Fall 2015

Problems in generalized linear model selection and predictive evaluation for binary outcomes

Patrick Ten Eyck
University of Iowa

PROBLEMS IN GENERALIZED LINEAR MODEL SELECTION AND PREDICTIVE EVALUATION FOR BINARY OUTCOMES

by

Patrick Ten Eyck

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics in the
Graduate College of The
University of Iowa

December 2015

Thesis Supervisor: Professor Joseph Cavanaugh

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Patrick Ten Eyck

has been approved by the Examining Committee
for the thesis requirement for the Doctor of
Philosophy degree in Biostatistics at the December 2015
graduation.

Thesis Committee: _____
Joseph Cavanaugh, Thesis Supervisor

_____
William Clarke

_____
Jacob Oleson

_____
Eric Foster

_____
Marizen Ramirez

To the man who has no limits,
who is whatever Gotham needs him to be

ACKNOWLEDGMENTS

First and foremost, I would like to thank my thesis advisor and close friend, Professor Joseph Cavanaugh, who has always had my best interests at heart. Even before I officially accepted my offer to attend the University of Iowa as a graduate student in Biostatistics, Joe went above and beyond to ensure that I was admitted under the most ideal of circumstances. As a student in several of his classes, I developed a level of appreciation for our field that I did not know was possible. The lectures that Joe presents are detailed, yet accessible, teaching students to ask questions if for nothing more than curiosity. As a thesis advisor, Joe has provided constant support and patience, encouraging me as our ideas found the bigger picture and reassuring me after we hit a dead end. I will never be able to fully express the gratitude I have towards Joe and his impact on my life. He has without a doubt played the most significant role in my entire academic career. Thank you, Joe!

I would also like to thank my thesis committee for their role in my time as a student here at Iowa. In my first year, I took a course taught by Professor Jacob Oleson. Like Joe, Jake has the invaluable talent of dissecting intricate concepts into their basic components, aiding his students in their learning and retention of the material. I was fortunate enough to spend three years working closely with Professor Marizen Ramirez on a study involving bullying in Iowa public schools. This collaborative effort provided real-world experience and prepared me for my next step in ways the classroom could not. As a first-time teaching assistant for Professor William Clarke, I got my initial glimpse of being in front of a class. Dr. Clarke showed me how to take mastery of material and share it with others who do not come from the same background. I was also assigned as a teaching assistant for Professor Eric Foster. Eric has inspired me to be my best as a fellow student, teaching mentor, and lifelong friend. Thank you, once again, to my outstanding

committee!

Furthermore, I would like to thank all of the Department of Biostatistics students, faculty, and staff for their support and help along the way. Every person has had a hand in making my experience at Iowa the best it could possibly be. I would like to single out Terry Kirk for her tireless efforts. Terry is eight steps ahead of the game and keeps everyone on track. During my recruitment weekend, everyone in the department told me an undeniable truth: Terry is the best!

I also want to extend my gratefulness to Professor James Deddens, who taught my first statistics course at the University of Cincinnati. He saw potential in me of which I was oblivious. He quickly took me under his wing as an advisee and pointed me in the direction that I am today. Without Dr. Deddens, my time here at Iowa may not have been. Thank you, Dr. Deddens!

Finally, thank you to the most important people in my life. My family. Mom and dad, you have been the foremost source of sound advice and unconditional love. You have shared in every high and low in my life and never wavered in support. A son could not have asked for better parents. Ashley, I am so happy to have seen you blossom into a wonderful woman. I truly believe that I have become a better man having had you as a sister. Tom, you are my big brother and my first role model. Your attitude and work ethic have taught me to weather the storm and come out stronger than before. Alex, you are only my younger brother in age. Your fearless approach to life has shown me how to not waste an opportunity when it is standing right in front of you. I love you guys!

ABSTRACT

This manuscript consists of three papers that formulate novel generalized linear modeling methodologies.

In Chapter 1, we introduce a variant of the traditional concordance statistic that is used to evaluate predictive discrimination for fitted logistic regression models. In computing the measures of concordance and discordance, this *adjusted c-statistic* utilizes the differences in predicted probabilities as weights for each event/non-event observation pair. Using simulations, we present an extensive comparison of the traditional and the adjusted c-statistic, which highlights the properties of the latter. We then illustrate the use of these measures in a modeling application.

In Chapter 2, we present the development and investigation of three model selection criteria based on cross-validatory analogues of the traditional and adjusted c-statistics. These criteria are designed to estimate three corresponding measures of predictive error: the *model misspecification prediction error*, the *fitting sample prediction error*, and the *sum of prediction errors*. We examine the properties of the selection criteria via an extensive simulation study, and illustrate their utility in a modeling application.

In Chapter 3, we propose and investigate an alternate approach to pseudo-likelihood model selection in the generalized linear mixed modeling framework. After outlining the problem with the natural approach to the computation of pseudo-likelihood model selection criteria, we propose a technique that circumvents this problem. The new approach can be implemented using a SAS macro that obtains and applies the pseudo-data from the full model to fitting candidate models based on all possible subsets of predictor variables. We justify the propriety of the resulting pseudo-likelihood selection criteria through an extensive simulation study, and highlight the use of the proposed method in a modeling application.

PUBLIC ABSTRACT

This thesis is comprised of three papers that formulate novel methodologies pertaining to model selection and the assessment of predictive effectiveness. In the first two papers, the methodologies are developed in the framework of the logistic regression model, and in the third, in the framework of the generalized linear mixed model (GLMM). The first two papers present contributions pertaining to the assessment of predictive effectiveness, and model evaluation and selection based on predictive efficacy. The second and third papers present novel model selection methodologies, the latter to accommodate GLMMs fitted using the pseudo-likelihood approach.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

PREFACE

This manuscript is comprised of three papers that formulate novel methodologies pertaining to model selection and the assessment of predictive effectiveness. In the first two papers, the methodologies are developed in the framework of the logistic regression model, and in the third, in the framework of the generalized linear mixed model (GLMM). The first two papers present contributions pertaining to the assessment of predictive effectiveness, and model evaluation and selection based on predictive efficacy. The second and third papers present novel model selection methodologies, the latter to accommodate GLMMs fitted using the pseudo-likelihood approach.

In Chapter 1, we introduce a variant of the traditional concordance statistic that is associated with logistic regression. In computing the measures of concordance and discordance, this *adjusted c-statistic* utilizes the differences in predicted probabilities as weights for each event/non-event observation pair. We argue that the inclusion of this additional information yields a more informative measure of predictive discrimination. Using simulations, we highlight the properties of the traditional and adjusted c-statistics, and present an extensive comparison of the two measures. We then employ these measures in a practical application based on modeling the occurrences of heart disease. We compare and contrast the traditional and adjusted c-statistics for each fitted model.

In Chapter 2, we present the development and investigation of three model selection criteria based on cross-validatory analogues of the traditional and adjusted c-statistics. These criteria are designed to estimate three corresponding measures of predictive error: the *model misspecification prediction error*, the *fitting sample prediction error*, and the *sum of prediction errors*. We aim to show that these

estimates make suitable model selection criteria under the logistic regression framework, balancing goodness-of-fit and parsimony, while achieving generalizability. We examine the properties of the selection criteria via an extensive simulation study designed as a factorial experiment. We then revisit the heart disease data set from Chapter 1 to illustrate and compare these criteria in a modeling application.

In Chapter 3, we propose and investigate an alternate approach to pseudo-likelihood model selection in the generalized linear mixed modeling framework. The problem with the natural approach to the computation of pseudo-likelihood model selection criteria is that the pseudo-data vary for each candidate model, leading to criteria based on fundamentally different goodness-of-fit statistics, rendering them incomparable. We propose a technique that circumvents this problem. This new approach can be implemented using a SAS macro that obtains and applies the pseudo-data from the full model to fitting candidate models based on all possible subsets of predictor variables. We justify the propriety of the resulting pseudo-likelihood selection criteria through an extensive study designed as a factorial experiment. We then illustrate this new method in a modeling application pertaining to bullying in public schools. The data for the application is taken from three waves of the Iowa Youth Survey.

The three chapters in this manuscript collectively address problems in generalized linear model selection and predictive evaluation for binary outcomes. The first two chapters are formulated in the framework of logistic regression and consider the problem of assessing predictive discrimination. The properties of the traditional and adjusted c-statistics lead to the development of criteria for selecting an optimal model. The third chapter proposes an appropriate method for computing and comparing selection criteria for generalized linear mixed models fit using the pseudo-likelihood approach.

CHAPTER 1
THE ADJUSTED CONCORDANCE STATISTIC

1.1 Introduction

Measures of model fit numerically characterize the extent to which inferential objectives of interest, such as prediction, are fulfilled by the fitted model. Predictive capability can be quantified for various fitted models in a candidate collection, allowing for model comparison, selection, and evaluation. In logistic regression, the c-statistic (or concordance statistic) is the most popular measure of predictive efficacy. It is essentially calculated by taking a ratio based on two counts, the number of concordant and discordant event/non-event observation pairs. These counts result from dichotomizing the relationship between the predicted probabilities for the observed event and non-event comprising each pair. However, with this approach, the size of the difference between the predicted probabilities is obscured by the dichotomization. In this chapter, we use this additional information to propose a more informative measure of predictive discrimination.

The purpose of this chapter is twofold. First, we introduce an adjusted variant of the traditional c-statistic that utilizes the differences in predicted probabilities as weights for each event/non-event observation pair. Second, we discuss and illustrate the properties for the adjusted c-statistic, and compare these to analogous properties for the traditional c-statistic.

The structure of this chapter is as follows. In section 2, we provide some background regarding logistic regression and the construction of the traditional c-statistic, along with an overview of its properties. In section 3, we propose an adjusted c-statistic using differences in predicted probabilities as weights, and discuss and investigate its properties. We provide simulated results, including an illustrative example comparing the traditional and adjusted c-statistics. In section 4, we

apply both the traditional and adjusted c-statistics in a modeling application and compare the results. Section 5 concludes.

1.2 Background

This section provides relevant background information on the logistic regression framework and associated measures of model fit, focusing on the c-statistic. We examine the construction of the c-statistic and discuss some general properties that will be highlighted in the comparison to its adjusted variant.

1.2.1 Logistic Regression

Logistic regression is the most popular modeling framework for data containing a Bernoulli (0/1) outcome. In fitting a logistic regression model, an assessment of predictive discrimination is often of interest. A common construct that facilitates this assessment is the receiver operating characteristic (ROC) curve. The area under the ROC curve (AUC) measures the probability that for a random event/non-event observation pair drawn from the population, the modeled probability for the event observation will be greater than that for the non-event observation (Hanley and McNeil, 1982). An AUC close to 1 indicates strong predictive discrimination while a value less than 0.5 indicates that the model is no more suitable a discriminating mechanism than randomly generating predicted outcomes using the event prevalence as the Bernoulli probability. The (0,0)-(1,1) line segment on an ROC curve, which corresponds to an AUC = 0.5, is referred to as the chance diagonal (Zhou, Obuchowski and McClish, 2002). The AUC is estimated by the c-statistic, or the sample proportion of event/non-event pairs that are ordered correctly by the fitted model.

1.2.2 Measures of Model Fit

The quality of the model fit for logistic regression can be assessed using several different measures. These can be calibration measures, such as goodness-of-fit tests based on agreement between observed outcomes and predictions (Hilden, Habbema and Bjerregaard, 1978), or discrimination measures of predictive fit based on concordant and discordant pairs (as formally defined in the next subsection).

The goodness-of-fit tests include deviance, Pearson chi-square, and Hosmer-Lemeshow (1980, 1982). These tests inform investigators if the functional form of the specified model is not appropriate for the data. The null hypothesis states that the functional form is correct; thus, a large test statistic indicates that the form is incongruent with the data.

Measures of predictive fit include sensitivity and specificity, the gamma concordance statistic, and the c-statistic. Classification tables allow investigators to evaluate the predictive discrimination of a fitted model through the construction of a 2×2 table of the observed and dichotomized predicted outcomes, based on a predetermined cutoff value for the estimated probabilities. These tables lead to estimates of the sensitivity (probability of an observed event predicted as an event) and specificity (probability of an observed non-event predicted as a non-event) for the fitted model. The gamma statistic is the ratio of (a) the difference between the number of concordant and discordant event/non-event pairs to (b) the total number of all event/non-event pairs. Its range is $[-1, 1]$, with a value closer to 1 indicating strong predictive discrimination and a value less than 0 indicating no predictive discrimination. The sample ROC curve is a plot of sensitivity vs. 1-specificity for all continuous cutoff probabilities in the range [0,1]. The c-statistic, which represents the area under the sample ROC curve, is the empirical probability that the fitted model correctly orders a randomly selected event/non-event pair of observations from the data. The c-statistic will be introduced later in this section.

1.2.3 Concordance and Discordance

The c-statistic and the gamma statistic are measures that are found using tallies of the concordant and discordant pairs. Consider a data set with n_1 observations where $y = 1$, and n_0 observations where $y = 0$. The total number of event/non-event pairs is given by $n_1 n_0$. A pair is said to be concordant if the predicted probability for the event outcome is greater than the predicted probability for the non-event outcome, $\hat{\pi}_{1i} > \hat{\pi}_{0j}$ for $i = 1, 2, ..., n_1$; $j = 1, 2, ..., n_0$. The total number of concordant pairs can be expressed as

$$C = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} 1_{[\hat{\pi}_{1i} > \hat{\pi}_{0j}]}.$$

A pair is said to be discordant if the predicted probability for the event outcome is less than the predicted probability for the non-event outcome, $\hat{\pi}_{1i} < \hat{\pi}_{0j}$ for $i = 1, 2, ..., n_1$; $j = 1, 2, ..., n_0$. The total number of discordant pairs can be expressed as

$$D = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} 1_{[\hat{\pi}_{1i} < \hat{\pi}_{0j}]}.$$

A pair is said to be tied if the predicted probability for the event outcome is equal to the predicted probability for the non-event outcome, $\hat{\pi}_{1i} = \hat{\pi}_{0j}$ for $i = 1, 2, ..., n_1$; $j = 1, 2, ..., n_0$. The total number of tied pairs can be written as

$$T = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} 1_{[\hat{\pi}_{1i} = \hat{\pi}_{0j}]}.$$

With these three count measures, we are able to calculate the c-statistic, providing an estimate for the AUC.

1.2.4 Traditional C-Statistic

According to Royston and Altman (2010), the c-statistic quantifies the ability of the model to discriminate between event and non-event observations. It is

expressed as the ratio of the number concordant pairs plus half the number of ties to the total number of all event/non-event pairs:

$$c = \frac{C + \frac{T}{2}}{C + D + T} = \frac{C + \frac{T}{2}}{n_1 n_0}.$$

Because the c-statistic estimates the AUC, it can be regarded as the empirical probability over all of the observations in a data set that the fitted model will yield a higher predicted probability for an observed event outcome than for an observed non-event outcome (Royston and Altman, 2010). Other interpretations include the average value of model sensitivity for all possible values of specificity (Zhou, Obuchowski and McClish, 2002), as well as the average value of model specificity for all possible values of sensitivity (Metz, 1986, 1989). Steyerberg, Vickers, Cook, et al. (2010) outline new measures that assess the performance of a predictive model, including variants of the c-statistic for survival analysis (Heagerty and Zheng, 2005; Gonen and Heller, 2005), reclassification tables (Cook, 2007), and integrated discrimination improvements (IDI) (Pencina, D'Agostino, D'Agostino and Vasan, 2008).

1.2.5 Properties and Simulated Results

The c-statistic has properties similar to the coefficient of determination (R^2), the most popular measure of predictive efficacy for linear regression. Specifically, the addition of predictor variables to the fitted model, significant or not, generally increases the value of the measure. Although the c-statistic can decrease with variable additions, such decreases are rare and tend to be quite small.

As with the AUC, the c-statistic takes on values over the range [0,1], although any value at or below 0.5 indicates that the fitted model is not an effective discriminating mechanism. In such a case, an equally effective approach would be randomly

generating the outcome using a Bernoulli distribution with the sample event prevalence as the event probability. Based on the value of the c-statistic, a simple set of guidelines adopted from Hosmer and Lemeshow (p. 162, 2000) for interpreting the quality of predictive discrimination for the fitted model is as follows:

$c \leq 0.5$	no discrimination
$0.5 < c < 0.7$	poor discrimination
$0.7 \leq c < 0.8$	acceptable discrimination
$0.8 \leq c < 0.9$	excellent discrimination
$0.9 \leq c \leq 1.0$	outstanding discrimination.

The form of the systematic component of a model and the size of the sample can influence the behavior of the c-statistic. An important characteristic of the systematic component is the signal of the model. Suppose that β_0 and X respectively denote the vector of regression parameters and a covariate vector for the generating model. For the purpose of definition, the elements of the covariate vector are regarded as random. The model signal is defined by $\text{Var}[X'\beta_0]$. This quantity plays a key role in the expectation of the c-statistic. An increase in the variability of the X variables and/or the size of the β_0 parameters leads to an increase in the signal. Let $e^{\lambda} = \text{Var}[X'\beta_0]$, so that λ denotes the log-signal.

Here, we provide a small simulated example to demonstrate how the behavior of the c-statistic is governed by the model signal. Our model contains a single predictor variable,

$$X \sim \text{Uniform}(-1, 1).$$

The outcome is

$$Y \sim \text{Bernoulli}(\pi),$$

where π is determined as follows:

$$\pi = \frac{1}{1 + e^{-\beta_0 X}}.$$

The parameter β_0 is generated as a function of the signal:

$$\beta_0 = \sqrt{3e^{\lambda}}.$$

In this example, the log-signal spans the range $[-5, 5]$. We consider the sample sizes $N = 25, 50, 100, 1000$. Using 100 replications per combination of sample size and log-signal, we illustrate the behavior of the c-statistic in Figure 1.1. Note that a low log-signal tends to yield a c-statistic closer to 0.5, while a high log-signal yields a c-statistic closer to 1. Figure 1.2 displays the means of the c-statistic over the range of the log-signal by the sample size. The means are quite similar for each sample size for a log-signal exceeding zero, whereas there appears to be some separation in the means for a log-signal less than zero, with smaller sample sizes yielding larger mean c-statistics. Due to the high variability of the c-statistic for a small sample size and a low log-signal, along with the meaningful lower bound of 0.5, the data are right skewed. Since the tails of the skewed data are longer and thicker for the smaller sample sizes, the corresponding means are pulled higher.

1.2.6 Predicted Probabilities as Weights: Motivating Example

Although the counts C, D, and T offer information about the conformity of the fitted model to the data at hand, essential information regarding the magnitude of the difference in predicted probabilities is being ignored in their calculation. Consider two pairs of event/non-event observations and their predicted probabilities: $(\hat{\pi}_{1i_1}, \hat{\pi}_{0j_1}) = (0.95, 0.10)$; $(\hat{\pi}_{1i_2}, \hat{\pi}_{0j_2}) = (0.52, 0.50)$ for $i_1, i_2 \in 1, \ldots, n_1$; $j_1, j_2 \in 1, \ldots, n_0$. Both pairs of observations are concordant since the predicted probabilities for the observed event outcomes are higher than those for their respective observed non-event outcomes. As a result, each pair counts equally in its contribution to C. However, the difference in predicted probabilities is substantially larger for the first pair, indicating the ability of the model to better discriminate that pair as opposed

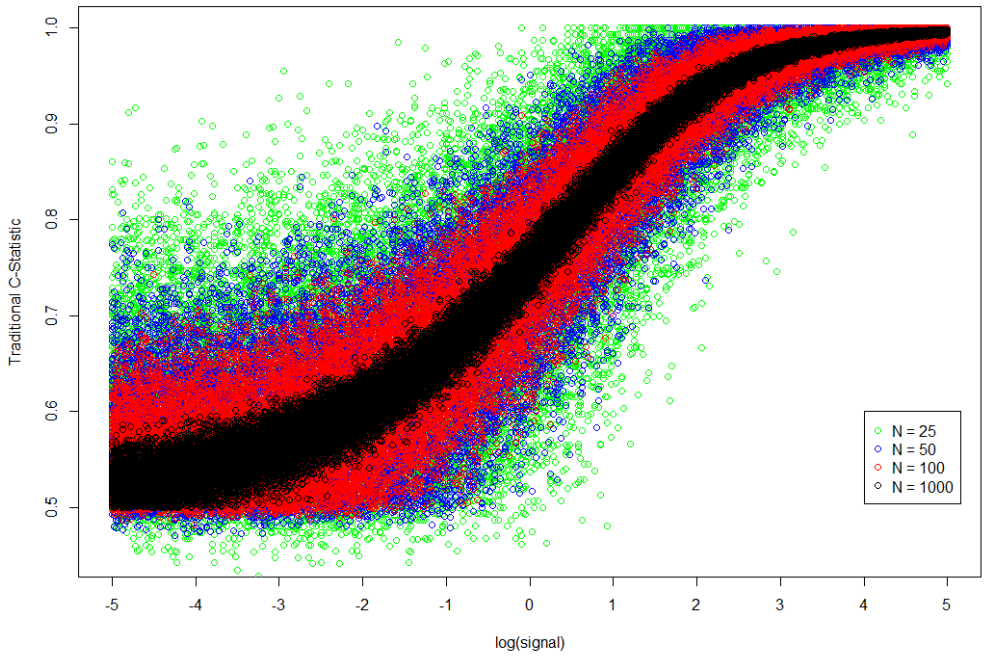

Figure 1.1: Plot of traditional c-statistic vs. log-signal by sample size.

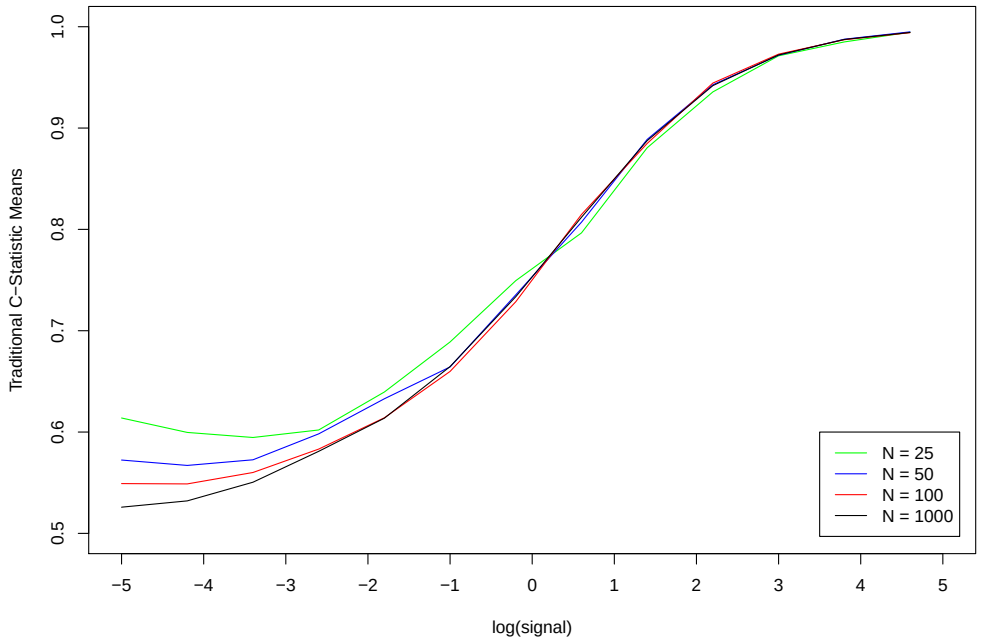

Figure 1.2: Plot of traditional c-statistic means vs. log-signal by sample size.

to the second. This additional information could be valuable in constructing a new version of the c-statistic.

In the introduction, we mentioned that information is ignored in calculating C, D, and T. If we reconsider the way in which we calculate these three measures, we can provide an adjusted variant of the c-statistic that may serve as a more informative measure of predictive discrimination for the fitted model.

1.3 New Method

This section introduces a variant of the c-statistic that utilizes the ignored information regarding differences in predicted probabilities for event/non-event pairs. After presenting properties and results regarding the adjusted c-statistic, an illustrative simulated example is provided to compare and contrast the traditional and adjusted c-statistics.

1.3.1 Adjusted C-Statistic

Statistical methodology occasionally employs a weighting scheme when calculating measures. This approach allows for observations to contribute proportionally to their assessed value in the fitted model. Examples of this technique include weighted least squares in fitting regression models and sample survey weights in computing representative summary statistics. Applying weights in the development of inferential constructs often adds a dimension of information, leading to more refined statistics.

The differences in predicted probabilities for the events and non-events across all observed event/non-event pairs are highly informative, since a larger difference indicates greater adherence to the principle of concordance or discordance. Utilizing these differences as weights in the formulation of C, D, and T and their associated statistics would provide information on similar scales and lead to a more nuanced

characterization of model discrimination. The equations for the weighted versions of C, D, and T are as follows:

$$C' = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} |\hat{\pi}_{1i} - \hat{\pi}_{0j}| 1_{[\hat{\pi}_{1i} > \hat{\pi}_{0j}]},$$

$$D' = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} |\hat{\pi}_{1i} - \hat{\pi}_{0j}| 1_{[\hat{\pi}_{1i} < \hat{\pi}_{0j}]},$$

$$T' = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} |\hat{\pi}_{1i} - \hat{\pi}_{0j}| 1_{[\hat{\pi}_{1i} = \hat{\pi}_{0j}]} = 0.$$

Noting that $T' = 0$, the calculation of the adjusted c-statistic then parallels the calculation of the traditional c-statistic:

$$c_{adj} = \frac{C'}{C' + D'}.$$

1.3.2 Properties and Simulated Results

The adjusted c-statistic resembles the traditional c-statistic; however, some subtle yet important differences exist between the measures. As with the traditional c-statistic, the adjusted c-statistic has properties similar to R^2. In particular, the addition of predictor variables tends to monotonically increase the value of the measure. Unlike the traditional c-statistic, however, the range of the adjusted c-statistic is (0.5,1], meaning that a fitted model with no predictive utility cannot yield a statistic ≤ 0.5. The fact that the lower bound of the adjusted c-statistic is 0.5 has been validated through simulation. Although a mathematical proof would be more compelling, formulating a rigorous argument has been elusive. The absolute lower bound of 0.5 for the adjusted c-statistic is a compelling property, since values of the traditional c-statistic that fall below 0.5 are difficult to reconcile and interpret.

The set of guidelines for assessing the discriminating ability of a fitted model based on the c-statistic, as outlined in the previous section, can also be applied to the adjusted c-statistic, with the caveat that the adjusted c-statistic cannot assume

a value at or below 0.5.

As we observed with the traditional c-statistic, the adjusted c-statistic is influenced by the sample size and signal. Once again, we provide a small simulated example to demonstrate how the behavior of the adjusted c-statistic is regulated by these factors. We consider the same set of conditions for the predictor variable, X, and the outcome, Y, as well as the same sample sizes and range of values for the log-signal. Based on 100 replications per combination of sample size and log-signal, the behavior of the adjusted c-statistic is illustrated in Figure 1.3. As we observed with the traditional formulation, a low log-signal tends to yield an adjusted c-statistic closer to 0.5, while a high log-signal yields a statistic closer to 1. The same relationship seen before between sample size and variance also exists for the adjusted c-statistic. Figure 1.4 displays the means of the adjusted c-statistic over the range of log-signal by sample size. We see the same pattern that we observed with the means of the traditional c-statistic. For a log-signal exceeding zero, the means are quite similar but there is separation of the means for a log-signal less than zero. Due to the high variability of the adjusted c-statistic for a small sample size and a low log-signal, along with the strict lower bound of 0.5, the data are right skewed. Again, since the tails of the skewed data are longer and thicker for the smaller sample sizes, the corresponding means are pulled higher.

1.3.3 Illustrative Example

At first glance, the figures for the traditional and adjusted c-statistics appear quite similar; thus, it may not be entirely clear exactly what is the relationship between the two measures. In order to better illustrate the relationship, we consider Figures 1.5 and 1.6, which use the c-statistic values from the previous two sets of figures. Each data point in Figure 1.5 corresponds to a pair of traditional and adjusted c-statistics resulting from a specific fitted model. The figures include a

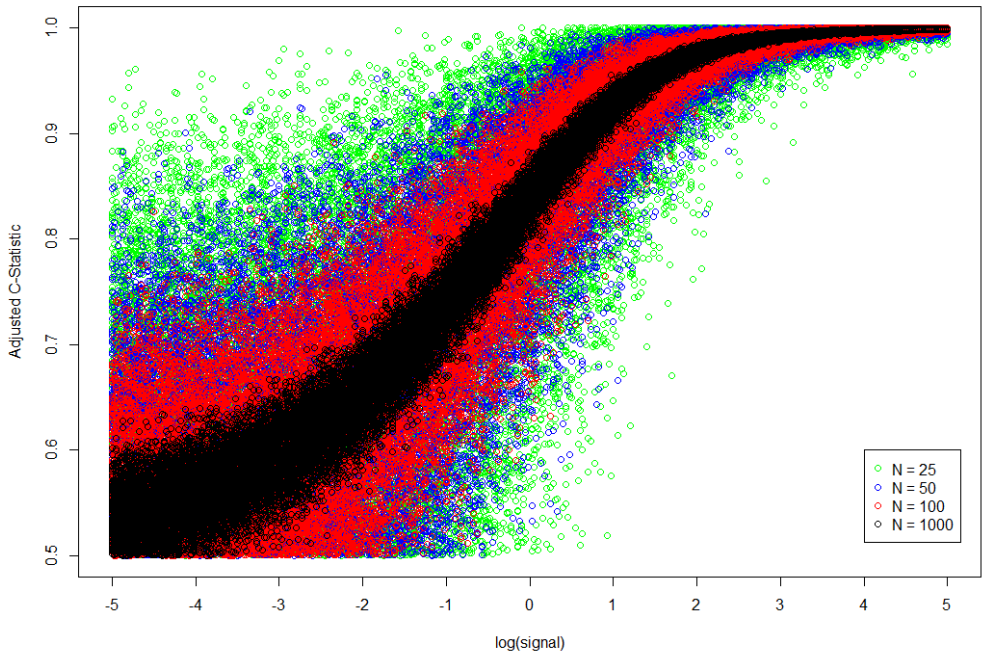

Figure 1.3: Plot of adjusted c-statistic vs. log-signal by sample size.

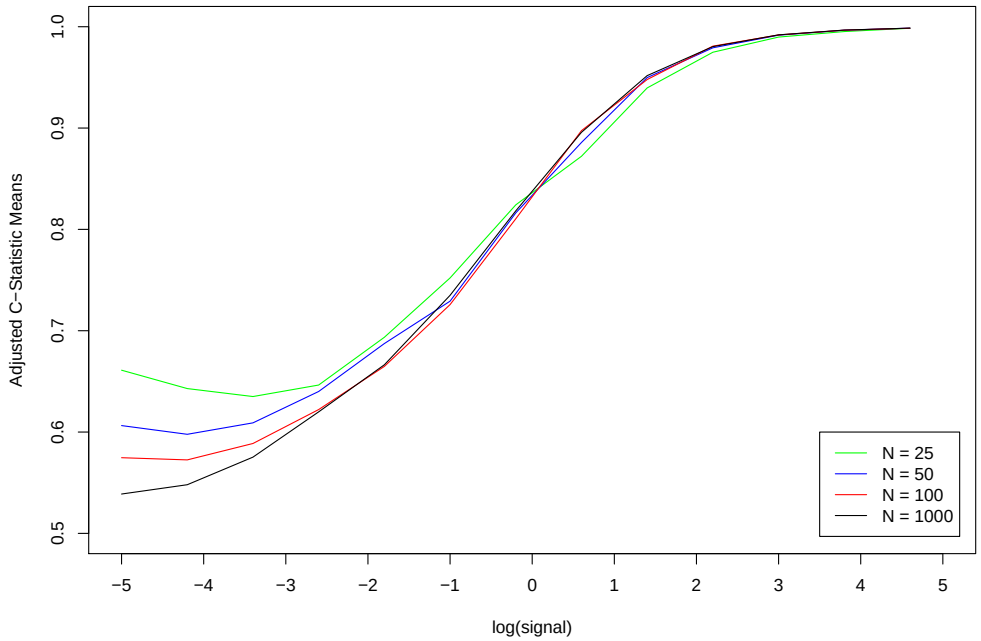

Figure 1.4: Plot of adjusted c-statistic means vs. log-signal by sample size.

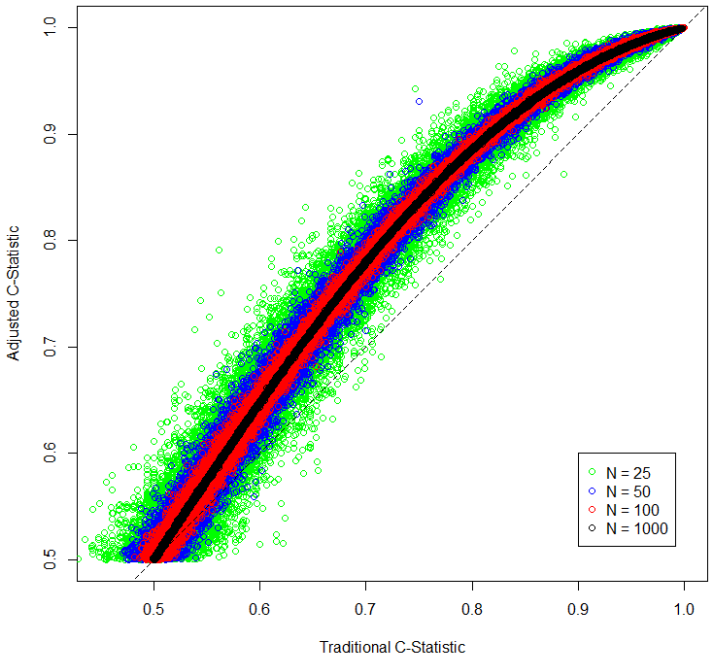

Figure 1.5: Plot of adjusted c-statistic vs. traditional c-statistic by sample size.

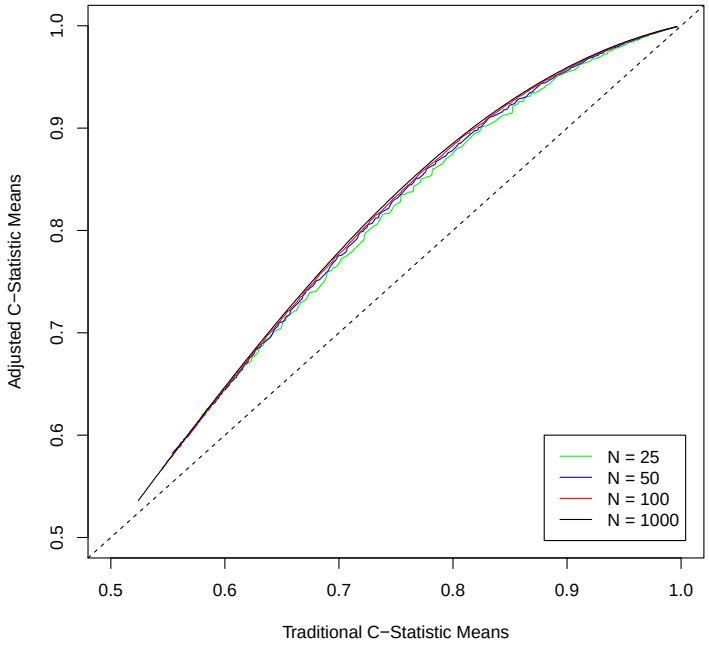

Figure 1.6: Plot of adjusted c-statistic means vs. traditional c-statistic means by sample size.

45° line which represents equal values for the traditional and adjusted c-statistics. When their values are near the edges of the ranges, the traditional and adjusted c-statistics are quite similar, both being close to 0.5 or 1. However, for values around the center of the range, the differences between the traditional and adjusted c-statistics are more pronounced, typically with the adjusted measures markedly larger than their traditional counterparts. This is emphasized by the plot of the means. The inflation of the adjusted c-statistic can be attributed to a tendency for the concordant pairs to yield larger weights than the discordant pairs.

Each sample size seems to be following a similar pattern, with smaller variance as the sample size increases. The c-statistic pairs where $N = 1000$ seem to form a clear curve that can be conceptualized as the expected relationship between the traditional and adjusted c-statistics. Curiously, based on our investigations, the expected relationship appears to be generalizable to any data set and corresponding fitted model. This expected relationship is the key to understanding how the weights applied to the adjusted c-statistic can further inform us as to the discriminating ability of the fitted model. In other words, the pair of traditional and adjusted c-statistics is substantially more valuable than either individual measure.

1.4 Application

Now that we have seen how the measures behave in a simulated setting, we can apply them in an actual modeling application and examine the results.

Heart disease, which refers to the narrowing or blocking of blood vessels and can induce heart attack or stroke, is the leading cause of death in the United States, claiming nearly 600,000 lives each year. Angiography, a medical imaging technique, allows doctors to visualize the inside of blood vessels so that narrowing can be assessed and a diagnosis about heart disease can be made.

The Cleveland database is a data set that contains 303 patient-specific records,

representing 14 attributes. The outcome of interest is the angiographic diagnosis of heart disease, which assumes the value 0 for patients with $< 50\%$ blood vessel diameter narrowing and 1 for patients with $> 50\%$ blood vessel diameter narrowing. The candidate predictor variables include age of the patient (`age`; in years, discrete), sex of the patient (`sex`; 0 = female, 1 = male), chest pain type (`cp`; 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic), resting systolic blood pressure on admission (`trestbps`; continuous), serum cholesterol (`chol`; continuous), fasting blood sugar indicator (`fbs`; 0 = false, 1 = true), resting electrocardiographic result (`restecg`; 0 = normal, 1 = ST-T wave abnormality, 2 = left ventricular hypertrophy), maximum heart rate achieved (`thalach`; continuous), exercise induced angina (`exang`; 0 = no, 1 = yes), ST depression induced by exercise (`oldpeak`; continuous), slope of peak exercise ST segment (`slope`; 1 = upsloping, 2 = flat, 3 = downsloping), number of major vessels colored by fluoroscopy (`ca`; discrete in [0,3]), and thallium stress test result (`thal`; 3 = normal, 6 = fixed defect, 7 = reversible defect).

Since we have 13 potential predictor variables, the all subsets setting will consider a candidate collection comprised of $2^{13}-1 = 8191$ models. Once each model is fitted, we calculate the traditional and adjusted c-statistics. Figure 1.7 shows the relationship between the calculated c-statistics for both approaches, with the large-sample expected relationship curve and a 45° line included for reference. Most of the models generate points that are near the expectation curve, underlining the generalizability of the relationship between the traditional and adjusted c-statistics.

Table 1.1 presents the traditional and adjusted c-statistics for three selected models from the Cleveland data set. Model 1 has an order of two, while models 2 and 3 have an order of five. Also, model 1 is nested within both models 2 and 3. Model 1 has a markedly smaller traditional c-statistic than models 2 and 3, which may not be initially viewed as surprising, since the inclusion of more predictors

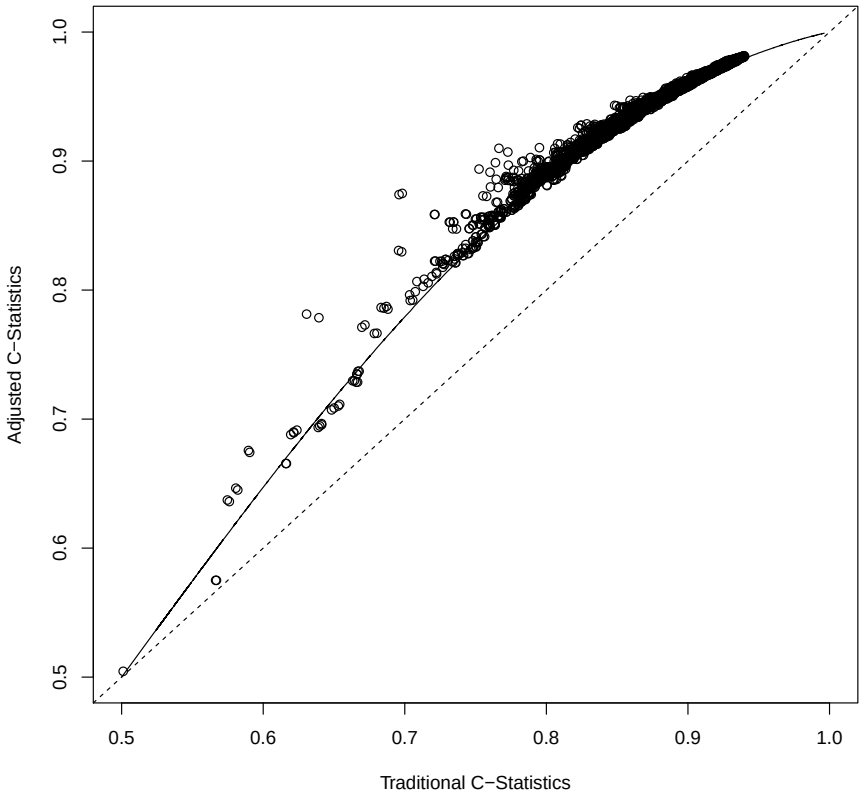

Figure 1.7: Plot of adjusted c-statistic vs. traditional c-statistic for Cleveland data set.

tends to increase the measure. Based on the implication regarding predictive discrimination arising from the use of traditional c-statistics, one may favor model 2 or 3 over model 1, despite their lack of parsimony compared to model 1. However, the adjusted c-statistics tell a different story. All three adjusted c-statistics are nearly the same, indicating that the three fitted models have nearly identical predictive discrimination. Based on the weighted measure, we are inclined to favor model 1 due to its parsimony.

The traditional c-statistic is a valuable measure but may misrepresent the ability of a fitted model to discriminate between events and non-events based on predictive probabilities. The adjusted c-statistic offers additional information about

Model Number	Traditional c	Adjusted c	Model Order	Model Predictors
1	0.6959	0.8740	2	thalach oldpeak
2	0.7833	0.8799	5	sex cp fbs thalach oldpeak
3	0.7858	0.8792	5	cp fbs thalach exang oldpeak

Table 1.1: Three selected models from Cleveland data set.

the fitted model that could lead to different conclusions. By examining both the traditional and adjusted c-statistics, we are able to more effectively assess the nature of predictive discrimination resulting from the fitted model. In the application at hand, when the magnitudes of the differences in predictive probabilities for event/non-event pairs are taken into account, a parsimonious two-variable model performs as well as two more complex five-variable models. These differences are important, because in reality, the actual probabilities from the fitted model are used to predict outcomes. (In this setting, an assessment of the likelihood of heart disease would be based on the predicted probability resulting from covariate information for a patient.) When only the ordering of the predictive probabilities are considered, the two five-variable models appear superior.

1.5 Summary Conclusion

This chapter has introduced an adjusted variant of the c-statistic in the logistic regression framework that utilizes information ignored in the construction of the traditional measure. While these two measures share similar properties and behaviors, by pairing the two measures together, we are able to add a dimension to the assessment of model predictive discrimination. Future work involves the development and investigation of an adjusted variant of the c-statistic for ordinal outcomes comprised of three or more levels.

CHAPTER 2
MODEL SELECTION CRITERIA BASED ON CROSS-VALIDATORY CONCORDANCE STATISTICS

2.1 Introduction

In regression frameworks, model selection procedures are often designed and implemented to emphasize prediction. Predictive capability can be quantified for various fitted models in a candidate collection, allowing for model comparison, selection, and evaluation. In logistic regression, the c-statistic (or concordance statistic) is the most popular measure of predictive efficacy. An adjusted c-statistic has been introduced in Chapter 1 that utilizes the differences in predicted probabilities for each of the concordant and discordant event/non-event observation pairs. In contrast, the traditional c-statistic focuses on the mere ordering of these probabilities. Although both measures inform investigators as to the predictive efficacy of a given fitted model, they cannot be used directly for the purpose of model selection because they tend to increase as complexity is added to the model.

The purpose of this chapter is twofold. First, we introduce measures, or parameters, that can serve as targets for the development of model selection criteria based on the traditional and adjusted c-statistics. Second, we propose and investigate estimators of these measures using cross-validated versions of the c-statistics. We aim to show that these estimators function as suitable model selection criteria in the logistic regression framework.

Analogous developments will be presented for both the traditional and adjusted c-statistics, allowing for a comparison of properties for each approach. The new criteria are designed to identify a model that balances goodness-of-fit and parsimony, and achieves generalizability. If we are inclined to believe we have access to the true model in a setting where we formulate candidate models based on all

possible subsets of predictors, this identified model will ideally feature the subset of predictors that defines the true model.

The structure of this chapter is as follows. In section 2, we provide some background regarding logistic regression, model selection, the traditional and adjusted c-statistics, and cross-validation methods. In section 3, we propose a new set of model selection criteria for logistic regression modeling. We first propose two unique measures for the purpose of model delineation. We discuss the intrinsic features of each, focusing on their relative sensitivities towards the detection of underfitting and overfitting. We then additively combine the measures to form a composite measure that reflects the strengths of both. We subsequently propose estimators for the target measures, and justify the propriety of these estimators in a small investigative simulation. In section 4, we present a simulation study to characterize and evaluate the behavior of the selection criteria based on empirical results. In section 5, we apply the new selection criteria in a modeling application and examine the results. Section 6 concludes.

2.2 Background

This section provides background information on the traditional and adjusted c-statistics. Additionally, it covers cross-validation, which plays an integral role in the formulation of the estimators for our model selection target measures. These measures will be introduced in the subsequent section.

2.2.1 Logistic Regression

Logistic regression is the most popular modeling framework for data containing a Bernoulli (0/1) outcome. In fitting a logistic regression model, an assessment of predictive discrimination is often of interest. A common construct that facilitates this assessment is the receiver operating characteristic (ROC) curve. The area under

the ROC curve (AUC) measures the probability that, for a random event/non-event observation pair drawn from the population, the modeled probability for the event observation will be greater than that for the non-event observation (Hanley and McNeil, 1982). An AUC close to 1 indicates strong predictive discrimination while a value less than 0.5 indicates that the model is no more suitable a discriminating mechanism than randomly generating predicted outcomes using the event prevalence as the Bernoulli probability. The (0,0)-(1,1) line segment on an ROC curve, which corresponds to an AUC = 0.5, is referred to as the chance diagonal (Zhou, Obuchowski and McClish, 2002). The AUC is estimated by the c-statistic, or the sample proportion of event/non-event pairs that are ordered correctly by the fitted model.

2.2.2 Model Selection Criteria

Statistical models are used to characterize the relationship between an outcome of interest and explanatory factors. Models condense information into an interpretable form, from which investigators can draw inferential conclusions. Modeling frameworks have been developed to handle outcomes that assume distributions of all varieties. Once fit, models can be applied to new data in order to predict new outcomes.

An optimal statistical model is characterized by three features: (1) parsimony, which refers to model simplicity; (2) goodness-of-fit, which indicates the conformity of the fitted model to the data at hand; (3) generalizability, which reflects the ability of the fitted model to predict or describe new outcomes. Parsimony and goodness-of-fit tend to pull in opposing directions with regards to model complexity, so it is important to strike a suitable balance between those two attributes, while still achieving generalizability.

In a model selection problem, an investigator strives to find the "best" model

from a collection of candidate models, where optimality may be defined based on adherence to the preceding principles. For theoretical and methodological developments pertaining to model selection, one needs to assume the existence of an underlying generating probabilistic mechanism. We will refer to this mechanism as a true model. In our development, we will assume that the true model is contained within the candidate collection. Although this is a strong assumption, it is commonly employed in model selection developments for either mathematical tractability or conceptual clarity. Here, we impose the assumption for the benefit of the latter.

Investigators frequently use model selection criteria in order to compare different candidate models and ascertain the one that best exemplifies the three optimality features. A common approach to the development of a model selection criterion is to estimate a measure that assesses the disparity between the fitted model under consideration and the true probabilistic mechanism. Such a measure is known as an *expected discrepancy*. One of the most popular and useful expected discrepancies is based on the Kullback-Leibler (K-L) information, a measure introduced by Kullback and Leibler (1951) and further investigated by Kullback (1968). This discrepancy serves as the basis for the ubiquitous Akaike (1973, 1974) information criterion (AIC) and its variants. One of the favorable properties of AIC is asymptotic efficiency, in the sense of Shibata (1980, 1981). Assuming that the generating model is of an infinite dimension and thus is not in the candidate collection, an efficient criterion will asymptotically select the fitted candidate model which minimizes the mean squared error of prediction (Cavanaugh, 1999). As outlined by Linhart and Zucchini (1986), a major deficiency of AIC arises in small to moderate sample-size applications, where AIC will often severely underestimate the K-L discrepancy and may tend to decrease as model complexity increases. Variants of AIC have been

proposed to address this deficiency and to relax the stringent model specification assumption under which the criterion is derived. These include corrected AIC (AICc) (Sugiura, 1978; Hurvich and Tsai, 1989), designed for small-sample settings, the Takeuchi (1976) information criterion (TIC), which relaxes the model specification assumption, CAIC (Bozdogon, 1987), which corrects for the lack of consistency, and a quasi-likelihood based measure for generalized linear models fit via generalized estimating equations (QIC) (Pan, 2001). Additional variants of AIC have been proposed based on complexity penalizations that are evaluated using a computationally intensive algorithm, including cross-validation (Stone, 1977; Davies, Neath and Cavanaugh, 2005), bootstrapping (Ishiguro, Sakamoto and Kitagawa, 1997; Cavanaugh and Shumway, 1997; Shibata, 1997), and Monte Carlo simulation (Hurvich, Shumway, and Tsai, 1990; Bengtsson and Cavanaugh, 2006).

The mean structure of a model is typically characterized using a linear combination of predictor variables. In order to ensure that an exhaustive model search has been conducted, investigators use the all possible subsets approach, which allows for comparison of fitted candidate models based on every combination of variables. For a set of p predictor variables, the all subsets approach considers the $2^p - 1$ collections of predictor variables, excluding the null model. Once all models have been fitted, corresponding model selection criteria are compared to determine which model is "best." For an effective criterion, in large sample settings where the underlying effects are all appreciable in magnitude, the true model will generally have the highest probability of being chosen out of the collection of candidate models. However, in such settings, it should be noted that model overfitting has a less detrimental impact on inferential objectives than underfitting.

2.2.3 Traditional and Adjusted C-Statistics

In logistic regression, the area under an ROC curve, or AUC, is a measure of predictive efficacy estimated by the c-statistic. The c-statistic is calculated using the predicted orderings of event/non-event pairs. Consider a data set with n_1 observations where $y = 1$, and n_0 observations where $y = 0$. The total number of event/non-event pairs is given by $n_1 n_0$.

A pair is said to be concordant if the predicted probability for the event outcome is greater than the predicted probability for the non-event outcome, $\hat{\pi}_{1i} > \hat{\pi}_{0j}$ for $i = 1, 2, ..., n_1$; $j = 1, 2, ..., n_0$. The total number of concordant pairs can be expressed as

$$C = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} 1_{[\hat{\pi}_{1i} > \hat{\pi}_{0j}]}.$$

A pair is said to be discordant if the predicted probability for the event outcome is less than the predicted probability for the non-event outcome, $\hat{\pi}_{1i} < \hat{\pi}_{0j}$ for $i = 1, 2, ..., n_1$; $j = 1, 2, ..., n_0$. The total number of discordant pairs can be expressed as

$$D = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} 1_{[\hat{\pi}_{1i} < \hat{\pi}_{0j}]}.$$

A pair is said to be tied if the predicted probability for the event outcome is equal to the predicted probability for the non-event outcome, $\hat{\pi}_{1i} = \hat{\pi}_{0j}$ for $i = 1, 2, ..., n_1$; $j = 1, 2, ..., n_0$. The total number of tied pairs can be expressed as

$$T = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} 1_{[\hat{\pi}_{1i} = \hat{\pi}_{0j}]}.$$

The traditional c-statistic is the ratio of the number of concordant pairs plus half the number of ties to the total number of all event/non-event pairs:

$$c = \frac{C + \frac{T}{2}}{C + D + T} = \frac{C + \frac{T}{2}}{n_1 n_0}.$$

As with the AUC, the c-statistic takes on values over the range [0,1], although any value at or below 0.5 indicates that the fitted model is not an effective discriminating mechanism.

A variant of this statistic that we proposed and investigated in Chapter 1 is the adjusted c-statistic, which applies the absolute difference in predicted probabilities as a weight for each event/non-event pair when calculating C, D, and T. The equations for the weighted versions of C, D, and T are as follows:

$$C' = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} |\hat{\pi}_{1i} - \hat{\pi}_{0j}| 1_{[\hat{\pi}_{1i} > \hat{\pi}_{0j}]},$$
$$D' = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} |\hat{\pi}_{1i} - \hat{\pi}_{0j}| 1_{[\hat{\pi}_{1i} < \hat{\pi}_{0j}]},$$
$$T' = \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} |\hat{\pi}_{1i} - \hat{\pi}_{0j}| 1_{[\hat{\pi}_{1i} = \hat{\pi}_{0j}]} = 0.$$

The adjusted c-statistic is calculated as follows:

$$c_{adj} = \frac{C'}{C' + D'}.$$

The range of the adjusted c-statistic is (0.5,1]. The meaningful lower bound for the traditional c-statistic is an absolute and exclusive lower bound for the adjusted c-statistic.

The adjusted c-statistic resembles the traditional c-statistic; however, as discussed in Chapter 1, some subtle yet important differences exist between the measures. As with the traditional c-statistic, the adjusted c-statistic has properties similar to R^2. In particular, the addition of predictor variables tends to monotonically increase the value of the measure. As a result of this tendency, neither c-statistic serves as a useful model selection criterion. However, alternate approaches to formulating the c-statistics, such as approaches based on cross-validation, allow for both c-statistics to play a more meaningful role in model selection.

2.2.4 Cross-Validation

Cross-validation is a technique for assessing the predictive efficacy of a model without new data. As with the bootstrap, the primary value of cross-validation is that the original data set is used for both fitting and validation, without splitting the sample into disjoint training and validation subsets that are used exclusively for these purposes. Both techniques employ numerous fitting samples and validation sets through multiple iterations. Although the techniques may require a substantial amount of computing, the computational demands are generally not prohibitive, even for moderate to large sample sizes. The use of bootstrapping and cross-validation for improving the performance of a prediction rule is investigated by Efron (1983, 1986).

Leave-k-out is a form of exhaustive cross-validation in which all combinations of k observations are removed from the training set and used as validation. The leave-one-out approach is the simplest and most often utilized approach, since it requires the least number of iterations of any leave-k-out cross-validation. The leave-one-out cross-validation approach will be used to provide c-statistics that can serve as the basis for estimators for our proposed target measures. These cross-validatory estimators can then be used as model selection criteria.

2.3 New Method

This section introduces two unique prediction error measures based on c-statistics for the purpose of model delineation. We discuss the intrinsic features of each, focusing on their relative sensitivities towards the detection of underfitting and overfitting. We then additively combine the measures to form a composite measure that reflects the strengths of both. We subsequently propose estimators for the target measures, and justify the propriety of these estimators in a small simulation

study.

2.3.1 Prediction Error Measures

To introduce our measures, we employ the following notation for the c-statistics: $c(data, \theta)$, where $\theta = (\beta_1, \beta_2, ..., \beta_p)$ is the regression parameter vector. The data are either the outcome vector for the fitting sample, Y, or a new outcome vector, Z, which serves as our validation data and follows from the same distribution as Y. The parameter θ is either the fitted model parameter vector estimator, $\hat{\theta}$, or the true parameter vector, θ_0.

In order to create new model selection criteria, we consider measures that are defined as expectations of differences of c-statistics. These measures, which are inherently non-negative, are referred to as *prediction errors*. A model that minimizes each measure is deemed optimal. The three target measures we propose in this section have been named the *model misspecification prediction error*, the *fitting sample prediction error*, and the *sum of prediction errors*.

2.3.1.1 Model Misspecification Prediction Error

The first measure that we introduce is the model misspecification prediction error (MMPE). Consider the positive difference $E[c(Z, \theta_0) - c(Z, \hat{\theta})]$. Both c-statistics are based on a new outcome, Z, but the first uses the true parameter vector, θ_0, while the second uses the fitted parameter vector, $\hat{\theta}$. This measure can be conceptualized as the error induced when the fitted model is used to predict new data as opposed to the true model. The MMPE shows more sensitivity to underfitting than overfitting due to its reflection of the bias in the predictions for the misspecified fitted model. For underspecified models, the fitted parameter vector, $\hat{\theta}$, is based on estimating a subset of the parameters in θ_0, and is attempting to explain an outcome without all of the necessary predictors, leading to biased predicted probabilities and

inaccurate predictive discrimination. For overspecified models, the fitted parameter vector contains estimates for all of the parameters in θ_0, as well as a few extraneous ones. Although the addition of predictors not essential to the model affects the variability of parameter estimators, unless the excess variability is substantial, the error measures should be smaller than those for any underspecified models. For a correctly specified model, $\hat{\theta}$ estimates the same exact parameters in θ_0, which should lead to a minimum MMPE. For large sample sizes, the error for correctly specified or overspecified models reduces to zero since $\hat{\theta}$ converges to θ_0.

2.3.1.2 Fitting Sample Prediction Error

The next measure that we propose is the fitting sample prediction error (FSPE). Consider the positive difference $E[c(Y, \hat{\theta}) - c(Z, \hat{\theta})]$. Both c-statistics are based on the fitted parameter, $\hat{\theta}$, but the first uses the fitting sample outcome, Y, and the second uses a new outcome, Z. We can conceptualize the FSPE as the measure of the overoptimism that arises when the fitted model is used to predict the data used in its own construction as opposed to new data. The FSPE is more sensitive to overfitting than underfitting due to its reflection of the conformity of the fitted model to the data at hand. For underspecified models, $\hat{\theta}$ does not contain estimates for all of the elements of θ_0, so we are not overly optimistic about the fitting sample yielding a substantially larger c-statistic than the new outcome. Therefore, the error incurred by using new data instead of the fitting sample is relatively small; in either case, predictive discrimination is dominated by bias. For overspecified models, the bias becomes immaterial. In this setting, the expected difference between $c(Y, \hat{\theta})$ and $c(Z, \hat{\theta})$ is inflated since the fitted parameter better conforms to the fitting outcome than the new outcome. As complexity is added to the model, there is an increase in variability for $c(Z, \hat{\theta})$ that is not reflected in $c(Y, \hat{\theta})$, increasing the error.

2.3.1.3 Sum of Prediction Errors

An estimator for either the MMPE or FSPE should be a sensible model selection criterion, but each error measure has a weakness when it comes to the identification of misspecified models. Fortunately, the strengths of each measure compensate for the shortcomings of the other. The MMPE is better at protecting from underfitting and the FSPE is better at protecting from overfitting. Since both measures are expected differences of c-statistics, they are comparable in scale. If we consider combining them additively, we can create a predictive error measure that may be favorable to either of its individual constituents. The sum of prediction errors (SUPER) is expressed as $E[c(Z,\theta_0) + c(Y,\hat{\theta}) - 2c(Z,\hat{\theta})]$. The improved protection from both underfitting and overfitting should result in a measure that exhibits a sharper minimum when the fitted model is correctly specified.

2.3.2 C-Statistics as Model Selection Criteria

The prediction error measures introduced in the previous section, most notably the SUPER, are ideal targets for model selection criteria due to the manner in which they reflect misspecification. Initially, it would appear that the simplest way to construct estimators of these measures is to begin by considering the statistics inside of the expectations.

In order to estimate the MMPE, expressed as $E[c(Z,\theta_0) - c(Z,\hat{\theta}))]$, we can consider the difference of the individual expectations of each c-statistic. The value $E[c(Z,\theta_0)]$ is inaccessible; however, since it does not depend on the fitted model, it is a constant that does not need to be estimated for the purpose of constructing a selection criterion. We therefore need only to estimate $E[c(Z,\hat{\theta})]$. An appropriate estimator would seemingly be $c(Z,\hat{\theta})$, which is certainly unbiased for its own expectation. However, as we will discuss in the next subsection, $c(Z,\hat{\theta})$ suffers from high variability, which limits its usefulness as a selection criterion. Moreover, a true

validation sample Z is generally unattainable. However, we can use a leave-one-out variant of the c-statistic, say $c_{L1O}(Y, \hat{\theta})$, which will serve as a valid estimator of $E[c(Z, \hat{\theta})]$ in large sample settings. For a given data set, the fitted model that yields the smallest $c(z, \theta_0) - c_{L1O}(y, \hat{\theta})$, or equivalently the largest value of $c_{L1O}(y, \hat{\theta})$, is deemed the best.

In order to estimate the FSPE, expressed as $E[c(Y, \hat{\theta}) - c(Z, \hat{\theta})]$, we can again consider the difference of the individual expectations of each c-statistic. Similar to the MMPE, we can use $c_{L1O}(Y, \hat{\theta})$ as an estimator of $E[c(Z, \hat{\theta})]$. The c-statistic $c(Y, \hat{\theta})$ serves as an estimator of its own expectation. Thus, we propose as an estimator for $E[c(Y, \hat{\theta}) - c_{L1O}(Y, \hat{\theta})]$ the statistic $c(Y, \hat{\theta}) - c_{L1O}(Y, \hat{\theta})$. For a given data set, the fitted model that yields the smallest value of $c(y, \hat{\theta}) - c_{L1O}(y, \hat{\theta})$ is deemed the best.

In order to estimate the SUPER, which is the sum of the MMPE and FSPE, we can simply take the sum of their estimators. Therefore, we estimate $E[c(Z, \theta_0) + c(Y, \hat{\theta}) - 2c(Z, \hat{\theta})]$ by using $c(Z, \theta_0) + c(Y, \hat{\theta}) - 2c_{L1O}(Y, \hat{\theta})$. As with the previous two measures, the model that yields the smallest $c(z, \theta_0) + c(y, \hat{\theta}) - 2c_{L1O}(y, \hat{\theta})$ is deemed the best.

Now that we have proposed estimators for the MMPE, FSPE, and SUPER, we can test their efficacy as model selection criteria in a simulation study. However, we need first establish that $c_{L1O}(Y, \hat{\theta})$ serves as a valid estimator of $E[c(Z, \hat{\theta})]$.

2.3.3 Investigative Simulation

The leave-one-out cross-validation c-statistic, $c_{L1O}(Y, \hat{\theta})$, is computed as follows. First, each case is sequentially deleted from the fitting sample of size N, and the model is fit to the resulting data set of size $N - 1$. With each model fit, the predicted probability is computed for the deleted case. Second, once all of the N models have been fit and the N case-deleted predicted probabilities have

been obtained, the $n_1 n_0$ event/non-event pairs are constructed, and the predicted probabilities are used in the computation of the statistic.

Our goal is to explore $c_{L1O}(Y, \hat{\theta})$ as an estimator of $E[c(Z, \hat{\theta})]$, which is involved in the construction of the MMPE, FSPE, and SUPER. Since the remaining part of the MMPE, $E[c(Z, \theta_0)]$, is constant, it need not be estimated for the purpose of constructing a model selection criterion. We will investigate $c_{L1O}(Y, \hat{\theta})$ as an estimator in terms of bias and model selection behaviors.

The natural estimator of $E[c(Z, \hat{\theta})]$ is $c(Z, \hat{\theta})$, which is exactly unbiased. However, the variability of $c(Z, \hat{\theta})$ is pronounced, rendering its accuracy to be poor. For simulation purposes, we will reduce the variability of $c(Z, \hat{\theta})$ through constructing an average based on replicated validation samples, say $\bar{c}_Z(\hat{\theta})$. Specifically, we define $\bar{c}_Z(\hat{\theta})$ as the average of 100 $c(Z, \hat{\theta})$ measures:

$$\bar{c}_Z(\hat{\theta}) = \sum_{i=1}^{100} \frac{1}{100} c(Z_i, \hat{\theta}),$$

where Z_i is the i^{th} randomly generated new outcome vector following the distribution of Y.

Using a simulated investigation, we will assess the difference in means and model selection counts of $\bar{c}_Z(\hat{\theta})$ and the leave-one-out cross-validation statistic, $c_{L1O}(Y, \hat{\theta})$, for both the traditional and adjusted c-statistics. In this simulation, we generate 1000 replications for sample sizes $N = 100, 500$. We consider 10 nested models with predictor sets

$$\{\{X_1\}, \{X_1, X_2\}, ..., \{X_1, X_2, ..., X_{10}\}\},$$

where

$$X_i \overset{iid}{\sim} \text{Uniform}(-1, 1) \text{ for } i = 1, 2, ..., 10.$$

The Bernoulli outcome

$$Y_i \overset{ind}{\sim} \text{Bernoulli}(\pi_i)$$

is based on the generating model

$$\text{logit}(\pi_i) = \beta_{01}x_{1i} + \beta_{02}x_{2i} + \beta_{03}x_{3i} + \beta_{04}x_{4i}.$$

Here, $\beta_{0i} = \sqrt{\frac{3}{p_0}e^{\lambda}} = \sqrt{\frac{3}{4}e^{\lambda}}$ for $i = 1, 2, 3, 4$. The parameters are set at this value so that we may fix the log-signal for the model; this quantity will be introduced later in the chapter.

For each of the 10 model orders, Table 2.1 contains the average of estimates for the difference $E[c(Z, \hat{\theta}) - c_{L1O}(Y, \hat{\theta})]$ based on $\bar{c}_Z(\hat{\theta}) - c_{L1O}(Y, \hat{\theta})$. This table shows the bias that is incurred by using $c_{L1O}(Y, \hat{\theta})$ as an estimator for $E[\bar{c}_Z(\hat{\theta})]$. We can see that the biases are small, particularly for the $N = 500$ case. For each sample size, the difference is smallest for the true probabilistic mechanism, while underspecified models tend to have the largest bias. For underspecified models, the differences calculated using the adjusted c-statistic tend to be larger than those using the traditional c-statistic; the converse is true for correctly and overspecified models. Due to the unavailability of the measure $c(Z, \hat{\theta})$ in most applications, $c_{L1O}(Y, \hat{\theta})$ appears to provide an appealing estimator for $E[\bar{c}_Z(\hat{\theta})]$. As validated through simulation, $c_{L1O}(Y, \hat{\theta})$ estimates $E[\bar{c}_Z(\hat{\theta})]$ with negligible bias.

Table 2.2 contains the model selection counts based on $\bar{c}_Z(\hat{\theta})$ and $c_{L1O}(y, \hat{\theta})$. A model will be identified as optimal for a replication if it yields the maximum c-statistic over all predictor sets. This table compares $c_{L1O}(y, \hat{\theta})$ as a model selection criterion to the "ideal" $\bar{c}_Z(\hat{\theta})$. Both measures exhibit protection from underfitting, but $c_{L1O}(y, \hat{\theta})$ tends to overfit more than $\bar{c}_Z(\hat{\theta})$. Both measures choose the true model as optimal a majority of the time, though $\bar{c}_Z(\hat{\theta})$ does this for about 20% more of the replications than $c_{L1O}(y, \hat{\theta})$. These tables indicate a strong similarity between the selection behaviors of the measures $\bar{c}_Z(\hat{\theta})$ and $c_{L1O}(y, \hat{\theta})$, allowing us to use $c_{L1O}(y, \hat{\theta})$ as a selection criterion, and as an estimator for the target $E[\bar{c}_Z(\hat{\theta})]$.

$n = 100$		
	Traditional	Adjusted
Order	$E[c(Z,\hat{\theta}) - c_{L1O}(Y,\hat{\theta})]$	$E[c(Z,\hat{\theta}) - c_{L1O}(Y,\hat{\theta})]$
1	0.03438	0.04700
2	0.02241	0.02542
3	0.01985	0.01680
4	0.01196	0.00707
5	0.01281	0.00765
6	0.01351	0.00826
7	0.01438	0.00894
8	0.01551	0.00973
9	0.01614	0.01011
10	0.01751	0.01111
$n = 500$		
	Traditional	Adjusted
Order	$E[c(Z,\hat{\theta}) - c_{L1O}(Y,\hat{\theta})]$	$E[c(Z,\hat{\theta}) - c_{L1O}(Y,\hat{\theta})]$
1	0.00761	0.00919
2	0.00534	0.00571
3	0.00433	0.00343
4	0.00236	0.00133
5	0.00239	0.00136
6	0.00237	0.00134
7	0.00242	0.00137
8	0.00251	0.00144
9	0.00254	0.00147
10	0.00256	0.00149

Table 2.1: Estimates of traditional and adjusted $E[c(Z,\hat{\theta}) - c_{L1O}(Y,\hat{\theta})]$ by order, 1000 replications.

$n = 100$				
Order	Traditional $\bar{c}_Z(\hat{\theta})$	Traditional $c_{L1O}(y, \hat{\theta})$	Adjusted $\bar{c}_Z(\hat{\theta})$	Adjusted $c_{L1O}(y, \hat{\theta})$
1	0	0	0	0
2	0	0	0	0
3	0	2	0	1
4	852	671	924	708
5	113	137	63	116
6	27	70	10	63
7	5	44	2	41
8	3	28	0	22
9	0	25	0	30
10	0	23	1	19
$n = 500$				
Order	Traditional $\bar{c}_Z(\hat{\theta})$	Traditional $c_{L1O}(y, \hat{\theta})$	Adjusted $\bar{c}_Z(\hat{\theta})$	Adjusted $c_{L1O}(y, \hat{\theta})$
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	915	678	942	730
5	73	131	47	116
6	11	83	9	68
7	1	43	1	41
8	0	28	1	19
9	0	20	0	16
10	0	17	0	13

Table 2.2: Model selection counts of traditional and adjusted $\bar{c}_Z(\hat{\theta})$ and $c_{L1O}(y, \hat{\theta})$ by order, 1000 replications.

2.4 Simulation Study

We compile and report a comprehensive two-part simulation study in order to assess and compare the performances of our newly proposed model selection criteria. By generating numerous replicated samples, and using these samples for model fitting and selection, we can characterize the general behaviors of the criteria. This simulation study considers two model selection settings and focuses on the criteria based on the MMPE, FSPE, and SUPER, as well as AIC. The criteria that target the predictive error measures will be calculated based on both the traditional and adjusted c-statistics, while AIC is included as an accepted standard criterion that is founded on predictive principles.

In a logistic regression framework, two principal factors govern the ability of a selection criterion to identify a model of appropriate structure: the sample size and the model signal. The signal is dictated by the variability of the predictors and the sizes of the effects, as reflected by the magnitudes of the regression coefficients. Formally, the signal of the model is defined by $e^{\lambda} = \text{Var}[X'\beta_0]$, for a random vector of predictors X and a generating parameter vector β_0.

The two settings that we consider are nested modeling and all subsets modeling. For each simulation set, 1000 samples are generated for each sample size and log-signal combination. The simulation study is designed as a factorial experiment, where the factors are the selection criterion (MMPE, FSPE, SUPER, or AIC), the type of c-statistic (traditional or adjusted), the sample size ($N = 100, 500$), and the log-signal ($\lambda = 0, 1, 2$).

The elements of the regression parameter vector for the generating model, $\beta_0 = (\beta_{01}, \beta_{02}, ..., \beta_{0p_0})'$, will be identical for all predictor variables (intercept: $\alpha_0 = 0$), and calculated based on the log-signal. Denoting the log-signal as λ and the

true number of parameters as p_0, the calculation for β_0 is as follows:

$$\lambda = \ln(\text{Var}[X'\beta_0])$$

$$e^{\lambda} = \text{Var}[X'\beta_0]$$

Since the predictors are $\overset{iid}{\sim}$ Uniform$(-1, 1)$, the preceding simplifies to

$$\begin{aligned} e^{\lambda} &= \text{Var}[\sum_{i=1}^{p_0} \beta_{0i} x_i] \\ &= \sum_{i=1}^{p_0} {\beta_{0i}}^2 \text{Var}[x_i] \\ &= p_0 {\beta_{0i}}^2 \text{Var}[x_i] \\ &= p_0 {\beta_{0i}}^2 \frac{(1-(-1))^2}{12} \\ &= \frac{p_0 {\beta_{0i}}^2}{3}. \end{aligned}$$

Solving for β_{0i}, we obtain

$$\beta_{0i} = \sqrt{\frac{3}{p_0} e^{\lambda}} \text{ for } i = 1, 2, ..., p_0.$$

The generating model is

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = x_i'\beta_0,$$

where $\pi_i = P(Y_i = 1)$. The actual values of Y_i are generated randomly using the **rbinom** function in **R**, where π_i is the input probability.

For each model, we calculate the traditional and adjusted c-statistics using the conventional and leave-one-out cross-validation approaches, as well as the AIC. Once we have these measures, we can calculate the estimates for the MMPE, FSPE, and SUPER.

For every candidate model order, we compute the means of each criterion by sample size and log-signal. We plot these means by model order to provide a visual representation of the behaviors of the criteria for underspecified, correctly specified, overspecified, and mixed misspecified models. (Mixed misspecified models contain

both legitimate and spurious predictors, yet do not contain all of the predictors represented in the true model.) Additionally, for each replicated sample, we record the optimal model selected by every criterion. We summarize the model selections in a table of counts. Such tables allow us to assess the ability of each criterion to pick the correct model, as well as to see which incorrectly specified models each criterion tends to favor. The figures featuring the criterion means are presented adjacent to the corresponding table of selection counts.

2.4.1 Nested Setting

The first part of the simulation study involves a nested modeling setting. Here, we generate data sets with 10 predictor variables

$$X_1, X_2, ..., X_{10} \overset{iid}{\sim} \text{Uniform}(-1, 1).$$

The Bernoulli outcome

$$Y_i \overset{ind}{\sim} \text{Bernoulli}(\pi_i)$$

is based on the generating model

$$\text{logit}(\pi_i) = \beta_{01}x_{1i} + \beta_{02}x_{2i} + \beta_{03}x_{3i} + \beta_{04}x_{4i},$$

where $\beta_{0i} = \sqrt{\frac{3}{p_0}e^{\lambda}} = \sqrt{\frac{3}{4}e^{\lambda}}$ for $i = 1, 2, 3, 4$.

Once we have the full data sets, we fit 10 nested models of orders 1 through 10. The models are as follows:

Order 1: $\text{logit}(\pi_i) = x_{1i}\beta_1$

Order 2: $\text{logit}(\pi_i) = x_{1i}\beta_1 + x_{2i}\beta_2$

$\vdots$

Order 10: $\text{logit}(\pi_i) = x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{10i}\beta_{10}$.

The nesting setting allows us to compare the criteria for models that are underspecified (candidate predictor set is a proper subset of the set of "true" predictors; i.e., the predictors for the generating model), correctly specified (candidate

predictor set is exactly the same as the set of true predictors), and overspecified (true predictor set is a proper subset of the candidate predictor set). Again, the simulation study is designed as a factorial experiment, where the factors are the selection criterion, the type of c-statistic, the sample size, and the log-signal. Table 2.3 lists the ID for each simulation set, along with the associated levels for three of the factors. The figures corresponding to sets N1 - N14 contain three curves; one for each log-signal. The criterion means over the 1000 replications are calculated for each model order and plotted, illustrating the general behavior of the measures. The tables corresponding to sets N1 - N14 feature model order selection counts by each log-signal. A model order is selected as optimal if it corresponds to the smallest value of the criterion over all of the fitted models in the candidate collection.

Set ID	Criterion	Type	Sample Size
N1	MMPE	Traditional	100
N2	MMPE	Traditional	500
N3	MMPE	Adjusted	100
N4	MMPE	Adjusted	500
N5	FSPE	Traditional	100
N6	FSPE	Traditional	500
N7	FSPE	Adjusted	100
N8	FSPE	Adjusted	500
N9	SUPER	Traditional	100
N10	SUPER	Traditional	500
N11	SUPER	Adjusted	100
N12	SUPER	Adjusted	500
N13	AIC		100
N14	AIC		500

Table 2.3: Factor levels for the 14 simulation sets presented for the nested simulation setting.

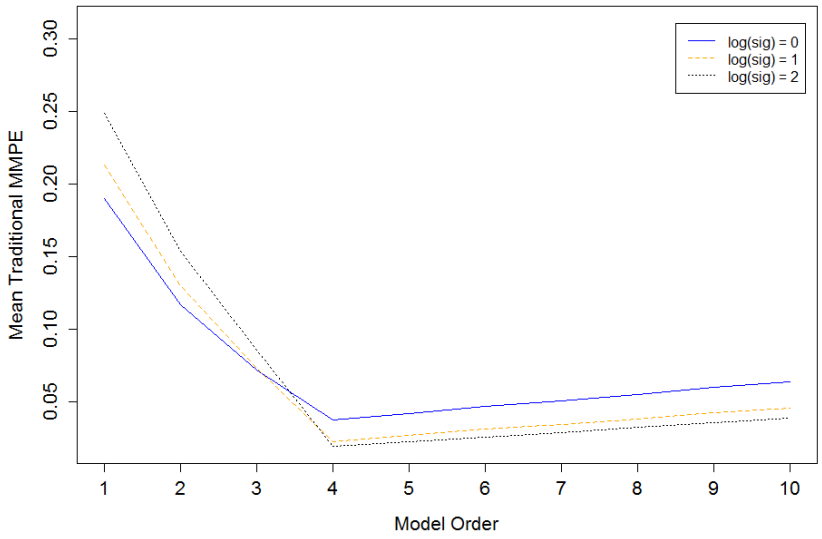

Figure 2.1: Set N1 - Means of traditional MMPE by model order, $N = 100$.

	log(signal)		
Model Order	0	1	2
1	5	0	0
2	21	2	0
3	120	22	2
4	459	624	671
5	129	129	137
6	78	76	70
7	65	53	44
8	46	39	28
9	40	25	25
10	37	30	23

Table 2.4: Set N1 - Counts of traditional MMPE by model order, $N = 100$.

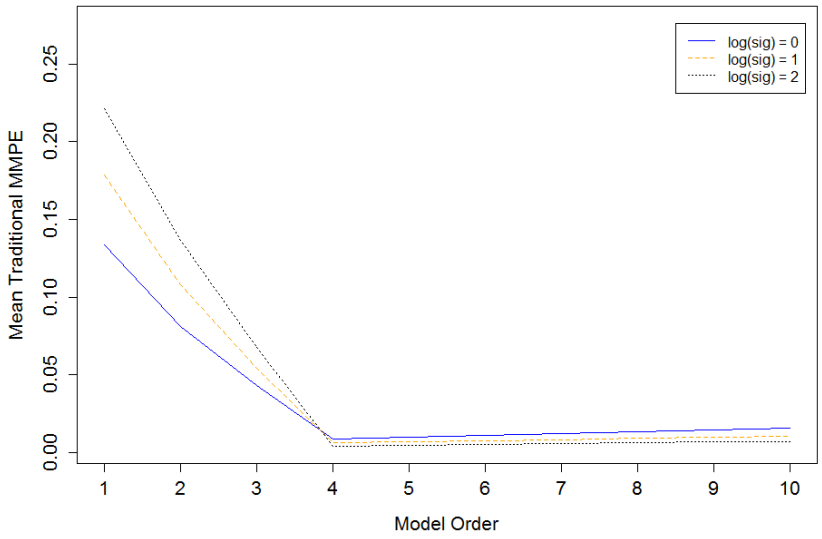

Figure 2.2: Set N2 - Means of traditional MMPE by model order, $N = 500$.

	log(signal)		
Model Order	0	1	2
1	0	0	0
2	0	0	0
3	0	0	0
4	663	661	678
5	121	144	131
6	67	75	83
7	52	46	43
8	48	35	28
9	30	18	20
10	19	21	17

Table 2.5: Set N2 - Counts of traditional MMPE by model order, $N = 500$.

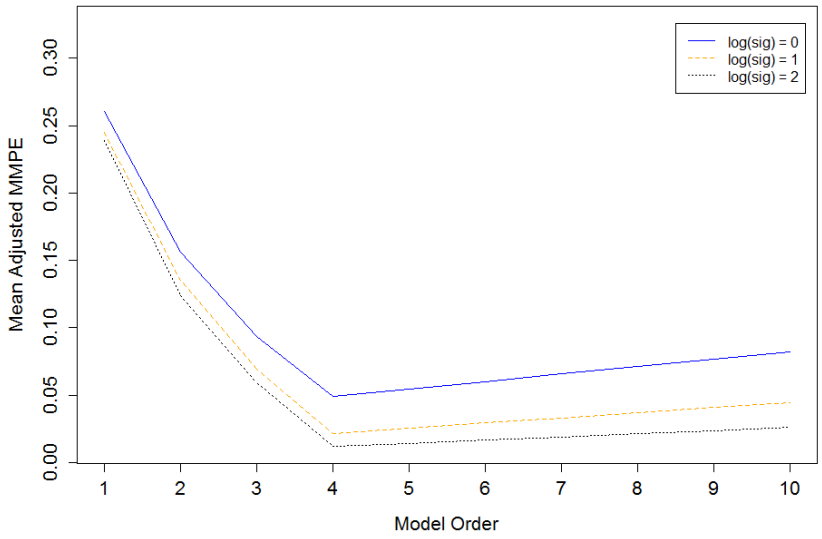

Figure 2.3: Set N3 - Means of adjusted MMPE by model order, $N = 100$.

	log(signal)		
Model Order	0	1	2
1	5	0	0
2	22	0	0
3	105	19	1
4	482	669	708
5	128	103	116
6	76	64	63
7	63	51	41
8	50	35	22
9	33	26	30
10	36	33	19

Table 2.6: Set N3 - Counts of adjusted MMPE by model order, $N = 100$.

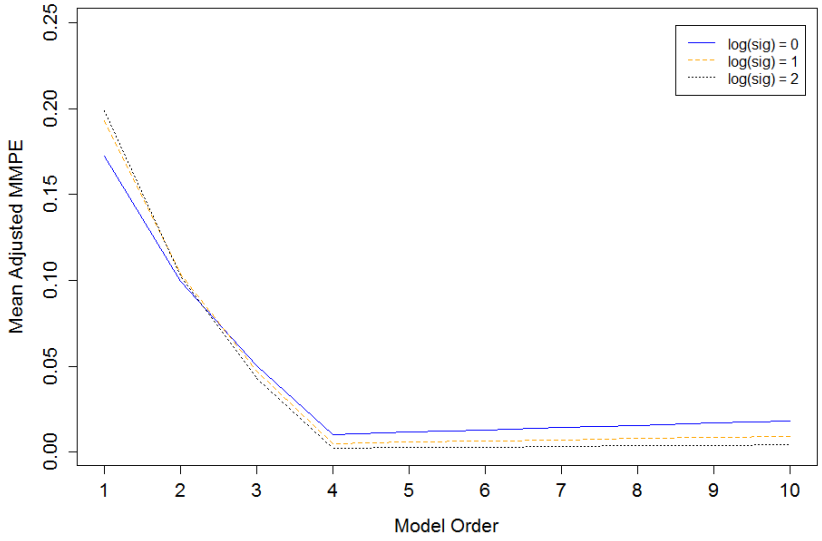

Figure 2.4: Set N4 - Means of adjusted MMPE by model order, $N = 500$.

	log(signal)		
Model Order	0	1	2
1	0	0	0
2	0	0	0
3	0	0	0
4	700	713	730
5	115	135	113
6	64	61	68
7	40	35	41
8	35	27	19
9	26	16	16
10	20	13	13

Table 2.7: Set N4 - Counts of adjusted MMPE by model order, $N = 500$.

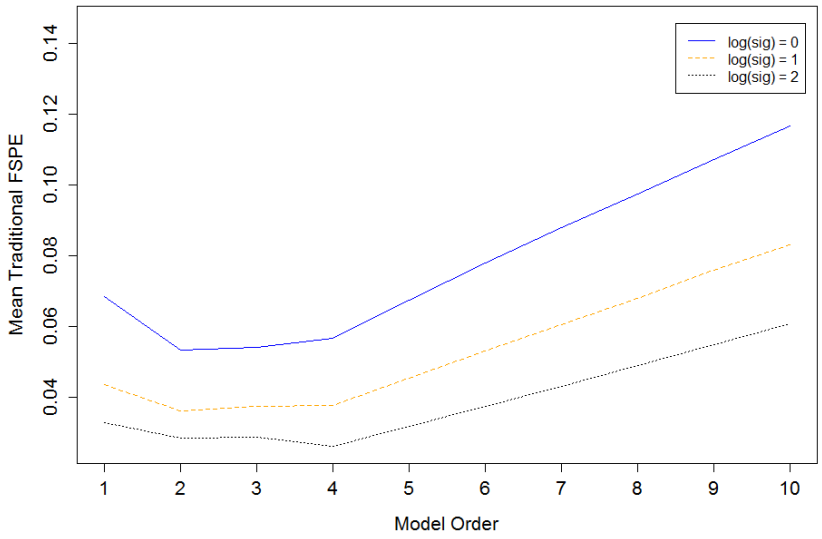

Figure 2.5: Set N5 - Means of traditional FSPE by model order, $N = 100$.

	log(signal)		
Model Order	0	1	2
1	404	317	208
2	286	296	236
3	198	171	159
4	107	210	379
5	4	6	16
6	1	0	2
7	0	0	0
8	0	0	0
9	0	0	0
10	0	0	0

Table 2.8: Set N5 - Counts of traditional FSPE by model order, $N = 100$.

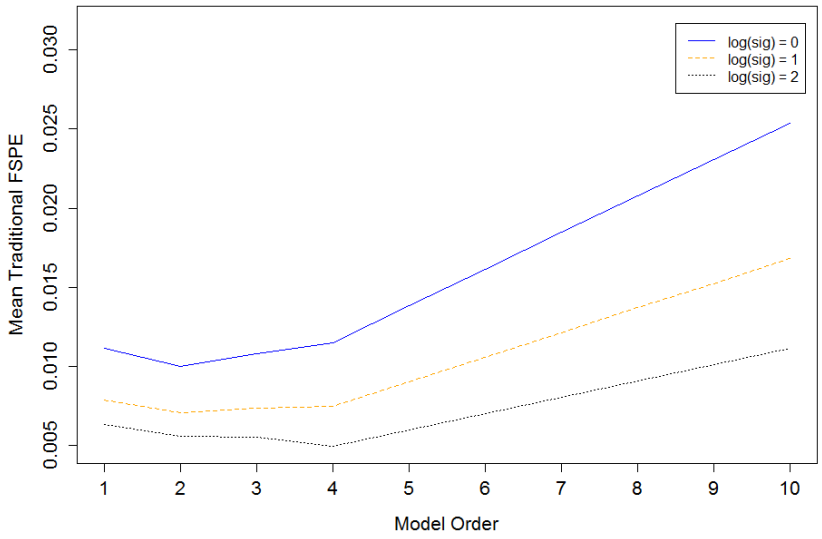

Figure 2.6: Set N6 - Means of traditional FSPE by model order, $N = 500$.

	log(signal)		
Model Order	0	1	2
1	306	182	32
2	485	466	131
3	143	172	104
4	66	180	731
5	0	0	2
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	0	0	0

Table 2.9: Set N6 - Counts of traditional FSPE by model order, $N = 500$.

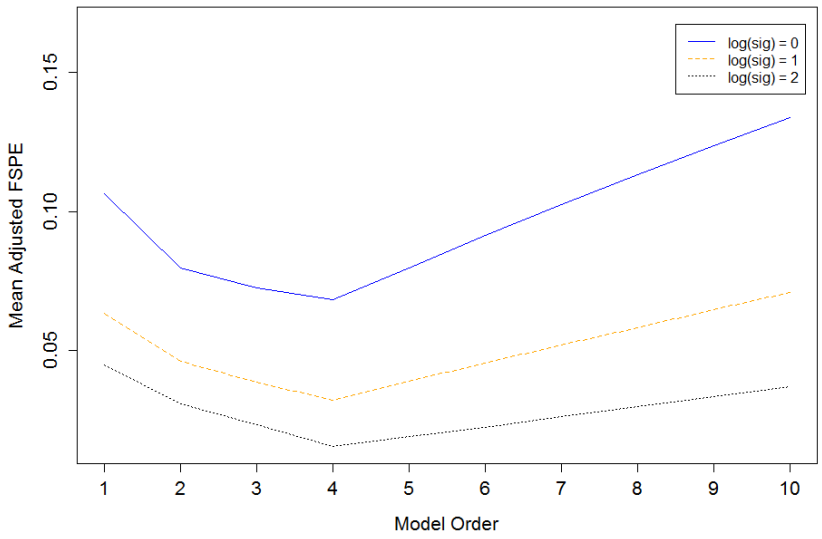

Figure 2.7: Set N7 - Means of adjusted FSPE by model order, $N = 100$.

	log(signal)		
Model Order	0	1	2
1	249	94	13
2	198	117	21
3	214	146	60
4	324	624	875
5	8	17	24
6	6	2	5
7	1	0	0
8	0	0	0
9	0	0	0
10	0	0	2

Table 2.10: Set N7 - Counts of adjusted FSPE by model order, $N = 100$.

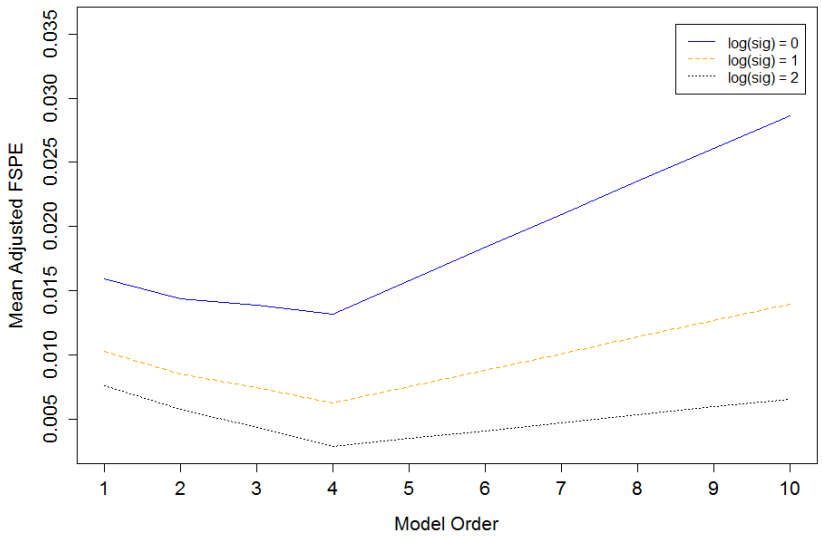

Figure 2.8: Set N8 - Means of adjusted FSPE by model order, $N = 500$.

	log(signal)		
Model Order	0	1	2
1	216	3	0
2	180	10	0
3	184	31	0
4	420	956	1000
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	0	0	0

Table 2.11: Set N8 - Counts of adjusted FSPE by model order, $N = 500$.

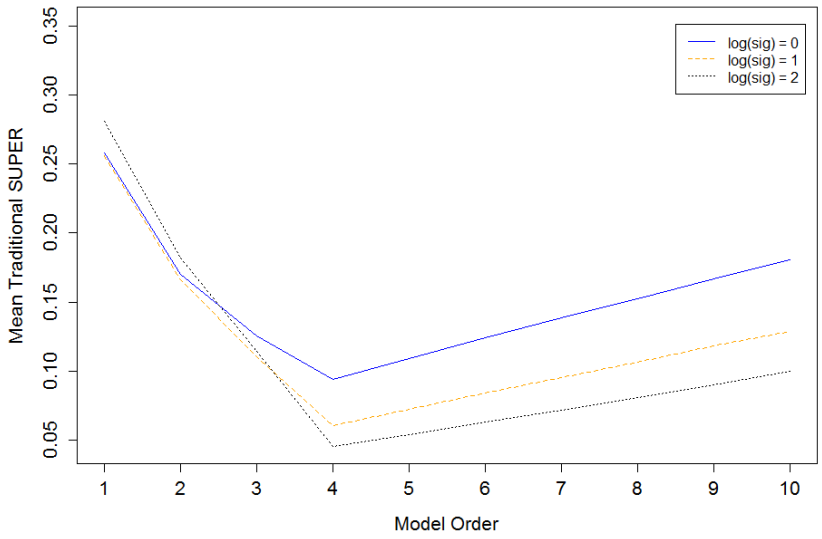

Figure 2.9: Set N9 - Means of traditional SUPER by model order, $N = 100$.

	log(signal)		
Model Order	0	1	2
1	25	1	0
2	67	13	0
3	194	58	7
4	539	759	831
5	94	95	105
6	29	36	31
7	23	21	13
8	15	5	6
9	9	5	4
10	5	7	3

Table 2.12: Set N9 - Counts of traditional SUPER by model order, $N = 100$.

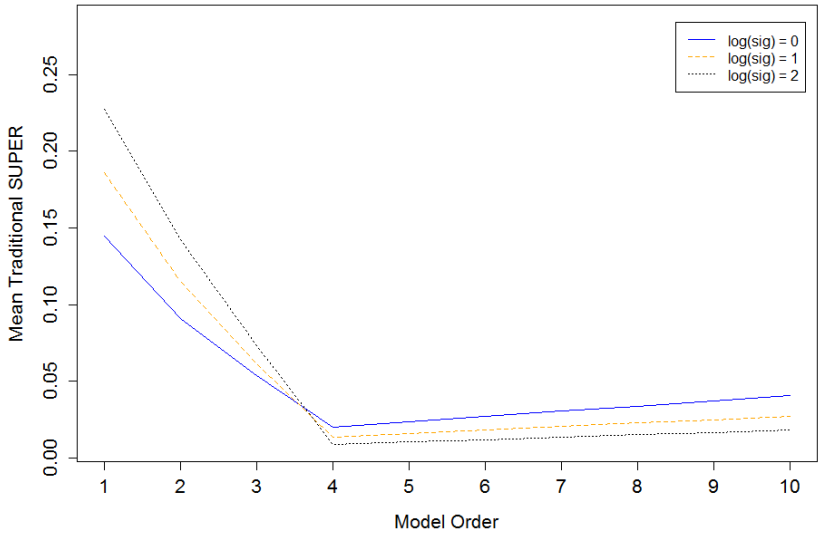

Figure 2.10: Set N10 - Means of traditional SUPER by model order, $N = 500$.

	log(signal)		
Model Order	0	1	2
1	0	0	0
2	0	0	0
3	1	0	0
4	886	907	916
5	73	65	54
6	25	22	19
7	8	4	7
8	4	1	4
9	2	1	0
10	1	0	0

Table 2.13: Set N10 - Counts of traditional SUPER by model order, $N = 500$.

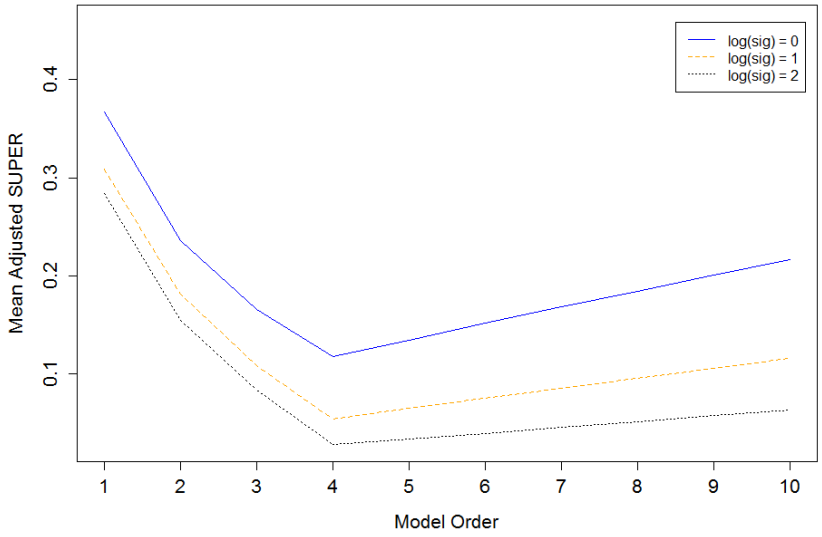

Figure 2.11: Set N11 - Means of adjusted SUPER by model order, $N = 100$.

	log(signal)		
Model Order	0	1	2
1	24	1	0
2	57	7	0
3	164	37	0
4	563	811	912
5	86	72	57
6	46	28	18
7	27	22	7
8	13	8	3
9	13	7	1
10	7	7	2

Table 2.14: Set N11 - Counts of adjusted SUPER by model order, $N = 100$.

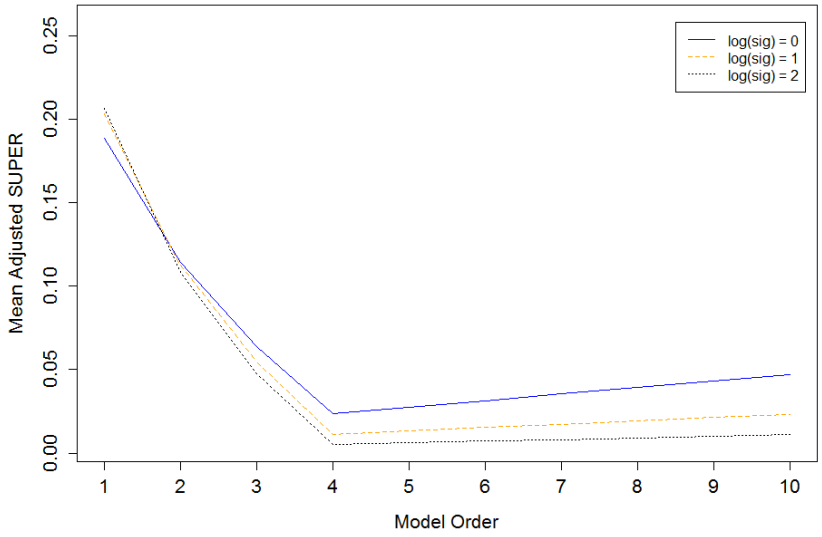

Figure 2.12: Set N12 - Means of adjusted SUPER by model order, $N = 500$.

	log(signal)		
Model Order	0	1	2
1	0	0	0
2	0	0	0
3	1	0	0
4	916	921	935
5	54	58	43
6	21	16	13
7	3	2	7
8	3	1	2
9	1	1	0
10	1	1	0

Table 2.15: Set N12 - Counts of adjusted SUPER by model order, $N = 500$.

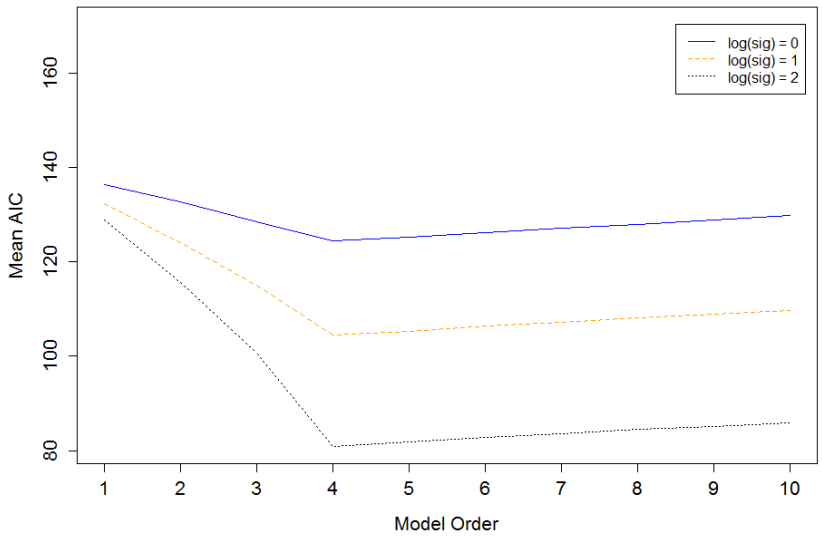

Figure 2.13: Set N13 - Means of AIC by model order, $N = 100$.

	log(signal)		
Model Order	0	1	2
1	21	0	0
2	40	1	0
3	110	15	0
4	534	687	663
5	109	108	110
6	69	62	73
7	46	44	44
8	30	33	34
9	23	19	34
10	18	31	42

Table 2.16: Set N13 - Counts of AIC by model order, $N = 100$.

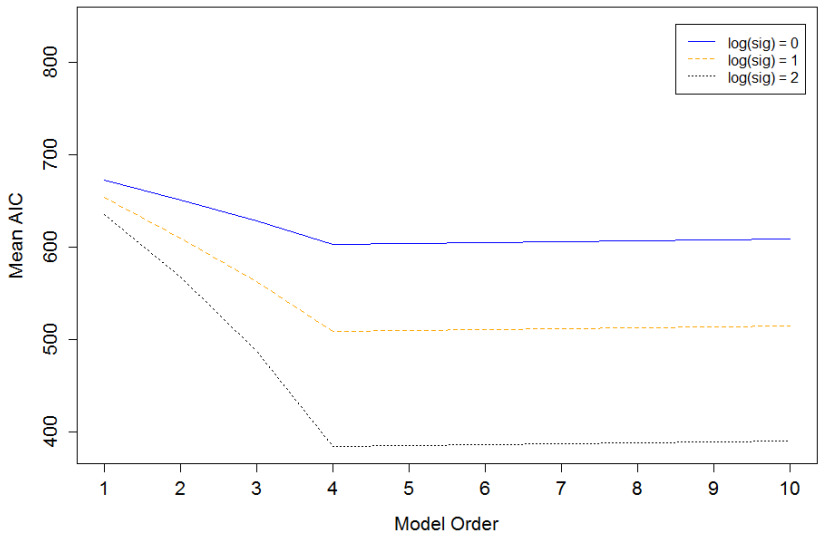

Figure 2.14: Set N14 - Means of AIC by model order, $N = 500$.

	log(signal)		
Model Order	0	1	2
1	0	0	0
2	0	0	0
3	0	0	0
4	725	703	717
5	106	132	113
6	62	69	70
7	36	32	43
8	32	30	23
9	23	17	21
10	16	17	13

Table 2.17: Set N14 - Counts of AIC by model order, $N = 500$.

When the log-signal is held constant and the sample size is increased, all of the model selection criteria show improved correct model selection counts. The same is true when the sample size is held constant and the log-signal is increased. The MMPE, FSPE, and SUPER all tend to perform better when calculated using the adjusted c-statistic rather than its traditional counterpart. This improved performance is most pronounced for the FSPE.

The means and selection counts for the MMPE reflect strong protection from underfitting. The mean criterion values for underspecified models are large compared to those for correctly specified or overspecified models. For each of the sets N1 - N4, the minimum mean values occur at the true order of four. The model counts exhibit frequent selection of the true probabilistic mechanism and minimal selections of models smaller in order. Sets N2 and N4 both show no underfitting for any of the log-signal values.

The means and selection counts for the FSPE reflect strong protection from overfitting. The mean criterion values for overspecified models are large compared to those for correctly specified or underspecified models. For sets N7 and N8, the minimum mean values occur at the true order of four, whereas for set N5 and N6, the minimums fall below the true order. The model counts reflect varied selection of the true probabilistic mechanism, with higher frequencies for larger sample sizes, higher log-signals, and the adjusted c-statistic. Sets N6 and N8 show virtually no overfitting.

The means and selection counts for the SUPER indicate the strong protection from underfitting exhibited by the MMPE as well as modest protection from overfitting as evident with the FSPE. The mean criterion values for underspecified models are large compared to those for correctly specified or overspecified models. For each of the sets N9 - N12, the minimum mean values occur at the true order. The model counts exhibit frequent selection of the true probabilistic mechanism and

minimal selections of models smaller in order. Sets N10 and N12 show virtually no underfitting.

Finally, the means and selection counts for AIC reflect strong protection from underfitting. The mean criterion values for underspecified models are large compared to those for correctly specified or overspecified models. For sets N13 and N14, the minimum means values correspond to the true model order four. The model counts exhibit frequent selection of the true probabilistic mechanism and minimal selections of models smaller in order. Set N14 shows no underfitting. The SUPER behaves very similarly to AIC, but does considerably better in selecting the correctly specified model.

2.4.2 All Subsets Setting

The second part of the simulation study involves an all subsets setting. Here, we generate data sets with 6 predictor variables

$$X_1, X_2, ..., X_6 \overset{iid}{\sim} \text{Uniform}(-1, 1).$$

The Bernoulli outcome

$$Y_i \overset{ind}{\sim} \text{Bernoulli}(\pi_i)$$

is based on the generating model

$$\text{logit}(\pi_i) = \beta_{01} x_{1i} + \beta_{02} x_{2i} + \beta_{03} x_{3i},$$

where $\beta_{0i} = \sqrt{\frac{3}{p_0} e^{\lambda}} = \sqrt{e^{\lambda}}$ for $i = 1, 2, 3$.

With the generated data sets, we fit candidate models based on all possible subsets of the predictor variables. Therefore, the candidate collection consists of $2^6 - 1 = 63$ models, since the null model is not included.

The all subsets setting allows us to compare the criteria for models that are underspecified (candidate predictor set is a proper subset of the set of true predictors), correctly specified (candidate predictor set is exactly the same as the set of

true predictors), overspecified (true predictor set is a proper subset of the candidate predictor set), and mixed misspecified (candidate predictor set is comprised of some but not all of the predictors in the true set, as well as some predictors not in the true set). Again, the simulation study is designed as a factorial experiment, where the factors are the selection criterion, the type of c-statistic, the sample size, and the log-signal. Table 2.18 lists the ID for each simulation set, along with the associated levels for three of the factors. The figures corresponding to sets AS1 - AS14 contain three colored sets of means; one for each log-signal. The criterion means over the 1000 replications are calculated for each model specification and plotted, illustrating the general behavior of the measures. The minimum mean criterion for each log-signal is indicated with a horizontal line. The tables corresponding to sets AS1 - AS14 feature model specification selection counts by each log-signal. A model specification is selected as optimal if it corresponds to the smallest value of the criterion over all of the fitted models in the candidate collection.

Set ID	Criterion	Type	Sample Size
AS1	MMPE	Traditional	100
AS2	MMPE	Traditional	500
AS3	MMPE	Adjusted	100
AS4	MMPE	Adjusted	500
AS5	FSPE	Traditional	100
AS6	FSPE	Traditional	500
AS7	FSPE	Adjusted	100
AS8	FSPE	Adjusted	500
AS9	SUPER	Traditional	100
AS10	SUPER	Traditional	500
AS11	SUPER	Adjusted	100
AS12	SUPER	Adjusted	500
AS13	AIC		100
AS14	AIC		500

Table 2.18: Factor levels for the 14 simulation sets presented for the all subsets simulation setting

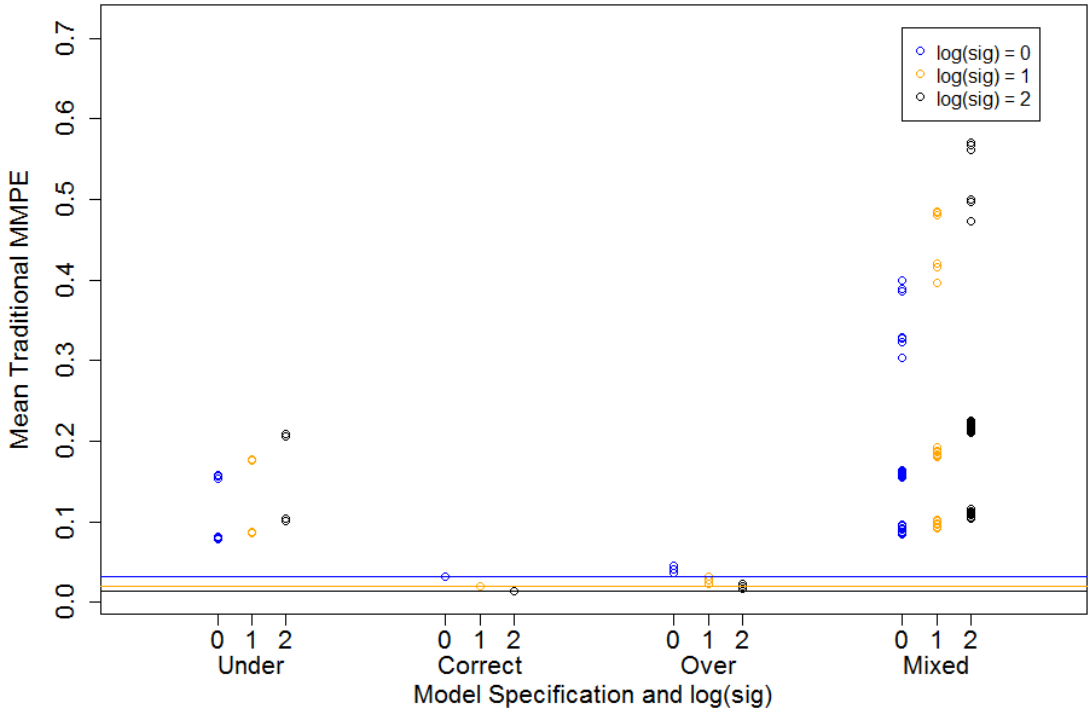

Figure 2.15: Set AS1 - Means of traditional MMPE by model specification, $N = 100$.

	log(signal)		
Model Specification	0	1	2
Under	129	11	0
Correct	343	528	536
Over	363	449	464
Mixed	165	12	0

Table 2.19: Set AS1 - Counts of traditional MMPE by model specification, $N = 100$.

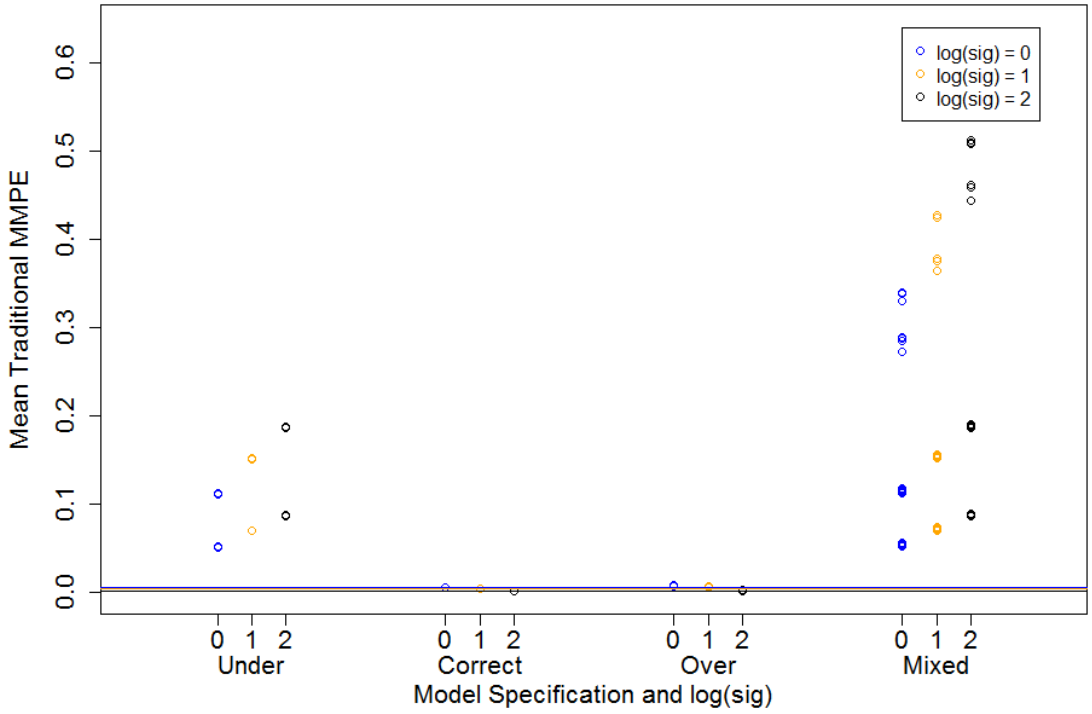

Figure 2.16: Set AS2 - Means of traditional MMPE by model specification, $N = 500$.

	log(signal)		
Model Specification	0	1	2
Under	0	0	0
Correct	548	546	600
Over	452	454	400
Mixed	0	0	0

Table 2.20: Set AS2 - Counts of traditional MMPE by model specification, $N = 500$.

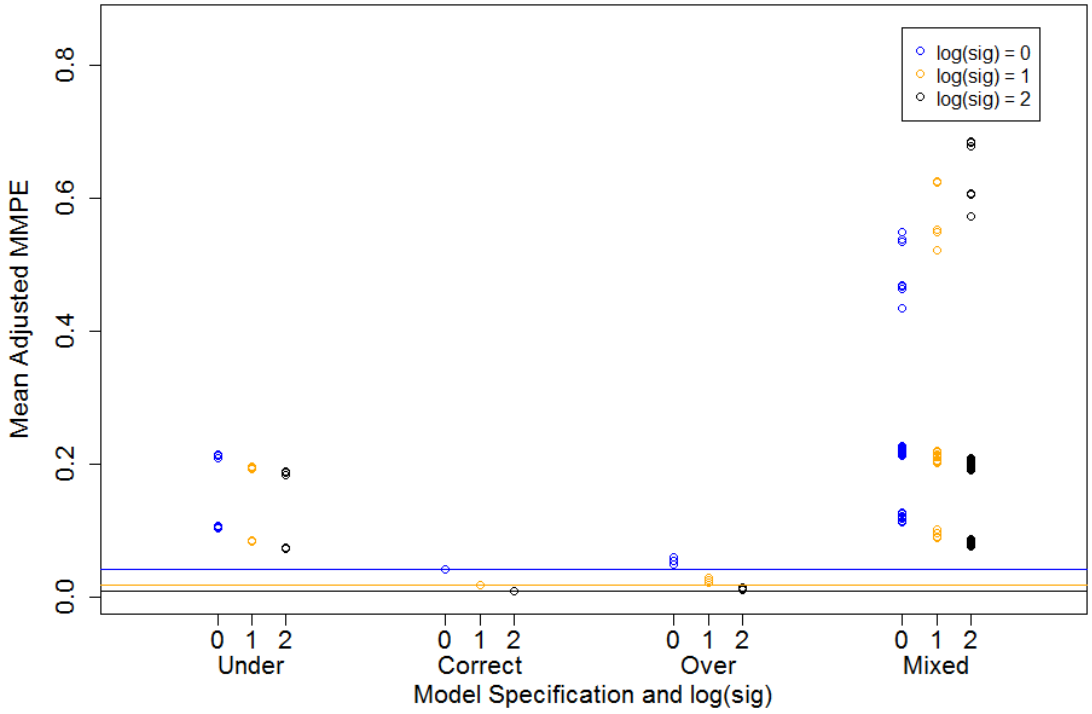

Figure 2.17: Set AS3 - Means of adjusted MMPE by model specification, $N = 100$.

	log(signal)		
Model Specification	0	1	2
Under	121	8	0
Correct	376	580	602
Over	353	409	398
Mixed	150	0	0

Table 2.21: Set AS3 - Counts of adjusted MMPE by model specification, $N = 100$.

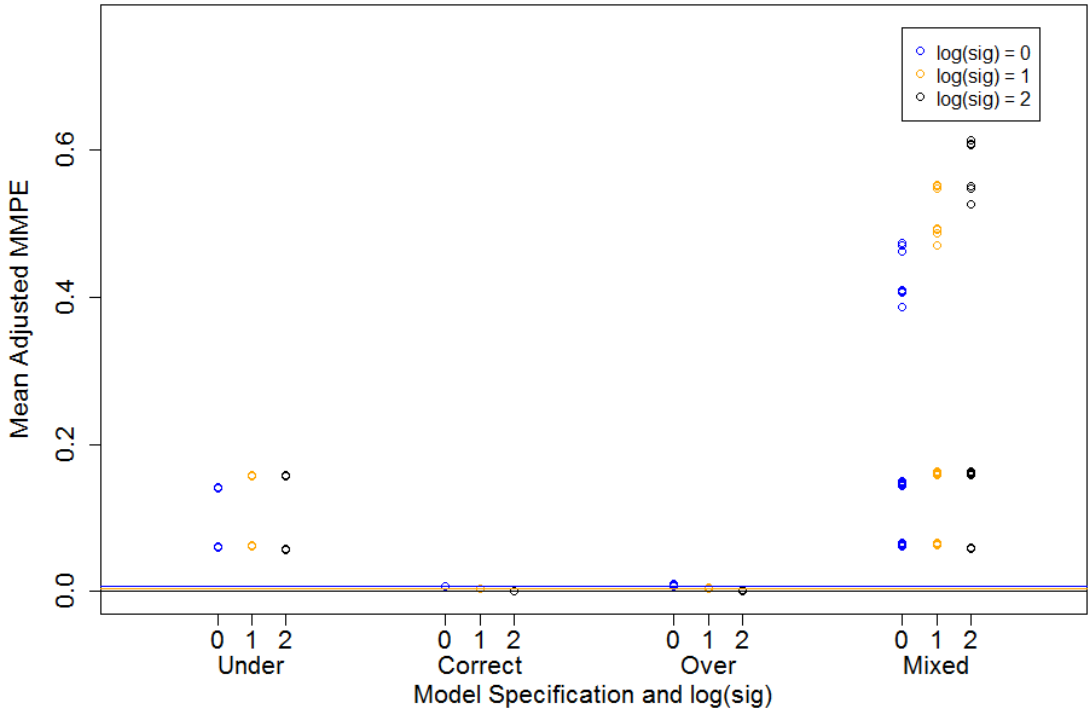

Figure 2.18: Set AS4 - Means of adjusted MMPE by model specification, $N = 500$.

	log(signal)		
Model Specification	0	1	2
Under	0	0	0
Correct	578	598	639
Over	422	402	361
Mixed	0	0	0

Table 2.22: Set AS4 - Counts of adjusted MMPE by model specification, $N = 500$.

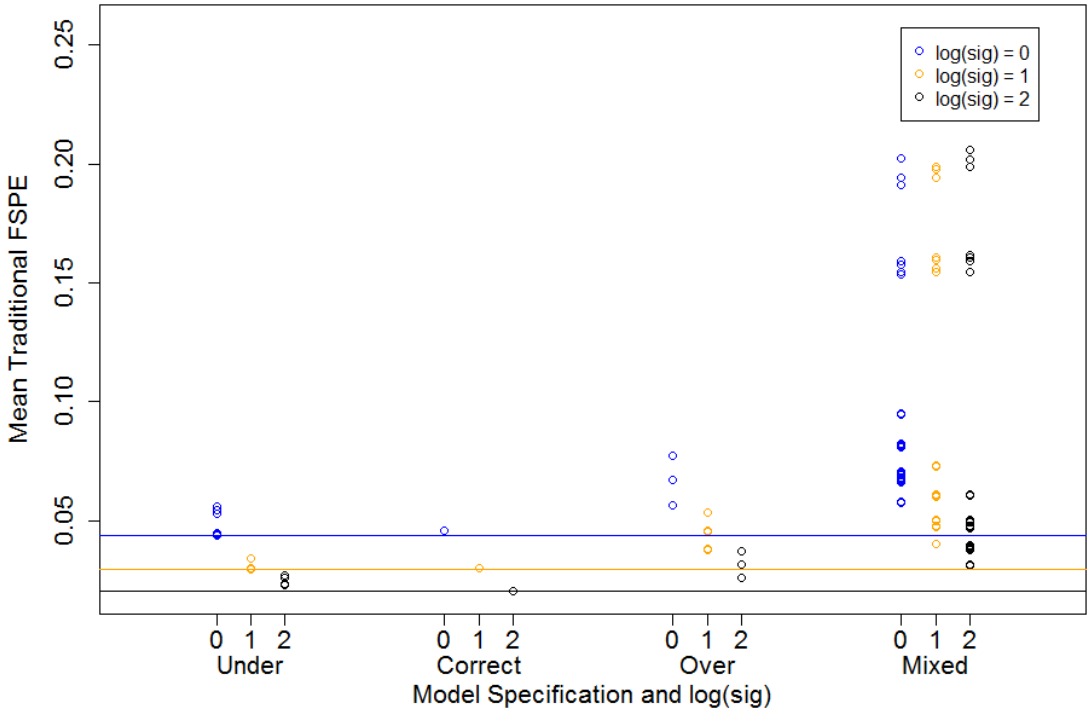

Figure 2.19: Set AS5 - Means of traditional FSPE by model specification, $N = 100$.

	log(signal)		
Model Specification	0	1	2
Under	937	907	688
Correct	12	71	420
Over	1	2	15
Mixed	50	20	0

Table 2.23: Set AS5 - Counts of traditional FSPE by model specification, $N = 100$.

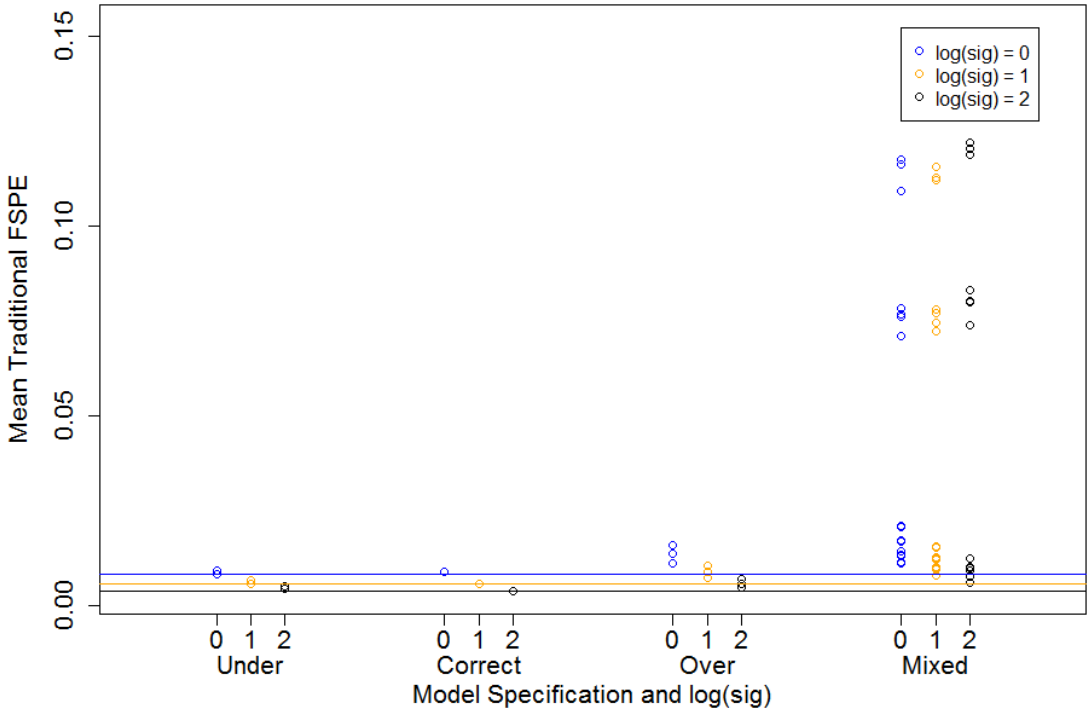

Figure 2.20: Set AS6 - Means of traditional FSPE by model specification, $N = 500$.

	log(signal)		
Model Specification	0	1	2
Under	992	894	274
Correct	8	106	724
Over	0	0	2
Mixed	0	0	0

Table 2.24: Set AS6 - Counts of traditional FSPE by model specification, $N = 500$.

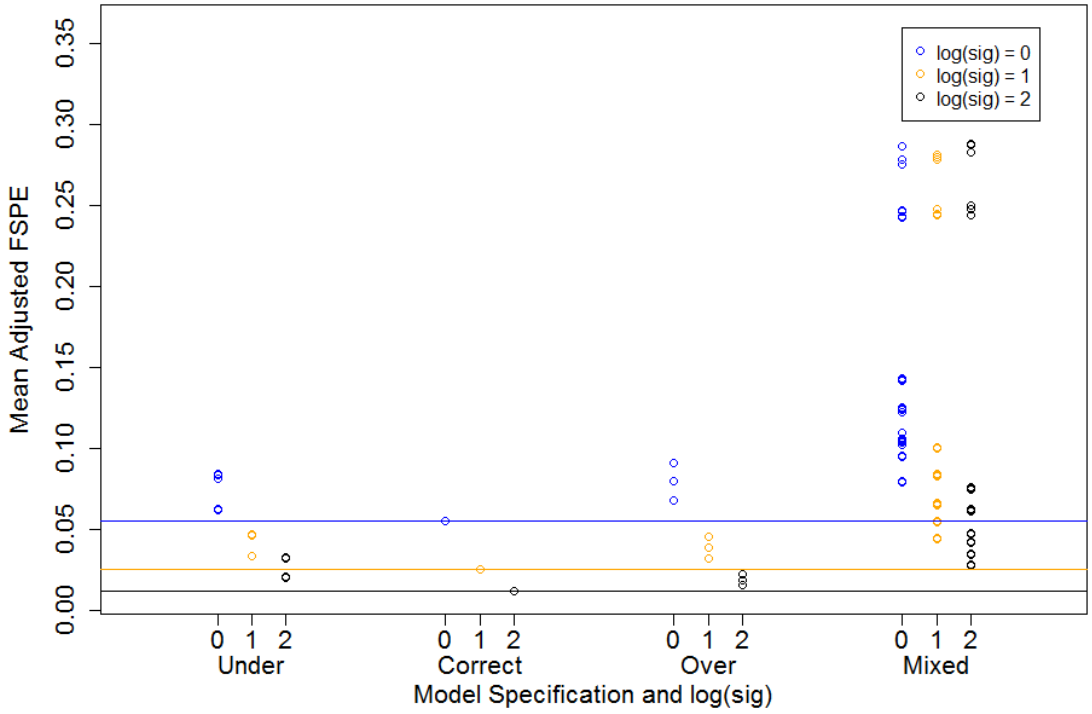

Figure 2.21: Set AS7 - Means of adjusted FSPE by model specification, $N = 100$.

	log(signal)		
Model Specification	0	1	2
Under	880	510	106
Correct	77	469	850
Over	4	1	41
Mixed	39	0	3

Table 2.25: Set AS7 - Counts of adjusted FSPE by model specification, $N = 100$.

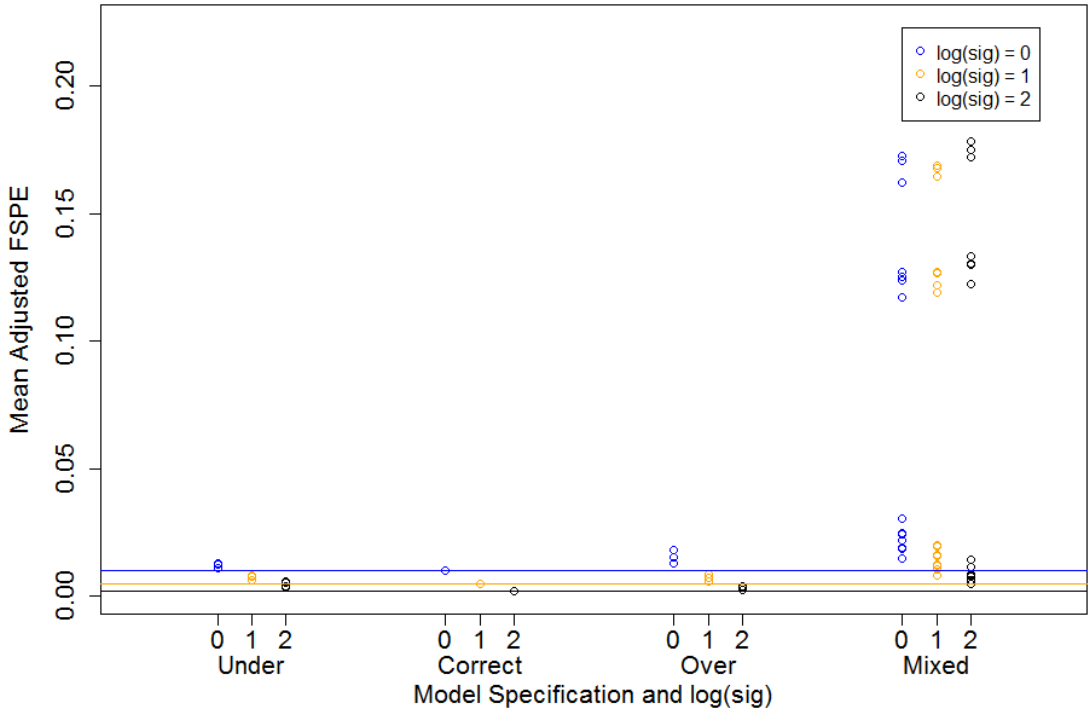

Figure 2.22: Set AS8 - Means of adjusted FSPE by model specification, $N = 500$.

	log(signal)		
Model Specification	0	1	2
Under	725	59	0
Correct	275	941	1000
Over	0	0	0
Mixed	0	0	0

Table 2.26: Set AS8 - Counts of adjusted FSPE by model specification, $N = 500$.

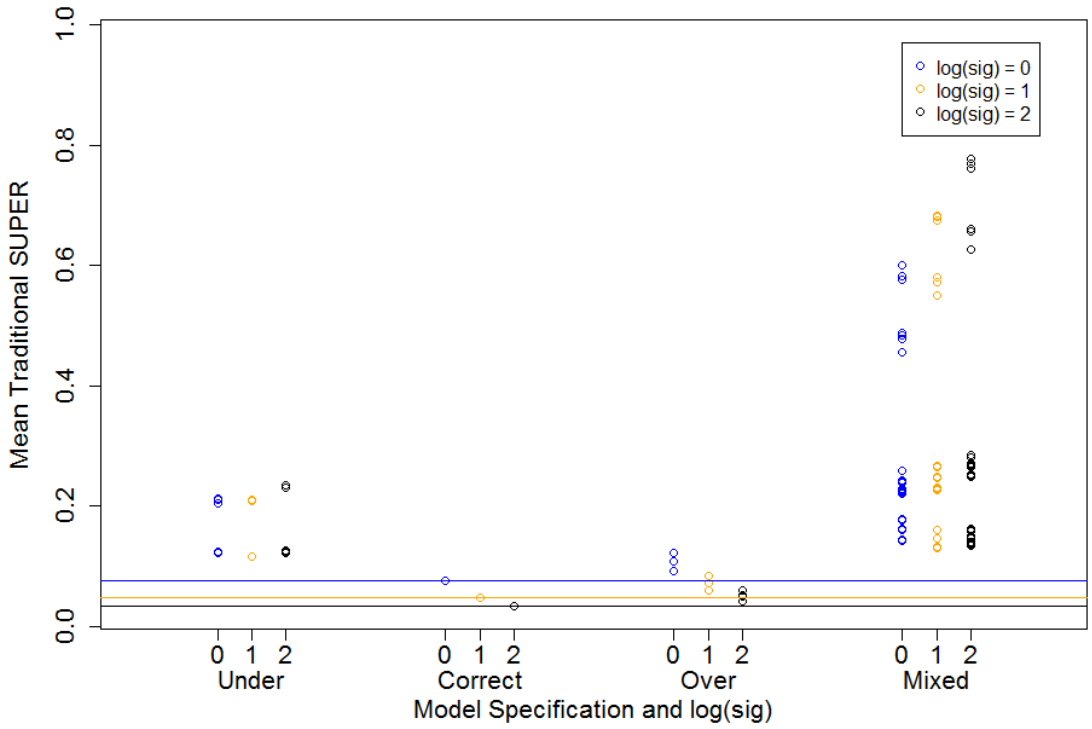

Figure 2.23: Set AS9 - Means of traditional SUPER by model specification, $N = 100$.

	log(signal)		
Model Specification	0	1	2
Under	317	54	3
Correct	365	659	697
Over	157	270	299
Mixed	161	17	1

Table 2.27: Set AS9 - Counts of traditional SUPER by model specification, $N = 100$.

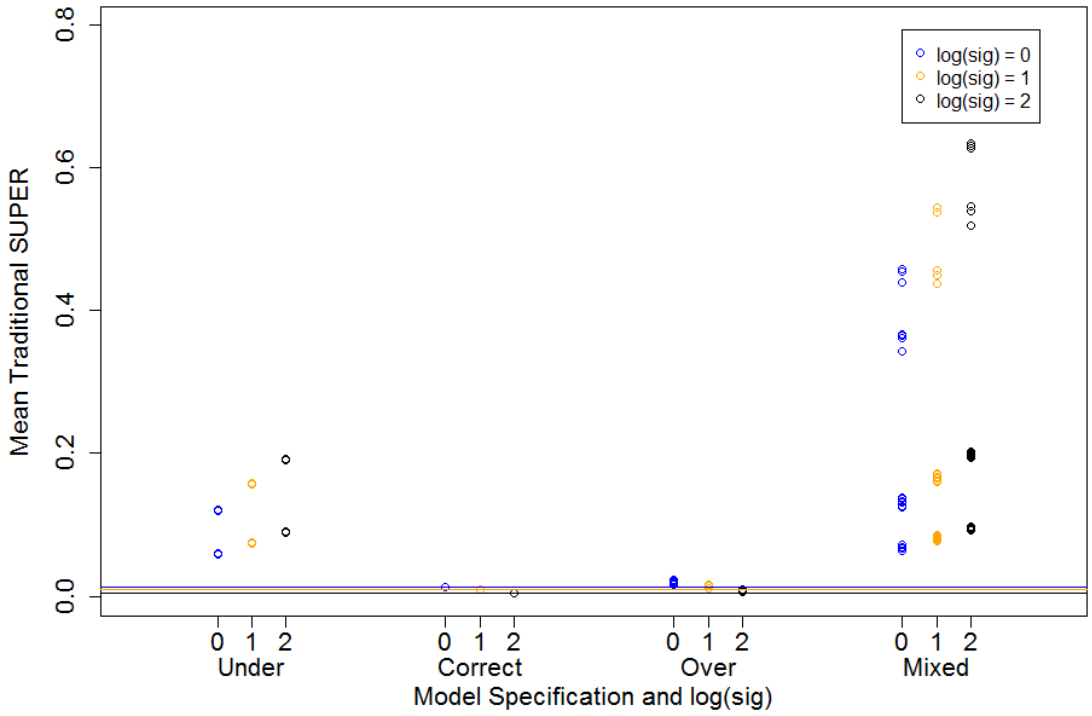

Figure 2.24: Set AS10 - Means of traditional SUPER by model specification, $N = 500$.

	log(signal)		
Model Specification	0	1	2
Under	0	0	0
Correct	799	803	840
Over	201	197	160
Mixed	0	0	0

Table 2.28: Set AS10 - Counts of traditional SUPER by model specification, $N = 500$.

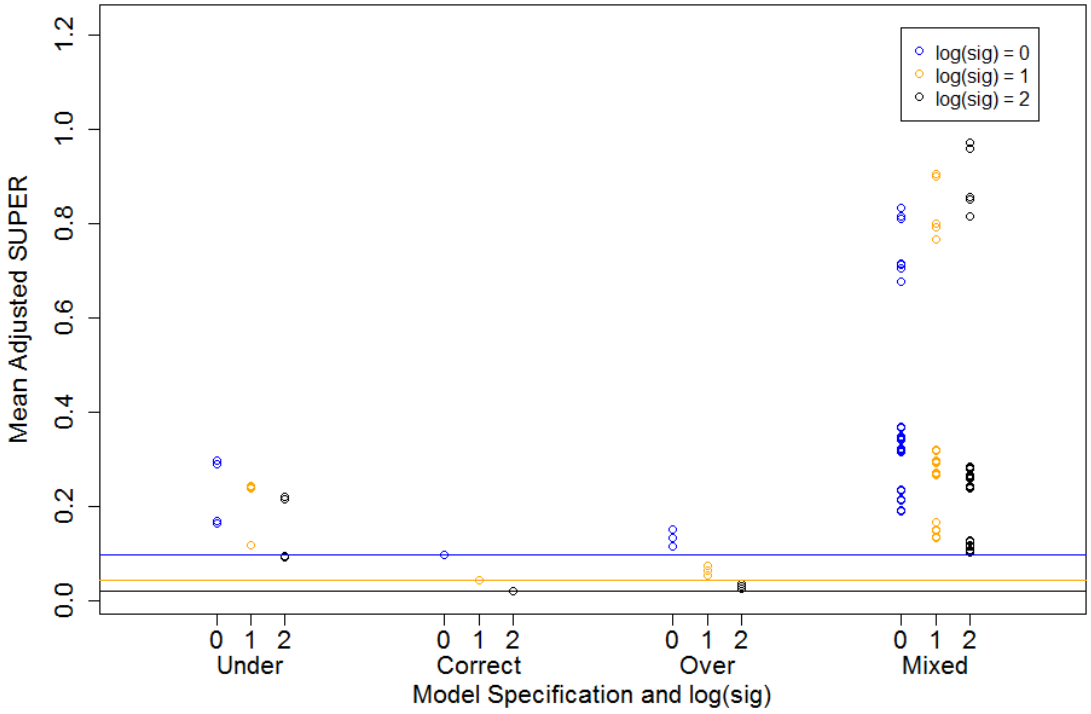

Figure 2.25: Set AS11 - Means of adjusted SUPER by model specification, $N = 100$.

	log(signal)		
Model Specification	0	1	2
Under	294	36	4
Correct	402	718	757
Over	169	220	238
Mixed	135	26	1

Table 2.29: Set AS11 - Counts of adjusted SUPER by model specification, $N = 100$.

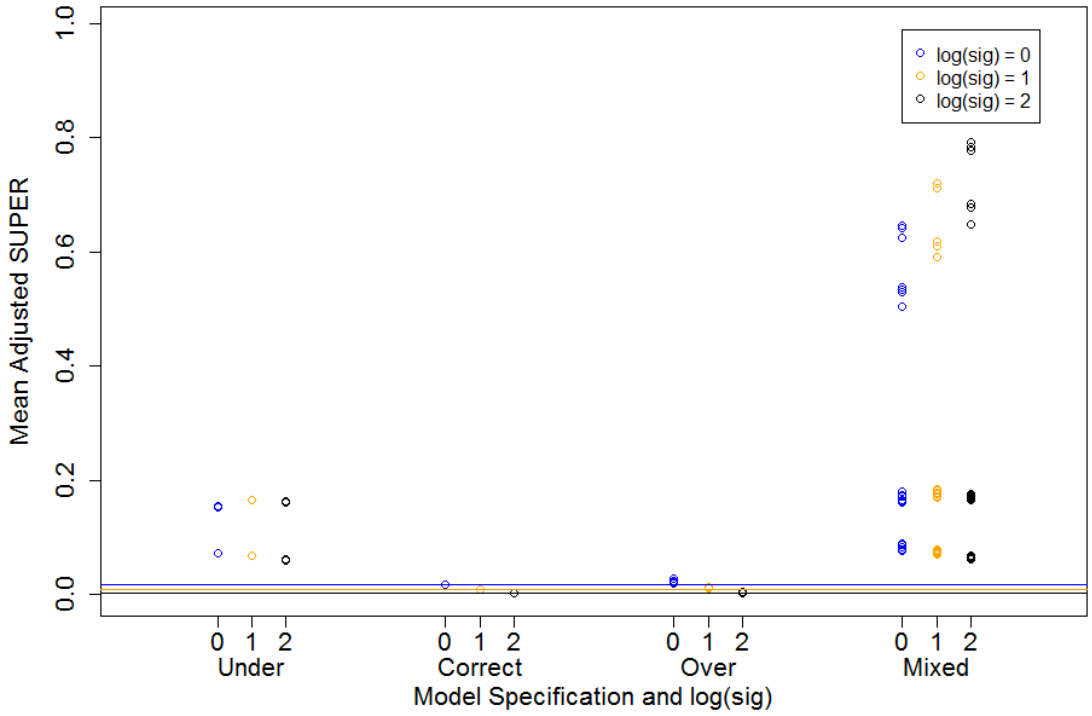

Figure 2.26: Set AS12 - Means of adjusted SUPER by model specification, $N = 500$.

	log(signal)		
Model Specification	0	1	2
Under	0	0	0
Correct	832	856	859
Over	168	144	141
Mixed	0	0	0

Table 2.30: Set AS12 - Counts of adjusted SUPER by model specification, $N = 500$.

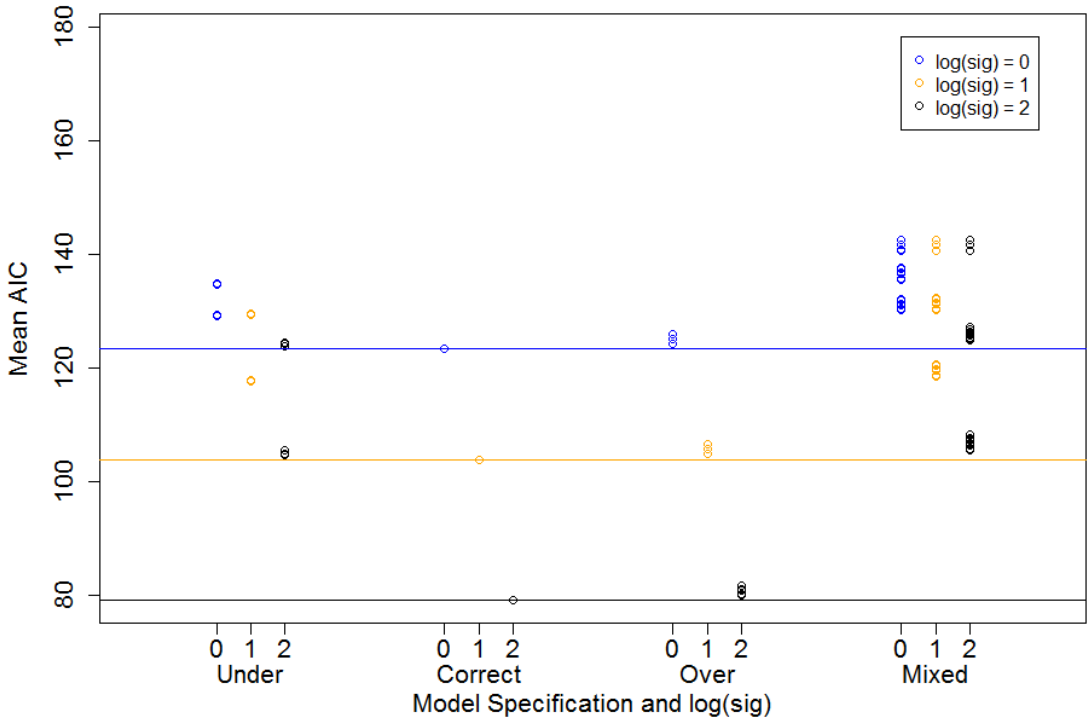

Figure 2.27: Set AS13 - Means of AIC
by model specification, $N = 100$.

	log(signal)		
Model Specification	0	1	2
Under	168	5	0
Correct	390	577	554
Over	298	410	446
Mixed	144	8	0

Table 2.31: Set AS13 - Counts of AIC
by model specification, $N = 100$.

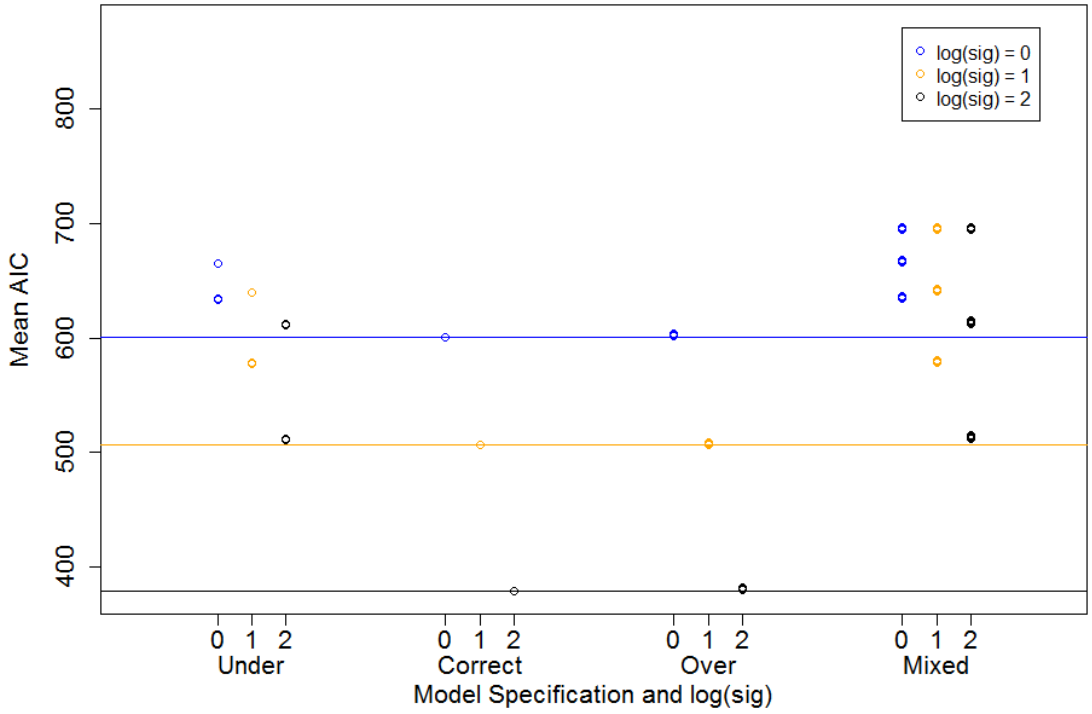

Figure 2.28: Set AS14 - Means of AIC
by model specification, $N = 500$.

	log(signal)		
Model Specification	0	1	2
Under	0	0	0
Correct	595	602	634
Over	405	398	366
Mixed	0	0	0

Table 2.32: Set AS14 - Counts of AIC
by model specification, $N = 500$.

Similar to the nested setting, when the log-signal is held constant and the sample size is increased, all of the model selection criteria show improved correct model selection counts. The same is true when the sample size is held constant and the log-signal is increased. The MMPE, FSPE, and SUPER all tend to perform better when calculated using the adjusted c-statistic rather than its traditional counterpart. This improved performance is most pronounced for the FSPE.

The means and selection counts for the MMPE reflect strong protection from choosing underspecified and mixed misspecified models. The mean criterion values for underspecified and mixed misspecified models are large compared to those for correctly specified or overspecified models. For each of the sets AS1 - AS4, the minimum mean values correspond to the correctly specified model. The model counts exhibit frequent selection of the true probabilistic mechanism and minimal selections of models that exclude some of the true predictors. Sets AS2 and AS4 both show no selections of underspecified or mixed misspecified models for any of the log-signal values.

The means and selection counts for the FSPE reflect strong protection from choosing overspecified and mixed misspecified models. The mean criterion values for overspecified models and mixed misspecified models are large compared to those for correctly specified or underspecified models. For sets AS7 and AS8, the minimum mean values correspond to the correctly specified model, whereas for set AS5 and AS6, the minimums correspond to underspecified models. The model counts reflect varied selection of the true probabilistic mechanism, with higher frequencies for larger sample sizes, higher log-signals, and the adjusted c-statistic. Sets AS6 and AS8 show virtually no selections of overspecified or mixed misspecified models for any of the log-signal values.

The means and selection counts for the SUPER indicate the strong protection from choosing underspecified and mixed misspecified models exhibited by the

MMPE as well as modest protection from choosing overspecified models as evident with the FSPE. The mean criterion values for underspecified and mixed misspecified models are large compared to those for correctly specified or overspecified models. For each of the sets AS9 - AS12, the minimum mean values correspond to the correctly specified model. The model counts exhibit frequent selection of the true probabilistic mechanism and minimal selections of models that exclude some of the true predictors. Sets AS10 and AS12 show no selections of underspecified or mixed misspecified models.

Finally, the means and selection counts for AIC reflect strong protection from choosing underspecified and mixed misspecified models. The mean criterion values for underspecified and mixed misspecified models are large compared to those for correctly specified or overspecified models. For sets AS13 and AS14, the minimum means values correspond to the correctly specified model. The model counts reflect frequent selection of the true probabilistic mechanism and minimal selections of models that exclude some of the true predictors. Set AS14 show no selections of underspecified or mixed misspecified models. The SUPER behaves very similarly to AIC, but does considerably better in selecting the correctly specified model.

2.5 Application

Now that we have characterized how our proposed selection criteria behave in a simulated setting, we can apply them in a modeling application and examine the results.

Heart disease, which refers to the narrowing or blocking of blood vessels and can induce heart attack or stroke, is the leading cause of death in the United States, claiming nearly 600,000 lives each year. Angiography, a medical imaging technique, allows doctors to visualize the inside of blood vessels so that narrowing can be assessed and a diagnosis about heart disease can be made.

The Cleveland database is a data set that contains 303 patient-specific records, representing 14 attributes. The outcome of interest is the angiographic diagnosis of heart disease, which assumes the value 0 for patients with $< 50\%$ blood vessel diameter narrowing and 1 for patients with $> 50\%$ blood vessel diameter narrowing. The candidate predictor variables include age of the patient (`age`; in years, discrete), sex of the patient (`sex`; 0 = female, 1 = male), chest pain type (`cp`; 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic), resting systolic blood pressure on admission (`trestbps`; continuous), serum cholesterol (`chol`; continuous), fasting blood sugar indicator (`fbs`; 0 = false, 1 = true), resting electrocardiographic result (`restecg`; 0 = normal, 1 = ST-T wave abnormality, 2 = left ventricular hypertrophy), maximum heart rate achieved (`thalach`; continuous), exercise induced angina (`exang`; 0 = no, 1 = yes), ST depression induced by exercise (`oldpeak`; continuous), slope of peak exercise ST segment (`slope`; 1 = upsloping, 2 = flat, 3 = downsloping), number of major vessels colored by fluoroscopy (`ca`; discrete in [0,3]), and thallium stress test result (`thal`; 3 = normal, 6 = fixed defect, 7 = reversible defect).

Since we have 13 potential predictor variables, the all subsets setting will consider a candidate collection comprised of $2^{13} - 1 = 8191$ models. Once each model is fitted, the estimates for the MMPE, FSPE, and SUPER will be calculated using traditional and adjusted c-statistics, and used to ascertain the models favored by each criterion. The selection results will also be compared to the models chosen by AIC. The model selections for the criteria are provided in Table 2.33.

Given the general behaviors of the criteria, it is not surprising that both of the MMPEs select models with higher orders, since these criteria tend to protect from underfitting. Conversely, we see that both of the FSPEs select models with lower orders, since these criteria tend to protect from overfitting.

Since each SUPER criterion is an additive combination of the corresponding

Criterion	Variables in Selected Model
Traditional MMPE	`trestpbs thalach oldpeak sex cp slope ca thal`
Adjusted MMPE	`oldpeak sex cp exang slope ca thal`
Traditional FSPE	`thalach`
Adjusted FSPE	`thalach thal`
Traditional SUPER	`trestbps sex cp exang slope ca thal`
Adjusted SUPER	`oldpeak sex cp exang slope ca thal`
AIC	`trestbps thalach oldpeak sex cp exang slope ca thal`

Table 2.33: Model selections by criterion.

MMPE and FSPE criteria, we might expect the models selected by each of the SUPER to be of intermediate order relative to the models selected by the corresponding MMPE and FSPE. Indeed, this is the case. However, we might also expect the variables in the selected models for each of the constituent criteria to be manifest in the selected models based on the composite criteria. This expectation is better fulfilled by the criteria based on the adjusted c-statistic than for those based on the traditional c-statistic.

For the traditional c-statistic, we note variables that were selected by at least one of the MMPE and FSPE, such as `thalach` and `oldpeak`, yet do not appear in the model chosen by the SUPER. Additionally, we note a variable, `exang`, that is in the model favored by the SUPER but is not in either of the models chosen by the MMPE or FSPE.

For the adjusted c-statistic, the model selected by the SUPER is the same as that selected by the MMPE. Similar to the traditional case, the model chosen by the FSPE contains the variable `thalach`, even though that variable is not in the model favored by the SUPER. However, the variable `thal` is included in the models selected by both the FSPE and SUPER.

AIC selects the model that is based on the union of the variables present in all six previous models, indicating that AIC assesses the increase in complexity resulting from the additional predictors as being offset by the improvement in fit.

For the seven criteria considered in this application, since the model selections in Table 2.33 differ, an obvious question arises: what candidate model might we deem as optimal? Obviously, any final decision should be based on both statistical and clinical significance. However, since we lack the informed perspective of a cardiologist, we cannot properly assess clinical significance. Despite this, we can use the selection results, guided by the simulation findings, to draw conclusions from a statistical standpoint. Based on the three fundamental properties of a good model, along with a recognition of the problem of predictive bias associated with underfitting, we are inclined to initially favor the models selected by the adjusted SUPER criterion and AIC. The adjusted SUPER places more emphasis on parsimony, whereas AIC places more on goodness-of-fit. The simulations illustrate the propensity of AIC to include some unnecessary variables that are not selected by the adjusted SUPER, so it is not surprising that AIC chooses a higher order model. This leads us to be skeptical as to the importance of the additional variables included in the model chosen by AIC. As a result, we recommend the model selected by the adjusted SUPER.

2.6 Summary Conclusion

In this chapter, we have introduced three target predictive error measures aimed to serve as the basis for model selection criteria in the logistic regression framework. Furthermore, we have provided estimates for these measures using leave-one-out cross-validated c-statistics, and have assessed the model selection capabilities of these estimates under simulated and applied settings. The estimates for

the MMPE and FSPE appear adequate as model selection criteria, and may warrant consideration in settings where the investigators are more concerned with the consequences of underfitting than overfitting (for the MMPE), or vise versa (for the FSPE). However, the estimates for the SUPER far outperformed the corresponding constituent criteria. Using AIC as a model selection criterion standard, the SUPER criterion did quite well at selecting the correct model in both the nested and all subsets frameworks.

Future work involves investigating convex combinations of the MMPEs and FSPEs to determine if some weighted combination of these measures, as governed by a tuning parameter, could yield a model selection criterion superior to the estimate of the SUPER. Such a criterion might allow one to better regulate the risks of underfitting and overfitting. This type of flexibility could be useful in certain modeling applications; however, the determination of an appropriate value of the tuning parameter presents a daunting challenge.

CHAPTER 3
AN ALTERNATE APPROACH TO PSEUDO-LIKELIHOOD MODEL SELECTION IN THE GLMM FRAMEWORK

3.1 Introduction

In regression frameworks, model selection encompasses a variety of approaches, including the use of diagnostics and criteria to evaluate a collection of candidate models and determine which provides the most appropriate fit to a given set of data. In the maximum likelihood framework, the use of information criteria based on the empirical likelihood is pervasive, where the empirical likelihood appears in the goodness-of-fit term of the criterion. To compare such criteria across different fitted models, the outcome data must be identical; otherwise, the likelihoods will not correspond.

The pseudo-likelihood method is a conventional fitting approach for the framework of the generalized linear mixed model (GLMM). With this method, *pseudo-data* are generated via a transformation of the outcome and used as a surrogate for the original response data. Pseudo-data are derived from a Taylor series expansion that utilizes both constructs from the candidate model and the original outcome. The purpose of this expansion is to offer a new outcome with an approximate normal distribution. In general, pseudo-data are inconsistent for different model specifications, leading to pseudo-likelihoods that are not comparable. Selection criteria based on the resulting goodness-of-fit statistics are fundamentally dissimilar, rendering comparisons invalid for evaluation.

The purpose of this chapter is twofold. First, we investigate the natural approach to model selection using pseudo-likelihood based information criteria under the GLMM framework. With this approach, the criteria are constructed using the

pseudo-data based on the candidate model at hand. Second, in this setting, we propose a new, improved method for the comparison of information criteria between candidate models. In SAS, the default method for the GLIMMIX (Generalized LInear MIXed Models) procedure is the natural, pseudo-likelihood based approach, leading to invalid model comparisons via information criteria. Our new method will be implemented using the GLIMMIX and MIXED (linear MIXED models) SAS procedures.

In SAS, the default fitting procedure for GLMMs utilizes the pseudo-likelihood. Under the natural approach to model selection in this setting, the use of information criteria, such as the Akaike information criterion (AIC; Akaike, 1973, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978), is problematic. This is due to their reliance on pseudo-data that are generated by transforming the original outcome vector into a new outcome, which is used to construct a Gaussian likelihood. Computing model selection criteria using pseudo-data from the specified model may seem natural, but this approach is fundamentally flawed. Any link function other than the identity link will lead to different pseudo-data under different fitted candidate models, violating the assumption that models under comparison share the same outcome. This violation is ignored under the default GLIMMIX procedure, thereby rendering model selection criteria of dubious utility. For most GLMMs, the identity link is neither the canonical link nor the most appropriate link. In order to ensure that use of model selection criteria is valid, we need to verify that the same pseudo-data are being used for all models under consideration. The simplest way to accomplish this objective is to use the full model (i.e., the model featuring all predictor variables under consideration) to obtain the pseudo-data, and to subsequently fit all subset candidate models with this generated outcome.

One common approach to generalized linear mixed modeling utilizes residual pseudo-likelihood. The residual maximum likelihood (REML) method is sometimes

preferred to the maximum likelihood (ML) method in that it accounts for fixed effects in the construction of the objective function, which reduces the bias in covariance parameter estimates, sometimes to zero. Both REML and ML can be implemented using the pseudo-likelihood method, but for the purposes of this chapter, we will focus exclusively on ML since AIC and BIC have been developed for this framework. Pseudo-likelihood is one of several GLMM estimation methods, including Gaussian quadrature and Laplace approximation, which are often necessary since integrating random effects out of the joint likelihood is typically intractable. The latter two methods are valuable in that they can approximate the marginal likelihood through numeric integration, which leads to traditional model selection criteria, unlike pseudo-likelihood. However, a disadvantage of numerical approximation is computational expensiveness. Conversely, maximizing a Gaussian likelihood based on pseudo-data, which are created using Taylor series for each iteration, is computationally efficient.

The structure of this chapter is as follows. In section 2, we provide some background regarding generalized linear mixed models and the associated criteria used in model selection, as well as pseudo-likelihood and the default implementation of AIC and BIC under the GLIMMIX procedure in SAS. We will highlight the problem with this approach using an investigative simulation, then propose a solution to this problem. In section 3, in the pseudo-likelihood setting, we propose a new approach for generalized linear mixed model selection using a heuristic justification. We provide a detailed account of the implementation via the MIXED and GLIMMIX procedures in SAS. In section 4, we use a simulation study to illustrate and compare the behavior of the selection criteria found using the default GLIMMIX procedure and the newly proposed technique. In section 5, we apply the new technique in a modeling application and examine the results. Section 6 concludes.

3.2 Background

In this section, we begin with a review of model selection criteria. We also examine the construction of the pseudo-likelihood and its impact on model selection criteria. Under the default implementation of the GLIMMIX procedure in SAS, we show that the comparison of information criteria for non-normal models is not appropriate and can lead to misguided model selections. We provide an investigative simulation, which illustrates model selection tendencies under the default implementation, then propose a more suitable implementation approach for this framework.

3.2.1 Model Selection Criteria

Statistical models are used to characterize the relationship between an outcome of interest and explanatory factors. Models condense information into an interpretable form, from which investigators can draw inferential conclusions. Modeling frameworks have been developed to handle outcomes that assume distributions of all varieties. Once fit, models can be applied to new data in order to predict new outcomes.

An optimal statistical model is characterized by three features: (1) parsimony, which refers to model simplicity; (2) goodness-of-fit, which indicates the conformity of the fitted model to the data at hand; (3) generalizability, which reflects the ability of the fitted model to predict or describe new outcomes. Parsimony and goodness-of-fit tend to pull in opposing directions with regards to model complexity, so it is important to strike a suitable balance between those two attributes, while still achieving generalizability.

In a model selection problem, an investigator strives to find the "best" model from a collection of candidate models, where optimality may be defined based on

adherence to the preceding principles. For theoretical and methodological developments pertaining to model selection, one needs to assume the existence of an underlying generating probabilistic mechanism. We will refer to this mechanism as a true model. In our development, we will assume that the true model is contained within the candidate collection. Although this is a strong assumption, it is commonly employed in model selection developments for either mathematical tractability or conceptual clarity. Here, we impose the assumption for the benefit of the latter.

Investigators frequently use model selection criteria in order to compare different candidate models and ascertain the one that best exemplifies the three optimality features. A common approach to the development of a model selection criterion is to estimate a measure that assesses the disparity between the fitted model under consideration and the true probabilistic mechanism. Such a measure is known as an *expected discrepancy*. One of the most popular and useful expected discrepancies is based on the Kullback-Leibler (K-L) information, a measure introduced by Kullback and Leibler (1951) and further investigated by Kullback (1968). This discrepancy serves as the basis for the ubiquitous Akaike (1973, 1974) information criterion (AIC) and its variants. One of the favorable properties of AIC is asymptotic efficiency, in the sense of Shibata (1980, 1981). Assuming that the generating model is of an infinite dimension and thus is not in the candidate collection, an efficient criterion will asymptotically select the fitted candidate model which minimizes the mean squared error of prediction (Cavanaugh, 1999). As outlined by Linhart and Zucchini (1986), a major deficiency of AIC arises in small to moderate sample-size applications, where AIC will often severely underestimate the K-L discrepancy and may tend to decrease as model complexity increases. Variants of AIC have been proposed to address this deficiency and to relax the stringent model specification assumption under which the criterion is derived. These include corrected AIC (AICc)

(Sugiura, 1978; Hurvich and Tsai, 1989), designed for small-sample settings, the Takeuchi (1976) information criterion (TIC), which relaxes the model specification assumption, CAIC (Bozdogon, 1987), which corrects for the lack of consistency, and a quasi-likelihood based measure for generalized linear models fit via generalized estimating equations (QIC) (Pan, 2001). Additional variants of AIC have been proposed based on complexity penalizations that are evaluated using a computationally intensive algorithm, including cross-validation (Stone, 1977; Davies, Neath and Cavanaugh, 2005), bootstrapping (Ishiguro, Sakamoto and Kitagawa, 1997; Cavanaugh and Shumway, 1997; Shibata, 1997), and Monte Carlo simulation (Hurvich, Shumway, and Tsai, 1990; Bengtsson and Cavanaugh, 2006).

Another common model selection criterion is BIC. One of the favorable properties of BIC is its consistency, which is characterized as follows. Suppose that the generating model is of a finite dimension, and that this model is represented in the candidate collection under consideration. A consistent criterion will asymptotically select the fitted candidate model having the correct structure with probability one. BIC was originally derived by Schwarz "for the case of independent, identically distributed observations, and linear models." Variants of BIC that generalize the criterion for other frameworks have been developed by Stone (1979), Kashyap (1982), Leonard (1982), Haughton (1988), and Cavanaugh and Neath (1999).

An advantage that AIC and BIC offer over common model comparison inferential techniques, such as the likelihood ratio test, is that the models under consideration do not need to be nested or even follow the same distribution. As long as the models are fitted using the same outcome, the selection criteria can be compared to determine the more appropriate fit. For a collection of candidate models, the model with the minimum AIC or BIC is deemed the most favorable. However, for the sake of parsimony, a model of lower order and within two units of the minimum information criterion is also considered a suitable selection. (See

Burnham and Anderson, 2002, p. 70, regarding AIC, and Kass and Raftery, 1995, p. 777, regarding BIC).

The conditional mean structure of a mixed model is characterized using a linear combination of predictor variables and random effects. For the purpose at hand, we assume that the selection of fixed effects is the focus. In order to ensure that an exhaustive model search has been conducted, investigators use the all possible subsets approach, which allows for comparison of every combination of variables. For a set of p predictor variables, the all subsets approach considers the $2^p - 1$ collections of predictor variables, excluding the null model. Once all models have been fitted, their corresponding model selection criteria are compared to determine which model is "best." For an effective criterion, in large sample settings where the underlying effects are all appreciable in magnitude, the true model will generally have the highest probability of being chosen out of the collection of candidate models. However, in such settings, model overfitting has a less detrimental impact on inferential objectives than underfitting.

3.2.2 Problem with Pseudo-Likelihood Criteria

As mentioned before, in order for AIC and BIC to serve as legitimate model selection criteria, the candidate class of models must share the same outcome. In fitting GLMMs, the pseudo-likelihood is determined for each model based on different pseudo-data. AIC and BIC use the maximum log pseudo-likelihood in place of the usual maximum log likelihood in the formulation of the goodness-of-fit term. If we wish to compare two candidate GLMMs using AIC or BIC, different pseudo-data will lead to criteria that cannot appropriately compare the two models. These notions will be developed more technically in the following subsections.

3.2.3 Generalized Linear Mixed Models

The GLMM framework is a natural extension of the linear mixed model (LMM) framework. GLMMs are the best tool for analyzing normal and nonnormal data that involve random effects, requiring only the specification of a conditional distribution for the outcome, a link function, and a covariance structure for the random effects (Bolker et al., 2009).

Under the GLMM framework, conditional on the random effect parameter vector γ, the $t_i \times 1$ response measurement vector Y_i is assumed to have a distribution in the exponential family, for $i = 1, ..., s$. We have the expression

$$g_i(E[Y_i|\gamma]) = \eta_i = X_i\beta + Z_i\gamma,$$

where X_i is the fixed effects design matrix, β is the $p \times 1$ fixed effect parameter vector, Z_i is the random effects design matrix, and $g_i(\cdot)$ is a strictly monotonic function from $\mathbb{R}^{T_i}$ to $\mathbb{R}^{T_i}$ that maps the elements of the conditional mean vector $E[Y_i|\gamma] = \mu_i$ to the elements of linear predictor $\eta_i = X_i\beta + Z_i\gamma$. The relevant variance/covariance matrices can be expressed as follows:

$$\text{Var}[\gamma] = G$$

and

$$\text{Var}[Y_i|\gamma] = A_i^{1/2} R_i A_i^{1/2}.$$

Here, $A_i^{1/2} = \text{diag}(\sqrt{\text{Var}[Y_{it}]}) = \text{diag}(\sqrt{\upsilon(\mu_{it})})$, where $\upsilon(\mu_{it})$ is the variance of Y_{it}, so that $A_i^{1/2} A_i^{1/2}$ represents the variance/covariance matrix of Y_i under independence and $R_i = \text{Corr}(\epsilon_i)$.

Using the preceding foundation for the GLMM framework, we can examine the methodology behind how a model is fit based on the pseudo-likelihood approach.

3.2.4 Pseudo-Likelihood Fitting Approach

Though GLMMs provide considerable opportunities for advancement and flexibility in modeling data, inference for these models is complicated by the integrals in the derivation of marginal likelihood estimating equations (Dean and Nielsen, 2007). Pseudo-likelihood is one of several fitting approaches in the GLMM framework. The other primary approaches include Gaussian quadrature and the Laplace approximation method. The advantage of using pseudo-likelihood over these other approaches is its computational efficiency. The qualifier "pseudo" is used because the likelihood is based on a linearized transformation of the data, which is assumed normal, and is not the actual likelihood based on the original data and its underlying distribution.

The following material is outlined in SAS/STAT 9.3 User's Guide: Mixed Modeling. Under the GLMM framework, we have

$$E[Y|\gamma] = g^{-1}(X\beta + Z\gamma) = g^{-1}(\eta) = \mu,$$

where $Y = (y_1', \ldots, y_s')'$, $\gamma \sim N(0, G)$, $E[Y|\gamma] = \mu$,

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_s \end{bmatrix}, Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_s \end{bmatrix},$$

and $\eta = (\eta_1', \ldots, \eta_s')'$. Here, $g(\cdot)$, is a link function from $\mathbb{R}^N$ to $\mathbb{R}^N$, defined analogously to $g_i(\cdot)$, that maps the conditional mean vector μ to the systematic component η. We also have $\text{Var}[Y|\gamma] = A^{1/2}RA^{1/2}$, where

$$A = \text{blk diag}(A_1, \ldots, A_s) \text{ and } R = \text{blk diag}(R_1, \ldots, R_s).$$

Following Wolfinger and O'Connell (1993), a first-order Taylor series of μ about iterated parameter estimates $\widetilde{\beta}$ and $\widetilde{\gamma}$ yields

$$g^{-1}(\eta) \doteq g^{-1}(\widetilde{\eta}) + \widetilde{\Delta}X(\beta - \widetilde{\beta}) + \widetilde{\Delta}Z(\gamma - \widetilde{\gamma})$$

where

$$\widetilde{\Delta} = \left(\frac{\partial g^{-1}(\eta)}{\partial \eta} \right)_{\widetilde{\beta},\widetilde{\gamma}}$$

is a diagonal matrix of derivatives of the conditional mean evaluated at the expansion locus. Rearranging terms yields the expression

$$\widetilde{\Delta}^{-1}(\mu - g^{-1}(\widetilde{\eta})) + X\widetilde{\beta} + Z\widetilde{\gamma} \doteq X\beta + Z\gamma. \tag{3.1}$$

With reference to the left side of the preceding approximation, we define the pseudo-data as

$$\widetilde{\Delta}^{-1}(Y - g^{-1}(\widetilde{\eta})) + X\widetilde{\beta} + Z\widetilde{\gamma} \equiv P. \tag{3.2}$$

Note that

$$\mathrm{Var}[P|\widetilde{\gamma}] = \widetilde{\Delta}^{-1} A^{-1/2} R A^{-1/2} \widetilde{\Delta}^{-1}.$$

Based on equations (3.1) and (3.2), one can thus consider the model

$$P = X\beta + Z\gamma + \epsilon, \tag{3.3}$$

which is a linear mixed model with pseudo-response data P, fixed effects, β, random effects γ, and $\mathrm{Var}[\epsilon] = \mathrm{Var}[P|\gamma]$.

Now define

$$V(\theta) = ZGZ' + \widetilde{\Delta}^{-1} A^{1/2} R A^{1/2} \widetilde{\Delta}^{-1}$$

as the marginal variance in the linear mixed pseudo-model, where θ is the $(q \times 1)$ parameter vector containing all of the unknown parameters in G and R. Based on the linearized model in equation (3.3), an objective function can be defined, assuming that the distribution of P is known. The pseudo-likelihood fitting procedure assumes that ϵ has a normal distribution. The log pseudo-likelihood is then defined

as

$$l(\theta, P) = -\frac{1}{2}\log[V(\theta)] - \frac{1}{2}R'V(\theta)^{-1}R - \frac{N}{2}\log\{2\pi\}$$

with $R = P - X(X'V^{-1}(\theta)X)^{-1}X'V^{-1}(\theta)P$. Here, N denotes the overall sample size. We assume that X is $N \times (p+1)$ of rank $(p+1)$ and that $X'V^{-1}(\theta)X$ is invertible. The objective function for minimization is $-2l(\theta, p)$. At convergence, the profiled parameters are estimated and the random effects are predicted as

$$\hat{\beta} = (X'V(\hat{\theta})^{-1}X)^{-1}X'V(\hat{\theta})^{-1}\hat{P}$$

$$\hat{\gamma} = \hat{G}Z'V(\hat{\theta})^{-1}\hat{R}.$$

The default approach for the GLIMMIX procedure fits GLMMs based on linearizations, which utilize Taylor series expansions to approximate the data using pseudo-data. In the iterative fitting routine, the pseudo-data are constructed using current regression and covariance parameter estimates. The GLMM is then approximated by a linear mixed model (LMM) based on the pseudo-data. The LMM fitting is itself an iterative process, yielding new parameter estimates that update the linearization, generating a new LMM. This process is complete once the LMM fits fail to change beyond a pre-specified tolerance.

The empirical pseudo-likelihood under the GLMM framework can seemingly be used to calculate model selection criteria in a similar manner to the empirical likelihood under the GLM framework. The apparent definitions for AIC and BIC are as follows:

$$\text{AIC} = -2\ell(\hat{\theta}, \hat{P}) + 2((p+1)+q),$$

$$\text{BIC} = -2\ell(\hat{\theta}, \hat{P}) + \log(N)((p+1)+q).$$

However, since model selection criteria constructed under the pseudo-likelihood approach utilize a different pseudo-data vector, P, for each model under consideration, comparisons of conformity between the models and the data are inappropriate.

3.2.5 Investigative Simulation

In order to illustrate the ineffectiveness of model selection via information criteria based on the default GLIMMIX fitting procedure, we consider a simulated data set in which the true, generating model is known. We can thereby illustrate the efficacy of a model selection criterion in delineating this true model from other candidates.

For a sample size of $N = 100$, let Y_{ij} represent the j^{th} observation on subject i, for $i = 1, 2, ..., 20$; $j = 1, 2, ..., 5$. Let

$$\{X_{1ij}, X_{2ij}, ..., X_{6ij}\} \overset{iid}{\sim} N(0, 1)$$

represent the j^{th} set of covariates on subject i with

$$\gamma_i \overset{iid}{\sim} N(0, 1)$$

as the random effect for subject i. The Bernoulli outcome

$$Y_{ij} \overset{ind}{\sim} \text{Bernoulli}(\pi_{ij})$$

is based on the generating model

$$\text{logit}(\pi_{ij}) = x_{1ij} + x_{2ij} + x_{3ij} + \gamma_i.$$

An evaluation of all possible subset models for the fixed effects, not including the null model, involves fitting $2^6 - 1 = 63$ models. We will consider four criteria for model selection: minimum AIC (minAIC), most parsimonious model within two units of minimum AIC (minAIC2), minimum BIC (minBIC), and most parsimonious model within two units of minimum BIC (minBIC2).

The model selections according to the four criteria are summarized in Table 3.1. As we can see from the table, all four criteria select the same order one model, which does not include any of the predictors in the generating model. This gross confusion of the actual mean structure is concerning as it shows the futility of the selection procedure for data that follows this modeling framework. Ideally, in settings where the sample size is large and the effects are all reasonable in size, we would

want a model selection procedure that identifies the correct model specification with higher probability than any other candidate model.

Selection Criteria	Selected Variables
minAIC	X_5
minAIC2	X_5
minBIC	X_5
minBIC2	X_5

Table 3.1: Model selections for four criteria using default GLMMMIX procedure.

3.2.6 Proposed Solution

In order to make the comparison of information criteria between models valid, we need to use the same outcome for all candidate models. The pseudo-data are a function of the predictor variables in the model and vary for different model specifications. In order to ensure that the outcome is the same for all models under consideration, we can generate the pseudo-data from the full model (i.e., the model that contains all predictors under consideration), and fit all other candidate models to that pseudo-data.

3.3 New Method

This section applies our proposed solution to model selection for the pseudo-likelihood setting under the GLMM framework, first by providing a heuristic justification, then by detailing the implementation in SAS via the GLIMMIX and MIXED procedures.

3.3.1 Heuristic Justification

In order to be able to make valid model comparisons using selection criteria such as AIC and BIC, the candidate models must be based on the same outcome data. Under the GLMM framework based on pseudo-likelihood estimation, unique pseudo-data are generated for each candidate model; the pseudo-data then serve as the basis for the construction of each model pseudo-likelihood. The unique pseudo-data lead to disparate pseudo-likelihoods, which are not commensurable, invalidating the comparison of AIC and BIC across candidate models.

Ideally, we would construct the pseudo-data based on the true model, which we do not know. However, if we assume that we have access to the predictors in the true model, we can generate the pseudo-data by using the full model, which subsumes the true model. Using this full model pseudo-data, a LMM can be fit with any subset of predictors from the full model. Since all models will share the same pseudo-data, information criteria can validly be compared for the purposes of model selection.

3.3.2 Implementation via SAS PROC MIXED/PROC GLIMMIX

The GLIMMIX procedure generates the pseudo-data, then fits the model with the normalized outcome, similar to the manner that the MIXED procedure would fit the model with the same transformed outcome. For an outcome of interest and a collection of candidate predictor variables, we fit the full model using GLIMMIX and output the predicted and residual components of the pseudo-data, given by $X\widetilde{\beta} + Z\widetilde{\gamma}$ and $\widetilde{\Delta}^{-1}(Y - g^{-1}(\widetilde{\eta}))$, respectively. Equation (3.2) shows that the sum of these components yields the pseudo-data, P. We then use the MIXED procedure with the full model pseudo-data to fit all candidate models of interest and generate the desired information criteria, which can be compared for the purposes of model selection.

Looking back at the investigative simulation presented in the previous section, we can apply our new model selection technique, which produces the results featured in Table 3.2. The most noticeable difference from the default technique is that all four criteria select models that include the three predictors in the generating model. Although three of the criteria select overspecified models, this is preferable to favoring a model that omits any of the true predictors, highlighting the value of our proposed technique. These selection results are corroborated in the next section, which presents a large scale simulation study covering four common modeling distributions.

Selection Criteria	Selected Variables
minAIC	X_1 X_2 X_3 X_4 X_6
minAIC2	X_1 X_2 X_3 X_4
minBIC	X_1 X_2 X_3 X_4
minBIC2	X_1 X_2 X_3

Table 3.2: Model selections for four criteria using new technique.

3.4 Simulation Study

We compile and report a comprehensive two-part simulation study in order to assess and compare the performances of pseudo-likelihood based model selection criteria under the GLMM framework using the natural approach and our newly proposed approach. By generating numerous replicated samples, and using these samples for model fitting and selection, we can characterize the general behaviors of the criteria under both approaches. This simulation study considers two model selection settings and focuses on the criteria AIC and BIC for outcomes following

Bernoulli, binomial ($n = 10$), Poisson, and gamma distributions.

The two settings that we consider are nested modeling and all subsets modeling. Each set in the simulation study is based on generating 1000 data samples of sample size $N = 100$. A single random effect intercept is included in the generating model, which effectively partitions the data set into 20 groupings (indexed by i) of 5 observations each. The regression parameters for the systematic component are set to be identical for all predictor variables, and the intercept is set to zero. The simulation study is designed as a factorial experiment, where the factors are the distribution (Bernoulli, binomial ($n = 10$), Poisson, or gamma), selection criterion (AIC or BIC), and modeling construction method for the pseudo-data (candidate model or full model).

For each replicated sample, we record the optimal model selected for each criterion. We summarize the model selections in a table of counts. Such tables allow us to assess the ability of each criterion to pick the correct model, as well as to see which types of incorrectly specified models each criterion tends to favor. Additionally, for every candidate model order, we compute the means of AIC and BIC using the techniques based on the candidate model construction (CMC) of pseudo-data and full model construction (FMC) of pseudo-data. The CMC method is the natural (yet incorrect) approach; the FMC method is the proposed approach. We plot these means by model order to provide a visual representation of the behaviors of the criteria for underspecified, correctly specified, overspecified, and mixed misspecified models. (Mixed misspecified models contain both legitimate and spurious predictors, yet do not contain all of the predictors represented in the true model.) The figures featuring the criterion means are presented adjacent to the corresponding table of selection counts.

3.4.1 Nested Setting

The first part of the simulation study involves a nested modeling setting. Here, we construct data sets with 10 fixed effect predictor variables

$$X_{1ij}, X_{2ij}, ..., X_{10ij} \overset{iid}{\sim} \mathrm{N}(0,1),$$

a random effect predictor

$$\gamma_i \overset{iid}{\sim} N(0,1),$$

and outcome Y_{ij} that is generated as seen in Table 3.3. Let $\mu_{ij} = E[Y_{ij}|\gamma_i]$. For the gamma distribution, the scale parameter is set equal to one, leaving the shape parameter equal to μ_{ij}.

Distribution	Generating Model	Link
Bernoulli	$Y_{ij}\|\gamma_i \overset{ind}{\sim} \text{Bernoulli}(\mu_{ij})$	logit
Binomial ($n = 10$)	$Y_{ij}\|\gamma_i \overset{ind}{\sim} \text{Binomial}(\mu_{ij})$	logit
Poisson	$Y_{ij}\|\gamma_i \overset{ind}{\sim} \text{Poisson}(\mu_{ij})$	log
Gamma	$Y_{ij}\|\gamma_i \overset{ind}{\sim} \text{Gamma}(\mu_{ij})$	log

Table 3.3: Generating models and link functions for each distribution.

Once we have the full data sets, we fit 10 nested models of orders 1 through 10. The models are as follows:

Order 1: $\mu_{ij} = g^{-1}(x_{1ij}\beta_1 + \gamma_i)$

Order 2: $\mu_{ij} = g^{-1}(x_{1ij}\beta_1 + x_{2ij}\beta_2 + \gamma_i)$

$\vdots$

Order 10: $\mu_{ij} = g^{-1}(x_{1ij}\beta_1 + x_{2ij}\beta_2 + ... + x_{10ij}\beta_{10} + \gamma_i)$,

where $g(\cdot)$ is the canonical link function for the generating distribution. The nested setting allows us to compare the criteria for models that are underspecified (candidate predictor set is a proper subset of the set of "true" predictors; i.e., the

predictors for the generating model), correctly specified (candidate predictor set is exactly the same as the set of true predictors), and overspecified (true predictor set is a proper subset of the candidate predictor set). Again, the simulation study is designed as a factorial experiment, where the factors are the distribution, selection criterion, and the modeling method for the construction of the pseudo-data. Similar to the investigative simulation, in addition to selecting models with a minimum AIC or BIC, we will also select the most parsimonious model within two units of the minimum AIC and BIC.

Table 3.4 lists the ID for each simulation set, along with the associated levels of the factors. The figures corresponding to sets N1 - N8 contain two curves; one for each modeling construction method. The AIC and BIC means of the 1000 replications are calculated for each model order and plotted, illustrating the general behavior of the criteria under each construction method. The tables corresponding to set N1 - N8 feature model order selection counts by each construction method.

Set ID	Distribution	Criterion
N1	Bernoulli	AIC
N2	Bernoulli	BIC
N3	Binomial (n=10)	AIC
N4	Binomial (n=10)	BIC
N5	Poisson	AIC
N6	Poisson	BIC
N7	Gamma	AIC
N8	Gamma	BIC

Table 3.4: Model factor levels for nested simulation setting.

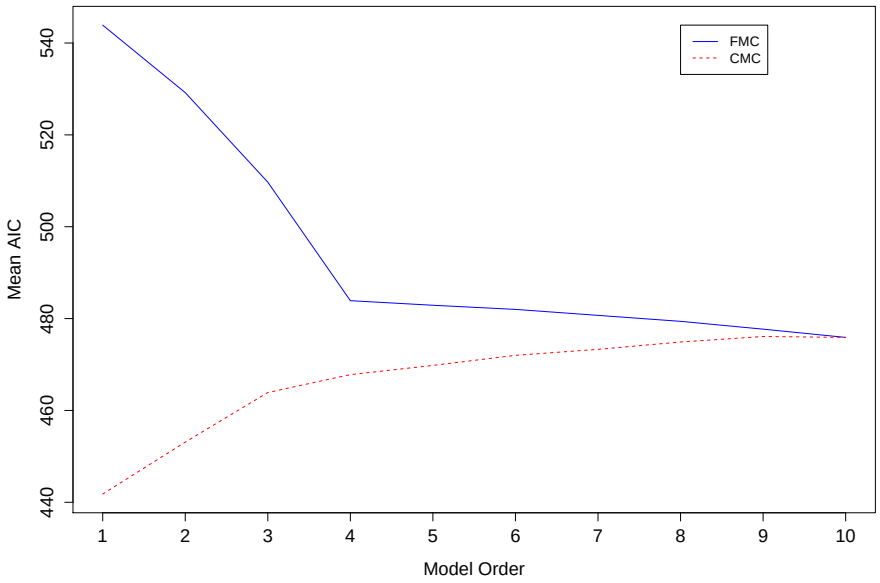

Figure 3.1: Set N1 - Bernoulli outcomes; nested setting.
Means of AIC by model order, $N = 100$.

Model Order	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
1	520	14	580	25
2	43	9	29	8
3	52	14	44	27
4	109	254	116	330
5	47	79	34	76
6	30	81	25	66
7	32	75	31	80
8	42	87	42	84
9	45	119	38	116
10	80	268	61	188

Table 3.5: Set N1 - Bernoulli outcomes; nested setting.
Counts of AIC selections by model order, $N = 100$.

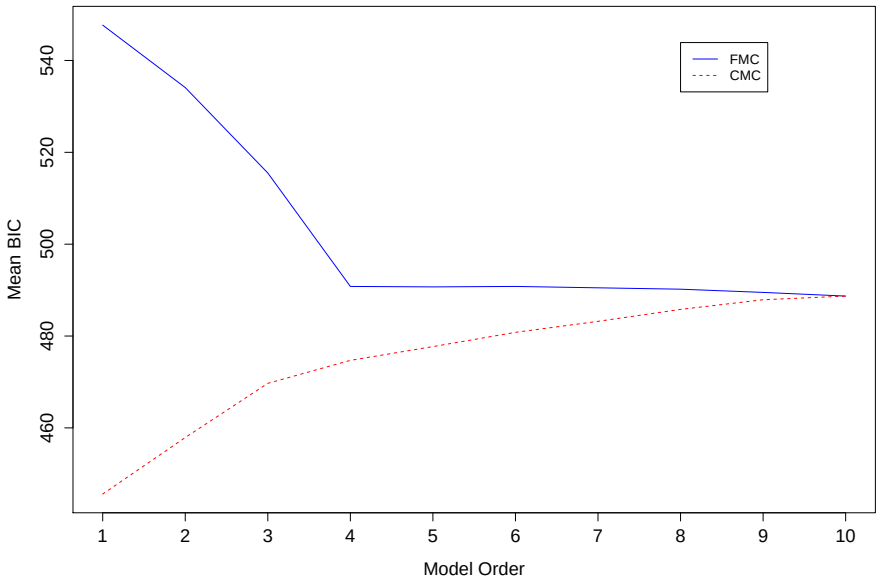

Figure 3.2: Set N2 - Bernoulli outcomes; nested setting.
Means of BIC by model order, $N = 100$.

Model Order	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
1	611	21	656	34
2	39	13	30	12
3	46	26	44	43
4	123	359	118	444
5	39	95	37	80
6	22	72	20	50
7	23	73	19	64
8	26	71	21	60
9	31	88	24	78
10	40	182	31	135

Table 3.6: Set N2 - Bernoulli outcomes; nested setting.
Counts of BIC selections by model order, $N = 100$.

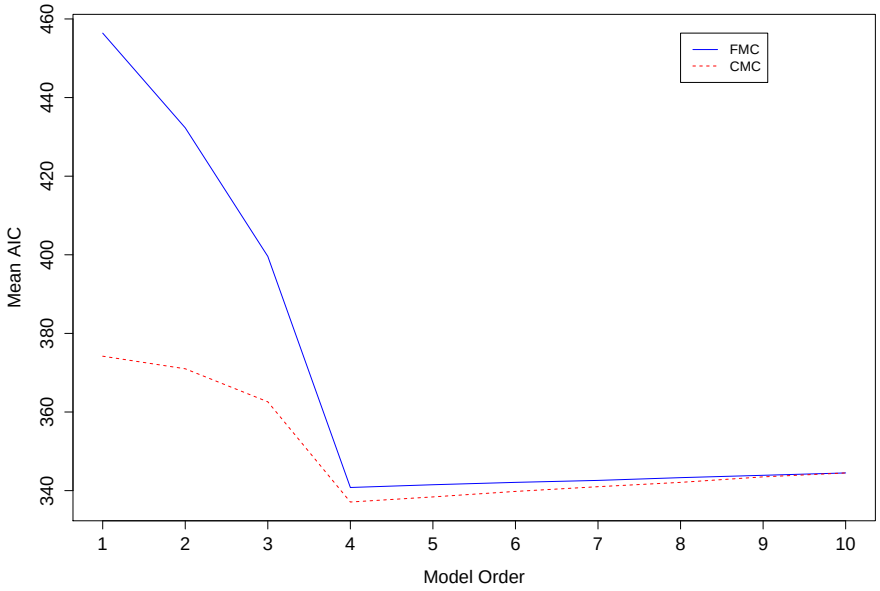

Figure 3.3: Set N3 - Binomial ($n = 10$) outcomes; nested setting. Means of AIC by model order, $N = 100$.

Model Order	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
1	24	5	31	6
2	53	1	49	2
3	110	1	125	1
4	404	559	504	737
5	101	123	84	82
6	86	88	58	43
7	74	60	50	41
8	43	54	37	32
9	44	54	29	27
10	61	55	33	29

Table 3.7: Set N3 - Binomial ($n = 10$) outcomes; nested setting. Counts of AIC selections by model order, $N = 100$.

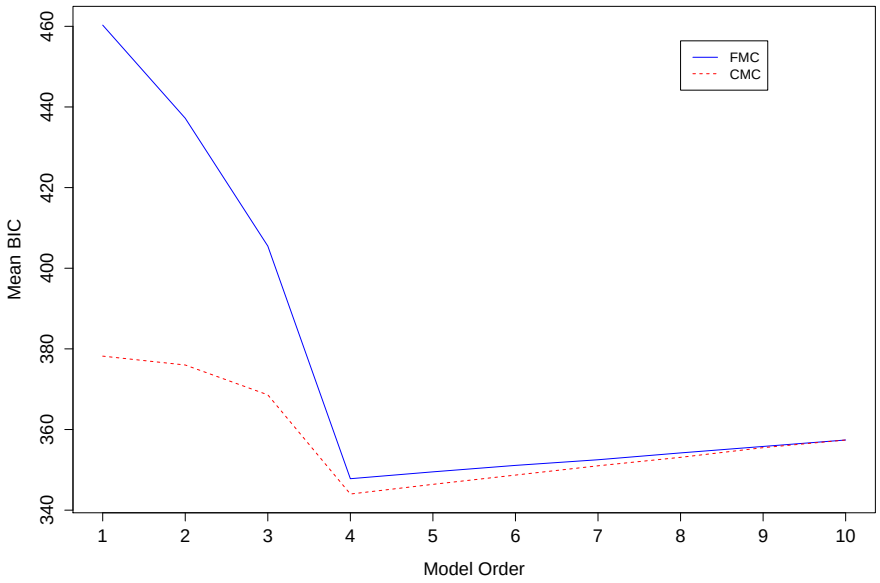

Figure 3.4: Set N4 - Binomial ($n = 10$) outcomes; nested setting. Means of BIC by model order, $N = 100$.

Model Order	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
1	30	6	35	9
2	52	2	50	1
3	130	2	138	1
4	502	746	590	865
5	113	110	75	53
6	59	46	37	24
7	44	32	28	23
8	27	27	23	11
9	23	14	13	6
10	20	15	11	7

Table 3.8: Set N4 - Binomial ($n = 10$) outcomes; nested setting. Counts of BIC selections by model order, $N = 100$.

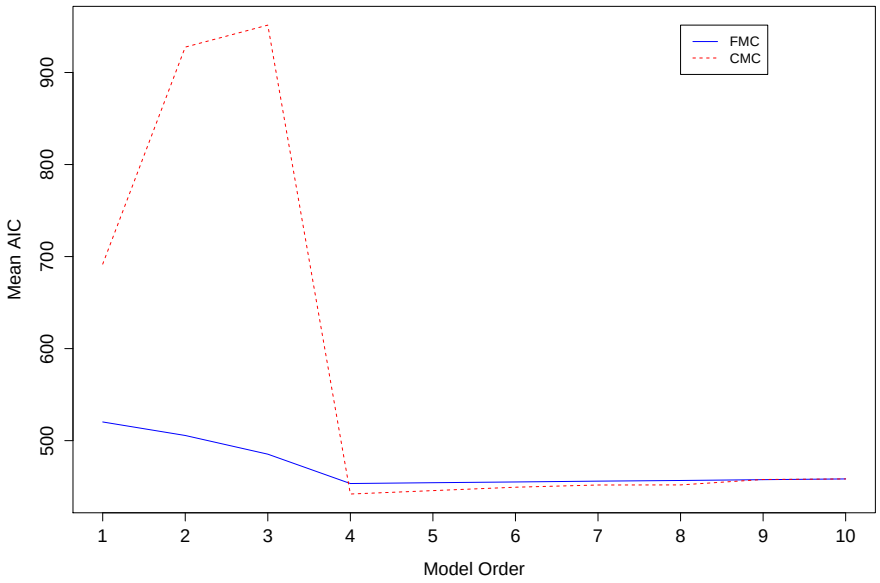

Figure 3.5: Set N5 - Poisson outcomes; nested setting.
Means of AIC by model order, $N = 100$.

Model Order	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
1	302	51	311	84
2	54	24	56	30
3	45	37	45	44
4	213	560	232	686
5	83	119	80	59
6	64	57	64	34
7	55	49	52	26
8	65	38	58	16
9	45	37	38	11
10	74	28	64	10

Table 3.9: Set N5 - Poisson outcomes; nested setting.
Counts of AIC selections by model order, $N = 100$.

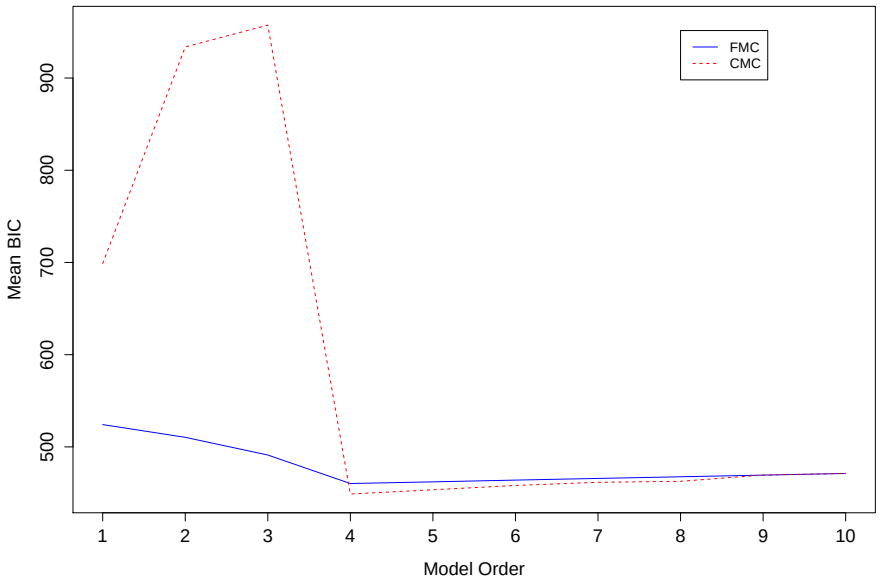

Figure 3.6: Set N6 - Poisson outcomes; nested setting.
Means of BIC by model order, $N = 100$.

Model Order	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
1	303	87	309	116
2	58	32	58	27
3	46	43	46	148
4	237	689	258	742
5	183	79	85	32
6	67	32	64	14
7	51	18	51	8
8	57	11	48	6
9	37	3	33	3
10	61	6	48	4

Table 3.10: Set N6 - Poisson outcomes; nested setting.
Counts of BIC selections by model order, $N = 100$.

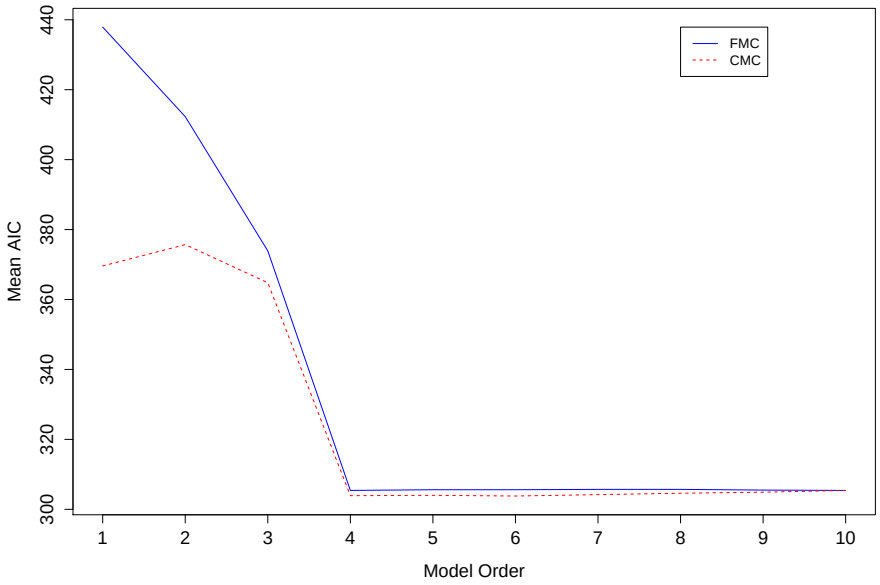

Figure 3.7: Set N7 - Gamma outcomes; nested setting.
Means of AIC by model order, $N = 100$.

Model Order	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
1	187	0	205	0
2	0	0	0	0
3	0	0	0	0
4	343	362	479	538
5	107	110	73	83
6	80	109	78	82
7	61	76	33	63
8	52	96	30	64
9	72	111	47	80
10	98	136	55	90

Table 3.11: Set N7 - Gamma outcomes; nested setting.
Counts of AIC selections by model order, $N = 100$.

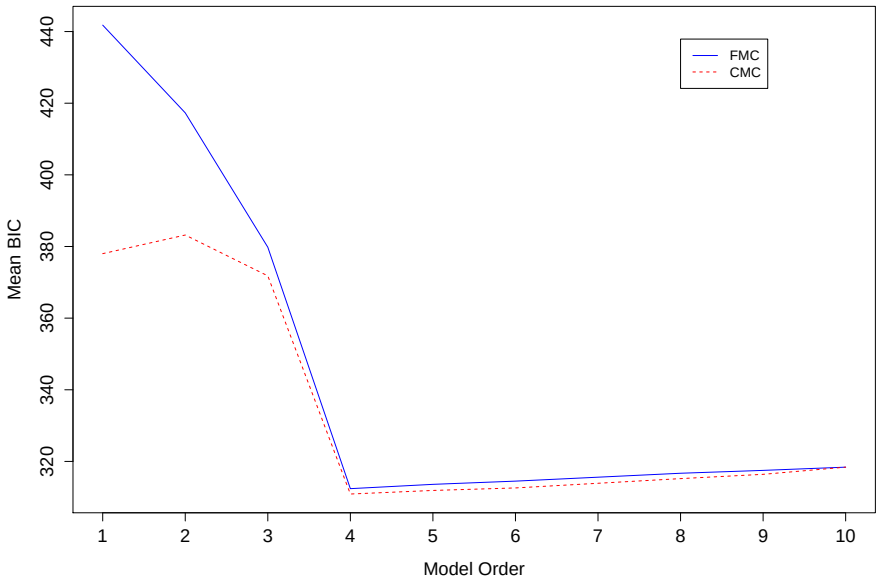

Figure 3.8: Set N8 - Gamma outcomes; nested setting.
Means of BIC by model order, $N = 100$.

Model Order	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
1	141	0	159	0
2	0	0	0	0
3	0	0	0	0
4	559	594	662	742
5	105	107	58	72
6	72	81	43	59
7	35	53	21	38
8	25	51	21	29
9	22	56	15	31
10	41	58	21	29

Table 3.12: Set N8 - Gamma outcomes; nested setting.
Counts of BIC selections by model order, $N = 100$.

For every configuration of the factorial design, the frequency of correctly specified model selections is greater for the FMC criteria than for the CMC criteria. For the FMC criteria, the mean patterns for the selection criteria under the binomial, Poisson, and gamma distributions are similar, exhibiting strong protection from underfitting and modest protection for overfitting, and minimum means corresponding to the generating model order. The CMC criteria for these three distributions also exhibit similar mean patterns. However, the protection from underfitting, particularly for models of order one, is not as strong, and the mean patterns for underfitted models are often paradoxical.

The Bernoulli distribution produces noticeably contrasting results for the CMC and FMC criteria. The FMC criteria tend to favor correctly specified and overspecified models while the CMC criteria favor underspecified models. The figures corroborate the results seen in the model selection counts. The means for the FMC criteria decrease as model order increases, indicating that the selection procedure protects from underfitting. The FMC criteria decrease substantially up to the true model order, where we see an "elbow", then decrease at a lower rate for larger orders. However, the means for the CMC criteria do not behave in this way. Over all model orders, the means for the CMC criteria increase as model order increases, leaving the criteria vulnerable to underfitting.

The binomial ($n = 10$) distribution models selection counts favor the FMC criteria, which rarely select an underfitted model. The figures show similar patterns for the CMC and FMC means curves, with greater protection from underfitting for the FMC criteria.

The Poisson distribution counts show that the CMC criteria commonly choose models of order one. The FMC criterion selection counts show modest protection from underfitting, but not to the degree seen with the binomial distribution. The figures illustrate some interesting differences between the CMC and FMC criteria.

The FMC criterion means follow a similar pattern as those from the binomial distribution, reflecting protection from underfitting. The CMC criterion means appear to also exhibit good protection from underfitting, except for models of order one. The means increase substantially for CMC from model order four to three, but then decrease for orders two and one.

The gamma distribution model selection counts exhibit no underfitting for the FMC criteria. The same is true for the CMC criteria except for model order one, which is selected with a concerning level of frequency. The figures show a similar pattern for the CMC and FMC criterion means as that which was observed with the Poisson distribution, but the CMC means achieve a maximum at model order two.

3.4.2 All Subsets Setting

The second part of the simulation study involves an all subsets modeling setting. Here, we construct data sets with 6 fixed effect predictor variables

$$X_{1ij}, X_{2ij}, ..., X_{6ij} \overset{iid}{\sim} \mathrm{N}(0, 1),$$

a random effect predictor

$$\gamma_i \overset{iid}{\sim} N(0, 1),$$

and outcome Y_{ij} that is generated as seen in Table 3.3 in the nested setting subsection.

With the generated data sets, we fit all possible subsets of fixed effect configurations, leading to $2^6 - 1 = 63$ models since the null model is not included. The all subsets setting allows us to compare the criteria for models that are underspecified (candidate predictor set is a proper subset of the set of true predictors), correctly specified (candidate predictor set is exactly the same as the set of true predictors), overspecified (true predictor set is a proper subset of the candidate predictor set), and mixed misspecified (candidate predictor set is comprised of some but not all of

the predictors in the true set, as well as some predictors not in the true set). Again, the simulation study is designed as a factorial experiment, where the factors are the distribution, selection criterion, and the modeling method for the construction of the pseudo-data. As with the investigative simulation and the previous simulation sets, in addition to selecting models with a minimum AIC or BIC, we will also select the most parsimonious model within two units of the minimum AIC and BIC.

Table 3.13 lists the ID for each simulation set, along with the associated levels of the factors. The figures corresponding to sets AS1 - AS8 feature two colored sets of points; one for each modeling construction method. The AIC and BIC means of the 1000 replications are calculated for each model specification and plotted, illustrating the general behavior of the criteria under each construction method. The minimum mean CMC and FMC model criteria are indicated with a horizontal line. The tables corresponding to sets AS1 - AS8 feature model specification selection counts by each construction method.

Set ID	Distribution	Criterion
AS1	Bernoulli	AIC
AS2	Bernoulli	BIC
AS3	Binomial (n=10)	AIC
AS4	Binomial (n=10)	BIC
AS5	Poisson	AIC
AS6	Poisson	BIC
AS7	Gamma	AIC
AS8	Gamma	BIC

Table 3.13: Model factor levels for all subsets simulation setting.

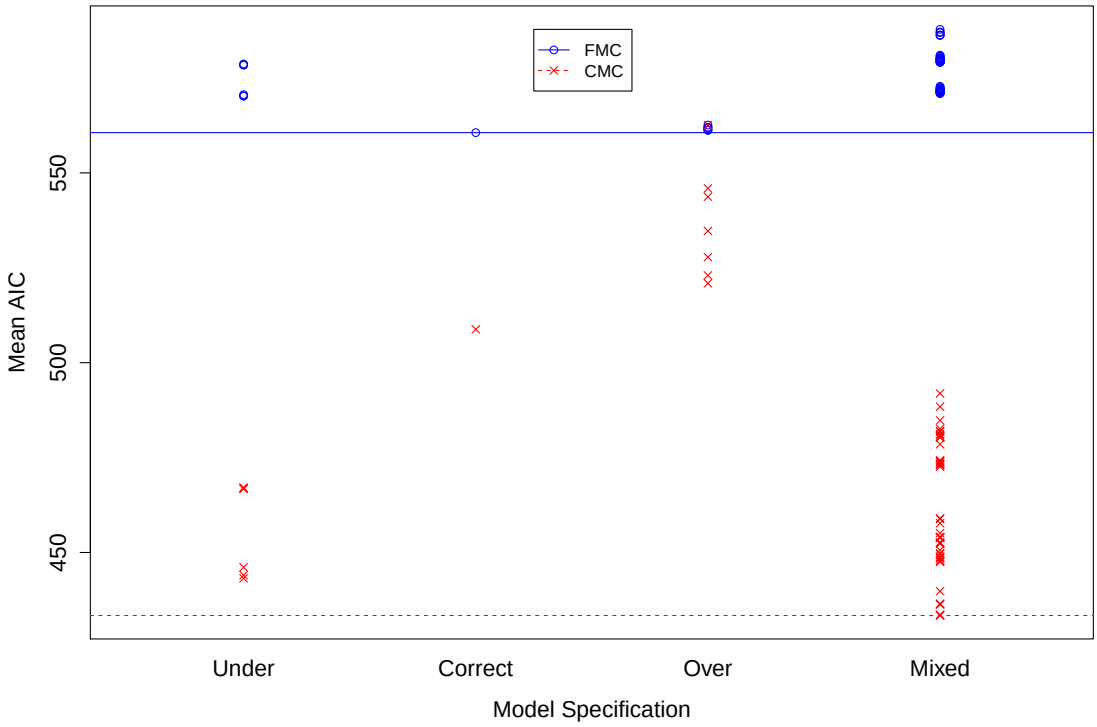

Figure 3.9: Set AS1 - Bernoulli outcomes; all subsets setting.
Means of AIC by model specification, $N = 100$.

Model Specification	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
Under	187	160	458	322
Correct	28	317	31	433
Over	36	377	17	162
Mixed	749	146	494	83

Table 3.14: Set AS1 - Bernoulli outcomes; all subsets setting.
Counts of AIC selections by model specification, $N = 100$.

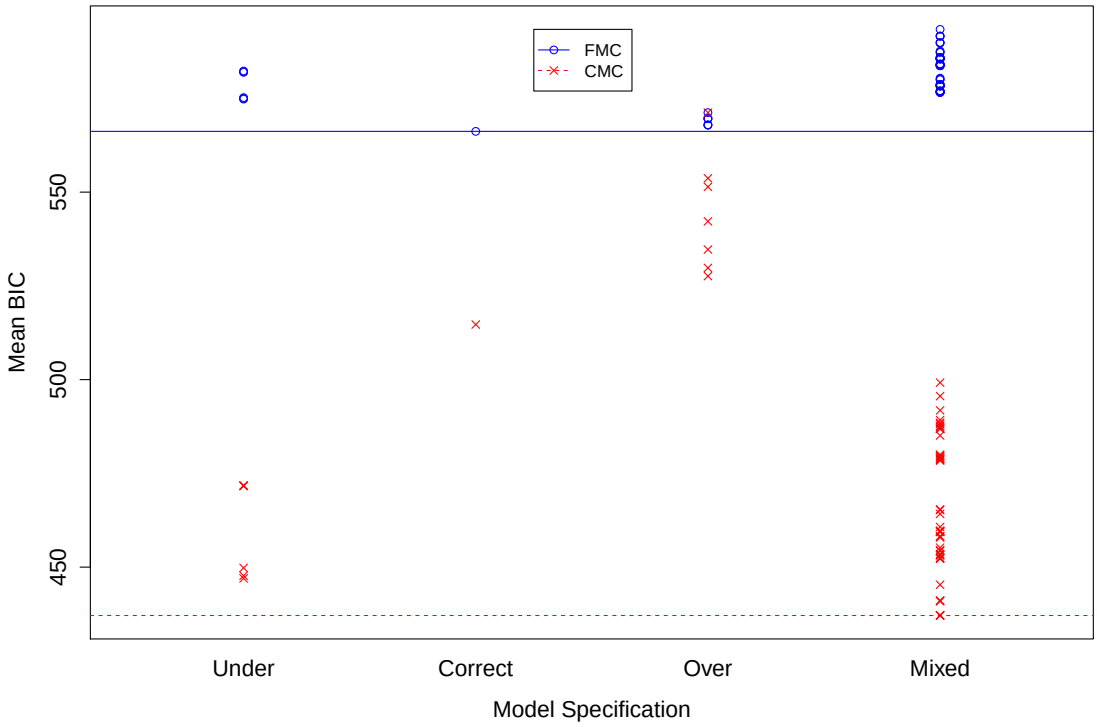

Figure 3.10: Set AS2 - Bernoulli outcomes; all subsets setting.
Means of BIC by model specification, $N = 100$.

Model Specification	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
Under	192	235	473	401
Correct	25	412	21	426
Over	16	221	9	99
Mixed	767	132	497	74

Table 3.15: Set AS2 - Bernoulli outcomes; all subsets setting.
Counts of BIC selections by model specification, $N = 100$.

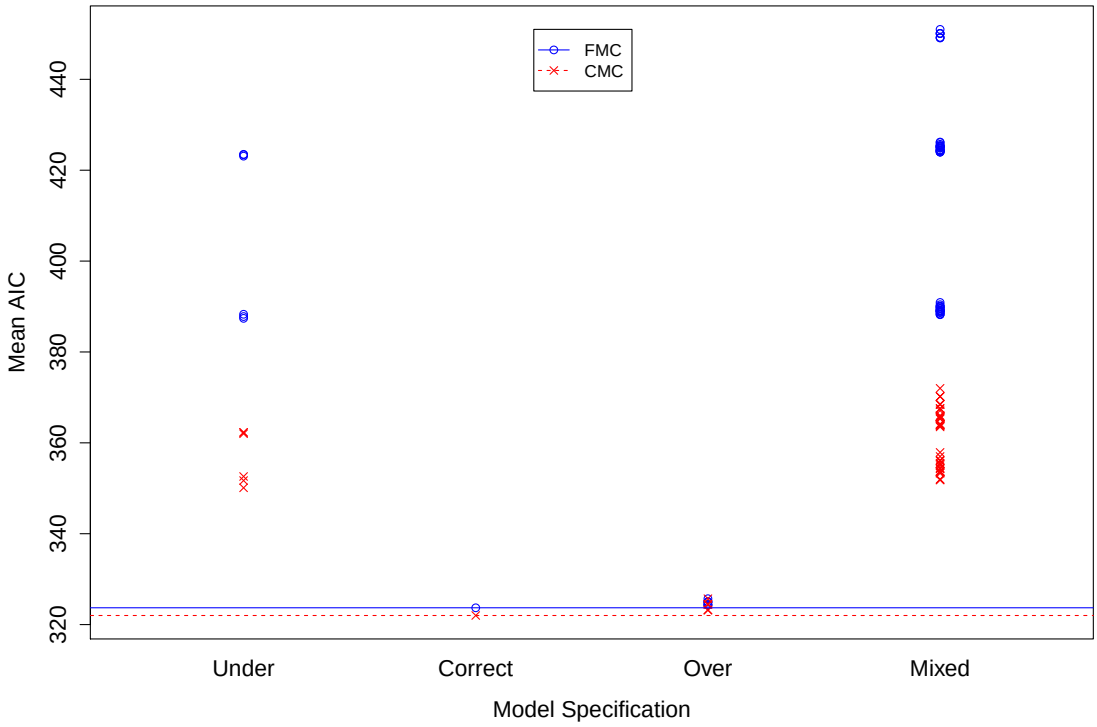

Figure 3.11: Set AS3 - Binomial ($n = 10$) outcomes; all subsets setting. Means of AIC by model specification, $N = 100$.

Model Specification	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
Under	106	1	147	4
Correct	343	472	518	755
Over	475	527	293	241
Mixed	76	0	42	0

Table 3.16: Set AS3 - Binomial ($n = 10$) outcomes; all subsets setting. Counts of AIC selections by model specification, $N = 100$.

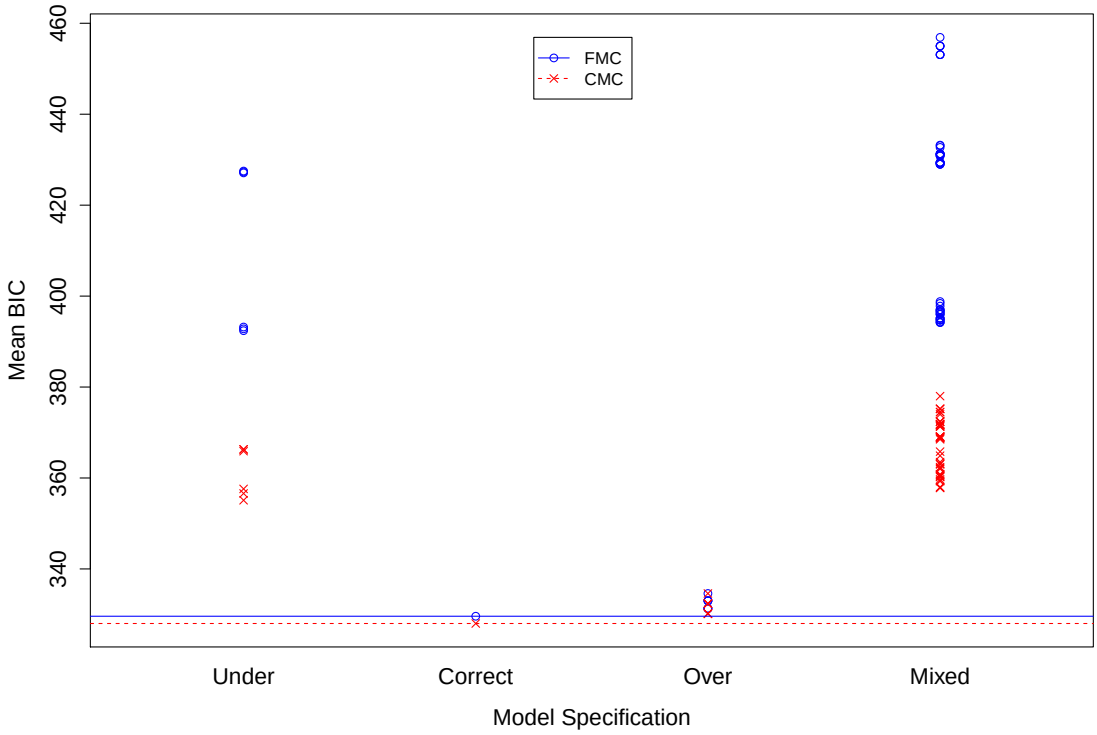

Figure 3.12: Set AS4 - Binomial ($n = 10$) outcomes; all subsets setting. Means of BIC by model specification, $N = 100$.

Model Specification	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
Under	136	3	175	5
Correct	443	646	576	842
Over	371	351	223	152
Mixed	50	0	26	1

Table 3.17: Set AS4 - Binomial ($n = 10$) outcomes; all subsets setting. Counts of BIC selections by model specification, $N = 100$.

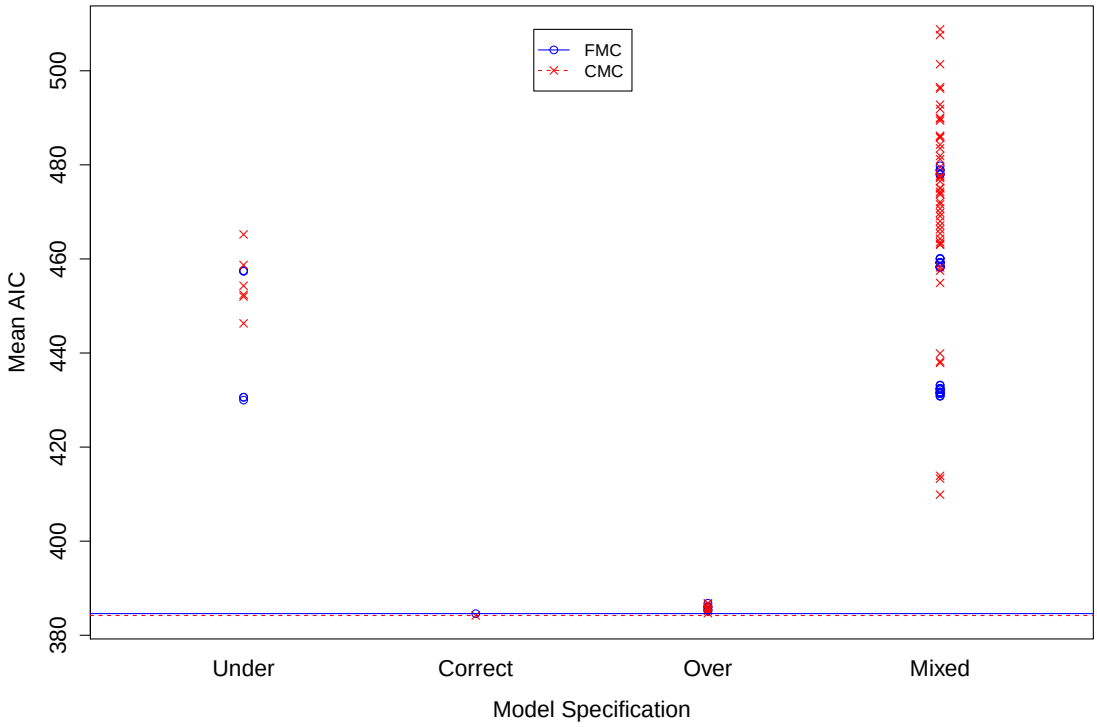

Figure 3.13: Set AS5 - Poisson outcomes; all subsets setting.
Means of AIC by model specification, $N = 100$.

Model Specification	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
Under	184	35	241	73
Correct	133	474	233	703
Over	399	462	290	211
Mixed	284	29	236	13

Table 3.18: Set AS5 - Poisson outcomes; all subsets setting.
Counts of AIC selections by model specification, $N = 100$.

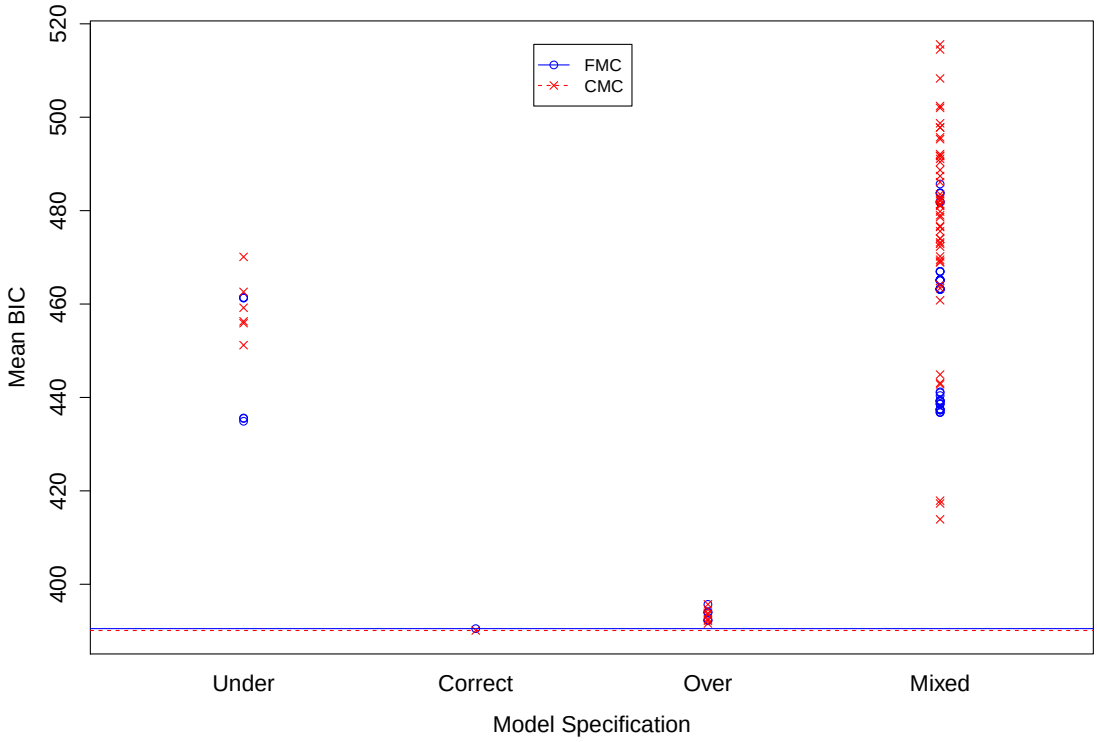

Figure 3.14: Set AS6 - Poisson outcomes; all subsets setting.
Means of BIC by model specification, $N = 100$.

Model Specification	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
Under	224	47	271	85
Correct	184	623	265	766
Over	337	301	242	136
Mixed	255	29	222	13

Table 3.19: Set AS6 - Poisson outcomes; all subsets setting.
Counts of BIC selections by model specification, $N = 100$.

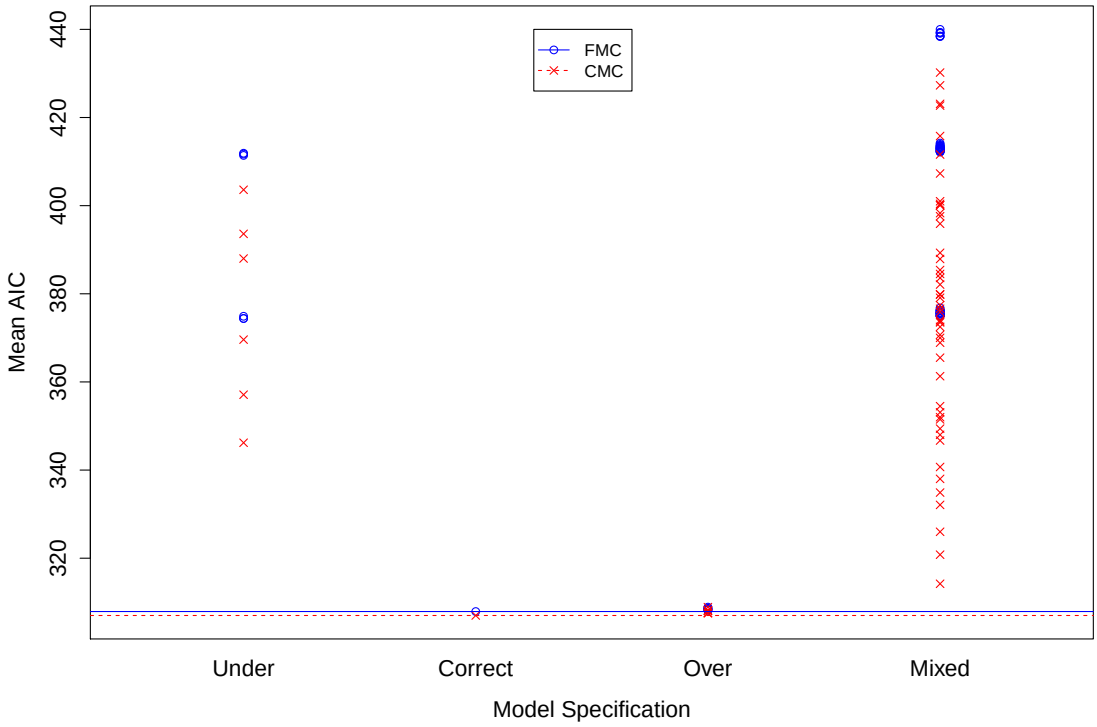

Figure 3.15: Set AS7 - Gamma outcomes; all subsets setting.
Means of AIC by model specification, $N = 100$.

Model Specification	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
Under	129	0	193	0
Correct	296	378	468	669
Over	575	622	339	331
Mixed	0	0	0	0

Table 3.20: Set AS7 - Gamma outcomes; all subsets setting.
Counts of AIC selections by model specification, $N = 100$.

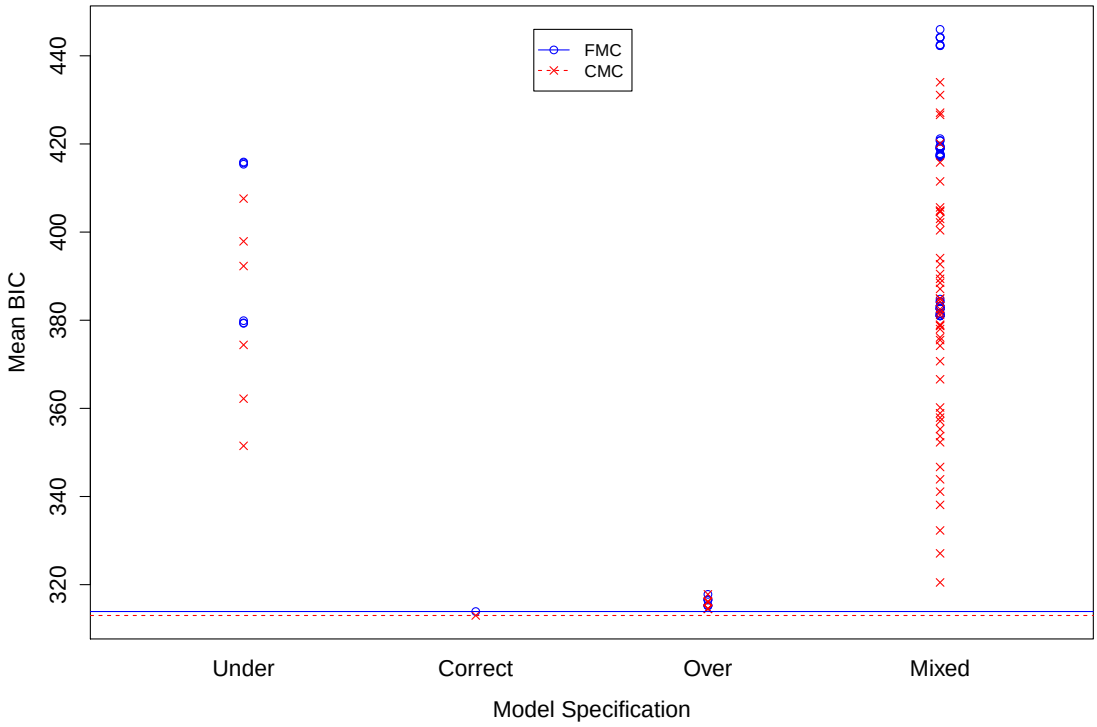

Figure 3.16: Set AS8 - Gamma outcomes; all subsets setting.
Means of BIC by model specification, $N = 100$.

Model Specification	Minimum CMC	Minimum FMC	Parsimonious w/i 2 CMC	Parsimonious w/i 2 FMC
Under	166	0	227	0
Correct	396	556	535	771
Over	438	444	238	229
Mixed	0	0	0	0

Table 3.21: Set AS8 - Gamma outcomes; all subsets setting.
Counts of BIC selections by model specification, $N = 100$.

Just as we observed in the nested setting, for every configuration of the factorial design, the frequency of correctly specified model selections is greater for the FMC criteria than for the CMC criteria. The mean patterns for the selection criteria under the binomial, Poisson, and gamma distributions are similar, exhibiting strong protection from choosing underspecified and mixed misspecified models, modest protection for choosing overspecified models, and minimum means corresponding to the correctly specified models.

Once again, the Bernoulli distribution produces noticeably contrasting results for the CMC and FMC criteria. The FMC criteria choose the generating model specification much more frequently than the CMC criteria, which mostly choose mixed misspecified models. The figures support the results reflected in the model selection counts. The means for the FMC criteria favor correctly specified and overspecified models, while the means for the CMC criteria favor underspecified and mixed misspecified models.

The binomial ($n = 10$) distribution model selection counts support the use of FMC criteria, which rarely choose an underspecified or mixed misspecified model. The figures exhibit similar patterns for the CMC and FMC means, reflecting more protection from choosing underspecified and mixed misspecified models for the FMC criteria.

The Poisson distribution counts show that the CMC criteria still tend to choose underspecified and mixed misspecified models more often than the FMC criteria. The FMC criterion selection counts reflect modest protection from choosing underspecified models, but not to the degree seen with the binomial distribution. The figures illustrate good protection from choosing underspecified and mixed misspecified models for both the CMC and FMC criteria, with higher means for the CMC, just as we had observed in the nested setting.

The gamma distribution model counts show no selections for underspecified or mixed misspecified models for the FMC criteria and no selections for mixed misspecified models for the CMC criteria. The figures indicate a similar pattern for the CMC and FMC criterion means as that which was seen with the Poisson distribution.

3.5 Application

Now that we have characterized how criteria based on the CMC and FMC approaches behave in a simulated setting, we can apply them in a modeling application and examine the results.

Bullying in schools is a pervasive and long-standing problem that has recently gained the attention of the media and public nationwide. Over the past several years, many efforts have been made to curb this behavior in students, including the passing of legislation to more specifically define bullying and to specify punishments that go along with offenses. In 2005, an anti-bullying law was passed in the state of Iowa in an effort to reduce the number of student-reported incidents of bullying.

The Iowa Youth Survey (IYS) is given statewide to sixth, eighth, and eleventh grade students in Iowa Public Schools every two to three years. The IYS covers an array of topics that are part of the everyday lives of students both inside and outside of school. A section of the survey contains questions regarding bullying, which is classified as four types: psychological (`psych`; experiences involving being excluded, ignored, or made the subject of lies/rumors), verbal (`verb`; experiences involving being made fun of or teased), physical (`phys`; experiences involving being hit, kicked, or shoved) and cyber (`cyber`; experiences involving being sent threatening or hurtful messages via email, social media, etc.).

The IYS data set considered for this application contains approximately 253,000 observations and 6 attributes, and includes surveys taken in 2005, 2008, and 2010.

The binary outcomes of interest are significant exposure to each of the four bullying types, where bullying is defined as three or more incidents of victimization in the month prior to taking the survey. The candidate fixed effect predictor variables include survey year (`surveyyear`; categorical in [2005, 2008, 2010]), grade of the student (`grade`; categorical in $[6^{th}, 8^{th}, 11^{th}]$), gender of the student (`gender`; 0 = female, 1 = male), ethnicity of the student (`ethnicity`; 0 = white, 1 = African American, 2 = Native American, 3 = Asian, 4 = Hispanic, 5 = other/mixed), living situation of the student (`livingsituation`; 0 = with parent(s), 1 = with grandparent(s), 2 = with foster parent(s), 3 = in shelter care, 4 = group home, 5 = independent living, 6 = other), and frequency of teacher intervention in bullying incidents (`intervene`; 0 = almost never, 1 = once in a while, 2 = sometimes, 3 = often or almost always). Following the work of Ramirez et al. (2015), the ordinal variable `intervene` is treated as quantitative as opposed to qualitative. A random effect variable, school district (`schooldist`; categorical containing 412 unique values), will be included in all of the models.

Since we have 6 potential fixed effect predictor variables, the all subsets setting will consider a candidate collection comprised of $2^6 - 1 = 31$ models. Once each model is fitted using both the CMC and FMC techniques, AIC and BIC will be calculated and used to ascertain the models favored by each criterion. The "best" model will be determined in the same way as in the simulation: minimum AIC (minAIC), most parsimonious model within two units of minimum AIC (minAIC2), minimum BIC (minBIC), and most parsimonious model within two units of minimum BIC (minBIC2). The selected models according to the CMC technique for the four criteria are provided in Table 3.22. The selected models according to the FMC technique for the four criteria are provided in Table 3.23.

There is a noticeable difference in model selections between the CMC and FMC techniques. Under the CMC, the model selections for each outcome and criterion

Outcome	Criterion	Selected Variables
Psych	minAIC	`surveyyear`
Psych	minAIC2	`surveyyear`
Psych	minBIC	`surveyyear`
Psych	minBIC2	`surveyyear`
Verb	minAIC	`gender`
Verb	minAIC2	`gender`
Verb	minBIC	`gender`
Verb	minBIC2	`gender`
Phys	minAIC	`surveyyear`
Phys	minAIC2	`surveyyear`
Phys	minBIC	`surveyyear`
Phys	minBIC2	`surveyyear`
Cyber	minAIC	`ethnicity`
Cyber	minAIC2	`ethnicity`
Cyber	minBIC	`ethnicity`
Cyber	minBIC2	`ethnicity`

Table 3.22: CMC model selections by criterion.

pairing are of order one. The four criteria select the same model for each different outcome. Under the FMC, the model selections for each outcome and criterion pairing are the full model with the exception of modeling verbal bullying and using the minBIC2 criterion, where the chosen model includes all candidate predictors except gender. These results strongly parallel the results from the investigative simulation presented earlier in the chapter. Specifically, although we are inclined to believe that the FMC criteria may be selecting overspecified models, the results are favorable compared to those for the CMC method, which tends to prefer simple models (often of order one) for outcomes following the Bernoulli distribution.

These selections may still raise questions as to which technique is truly doing

Outcome	Criterion	Selected Variables
Psych	minAIC	surveyyear grade gender ethnicity livingsituation intervene
Psych	minAIC2	surveyyear grade gender ethnicity livingsituation intervene
Psych	minBIC	surveyyear grade gender ethnicity livingsituation intervene
Psych	minBIC2	surveyyear grade gender ethnicity livingsituation intervene
Verb	minAIC	surveyyear grade gender ethnicity livingsituation intervene
Verb	minAIC2	surveyyear grade gender ethnicity livingsituation intervene
Verb	minBIC	surveyyear grade gender ethnicity livingsituation intervene
Verb	minBIC2	surveyyear grade ethnicity livingsituation intervene
Phys	minAIC	surveyyear grade gender ethnicity livingsituation intervene
Phys	minAIC2	surveyyear grade gender ethnicity livingsituation intervene
Phys	minBIC	surveyyear grade gender ethnicity livingsituation intervene
Phys	minBIC2	surveyyear grade gender ethnicity livingsituation intervene
Cyber	minAIC	surveyyear grade gender ethnicity livingsituation intervene
Cyber	minAIC2	surveyyear grade gender ethnicity livingsituation intervene
Cyber	minBIC	surveyyear grade gender ethnicity livingsituation intervene
Cyber	minBIC2	surveyyear grade gender ethnicity livingsituation intervene

Table 3.23: FMC model selections by criterion.

better at representing the data and identifying optimal models. As a supplementary comparison, the models have been re-fitted under the GLM framework, omitting the random effects. Although we are intentionally choosing to ignore the clustering adjustments that these effects provide, based on previous work and additional analyses not reported here, the random effects exhibit only a small degree of variation between school districts, indicating that we are not losing pertinent information through their exclusion. The selected models for the four criteria are provided in Table 3.24. Of the 16 model selections considered, 13 match the previous results for the FMC technique. The remaining three models only differ in selection by the same single variable. This agreement indicates that the FMC criterion selections are quite appropriate, making them favorable to the CMC criterion selections. This

finding further supports the conclusions drawn from the simulation study.

Outcome	Criterion	Selected Variables
Psych	minAIC	surveyyear grade gender ethnicity livingsituation intervene
Psych	minAIC2	surveyyear grade gender ethnicity livingsituation intervene
Psych	minBIC	surveyyear grade gender ethnicity livingsituation intervene
Psych	minBIC2	surveyyear grade gender ethnicity livingsituation intervene
Verb	minAIC	surveyyear grade ethnicity livingsituation intervene
Verb	minAIC2	surveyyear grade ethnicity livingsituation intervene
Verb	minBIC	surveyyear grade ethnicity livingsituation intervene
Verb	minBIC2	surveyyear grade ethnicity livingsituation intervene
Phys	minAIC	surveyyear grade gender ethnicity livingsituation intervene
Phys	minAIC2	surveyyear grade gender ethnicity livingsituation intervene
Phys	minBIC	surveyyear grade gender ethnicity livingsituation intervene
Phys	minBIC2	surveyyear grade gender ethnicity livingsituation intervene
Cyber	minAIC	surveyyear grade gender ethnicity livingsituation intervene
Cyber	minAIC2	surveyyear grade gender ethnicity livingsituation intervene
Cyber	minBIC	surveyyear grade gender ethnicity livingsituation intervene
Cyber	minBIC2	surveyyear grade gender ethnicity livingsituation intervene

Table 3.24: No random effects model selections by criterion.

3.6 Summary Conclusion

This chapter has introduced a new technique for constructing GLMM criteria used for model selection in the pseudo-likelihood framework. In SAS, the technique can be implemented using the GLIMMIX and MIXED procedures. Compared to the default criterion construction with the GLIMMIX procedure, the new technique shows considerable improvement in model selection as illustrated in the simulation portion of this chapter. Under the assumption that the candidate collection contains the generating model, the criteria based on the new technique select the generating model much more frequently than the criteria based on the GLIMMIX default

construction.

A SAS macro has been written by the author that fits models with all possible subsets of predictor variables and selects an appropriate model based on criteria constructed using the new technique. This macro is generalizable for any set of p variables, providing the $2^p - 1$ fitted candidate models based on predictor subsets.

Future work involves comparison of the CMC and FMC criteria for additional distributions not considered in this chapter. Additionally, we will further characterize the behavior of the CMC and FMC criteria based on GLMMs with more complex random effects than merely the random intercept.

REFERENCES

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in: B. N. Petrov and F. Csaki, (Eds.), *2nd International Symposium on Information Theory* (Akademia Kiado, Budapest), 267–281.

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**, 716–723.

Bengtsson, T. and Cavanaugh, J. E. (2006), An improved Akaike information criterion for state-space model selection, *Computational Statistics and Data Analysis* **50**, 2635–2654.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. and White, J. S. S. (2009), Generalized linear mixed models: a practical guide for ecology and evolution, *Trends in ecology & evolution* **24(3)**, 127–135.

Bozdogan, H. (1987), Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions, *Psychometrika* **52**, 345–370.

Burnham, K. P. and Anderson, D. R. (2002), *Model Selection and Multimodel Inference* (Springer-Verlag, New York).

Cavanaugh, J. E. (1999), A large–sample model selection criterion based on Kullback's symmetric divergence, *Statistics & Probability Letters* **44**, 333–344.

Cavanaugh, J. E. and Shumway, R. H. (1997), A bootstrap variant of AIC for state-space model selection, *Statistica Sinica* **7**, 473–496.

Cook, N. R. (2007), Use and misure of the receiver operating characteristic curve in risk prediction, *Circulation* **115(7)**, 928–935.

Davies, S. L., Neath, A. A. and Cavanaugh J. E. (2005), Cross validation model selection criteria for linear regression based on the Kullback-Leibler discrepancy, *Statistical Methodology* **2**, 249–266.

Dean, C. B. and Nielsen (2007), Generalized linear mixed models: a review and some extensions, *Lifetime Data Analysis* **13**, 497–512.

Efron, B. (1983), Estimating the error rate of a prediction rule: Improvement on cross-validation, *Journal of the American Statistical Association* **78**, 316–331.

Efron, B. (1986), How biased is the apparent error rate of a prediction rule?, *Journal of the American Statistical Association* **81**, 461–470.

Gonen, M. and Heller, G. (2005), Concordance probability and discriminatory power in proportional hazards regression, *Biometrika* **92(4)**, 965–970.

Hanley, J. A. and McNeil, B. J. (1982), The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* **143**, 29–36.

Haughton, D. M. A. (1988), On the choice of a model to fit data from an exponential family, *The Annals of Statistics* **6**, 342–355.

Heagerty, P. J. and Zheng, Y. (2005), Survival model predictive accuracy and ROC curves, *Biometrics* **61**, 92–105.

Hilden, J., Habbema, J.D. and Bjerregaard, B. (1978), The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities, *Methods of Information in Medicine* **17(4)**, 227–237.

Hosmer, D. W. and Lemeshow, S. (2000), *Applied Logistic Regression* (John Wiley & Sons, Inc., New York).

Hosmer, D. W. and Lemeshow, S. (1980), A goodness-of-fit test for the multiple logistic regression model, *Communications in Statistics* **A10**, 1043–1069.

Hurvich, C. M., Shumway, R. H. and Tsai, C. L. (1990), Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples, *Biometrika* **77**, 709–719.

Hurvich, C. M. and Tsai, C. L. (1989), Regression and time series model selection in small samples, *Biometrika* **76**, 297–307.

Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997), Bootstrapping log likelihood and EIC, an extension of AIC, *Annals of the Institute of Statistical Mathematics* **49**, 411–434.

Kashyap, R. L. (1982), Optimal choice of AR and MA parts in autoregressive moving-average models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **4**, 99–104.

Kass, R. E. and Raftery, A. E. (1995), Bayes factors, *Journal of the American Statistical Association* **90**, 773–795.

Kullback, S. (1968), *Information Theory and Statistics* (Dover, New York).

Kullback, S. and Leibler, R. A. (1951), On information and sufficiency, *The Annals of Mathematical Statistics* **22**, 79–86.

Lemeshow, S. and Hosmer, D. W. (1982), A review of goodness of fit statistics for use in the development of logistic regression models, *American journal of epidemiology* **115(1)**, 92–106.

Leonard, T. (1982), Comments on 'A simple predictive density function,' by M LeJeune and GD Faulkenberry, *Journal of the American Statistical Association* **77**, 657–658.

Linhart, H. and Zucchini, W. (1986), *Model Selection* (Wiley, New York).

Metz, C. E. (1986), ROC methodology in radiologic imaging, *Investigative Radiology* **21**, 720–733.

Metz, C. E. (1989), Some practical issues of experimental design and data analysis in radiologic ROC studies, *Investigative Radiology* **24**, 234–245.

Pan, W. (2001), Akaike's information criterion in generalized estimating equations, *Biometrics* **57**, 120–125.

Pencina, M. J., D'Agostino, R. B. Sr., D'Agostino, R. B. Jr. and Vasan, R. S. (2008), Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond, *Statistics in Medicine* **27(2)**, 157–172.

Ramirez, M., Ten Eyck, P., Peek-Asa, C., Onwuachi-Willig, A. and Cavanaugh, J. (2015), Effectiveness of Iowas anti-bullying law in preventing bullying, (Submitted).

Royston, P. and Altman, D. G. (2010), Visualizing and assessing discrimination in the logistic regression model, *Statistics in Medicine* **29**, 2508–2520.

Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics* **6**, 461–464.

Shibata, R. (1980), Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *The Annals of Statistics* **8(1)**, 147–164.

Shibata, R. (1981), An optimal selection of regression variables, *Biometrika* **68**, 45–54.

Shibata, R. (1997), Bootstrap estimate of Kullback-Leibler information for model selection, *Statistica Sinica* **7**, 375–394.

Steyerberg E. W., Vickers A. J., Cook N. R., et al. (2010), Assessing the performance of prediction models: a framework for some traditional and novel measures, *Epidemiology* **21(1)**, 128–138.

Stone, M. (1977), An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society, Series B* **39**, 44–47.

Stone, M. (1979), Comments on model selection criteria of Akaike and Schwarz, *Journal of the Royal Statistical Society, Series B* **41**, 276–278.

Sugiura, N. (1978), Further analysis of the data by Akaike's information criterion and the finite corrections, *Communications in Statistics* **A7**, 13–26.

Takeuchi, K. (1976), Distribution of information statistics and criteria for adequacy of models, *Mathematical Sciences* **153**, 12–18 (in Japanese).

Zhou, X. H., Obuchowski, N. A. and McClish, D. K. (2002), *Statistical Methods in Diagnostic Medicine* (John Wiley & Sons, Inc., New York).