

## ABSTRACT

GOTWALT, CHRISTOPHER MICHAEL. Model-Robust Interval Estimation. (Under the direction of Dennis Boos.)

Confidence intervals are one of the most useful statistical tools. This dissertation is a study of several methods for forming confidence intervals that are insensitive to model assumptions, provided that the mean model for the data is not misspecified. The most commonly used robust confidence interval, the generalized Wald interval, is known to be liberal in small sample situations. We investigate several alternatives to the generalized Wald interval that are shown to often have superior performance: the generalized score interval; the robust profile likelihood interval; a new, modified generalized score interval that we call the GS2 interval; and we investigate a bootstrap calibration of the generalized Wald interval. We also introduce a new general procedure, length-optimal interval estimation, that takes an existing equal-tail confidence procedure and creates a new one whose length is shorter than the original. Surprisingly, in simulations we see that these shorter intervals are shown to sometimes enjoy higher coverage than their standard counterparts.

# **Model-Robust Interval Estimation**

by

**Christopher Michael Gotwalt**

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

in

**STATISTICS**

in the

GRADUATE SCHOOL

at

NC STATE UNIVERSITY

2003

---

Professor Dennis D. Boos  
Chair of Advisory Committee

---

Professor John Monahan

---

Professor Leonard Stefanski

---

Professor Charlie Smith

*To my wife, Jessica*

## Biography

Christopher Michael Gotwalt was born in York, Pennsylvania on January 2, 1974. He finished his secondary education at Ridgewood High School in New Port Richey, Florida in 1992.

Upon graduation, he was awarded the Florida Undergraduate Scholars Fund Scholarship. Christopher attended the University of South Florida in the USF Honors Program from August 1992 until May 1994. In October 1994, he received the Superior Academic Achievement Award at the USF Presidential Inauguration and Honors Convocation. He studied at the University of Florida from August 1994 until August 1997, whereupon he graduated with Honors, earning a Bachelor of Arts in Mathematics with a Minor in Physics.

In 1997, he began graduate studies in statistics at North Carolina State University. He was awarded the Mendenhall Teaching Fellowship in May 1999, and the following summer developed the MLAB educational statistics software toolkit for Matlab. From August 1999 to August 2000, Christopher was a Graduate Industrial Trainee at Analytical Sciences, Inc., where he developed SAS System programs for use by toxicology researchers at the NIH and NIEHS. Christopher received a Master of Science degree in Statistics from NCSU in May 2000, and was selected as a recipient of the Gertrude M. Cox Award as Outstanding Ph.D. Candidate. In August 2000, Christopher was awarded a GAANN Fellowship in Computational Science. He became the inaugural SAS Computational Statistics Fellow in January 2001. From May 2001 to May 2003, he was a student developer at the SAS Institute, where he developed statistical software for reliability and survival analysis, neural networks, and psychometric ability testing.

Christopher will start a position of Systems Developer at the SAS Institute.

## Acknowledgements

First and foremost, I would like to thank my patient wife, Jessica, without whose love and support this dissertation could not have been completed. I would also like to thank my parents, Rick and Judy Gotwalt, for their support through these years, and for giving me the inspiration and drive necessary to come this far.

I would like to extend my gratitude to my advisor, Professor Dennis Boos, for giving me this challenging topic, and for his time and guidance during the completion of my dissertation. I would like to thank my committee members, Professors John Monahan, Leonard Stefanski, and Charlie Smith for their helpful comments during the research process. I would also like to thank Dr. Sastry Pantula for his support throughout my graduate career, and for pushing me to be the best I can be.

I would also like to acknowledge John Sall, Executive Vice President and Co-Founder of the SAS Institute, for sponsoring the SAS Computational Statistics Fellowship, which provided the funding for this research, and enabled me to take the course work in applied mathematics and computer science which gave me the broad foundation necessary for a successful career as a statistical software developer.

Last, but not least, I would like to offer my sincerest gratitude to Dr. Juan E. Sanchez for all the help and support that he has given me over the years.

One cannot succeed alone, and I have been very fortunate in that I have had so many kind people help me along the way. I hope that I am given the opportunity to pass along the good will these people have given me.

# Contents

|                       |            |
|-----------------------|------------|
| <b>List of Tables</b> | <b>vii</b> |
|-----------------------|------------|

|                        |          |
|------------------------|----------|
| <b>List of Figures</b> | <b>x</b> |
|------------------------|----------|

|                                                                                                             |           |
|-------------------------------------------------------------------------------------------------------------|-----------|
| <b>1 Generalized Wald, Score, and Robust Profile Likelihood Confidence Intervals</b>                        | <b>1</b>  |
| 1.1 Introduction . . . . .                                                                                  | 1         |
| 1.2 Notation . . . . .                                                                                      | 3         |
| 1.3 Generalized Wald Confidence Intervals . . . . .                                                         | 5         |
| 1.4 Generalized Score Confidence Intervals . . . . .                                                        | 6         |
| 1.5 Robust Profile Likelihood Confidence Intervals . . . . .                                                | 10        |
| 1.6 Numerical Computation of Generalized Score and Robust Profile Likelihood Confidence Intervals . . . . . | 12        |
| 1.7 Simulations . . . . .                                                                                   | 14        |
| 1.7.1 Logistic Regression GEE . . . . .                                                                     | 15        |
| 1.7.2 Poisson Regression GEE . . . . .                                                                      | 21        |
| 1.7.3 Huber Robust Linear Regression . . . . .                                                              | 26        |
| 1.7.4 Simple Linear Regression With Covariate Measurement Error . . . . .                                   | 33        |
| <b>2 Length-Optimal Interval Estimation</b>                                                                 | <b>36</b> |
| 2.1 Motivation and Description . . . . .                                                                    | 36        |
| 2.2 Example: Binomial Proportion . . . . .                                                                  | 41        |
| 2.3 Theory . . . . .                                                                                        | 47        |
| 2.4 Simulations . . . . .                                                                                   | 59        |
| 2.4.1 Logistic GEE Simulations . . . . .                                                                    | 59        |
| 2.4.2 Poisson GEE simulations . . . . .                                                                     | 64        |
| 2.4.3 Huber Robust Linear Regression . . . . .                                                              | 67        |
| 2.4.4 Simple Linear Measurement Error Models . . . . .                                                      | 69        |
| <b>3 The GS2 Confidence Interval</b>                                                                        | <b>72</b> |
| 3.1 Introduction . . . . .                                                                                  | 72        |
| 3.2 Motivation . . . . .                                                                                    | 73        |
| 3.3 Theory . . . . .                                                                                        | 76        |
| 3.4 Simulations . . . . .                                                                                   | 83        |

|          |                                                                                   |            |
|----------|-----------------------------------------------------------------------------------|------------|
| 3.4.1    | Simple Linear Logistic GEE With Exchangeable Working Correlation Matrix . . . . . | 83         |
| 3.4.2    | GUIDE Data Example and Simulations . . . . .                                      | 86         |
| 3.4.3    | Simple Linear Poisson GEE With AR(1) Working Correlation Matrix . . . . .         | 89         |
| 3.4.4    | Epilepsy Data Example and Simulations . . . . .                                   | 92         |
| <b>4</b> | <b>Bootstrap Calibrated Generalized Wald Confidence Intervals</b>                 | <b>94</b>  |
| 4.1      | Introduction . . . . .                                                            | 94         |
| 4.2      | Bootstrap Confidence Intervals . . . . .                                          | 94         |
| 4.2.1    | Bias-Corrected and Accelerated Bootstrap Confidence Intervals . . . . .           | 95         |
| 4.2.2    | Bootstrap Calibrated Generalized Wald Confidence Intervals . . . . .              | 97         |
| 4.3      | Simulations . . . . .                                                             | 99         |
| 4.3.1    | Logistic GEE Simulations . . . . .                                                | 99         |
| 4.3.2    | Poisson GEE Simulations . . . . .                                                 | 102        |
| <b>5</b> | <b>Conclusions and Future Work</b>                                                | <b>104</b> |
| <b>A</b> | <b>A Serially Correlated Poisson-Type Random Number Generation Scheme</b>         | <b>107</b> |
|          | <b>Bibliography</b>                                                               | <b>111</b> |

# List of Tables

|      |                                                                                                                                                                           |    |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1  | Coverage and length for generalized Wald, score, and robust likelihood intervals for Bernoulli data under a simple linear logistic regression model. . . . .              | 17 |
| 1.2  | Proportion of infinite length generalized score and robust likelihood intervals for Bernoulli data under simple a linear logistic regression model. . . . .               | 18 |
| 1.3  | Logistic regression MLE's and approximate 95% GEE confidence intervals for the GUIDE dataset. . . . .                                                                     | 19 |
| 1.4  | Coverage and length for generalized Wald, score, and robust likelihood intervals for clustered Bernoulli data simulated from the GUIDE dataset. . . . .                   | 20 |
| 1.5  | Proportion of infinite length generalized score and robust likelihood intervals for clustered Bernoulli data simulated from the GUIDE dataset. . . . .                    | 21 |
| 1.6  | Coverage and length for generalized Wald, score, and robust likelihood intervals for serially correlated Poisson data under a log linear regression model. . . . .        | 23 |
| 1.7  | Proportion of infinite length generalized score and robust likelihood intervals for serially correlated Poisson data under a log linear regression model. . . . .         | 24 |
| 1.8  | Poisson regression MLE's and approximate 95% GEE confidence intervals for the Epilepsy dataset. . . . .                                                                   | 25 |
| 1.9  | Coverage and length for generalized Wald, score, and robust likelihood intervals for serially correlated Poisson data using the Epilepsy data covariates and MLE. . . . . | 26 |
| 1.10 | Coverage and length for generalized Wald, score, and robust likelihood intervals for iid data under a simple linear Huber regression model. . . . .                       | 32 |
| 1.11 | Proportion of infinite length generalized score and robust likelihood intervals for iid data under a simple linear Huber regression model. . . . .                        | 32 |
| 1.12 | Coverage and length for generalized Wald, score, and robust likelihood intervals from simple linear measurement error model. . . . .                                      | 34 |
| 1.13 | Coverage and length for generalized Wald, score, and robust likelihood intervals from simple linear measurement error model. . . . .                                      | 35 |
| 1.14 | Proportion of infinite length generalized score and robust likelihood intervals from simple linear measurement error model. . . . .                                       | 35 |
| 2.1  | Length-optimal and equal-tail treatment effect 95% confidence intervals for cluster subset of the Epilepsy data. . . . .                                                  | 41 |

|      |                                                                                                                                                                                   |    |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.2  | Coverage and length for length-optimal generalized score, and robust likelihood intervals for Bernoulli data under a simple linear logistic regression model. . . . .             | 61 |
| 2.3  | Proportion of infinite length length-optimal generalized score and robust likelihood intervals for Bernoulli data under simple a linear logistic regression model. . . . .        | 61 |
| 2.4  | Equal-tail and length-optimal approximate 95% GEE confidence intervals for the GUIDE dataset. . . . .                                                                             | 62 |
| 2.5  | Coverage and length for length-optimal generalized score and robust likelihood intervals for clustered Bernoulli data simulated from the GUIDE dataset. . . . .                   | 62 |
| 2.6  | Proportion of infinite length length-optimal generalized score and robust likelihood intervals for clustered Bernoulli data simulated from the GUIDE dataset. . . . .             | 63 |
| 2.7  | Coverage and length for length-optimal generalized score and robust likelihood intervals for serially correlated Poisson data under a log linear regression model. . . . .        | 65 |
| 2.8  | Proportion of infinite length length-optimal generalized score and robust likelihood intervals for serially correlated Poisson data under a log linear regression model. . . . .  | 65 |
| 2.9  | Equal-tail and length-optimal approximate 95% GEE confidence intervals for the Epilepsy dataset. . . . .                                                                          | 66 |
| 2.10 | Coverage and length for length-optimal generalized score and robust likelihood intervals for serially correlated Poisson data using the Epilepsy data covariates and MLE. . . . . | 66 |
| 2.11 | Coverage and length for length-optimal generalized score and robust likelihood intervals for iid data under a simple linear Huber regression model. . . . .                       | 67 |
| 2.12 | Coverage and length for length-optimal generalized score and robust likelihood intervals for iid data under a simple linear Huber regression model. . . . .                       | 68 |
| 2.13 | Proportion of infinite length length-optimal generalized score and robust likelihood intervals for iid data under a simple linear Huber regression model. . . . .                 | 68 |
| 2.14 | Coverage and length for length-optimal generalized score and robust likelihood intervals from simple linear measurement error model. . . . .                                      | 69 |
| 2.15 | Coverage and length for length-optimal generalized score and robust likelihood intervals from simple linear measurement error model. . . . .                                      | 70 |
| 2.16 | Proportion of infinite length length-optimal generalized score and robust likelihood intervals from simple linear measurement error model. . . . .                                | 71 |
| 3.1  | Coverage and length for generalized Wald, score, and GS2 intervals for Bernoulli data under a simple linear logistic regression model. . . . .                                    | 84 |
| 3.2  | Proportion of infinite length generalized score and GS2 intervals for Bernoulli data under simple a linear logistic regression model. . . . .                                     | 85 |
| 3.3  | Logistic GEE with exchangeable working correlation matrix estimates and approximate 95% confidence intervals parameters from the GUIDE dataset. . . . .                           | 86 |
| 3.4  | Coverage and length for generalized Wald, score, and GS2 intervals for clustered Bernoulli data simulated from the GUIDE dataset. . . . .                                         | 87 |
| 3.5  | Proportion of infinite length generalized score and GS2 intervals for clustered Bernoulli data simulated from the GUIDE dataset. . . . .                                          | 88 |
| 3.6  | Coverage and length for generalized Wald, score, and GS2 intervals for serially correlated Poisson data under a log linear regression model. . . . .                              | 90 |

|     |                                                                                                                                                                   |     |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.7 | Proportion of infinite length generalized score and GS2 intervals for serially correlated Poisson data under a log linear regression model. . . . .               | 91  |
| 3.8 | Poisson GEE AR(1) working correlation matrix estimates and approximate 95% GEE confidence intervals for the Epilepsy dataset. . . . .                             | 92  |
| 3.9 | Coverage and length for generalized Wald, score, and GS2 intervals for serially correlated Poisson data using the Epilepsy data covariates and MLE. . . . .       | 93  |
| 4.1 | Coverage and length for calibrated generalized Wald and $BC_a$ intervals for Bernoulli data under a simple linear logistic regression model. . . . .              | 100 |
| 4.2 | Coverage and length for calibrated generalized Wald and $BC_a$ intervals for clustered Bernoulli data simulated from the GUIDE dataset. . . . .                   | 101 |
| 4.3 | Coverage and length for calibrated generalized Wald and $BC_a$ intervals for serially correlated Poisson data under a log linear regression model. . . . .        | 102 |
| 4.4 | Coverage and length for calibrated generalized Wald and $BC_a$ intervals for serially correlated Poisson data using the Epilepsy data covariates and MLE. . . . . | 103 |

# List of Figures

|     |                                                                                                                                                                       |    |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1 | Confidence curve for a 33 cluster subset of the Epilepsy data. . . . .                                                                                                | 40 |
| 2.2 | Coverage plots for binomial proportion $n=5$ . . . . .                                                                                                                | 45 |
| 2.3 | Coverage plots for binomial proportion $n=10$ . . . . .                                                                                                               | 45 |
| 2.4 | Coverage plots for binomial proportion $n=20$ . . . . .                                                                                                               | 46 |
| 3.1 | Numerator and denominator of the generalized score test statistic as a function<br>of the null hypothesis value for a 33 cluster subset of the Epilepsy data. . . . . | 75 |

## Chapter 1

# Generalized Wald, Score, and Robust Profile Likelihood Confidence Intervals

### 1.1 Introduction

In likelihood inference there are three standard methods of generating confidence intervals for a parameter of interest. These three methods are derived by simply inverting the three corresponding likelihood-based hypothesis tests: the Wald test, the score test, and the likelihood ratio test. These confidence intervals require that the assumed probability model for the data is the one that actually generated the data; in other words, the probability model is not misspecified. If the assumed likelihood is misspecified, then the error rates for these confidence intervals are very likely to be significantly higher than their nominal error rates.

These standard intervals rely on the theory which states that Wald, score, and likelihood ratio tests are all asymptotically distributed as  $\chi^2$  random variables. If the model is incorrect this asymptotic result may not hold, often leading to higher than nominal error rates. In the case of the Wald and score tests, there are relatively simple modifications of the test statistics that can be made so that the adjusted statistics follow an asymptotic  $\chi^2$  distribution even when the assumed model is incorrect. However, in the case of the likelihood ratio test, this is not in general possible. Under model misspecification, the likelihood ratio test statistic is distributed asymptotically as a linear combination of  $p - p_0$   $\chi_1^2$  random variables (chi-squared with 1 degree of freedom), where  $p$  and  $p_0$  are the dimensionalities of the full parameter space and the parameter space under the null hypothesis, respectively (Kent 1982). The task of finding critical values for such a distribution is somewhat computationally challenging, and application-level software to compute such critical values does not appear to be generally available, which is possibly why it is uncommon to find the likelihood ratio test used in situations where the model is expected to be incorrect. However, in the case of confidence intervals, we are only interested in the situation where  $p - p_0 = 1$ . The likelihood test then follows the distribution of a  $\chi_1^2$  multiplied by a scalar constant that is not difficult to estimate. This fact appears to be quite unappreciated at this time. It suggests a procedure for generating confidence intervals that is in many ways superior to robust versions of the Wald and score confidence intervals. In this chapter we discuss and compare the generalized, robust versions of the Wald, score, and profile likelihood intervals.

## 1.2 Notation

Let the data be denoted by  $\mathbf{Y} = \{y_i\}_{i=1}^n$ , and suppose that the assumed model for the data is that the individual data points,  $y_i$ , are independent random variables that have density  $y_i \sim f(y_i|\theta, \mathbf{x}_i)$ , where  $\theta$  is an unknown parameter of dimension  $p$ , and each  $\mathbf{x}_i$  is a known vector of covariates whose dimension is less than or equal to  $p$ . The true density that generated the complete data  $\mathbf{Y}$  is  $g(\mathbf{Y}|\theta, \mathbf{X})$ , and is unknown. In particular,  $g$  is not necessarily equal to  $\prod f(y_i|\theta, \mathbf{x}_i)$  for any value of  $\theta$ . We assume that regularity conditions like those in White (1982) hold for  $f(y|\theta, \mathbf{x})$  and  $g(\mathbf{Y}|\theta, \mathbf{X})$ , so that the maximum likelihood estimator (MLE) of  $\theta$  under model  $f$ ,  $\hat{\theta}_n$ , is strongly consistent for  $\theta_t$ , the true value of  $\theta$ , and is asymptotically normally distributed. The model log likelihood is,

$$l_n(\theta|\mathbf{Y}) = \sum_{i=1}^n \log f(y_i|\theta, \mathbf{x}_i). \quad (1.1)$$

We denote the gradient of the log likelihood, or score function, by,

$$S(\theta) = \frac{\partial}{\partial \theta} l_n(\theta|\mathbf{Y}). \quad (1.2)$$

There are two information-type matrices that will be of use. These are, assuming that the necessary limits exist,

$$I(\theta) = - \lim_{n \rightarrow \infty} E_g \left[ \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta \partial \theta^T} \log f(y_i|\theta, \mathbf{x}_i) \right] \quad (1.3)$$

and,

$$D(\theta) = \lim_{n \rightarrow \infty} E_g \left[ \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(y_i | \theta, \mathbf{x}_i) \frac{\partial}{\partial \theta^T} \log f(y_i | \theta, \mathbf{x}_i) \right]. \quad (1.4)$$

As  $g$  is unknown in practical situations, one commonly uses empirical estimators for the two preceding quantities,

$$\hat{I}(\theta) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta \partial \theta^T} \log f(y_i | \theta, \mathbf{x}_i), \quad (1.5)$$

and

$$\hat{D}(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(y_i | \theta, \mathbf{x}_i) \frac{\partial}{\partial \theta^T} \log f(y_i | \theta, \mathbf{x}_i). \quad (1.6)$$

Note that, strictly speaking, these are not truly estimates, but are estimates of a matrix function that takes  $\theta$  as its argument.

Because our interest lies in confidence intervals, we partition  $\theta$  into

$$\theta = \begin{pmatrix} \psi \\ \eta \end{pmatrix}, \quad (1.7)$$

where  $\psi$  is the scalar parameter of interest and  $\eta$  is the vector containing the remaining  $p - 1$  parameters. When vector or matrix quantities are seen with subscripts such as  $S_1$  or  $I_{12}$ , we assume that the full vector or matrix has been partitioned conformally with the  $(\psi, \eta)$  partitioning of  $\theta$ .

The restricted MLE of  $\eta$  under the constraint that  $\psi$  is held fixed,  $\tilde{\eta}(\psi)$ , figures prominently in the calculation of generalized score and robust likelihood intervals, and is computed by solving

$$S_2(\psi, \tilde{\eta}(\psi)) = 0. \quad (1.8)$$

### 1.3 Generalized Wald Confidence Intervals

Because of its simplicity, the generalized Wald confidence interval has been the most commonly used confidence interval that is robust to misspecification. Once parameter estimation is completed, the generalized Wald interval is not difficult to compute, and has a simple, easily interpretable, form.

However, there are some serious drawbacks to these intervals. Although under very broad conditions the generalized Wald interval has an asymptotic coverage rate equal to the nominal  $1 - \alpha$  rate, in practical small sample situations the actual coverage may be significantly smaller than stated. The generalized Wald interval also lacks the parameterization invariance that the other two likelihood-based intervals, the generalized score and robust profile likelihood intervals, often share.

The generalized Wald interval is motivated by the asymptotic normality of the MLE under possible model misspecification. Under broad conditions, the MLE,  $\hat{\theta}_n$ , has an asymptotic normal distribution (Stefanski and Boos 2002)

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I^{-1}(\theta)D(\theta)I^{-1}(\theta)). \quad (1.9)$$

This convergence suggests an obvious form for the generalized Wald statistic for testing

$$H_0 : \psi = \psi_0,$$

$$T_{GW}(\psi_0) = \frac{n(\hat{\psi} - \psi_0)^2}{\{\hat{I}^{-1}(\hat{\theta})\hat{D}(\hat{\theta})\hat{I}^{-1}(\hat{\theta})\}_{11}}. \quad (1.10)$$

This generalized Wald test statistic, which has a null asymptotic  $\chi_1^2$  distribution, is easily inverted to obtain the corresponding confidence interval,

$$\hat{\psi}_n \pm \sqrt{\{\hat{I}^{-1}(\hat{\theta})\hat{D}(\hat{\theta})\hat{I}^{-1}(\hat{\theta})\}_{11}} \frac{z_{1-\alpha/2}}{\sqrt{n}}, \quad (1.11)$$

where  $z_{1-\alpha/2}$  is the  $1 - \alpha/2$  quantile of the standard normal distribution. In order to improve the generalized Wald interval's coverage we will often use critical values from the  $t$ -distribution with  $n - p$  degrees of freedom, which can dramatically improve its small sample performance. We will call such intervals generalized Wald- $t$  confidence intervals.

The quantity  $\hat{I}^{-1}(\hat{\theta})\hat{D}(\hat{\theta})\hat{I}^{-1}(\hat{\theta})$  is commonly referred to as the *sandwich matrix*, and plays a central role in robust likelihood inference.

## 1.4 Generalized Score Confidence Intervals

The generalized score interval arises from inverting the generalized score statistic (Boos 1992) for the hypothesis  $H_0: \psi = \psi_0$ ,

$$\begin{aligned} T_{GS}(\psi_0) &= n^{-1}S^T(\tilde{\theta})\hat{I}^{-1}(\tilde{\theta})\mathbf{e}_1(\mathbf{e}_1^T\hat{I}^{-1}(\tilde{\theta})\hat{D}(\tilde{\theta})\hat{I}^{-1}(\tilde{\theta})\mathbf{e}_1)^{-1}\mathbf{e}_1^T\hat{I}^{-1}(\tilde{\theta})S(\tilde{\theta}) \\ &= n^{-1}S^T(\tilde{\theta})\hat{I}^{-1}(\tilde{\theta})S(\tilde{\theta}) \cdot \frac{\{\hat{I}^{-1}(\tilde{\theta})\}_{11}}{\{\hat{I}^{-1}(\tilde{\theta})\hat{D}(\tilde{\theta})\hat{I}^{-1}(\tilde{\theta})\}_{11}}, \end{aligned} \quad (1.12)$$

where  $\tilde{\theta} = (\psi_0, \eta(\tilde{\psi}_0))^T$ , and  $\mathbf{e}_1$  is the unit vector whose first element is one, and the remaining elements are zero.

The first representation is the form of the fully general, hypothesis via constraint, version of the generalized score test statistic (Boos 1992), for the special case of  $H_0 : \psi = \psi_0$ . The second form is reduced from the first and highlights its relationship to Rao's original score test statistic,  $n^{-1}S^T I^{-1} S$ . The generalized score test statistic is equal to Rao's score test multiplied by a scalar robust correction constant. This form, a non-robust test statistic multiplied by a robust correction, is similar of the robust likelihood ratio test, that we introduce in the next section.

In many common situations, the generalized score test is invariant to differentiable parameter transformations. We are easily able to verify this using an argument similar to one given in Stafford (1996). A sufficient condition for the generalized score test to be invariant is that the observed and expected information matrices coincide. This is the case, for example, in generalized linear models with canonical link function, since the second derivative matrix of the loglikelihood is functionally independent of  $\mathbf{Y}$ . Suppose that  $(\psi, \eta^T)$  is an invertible and twice differentiable reparameterization of  $(\tau, \lambda^T) = (\tau(\psi), \lambda^T(\psi, \eta))$ . Notice that the two parameters  $\psi$  and  $\tau$  are measuring the same quantity on different scales of measurement whose relationship may be nonlinear. This means that  $H_0 : \psi = \psi_0$  and  $H_0 : \tau = \tau(\psi_0)$  are testing the same hypothesis. If we differentiate the reparameterized likelihood,  $l_n(\tau(\psi), \lambda(\psi, \eta) | \mathbf{Y})$ , and take the necessary expectations for the computation of information matrices, we obtain the  $\psi$ -scale score vector and information matrices,

$$S_\psi = J(\psi, \eta) S(\tau(\psi), \lambda(\psi, \eta)), \quad (1.13)$$

$$I_\psi = J(\psi, \eta) I(\tau(\psi), \lambda(\psi, \eta)) J(\psi, \eta)^T, \quad (1.14)$$

and

$$D_\psi = J(\psi, \eta) D(\tau(\psi), \lambda(\psi, \eta)) J(\psi, \eta)^T, \quad (1.15)$$

where  $J$  is the Jacobian of the reparameterization,

$$J(\psi, \eta) = \begin{pmatrix} \frac{\partial \tau}{\partial \psi} & \mathbf{0}^T \\ \frac{\partial \lambda}{\partial \psi} & \frac{\partial \lambda}{\partial \eta^T} \end{pmatrix}. \quad (1.16)$$

Now, Rao's score test statistic is invariant because,

$$\begin{aligned} (JS)^T (JIJ^T)^{-1} (JS) &= S^T J^T J^{-T} I^{-1} J^{-1} JS \\ &= S^T I^{-1} S, \end{aligned} \quad (1.17)$$

establishes that the  $\psi$ -scale and  $\tau$ -scale score test statistics coincide. Partitioned-matrix algebra shows the invariance of the correction constant, although individually  $\hat{I}(\theta)$  and  $\hat{D}(\theta)$  are not invariant.

$$\begin{aligned} \frac{\{J^{-T} I^{-1} J^{-1}\}_{11}}{\{(J^{-T} I^{-1} J^{-1})(J D J^T)(J^{-T} I^{-1} J^{-1})\}_{11}} &= \frac{\{J^{-T} I^{-1} J^{-1}\}_{11}}{\{J^{-T} I^{-1} D I^{-1} J^{-T}\}_{11}} \\ &= \frac{J_{11}^2 \{I^{-1}\}_{11}}{J_{11}^2 \{I^{-1} D I^{-1}\}_{11}} \\ &= \frac{\{I^{-1}\}_{11}}{\{I^{-1} D I^{-1}\}_{11}}. \end{aligned} \quad (1.18)$$

This demonstrates that the generalized score test statistics in the  $\psi$ -scale and the  $\tau$ -scale have the same value, proving the invariance of the generalized score test statistic.

If  $\psi_0$  is the true value of  $\psi$ , then the  $T_{GS}$  has an asymptotic  $\chi_1^2$  distribution, like the generalized Wald statistic. This test statistic is derived using the asymptotic distribution of  $n^{-1}S(\theta)$ , which is a mean of independent random variables, and therefore, under broad regularity conditions, is asymptotically normally distributed by the Central Limit Theorem. The Wald statistic also implicitly assumes that  $n^{-1}S(\theta)$  is normally distributed, but in addition requires the use of the delta theorem to obtain the asymptotic normality of  $\hat{\theta}$ , as it is the solution of the equation  $S(\hat{\theta}) = 0$ . Thus, one would expect that the  $\chi_1^2$  approximation to the distribution of the generalized score statistic would be more accurate than that of the generalized Wald statistic. The generalized score statistic contains only restricted estimates of  $\theta$  under the null hypothesis. This usually gives it a slight computational advantage over the Wald and likelihood ratio tests, because there is one less parameter that needs to be estimated. Unfortunately, this computational advantage does not carry over to generalized score confidence intervals. This is because obtaining the generalized score confidence interval amounts to solving  $T_{GS}(\psi) - z_{1-\alpha}^2 = 0$  twice: once each for the upper and lower endpoints of the interval. As  $T_{GS}$  is generally a nonlinear function of  $\psi$ , an iterative procedure is required, and  $\hat{\eta}(\psi)$  will have to be recomputed during each iteration.

As the simulations later in this chapter show, in small samples the generalized score confidence intervals maintain their nominal coverage rate better than the generalized Wald intervals, but in small samples they can be highly inefficient. These confidence intervals can be many times wider than generalized Wald intervals, and in some situations can be infinite in length with high frequency. This surprising loss of efficiency led us to investigate the robust

profile likelihood confidence interval.

## 1.5 Robust Profile Likelihood Confidence Intervals

The robust profile likelihood confidence interval is the least well known of the likelihood-based intervals, possibly because the critical values of the asymptotic distribution of the robust likelihood ratio test initially appear to be difficult and unwieldy to compute. A somewhat closer look reveals that robust likelihood confidence intervals are no more difficult to compute than generalized score confidence intervals, however. The idea behind these intervals goes back to Foutz and Srivastava (1977), and Kent (1982), who independently derived the asymptotic distribution of the likelihood ratio test under model misspecification. Unlike the generalized Wald and score statistics, except for the scalar parameter of interest case, there is no direct modification of the likelihood ratio statistic that generally gives it an asymptotic  $\chi^2$  distribution. In place of such a modification, Foutz et al, and Kent found the asymptotic distribution of the likelihood ratio test under misspecification. The asymptotic form of the likelihood ratio statistic for  $H_0 : \psi = \psi_0$ , is

$$2(l_n(\hat{\theta}) - l_n(\psi_0, \tilde{\eta}(\psi_0))). \quad (1.19)$$

Assuming for the moment that  $\psi$  has dimension  $p_0$  between 1 and  $p$ , Kent (1982) gives the asymptotic distribution of the test statistic to be that of

$$\sum_{i=1}^{p_0} c_i X_i,$$

where the  $X_i$  are independent  $\chi_1^2$  random variables, and the  $c_i$  are the eigenvalues of the matrix

$$\{\{I(\theta)^{-1}\}_{11}\}^{-1}\{I(\theta)^{-1}D(\theta)I(\theta)^{-1}\}_{11}. \quad (1.20)$$

Finding critical values for a linear combination of  $\chi^2$  random variables can be done using Imhof's algorithm (Imhof 1961), but in the case of confidence intervals, this is unnecessary, as the asymptotic distribution of the likelihood ratio test is just a scalar constant multiplied by a  $\chi_1^2$  random variable. This correction constant  $c$  is simply the robust estimate of the variance of  $\psi$  divided by the model based variance estimate,

$$c = \frac{\{I(\theta)^{-1}D(\theta)I(\theta)^{-1}\}_{11}}{\{I(\theta)^{-1}\}_{11}}, \quad (1.21)$$

and can be consistently estimated by simply inserting the MLE of  $\hat{\theta}$ ,

$$\hat{c} = \frac{\{I(\hat{\theta})^{-1}D(\hat{\theta})I(\hat{\theta})^{-1}\}_{11}}{\{I(\hat{\theta})^{-1}\}_{11}}. \quad (1.22)$$

So, for scalar  $\psi$ , the robust likelihood ratio test statistic is defined to be

$$T_{RLR}(\psi_0) = \frac{2(l_n(\hat{\theta}_n) - l_n(\psi_0, \tilde{\eta}(\psi_0)))}{\hat{c}}, \quad (1.23)$$

and has an asymptotic  $\chi_1^2$  distribution under  $H_0 : \psi = \psi_0$ .

Like the generalized score interval, the robust profile likelihood interval is invariant under all differentiable reparameterizations that preserve  $\psi$  (Stafford 1996). This is easily seen by noting that the correction constant is invariant, by the same argument made in the

demonstration of the generalized score test's invariance, and also by the well-known invariance of the likelihood ratio test statistic.

The simulations later in this chapter indicate that the robust profile likelihood interval frequently outperforms both the Wald and generalized score confidence intervals. Its coverage of the true parameter value is closer to nominal than the Wald interval, while lacking the loss of efficiency and conservatism of the generalized score interval.

## 1.6 Numerical Computation of Generalized Score and Robust Profile Likelihood Confidence Intervals

The computation of the endpoints of profile likelihood and generalized score confidence intervals, while simple in principle, requires some care to ensure that convergence occurs whenever possible. The basic problem is this: given a test statistic  $T(\psi, \tilde{\eta}(\psi))$  whose null distribution is  $\chi_1^2$ , find two solutions  $(\psi^L, \psi^U)$  to the scalar equation

$$T(\psi, \tilde{\eta}(\psi)) - z_{1-\alpha/2}^2 = 0. \quad (1.24)$$

There are several difficulties that an algorithm for finding  $(\psi^L, \psi^U)$  must overcome. First and foremost, the algorithm should employ a root bracketing scheme to ensure that the same root is not found twice. Root bracketing algorithms maintain a bracket, or interval, that is known to contain the root of an equation. At every iteration, the bracket's length is decreased until it is smaller than some prespecified tolerance. The bisection method is the simplest and most well known of the root bracketing schemes, although there are much faster alternatives. The task of

finding an initial bracket for each of the confidence interval's endpoints is made easier because we know *a priori* that  $\psi^L < \hat{\psi} < \psi^U$ , as all likelihood-based tests accept the null hypothesis,  $H_0 : \psi = \hat{\psi}$ . An initial line search is necessary, however, to completely bracket the root, and to ensure that both of the two roots exist. Unlike Wald type confidence intervals, generalized score and robust profile confidence intervals are not necessarily finite in length. In fact, if one were to speak carefully, one would speak instead of generalized score and likelihood confidence *regions*, as opposed to intervals, because the set of values of  $\psi$  for which test  $T$  accepts the null hypothesis,  $\{\psi | T(\psi, \tilde{\eta}(\psi)) < z_{1-\alpha/2}^2\}$ , is not guaranteed to be an interval at all, even an infinite one. However, in the experience of the author, generalized score or profile likelihood confidence regions that are not intervals (finite or infinite in length) are very rare for most models. Furthermore, the asymptotic equivalence of the generalized Wald test, the generalized score, and robust likelihood ratio tests (in the case of a single parameter of interest) ensures that the probability that these confidence regions are not intervals goes to zero as the sample size becomes large. A final consideration is a slight preference for derivative-free methods for obtaining  $(\psi^L, \psi^U)$ . Although the derivative of the profile likelihood is not difficult to derive, the derivative of the generalized score test involves quantities like vector derivatives of the inverse of a matrix, a tensor of degree three, which would be convenient to avoid computing whenever possible.

Brent's scalar root finding method (Brent 1973) provides a solution that meets all the demands stated in the preceding paragraph, and is fast. In fact, if the derivative is at least as difficult to compute as the function whose root is being sought, Brent's method is even faster than Newton's method (Press, Teukolsky, Vetterling, and Flannery 1988). The details of Brent's method are too complicated to be reviewed here, but generally speaking, it is a carefully

constructed combination of bisection, regula falsi, and inverse quadratic interpolation that is put together in such a way that it converges quickly for well behaved problems, and in most cases does no worse than bisection on hard problems.

Thus, an effective strategy for computing the endpoints of generalized score and robust profile likelihood confidence intervals is to first perform a line search to find where  $T(\psi, \tilde{\eta}(\psi)) - z_{1-\alpha/2}^2$  changes sign from negative to positive. Without loss of generality, we will discuss the computation of upper endpoints, as the method for computing lower endpoints is completely analagous. We know from the outset that  $T(\hat{\psi}, \tilde{\eta}(\hat{\psi})) = 0$ , so that for both endpoints the side of the root bracket that corresponds to where  $T(\psi, \tilde{\eta}(\psi)) - z_{1-\alpha/2}^2 < 0$  can be initially set to  $\hat{\psi}$ . Then, we initialize  $\psi$  to be equal to the upper generalized Wald endpoint and compute  $T(\psi, \tilde{\eta}(\psi)) - z_{1-\alpha/2}^2$ . Because of the asymptotic equivalence of these likelihood-based tests, the generalized Wald upper endpoint makes a perfectly reasonable, easily computed, starting value. We continue incrementing  $\psi$  by the standard error of the MLE, and computing  $T(\psi, \tilde{\eta}(\psi)) - z_{1-\alpha/2}^2$  until this quantity becomes positive. Once this happens, a bracket that contains the upper endpoint of the confidence interval has been found, and one can now employ Brent's method (or bisection, or any other root finding algorithm) to obtain a solution to the equation. However, if after a predetermined number of steps, a value of  $\psi$  that yields a positive value of  $T(\psi, \tilde{\eta}(\psi)) - z_{1-\alpha/2}^2$  has not been found, declare the endpoint to be infinite in value.

## 1.7 Simulations

In this section are the results of simulations that compare the generalized Wald, generalized score, and robust profile likelihood intervals in a variety of situations. We focus on

logistic and log-linear GEE models, Huber robust linear regression models, and linear measurement error models.

Generalized estimating equations (GEE) models are an important class of models that extend the applicability of generalized linear models to data that are correlated within clusters (Liang and Zeger 1986). The standard method for hypothesis testing and generating confidence intervals for parameters of a GEE model is the generalized Wald test. However, it is well known (Gunsolley, Getchell, and Chinchilli 1995) that Wald methods do not perform well in datasets where the number of clusters is small. In this section we give the results from a number of simulations that provide evidence that the generalized score and robust profile likelihood confidence intervals offer a significant improvement over the standard generalized Wald interval. It also becomes clear that the generalized score interval is often less efficient than the robust profile likelihood interval, as measured by its length.

### 1.7.1 Logistic Regression GEE

We investigate, via simulation, the small sample properties of confidence intervals for parameters that describe the mean of binary data that is correlated in clusters. Two situations are examined. The first is a series of simulations in the case of simple linear logistic regression, and the second set of simulations are motivated by the data from a study (Preisser and Qaqish 1999) of urinary incontinence.

#### Simple Linear Logistic GEE Simulations

In each of the following simulations we generate  $N_S = 2000$  data sets in the following manner. For each dataset there are a total of  $K$  clusters of size 4, where  $K$  is 5, 10 or 20. We

represent the data by  $\{y_{ij}\}$  where  $i = 1, 2, \dots, K$  and  $j = 1, 2, 3, 4$ . The data for each cluster is then created using Al-Osh and Lee's algorithm for generating correlated binary data, given a specific mean and correlation structure (Al-Osh and Lee 2001). The observations within the cluster are individually *Bernoulli* random variables whose *log* odds of equalling 1 are  $\eta + \psi x_{ij}$ , where the  $x_{ij}$  are independently drawn *Uniform* $(-1, 1)$  random variables, and the values  $\eta$  and  $\psi$  are known constants. The observations within the cluster have common correlation,  $\gamma$ , where the value of  $\gamma$  is either 0, .2, or .5, depending on the simulation.

We estimate  $\theta = (\psi, \eta)^T$  with a GEE model using the independence working correlation matrix, that amounts to ordinary simple linear logistic regression using the  $x_{ij}$  as covariates. We then use the estimate of  $\theta$  to form 95% (equal tail) generalized Wald, generalized score, and robust profile likelihood confidence intervals for  $\psi$ . Because of the overall poor performance of the generalized Wald interval, we also report simulation results for the generalized Wald interval using *t*-distribution critical values in place of the more commonly used normal-theory critical values. By analogy with linear models, the degrees of freedom for the *t*-distribution critical values is set to the number of clusters minus the number of estimated parameters. Using these critical values dramatically improves the coverage of the generalized Wald interval.

The size of these simulations,  $N_S = 2000$ , was chosen so that for values near .95 the estimated coverage probabilities should be known to within about .01 of the true value with 95% certainty. The proportion of the confidence intervals that covered the true value of  $\psi$  are reported, along with median length of the intervals. The median length was chosen to be reported, rather than the mean length, because the possibility of infinite intervals renders the mean a useless measure of *typical* length. We also report the proportion of intervals that are infinite in length. These infinite intervals occur with surprising frequency in some situations,

and is a manifestation of a significant loss of inferential power.

Table 1.1: Coverage and length for generalized Wald, score, and robust likelihood intervals for Bernoulli data under a simple linear logistic regression model.

| $(\psi, \eta)$ | # of Clusters $K =$ |                | $\rho = 0$ |     |     | $\rho = 0.2$ |     |     | $\rho = 0.5$ |     |     |
|----------------|---------------------|----------------|------------|-----|-----|--------------|-----|-----|--------------|-----|-----|
|                |                     |                | 5          | 10  | 20  | 5            | 10  | 20  | 5            | 10  | 20  |
| (0,0)          | Coverage            | Gen. Wald      | .87        | .92 | .93 | .85          | .92 | .93 | .82          | .91 | .93 |
|                |                     | Gen. Wald- $t$ | .96        | .94 | .94 | .97          | .93 | .95 | .94          | .95 | .95 |
|                |                     | Gen. Score     | .98        | .96 | .95 | .98          | .96 | .95 | .98          | .95 | .94 |
|                |                     | Rob. Like.     | .95        | .95 | .96 | .95          | .96 | .95 | .95          | .95 | .95 |
|                | Median              | Gen. Wald      | 2.9        | 2.1 | 1.5 | 2.9          | 2.1 | 1.5 | 2.7          | 2.1 | 1.5 |
|                | Length              | Gen. Wald- $t$ | 4.8        | 2.5 | 1.6 | 4.8          | 2.5 | 1.6 | 4.5          | 2.5 | 1.7 |
|                |                     | Gen. Score     | $\infty$   | 3.4 | 1.8 | $\infty$     | 3.4 | 1.8 | $\infty$     | 3.5 | 1.8 |
|                |                     | Rob. Like.     | 3.7        | 2.4 | 1.6 | 3.7          | 2.4 | 1.6 | 3.7          | 2.4 | 1.6 |
|                | Coverage            | Gen. Wald      | .87        | .93 | .94 | .83          | .92 | .93 | .78          | .89 | .93 |
|                |                     | Gen. Wald- $t$ | .97        | .96 | .94 | .97          | .94 | .95 | .94          | .93 | .94 |
|                |                     | Gen. Score     | .97        | .96 | .95 | .96          | .95 | .95 | .96          | .93 | .94 |
|                |                     | Rob. Like.     | .95        | .96 | .95 | .93          | .96 | .96 | .88          | .91 | .94 |
|                | Median              | Gen. Wald      | 3.2        | 2.3 | 1.6 | 3.0          | 2.3 | 1.6 | 3.8          | 3.0 | 2.2 |
|                | Length              | Gen. Wald- $t$ | 5.4        | 2.8 | 1.7 | 5.1          | 2.7 | 1.7 | 7.0          | 3.6 | 2.3 |
|                |                     | Gen. Score     | $\infty$   | 3.8 | 2.0 | $\infty$     | 3.9 | 1.9 | $\infty$     | 5.7 | 2.7 |
|                |                     | Rob. Like.     | 3.9        | 2.6 | 1.7 | 4.0          | 2.6 | 1.7 | 4.9          | 3.2 | 2.2 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Cluster size=4.

There are two things that are immediately seen by glancing at the simulation results in Tables 1.1 and 1.2. The first is that the generalized Wald interval's coverage rate is much too low. In none of these simulations does the generalized Wald interval's coverage rate exceed .94, and in the smallest sample sizes it is below .87. This result is not really surprising, however, since it is known that Wald tests can perform very poorly in small samples. The second striking result is the extreme conservatism of the generalized score interval, especially in small samples. When  $K = 5$  it is generally infinite in length about 90% of the time. However, when the number of clusters is 20, its median length is about 10% greater than that of the robust profile

Table 1.2: Proportion of infinite length generalized score and robust likelihood intervals for Bernoulli data under simple a linear logistic regression model.

|                |            |                     | $\rho = 0$ |     |    | $\rho = 0.2$ |     |    | $\rho = 0.5$ |     |     |
|----------------|------------|---------------------|------------|-----|----|--------------|-----|----|--------------|-----|-----|
| $(\psi, \eta)$ |            | # of Clusters $K =$ | 5          | 10  | 20 | 5            | 10  | 20 | 5            | 10  | 20  |
| (0,0)          | Proportion | Gen. Score          | .87        | .01 | 0  | .90          | .01 | 0  | .92          | .05 | 0   |
|                | Infinite   | Rob. Like.          | 0          | 0   | 0  | 0            | 0   | 0  | .01          | 0   | 0   |
| (1,0)          | Proportion | Gen. Score          | .91        | .04 | 0  | .92          | .06 | 0  | .96          | .16 | .01 |
|                | Infinite   | Rob. Like.          | 0          | 0   | 0  | 0            | 0   | 0  | 0            | 0   | 0   |

interval's, while its coverage is generally lower than that of the robust profile interval's.

The robust profile likelihood interval and the generalized Wald interval using  $t$  critical values perform well, and are quite competitive. Except for the  $(\psi = 1, \rho = .5)$  case, the robust profile likelihood interval maintains its coverage well. The generalized Wald- $t$  interval also maintains its coverage well, but appears to be less efficient than the robust likelihood interval, as its median length in these simulations is uniformly longer than the robust profile likelihood's in all the  $K = 5$  and  $K = 10$  situations.

## GUIDE Data Example and Simulations

The next group of simulations is motivated by the relatively recent Guidelines for Urinary Incontinence Discussion and Evaluation (GUIDE) study (Preisser and Qaqish 1999). In this study, researchers contacted 137 patients from 38 medical practices that suffered from various degrees of incontinence, and were asked whether their incontinence interfered with their day to day activities, or otherwise bothered them. Their response to this question was considered to be the independent variable of interest (1=bothered), and there were five other covariates: a sex indicator (SEX), the scaled and centered ages (AGE), the average number of leaking accidents per day (DAYACC), a severity indicator (SEVERE), and the average number

of times per day the patient uses the restroom to urinate (TOILET). The data are assumed to be correlated within each medical practice. The data from one of the clusters was removed because of missing values in one of the covariates for all patients in the cluster, leaving a total of 37 clusters. The cluster sizes ranged from 1 to 8, and an analysis of the data using a GEE model with the exchangeable working correlation matrix yielded an estimate of the within-cluster correlation equal to .17.

The data in the simulations were generated in a manner similar to the previous simulations. We generate simulation data using parameter values equal to the logistic regression MLE for the complete data set, which are displayed in Table 1.3 with their corresponding 95% confidence intervals.

Table 1.3: Logistic regression MLE's and approximate 95% GEE confidence intervals for the GUIDE dataset.

| Parameter | MLE  | Gen. Wald  | Gen. Wald- $t$ | Gen. Score  | Rob. Like   |
|-----------|------|------------|----------------|-------------|-------------|
| SEX       | -.62 | (-2.0,.72) | (-2.0,.77)     | (-2.0,1.7)  | (-2.0,.75)  |
| AGE       | -.70 | (-1.8,.44) | (-1.9,.49)     | (-2.0,.67)  | (-1.9,.45)  |
| DAYACC    | .43  | (.23,.63)  | (.23,.64)      | (.26,.74)   | (.25,.65)   |
| SEVERE    | .86  | (.13,1.6)  | (.10,1.6)      | (.069,1.9)  | (.15,1.6)   |
| TOILET    | .080 | (-.10,.26) | (-.10,.26)     | (-.074,.42) | (-.095,.27) |

We provide simulation results for confidence intervals about the SEX and AGE parameters in Tables 1.4 and 1.5. Two sample sizes are investigated, the full dataset (ignoring the clusters with missing covariate information) of  $K = 37$  clusters, and also a subset of the data consisting of  $K = 20$  clusters. Two values of the within cluster correlation are investigated,  $\rho = 0$  and  $\rho = .2$ . For these values of the regression parameters, it is not possible to generate from a distribution whose within cluster correlation is much larger than .2, due to constraints

imposed on the correlation by having different observations within a cluster having different success probabilities.

Table 1.4: Coverage and length for generalized Wald, score, and robust likelihood intervals for clustered Bernoulli data simulated from the GUIDE dataset.

| Parameter | # of Clusters $K =$ |                | $\rho = 0$ |     | $\rho = 0.2$ |     |
|-----------|---------------------|----------------|------------|-----|--------------|-----|
|           |                     |                | 20         | 37  | 20           | 37  |
| SEX       | Coverage            | Gen. Wald      | .89        | .91 | .85          | .92 |
|           |                     | Gen. Wald- $t$ | .90        | .93 | .88          | .93 |
|           |                     | Gen. Score     | .97        | .95 | .94          | .95 |
|           |                     | Rob. Like.     | .95        | .95 | .93          | .95 |
|           | Median              | Gen. Wald      | 3.3        | 2.4 | 3.3          | 2.1 |
|           |                     | Gen. Wald- $t$ | 3.6        | 2.5 | 3.6          | 2.5 |
|           | Length              | Gen. Score     | 22.3       | 3.3 | 23.2         | 3.6 |
|           |                     | Rob. Like.     | 3.9        | 2.6 | 4.1          | 2.7 |
|           | Coverage            | Gen. Wald      | .87        | .94 | .90          | .91 |
|           |                     | Gen. Wald- $t$ | .91        | .95 | .93          | .92 |
|           |                     | Gen. Score     | .93        | .96 | .94          | .92 |
|           |                     | Rob. Like.     | .93        | .95 | .95          | .93 |
|           | Median              | Gen. Wald      | 3.3        | 2.4 | 3.2          | 2.4 |
|           |                     | Gen. Wald- $t$ | 3.5        | 2.3 | 3.6          | 2.5 |
|           | Length              | Gen. Score     | 5.3        | 2.7 | 5.7          | 2.9 |
|           |                     | Rob. Like.     | 3.6        | 2.4 | 3.7          | 2.5 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Table 1.5: Proportion of infinite length generalized score and robust likelihood intervals for clustered Bernoulli data simulated from the GUIDE dataset.

|           |            |                     | $\rho = 0$ |     | $\rho = 0.2$ |     |
|-----------|------------|---------------------|------------|-----|--------------|-----|
| Parameter |            | # of Clusters $K =$ | 20         | 37  | 20           | 37  |
| SEX       | Proportion | Gen. Score          | .50        | .02 | .56          | .03 |
|           | Infinite   | Rob. Like.          | .03        | 0   | 0            | 0   |
| AGE       | Proportion | Gen. Score          | .04        | 0   | .10          | 0   |
|           | Infinite   | Rob. Like.          | 0          | 0   | 0            | 0   |

In these simulations the difference in the performances of the generalized Wald- $t$  and robust profile intervals is more pronounced. Here the robust profile interval clearly outperforms the other confidence intervals. Again we see the inefficiency of the generalized score interval. When  $K = 20$  the generalized score interval for the *SEX* effect was infinite 50% of the time.

### 1.7.2 Poisson Regression GEE

As in the previous section, we provide simulation results from two groups of simulations. The first group of simulation results is from simple linear Poisson regression where the data are correlated in clusters, and the second group is motivated by the data from a study of epileptic seizures (Thall and Vail 1990).

#### Simple Linear Poisson GEE Simulations

Like the logistic regression simulations we generate  $N_S = 2000$  datasets for each simulation. For each dataset there are a total of  $K$  clusters of size 4, where  $K$  is 10, 20 or 50. We represent the data by  $\{y_{ij}\}$  where  $i = 1, 2, \dots, K$  and  $j = 1, 2, 3, 4$ . To investigate the small sample properties of the confidence intervals under model misspecification, we create our

random samples using the algorithm described in Appendix A, which generates autocorrelated Poisson-type random variates.

We estimate  $\theta = (\psi, \eta)^T$  with a GEE model using the independence working correlation matrix, essentially fitting a simple linear Poisson regression using the  $x_{ij}$  as covariates. We then use the estimate of  $\theta$  to form the 95% generalized Wald, the generalized Wald- $t$ , generalized score, and robust profile likelihood confidence intervals for  $\psi$ . The results of these simulations are summarized in Tables 1.6 and 1.7 below.

Table 1.6: Coverage and length for generalized Wald, score, and robust likelihood intervals for serially correlated Poisson data under a log linear regression model.

|                |          |                     | $\rho = 0$ |     |     | $\rho = .2$ |     |     | $\rho = .5$ |     |     |
|----------------|----------|---------------------|------------|-----|-----|-------------|-----|-----|-------------|-----|-----|
| $(\psi, \eta)$ |          | # of Clusters $K =$ | 10         | 20  | 50  | 10          | 20  | 50  | 10          | 20  | 50  |
| (0,0)          | Coverage | Gen. Wald           | .90        | .92 | .93 | .89         | .92 | .95 | .88         | .91 | .94 |
|                |          | Gen. Wald- $t$      | .93        | .94 | .94 | .93         | .94 | .95 | .92         | .93 | .94 |
|                |          | Gen. Score          | .98        | .96 | .95 | .97         | .96 | .96 | .96         | .95 | .96 |
|                |          | Rob. Like.          | .96        | .96 | .95 | .95         | .95 | .96 | .92         | .94 | .94 |
|                | Median   | Gen. Wald           | .98        | .72 | .47 | 1.0         | .77 | .50 | 1.1         | .86 | .56 |
|                | Length   | Gen. Wald- $t$      | 1.1        | .77 | .48 | 1.2         | .83 | .52 | 1.3         | .92 | .58 |
|                |          | Gen. Score          | 1.5        | .86 | .50 | 1.8         | .94 | .54 | 2.1         | 1.1 | .61 |
|                |          | Rob. Like.          | 1.1        | .79 | .50 | 1.2         | .82 | .51 | 1.2         | .88 | .57 |
|                | Coverage | Gen. Wald           | .91        | .93 | .94 | .89         | .92 | .94 | .88         | .91 | .93 |
|                |          | Gen. Wald- $t$      | .93        | .93 | .94 | .93         | .93 | .94 | .92         | .93 | .93 |
|                |          | Gen. Score          | .96        | .95 | .95 | .96         | .95 | .95 | .96         | .95 | .94 |
|                |          | Rob. Like.          | .97        | .96 | .96 | .94         | .94 | .95 | .92         | .94 | .94 |
|                | Median   | Gen. Wald           | .99        | .73 | .48 | 1.1         | .78 | .51 | 1.2         | .88 | .57 |
|                | Length   | Gen. Wald- $t$      | 1.2        | .79 | .49 | 1.2         | .84 | .53 | 1.4         | .94 | .59 |
|                |          | Gen. Score          | 1.5        | .87 | .51 | 1.7         | .94 | .55 | 2.1         | 1.1 | .63 |
|                |          | Rob. Like.          | 1.2        | .80 | .50 | 1.2         | .83 | .52 | 1.2         | .90 | .58 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Cluster size=4.

Table 1.7: Proportion of infinite length generalized score and robust likelihood intervals for serially correlated Poisson data under a log linear regression model.

|                |            |                     | $\rho = 0$ |    |    | $\rho = .2$ |    |    | $\rho = .5$ |    |    |
|----------------|------------|---------------------|------------|----|----|-------------|----|----|-------------|----|----|
| $(\psi, \eta)$ |            | # of Clusters $K =$ | 10         | 20 | 50 | 10          | 20 | 50 | 10          | 20 | 50 |
| (0,0)          | Proportion | Gen. Score          | .01        | 0  | 0  | .01         | 0  | 0  | .06         | 0  | 0  |
|                | Infinite   | Rob. Like.          | 0          | 0  | 0  | 0           | 0  | 0  | 0           | 0  | 0  |
| (1,0)          | Proportion | Gen. Score          | 0          | 0  | 0  | .03         | 0  | 0  | .05         | 0  | 0  |
|                | Infinite   | Rob. Like.          | 0          | 0  | 0  | 0           | 0  | 0  | 0           | 0  | 0  |

The results of these simulations are quite clear. The generalized Wald interval, using normal critical values performs quite poorly when the sample size is small, and is improved significantly by the use of  $t$ -critical values. In this set of simulations, the robust profile interval performs reasonably well, except when the sample size is small and the autocorrelation high. The generalized score interval consistently has good coverage as well, but is somewhat inefficient relative to the robust profile likelihood, and is occasionally infinite when the sample size is  $K = 10$ , while no infinite length robust profile intervals were observed.

### Epilepsy Data Example and Simulations

The data in the following example comes from a study of the effect of the anti-seizure drug, progabide, on epileptics (Thall and Vail 1990). In this experiment 59 patients were randomized into treatment and placebo (standard chemotherapy) groups. A baseline two-week count of seizures was recorded for the period before treatment was begun, and four other seizure counts were recorded for each patient at intervals of two weeks over the course of the eight-week study. Parameter estimates were obtained by fitting a Poisson regression model using treatment (TREATMENT), age (AGE), and an indicator of baseline vs. study period seizure count as covariates (BASELINE). These parameter estimates, and their corresponding 95% confidence

intervals are given in Table 1.8.

Table 1.8: Poisson regression MLE's and approximate 95% GEE confidence intervals for the Epilepsy dataset.

| Parameter | MLE  | Gen. Wald   | Gen. Wald- <i>t</i> | Gen. Score  | Rob. Like   |
|-----------|------|-------------|---------------------|-------------|-------------|
| TREATMENT | -0.8 | (-.76,.59)  | (-.78,.60)          | (-1.0,.58)  | (-.77,.60)  |
| AGE       | -.44 | (-1.4,.49)  | (-1.4,.51)          | (-1.5,.86)  | (-1.4,.46)  |
| BASELINE  | -1.3 | (-1.6,-.97) | (-1.6,-.97)         | (-1.7,-.99) | (-1.6,-.98) |

In the following simulations, we used a method analogous to the simple linear Poisson GEE simulations, using the same method to generate counts that are serially correlated within a patient. The simulated data were generated using the correlated parameter values equal to the study data's MLE. In each simulation, 2000 simulation replications of the data were created. Two sample sizes were used, the full set of 59 observations, and a subset of 30 patients with 15 observations from each treatment group. Table 1.9 summarizes the results of these simulations.

Table 1.9: Coverage and length for generalized Wald, score, and robust likelihood intervals for serially correlated Poisson data using the Epilepsy data covariates and MLE.

| Parameter | # of Clusters $K =$ |                | $\rho = 0$ |     | $\rho = .2$ |     | $\rho = .5$ |     |
|-----------|---------------------|----------------|------------|-----|-------------|-----|-------------|-----|
|           |                     |                | 30         | 59  | 30          | 59  | 30          | 59  |
| TREATMENT | Coverage            | Gen. Wald      | .93        | .93 | .91         | .93 | .91         | .92 |
|           |                     | Gen. Wald- $t$ | .94        | .93 | .92         | .93 | .92         | .93 |
|           |                     | Gen. Score     | .94        | .94 | .95         | .94 | .95         | .94 |
|           |                     | Rob. Like.     | .96        | .95 | .92         | .93 | .91         | .92 |
|           | Median              | Gen. Wald      | .17        | .12 | .68         | .51 | 1.3         | .98 |
|           |                     | Gen. Wald- $t$ | .18        | .13 | .72         | .52 | 1.4         | .99 |
|           |                     | Gen. Score     | .18        | .13 | .76         | .54 | 1.6         | 1.1 |
|           |                     | Rob. Like.     | .18        | .13 | .69         | .51 | 1.3         | .98 |
|           | Length              | Gen. Wald      | .91        | .93 | .90         | .92 | .90         | .92 |
|           |                     | Gen. Wald- $t$ | .92        | .93 | .92         | .93 | .90         | .93 |
|           |                     | Gen. Score     | .95        | .95 | .95         | .94 | .96         | .95 |
|           |                     | Rob. Like.     | .96        | .95 | .91         | .92 | .90         | .92 |
| AGE       | Coverage            | Gen. Wald      | .91        | .93 | .90         | .92 | .90         | .92 |
|           |                     | Gen. Wald- $t$ | .92        | .93 | .92         | .93 | .90         | .93 |
|           |                     | Gen. Score     | .95        | .95 | .95         | .94 | .96         | .95 |
|           |                     | Rob. Like.     | .96        | .95 | .91         | .92 | .90         | .92 |
|           | Median              | Gen. Wald      | .27        | .24 | 1.1         | .93 | 2.0         | 1.7 |
|           |                     | Gen. Wald- $t$ | .29        | .24 | 1.1         | .95 | 2.1         | 1.8 |
|           |                     | Gen. Score     | .39        | .27 | 1.6         | 1.1 | 3.2         | 2.1 |
|           |                     | Rob. Like.     | .32        | .25 | 1.1         | .93 | 2.0         | 1.8 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Cluster size=5. Less than .005 of each of the intervals were reported as infinite.

The simulation results are consistent with the simple linear Poisson GEE simulations.

The generalized score interval maintains its coverage better than the other intervals, especially as the within cluster correlation becomes larger. No infinite length generalized score intervals were observed. These observations lead us to recommend that the generalized score confidence interval be used in the analysis of correlated Poisson-type data.

### 1.7.3 Huber Robust Linear Regression

We investigate the performance of these confidence intervals for robust linear regression models using the Huber  $\psi$ -function approach (Huber 1981). The model for the data is  $y_i =$

$\mathbf{x}_i^T \beta + \epsilon_i$ , where the  $\epsilon$ 's are iid with median 0, but are not assumed to be normally distributed.

$\beta = (\beta_1, \beta_2)$  has dimension  $p$ , and  $\beta_1$  is the scalar parameter of interest. Parameter estimation of the mean parameters,  $\beta$ , proceeds by solving,

$$\sum_{i=1}^n \psi \left( \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\sigma} \right) \mathbf{x}_i = 0, \quad (1.25)$$

where Huber's  $\psi$ -function is given by,

$$\psi(z) = \begin{cases} z & \text{if } |z| < k \\ k \cdot \text{sign}(z) & \text{otherwise.} \end{cases} \quad (1.26)$$

For all of the simulations in this study, we use  $k = 1$ . Solving the above estimating equation is equivalent to maximum likelihood estimation, where the density of the errors has the functional form  $\exp(-\rho(z)/\sigma)$ , where  $\rho(z)$  is defined to be,

$$\rho(z) = \begin{cases} \frac{z^2}{2} & \text{if } |z| < k \\ k|z| - \frac{k^2}{2} & \text{otherwise.} \end{cases} \quad (1.27)$$

We estimate the scale parameter,  $\sigma$ , by solving

$$\sum_{i=1}^n \chi \left( \frac{y_i - \mathbf{x}_i^T \beta}{\hat{\sigma}} \right) - (n - p) E_{\Phi}(\chi) = 0 \quad (1.28)$$

where  $\chi(z) = z\psi(z) - \rho(z)$ , and  $E_{\Phi}$  represents expectation with respect to the standard normal distribution.

Unlike the previous simulations, we do not compute a restricted estimate of  $\sigma$  while computing the robust likelihood and generalized score confidence intervals, because preliminary

simulations indicated that these confidence intervals suffered a loss of efficiency when this was done. However, restricted estimates of  $\beta_2$  are computed.

Robust likelihood ratio tests of ANOVA models using Huber's  $\psi$ -function were investigated by (Schrader and Hettmansperger 1980). In their paper, Schrader and Hettmansperger develop a modeling strategy completely analogous to standard least squares ANOVA using robust likelihood ratio tests. Although they describe a complete system for likelihood testing of robust linear models, I have not seen their approach generalized to interval estimation.

Their approach differs, somewhat, from the one we used for GEE models, however. Rather than use the purely empirical estimators for  $I$  and  $D$ ,

$$\frac{\hat{\sigma}}{n} \sum_{i=1}^n \psi' \left( \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}} \right) \mathbf{x}_i \mathbf{x}_i^T,$$

and

$$\frac{\hat{\sigma}}{n} \sum_{i=1}^n \psi^2 \left( \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}} \right) \mathbf{x}_i \mathbf{x}_i^T,$$

respectively, they use estimates of  $I$  and  $D$  that assume that the data are identically distributed,

$$\frac{\hat{\sigma}}{n} \sum_{i=1}^n \psi' \left( \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}} \right) \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T,$$

and

$$\frac{\hat{\sigma}}{n} \sum_{i=1}^n \psi^2 \left( \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}} \right) \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T,$$

respectively, where  $\psi'(z) = \partial\psi(z)/\partial z$ . Notice that these estimators for  $I$  and  $D$  imply that a

consistent estimator for the likelihood ratio test's robust correction constant is

$$\hat{c}_\psi = \xi \frac{n}{n-p} \frac{\sum_{i=1}^n \psi^2 \left( \frac{y(i) - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}} \right)}{\sum_{i=1}^n \psi' \left( \frac{y(i) - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}} \right)},$$

where  $\xi$  is a bias correction that is given by,

$$\xi = 1 + \frac{p}{n} \frac{n - \sum_{i=1}^n \psi' \left( \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}} \right)}{\sum_{i=1}^n \psi' \left( \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}} \right)}.$$

Taking the analogy with ANOVA further, they use the  $F$  distribution as the reference distribution for the robust likelihood ratio test instead of the  $\chi^2$  distribution.

Schrader and Hettmansperger also investigate a bias corrected version of the generalized Wald test statistic that, like their version of the robust likelihood ratio test, assumes that the data are iid. Inverting this generalized Wald test yields a bias corrected version of the generalized Wald interval,

$$\hat{\beta}_1 \pm t_{n-p}(1 - \alpha/2) \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{n}{n-p} \xi \frac{\sum_{i=1}^n \psi^2 \left( \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}} \right)}{\left( \sum_{i=1}^n \psi' \left( \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}} \right) \right)^2} (\mathbf{X}^T \mathbf{X})_{1,1}^{-1}}.$$

Like their robust likelihood ratio test, Schrader and Hettmansperger use a  $t$  distribution to obtain critical values for the Wald test.

In each of the simulations below, we report coverages and median lengths for 5 types of confidence intervals: the ordinary least squares  $t$  confidence interval (OLS  $t$ ), the Schrader-Hettmansperger generalized Wald interval (SH-Gen. Wald) and the Schrader Hettmanperger robust likelihood intervals (SH-Rob. Like) described above, as well as the 'standard' generalized

score and robust profile likelihood intervals.

In each simulation we generate  $N = 2000$  datasets from the model  $y_i = \mu + \beta x_i + e_i$ ,  $i = 1, 2, \dots, n$ , where  $n$  is equal to 10, 20, or 50, and the  $x_i$  are equally spaced in the interval  $(-1, 1)$ . In the first group of simulations  $\beta = \mu = 0$ , and the  $e_i$ 's are iid, but are generated from one of three distributions: *Normal*(0, 1), *Cauchy*, or a mixture distribution where each  $e_i$  has a 75% probability of coming from a *Normal*(0, 1) distribution, and a 25% probability of being a *Cauchy* random variate. In this way, we can see the loss of efficiency when the data are normally distributed, and compare the performance of the various confidence intervals when the data are heavy tailed, or prone to outliers.

For the second group of simulations  $\beta = \mu = 1$ , and the  $e_i$ 's are independent, but no longer identically distributed. This allows us to determine whether or not the performance of the Schrader-Hettmansperger intervals is sensitive to the assumption that the errors are iid. Here the  $e_i$  are distributed as either a *Normal*(0, 1), a *Cauchy*, or a 75%-25% mixture of the two, multiplied by a scale parameter,  $\sigma_i$ , that varies with the  $x_i$ 's according to the rule  $\sigma_i = .1 + |\mu + x_i\beta|$ . Tables 1.10 and 1.11 show the results of these simulations.

As measured by coverage, all the confidence intervals perform quite well, with coverages often at or exceeding 95% in most of the simulations. We are therefore compelled to make our recommendation based on the lengths of the intervals. As one would expect, the OLS  $t$  interval performs the best when the error are iid from a *Normal*(0, 1) distribution, but is inefficient when the distribution of the errors has heavy tails. Like the previous simulations, the generalized score interval is often inefficient when the sample size is small, and is occasionally infinite in length. The robust profile likelihood interval (the general version that does not assume that the errors are iid) is also conservative, with coverage rates often exceeding 95%.

The Schrader and Hettmansperger (SH) version of the robust likelihood interval is the most efficient, and maintains its coverage well, even when the errors are not iid. The SH-robust profile interval is generally shorter than the SH-generalized Wald interval, and is the one we choose to recommend for use in practice.

Table 1.10: Coverage and length for generalized Wald, score, and robust likelihood intervals for iid data under a simple linear Huber regression model.

|                            |          | Dist.         | <i>Normal</i> |     |     | <i>Cauchy</i> |     |     | <i>Mixture</i> |     |     |
|----------------------------|----------|---------------|---------------|-----|-----|---------------|-----|-----|----------------|-----|-----|
| $(\mu, \beta, \sigma)$     |          | # Obs. $K =$  | 5             | 10  | 20  | 5             | 10  | 20  | 5              | 10  | 20  |
| (0,0,1)                    | Coverage | OLS $t$       | .94           | .94 | .95 | .97           | .96 | .96 | .96            | .97 | .96 |
|                            |          | SH-Gen. Wald  | .93           | .94 | .95 | .96           | .96 | .95 | .96            | .95 | .95 |
|                            |          | SH-Rob. Like. | .93           | .94 | .95 | .96           | .95 | .95 | .96            | .96 | .95 |
|                            |          | Gen. Score    | .95           | .95 | .95 | .97           | .95 | .95 | .96            | .96 | .95 |
|                            |          | Rob. Like.    | .95           | .96 | .98 | .97           | .97 | .97 | .97            | .95 | .97 |
|                            | Median   | OLS- $t$      | 2.1           | 1.5 | .94 | 8.1           | 7.8 | 7.9 | 6.2            | 6.1 | 5.7 |
|                            | Length   | SH-Gen. Wald  | 2.3           | 1.6 | 1.0 | 4.7           | 3.0 | 1.7 | 3.6            | 2.3 | 1.4 |
|                            |          | SH-Rob. Like. | 2.2           | 1.5 | .99 | 4.6           | 2.9 | 1.7 | 3.5            | 2.2 | 1.3 |
|                            |          | Gen. Score    | 2.7           | 1.6 | 1.0 | 8.4           | 3.6 | 1.8 | 6.6            | 2.8 | 1.4 |
|                            |          | Rob. Like.    | 2.4           | 1.8 | 1.1 | 4.9           | 3.2 | 1.9 | 3.8            | 2.5 | 1.5 |
| (1,1, $\sigma_i^\dagger$ ) | Coverage | OLS $t$       | .95           | .94 | .95 | .97           | .97 | .97 | .96            | .96 | .97 |
|                            |          | SH-Gen. Wald  | .95           | .95 | .95 | .97           | .95 | .95 | .95            | .95 | .94 |
|                            |          | SH-Rob. Like. | .94           | .95 | .95 | .96           | .95 | .95 | .95            | .95 | .94 |
|                            |          | Gen. Score    | .95           | .94 | .95 | .96           | .94 | .94 | .95            | .97 | .95 |
|                            |          | Rob. Like.    | .95           | .96 | .97 | .97           | .97 | .97 | .95            | .94 | .97 |
|                            | Median   | OLS- $t$      | 2.3           | 1.6 | 1.0 | 8.1           | 8.3 | 8.4 | 2.9            | 2.4 | 2.2 |
|                            | Length   | SH-Gen. Wald  | 2.3           | 1.6 | 1.0 | 4.6           | 2.9 | 1.7 | 2.5            | 1.7 | 1.0 |
|                            |          | SH-Rob. Like. | 2.2           | 1.6 | .98 | 4.5           | 2.8 | 1.7 | 2.4            | 1.6 | 1.0 |
|                            |          | Gen. Score    | 2.2           | 1.7 | 1.0 | 8.1           | 3.5 | 1.8 | 3.4            | 1.8 | 1.1 |
|                            |          | Rob. Like.    | 2.4           | 1.7 | 1.1 | 4.6           | 3.1 | 1.9 | 2.6            | 1.8 | 1.1 |

Note: Estimated coverages have approximate standard error .005.

$^\dagger \sigma_i = .1 + |\mu + x_i|$ .

Table 1.11: Proportion of infinite length generalized score and robust likelihood intervals for iid data under a simple linear Huber regression model.

|                            |            | Dist.          | <i>Normal</i> |    |    | <i>Cauchy</i> |     |    | <i>Mixture</i> |     |    |
|----------------------------|------------|----------------|---------------|----|----|---------------|-----|----|----------------|-----|----|
| $(\mu, \beta, \sigma)$     |            | # of Obs $K =$ | 10            | 20 | 50 | 10            | 20  | 50 | 10             | 20  | 50 |
| (0,0,1)                    | Proportion | Gen. Score     | .01           | 0  | 0  | .12           | .01 | 0  | .12            | 0   | 0  |
| (1,1, $\sigma_i^\dagger$ ) | Infinite   | Gen. Score     | 0             | 0  | 0  | .12           | 0   | 0  | .06            | .01 | 0  |

### 1.7.4 Simple Linear Regression With Covariate Measurement Error

We investigate the performance of the generalized Wald, generalized score, and robust profile likelihood confidence intervals in the context of simple linear measurement error models. In these models we assume that the data consist of pairs  $(y, u)$  where  $y$  is the independent random variable of interest, and  $u$  plays the role of the covariate. In addition, it is assumed, however, that there is an unobserved random variable,  $x$ , such that  $x = \mu_x + \epsilon_x$  where  $\epsilon_x \sim \text{Normal}(0, \sigma_x^2)$ . The observed variables,  $y$  and  $u$ , are assumed to have model,  $y = \mu + \beta x + \epsilon$  and  $u = x + \epsilon_u$ , where  $\epsilon$  and  $\epsilon_u$  are independent of one another, as well as with  $\epsilon_x$ , and normally distributed, each with mean 0, and variances  $\sigma^2$  and  $\sigma_u^2$ , respectively. Unfortunately this model is not completely identified, which means that one cannot estimate all six parameters,  $(\mu, \beta, \mu_x, \sigma^2, \sigma_x^2, \sigma_u^2)$ . To circumvent this difficulty we assume that the measurement error variance,  $\sigma_u^2$ , is known *a priori*. We form the likelihood-based on the model described above and use it to find global and restricted estimates of the parameters.

To examine the effects of model misspecification we examine four situations. In each of these the identified scale parameters satisfy  $\sigma^2 = \sigma_x^2 = 1$ , and the distributions of  $\epsilon$  and  $\epsilon_x$  are either  $\text{Normal}(0, 1)$ , or scaled  $t_5$  so that the error variance is 1. The mean model parameters have true values,  $\mu = \mu_x = 0$ , and  $\beta = 1$ . The measurement error variance,  $\sigma_u^2$  is varied, taking on values .1, .25, and .5. We provide simulation results for all four possible combinations of  $N(0, 1)$  and  $t_5$  distributions for  $\epsilon$  and  $\epsilon_x$ . Three sample sizes are investigated,  $n = 20, 50$ , and 100. In each simulation,  $N_S=2000$  independent samples are drawn, and for each the 95% generalized Wald, generalized score and robust profile likelihood confidence intervals for the slope parameter,  $\beta$  are formed. In agreement with standard practice, the generalized Wald

interval is created using  $t$  critical values with  $n - 2$  degrees of freedom, but the generalized score and robust profile likelihood confidence intervals simply use  $\chi_1^2$  critical values. Tables 1.12, 1.13, and 1.14 below contain the results of these simulations.

Table 1.12: Coverage and length for generalized Wald, score, and robust likelihood intervals from simple linear measurement error model.

| Distribution of          |          |                 | $\sigma_u^2 = .1$ |     |     | $\sigma_u^2 = 0.25$ |     |     | $\sigma_u^2 = 0.5$ |     |     |
|--------------------------|----------|-----------------|-------------------|-----|-----|---------------------|-----|-----|--------------------|-----|-----|
| $(\epsilon_x, \epsilon)$ |          | # of Obs. $n =$ | 20                | 50  | 100 | 20                  | 50  | 100 | 20                 | 50  | 100 |
| $N(0, 1), N(0, 1)$       | Coverage | Gen. Wald       | .88               | .92 | .94 | .90                 | .94 | .95 | .93                | .94 | .94 |
|                          |          | Gen. Score      | .95               | .94 | .95 | .94                 | .95 | .95 | .93                | .93 | .94 |
|                          |          | Rob. Like.      | .93               | .95 | .96 | .95                 | .96 | .96 | .95                | .95 | .95 |
|                          | Median   | Gen. Wald       | .88               | .59 | .42 | 1.0                 | .68 | .49 | 1.4                | .86 | .62 |
|                          |          | Gen. Score      | 1.5               | .68 | .46 | 2.2                 | .83 | .54 | 6.7                | 1.2 | .71 |
|                          |          | Rob. Like.      | 1.1               | .64 | .45 | 1.3                 | .76 | .52 | 2.0                | 1.0 | .67 |
|                          | Length   | Gen. Wald       | .88               | .59 | .42 | 1.0                 | .68 | .49 | 1.4                | .86 | .62 |
|                          |          | Gen. Score      | 1.5               | .68 | .46 | 2.2                 | .83 | .54 | 6.7                | 1.2 | .71 |
|                          |          | Rob. Like.      | 1.1               | .64 | .45 | 1.3                 | .76 | .52 | 2.0                | 1.0 | .67 |
| $N(0, 1), t_5$           | Coverage | Gen. Wald       | .88               | .91 | .93 | .91                 | .91 | .93 | .93                | .95 | .94 |
|                          |          | Gen. Score      | .96               | .96 | .95 | .95                 | .95 | .95 | .93                | .94 | .93 |
|                          |          | Rob. Like.      | .95               | .96 | .96 | .94                 | .95 | .96 | .95                | .97 | .96 |
|                          | Median   | Gen. Wald       | .88               | .56 | .41 | 1.1                 | .69 | .95 | 1.4                | .87 | .61 |
|                          |          | Gen. Score      | 1.4               | .66 | .45 | 2.1                 | .91 | .53 | 6.6                | 1.2 | .70 |
|                          |          | Rob. Like.      | 1.0               | .64 | .45 | 1.3                 | .74 | .52 | 2.0                | .98 | .66 |
|                          | Length   | Gen. Wald       | .88               | .56 | .41 | 1.1                 | .69 | .95 | 1.4                | .87 | .61 |
|                          |          | Gen. Score      | 1.4               | .66 | .45 | 2.1                 | .91 | .53 | 6.6                | 1.2 | .70 |
|                          |          | Rob. Like.      | 1.0               | .64 | .45 | 1.3                 | .74 | .52 | 2.0                | .98 | .66 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Here we see that the generalized Wald confidence interval is liberal, while the generalized score interval is quite inefficient, and in some circumstances it is frequently infinite in length. The robust profile likelihood confidence interval performs well, as measured its coverage and median length.

Table 1.13: Coverage and length for generalized Wald, score, and robust likelihood intervals from simple linear measurement error model.

| Distribution of          |          |                 | $\sigma_u^2 = .1$ |     |     | $\sigma_u^2 = 0.25$ |     |     | $\sigma_u^2 = 0.5$ |     |     |
|--------------------------|----------|-----------------|-------------------|-----|-----|---------------------|-----|-----|--------------------|-----|-----|
| $(\epsilon_x, \epsilon)$ |          | # of Obs. $n =$ | 20                | 50  | 100 | 20                  | 50  | 100 | 20                 | 50  | 100 |
| $t_5, N(0, 1)$           | Coverage | Gen. Wald       | .87               | .92 | .94 | .90                 | .92 | .94 | .92                | .93 | .94 |
|                          |          | Gen. Score      | .96               | .96 | .96 | .95                 | .94 | .95 | .94                | .94 | .93 |
|                          |          | Rob. Like.      | .94               | .95 | .96 | .95                 | .95 | .96 | .95                | .96 | .94 |
|                          | Length   | Gen. Wald       | .90               | .60 | .42 | 1.1                 | .70 | .49 | 1.5                | .91 | .63 |
|                          |          | Gen. Score      | 2.0               | .77 | .48 | 4.0                 | .97 | .58 | 8.2                | 1.5 | .77 |
|                          |          | Rob. Like.      | 1.1               | .67 | .46 | 1.4                 | .78 | .53 | 2.2                | 1.0 | .67 |
| $t_5, t_5$               | Coverage | Gen. Wald       | .88               | .91 | .93 | .91                 | .91 | .93 | .93                | .95 | .94 |
|                          |          | Gen. Score      | .96               | .95 | .95 | .95                 | .94 | .94 | .96                | .93 | .93 |
|                          |          | Rob. Like.      | .94               | .95 | .96 | .95                 | .95 | .95 | .95                | .95 | .94 |
|                          | Length   | Gen. Wald       | .88               | .56 | .41 | 1.1                 | .69 | .95 | 1.4                | .87 | .61 |
|                          |          | Gen. Score      | 1.9               | .75 | .47 | 4.0                 | .96 | .57 | 8.0                | 1.5 | .77 |
|                          |          | Rob. Like.      | 1.0               | .65 | .45 | 1.3                 | .78 | .53 | 2.0                | 1.0 | .67 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Table 1.14: Proportion of infinite length generalized score and robust likelihood intervals from simple linear measurement error model.

| Distribution of          |            |                 | $\sigma_u^2 = .1$ |     |     | $\sigma_u^2 = 0.25$ |     |     | $\sigma_u^2 = 0.5$ |     |     |
|--------------------------|------------|-----------------|-------------------|-----|-----|---------------------|-----|-----|--------------------|-----|-----|
| $(\epsilon_x, \epsilon)$ |            | # of Obs. $n =$ | 20                | 50  | 100 | 20                  | 50  | 100 | 20                 | 50  | 100 |
| $N(0, 1), N(0, 1)$       | Proportion | Gen. Score      | .05               | 0   | 0   | .06                 | 0   | 0   | .12                | 0   | 0   |
|                          | Infinite   | Rob. Like.      | 0                 | 0   | 0   | 0                   | 0   | 0   | .05                | 0   | 0   |
| $N(0, 1), t_5$           | Proportion | Gen. Score      | .04               | 0   | 0   | .07                 | 0   | 0   | .15                | 0   | 0   |
|                          | Infinite   | Rob. Like.      | 0                 | 0   | 0   | .01                 | 0   | 0   | .05                | 0   | 0   |
| $t_5, N(0, 1)$           | Proportion | Gen. Score      | .18               | .03 | .01 | .21                 | .02 | .01 | .22                | .05 | 0   |
|                          | Infinite   | Rob. Like.      | 0                 | 0   | 0   | .01                 | 0   | 0   | .09                | 0   | 0   |
| $t_5, t_5$               | Proportion | Gen. Score      | .17               | .02 | .01 | .20                 | .03 | .01 | .23                | .05 | .01 |
|                          | Infinite   | Rob. Like.      | 0                 | 0   | 0   | .01                 | 0   | 0   | .09                | 0   | 0   |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

## Chapter 2

# Length-Optimal Interval Estimation

In this chapter we introduce a general procedure for taking an existing confidence interval procedure and creating a new one. We call this new confidence interval procedure *length-optimal interval estimation*, as it finds the shortest asymptotically level- $\alpha$  confidence interval in an infinite class of asymptotically level- $\alpha$  confidence intervals. Some theory is developed for the case of the single parameter likelihood intervals that demonstrates that the length-optimal profile likelihood intervals are asymptotically equivalent to the more common equal-tail likelihood interval, indicating that the length-optimal intervals retain their asymptotic error rate of  $1 - \alpha$ . Simulation results are provided that show that in some circumstances the small sample coverage rate of these confidence intervals is higher than their standard counterparts.

### 2.1 Motivation and Description

In the previous chapter, we saw that the generalized score interval often has superior coverage when compared to the generalized Wald and robust profile likelihood intervals. How-

ever, the generalized score interval was observed to be inefficient, and in some circumstances suffered from the defect of being infinite in length a significant proportion of the time. Length-optimal confidence interval procedures were motivated by an attempt to correct these problems of the generalized score interval. As we shall see, the length-optimal approach is only modestly successful in reducing the conservatism of the generalized score interval, but when we apply the procedure to the robust profile likelihood and generalized score intervals in simulations we see the surprising result that the length-optimal versions of the intervals often enjoy higher coverage than their equal-tail counterparts.

Confidence intervals are generally formed by inverting a two-sided level- $\alpha$  hypothesis test. That is, the endpoints of the confidence interval are found by inverting two level- $\alpha/2$  one-sided hypothesis tests,  $H_0 : \psi \leq \psi_0$ , and  $H_0 : \psi \geq \psi_0$ . Since the error rate is allocated equally to each side, such intervals are sometimes called equal-tail confidence intervals. For a given test statistic,  $T(\psi_0)$ , with an asymptotic  $\chi_1^2$  distribution, the endpoints of the equal-tail confidence interval,  $(\psi_L, \psi_U)$ , are obtained by solving  $T(\psi_L) = z_{\alpha/2}^2$  and  $T(\psi_U) = z_{\alpha/2}^2$  subject to the constraint that  $\psi_L < \hat{\psi} < \psi_U$ . When one or more of these two equations has no solution, the confidence interval will have infinite length. This occurs when, for instance, for all  $\psi_0 < \hat{\psi}$  the value of the test statistic never exceeds  $z_{\alpha/2}^2$ , or, equivalently, the p-value of the hypothesis test never drops below  $\alpha/2$ .

To remedy this situation, we can reallocate the error rate of the confidence intervals by introducing a new parameter,  $\epsilon$ , that takes on values in  $(0, 1)$ . Rather than place an error rate of  $\alpha/2$  to each side of the interval, we can distribute  $\epsilon\alpha$  to the lower side of the interval, and  $(1 - \epsilon)\alpha$  to the upper side of the interval. We call these  $\epsilon$ -intervals, and their endpoints are found by solving  $T(\psi_L(\epsilon)) = z_{\epsilon\alpha}^2$  and  $T(\psi_U(\epsilon)) = z_{(1-\epsilon)\alpha}^2$ , again subject to the constraint

that  $\psi_L(\epsilon) < \hat{\psi} < \psi_U(\epsilon)$ . Note that the equal-tail interval is simply the  $\epsilon$ -interval with  $\epsilon = 1/2$ . In the case of an interval that is left-infinite, if one allows the error rate to increase on the lower side of the interval, the critical value that the test statistic must be solved for decreases, thereby making it easier for a solution to be found.

As a means of selecting which value of  $\epsilon$  to use, we allow the data to determine  $\epsilon$  by choosing the value of  $\epsilon$  that yields the minimum length  $\epsilon$  interval,  $\epsilon^*$ :

$$\epsilon^* = \arg \inf \{ \psi_U(\epsilon) - \psi_L(\epsilon) \}. \quad (2.1)$$

This  $\epsilon^*$ -interval,  $(\psi_L(\epsilon^*), \psi_U(\epsilon^*))$ , is the length-optimal confidence interval based on test statistic  $T$ . Although length-optimal intervals were motivated as attempt to remedy infinite length generalized score intervals, the procedure is general and can be applied to almost any interval procedure, regardless of whether or not the equal-tail interval is infinite.

The length-optimal generalized Wald confidence interval is simply the equal-tail interval. This is not difficult to see after an application of the first-derivative test on the length of the  $\epsilon$ -generalized Wald confidence intervals,  $v_\theta(z_{1-(1-\epsilon)\alpha} + z_{\epsilon\alpha})$ . In most cases, the endpoints of length-optimal versions of generalized score and robust profile likelihood intervals are not expressible in terms of elementary functions, and require the use of a nonlinear optimization algorithm to be obtained. Brent's derivative free minimization algorithm (Brent 1973) has proven to be an effective means of obtaining length-optimal confidence intervals, as it has guaranteed convergence under very mild assumptions, and it only requires one to have a routine that computes profile likelihood or score intervals given a particular value of  $\epsilon$ .

As an illustration of the length-optimal approach to interval estimation, consider the

following example. Figure 2.1 contains the treatment effect generalized score and generalized Wald confidence curves, plots of the p-values of the test statistics as a function of the null hypothesis value of  $\psi$  (Birnbaum 1961), for a subset of the Epilepsy study data described in the previous chapter. The subset consists of 33 of the total of 59 clusters, and is notable in that the generalized score confidence interval,  $(-\infty, .43)$ , for this subset of the data has an infinite lower tail. Intuitively, the 95% equal-tail confidence intervals for the respective tests are the regions where the plotted p-value remains above the .05 mark, which is drawn on the graph to facilitate visualization of confidence intervals. The interval endpoints, if they exist, are the points on the graph where the confidence curve crosses .05 on the  $y$  axis. We see that the left side of the generalized score test's confidence curve fails to fall below .05, and appears to flatten out just above the .05 level, providing a pictorial explanation of why the generalized score interval's lower endpoint is infinite. The generalized Wald confidence curve is much more like one would expect. It is symmetric about the MLE and crosses the .05 level at two places.

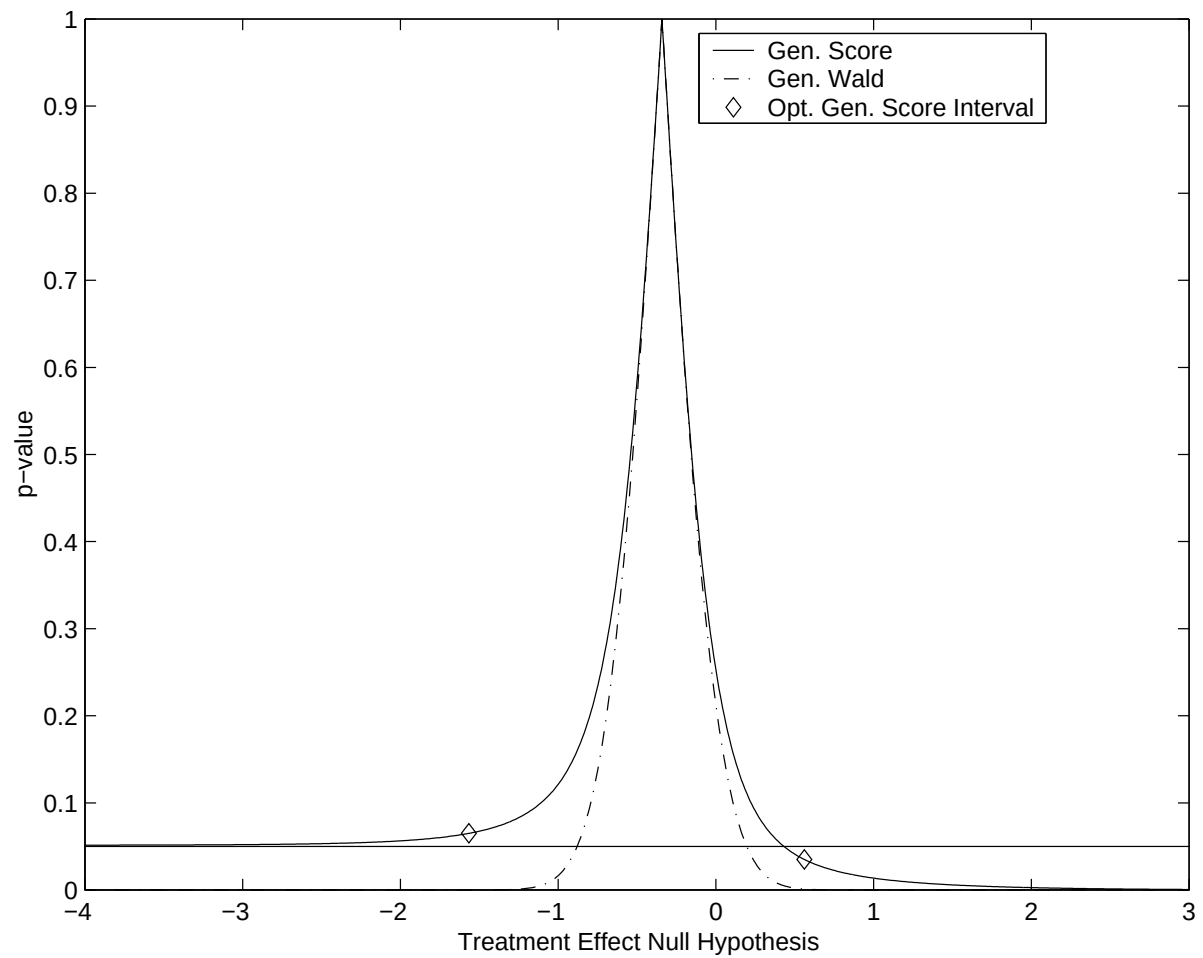


Figure 2.1: Confidence curve for a 33 cluster subset of the Epilepsy data. Solid line and dotted lines are the generalized score and Wald confidence curves, respectively. Length-optimal generalized score interval endpoints are marked with diamonds.

When we apply the length-optimal procedure to the treatment effect generalized score test, we find that the optimal length generalized score confidence interval is obtained at  $\epsilon^* = .65$ . This value of  $\epsilon^*$  shifts more of the error rate to left side of the interval, thereby lowering the critical value that the test statistic must be solved for on that side, and thus increasing the likelihood that a solution can be found. As the diamonds on the graph indicate, the overall effect is that both length-optimal endpoints are shifted to the right from the equal-tail generalized score confidence interval's endpoints. A similar, but less dramatic, effect is seen in length-optimal robust profile likelihood interval which has  $\epsilon^* = .57$ . The generalized score and robust profile likelihood intervals for this subset of the Epilepsy data, along with their length optimal versions and the generalized Wald interval, are given in Table 2.1.

Table 2.1: Length-optimal and equal-tail treatment effect 95% confidence intervals for 33 cluster subset of the Epilepsy data.

| Interval                  | equal-tail       | length-optimal |
|---------------------------|------------------|----------------|
| Generalized Score         | $(-\infty, .43)$ | $(-1.6, .56)$  |
| Robust Profile Likelihood | $(-.91, .17)$    | $(-.90, .18)$  |
| Generalized Wald          | $(-.87, .20)$    |                |

## 2.2 Example: Binomial Proportion

In this section we shift our attention, momentarily, away from model-robust confidence intervals, to the (non-robust) interval estimation of a binomial proportion. The length-optimal confidence interval estimation procedure is quite general, and not restricted to the interval estimation of parameters when the model is misspecified. The context of the binomial parameter is interesting because it is simple enough that we are able to compare the performance of competing confidence intervals exactly for all possible values of the parameter, and complex

enough that the results are interesting and nontrivial, unlike many other simple situations, such as linear least squares, in which case the length optimal intervals are the same as their equal-tail counterparts.

Interval estimation of a binomial proportion is, perhaps, the most deceptively simple problem in the field of statistics. Even quite recently, articles have been published on this and closely related topics, such as Agresti and Caffo (2000); Agresti and Coull (1998); Brown et al. (2002). Let  $p$  represent the *true* probability of success, and let  $y$  represent the number of successes out of  $n$  independent trials. Also, define  $\hat{p} = y/n$ , the proportion of successes in the data, to be the estimate of  $p$ . First-order asymptotic theory suggests that the Wald interval,

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (2.2)$$

would be adequate for moderately large sample sizes. However, for  $p = .01$  the smallest sample size for which the Wald interval's coverage exceeds .94 for all  $N \geq n$  is  $n = 7963$ , and for  $p = .1$  the smallest sample size for this condition to be obtained is  $n = 646$  (Agresti and Caffo 2000). The Wald interval's coverage rate converges to .95 incredibly slowly for values of  $p$  near 0 or 1.

Recently, a new addition to the plethora of confidence intervals for the binomial proportion was given in the literature (Agresti and Coull 1998) that is both very simple to compute and has good coverage properties. This simple interval, which we will refer to as the Agresti-Coull confidence interval, is motivated as an approximation to the score interval, but has the simple form of the Wald interval. To compute the Agresti-Coull interval, one computes the Wald interval for a data set having two more successes and failures than the observed sample.

That is, defining  $\tilde{p} = (y + 2)/(n + 4)$  and  $\tilde{n} = n + 4$ , the Agresti-Coull interval is simply,

$$\tilde{p} \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}}. \quad (2.3)$$

We apply the length-optimal procedure to the profile likelihood interval, and show that one version of it performs clearly better than the equal-tail version, and is arguably better than the Agresti-Coull interval. This example illustrates the importance of the choice of scale to the performance of the length-optimal approach, highlighting its lack of invariance to nonlinear parameter transformations. Length-optimal confidence interval procedures can improve coverage when the length is minimized on one scale, but coverage can deteriorate on another. In this example, we look at length-optimal likelihood interval estimation of the binomial proportion parameter where the length is minimized on two scales, the natural binomial proportion scale, and on the log odds scale,  $\theta = \log(p/(1 - p))$ .

The equal-tail profile likelihood interval's endpoints are defined to be the two solutions,  $(p_L, p_U)$  to the equation,

$$2 \left[ y \log \left( \frac{\hat{p}}{1 - \hat{p}} \right) + y \log \left( \frac{1 - p_0}{p_0} \right) + n \log \left( \frac{1 - \hat{p}}{1 - p_0} \right) \right] - z_{\alpha/2}^2 = 0. \quad (2.4)$$

The length-optimal likelihood interval on the  $p$ -scale is the interval that minimizes  $p_U(\epsilon) - p_L(\epsilon)$ , where

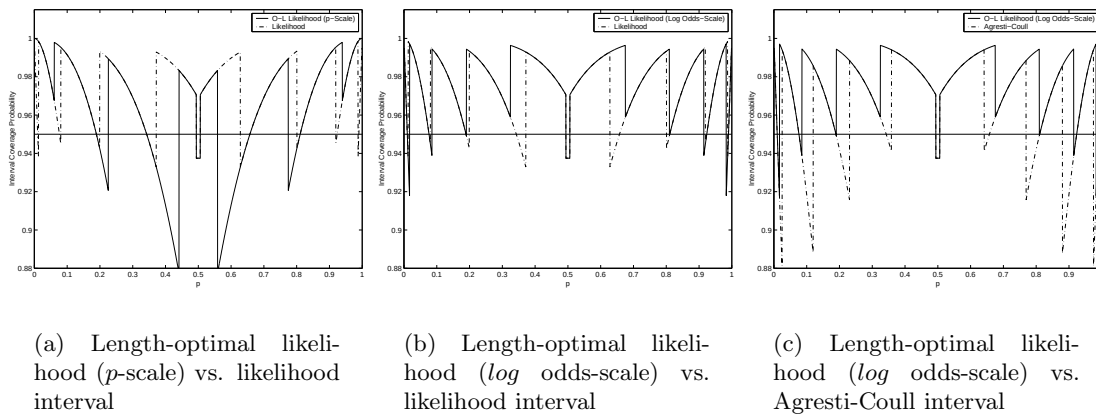
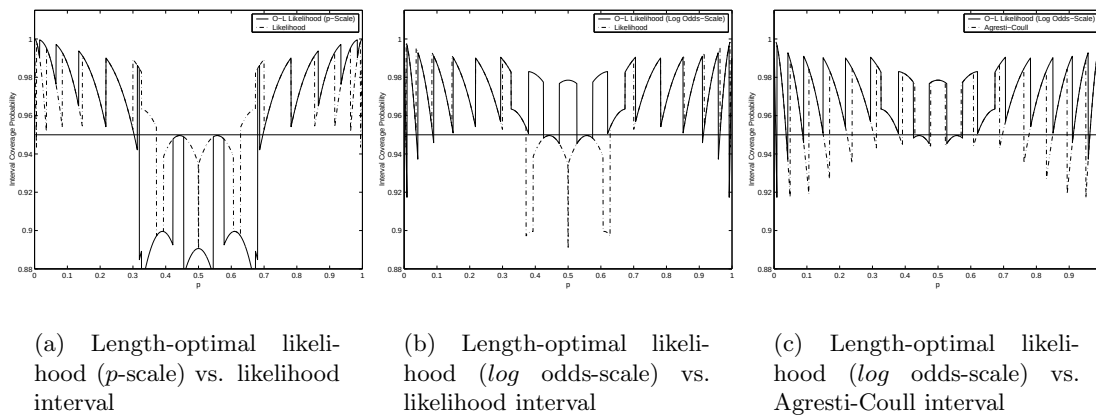
$$\begin{aligned} 2 \left[ y \log \left( \frac{\hat{p}}{1 - \hat{p}} \right) + y \log \left( \frac{1 - p_L(\epsilon)}{p_L(\epsilon)} \right) + n \log \left( \frac{1 - \hat{p}}{1 - p_L(\epsilon)} \right) \right] - z_{\epsilon\alpha}^2 &= 0 \\ 2 \left[ y \log \left( \frac{\hat{p}}{1 - \hat{p}} \right) + y \log \left( \frac{1 - p_U(\epsilon)}{p_U(\epsilon)} \right) + n \log \left( \frac{1 - \hat{p}}{1 - p_U(\epsilon)} \right) \right] - z_{(1-\epsilon)\alpha}^2 &= 0. \end{aligned} \quad (2.5)$$

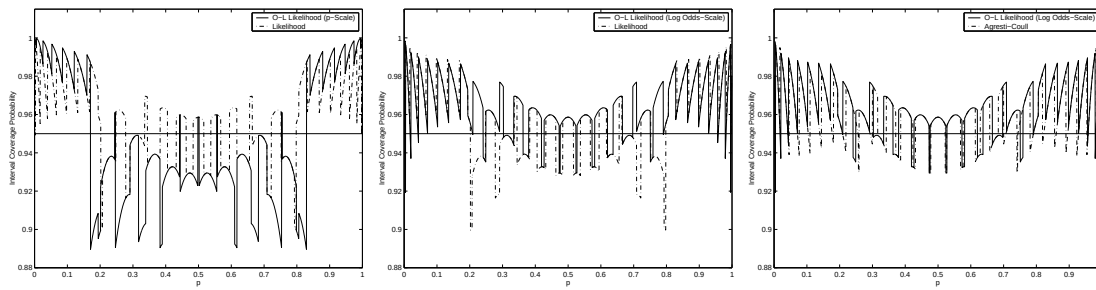
and the log odds-scale length-optimal interval minimizes the the length of the interval on the log-odds scale,

$$\log(p_U(\epsilon)/(1 - p_U(\epsilon)) - \log(p_L(\epsilon)/(1 - p_L(\epsilon))), \quad (2.6)$$

which is possible because, for fixed values of  $\epsilon$ ,  $\epsilon$ -likelihood intervals inherit the invariance of the likelihood ratio test.

In Figures 2.2, 2.3, and 2.4, we see plots comparing the coverages of confidence intervals as a function of the true value of  $p$  for three sample sizes,  $n = 5, 10, 20$ . In the leftmost plots we see the coverages of the  $p$ -scale length-optimal likelihood confidence interval and the equal-tail interval. For all three sample sizes, we see that the  $p$ -scale length-optimal interval's coverage probability is lower than that of the equal-tail version. The middle plots show that the coverage of the log odds scale length-optimal likelihood interval is almost uniformly superior to the equal-tail likelihood interval. The plots on the right compare the Agresti-Coull interval to the log odds scale length-optimal likelihood interval. When the sample size is  $n = 20$ , the two intervals' performance is quite similar. However, at the smaller sample sizes,  $n = 5$  and  $10$ , we see that the length-optimal interval's coverage is below .95 less frequently than the Agresti-Coull interval, and almost uniformly enjoys higher coverage. Although it is somewhat disappointing that the length-optimal interval's coverage is rather dependent on the parameterization, it is encouraging that the length-optimal interval performs well on the log odds scale, as this is the scale on which logistic regression parameters are defined, and where further generalization is possible.

Figure 2.2: Coverage plots for binomial proportion  $n=5$ Figure 2.3: Coverage plots for binomial proportion  $n=10$



(a) Length-optimal likelihood ( $p$ -scale) vs. likelihood interval

(b) Length-optimal likelihood ( $\log$  odds-scale) vs. likelihood interval

(c) Length-optimal ( $\log$  odds-scale) likelihood vs. Agresti-Coull interval

Figure 2.4: Coverage plots for binomial proportion  $n=20$

## 2.3 Theory

We now develop theoretical results for the length-optimal univariate likelihood ratio intervals, which indicate that they are asymptotically equivalent to their equal-tail counterparts. The proofs of more general results for the multiparameter and robust likelihood ratio follow essentially the same outline as the proofs contained herein, but are notationally far more tedious and complicated.

**Definitions** Let  $\mathbf{Y} = \{y_i\}_{i=1}^n$  be a random sample from a distribution with pdf  $f(y|\theta_0)$ , where  $f$  is a member of a class of distributions indexed by a univariate parameter  $\theta$ , and  $\theta_0$  is the value of  $\theta$  from which  $\mathbf{Y}$  was generated. Define  $l_n(\theta) = \sum_{i=1}^n \log f(y_i|\theta)$ . Let  $\hat{\theta}_n$  be the maximum likelihood estimator of  $\theta$ . Let  $v_\theta = -1/E_\theta(\partial^2/\partial\theta^2 \log f(y|\theta))$  be the normalized asymptotic variance of the MLE. Define  $\hat{\theta}_n^U(\alpha)$  be the upper level- $\alpha$  endpoint of the likelihood interval for  $\theta$ . That is,  $\hat{\theta}_n^U(\alpha)$  solves:

$$2(l(\hat{\theta}_n) - l(\hat{\theta}_n^U(\alpha))) - z_{1-\alpha/2}^2 = 0,$$

subject to the constraint that  $\hat{\theta}_n^U(\alpha) \geq \hat{\theta}$ . Similarly, let  $\hat{\theta}_n^L(\alpha)$  to be the lower level- $\alpha$  endpoint of the likelihood interval for  $\theta$ , which is defined to satisfy,

$$2(l(\hat{\theta}_n) - l(\hat{\theta}_n^L(\alpha))) - z_{1-\alpha/2}^2 = 0$$

subject to  $\hat{\theta}_n^L(\alpha) \leq \hat{\theta}$ .

As described in section 2.1, we generalize the definition of the likelihood interval by allowing the nominal miss rate to differ on the two sides of the interval, creating an infinite class

of likelihood based intervals indexed by a parameter,  $\epsilon \in [0, 1]$ , that indicates the proportion of  $\alpha$  partitioned to the left side of the interval. More specifically, for a particular value of  $\epsilon$  let the  $\epsilon$ -likelihood profile interval of level- $\alpha$  be the confidence interval whose endpoints,  $(\hat{\theta}_n^L(\alpha, \epsilon), \hat{\theta}_n^U(\alpha, \epsilon))$ , solve

$$2(l(\hat{\theta}_n) - l(\hat{\theta}_n^U(\alpha, \epsilon))) - z_{1-(1-\epsilon)\alpha}^2 = 0, \quad (2.7)$$

and

$$2(l(\hat{\theta}_n) - l(\hat{\theta}_n^L(\alpha, \epsilon))) - z_{1-\epsilon\alpha}^2 = 0. \quad (2.8)$$

The more common, equal-tail, likelihood confidence interval corresponds to the  $\epsilon$ -interval with  $\epsilon = 1/2$ . For fixed values of  $\epsilon$ , these confidence intervals have asymptotic coverage  $1 - \alpha$  with the proportion of misses on the left being  $\epsilon\alpha$ , and with  $(1 - \epsilon)\alpha$  being the asymptotic proportion of misses on the right.

The length-optimal likelihood interval is the  $\epsilon_n^*$ -likelihood interval where  $\epsilon_n^*$  is the value of  $\epsilon$  that minimizes the length of the  $\epsilon$ -likelihood intervals,

$$\hat{\theta}_n^U(\alpha, \epsilon_n^*) - \hat{\theta}_n^L(\alpha, \epsilon_n^*). \quad (2.9)$$

In order for this interval to be considered valid for use in practical data analysis situations, it should be shown that the interval has asymptotic coverage  $1 - \alpha$ . The way that this is done is by proving that the sequence of values of  $\epsilon$  parameter that correspond to the shortest interval converges in probability to  $1/2$ .

Throughout the ensuing discussion the following regularity conditions will be made about  $\log(f(y|\theta))$ . These are essentially the same as those in (Serfling 1980) with the inclusion that the third derivative of the loglikelihood is continuous. These conditions are sufficient for the MLE,  $\hat{\theta}$ , to be strongly consistent for the true value of the parameter,  $\theta_0$ .

**A 2.1** *The parameter space,  $\Theta$ , is an interval that is possibly infinite in length. Also,  $\log(f(y|\theta))$  is four times continuously differentiable.*

**A 2.2** *For every  $\theta_0 \in \Theta$  there exist functions  $h_1(y, \theta_0)$ ,  $h_2(y, \theta_0)$ ,  $h_3(y, \theta_0)$ , and  $h_4(y, \theta_0)$  such that for all  $\theta$  in an open neighborhood of  $\theta_0$ , we have for all  $y$ ,*

$$\begin{aligned} \left| \frac{\partial}{\partial \theta} \log(f(y|\theta)) \right| &\leq h_1(y, \theta_0), & \left| \frac{\partial^2}{\partial \theta^2} \log(f(y|\theta)) \right| &\leq h_2(y, \theta_0), \\ \left| \frac{\partial^3}{\partial \theta^3} \log(f(y|\theta)) \right| &\leq h_3(y, \theta_0), & \left| \frac{\partial^4}{\partial \theta^4} \log(f(y|\theta)) \right| &\leq h_4(y, \theta_0), \end{aligned}$$

*and in addition,*

$$\begin{aligned} \int h_1(y, \theta_0) dy &< \infty, & \int h_2(y, \theta_0) dy &< \infty, \\ E_{\theta} h_3(y, \theta_0) &< \infty, & E_{\theta} h_4(y, \theta_0) &< \infty. \end{aligned}$$

**A 2.3** *For all  $\theta \in \Theta$ ,*

$$0 < E_{\theta} \left( \frac{\partial}{\partial \theta} \log f(y|\theta) \right)^2 < \infty,$$

and

$$0 < E_{\theta} \left( \frac{\partial^2}{\partial \theta^2} \log f(y|\theta) \right)^2 < \infty.$$

To prove that the length-optimal interval has an asymptotically correct coverage rate,  $1 - \alpha$ , there are several preliminary lemmas that must be established. The first lemma concerns the asymptotic behavior of the observed information when the parameter estimate is substituted for the true value.

**Lemma 2.1** *Suppose  $y_1, y_2, \dots, y_n$  are iid with pdf  $f(y|\theta)$ , where  $f$  satisfies regularity conditions A 2.1, A 2.2, and A 2.3. Then the quantity,*

$$\frac{1}{n} \frac{\partial^2}{\partial \theta^2} l_n(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log f(y_i|\hat{\theta}),$$

*is asymptotically normally distributed with a mean of  $-1/v_{\theta}$  and variance,*

$$\frac{1}{n} \text{Var} \left( \frac{\partial^2}{\partial \theta^2} l(y|\theta_0) + v_{\theta} E \left( \frac{\partial^3}{\partial \theta^3} l(y|\theta_0) \right) \frac{\partial}{\partial \theta} l(y|\theta_0) \right).$$

**Proof** For the sake of a slight notational simplification, let  $\log f(y|\theta)$  be represented by  $l(y|\theta)$ .

We expand the observed information about  $\theta_0$ ,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} l(y_i | \hat{\theta}) &= \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} l(y_i | \theta_0) + \frac{1}{n} \sum_{i=1}^n \frac{\partial^3}{\partial \theta^3} l(y_i | \theta_0) (\hat{\theta} - \theta_0) \\
&\quad + \frac{1}{2n} \sum_{i=1}^n \frac{\partial^4}{\partial \theta^4} l(y_i | \theta^*) (\hat{\theta} - \theta_0)^2,
\end{aligned} \tag{2.10}$$

where  $\theta^*$  lies between  $\theta_0$  and  $\hat{\theta}$ . Adding and subtracting  $E \frac{\partial^3}{\partial \theta^3} l(y | \theta_0)$ , and  $n^{-1} v_\theta \sum \frac{\partial}{\partial \theta} l(y_i | \theta_0)$  in the second term on the right gives,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} l(y_i | \hat{\theta}) &= \frac{1}{n} \sum_{i=1}^n \left( \frac{\partial^2}{\partial \theta^2} l(y_i | \theta_0) + E \left( \frac{\partial^3}{\partial \theta^3} l(y | \theta_0) \right) v_\theta \frac{\partial}{\partial \theta} l(y_i | \theta_0) \right) \\
&\quad + E \left( \frac{\partial^3}{\partial \theta^3} l(y | \theta_0) \right) \left( \hat{\theta} - \theta_0 - \frac{v_\theta}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} l(y_i | \theta_0) \right) \\
&\quad + \frac{1}{n} \sum_{i=1}^n \left( \frac{\partial^3}{\partial \theta^3} l(y_i | \theta_0) - E \left( \frac{\partial^3}{\partial \theta^3} l(y | \theta_0) \right) \right) (\hat{\theta} - \theta_0) \\
&\quad + \frac{1}{2n} \sum_{i=1}^n \frac{\partial^4}{\partial \theta^4} l(y_i | \theta^*) (\hat{\theta} - \theta_0)^2.
\end{aligned} \tag{2.11}$$

The expression in the second line is  $o_p(n^{-1/2})$  by Lemma B of Serfling (1980), p. 153. The third line is also  $o_p(n^{-1/2})$  by Slutsky's Theorem, because  $\sqrt{n}(\hat{\theta} - \theta_0)$  is asymptotically normally distributed, and  $n^{-1} \sum (\frac{\partial^3}{\partial \theta^3} l(y_i | \theta_0) - E(\frac{\partial^3}{\partial \theta^3} l(y | \theta_0))) = o_p(1)$ . The expression in the last line is also  $o_p(n^{-1/2})$ . To see this, observe that since  $\hat{\theta} \xrightarrow{wp1} \theta_0$ , there is an  $N$  such that for all  $n > N$  we have,

$$\sqrt{n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\partial^4}{\partial \theta^4} l(y_i | \theta^*) \right| (\hat{\theta} - \theta_0)^2 < \sqrt{n} (\hat{\theta} - \theta_0) \frac{1}{n} \sum_{i=1}^n h_4(y_i) (\hat{\theta} - \theta_0). \tag{2.12}$$

Now we know that,

$$\sqrt{n}(\hat{\theta} - \theta_0) \frac{1}{n} \sum_{i=1}^n h_4(y_i) \xrightarrow{p} 0. \quad (2.13)$$

This, along with the asymptotic normality of the MLE imply, via Slutsky's Theorem, that the last line of the expansion is  $o_p(n^{-1/2})$ . The result then follows from the CLT.  $\square$

The next two lemmas describe the asymptotic behavior of the score function and likelihood ratio test statistic as a function of Wald-type interval endpoints.

**Lemma 2.2** *Suppose that  $y_1, y_2, \dots, y_n$  are iid with pdf  $f(y|\theta)$ , where  $f$  satisfies regularity conditions A 2.1, A 2.2, and A 2.3. Then for any real  $z$  we have,*

$$n^{-1/2} \frac{\partial}{\partial \theta} l_n \left( \hat{\theta} + \frac{z\sqrt{v_\theta}}{\sqrt{n}} \right) = -\frac{z}{\sqrt{v_\theta}} + O_p(n^{-1/2}).$$

**Proof** Taking a Taylor series about  $\hat{\theta}$  gives, for some  $\theta^*$  between  $\hat{\theta}$  and  $\hat{\theta} + z\sqrt{v_\theta}/\sqrt{n}$

$$\begin{aligned} n^{-1/2} \frac{\partial}{\partial \theta} l_n \left( \hat{\theta} + \frac{z\sqrt{v_\theta}}{\sqrt{n}} \right) &= \frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} l_n(\hat{\theta}) + \frac{1}{\sqrt{n}} \frac{\partial^2}{\partial \theta^2} l_n(\hat{\theta}) \left( z\sqrt{v_\theta} \frac{1}{\sqrt{n}} \right) \\ &\quad + \frac{1}{2\sqrt{n}} \frac{\partial^3}{\partial \theta^3} l_n(\theta^*) \left( z^2 v_\theta \frac{1}{n} \right) \\ &= 0 + \frac{1}{n} \frac{\partial^2}{\partial \theta^2} l_n(\hat{\theta}) z\sqrt{v_\theta} + \frac{1}{2n} \frac{\partial^3}{\partial \theta^3} l_n(\theta^*) \left( z^2 v_\theta \frac{1}{n} \right) \\ &= -\frac{z}{\sqrt{v_\theta}} + O_p(n^{-1/2}). \end{aligned} \quad (2.14)$$

The final line holds because of Lemma 2.1, and the regularity conditions ensure that

$$n^{-1} \partial^2 / \partial \theta^2 l_n(\hat{\theta}) \xrightarrow{p} -1/v_\theta.$$

This demonstrates that the first term converges in probability to  $-z/\sqrt{v_\theta}$ .  $\square$

**Lemma 2.3** *Suppose that  $y_1, y_2, \dots, y_n$  are iid with pdf  $f(y|\theta)$ , where  $f$  satisfies regularity conditions A 2.1, A 2.2, and A 2.3. Then we have,*

$$2 \left( l_n(\hat{\theta}) - l_n \left( \hat{\theta} + \frac{z\sqrt{v_\theta}}{\sqrt{n}} \right) \right) = z^2 + O_p(n^{-1/2}).$$

**Proof** Along the same lines of the proof of Lemma 2.2, we have the expansion,

$$\begin{aligned} 2 \left( l_n(\hat{\theta}) - l_n \left( \hat{\theta} + \frac{z\sqrt{v_\theta}}{\sqrt{n}} \right) \right) &= 0 - 2 \frac{\partial l_n(\hat{\theta})}{\partial \theta} \frac{z\sqrt{v_\theta}}{\sqrt{n}} - \frac{1}{n} \frac{\partial^2 l_n(\hat{\theta})}{\partial \theta^2} z^2 v_\theta - \frac{1}{n} \frac{1}{3} \frac{\partial^3 l_n(\hat{\theta})}{\partial \theta^3} \frac{z^3 v_\theta^{\frac{3}{2}}}{\sqrt{n}} \\ &= -\frac{1}{n} \frac{\partial^2 l_n(\hat{\theta})}{\partial \theta^2} z^2 v_\theta - \frac{1}{n} \frac{1}{3} \frac{\partial^3 l_n(\theta^*)}{\partial \theta^3} \frac{z^3 v_\theta^{\frac{3}{2}}}{\sqrt{n}} \\ &\xrightarrow{p} z^2 + O_p(n^{-\frac{1}{2}}). \end{aligned} \tag{2.15}$$

The result follows for essentially the same reasons as the results of Lemma 2.  $\square$

The fourth lemma establishes the first order equivalence of the  $\epsilon$ -versions of the likelihood and Wald-type intervals. Here we show that their endpoints differ by  $o_p(n^{-1/2})$ , but the argument can be easily generalized to  $o_p(n^{1/k})$ , for any fixed  $k > 0$ .

**Lemma 2.4** *Suppose that  $y_1, y_2, \dots, y_n$  are iid with pdf  $f(y|\theta)$ , where  $f$  satisfies regularity conditions A 2.1, A 2.2, and A 2.3. Then the endpoints of the  $\epsilon$ -likelihood interval are equivalent to the  $\epsilon$ -Wald-type interval endpoints, to order  $o_p(n^{-1/2})$ . That is,*

$$\begin{aligned} \hat{\theta}_n^U(\alpha, \epsilon) &= \hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha}\sqrt{v_\theta}}{\sqrt{n}} + o_p(n^{-\frac{1}{2}}) \\ \hat{\theta}_n^L(\alpha, \epsilon) &= \hat{\theta} + \frac{z_{1-\epsilon\alpha}\sqrt{v_\theta}}{\sqrt{n}} + o_p(n^{-\frac{1}{2}}). \end{aligned}$$

**Proof** Let  $\delta$  be given. Consider the following expansion of the LRT statistic about

$$\hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha}\sqrt{v_\theta}}{\sqrt{n}}.$$

$$\begin{aligned} n^{\frac{1}{4}} \left( 2(l_n(\hat{\theta}) - l_n(\hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha}\sqrt{v_\theta}}{\sqrt{n}} + \frac{\delta}{n^{\frac{3}{4}}})) - z_{1-(1-\epsilon)\alpha}^2 \right) &= n^{\frac{1}{4}} \left( 2(l_n(\hat{\theta}) - l_n(\hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha}\sqrt{v_\theta}}{\sqrt{n}})) \right. \\ &\quad \left. - z_{1-(1-\epsilon)\alpha}^2 \right) \\ &\quad - n^{\frac{1}{4}} 2 \frac{\partial}{\partial \theta} l_n \left( \hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha}\sqrt{v_\theta}}{\sqrt{n}} \right) \frac{\delta}{n^{\frac{3}{4}}} \\ &\quad - n^{\frac{1}{4}} \frac{\partial^2}{\partial \theta^2} l_n(\theta^*) \frac{\delta^2}{n^{\frac{6}{4}}}. \end{aligned} \tag{2.16}$$

Where  $\theta^*$  lies between  $\hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha}\sqrt{v_\theta}}{\sqrt{n}}$  and  $\hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha}\sqrt{v_\theta}}{\sqrt{n}} + \frac{\delta}{n^{-3/4}}$ . Now, if we apply

Lemma 2.2 and Lemma 2.3 to the first two right hand expressions we have,

$$\begin{aligned} n^{\frac{1}{4}} \left( 2 \left( l_n(\hat{\theta}) - l_n \left( \hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha}\sqrt{v_\theta}}{\sqrt{n}} + \frac{\delta}{n^{\frac{3}{4}}} \right) \right) - z_{1-(1-\epsilon)\alpha}^2 \right) &= n^{\frac{1}{4}} O_p(n^{-\frac{1}{2}}) + 2 \frac{z_{1-(1-\epsilon)\alpha}}{\sqrt{v_\theta}} \delta + O_p(n^{-\frac{1}{2}}) \\ &\quad - \frac{1}{n} \frac{\partial^2}{\partial \theta^2} l_n(\theta^*) \frac{\delta^2}{n^{\frac{1}{4}}}. \end{aligned}$$

Then, as in the proofs of the Lemmas 2.2 and 2.3, the last term on the right hand side may be described as  $O_p(1)O_p(n^{-1/4}) = O_p(n^{-1/4})$  enabling us to write,

$$n^{\frac{1}{4}} \left( 2 \left( l_n(\hat{\theta}) - l_n \left( \hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha}\sqrt{v_\theta}}{\sqrt{n}} + \frac{\delta}{n^{\frac{3}{4}}} \right) \right) - z_{1-(1-\epsilon)\alpha}^2 \right) = 2 \frac{z_{1-(1-\epsilon)\alpha}}{\sqrt{v_\theta}} \delta + O_p(n^{-\frac{1}{4}}).$$

Up to this point we have assumed only that  $\delta$  is a real number. From this point on, we assume that  $\delta > 0$ . Then we also have the similar expression,

$$n^{\frac{1}{4}} \left( 2 \left( l_n(\hat{\theta}) - l_n \left( \hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha}\sqrt{v_\theta}}{\sqrt{n}} - \frac{\delta}{n^{\frac{3}{4}}} \right) \right) - z_{1-(1-\epsilon)\alpha}^2 \right) = -2 \frac{z_{1-(1-\epsilon)\alpha}}{\sqrt{v_\theta}} \delta + O_p(n^{-\frac{1}{4}}).$$

This demonstrates that, for any  $\delta > 0$ , the limit of the probabilities that there is a  $-\delta < \delta^* < \delta$  such that,

$$n^{\frac{1}{4}} \left( 2 \left( l_n(\hat{\theta}) - l_n \left( \hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha} \sqrt{v_{\theta}}}{\sqrt{n}} + \frac{\delta^*}{n^{\frac{3}{4}}} \right) \right) - z_{1-(1-\epsilon)\alpha}^2 \right) = 0$$

is 1. This is due to the assumed continuity of the loglikelihood, and the fact that the expression on the left hand side above considered as a function of  $\delta^*$  changes sign in the interval  $(-\delta, \delta)$ . Because a root of this form has been shown to exist within  $\frac{\delta}{n^{\frac{3}{4}}}$  of  $\hat{\theta} + z_{1-(1-\epsilon)\alpha} \frac{v_{\theta}}{\sqrt{n}}$ , we may assert that

$$\begin{aligned} \hat{\theta}_n^U(\alpha, \epsilon) &= \hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha} \sqrt{v_{\theta}}}{\sqrt{n}} + O_p(n^{-\frac{3}{4}}) \\ &= \hat{\theta} + \frac{z_{1-(1-\epsilon)\alpha} \sqrt{v_{\theta}}}{\sqrt{n}} + o_p(n^{-\frac{1}{2}}). \end{aligned} \tag{2.17}$$

The proof of the second part of the lemma concerning  $\hat{\theta}_n^L(\alpha, \epsilon)$  is identical to the proof for  $\hat{\theta}_n^U(\alpha, \epsilon)$ .  $\square$

Now define  $\epsilon_n^*$  to be the  $\epsilon$  that corresponds to the shortest interval. That is, the value of  $\epsilon$  that minimizes

$$\hat{\theta}_n^U(\alpha, \epsilon) - \hat{\theta}_n^L(\alpha, \epsilon).$$

We are now ready to establish that  $\epsilon_n^*$  is consistent for  $1/2$ . The essential idea of the proof is that for any two values  $\epsilon_1$  and  $\epsilon_2$ , the  $\epsilon$ -interval whose value of  $\epsilon$  is closer to  $1/2$  is

asymptotically shorter than the other.

**Lemma 2.5** *Suppose that  $y_1, y_2, \dots, y_n$  are iid with pdf  $f(y|\theta)$ , where  $f$  satisfies regularity conditions A 2.1, A 2.2, and A 2.3 hold. Then  $\epsilon_n^* - 1/2 \xrightarrow{p} 0$ .*

**Proof** Let  $\epsilon_1 \in (0, 1)$  be not equal to  $1/2$ , but otherwise arbitrary. Note that if we had allowed  $\epsilon_1 = 0$  or  $\epsilon_1 = 1$  then either  $z_{1-(1-\epsilon_1)\alpha}^2$  or  $z_{\epsilon_1\alpha}^2$  is infinite in value, so that the corresponding  $\epsilon$ -likelihood interval must be infinite in length. So, the minimum length interval can never occur at either of those values of  $\epsilon_1$ .

Now, by Lemma 2.4 we have,

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n^U(\alpha, \epsilon_1) - \hat{\theta}_n^L(\alpha, \epsilon_1)) &= \sqrt{v_\theta}(z_{1-(1-\epsilon_1)\alpha} - z_{\epsilon_1\alpha}) + o_p(1) \\ &= \sqrt{v_\theta}(z_{1-\alpha+\epsilon_1\alpha} + z_{1-\epsilon_1\alpha}) + o_p(1). \end{aligned} \tag{2.18}$$

Performing the first and second derivative tests on the function,  $f(\epsilon) = z_{1-\alpha+\epsilon\alpha} + z_{1-\epsilon\alpha}$ , shows that  $f(\epsilon)$  is strictly convex and has its unique minimum at  $\epsilon = 1/2$ . Thus, for every  $\epsilon_2$  such that  $0 \leq |\epsilon_2 - 1/2| < |\epsilon_1 - 1/2|$ , and  $0 \leq (\epsilon_1 - 1/2)(\epsilon_2 - 1/2)$  we have,

$$\begin{aligned} \sqrt{n}((\hat{\theta}_n^U(\alpha, \epsilon_1) - \hat{\theta}_n^L(\alpha, \epsilon_1)) - (\hat{\theta}_n^U(\alpha, \epsilon_2) - \hat{\theta}_n^L(\alpha, \epsilon_2))) &= \sqrt{v_\theta}(z_{1-\alpha+\epsilon_1\alpha} + z_{\epsilon_1\alpha} \\ &\quad - (z_{1-\alpha+\epsilon_2\alpha} + z_{1-\epsilon_2\alpha})) + o_p(1). \end{aligned} \tag{2.19}$$

The right hand side of 2.19 is, by the preceding comments, a fixed positive constant plus an expression that is converging in probability to zero. This indicates that,

$$\lim_{n \rightarrow \infty} P(\sqrt{n}(\hat{\theta}_n^U(\alpha, \epsilon_2) - \hat{\theta}_n^L(\alpha, \epsilon_2)) < \sqrt{n}(\hat{\theta}_n^U(\alpha, \epsilon_1) - \hat{\theta}_n^L(\alpha, \epsilon_1))) = 1.$$

This demonstrates that for every  $\epsilon_1 \in (0, 1)$  with  $\epsilon_1 \neq 1/2$ , the probability that there is another  $\epsilon_2$  that is closer to  $1/2$  such that the length of the  $\epsilon_2$  interval is shorter than the  $\epsilon_1$  interval goes to one as  $n \rightarrow \infty$ . This is equivalent to the assertion that  $\epsilon_n^* \xrightarrow{p} 1/2$ .  $\square$

Theorem 2.1 establishes the asymptotic equivalence of the length-optimal likelihood interval and the Wald-type interval. The corollary that follows shows that the length-optimal interval's coverage is asymptotically correct.

**Theorem 2.1** *Suppose that  $y_1, y_2, \dots, y_n$  are iid with pdf  $f(y|\theta)$ , where  $f$  satisfies regularity conditions A 2.1, A 2.2, and A 2.3 hold. Then the length-optimal likelihood confidence interval endpoints is asymptotically equivalent to the equal-tail Wald-type interval endpoints, to order  $o_p(n^{-1/2})$ . That is,*

$$\sqrt{n} \left( \hat{\theta}_n^U(\alpha, \epsilon_n^*) - \left( \hat{\theta} + z_{1-\alpha/2} \frac{\sqrt{v_\theta}}{\sqrt{n}} \right) \right) = o_p(1),$$

and

$$\sqrt{n} \left( \hat{\theta}_n^L(\alpha, \epsilon_n^*) - \left( \hat{\theta} + z_{\alpha/2} \frac{\sqrt{v_\theta}}{\sqrt{n}} \right) \right) = o_p(1).$$

**Proof** For all  $\epsilon \in (0, 1)$  we have, by Lemma 2.5,

$$\begin{aligned} \sqrt{n} \left( \hat{\theta}_n^U(\alpha, \epsilon) - \left( \hat{\theta} + z_{1-\alpha/2} \frac{\sqrt{v_\theta}}{\sqrt{n}} \right) \right) &= \sqrt{n} \left( \left( \hat{\theta} + z_{1-(1-\epsilon)\alpha} \frac{\sqrt{v_\theta}}{\sqrt{n}} \right) + o_p(n^{-1/2}) - \left( \hat{\theta} + z_{1-\alpha/2} \frac{\sqrt{v_\theta}}{\sqrt{n}} \right) \right) \\ &= \sqrt{v_\theta} (z_{1-(1-\epsilon)\alpha} - z_{1-\alpha/2}) + o_p(1). \end{aligned} \tag{2.20}$$

Now, inserting  $\epsilon_n^*$  gives,

$$\sqrt{n} \left( \hat{\theta}_n^U(\alpha, \epsilon_n^*) - \left( \hat{\theta} + z_{1-\alpha/2} \frac{\sqrt{v_\theta}}{\sqrt{n}} \right) \right) = \sqrt{v_\theta} (z_{1-(1-\epsilon_n^*)\alpha} - z_{1-\alpha/2}) + o_p(1).$$

Lemma 2.5 indicates that  $\epsilon_n^* \xrightarrow{p} 1/2$ . Hence,  $z_{1-(1-\epsilon_n^*)\alpha} - z_{1-\alpha/2} = o_p(1)$ . This implies that,

$$\begin{aligned} \sqrt{n} \left( \hat{\theta}_n^U(\alpha, \epsilon_n^*) - \left( \hat{\theta} + z_{1-\alpha/2} \frac{\sqrt{v_\theta}}{\sqrt{n}} \right) \right) &= o_p(1) + o_p(1) \\ &= o_p(1). \end{aligned}$$

This completes the proof of the theorem for the upper endpoint,  $\hat{\theta}_n^U(\alpha, \epsilon_n^*)$ . The proof for  $\hat{\theta}_n^L(\alpha, \epsilon_n^*)$  is identical to the argument above, so it is omitted.  $\square$

**Corollary 2.1** *Suppose that  $y_1, y_2, \dots, y_n$  are iid with pdf  $f(y|\theta)$ , where  $f$  satisfies regularity conditions A 2.1, A 2.2, and A 2.3. Then the asymptotic coverage of the length optimal likelihood interval is  $1 - \alpha$ .*

**Proof** This is a simple application of Slutsky's Theorem, the asymptotic normality of the MLE,

and the Theorem 2.1.

$$\begin{aligned}
\lim_{n \rightarrow \infty} P(\hat{\theta}_n^L(\alpha, \epsilon_n^*) \leq \theta \leq \hat{\theta}_n^U(\alpha, \epsilon_n^*)) &= \lim_{n \rightarrow \infty} P(\hat{\theta} + z_{\alpha/2} \frac{\sqrt{v_\theta}}{\sqrt{n}} + o_p(n^{-1/2}) \leq \theta \leq \hat{\theta} + z_{1-\alpha/2} \frac{\sqrt{v_\theta}}{\sqrt{n}} + o_p(n^{-1/2})) \\
&= \lim_{n \rightarrow \infty} P(-z_{1-\alpha/2} \leq \sqrt{n} \frac{\hat{\theta} - \theta}{\sqrt{v_\theta}} \leq z_{1-\alpha/2}) \\
&= 1 - \alpha.
\end{aligned} \tag{2.21}$$

The right hand side of the first line follows from Theorem 2.1, and the second line by Slutsky's Theorem. The third line is the application of the asymptotic normality of the MLE.  $\square$

## 2.4 Simulations

In this section we apply the length-optimal approach to the simulations in Chapter 1. When those simulations were run, the length-optimal generalized score and robust profile likelihood intervals were also computed. We find that the length optimal approach often increases the coverage probabilities of parameters in logistic GEE models and simple linear measurement error models, but has little effect in the Huber regression and Poisson GEE models. We also see that the length-optimal approach is only modestly successful in reducing the frequency with which the generalized score interval is infinite.

### 2.4.1 Logistic GEE Simulations

In the case of logistic GEE models, the length-optimal generalized score and robust likelihood intervals appear to generally have higher coverage than their equal-tail counterparts. This effect is most pronounced when the sample size is small, as the theory of the previous

section would suggest. For the case of the simple linear logistic GEE simulations, when the true value of  $\psi = 1$ ,  $K = 5$ , and  $\rho = .5$ , the robust likelihood interval's coverage is .88 for the equal-tail version, but the length-optimal robust likelihood interval is much higher at .92. Surprisingly, the effect of the length-optimal procedure on the median length of the robust likelihood interval was quite small, as the median lengths of the equal-tail and length-optimal robust likelihood intervals were generally the same through the third significant figure.

We also see that the generalized score interval's length-optimal version has higher coverage than the equal-tail generalized score interval, but when we examine its median lengths we see that this interval still appears to be rather inefficient. As one would guess, the length-optimal approach is able to significantly reduce the frequency of infinite length intervals, but in the smallest samples this frequency is still disappointingly high.

The GUIDE data example's simulation results were similar to those of the simple linear logistic GEE simulations', with the length-optimal robust profile performing the best overall. In Table 2.4, the length-optimal versions of the generalized score and robust profile likelihood intervals are tabulated with their equal-tail counterparts. Tables 2.5 and 2.6 list the results of the GUIDE data simulations.



Table 2.4: Equal-tail and length-optimal approximate 95% GEE confidence intervals for the GUIDE dataset.

| Parameter | Gen. Score  | Opt. Gen. Score | Rob. Like   | Opt. Rob. Like |
|-----------|-------------|-----------------|-------------|----------------|
| SEX       | (-2.0,1.7)  | (-2.2,1.3.)     | (-2.0,.75)  | (-2.0,.74)     |
| AGE       | (-2.0,.67)  | (-2.0,.66)      | (-1.9,.45)  | (-1.9,.46)     |
| DAYACC    | (.26,.74)   | (.24,.69)       | (.25,.65)   | (.24,.64)      |
| SEVERE    | (.069,1.9)  | (.0019,.1.8)    | (.15,1.6)   | (.14,.16)      |
| TOILET    | (-.074,.42) | (-.099,.37)     | (-.095,.27) | (-.097,.26)    |

Table 2.5: Coverage and length for length-optimal generalized score and robust likelihood intervals for clustered Bernoulli data simulated from the GUIDE dataset.

| Parameter | # of Clusters $K =$ |                 | $\rho = 0$ |     | $\rho = 0.2$ |     |
|-----------|---------------------|-----------------|------------|-----|--------------|-----|
|           |                     |                 | 20         | 37  | 20           | 37  |
| SEX       | Coverage            | Gen. Score      | .97        | .95 | .94          | .95 |
|           |                     | Opt. Gen. Score | .99        | .97 | .96          | .97 |
|           |                     | Rob. Like.      | .95        | .95 | .93          | .95 |
|           |                     | Opt. Rob. Like. | .96        | .96 | .94          | .96 |
|           | Median              | Gen. Score      | 22.3       | 3.3 | $\infty$     | 3.6 |
|           |                     | Opt. Gen. Score | 10.9       | 3.3 | 13.2         | 3.6 |
|           | Length              | Rob. Like.      | 3.9        | 2.6 | 4.1          | 2.7 |
|           |                     | Opt. Rob. Like. | 3.9        | 2.6 | 4.1          | 2.7 |
|           | Coverage            | Gen. Score      | .93        | .96 | .94          | .92 |
|           |                     | Opt. Gen. Score | .96        | .97 | .96          | .94 |
|           |                     | Rob. Like.      | .93        | .95 | .95          | .93 |
|           |                     | Opt. Rob. Like. | .94        | .95 | .95          | .94 |
| AGE       | Median              | Gen. Score      | 5.3        | 2.7 | 5.7          | 2.9 |
|           | Length              | Opt. Gen. Score | 5.1        | 2.7 | 5.4          | 2.6 |
|           |                     | Rob. Like.      | 3.6        | 2.4 | 3.7          | 2.5 |
|           |                     | Opt. Rob. Like. | 3.6        | 2.4 | 3.7          | 2.5 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Table 2.6: Proportion of infinite length length-optimal generalized score and robust likelihood intervals for clustered Bernoulli data simulated from the GUIDE dataset.

| Parameter |            | # of Clusters $K =$ | $\rho = 0$ |     | $\rho = 0.2$ |     |
|-----------|------------|---------------------|------------|-----|--------------|-----|
|           |            |                     | 20         | 37  | 20           | 37  |
| SEX       | Proportion | Gen. Score          | .50        | .02 | .56          | .03 |
|           | Infinite   | Opt. Gen. Score     | .28        | 0   | .42          | 0   |
|           |            | Rob. Like.          | .03        | 0   | 0            | 0   |
|           |            | Opt. Rob. Like.     | 0          | 0   | 0            | 0   |
| AGE       | Proportion | Gen. Score          | .04        | 0   | .10          | 0   |
|           | Infinite   | Opt. Gen. Score     | .01        | 0   | .05          | 0   |
|           |            | Rob. Like.          | 0          | 0   | 0            | 0   |
|           |            | Opt. Rob. Like.     | 0          | 0   | 0            | 0   |

### 2.4.2 Poisson GEE simulations

In the Poisson GEE simulations, the effect of the length-optimal approach was less significant than in the logistic GEE case. The length-optimal approach had almost no effect, for better or for worse, on the robust profile likelihood interval. It does have some positive effect on the generalized score interval, reducing somewhat the frequency of infinite intervals and median length as compared to the equal-tail generalized score interval, and somewhat increasing its coverage probability.

The tables below are results for the Epilepsy dataset. Table 2.9 compares the length-optimal and equal-tail intervals for the Poisson regression parameters, while Table 2.10 contains simulation results. Like the results of the simple linear Poisson regression example, the length-optimal approach appears to have little effect.



Table 2.9: Equal-tail and length-optimal approximate 95% GEE confidence intervals for the Epilepsy dataset.

| Parameter | Gen. Score  | Opt. Gen. Score | Rob. Like   | Opt. Rob. Like |
|-----------|-------------|-----------------|-------------|----------------|
| TREATMENT | (-1.0,.58)  | (-.93,.58)      | (-.77,.60)  | (-.77,.60)     |
| AGE       | (-1.5,.86)  | (-1.6,.73)      | (-1.4,.46)  | (-1.4,.47)     |
| BASELINE  | (-1.7,-.99) | (-1.7,-.97)     | (-1.6,-.98) | (-1.6,-.98)    |

Table 2.10: Coverage and length for length-optimal generalized score and robust likelihood intervals for serially correlated Poisson data using the Epilepsy data covariates and MLE.

|           |          |                     | $\rho = 0$ |     | $\rho = .2$ |     | $\rho = .5$ |     |
|-----------|----------|---------------------|------------|-----|-------------|-----|-------------|-----|
| Parameter |          | # of Clusters $K =$ | 30         | 59  | 30          | 59  | 30          | 59  |
| TREATMENT | Coverage | Gen. Score          | .94        | .94 | .95         | .94 | .95         | .94 |
|           |          | Opt. Gen. Score     | .94        | .94 | .95         | .94 | .95         | .94 |
|           |          | Rob. Like.          | .96        | .95 | .92         | .93 | .91         | .92 |
|           |          | Opt. Rob. Like.     | .96        | .95 | .92         | .93 | .91         | .93 |
|           | Median   | Gen. Score          | .18        | .13 | .76         | .54 | 1.6         | 1.1 |
|           |          | Opt. Gen. Score     | .18        | .13 | .76         | .54 | 1.6         | 1.1 |
|           | Length   | Rob. Like.          | .18        | .13 | .69         | .51 | 1.3         | .98 |
|           |          | Opt. Rob. Like.     | .18        | .13 | .69         | .51 | 1.3         | .98 |
|           | Coverage | Gen. Score          | .95        | .95 | .95         | .94 | .96         | .95 |
|           |          | Opt. Gen. Score     | .95        | .95 | .96         | .95 | .96         | .96 |
|           |          | Rob. Like.          | .96        | .95 | .91         | .92 | .90         | .92 |
|           |          | Opt. Rob. Like.     | .96        | .95 | .91         | .92 | .90         | .92 |
| AGE       | Median   | Gen. Score          | .39        | .27 | 1.6         | 1.1 | 3.2         | 2.1 |
|           |          | Opt. Gen. Score     | .38        | .27 | 1.5         | 1.1 | 3.2         | 2.1 |
|           | Length   | Rob. Like.          | .32        | .25 | 1.1         | .93 | 2.0         | 1.8 |
|           |          | Opt. Rob. Like.     | .32        | .25 | 1.1         | .93 | 2.0         | 1.8 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Cluster size=5. Less than .005 of each of the intervals were reported as infinite.

### 2.4.3 Huber Robust Linear Regression

The length-optimal approach appears to have virtually no effect on the intervals in the case of robust linear regression, except that it reduces the frequency with which the generalized score interval is infinite in length. By the other measures of coverage probability and median length, the length-optimal and equal-tail versions of all the intervals investigated here are practically identical.

Table 2.11: Coverage and length for length-optimal generalized score and robust likelihood intervals for iid data under a simple linear Huber regression model.

| $(\mu, \beta, \sigma)$ |                  | Dist.<br># Obs. $K =$ | <i>Normal</i> |     |     | <i>Cauchy</i> |     |     | <i>Mixture</i> |     |     |
|------------------------|------------------|-----------------------|---------------|-----|-----|---------------|-----|-----|----------------|-----|-----|
|                        |                  |                       | 5             | 10  | 20  | 5             | 10  | 20  | 5              | 10  | 20  |
| (0,0,1)                | Coverage         | SH-Rob. Like.         | .93           | .94 | .95 | .96           | .95 | .95 | .96            | .96 | .95 |
|                        |                  | Opt. SH-Rob. Like.    | .93           | .94 | .95 | .96           | .95 | .95 | .96            | .96 | .95 |
|                        | Coverage         | Gen. Score            | .95           | .95 | .95 | .97           | .95 | .95 | .96            | .96 | .95 |
|                        |                  | Opt. Gen. Score       | .95           | .95 | .95 | .97           | .95 | .95 | .96            | .96 | .95 |
|                        |                  | Rob. Like.            | .95           | .96 | .98 | .97           | .97 | .97 | .97            | .95 | .97 |
|                        |                  | Opt. Rob. Like.       | .95           | .96 | .98 | .97           | .97 | .97 | .97            | .95 | .97 |
|                        | Median<br>Length | SH-Rob. Like.         | 2.2           | 1.5 | .99 | 4.6           | 2.9 | 1.7 | 3.5            | 2.2 | 1.3 |
|                        |                  | Opt. SH-Rob. Like.    | 2.2           | 1.5 | .99 | 4.6           | 2.9 | 1.7 | 3.5            | 2.2 | 1.3 |
|                        |                  | Gen. Score            | 2.7           | 1.6 | 1.0 | 8.4           | 3.6 | 1.8 | 6.6            | 2.8 | 1.4 |
|                        |                  | Opt. Gen. Score       | 2.7           | 1.6 | 1.0 | 8.4           | 3.6 | 1.8 | 6.6            | 2.8 | 1.4 |
|                        |                  | Rob. Like.            | 2.4           | 1.8 | 1.1 | 4.9           | 3.2 | 1.9 | 3.8            | 2.5 | 1.5 |
|                        |                  | Opt. Rob. Like.       | 2.4           | 1.8 | 1.1 | 4.9           | 3.2 | 1.9 | 3.8            | 2.5 | 1.5 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Table 2.12: Coverage and length for length-optimal generalized score and robust likelihood intervals for iid data under a simple linear Huber regression model.

|                            |                  | Dist.              | <i>Normal</i> |     |     | <i>Cauchy</i> |     |     | <i>Mixture</i> |     |     |
|----------------------------|------------------|--------------------|---------------|-----|-----|---------------|-----|-----|----------------|-----|-----|
| $(\mu, \beta, \sigma)$     |                  | # Obs. $K =$       | 5             | 10  | 20  | 5             | 10  | 20  | 5              | 10  | 20  |
| $(1, 1, \sigma_i^\dagger)$ | Coverage         | SH-Rob. Like.      | .94           | .95 | .95 | .96           | .95 | .95 | .95            | .95 | .94 |
|                            |                  | Opt. SH-Rob. Like. | .94           | .95 | .95 | .96           | .95 | .95 | .95            | .95 | .94 |
|                            |                  | Gen. Score         | .95           | .94 | .95 | .96           | .94 | .94 | .95            | .97 | .95 |
|                            |                  | Opt. Gen. Score    | .95           | .94 | .95 | .96           | .94 | .94 | .95            | .97 | .95 |
|                            |                  | Rob. Like.         | .95           | .96 | .97 | .97           | .97 | .97 | .95            | .94 | .97 |
|                            |                  | Opt. Rob. Like.    | .95           | .96 | .97 | .97           | .97 | .97 | .95            | .94 | .97 |
|                            | Median<br>Length | SH-Rob. Like.      | 2.2           | 1.6 | .98 | 4.5           | 2.8 | 1.7 | 2.4            | 1.6 | 1.0 |
|                            |                  | Opt. SH-Rob. Like. | 2.2           | 1.6 | .98 | 4.5           | 2.8 | 1.7 | 2.4            | 1.6 | 1.0 |
|                            |                  | Gen. Score         | 2.2           | 1.7 | 1.0 | 8.1           | 3.5 | 1.8 | 3.4            | 1.8 | 1.1 |
|                            |                  | Opt. Gen. Score    | 2.2           | 1.7 | 1.0 | 8.1           | 3.5 | 1.8 | 3.4            | 1.8 | 1.1 |
|                            |                  | Rob. Like.         | 2.4           | 1.7 | 1.1 | 4.6           | 3.1 | 1.9 | 2.6            | 1.8 | 1.1 |
|                            |                  | Opt. Rob. Like.    | 2.4           | 1.7 | 1.1 | 4.6           | 3.1 | 1.9 | 2.6            | 1.8 | 1.1 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

$^\dagger \sigma_i = .1 + |\mu + x_i|$ .

Table 2.13: Proportion of infinite length length-optimal generalized score and robust likelihood intervals for iid data under a simple linear Huber regression model.

|                            |            | Dist.           | <i>Normal</i> |    |    | <i>Cauchy</i> |     |    | <i>Mixture</i> |     |    |
|----------------------------|------------|-----------------|---------------|----|----|---------------|-----|----|----------------|-----|----|
| $(\mu, \beta, \sigma)$     |            | # of Obs. $K =$ | 10            | 20 | 50 | 10            | 20  | 50 | 10             | 20  | 50 |
| $(0, 0, 1)$                | Proportion | Gen. Score      | .01           | 0  | 0  | .12           | .01 | 0  | .12            | 0   | 0  |
|                            | Infinite   | Opt. Gen. Score | 0             | 0  | 0  | .07           | 0   | 0  | .07            | 0   | 0  |
| $(1, 1, \sigma_i^\dagger)$ | Proportion | Gen. Score      | 0             | 0  | 0  | .12           | 0   | 0  | .06            | .01 | 0  |
|                            | Infinite   | Opt. Gen. Score | 0             | 0  | 0  | .07           | 0   | 0  | .03            | .01 | 0  |

Note: The incidence of infinite length robust likelihood-type intervals did not exceed .005.

$^\dagger \sigma_i = .1 + |\mu + x_i|$ .

### 2.4.4 Simple Linear Measurement Error Models

The length-optimal generalized score and robust profile intervals are both slightly superior to their equal-tail version for the case of simple linear measurement error models. The improvement appears to increase with the measurement error variance. This effect is plausible because as the measurement error increases, we are in a sense moving farther away from least squares, where, in the case of the robust likelihood interval, the length-optimal interval coincides with its equal-tail version. As in the logistic GEE simulations, the improvements over the equal-tail versions of the intervals are manifested more in the coverage probabilities than with the median lengths.

Table 2.14: Coverage and length for length-optimal generalized score and robust likelihood intervals from simple linear measurement error model.

| Distribution of          |          |                 | $\sigma_u^2 = .1$ |     |     | $\sigma_u^2 = 0.25$ |     |     | $\sigma_u^2 = 0.5$ |     |     |
|--------------------------|----------|-----------------|-------------------|-----|-----|---------------------|-----|-----|--------------------|-----|-----|
| $(\epsilon_x, \epsilon)$ |          | # of Obs. $n =$ | 20                | 50  | 100 | 20                  | 50  | 100 | 20                 | 50  | 100 |
| $N(0, 1), N(0, 1)$       | Coverage | Gen. Score      | .95               | .94 | .95 | .94                 | .95 | .95 | .93                | .93 | .94 |
|                          |          | Opt. Gen. Score | .97               | .95 | .96 | .95                 | .96 | .96 | .93                | .96 | .96 |
|                          |          | Rob. Like.      | .93               | .95 | .96 | .95                 | .96 | .96 | .95                | .95 | .95 |
|                          |          | Opt. Rob. Like. | .94               | .95 | .96 | .96                 | .96 | .96 | .96                | .96 | .96 |
|                          | Median   | Gen. Score      | 1.5               | .68 | .46 | 2.2                 | .83 | .54 | 6.7                | 1.2 | .71 |
|                          |          | Opt. Gen. Score | 1.4               | .68 | .45 | 1.9                 | .82 | .54 | 5.6                | 1.1 | .69 |
|                          |          | Rob. Like.      | 1.1               | .64 | .45 | 1.3                 | .76 | .52 | 2.0                | 1.0 | .67 |
|                          |          | Opt. Rob. Like. | 1.1               | .64 | .45 | 1.3                 | .76 | .52 | 1.9                | 1.0 | .66 |
|                          | Length   | Gen. Score      | 1.5               | .68 | .46 | 2.2                 | .83 | .54 | 6.7                | 1.2 | .71 |
|                          |          | Opt. Gen. Score | 1.4               | .68 | .45 | 1.9                 | .82 | .54 | 5.6                | 1.1 | .69 |
|                          |          | Rob. Like.      | 1.1               | .64 | .45 | 1.3                 | .76 | .52 | 2.0                | 1.0 | .67 |
|                          |          | Opt. Rob. Like. | 1.1               | .64 | .45 | 1.3                 | .76 | .52 | 1.9                | 1.0 | .66 |
| $N(0, 1), t_5$           | Coverage | Gen. Score      | .96               | .96 | .95 | .95                 | .95 | .95 | .93                | .94 | .93 |
|                          |          | Opt. Gen. Score | .97               | .96 | .95 | .96                 | .96 | .96 | .93                | .97 | .96 |
|                          |          | Rob. Like.      | .95               | .96 | .96 | .94                 | .95 | .96 | .95                | .97 | .96 |
|                          |          | Opt. Rob. Like. | .95               | .96 | .96 | .94                 | .96 | .96 | .96                | .97 | .96 |
|                          | Median   | Gen. Score      | 1.4               | .66 | .45 | 2.1                 | .91 | .53 | 6.6                | 1.2 | .70 |
|                          |          | Opt. Gen. Score | 1.3               | .66 | .45 | 1.9                 | .91 | .53 | 5.5                | 1.1 | .68 |
|                          |          | Rob. Like.      | 1.0               | .64 | .45 | 1.3                 | .74 | .52 | 2.0                | .98 | .66 |
|                          |          | Opt. Rob. Like. | 1.0               | .64 | .45 | 1.2                 | .74 | .52 | 1.9                | .97 | .66 |
|                          | Length   | Gen. Score      | 1.4               | .66 | .45 | 2.1                 | .91 | .53 | 6.6                | 1.2 | .70 |
|                          |          | Opt. Gen. Score | 1.3               | .66 | .45 | 1.9                 | .91 | .53 | 5.5                | 1.1 | .68 |
|                          |          | Rob. Like.      | 1.0               | .64 | .45 | 1.3                 | .74 | .52 | 2.0                | .98 | .66 |
|                          |          | Opt. Rob. Like. | 1.0               | .64 | .45 | 1.2                 | .74 | .52 | 1.9                | .97 | .66 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$

Table 2.15: Coverage and length for length-optimal generalized score and robust likelihood intervals from simple linear measurement error model.

| Distribution of          |          |                 | $\sigma_u^2 = .1$ |     |     | $\sigma_u^2 = 0.25$ |     |     | $\sigma_u^2 = 0.5$ |     |     |
|--------------------------|----------|-----------------|-------------------|-----|-----|---------------------|-----|-----|--------------------|-----|-----|
| $(\epsilon_x, \epsilon)$ |          | # of Obs. $n =$ | 20                | 50  | 100 | 20                  | 50  | 100 | 20                 | 50  | 100 |
| $t_5, N(0, 1)$           | Coverage | Gen. Score      | .96               | .96 | .96 | .95                 | .94 | .95 | .94                | .94 | .93 |
|                          |          | Opt. Gen. Score | .97               | .96 | .96 | .95                 | .96 | .96 | .94                | .96 | .95 |
|                          |          | Rob. Like.      | .94               | .95 | .96 | .95                 | .95 | .96 | .95                | .96 | .94 |
|                          |          | Opt. Rob. Like. | .94               | .95 | .96 | .96                 | .95 | .96 | .96                | .97 | .96 |
|                          | Median   | Gen. Score      | 2.0               | .77 | .48 | 4.0                 | .97 | .58 | 8.2                | 1.5 | .77 |
|                          |          | Opt. Gen. Score | 1.8               | .76 | .48 | 3.0                 | .94 | .58 | 7.3                | 1.4 | .74 |
|                          |          | Rob. Like.      | 1.1               | .67 | .46 | 1.4                 | .78 | .53 | 2.2                | 1.0 | .67 |
|                          |          | Opt. Rob. Like. | 1.1               | .67 | .46 | 1.4                 | .78 | .53 | 2.0                | 1.0 | .67 |
|                          | Length   | Gen. Score      | 1.9               | .75 | .47 | 4.0                 | .96 | .57 | 8.0                | 1.5 | .77 |
|                          |          | Opt. Gen. Score | 1.8               | .73 | .47 | 2.9                 | .93 | .56 | 7.0                | 1.3 | .74 |
|                          |          | Rob. Like.      | 1.0               | .65 | .45 | 1.3                 | .78 | .53 | 2.0                | 1.0 | .67 |
|                          |          | Opt. Rob. Like. | 1.0               | .65 | .45 | 1.3                 | .78 | .53 | 1.9                | 1.0 | .67 |
| $t_5, t_5$               | Coverage | Gen. Score      | .96               | .95 | .95 | .95                 | .94 | .94 | .96                | .93 | .93 |
|                          |          | Opt. Gen. Score | .97               | .96 | .96 | .96                 | .96 | .95 | .96                | .96 | .95 |
|                          |          | Rob. Like.      | .94               | .95 | .96 | .95                 | .95 | .95 | .95                | .95 | .94 |
|                          |          | Opt. Rob. Like. | .95               | .95 | .96 | .97                 | .95 | .95 | .96                | .97 | .95 |
|                          | Median   | Gen. Score      | 1.9               | .75 | .47 | 4.0                 | .96 | .57 | 8.0                | 1.5 | .77 |
|                          |          | Opt. Gen. Score | 1.8               | .73 | .47 | 2.9                 | .93 | .56 | 7.0                | 1.3 | .74 |
|                          |          | Rob. Like.      | 1.0               | .65 | .45 | 1.3                 | .78 | .53 | 2.0                | 1.0 | .67 |
|                          |          | Opt. Rob. Like. | 1.0               | .65 | .45 | 1.3                 | .78 | .53 | 1.9                | 1.0 | .67 |
|                          | Length   | Gen. Score      | 1.9               | .75 | .47 | 4.0                 | .96 | .57 | 8.0                | 1.5 | .77 |
|                          |          | Opt. Gen. Score | 1.8               | .73 | .47 | 2.9                 | .93 | .56 | 7.0                | 1.3 | .74 |
|                          |          | Rob. Like.      | 1.0               | .65 | .45 | 1.3                 | .78 | .53 | 2.0                | 1.0 | .67 |
|                          |          | Opt. Rob. Like. | 1.0               | .65 | .45 | 1.3                 | .78 | .53 | 1.9                | 1.0 | .67 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$



## Chapter 3

# The GS2 Confidence Interval

### 3.1 Introduction

In the previous two chapters we have investigated interval estimation methods for models based on a likelihood. In many of the situations we have examined, the robust profile likelihood interval has been shown, via simulation, to be superior to the generalized Wald and score confidence intervals. When one is using a pure estimating equations approach and there is no likelihood, the robust profile interval is not available, but the definitions of the generalized Wald and generalized score intervals are easily modified to accomodate this general situation (Boos 1992). However, one would expect to see the liberalism of the generalized Wald interval and the inefficiency and conservatism of the generalized score interval to be present in estimating equations situations.

In this chapter we introduce a new confidence interval that is based on a modified version of the generalized score test. This new generalized score interval, which, for lack of a better term, we call the GS2 confidence interval, is motivated, like the length optimal interval

of Chapter 2, as method of reducing the conservatism of the generalized score interval. The idea behind the interval is quite simple. The GS2 statistic is just the first element of the score function squared,  $S_1^2(\psi, \hat{\eta})$ , divided by an estimate of its variance that relies only on the global estimates  $\hat{\psi}$  and  $\hat{\eta}$ . Thus, the GS2 confidence interval has the advantage that it is easier to compute than the generalized score interval, because the GS2 interval does not require the computation of restricted estimates of the nuisance parameters. Using simulations, we investigate the small sample properties of the GS2 confidence interval by applying it to the logistic and Poisson GEE examples from the previous chapters, using GEE models with the exchangeable and AR(1) working correlation matrices for the logistic and Poisson data, respectively. In the simulation results, we see that the GS2 confidence interval is preferable to the generalized score confidence interval, as its coverage is generally close to nominal, and is infinite in length much less frequently and usually has smaller median length.

## 3.2 Motivation

In order to identify one of the sources of the conservatism and inefficiency of the generalized score interval, we take another look at the form of the generalized score test statistic. Generalized score test statistics for hypotheses of the form  $H_0 : \psi = \psi_0$  may be written as,

$$\begin{aligned} T_{GS}(\psi_0) &= n^{-1} S^T(\tilde{\theta}) \hat{I}^{-1}(\tilde{\theta}) S(\tilde{\theta}) \cdot \frac{\{\hat{I}^{-1}(\tilde{\theta})\}_{11}}{\{\hat{I}^{-1}(\tilde{\theta}) \hat{D}(\tilde{\theta}) \hat{I}^{-1}(\tilde{\theta})\}_{11}} \\ &= n^{-1} \frac{S_1^2(\tilde{\theta})}{\{\hat{I}^{-1}(\tilde{\theta})\}_{11}^{-2} \{\hat{I}^{-1}(\tilde{\theta}) \hat{D}(\tilde{\theta}) \hat{I}^{-1}(\tilde{\theta})\}_{11}}, \end{aligned} \tag{3.1}$$

where  $\tilde{\theta} = (\psi_0, \eta^T(\tilde{\psi}_0))^T$ . The expression in the second line is a decomposition of the generalized score test statistic into a *mean* part and a *variance* part, corresponding to the numerator and denominator of the expression, respectively. This is seen by noting that the expected value of  $\sqrt{n}S(\theta) = 0$ , and a Taylor series argument shows that the expression in the denominator is the variance of  $\sqrt{n}S_1(\tilde{\theta})$ .

In some situations, the denominator, or variance part, of the generalized score test statistic increases as the value of the null hypothesis moves away from  $\hat{\psi}$  at a rate that is great enough that the overall value of the test statistic is unable to exceed the specified critical value, causing the confidence interval to be infinite in length. Figure 3.1 is a plot of the numerator and denominator of the generalized score test statistic for the subset of the Epilepsy data whose generalized score confidence interval for the treatment effect was left infinite,  $(-\infty, .43)$ . We see that the denominator part of the test statistic is increasing as the null hypothesis value moves away from the estimate of the treatment effect,  $-0.34$ , and that the plot of the numerator's value flattens out on the left side of the graph. This suggests that one way to reduce the conservatism of the generalized score interval is to not allow the denominator of the test statistic to vary with the null hypothesis; instead we use unrestricted estimates in the denominator part of the generalized score test statistic. The GS2 interval is obtained by inverting a modified generalized score test statistic that places  $\hat{\theta}$  in its denominator, but also goes further by not using restricted estimates of the nuisance parameter in the numerator. This changes the asymptotics of  $n^{-1/2}S_1(\psi_0, \hat{\eta})$  somewhat, so the functional form of the GS2 test statistic, which we describe formally in the next section, differs from that of the standard generalized score test statistic. In the current example, the GS2 treatment effect confidence interval is  $(-1.3, .13)$ , which is more informative than the generalized score interval,  $(-\infty, .43)$ .

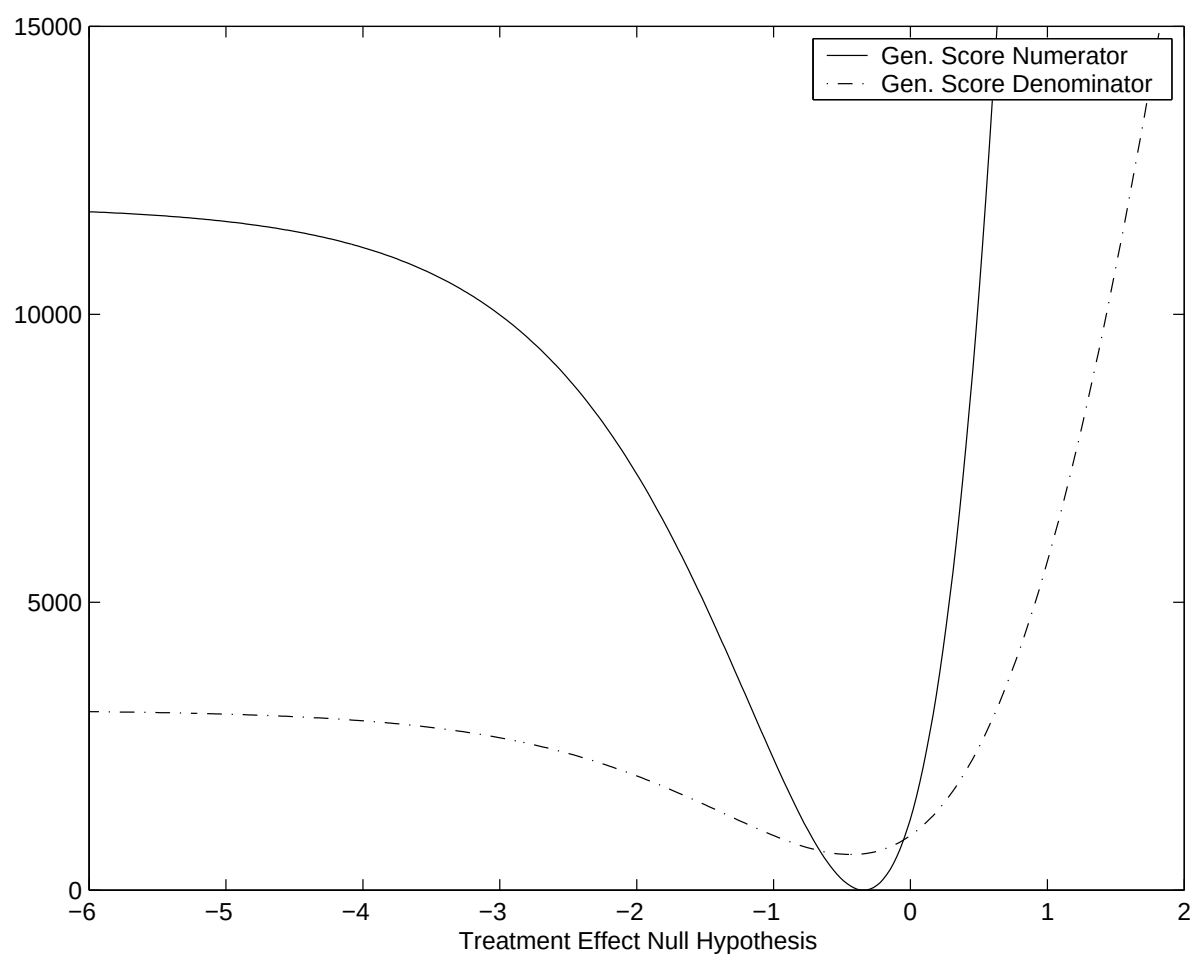


Figure 3.1: Numerator and denominator of the generalized score test statistic as a function of the null hypothesis value for a 33 cluster subset of the Epilepsy data. Solid and dotted lines indicate numerator and denominator values, respectively.

### 3.3 Theory

We use a slightly different notation than in previous chapters, to reflect that we are no longer necessarily working in a situation where there is a likelihood. We denote the estimating equation by,

$$S(\theta) = \sum_{i=1}^n U(y_i|\theta, \mathbf{x}_i), \quad (3.2)$$

where  $U$  is the function that makes up the individual observational components of the estimating equation. The estimate of  $\theta$ ,  $\hat{\theta} = (\hat{\psi}, \hat{\eta}^T)^T$ , is defined to be the solution to  $S(\hat{\theta}) = \mathbf{0}$ .

Under mild regularity conditions, the asymptotic variance of  $\hat{\theta}$  is  $n^{-1}A^{-1}(\theta_0)D(\theta_0)A^{-1}(\theta_0)$ , where matrices  $A$  and  $D$  are defined as

$$A(\theta) = -E_g \left[ \frac{\partial}{\partial \theta^T} U(y|\theta, \mathbf{x}) \right], \quad (3.3)$$

and

$$D(\theta) = E_g [U(y|\theta, \mathbf{x})U(y|\theta, \mathbf{x})^T]. \quad (3.4)$$

We use  $g(\mathbf{Y}, \mathbf{X}|\theta_0)$  to denote the unknown generating distribution for the data. We assume that the  $x_i$ 's are random, and not necessarily independent of the  $y_i$ . We also assume that the observation and covariate pairs  $(y_i, x_i)$ , of different observations are independent of one another, and that they are identically distributed. Because the true distribution  $g$  is unknown, estimates for  $A$  and  $D$  need to be employed,

$$\hat{A}(\theta) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta^T} U(y_i|\theta, \mathbf{x}_i), \quad (3.5)$$

and

$$\hat{D}(\theta) = \frac{1}{n} \sum_{i=1}^n U(y_i|\theta, \mathbf{x}_i) U(y_i|\theta, \mathbf{x}_i)^T. \quad (3.6)$$

It is important to point out that  $\hat{A}(\theta)$  and  $\hat{D}(\theta)$  are not, strictly speaking estimates, but are estimates of  $A(\theta)$  and  $D(\theta)$ , considered as *functions* of  $\theta$ .

The generalized Wald and generalized score tests of  $H_0 : \psi = \psi_0$ , in the context of estimating equations, are defined in the same manner as in likelihood situations, with the  $\hat{A}$  replacing the model information matrix in all expressions,

$$T_{GW}(\psi_0) = \frac{(\psi_0 - \hat{\psi})^2}{\{\hat{A}^{-1}(\hat{\theta}) \hat{D}(\hat{\theta}) \hat{A}^{-1}(\hat{\theta})\}_{11}}, \quad (3.7)$$

and

$$T_{GS}(\psi_0) = n^{-1} S^T(\tilde{\theta}) \hat{A}^{-1}(\tilde{\theta}) S(\tilde{\theta}) \cdot \frac{\{\hat{A}^{-1}(\tilde{\theta})\}_{11}}{\{\hat{A}^{-1}(\tilde{\theta}) \hat{D}(\tilde{\theta}) \hat{A}^{-1}(\tilde{\theta})\}_{11}}. \quad (3.8)$$

The test statistic that is inverted in the formation of GS2 intervals has the following form,

$$T_{GS2}(\psi_0) = n^{-1} \frac{S_1^2(\psi_0, \hat{\eta})}{\{\hat{A}(\psi_0, \hat{\eta})\}_{11}^2 \{\hat{A}^{-1}(\hat{\theta}) \hat{D}(\hat{\theta}) \hat{A}^{-1}(\hat{\theta})\}_{11}}. \quad (3.9)$$

The GS2 test statistic is easier to compute than the generalized score test statistic. The denominator is constant with respect to  $\psi_0$ , as only unrestricted estimates of the parameter,  $\hat{\theta}$ , are employed in that part of the test statistic. In the numerator, unrestricted estimates of the nuisance parameter  $\hat{\eta}$  are used, but the  $\psi$  component of the parameter vector varies with the

null hypothesis value,  $\psi_0$ . For the creation of GS2 confidence intervals, we suggest using  $F_{1,n-p}$  critical values because in preliminary simulations GS2 intervals using  $F_{1,n-p}$  critical values had coverages that were closer to nominal than when  $\chi_1^2$  critical values were employed.

In the ensuing discussion, we demonstrate that the GS2 statistic is asymptotically  $\chi_1^2$  distributed. In order to obtain this result, we make use of several assumptions in the form of regularity conditions. Note that these conditions imply the existence of a sequence of solutions,  $\hat{\theta}_n$ , that is consistent for  $\theta_0$ , and we implicitly assume that  $\{\hat{\theta}_n\}$  is such a sequence throughout this section. These conditions also imply that  $\sqrt{n}(\hat{\theta} - \theta_0)$  is asymptotically normally distributed with mean 0 and variance  $A^{-1}(\theta_0)D(\theta_0)A^{-1}(\theta_0)$  (Van Der Vaart 1998), page 68. These regularity conditions are similar to those found in Presnell and Boos (2002).

**A 3.1** *The parameter space,  $\Theta$ , is a closed, connected subset of  $\mathbb{R}^p$ . The true value of the parameter,  $\theta_0$  lies in the interior of  $\Theta$ . The support of the random variables,  $(y, x)$ , does not depend on  $\theta_0$ .*

**A 3.2** *Each element of  $U$  is twice continuously differentiable, and satisfies*

$$E_g(U_j(y|\theta_0, x)) = 0.$$

**A 3.3** *The following expectations exist and are finite for all indices  $j, k, l \in \{1, 2, \dots, p\}$ :*

$$E_g \left| \frac{\partial}{\partial \theta_k} U_j(y|\theta, x) \right| < \infty,$$

$$E_g \left| \frac{\partial}{\partial \theta_k} U_j(y|\theta, x) \right|^2 < \infty,$$

$$E_g |U_j(y|\theta, x)U_k(y|\theta, x)| < \infty,$$

and

$$E_g |U_j(y|\theta, x)U_k(y|\theta, x)|^2 < \infty.$$

This ensures that  $\hat{A}(\theta_0) \xrightarrow{p} A(\theta_0)$ , by the SLLN.

**A 3.4** *There exists a function  $G(y, x)$ , such that for all  $\theta$  in an open neighborhood of  $\theta_0$ , and for all indices  $j, k, l \in \{1, 2, \dots, p\}$ ,*

$$\left| \frac{\partial^2}{\partial \theta_k \partial \theta_l} U_j(y|\theta, x) \right| < G(y, x), \quad (3.10)$$

$$\left| \frac{\partial}{\partial \theta_l} U_j(y|\theta, x) U_k(y|\theta, x) \right| < G(y, x). \quad (3.11)$$

Furthermore,  $G(y, x)$  satisfies  $E_g(G(y, x)) < \infty$ .

The fact that the same dominating function is used for different derivatives of  $U$  and all combinations of indices in A 3.4 is no loss of generality. This is because, if different dominating functions were used, we could construct  $G$  by setting it the pointwise maximum of this finite, fixed number of functions, and  $G$  defined in this way would be nonnegative, integrable, and would dominate all the necessary different derivatives of  $U$  (Presnell and Boos 2002).

We begin with a preliminary lemma concerning the convergence of the denominator of the GS2 statistic.

**Lemma 3.1** *Suppose that the data,  $(y_i, x_i)$ , are iid with pdf  $g(y_i, x_i|\theta_0)$  such that regularity*

conditions A 3.1, A 3.2, A 3.3, and A 3.4 hold. Then

$$\hat{A}_{11}^2(\hat{\theta})\{\hat{A}^{-1}(\hat{\theta})\hat{D}(\hat{\theta})\hat{A}^{-1}(\hat{\theta})\}_{11} \xrightarrow{p} A_{11}^2(\theta_0)\{A^{-1}(\theta_0)D(\theta_0)A^{-1}(\theta_0)\}_{11}.$$

**Proof** We begin by taking a Taylor series of the individual elements of  $\hat{A}(\hat{\theta})$  about  $\theta_0$ ,

$$\begin{aligned}\hat{A}_{jk}(\hat{\theta}) &= -\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta_k} U_j(y|\hat{\theta}, x) \\ &= -\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta_k} U_j(y|\theta_0, x) - \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta^T} \frac{\partial}{\partial \theta_k} U_j(y|\theta^*, x)(\hat{\theta} - \theta_0),\end{aligned}\tag{3.12}$$

where  $\theta^*$  lies on the line segment connecting  $\hat{\theta}$  and  $\theta_0$ . The first term in the expansion is converging *wpl* to  $A_{jk}(\theta_0)$ , by the SLLN. The second term is  $O_p(n^{-1/2})$ , since

$$\begin{aligned}\left| \frac{\sqrt{n}}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta^T} \frac{\partial}{\partial \theta_k} U_j(y|\theta^*, x)(\hat{\theta} - \theta_0) \right| &= \left| \frac{\sqrt{n}}{n} \sum_{l=1}^p \sum_{i=1}^n \frac{\partial}{\partial \theta_l} \frac{\partial}{\partial \theta_k} U_j(y|\theta^*, x)(\hat{\theta}_l - \theta_{0l}) \right| \\ &\leq \sqrt{n} \sum_{l=1}^p \sum_{i=1}^n \left| \frac{1}{n} \frac{\partial}{\partial \theta_l} \frac{\partial}{\partial \theta_k} U_j(y|\theta^*, x) \right| |(\hat{\theta}_l - \theta_{0l})| \\ &\leq \frac{1}{n} \sum_{i=1}^n G(y_i, x_i) \sum_{l=1}^p \left| \sqrt{n}(\hat{\theta}_l - \theta_{0l}) \right|.\end{aligned}\tag{3.13}$$

The right hand side of the final line of 3.13 is  $O_p(1)$ , because it is the product of a  $p$ -term sum of  $O_p(1)$  random variables multiplied by  $n^{-1} \sum_{i=1}^n G(y_i, x_i)$ , which converges *wpl* to its expectation. Thus,

$$\begin{aligned}\hat{A}_{jk}(\hat{\theta}) &= A_{jk}(\theta_0) + o_p(1) + O_p(n^{-1/2}), \\ &= A_{jk}(\theta_0) + o_p(1)\end{aligned}$$

which gives us the result that  $\hat{A}_{jk}(\hat{\theta}) \xrightarrow{p} A_{jk}(\theta_0)$ .

Using nearly identical reasoning, one can show that  $\hat{D}_{jk}(\hat{\theta}) \xrightarrow{p} D_{jk}(\theta_0)$ , but for the sake of brevity we omit the proof.

Now,  $\hat{A}_{11}^2(\hat{\theta})\{\hat{A}^{-1}(\hat{\theta})\hat{D}(\hat{\theta})\hat{A}^{-1}(\hat{\theta})\}_{11}$  is a continuous function of the  $p(3p-1)/2$  random variables  $\{\hat{A}_{jk}(\hat{\theta})\}_{j=1k=1}^p$ , and  $\{\hat{D}_{jk}(\hat{\theta})\}_{j=1k=j}^p$ . As these variables are converging in probability to  $\{A_{jk}(\theta_0)\}_{j=1k=1}^p$ , and  $\{D_{jk}(\theta_0)\}_{j=1k=j}^p$ , respectively, we obtain the result that

$$\hat{A}_{11}^2(\hat{\theta})\{\hat{A}^{-1}(\hat{\theta})\hat{D}(\hat{\theta})\hat{A}^{-1}(\hat{\theta})\}_{11} \xrightarrow{p} A_{11}^2(\theta_0)\{A^{-1}(\theta_0)D(\theta_0)A^{-1}(\theta_0)\}_{11},$$

completing the proof.  $\square$

**Theorem 3.1** *Suppose that the data,  $(y_i, x_i)$ , are iid with pdf  $g(y_i, x_i|\theta_0)$  such that regularity conditions A 3.1, A 3.2, A 3.3, and A 3.4 hold. Then  $T_{GS2}(\psi_0)$  given in 3.9 is asymptotically  $\chi_1^2$  distributed.*

**Proof** We first show that  $n^{-1/2}S(\psi_i, \hat{\eta})$  is asymptotically distributed as  $N(0, A_{11}^2\{A^{-1}DA^{-1}\}_{11})$ . Consider the following Taylor series expansion of  $S(\psi_0, \hat{\eta})$  about  $\hat{\psi}$ ,

$$\begin{aligned} n^{-1/2} \sum_{i=1}^n U_1(y_i|\psi_0, \hat{\eta}, x_i) &= n^{-1/2} \sum_{i=1}^n U_1(y_i|\hat{\psi}, \hat{\eta}, x_i) + \frac{1}{n} \sum_{i=1}^n \left[ \frac{\partial}{\partial \psi} U_1(y_i|\hat{\psi}, \hat{\eta}, x_i) \right] \sqrt{n}(\psi_0 - \hat{\psi}) \\ &\quad + \frac{1}{2n} \sum_{i=1}^n \left[ \frac{\partial^2}{\partial \psi^2} U_1(y_i|\psi^*, \hat{\eta}, x_i) \right] \sqrt{n}(\psi_0 - \hat{\psi})^2, \end{aligned} \tag{3.14}$$

where  $\psi^*$  lies between  $\psi_0$  and  $\hat{\psi}$ .

Observe that the first component of the right hand side of 3.14 is identically equal to zero, because it is the first component of the estimating equation with the  $\hat{\theta}$  inserted for  $\theta$ .

Because of regularity condition A 3.4, and the fact that  $\sqrt{n}(\psi_0 - \hat{\psi}) = O_p(1)$  by Theorem 5.41 of Van Der Vaart (1998) page 68, we see that the last term on the right hand side is  $O_p(n^{-1/2})$  by an argument similar to that of Lemma 3.1. After expanding the second term of 3.14 about  $\theta_0$  we have

$$\begin{aligned} n^{-1/2} \sum_{i=1}^n U_1(y_i|\psi_0, \hat{\eta}, x_i) &= \frac{1}{n} \sum_{i=1}^n \left[ \frac{\partial}{\partial \psi} U_1(y_i|\theta_0, x_i) \right] \sqrt{n}(\psi_0 - \hat{\psi}) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left[ \frac{\partial^2}{\partial \psi \partial \theta^T} U_1(y_i|\theta^{**}, x_i) \right] \sqrt{n}(\hat{\theta} - \theta_0)(\psi_0 - \hat{\psi}) \\ &\quad + O_p(n^{-1/2}), \end{aligned} \quad (3.15)$$

where  $\theta^{**}$  lies on the line segment between  $\theta$  and  $\hat{\theta}$ , we see that the second term on the right hand side of the above expression is also  $O_p(n^{-1/2})$ , again by the same argument used in Lemma 3.1.

Then, by adding and subtracting  $A_{11}$  in the first term of the expansion,

$$\begin{aligned} n^{-1/2} \frac{1}{n} \sum_{i=1}^n U_1(y_i|\psi_0, \hat{\eta}, x_i) &= -A_{11} \sqrt{n}(\psi_0 - \hat{\psi}) \\ &\quad + \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \psi} U_1(y_i|\theta_0, x_i) + A_{11} \right) \sqrt{n}(\psi_0 - \hat{\psi}) \\ &\quad + O_p(n^{-1/2}), \end{aligned} \quad (3.16)$$

we recognize that the second term on the right is  $o_p(1)$ . This gives us the result that,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n U_1(y_i|\psi_0, \hat{\eta}, x_i) = -A_{11} \sqrt{n}(\psi_0 - \hat{\psi}) + o_p(1). \quad (3.17)$$

The result then follows from fact that  $\psi_0 - \hat{\psi}$  is asymptotically  $N(0, \{A^{-1}DA^{-1}\}_{11})$ , Lemma 3.1, and Slutsky's Theorem.  $\square$

## 3.4 Simulations

We repeat the logistic and Poisson GEE simulations of the previous chapters. Unlike in previous simulations, we estimate the parameters using working correlation matrices that are not equal to the identity matrix, so the estimating procedure is no longer equivalent to one under any likelihood model. We also apply the length-optimal procedure to the generalized score and GS2 intervals, and see that it improves the GS2 and generalized score intervals in the context of binary data, but has little effect on these confidence intervals for the Poisson data simulations.

### 3.4.1 Simple Linear Logistic GEE With Exchangeable Working Correlation Matrix

In the Tables 3.1 and 3.3 below we see the results for simple linear regression GEE simulations using the working correlation matrix. The GS2 interval performs consistently well, in terms of its coverage, except when  $\psi = 1$  and the within cluster correlation is  $\rho = .5$ . In this situation, however, the other intervals perform even more poorly. When the sample size is very small,  $K = 05$ , it is clear that the GS2 interval is quite inefficient, but in larger sample sizes it is clearly preferable to the other alternatives. The generalized Wald- $t$  interval's coverage is lower than that of the GS2 interval, while the generalized score's coverage is at or above the nominal rate, it is infinite in length too often for it to be considered competitive with the GS2 interval. The GS2 interval, while still infinite in length on occasion, is infinite with substantially less frequency than the generalized score interval, and has a generally lower median length. The length-optimal GS2 interval's coverage is better than that of the equal-tail GS2 confidence

interval, and is recommended whenever sufficient computing resources are available.

Table 3.1: Coverage and length for generalized Wald, score, and GS2 intervals for Bernoulli data under a simple linear logistic regression model.

| $(\psi, \eta)$ | # of Clusters $K =$ |                 | $\rho = 0$ |     |     | $\rho = 0.2$ |          |     | $\rho = 0.5$ |          |     |
|----------------|---------------------|-----------------|------------|-----|-----|--------------|----------|-----|--------------|----------|-----|
|                |                     |                 | 5          | 10  | 20  | 5            | 10       | 20  | 5            | 10       | 20  |
| (0,0)          | Coverage            | Gen. Wald       | .87        | .90 | .93 | .84          | .89      | .93 | .85          | .91      | .94 |
|                |                     | Gen. Wald- $t$  | .95        | .93 | .94 | .94          | .93      | .94 | .93          | .94      | .95 |
|                |                     | Gen. Score      | .98        | .95 | .94 | .98          | .95      | .95 | .99          | .97      | .96 |
|                |                     | Opt. Gen. Score | .99        | .97 | .96 | .99          | .97      | .96 | .99          | .98      | .96 |
|                |                     | GS2             | .97        | .94 | .95 | .97          | .94      | .95 | .96          | .96      | .95 |
|                |                     | Opt. GS2        | .96        | .96 | .96 | .96          | .96      | .96 | .96          | .96      | .96 |
|                | Median Length       | Gen. Wald       | 2.9        | 2.1 | 1.5 | 2.6          | 2.0      | 1.4 | 2.1          | 1.7      | 1.2 |
|                |                     | Gen. Wald- $t$  | 4.1        | 2.4 | 1.6 | 3.7          | 2.3      | 1.5 | 3.0          | 2.0      | 1.3 |
|                |                     | Gen. Score      | $\infty$   | 3.3 | 1.8 | $\infty$     | 3.6      | 1.7 | $\infty$     | $\infty$ | 1.5 |
|                |                     | Opt. Gen. Score | $\infty$   | 3.3 | 1.8 | $\infty$     | 3.3      | 1.7 | $\infty$     | $\infty$ | 1.5 |
|                |                     | GS2             | 15.6       | 3.0 | 1.8 | 13.0         | 2.9      | 1.7 | 6.2          | 2.4      | 1.4 |
|                |                     | Opt. GS2        | 12.9       | 2.9 | 1.7 | 9.1          | 2.9      | 1.7 | 5.6          | 2.4      | 1.4 |
| (1,0)          | Coverage            | Gen. Wald       | .84        | .89 | .92 | .81          | .91      | .92 | .71          | .87      | .92 |
|                |                     | Gen. Wald- $t$  | .92        | .94 | .94 | .90          | .94      | .94 | .79          | .90      | .93 |
|                |                     | Gen. Score      | .97        | .96 | .94 | .99          | .96      | .94 | .95          | .95      | .95 |
|                |                     | Opt. Gen. Score | .99        | .94 | .96 | .99          | .97      | .96 | .99          | .95      | .96 |
|                |                     | GS2             | .95        | .95 | .94 | .93          | .95      | .94 | .83          | .91      | .94 |
|                |                     | Opt. GS2        | .95        | .95 | .96 | .93          | .96      | .95 | .83          | .92      | .95 |
|                | Median Length       | Gen. Wald       | 3.1        | 2.2 | 1.6 | 2.7          | 2.0      | 1.5 | 2.9          | 2.1      | 1.5 |
|                |                     | Gen. Wald- $t$  | 4.5        | 2.6 | 1.7 | 3.8          | 2.3      | 1.6 | 4.0          | 2.4      | 1.6 |
|                |                     | Gen. Score      | $\infty$   | 3.8 | 1.9 | $\infty$     | $\infty$ | 1.8 | $\infty$     | $\infty$ | 2.0 |
|                |                     | Opt. Gen. Score | $\infty$   | 3.6 | 1.9 | $\infty$     | 4.1      | 1.8 | $\infty$     | $\infty$ | 1.9 |
|                |                     | GS2             | 16.5       | 3.3 | 1.9 | 13.6         | 3.0      | 1.7 | 12.7         | 3.1      | 1.8 |
|                |                     | Opt. GS2        | 14.5       | 3.2 | 1.9 | 11.1         | 3.0      | 1.7 | 9.9          | 3.0      | 1.8 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Cluster size=4.

Table 3.2: Proportion of infinite length generalized score and GS2 intervals for Bernoulli data under simple a linear logistic regression model.

| $(\psi, \eta)$ |            | # of Clusters $K =$ | $\rho = 0$ |     |    | $\rho = 0.2$ |     |    | $\rho = 0.5$ |     |     |
|----------------|------------|---------------------|------------|-----|----|--------------|-----|----|--------------|-----|-----|
|                |            |                     | 5          | 10  | 20 | 5            | 10  | 20 | 5            | 10  | 20  |
| (0,0)          | Proportion | Gen. Score          | .99        | .11 | 0  | .99          | .29 | 0  | .99          | .63 | .01 |
|                | Infinite   | Opt. Gen. Score     | .96        | .06 | 0  | .95          | .20 | 0  | .99          | .53 | 0   |
|                |            | GS2                 | .01        | 0   | 0  | .01          | 0   | 0  | 0            | 0   | 0   |
|                |            | Opt. GS2            | 0          | 0   | 0  | .01          | 0   | 0  | 0            | 0   | 0   |
| (1,0)          | Proportion | Gen. Score          | .99        | .20 | 0  | .99          | .54 | 0  | .99          | .80 | .23 |
|                | Infinite   | Gen. Score          | .98        | .10 | 0  | .96          | .39 | 0  | .99          | .70 | .12 |
|                |            | GS2                 | .02        | 0   | 0  | .03          | 0   | 0  | .04          | 0   | 0   |
|                |            | Opt. GS2            | 0          | 0   | 0  | .02          | 0   | 0  | .03          | 0   | 0   |

### 3.4.2 GUIDE Data Example and Simulations

The differences between the confidence intervals under consideration are more pronounced when we apply them to the GUIDE data. Table 3.3 lists the various confidence intervals for the parameters in the analysis of the GUIDE data set.

Table 3.3: Logistic GEE with exchangeable working correlation matrix estimates and approximate 95% confidence intervals parameters from the GUIDE dataset.

| Parameter | Estimate | Gen. Wald- $t$ | Gen. Score   | Opt. Gen. Score | GS2         | Opt. GS2    |
|-----------|----------|----------------|--------------|-----------------|-------------|-------------|
| SEX       | -.78     | (-2.0, .48)    | (-2.1, 1.4)  | (-2.2, 1.0)     | (-2.9, .52) | (-2.2, 1.0) |
| AGE       | -.75     | (-1.9, .36)    | (-2.0, .58)  | (-2.0, .57)     | (-2.5, .42) | (-2.3, .55) |
| DAYACC    | .39      | (.20, .57)     | (.23, .71)   | (.20, .65)      | (.22, .76)  | (-.19, .68) |
| SEVERE    | .82      | (.12, 1.6)     | (.053, 1.8)  | (0.0, 1.8)      | (-.14, 1.8) | (-.14, 1.8) |
| TOILET    | .076     | (-.096, .25)   | (-.067, .39) | (-.088, .35)    | (-.15, .30) | (-.15, .30) |

In the simulation results, we see that the generalized Wald- $t$  interval, which was competitive in the simple linear simulations, performs rather poorly in this context, especially with regard to the SEX effect parameter. The generalized score interval is clearly inefficient, and is certainly not preferable to the GS2 interval. The GS2 performs well, except that when the number of clusters is 20, the SEX effect parameter's confidence interval is infinite about 20%-30% of the time. The length-optimal interval is able to overcome this moderately well, as it is infinite in length 10%-15% of the time, and also has higher coverage than the equal-tail GS2 interval.

Table 3.4: Coverage and length for generalized Wald, score, and GS2 intervals for clustered Bernoulli data simulated from the GUIDE dataset.

| Parameter | # of Clusters $K =$ |                 | $\rho = 0$ |     | $\rho = 0.2$ |     |
|-----------|---------------------|-----------------|------------|-----|--------------|-----|
|           |                     |                 | 20         | 37  | 20           | 37  |
| SEX       | Coverage            | Gen. Wald       | .86        | .90 | .83          | .90 |
|           |                     | Gen. Wald- $t$  | .89        | .91 | .85          | .91 |
|           |                     | Gen. Score      | .96        | .94 | .97          | .95 |
|           |                     | Opt. Gen. Score | .94        | .94 | .97          | .97 |
|           |                     | GS2             | .94        | .94 | .92          | .95 |
|           |                     | Opt. GS2        | .95        | .95 | .93          | .95 |
|           | Median              | Gen. Wald       | 3.1        | 2.2 | 2.7          | 2.2 |
|           |                     | Gen. Wald- $t$  | 3.4        | 2.3 | 2.9          | 2.3 |
|           |                     | Gen. Score      | $\infty$   | 3.2 | $\infty$     | 3.5 |
|           |                     | Opt. Gen. Score | $\infty$   | 3.1 | $\infty$     | 3.4 |
|           |                     | GS2             | 5.2        | 2.8 | 4.2          | 2.8 |
|           |                     | Opt. GS2        | 4.9        | 2.7 | 4.1          | 2.8 |
| AGE       | Coverage            | Gen. Wald       | .90        | .92 | .88          | .93 |
|           |                     | Gen. Wald- $t$  | .92        | .93 | .90          | .93 |
|           |                     | Gen. Score      | .93        | .94 | .91          | .95 |
|           |                     | Opt. Gen. Score | .96        | .94 | .99          | .96 |
|           |                     | GS2             | .94        | .94 | .94          | .95 |
|           |                     | Opt. GS2        | .95        | .96 | .94          | .96 |
|           | Median              | Gen. Wald       | 3.1        | 2.2 | 2.8          | 2.1 |
|           |                     | Gen. Wald- $t$  | 3.4        | 2.3 | 3.0          | 2.2 |
|           |                     | Gen. Score      | 5.2        | 2.7 | $\infty$     | 2.6 |
|           |                     | Opt. Gen. Score | 4.9        | 2.6 | 6.9          | 2.6 |
|           |                     | GS2             | 5.2        | 2.7 | 5.0          | 2.6 |
|           |                     | Opt. GS2        | 5.0        | 2.7 | 4.7          | 2.5 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Table 3.5: Proportion of infinite length generalized score and GS2 intervals for clustered Bernoulli data simulated from the GUIDE dataset.

|           |            |                     | $\rho = 0$ |     | $\rho = 0.2$ |     |
|-----------|------------|---------------------|------------|-----|--------------|-----|
| Parameter |            | # of Clusters $K =$ | 20         | 37  | 20           | 37  |
| SEX       | Proportion | Gen. Score          | .78        | .03 | .89          | .18 |
|           | Infinite   | Opt. Gen. Score     | .65        | .01 | .80          | .10 |
|           |            | GS2                 | .26        | .01 | .19          | .03 |
|           |            | Opt. GS2            | .14        | 0   | .11          | .01 |
| AGE       | Proportion | Gen. Score          | .21        | .01 | .91          | .06 |
|           | Infinite   | Opt. Gen. Score     | .12        | 0   | .47          | .05 |
|           |            | GS2                 | 0          | 0   | 0            | 0   |
|           |            | Opt. GS2            | 0          | 0   | 0            | 0   |

### 3.4.3 Simple Linear Poisson GEE With AR(1) Working Correlation Matrix

Tables 3.6 and 3.7 contain results from simulations using the same data generating mechanism as in the simple linear Poisson GEE simulations of the previous two chapters. The data are serially correlated in clusters, but are marginally Poisson distributed conditional on the prior observation. In this instance, the parameters are estimated using a GEE model using the autoregressive, or AR(1), working correlation matrix. We see in the results below that the GS2 interval performs consistently well, better than the generalized Wald- $t$  interval, while not appearing to suffer from a lack of efficiency. This is in contrast to the generalized score interval, which has coverages generally near or at nominal, but is clearly inefficient in comparison to the GS2. The optimal-length GS2 interval performs quite similarly to the equal-tail version, and the gain in efficiency is probably not worth the extra amount of computational effort.

Table 3.6: Coverage and length for generalized Wald, score, and GS2 intervals for serially correlated Poisson data under a log linear regression model.

|                |          |                     | $\rho = 0$ |     |     | $\rho = .2$ |     |     | $\rho = .5$ |     |     |
|----------------|----------|---------------------|------------|-----|-----|-------------|-----|-----|-------------|-----|-----|
| $(\psi, \eta)$ |          | # of Clusters $K =$ | 10         | 20  | 50  | 10          | 20  | 50  | 10          | 20  | 50  |
| (0,0)          | Coverage | Gen. Wald           | .89        | .93 | .93 | .88         | .92 | .94 | .87         | .93 | .94 |
|                |          | Gen. Wald- $t$      | .92        | .94 | .94 | .92         | .94 | .94 | .91         | .94 | .94 |
|                |          | Gen. Score          | .96        | .96 | .94 | .95         | .96 | .95 | .96         | .96 | .95 |
|                |          | Opt. Gen. Score     | .96        | .96 | .94 | .97         | .96 | .95 | .97         | .97 | .95 |
|                |          | GS2                 | .97        | .96 | .94 | .96         | .96 | .95 | .95         | .96 | .95 |
|                |          | Opt. GS2            | .96        | .96 | .94 | .95         | .95 | .95 | .95         | .96 | .95 |
|                | Length   | Gen. Wald           | .96        | .72 | .47 | 1.0         | .75 | .49 | 1.1         | .80 | .52 |
|                |          | Gen. Wald- $t$      | 1.1        | .77 | .48 | 1.2         | .80 | .51 | 1.2         | .86 | .54 |
|                |          | Gen. Score          | 1.5        | .86 | .50 | 1.6         | .91 | .53 | 2.0         | 1.0 | .57 |
|                |          | Opt. Gen. Score     | 1.5        | .85 | .50 | 1.6         | .90 | .53 | 1.8         | 1.0 | .57 |
|                |          | GS2                 | 1.2        | .80 | .49 | 1.3         | .84 | .52 | 1.4         | .89 | .54 |
|                |          | Opt. GS2            | 1.2        | .80 | .49 | 1.3         | .84 | .52 | 1.3         | .89 | .54 |
| (1,0)          | Coverage | Gen. Wald           | .89        | .91 | .94 | .88         | .92 | .93 | .88         | .91 | .94 |
|                |          | Gen. Wald- $t$      | .93        | .93 | .95 | .93         | .93 | .94 | .92         | .93 | .94 |
|                |          | Gen. Score          | .96        | .95 | .95 | .95         | .94 | .95 | .95         | .95 | .95 |
|                |          | Opt. Gen. Score     | .97        | .95 | .96 | .97         | .95 | .94 | .96         | .95 | .95 |
|                |          | GS2                 | .96        | .94 | .96 | .96         | .95 | .95 | .95         | .95 | .95 |
|                |          | Opt. GS2            | .96        | .94 | .96 | .96         | .95 | .94 | .95         | .95 | .95 |
|                | Length   | Gen. Wald           | .97        | .73 | .47 | 1.0         | .78 | .50 | 1.1         | .82 | .54 |
|                |          | Gen. Wald- $t$      | 1.1        | .78 | .48 | 1.2         | .83 | .52 | 1.2         | .88 | .55 |
|                |          | Gen. Score          | 1.5        | .86 | .50 | 1.7         | .93 | .54 | 1.9         | 1.0 | .58 |
|                |          | Opt. Gen. Score     | 1.5        | .86 | .50 | 1.6         | .93 | .54 | 1.8         | 1.0 | .58 |
|                |          | GS2                 | 1.3        | .83 | .49 | 1.4         | .88 | .53 | 1.4         | .93 | .56 |
|                |          | Opt. GS2            | 1.3        | .82 | .49 | 1.3         | .88 | .53 | 1.4         | .92 | .56 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Cluster size=4.



### 3.4.4 Epilepsy Data Example and Simulations

In Table 3.8 we see the 95% confidence intervals for the Epilepsy data for the parameters of interest when the estimation proceeds using a GEE model using the AR(1) working correlation matrix.

Table 3.8: Poisson GEE AR(1) working correlation matrix estimates and approximate 95% GEE confidence intervals for the Epilepsy dataset.

| Parameter | MLE  | Gen. Wald- $t$ | Gen. Score   | Opt. Gen. Score | GS2         | Opt. GS2    |
|-----------|------|----------------|--------------|-----------------|-------------|-------------|
| TREATMENT | -.13 | (-.63,.38)     | (-.89,.32)   | (-.77,.39)      | (-.88,.31)  | (-.77,.37)  |
| AGE       | -.81 | (-1.5,-.092)   | (-2.0,-.083) | (-1.9,-.058)    | (-1.8,-.13) | (-1.8,-.12) |
| BASELINE  | -1.2 | (-1.5,-.99)    | (-1.5,-.94)  | (-1.5,-.94)     | (-1.6,-.99) | (-1.5,-.97) |

In our final batch of simulation results for the GS2 interval, we see that, like in the preceding simple linear Poisson GEE simulations, the GS2 interval has better coverage than the generalized Wald- $t$  interval, while not having an unacceptably high median length. The GS2 interval's coverage is almost uniformly superior to that of the generalized score interval, also. Surprisingly, for the in the case of the treatment effect confidence intervals, the GS2 interval is 1%-5% longer than the the generalized score interval. This is acceptable, however, as the GS2 interval's coverage is closer to 95% than that of the generalized score interval, in that situation. Like the previous simulations, it appears that little is gained by using the length-optimal GS2 interval over the equal-tail version.

Table 3.9: Coverage and length for generalized Wald, score, and GS2 intervals for serially correlated Poisson data using the Epilepsy data covariates and MLE.

|           |               |                     | $\rho = 0$ |     | $\rho = .2$ |     | $\rho = .5$ |     |
|-----------|---------------|---------------------|------------|-----|-------------|-----|-------------|-----|
| Parameter |               | # of Clusters $K =$ | 30         | 59  | 30          | 59  | 30          | 59  |
| TREATMENT | Coverage      | Gen. Wald           | .93        | .94 | .92         | .94 | .91         | .92 |
|           |               | Gen. Wald- $t$      | .94        | .94 | .93         | .94 | .92         | .93 |
|           |               | Gen. Score          | .94        | .95 | .94         | .94 | .93         | .94 |
|           |               | Opt. Gen. Score     | .95        | .95 | .94         | .94 | .93         | .94 |
|           |               | GS2                 | .95        | .95 | .95         | .95 | .94         | .95 |
|           |               | Opt. GS2            | .96        | .95 | .95         | .95 | .94         | .95 |
|           | Median Length | Gen. Wald           | .24        | .17 | .28         | .20 | .36         | .26 |
|           |               | Gen. Wald- $t$      | .25        | .18 | .29         | .21 | .38         | .27 |
|           |               | Gen. Score          | .26        | .18 | .30         | .21 | .39         | .27 |
|           |               | Opt. Gen. Score     | .26        | .18 | .30         | .21 | .39         | .27 |
|           |               | GS2                 | .27        | .18 | .31         | .21 | .41         | .28 |
|           |               | Opt. GS2            | .27        | .18 | .31         | .21 | .41         | .28 |
| AGE       | Coverage      | Gen. Wald           | .91        | .93 | .88         | .92 | .88         | .92 |
|           |               | Gen. Wald- $t$      | .92        | .93 | .90         | .93 | .89         | .92 |
|           |               | Gen. Score          | .96        | .95 | .93         | .94 | .92         | .94 |
|           |               | Opt. Gen. Score     | .97        | .95 | .93         | .94 | .93         | .94 |
|           |               | GS2                 | .95        | .95 | .93         | .94 | .93         | .94 |
|           |               | Opt. GS2            | .95        | .95 | .93         | .94 | .93         | .94 |
|           | Median Length | Gen. Wald           | .27        | .23 | .31         | .26 | .37         | .32 |
|           |               | Gen. Wald- $t$      | .29        | .24 | .33         | .27 | .39         | .32 |
|           |               | Gen. Score          | .39        | .26 | .42         | .30 | .51         | .36 |
|           |               | Opt. Gen. Score     | .37        | .26 | .41         | .30 | .49         | .36 |
|           |               | GS2                 | .31        | .25 | .35         | .28 | .42         | .33 |
|           |               | Opt. GS2            | .31        | .25 | .35         | .28 | .42         | .33 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

Cluster size=5. No infinite intervals were reported.

## Chapter 4

# Bootstrap Calibrated Generalized Wald Confidence Intervals

### 4.1 Introduction

In this chapter we investigate the small sample properties, via simulation, of two computationally intensive methods of generating confidence intervals using the bootstrap resampling method. We compare these confidence intervals to the generalized Wald and GS2 confidence intervals, and show that one of the bootstrap intervals, the calibrated generalized Wald interval is often superior to the GS2 confidence interval.

### 4.2 Bootstrap Confidence Intervals

Since its introduction (Efron 1979), the bootstrap has received a tremendous amount of attention. The main appeal bootstrap methods is that they can sometimes significantly im-

prove standard statistical methods, or suggest new methods that perform better than standard techniques. One of the main advantages of the bootstrap is that in many cases implementing the bootstrap is quite simple. Often, computer programs that implement these methods can be written with only a few more lines of code than other, standard methods. However, most bootstrap methods are hundreds, if not thousands, of times more computationally intensive than classical methods such as maximum likelihood. The basic idea of the bootstrap methodology is that one uses sampling from an estimate of the distribution function of the data to gain insight into, and correct for, the small sample behavior of a statistic. We focus on two bootstrap confidence interval estimation methods. One of these confidence intervals, the  $BC_a$  interval (Efron 1987), is generally regarded as being the premier bootstrap confidence interval. The other confidence interval is a bootstrap calibrated generalized Wald interval. Because we are interested in model-robust interval estimates, we will focus on non-parametric bootstrap methods, as parametric bootstrap techniques require that the generating distribution for the data is known to be a member of a particular parametric family.

#### 4.2.1 Bias-Corrected and Accelerated Bootstrap Confidence Intervals

The  $BC_a$ , or bias-corrected and accelerated, confidence interval is an improved version of an earlier bootstrap confidence interval, the percentile interval (Efron 1981), which for ease of exposition we will introduce first. The percentile interval for a parameter,  $\psi$ , is computed via a simple algorithmic procedure, as follows.

First, generate  $B$  bootstrap replicates of the data,  $\mathbf{Y}_b^*$ , where  $b = 1, 2, \dots, B$ , by independently drawing samples, with replacement, of size  $n$  from the original data,  $\mathbf{Y}$ . Note that this is equivalent to drawing  $B$  samples of size  $n$  from the empirical cdf of the data. From

each  $\mathbf{Y}_b^*$ , compute an estimate of the parameter of interest,  $\hat{\psi}_b^*$ . One then views this collection of bootstrap replications of  $\hat{\psi}$  as a random sample from the actual sampling distribution of  $\hat{\psi}$ . This suggests that a level- $\alpha$  confidence interval, the percentile interval, for  $\psi$  is  $(\hat{\psi}_{(B \cdot \alpha/2)}^*, \hat{\psi}_{(B - B \cdot \alpha/2)}^*)$ , where  $\hat{\psi}_{(B \cdot \alpha/2)}^*$ , and  $\hat{\psi}_{(B - B \cdot \alpha/2)}^*$  are the  $B \cdot \alpha/2$ -th and  $B - B \cdot \alpha/2$ -th order statistics of the bootstrap replicates of  $\hat{\psi}$ . Unfortunately, the percentile interval's coverage can be unacceptably low when the sample size is small for some simple problems, such as interval estimation for the variance of a normal population (Schenker 1985).

The  $BC_a$  confidence interval was developed in response to the shortcomings of the percentile interval. The  $BC_a$  interval is similar to the percentile interval, in that its upper and lower bounds are order statistics of the resamples of  $\hat{\psi}$ , but chooses different pair of order statistics to report as the interval's bounds. The  $BC_a$  interval has the form  $(\hat{\psi}_{(B \cdot \alpha_1)}^*, \hat{\psi}_{(B \cdot \alpha_2)}^*)$ , where

$$\alpha_1 = \Phi \left( z_0 + \frac{z_0 + z_{\alpha/2}}{1 - a(z_0 + z_{\alpha/2})} \right),$$

and

$$\alpha_2 = \Phi \left( z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})} \right).$$

The constants,  $z_0$ , and  $a$ , are estimated from the data as follows,

$$z_0 = \Phi^{-1} \left( \frac{\#\{\hat{\psi}_b^* < \hat{\psi}\}}{B} \right),$$

and

$$a = \frac{\sum_{i=1}^n (\hat{\psi}^{jack} - \hat{\psi}_i^{jack})^3}{6 \{ \sum_{i=1}^n (\hat{\psi}^{jack} - \hat{\psi}_i^{jack})^2 \}^{3/2}},$$

where  $\hat{\psi}_i^{jack}$  is the estimate of  $\psi$  with the  $i$ -th observation removed from the sample, and  $\hat{\psi}^{jack}$

is the mean of the  $\hat{\psi}_i^{jack}$ .

In some circumstances, the coverage probability of the  $BC_a$  interval converges to the nominal rate an order of magnitude faster than classical intervals, such as the Wald-type, score-type, and profile likelihood-type intervals we have been focusing on in this research (Hall 1988). That is, the actual coverage probability of the  $BC_a$  interval is, under certain relatively strong regularity conditions,  $1 - \alpha + O_p(n^{-1})$ , whereas the coverage probability of the classical intervals is  $1 - \alpha + O_p(n^{-1/2})$ . However, the proof that the coverage probability of the  $BC_a$  interval converges to  $1 - \alpha$  at the rate of  $O_p(n^{-1})$  relies on Edgeworth expansions that assume that the distribution of  $y$  is not discrete. This suggests the possibility that the classical methods of the previous chapters may, in fact, be competitive, or even outperform the  $BC_a$  interval in the logistic and Poisson GEE situations of the previous chapter.

#### 4.2.2 Bootstrap Calibrated Generalized Wald Confidence Intervals

Bootstrap calibration is a general method that applies the bootstrap to a confidence interval procedure, and directly attempts to improve its coverage (Loh 1987). Interval calibration can, in principle, be applied to nearly any interval procedure, though most of the literature on calibration appears to be focused on the calibration of bootstrap confidence intervals, such as the the  $BC_a$ , via iterated or double bootstrapping. The idea of applying interval calibration to robust non-bootstrap-based intervals, such as the generalized Wald interval, appears to have been ignored. Using Edgeworth and Cornish-Fisher expansions, one is able to show that, in some situations, bootstrap calibration is able to increase the coverage accuracy of a (two-sided) interval procedure by two orders of magnitude (Martin 1990). This would imply that under certain regularity conditions, the calibration procedure would improve the coverage rate of the

generalized Wald interval from  $1 - \alpha + O_p(n^{-1/2})$  to  $1 - \alpha + O_p(n^{-3/2})$ . In the examples we consider here, this is certainly overoptimistic, since we apply the procedure in situations where the data are discrete, where the expansions in (Martin 1990) are not valid. Nevertheless, it is interesting to see whether bootstrap calibration is able to significantly improve the coverage of the generalized Wald interval in these situations. We find that it is competitive, as measured by coverage and median length, with the GS2 and  $BC_a$  confidence intervals.

We now describe bootstrap calibration as it is applied to the generalized Wald interval. Because of the need to compute a large number of resampled estimates of  $\hat{\theta}$ , the procedure is very computationally intensive. The idea is simple, though, and easy to implement in practice.

As in the computation of the  $BC_a$  interval, compute  $B$  bootstrap resamples of the data,  $Y_b^*$ . For each of these resampled data sets, compute the bootstrap replicates of  $\hat{\psi}$  and its sandwich matrix-based standard error,  $\hat{\psi}_b^*$  and  $\hat{s}e_b^* = \sqrt{\{\hat{A}_b^{*-1}(\hat{\theta}_b^*)\hat{D}_b^*(\hat{\theta}_b^*)\hat{A}_b^{*-1}(\hat{\theta}_b^*)\}_{11}}$ . The quantities,  $\hat{A}_b^*$  and  $\hat{D}_b^*$ , are the  $\hat{A}$  and  $\hat{D}$  matrices computed for the resampled dataset. Then find the value,  $\alpha^*$ , such that the proportion of the bootstrap replicate generalized Wald confidence intervals that cover  $\hat{\psi}$  is precisely  $\alpha$ :

$$\alpha^* = \arg_{\beta \in (0, .5)} \left\{ \frac{\#(\hat{\psi}_b^* - z_{1-\beta/2}\hat{s}e_b^* \leq \hat{\psi} \leq \hat{\psi}_b^* + z_{1-\beta/2}\hat{s}e_b^*)}{B} = \alpha \right\}. \quad (4.1)$$

This is most easily accomplished by first accumulating the  $B$  values of  $\hat{\psi}_b^*$  and  $\hat{s}e_b^*$ , and then using bisection or Brent's root finding method (Brent 1973) to find  $\alpha^*$ . We then declare the calibrated generalized Wald interval to be

$$\hat{\psi} \pm z_{1-\alpha^*/2} \sqrt{\{\hat{A}^{-1}(\hat{\theta})\hat{D}(\hat{\theta})\{\hat{A}^{-1}(\hat{\theta})\}_{11}}}. \quad (4.2)$$

## 4.3 Simulations

We repeated the non-identity working correlation matrix GEE simulations of the previous chapter to investigate the small sample properties of the  $BC_a$  and calibrated generalized Wald confidence intervals, and compare these intervals to the generalized Wald and GS2 confidence intervals described in the previous chapter. Because of the computational demands of bootstrap, the simulations were implemented in C++ using the Template Numerical Toolkit (TNT) numerical linear algebra libraries, and C++ ported versions some of John Monahan's FORTRAN statistical subroutines (Monahan 2001). In each of the simulations, there were a total of 2000 simulation replicates. The bootstrap intervals for each simulation replicate were computed using  $B = 2000$  bootstrap samples.

### 4.3.1 Logistic GEE Simulations

In Table 4.1 we compare the generalized Wald- $t$  and GS2 confidence intervals to the calibrated generalized Wald and  $BC_a$  intervals in the case of a simple linear logistic GEE using the exchangeable working correlation matrix. We see that the  $BC_a$  interval's performance is often poorer than that of the generalized Wald- $t$ . The calibrated generalized Wald interval, however, has coverages that are almost uniformly superior to that of the GS2 interval, and has a competitive median length.

The GUIDE data bootstrap simulation results are given in Table 4.2 below. Like the previous simulations, the  $BC_a$  confidence interval performs rather poorly. Surprisingly, in this situation it appears to be quite conservative, while in the simple linear logistic GEE simulations it often suffered from being too liberal. The coverage rate of the calibrated generalized Wald

Table 4.1: Coverage and length for calibrated generalized Wald and  $BC_a$  intervals for Bernoulli data under a simple linear logistic regression model.

| $(\psi, \eta)$ | # of Clusters $K =$ |                  | $\rho = 0$ |     |     | $\rho = 0.2$ |     |     | $\rho = 0.5$ |     |     |
|----------------|---------------------|------------------|------------|-----|-----|--------------|-----|-----|--------------|-----|-----|
|                |                     |                  | 5          | 10  | 20  | 5            | 10  | 20  | 5            | 10  | 20  |
| (0,0)          | Coverage            | Gen. Wald- $t$   | .95        | .93 | .94 | .94          | .93 | .94 | .93          | .94 | .95 |
|                |                     | GS2              | .97        | .94 | .95 | .97          | .94 | .95 | .96          | .96 | .95 |
|                |                     | Calib. Gen. Wald | .95        | .96 | .96 | .97          | .96 | .96 | .98          | .97 | .97 |
|                |                     | $BC_a$           | .79        | .91 | .95 | .85          | .91 | .95 | .88          | .88 | .90 |
|                | Median              | Gen. Wald- $t$   | 4.1        | 2.4 | 1.6 | 3.7          | 2.3 | 1.5 | 3.0          | 2.0 | 1.3 |
|                | Length              | GS2              | 15.6       | 3.0 | 1.8 | 13.0         | 2.9 | 1.7 | 6.2          | 2.4 | 1.4 |
|                |                     | Calib. Gen. Wald | 5.3        | 2.6 | 1.7 | 6.2          | 2.5 | 1.7 | 6.8          | 2.4 | 1.4 |
|                |                     | $BC_a$           | 4.6        | 2.5 | 1.7 | 4.8          | 2.4 | 1.7 | 4.4          | 2.1 | 1.3 |
|                |                     |                  |            |     |     |              |     |     |              |     |     |
|                | (1,0)               | Gen. Wald- $t$   | .92        | .94 | .94 | .90          | .94 | .94 | .79          | .90 | .93 |
|                |                     | GS2              | .95        | .95 | .94 | .93          | .95 | .94 | .83          | .91 | .94 |
|                |                     | Calib. Gen. Wald | .95        | .96 | .96 | .96          | .96 | .96 | .96          | .96 | .96 |
|                |                     | $BC_a$           | .80        | .92 | .95 | .84          | .93 | .95 | .89          | .97 | .95 |
|                | Median              | Gen. Wald- $t$   | 4.5        | 2.6 | 1.7 | 3.8          | 2.3 | 1.6 | 4.0          | 2.4 | 1.6 |
|                | Length              | GS2              | 16.5       | 3.3 | 1.9 | 13.6         | 3.0 | 1.7 | 12.7         | 3.1 | 1.8 |
|                |                     | Calib. Gen. Wald | 5.9        | 3.8 | 1.8 | 6.5          | 2.8 | 1.7 | 6.6          | 3.2 | 1.8 |
|                |                     | $BC_a$           | 4.9        | 2.7 | 1.8 | 5.2          | 2.6 | 1.7 | 4.5          | 2.8 | 1.7 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ . Cluster size=4.

interval is uniformly acceptable, and appears to also be more efficient than the GS2 interval, when the GS2 interval's coverage is at least .94.

These observations lead us to conclude that the calibrated generalized Wald confidence interval should be the method of choice for logistic GEE models, provided that sufficient computing resources are available.

Table 4.2: Coverage and length for calibrated generalized Wald and  $BC_a$  intervals for clustered Bernoulli data simulated from the GUIDE dataset.

| Parameter |          | # of Clusters $K =$ | $\rho = 0$ |     | $\rho = 0.2$ |     |
|-----------|----------|---------------------|------------|-----|--------------|-----|
|           |          |                     | 20         | 37  | 20           | 37  |
| SEX       | Coverage | Gen. Wald- $t$      | .89        | .91 | .85          | .91 |
|           |          | GS2                 | .94        | .94 | .92          | .95 |
|           |          | Calib. Gen. Wald    | .96        | .96 | .96          | .96 |
|           |          | $BC_a$              | .99        | .97 | .99          | .97 |
|           | Median   | Gen. Wald- $t$      | 3.4        | 2.3 | 2.9          | 2.3 |
|           |          | GS2                 | 5.2        | 2.8 | 4.2          | 2.8 |
|           |          | Calib. Gen. Wald    | 4.8        | 2.6 | 4.9          | 2.7 |
|           |          | $BC_a$              | 18.6       | 3.2 | 16.6         | 3.1 |
|           | Length   | Gen. Wald- $t$      | 3.4        | 2.3 | 2.9          | 2.3 |
|           |          | GS2                 | 5.2        | 2.8 | 4.2          | 2.8 |
|           |          | Calib. Gen. Wald    | 4.8        | 2.6 | 4.9          | 2.7 |
|           |          | $BC_a$              | 18.6       | 3.2 | 16.6         | 3.1 |
| AGE       | Coverage | Gen. Wald- $t$      | .92        | .93 | .90          | .93 |
|           |          | GS2                 | .94        | .94 | .94          | .95 |
|           |          | Calib. Gen. Wald    | .95        | .95 | .96          | .96 |
|           |          | $BC_a$              | .98        | .96 | .98          | .97 |
|           | Median   | Gen. Wald- $t$      | 3.4        | 2.3 | 3.0          | 2.2 |
|           |          | GS2                 | 5.2        | 2.7 | 5.0          | 2.6 |
|           |          | Calib. Gen. Wald    | 3.9        | 2.5 | 3.8          | 2.5 |
|           |          | $BC_a$              | 5.7        | 2.8 | 5.6          | 2.8 |
|           | Length   | Gen. Wald- $t$      | 3.4        | 2.3 | 3.0          | 2.2 |
|           |          | GS2                 | 5.2        | 2.7 | 5.0          | 2.6 |
|           |          | Calib. Gen. Wald    | 3.9        | 2.5 | 3.8          | 2.5 |
|           |          | $BC_a$              | 5.7        | 2.8 | 5.6          | 2.8 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .

### 4.3.2 Poisson GEE Simulations

Table 4.3 shows simulation results comparing the bootstrap confidence intervals to the generalized Wald- $t$  and GS2 confidence intervals of the previous chapter. We see that the two best performing confidence intervals are the GS2 and the calibrated generalized Wald confidence intervals, but neither one consistently outperforms the other.

Table 4.3: Coverage and length for calibrated generalized Wald and  $BC_a$  intervals for serially correlated Poisson data under a log linear regression model.

| $(\psi, \eta)$ |          | # of Clusters $K =$ | $\rho = 0$ |     |     | $\rho = .2$ |     |     | $\rho = .5$ |     |     |
|----------------|----------|---------------------|------------|-----|-----|-------------|-----|-----|-------------|-----|-----|
|                |          |                     | 10         | 20  | 50  | 10          | 20  | 50  | 10          | 20  | 50  |
| (0,0)          | Coverage | Gen. Wald- $t$      | .92        | .94 | .94 | .92         | .94 | .94 | .91         | .94 | .94 |
|                |          | GS2                 | .97        | .96 | .94 | .96         | .96 | .95 | .95         | .96 | .95 |
|                |          | Calib. Gen. Wald    | .95        | .95 | .94 | .95         | .95 | .95 | .94         | .95 | .95 |
|                |          | $BC_a$              | .90        | .94 | .93 | .91         | .93 | .95 | .90         | .94 | .95 |
|                | Median   | Gen. Wald- $t$      | 1.1        | .77 | .48 | 1.2         | .80 | .51 | 1.2         | .86 | .54 |
|                |          | GS2                 | 1.2        | .80 | .49 | 1.3         | .84 | .52 | 1.4         | .89 | .54 |
|                |          | Calib. Gen. Wald    | 1.3        | .82 | .49 | 1.3         | .83 | .50 | 1.3         | .85 | .54 |
|                |          | $BC_a$              | 1.1        | .76 | .48 | 1.1         | .77 | .50 | 1.1         | .84 | .56 |
|                | Length   | Gen. Wald- $t$      | 1.1        | .77 | .48 | 1.2         | .80 | .51 | 1.2         | .86 | .54 |
|                |          | GS2                 | 1.2        | .80 | .49 | 1.3         | .84 | .52 | 1.4         | .89 | .54 |
|                |          | Calib. Gen. Wald    | 1.3        | .82 | .49 | 1.3         | .83 | .50 | 1.3         | .85 | .54 |
|                |          | $BC_a$              | 1.1        | .76 | .48 | 1.1         | .77 | .50 | 1.1         | .84 | .56 |
| (1,0)          | Coverage | Gen. Wald- $t$      | .93        | .93 | .95 | .93         | .93 | .94 | .92         | .93 | .94 |
|                |          | GS2                 | .96        | .94 | .96 | .96         | .95 | .95 | .95         | .95 | .95 |
|                |          | Calib. Gen. Wald    | .96        | .95 | .94 | .95         | .94 | .95 | .95         | .94 | .95 |
|                |          | $BC_a$              | .90        | .92 | .94 | .89         | .92 | .94 | .89         | .91 | .93 |
|                | Median   | Gen. Wald- $t$      | 1.1        | .78 | .48 | 1.2         | .83 | .52 | 1.2         | .88 | .55 |
|                |          | GS2                 | 1.3        | .83 | .49 | 1.4         | .88 | .53 | 1.4         | .93 | .56 |
|                |          | Calib. Gen. Wald    | 1.3        | .84 | .50 | 1.4         | .84 | .51 | 1.5         | .90 | .54 |
|                |          | $BC_a$              | 1.1        | .77 | .49 | 1.1         | .79 | .49 | 1.3         | .86 | .53 |
|                | Length   | Gen. Wald- $t$      | 1.1        | .78 | .48 | 1.2         | .83 | .52 | 1.2         | .88 | .55 |
|                |          | GS2                 | 1.3        | .83 | .49 | 1.4         | .88 | .53 | 1.4         | .93 | .56 |
|                |          | Calib. Gen. Wald    | 1.3        | .84 | .50 | 1.4         | .84 | .51 | 1.5         | .90 | .54 |
|                |          | $BC_a$              | 1.1        | .77 | .49 | 1.1         | .79 | .49 | 1.3         | .86 | .53 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ . Cluster size=4.

In Table 4.4 we see the bootstrap simulation results for the Epilepsy data. The calibrated generalized Wald interval has superior coverage compared to the GS2 interval. The  $BC_a$  interval's coverage was never higher than .94 in any of the simulations, and, therefore, is certainly not preferable to either the GS2 or calibrated generalized Wald intervals. Overall,

we conclude that the calibrated generalized Wald interval is slightly preferable to the GS2 confidence interval, because its coverage is generally in acceptable range between .94 and .96.

Table 4.4: Coverage and length for calibrated generalized Wald and  $BC_a$  intervals for serially correlated Poisson data using the Epilepsy data covariates and MLE.

|           |          |                     | $\rho = 0$ |     | $\rho = .2$ |     | $\rho = .5$ |     |
|-----------|----------|---------------------|------------|-----|-------------|-----|-------------|-----|
| Parameter |          | # of Clusters $K =$ | 30         | 59  | 30          | 59  | 30          | 59  |
| TREATMENT | Coverage | Gen. Wald- $t$      | .94        | .94 | .93         | .94 | .92         | .93 |
|           |          | GS2                 | .95        | .95 | .95         | .95 | .94         | .95 |
|           |          | Calib. Gen. Wald    | .95        | .95 | .96         | .95 | .96         | .96 |
|           |          | $BC_a$              | .94        | .94 | .94         | .94 | .93         | .94 |
|           | Median   | Gen. Wald- $t$      | .25        | .18 | .29         | .21 | .38         | .27 |
|           |          | GS2                 | .27        | .18 | .31         | .21 | .41         | .28 |
|           |          | Calib. Gen. Wald    | .26        | .18 | .31         | .21 | .47         | .30 |
|           |          | $BC_a$              | .25        | .18 | .30         | .21 | .45         | .29 |
|           | Length   | Gen. Wald- $t$      | .25        | .18 | .29         | .21 | .38         | .27 |
|           |          | GS2                 | .27        | .18 | .31         | .21 | .41         | .28 |
|           |          | Calib. Gen. Wald    | .26        | .18 | .31         | .21 | .47         | .30 |
|           |          | $BC_a$              | .25        | .18 | .30         | .21 | .45         | .29 |
| AGE       | Coverage | Gen. Wald- $t$      | .92        | .93 | .90         | .93 | .89         | .92 |
|           |          | GS2                 | .95        | .95 | .93         | .94 | .93         | .94 |
|           |          | Calib. Gen. Wald    | .94        | .95 | .95         | .95 | .96         | .96 |
|           |          | $BC_a$              | .93        | .94 | .94         | .94 | .93         | .94 |
|           | Median   | Gen. Wald- $t$      | .29        | .24 | .33         | .27 | .39         | .32 |
|           |          | GS2                 | .31        | .25 | .35         | .28 | .42         | .33 |
|           |          | Calib. Gen. Wald    | .33        | .26 | .38         | .29 | .59         | .39 |
|           |          | $BC_a$              | .32        | .25 | .37         | .28 | .67         | .43 |
|           | Length   | Gen. Wald- $t$      | .29        | .24 | .33         | .27 | .39         | .32 |
|           |          | GS2                 | .31        | .25 | .35         | .28 | .42         | .33 |
|           |          | Calib. Gen. Wald    | .33        | .26 | .38         | .29 | .59         | .39 |
|           |          | $BC_a$              | .32        | .25 | .37         | .28 | .67         | .43 |

Note: Estimated coverages have approximate standard error  $\sqrt{(.95)(.05)/2000} = .005$ .  
Cluster size=5.

## Chapter 5

# Conclusions and Future Work

In this dissertation, we have investigated several methods of interval estimation that are insensitive to model assumptions. That is, these confidence intervals maintain their asymptotic nominal coverage rate, even when the model is misspecified.

In the first chapter, we investigated three likelihood-based robust confidence intervals, the generalized Wald, generalized score, and robust profile likelihood confidence intervals. We found that for logistic GEE models with the independence working correlation matrix, Huber robust regression models, and measurement error models, the robust profile likelihood interval is superior to the generalized Wald and generalized score confidence intervals, as measured by coverage probability and median length. Surprisingly, the generalized score interval was found to be unacceptably conservative and inefficient in small samples. The most striking manifestation of this conservatism was the frequency with which the generalized score interval was infinite in length, which was sometimes as high as 90%. In Poisson GEE models with the independence working correlation matrix, the conservatism of the generalized score interval was less pronounced, and its nearly nominal coverage rate leads us to recommend its use in the

analysis of Poisson-type data.

Next, we introduced a new form of interval estimation, length-optimal confidence intervals, which is a general methodology for modifying score and likelihood-type confidence intervals that yields intervals with shorter length. Surprisingly, the confidence intervals formed using this method were shown in simulations to enjoy higher coverage rates in the context of binary data and linear measurement error models. However, these confidence intervals were only slightly different from their standard, equal-tail, counterparts in Huber regression and Poisson regression models. We developed theoretical results that indicate that the length-optimal version of the non-robust likelihood interval, in the absence of nuisance parameters, is asymptotically equivalent to the standard likelihood interval. This is shown to imply that the length-optimal likelihood interval has asymptotically correct coverage. One avenue of future work would be to establish a more general result for the robust profile interval when nuisance parameters are present.

In Chapter 3, we developed a new type of generalized score interval, the GS2 confidence interval. Because the robust profile likelihood interval is unavailable in pure estimating equations situations, an alternative to the generalized score and generalized Wald intervals is desirable. Simulations indicated that the GS2 interval is superior to these other intervals, in that its coverage rate was near nominal, and did not suffer as severely from the defects of the generalized score interval. This indicates that the GS2 interval is preferable to the generalized score and Wald intervals in the context of estimating equations.

In the final chapter, we did a small simulation study of a bootstrap calibrated version of the generalized Wald interval. This bootstrap calibration was shown to significantly improve the generalized Wald interval. The improvement is such that the calibrated generalized Wald

interval is at least competitive, if not superior to, the GS2 interval. The performance of the standard bootstrap confidence interval, the  $BC_a$  interval, was inconsistent in the situations investigated here. In future work, it would be interesting to investigate a double-bootstrap calibration of the generalized Wald interval. This additional layer of bootstrapping may yield even more accurate confidence intervals with competitive median length. It would also be interesting to apply the calibrated generalized Wald interval in situations where other bootstrap procedures have been shown to fail, such as in the interval estimation of variance parameters.

As a matter of expediency, we developed a new method of generating autocorrelated Poisson-type random variables with a specific mean structure and serial correlation for use in generating data for some of our simulations. In the concluding phase of this research, we realized that this scheme suggests a likelihood for analyzing autocorrelated Poisson-type data. It be interesting to compare the performance of the maximum likelihood estimator for this likelihood to that of the estimator arising from the Poisson GEE model using the AR(1) working correlation matrix, to see if the maximum likelihood estimator is more efficient.

## Appendix A

# A Serially Correlated Poisson-Type Random Number Generation Scheme

In this appendix we describe a random number generating scheme for generating serially correlated Poisson-type data.

Suppose that one wishes to generate a sequence of Poisson-type random variables,  $\{y_i\}_{i=0}^{\infty}$  with autocorrelation  $0 \leq \rho < 1$ , such that  $E(y_i) = \lambda_i > 0$ . This is possible, given certain constraints on the  $\lambda_i$ , by the following algorithm.

1. Generate  $y_0 \sim \text{Poisson}(\lambda_0)$ .
2. For  $i = 0, 1, \dots, \infty$ , generate  $y_{i+1}$  from  $y_{i+1}|y_i \sim \text{Poisson}(\rho\sqrt{\frac{\lambda_{i+1}}{\lambda_i}}y_i + \lambda_{i+1} - \rho\sqrt{\lambda_i\lambda_{i+1}})$ .

The theorem below describes describes the means, variances, and lagged autocorrelations for the sequence,  $\{y_i\}$ .

**Theorem A.1** Let  $0 \leq \rho < 1$  be given. Also, let  $\{\lambda_i\}_{i=0}^{\infty}$  be a sequence of non-negative real numbers such that  $\rho\sqrt{\lambda_i\lambda_{i+1}} < \lambda_i$ , for all  $i = 0, 1, \dots, \infty$ . Let  $\{y_i\}_{i=0}^{\infty}$  be a sequence of random variables generated using the algorithm described above. Then,

$$E(y_i) = \lambda_i,$$

$$Var(y_i) = \lambda_i \sum_{j=0}^i \rho^{2j} = \lambda_i \frac{1 - \rho^{2(i+1)}}{1 - \rho^2},$$

and

$$Corr(y_i, y_{i+k}) = \rho^k \sqrt{\frac{\sum_{j=0}^i \rho^{2j}}{\sum_{j=0}^{i+k} \rho^{2j}}} = \rho^k \sqrt{\frac{1 - \rho^{2(i+1)}}{1 - \rho^{2(i+k+1)}}}.$$

**Proof** We use induction to prove the statements about the means and variances of the  $y_i$  are correct. Since  $y_0$  is  $Poisson(\lambda_0)$  distributed,  $E(y_0) = Var(y_0) = \lambda_0$ , so the expressions for the mean and variance are correct for  $i = 0$ . Now, assume, by way of induction, that the theorem is true for some  $i \geq 0$ . Then, we see that the expression for the  $E(y_i)$  is correct, as,

$$\begin{aligned} E(y_{i+1}) &= E(E(y_{i+1}|y_i)) \\ &= E\left(\rho\sqrt{\frac{\lambda_{i+1}}{\lambda_i}}y_i + \lambda_{i+1} - \rho\sqrt{\lambda_i\lambda_{i+1}}\right) \\ &= \rho\sqrt{\frac{\lambda_{i+1}}{\lambda_i}}\lambda_i + \lambda_{i+1} - \rho\sqrt{\lambda_i\lambda_{i+1}} \\ &= \lambda_i. \end{aligned}$$

The expression for  $Var(y_i)$  is proven to be correct by seeing that

$$\begin{aligned}
Var(y_{i+1}) &= Var(E(y_{i+1}|y_i)) + E(Var(y_{i+1}|y_i)) \\
&= Var(\rho\sqrt{\frac{\lambda_{i+1}}{\lambda_i}}y_i) + E(\rho\sqrt{\frac{\lambda_{i+1}}{\lambda_i}}y_i + \lambda_{i+1} - \rho\sqrt{\lambda_{i+1}\lambda_i}) \\
&= \rho^2\frac{\lambda_{i+1}}{\lambda_i}\lambda_i\sum_{j=0}^i\rho^{2j} + \lambda_i \\
&= \lambda_i(1 + \rho^2\sum_{j=0}^i\rho^{2j}) \\
&= \lambda_i\sum_{j=0}^{i+1}\rho^{2j}.
\end{aligned}$$

Before proving the result concerning the autocorrelations, we first prove that

$$E(y_{i+k}|y_i) = \rho^k\sqrt{\frac{\lambda_{i+k}}{\lambda_i}}y_i + \lambda_{i+k} - \rho^k\sqrt{\lambda_i\lambda_{i+k}}, \quad (\text{A.1})$$

by inducting on  $k$ . This result is automatically true for all  $k = 1$ , because of the nature of the way the  $y_i$  are generated. Then, assuming that the result is true for  $k - 1$ , we have

$$\begin{aligned}
E(y_{i+k}|y_i) &= E(E(y_{i+k}|y_{i+1})|y_i) \\
&= E(\rho^{k-1}\sqrt{\frac{\lambda_{i+k}}{\lambda_{i+1}}}y_{i+1} + \lambda_{i+k} - \rho^{k-1}\sqrt{\lambda_{i+1}\lambda_{i+k}}|y_i) \\
&= \rho^{k-1}\sqrt{\frac{\lambda_{i+k}}{\lambda_{i+1}}}(\rho\sqrt{\frac{\lambda_{i+1}}{\lambda_i}}y_i + \lambda_{i+1} - \rho\sqrt{\lambda_i\lambda_{i+1}}) + \lambda_{i+k} - \rho^{k-1}\sqrt{\lambda_{i+1}\lambda_{i+k}} \\
&= \rho^k\sqrt{\frac{\lambda_{i+k}}{\lambda_i}}y_i + \lambda_{i+k} - \rho^k\sqrt{\lambda_i\lambda_{i+k}},
\end{aligned}$$

thus proving the result. Now,

$$\begin{aligned}
E(y_i y_{i+k}) &= E(y_i E(y_{i+k} | y_i)) \\
&= E(y_i (\rho^k \sqrt{\frac{\lambda_{i+k}}{\lambda_i}} y_i + \lambda_{i+k} - \rho^k \sqrt{\lambda_i \lambda_{i+k}})) \\
&= \rho^k \sqrt{\frac{\lambda_{i+k}}{\lambda_i}} (\lambda_i^2 + \lambda_i \sum_{j=0}^i \rho^{2j} + \lambda_i \lambda_{i+k} - \rho^k \lambda_i \sqrt{\lambda_i \lambda_{i+k}}) \\
&= \rho^k \sqrt{\lambda_i \lambda_{i+k}} \sum_{j=0}^i \rho^{2j} + \lambda_i \lambda_{i+k}.
\end{aligned}$$

Thus,

$$\begin{aligned}
Corr(y_i, y_{i+k}) &= \frac{E(y_i y_{i+k}) - E(y_i)E(y_{i+k})}{\sqrt{Var(y_i)Var(y_{i+k})}} \\
&= \frac{\rho^k \sqrt{\lambda_i \lambda_{i+k}} \sum_{j=0}^i \rho^{2j} + \lambda_i \lambda_{i+k} - \lambda_i \lambda_{i+k}}{\sqrt{\lambda_i \sum_{j=0}^i \rho^{2j} \lambda_{i+k} \sum_{j=0}^{i+k} \rho^{2j}}} \\
&= \rho^k \sqrt{\frac{\sum_{j=0}^i \rho^{2j}}{\sum_{j=0}^{i+k} \rho^{2j}}} \\
&= \rho^k \sqrt{\frac{1 - \rho^{2(i+1)}}{1 - \rho^{2(i+k+1)}}},
\end{aligned}$$

which completes the proof.  $\square$

The theorem shows that the individual  $y_i$  have an overdispersed poisson-type distribution with expectation  $\lambda_i$ , and variance inflation factor  $(1 - \rho^{2(i+1)})/(1 - \rho^2)$ . Also, it shows that  $Corr(y_i, y_{i+1}) = \rho \sqrt{(1 - \rho^{2(i+1)})/(1 - \rho^{2(i+2)})}$ , which converges rapidly to  $\rho$  as  $i$  increases.

# Bibliography

- Agresti, A. and B. Caffo (2000). Simple and effective confidence intervals and differences of proportions result from adding two successes and failures. *American Statistician* 54(4), 280–288.
- Agresti, A. and B. Coull (1998). Approximate is better than exact for interval estimation of binomial proportions. *American Statistician* 52(4), 119–126.
- Al-Osh, M. and S. Lee (2001). A simple approach for generating correlated binary covariates. *Journal of Statistical Computation and Simulation* 70(2), 231–255.
- Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* 74(3), 457–468.
- Birnbaum, A. (1961). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of the American Statistical Association* 56, 246–249.
- Boos, D. (1992). On generalized score tests. *American Statistician* 46(4), 327–333.
- Brent, R. (1973). *Algorithms For Minimization Without Derivatives*. Upper Saddle River, New Jersey: Prentice-Hall.
- Brown, L., T. Cai, A. DasGupta, and A. Agresti (2002). Interval estimation for a binomial

- proportion. *Statistical Science* 16(2), 101–133.
- Diciccio, T. and B. Efron (1996). Bootstrap confidence intervals. *Statistical Science* 11(3), 189–212.
- Diciccio, T. and J. Romano (1988). A review of confidence intervals. *Journal of the Royal Statistical Society, Series B* 50(3), 338–354.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* 7(1), 1–26.
- Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics* 9, 139–172.
- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* 82(397), 171–185.
- Efron, B. and R. Tibshirani (1981). *An Introduction To the Bootstrap*. New York: Chapman and Hall.
- Foutz, R. V. and R. C. Srivastava (1977). The performance of the likelihood ratio test when the model is incorrect. *The Annals of Statistics* 5(6), 1183–1194.
- Gunsolley, J., C. Getchell, and V. Chinchilli (1995). Small sample characteristics of generalized estimating equations. *Communications in Statistics - Simulation and Computation* 24(4), 869–878.
- Hall, P. (1986). On the bootstrap and confidence intervals. *The Annals of Statistics* 14(4), 1431–1452.
- Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of*

*Statistics* 16(4), 927–953.

Hall, P. and M. Martin (1988). On bootstrap resampling and iteration. *Biometrika* 75(4), 661–671.

Huber, P. (1967). The behavior of the maximum likelihood estimator under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (1 ed.), pp. 221–223. University of California Press.

Huber, P. (1981). *Robust Statistics*. New York: Wiley Series in Probability And Mathematical Statistics.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* 48(3), 419–426.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika* 69(1), 19–27.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73(1), 13–22.

Loh, W. (1987). Calibrating confidence coefficients. *Journal of the American Statistical Association* 82(397), 155–162.

Martin, M. (1990). On bootstrap iteration for coverage correction. *Journal of the American Statistical Association* 85(412), 1105–1118.

Monahan, J. (2001). *Numerical Methods of Statistics*. New York: Cambridge University Press.

Preisser, J. S. and B. F. Qaqish (1999). Robust regression for clustered data with application to binary responses. *Biometrics* 55(2), 574–579.

Presnell, B. and D. Boos (2002). The In-and-Out-of-Sample (IOS) likelihood ratio test for

model misspecification. *Institute of Statistics Mimeograph Series 2536*.

Press, W., S. Teukolsky, W. Vetterling, and B. Flannery (1988). *Numerical Recipes In C*. Cambridge University Press.

Schenker, N. (1985). Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association* 80, 360–361.

Schrader, R. M. and T. Hettmansperger (1980). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika* 67(1), 93–101.

Serfling, R. (1980). *Approximation Theorems Of Mathematical Statistics*. New York: John Wiley and Sons, Inc.

Stafford, J. (1996). A robust adjustment of the profile likelihood. *The Annals of Statistics* 24(1), 336–352.

Stefanski, L. and D. Boos (2002). The calculus of M-estimation. *American Statistician* 56(1), 29–38.

Thall, P. and S. Vail (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* 46(3), 657–651.

Van Der Vaart, A. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50(1), 1–25.