our confidence intervals.

4.2 Exploration of Pivotal Inference

Given a pivotal quantity, any monotonic transformation results in yet another pivotal quantity. For example the pivotal quantity described in line (4.7) has an approximately normal distribution. However, if we take $\exp(t(y;\theta))$, then we have an approximately log-normally distributed pivot. As we will see in this section, the pivot we use does not matter in frequentist inference but it can have a big impact in pseudo-Bayesian inference.

In frequentist analysis, the use of different pivots, modulo monotonic transformations, does not affect our conclusions. Suppose we have data $y_1, \ldots, y_n$. We also have a parameter of interest θ. We create a pivotal quantity $t(y,\theta) \sim P_0$ whose distribution, P_0, is independent of θ. We then create a probability statement of the form

$$\Pr(p_{0,\alpha/2} \leq t(y,\theta) \leq p_{0,1-\alpha/2}) = 1 - \alpha \tag{4.13}$$

where α is the significance level we wish to use, and $p_{0,q}$ is the $100 \times q\%$ quantile of the P_0 distribution. Finding θ that solves for the boundaries of the event $p_{0,\alpha/2} \leq t(y,\theta) \leq p_{0,1-\alpha/2}$ is called "inverting the pivot" and gives us the confidence interval we want.

Suppose we create another pivot $u(y,\theta) = g(t(y,\theta))$ where $g(\cdot)$ is a monotonic transfor-

mation. Suppose also that the transformation is increasing. We can write the distribution of $u(y,\theta)$ as P_1, which has relevant quantiles $p_{1,\alpha/2}$ and $p_{1,1-\alpha/2}$, such that

$$\Pr(p_{1,\alpha/2} \leq u(y,\theta) \leq p_{1,1-\alpha/2}) = 1-\alpha. \tag{4.14}$$

Now, consider the $100q\%$ quantile, $p_{1,q}$, of the distribution P_1

$$\begin{aligned} q &= \Pr(u(y,\theta) \leq p_{1,q}) && (4.15)\\ &= \Pr(g(t(y,\theta)) \leq p_{1,q}) && (4.16)\\ &= \Pr(t(y,\theta) \leq g^{-1}(p_{1,q})). && (4.17) \end{aligned}$$

where the last line comes from the monotonicity of $g(\cdot)$. But from the distribution P_0 we also have

$$q = \Pr(t(y,\theta) \leq p_{0,q}). \tag{4.18}$$

This means that for all $q \in [0,1]$,

$$\Pr(t(y,\theta) \leq g^{-1}(p_{1,q})) = \Pr(t(y,\theta) \leq p_{0,q}). \tag{4.19}$$

But this means that $g^{-1}(p_{1,q}) = p_{0,q}$. So inverting the $u(y,\theta)$ pivot means we find θ such that

$$\begin{aligned} p_{1,\alpha/2} \leq u(y,\theta) \leq p_{1,1-\alpha/2} \quad &\text{or} && (4.20)\\ p_{1,\alpha/2} \leq g(t(y,\theta)) \leq p_{1,1-\alpha/2} \quad &\text{or} && (4.21)\\ g^{-1}(p_{1,\alpha/2}) \leq t(y,\theta) \leq g^{-1}(p_{1,1-\alpha/2}) \quad &\text{or} && (4.22)\\ p_{0,\alpha/2} \leq t(y,\theta) \leq p_{0,1-\alpha/2}. && (4.23) \end{aligned}$$

This means that the confidence intervals generated by either pivot are the same. We have

similar computations and the same conclusion if our monotonic transformation is decreasing.

When performing Bayesian analysis we get a different result. Consider what happens if we take a monotonic transformation of our pivot $u(y,\theta) = g(t(y,\theta))$. The new pseudo-likelihood generated by this transformed pivot is given by the usual density calculus

$$p_1(u) = p_0(g^{-1}(u)) \left| \frac{1}{g'(g^{-1}(u))} \right| \tag{4.24}$$

We can re-write this as a function of y and θ

$$p_1(u(y,\theta)) = p_0(g^{-1}(u(y,\theta))) \left| \frac{1}{g'(g^{-1}(u(y,\theta)))} \right| \tag{4.25}$$

and recognizing that $g^{-1}(u) = t$ we have

$$p_1(u(y,\theta)) = p_0(t(y,\theta)) \left| \frac{1}{g'(t(y,\theta))} \right| \tag{4.26}$$

but $p_0(t(y,\theta))$ is the pseudo-likelihood of $t(y,\theta)$. So all we have done is to add in a Jacobian factor to our pseudo-likelihood. But this means that our pseudo-posterior might also differ by this Jacobian factor, depending on whether or not θ appears in $g'(t(y,\theta))$. The choice of pivot appears as if it might alter the inference we perofrm, a clear difference from the frequentist approach. If the choice of pivot affects inference, then the immediate questions raised are "which pivot should we pick for optimal inference?", and "are there any criteria we can use to select such a pivot?" Thus, it would be helpful to have some principle by which we can select a pivot to use for a pseudo-Bayesian analysis.

We will look at a simulation example to see if the choice of pivot does indeed make a difference in inference, and to see if we might get an idea of how to choose which pivot to use. We do not need to worry about misspecified models just yet. We are more concerned at the moment with using pivots in a Bayesian context.

We generated count data from a Poisson distribution with mean parameter $\theta = 3$. The

likelihood and log-likelihood of this distribution are

$$L(\mathbf{y} \mid \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{y_i}}{y_i!} \tag{4.27}$$

$$l(\mathbf{y} \mid \theta) = -n\theta + \sum_{i=1}^{n} y_i \log(\theta) - \sum_{i=1}^{n} y_i! \tag{4.28}$$

The pivots we used were our original pivot, based on the likelihood, and the exponential of this pivot. That is we have $t(\mathbf{y}, \theta)$ and $u(\mathbf{y}, \theta)$ as

$$t(\mathbf{y}, \theta) = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} l_\theta(y_i, \theta)^2\right]^{-1/2} \frac{1}{n}\sum_{i=1}^{n} l_\theta(y_i, \theta) \tag{4.29}$$

$$= \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} \left(\frac{y_i}{\theta} - 1\right)^2\right]^{-1/2} \frac{1}{n}\sum_{i=1}^{n} \left(\frac{y_i}{\theta} - 1\right) \tag{4.30}$$

$$u(\mathbf{y}, \theta) = e^{t(\mathbf{y},\theta)} \tag{4.31}$$

so that $g(t) = \exp(t)$. We have that their approximate distributions are given by

$$t(\mathbf{y}, \theta) \dot{\sim} N(0, 1) \tag{4.32}$$

$$u(\mathbf{y}, \theta) \dot{\sim} \ln N(0, 1). \tag{4.33}$$

We will refer to them as the "normal" and "log-normal" pivots for ease of discussion. Their pseudo-likelihoods are

$$p_0(t(\mathbf{y}, \theta)) \propto e^{-t(\mathbf{y},\theta)^2} \tag{4.34}$$

$$p_1(u(\mathbf{y}, \theta)) \propto e^{-t(\mathbf{y},\theta)^2} e^{-t(\mathbf{y},\theta)} = e^{-t(\mathbf{y},\theta)^2 - t(\mathbf{y},\theta)} \tag{4.35}$$

where line (4.35) comes from line (4.26) and the fact that $g'(t) = \exp(t)$. For both pivots, we used a Gamma(shape=1, scale=4) prior distribution on θ as a relatively diffuse prior. That

density is written as

$$\pi(\theta) = \frac{1}{\Gamma(1)4^1}\theta^{1-1}e^{-\frac{\theta}{4}}. \tag{4.36}$$

This is also an Exponential distribution, but thinking of it as a Gamma will make it conceptually easier to do the proper Bayesian analysis later. Multiplying the prior by our pseudo-likelihoods, we get pseudo-posteriors described by

$$\pi_0(\theta \mid \mathbf{y}) \propto e^{-t(\mathbf{y},\theta)^2-\frac{\theta}{4}} \tag{4.37}$$

$$\pi_1(\theta \mid \mathbf{y}) \propto e^{-t(\mathbf{y},\theta)^2-t(\mathbf{y},\theta)-\frac{\theta}{4}} \tag{4.38}$$

where line (4.37) is proportional to the pseudo-posterior using the pseudo-likelihood from the normal pivot and line (4.38) is proportional to the pseudo-posterior using the pseudo-likelihood from the log-normal pivot.

Before running the full simulation study, we generated test runs at sample sizes 10 and $1,000$ for both the normal and log-normal pivots to check the behavior of Markov chain Monte Carlo (MCMC) simulation as an approximation to the distributions of our pseudo-posteriors. We let the burn-in be $1,000$ points, then kept the next $10,000$ points and thinned those by 10 for a MCMC chain of length $1,000$. We looked at plots of the chains generated and ran the `acf` and `effectiveSize` diagnostics from the `R` package `coda`. Plots of the chains can be seen in Figure 4.1, and they look like they have reached stationarity. In general we hope to see the "fuzzy caterpillar" graph, and that is what we see here. The calculated `effectiveSize` was at least $1,000$ for each chain which matches their actual lengths. Autocorrelation graphs are shown in Figure 4.2, and we see very little autocorrelation between points for each run, so they roughly simulate independent draws from the pseudo-posterior distributions. Overall, the diagnostics do not indicate any serious problems with our Markov chains.

Given that the diagnostics indicated that our MCMC chains would provide efficient approximations to the pseudo-posterior distributions, we proceeded with a full simulation study

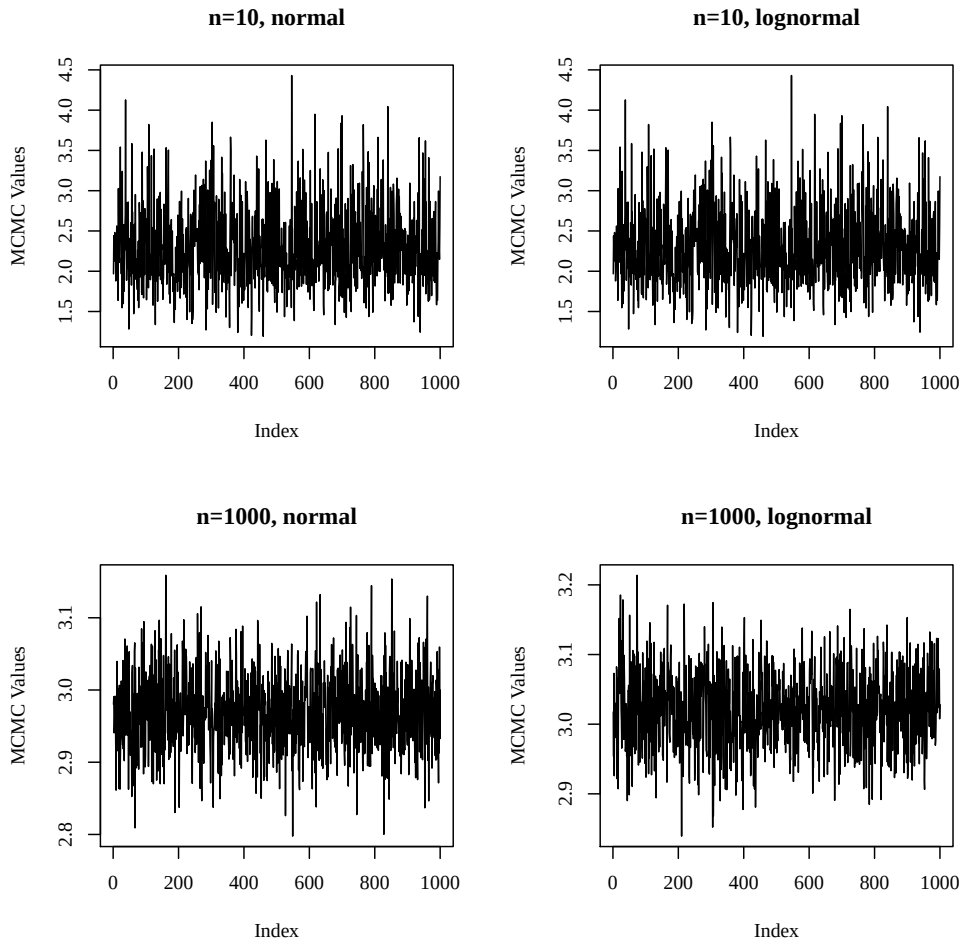

Figure 4.1: MCMC runs from differnt pivots at different sample sizes. Sample sizes of simulated data are 10 and $1,000$. Pivots used are normal and log-normal.

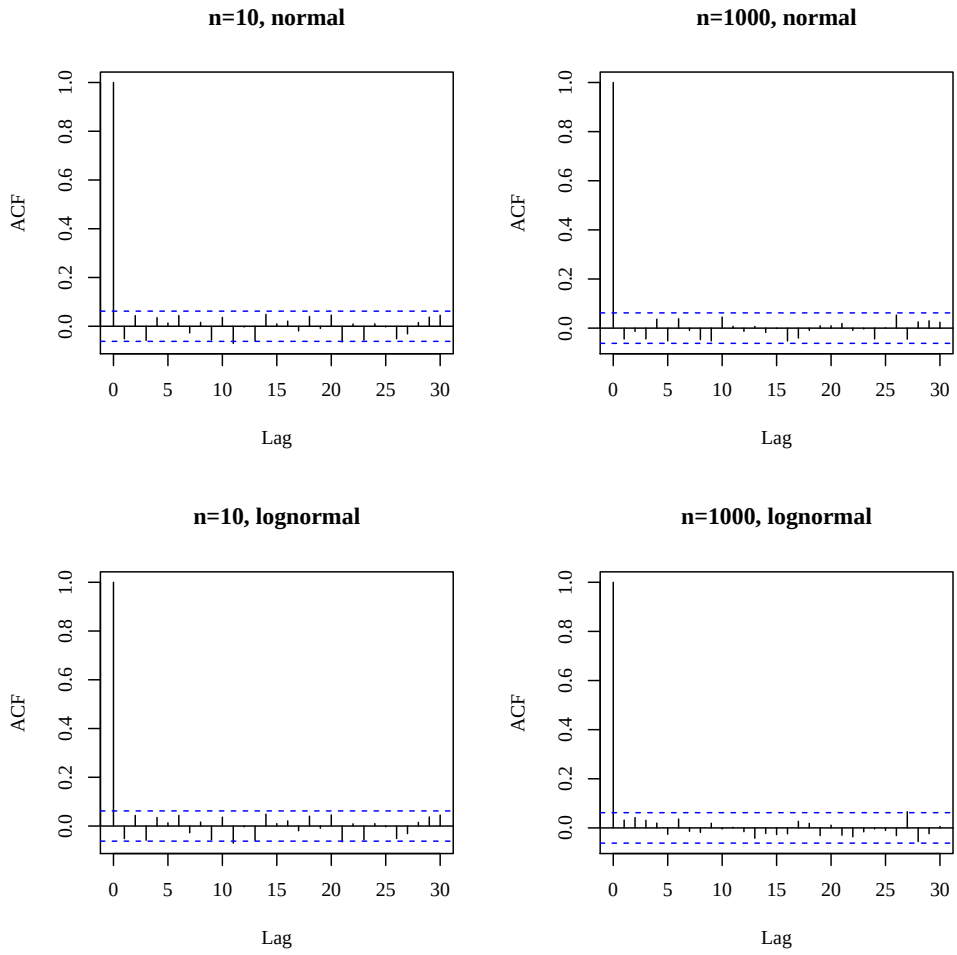

Figure 4.2: Autocorrelation graphs from MCMC chains. Sample sizes of simulated data are 10 and $1,000$. Pivots used are normal and log-normal.

Pivot	n=10	n=1,000
Normal	95.3%	95.7%
log-Normal	83.4%	85.1%

Table 4.1: Coverage rates from pseudo-Bayesian analysis using both approximately normally distributed and approximately log-normally distributed pivots. Coverage was designed to be 95% in all cases. Sample sizes tested were 10 and $1,000$.

to check coverage rates of confidence intervals generated by both the normal and log-normal pivots. We used the same priors, pseudo-likelihoods, and pseudo-posteriors as described beginning in line (4.27) and ending in line (4.38). As in the test run, sample sizes were set to 10 and $1,000$. We again used the Metropolis-Hastings algorithm to generate draws from these distributions. As before, we used a $1,000$ value burn-in period, kept the subsequent $10,000$ values in the chain, and then thinned by 10 to get $1,000$ simulated values from the two pseudo-posterior distributions. The 2.5% and 97.5% quantiles from each thinned chain of $1,000$ values generated one confidence interval for the true parameter. We generated $1,000$ confidence intervals in this fashion for each sample size and each pivot. The confidence level was set at 95% throughout.

The coverage results of the simulation study can be seen in Table 4.1. In both sample sizes, the approximately normally distributed pivot generated confidence interval coverage very close to the nominal 95%. The approximately log-normally distributed pivot generated confidence interval coverage that was closer to 85%. These coverage rates were similar at both sample sizes of 10 and $1,000$. It appears as if choice of pivot can affect coverage rates, and that the coverage rate might not necessarily improve with increased sample size.

To investigate why the log-normal pivot provided less accurate coverage of the true parameter than the normal pivot, we looked at the midpoints of the confidence intervals and their lengths. Looking at the midpoints gives a good indication of whether the intervals are biased away from the true value. Ideally, the distribution of confidence interval midpoints would be centered around the true value. Confidence interval lengths are a measurement

of the variability in the midpoints. The more variable our midpoints are, the larger our confidence interval lengths should be to compensate and yield correct coverage rates.

Investigating these midpoints and lengths we discovered that there was a difference in the distribution of midpoints of the confidence intervals generated by the normal and lognormal pivots. The confidence intervals based on normal pivots had a distribution of midpoints that was pretty symmetric around the true value. The distribution of the midpoints generated from the log-normal pivot, on the other hand, were centered at a value larger than the truth. The histograms of these midpoints can be seen in Figure 4.3. It appears as if the confidence intervals generated by the log-normal pivot are centered too high for both sample sizes tested. This would not necessarily be a problem if the confidence interval lengths are longer to adjust for that bias.

Checking the distribution of confidence interval lengths gives us a more complete understanding of what is happening with the confidence interval coverage rates. The distribution of lengths of confidence intervals did not change much between the normal and log-normal pivots. At $n = 10$, the lengths of confidence intervals from the normal pivot are centered around 2.1. The corresponding lengths of confidence intervals from the log-normal pivot are centered around 2.5. Though the log-normal pivots are larger, this is not enough to overcome the bias of their midpoints. Both sets of confidence interval lengths did, however, decrease with sample size which corresponds with the fact that the midpoints were less variable at larger n and overall coverage didn't change from small n to large n. Histograms of the lengths can be seen in Figure 4.4.

Our analysis indicates that it is the bias of the confidence intervals that is causing undercoverage with the log-normally generated intervals. From these results, we decided to analyze the pseudo-likelihoods themselves to see if we could quantify the bias. Granted, we are interested in the bias in the pseudo-posterior distributions, but our pseudo-posteriors only differ in the choice of pseudo-likelihood, so we start there.

We computed the modes of the normal pivot pseudo-likelihood and of the log-normal pivot pseudo-likelihood. The mode of pseudo-likelihood for the normally distributed pivot

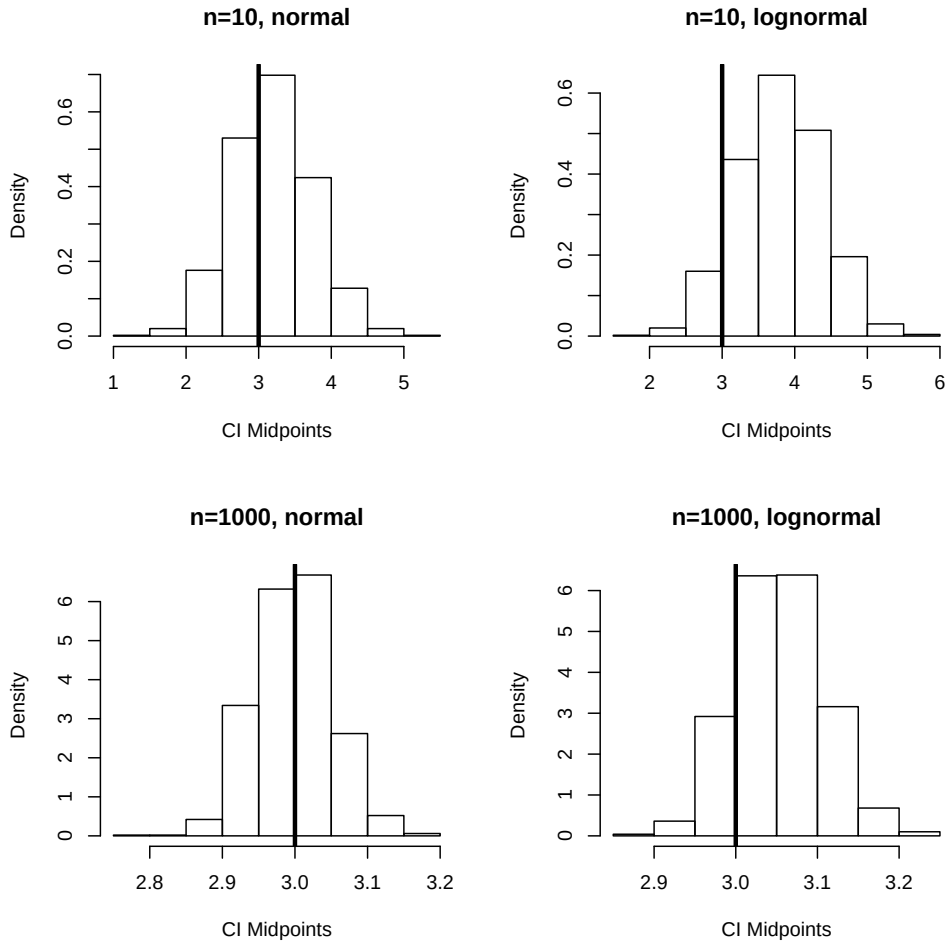

Figure 4.3: Histograms of the midpoints from confidence intervals generated in the pseudo-Bayesian procedure. The true value of the parameter of interest is three, as deonted by the thick lines.

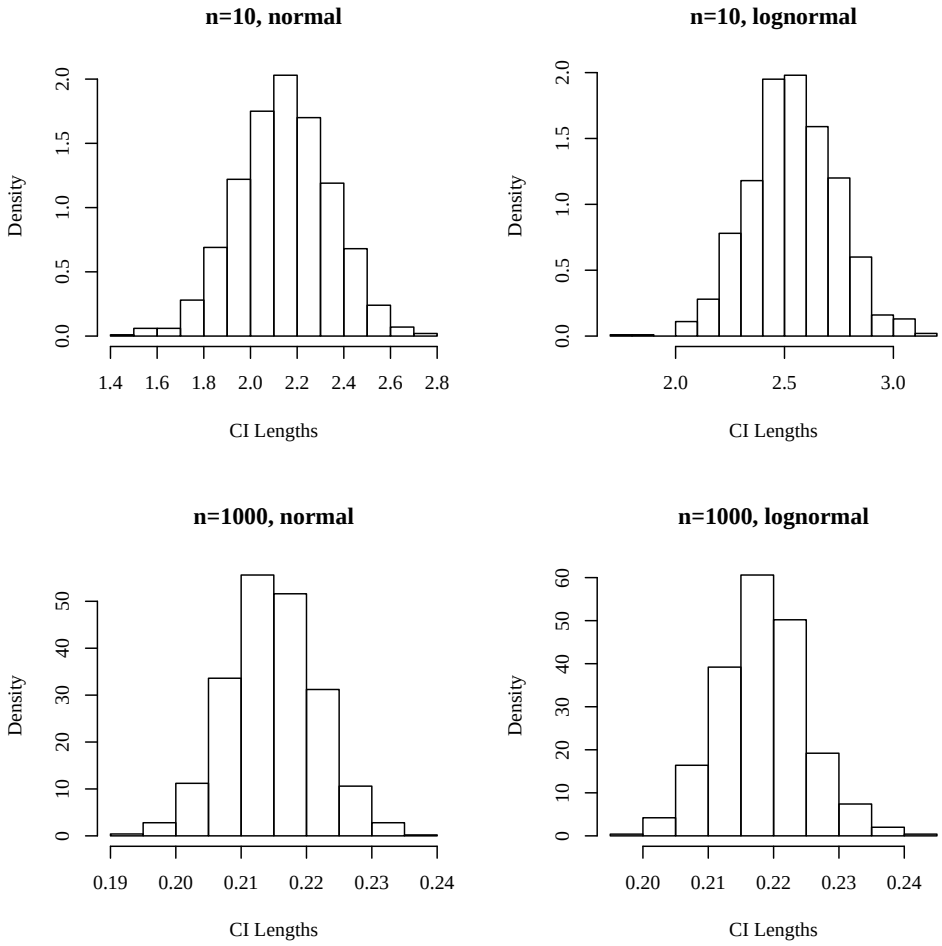

Figure 4.4: Histograms of the lengths from confidence intervals generated in the pseudo-Bayesian procedure. The lengths are pretty consistent across pivots, and decrease with sample size.

is $\bar{Y}$, which is also the MLE, and is unbiased for the true parameter value. The mode of the pseudo-likelihood for the lognormally distributed pivot in this example is computed to be $\bar{Y} + \frac{1}{2n} + \sqrt{(\frac{1}{4n})^2 + \frac{\bar{Y}}{n}}$, which is biased high with respect to the true θ, though it is asymptotically unbiased. Since the prior was the same for each pivot, this explains the biases shown in the histograms in Figure 4.3 and why those biases appear to decrease in the larger sample size.

To summarize this section, we have shown that the inference based on pseudo-likelihoods can vary depending upon choice of pivot. Particularly, the mode of the pseudo-likelihood and hence the pseudo-posterior might be biased depending on choice of pivot, which can affect the centering of confidence intervals. This can reduce the true coverage percentage of those intervals.

4.3 Pivot Selection Principle

We have seen in the last section that the mode of the pseudo-likelihood can be biased away from the true parameter value. In this section we show that our likelihood pivot will have a pseudo-likelihood whose mode is at $\hat{\theta}_{MMLE}$. While not always an unbiased estimator of the pseudo-true parameter, $\hat{\theta}_{MMLE} \rightarrow_{a.s.} \theta^*$ and is thus a consistent estimator of the pseudo-true parameter. The fact that the mode of the pseudo-likelihood and the original misspecified likelihood are at the same point with respect to θ forms the basis of our selection principle.

In the previous section, we stated that the pseudo-likelihood based on the pivot using the Poisson likelihood had its mode at $\hat{\theta}_{MMLE}$. Let us now analyze the pseudo-likelihood for our pivot in the general case and see if that always holds true. Recall we have

$$t(y,\theta) = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} l_\theta(y_i,\theta)^2\right]^{-1/2} \frac{1}{n}\sum_{i=1}^{n} l_\theta(y_i,\theta) \dot{\sim} N(0,1) \tag{4.39}$$

The pseudo-likelihood of this pivot is

$$p_0(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \tag{4.40}$$

which is the standard normal denstiy and thus has its mode at $t = 0$. But, $t = 0$ implies

$$0 = \sqrt{n} \left[\frac{1}{n} \sum_{i=1}^{n} l_\theta(y_i, \theta)^2 \right]^{-1/2} \frac{1}{n} \sum_{i=1}^{n} l_\theta(y_i, \theta) \tag{4.41}$$

$$\Rightarrow \tag{4.42}$$

$$0 = \frac{1}{n} \sum_{i=1}^{n} l_\theta(y_i, \theta). \tag{4.43}$$

The θ value that satisfies this requirement is, by definition, $\theta = \hat{\theta}_{MMLE}$. So the mode of the pseudo-likelihood of our pivot as a function of θ and in the general case occurs at $\hat{\theta}_{MMLE}$. From this insight, we compose our selection principle for choosing a pivot in our pseudo-Bayesian procedure. If we use a working model with likelihood

$$\prod_{i=1}^{n} L(y_i \mid \theta) \tag{4.44}$$

we should choose a pivotal quantity $t(\mathbf{y}, \theta)$ whose pseudo-likelihood, $p_0(t(\mathbf{y}, \theta))$, is maximized at the same θ-value that maximizes the working likelihood, line (4.44). This value is $\theta = \hat{\theta}_{MMLE}$.

This principle justifies the use of our pivot in our pseudo-Bayesian analysis. It also gives us a principle that we can use to decide if another pivot might also be a good choice for our pseudo-Bayesian procedure. Since any monotonic transformation of a pivot is also a pivot, we want to know which transformations of our pivot also result in pivots that satisfy the principle of preservation of modal values. We know from the last section that not all transformations preserve this quality, as the exponential of our pivot did not have its mode at $\hat{\theta}_{MMLE}$. Our analysis of the general case can help us determine what class of transformations

preserve this modal property that undergirds our selection principle.

Consider a monotonic transformation of our pivot, $u = g(t)$. If $g(t)$ is increasing, then we can simplify the pseudo-likelihood of u as

$$p_u(u) = p_0(g^{-1}(u))\frac{1}{g'(g^{-1}(u))} \tag{4.45}$$

where the absolute value bars have been removed since they are superfluous. Writing this as a function of y and θ, we have

$$p_u^*(u(y,\theta)) = p_0^*(g^{-1}(u(y,\theta)))\frac{1}{g'(g^{-1}(u(y,\theta)))} \tag{4.46}$$

$$= p_0^*(t(y,\theta))\frac{1}{g'(t(y,\theta))} \tag{4.47}$$

where the derivative of $g(\cdot)$ is taken with respect to t not θ. To find the mode of our new pseudo-likelihood, we take the derivative with respect to θ and then set that equal to zero. Note that primes in what follows represent derivatives of functions that are taken with respect to the immediate argument of that function. For example, $g'(t(y,\theta))$ is the derivative taken with respect to t and $t'(y,\theta)$ is the derivative taken with respect to θ. So the full partial derivative $\frac{\partial}{\partial\theta}g(t(y;\theta))$ is expressed as $g'(t(y,\theta))t'(y,\theta)$.

$$\frac{\partial}{\partial\theta}p_u^*(u(y,\theta)) = \frac{\partial}{\partial\theta}\left[p_0^*(t(y,\theta))\frac{1}{g'(t(y,\theta))}\right] \tag{4.48}$$

$$= \frac{p_0'^*(t(y,\theta))t'(y,\theta)g'(t(y,\theta)) - p_0^*(t(y,\theta))g''(t(y,\theta))t'(y,\theta)}{g'(t(y,\theta))^2} \tag{4.49}$$

$$0 = \frac{p_0'^*(t(y,\theta))t'(y,\theta)g'(t(y,\theta)) - p_0^*(t(y,\theta))g''(t(y,\theta))t'(y,\theta)}{g'(t(y,\theta))^2} \tag{4.50}$$

$$\Rightarrow \tag{4.51}$$

$$0 = p_0'^*(t(y,\theta))t'(y,\theta)g'(t(y,\theta)) - p_0^*(t(y,\theta))g''(t(y,\theta))t'(y,\theta) \tag{4.52}$$

Continuing with the algebra gives us the following criterion

$$0 = p_0'^*(t(y,\theta))t'(y,\theta)g'(t(y,\theta)) - p_0^*(t(y,\theta))g''(t(y,\theta))t'(y,\theta) \tag{4.53}$$

$$p_0'^*(t(y,\theta))t'(y,\theta)g'(t(y,\theta)) = p_0^*(t(y,\theta))g''(t(y,\theta))t'(y,\theta) \tag{4.54}$$

$$\frac{p_0'^*(t(y,\theta))t'(y,\theta)}{p_0^*(t(y,\theta))} = \frac{g''(t(y,\theta))t'(y,\theta)}{g'(t(y,\theta))} \tag{4.55}$$

$$\frac{d}{d\theta}\log[p_0^*(t(y,\theta))] = \frac{d}{d\theta}\log[g'(t(y,\theta))]. \tag{4.56}$$

assuming that $g'(t(y,\theta))$ is not identically zero. But $g'(t(y,\theta))^2$ is identically zero only if $g(t)$ is a constant function. We will exclude those cases since they make for very uninteresting transformations. The mode for the pseudo-likelihood of our transformed pivot is the θ value that satisfies (4.56). As a side note, we see that the left-hand side of the equation is defined entirely in terms of the pseudo-likelihood of our original pivot, while the right-hand side is defined entirely in terms of the transformation function.

Returning now to the pivot determined by the misspecified model, we showed earlier in this section that the pseudo-likelihood of this pivot has its modal value at $\theta = \hat{\theta}_{MMLE}$. By the definition of a mode this causes the left-hand side of (4.56) to be zero, meaning that any transformation of our pivot that also results in the mode of the transformed pivot's pseudo-likelihood being at $\hat{\theta}_{MMLE}$ must satisfy

$$0 = \frac{d}{d\theta}\log[g'(t(y,\theta))]\Big|_{\theta=\hat{\theta}_{MMLE}}. \tag{4.57}$$

There are many transformations that satisfy this criterion. One particular class of such transformations is the class of linear transformations $g(t) = at + b$ for constants a and b. $g'(t) = a$ is then a constant, and the derivative of the log of that with respect to θ is always zero. Since the original pivot was approximately normally distributed, linear transformations of that pivot are also approximately normally distributed. So one class of pivots created from our misspecified likelihoods whose pseudo-likelihoods have modes at $\hat{\theta}_{MMLE}$ are our pivots

and linear transformations of them. Based on our principle above, this class is a good choice of pivot to use for pseudo-Bayesian analysis.

4.4 Proper Bayesian vs. Pseudo-Bayesian Simulation Example

In this section we investigate our pseudo-Bayesian analysis in competition with standard Bayesian analysis performed with a misspecified likelihood. The setting is similar to the previous example. We have discrete data that are, perhaps, counts of objects or events. One might naturally assume a Poisson data-generating distribution, especially if we have reason to believe that the real world characteristics of the Poisson distribution are met. For example the counts might represent data we assume is approximately binomial with n very large and p very small. In that case the Poisson distribution is a well-known approximation of the binomial. Alternatively, it might be reasonable to assume that the data occur within the context of a Poisson process. Regardless, given that we are also working with small sample sizes, we might allow for some amount of overdispersion due to natural variability. However it is impossible to know whether the overdispersion is due to small sample variation or true overdispersion. In this section, we compare a proper Bayesian analysis done with a Poisson likelihood to our pseudo-Bayesian analysis done with our pivot from a Poisson likelihood. The parameter of interest is the mean parameter of the assumed Poisson model. We use the same prior distribution on the parameter of interest in each case, so the main difference in the methods is the use of the likelihood or pseudo-likelihood function.

We generate data from a Negative Binomial distribution with mean 10 and dispersion parameter 10 as well. This is equivalent to a negative binomial with success probability 0.5 and $n = 10$ or mean 10 and variance 20. For the standard Bayesian analysis, we assume a misspecified Poisson likelihood on our data with density function $f(y;\theta) = e^{-\theta}\theta^y/y!$. We use a Gamma($k = 1, \lambda = 4$) prior distribution, where the parameters are shape and scale respectively. The exact form of the prior density is $\pi(\theta; k, \lambda) = \theta^{k-1}e^{-\theta/\lambda}\lambda^{-k}/\Gamma(k)$. Since the family of Gamma distributions is the conjugate prior for a Poisson likelihood model, the posterior is also a Gamma distribution. In particular, it will be a Gamma($1 + \sum y_i, 4/(1 +$

$4n)$), where $\sum y_i$ is the sum of the data and n is the sample size. Confidence intervals can be generated by computing the relevant quantiles of this Gamma distribution.

We wish to compare the proper Bayesian analysis above with pseudo-Bayesian analysis using our prior as performed in the previous section. For our prior distribution, we will use the Gamma$(1, 4)$ distribution described in the previous paragraph. In order to compute the pseudo-posterior, we'll use a Metropolis-Hastings algorithm to generate the MCMC chains as we did in the previous section. We will let the chain burn in for 1000 values, then keep the next $10,000$ values, thinning by 10 to get $1,000$ values that we will use to approximate the pseudo-posterior distribution of $\theta \mid y$. As before, we wil run some test chains and check diagnostics first before doing the full simulation study. We want to see if misspecifying the model affects our Markov chain efficiency.

We ran test chains for each sample size from $n = 10$ to $n = 100$. The resulting MCMC chains can be seen in Figure 4.5. Again, we are looking for the "fuzzy caterpillar" graphs, and these look pretty good. Additionally, we ran the `acf` function from `R` package `coda`. The results of that can be seen in Figure 4.6. Here too, we see that autocorrelation is not that serious a problem across all runs. Autocorrelation in the sample size $n = 10$ chain is the worst of all sample sizes tested, but it is not so bad that our analysis is completely invalidated. We also ran the `effectiveSize` function on each of the thinned test chains. The results can be seen in Table 4.2. The computed effective sizes are not as good as in the last section, but again, they are not terrible. All but two computed effective sizes are above 900. The smallest is 600, and the second smallest is almost 700. We note that the smallest computed effective size is for the smallest sample size tested. Throughout our work, the very smallest sample sizes are the ones that give us the most difficulty, so this is not surprising. It does appear that the efficiency with which the Markov chains approximate our pseudo-posterior is less than in the last section. This is likely due to the fact that we have misspecified our model in this section. However, the Markov chains wil still give us a reasonable approximation to the distribution of our pseudo-posteriors.

We proceed with our pseudo-Bayesian analysis and the comparison with a proper Bayesian

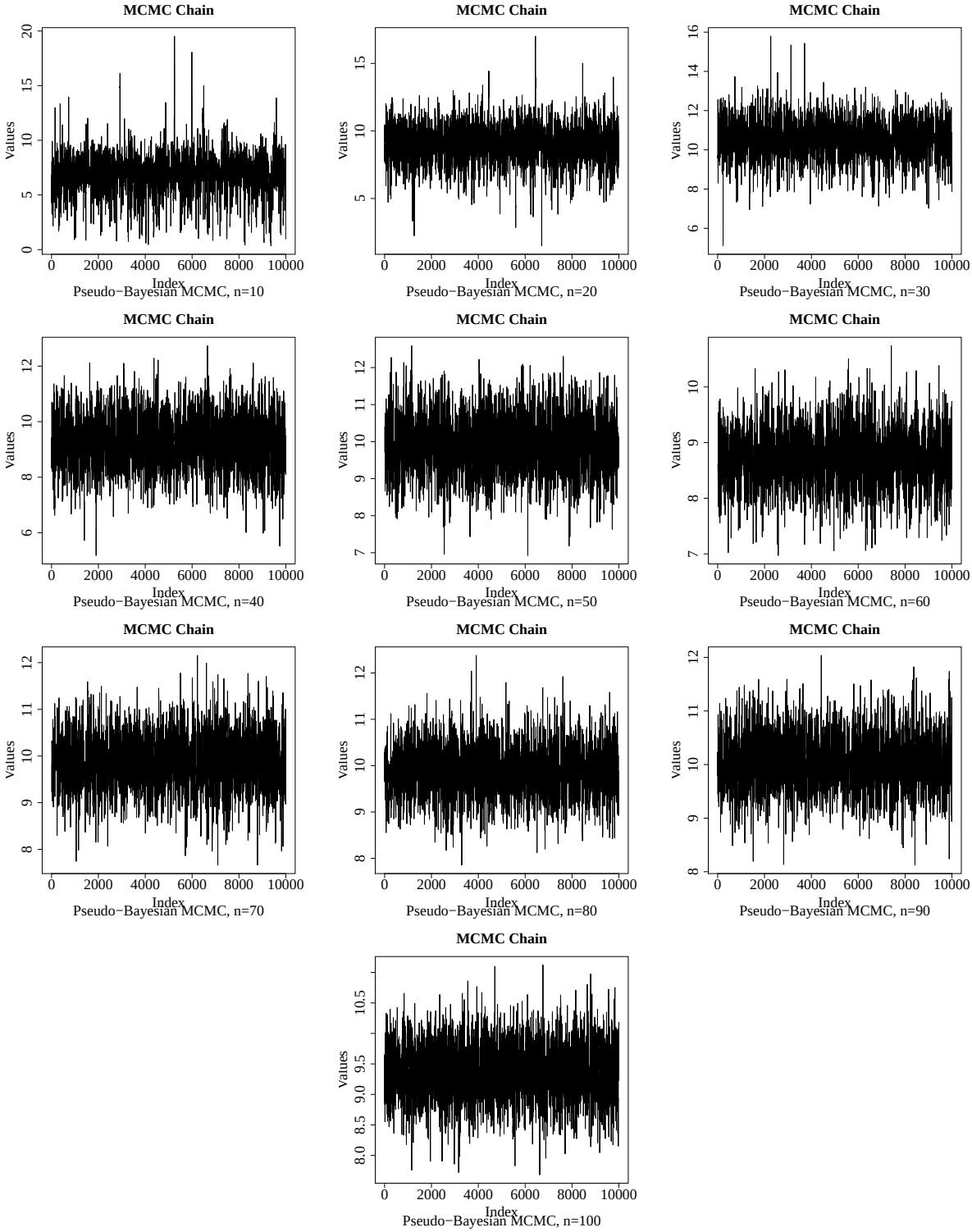

Figure 4.5: Sample MCMC runs for pseudo-Bayesian inference by sample size. Burn-in was $1,000$ values. The next $10,000$ values were kept, and then thinned by a factor of 10 to yield $1,000$ values in each chain.

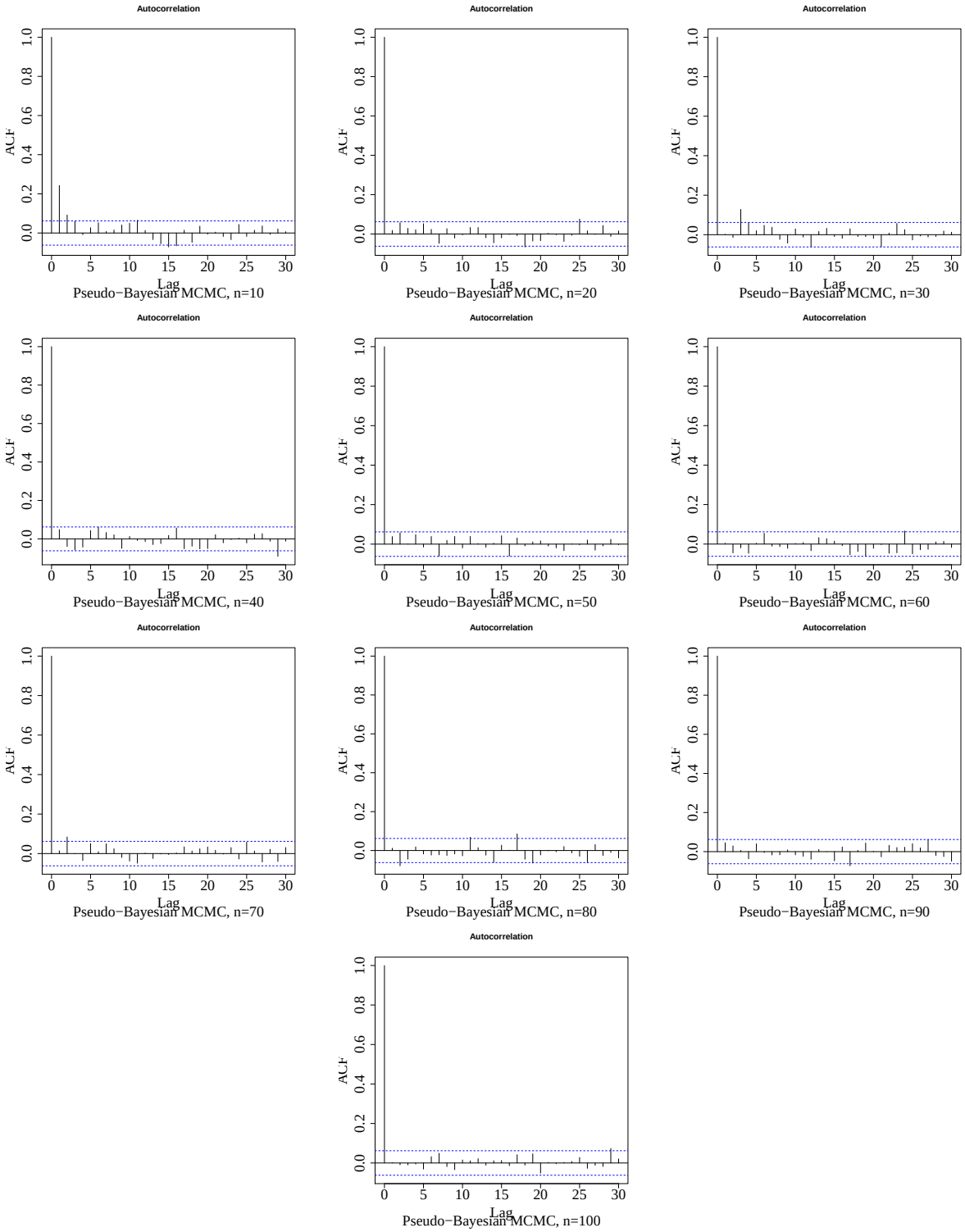

Figure 4.6: Autocorrelation graphs for the sample MCMC chains.

Sample Size	10	20	30	40	50	60	70	80	90	100
effectiveSize	609	1000	697	1098	827	1000	819	1141	912	1000

Table 4.2

analysis via a simulation study. We generated $1,000$ Negative Binomial datasets at each sample size tested from $n = 10$ to $n = 100$. For each dataset, we computed the 2.5% and 97.5% quantiles of the proper posterior distribution for our proper Bayesian confidence interval using the Gamma$(1, 4)$ prior earlier discussed. Given that the Gamma distribution is the conjugate distribution for a Poisson likelihood, this means that the misspecified posteriors were Gamma$(1 + \sum y_i, \frac{4}{1+4n})$. We also computed confidence intervals using the pseudo-posterior distribution resulting from the pseudo-likelihood and the same Gamma$(1, 4)$ prior. For each simulated dataset, we ran an MCMC chain via a Metropolis-Hastings algorithm to generate $1,000$ points from the pseudo-posterior distribution. We then used the 2.5% and 97.5% quantiles of the chain to generate a confidence interval based on that dataset. For each dataset, we determined whether the true mean of 10 was covered by the the proper Bayesian and pseudo-Bayesian intervals and reported the coverage rate for each method by sample size. The results can be seen in Figure 4.7.

The results for our pseudo-Bayesian method compare favorably to the results of the proper Bayesian interval coverage. As can be seen from the graph, the proper Bayesian confidence intervals only cover the true parameter around 85% or less of the time. The pseudo-Bayesian analysis, however, hovers around the true 95% coverage throughout all sample sizes tested. To see what might be causing the difference, we look at histograms of midpoints and lengths of each set of confidence intervals. Boxplots of midpoints can be seen in Figures 4.8. while boxplots of confidence interval lengths can be seen in Figures 4.9. When looking at midpoints, we do not see a huge difference between the two methods. The pseudo-Bayesian pivot method at sample size $n = 10$ looks to be biased low, but after that the midpoints cluster around the true value of $\theta = 10$. What is different is the lengths of confidence intervals generated by the

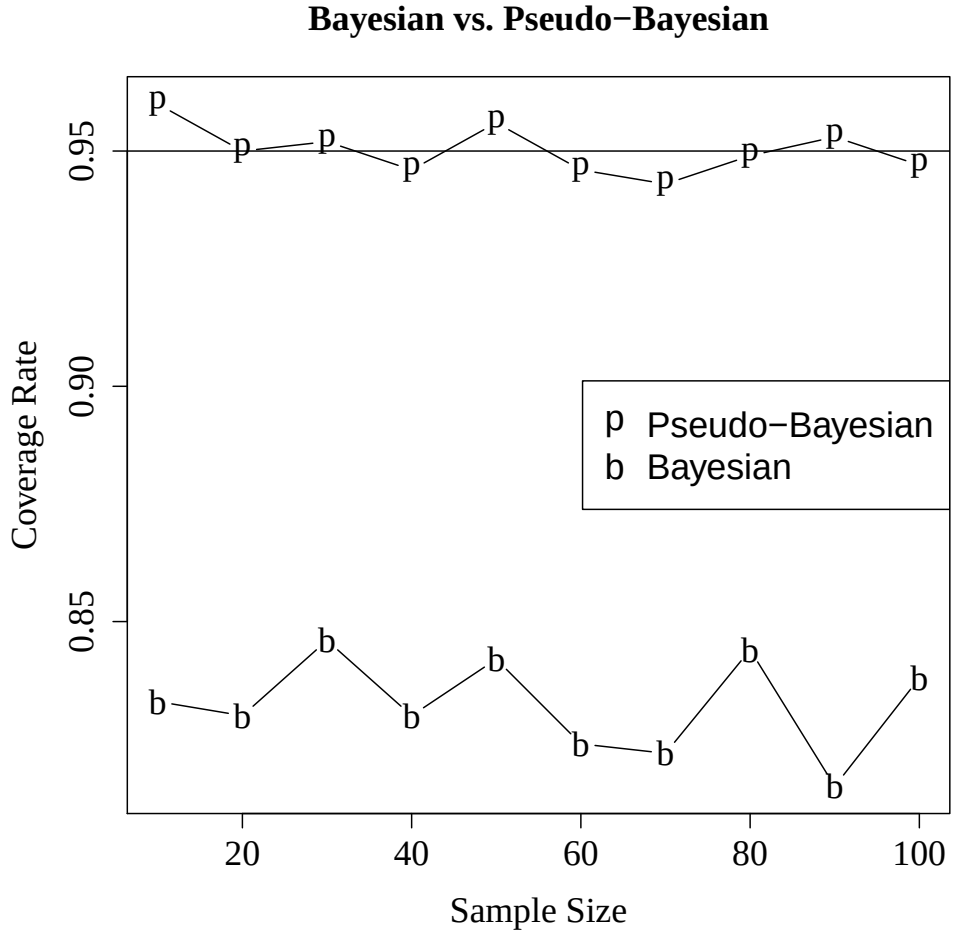

Figure 4.7: Coverage Rates comparing proper Bayesian versus pseudo-Bayesian by sample size. $1,000$ samples were generated at each sample size.

two methods. In particular the pseudo-Bayesian confidence intervals are longer than their corresponding proper Bayesian confidence intervals for any given sample size. This appears to be the cause of the difference in coverage rates between the two methods. Undercoverage in the proper Bayesian confidence intervals is likely due to the mean-variance relationship in the assumed Poisson distribution. For Poisson distributed random variables, the mean and variance are equal. In our simulation the variance was twice the mean, 20 versus 10. It is likely that the posterior distribution in the proper Bayesian procedure is not dispersed enough to account for the variance in the true data-generating distribution.

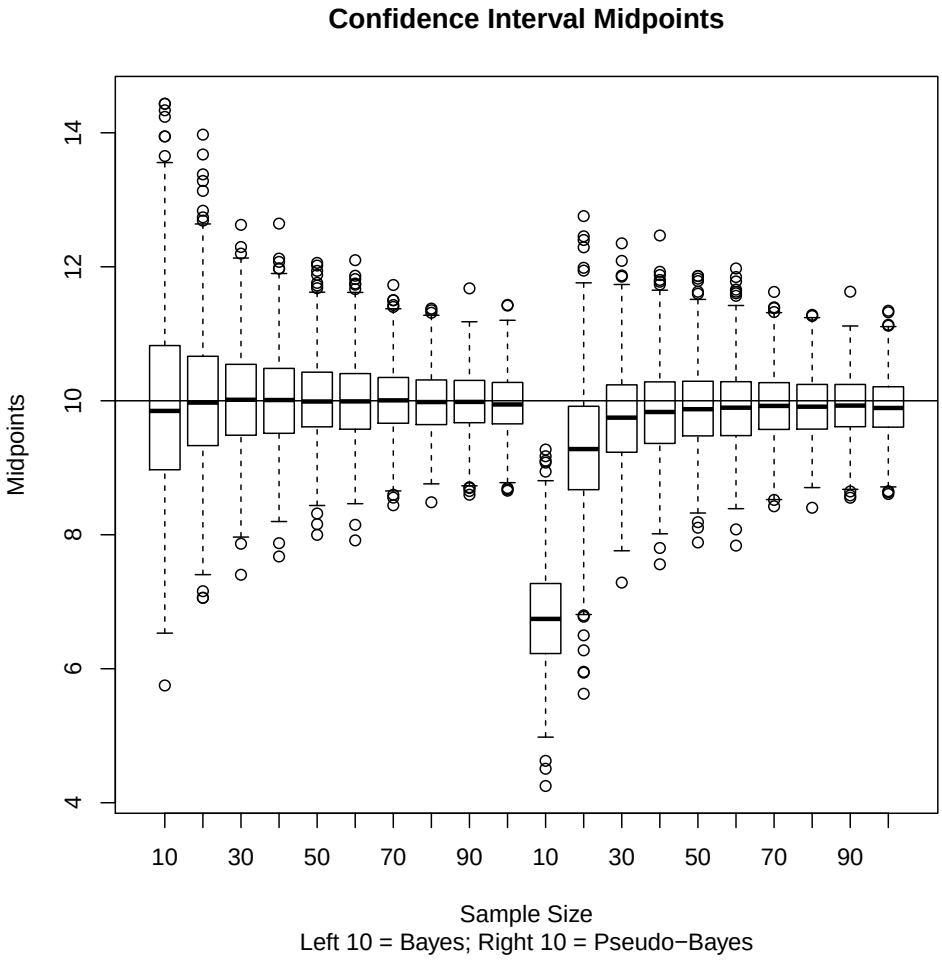

Figure 4.8: Midpoints of confidence intervals from the simulation study for both the proper Bayesian and pseudo-Bayesian procedures. Note that all boxes except for the pseudo-Bayesian box for $n = 10$ are roughly centered at the true value of $\theta = 10$.

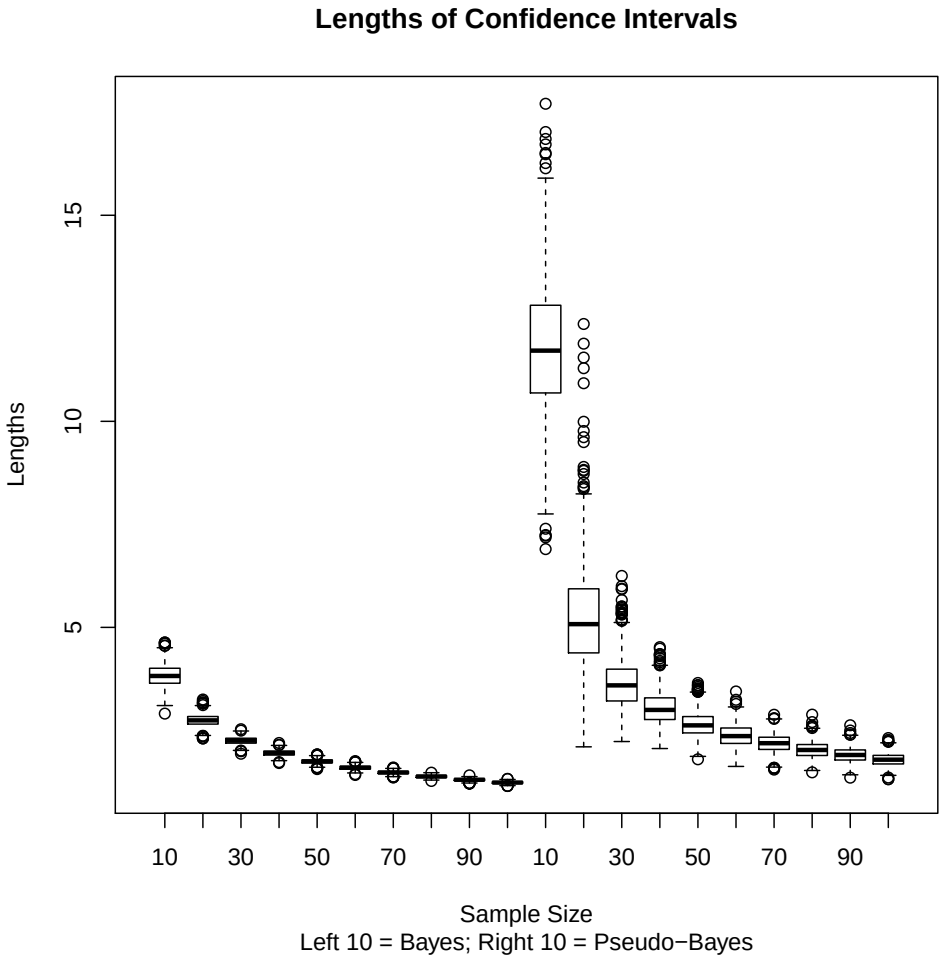

Figure 4.9: Boxplots of lengths of confidence intervals from the simulation study for both the proper Bayesian and pseudo-Bayesian procedures. Note that for any given sample size, the pseudo-Bayesian confidence intervals are longer than their proper Bayesian counterparts.

Chapter 5

DATASET EXAMPLES

Up until now the simulations we have run used simple, artifical datasets. At this point we would like to see how our pivot method compares to other sandwich-based methods on small sample data that we might find in the real world. If we simply obtained a small, real world dataset and computed confidence regions using our pivot and the sandwich, we would have no way of knowing how well they worked. We would not even know whether they covered the pseudo-true parameter or not. In order to test our method we would need to know the pseudo-true parameter; we would have to generate confidence regions from multiple, small random samples; and we would have to compare coverage rates. What we will do in this chapter is perform simulation studies where our data-generating population is actually a large, real world dataset. This will allow us to compare the covarage rates of confidence regions generated by our pivot and the sandwich when the true data-generating population is something we might find in the real world.

We obtained a large dataset from the National Health and Nutrition Examination Survey (NHANES) to use as our population data. In each section of this chapter, we will choose a different model as our working model for that section's simulation study. Fitting a particular working model to the population dataset will give us the pseudo-true parameter for that model. We will then run simulations where we sample small sub-datasets from the larger, population dataset. Doing so simulates the act of obtaining a small sample dataset from a large, real world population. We will then fit the section's model to these sub-datasets, and generate confidence regions for the pseudo-true parameter. Comparing coverage rates between confidence regions generated by our pivot and by the sandwich will tell us if we have improved upon the small sample, sandwich-based confidence region coverage.

In the first section we fit a multiple regression model to our data. We model one response variable that is dependent on five covariates. This is similar to what we did in Section 3.3, but now we use real world data and have more covariates. In the second section we add survey weights to our experiment. We use the publicly-provided weights to estimate the parameter values for the entire U.S. population, and then use those same weights to sub-sample from our large sample. We use our pivot to generate inference from these sub-samples. In the final section, we fit a semiparametric copula model to two of our data variables. This is very similar to the work done in Subsection 2.3.4, except again we are using real world data.

5.1 Dataset as "Truth" - Regression

In this section we look at an example using a real dataset for our data-generating population. We will select several variables from the dataset and use a multiple linear regression model as our working model for this data. We will then run a simulation study by taking small samples from the entire dataset, fitting the model to each small sample, and computing confidence region coverage of the pseudo-true parameter using our pivot-based method and sandwich-based methods. The pseudo-true parameter itself can be computed by fitting our model to the entire dataset, as that dataset functions as our population. The purpose of this example is to see how our pivot and the sandwich compare in a more realistic regression setting than earlier simulation examples. This is why we use a real dataset as our population.

The population data comes from the NHANES dataset, 2011-2012. We are using the following variables: Body Mass Index (BMI), total kilocalories consumed, sugar consumed, fat consumed, a composite inactivity measure, and gender. The height and weight measurements comprising BMI were made in mobile examination centers by trained workers. The values for kilocalories, sugar, and fat were those amounts consumed the day prior to the interview with each respondent. Sugar and fat are measured in grams. The inactivity measures were reported by the repondent. We use the sum of the total hours spent watching TV or video and the total hours of non-work-related computer use. Each of those was an estimated amount of the total time spent on those activities during the last 30 days prior to

the interview.

BMI is a widely used measure of health and wellness. It is computed by taking an individual's weight and dividing by the square of her height and is reported in units of $\frac{kg}{m^2}$. This measure adjusts for the fact that tall people will naturally weigh more than short people, all other things being equal. The Centers for Disease Control (CDC) list the following BMI standards: < 18.5, underweight; $18.5 - 24.9$, normal or healthy weight; $25.0 - 29.9$, overweight, > 30.0, obese [CDC]. Though BMI standards are a general measure of wellness, they are not a perfect measure of health. For example a well-muscled athlete and a couch potato might have the same BMI but much different levels of overall health. However, for the majority of the population, BMI is a reasonable measure of health and wellness.

Our model in this section will be a multiple linear regression model with BMI as the response variable and the nutrition-related measurements and inactivity measurement as independent variables. Concerning the appropriateness of these variables, we do not expect that the specific quantities of food consumed on the day prior to the interview will greatly affect weight and hence BMI. It is more reasonable that one's weight, and BMI, are more closely related to average food consumption, with possibly some extra emphasis on recent consumption, than food consumed on any given day. If we assume, however, that the amount of food an individual consumed on the day prior to the interview is representative of their average consumption, then it is reasonable to use this data in our model. We also expect weight, and BMI, to be a function of average activity levels. A person with a less active lifestyle, all else being equal, will likely have a larger weight and BMI than a more active person. The inactivity measurement we use is calculated based on the 30-day period prior to the interview. Similar to the nutrition variables, if we assume the prior month was typical for individuals' activity levels, then it is reasonable to use this measurement in our model as well. This model might be of interest to researchers because it could give clues regarding which habits affect BMI most. This could drive research to study which diet and exercise changes have the greatest effect on BMI and consequently overall health. Further, we are quite positive that there are factors other than food and inactivity that affect one's overall BMI

measurement. However, since we are interested in performing inference with misspecified models, this is actually appropriate for our simulation study.

The working model is given below

$$\mathrm{B}_i = \theta_0 + \theta_1 \times \mathrm{K}_i + \theta_2 \times \mathrm{S}_i + \theta_3 \times \mathrm{F}_i + \theta_4 \times \mathrm{I}_i + \theta_5 \times \mathrm{G}_i + \epsilon_i \tag{5.1}$$

$$\epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2) \tag{5.2}$$

with "B" being the BMI measurement, "K" being the kilocalorie consumption, "S" being the amount of sugar consumed, "F" being the amount of fat consumed, "I" being the inactivity measurement, "G" being an indicator variable coded 1 for "female" and 0 for "male", and ϵ being the random error term. We also write $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$. The pivot used is similar in structure to the one studied in Section 3.3. The big difference here is that we have more than one covariate in this example. To account for the full set of parameters, the pivot will have an approximate χ^2 distribution with six degrees of freedom. To construct the pivot, we will need the log-likelihood with respect to $\boldsymbol{\theta}$ which is

$$\sum_{i=1}^{n} l(y_i \mid \boldsymbol{\theta}) = \frac{-1}{2\sigma^2} \sum_{i=1}^{n} [\mathrm{B}_i - (\theta_0 + \theta_1 \times \mathrm{K}_i + \theta_2 \times \mathrm{S}_i + \theta_3 \times \mathrm{F}_i + \theta_4 \times \mathrm{I}_i + \theta_5 \times \mathrm{G}_i)]^2 \tag{5.3}$$

$$= \frac{-1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 \tag{5.4}$$

where we are using ϵ_i, which is a function of $\boldsymbol{\theta}$, as shorthand for $\mathrm{B}_i - (\theta_0 + \theta_1 \times \mathrm{K}_i + \theta_2 \times \mathrm{S}_i + \theta_3 \times \mathrm{F}_i + \theta_4 \times \mathrm{I}_i + \theta_5 \times \mathrm{G}_i)$. The gradient of the log-likelihood is the column vector

$$\sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{\theta}) = \frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} \epsilon_i, \sum_{i=1}^{n} \mathrm{K}_i \epsilon_i, \sum_{i=1}^{n} \mathrm{S}_i \epsilon_i, \sum_{i=1}^{n} \mathrm{F}_i \epsilon_i, \sum_{i=1}^{n} \mathrm{I}_i \epsilon_i, \sum_{i=1}^{n} \mathrm{G}_i \epsilon_i \right]^T . \tag{5.5}$$

We can now define the $\mathbf{B}_n(\boldsymbol{\theta})$ matrix as

$$\mathbf{B}_n(\boldsymbol{\theta}) = \frac{1}{4n\sigma^2}\sum_{i=1}^{n}\begin{pmatrix} \epsilon_i^2 & \mathrm{K}_i\epsilon_i^2 & \mathrm{S}_i\epsilon_i^2 & \mathrm{F}_i\epsilon_i^2 & \mathrm{I}_i\epsilon_i^2 & \mathrm{G}_i\epsilon_i^2 \\ \mathrm{K}_i\epsilon_i^2 & \mathrm{K}_i^2\epsilon_i^2 & \mathrm{K}_i\mathrm{S}_i\epsilon_i^2 & \mathrm{K}_i\mathrm{F}_i\epsilon_i^2 & \mathrm{K}_i\mathrm{I}_i\epsilon_i^2 & \mathrm{K}_i\mathrm{G}_i\epsilon_i^2 \\ \mathrm{S}_i\epsilon_i^2 & \mathrm{S}_i\mathrm{K}_i\epsilon_i^2 & \mathrm{S}_i^2\epsilon_i^2 & \mathrm{S}_i\mathrm{F}_i\epsilon_i^2 & \mathrm{S}_i\mathrm{I}_i\epsilon_i^2 & \mathrm{S}_i\mathrm{G}_i\epsilon_i^2 \\ \mathrm{F}_i\epsilon_i^2 & \mathrm{F}_i\mathrm{K}_i\epsilon_i^2 & \mathrm{F}_i\mathrm{S}_i\epsilon_i^2 & \mathrm{F}_i^2\epsilon_i^2 & \mathrm{F}_i\mathrm{I}_i\epsilon_i^2 & \mathrm{F}_i\mathrm{G}_i\epsilon_i^2 \\ \mathrm{I}_i\epsilon_i^2 & \mathrm{I}_i\mathrm{K}_i\epsilon_i^2 & \mathrm{I}_i\mathrm{S}_i\epsilon_i^2 & \mathrm{I}_i\mathrm{F}_i\epsilon_i^2 & \mathrm{I}_i\epsilon_i^2 & \mathrm{I}_i\mathrm{G}_i\epsilon_i^2 \\ \mathrm{G}_i\epsilon_i^2 & \mathrm{G}_i\mathrm{K}_i\epsilon_i^2 & \mathrm{G}_i\mathrm{S}_i\epsilon_i^2 & \mathrm{G}_i\mathrm{F}_i\epsilon_i^2 & \mathrm{G}_i\mathrm{I}_i\epsilon_i^2 & \mathrm{G}_i^2\epsilon_i^2 \end{pmatrix}. \tag{5.6}$$

The pivot inequality as used throughout this dissertation is then given by

$$n\left[\frac{1}{n}\sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i,\boldsymbol{\theta})\right]\mathbf{B}_n(\boldsymbol{\theta})^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} l_{\boldsymbol{\theta}}(y_i,\boldsymbol{\theta})\right] \leq \chi^2_{6,1-\alpha} \tag{5.7}$$

For the sandwich-based confidence regions, we use the usual inequality

$$n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\mathbf{A}_n(\hat{\boldsymbol{\theta}})^{-1}\mathbf{B}_n(\hat{\boldsymbol{\theta}})\mathbf{A}_n(\hat{\boldsymbol{\theta}})^{-1}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \leq \chi^2_{6,1-\alpha} \tag{5.8}$$

with the $\mathbf{A}_n(\hat{\boldsymbol{\theta}})$ matrix defined as

$$\mathbf{A}_n(\hat{\boldsymbol{\theta}}) = \frac{-1}{2n\sigma^2}\sum_{i=1}^{n}\begin{pmatrix} 1 & \mathrm{K}_i & \mathrm{S}_i & \mathrm{F}_i & \mathrm{I}_i & \mathrm{G}_i \\ \mathrm{K}_i & \mathrm{K}_i^2 & \mathrm{K}_i\mathrm{S}_i & \mathrm{K}_i\mathrm{F}_i & \mathrm{K}_i\mathrm{I}_i & \mathrm{K}_i\mathrm{G}_i \\ \mathrm{S}_i & \mathrm{S}_i\mathrm{K}_i & \mathrm{S}_i^2 & \mathrm{S}_i\mathrm{F}_i & \mathrm{S}_i\mathrm{I}_i & \mathrm{S}_i\mathrm{G}_i \\ \mathrm{F}_i & \mathrm{F}_i\mathrm{K}_i & \mathrm{F}_i\mathrm{S}_i & \mathrm{F}_i^2 & \mathrm{F}_i\mathrm{I}_i & \mathrm{F}_i\mathrm{G}_i \\ \mathrm{I}_i & \mathrm{I}_i\mathrm{K}_i & \mathrm{I}_i\mathrm{S}_i & \mathrm{I}_i\mathrm{F}_i & \mathrm{I}_i^2 & \mathrm{I}_i\mathrm{G}_i \\ \mathrm{G}_i & \mathrm{G}_i\mathrm{K}_i & \mathrm{G}_i\mathrm{S}_i & \mathrm{G}_i\mathrm{F}_i & \mathrm{G}_i\mathrm{I}_i & \mathrm{G}_i^2 \end{pmatrix} \tag{5.9}$$

After cleaning the data to purge non-responses, we had 7804 rows of data remaining from our original set of 8737. This is a reduction of 10.6% of the original data. This is at the upper limit of what is generally acceptible in terms of eliminating data rows without having to worry too much about the nature of the non-responders. We let this be our population.

We then ran a linear regression of BMI on all the listed explanatory variables to get our pseudo-true parameter values. From our population we sampled $10,000$ datasets each with sample sizes ranging from 30 to 100, ran regressions using those samples, and computed confidence region coverage for all covariate coefficients together. The results can be see in Figure 5.1.

The results are similar to what we have seen in our previous simulation studies. The standard simple linear regression confidence regions procedure undercovers the pseudo-true parameter vector, providing just under 93% coverage even at the higher sample sizes. The original sandwich-based confidence region does particularly poorly at first, but improves as sample size increases. The ad hoc methods known as hc1, hc2, and hc3 perform slightly better than the sandwich, again because they start with the sandwich estimate of variance and increase that estimate by different amounts. Our pivot-based confidence regions start out with slightly high coverage, but quickly drop to 95%, and this is the only method that provides 95% coverage within this study. The hc4 and hc5 methods are not usable in a multi-parameter confidence region analysis.

5.2 Dataset as Weighted Sample - Regression

In this section we add survey weights to our analysis. Instead of using the large dataset as our population, we will rightly treat it as a large sample from the larger U.S. population. We will use the weights to run a weighted regression on the large sample in order to estimate the pseudo-true parameter values for the entire U.S. population. Then, we will also use the inverse of the weights, assuming they are proportional to the probability of selection, to take sub-samples from our large sample, to simulate sampling from the entire U.S. population.

The main area of concern here is that the weights we are using are not true design weights, so their inverses are not proportional to the proability of being sampled. The publicly-provided weights from NHANES are computed by taking the design weights and adjusting them further to account for non-response and then poststratified to make the final sample more representative of the overall U.S. population. For these reasons, we should

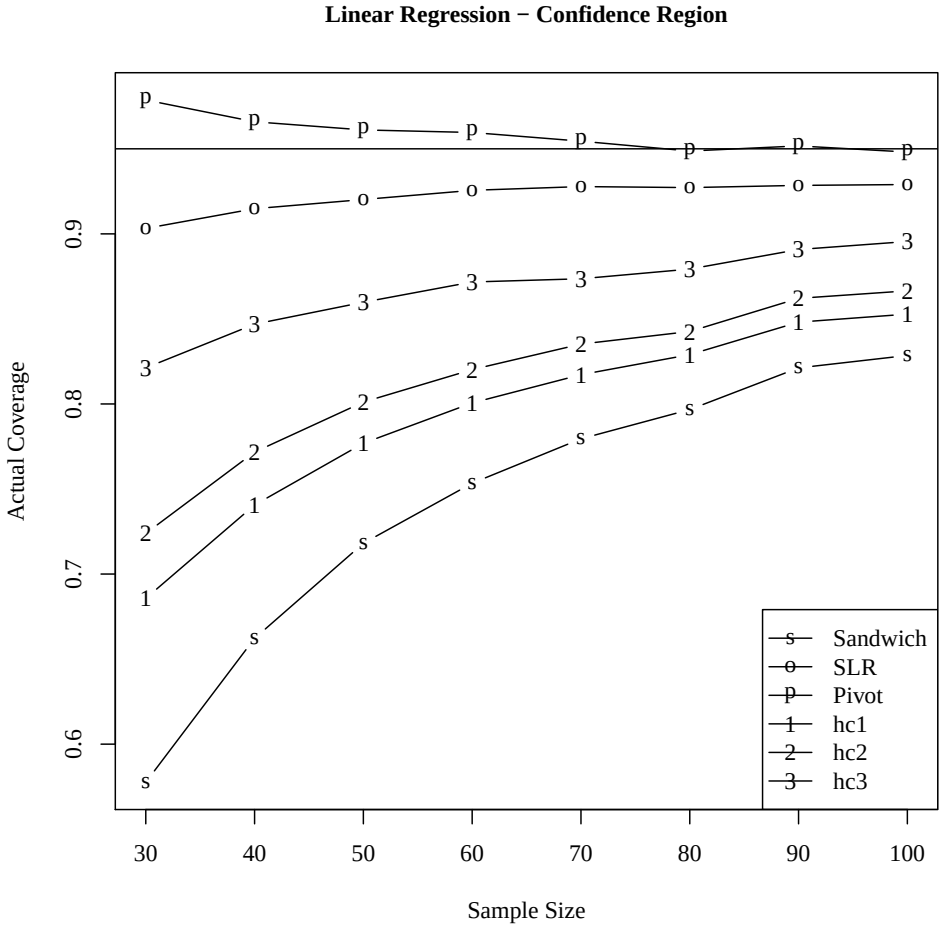

Figure 5.1: Graph showing confidence interval coverage of the pseudo-true parameter vector value in the NHANES data using various methods. Here we use the dataset as the population and perform standard multiple regression analysis. The horizontal line is the 95% line.

ideally use true design weights instead. However, the design weights are not available to the public [Mirel et al., 2013]. Additionally, the publicly-provided weights do use the design weights as the foundation for their computation. In lieu of having the true design weights, using the publicly-provided weights is the best approximation to the design weights we can make.

Sampling using weights does have a benefit: we can use unweighted regression to estimate the pseudo-true parameters. When we sub-sample using the design weights to generate the probability of selection, our sub-sample is automatically representative of the overall population. Since our sub-sample is representative of the entire population, we only need to use unweighted regression techniques to analyze it. This means we can use all the previously developed pivot and sandwich results to perform inference.

The only other issue of concern is that of independence. Our sampling is clearly not independent. But our sub-samples will be so small relative to the large sample, that making the assumption of independence will not cause serious problems. If we were sampling with simple random samples, then we can assume independence if the samples are 10% or less of the overall population. Our sub-samples will be at most 1.2% of the larger sample. We are not simple random sampling, but the largest sampling probability is the same order of magnitude as simple random sampling, so again we can assume independence and be confident that our results will be approximately the same as if we accurately accounted for dependence.

For this example we model BMI as a function of waist measurement, both of which were measured at mobile examination centers. For any given height, BMI is exactly determined by weight, and we are using waistline as a proxy for weight. As in Section 5.1, there are likely other factors that could be included in our model, such as weight itself, but we are interested in performing inference on a misspecified model, so this simple model will suffice for that purpose.

The working model we use for this study is

$$B_i = \theta_0 + \theta_1 \times X_i + \epsilon_i \tag{5.10}$$

$$\epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2). \tag{5.11}$$

where "B_i" reresents the BMI of respondent i and "X_i" represents the corresponding waist measurement. Since we are assuming a simple linear regression model, the pivot and sandwich that we use are exactly the same as in Subsection 3.3.

To generate the pseudo-true parameters for our simulation study, we used weighted regression with the provided weights to get estimates of the pseudo-true U.S. population parameters in this model. This resulted in parameters $\theta_0 = -3.459591$ and $\theta_1 = 0.328860$.

To generate sub-samples from our sample, we used the weights as if they were design weights. That is, if the weight for respondent i is w_i, then the probability of selecting respondent i in our sub-sample is proportional to $\frac{1}{w_i}$. We used the inverses of the provided weights, scaled so they sum to one, as the selection probabilities in our sub-sampling.

For our simulation study, we used sample sizes from $n = 10$ to $n = 100$ and generated $10,000$ samples at each sample size. For each sample, we computed the maximum likelihood estimates of the coefficients. For each sample size, we computed coverage rates from confidence regions created using standard maximum likelihood estimation, the sandwich, the hc1-3 ad hoc adjustments to the sandwich, and the pivot method. Results can be seen in Figure 5.2.

These results match what we have seen in our previous simulation studies. Using ordinary least squares methods to perform inference, we see that coverage of the pseudo-true parameters hovers at about 85% throughout all sample sizes tested. The sandwich-based confidence regions show coverage starting at around 70% at $n = 10$ and increase to about 90% by the end of this study. Confidence regions generated by adjustments to the sandwich show slightly better coverage rates than the sandwich-based regions. The pivot performs best of all. It shows higher than expected coverage in the smallest sample size, but quickly

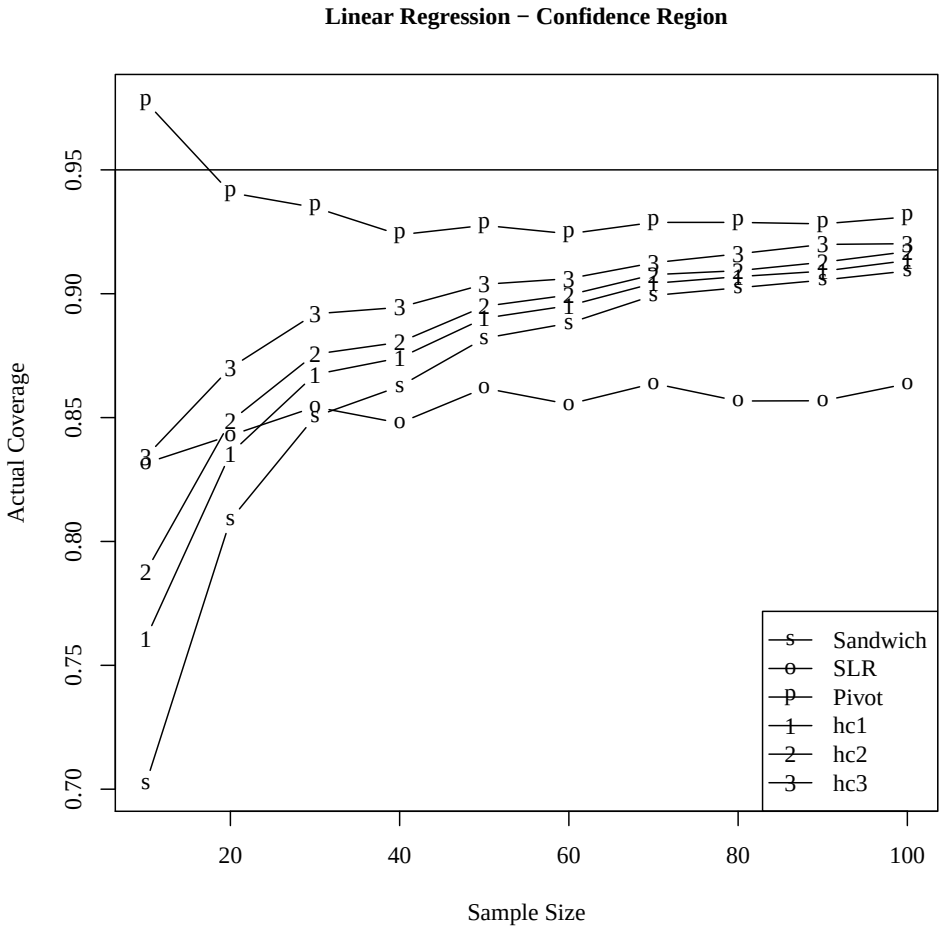

Figure 5.2: Graph showing confidence interval coverage of the pseudo-true parameter vector value in the NHANES data using various methods. Our simple linear regression analysis used examination weights to sub-sample the respondents. The horizontal line is the 95% line.

drops down to about $92 - 93\%$ for the remainder of the simulation. There is some concern that the pivot-based confidence regions appear to provide coverage rates that never rise back to 95%. To verify that the pivot method does not continue to under-cover the pseudo-true parameter, we extended the simulations out to sample sizes up to 800, since beyond that we would need to account for the fact that we are pulling from a finite population. The pivot-based confidence regions showed coverage rates that curved back up to 95% in larger sample sizes and were no worse than 93% in any sample size tested. The results of that extended sample size simulation can be seen in Figure 5.3.

5.3 Dataset as "Truth" - Semiparametric Copula Model

In this section we perform a third simulation experiment with our real world data. The context here is semiparametric copula estimation. This differs from the previous sections in three ways. First, we are only dealing with one parameter instead of multiple parameters. Working in one dimension simplifies some of the computations we require. Second, we are performing a semiparametric analysis. This will necessarily increase the uncertainty of our estimates, and unfortunately complicates the computations required. Third, we are not using a regression model. We are using a parametric copula to model the dependency relationship between two of our data variables while using empirical approximations for the margins.

The data we use are the BMI and waist data from the previous section. As mentioned before, BMI is a function of height and weight. We would expect that there is probably a relationship between waist measurement and BMI. If the relationship is strong enough, perhaps we can replace BMI with waist measurement in some follow-on analysis. There are several benefits. The measuring device, a tape measure, is cheaper, smaller, less expensive and less likely to break down than a scale, and only one measurement needs to be taken.

To check and see if a Gaussian model is reasonable, we plot the data. The plot of data after conversion to empirical quantiles can be seen in Figure 5.4. If the data were Gaussian, it would look like an oval that was denser in the middle than on the edges. What we see in Figure 5.4 can be described as either a broom, a vacuum cleaner, or a rocket ship, but it

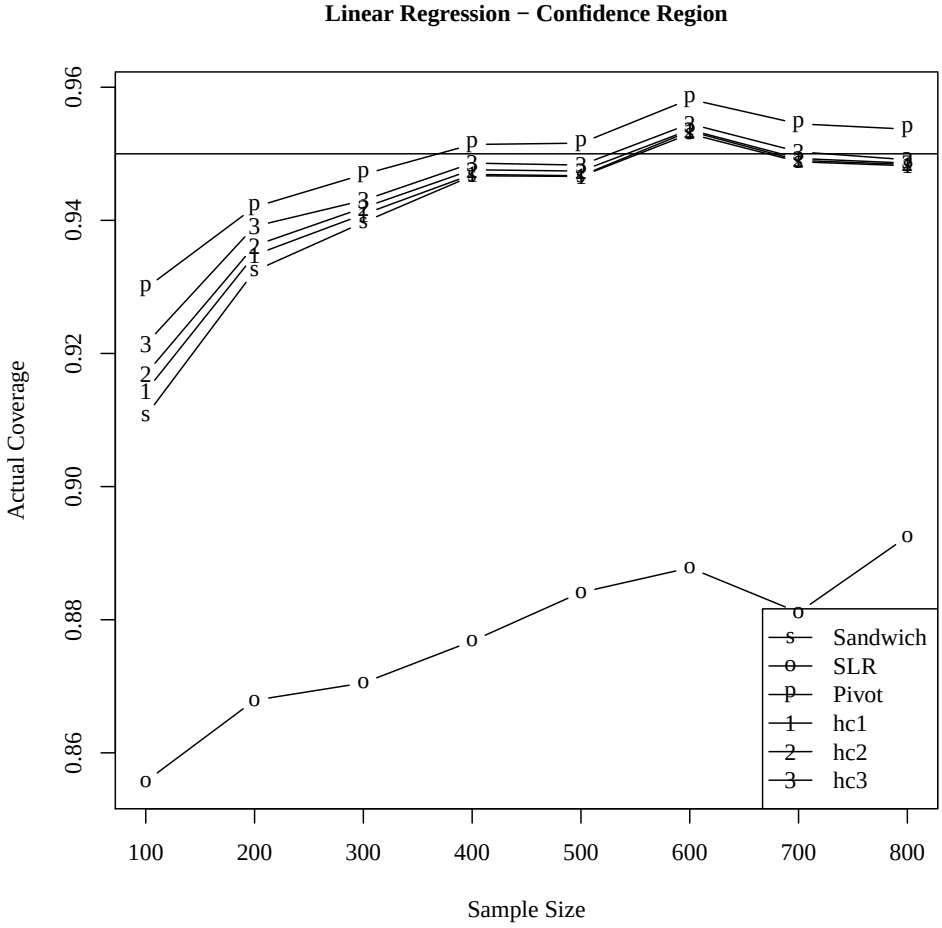

Figure 5.3: Graph showing confidence interval coverage at larger sample sizes of the pseudo-true parameter vector value in the NHANES data using various methods. Our simple linear regression analysis used examination weights to sub-sample the respondents. The horizontal line is the 95% line. The ordinary least squares coverage rate is not shown, as it does not improve upon 70% coverage over all sample sizes tested. Note the difference in scale from Figure 5.2.

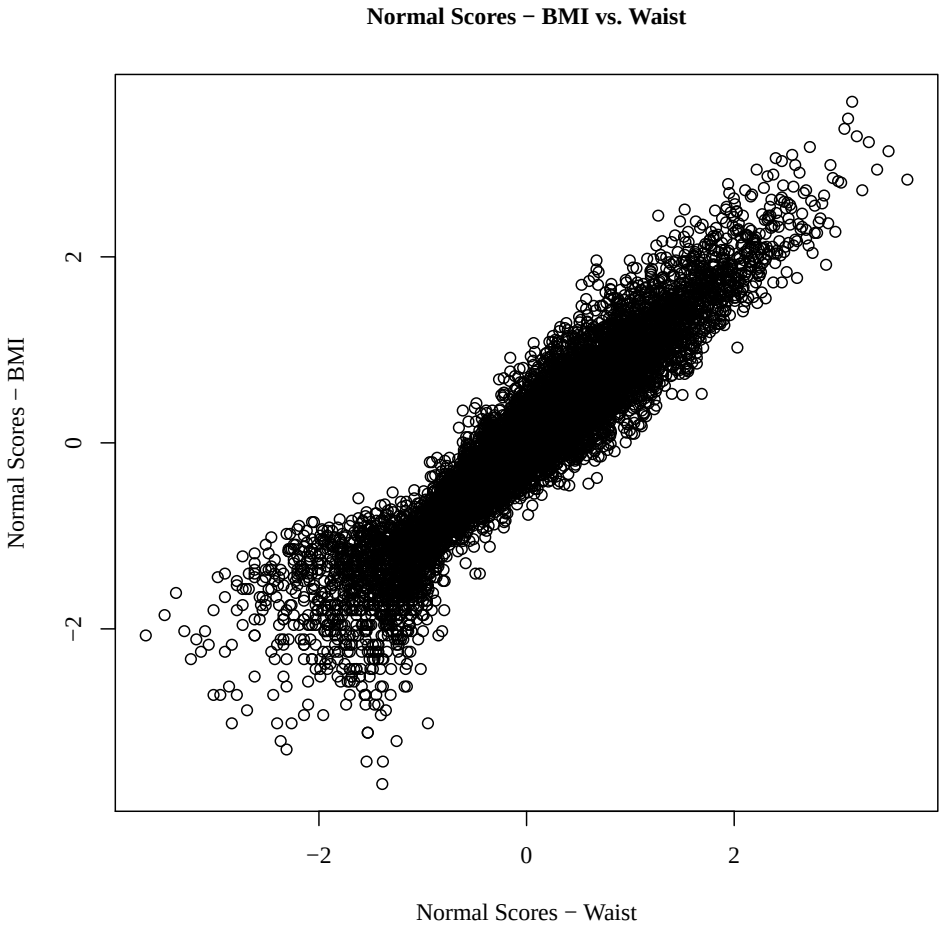

Figure 5.4: Graph showing a plot of the normal scores of the data used in the semiparametric copula model estimation. Data were converted to ranks, then divided by one more than the number of data points. Finally, the standard normal quantiles of those values were computed and then plotted against each other.

is definitly not an oval. For this reason, the data probably do not follow a Gaussian copula model. But we want a misspecified model for our test so it will work well for this example.

The Gaussian semiparametric copula model is the same as the one investigated in Subsection 2.3.4. We start by taking the data $(x_{1,1}, x_{2,1}), \ldots (x_{1,n}, x_{2,n})$, and compute empirical quantiles for each margin separately

$$\hat{u}_{jt} = \frac{1}{n+1} \sum_{s=1}^{n} I(x_{js} \leq x_{jt}), \qquad j = 1, 2. \tag{5.12}$$

Our working model is the Gaussian copula density

$$c_G(u_1, u_2, \theta) = \frac{1}{\sqrt{1-\theta^2}} \exp\left(\frac{-\theta^2(\Phi^{-1}(u_1)^2 + \Phi^{-1}(u_2)^2) + 2\theta\Phi^{-1}(u_1)\Phi^{-1}(u_2)}{2(1-\theta^2)} \right) \tag{5.13}$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function, and θ is the standard correlation coefficient. The derivatives of the log-likelihood that we will need for the pivot and sandwich are

$$\begin{aligned}
\sum_{i=1}^{n} l_\theta(u_{1,i}, u_{2,i}, \theta) &= \frac{\theta}{1-\theta^2} - \frac{\theta}{(1-\theta^2)^2} \sum_{i=1}^{n} [\Phi^{-1}(u_{1,i})^2 + \Phi^{-1}(u_{2,i})^2] \\
&\quad + \frac{1+\theta^2}{(1-\theta^2)^2} \sum_{i=1}^{n} \Phi^{-1}(u_{1,i})\Phi^{-1}(u_{2,i}) \qquad (5.14) \\
\sum_{i=1}^{n} l_{\theta,1}(u_{1,i}, u_{2,i}, \theta) &= \sum_{i=1}^{n} \frac{(1+\theta^2)\Phi^{-1}(u_{2,i}) - 2\theta\Phi^{-1}(u_{1,i})}{(1-\theta^2)^2 \phi(\Phi^{-1}(u_{1,i}))} \qquad (5.15) \\
\sum_{i=1}^{n} l_{\theta,\theta}(u_{1,i}, u_{2,i}, \theta) &= \frac{1+\theta^2}{(1-\theta^2)^2} - \frac{(1+3\theta^2)}{(1-\theta^2)^3} \sum_{i=1}^{n} \left[\Phi^{-1}(u_{1,i})^2 + \Phi^{-1}(u_{2,i})^2\right] \\
&\quad + \frac{2\theta(3+\theta^2)}{(1-\theta^2)^3} \sum_{i=1}^{n} \Phi^{-1}(u_{1,i})\Phi^{-1}(u_{2,i}) \qquad (5.16)
\end{aligned}$$

where $\phi(\cdot)$ is the density function of the standard normal distribution, and as a reminder, is also the derivative of the cumulative distribution function. These formula give us the derivative of the log-likelihood for the pivot, line (5.14), and the $A_n(\theta)$ quantity needed for

the sandwich, line (5.16). We will also use the modified $B_n(\theta)$ estimator as detailed in Chen and Fan [2005] for both the sandwich and pivot confidence intervals.

To describe $B_n(\theta)$, we first look at what the theoretical quantity should be. Then we can determine an approximation for it. The $B(\theta)$ quantity is more complicated than in fully parametric cases. It is defined as

$$B(\theta) = \text{Var}\left[l_\theta(u_1, u_2, \theta) + W_1(u_1, \theta) + W_2(u_2, \theta)\right] \tag{5.17}$$

with the variance also being taken under the true distribution. The W_1 and W_2 terms are included to account for the extra variability that arises from using the empirical distribution, $\hat{u}_{jt}$, $j = 1, 2$; $t = 1, \ldots, n$, on the marginals instead of the actual data, x_{jt}, $j = 1, 2$; $t = 1, \ldots, n$. We define the function W_j for $j = 1, 2$ as

$$W_j(u_{j,t}, \theta) = \text{E}\left[l_{\theta,j}(u_{1,s}, u_{2,s}, \theta)\{I(u_{j,t} \leq u_{j,s}) - u_{j,s}\} \mid u_{j,t}\right] \tag{5.18}$$

where $l_{\theta,j} = \frac{\partial^2 l}{\partial\theta\partial u_j}$ and $I(\cdot)$ is an indicator function.

Defining $B_n(\theta)$ requires using the empirical esimate of line (5.17). We start with $l_\theta(\hat{u}_{1,t}, \hat{u}_{2,t}, \theta)$ where $t = 1, \ldots, n$ and then add in estimates of $W_1(\hat{u}_{1,t}, \theta) + W_2(\hat{u}_{2,t}, \theta)$. We use

$$\hat{W}_1(\hat{u}_{1,t}, \theta) + \hat{W}_2(\hat{u}_{2,t}, \theta) = \frac{1}{n}\sum_{j=t}^{n} l_{\theta,1}\left(\frac{j}{n+1}, \frac{S_j}{n+1}, \theta\right) + \frac{1}{n}\sum_{S_j \geq S_t}^{n} l_{\theta,2}\left(\frac{j}{n+1}, \frac{S_j}{n+1}, \theta\right) \tag{5.19}$$

where S_j is the rank of $u_{2,j}$ among the u_2 data when $u_{1,j}$ is the j^{th} order statistic among the u_1 data. For any given dataset, we can then compute $B_n(\theta)$ as the sample variance of the n values $l_\theta(\hat{u}_{1,t}, \hat{u}_{2,t}, \theta) + \hat{W}_1(\hat{u}_{1,t}, \theta) + \hat{W}_2(\hat{u}_{2,t}, \theta)$, $t = 1, \ldots, n$. The pivot and its approximate distribution are

$$\sqrt{n}B_n(\theta)^{-1/2}\frac{1}{n}\sum_{t=1}^{n} l_\theta(\theta, \hat{u}_{1,t}, \hat{u}_{2,t}) \dot{\sim} N(0, 1). \tag{5.20}$$

To compute the sandwich, we have $B_n(\hat{\theta})$ as the sample variance of $l_\theta(\hat{u}_{1,t}, \hat{u}_{2,t}, \hat{\theta}) + \hat{W}_1(\hat{u}_{1,t}, \hat{\theta}) + \hat{W}_2(\hat{u}_{2,t}, \hat{\theta}),\ t = 1, \ldots, n$. The only difference in this quantity from the previous paragraph is the setting of $\theta = \hat{\theta}$. For the "bread", we have $A_n(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} l_{\theta,\theta}(\hat{u}_{1,i}, \hat{u}_{2,i}, \hat{\theta})$. The sandwich estimate of variance is then computed as $A_n(\hat{\theta})^{-1} B_n(\hat{\theta}) A_n(\hat{\theta})^{-1}$.

From the population data, we generated $10,000$ sample datasets at each sample size tested between 20 and 100. Using the empirical quantiles of the marginals, we then computed the maximum misspecified likelihood estimate of the Gaussian copula parameter. Confidence intervals were generated using both the sandwich and our pivot. Confidence interval coverage of the pseudo-true parameter value, computed using the entire dataset to be $\theta^* = 0.9366$, was then evaluated for both the pivot-based intervals and the sandwich-based intervals. Results can be seen in Figure 5.5.

This is the one surprising set of results so far. Namely, the pivot doesn't seem to be doing any better than the sandwich at estimating the pseudo-true parameter, and neither of them do particularly well. One possible explanation is that the pseudo-true value is so high, 0.9366, that the pivot confidence intervals and sandwich confidence intervals cover at approximately the same rate. An indication of this behavior was seen in the semiparametric example in Subsection 2.3.4. In those simulations, our pivot confidence intervals and sandwich confidence intervals showed coverage rates that were closer to each other at the extreme values of the parameter than at values close to independence, $\theta = 0$. Coverage rates were also worse at higher pseudo-true values than at values close to independence. This trend held true for all sample sizes tested. The most extreme values of the pseudo-true parameter in that example, however, were only $\pm 1/\pi \approx \pm 0.318$. If the coverage gap continues to decrease for larger pseudo-true values and coverage rates continue to worsen for higher pseudo-true values, we might have seen this exact behavior in simulations where the pseudo-true parameters are closer to ± 1, like the one in this section.

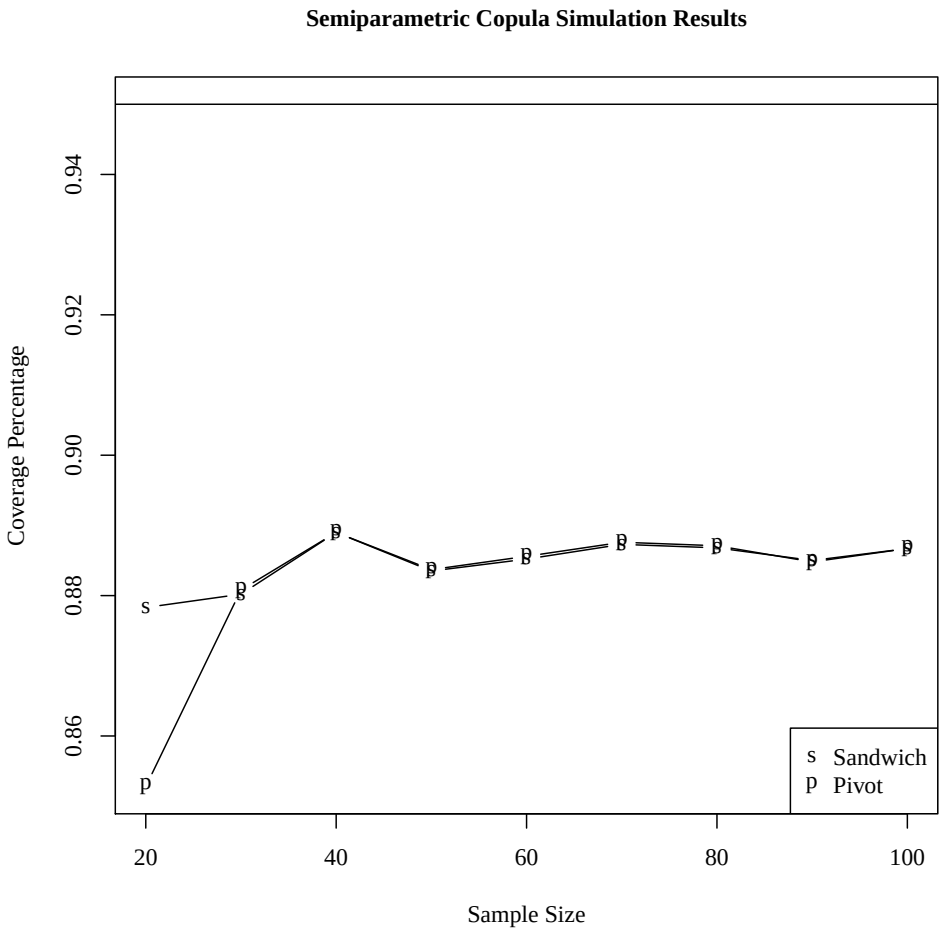

Figure 5.5: Graph showing confidence interval coverage of the pseudo-true parameter value for the semiparametric copula model example.

Chapter 6

CONCLUSION AND DISCUSSION

We have presented a pivot-based method of inference that outperforms the sandwich and its variants in small sample sizes. We further showed that this method incurs no efficiency loss at large sample sizes, when compared to sandwich-based inference. Several simulation examples were presented which strengthen our argument that making fewer approximating assumptions leads to better inference. All results were extended from the univariate case to the multivariate case. Additionally, we explored the use of pivots in a Bayesian context. This resulted in a principle that can be used to select a best class of pivots to use in our suggested pseudo-Bayesian analysis. In a simple simulation example, our pseudo-Bayesian analysis compared favorably to a proper Bayesian analysis when the likelihood had been misspecified. Finally, we used our pivot on a real world dataset and saw that confidence region coverage compared favorably to the coverage of sandwich-based methods.

There are many applications of the sandwich that were not explicitly tested in this dissertation. A more thorough comparison of sandwich variants to our pivot would likely generate several volumes of material. Scenarios with dependent data, time series, general estimating equations, and many other methods could be considered. More analysis could also help determine under what general and specific circumstances the pivot does and does not work well.

A natural extension to the Bayesian work would be to explore further what happens when the likelihood has been misspecified in a Bayesian model. Does pseudo-Bayesian inference with the pivot consistently compare favorably to a proper Bayesian inference using an incorrect likelihood? How sensitive is pseudo-Bayesian analysis to prior selection? In a broader context, what is the penalty for using an incorrect likelihood in a Bayesian analysis?

These questions are larger than can be covered in the current dissertation, but the author anticipates that exploring them will provide ample material for future publication.

An additional issue was noted in the real world copula simulation. There, the sandwich- and pivot-based confidence intervals performed equally well for all but the smallest sample size tested. This was likely due to the pseudo-true parameter lying close to the boundary of the parameter space. According to the standard regularity assumptions, the pseudo-true parameter should be an interior point to the parameter space. It would be interesting to investigate what happens when that interior point is within some small distance, $\epsilon > 0$, of the boundary of the parameter space. Perhaps the different methods of inference perform more equally than at points deeper in the interior of the parameter space.

When performing this sort of work, especially in the multivariate setting described in this dissertation, one should be careful of numerical stability. Frequently, the author had to experiment to find numerically stable codings in order to ensure accurate results. This would be critical in developing software packages to automate the work in this dissertation, or if an analyst wished to write his own code to perform the same work.

Another caveat is that at extremely small sample sizes, and even with careful coding, no method works very well. The author did not simulate any sample sizes less than 5, and even $n = 10$ was considered quite low by members of his dissertation committee. It should be remembered that the inference one makes is dependent on the information in the data about the parameter of interest. Extremely small amounts of data necessarily imply extremely limited information about the parameter of interest.

Appendix A

UNIVARIATE RESULTS

We will show that $B(\theta^*)$ and $A(\theta^*)$ are positive on Θ. We also need to show that $A_n(\theta^*)$ is positive on this same subset, but due to the uniform central limit theorems, it suffices to show that $A(\theta^*)$ is positive on the required subset. Hence, for large enough n, $A_n(\theta^*)$ is also positive on that subset.

Lemma 1. *Under the assumptions of [White, 1982],* $B(\theta^*) > 0$.

Proof. By definition,

$$B(\theta) = E[l_\theta(y;\theta)^2] \tag{A.1}$$

where the expectation is taken with respect to the true density. Since we're taking the expectation of the square of a function, we have that

$$B(\theta) \geq 0 \tag{A.2}$$

for all possible $\theta \in \Theta$.

Assumption A6 in [White, 1982] states that $B(\theta^*)$ is nonsingular. For the one-dimensional case, that means that $B(\theta^*) \neq 0$. Combined with (A.2), we have $B(\theta^*) > 0$. □

Lemma 2. *Under the assumptions of [White, 1982],* $A(\theta^*) > 0$.

Proof. By definition,

$$A(\theta) = -E[l_{\theta\theta}(y;\theta)]. \tag{A.3}$$

The case when this is positive is the same as when the expectation of $l_{\theta\theta}$ is negative. Suppose the contrary, namely that θ^* is such that $A(\theta^*) < 0$, or the expectation of $l_{\theta\theta}$ is positive. Since $A_n(\hat{\theta}) \rightarrow_{a.s.} A(\theta^*)$, for n large enough, we have that $A_n(\hat{\theta}) < 0$. In other words, the second derivative of the log-likelihood is positive, or in other words that the derivative of the score equation evaluated at θ^* is positive. But this would mean that $\hat{\theta}$ is a minimum of the likelihood equation, and not a maximum. Hence, we restrict ourselves to the cases where $A(\theta^*) \geq 0$.

Assumption A6 in [White, 1982] further states that θ^* is a "regular" point of $A(\theta)$. White [1982] then defines a regular point as one in which $A(\theta)$ has constant rank in an open neighborhood of that point. For the one-dimensional case, we have that if $A(\theta) = 0$, then $A(\theta)$ has rank 0. If $A(\theta) \neq 0$, then $A(\theta)$ has rank 1. So, if $A(\theta^*) \geq 0$, then for the constant rank assumption to hold we must have $A(\theta^*) > 0$ or $A(\theta^*)$ is identically 0 in a neighborhood of θ_*. But if the latter condition holds, then $E[l_\theta]$ is constant in a neighborhood of θ^*, and it no longer has a unique minimum at θ^*. Hence, $A(\theta^*) > 0$. □

We now turn our attention to $\theta_{1,n}$ and $\theta_{2,n}$. We will show that $\theta_{1,n} \rightarrow_{a.s.} \theta^*$ and $\theta_{2,n} \rightarrow_{a.s.} \theta^*$.

Lemma 3. *Under the assumptions of [White, 1982],* $\theta_{2,n} \rightarrow_{a.s.} \theta^*$.

Proof. We can solve directly for $\theta_{2,n}$ in terms of other quantities:

$$\theta_{2,n} = \hat{\theta}_n - \frac{A_n(\hat{\theta}_n)^{-1} B_n(\hat{\theta}_n)^{1/2} z_{1-\alpha}}{\sqrt{n}}. \tag{A.4}$$

We know from [White, 1982] that $\hat{\theta}_n \rightarrow_{a.s.} \theta^*$. By Lemma 4 of [Amemiya, 1973], this means that $B_n(\hat{\theta}_n) \rightarrow_{a.s.} B(\theta^*)$ and $A_n(\hat{\theta}_n) \rightarrow_{a.s.} A(\theta^*)$. By Mann-Wald this implies that $B_n(\hat{\theta}_n)^{1/2} \rightarrow_{a.s.} B(\theta^*)^{1/2}$ and $A_n(\hat{\theta}_n)^{-1} \rightarrow_{a.s.} A(\theta^*)^{-1}$. $z_{1-\alpha}$ is a constant. And $\sqrt{n}$ grows arbitrarily large with increasing n. Thus, we have $\theta_{2,n} \rightarrow_{a.s.} \theta^* - 0 = \theta^*$. □

Lemma 4. *For all* $\epsilon > 0$ *there exists* N, A_{max}, A_{min}, B_{max}, *and* B_{min} *such that for all*

$n > N$,

$$A_{min} - \epsilon \leq A_n(\bar{\theta}_n)^{-1} \leq A_{max} + \epsilon \tag{A.5}$$

$$B_{min} - \epsilon \leq B_n(\theta_{1,n})^{1/2} \leq B_{max} + \epsilon. \tag{A.6}$$

Proof. At this point, we remind the reader of some of the standard regularity assumptions generally made in this subject. Specifically, we will assume that the θ-space, Θ, is a compact subset of Euclidean space. We will also rely on Assumption A5 in [White, 1982] to ensure the continuity of $B(\theta)$ and $A(\theta)$. Furthermore, we will be able to apply a uniform law of large numbers to $B_n(\theta)$ and $A_n(\theta)$. A full list of standard regularity assumptions can be found in [White, 1982].

Additionaly, we need to be working in a convex set. Technically, Θ is only a compact set in Euclidean space. But, being closed and bounded, we can enclose Θ within a possibly larger, convex, still compact set Θ_{convex}. We will work within Θ_{convex} to set our bounds. We also need to assume that $B_n(\theta)^{1/2}$, $B(\theta)^{1/2}$, $A_n(\theta)^{-1}$ and $A(\theta)^{-1}$ are continuous functions on Θ_{convex}.

Now, since $B_n(\theta)^{1/2}$, $B(\theta)^{1/2}$, $A_n(\theta)^{-1}$ and $A(\theta)^{-1}$ are continuous functions on $\theta \in \Theta_{convex}$ and Θ_{convex} is compact, those functions each map to compact subsets of $\mathbb{R}$, thus they attain their maximum and minimum on the image sets. That is, there are values $\theta_{min,B^{1/2}}$ and $\theta_{max,B^{1/2}}$ such that $0 < B(\theta_{min,B^{1/2}})^{1/2} \leq B(\theta)^{1/2} \leq B(\theta_{max,B^{1/2}})^{1/2} < \infty$, $\forall \theta \in \Theta_{convex}$. We can define $\theta_{min,A^{-1}}$, $\theta_{max,A^{-1}}$, $\theta_{min,A_n^{-1}}$, and $\theta_{max,A_n^{-1}}$, $\theta_{min,B_n^{1/2}}$, and $\theta_{max,B_n^{1/2}}$ similarly.

We will proceed to show the details for the bound on $A_n(\bar{\theta}_n)^{-1}$ knowing that the argument is similar for $B_n(\theta_{1,n})^{1/2}$. From the convexity of Θ_{convex}, we know that $\bar{\theta}_n = (1-b)\hat{\theta}_n + b\theta_{1,n} \in \Theta_{convex}$ for some $b \in [0,1]$, and hence $A_n((1-b)\hat{\theta}_n + b\theta_{1,n})^{-1} \leq A_n(\theta_{max,A_n^{-1}})^{-1}$. Since we can apply the uniform law of large numbers to $A_n(\theta)^{-1}$, we have that

$$\sup_{\theta \in \Theta_{convex}} |A_n(\theta)^{-1} - A(\theta)^{-1}| \rightarrow_{a.s.} 0. \tag{A.7}$$

What this means is that with probability one, $\forall \epsilon > 0$, $\exists N_\epsilon$ such that $\forall n \geq N_\epsilon$, $\forall \theta \in \Theta_{convex}$, $|A_n(\theta) - A(\theta)| < \epsilon$. So, if we fix an $\epsilon > 0$, there is some N_ϵ such that for all $n \geq N_\epsilon$ we have

$$A_n((1-b)\hat{\theta}_n + b\theta_{1,n})^{-1} \leq A_n(\theta_{max,A_n^{-1}})^{-1} \tag{A.8}$$

$$\leq A(\theta_{max,A_n^{-1}})^{-1} + \epsilon \tag{A.9}$$

$$\leq A(\theta_{max,A^{-1}})^{-1} + \epsilon. \tag{A.10}$$

Line (A.8) comes from the definition of $\theta_{max,A_n^{-1}}$. Line (A.9) is from the uniform law of large numbers. And line (A.10) is from the definition of $\theta_{max,A^{-1}}$.

Now, we can say that for a fixed $\epsilon > 0$, there exists an $N_{\epsilon,B^{1/2}}$ for $B_n(\theta)^{1/2}$ and an $N_{\epsilon,A_n^{-1}}$ for $A(\theta)^{-1}$ such that for $n \geq max(N_{\epsilon,B^{1/2}}, N_{\epsilon,A_n^{-1}})$

$$A_n((1-b)\hat{\theta}_n + b\theta_{1,n})^{-1} \leq A(\theta_{max,A^{-1}})^{-1} + \epsilon \tag{A.11}$$

$$B_n(\theta_{1,n})^{1/2} \leq B(\theta_{max,B^{1/2}})^{1/2} + \epsilon. \tag{A.12}$$

By the same style of argument for n sufficiently large, we have additionally that

$$A_n((1-b)\hat{\theta}_n + b\theta_{1,n})^{-1} \geq A(\theta_{min,A^{-1}})^{-1} - \epsilon \tag{A.13}$$

$$B_n(\theta_{1,n})^{1/2} \geq B(\theta_{min,B^{1/2}})^{1/2} - \epsilon. \tag{A.14}$$

If we let $A_{max} = A(\theta_{max,A^{-1}})^{-1}$, $B_{max} = B(\theta_{max,B^{1/2}})^{1/2}$, $A_{min} = A(\theta_{min,A^{-1}})^{-1}$, $B_{min} = B(\theta_{min,B^{1/2}})^{1/2}$ then the result is proved. □

Lemma 5. $\theta_{1,n} \rightarrow_{a.s.} \theta^*$

Proof. $\theta_{1,n}$ solves the following equation

$$\sqrt{n} B_n(\theta_{1,n})^{-1/2} \frac{1}{n} \sum_{i=1}^{n} l_\theta(y_i; \theta_{1,n}) = z_{1-\alpha}. \tag{A.15}$$

Going back to Jennrich's Lemma, we see that the result still holds replacing θ^* with $\theta_{1,n}$.

Thus, we can investigate

$$\sqrt{n}B_n(\theta_{1,n})^{-1/2}A_n(\bar{\theta})(\hat{\theta}-\theta_{1,n}) = z_{1-\alpha} \tag{A.16}$$

instead.

Since $\bar{\theta}$ lies on the line segment containing $\hat{\theta}$ and $\theta_{1,n}$, we can write $\bar{\theta} = (1-b)\hat{\theta} + b\theta_{1,n}$ for some $b \in [0,1]$. This means our equation can be re-written as

$$\sqrt{n}B_n(\theta_{1,n})^{-1/2}A_n((1-b)\hat{\theta} + b\theta_{1,n})(\hat{\theta}-\theta_{1,n}) = z_{1-\alpha}, \qquad b \in [0,1]. \tag{A.17}$$

Clearly, we cannot solve this equation directly for $\theta_{1,n}$ in the general case. Instead we perform a linear search to find the value of $\theta_{1,n}$ that satisfies Equation (A.17).

We do some minor algebra to the previous equation

$$\hat{\theta}_n - \theta_{1,n} = \frac{z_{1-\alpha}}{\sqrt{n}}B_n(\theta_{1,n})^{1/2}A_n((1-b)\hat{\theta}_n + b\theta_{1,n})^{-1} \tag{A.18}$$

This means that for n sufficiently large, we have

$$-\infty < \frac{z_{1-\alpha}}{\sqrt{n}}(B(\theta_{min,B^{1/2}})^{1/2} - \epsilon)(A(\theta_{min,A^{-1}})^{-1} - \epsilon) \tag{A.19}$$

$$\leq \hat{\theta}_n - \theta_{1,n} \leq \frac{z_{1-\alpha}}{\sqrt{n}}(B(\theta_{max,B^{1/2}})^{1/2} + \epsilon)(A(\theta_{max,A^{-1}})^{-1} + \epsilon) < \infty. \tag{A.20}$$

But the products $z_{1-\alpha}(B(\theta_{min,B^{1/2}})^{1/2} - \epsilon)(A(\theta_{min,A^{-1}})^{-1} - \epsilon)$, and $z_{1-\alpha}(B(\theta_{max,B^{1/2}})^{1/2} + \epsilon)(A(\theta_{max,A^{-1}})^{-1} + \epsilon)$ are constants, so the middle term converges to zero as $n \to \infty$ by the Squeeze Theorem. And since $\hat{\theta}_n \to_{a.s.} \theta^*$, we have that $\theta_{1,n} \to_{a.s.} \theta^*$ as well. □

Appendix B

MULTIVARIATE RESULTS

Lemma 6. *Suppose* $\mathbf{A}_1, \ldots, \mathbf{A}_k, \mathbf{A}$ *are matrices such that* $\|\mathbf{A}_i - \mathbf{A}\| < \epsilon$ *for* $i = 1, \ldots, k$ *and for some* $\epsilon > 0$. *Let* $\omega_1, \ldots, \omega_k$ *be weights such that* $\omega_i \geq 0$ *for* $i = 1, \ldots, k$ *and* $\sum_{i=1}^k \omega_i = 1$. *Then* $\| \sum_{i=1}^k \omega_i \mathbf{A}_i - \mathbf{A}\| < \epsilon$.

Proof. We have

$$\| \sum_{i=1}^k \omega_i \mathbf{A}_i - \mathbf{A}\| = \| \sum_{i=1}^k \omega_i (\mathbf{A}_i - \mathbf{A})\| \tag{B.1}$$

$$\leq \sum_{i=1}^k \|\omega_i (\mathbf{A}_i - \mathbf{A})\| \tag{B.2}$$

$$= \sum_{i=1}^k \omega_i \|(\mathbf{A}_i - \mathbf{A})\| \tag{B.3}$$

$$\leq \sum_{i=1}^k \omega_i \epsilon \tag{B.4}$$

$$= \epsilon \tag{B.5}$$

□

Lemma 7. *For every* $\epsilon > 0$ *there is a ball of radius* δ *around* $\boldsymbol{\theta}^*$, *called* $Ball_\delta(\boldsymbol{\theta}^*)$, *such that if* $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k \in Ball_\delta(\boldsymbol{\theta}^*)$ *and* $\mathbf{A}(\bar{\boldsymbol{\theta}}) = \sum_{i=1}^k \omega_i \mathbf{A}(\boldsymbol{\theta}_i)$ *a convex combination, then* $\|\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\boldsymbol{\theta}^*) - \mathbf{A}(\bar{\boldsymbol{\theta}})\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\bar{\boldsymbol{\theta}})\| < \epsilon$.

Proof. First, we note the constants $A^* = \|\mathbf{A}(\boldsymbol{\theta}^*)\|$ and $B^* = \|\mathbf{B}(\boldsymbol{\theta}^*)\|$.

Next, we use the continuity of $\mathbf{B}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ to see that for all $\epsilon > 0$, there is a $\delta_B > 0$ such that if $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta_B$ then $\|\mathbf{B}(\boldsymbol{\theta}) - \mathbf{B}(\boldsymbol{\theta}^*)\| < \epsilon/(3A^*)$.

We now set $B_{bnd} = \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|<\delta_B} \|\mathbf{B}(\boldsymbol{\theta})\|$ and similarly set $A_{bnd} = \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|<\delta_B} \|\mathbf{A}(\boldsymbol{\theta})\|$.

This implies that for all $\epsilon > 0$, there is a $\delta_{B'} > 0$ such that if $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta_{B'}$ then $\|\mathbf{B}(\boldsymbol{\theta}) - \mathbf{B}(\boldsymbol{\theta}^*)\| < \epsilon/(3A^*A_{bnd})$

Now use the continuity of $\mathbf{A}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ to see that for all $\epsilon > 0$, there is a $\delta_A > 0$ such that if $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta_A$ then $\|\mathbf{A}(\boldsymbol{\theta}) - \mathbf{A}(\boldsymbol{\theta}^*)\| < \epsilon/(3 * \max(A^* \cdot B^*, A_{bnd} \cdot B_{bnd}))$.

Pick $\epsilon > 0$, and let $\delta = \min(\delta_A, \delta_B, \delta_{B'})$. This means that $B_{bnd} \geq \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|<\delta} \|\mathbf{B}(\boldsymbol{\theta})\|$ and likewise $A_{bnd} \geq \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|<\delta} \|\mathbf{A}(\boldsymbol{\theta})\|$.

Consider $Ball_\delta(\boldsymbol{\theta}^*) = \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta\}$. Now select any $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k \in Ball_\delta(\boldsymbol{\theta}^*)$. And let $\mathbf{A}(\bar{\boldsymbol{\theta}}) = \sum_{i=1}^k \omega_i \mathbf{A}(\boldsymbol{\theta}_i)$ a convex combination. Then we have

$$\|\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\boldsymbol{\theta}^*) - \mathbf{A}(\bar{\boldsymbol{\theta}})\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\bar{\boldsymbol{\theta}})\| \tag{B.6}$$

$$= \|\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\boldsymbol{\theta}^*) - \mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\bar{\boldsymbol{\theta}}) + \mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\bar{\boldsymbol{\theta}}) - \mathbf{A}(\bar{\boldsymbol{\theta}})\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\bar{\boldsymbol{\theta}})\| \tag{B.7}$$

$$\leq \|\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\boldsymbol{\theta}^*) - \mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\bar{\boldsymbol{\theta}})\| + \|\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\bar{\boldsymbol{\theta}}) - \mathbf{A}(\bar{\boldsymbol{\theta}})\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\bar{\boldsymbol{\theta}})\| \tag{B.8}$$

$$\leq \|\mathbf{A}(\boldsymbol{\theta}^*)\|\|\mathbf{B}(\boldsymbol{\theta}^*)\|\|\mathbf{A}(\boldsymbol{\theta}^*) - \mathbf{A}(\bar{\boldsymbol{\theta}})\| + \|\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*) - \mathbf{A}(\bar{\boldsymbol{\theta}})\mathbf{B}(\boldsymbol{\theta}_0)\|\|\mathbf{A}(\bar{\boldsymbol{\theta}})\| \tag{B.9}$$

$$\leq A^*B^* \frac{\epsilon}{3 * \max(A^*B^*, A_{bnd}B_{bnd})} \tag{B.10}$$

$$+ \|\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*) - \mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}_0) + \mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}_0) - \mathbf{A}(\bar{\boldsymbol{\theta}})\mathbf{B}(\boldsymbol{\theta}_0)\|\|\mathbf{A}(\bar{\boldsymbol{\theta}})\| \tag{B.11}$$

$$\leq \frac{\epsilon}{3} + \|\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*) - \mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}_0)\|\|\mathbf{A}(\bar{\boldsymbol{\theta}})\| + \|\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}_0) - \mathbf{A}(\bar{\boldsymbol{\theta}})\mathbf{B}(\boldsymbol{\theta}_0)\|\|\mathbf{A}(\bar{\boldsymbol{\theta}})\| \tag{B.12}$$

$$\leq \frac{\epsilon}{3} + \|\mathbf{A}(\boldsymbol{\theta}^*)\|\|\mathbf{B}(\boldsymbol{\theta}^*) - \mathbf{B}(\boldsymbol{\theta}_0)\|\|\mathbf{A}(\bar{\boldsymbol{\theta}})\| + \|\mathbf{A}(\boldsymbol{\theta}^*) - \mathbf{A}(\bar{\boldsymbol{\theta}})\|\|\mathbf{B}(-\theta_0)\|\|\mathbf{A}(\bar{\boldsymbol{\theta}})\| \tag{B.13}$$

$$\leq \frac{\epsilon}{3} + A^* \frac{\epsilon}{3A^*A_{bnd}} A_{bnd} + \frac{\epsilon}{3 * \max(A^*B^*, A_{bnd}B_{bnd})} B_{bnd}A_{bnd} \tag{B.14}$$

$$\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \tag{B.15}$$

$$= \epsilon \tag{B.16}$$

and the proof is done. □

Lemma 8. *With probability one, for all $\epsilon > 0$ there exists a $\delta > 0$ and an N such that for all $n > N$ and all $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ with $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\| < \delta$ for $i = 0, 1, \ldots, k$, $\|\mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\boldsymbol{\theta}_0)\mathbf{A}_n(\bar{\boldsymbol{\theta}}) - \mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\boldsymbol{\theta}^*)\| < \epsilon$ for $\mathbf{A}_n(\bar{\boldsymbol{\theta}})$ defined above.*

Proof. We have by [Jennrich, 1969], Theorem 2, that the almost sure convergence of $\mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\boldsymbol{\theta}_0)\mathbf{A}_n(\bar{\boldsymbol{\theta}})$ to $\mathbf{A}(\bar{\boldsymbol{\theta}})\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\bar{\boldsymbol{\theta}})$ is uniform in $\boldsymbol{\theta}$.

Consequently, pick $\epsilon > 0$. With probability one, there is an N such that for all $n > N$, and for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\|\mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\boldsymbol{\theta}_0)\mathbf{A}_n(\bar{\boldsymbol{\theta}}) - \mathbf{A}(\bar{\boldsymbol{\theta}})\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\bar{\boldsymbol{\theta}})\| < \epsilon/2$ with $\mathbf{A}_n(\bar{\boldsymbol{\theta}})$ as defined above.

For that same $\epsilon > 0$ by Lemma 7 there is a $\delta > 0$ such that for all $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ with $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\| < \delta$ for $i = 0, 1, \ldots, k$, $\|\mathbf{A}(\bar{\boldsymbol{\theta}})\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\bar{\boldsymbol{\theta}}) - \mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\boldsymbol{\theta}^*)\| < \epsilon/2$

Combining the two inequalities, we see that with probability one, for any $\epsilon > 0$ there is both $\delta > 0$ and N such that for all $n > N$ and for all $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ with $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\| < \delta$ for $i = 0, 1, \ldots, k$, we have $\|\mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\boldsymbol{\theta}_0)\mathbf{A}_n(\bar{\boldsymbol{\theta}}) - \mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\boldsymbol{\theta}^*)\| < \epsilon$. □

Lemma 9. *Suppose $\mathbf{A}$ is a $p \times p$ matrix, and $\mathbf{x}$ is a p-dimensional vector. Then $(-1)\|\mathbf{A}\| \cdot p^2 \cdot \|\mathbf{x}\|^2 \leq \mathbf{x}^T\mathbf{A}\mathbf{x} \leq \|\mathbf{A}\| \cdot p^2 \cdot \|\mathbf{x}\|^2$*

Proof. We will prove the right-hand inequality. The left-hand inequality is done similarly.

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i,j} a_{i,j}x_i x_j \tag{B.17}$$

$$\leq |\sum_{i,j} a_{i,j}x_i x_j| \tag{B.18}$$

$$\leq \sum_{i,j} |a_{i,j}x_i x_j| \tag{B.19}$$

$$\leq \sum_{i,j} \max_{i,j}|a_{i,j}| \cdot |x_i x_j| \tag{B.20}$$

$$= \max_{i,j}|a_{i,j}| \left(\sum_i |x_i|\right)^2 \tag{B.21}$$

$$\leq \max_{i,j}|a_{i,j}| \cdot p^2 \cdot \max_i |x_i|^2 \tag{B.22}$$

$$\leq \|\mathbf{A}\| \cdot p^2 \cdot \|\mathbf{x}\|^2 \tag{B.23}$$

□

We require that the eigenvalues of all relevant inverse sandwich matrix variations be finite and positive, and specifically bounded away from zero. First we look at $\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)^{-1}\mathbf{A}(\boldsymbol{\theta}^*)$. The standard regularity assumptions in [White, 1982] state that $\mathbf{A}(\boldsymbol{\theta}^*)$ is full rank and that $\mathbf{B}(\boldsymbol{\theta}^*)$ is invertible. That means that $\mathbf{B}(\boldsymbol{\theta}^*)$ is positive definite. Which implies that the asymptotic inverse sandwich, $\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)^{-1}\mathbf{A}(\boldsymbol{\theta}^*)$, is also positive definite. This means that the asymptotic inverse sandwich has eigenvalues that are all bounded and positive. In particular, there is an eigenvalue with smallest magnitude, which we will call λ^*_{min} and an eigenvalue with largest magnitude, which we will call λ^*_{max}. For completeness, we have $0 < \lambda^*_{min} \leq \lambda^*_{max} < \infty$

Bounding the eigenvalues of $\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)$ is now relatively easy. Since we have $\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) \rightarrow_{a.s.} \mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)^{-1}\mathbf{A}(\boldsymbol{\theta}^*)$, we can bound the eigenvalues of $\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)$ arbitrarily closely to the eigenvalues of $\mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)^{-1}\mathbf{A}(\boldsymbol{\theta}^*)$. For any $\epsilon > 0$ with $\epsilon < \lambda^*_{min}$ we can then bound the range of eigenvalues of $\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)$

to be between $\lambda^*_{min} - \epsilon$ and $\lambda^*_{max} + \epsilon$, for n sufficiently large enough.

To bound the eigenvalues of $\mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\boldsymbol{\theta})^{-1}\mathbf{A}_n(\bar{\boldsymbol{\theta}})$, we need another few steps. These are done with Lemmas 6, 7, and 8, which give us the results we need. Namely that with probability one, for all $\epsilon > 0$ there is both a $\delta > 0$ and N such that for all possible combinations of $\boldsymbol{\theta}$ in the ball of radius δ around $\boldsymbol{\theta}^*$ and for all $n > N$, $\|\mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\boldsymbol{\theta}_0)^{-1}\mathbf{A}_n(\bar{\boldsymbol{\theta}}) - \mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)^{-1}\mathbf{A}(\boldsymbol{\theta}^*)\| < \epsilon$. This means that with a judicious choice of ϵ ($\epsilon = \lambda^*_{min}/2$ for example), we can bound the eigenvalues of $\mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\boldsymbol{\theta}_0)^{-1}\mathbf{A}_n(\bar{\boldsymbol{\theta}})$ to be positive and away from zero using Theorem 5.3 in [Zhan, 2013].

Lemma 10. *For any sequence of solutions, $\{\breve{\boldsymbol{\theta}}_{n,PV}\} = \{\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,PV}\}$, to the pivot equality, $\|\tilde{\boldsymbol{\theta}}_{n,PV}\| = O(1/\sqrt{n})$ with probability one.*

Proof. Select ϵ such that $0 < \epsilon < \lambda^*_{min}/2$. With probability one, there is a $\delta > 0$ and an N_ϵ such that for all combinations of $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ with $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\| < \delta,\ i = 0, 1, \ldots, k$ and all $n > N$ we have $\|\mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\boldsymbol{\theta}_0)^{-1}\mathbf{A}_n(\bar{\boldsymbol{\theta}}) - \mathbf{A}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)^{-1}\mathbf{A}(\boldsymbol{\theta}^*)\| < \epsilon < \lambda^*_{min}/2$. This implies that the eigenvalues $\lambda_1, \ldots, \lambda_p$ of $\mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\boldsymbol{\theta}_0)^{-1}\mathbf{A}_n(\bar{\boldsymbol{\theta}})$ satisfy $0 < \lambda^*_{min}/2 < \lambda_j < \lambda^*_{max} + \lambda^*_{min}/2 < \infty,\ j = 1, \ldots, p$. Let us simplify this and say that there are bounds $0 < \lambda_{lo}$ and $\lambda_{hi} < \infty$ such that $0 < \lambda_{lo} < \lambda_j < \lambda_{hi} < \infty,\ j = 1, \ldots, p$.

By the almost sure convergence of $\hat{\boldsymbol{\theta}}_n$ to $\boldsymbol{\theta}^*$, we have with probability one that there exists an N_δ such that for all $n > N_\delta$, $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| < \delta/2$, where δ is defined in the previous paragraph. This means that with probability one and for $n > N_\delta$, the $\delta/2$ ball around $\hat{\boldsymbol{\theta}}_n$ is contained within the δ ball around $\boldsymbol{\theta}^*$. We can take $N_{max} = \max(N_\epsilon, N_\delta)$ so that all of the results in this and the previous paragraph hold.

Let $n > N_{max}$, and select a point on the boundary of the $\delta/2$ ball around $\hat{\boldsymbol{\theta}}_n$. Let's refer to this point as $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n$ because we want to emphasize the fact that $\|\tilde{\boldsymbol{\theta}}_n\| = \delta/2$. We can plug this point in for $\boldsymbol{\theta}$ in the pivotal quadratic form. Recall we also use it and $\hat{\boldsymbol{\theta}}_n$ and as the points to define $\mathbf{A}_n(\bar{\boldsymbol{\theta}})$. Using results on the bounds of quadratic forms of symmetric

positive definite matrices we have

$$\lambda_{lo}\|\tilde{\boldsymbol{\theta}}_n\|^2 \leq (\tilde{\boldsymbol{\theta}}_n)^T \mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\boldsymbol{\theta}_0)^{-1}\mathbf{A}_n(\bar{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_n) \leq \lambda_{hi}\|\tilde{\boldsymbol{\theta}}_n\|^2 \tag{B.24}$$

$$\lambda_{lo}\frac{\delta^2}{4} \leq (\tilde{\boldsymbol{\theta}}_n)^T \mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\boldsymbol{\theta}_0)^{-1}\mathbf{A}_n(\bar{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_n) \leq \lambda_{hi}\frac{\delta^2}{4}. \tag{B.25}$$

which is true regardless of the direction of $\tilde{\boldsymbol{\theta}}_n$, hence it is uniformly true in all directions away from $\hat{\boldsymbol{\theta}}_n$.

We now compare (B.25) with the main pivot inequality in (3.13). As n increases, we see that the lower bound in (B.25) is a constant. However as n increases in (3.13), it will eventually be the case that $\frac{\chi^2_{p,1-\alpha}}{n} < \lambda_{lo}\frac{\delta^2}{4}$, namely when $n > \frac{4\chi^2_{p,1-\alpha}}{\lambda_{lo}\delta^2}$. Since $\hat{\boldsymbol{\theta}}_n$ makes the pivotal inequality equal to zero, and every point on the $\delta/2$ ball makes the quadratic form evaluate to more than $\chi^2_{p,1-\alpha}/n$, we can invoke the intermediate value theorem to conclude that all solutions to the pivotal inequality will lie within the $\delta/2$ ball around $\hat{\boldsymbol{\theta}}_n$, and that they do exist. We can now take $N = \max(\frac{4\chi^2_{p,1-\alpha}}{\lambda_{lo}\delta^2}, N_{max})$.

Define N and δ as we have done. We consider a solution $\breve{\boldsymbol{\theta}}_{n,PV} = \hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_{n,PV}$ to the pivotal equality in an arbitrary direction from $\hat{\boldsymbol{\theta}}_n$, and specifically look at the equality

$$(\tilde{\boldsymbol{\theta}}_{n,PV})^T \mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\breve{\boldsymbol{\theta}}_{n,PV})^{-1}\mathbf{A}_n(\bar{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_{n,PV}) = \frac{\chi^2_{p,1-\alpha}}{n}. \tag{B.26}$$

Given the eigenvalue-based bounds above and using knowledge of quadratic forms, we can conclude that for $n > N$

$$\lambda_{lo}\|\tilde{\boldsymbol{\theta}}_{n,PV}\|^2 \leq (\tilde{\boldsymbol{\theta}}_{n,PV})^T \mathbf{A}_n(\bar{\boldsymbol{\theta}})\mathbf{B}_n(\breve{\boldsymbol{\theta}}_{n,PV})^{-1}\mathbf{A}_n(\bar{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_{n,PV}) \leq \lambda_{hi}\|\tilde{\boldsymbol{\theta}}_{n,PV}\|^2 \tag{B.27}$$

$$\lambda_{lo}\|\tilde{\boldsymbol{\theta}}_{n,PV}\|^2 \leq \frac{\chi^2_{p,1-\alpha}}{n} \leq \lambda_{hi}\|\tilde{\boldsymbol{\theta}}_{n,PV}\|^2 \tag{B.28}$$

and solving for $\|\tilde{\boldsymbol{\theta}}_{n,PV}\|$ we have

$$\sqrt{\frac{\chi^2_{p,1-\alpha}}{n\lambda_{hi}}} \leq \|\tilde{\boldsymbol{\theta}}_{n,PV}\| \leq \sqrt{\frac{\chi^2_{p,1-\alpha}}{n\lambda_{lo}}} \tag{B.29}$$

which gives us the result that $\|\tilde{\boldsymbol{\theta}}_{n,PV}\| = O(1/\sqrt{n})$ with probability one. □

Appendix C

CODE

The following R code will generate simulations like the one in this disseration. This code will provide results for the pivot, the sandwich, all five ad hocs, and standard regression confidence intervals.

```
n<-10*c(1:10) ##c(5, 10, 15, 20, 25, 30, 1000)
num.runs<-10000
theta<-3
theta.star<-3
coverage<-matrix(0, nrow=length(n), ncol=8)
sigma.star<- 1+sqrt(2/pi) ## based on data-generating model
ci.len<-matrix(0, nrow=length(n), ncol=8)

ci.piv.keep<-array(data=0, dim=c(length(n), num.runs, 2))

## introduce heteroscedasticity

set.seed(343)
for(i in 1:length(n))
{
  for(j in 1:num.runs)
  {
    x.dat<-rnorm(n[i], mean=0, sd=1)
    ## x.dat<-runif(n[i], min=-1*sqrt(3), max=sqrt(3)) ## see if it
```

```
dat<-rnorm(n[i], mean= theta*x.dat, sd=sqrt(1+abs(x.dat)))
theta.hat<-sum(dat*x.dat)/sum(x.dat^2)

## asymptotic sandwich variance...needs to be divided by n
var.sandwich<-mean((dat*x.dat - theta.hat*x.dat^2)^2) /
  (mean(x.dat^2))^2
ci.sand<-theta.hat + sqrt(var.sandwich/n[i]) *
  qnorm(c(0.025, 0.975))

pivot.lb<-function(star)
{
  abs(
    ((mean(x.dat^2))/sigma.star)*sqrt(n[i])*(theta.hat - star) -
      qnorm(0.025, mean=0, sd=sqrt(mean((x.dat*dat -
        star*x.dat^2)^2)/sigma.star^2))
    )
}

pivot.ub<-function(star)
{
  abs(
    ((mean(x.dat^2))/sigma.star)*sqrt(n[i])*(theta.hat - star) -
      qnorm(0.975, mean=0, sd=sqrt(mean((x.dat*dat -
        star*x.dat^2)^2)/sigma.star^2))
    )
}
```

```
ci.temp<-c(optimize(pivot.lb, interval=c(-12, 18))$minimum,
  optimize(pivot.ub, interval=c(-12, 18))$minimum)
ci.piv<-c(min(ci.temp), max(ci.temp))

ci.piv.keep[i,j,]<-ci.piv

## ad hoc-ery
## need 5 different ad hocs
xtxi<-1/sum(x.dat^2)
u.hat<-dat - x.dat * theta.hat
omega.hat<- diag(as.vector(u.hat)^2)
hat.mat<-xtxi * tcrossprod(x.dat)

hc1<-var.sandwich/(n[i] - 1)

denom<- 1 - diag(hat.mat)
omega.tilde<-diag(as.vector(u.hat^2 / denom))

hc2<-xtxi^2 * t(x.dat) %*% omega.tilde %*% x.dat

u.star<-u.hat / denom
omega.star<-diag(as.vector(u.star^2))

hc3.a<-(n[i] - 1)/n[i] * xtxi^2 * t(x.dat) %*% omega.star %*% x.dat
hc3.b<-(n[i] - 1)/n[i]^2 * xtxi^2 * t(x.dat) %*%
  tcrossprod(as.vector(u.star)) %*%  x.dat
```

```
hc3<-hc3.a - hc3.b ## efron: (1-hat)^2

## hc4 work
a.i<-xtxi * x.dat
a.i2<-a.i^2

temp.mat<-tcrossprod(as.vector(a.i2))
var.vsandu1<-sum(diag(temp.mat))  ## trace, sum of 4th powers
temp.mat2<-temp.mat - diag(diag(temp.mat))
hat.denom<-sqrt(tcrossprod(1-hat.mat))
hat.tilde<-hat.mat / hat.denom
var.vsandu2<-sum(temp.mat2 * hat.tilde^2)/2
var.vsandu<-var.vsandu1 + var.vsandu2

find.p<-function(p.til)
{
  z.p.til<-qnorm(p.til)
  temp<-p.til - dnorm(z.p.til)*var.vsandu*(z.p.til^3 +
    z.p.til)/4
  abs(0.975 - temp)
}
p.new<-optimize(find.p, interval=c(0.975, 0.99999999))$minimum
z.new<-qnorm(p.new)  ## use this with sandwich to compute CI

## hc5 work
b<-0.75 ## bound in the H matrix
big.u<-x.dat * u.hat
```

```
H<-numeric(n[i])
psi<-numeric(n[i])
HUUH<-0
omega<-numeric(n[i])
UU<-numeric(n[i])
for(k in 1:n[i])
{
  UU[k]<-big.u[k]^2
  omega[i]<-x.dat[k]^2
  H[k]<-1/sqrt(1 - min(b,omega[k] * xtxi))
  psi[k]<-H[k]^2 * UU[k]
  HUUH<-HUUH + psi[k]
}
Va<-xtxi^2 * HUUH ## use this in the CI computation

psi.big<-matrix(0, nrow=n[i], ncol=n[i])
o.hat<-matrix(0, nrow=n[i], ncol=1)
F.mat<-matrix(0, nrow=1, ncol=n[i]) ## t(F) in paper
G.mat<-matrix(0, nrow=n[i], ncol=n[i])
M.mat0<-matrix(0, nrow=n[i], ncol=n[i])

for(k in 1:n[i])
{
  o.hat[k,1]<-omega[k]
  F.mat[1,k]<-1
  M.mat0[k,k]<-H[k]^2 * xtxi^2 * 1 ## that last is "c"
  psi.big[k,k]<-psi[k]
```

```
}
G.mat<-diag(rep(1,n[i])) - o.hat %*% F.mat * xtxi
B.one0<-t(G.mat) %*% M.mat0 %*% G.mat
d.mat0<-psi.big %*% B.one0
d.hat0<-sum(diag(d.mat0))^2 / sum(diag(d.mat0 %*% d.mat0))

ci.hc1<-theta.hat + qnorm(c(0.025, 0.975))*sqrt(hc1)
ci.hc2<-theta.hat + qnorm(c(0.025, 0.975))*sqrt(hc2)
ci.hc3<-theta.hat + qnorm(c(0.025, 0.975))*sqrt(hc3)
ci.hc4<-theta.hat + c(-1,1)*z.new*sqrt(hc2)
ci.hc5<-theta.hat + c(-1,1)*qt(0.975,d.hat0)*sqrt(Va)

## end ad hoc-ery

## actual stat100 standard errors...don't need to be divided by n
var.stat100<-1/(n[i]-1) * sum((dat - theta.hat*x.dat)^2) /
    sum(x.dat^2)
ci.stat100<-theta.hat + sqrt(var.stat100) * qt(c(0.025, 0.975),
    df=n[i]-1)

ci.len[i,1]<-ci.len[i,1] + (ci.sand[2]-ci.sand[1])/num.runs
ci.len[i,2]<-ci.len[i,2] + (ci.piv[2]-ci.piv[1])/num.runs
ci.len[i,3]<-ci.len[i,3] + (ci.stat100[2]-ci.stat100[1])/num.runs
ci.len[i,4]<-ci.len[i,4] + (ci.hc1[2] - ci.hc1[1])/num.runs
ci.len[i,5]<-ci.len[i,5] + (ci.hc2[2] - ci.hc2[1])/num.runs
ci.len[i,6]<-ci.len[i,6] + (ci.hc3[2] - ci.hc3[1])/num.runs
ci.len[i,7]<-ci.len[i,7] + (ci.hc4[2] - ci.hc4[1])/num.runs
```

```
    ci.len[i,8]<-ci.len[i,8] + (ci.hc5[2] - ci.hc5[1])/num.runs

    if(theta.star>ci.sand[1] && theta.star<ci.sand[2])
        {coverage[i,1]<-coverage[i,1]+1/num.runs}
    if(theta.star>ci.piv[1] && theta.star<ci.piv[2])
        {coverage[i,2]<-coverage[i,2]+1/num.runs}
    if(theta.star>ci.stat100[1] && theta.star<ci.stat100[2])
        {coverage[i,3]<-coverage[i,3]+1/num.runs}
    if(theta.star>ci.hc1[1] && theta.star<ci.hc1[2])
        {coverage[i,4]<-coverage[i,4]+1/num.runs}
    if(theta.star>ci.hc2[1] && theta.star<ci.hc2[2])
        {coverage[i,5]<-coverage[i,5]+1/num.runs}
    if(theta.star>ci.hc3[1] && theta.star<ci.hc3[2])
        {coverage[i,6]<-coverage[i,6]+1/num.runs}
    if(theta.star>ci.hc4[1] && theta.star<ci.hc4[2])
        {coverage[i,7]<-coverage[i,7]+1/num.runs}
    if(theta.star>ci.hc5[1] && theta.star<ci.hc5[2])
        {coverage[i,8]<-coverage[i,8]+1/num.runs}
    if(j %% 100 == 0){print(paste("Size ", n[i], " run ", j))}
  } ## j = number of runs
} ## i = n = sample size
```

The following code will generate MCMC runs like the one I used in the Bayesian sections of this disseration.

```
theta<-3
samp.size<-c(10, 1000)
tune1<-c(2, 0.03)
tune2<-c(2, 0.03)
```

```
num.runs<-1000

to.keep<-1000
thinby<-10
num.runs.MCMC<-to.keep*thinby
burnin<-1000

indices<-c(1:to.keep)*thinby

theta1.MCMC<-numeric(num.runs.MCMC)
theta2.MCMC<-numeric(num.runs.MCMC)

theta1.MCMC.keep<-matrix(0, nrow=num.runs.MCMC, ncol=2)
theta2.MCMC.keep<-matrix(0, nrow=num.runs.MCMC, ncol=2)

theta1.ci.array<-array(0, dim=c(length(samp.size), num.runs, 2) )
theta2.ci.array<-array(0, dim=c(length(samp.size), num.runs, 2) )

acc1<-0
acc2<-0

ptm<-proc.time()
set.seed(343)

for(i in 1:length(samp.size))
{
```

```
for(j in 1:num.runs)
{
  x.dat<-rpois(samp.size[i], theta)

  x.piv<-function(THETA)
  {
    (mean(x.dat) - THETA)/sqrt(THETA/samp.size[i])
  }

  x.logpiv<-function(THETA)
  {
    exp(x.piv(THETA))
  }

  ## start MCMC process
  theta.0<-mean(x.dat)
  ## burnin, pivot 1
  for(k in 1:burnin)
  {
    theta.s<-rgamma(1, shape=theta.0^2/tune1[i],
      scale=tune1[i]/theta.0)
    r<-dnorm(x.piv(theta.s))/dnorm(x.piv(theta.0)) *
      dgamma(theta.0, shape=theta.s^2/tune1[i],
        scale=tune1[i]/theta.s) /
      dgamma(theta.s, shape=theta.0^2/tune1[i],
        scale=tune1[i]/theta.0) *
      dgamma(theta.s, shape=1, scale=4) /
```

```
        dgamma(theta.0, shape=1, scale=4)
    u<-runif(1)
    if(u<r)
    {
      theta.0<-theta.s
    }
  }
  ## MCMC pivot 1
  for(k in 2:num.runs.MCMC)
  {
    theta.s<-rgamma(1, shape=theta.0^2/tune1[i],
      scale=tune1[i]/theta.0)
    r<-dnorm(x.piv(theta.s))/dnorm(x.piv(theta.0)) *
      dgamma(theta.0, shape=theta.s^2/tune1[i],
          scale=tune1[i]/theta.s) /
      dgamma(theta.s, shape=theta.0^2/tune1[i],
          scale=tune1[i]/theta.0) *
      dgamma(theta.s, shape=1, scale=4) /
          dgamma(theta.0, shape=1, scale=4)

    u<-runif(1)
    if(u<r)
    {
      theta1.MCMC[k]<-theta.s
      theta.0<-theta.s
      acc1<-acc1+1/num.runs.MCMC
    } else
```

```
  {
    theta1.MCMC[k]<-theta.0
  }
}

theta.0<-mean(x.dat)
## burnin pivot 2
for(k in 1:burnin)
{
  theta.s<-rgamma(1, shape=theta.0^2/tune2[i],
    scale=tune2[i]/theta.0)
  r<-dlnorm(x.logpiv(theta.s))/dlnorm(x.logpiv(theta.0)) *
    dgamma(theta.0, shape=theta.s^2/tune2[i],
        scale=tune2[i]/theta.s) /
    dgamma(theta.s, shape=theta.0^2/tune2[i],
        scale=tune2[i]/theta.0) *
    dgamma(theta.s, shape=1, scale=4) /
        dgamma(theta.0, shape=1, scale=4)

  u<-runif(1)
  if(u<r)
  {
    theta.0<-theta.s
  }
}
## MCMC pivot 2
for(k in 2:num.runs.MCMC)
```

```
{
  theta.s<-rgamma(1, shape=theta.0^2/tune2[i],
    scale=tune2[i]/theta.0)
  r<-dlnorm(x.logpiv(theta.s))/dlnorm(x.logpiv(theta.0)) *
    dgamma(theta.0, shape=theta.s^2/tune2[i],
        scale=tune2[i]/theta.s) /
    dgamma(theta.s, shape=theta.0^2/tune2[i],
        scale=tune2[i]/theta.0) *
    dgamma(theta.s, shape=1, scale=4) /
        dgamma(theta.0, shape=1, scale=4)

  u<-runif(1)
  if(u<r)
  {
    theta2.MCMC[k]<-theta.s
    theta.0<-theta.s
    acc2<-acc2+1/num.runs.MCMC
  } else
  {
    theta2.MCMC[k]<-theta.0
  }
}

theta1.thin<-theta1.MCMC[indices]
theta2.thin<-theta2.MCMC[indices]

theta1.ci<-c(theta1.thin[which(25==rank(theta1.thin,
```

```
      ties.method="first"))],
      theta1.thin[which(975==rank(theta1.thin,
        ties.method="first"))])
    theta2.ci<-c(theta2.thin[which(25==rank(theta2.thin,
      ties.method="first"))],
        theta2.thin[which(975==rank(theta2.thin,
          ties.method="first"))])

    theta1.ci.array[i,j,]<-theta1.ci
    theta2.ci.array[i,j,]<-theta2.ci
    if(0 == j %% 10)
    {
      print(paste("size ", samp.size[i], " run ", j))
    }
  } ## j, number of runs

  theta1.MCMC.keep[,i]<-theta1.MCMC
  theta2.MCMC.keep[,i]<-theta2.MCMC
} ## i, sample sizes
```

Appendix D

ASSUMPTIONS

The following assumptions are those of [White, 1982], and are fairly standard when dealing with misspecified models. They are reproduced here using [White, 1982]'s original notation.

Assumption 1. *We assume that the data vectors, U_t, are independent and randomly distributed according to a common joint distribution function $G(u)$ on Ω, a measurable Euclidean space with measure ν. We let $g(u)$ be the Radon-Nikodym density, $g(u) = dG(u)/d\nu$.*

Assumption 2. *We will work with a family of distributions $F(u, \theta)$ which have Radon-Nikodym densities $f(u, \theta) = dF(u, \theta)/d\nu$. We assume $\theta \in \Theta$, a compact subset of a p-dimensional Euclidean space. We also assume that $f(u, \theta)$ is measurable in u for every $\theta \in \Theta$ and continuous in θ for every $u \in \Omega$.*

Assumption 3. *(a) $E(\log g(U_t))$ exists and $\mid \log f(u, \theta) \mid \leq m(u)\ \forall \theta \in \Theta$, where m is integrable with respect to G; (b) $I(g : f, \theta)$, the Kullback-Leibler Information Criterion, has a unique minimum at $\theta_* \in \Theta$.*

Assumption 4. *$\partial \log f(u, \theta)/\partial \theta_i$, $i = 1, \ldots, p$ are measurable functions of u for each $\theta \in \Theta$ and continuously differentiable functions of θ for each $u \in \Omega$.*

Assumption 5. *$\mid \partial^2 \log f(u, \theta)/\partial \theta_i \partial \theta_j \mid$ and $\mid \partial \log f(u, \theta)/\partial \theta_i \cdot \partial \log f(u, \theta)/\partial \theta_j \mid$ $i, j = 1, \ldots, p$ are dominated by functions integrable with respect to G for all $u \in \Omega$ and $\theta \in \Theta$.*

Assumption 6. *(a) θ_* is interior to Θ; (b) $\mathbf{B}(\theta_*)$ is nonsingular; (c) θ_* is a regular point of $\mathbf{A}(\theta)$ (i.e., $\mathbf{A}(\theta)$ has constant rank in a neighborhood of θ_*.)*

Assumption 7. $\mid \partial[\partial f(u,\theta)/\partial\theta_i \cdot f(u,\theta)]/\partial\theta_j \mid$ *for* $i,j = 1,\ldots,p$ *are dominated by functions integrable with respect to* ν $\forall\theta \in \Theta$*, and the minimal support of* $f(u,\theta)$ *does not depend on* θ*.*

Assumption 8. $\partial d_l(u,\theta)/\partial\theta_k$ $l = 1,\ldots,q$*,* $k = 1,\ldots,p$ *exist and are continuous functions of* θ *for each* u*.*

Assumption 9. $\mid d_l(u,\theta)d_m(u,\theta) \mid$*,* $\mid \partial d_l(u,\theta)/\partial\theta_k \mid$*, and* $\mid d_l(u,\theta) \cdot \partial \log f(u,\theta)/\partial\theta_k \mid$*,* $k = 1,\ldots,p$*,* $l,m = 1,\ldots,q$ *are dominated by functions integrable with respect to* G *for all* u *and* $\theta \in \Theta$*.*

Assumption 10. $\mathbf{V}(\theta_*)$ *is nonsingular.*

Assumption 11. h *satisfies assummptions A2-A6, and if* $g(u) = f(u,\theta_0)$ *for any* $\theta_0' = (\beta_0', \psi_0') \in \Theta$*, then* $\gamma_*' = (\beta_0', \alpha_*')$ *for* $\gamma_* \in \Gamma$*.*

Assumption 12. $\mathbf{S}(\theta_*, \gamma_*)$ *is nonsingular.*

BIBLIOGRAPHY

H. Akaike. Information theory and an extension of the maximum likelihood principle. *IEEE International Symposium on Information Theory, Proceedings*, pages 267–281, 1973.

Takeshi Amemiya. Regression analysis when the dependent variable is truncated normal. *Econometrica*, 41(6):997–1016, 1973.

S.F. Arnold. *Mathematical Statistics*. Prentice-Hall, 1990. ISBN 9780135610510. URL `https://books.google.com/books?id=EBgnAQAAIAAJ`.

F Bertolino and W Racugno. Analysis of the linear correlation coefficient using pseudo-likelihoods. *Journal of the Italian Statistical Society*, 1(1):33–50, 1992.

F Bertolino and W Racugno. Robust bayesian analysis in analysis of variance and the $\chi\hat{}2$-test by using marginal likelihoods. *The Statistician*, pages 191–201, 1994.

Subha Bhuchongkul. A class of nonparametric tests for independence in bivariate populations. *The Annals of Mathematical Statistics*, pages 138–149, 1964.

Norman Breslow. Tests of hypotheses in overdispersed poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association*, 85(410):565–571, 1990.

George Casella and Roger L Berger. *Statistical inference*, volume 70. Duxbury Press Belmont, CA, 1990.

CDC. About adult bmi. `http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi`. Accessed: 2015-07-05.

Xiaohong Chen and Yanqin Fan. Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection. *Canadian Journal of Statistics*, 33(3):389–414, 2005.

Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of longitudinal data, 2nd ed.* Oxford University Press, 2009.

B Efron. Discussion: jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, pages 1301–1304, 1986.

Bradley Efron. *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM, 1982.

Bradley Efron. Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80 (1):3–26, 1993.

Michael P Fay and Barry I Graubard. Small-sample adjustments for wald-type tests using sandwich estimators. *Biometrics*, 57(4):1198–1206, 2001.

D. Firth. Discussion of the paper by liang, zeger & qaqish multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, Series B*, 54:24–26, 1992.

M Furi and M Martelli. On the mean value theorem, inequality, and inclusion. *The American Mathematical Monthly*, 98(9):840–846, 1991.

C. Genest, K. Ghoudi, and L.P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995.

Luca Greco, Walter Racugno, and Laura Ventura. Robust likelihood functions in bayesian inference. *Journal of Statistical Planning and Inference*, 138(5):1258–1270, 2008.

James W Hardin. The sandwich estimate of variance. *Advances in Econometrics*, 17:45–73, 2003.

David V Hinkley. Jackknifing in unbalanced situations. *Technometrics*, 19(3):285–292, 1977.

Susan D Horn, Roger A Horn, and David B Duncan. Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 70(350):380–385, 1975.

Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233, 1967.

Peter J Huber and Elvezio M Ronchetti. *Robust statistics*. Wiley, 2009.

Robert I Jennrich. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643, 1969.

Harry Joe. *Multivariate models and dependence concepts*, volume 73. CRC Press, 1997.

Göran Kauermann and Raymond J Carroll. The sandwich variance estimator: efficiency properties and coverage probability of confidence intervals. 2000.

Göran Kauermann and Raymond J Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396, 2001.

Erich Leo Lehmann. *Elements of large-sample theory*. Springer Science & Business Media, 1999.

Kung-Yee Liang, Scott L Zeger, and Bahjat Qaqish. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–40, 1992.

Danyu Y Lin and Lee-Jen Wei. The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):1074–1078, 1989.

James G MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29 (3):305–325, 1985.

Peter McCullagh. Discussion of the paper by liang, zeger & qaqish multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, Series B*, 54:33–34, 1992.

LB Mirel, LK Mohadjer, SM Dohrmann, J Clark, VL Burt, CL Johnson, and LR Curtin. National health and nutrition examination survey: estimation procedures, 2007-2010. *Vital and health statistics. Series 2, Data evaluation and methods research*, (159):1–25, 2013.

John F Monahan and Dennis D Boos. Proper likelihoods for bayesian analysis. *Biometrika*, 79(2):271–278, 1992.

Roger B Nelsen. *An introduction to copulas*. Springer Verlag, 2006.

MI Parzen, LJ Wei, and Z Ying. A resampling method based on pivotal estimating functions. *Biometrika*, 81(2):341–350, 1994.

FH Ruymgaart. Asymptotic normality of nonparametric tests for independence. *The Annals of Statistics*, pages 892–910, 1974.

Halbert White. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, 76(374):419–433, 1981.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

Chien-Fu Jeff Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, pages 1261–1295, 1986.

Xingzhi Zhan. *Matrix theory*, volume 147. American Mathematical Soc., 2013.