Fall 2009

Model selection criteria in the presence of missing data based on the Kullback-Leibler discrepancy

JonDavid Sparks
University of Iowa

MODEL SELECTION CRITERIA IN THE PRESENCE OF MISSING DATA BASED ON THE KULLBACK-LEIBLER DISCREPANCY

by

JonDavid Sparks

An Abstract

Of 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 2009

Thesis Supervisor: Professor Joseph Cavanaugh

ABSTRACT

An important challenge in statistical modeling involves determining an appropriate structural form for a model to be used in making inferences and predictions. Missing data is a very common occurrence in most research settings and can easily complicate the model selection problem. Many useful procedures have been developed to estimate parameters and standard errors in the presence of missing data; however, few methods exist for determining the actual structural form of a model when the data is incomplete.

In this dissertation, we propose model selection criteria based on the Kullback-Leibler discrepancy that can be used in the presence of missing data. The criteria are developed by accounting for missing data using principles related to the expectation maximization (EM) algorithm and bootstrap methods. We formulate the criteria for three specific modeling frameworks: for the normal multivariate linear regression model, a generalized linear model, and a normal longitudinal regression model. In each framework, a simulation study is presented to investigate the performance of the criteria relative to their traditional counterparts. We consider a setting where the missingness is confined to the outcome, and also a setting where the missingness may occur in the outcome and/or the covariates. The results from the simulation studies indicate that our criteria provide better protection against underfitting than their traditional analogues.

We outline the implementation of our methodology for a general discrepancy measure. An application is presented where the proposed criteria are utilized in a study that evaluates the driving performance of individuals with Parkinson's disease under low contrast (fog) conditions in a driving simulator.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

MODEL SELECTION CRITERIA IN THE PRESENCE OF MISSING DATA BASED ON THE KULLBACK-LEIBLER DISCREPANCY

by

JonDavid Sparks

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 2009

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

JonDavid Sparks

has been approved by the Examining Committee
for the thesis requirement for the Doctor of
Philosophy degree in Biostatistics at the December 2009
graduation.

Thesis Committee: _____
Joseph Cavanaugh, Thesis Supervisor

_____
Jeffrey Dawson

_____
Dawei Liu

_____
Jacob Oleson

_____
Philip Polgreen

ACKNOWLEDGMENTS

I would like to sincerely thank my advisor Dr. Joseph Cavanaugh for collaborating with me on this dissertation. It was always a pleasure to meet with Joe during our marathon meetings, and I am very grateful for the time and energy that he committed to this project. Joe was an excellent advisor and was able to bring out the best of my talents and abilities.

I am also very appreciative of the other members of my dissertation committee: Dr. Jeffrey Dawson, Dr. Dawei Liu, Dr. Jacob Oleson and Dr. Philip Polgreen. Their contributions and insights greatly improved the manuscript. In addition, I kindly acknowledge Dr. Ergun Uc, for allowing me use of his data for the application.

I would also like to express gratitude to the faculty and students at the University of Iowa. I have been provided with an excellent education, along with many opportunities to be involved in practical applications, that will prepare me for my career.

Most importantly, I am very grateful for my wife Sarah, for her love and assistance during these challenging academic years. Her encouragement and patience has always been appreciated. I am thankful for my son Rocco who always brightens my days and keeps life in perspective. I would also like to thank my parents and family for their examples and support, and for teaching me the most important things in life that have allowed me to succeed.

ABSTRACT

An important challenge in statistical modeling involves determining an appropriate structural form for a model to be used in making inferences and predictions. Missing data is a very common occurrence in most research settings and can easily complicate the model selection problem. Many useful procedures have been developed to estimate parameters and standard errors in the presence of missing data; however, few methods exist for determining the actual structural form of a model when the data is incomplete.

In this dissertation, we propose model selection criteria based on the Kullback-Leibler discrepancy that can be used in the presence of missing data. The criteria are developed by accounting for missing data using principles related to the expectation maximization (EM) algorithm and bootstrap methods. We formulate the criteria for three specific modeling frameworks: for the normal multivariate linear regression model, a generalized linear model, and a normal longitudinal regression model. In each framework, a simulation study is presented to investigate the performance of the criteria relative to their traditional counterparts. We consider a setting where the missingness is confined to the outcome, and also a setting where the missingness may occur in the outcome and/or the covariates. The results from the simulation studies indicate that our criteria provide better protection against underfitting than their traditional analogues.

We outline the implementation of our methodology for a general discrepancy measure. An application is presented where the proposed criteria are utilized in a study that evaluates the driving performance of individuals with Parkinson's disease under low contrast (fog) conditions in a driving simulator.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

Model selection is one of the key areas of investigation in the development and application of statistical methodology. An important component of any statistical modeling problem consists of determining an appropriate structural form for the model. Improper model specification may substantially impact both the estimators of the model parameters and the predictors of the response variable. Underspecification refers to a situation where a model does not include all of the necessary components and may lead to results which are severely biased. Overspecification refers to a situation where a model includes the necessary components, in addition to some possibly spurious components, and may lead to results with unnecessarily high variability.

The determination of a suitable structure can often be facilitated by the use of a model selection criterion. A group of candidate models is often postulated that could potentially describe the phenomenon under study. A model selection criterion can be used to assign scores to each of the fitted candidate models in order to assist the analyst in selecting a model for inference.

The model selection problem is easily complicated by the presence of missing data, which is a common occurrence in many research settings. If a structural form for a model is selected, methods such as imputation approaches and the expectation maximization (EM) algorithm have been developed to estimate parameters and standard errors when the data is incomplete. However, choosing the structural

form of a model may be very difficult in the presence of missing data.

1.1 Objectives of Dissertation

A discrepancy-based model selection criterion is often formulated by constructing an approximately unbiased estimator of an expected overall discrepancy, a measure that gauges the separation between the true model and a fitted candidate model. The expected discrepancy reflects how well, on average, the fitted candidate model predicts "new" data generated under the true model.

The general form of a model selection criterion consists of the summation of a goodness-of-fit term and a penalty term. The natural estimator of the expected discrepancy, the estimated discrepancy, corresponds to the goodness-of-fit term in the selection criterion. The estimated discrepancy reflects how well the fitted candidate model predicts the data used in its own construction. Thus, the estimated discrepancy yields an overly optimistic assessment of how effectively the fitted model predicts new data. It therefore serves as a negatively biased estimator of the expected discrepancy. Correcting for this bias leads to the penalty term of the model selection criterion.

The penalty term approximates the average amount by which the expected discrepancy is underestimated by the estimated discrepancy, a quantity known as the expected optimism. In situations where there is no missing data, classical and computational approaches have been developed to approximate or estimate the expected optimism. Classical approaches often lead to simplistic penalty terms based on the sample size and the dimension of the fitted candidate model. Such approaches generally involve large-sample arguments, restrictive assumptions on the form of the candidate model, or both. Other more computational approaches, such as bootstrapping methods, facilitate the development of flexible and accurate data-based estimators of the expected optimism.

In the presence of missing data, the goodness-of-fit term used in the construction of the model selection criterion may be problematic to compute. Furthermore, the classical and computational approaches to estimate the expected optimism may require modification.

Often times, if one is presented a data set with missing values, a simple yet naive approach would be to simply use the cases that contain no missing data in selecting a statistical model. This is known as listwise deletion or a complete case analysis, and would be the default in many statistical packages such as SAS and R. The model selection criteria based on this paradigm would target what we refer to as the fully observed data discrepancy, which ignores potentially useful information that is contained in the cases that have some missing data.

A more desirable paradigm may be to create model selection criteria based on what we refer to as the complete data discrepancy, an idealized disparity measure constructed by averaging over both observed and unobserved data. Such criteria utilize all of the available observed information in the data set, and use the fitted candidate model to impute values for the unobserved data.

The main objective of this dissertation is to develop model selection criteria that can effectively determine the structural form of a model in the presence of missing data. We address this objective by developing model selection criteria that target the complete data discrepancy. This is accomplished by first accounting for the missing data that occurs in the goodness-of-fit term, which is done by imputing missing values using parameter estimates from the EM algorithm and techniques related to the bootstrap. Secondly, procedures that are analogous to the classical and bootstrap based methods of estimating the expected optimism are provided and justified.

We now review some model selection criteria that are pertinent to our work.

The Akaike (1973) information criterion (AIC) is the most well known and commonly used model selection criterion, and was the first to gain wide spread acceptance as a model selection tool. AIC estimates the expected Kullback-Leibler discrepancy (Kullback, 1968) between the model generating the data and a fitted candidate model. The criterion is justified in a very general framework using asymptotic arguments, and as a result, offers a crude estimate of the expected Kullback-Leibler discrepancy. Much work has been done to improve the estimating properties of AIC. A criterion called corrected AIC (AICc), has been developed in certain modeling frameworks. AICc is designed to provide a better estimate of the expected Kullback-Leibler discrepancy in small-sample settings (Hurvich, Shumway and Tsai, 1990; Hurvich and Tsai, 1993; Bedrick and Tsai, 1994; Hurvich and Tsai, 1995). There also exist bootstrap based variants of AIC. Two such variants are the extended information criterion, EIC (Efron, 1983, 1986; Ishiguro, Sakamoto and Kitagawa, 1997; Konishi and Kitagawa, 1996, 2008) and AICb (Cavanaugh and Shumway, 1997; Shang and Cavanaugh, 2008).

1.2 Review of Relevant Literature

Several approaches based on constructs and parameter estimates from the EM algorithm have been proposed to perform model selection in the presence of missing data. Cavanaugh and Shumway (1998) develop an AIC-type criterion (AICcd), which is related to a criterion introduced by Shimodaira (1994) known as the predictive divergence for indirect observation models (PDIO). Cavanaugh and Shumway use the expected complete data log-likelihood, which serves as the basis for the E-step of the EM algorithm, for the goodness-of-fit term, whereas Shimodaira uses the likelihood of the incomplete data. Seghouane, Bekara and Fleury (2005) use the motivation for AICcd in developing an information criterion based on the Kullback (1968) symmetric divergence (Cavanaugh, 1999). Claeskens and Consentino

(2008) propose variants of AIC and the Takeuchi (1976) information criterion (TIC) that can accommodate models with missing covariate data, based on the EM algorithm and the method of weights (Ibrahim, 1990). Ibrahim, Zhu and Tang (2008) also use output from the EM algorithm to develop versions of AIC and Schwarz's (1978) Bayesian information criterion (BIC). Hens, Aerts and Molenberghs (2006) re-weight the cases that contain no missing data by their inverse selection probabilities to develop a variant of AIC for regression models. Another EM algorithm based model selection approach is given by Bueso, Qian and Angulo (1999).

Other approaches based on Bayesian model selection and multiple imputation have been proposed. Yang, Belin and Boscardin (2005) develop two methods. The first approach is to create multiply imputed data sets and separately apply Bayesian variable selection methods to each data set. The second approach is to combine the imputation and variable selection steps into a single Gibbs sampling process. Celeux, Forbes, Robert and Titterington (2006) propose deviance information criterion (DIC) (Spiegelhalter, Best, Carlin and van der Linde, 2002) constructions for Bayesian model selection. Wood, White and Royston (2008) consider stacking multiply imputed data sets into one large data set and then performing model selection. Heymans, Van Buuren, Knol, van Mechelen and de Vet (2007) investigate the effects of taking bootstrap samples of the multiply imputed data sets, followed by the application of model selection methods on each data set.

1.3 Outline of Dissertation

Chapter 1 provides an introduction to this dissertation. Chapter 2 gives background material that serves as a foundation for the remainder of the dissertation. It includes a description of the model selection framework, along with an overview of some missing data principles and common statistical approaches that are used to account for missing data. The chapter concludes with a brief outline of the

bootstrap.

In Chapter 3, the model selection criteria AIC, AICc, AICb, and EIC are outlined in a general setting. Variants of these criteria that can be used for incomplete data are proposed and developed. The chapter also includes a brief discussion contrasting the newly proposed methodology to techniques in multiple imputation.

In Chapter 4, variants of AIC, AICc, AICb, and EIC are first proposed in the framework of the normal multivariate linear regression model with missing data in the outcome. The criteria are then developed under the setting of missing data in the outcome and/or covariates, by making some additional assumptions and necessary adjustments. Simulation studies are conducted in both settings to evaluate and compare the selection performances and biases of the proposed criteria to those that would commonly be used in practice.

In Chapter 5, variants of AIC, AICb, and EIC are developed in a generalized linear modeling framework with missing data in the outcome. In Chapter 6, variants of AIC, AICc, AICb, and EIC are proposed in the framework of a normal longitudinal regression model with missing data in the outcome. Simulation studies are conducted in both frameworks in order to evaluate and compare the selection performances and biases of the proposed criteria to those that would commonly be used in practice. In both chapters, we also include a brief discussion of the distributional assumptions that would be necessary to develop criteria under the setting of missing data in the outcome and/or covariates.

In Chapter 7, a general outline is given on how to construct criteria that can be used in incomplete data settings for other types of discrepancies. In Chapter 8, an application is given that utilizes the proposed criteria in a study that evaluates the driving performance of individuals with Parkinson's disease under low contrast (fog) conditions in a driving simulator.

CHAPTER 2
PRELIMINARIES

This chapter contains useful concepts and principles that serve as a foundation for the remainder of the dissertation. Included in this chapter is a brief review of the model selection framework, a discussion of missing data principles, and an overview of the bootstrap.

2.1 Model Selection Framework and Principles

We begin with a brief description of the model selection problem. A statistical model is often constructed in order to make conclusions, predictions, or inferences from data. Consider a collection of data $Y = \{y_1, y_2, \ldots, y_n\}$, where the y_is may be scalars or vectors. Suppose the data Y has been generated according to an unknown parametric model. Associated with the unknown parametric model is a likelihood, denoted as $L(\theta_o|Y)$, and an equivalent joint density, denoted as $f(Y|\theta_o)$. One can think of the model in terms of its structural and probabilistic representation, or in terms of its corresponding density or likelihood. The model $L(\theta_o|Y)$ (or equivalently, $f(Y|\theta_o)$) is often referred to as the *true* or *generating model*. We endeavor to find a fitted parametric model which provides a suitable approximation to $L(\theta_o|Y)$.

A parametric family of *candidate* or *approximating models* is often proposed that contains a collection of models of various dimensions and structures. A candidate or approximating model is a model that could potentially be used to describe the data. Let $L(\theta|Y)$ (or equivalently, $f(Y|\theta)$) denote a candidate model, where θ denotes a k-dimensional parameter vector. Let $\hat{\theta}$ denote a set of parameter estimates that are obtained by maximizing the likelihood $L(\theta|Y)$, and let $L(\hat{\theta}|Y)$ denote the corresponding *fitted candidate model*. The objective of model selection is to find the fitted candidate model that is "nearest" to the true model. More specifically,

we search for the fitted candidate model $L(\hat{\theta}|Y)$ among the class of candidate models that "best" approximates the true model $L(\theta_o|Y)$. The objective can often be accomplished by using a model selection criterion.

The model selection framework presented addresses the problem of finding the best fitted candidate model. One may wonder how to determine whether the chosen fitted candidate model is actually a good model. An accepted belief is that a good model will be generalizable, meaning that the model is capable of describing or predicting future observations with a high degree of certainty. Konishi and Kitagawa (2008, p.2) say that, "Akaike considered that the purpose of statistical modeling is not to accurately describe current data or to infer the 'true distribution.' Rather, he thought that the purpose of statistical modeling is to predict future data as accurately as possible."

A good model will often strike a balance between two modeling objectives known as goodness-of-fit and parsimony. Goodness-of-fit refers to the extent to which a fitted candidate model will conform to the data that was used in its own construction. A goodness-of-fit term often includes a measure that reflects the discrepancy between the values used to construct the model, and the expected values under the fitted model. It may be possible to fit a model that conforms to the data very well, but does so because the model is excessively complicated and possibly difficult to interpret. For this reason, the principle of parsimony must be considered.

The idea of parsimony is a consequence of Occam's Razor, which is a principle credited to the medieval English philosopher William of Ockham (1285-1349). Occam's Razor states that given two or more competing explanations for a phenomena, none of which can be discounted, the simplest explanation is to be preferred. Relating this principle to model selection, within a candidate collection of fitted models, the simplest model that adequately fits the data should be preferred. A

key objective in model selection is achieving a balance between goodness-of-fit and parsimony.

The concepts of underfitting and overfitting are also pertinent in determining the quality of a model. Both concepts are defined in terms of the true model. Suppose the true model belongs to the candidate class. A candidate model that has the same structure as the true model is called *correctly specified*. The resulting fitted candidate model would be *correctly fit*. A candidate model that provides an incomplete representation of the true model, perhaps because it does not include necessary covariates, is called *underspecified*. The resulting fitted candidate model would be *underfitted*, and choosing such a model is referred to as *underfitting*. A candidate model that is more complex than the true model, perhaps because it contains extraneous covariates, is called *overspecified*. The resulting fitted candidate model would be *overfitted*, and choosing such a model is referred to as *overfitting*.

Both underfitting and overfitting can lead to problems in statistical modeling. Underfitting may lead to results that are biased. Overfitting may lead to results with unnecessarily high variability. Burnham and Andersen (2002, p.17), in reference to Shibata (1989), state "While one must worry about errors due to both underfitting and overfitting, it seems that modest overfitting is less damaging than underfitting." This may be conceptualized by realizing that it may be less damaging to additionally include an irrelevant covariate to a correctly specified model (overfitting), whereas the failure to include an important covariate in a model (underfitting) could be more problematic. As previously stated, a key objective in model selection is achieving a balance between goodness-of-fit and parsimony, which is analogous to maintaining a balance of bias and variance reduction.

In summary, from a theoretical standpoint, a goal in model selection is to choose a candidate model that best approximates the true model. In many statistical modeling applications, especially those in the biomedical and health sciences,

the notion of any model being correct or true is difficult to defend. However, in theoretical frameworks, one must acknowledge that there is a probabilistic mechanism that generated the data in order to proceed. From a practical point of view, a more realistic goal in model selection is to attempt to capture the most salient features of the true model using the best fitted candidate model. After all, as George Box pointed out, "all models are wrong, some are useful."

2.2 Principles of Missing Data

Missing data is an extremely common occurrence in most research settings. In many statistical analyses, the consequences of missing data are often not fully understood or addressed. Knowing why data is missing, considering how to account for the missing data, and determining the effect of missing data on the original research hypotheses are issues that need to be addressed when a statistical analysis on a data set with missing values is conducted.

Suppose again that we have a collection of data $Y = \{y_1, y_2, \ldots, y_n\}$, where the y_is may be scalars or vectors, and the y_is are independent. Now, suppose some of the elements in Y contain missing values. A common representation of Y is written as $Y = (Y_{obs}, Y_{mis})$. The term Y_{obs} corresponds to all of the observed elements of Y, and is referred to as the *observed data*. The term Y_{mis} corresponds to all of the missing elements of Y, and is referred to as the *missing data*. The data $Y = (Y_{obs}, Y_{mis})$ is often called the *complete data*, whereas the data Y_{obs} is often called the *incomplete data*. The depiction of Y as (Y_{obs}, Y_{mis}) is a conventional representation of Y that is found in many textbooks (Schafer, 1997; Little and Rubin, 2002; McKnight, McKnight, Sidani and Figueredo, 2007).

We introduce another useful delineation of Y as $Y = (Y_{fobs}, Y_{pobs})$. The term Y_{fobs} corresponds to all of the cases which contain no missing data, and is referred to as the *fully observed data* since all of the cases are fully observed. The term Y_{pobs}

corresponds to all of the cases where at least one element is missing, and is referred to as the *partially observed data* since all of the cases are partially observed.

To emphasize the distinction between the representations of $Y = (Y_{obs}, Y_{mis})$ and $Y = (Y_{fobs}, Y_{pobs})$, we first outline what is known as the missing data indicator matrix M. Let M denote a collection of indicator variables that has the same number of entries as Y, where the elements of M are either a 1 or 0. Let an element of M be 1 if the corresponding element of Y is observed, and let an element be 0 if the corresponding element of Y is missing.

The difference between the two delineations of Y is illustrated in the following simple example. Suppose $Y = [y_1, y_2, \ldots, y_6]'$ represents a matrix of data that has been collected for six cases. Let the y_is be vectors of length four where some of the elements are missing. The missing data indicator matrix M for these six cases is given as follows:

y_1	1	1	1	1
y_2	1	1	1	1
y_3	1	1	1	1
y_4	1	1	0	1
y_5	1	0	0	1
y_6	0	1	1	0

For the representation of $Y = (Y_{obs}, Y_{mis})$, Y_{obs} would be that part of Y which corresponds to all of the 1's, while Y_{mis} would be that part of Y which corresponds to all of the 0's. For the representation of $Y = (Y_{fobs}, Y_{pobs})$, Y_{fobs} would correspond to the cases $y_1, y_2, \text{and}, y_3$ since these cases are fully observed, while Y_{pobs} would correspond to the cases $y_4, y_5, \text{and}, y_6$ since these cases are partially observed. Throughout the dissertation, we will use both delineations of Y and attempt to clearly define which we are using.

The effectiveness of tools used to account for missing data may depend on

what is known as the missing data mechanism. The missing data mechanism is determined by the conditional distribution of M given $Y = (Y_{obs}, Y_{mis})$, say $f(M|Y_{obs}, Y_{mis}, \phi)$, where ϕ denotes the unknown parameters that characterize the relationship between Y and M.

Rubin (1976) defines three different types of missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Ideally, in a study that contains missing data, one would have data that can be classified as MCAR. In a MCAR setting, there is no relationship between M and $Y = (Y_{obs}, Y_{mis})$, so the missingness does not depend on the values of either the observed or missing data. This is the most restrictive assumption, which can be written as

$$f(M|Y_{obs}, Y_{mis}, \phi) = f(M|\phi), \quad \text{for all } Y_{obs}, Y_{mis}, \phi.$$

A less restrictive assumption is the MAR mechanism. In a MAR mechanism, there is a relationship between M and Y_{obs}, but not between M and Y_{mis}. Hence, the missingness depends on the values of the observed data, but not the values of the missing data. This mechanism can be expressed as

$$f(M|Y_{obs}, Y_{mis}, \phi) = f(M|Y_{obs}, \phi), \quad \text{for all } Y_{mis}, \phi.$$

Lastly, if the missing data mechanism cannot be classified as MAR or MCAR, it is classified as NMAR. In the NMAR mechanism, there is a relationship between M and Y_{mis}, so the missingness depends on the missing values in Y.

Little and Rubin (2002, p.8) make the assumption throughout their book that the missing data indicators in M contain meaningful information for the analysis. This may lead to the decision of whether to include M in a statistical model that makes inference on θ. In NMAR settings, it would be necessary to model M, but doing so is not a trivial task and the effect on parameter estimates is less clear.

Little and Rubin (2002, p.119) claim that the missing data mechanism is ignorable for likelihood inference, that is M does not need to be modeled, if the data can be classified as MAR and the parameters ϕ and θ are distinct. In MAR settings, there exists a systematic process that underlies the missing data that can be modeled using Y_{obs}. There should be no parameter estimation bias if the mechanism can be classified as MCAR; thus, it is not necessary to model M. Throughout this dissertation we will work under the assumption that the missing data mechanism is MCAR or MAR, and the parameters ϕ and θ are distinct, so the missing data mechanism can be assumed to be ignorable.

2.3 Statistical Approaches to Account for Missing Data

Various approaches have been developed to provide parameter estimates in incomplete data settings. In the next subsections, (2.3.1), (2.3.2), and (2.3.3) we will discuss data deletion methods, imputation methods, and the EM algorithm. These approaches can all lead to reasonable conclusions in certain settings. However, to reiterate the words of R.A. Fisher, "The best solution to handle missing data is to have none." (See McKnight, McKnight, Sidani and Figueredo, 2007, p.vii.)

2.3.1 Complete and Available Case Analysis

A common approach to account for missing data is to use only the cases that are fully observed, which is often referred to as *listwise deletion*, *case deletion*, or a *complete case analysis*. Recall the delineation of the fully and partially observed cases as $Y = (Y_{fobs}, Y_{pobs})$. A complete case analysis ignores Y_{pobs} and considers only Y_{fobs}. The main advantage of a complete case analysis is convenience, since the planned statistical methods can then be applied to a data set with no missing values. The approach may be justifiable in MCAR settings when the amount of

missing data is small, which results only in a loss of precision. It could also be considered when there is a very large sample size relative to a small portion of missing information. Many statistical packages (such as SAS and R) employ a complete case analysis as a default. It should be emphasized that the terminology of a complete case analysis should not be confused with the complete data Y, which is comprised of both the observed and the missing elements.

Little and Rubin (2002, p.41) mention the following disadvantages of a complete case analysis: a loss of precision in parameter estimates, and an increase in bias when the missing data mechanism is not MCAR. In MAR or NMAR settings, the sample of cases in Y_{fobs} may not represent the intended study population of interest. In complete case analyses, note that the number of fully observed cases can quickly diminish as the amount of missingness increases. Suppose that a study contains 10 independent variables with each variable having a probability of missingness of only 5%. The sample size used in a complete case analysis would quickly be reduced by approximately 40% of the original sample size.

Another common technique uses all of the available observed data for a variable of interest, and is referred to as an *available case analysis*. This approach is often considered in many univariate analyses, and may be more attractive than a complete case analysis since less data is deleted. A disadvantage of an available case analysis is that the sample size changes from variable to variable, which presents possible concerns for comparing summary statistics across variables. For example, if variances and covariances between variables are calculated from different samples, the resulting correlations could lie outside the range of $[-1, 1]$. Verbeke and Molenberghs (2000) claim that an available case analysis is only valid under MCAR settings, and they also suggest that using the available cases is more efficient than using the complete cases since more information is incorporated. However, based on conclusions from past simulation studies of others, Little and Rubin (2002, p.55)

suggest that either a complete case analysis or an available case analysis can be superior. In general, neither practice is advocated, since alternative methods exist that can more effectively account for the missing data.

2.3.2 Imputation Methods

Another common approach to account for missing data is to "fill in" the missing elements of Y with reasonable values. Such methods are called imputation methods. The missing values could be filled in one time (single imputation), or several times (multiple imputation). After the missing values are imputed, an analysis is conducted on the data set(s) as if the imputed values were originally observed.

One of the most simple imputation methods is unconditional mean imputation. In this approach, a missing value for a particular variable is imputed with the mean of the observed values for the variable under consideration. Another common imputation method is conditional mean imputation. In this approach, a regression model is fit using all of the cases that contain no missing data. From this model, predictions can be directly inserted for the missing values, or they can be used as a basis for creating a distribution from which draws can be made. Two other approaches common to sampling surveys are hot deck and cold deck imputation. In hot deck imputation, missing values are replaced with draws from similarly responding units. In cold deck imputation, missing values are replaced by a constant value from a similar survey that was previously conducted, or from another external source. The aforementioned imputation methods are often easy to implement. However, for the conditional and unconditional mean approaches, values that are far from the mean or predicted mean tend to be underrepresented. This leads to underestimation of the variance and covariance parameters, which can inflate test statistics. Also, single imputation methods generally do not account for the variability that occurs due to imputation.

Rubin (1987) introduced the notion of multiple imputation where missing elements are replaced d times (usually from 3-10) to create d multiply imputed data sets. In general, the objective is to generate multiple imputations that are "proper", meaning they are independent draws from the posterior predictive distribution of the missing data under a complete data model and a prior distribution. The missing values from data sets with arbitrary patterns of missingness are often generated through a Gibbs sampling algorithm (Geman and Geman, 1984; Gelfand and Smith 1990), a Markov Chain Monte Carlo (MCMC) method (Schafer, 1997) that assumes multivariate normality, or an imputation software package called IVEWARE uses a sequence of regression models in creating imputed values (Raghunathan, Lepkowski, Van Hoewyk and Solenberger, 2001). The desired statistical analysis is conducted on each of the d multiply imputed data sets as if there was no missing data, and the results are combined. We follow the explanation of Rubin in outlining the procedure.

Let $\hat{\theta}_d$ and $\hat{V}_d$ denote the parameter and variance estimates of θ, respectively, from the multiply imputed data sets $d = 1, 2, \ldots, D$. The multiple imputation estimate of θ is

$$\bar{\theta} = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d.$$

The associated variance estimate of $\bar{\theta}$ is comprised of two components known as the average within-imputation variance,

$$W = \frac{1}{D} \sum_{d=1}^{D} \hat{V}_d,$$

and the between-imputation variance,

$$B = \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\theta}_d - \bar{\theta})^2.$$

The overall variability of $\bar{\theta}$ is then defined as

$$T = W + \frac{D+1}{D}B,$$

where $(D+1)/D$ is an adjustment for a finite number of multiply imputed data sets.

Multiple imputation has many attractive properties. These properties include an applicability to many different missing data patterns and readily available adjustments for imputation uncertainty. Furthermore, valid parameter estimates and standard errors are attainable. However, the imputed values must be "proper" in order for Rubin's formulas to be valid (Schafer, 1997). An estimate of the amount of information that is lost due to missing data can also be obtained.

2.3.3 EM algorithm

Dempster, Laird and Rubin (1977) introduced the expectation maximization (EM) algorithm as a procedure for obtaining maximum likelihood estimates in incomplete data settings. The EM algorithm is applicable in a large variety of conventional missing data settings, and can even be implemented for unconventional settings where the missing data corresponds to latent variables, such as variance component estimation for a mixed model. The objective of the EM algorithm is to obtain an estimate of θ by maximizing the incomplete data log-likelihood, $\log L(\theta|Y_{obs})$. However, the incomplete data log-likelihood may often be cumbersome to work with directly and this maximization may be difficult to accomplish.

The EM algorithm is an iterative procedure that involves 2 steps: the expectation (E) step and the maximization (M) step. The complete data log-likelihood is defined as $\log L(\theta|Y_{obs}, Y_{mis})$. The E-step takes the expectation of the complete data log-likelihood with respect to the distribution $f(Y_{mis}|Y_{obs}, \theta^{(t)})$, where $\theta^{(t)}$ is

the current parameter estimate of θ. Thus, the E-step is based on the evaluation of

$$Q(\theta|\theta^{(t)}) = \int \log L(\theta|Y_{obs}, Y_{mis}) f(Y_{mis}|Y_{obs}, \theta^{(t)}) dY_{mis}.$$

The M-step is found by maximizing $Q(\theta|\theta^{(t)})$ with respect to θ to obtain $\theta^{(t+1)}$. Iteration occurs between the E and M steps until convergence.

A drawback of the EM algorithm is that convergence can be slow when there are large amounts of missing data. However, under general conditions, the EM algorithm increases the incomplete data log-likelihood at each iteration. Moreover, if the sequence of $\theta^{(t)}$ converges, it will do so to a local maximum or saddle point of $\log L(\theta|Y_{obs})$. (See Little and Rubin, 2002.) Similar to multiple imputation, the EM algorithm solves an incomplete data problem by repeatedly solving the complete data problem (Schafer, 1997).

2.4 Overview of the Bootstrap

The bootstrap is a computationally-intensive resampling technique that was first introduced by Efron (1979). A comprehensive introduction to the bootstrap is provided by Efron and Tibshirani (1993), who present many bootstrap applications in a variety of settings. The bootstrap is often employed as a way of obtaining an empirical representation of the sampling distribution. It is particularly useful when necessary distributional assumptions are violated that may be needed in order to analytically characterize the sampling distribution.

The general idea behind the bootstrap is easily described. Based on the original sample, one obtains a bootstrap sample from which the desired statistic or parameter estimate is calculated. The process is repeated many times to obtain a collection of replicates of the statistic or parameter estimate of interest. Using this collection of replicates, one may obtain confidence intervals, perform hypotheses tests, or make other inferences.

Different approaches exists for obtaining a bootstrap sample. In the regression framework, we will outline three popular methods: the *non-parametric bootstrap*, the *parametric bootstrap*, and the *semi-parametric bootstrap*.

Assume data is obtained of the form $\{(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\}$, where the (y_i, x_i) are independent and identically distributed according to some unknown distribution F. Suppose that each y_i is a scalar response, and each x_i is a vector of covariates.

The empirical distribution of the original sample is a discrete distribution, say $\hat{F}$, which assigns probability $1/n$ to each (y_i, x_i). In non-parametric bootstrap sampling, the bootstrap data (y_i^*, x_i^*) are obtained by generating a random sample of size n drawn with replacement from $\hat{F}$.

Generating a parametric bootstrap sample requires the parametric estimation of F, which can be accomplished by defining a parametric model for the data. In the regression setting, the relationship between y_i and x_i is hypothesized to be of the form

$$y_i = x_i'\beta + e_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{2.1}$$

where β is a vector of unknown regression parameters corresponding to the covariate vector x_i, and the e_i are unobserved errors assumed to follow a $N(0, \sigma^2)$ distribution. Estimates of β and σ^2 can be obtained by fitting the model using the method of maximum likelihood. Let these estimates be respectively denoted by $\hat{\beta}$ and $\hat{\sigma}^2$. The bootstrap responses y_i^* in a parametric bootstrap sample are generated on a case-by-case basis by taking draws from the distribution

$$N(x_i'\hat{\beta}, \hat{\sigma}^2), \quad \text{for } i = 1, 2, \ldots, n.$$

Note that the mean value from which each bootstrap response y_i^* is drawn varies with each case.

We outline the semi-parametric bootstrap by again assuming that the relationship between y_i and x_i is of the form (2.1). However, we only assume that the e_i are a random sample of unobserved errors from an unknown distribution with mean 0. An estimate of β can be obtained using the method of least squares or some other fitting procedure. Let $\hat{\beta}$ denote this estimate. The residuals can then be calculated as

$$\hat{e}_i = y_i - x_i'\hat{\beta}, \quad \text{for } i = 1, 2, \ldots, n.$$

A sample of size n is drawn with replacement from the collection of residuals. We denote this sample as $(e_1^*, e_2^*, \ldots, e_n^*)$. The bootstrap responses y_i^* in a semi-parametric bootstrap sample are obtained as

$$y_i^* = x_i'\hat{\beta} + e_i^*, \quad \text{for } i = 1, 2, \ldots, n.$$

CHAPTER 3
KULLBACK-LEIBLER DISCREPANCY BASED MODEL SELECTION CRITERIA

In this chapter the Kullback-Leibler (KL) discrepancy will be formally defined. This measure serves as the basis for the Akaike information criterion (AIC) and AIC based variants. A characterization of the discrepancies based on the complete data and the fully observed data is also provided. An outline of the criteria that target the fully observed data discrepancy is included. We then propose new selection criteria that estimate the complete data discrepancy. We conclude with a brief discussion contrasting the newly proposed methodology to techniques in multiple imputation.

3.1 Kullback-Leibler Based Criteria With No Missing Data

Suppose again that we have a collection of data $Y = \{y_1, y_2, \ldots, y_n\}$, where the y_is may be scalars or vectors. Let $L(\theta_o|Y)$ (or equivalently, $f(Y|\theta_o)$) denote an unknown parametric model that presumably generated Y. Let θ denote a k-dimensional parameter vector for the candidate model that could be used to describe the data Y. Let $L(\theta|Y)$ (or equivalently, $f(Y|\theta)$) denote the candidate model. For now, assume that Y contains *no* missing values.

The Kullback-Leibler (KL) discrepancy, also known as directed-divergence or information, is a well-known measure of separation between two statistical models. The KL discrepancy between the true or generating model $L(\theta_o|Y)$ and the candidate model $L(\theta|Y)$ is defined as

$$d(\theta, \theta_o) = E_o\{-2 \log L(\theta|Y)\},$$

where E_o denotes the expectation under the true model, and $L(\theta|Y)$ represents the likelihood corresponding to the candidate model.

Let $\hat{\theta}$ denote a set of estimates for θ obtained by maximizing the likelihood for the candidate model $L(\theta|Y)$. The fitted candidate model is represented by $L(\hat{\theta}|Y)$. The *overall KL discrepancy*

$$d(\hat{\theta}, \theta_o) = E_o\{-2 \log L(\theta|Y)\}|_{\theta=\hat{\theta}}, \tag{3.1}$$

would then provide a useful measure of separation between the true model and the fitted candidate model.

A model selection criterion is often formulated by creating a statistic that has an expectation equal to (or approximately equal to) the expectation of (3.1). This measure is called the *expected overall KL discrepancy*, denoted as

$$D(k, \theta_o) = E_o\{d(\hat{\theta}, \theta_o)\} = E_o\{E_o\{-2 \log L(\theta|Y)\}|_{\theta=\hat{\theta}}\}. \tag{3.2}$$

Evaluating (3.1) and (3.2) is not possible, since doing so requires knowledge of θ_o, which is unknown. Akaike (1973), however, noted that $-2 \log L(\hat{\theta}|Y)$ serves as a negatively biased estimator of (3.1). In order to utilize this estimator, we must investigate this bias. To do so, consider writing (3.2) as

$$\begin{aligned} D(k, \theta_o) &= E_o\{E_o\{-2 \log L(\theta|Y)\}|_{\theta=\hat{\theta}}\} \\ &= E_o\{-2 \log L(\hat{\theta}|Y)\} \\ &\quad + [E_o\{E_o\{-2 \log L(\theta|Y)\}|_{\theta=\hat{\theta}}\} - E_o\{-2 \log L(\hat{\theta}|Y)\}]. \end{aligned} \tag{3.3}$$

The bracketed quantity (3.3) is often referred to as the *expected optimism* in judging the fit of a model using the same data as that which was used to construct the fit (Efron, 1983, 1986). The expected optimism is positive and must be evaluated or approximated to correct for the negative bias incurred when $-2 \log L(\hat{\theta}|Y)$ is used as an estimator of (3.1). The bias correction or estimator of the expected optimism, say $\widehat{eo}$, is then added to $-2 \log L(\hat{\theta}|Y)$ to produce an approximately

unbiased estimator of the expected overall KL discrepancy $D(k, \theta_o)$:

$$-2 \log L(\hat{\theta}|Y) + \widehat{eo}.$$

The measure $-2 \log L(\hat{\theta}|Y)$ is often referred to as the goodness-of-fit term, while the estimate $\widehat{eo}$ is often referred to as the penalty term. Estimates of the expected optimism can be constructed using a variety of methods, including asymptotic simplifications, Monte Carlo simulation, cross-validation, and bootstrapping. In this chapter, we will briefly outline two classical and two bootstrap based estimators of the expected optimism.

The estimator of the expected optimism for AIC was derived by Akaike using the large-sample properties of maximum likelihood estimators. The derivation of AIC, which shows that AIC is an asymptotically unbiased estimator of $D(k, \theta_o)$, assumes that the candidate model is correctly specified or overspecified. To justify the asymptotic unbiasedness of AIC, consider writing (3.2) as

$$\begin{aligned} D(k, \theta_o) &= E_o\{d(\hat{\theta}, \theta_o)\} \\ &= E_o\{-2 \log L(\hat{\theta}|Y)\} \\ &\quad + [E_o\{-2 \log L(\theta_o|Y)\} - E_o\{-2 \log L(\hat{\theta}|Y)\}] \qquad (3.4) \\ &\quad + [E_o\{d(\hat{\theta}, \theta_o)\} - E_o\{-2 \log L(\theta_o|Y)\}]. \qquad (3.5) \end{aligned}$$

Assuming the necessary regularity conditions to ensure the large sample properties of maximum likelihood estimators, one can establish that the terms (3.4) and (3.5) are each within $o(1)$ of k (Cavanaugh, 1997). Thus, the estimator is given by twice the dimension of θ: i.e., if k denotes the number of functionally independent parameters in the candidate model $L(\theta|Y)$, the estimator is $2k$. Under the appropriate conditions, the expected value of

$$\text{AIC} = -2 \log L(\hat{\theta}|Y) + 2k$$

should be asymptotically near the expected value of (3.1). More specifically, one can establish that

$$E_o\{\text{AIC}\} + o(1) = D(k, \theta_o).$$

AIC provides us with an approximately unbiased estimator of $D(k, \theta_o)$ in settings where n is large and k is comparatively small. Yet in other settings, $2k$ may be much smaller than the expected optimism (3.3), making AIC substantially negatively biased as an estimator of $D(k, \theta_o)$. To correct for the negative bias, other methods have been proposed to evaluate the expected optimism. One approach is to consider a specific modeling framework and to exactly evaluate or approximate the expected optimism based on the properties of maximum likelihood estimators for the setting. The resulting criterion under this approach is called the "corrected" Akaike information criterion, AICc. Sugiura (1978) noted that the expected optimism can be exactly evaluated in the framework of normal univariate linear regression models. Bedrick and Tsai (1994) showed that the expected optimism can be exactly evaluated in the framework of normal multivariate linear regression models. AICc has also been extended to a number of additional modeling frameworks, which include nonlinear regression models and autoregressive models (Hurvich and Tsai, 1989), autoregressive moving-average models (Hurvich, Shumway and Tsai, 1990), vector autoregressive models (Hurvich and Tsai, 1993), generalized linear models with a dispersion parameter (Hurvich and Tsai, 1995), and models for longitudinal data analysis under the assumption of a known covariance structure (Azari, Li and Tsai, 2006). The derivations assume that the candidate model is correctly specified or overspecified.

We introduce AICc in the context of the normal multivariate linear regression model. Consider the normal multivariate linear regression model

$$Y = X\beta + U.$$

Here, Y is an $n \times m$ response matrix of independent rows corresponding to m response variables for n cases, X is an $n \times p$ known design matrix of full column rank p, β is a $p \times m$ matrix of unknown regression parameters, and U is an $n \times m$ error matrix comprised of independent rows which are normal random vectors, each with a mean of 0 and a covariance matrix Σ.

Bedrick and Tsai (1994) show that in this setting, an exact estimator of the expected optimism is given as $2k(n/(n-m-p-1))$. If the candidate model is correctly specified or overspecified, the expected value of

$$\text{AICc} = -2\log L(\hat{\theta}|Y) + 2k(n/(n-m-p-1))$$

should be exactly equal to the expected value of (3.1). More specifically, one can establish that

$$E_o\{\text{AICc}\} = D(k, \theta_o);$$

thus, AICc is an exactly unbiased estimator of $D(k, \theta_o)$. (The preceding property holds up to $o(1)$ for other modeling frameworks in which AICc has been justified and developed.)

Bootstrapping methods have also been used to develop estimators of the expected optimism. In some situations, the estimators based on bootstrap methods may have some advantages over the estimators used in AIC and AICc. AIC was established using large-sample arguments; hence, the estimator of the expected optimism may not hold in smaller sample settings. AICc has only been established in certain modeling frameworks and is therefore not as universally applicable as AIC. The estimators of the expected optimism used in AIC and AICc were derived under the strict assumption that the candidate model is correctly specified or overfit. The behavior of AIC and AICc when this condition is not met has been investigated by Hurvich and Tsai (1991). In general, the validity of bootstrap estimators does not

depend on this assumption (Shibata, 1997). Bootstrap methods can be applied in a large variety of modeling frameworks and may provide better estimators of the expected optimism when the sample size is small.

The idea of using the bootstrap to improve the performance of a model selection rule was first introduced by Efron (1983, 1986). Instead of analytically deriving the expected optimism for a certain type of candidate model under stringent (and perhaps unrealistic) conditions, the bias is estimated through numerical approximations that involve the bootstrap. In what follows, we will outline two different ways to estimate the expected optimism using the bootstrap. The first approach we will discuss was developed by Ishiguro, Sakamoto and Kitagawa (1997), who advocated an estimator of the expected optimism that is used in the extended information criterion (EIC). This estimator is based on Efron's methodology.

Recall that the expected optimism (3.3) can be written in terms of the density as the difference of two components:

$$E_o\{E_o\{-2\log f(Y|\theta)\}|_{\theta=\hat{\theta}}\} \tag{3.6}$$

$$-E_o\{-2\log f(Y|\hat{\theta})\}. \tag{3.7}$$

Following Efron (1983, 1986) and Konishi and Kitagawa (2008), assume that the bootstrap distribution corresponds to the empirical distribution $\hat{F}(Y)$ of the original sample, which assigns probability $1/n$ to each of the cases $\{y_1, y_2, \ldots, y_n\}$. The bootstrap sample of Y is obtained by generating a random sample of size n drawn with replacement from $\hat{F}(Y)$ and is denoted as Y^*. Let $\hat{\theta}^*$ denote the bootstrap replicate of the MLE of θ based on Y^*. In constructing the bootstrap analogues of (3.6) and (3.7), we can apply a technique that Efron referred to as the plug-in principle. To illustrate this idea, let $F(Y|\theta_o)$ represent the cumulative distribution function corresponding to $f(Y|\theta_o)$. The true cumulative distribution function $F(Y|\theta_o)$ is replaced with the empirical distribution $\hat{F}(Y)$ of the original

sample. In conjunction with this replacement, the other components in (3.6) and (3.7) are replaced as follows:

$$F(Y|\theta_o) \longrightarrow \hat{F}(Y),$$
$$Y \sim F(Y|\theta_o) \longrightarrow Y^* \sim \hat{F}(Y),$$
$$E_o \longrightarrow E_*,$$
$$\hat{\theta} \longrightarrow \hat{\theta}^*.$$

Here, * corresponds to the bootstrap sample, and E_* represents the expectation taken with respect to the distribution of the bootstrap sample. The bootstrap analogues of (3.6) and (3.7) are formulated by considering the distribution of the bootstrap sample as the true distribution. After appropriate substitutions, these analogues are represented as

$$E_*\{E_*\{-2\log f(Y^*|\theta)\}|_{\theta=\hat{\theta}^*}\} \tag{3.8}$$
$$-E_*\{-2\log f(Y^*|\hat{\theta}^*)\}. \tag{3.9}$$

One of the most useful features of the bootstrap is that the expectations in (3.8) and (3.9) can be approximated numerically using Monte Carlo simulation, since the empirical distribution of the sample is known. Let $\{Y^*(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples. Let $\{\hat{\theta}^*(b)|b = 1, 2, \ldots, B\}$ represent a set of B bootstrap replicates of MLEs of θ corresponding to the B bootstrap samples.

Consider first the bootstrap expectation (3.9). This expectation can be estimated by

$$\frac{1}{B}\sum_{b=1}^{B} -2\log f(Y^*(b)|\hat{\theta}^*(b)). \tag{3.10}$$

For (3.8), we first focus on the inner expectation. It can be shown that

$$\begin{aligned} E_*\{-2\log f(Y^*|\theta)\} &= \frac{1}{n}\sum_{i=1}^{n} -2\log f_i(y_i|\theta) \\ &= -2\log f(Y|\theta), \end{aligned} \tag{3.11}$$

where f_i represents the density for case i under the candidate model. The result from (3.11) holds true when the bootstrap sample is generated by taking a sample with replacement from the empirical distribution $\hat{F}(Y)$ (i.e. a non-parametric bootstrap). Shibata (1997) shows that under a normal linear regression model setting, the result (3.11) also holds true under parametric or semi-parametric bootstrapping.

Using (3.11), one can express (3.8) as

$$E_*\{-2\log f(Y|\theta)|_{\theta=\hat{\theta}^*}\} = E_*\{-2\log f(Y|\hat{\theta}^*)\},$$

which can be estimated by

$$\frac{1}{B}\sum_{b=1}^{B} -2\log f(Y|\hat{\theta}^*(b)). \tag{3.12}$$

Employing the Monte Carlo approximations given in (3.10) and (3.12), a bootstrap based estimator of the expected optimism in terms of the likelihood is provided by

$$\frac{1}{B}\sum_{b=1}^{B}\Big\{ -2\log L(\hat{\theta}^*(b)|Y) - \Big(-2\log L(\hat{\theta}^*(b)|Y^*(b))\Big)\Big\}.$$

The model selection criterion based on this estimator is the extended information criterion, EIC. EIC is given by

$$\text{EIC} = -2\log L(\hat{\theta}|Y) + \frac{1}{B}\sum_{b=1}^{B}\Big\{ -2\log L(\hat{\theta}^*(b)|Y) - \Big(-2\log L(\hat{\theta}^*(b)|Y^*(b))\Big)\Big\}$$

Another estimate of the expected optimism using bootstrap methodology was proposed by Cavanaugh and Shumway (1997) in the context of the state-space modeling framework. They show that under suitable conditions, the difference

between

$$2[E_*\{-2\log L(\hat{\theta}^*|Y)\} - \{-2\log L(\hat{\theta}|Y)\}]$$

and the expected optimism (3.3) converges to zero (as $n \to \infty$). This follows from the observation that (3.3) can be decomposed into the sum of

$$E_o\{E_o\{-2\log L(\theta|Y)\}|_{\theta=\hat{\theta}}\} - E_o\{-2\log L(\theta_o|Y)\} \tag{3.13}$$

and

$$E_o\{-2\log L(\theta_o|Y)\} - E_o\{-2\log L(\hat{\theta}|Y)\}, \tag{3.14}$$

and that the difference between

$$E_*\{-2\log L(\hat{\theta}^*|Y)\} - \{-2\log L(\hat{\theta}|Y)\} \tag{3.15}$$

and either (3.13) or (3.14) tends to zero (as $n \to \infty$).

Now by the law of large numbers, as $B \to \infty$,

$$\frac{1}{B}\sum_{b=1}^{B} -2\log L(\hat{\theta}^*(b)|Y)$$

converges to

$$E_*\{-2\log L(\hat{\theta}^*|Y)\}.$$

Thus, for $B \to \infty$,

$$\frac{1}{B}\sum_{i=1}^{B}\Big\{-2\log L(\hat{\theta}^*(b)|Y)\Big\} - \Big(-2\log L(\hat{\theta}|Y)\Big)$$

is asymptotically the same as (3.15). This leads to another large-sample bootstrap based estimator of the expected optimism:

$$2\left[\frac{1}{B}\sum_{b=1}^{B}\Big\{-2\log L(\hat{\theta}^*(b)|Y)\Big\} - \Big(-2\log L(\hat{\theta}|Y)\Big)\right].$$

The model selection criterion based on this estimator is AICb, which is given by

$$\text{AICb} = -2\log L(\hat{\theta}|Y) + 2\left[\frac{1}{B}\sum_{b=1}^{B}\Big\{-2\log L(\hat{\theta^*}(b)|Y)\Big\} - \Big(-2\log L(\hat{\theta}|Y)\Big)\right].$$

Cavanaugh and Shumway justify AICb in a state-space modeling framework, but the criterion can be applied in a very general context. Shibata (1997) has established the asymptotic equivalence of AICb and EIC under a general set of conditions, and has indicated the existence of other asymptotically equivalent bootstrap based AIC variants.

3.2 Complete vs. Fully Observed Data Discrepancy

In the preceding section (3.1) we defined Y as a collection of data $Y = \{y_1, y_2, \ldots, y_n\}$, where the y_is were scalars or vectors containing no missing data. The four model selection criteria, AIC, AICc, EIC, and AICb were all outlined under this assumption. The model selection criteria have expectations which are equal to, or approximately equal to, the expected KL discrepancy (3.2). For the remainder of the chapter, we now assume that some of the elements of Y are missing.

We define the KL discrepancy in two different ways, by using either the complete data, $Y = (Y_{fobs}, Y_{pobs})$, or the fully observed data, Y_{fobs}. Assuming that the cases comprising Y are independent, the following decomposition is valid:

$$-2\log L(\theta|Y) = -2\log L(\theta|Y_{fobs}) + (-2\log L(\theta|Y_{pobs})).$$

The *complete data KL discrepancy* between $L(\theta_o|Y)$ and $L(\theta|Y)$ is defined as

$$\begin{aligned} d_{comp}(\theta,\theta_o) &= E_o\{-2\log L(\theta|Y)\} \\ &= E_o\{-2\log L(\theta|Y_{fobs})\} + E_o\{-2\log L(\theta|Y_{pobs})\}, \end{aligned} \tag{3.16}$$

where E_o denotes the expectation under the true model, and $L(\theta|Y)$ represents the likelihood corresponding to the candidate model based on the complete data.

The *fully observed data KL discrepancy* between $L(\theta_o|Y_{fobs})$ and $L(\theta|Y_{fobs})$ is defined as

$$d_{fobs}(\theta, \theta_o) = E_o\{-2 \log L(\theta|Y_{fobs})\}, \tag{3.17}$$

where E_o denotes the expectation under the true model, and $L(\theta|Y_{fobs})$ represents the likelihood corresponding to the candidate model based on only the fully observed data.

Subsequently, model selection criteria can be developed based on either (3.16) or (3.17). The fully observed data KL discrepancy ignores potentially useful information that could be contained in the partially observed cases Y_{pobs}, while the complete data KL discrepancy takes into account this information. Model selection criteria calculated in a complete case analysis would be based on the fully observed data KL discrepancy (3.17), but it may be advantageous to define model selection criteria based on the complete data KL discrepancy (3.16).

The relationship between (3.16) and (3.17) provides an important insight into why it may be preferable to base model selection on the former as opposed to the latter. To exhibit this relationship, one can establish using Jensen's inequality that for any Y_{pobs} and any θ,

$$E_o\{-2 \log L(\theta|Y_{pobs})\} \geq E_o\{-2 \log L(\theta_o|Y_{pobs})\}. \tag{3.18}$$

If we then define $k(\theta_o) = E_o\{-2 \log L(\theta_o|Y_{pobs})\}$, by (3.16) and (3.18), we have for any θ

$$d_{comp}(\theta, \theta_o) \geq d_{fobs}(\theta, \theta_o) + k(\theta_o). \tag{3.19}$$

Thus as a function of θ, the complete data KL discrepancy $d_{comp}(\theta, \theta_o)$ is always at least as great as the fully observed data KL discrepancy $d_{fobs}(\theta, \theta_o)$, adjusted by the constant $k(\theta_o)$.

From (3.16) and (3.19), we can infer that the complete data KL discrepancy

is potentially more sensitive than the fully observed data KL discrepancy and may be preferable for assessing the separation between the candidate model and the true model. Consequently, in the presence of missing data, it may be desirable to create model selection criteria based on the complete data KL discrepancy.

In the next section (3.3), we will briefly outline criteria that are based on the fully observed data KL discrepancy. In section (3.4), we will develop and propose criteria that are based on the complete data KL discrepancy, since it may be a more beneficial target discrepancy.

3.3 Kullback-Leibler Based Criteria Using the Fully Observed Data

Recall the fully observed data KL discrepancy between $L(\theta_o|Y_{fobs})$ and $L(\theta|Y_{fobs})$ given in (3.17). Consider a set of maximum likelihood estimates, $\hat{\theta}_{fobs}$, obtained by maximizing the likelihood $L(\theta|Y_{fobs})$. The *overall fully observed data KL discrepancy* is given by

$$d_{fobs}(\hat{\theta}_{fobs}, \theta_o) = E_o\{-2\log L(\theta|Y_{fobs})\}|_{\theta=\hat{\theta}_{fobs}}. \tag{3.20}$$

The four model selection criteria derived in section (3.1), under the assumption of no missing data, are available in an incomplete data setting using Y_{fobs} in place of Y. We shall refer to these criteria as the *fully observed data criteria*.

AIC constructed using the fully observed data is given by

$$\text{AIC}_{fobs} = -2\log L(\hat{\theta}_{fobs}|Y_{fobs}) + 2k.$$

Let n_{cc} denote the number of cases that are fully observed (i.e., complete cases). Under the multivariate linear regression model framework outlined in section (3.1), the fully observed version of AICc is defined as

$$\text{AICc}_{fobs} = -2\log L(\hat{\theta}_{fobs}|Y_{fobs}) + 2k(n_{cc}/(n_{cc} - m - p - 1)).$$

To construct the fully observed data bootstrap criteria, suppose that we obtain a collection of B bootstrap samples. Let $\{Y^*_{fobs}(b)|b = 1, 2, \ldots, B\}$ represent this collection, with each sample of size n_{cc}. Let $\{\hat{\theta}^*_{fobs}(b)|b = 1, 2, \ldots, B\}$ represent a set of B bootstrap replicates of MLEs of θ corresponding to the B bootstrap samples.

The fully observed data analogue of EIC is given as

$$\begin{aligned}\mathrm{EIC}_{fobs} &= -2\log L(\hat{\theta}_{fobs}|Y_{fobs}) \\ &\quad + \frac{1}{B}\sum_{b=1}^{B}\Big\{ -2\log L(\hat{\theta}^*_{fobs}(b)|Y_{fobs}) - \Big(-2\log L(\hat{\theta}^*_{fobs}(b)|Y^*_{fobs}(b))\Big)\Big\}.\end{aligned}$$

The version of AICb based on the fully observed data is represented as

$$\begin{aligned}\mathrm{AICb}_{fobs} &= -2\log L(\hat{\theta}_{obs}|Y_{fobs}) \\ &\quad + 2\left[\frac{1}{B}\sum_{b=1}^{B}\Big\{ -2\log L(\hat{\theta}^*_{fobs}(b)|Y_{fobs})\Big\} - \Big(-2\log L(\hat{\theta}_{fobs}|Y_{fobs})\Big)\right].\end{aligned}$$

The four criteria given in this section are what would be used in a complete case analysis. These criteria should have expectations which are equal to (or approximately equal to) the expectation of the overall fully observed data KL discrepancy (3.20).

3.4 Kullback-Leibler Based Criteria Using the Complete Data

In this section, four criteria are proposed that target the complete data KL discrepancy. We refer to such criteria as *complete data criteria*. Recall the complete data KL discrepancy between $L(\theta_o|Y)$ and $L(\theta|Y)$ given in (3.16). Consider a set of maximum likelihood estimates, $\hat{\theta}$, that are calculated using Y. The *overall complete data KL discrepancy* is given by

$$d_{comp}(\hat{\theta}, \theta_o) = E_o\{-2\log L(\theta|Y)\}|_{\theta=\hat{\theta}}. \tag{3.21}$$

More formally stated, our goal is to propose model selection criteria that have

an expectation which approximates the expectation of (3.21). This measure is called the *expected overall complete data KL discrepancy*, written as

$$\begin{aligned} D_{comp}(k, \theta_o) &= E_o\{d_{comp}(\hat{\theta}, \theta_o)\} \\ &= E_o\{E_o\{-2 \log L(\theta|Y)\}|_{\theta=\hat{\theta}}\} \\ &= E_o\{-2 \log L(\hat{\theta}|Y)\} \qquad (3.22) \\ &\quad + [E_o\{E_o\{-2 \log L(\theta|Y)\}|_{\theta=\hat{\theta}}\} - E_o\{-2 \log L(\hat{\theta}|Y)\}]. \qquad (3.23) \end{aligned}$$

In developing criteria that have an expectation which approximates $D_{comp}(k, \theta_o)$, we first propose a complete data goodness-of-fit term which approximates (3.22). We then propose four estimators of the expected optimism which approximate (3.23).

A natural estimator of $E_o\{-2 \log L(\hat{\theta}|Y)\}$ is simply $-2 \log L(\hat{\theta}|Y)$, the goodness-of-fit term. However, in the present setting, this estimate is inaccessible since some of the elements of Y are missing. Our goal is to construct an analogue of $-2 \log L(\hat{\theta}|Y)$ based on imputing the missing elements of Y in an appropriate fashion. One important point of emphasis: in the goodness-of-fit term, the maximum likelihood estimate for a candidate model should be based on the data that appears in the likelihood.

Using the conventional representation of $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} contains all of the elements of Y that are observed and Y_{mis} is the collection of elements that are not seen, the following decomposition in terms of the densities is valid:

$$-2 \log f(Y|\theta) = -2 \log f(Y_{obs}|\theta) + -2 \log f(Y_{mis}|Y_{obs}, \theta).$$

Ideally, Y_{mis} would be imputed as a draw from the conditional distribution of the missing data given the observed data using the true model parameter θ_o, denoted as $f(Y_{mis}|Y_{obs}, \theta_o)$. The maximum likelihood estimate could then be constructed using the observed data Y_{obs} and the imputed values for Y_{mis}. However, the preceding proposal is not possible since the true parameter θ_o is not available.

We propose using an estimator of θ obtained from the EM algorithm, along with techniques related to the bootstrap, in developing a complete data analogue of $-2\log L(\hat{\theta}|Y)$. As described in section (2.3.3), the EM algorithm is a useful procedure for estimating model parameters when the data is incomplete. Loosely speaking, the E-step calculates the conditional expectation of the missing data given the observed data at the current values of the parameter estimates. A bootstrap related approach in the spirit of the E-step could be considered by replacing the missing data with a draw from the conditional distribution of the missing data given the observed data, using the parameter estimates from the EM algorithm. Little and Rubin (2002) suggest that imputing missing values by taking a draw from the conditional distribution of the missing data given the observed data is generally preferred to simply imputing the mean from the conditional distribution of the missing data given the observed data.

The proposed methodology is now formally provided. Let $\hat{\theta}_{obs}$ denote the estimator of θ obtained using the EM algorithm for a particular candidate model, where the subscript *obs* reminds us that this estimator maximizes the observed data log-likelihood $L(\theta|Y_{obs})$. Let $\{Y^*_{mis}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y_{mis}, generated by taking draws from the conditional distribution $f(Y_{mis}|Y_{obs}, \hat{\theta}_{obs})$. Let $\{Y^*_{par}(b) = (Y_{obs}, Y^*_{mis}(b))|b = 1, 2, \ldots, B\}$ represent a collection of B *partially* bootstrapped samples. We refer to the samples as being partially bootstrapped since only the elements of Y_{mis} are bootstrapped, while the observed elements Y_{obs} are not. Let $\{\hat{\theta}^*_{par}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B partial bootstrap replicates of MLEs of θ corresponding to the B partially bootstrapped samples. Note that the distribution from which the bootstrap samples of Y_{mis} are drawn will vary with each candidate model under consideration.

We have proposed a bootstrap related approach using the EM algorithm to address the missing elements of Y. The maximum likelihood estimate for a candidate

model is based on the observed data and the bootstrapped samples of the missing elements. The complete data goodness-of-fit term using partial bootstrapping is now proposed as

$$\text{GOF}_{comp} = \frac{1}{B}\sum_{b=1}^{B} -2\log L(\hat{\theta}^*_{par}(b)|Y^*_{par}(b)). \tag{3.24}$$

This term is equivalent to the standard goodness-of-fit term, $-2\log L(\hat{\theta}|Y)$, when there is no missing data. The complete data goodness-of-fit term will be used for all four of the complete data analogues of AIC, AICc, EIC, and AICb. We now propose four estimators of the expected optimism that correspond to the respective complete data criteria.

In settings where there is no missing data, recall that Akaike showed $2k$ serves as an estimator of the expected optimism. This estimator is used for AIC. The derivation uses the large-sample properties of maximum likelihood estimators, and assumes that the candidate model is correctly specified or overfit. If the degree of missingness of Y is not too extreme, the large-sample results should still hold. Intuitively, a large-sample argument would not be affected by small or moderate amounts of missingness; however, in situations with large amounts of missingness relative to the sample size, $2k$ may not provide a good estimator of the expected optimism. The complete data analogue of AIC is proposed as

$$\text{AIC}_{comp} = \frac{1}{B}\sum_{b=1}^{B}\Big\{ -2\log L(\hat{\theta}^*_{par}(b)|Y^*_{par}(b))\Big\} + 2k.$$

In settings where there is no missing data, recall that Bedrick and Tsai showed in the framework of normal multivariate linear regression models that the expected optimism can be exactly evaluated as $2k(n/(n-m-p-1))$, which is the estimator used in AICc. Again, the derivation assumes that the candidate model is correctly specified or overfit. If the degree of missingness of Y is not too extreme, the result should still approximately hold; however, the estimator will no longer be exact. The

complete data analogue of AICc is given by

$$\text{AICc}_{comp} = \frac{1}{B}\sum_{b=1}^{B}\Big\{ -2\log L(\hat{\theta}^*_{par}(b)|Y^*_{par}(b))\Big\} + 2k(n/(n-m-p-1)).$$

The bootstrap based estimators of the expected optimism used for EIC and AICb were derived in section (3.1) in settings where there is no missing data. The estimators both depend on Y, and thus are unattainable if some of the elements of Y are missing. Recall that in the complete data goodness-of-fit term (3.24), the missing elements of Y are replaced with bootstrap samples obtained by taking draws from $f(Y_{mis}|Y_{obs},\hat{\theta}_{obs})$. In conjunction with the complete data goodness-of-fit term, Y can be replaced with the collection of B partially bootstrapped samples $\{Y^*_{par}(b) = (Y_{obs}, Y^*_{mis}(b))|b = 1,2,\ldots,B\}$ as needed in the bootstrap based estimators.

The maximum likelihood estimates obtained using the EM algorithm can also be used to create parametric bootstrap samples of the entire sample $Y = (Y_{obs}, Y_{mis})$, which are needed for the bootstrap based estimators of the expected optimism. Let $\{Y^*(b)|b = 1,2,\ldots,B\}$ represent a collection of B bootstrap samples generated by taking draws from the fitted candidate model $f(Y|\hat{\theta}_{obs})$. Let $\{\hat{\theta}^*(b)|b = 1,2,\ldots,B\}$ represent a collection of B bootstrap replicates of MLEs of θ corresponding to the B bootstrap samples. In this situation, the bootstrap samples are of the entire sample, consisting of both the observed and missing elements.

Note that the bootstrap samples previously described in this section are generated using a parametric bootstrap. A semi-parametric bootstrap could also be used in some situations, but we have not extensively examined this approach. A non-parametric bootstrap would still likely result in a sample containing missing data since each case is drawn with equal probability from the empirical distribution; hence, the missing data issues would still need to be addressed. Little and Rubin (2002) propose a method to generate a bootstrap sample in incomplete data settings based on non-parametric bootstrap ideas, which we have not investigated.

They obtain a non-parametric bootstrap sample based on the original incomplete data and then fill in the missing values using an imputation procedure. The desired estimates are calculated, and the process is repeated.

Recall that in settings where there is no missing data, EIC employs the following estimator of the expected optimism:

$$\frac{1}{B}\sum_{b=1}^{B}\Big\{-2\log L(\hat{\theta}^*(b)|Y)-\Big(-2\log L(\hat{\theta}^*(b)|Y^*(b))\Big)\Big\}.$$

In constructing the complete data analogue of this estimator, the Y on the left-hand side of the estimator is replaced with the B partially bootstrapped samples as previously described. The right-hand side of the expected optimism needs no modifications since both the bootstrap samples and the bootstrap estimators are already represented in this term. The complete data analogue of EIC is therefore given by

$$\begin{aligned}\mathrm{EIC}_{comp} = &\frac{1}{B}\sum_{b=1}^{B}\Big\{-2\log L(\hat{\theta}^*_{par}(b)|Y^*_{par}(b))\Big\}\\ &+\frac{1}{B}\sum_{b=1}^{B}\Big\{-2\log L(\hat{\theta}^*(b)|Y^*_{par}(b))-\Big(-2\log L(\hat{\theta}^*(b)|Y^*(b))\Big)\Big\}.\end{aligned}$$

Also, recall that in settings where there is no missing data, Cavanaugh and Shumway develop an estimator of the expected optimism used for AICb:

$$2\left[\frac{1}{B}\sum_{b=1}^{B}\Big\{-2\log L(\hat{\theta}^*(b)|Y)\Big\}-\Big(-2\log L(\hat{\theta}|Y)\Big)\right].$$

In constructing the complete data analogue of this estimator, the Y on the left-hand side of the estimator is replaced with the B partially bootstrapped samples as previously described. The right-hand side of the expected optimism is replaced with the complete data goodness-of-fit term, which is analogous to how AICb is constructed when there is no missing data. The complete data analogue of AICb is

therefore given by

$$\text{AICb}_{comp} = \frac{1}{B}\sum_{b=1}^{B}\Big\{ -2\log L(\hat{\theta}^*_{par}(b)|Y^*_{par}(b))\Big\} + 2\left[\frac{1}{B}\sum_{b=1}^{B}\Big\{ -2\log L(\hat{\theta}^*(b)|Y^*_{par}(b)) - \Big(-2\log L(\hat{\theta}^*_{par}(b)|Y^*_{par}(b))\Big)\Big\}\right].$$

We have now proposed the complete data analogues of AIC, AICc, EIC, and AICb. All of the complete data criteria are equivalent to the standard criteria when there is no missing data. The complete data criteria, under the proper assumptions, should have expectations which are approximately equal to the expected complete data KL discrepancy $D_{comp}(k, \theta_o)$.

The complete data criteria can be used to facilitate the selection of a model from a set of potential candidate models. Once a model is selected for inference, we advocate using the estimator of θ obtained via the EM algorithm in constructing the fitted candidate model and in making subsequent inferences. Our goal is to develop model selection criteria that can effectively determine the structural form of a model in the presence of missing. We have not extensively studied the sampling distribution and statistical properties of $\hat{\theta}^*_{par}$, and do not advocate its use in constructing the fitted candidate model.

3.5 Comparison to Multiple Imputation

Multiple imputation was discussed in section (2.3.2) as a useful technique that can provide parameter estimates and standard errors in the presence of missing data. For our proposed methodology, values are imputed for the missing elements many times, which certainly bears a resemblance to multiple imputation. However, there are some fundamental differences between our imputation approach and multiple imputation.

One of the main differences between multiple imputation and our approach is in the generation of the imputed values. In multiple imputation, the missing

values are often imputed with iterative procedures, such as a Gibbs sampling algorithm (as mentioned in subsection (2.3.2)). These procedures use methods that fill in the missing data, and update the estimated parameter values based on the observed data and the completed data. The iterative process is repeated until stable imputed values occur. Our method simply draws values to impute using the distribution of the fitted candidate model, and does not include the sequential aspect of updating estimated parameter values in order to create the draws. In principle, in a multiple imputation procedure, the multiply imputed data sets could be generated many ways; however, the underlying imputation methodology is built on a Bayesian framework. Our proposed methodology uses parameter estimates from the EM algorithm and is based on frequentist methods.

Multiple imputation provides a means of obtaining valid parameter estimates. However, as previously mentioned, the imputed values must be "proper" in order for Rubin's formulas given in section (2.3.2) to be valid (Schafer, 1997). We advocate the use of the parameter estimates from the EM algorithm after our techniques have been applied in selecting the structural form of a model, since the properties of the EM algorithm have been well studied and justified.

Multiple imputation is often seen as a two phase process. The first phase is the imputation phase and the second phase is the analysis phase, where both phases could be completed by different people or organizations. The two phase approach is generally seen as an advantage since the missing data issues are entirely confined to the first phase. When the model used by the imputer and the model used by the analyst have assumptions that agree, valid inferences are obtained via multiple imputation. Schafer (1997, p.140), states "The validity of multiple-imputation inferences when the imputer's and analyst's models differ has been the subject of recent controversy." The controversy is based on understanding the effects on inference when the analyst assumes more than the imputer, or vice versa. In our

methodology, we impute the missing elements using the fitted candidate model of interest, and the complete data criterion value for that candidate model is then calculated. This process is repeated for each candidate model, which essentially combines the two-phase approach and forces the same assumptions for both the analyst and imputer. In principle, our methodology could be implemented using a similar two-phase approach, where a collection of imputed data sets would be generated from a single fitted candidate model (e.g., the largest model in the candidate family). The imputed data sets could then be used in calculating the complete data criterion values for each candidate model, although we have not extensively studied this approach. Similarly in multiple imputation, multiply imputed data sets could be generated and used from each of the candidate models, which is more analogous to our approach.

Multiple imputation is an exciting and popular technique for addressing missing data. Combining some of our ideas with multiple imputation is certainly an area of future research. We chose to use our bootstrapping approach based on the EM algorithm since many frequentist model selection methods for incomplete data have been developed on principles relating to the EM algorithm. The bootstrapping approach also has a natural extension to the criteria AICb and EIC, which feature estimators of the expected optimism based on the bootstrap.

CHAPTER 4
KULLBACK-LEIBLER DISCREPANCY BASED CRITERIA: LINEAR MODELS FRAMEWORK

In this chapter, the proposed methodology is implemented in a linear models framework using a normal multivariate linear regression model. Two important scenarios are considered: missing data in the outcome, and missing data in the outcome and/or covariates. The latter scenario requires some additional distributional assumptions that accommodate the possibility of missing data in the covariates, as well as some methodological modifications to the proposed criteria. A simulation study is presented and discussed for both scenarios.

4.1 Normal Multivariate Linear Regression Model

The normal multivariate linear regression model is often used for exploring linear relationships between covariates and a collection of response outcomes. Suppose we have data Y that has been generated from the normal multivariate linear regression model

$$Y = X_o\beta_o + U_o, \tag{4.1}$$

which denotes the true model. Here, Y is an $n \times m$ response matrix of independent rows corresponding to m response outcomes for n cases, X_o is an $n \times p_o$ known design matrix of full column rank, β_o is a $p_o \times m$ matrix of regression parameters, and U_o is an $n \times m$ error matrix comprised of independent rows which are normal random vectors, each with a mean of 0 and a covariance matrix Σ_o.

Suppose that a normal multivariate linear regression candidate model is postulated of the form

$$Y = X\beta + U. \tag{4.2}$$

Here, Y is as previously described, X is an $n \times p$ known design matrix of full column rank, β is an $p \times m$ matrix of regression parameters, and U is an $n \times m$ error matrix comprised of independent rows which are normal random vectors, each with a mean of 0 and a covariance matrix Σ.

The log-likelihood for the candidate model (4.2), ignoring the constant terms, is given by

$$\log L(\beta, \Sigma|Y) \propto -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\Big((Y - X\beta)\Sigma^{-1}(Y - X\beta)'\Big).$$

Fitting the candidate model requires estimation of β and Σ. Using the method of maximum likelihood, it can be shown that the MLE of β is given by

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{4.3}$$

The MLE of Σ is given by

$$\hat{\Sigma} = \frac{1}{n}(Y - X\hat{\beta})'(Y - X\hat{\beta}), \tag{4.4}$$

which is a negatively biased estimator of Σ.

The normal multivariate linear regression model allows for a $p \times 1$ vector of regression parameters for each of the m response outcomes. This is advantageous in situations where the response outcomes are potentially related, yet one does not want to necessarily model each response outcome using the same regression parameters. For example, suppose that systolic blood pressure, diastolic blood pressure, and a person's weight are three response outcomes of interest. It would be desirable to allow different regression parameters for each outcome, even though the outcomes are potentially related, as opposed to forcing the same linear relationship between each outcome and the covariates.

The main advantage of fitting a normal multivariate linear regression model over fitting separate normal univariate regression models for each of the m outcomes is that hypotheses can be tested between regression parameters comprising the β

matrix. Assume a hypothesis test of the form $a'\beta c = d$ is of interest, where a is a $p \times m$ matrix of known constants, and c and d are m-dimensional vectors of known constants. Such a test can be accommodated through a variety of procedures: e.g. Wilk's Lambda, Pillai trace, Hotelling-Lawley trace, or Roy's maximum root.

Suppose a particular case has missing data for at least one of the outcome variables and/or missing data in the covariates. Such a case will not be included in fitting a candidate model by default in most statistical packages. Hence, a complete case analysis is often the recourse of analysts in a normal multivariate linear regression model setting with missing data.

In chapter 6, we will introduce the normal longitudinal regression model, which allows for a different set of covariate values for each response outcome. In this setting, the outcome data is typically collected in an $nm \times 1$ vector, where often the m response outcomes for a case are the same variable repeatedly measured at m distinct time points. Both the normal longitudinal regression model and the normal multivariate linear regression model are valuable modeling techniques that have applicability in a wide variety of practical settings.

4.2 Missing Data in the Outcome

In the following subsections, (4.2.1), (4.2.2), and (4.2.3), we consider a situation in which there is missing data in the outcome matrix Y. The elements of the design matrix X are assumed to contain *no* missing values.

4.2.1 Criteria Using the Fully Observed Data

As previously stated, a common approach to selecting a statistical model in missing data settings is to simply use a complete case analysis. Recall that the fully and partially observed cases comprising Y can be delineated as $Y = (Y_{fobs}, Y_{pobs})$. The design matrix X in the candidate model (4.2) can also be delineated as $X =$

(X_{fobs}, X_{pobs}), which corresponds to $Y = (Y_{fobs}, Y_{pobs})$. Similarly, the design matrix X_o in the true model (4.1) can be delineated as $X_o = (X_{fobs,o}, X_{pobs,o})$. Let n_{cc} denote the number of complete cases, or equivalently, the number of rows in Y_{fobs}.

A candidate model analogous to (4.2) can be postulated by using Y_{fobs} and X_{fobs} in place of Y and X. The fully observed data log-likelihood for this candidate model is given as

$$\log L(\beta, \Sigma | Y_{fobs}) \propto -\frac{n_{cc}}{2} \log |\Sigma| - \frac{1}{2} \mathrm{tr}\Big((Y_{fobs} - X_{fobs}\beta)\Sigma^{-1}(Y_{fobs} - X_{fobs}\beta)' \Big). \quad (4.5)$$

It can be shown that the MLEs of β and Σ corresponding to (4.5) are given by

$$\hat{\beta}_{fobs} = (X'_{fobs}X_{fobs})^{-1}X'_{fobs}Y_{fobs}, \quad \text{and}$$
$$\hat{\Sigma}_{fobs} = \frac{1}{n_{cc}}(Y_{fobs} - X_{fobs}\hat{\beta}_{fobs})'(Y_{fobs} - X_{fobs}\hat{\beta}_{fobs}).$$

In this setting, the fully observed data KL discrepancy is given by

$$\begin{aligned} & E_o\{-2 \log L(\beta, \Sigma | Y_{fobs})\} \quad (4.6) \\ & = n_{cc} \log |\Sigma| + n_{cc} \mathrm{tr}(\Sigma^{-1}\Sigma_o) \\ & + \mathrm{tr}\Big((X_{fobs,o}\beta_o - X_{fobs}\beta)\Sigma^{-1}(X_{fobs,o}\beta_o - X_{fobs}\beta)' \Big). \end{aligned}$$

To define the overall fully observed data KL discrepancy, which is estimated by the fully observed data criteria, we substitute $\hat{\beta}_{fobs}$ and $\hat{\Sigma}_{fobs}$ as estimators of β and Σ in (4.6). The expected overall fully observed data KL discrepancy is found by averaging the preceding over the sampling distribution of the estimators based on Y_{fobs}.

The fully observed data goodness-of-fit term for the normal multivariate linear regression model simplifies to

$$-2 \log L(\hat{\beta}_{fobs}, \hat{\Sigma}_{fobs} | Y_{fobs}) = n_{cc} \log |\hat{\Sigma}_{fobs}| + m n_{cc}.$$

The four fully observed data model selection criteria outlined in section (3.3) are

now given in the normal multivariate linear regression model setting. Let $k = mp + .5m(m+1)$ denote the number of functionally independent regression parameters in the candidate model. The fully observed data versions of AIC and AICc are given as

$$\text{AIC}_{fobs} = n_{cc} \log |\hat{\Sigma}_{fobs}| + mn_{cc} + 2k, \quad \text{and}$$
$$\text{AICc}_{fobs} = n_{cc} \log |\hat{\Sigma}_{fobs}| + mn_{cc} + 2k(n_{cc}/(n_{cc} - m - p - 1)).$$

To construct the estimators of the expected optimism for the fully observed data bootstrap criteria, bootstrap samples of Y_{fobs} must be obtained. The bootstrap samples could be compiled using either a parametric, semi-parametric, or non-parametric bootstrap. We will outline the generation of the bootstrap samples using a parametric bootstrap, since this is how the bootstrap samples were obtained in the simulation study to be presented in subsection (4.2.3). Thus, the bootstrap samples for the complete data and the fully observed data criteria are both generated using a parametric bootstrap.

Let $y_{fobs,i}$ denote the m-dimensional outcome vector, and let $x_{fobs,i}$ denote the p-dimensional covariate vector, for the fully observed case i. The bootstrap responses $y^*_{fobs,i}$ in a bootstrap sample Y^*_{fobs} are generated case-by-case by taking draws from the distribution of the fitted candidate model, depicted as

$$N((x'_{fobs,i}\hat{\beta}_{fobs})', \hat{\Sigma}_{fobs}), \quad \text{for } i = 1, 2, \ldots, n_{cc}.$$

Let $\{Y^*_{fobs}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y_{fobs} generated as described. Let $\{\hat{\beta}^*_{fobs}(b)|b = 1, 2, \ldots, B\}$ and $\{\hat{\Sigma}^*_{fobs}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap replicates of MLEs of β and Σ, respectively, corresponding to the B bootstrap samples of Y_{fobs}.

The fully observed data version of EIC simplifies to

$$\begin{aligned}\mathrm{EIC}_{fobs} &= n_{cc}\log|\hat{\Sigma}_{fobs}| + mn_{cc} \\ &+ \frac{1}{B}\sum_{b=1}^{B}\Big\{\mathrm{tr}\Big((Y_{fobs} - X_{fobs}\hat{\beta}^{*}_{fobs}(b))\hat{\Sigma}^{*}_{fobs}(b)^{-1}(Y_{fobs} - X_{fobs}\hat{\beta}^{*}_{fobs}(b))'\Big) - mn_{cc}\Big\}.\end{aligned}$$

The fully observed data version of AICb simplifies to

$$\begin{aligned}\mathrm{AICb}_{fobs} &= n_{cc}\log|\hat{\Sigma}_{fobs}| + mn_{cc} \\ &+ 2\Bigg[\frac{1}{B}\sum_{b=1}^{B}\Big\{\mathrm{tr}\Big((Y_{fobs} - X_{fobs}\hat{\beta}^{*}_{fobs}(b))\hat{\Sigma}^{*}_{fobs}(b)^{-1}(Y_{fobs} - X_{fobs}\hat{\beta}^{*}_{fobs}(b))'\Big) \\ &\qquad + n_{cc}\log|\hat{\Sigma}^{*}_{fobs}(b)|\Big\} - \Big(n_{cc}\log|\hat{\Sigma}_{fobs}| + mn_{cc}\Big)\Bigg].\end{aligned}$$

4.2.2 Criteria Using the Complete Data

In the normal multivariate linear regression model setting, the complete data analogues of AIC, AICc, EIC, and AICb are now developed. In this modeling framework, the complete data KL discrepancy is given by

$$\begin{aligned}E_o\{-2\log L(\beta,\Sigma|Y)\} &= n\log|\Sigma| + n\mathrm{tr}(\Sigma^{-1}\Sigma_o) \\ &\quad + \mathrm{tr}\Big((X_o\beta_o - X\beta)\Sigma^{-1}(X_o\beta_o - X\beta)'\Big). \qquad (4.7)\end{aligned}$$

To define the overall complete data KL discrepancy, which is estimated by the complete data criteria, we substitute the maximum likelihood estimators based on Y for β and Σ in (4.7). The expected overall complete data KL discrepancy is found by averaging the preceding over the sampling distribution of the estimators based on Y.

Following the proposed methodology of section (3.4), the complete data goodness-of-fit term can now be calculated. Recalling the delineation of $Y = (Y_{obs}, Y_{mis})$, bootstrap samples of Y_{mis} must be collected in order to construct the complete data goodness-of-fit term. Let $y_i = (y'_{obs,i}, y'_{mis,i})'$ denote the observed and missing

elements, respectively, for the m-dimensional outcome vector for case i, and let x_i denote the corresponding p-dimensional vector of covariates. For a particular fitted candidate model, the distribution of y_i is estimated as $N((x_i'\hat{\beta})', \hat{\Sigma})$, where the estimators $\hat{\beta}$ and $\hat{\Sigma}$ are obtained via the EM algorithm. This distribution can be partitioned corresponding to the observed and missing elements of y_i, and is given by

$$(y_{obs,i}', y_{mis,i}')' \sim N\left((x_i'(\hat{\beta}_{obs,i}, \hat{\beta}_{mis,i}))', \begin{pmatrix} \hat{\Sigma}_{obs,i} & \hat{\Sigma}_{obsmis,i} \\ \hat{\Sigma}_{misobs,i} & \hat{\Sigma}_{mis,i} \end{pmatrix}\right), \text{ for } i = 1, 2, \ldots, n. \tag{4.8}$$

The conditional distribution of $y_{mis,i}|y_{obs,i}$ can now be obtained by using (4.8). In a bootstrap sample Y_{mis}^*, the bootstrap vectors $y_{mis,i}^*$ are generated for those cases that have at least one missing element in Y by taking a draw from the conditional distribution of $y_{mis,i}|y_{obs,i}$, depicted as

$$\begin{aligned} &N(\hat{\mu}_{y_{mis,i}|y_{obs,i}}, \hat{\Sigma}_{y_{mis,i}|y_{obs,i}}), \quad \text{where} \\ &\hat{\mu}_{y_{mis,i}|y_{obs,i}} = (x_i'\hat{\beta}_{mis,i})' + \hat{\Sigma}_{misobs,i}\hat{\Sigma}_{obs,i}^{-1}(y_{obs,i} - (x_i'\hat{\beta}_{obs,i})'), \quad \text{and} \\ &\hat{\Sigma}_{y_{mis,i}|y_{obs,i}} = \hat{\Sigma}_{mis,i} - \hat{\Sigma}_{misobs,i}\hat{\Sigma}_{obs,i}^{-1}\hat{\Sigma}_{obsmis,i}, \text{ for } i = 1, 2, \ldots, n. \end{aligned}$$

Let $\{Y_{mis}^*(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y_{mis} generated as described. Let $\{Y_{par}^*(b) = (Y_{obs}, Y_{mis}^*(b))|b = 1, 2, \ldots, B\}$ represent a collection of B partially bootstrapped samples of Y. Let $\{\hat{\beta}_{par}^*(b)|b = 1, 2, \ldots, B\}$ and $\{\hat{\Sigma}_{par}^*(b)|b = 1, 2, \ldots, B\}$ represent a collection of B partial bootstrap replicates of MLEs of β and Σ, corresponding to the B partially bootstrapped samples of Y.

The complete data goodness-of-fit term is now given as

$$\begin{aligned} \text{GOF}_{comp} &= \frac{1}{B}\sum_{b=1}^{B} -2\log L(\hat{\beta}_{par}^*(b), \hat{\Sigma}_{par}^*(b)|Y_{par}^*(b)) \\ &= \frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}_{par}^*(b)| + mn\Big\}. \end{aligned} \tag{4.9}$$

The four complete data criteria proposed in section (3.4) are now given, which should have expectations that are approximately equal to the expected complete data KL discrepancy. Let $k = mp + .5m(m+1)$ denote the number of functionally independent regression parameters in the candidate model. The complete data analogues of AIC and AICc are

$$\text{AIC}_{comp} = \frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}^*_{par}(b)| + mn\Big\} + 2k, \quad \text{and} \tag{4.10}$$

$$\text{AICc}_{comp} = \frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}^*_{par}(b)| + mn\Big\} + 2k(n/(n-m-p-1)). \tag{4.11}$$

In order to construct the estimators of the expected optimism in the complete data bootstrap criteria, bootstrap samples of both the observed and missing elements in Y must be obtained. The bootstrap vectors y_i^* in a bootstrap sample Y^* are generated case-by-case by taking draws from the distribution of the fitted candidate model given in (4.8). Let $\{Y^*(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y generated as described. Let $\{\hat{\beta}^*(b)|b = 1, 2, \ldots, B\}$ and $\{\hat{\Sigma}^*(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap replicates of MLEs of β and Σ, corresponding to the B bootstrap samples of Y.

The complete data analogue of EIC simplifies to

$$\begin{aligned}\text{EIC}_{comp} = &\frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}^*_{par}(b)| + mn\Big\} \\ &+ \frac{1}{B}\sum_{b=1}^{B}\Big\{\text{tr}\Big((Y^*_{par}(b) - X\hat{\beta}^*(b))\hat{\Sigma}^*(b)^{-1}(Y^*_{par}(b) - X\hat{\beta}^*(b))'\Big) - mn\Big\}.\end{aligned} \tag{4.12}$$

The complete data analogue of AICb simplifies to

$$\begin{aligned}\text{AICb}_{comp} = &\frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}^*_{par}(b)| + mn\Big\} \\ &+ 2\Bigg[\frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}^*(b)| + \text{tr}\Big((Y^*_{par}(b) - X\hat{\beta}^*(b))\hat{\Sigma}^*(b)^{-1}(Y^*_{par}(b) - X\hat{\beta}^*(b))'\Big) \\ &\quad - (n\log|\hat{\Sigma}^*_{par}(b)| + mn)\Big\}\Bigg]. \qquad (4.13)\end{aligned}$$

4.2.3 Simulation Study

Consider a setting where a sample of size n is generated according to model (4.1). Suppose our objective is to search among a class of nested candidate models for the fitted candidate model that best approximates (4.1). The candidate models are assumed to be of the form (4.2).

We consider a setting where Y is an $n \times 2$ matrix, so the rows of Y represent independent bivariate data pairs. Let y_1 denote the response represented in the first column of Y, and let y_2 denote the response represented in the second column of Y.

In model (4.2), recall that X is an $n \times p$ known design matrix of full column rank p. Consider a class of 8 nested candidate models where the number of covariates range from 1 to 8, corresponding to design matrices having from $p = 2$ to $p = 9$ columns. We will refer to p as the order of the candidate model. For the simulation study, the first column of X was taken to be a vector of ones. The covariate values of X were generated independently from a $N(0, 5)$ distribution. Generating the covariates for the design matrices in this simplistic fashion ensures the simulation results are not unduly influenced by such factors as multicollinearity and high-leverage cases.

Simulation sets with 500 samples of data of size $n = 30$ or $n = 60$ were

generated from a model with one of two types of covariance structures,

$$Y = X_o\beta_o + U_o, \text{ where } \Sigma_o = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} \text{ or } \Sigma_o = \begin{bmatrix} 10 & 8 \\ 8 & 10 \end{bmatrix},$$

and where β_o is a 5×2 matrix consisting of all ones. In the first covariance structure, $\text{corr}(y_1, y_2) = 0$, and in the second covariance structure, $\text{corr}(y_1, y_2) = .8$.

The order of the true model, or true order p_o, is five. In the nested model setting that we are considering, the true model includes the first four covariates. An underspecified model would include the first covariate, the first two covariates, or the first three covariates. An overspecified model would include the first five covariates, the first six covariates, the first seven covariates, or all eight covariates. All candidate models include an intercept.

For some of the cases, y_1 or y_2 were randomly discarded so as to create a situation where the missing data in Y are MAR. Let $\Pr(y_1 \text{ mis})$ denote the probability that for a particular case, y_1 is discarded and y_2 is retained, and let $\Pr(y_2 \text{ mis})$ denote the probability that for a particular case, y_2 is discarded and y_1 is retained. Thus, in a simulation set where $(\Pr(y_1 \text{ mis}), \Pr(y_2 \text{ mis})) = (0.15, 0.15)$, for each sample, one would expect roughly 15% of the pairs to have y_1 missing and y_2 observed, 15% of the pairs to have y_2 missing and y_1 observed, and 70% of the pairs to have y_1 and y_2 both observed. We considered discard probabilities of $(\Pr(y_1 \text{ mis}), \Pr(y_2 \text{ mis}))$ set at (0.0, 0.0), (0.075, 0.075), and (0.15, 0.15).

In summary, simulation results from a total of 12 different settings are investigated. We consider all possible combinations of the following factors: two different sample sizes of $n = 30$ and $n = 60$, two different correlation levels of $\text{corr}(y_1, y_2) = 0$ and $\text{corr}(y_1, y_2) = .8$, and three levels of missingness of $(\Pr(y_1 \text{ mis}), \Pr(y_2 \text{ mis}))$ set at (0.0, 0.0), (0.075, 0.075), and (0.15, 0.15).

For each of the 12 settings, the 8 candidate models were fit for each sample. The complete data and fully observed data analogues of AIC, AICc, EIC, and AICb

were then calculated for every set of fitted candidate models. The overall complete data KL discrepancy, which is found by substituting the maximum likelihood estimators based on Y for β and Σ in (4.7), along with the overall fully observed data KL discrepancy, which is obtained by substituting the maximum likelihood estimators based on Y_{fobs} for β and Σ in (4.6), were also computed under the 12 settings for each sample and candidate model. The model selected by each of the fully observed data criteria and complete data criteria, along with each of the target discrepancies, was determined. For every candidate model, a total of 500 bootstrap samples were generated for EIC and AICb, along with 500 partially bootstrapped samples for the complete data criteria.

We use two performance measures to characterize the effectiveness of the complete and fully observed data criteria:

- the distribution of the model selections for each of the criterion, and
- the average criterion values for each fitted candidate model, which determines how well each criterion estimates its intended target: i.e., the expected KL discrepancy based on either the complete or fully observed data.

The results from the 12 simulation settings are summarized in figures (4.1), (4.2), and (4.3), and in tables (4.1), (4.2), (4.3), and (4.4).

Figures (4.1), (4.2), and (4.3) illustrate the effectiveness of the criteria in estimating their target discrepancies. For $n = 30$, the mean values of AICc_{comp} and EIC_{comp} are roughly equivalent and the corresponding criterion curves cannot be differentiated. In general, these criteria are closest to the complete data target among the complete data criteria. The slopes of the AICc_{comp} and EIC_{comp} curves generally parallel the complete data target. For overfitted models, the mean value of AICb_{comp} is generally a little too high and the slope of the AICb_{comp} curve is too steep. However, this positive bias appears to aid in the selection of the true model

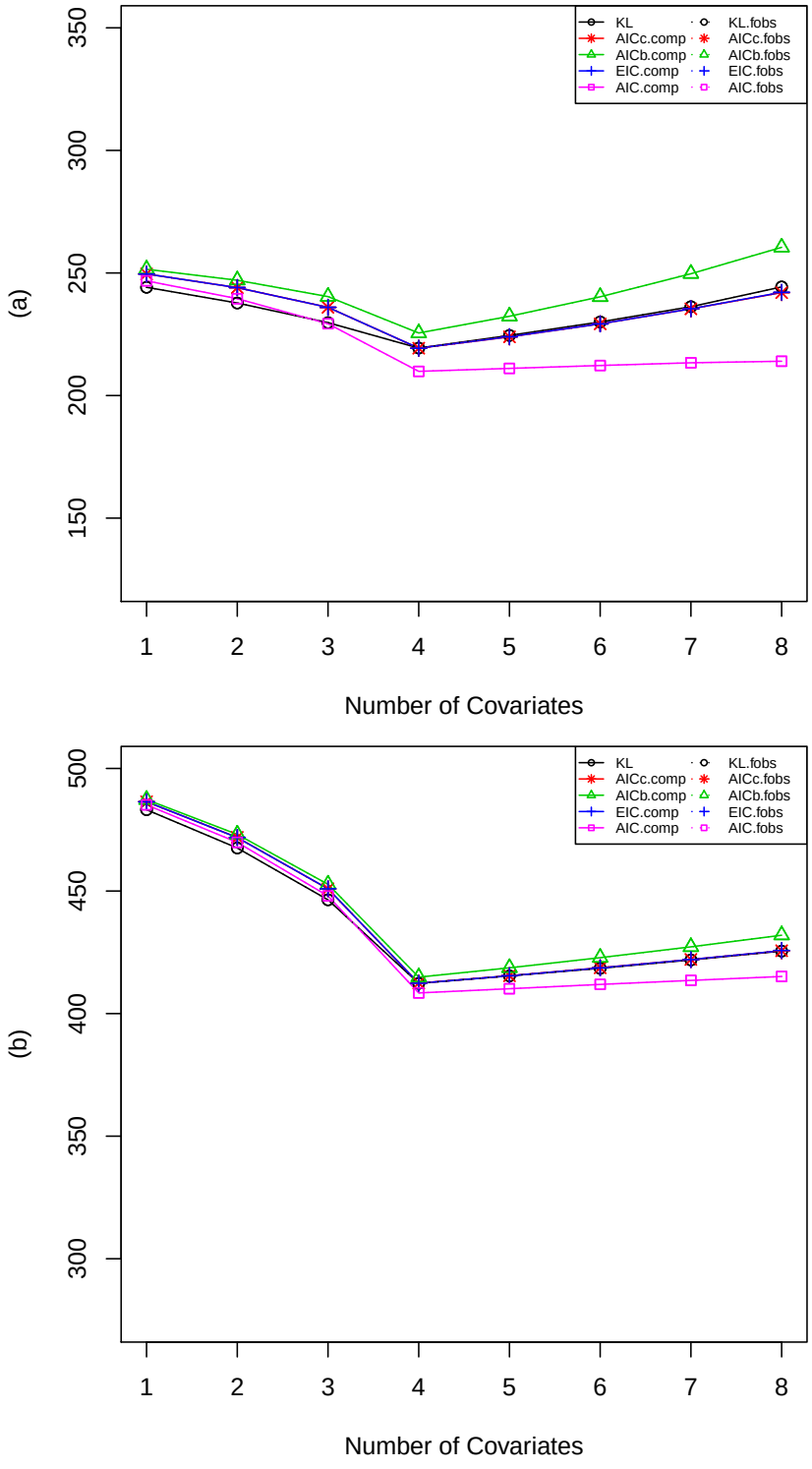

Figure 4.1: Average criterion values under bivariate linear regression setting with missing data in the outcome: $(\Pr(y_1 \text{ mis}), \Pr(y_2 \text{ mis})) = (0, 0), \text{corr}(y_1, y_2) = 0$. (a) n=30; (b) n=60.

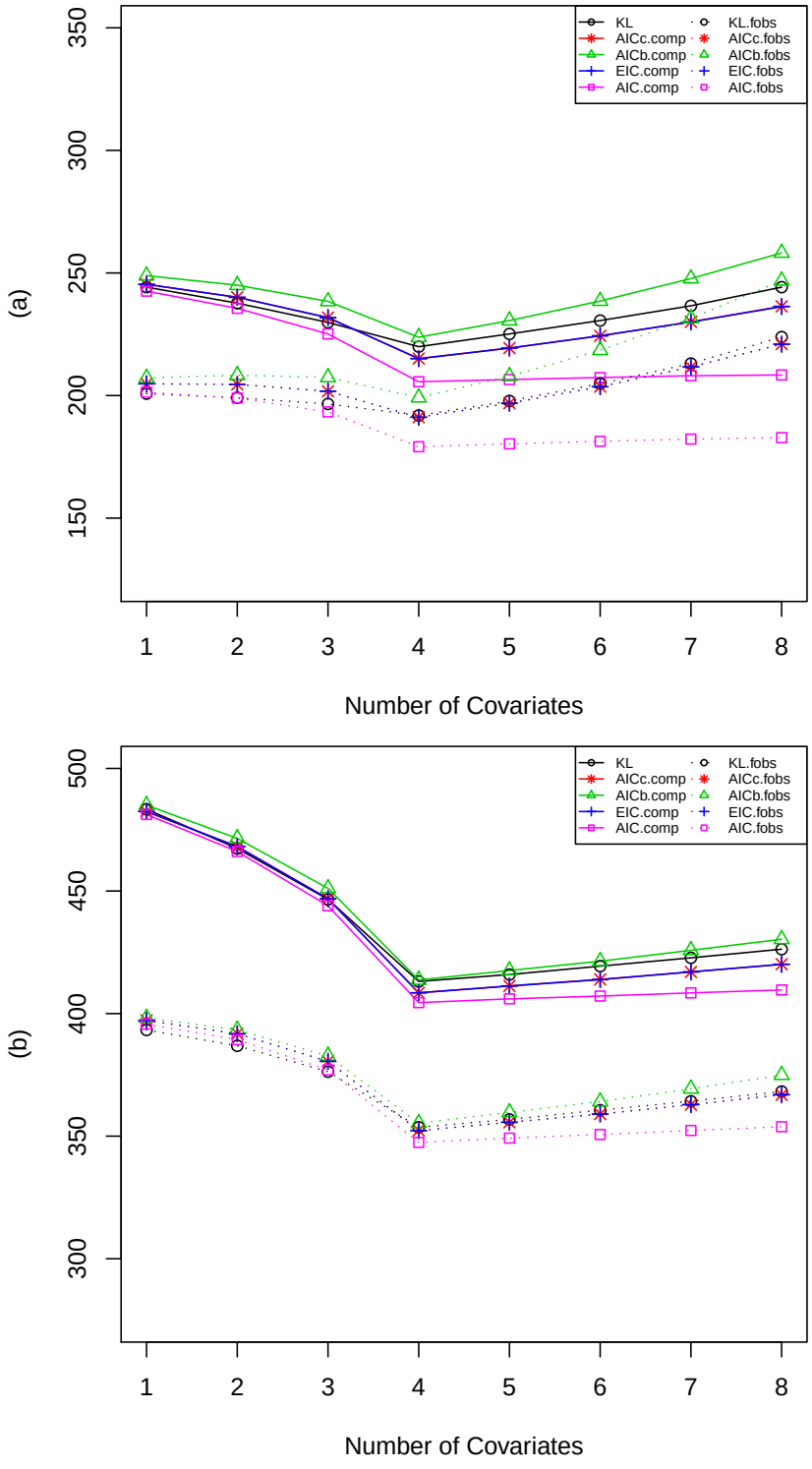

Figure 4.2: Average criterion values under bivariate linear regression setting with missing data in the outcome: $(\Pr(y_1 \text{ mis}), \Pr(y_2 \text{ mis})) = (.075, .075), \text{corr}(y_1, y_2) = 0$. (a) n=30; (b) n=60.

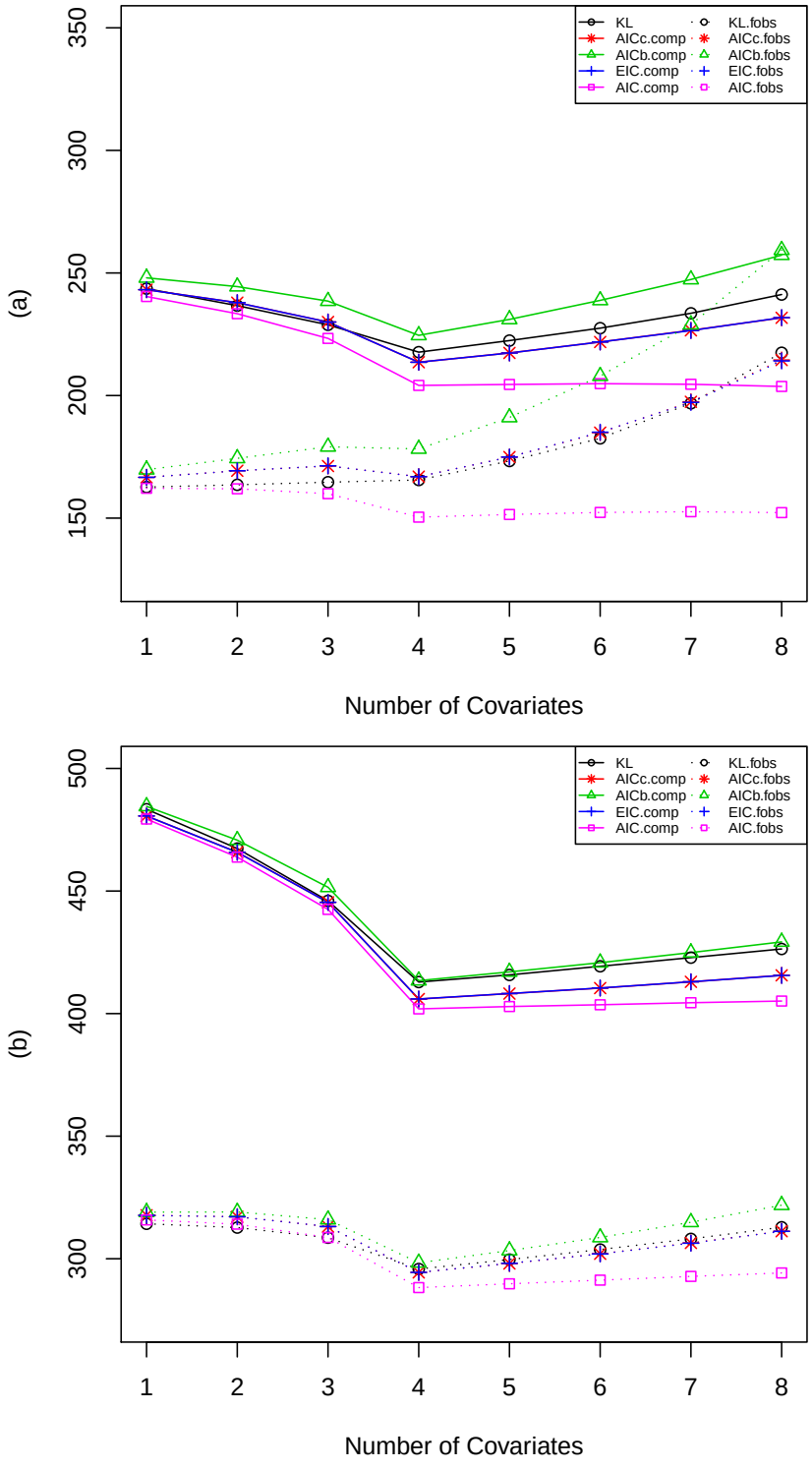

Figure 4.3: Average criterion values under bivariate linear regression setting with missing data in the outcome: $(\Pr(y_1 \text{ mis}), \Pr(y_2 \text{ mis})) = (.15, .15), \text{corr}(y_1, y_2) = 0$. (a) n=30; (b) n=60.

Table 4.1: Frequency of criterion selections under bivariate linear regression setting with missing data in the outcome: $\text{corr}(y_1, y_2) = 0, n = 30$.

$\Pr(y_1$ mis),		Complete Data					Fully Observed Data				
$\Pr(y_2$ mis)	Covariates	KL	AICc	AIC	EIC	AICb	KL	AICc	AIC	EIC	AICb
(0, 0)	1	5	1	0	1	1	5	1	0	1	1
	2	9	0	0	0	1	9	0	0	1	2
	3	37	7	2	8	15	37	7	2	10	22
	4	449	451	289	445	463	449	451	289	436	454
	5	0	33	75	36	18	0	33	75	44	19
	6	0	6	42	8	2	0	6	42	7	2
	7	0	1	32	1	0	0	1	32	0	0
	8	0	1	60	1	0	0	1	60	1	0
(.075, .075)	1	3	0	0	0	2	43	30	1	33	92
	2	12	2	0	2	5	28	11	2	14	25
	3	45	5	2	6	18	78	22	4	21	31
	4	440	439	257	441	451	351	414	288	402	344
	5	0	35	79	34	21	0	18	63	25	8
	6	0	13	48	12	3	0	3	51	3	0
	7	0	3	42	2	0	0	1	40	1	0
	8	0	3	72	3	0	0	1	51	1	0
(.15, .15)	1	3	0	0	0	4	129	215	18	213	351
	2	7	2	0	2	8	67	46	7	51	45
	3	35	8	2	10	26	62	28	9	29	17
	4	455	399	203	394	430	242	205	222	197	85
	5	0	47	73	50	22	0	5	72	9	2
	6	0	14	43	14	2	0	1	36	0	0
	7	0	10	69	13	4	0	0	45	1	0
	8	0	20	110	17	4	0	0	91	0	0

Table 4.2: Frequency of criterion selections under bivariate linear regression setting with missing data in the outcome: $\text{corr}(y_1, y_2) = .8, n = 30$.

$\Pr(y_1$ mis),		Complete Data					Fully Observed Data				
$\Pr(y_2$ mis)	Covariates	KL	AICc	AIC	EIC	AICb	KL	AICc	AIC	EIC	AICb
(0, 0)	1	17	9	1	14	26	17	9	1	12	28
	2	24	16	1	13	36	24	16	1	15	31
	3	52	56	18	52	65	52	56	18	56	74
	4	407	381	291	376	359	407	381	291	380	355
	5	0	31	72	38	14	0	31	72	31	9
	6	0	4	48	5	0	0	4	48	2	1
	7	0	3	36	2	0	0	3	36	4	2
	8	0	0	33	0	0	0	0	33	0	0
(.075, .075)	1	10	9	0	8	23	76	116	18	116	216
	2	25	14	2	13	38	53	53	12	53	73
	3	61	31	8	40	62	65	52	21	57	57
	4	404	388	240	376	353	306	251	244	245	146
	5	0	43	86	48	22	0	24	76	25	7
	6	0	11	49	9	2	0	2	34	3	1
	7	0	3	50	4	0	0	2	39	1	0
	8	0	1	65	2	0	0	0	56	0	0
(.15, .15)	1	9	3	0	5	30	225	331	66	328	425
	2	19	6	1	8	26	69	52	19	53	44
	3	55	39	6	35	65	45	31	28	29	9
	4	417	360	171	357	344	161	79	165	80	21
	5	0	52	64	54	24	0	6	50	9	1
	6	0	20	51	20	8	0	1	38	1	0
	7	0	12	96	16	3	0	0	60	0	0
	8	0	8	111	5	0	0	0	74	0	0

Table 4.3: Frequency of criterion selections under bivariate linear regression setting with missing data in the outcome: $\text{corr}(y_1, y_2) = 0, n = 60$.

$\Pr(y_1$ mis),		Complete Data					Fully Observed Data				
$\Pr(y_2$ mis)	Covariates	KL	AICc	AIC	EIC	AICb	KL	AICc	AIC	EIC	AICb
(0, 0)	1	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0
	4	500	436	370	419	456	500	436	370	423	455
	5	0	43	63	56	32	0	43	63	55	35
	6	0	12	29	15	8	0	12	29	12	7
	7	0	7	16	7	2	0	7	16	7	2
	8	0	2	22	3	2	0	2	22	3	1
(.075, .075)	1	0	0	0	0	0	1	0	0	0	0
	2	0	0	0	0	0	1	0	0	0	0
	3	0	0	0	0	0	2	0	0	0	0
	4	500	419	337	403	447	496	447	372	439	466
	5	0	38	58	53	32	0	33	56	37	26
	6	0	33	49	33	18	0	17	42	18	7
	7	0	5	27	7	2	0	2	16	5	1
	8	0	5	29	4	1	0	1	14	1	0
(.15, .15)	1	0	0	0	0	0	8	1	0	0	6
	2	0	0	0	0	0	9	2	0	1	3
	3	1	0	0	0	0	18	3	3	5	3
	4	499	372	274	342	429	465	445	341	430	456
	5	0	50	68	79	49	0	30	68	43	23
	6	0	40	62	38	17	0	14	34	16	9
	7	0	23	43	22	3	0	4	29	3	0
	8	0	15	53	19	2	0	1	25	2	0

Table 4.4: Frequency of criterion selections under bivariate linear regression setting with missing data in the outcome: $\text{corr}(y_1, y_2) = .8, n = 60$.

$\Pr(y_1$ mis),		Complete Data					Fully Observed Data				
$\Pr(y_2$ mis)	Covariates	KL	AICc	AIC	EIC	AICb	KL	AICc	AIC	EIC	AICb
(0, 0)	1	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0
	3	5	1	0	1	2	5	1	0	0	2
	4	495	434	353	427	452	495	434	353	421	450
	5	0	49	78	54	38	0	49	78	55	39
	6	0	9	30	9	5	0	9	30	16	6
	7	0	6	22	7	3	0	6	22	5	3
	8	0	1	17	2	0	0	1	17	3	0
(.075, .075)	1	0	0	0	0	0	2	1	0	1	2
	2	2	0	0	0	0	8	0	0	0	0
	3	2	2	0	2	2	15	6	2	8	9
	4	496	407	326	379	435	475	436	349	420	447
	5	0	46	63	72	46	0	43	58	51	33
	6	0	25	35	24	12	0	10	34	13	9
	7	0	10	37	13	3	0	3	29	4	0
	8	0	10	39	10	2	0	1	28	3	0
(.15, .15)	1	0	0	0	0	0	41	41	12	46	72
	2	1	0	0	0	0	17	13	7	15	21
	3	4	0	0	1	3	33	19	8	21	27
	4	495	352	263	332	407	409	379	320	361	351
	5	0	76	80	88	58	0	38	64	46	27
	6	0	37	53	40	22	0	7	43	7	2
	7	0	19	45	21	8	0	3	24	4	0
	8	0	16	59	18	2	0	0	22	0	0

since overspecified models are less likely to be chosen. For $n = 60$, the slope of the AICb_{comp} curve reflects an attenuation in bias, and the mean value of AICb_{comp} is the closest to the complete data target among the complete data criteria. The mean value of AIC_{comp} is generally too low at $n = 30$, but as the sample size increases to $n = 60$, the mean value more closely approaches the complete data target. Similar trends in the mean criterion values hold for the fully observed data criteria, yet the trends are often more marked.

The figures representing the average criterion values for the settings when $\text{corr}(y_1, y_2)$=.8 are not included since the figures bear a strong resemblance to figures (4.1), (4.2), and (4.3). For the models that are correctly specified or overspecified, the slopes of the criterion curves are very similar at both correlation levels. For the models that are underspecified, the slopes are flatter for the high correlation level of $\text{corr}(y_1, y_2)$=.8, which results in a tendency for the criteria to select more underspecified models.

Tables (4.1), (4.2), (4.3), and (4.4) show the distribution of the model selections for each of the criterion. With $n = 30$, the complete data criteria, with the exception of AIC_{comp}, consistently outperform their fully observed data criteria counterparts at all settings in selecting the correct model. The improvement of the complete data criteria over the fully observed data criteria is greatest at the high level of correlation and at higher percentages of missingness. The selections of the complete data criteria and the fully observed data criteria are both worse at the high level of correlation, compared to the level of no correlation. However, the complete data criteria demonstrate less degradation than the fully observed data criteria as the correlation increases. Also, as the percentage of missing data increases, the performance of the fully observed data criteria deteriorates more rapidly than that of the complete data criteria. The fully observed discrepancy target is greatly affected by an increase in the amount of missingness; consequently, the criteria that

estimate this target tend to under perform in selecting the correct order. The fully observed data criteria exhibit a pronounced tendency to select more models that are underspecified as the amount of missingness increases. The complete data criteria protect against underfitting, and are less affected by an increase in the amount of missingness.

With $n = 60$ and $\text{corr}(y_1, y_2) = 0$, the fully observed data criteria outperform the complete data criteria. With $n = 60$ and $\text{corr}(y_1, y_2) = .8$, most of the fully observed data criteria perform marginally better than the complete data criteria. However, the fully observed data criteria still show a tendency to select models that are underspecified as the amount of missingness increases. The complete data criteria protect against underfitting, and do not exhibit a higher propensity to underfit as the amount of missingess increases. The criterion AICb_{comp} outperforms its fully observed counterpart at the highest level of missingness.

In summary, it reasonable to conclude that in small sample settings, or in other situations where less information is contained in the data (e.g, large error variances/correlations relative to the variance of the covariates), failing to use the information that is contained in the partially observed cases may result in a tendency to select an underspecified model. In these settings, the complete data criteria may provide more reliable selection results than the fully observed data criteria.

4.3 Missing Data in the Outcome and/or Covariates

In the previous section (4.2), for the normal multivariate linear regression framework, we presented the complete data criteria that can be used in selecting a model when there is missing data in the outcome. A common occurrence in many research settings is to have missing data in the outcome and/or covariates. Potentially, this may be a much more serious problem since the number of complete cases could diminish very quickly with even a small amount of missing data in a few

of the covariates. In the following subsections, (4.3.1), (4.3.2), (4.3.3), and (4.3.4), we consider a setting where there is missing data in the outcome and/or covariates.

4.3.1 Criteria Using the Fully Observed Data

As mentioned in subsection (4.2.1), a common method for selecting the structural form of a statistical model with missing data in the outcome is to use a complete case analysis. In multivariate linear regression, this approach is often employed when there is missing data in the outcome and/or covariates. Again, the outcome Y can be delineated as $Y = (Y_{fobs}, Y_{pobs})$. Accordingly, the design matrix X can be delineated as $X = (X_{fobs}, X_{pobs})$, which corresponds to the fully and partially observed cases. Here, Y_{fobs} and X_{fobs} contain the data for the n_{cc} complete cases. The criteria AIC, AICc, EIC, and AICb can be computed using the complete cases and are available as outlined in subsection (4.2.1), with the distinction that a complete case is now determined by a case with no missing data in both the outcome and the covariates.

4.3.2 Model Assumptions and the Sweep Algorithm

Recall the normal multivariate linear regression candidate model (4.2) presented at the beginning of the chapter. This model makes the assumption that the error matrix U is an $n \times m$ matrix comprised of independent rows which are normal random vectors, each with a mean of 0 and a covariance matrix Σ.

In the presence of missing data in the outcome and/or covariates, further distributional assumptions are necessary in order to utilize our proposed methodology. In the present setting, Little and Rubin (2002, p.239) propose employing the assumption of multivariate normality for the outcome and the covariates when either contains missing data. More specifically, for case i, recall that y_i represents the m-dimensional outcome vector from Y, and x_i is the corresponding p-dimensional

covariate vector from X. The distribution of the outcome and covariates can be partitioned as

$$(y_i', x_i')' \sim N\left((\mu_Y', \mu_X')', \begin{pmatrix} \Sigma_Y & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_X \end{pmatrix}\right), \quad \text{for } i = 1, 2, \ldots, n.$$

Note that generally the first column of X is a vector of ones corresponding to the intercept. If this is the case, the first column of X would not be included in the preceding multivariate normality assumption.

The EM algorithm can be used to obtain MLEs for the multivariate normal distribution. The MLEs can be arranged in what is denoted as the estimated augmented covariance matrix, which is shown as

$$\begin{bmatrix} -1 & \hat{\mu}_Y' & \hat{\mu}_X' \\ \hat{\mu}_Y & \hat{\Sigma}_Y & \hat{\Sigma}_{YX} \\ \hat{\mu}_X & \hat{\Sigma}_{XY} & \hat{\Sigma}_X \end{bmatrix}.$$

Returning to the candidate model of interest in (4.2), recall that fitting the candidate model requires estimation of β and Σ. The estimated augmented covariance matrix, along with what is known as the sweep operator (Beaton, 1964; Dempster, 1969), can be used to provide parameter estimators for the candidate model. The sweep operator has a close connection with multivariate linear regression. The parameter estimators are found by sweeping out the parameters in the estimated augmented covariance matrix that are associated with the covariates in the candidate model. Using the described approach, the estimator of β is given by

$$\hat{\beta} = \begin{pmatrix} \hat{\mu}_Y' - \hat{\mu}_X' \hat{\Sigma}_X^{-1} \hat{\Sigma}_{XY} \\ \hat{\Sigma}_X^{-1} \hat{\Sigma}_{XY} \end{pmatrix}.$$

The estimator of Σ is given by

$$\hat{\Sigma} = \hat{\Sigma}_Y - \hat{\Sigma}_{YX} \hat{\Sigma}_X^{-1} \hat{\Sigma}_{XY}.$$

Under the appropriate conditions, the parameter estimators in the estimated augmented covariance matrix are MLEs which are invariant under transformation; thus, the parameter estimators of $\hat{\beta}$ and $\hat{\Sigma}$ are also MLEs.

The assumption of multivariate normality of the outcome and covariates may seem somewhat restrictive. Little and Rubin (2002, p.239) explain that in certain situations, the assumption can be relaxed and yet MLEs can still be provided. A less restrictive framework could accommodate settings where some of the covariates are categorical variables or other variable types.

4.3.3 Criteria Using the Complete Data

A method that allows for fitting a candidate model of the form (4.2) in the presence of missing data in the outcome and/or covariates was described in the preceding subsection (4.3.2). We now outline how the four complete data criteria are calculated in the presence of missing data in the outcome and/or covariates. The criteria estimate the overall complete data KL discrepancy, which was described in subsection (4.2.2). In general, the formulas for the complete data analogues of AIC, AICc, EIC, and AICb are very similar to those given in subsection (4.2.2). However, we outline the necessary adjustments that need to be made in order to accommodate the possibility of missing data in the outcome and/or covariates.

Recall that Y can be delineated as $Y = (Y_{obs}, Y_{mis})$, and the design matrix X can also be delineated as $X = (X_{obs}, X_{mis})$, corresponding to the observed and the missing elements. For simplicity in notation, we define $Z = (Y, X)$. The observed and missing elements in Z can now be partitioned as $Z = (Z_{obs}, Z_{mis})$, where $Z_{obs} = (Y_{obs}, X_{obs})$ and $Z_{mis} = (Y_{mis}, X_{mis})$. In order to construct the complete data goodness-of-fit term, bootstrap samples of $Z_{mis} = (Y_{mis}, X_{mis})$ must be obtained.

Let $z_i = (z'_{obs,i}, z'_{mis,i})'$ denote the observed and missing elements, respectively, for case i from Z. Assuming a multivariate normal distribution for the outcome

and covariates in the candidate model, the EM algorithm can be used to obtain parameter estimates. This distribution can be partitioned corresponding to the observed and missing elements of z_i, and is given by

$$(z'_{obs,i}, z'_{mis,i})' \sim N\left((\hat{\mu}'_{z_{obs,i}}, \hat{\mu}'_{z_{mis,i}})', \begin{pmatrix} \hat{\Sigma}_{z_{obs,i}} & \hat{\Sigma}_{z_{obsmis,i}} \\ \hat{\Sigma}_{z_{misobs,i}} & \hat{\Sigma}_{z_{mis,i}} \end{pmatrix}\right), \quad \text{for } i = 1, 2, \ldots, n. \tag{4.14}$$

The conditional distribution of $z_{mis,i}|z_{obs,i}$ can now be obtained by using (4.14). The bootstrap vectors $z^*_{mis,i}$ in a bootstrap sample $Z^*_{mis} = (Y^*_{mis}, X^*_{mis})$ are generated for those cases that have at least one missing element in Y or X by taking a draw from the conditional distribution of $z_{mis,i}|z_{obs,i}$, depicted as

$$\begin{aligned} &N(\hat{\mu}_{z_{mis,i}|z_{obs,i}}, \hat{\Sigma}_{z_{mis,i}|z_{obs,i}}), \quad \text{where} \\ &\hat{\mu}_{z_{mis,i}|z_{obs,i}} = \hat{\mu}_{z_{mis,i}} + \hat{\Sigma}_{z_{misobs,i}}\hat{\Sigma}^{-1}_{z_{obs,i}}(z_{obs,i} - \hat{\mu}_{z_{obs,i}}), \quad \text{and} \\ &\hat{\Sigma}_{z_{mis,i}|z_{obs,i}} = \hat{\Sigma}_{z_{mis,i}} - \hat{\Sigma}_{z_{misobs,i}}\hat{\Sigma}^{-1}_{z_{obs,i}}\hat{\Sigma}_{z_{obsmis,i}}, \quad \text{for } i = 1, 2, \ldots, n. \end{aligned}$$

Let $\{Y^*_{mis}(b)|b = 1, 2, \ldots, B\}$ and $\{X^*_{mis}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y_{mis} and X_{mis} generated as described. Let $\{Y^*_{par}(b) = (Y_{obs}, Y^*_{mis}(b))|b = 1, 2, \ldots, B\}$ and $\{X^*_{par}(b) = (X_{obs}, X^*_{mis}(b))|b = 1, 2, \ldots, B\}$ represent a collection of B partially bootstrapped samples of Y and X. Let $\{\hat{\beta}^*_{par}(b)|b = 1, 2, \ldots, B\}$ and $\{\hat{\Sigma}^*_{par}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B partial bootstrap replicates of MLEs of β and Σ calculated by using the B partially bootstrapped samples of Y and X.

The components needed to construct the complete data goodness-of-fit term have now been provided. The formula for the complete data goodness-of-fit term with missing data in the outcome and/or covariates is identical to equation (4.9), but here we use the collection of B partial bootstrap replicates of MLEs of β and Σ as described in the previous paragraphs.

The formula for the complete data analogues of AIC and AICc with missing

data in the outcome and/or covariates are identical to (4.10) and (4.11), respectively, but again we use the collection of B partial bootstrap replicates of MLEs of β and Σ as previously described.

In a typical normal multivariate linear regression model setting, the covariate values of X are often treated as "fixed" known values. Following this logic, the partially bootstrapped samples of X, $\{X^*_{par}(b) = (X_{obs}, X^*_{mis}(b))|b = 1, 2, \ldots, B\}$, can be treated as fixed in order to generate the bootstrap samples of Y. Let $x^*_{par,i}(b)$ represent the m-dimensional vector for case i from the b^{th} partially bootstrapped sample $X^*_{par}(b)$. This vector would contain all of the observed covariate data for case i, in addition to the bootstrapped elements for the missing covariate data. In a bootstrap sample Y^*, the bootstrap vectors y^*_i are generated case-by-case by taking draws from the distribution of

$$N((x^*_{par,i}(b))'\hat{\beta}, \hat{\Sigma}) \quad \text{for } i = 1, 2, \ldots, n,$$

using the parameter estimators from the fitted candidate model. Let $\{Y^*(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y generated as described. Let $\{\hat{\beta}^*(b)|b = 1, 2, \ldots, B\}$ and $\{\hat{\Sigma}^*(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap replicates of MLEs of β and Σ, calculated using the B bootstrap samples of Y and the B partially bootstrapped samples of X.

The formulas for the complete data analogues of EIC and AICb with missing data in the outcome and/or covariates are very similar to equations (4.12) and (4.13) given in subsection (4.2.2), where missing data is confined to the outcome, but are included here for completeness.

The complete data analogue of EIC simplifies to

$$\begin{aligned}\mathrm{EIC}_{comp} &= \frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}^*_{par}(b)| + mn\Big\} \\ &+ \frac{1}{B}\sum_{b=1}^{B}\Big\{\mathrm{tr}\Big((Y^*_{par}(b) - X^*_{par}(b)\hat{\beta}^*(b))\hat{\Sigma}^*(b)^{-1}(Y^*_{par}(b) - X^*_{par}(b)\hat{\beta}^*(b))'\Big) - mn\Big\}.\end{aligned}$$

The complete data analogue of AICb simplifies to

$$\begin{aligned}\mathrm{AICb}_{comp} &= \frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}^*_{par}(b)| + mn\Big\} \\ &\quad + 2\Bigg[\frac{1}{B}\sum_{b=1}^{B}\Big\{\mathrm{tr}\Big((Y^*_{par}(b) - X^*_{par}(b)\hat{\beta}^*(b))\hat{\Sigma}^*(b)^{-1}(Y^*_{par}(b) - X^*_{par}(b)\hat{\beta}^*(b))'\Big) \\ &\qquad + n\log|\hat{\Sigma}^*(b)| - (n\log|\hat{\Sigma}^*_{par}(b)| + mn)\Big\}\Bigg].\end{aligned}$$

4.3.4 Simulation Study

As in the simulation study presented in subsection (4.2.3), consider a setting where a sample of size n is generated according to the true model (4.1). We will employ a framework where the outcome Y is an n-dimensional vector; hence, the true model and candidate models are simply univariate linear regression models. Suppose our objective is to search among a class of nested candidate models for the fitted candidate model that best approximates (4.1), where candidate models are assumed to be of the form (4.2).

We again consider a class of 8 nested candidate models, where the number of covariates range from 1 to 8, corresponding to design matrices ranging in size from $p = 2$ to $p = 9$ columns. The design matrix X was constructed in the exact same fashion as the simulation study in subsection (4.2.3), where the first column of X was taken to be a vector of ones, and the covariate values of X were generated independently from a $N(0, 5)$ distribution.

Simulation sets with 500 samples of data of size $n = 30$ or $n = 60$ were

generated from a model of the form

$$Y = X_o\beta_o + U_o,$$

where U_o is an n-dimensional error vector comprised of independent normal random variables, each with a mean of 0 and a variance of σ_o^2, and β_o is a 5×1 vector consisting of all ones. We generated simulation sets with either $\sigma_o^2 = 10$ or $\sigma_o^2 = 15$.

The order of the true model, or true order p_o, is five. In the nested model setting that we are considering, the true model includes the first four covariates. An underspecified model would include the first covariate, the first two covariates, or the first three covariates. An overspecified model would include the first five covariates, the first six covariates, the first seven covariates, or all eight covariates. All candidate models include an intercept.

Let x_1 and x_2 respectively denote the covariate values for the first and second covariates. These covariates are represented in the second and third columns of the design matrix X. For some of the cases, x_1 or x_2 were randomly discarded so as to create a situation where the missing data in X are MAR. Let $\Pr(x_1 \text{ mis})$ denote the probability that for a particular case, x_1 is discarded and x_2 is retained, and let $\Pr(x_2 \text{ mis})$ denote the probability that for a particular case, x_2 is discarded and x_1 is retained. We considered discard probabilities of $(\Pr(x_1 \text{ mis}), \Pr(x_2 \text{ mis}))$ set at (0.0, 0.0), (0.075, 0.075), and (0.15, 0.15).

In summary, simulation results from a total of 12 different settings are provided. We consider all possible combinations of the following factors: two different sample sizes of $n = 30$ and $n = 60$, two different error variance levels of $\sigma_o^2 = 10$ and $\sigma_o^2 = 15$, and three levels of missingness of $(\Pr(x_1 \text{ mis}), \Pr(x_2 \text{ mis}))$ set at (0.0, 0.0), (0.075, 0.075), and (0.15, 0.15).

As in the simulation study presented in subsection (4.2.3), for each of the 12 settings, the 8 candidate models were fit for each sample. The complete data and

fully observed data analogues of AIC, AICc, EIC, and AICb were then calculated for every set of fitted candidate models. The overall complete data KL discrepancy, which is found by substituting the maximum likelihood estimators based on Y for β and Σ in (4.7), along with the overall fully observed data KL discrepancy, which is obtained by substituting the maximum likelihood estimators based on Y_{fobs} for β and Σ in (4.6), were also computed under the 12 settings for each sample and candidate model. The model selected by each of the fully observed data criteria and complete data criteria, along with the selections of both discrepancies, was determined. For every candidate model, a total of 500 bootstrap samples were generated for EIC and AICb, along with 500 partially bootstrapped samples for the complete data criteria.

To characterize the effectiveness of the complete and fully observed data criteria, we use the same two performance measures given in section (4.2.3): the average criterion values and distribution of model selections. The results from the 12 simulation settings are summarized in figures (4.4), (4.5), and (4.6), and in tables (4.5), (4.6), (4.7), and (4.8).

Figures (4.4), (4.5), and (4.6) illustrate the effectiveness of the criteria in estimating their target discrepancies. In general, the mean values of the complete data criteria, and the slopes of the criterion curves relative to the complete data target, bear a strong resemblance to the results under the settings of the simulation study in subsection (4.2.3). For $n = 30$, the mean values of AICc_{comp} and EIC_{comp} are closest to the complete data target. For the overfitted models, the mean value of AICb_{comp} is generally a little too high and the slope of the AICb_{comp} curve is slightly steeper than the complete data target. When $n = 60$, all of the complete data criteria estimate the complete data target reasonably well, with the exception of AIC_{comp}. Similar trends in the mean criterion values and slopes of the curves hold for the fully observed data criteria, yet the trends are often more pronounced,

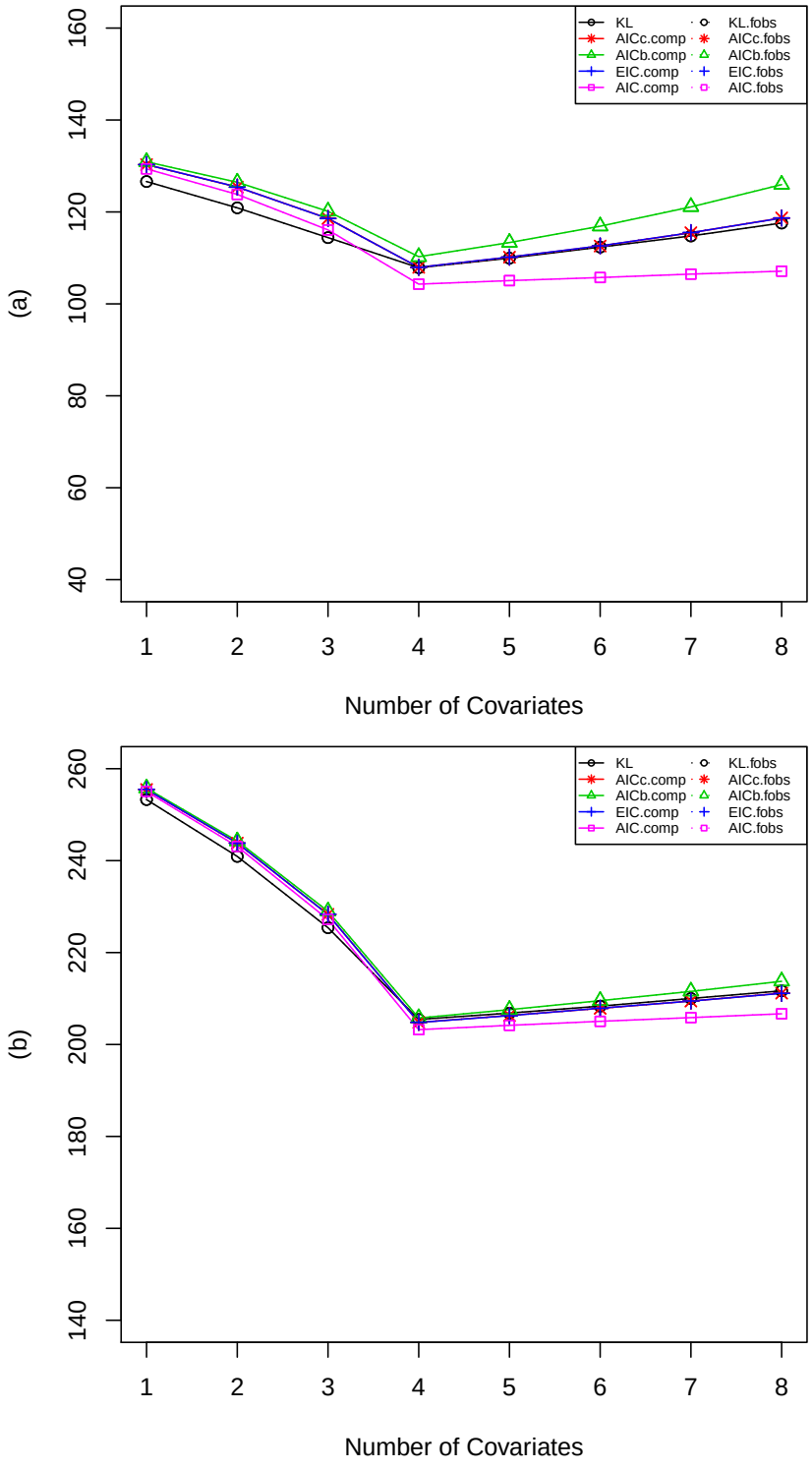

Figure 4.4: Average criterion values under univariate linear regression setting with missing data in the covariates: $(\Pr(x_1 \text{ mis}), \Pr(x_2 \text{ mis})) = (0, 0)$, $\sigma_o^2 = 10$. (a) n=30; (b) n=60.

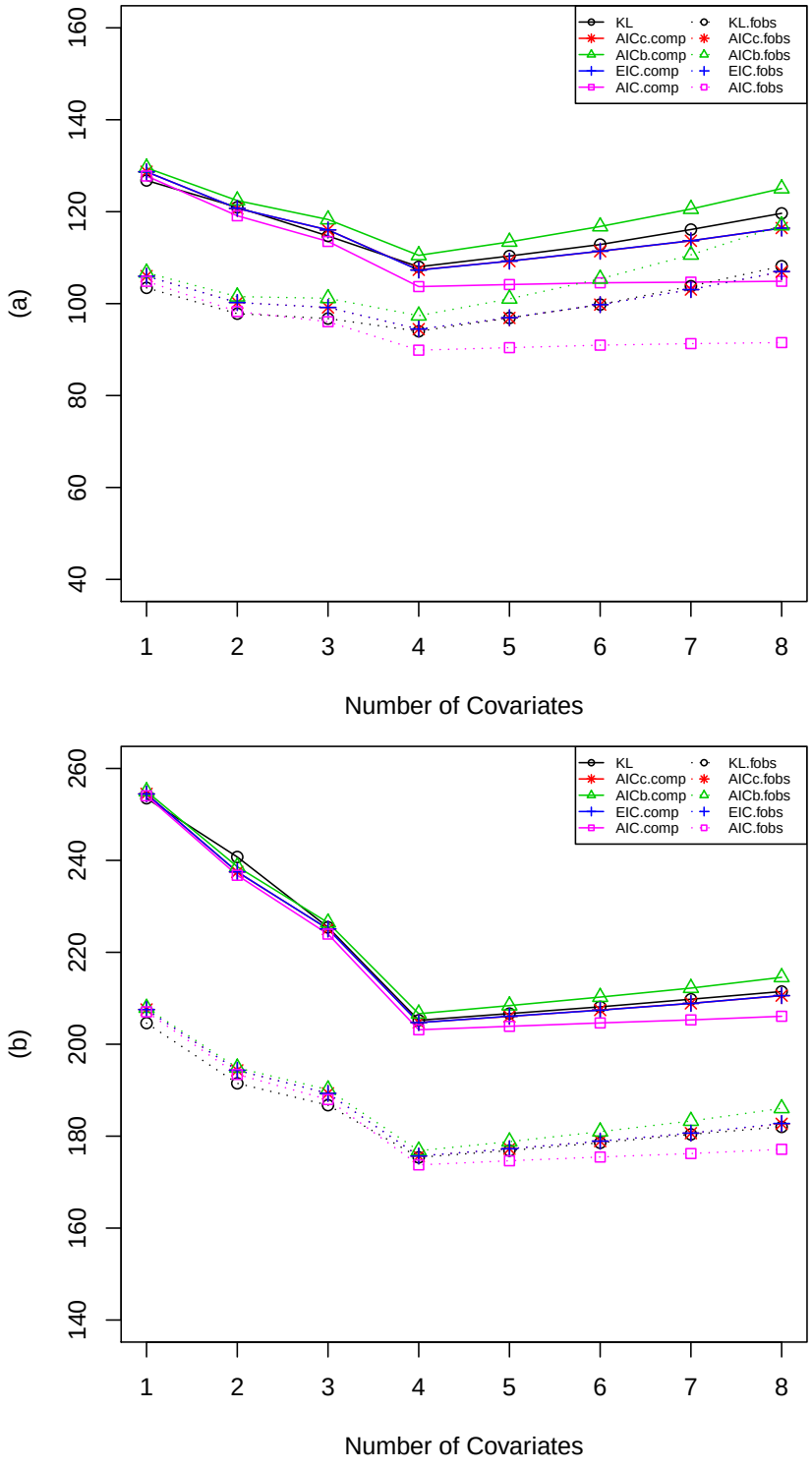

Figure 4.5: Average criterion values under univariate linear regression setting with missing data in the covariates: $(\Pr(x_1 \text{ mis}), \Pr(x_2 \text{ mis})) = (.075, .075)$, $\sigma_o^2 = 10$. (a) n=30; (b) n=60.

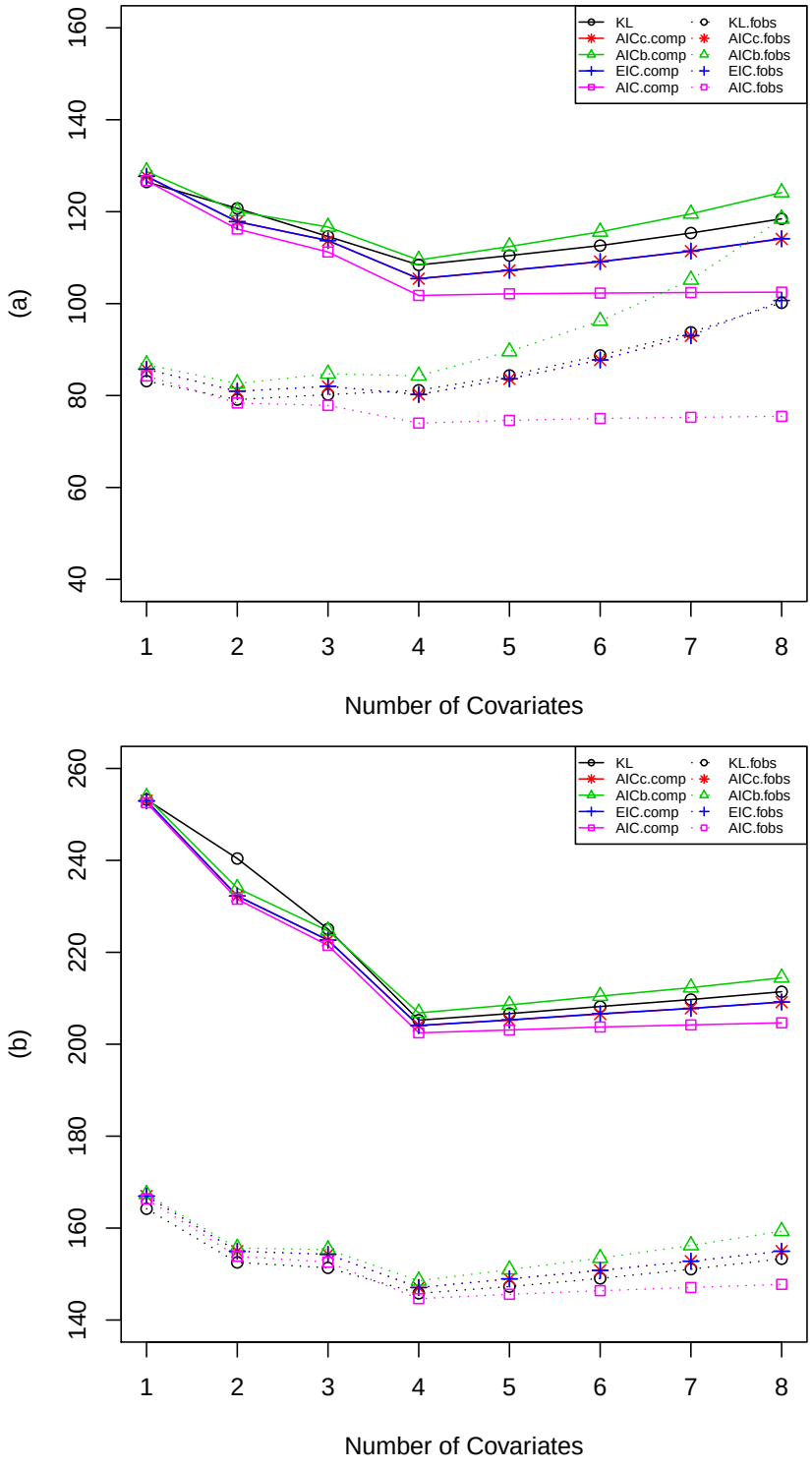

Figure 4.6: Average criterion values under univariate linear regression setting with missing data in the covariates: $(\Pr(x_1 \text{ mis}), \Pr(x_2 \text{ mis})) = (.15, .15)$, $\sigma_o^2 = 10$. (a) n=30; (b) n=60.

Table 4.5: Frequency of criterion selections under univariate linear regression setting with missing data in the covariates: $\sigma_o^2 = 10$, $n = 30$.

Pr(x_1 mis),		Complete Data					Fully Observed Data				
Pr(x_2 mis)	Covariates	KL	AICc	AIC	EIC	AICb	KL	AICc	AIC	EIC	AICb
(0, 0)	1	2	1	1	1	2	2	1	1	1	1
	2	4	4	1	4	5	4	4	1	2	4
	3	37	22	8	23	31	37	22	8	23	33
	4	457	402	308	395	418	457	402	308	398	418
	5	0	44	68	49	31	0	44	68	47	31
	6	0	16	45	18	10	0	16	45	18	11
	7	0	8	32	7	2	0	8	32	5	1
	8	0	3	37	3	1	0	3	37	6	1
(.075, .075)	1	3	0	0	0	3	9	13	1	11	25
	2	6	6	2	4	10	39	62	23	62	87
	3	35	30	13	32	44	53	54	22	61	77
	4	456	353	245	351	377	399	298	232	296	276
	5	0	58	71	60	42	0	47	69	42	28
	6	0	24	44	21	14	0	15	35	15	4
	7	0	18	58	19	9	0	8	52	10	3
	8	0	11	67	13	1	0	3	66	3	0
(.15, .15)	1	9	1	0	2	4	40	53	14	48	79
	2	9	16	1	13	21	123	187	75	189	220
	3	26	40	19	40	63	70	49	30	50	46
	4	456	322	223	317	341	267	184	160	183	143
	5	0	49	55	55	37	0	17	59	18	9
	6	0	36	53	31	22	0	7	46	7	3
	7	0	16	58	23	7	0	1	49	2	0
	8	0	20	91	19	5	0	2	67	3	0

Table 4.6: Frequency of criterion selections under univariate linear regression setting with missing data in the covariates: $\sigma_o^2 = 15$, $n = 30$.

Pr(x_1 mis),		Complete Data					Fully Observed Data				
Pr(x_2 mis)	Covariates	KL	AICc	AIC	EIC	AICb	KL	AICc	AIC	EIC	AICb
(0, 0)	1	3	5	0	5	9	3	5	0	5	9
	2	16	17	3	20	24	16	17	3	14	25
	3	56	43	26	46	60	56	43	26	48	63
	4	425	367	281	351	365	425	367	281	357	359
	5	0	39	67	49	33	0	39	67	43	30
	6	0	14	49	16	4	0	14	49	20	10
	7	0	7	37	7	2	0	7	37	9	1
	8	0	8	37	6	3	0	8	37	4	3
(.075, .075)	1	2	5	3	6	14	28	35	15	39	66
	2	12	33	11	28	46	68	102	36	96	132
	3	58	52	33	57	69	74	68	47	75	80
	4	428	337	229	322	323	330	250	217	240	202
	5	0	42	67	52	29	0	28	52	36	15
	6	0	13	44	13	8	0	8	40	7	3
	7	0	11	52	13	6	0	7	42	6	2
	8	0	7	61	9	5	0	2	51	1	0
(.15, .15)	1	3	8	2	8	11	56	111	25	109	156
	2	20	27	9	34	61	164	175	78	177	216
	3	44	59	31	57	78	63	67	57	71	51
	4	433	296	202	287	297	217	120	137	112	68
	5	0	49	59	49	31	0	18	45	22	5
	6	0	24	49	29	11	0	7	37	7	4
	7	0	22	67	21	8	0	1	60	1	0
	8	0	15	81	15	3	0	1	61	1	0

Table 4.7: Frequency of criterion selections under univariate linear regression setting with missing data in the covariates: $\sigma_o^2 = 10$, $n = 60$.

Pr(x_1 mis),		Complete Data					Fully Observed Data				
Pr(x_2 mis)	Covariates	KL	AICc	AIC	EIC	AICb	KL	AICc	AIC	EIC	AICb
(0, 0)	1	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0
	3	2	0	0	0	0	2	0	0	0	0
	4	498	405	345	374	418	498	405	345	384	419
	5	0	45	64	69	46	0	45	64	65	48
	6	0	28	42	34	23	0	28	42	24	19
	7	0	16	26	17	11	0	16	26	21	10
	8	0	6	23	6	2	0	6	23	6	4
(.075, .075)	1	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	4	0	0	0	1
	3	2	0	0	0	0	6	4	2	5	8
	4	498	392	316	359	407	490	414	352	385	418
	5	0	45	67	69	52	0	41	53	63	42
	6	0	31	43	39	25	0	25	37	25	19
	7	0	16	39	22	10	0	8	29	14	7
	8	0	16	35	11	6	0	8	27	8	5
(.15, .15)	1	0	0	0	0	0	1	0	0	0	0
	2	0	0	0	0	0	19	40	14	42	60
	3	3	2	2	2	2	23	25	13	26	29
	4	497	338	257	322	376	457	364	304	359	363
	5	0	63	76	74	63	0	39	57	37	27
	6	0	42	56	40	23	0	16	45	21	9
	7	0	32	54	36	25	0	9	36	8	7
	8	0	23	55	26	11	0	7	31	7	5

Table 4.8: Frequency of criterion selections under univariate linear regression setting with missing data in the covariates: $\sigma_o^2 = 15$, $n = 60$.

Pr(x_1 mis),		Complete Data					Fully Observed Data				
Pr(x_2 mis)	Covariates	KL	AICc	AIC	EIC	AICb	KL	AICc	AIC	EIC	AICb
(0, 0)	1	0	1	0	1	0	0	1	0	1	1
	2	0	0	0	0	0	0	0	0	0	0
	3	0	1	0	2	2	0	1	0	1	2
	4	500	386	321	363	396	500	386	321	378	402
	5	0	61	70	68	62	0	61	70	60	54
	6	0	23	37	31	17	0	23	37	27	19
	7	0	21	42	24	16	0	21	42	24	16
	8	0	7	30	11	7	0	7	30	9	6
(.075, .075)	1	0	0	0	0	0	0	2	1	2	2
	2	0	0	0	0	0	4	14	12	12	17
	3	2	1	1	1	2	15	8	6	10	13
	4	498	372	315	367	397	481	380	321	374	391
	5	0	56	64	63	57	0	54	56	53	44
	6	0	42	53	38	30	0	29	46	34	24
	7	0	19	37	20	8	0	9	32	11	7
	8	0	10	30	11	6	0	4	26	4	2
(.15, .15)	1	0	0	0	0	0	7	9	3	8	11
	2	0	0	0	0	0	38	93	46	86	112
	3	0	7	5	10	13	29	24	25	32	29
	4	500	362	293	340	377	426	304	276	289	297
	5	0	57	63	71	63	0	43	57	54	37
	6	0	36	50	41	26	0	17	34	19	9
	7	0	23	36	21	14	0	5	31	6	4
	8	0	15	53	17	7	0	5	28	6	1

especially when $n = 30$.

The figures representing the average criterion values for the settings where $\sigma_o^2 = 15$ are not included, since the figures share a great likeness to figures (4.4), (4.5), and (4.6). The slopes of the criterion curves for the correctly specified or overspecified models are very similar at both $\sigma_o^2 = 10$ and $\sigma_o^2 = 15$. For the models that are underspecified, the slopes are flatter for $\sigma_o^2 = 15$, which results in a tendency for the criteria to select more underspecified models.

Tables (4.5), (4.6), (4.7), and (4.8) show the distribution of the model selections for each of the criterion. Again, the trends in selections mimic those of the simulation study presented in subsection (4.2.3). At the smaller sample size of $n = 30$, the complete data criteria demonstrate the ability to select correctly specified models more frequently than their fully observed data counterparts. As the amount of missingness and the variance increases, the fully observed data criteria exhibit a marked propensity to select more models that are underspecified.

With $n = 60$ and $\sigma_o^2 = 10$, the fully observed data criteria generally perform marginally better than the complete data criteria in terms of selecting the correct model. However, the fully observed data criteria tend to select more underspecified models as the amount of missingess increases. At the highest level of missingness, the strength of the complete data criteria is evident in the protection against underfitting.

With $n = 60$ and $\sigma_o^2 = 15$, the complete data criteria and fully observed data criteria perform similarly as far as selecting the correct model at the moderate level of missingness. However, at the highest level of missingess, the compete data criteria select the correct model more often and select fewer underspecified models.

We conclude that in small sample settings or other situations where less information is contained in the data (e.g., large error variance compared to the variance of the covariates), calculating model selection criteria using only the fully observed

data may result in an increased likelihood of selecting an underspecified model. The complete data criteria may provide better selection results than the fully observed data criteria in these settings.

CHAPTER 5
KULLBACK-LEIBLER DISCREPANCY BASED CRITERIA: GENERALIZED LINEAR MODELS FRAMEWORK

The generalized linear modeling framework encompasses a large variety of statistical applications and can accommodate many types of continuous, discrete, and categorical response data. To investigate the utility of the proposed model selection criteria within this framework, the criteria are developed and implemented in the setting of a baseline-category logit model with missing data in the outcome. A simulation study is presented and evaluated in this setting. We also discuss how to develop the complete data criteria in the presence of missing data in the outcome and/or covariates.

5.1 Baseline-Category Logit Model

Suppose an $n \times 2$ dimensional data matrix Z has been collected for n cases. Here, Z is a matrix of independent rows, where each row is a pair of potentially correlated binary random variables that can assume a value of either 1 or 0. Let $z_i = (z_{i1}, z_{i2})$ denote the outcome pair for case i from Z. A multinomial quadruple $y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})$ can be defined that designates the four possible outcomes for $z_i = (z_{i1}, z_{i2})$, where one of the four elements in y_i is 1 and the other three elements are 0. The data can be represented in terms of either z_i or the corresponding y_i, as illustrated in the following 2×2 table:

	$z_{i2} = 1$	$z_{i2} = 0$
$z_{i1} = 1$	y_{i1}	y_{i2}
$z_{i1} = 0$	y_{i3}	y_{i4}

The position in the vector y_i that is equal to 1 is determined by the values of z_{i1} and z_{i2}. For example, if $z_{i1} = 1$ and $z_{i2} = 1$ then $y_i = (y_{i1} = 1, y_{i2} = 0, y_{i3} = 0, y_{i4} = 0)$. The outcomes of the multinomial trials for the n cases can be collected in an $n \times 4$

matrix Y. In summary, each row in Z is a pair of potentially correlated binary random variables, and each row in Y is a quadruple representing the outcome of a multinomial trial.

Suppose that each y_i in Y is generated from the distribution

$$Multinomial\Big(1, (\pi_{o,i1}, \pi_{o,i2}, \pi_{o,i3}, \pi_{o,i4})\Big), \quad \text{where} \tag{5.1}$$
$$\pi_{o,ij} = \frac{exp(x'_{o,i}\beta_{o,j})}{1 + \sum_{h=1}^{3} exp(x'_{o,i}\beta_{o,h})}, \quad \text{for } i = 1, 2, \ldots, n; j = 1, 2, 3.$$

Here, $x_{o,i}$ denotes the p_o-length vector of covariates for case i, which can be collected to form an $n \times p_o$ known design matrix X_o of full column rank; $\beta_{o,1}$, $\beta_{o,2}$, and $\beta_{o,3}$ each denote a p_o-length vector of regression parameters, which can be collected to form a $p_o \times 3$ matrix $\beta_o = (\beta_{o,1}, \beta_{o,2}, \beta_{o,3})$. Note that $\pi_{o,i4} = 1 - \pi_{o,i1} - \pi_{o,i2} - \pi_{o,i3}$.

A baseline-category logit model which denotes the true model is given as

$$\log \frac{\pi_{o,ij}}{\pi_{o,i4}} = x'_{o,i}\beta_{o,j}, \quad \text{for } i = 1, 2, \ldots, n; j = 1, 2, 3. \tag{5.2}$$

Since the true model is unknown, a baseline-category logit candidate model for modeling the data Y generated from (5.1) is postulated. This model can be written as

$$\log \frac{\pi_{ij}}{\pi_{i4}} = x'_i\beta_j, \quad \text{where} \tag{5.3}$$
$$\pi_{ij} = \frac{\exp(x'_i\beta_j)}{1 + \sum_{h=1}^{3} \exp(x'_i\beta_h)}, \quad \text{for } i = 1, 2, \ldots, n; j = 1, 2, 3.$$

Here, x_i denotes the p-length vector of covariates for case i, which can be collected to form an $n \times p$ known design matrix X of full column rank; β_1, β_2, and β_3 each denote a p-length vector of regression parameters, which can be collected to form a $p \times 3$ matrix $\beta = (\beta_1, \beta_2, \beta_3)$. Note that $\pi_{i4} = 1 - \pi_{i1} - \pi_{i2} - \pi_{i3}$.

The log-likelihood for the baseline-category logit candidate model (5.3) is found by substituting the expressions for π_{ij} under the candidate model into the log-likelihood for a multinomial distribution based on n trials (see Agresti, 2002,

p.272-273). The log-likelihood has the form

$$\sum_{i=1}^{n}\sum_{j=1}^{4} y_{ij}\log\pi_{ij} = \sum_{i=1}^{n}\left[\sum_{j=1}^{3} y_{ij}(x_i'\beta_j) - \log\left(1+\sum_{j=1}^{3}\exp(x_i'\beta_j)\right)\right].$$

Fitting the candidate (5.3) model requires estimation of β. The maximum likelihood estimator of β can be found by using the Newton-Raphson algorithm.

The baseline-category logit model is a useful approach for modeling a collection of pairs of binary random variables by using the outcomes of the corresponding multinomial trials. Following Agresti (2002), the logits of each of the outcome categories (y_{i1}, y_{i2}, y_{i3}) are paired with the baseline category of y_{i4}, where the logits are denoted as $\log\pi_{i1}/\pi_{i4}$, $\log\pi_{i2}/\pi_{i4}$, $\log\pi_{i3}/\pi_{i4}$, for $i = 1, 2, \ldots, n$. The baseline-category logit model determines the effect of the covariates on the logits.

Suppose a case has missing data for at least one of the binary variables and/or missing data in the covariates. Such a case will not be included in fitting a candidate model by default in most statistical packages. Hence, a complete case analysis is often the recourse of analysts in a baseline-category logit model setting with missing data.

5.2 Missing Data in the Outcome

In the following subsections, (5.2.1), (5.2.2), and (5.2.3) we consider a situation in which an outcome in each pair of binary variables could potentially be missing, and we assume that design matrix X contains *no* missing values. If a given case has a missing element for either z_{i1} or z_{i2}, but not both, the case falls under one of four possible missing data scenarios: $(z_{i1} = ., z_{i2} = 1)$, $(z_{i1} = ., z_{i2} = 0)$, $(z_{i1} = 0, z_{i2} = .)$, or $(z_{i1} = 1, z_{i2} = .)$. (Here, a period denotes a missing value.) The observed information in the correlated binary random variables gives partial information for the outcome of the corresponding multinomial trial. For example, consider the fourth missing data scenario of $(z_{i1} = 1, z_{i2} = .)$. By examining the 2×2 table in

section (5.1), one can determine that the corresponding multinomial quadruple is $y_i = (y_{i1} = ., y_{i2} = ., y_{i3} = 0, y_{i4} = 0)$. A value of 1 would be in position y_{i1} or y_{i2}, determined by the value of z_{i2} if it was hypothetically observed.

5.2.1 Criteria Using the Fully Observed Data

As previously stated, a common approach for fitting a baseline-category logit model in the presence of missing data is to simply use a complete case analysis. The fully and partially observed cases can be delineated as $Z = (Z_{fobs}, Z_{pobs})$, or in terms of the corresponding Y as $Y = (Y_{fobs}, Y_{pobs})$. In this situation, Z_{fobs} contains the outcome data for all of the subjects who have observed elements for both binary variables, and Z_{pobs} contains the outcome data for all of the cases who have observed elements for only one of the binary variables. Similarly, the data Y_{fobs} contains the outcomes of the multinomial trials for all cases who have observed elements for both binary variables. The data Y_{pobs} consists of a collection of partially observed outcomes of the multinomial trials, where each quadruple has two zero values and two missing values, according to the value of the binary variable that is observed. The design matrix X in the candidate model (5.3) can be delineated as $X = (X_{fobs}, X_{pobs})$, and the design matrix X_o in the true model (5.2) can be similarly delineated as $X_o = (X_{fobs,o}, X_{pobs,o})$. The delineations of X and X_o correspond to $Y = (Y_{fobs}, Y_{pobs})$, or equivalently, $Z = (Z_{fobs}, Z_{pobs})$. Let n_{cc} denote the number of complete cases.

A candidate model analogous to equation (5.3) can be postulated by using Y_{fobs} and X_{fobs} in place of Y and X. Let $y_{fobs,ij}$ denote the element for case i and column j from the outcome matrix Y_{fobs}, and let $x_{fobs,i}$ denote the respective p-length vector of covariates from X_{fobs}. The fully observed data log likelihood for

the candidate model is given by

$$\sum_{i=1}^{n_{cc}}\sum_{j=1}^{4} y_{fobs,ij}\log\pi_{ij} = \sum_{i=1}^{n_{cc}}\left[\sum_{j=1}^{3} y_{fobs,ij}(x'_{fobs,i}\beta_j) - \log\left(1+\sum_{j=1}^{3}\exp(x'_{fobs,i}\beta_j)\right)\right]. \tag{5.4}$$

The maximum likelihood estimate of β corresponding to (5.4) can be found by using the Newton-Raphson algorithm, and can be depicted in terms of its column vectors as $\hat{\beta}_{fobs} = (\hat{\beta}_{fobs,1},\, \hat{\beta}_{fobs,2},\, \hat{\beta}_{fobs,3})$.

In this setting, the fully observed data KL discrepancy is given by

$$E_o\left\{-2\sum_{i=1}^{n_{cc}}\left[\sum_{j=1}^{3} y_{fobs,ij}(x'_{fobs,i}\beta_j) - \log\left(1+\sum_{j=1}^{3}\exp(x'_{fobs,i}\beta_j)\right)\right]\right\} \tag{5.5}$$

$$= -2\sum_{i=1}^{n_{cc}}\left[\sum_{j=1}^{3}\pi_{o,ij}(x'_{fobs,i}\beta_j) - \log\left(1+\sum_{j=1}^{3}\exp(x'_{fobs,i}\beta_j)\right)\right], \quad \text{where}$$

$$\pi_{o,ij} = \frac{exp(x'_{fobs,o,i}\beta_{o,j})}{1+\sum_{h=1}^{3} exp(x'_{fobs,o,i}\beta_{o,h})}, \quad \text{for } i = 1,2,\ldots,n_{cc}; j = 1,2,3.$$

To define the overall fully observed data KL discrepancy, which is estimated by the fully observed data criteria, we substitute $\hat{\beta}_{fobs}$ as an estimator of β in (5.5). The expected overall fully observed data KL discrepancy is found by averaging the preceding over the sampling distribution of the estimator based on Y_{fobs}.

The fully observed data model selection criteria outlined in section (3.3) are now given in the baseline-category logit model setting. Let $k = 3p$ denote the number of functionally independent parameters in the candidate model. The fully observed data criterion of AIC is given by

$$\text{AIC}_{fobs} = -2\sum_{i=1}^{n_{cc}}\left[\sum_{j=1}^{3} y_{fobs,ij}(x'_{fobs,i}\hat{\beta}_{fobs,j}) - \log\left(1+\sum_{j=1}^{3}\exp(x'_{fobs,i}\hat{\beta}_{fobs,j})\right)\right] + 2k.$$

Note that AICc has not been justified in the baseline-category logit model setting.

To construct the estimators of the expected optimism for the fully observed data bootstrap criteria, bootstrap samples of Y_{fobs} must be collected. The bootstrap samples could be obtained using a parametric, semi-parametric, or non-parametric

bootstrap. We will outline the generation of the bootstrap samples using a parametric bootstrap since this is how the bootstrap samples were obtained in the simulation study to be described in subsection (5.2.3).

In a parametric bootstrap sample Y^*_{fobs}, the bootstrap multinomial trials $y^*_{fobs,i} = (y^*_{fobs,i1}, y^*_{fobs,i2}, y^*_{fobs,i3}, y^*_{fobs,i4})$, are generated case-by-case from the multinomial distribution evaluated at the parameter estimates from the fitted candidate model. This distribution is given by

$$Multinomial\Big(1, (\hat{\pi}_{fobs,i1}, \hat{\pi}_{fobs,i2}, \hat{\pi}_{fobs,i3}, \hat{\pi}_{fobs,i4})\Big), \quad \text{where}$$

$$\hat{\pi}_{fobs,ij} = \frac{exp(x'_{fobs,i}\hat{\beta}_{fobs,j})}{1 + \sum_{h=1}^{3} exp(x'_{fobs,i}\hat{\beta}_{fobs,h})}, \quad \text{for } i = 1, 2, \ldots, n_{cc}; j = 1, 2, 3.$$

Note that $\hat{\pi}_{fobs,i4} = 1 - \hat{\pi}_{fobs,i1} - \hat{\pi}_{fobs,i2} - \hat{\pi}_{fobs,i3}$. Let $\{Y^*_{fobs}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y_{fobs} generated as described. Let $\{\hat{\beta}^*_{fobs}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap replicates of MLEs of β corresponding to the B bootstrap samples. The b^{th} bootstrap replicate of the MLE of β can be written in terms of its column vectors as $\hat{\beta}^*_{fobs}(b) = (\hat{\beta}^*_{fobs,1}(b)$, $\hat{\beta}^*_{fobs,2}(b)$, $\hat{\beta}^*_{fobs,3}(b))$.

The fully observed data version of EIC is

$$\begin{aligned}
\text{EIC}_{fobs} = &-2\sum_{i=1}^{n_{cc}}\left[\sum_{j=1}^{3} y_{fobs,ij}(x'_{fobs,i}\hat{\beta}_{fobs,j}) - \log\left(1 + \sum_{j=1}^{3}\exp(x'_{fobs,i}\hat{\beta}_{fobs,j})\right)\right] \\
&+\frac{1}{B}\sum_{b=1}^{B}\left\{-2\sum_{i=1}^{n_{cc}}\left[\sum_{j=1}^{3} y_{fobs,ij}(x'_{fobs,i}\hat{\beta}^*_{fobs,j}(b)) - \log\left(1 + \sum_{j=1}^{3}\exp(x'_{fobs,i}\hat{\beta}^*_{fobs,j}(b))\right)\right]\right. \\
&\left.-\left(-2\sum_{i=1}^{n_{cc}}\left[\sum_{j=1}^{3} y^*_{fobs,ij}(b)(x'_{fobs,i}\hat{\beta}^*_{fobs,j}(b)) - \log\left(1 + \sum_{j=1}^{3}\exp(x'_{fobs,i}\hat{\beta}^*_{fobs,j}(b))\right)\right]\right)\right\}.
\end{aligned}$$

The fully observed data version of AICb is

$$\text{AICb}_{fobs} = -2\sum_{i=1}^{n_{cc}}\left[\sum_{j=1}^{3} y_{fobs,ij}(x'_{fobs,i}\hat{\beta}_{fobs,j}) - \log\left(1+\sum_{j=1}^{3}\exp(x'_{fobs,i}\hat{\beta}_{fobs,j})\right)\right]$$
$$+\frac{2}{B}\sum_{b=1}^{B}\left\{-2\sum_{i=1}^{n_{cc}}\left[\sum_{j=1}^{3} y_{fobs,ij}(x'_{fobs,i}\hat{\beta}^{*}_{fobs,j}(b)) - \log\left(1+\sum_{j=1}^{3}\exp(x'_{fobs,i}\hat{\beta}^{*}_{fobs,j}(b))\right)\right]\right.$$
$$\left.-\left(-2\sum_{i=1}^{n_{cc}}\left[\sum_{j=1}^{3} y_{fobs,ij}(x'_{fobs,i}\hat{\beta}_{fobs,j}) - \log\left(1+\sum_{j=1}^{3}\exp(x'_{fobs,i}\hat{\beta}_{fobs,j})\right)\right]\right)\right\}.$$

5.2.2 Criteria Using the Complete Data

The complete data analogues of AIC, EIC, and AICb are now developed for the baseline-category logit model. In this modeling framework, the complete data KL discrepancy is given by

$$E_o\left\{-2\sum_{i=1}^{n}\left[\sum_{j=1}^{3} y_{ij}(x'_i\beta_j) - \log\left(1+\sum_{j=1}^{3}\exp(x'_i\beta_j)\right)\right]\right\} \tag{5.6}$$
$$= -2\sum_{i=1}^{n}\left[\sum_{j=1}^{3}\pi_{o,ij}(x'_i\beta_j) - \log\left(1+\sum_{j=1}^{3}\exp(x'_i\beta_j)\right)\right], \quad \text{where}$$
$$\pi_{o,ij} = \frac{exp(x'_{o,i}\beta_{o,j})}{1+\sum_{h=1}^{3} exp(x'_{o,i}\beta_{o,h})}, \quad \text{for } i = 1,2,\ldots,n; j = 1,2,3,4.$$

To define the overall complete data KL discrepancy, which is estimated by the complete data criteria, we substitute the maximum likelihood estimator based on Y for β in (5.6). The expected overall complete data KL discrepancy is found by averaging the preceding over the sampling distribution of the estimator based on Y.

For a particular candidate model of the form (5.3), let $\hat{\beta}$ denote the maximum likelihood estimator of β obtained via the EM algorithm. The estimator $\hat{\beta}$ can be represented in terms of its column vectors as $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$. The fitted candidate model is obtained by using $\hat{\beta}$ as the parameter value in the candidate model.

Following the proposed methodology of section (3.4), the complete data goodness-of-fit term can be calculated as follows. Recalling the delineation of $Y = (Y_{obs}, Y_{mis})$,

bootstrap samples of Y_{mis} must be collected in order to construct the complete data goodness-of-fit term. For each of the four missing data scenarios, the corresponding multinomial quadruple can be expressed as $y_i = (y'_{obs,i}, y'_{mis,i})'$, where the two elements in $y_{obs,i}$ are both 0, and the two elements in $y_{mis,i}$ are both missing. For a particular fitted candidate model, the distribution of $y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})$ is given by

$$Multinomial\Big(1, (\hat{\pi}_{i1}, \hat{\pi}_{i2}, \hat{\pi}_{i3}, \hat{\pi}_{i4})\Big), \quad \text{where}$$
$$\hat{\pi}_{ij} = \frac{exp(x'_i\hat{\beta}_j)}{1 + \sum_{h=1}^{3} exp(x'_i\hat{\beta}_h)}, \quad \text{for } i = 1, 2, \ldots, n; j = 1, 2, 3. \tag{5.7}$$

Note that $\hat{\pi}_{i4} = 1 - \hat{\pi}_{i1} - \hat{\pi}_{i2} - \hat{\pi}_{i3}$.

We will illustrate the generation of the bootstrap samples of Y_{mis} by considering the fourth missing data scenario of $(z_{i1} = 1, z_{i2} = .)$. For this scenario, recall that the corresponding multinomial quadruple is $y_i = (y_{i1} = ., y_{i2} = ., y_{i3} = 0, y_{i4} = 0)$, where $y_{i1} = 1$ and $y_{i2} = 0$ if z_{i2} was hypothetically observed as 1, or $y_{i1} = 0$ and $y_{i2} = 1$ if z_{i2} was hypothetically observed as 0. The conditional distribution of $y_{mis,i}|y_{obs,i}$ under this scenario, as well as the other missing data scenarios, can be obtained using (5.7). In a bootstrap sample Y^*_{mis}, the bootstrap vectors $y^*_{mis,i}$ are generated for the cases with the considered missing data scenario by taking a draw from the conditional distribution of $y_{mis,i1}, y_{mis,i2}|y_{obs,i3} = 0, y_{obs,i4} = 0$. This distribution is given by

$$Multinomial\Big(1, (\hat{\pi}_{i1}/(\hat{\pi}_{i1} + \hat{\pi}_{i2}), \hat{\pi}_{i2}/(\hat{\pi}_{i1} + \hat{\pi}_{i2}))\Big).$$

The bootstrap vectors $y^*_{mis,i}$ in a bootstrap sample Y^*_{mis} for the cases under the other three missing data scenarios can be generated by taking a draw from the conditional distribution of $y_{mis,i}|y_{obs,i}$, in a similar fashion. Let $\{Y^*_{mis}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y_{mis} generated as described. Let

$\{Y^*_{par}(b) = (Y_{obs}, Y^*_{mis}(b)) | b = 1, 2, \ldots, B\}$ represent a collection of B partially bootstrapped samples of Y. Let $y^*_{par,ij}(b)$ denote the value for case i and column j from $Y^*_{par}(b)$. Let $\{\hat{\beta}^*_{par}(b) | b = 1, 2, \ldots, B\}$ represent a collection of B partial bootstrap replicates of MLEs of β, corresponding to the B partially bootstrapped samples of Y. The b^{th} partial bootstrap replicate of the MLE of β can be written in terms of its column vectors as $\hat{\beta}^*_{par}(b) = (\hat{\beta}^*_{par,1}(b),\ \hat{\beta}^*_{par,2}(b),\ \hat{\beta}^*_{par,3}(b))$.

The complete data goodness-of-fit term, say GOF_{comp}, is now given as

$$\frac{1}{B}\sum_{b=1}^{B}\left\{ -2\sum_{i=1}^{n}\left[\sum_{j=1}^{3} y^*_{par,ij}(b)(x_i'\hat{\beta}^*_{par,j}(b)) - \log\left(1 + \sum_{j=1}^{3}\exp(x_i'\hat{\beta}^*_{par,j}(b))\right)\right]\right\}.$$

The three complete data criteria proposed in section (3.4) are now developed. These criteria should have expectations that are approximately equal to the expected complete data KL discrepancy. Again, let $k = 3p$ denote the number of functionally independent parameters in the candidate model (5.3). The complete data analogue of AIC, say AIC_{comp}, is

$$\frac{1}{B}\sum_{b=1}^{B}\left\{ -2\sum_{i=1}^{n}\left[\sum_{j=1}^{3} y^*_{par,ij}(b)(x_i'\hat{\beta}^*_{par,j}(b)) - \log\left(1 + \sum_{j=1}^{3}\exp(x_i'\hat{\beta}^*_{par,j}(b))\right)\right]\right\} + 2k.$$

In order to construct the estimators of the expected optimism in the complete data bootstrap criteria, bootstrap samples of both the observed and missing elements in Y must be obtained. The bootstrap responses $y^*_i = (y^*_{i1},\ y^*_{i2},\ y^*_{i3},\ y^*_{i4})$ in a parametric bootstrap sample Y^* are generated case-by-case from the multinomial distribution given in (5.7). Let $\{Y^*(b) | b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y generated as described. Let $\{\hat{\beta}^*(b) | b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap replicates of MLEs of β, corresponding to the B bootstrap samples of Y. The b^{th} bootstrap replicate of the MLE of β can be written in terms of its column vectors as $\hat{\beta}^*(b) = (\hat{\beta}^*_1(b),\ \hat{\beta}^*_2(b),\ \hat{\beta}^*_3(b))$.

The complete data analogue of EIC, say EIC_{comp}, is represented as

$$\frac{1}{B}\sum_{b=1}^{B}\left\{-2\sum_{i=1}^{n}\left[\sum_{j=1}^{3}y^*_{par,ij}(b)(x_i'\hat{\beta}^*_{par,j}(b))-\log\left(1+\sum_{j=1}^{3}\exp(x_i'\hat{\beta}^*_{par,j}(b))\right)\right]\right\}$$
$$+\frac{1}{B}\sum_{b=1}^{B}\left\{-2\sum_{i=1}^{n_{cc}}\left[\sum_{j=1}^{3}y^*_{par,ij}(b)(x_i'\hat{\beta}^*_{j}(b))-\log\left(1+\sum_{j=1}^{3}\exp(x_i'\hat{\beta}^*_{j}(b))\right)\right]\right.$$
$$\left.-\left(-2\sum_{i=1}^{n}\left[\sum_{j=1}^{3}y^*_{ij}(b)(x_i'\hat{\beta}^*_{j}(b))-\log\left(1+\sum_{j=1}^{3}\exp(x_i'\hat{\beta}^*_{j}(b))\right)\right]\right)\right\}.$$

The complete data analogue of AICb, say AICb_{comp}, is given as

$$\frac{1}{B}\sum_{b=1}^{B}\left\{-2\sum_{i=1}^{n}\left[\sum_{j=1}^{3}y^*_{par,ij}(b)(x_i'\hat{\beta}^*_{par,j}(b))-\log\left(1+\sum_{j=1}^{3}\exp(x_i'\hat{\beta}^*_{par,j}(b))\right)\right]\right\}$$
$$+\frac{2}{B}\sum_{b=1}^{B}\left\{-2\sum_{i=1}^{n}\left[\sum_{j=1}^{3}y^*_{par,ij}(b)(x_i'\hat{\beta}^*_{j}(b))-\log\left(1+\sum_{j=1}^{3}\exp(x_i'\hat{\beta}^*_{j}(b))\right)\right]\right.$$
$$\left.-\left(-2\sum_{i=1}^{n}\left[\sum_{j=1}^{3}y^*_{par,ij}(b)(x_i'\hat{\beta}^*_{par,j}(b))-\log\left(1+\sum_{j=1}^{3}\exp(x_i'\hat{\beta}^*_{par,j}(b))\right)\right]\right)\right\}.$$

5.2.3 Simulation Study

Consider a setting where a sample of n multinomial trials are generated from the distribution (5.1). Suppose our goal is to search among a class of nested candidate models for the fitted candidate model that best approximates the true model (5.2). We assume the candidate models are of the form (5.3).

In model (5.3), recall that x_i denotes a p-length vector of covariates for case i from the $n \times p$ design matrix X. We consider a class of five nested candidate models, where the number of covariates range from 1 to 5, corresponding to design matrices having from $p = 2$ to $p = 6$ columns. We will refer to p as the order of the candidate model. The first column of X was taken to be a vector of ones. The covariate values of X were generated independently from a $N(0, 1)$ distribution.

The simulations sets are based on 200 samples, where the samples are of size $n = 90$ or $n = 135$. Each sample was generated from a multinomial distribution of

the form (5.1), where β_o is a 4×3 matrix. The columns of β_o, represented as $\beta_{o,1}$, $\beta_{o,2}$, $\beta_{o,3}$ are vectors of length four whose elements are all ones.

The order of the true model, or true order p_o, is 4. In the nested model setting that we are considering, the true model includes the first three covariates. An underspecified model would include the first covariate, or the first two covariates. An overspecified model would include the first four covariates, or all five covariates. All candidate models include an intercept.

Throughout this chapter, we have emphasized that the outcome can be viewed in terms of Y, where each row in Y is a quadruple that represents the outcome of a multinomial trial, or Z, where each row in Z is a pair of potentially correlated binary variables. In order to create a missing data setting, we randomly discarded the data in terms of Z so as to create a situation where the missing data in Z are MAR. Recall that $z_i = (z_{i1}, z_{i2})$ denotes the outcome pair for case i from Z. Let $\Pr(z_{i1} \text{ mis})$ denote the probability that for a particular case, z_{i1} is discarded and z_{i2} is retained, and let $\Pr(z_{i2} \text{ mis})$ denote the probability that for a particular case, z_{i2} is discarded and z_{i1} is retained. We considered discard probabilities of $(\Pr(z_{i1} \text{ mis}), \Pr(z_{i2} \text{ mis}))$ set at (0.0, 0.0), (0.075, 0.075), and (0.15, 0.15).

In summary, simulation results from a total of six different settings are investigated. We consider all possible combinations of the following factors: two different sample sizes of $n = 90$ and $n = 135$, and three levels of missingness of $(\Pr(z_{i1} \text{ mis}), \Pr(z_{i2} \text{ mis}))$ set at (0.0, 0.0), (0.075, 0.075), and (0.15, 0.15).

For each of the six settings, the five candidate models were fit for each sample. The complete data and fully observed data analogues of AIC, EIC, and AICb were then calculated for every set of fitted candidate models. The overall complete data KL discrepancy, which is found by substituting the maximum likelihood estimator based on Y for β in (5.6), along with the overall fully observed data KL discrepancy, which is obtained by substituting the maximum likelihood estimator based on Y_{fobs}

for β in (5.5), were also computed for every set of fitted candidate models. The model selected by each of the fully observed and complete data criteria, along with the selections of both discrepancies, was determined. For every candidate model, a total of 300 bootstrap samples were generated for EIC and AICb, along with 300 partially bootstrapped samples for the complete data criteria.

To characterize the effectiveness of the complete and fully observed data criteria, we use the two performance measures given in section (4.2.3): the average criterion values and distribution of model selections. The results from the six simulation settings are summarized in figures (5.1), (5.2), and (5.3), and in tables (5.1) and (5.2).

Figures (5.1), (5.2), and (5.3) illustrate the effectiveness of the criteria in estimating their target discrepancies. In settings with no missing data (see figure 5.1), EIC estimates the KL target very well at both sample sizes. For the correctly specified or overspecified models, when $n = 90$, the mean values of AIC are too low, and the mean values of AICb_{comp} are a bit high. For $n = 135$, the estimation of the KL target improves for both AIC and AICb_{comp}.

The curves for the complete data criteria for both the moderate level of missingness (see figure 5.2), and the highest percentage of missingness (see figure 5.3), share some similarities. For the correctly specified or overspecified models, when $n = 90$, the mean values of AICb_{comp} are a little bit higher than the complete data target, and the mean values of AIC_{comp} and EIC_{comp} are both lower than the complete data target. However, when $n = 135$, the mean values of AICb_{comp} are nearly identical to that of the complete data target, and both EIC_{comp} and AIC_{comp} have mean values that are still too low. The slopes of the curves for both EIC_{comp} and AICb_{comp} at both sample sizes generally parallel the complete data target well for the correctly specified or overspecified models. The slope of the curve for AIC_{comp} is too flat at $n = 90$, which will result in the selection of more overspecified models,

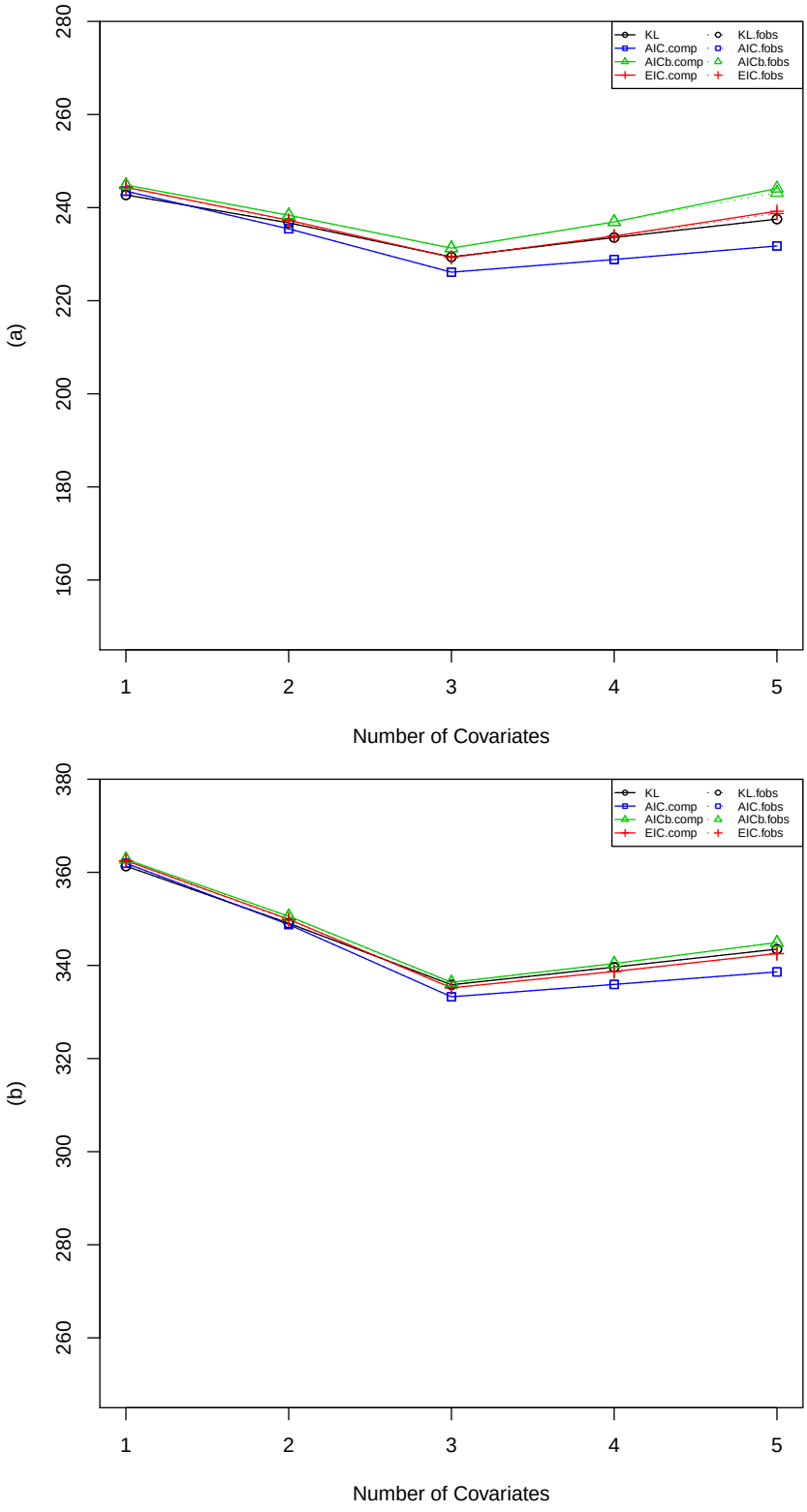

Figure 5.1: Average criterion values under baseline-category logit model setting with missing data in the outcome: $(\Pr(z_{i1} \text{ mis}), \Pr(z_{i2} \text{ mis})) = (0, 0)$. (a) $n = 90$; (b) $n = 135$.

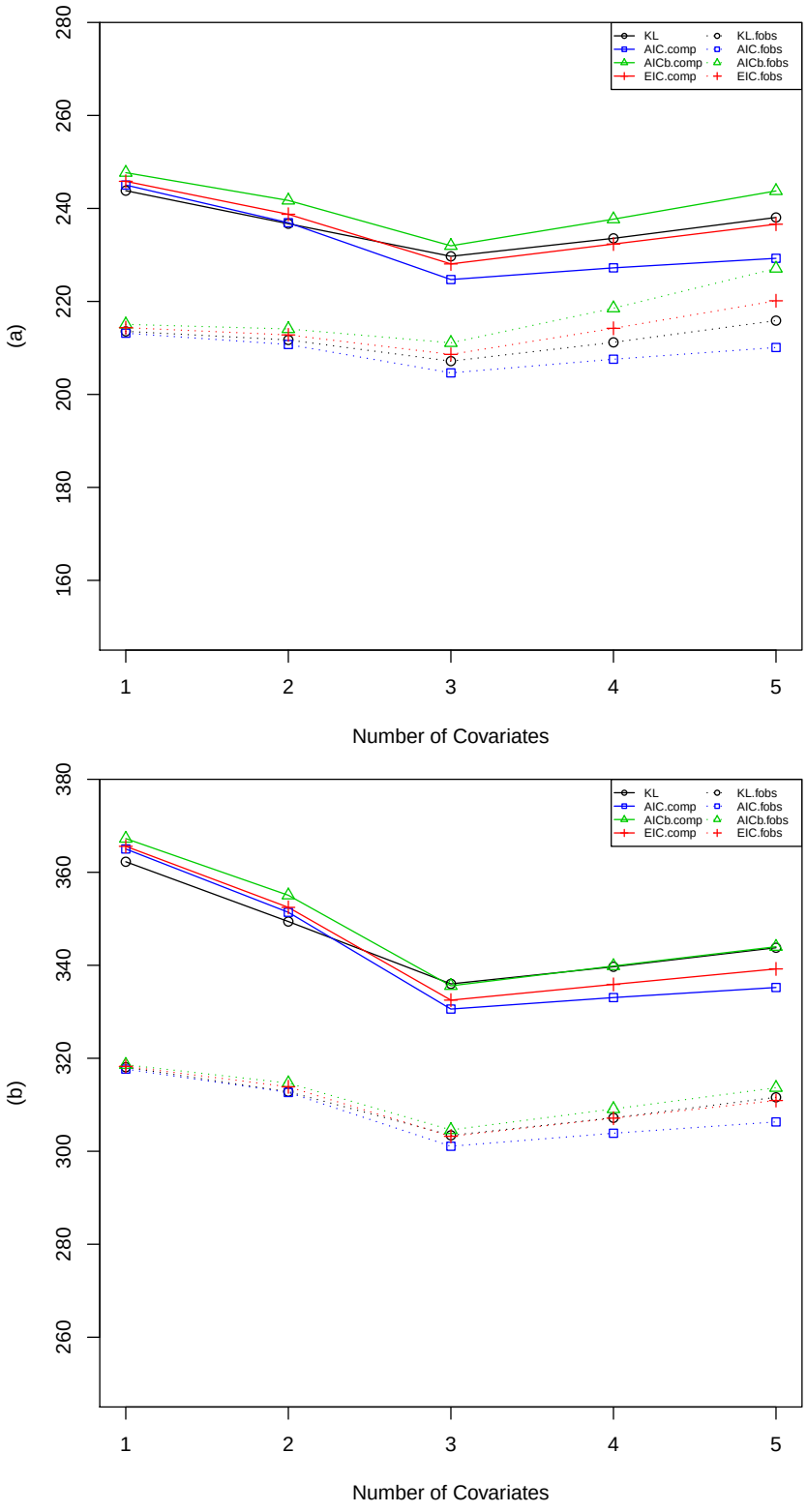

Figure 5.2: Average criterion values under baseline-category logit model setting with missing data in the outcome: $(\Pr(z_{i1} \text{ mis}), \Pr(z_{i2} \text{ mis})) = (.075, .075)$. (a) $n = 90$; (b) $n = 135$.

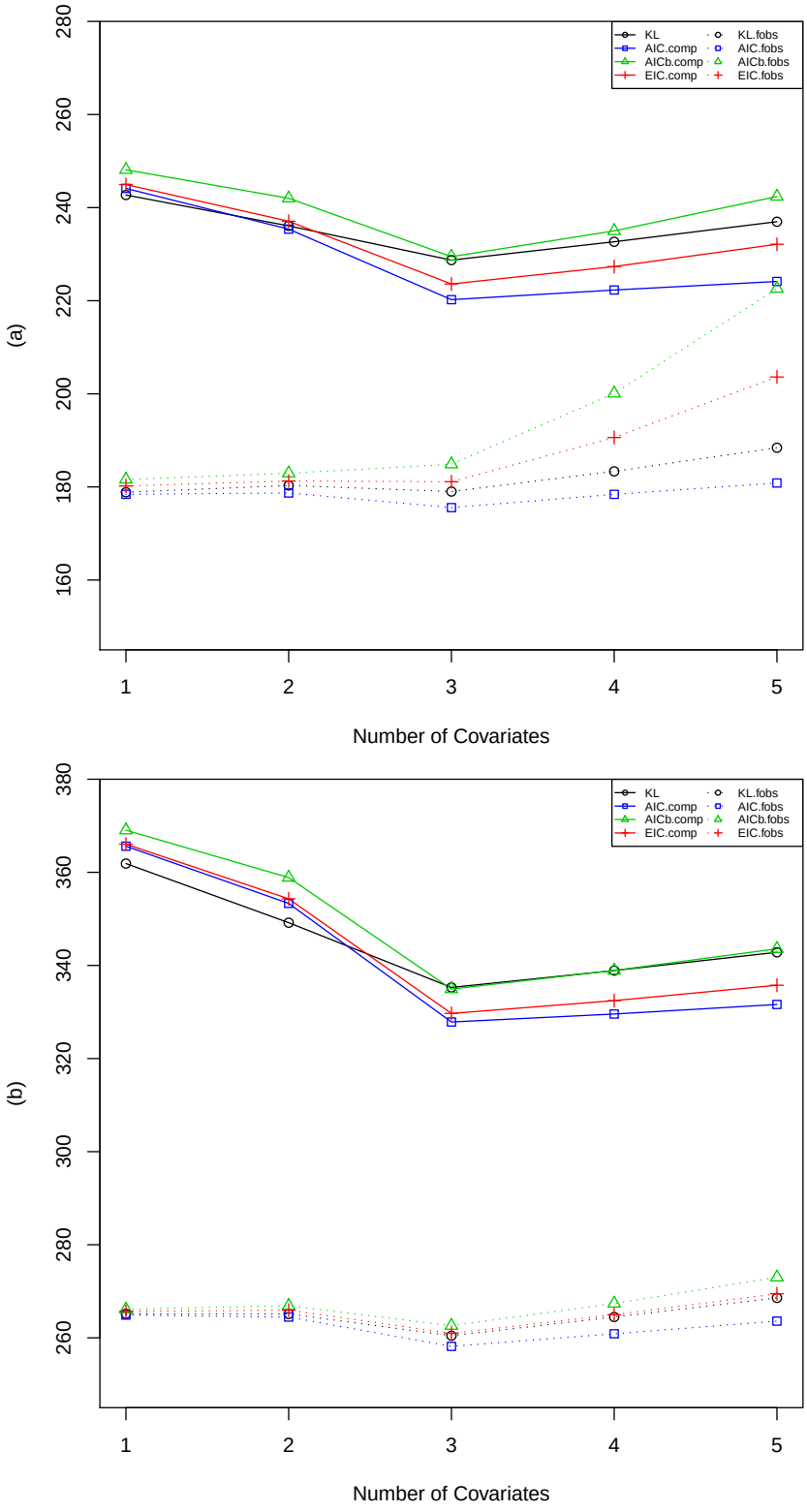

Figure 5.3: Average criterion values under baseline-category logit model setting with missing data in the outcome: $(\Pr(z_{i1} \text{ mis}), \Pr(z_{i2} \text{ mis})) = (.15, .15)$. (a) $n = 90$; (b) $n = 135$.

Table 5.1: Frequency of criterion selections under baseline-category logit model setting with missing data in the outcome: $n = 90$.

$\Pr(z_{i1}$ mis),		Complete Data				Fully Observed Data			
$\Pr(z_{i2}$ mis)	Covariates	KL	AIC	EIC	AICb	KL	AIC	EIC	AICb
(0, 0)	1	5	3	5	9	5	3	7	9
	2	13	15	23	24	13	15	22	29
	3	182	144	158	162	182	144	157	156
	4	0	30	10	5	0	30	10	5
	5	0	8	4	0	0	8	4	1
(.075, .075)	1	0	1	2	3	13	13	33	47
	2	12	7	10	15	21	25	33	41
	3	188	147	161	169	166	129	125	108
	4	0	25	21	12	0	17	8	3
	5	0	20	6	1	0	16	1	1
(.15, .15)	1	2	3	3	4	61	59	100	130
	2	10	1	12	17	20	21	26	26
	3	188	142	156	167	119	101	69	42
	4	0	30	21	11	0	12	5	2
	5	0	24	8	1	0	7	0	0

Table 5.2: Frequency of criterion selections under baseline-category logit model setting with missing data in the outcome: $n = 135$.

$\Pr(z_{i1}$ mis),		Complete Data				Fully Observed Data			
$\Pr(z_{i2}$ mis)	Covariates	KL	AIC	EIC	AICb	KL	AIC	EIC	AICb
(0, 0)	1	0	0	0	0	0	0	0	0
	2	4	3	3	4	4	3	5	6
	3	196	163	175	180	196	163	162	173
	4	0	22	19	15	0	22	26	18
	5	0	12	3	1	0	12	7	3
(.075, .075)	1	0	0	0	0	1	1	2	5
	2	1	0	3	3	5	8	14	15
	3	199	154	167	175	194	155	160	163
	4	0	26	21	15	0	20	17	12
	5	0	20	9	7	0	16	7	5
(.15, .15)	1	0	0	0	2	21	34	54	63
	2	1	1	1	2	13	15	18	21
	3	199	146	164	172	166	124	116	106
	4	0	33	23	18	0	14	8	8
	5	0	20	12	6	0	13	4	2

while the slope of the curve for AICb_{comp} is too steep, which will generally facilitate the selection of fewer models that are overspecified.

In general, the mean values of all complete data criteria tend to be higher than the complete data target for underspecified models. Some of the trends in the mean criterion values for the fully observed data criteria are similar to those of the complete data criteria; yet, the trends are often more pronounced, especially when $n = 90$.

Tables (5.1) and (5.2), show the distribution of the model selections for each of the criteria. When $n = 90$ (see table 5.1), the complete data criteria consistently outperform the fully observed data criteria at both levels of missingness. The difference in correct model selections between the complete and fully observed data criteria is greatest at the highest level of missingess. As the amount of missingness increases, the tendency of the fully observed data criteria is to select underspecified models, yet the complete data criteria protect against this tendency. The fully observed data target is affected by an increase in the amount of missingness, as indicated by a decrease in the number of correctly specified model selections; consequently, the criteria that estimate this target tend to under perform in selecting the correct model. Of the complete data criteria, EIC_{comp} and AICb_{comp} both select more correctly specified models than AIC_{comp}.

With $n = 135$ (see table 5.2), the complete data criteria again exhibit more desirable selection results than the fully observed data criteria at the highest level of missingness. The complete and fully observed data criteria perform similarly at the moderate level of missingness; however, the complete data criteria protect against the slight tendency to select underspecified models that the fully observed data criteria exhibit.

In summary, in the investigated settings, the results from the simulation study suggest that the complete data criteria generally outperform their fully observed

data counterparts. The utility of the complete data criteria will be greatest with moderate or small sample sizes when there is less information in the data. In such scenarios, employing model selection criteria using only the fully observed data could possibly result in selecting an underspecified model. Thus, the complete data criteria may provide more reliable selections than the fully observed data criteria.

5.3 Missing Data in the Outcome and/or Covariates

In subsection (5.2.2), we demonstrated how to calculate the complete data criteria if one of the elements in the pair of binary variables for the cases in Z was missing. We now consider two additional missing data scenarios:

(1) missing data only in the covariates, and

(2) missing data in the covariates and in one of the elements in the pair of binary variables.

In this section, we will outline the distributions that would be necessary to generate the bootstrap samples needed for the complete data criteria. We will view the response data in terms of the collection of multinomial quadruples contained in Y.

Schafer (1997) provides a general treatment of mixed data that consists of a combination of continuous and categorical variables. We briefly outline the relevant methodology in the context of candidate model (5.3). Assume that the p-length vector of covariates x_i is distributed as

$$x_i|y_i \sim N(\mu_d, \Sigma), \quad \text{for } i = 1, 2, \ldots, n; d = 1, 2, 3, 4. \tag{5.8}$$

Here, μ_d is a vector of length p, where the subscript d indicates the position of the 1 for multinomial trial y_i (e.g. $y_i = (0, 0, 1, 0)$ corresponds to $d = 3$); Σ is a $p \times p$ covariance matrix. Note that for (5.8), a mean vector is assumed for the covariate data corresponding to all cases that have the same multinomial trial outcome.

Let us first consider missing data scenario (1) where the missing data is restricted only to the covariates. For case i, the observed and missing covariates can be delineated as $x_i = (x'_{obs,i}, x'_{mis,i})'$. Using (5.8), the conditional distribution of $x_{mis,i}|x_{obs,i}, y_i$ can also be shown to be multivariate normal. Parameter estimates for the conditional distribution can be obtained via the EM algorithm and the sweep operator (Schafer, 1997, p.350), from which a bootstrap vector $x^*_{mis,i}$ can be drawn.

The distributions needed for missing data scenario (2) are now given. For the outcome, we now have only the partially observed information in a multinomial trial y_i since one of the elements in the pair of binary variables is assumed to be missing. Recall that for case i, the observed and missing elements in y_i can be delineated as $y_i = (y'_{obs,i}, y'_{mis,i})'$. Similar to the treatment of missing covariates under scenario (1), the EM algorithm and sweep operator can be used to obtain the multivariate normal conditional distribution of $x_{mis,i}|x_{obs,i}, y_{obs,i}$, which can be used to generate a bootstrap vector $x^*_{mis,i}$. The conditional distribution $y_{mis,i}|x_{obs,i}, y_{obs,i}$ needed to obtain the bootstrap vectors of the missing elements in a multinomial trial is more complicated to describe, and we refer the reader to Schafer (1997, p.350) for specific details. A bootstrap vector $y^*_{mis,i}$ is generated by taking a draw from the preceding distribution.

Under both missing data scenarios, the bootstrap samples of the missing data can be combined with the observed data to create the partially bootstrapped samples. The complete data goodness-of-fit term is evaluated based on the partially bootstrapped samples.

In order to construct the estimators of the expected optimism for the complete data analogues of AICb and EIC, bootstrap samples of both the observed and missing elements in y_i must be obtained. Under both missing data scenarios, one way of obtaining these bootstrap samples can be derived using a distributional result in Schafer (1997). Schafer evaluates the probability of a case falling into a particular

cell for a combination of categorical variables, which can be used in our setting to provide the probabilities needed to obtain a bootstrap multinomial quadruple y_i^*.

All of the distributions necessary for the construction of the complete data criteria have been established. One can follow the ideas of subsection (4.3.3) in calculating the criteria.

CHAPTER 6
KULLBACK-LEIBLER DISCREPANCY BASED CRITERIA: LONGITUDINAL DATA ANALYSIS FRAMEWORK

Longitudinal data frequently arises in many medical and biological applications. In this chapter, we consider an important modeling framework that is often utilized for longitudinal data analysis. The proposed methodology is developed using a normal longitudinal regression model with missing data in the outcome. A simulation study is presented and discussed for this setting. Furthermore, we briefly outline how criteria based on the complete data can be calculated which accommodate missing data in the outcome and/or covariates.

6.1 Normal Longitudinal Regression Model

The normal longitudinal regression model is frequently used to determine the linear relationship between covariates and a vector of response outcomes. Often, the vector of response outcomes consists of repeated measurements at different time points. Let the collection of outcomes for n cases be represented as $Y = (y_1', y_2', \ldots, y_n')'$, where y_i is a $t_i \times 1$ response vector for $i = 1, 2, \ldots, n$. For simplicity, we will assume that $t_i = t$. Let $N = nt$ denote the total number of responses in Y; thus, Y is an $N \times 1$ vector.

Suppose that each y_i in Y has been generated according to the model

$$y_i = X_{o,i}\beta_o + e_{o,i}, \quad \text{for } i = 1, 2, \ldots n. \tag{6.1}$$

Hence, (6.1) denotes the true model. Here, $X_{o,i}$ is a $t \times p_o$ known design matrix, β_o is a $p_o \times 1$ vector of regression parameters, and $e_{o,i}$ is a $t \times 1$ error vector assumed to follow a multivariate normal distribution with mean vector 0 and a $t \times t$ covariance matrix Σ_o.

Suppose that a candidate model for the entire sample $Y = (y_1', y_2', \ldots, y_n')'$ is

postulated of the form

$$y_i = X_i\beta + e_i, \quad \text{for } i = 1, 2, \ldots n. \tag{6.2}$$

Here, y_i is as previously described, X_i is a $t \times p$ known design matrix, β is a $p \times 1$ vector of regression parameters, and e_i is a $t \times 1$ error vector assumed to follow a multivariate normal distribution with mean vector 0 and a $t \times t$ covariance matrix Σ. As outlined in Little and Rubin (2002, p.242), the candidate model (6.2) permits a variety of structures for the covariance matrix Σ. These structures include the following.

- Independence: $\Sigma = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_t^2)$, where each element σ_i^2 is a scalar.
- Compound symmetry: $\Sigma = \sigma^2 J + \sigma_1 I$, where J is a $t \times t$ matrix of ones, I is a $t \times t$ identity matrix, and σ^2 and σ_1 are both scalars.
- Random effects: $\Sigma = ZDZ' + \sigma^2 I$, where Z is a $t \times q$ known matrix, D is a $q \times q$ matrix comprised of variance components for random effects, I is a $t \times t$ identity matrix, and σ^2 is a scalar.
- Unstructured: $\Sigma = (\sigma_{ab})$, where σ_{ab} denotes the covariance parameter in the a^{th} row and b^{th} column, and all parameters in the upper triangular matrix (or lower triangular matrix) are allowed to vary.

For the model selection criteria that will be given in subsections (6.2.1) and (6.2.2), we assume the same covariance structure for all candidate models under consideration. Shang and Cavanaugh (2008) consider the use of EIC and AICb in settings with no missing data, when different random effects covariance structures are considered. There results show that the bootstrap based criteria can be used to effectively delineate between different covariance structures. Hence, the bootstrap versions of our complete data criteria, to be developed in subsection (6.2.2), may show promise for covariance structure selection in missing data applications.

The log-likelihood for the candidate model (6.2), ignoring the constant terms, is given by

$$\log L(\beta, \Sigma|Y) \propto \sum_{i=1}^{n} \Big(-\frac{1}{2}\log|\Sigma| - \frac{1}{2}(y_i - X_i\beta)'\Sigma^{-1}(y_i - X_i\beta)\Big).$$

Fitting the candidate model requires estimation of β and Σ. Generally, maximum likelihood estimation of β and Σ is not available in closed form and iterative procedures are often employed. For example, following Little and Rubin (2002, p.180), we outline a method of parameter estimation under the assumption that Σ is an unstructured covariance matrix. Suppose the q^{th} iterate of Σ is denoted as $\Sigma = \Sigma^{(q)}$. Given $\Sigma^{(q)}$, it can be shown that the MLE of β for the $(q+1)^{th}$ iteration is

$$\beta^{(q+1)} = \left(\sum_{i=1}^{n} X_i'(\Sigma^{(q)})^{-1}X_i \right)^{-1} \left(\sum_{i=1}^{n} X_i'(\Sigma^{(q)})^{-1}y_i \right). \tag{6.3}$$

Given $\beta^{(q+1)}$, the MLE of Σ for the $(q+1)^{th}$ iteration is

$$\Sigma^{(q+1)} = \frac{1}{n}\sum_{i=1}^{n} \Big(y_i - X_i\beta^{(q+1)}\Big)\Big(y_i - X_i\beta^{(q+1)}\Big)'. \tag{6.4}$$

For covariance matrices that are not unstructured, equation (6.4) would need to be appropriately modified.

The normal longitudinal regression model introduced in this section has some fundamental differences from the normal multivariate linear regression model introduced in section (4.1). The normal longitudinal regression model allows for a different set of covariate measurements for each outcome in the response vector, whereas the normal multivariate linear regression model requires the same set of covariate measurements for each of the potentially related outcomes. Due to this difference, the normal longitudinal regression model assumes the same $p\times 1$ vector of regression parameters for each outcome in the response vector, whereas the normal multivariate linear regression model allows for a $p \times 1$ vector of regression parameters for each of the potentially related outcomes. Both models have applicability

in a wide variety of practical settings and the choice between the two models will depend on the assumptions the analyst is willing to make, along with the nature of the data.

In the previous chapters, we have discussed how a complete case analysis is a common approach to selecting a model in incomplete data settings since it is the default in most statistical packages. We now delineate how statistical packages generally handle missing data for a normal longitudinal regression model, as compared to the normal multivariate linear regression model of Chapter 4 and the baseline-category logit model of Chapter 5. Consider, as an example, the structure of a typical data file in a longitudinal data setting. Each row in the file generally consists of an outcome and the associated covariates for a given case and time point. If a row contains at least one missing element, statistical packages will not use the row in fitting a model. However, for longitudinal data settings, the failure to utilize a row in the data file does not eliminate the use of the rest of the data from the other time points for the case. This is a different use of the data than the normal multivariate linear regression model and baseline-category logit model, where a missing value for one of the elements in the response outcome or covariates results in not using any of the data for the case. Thus, the normal longitudinal regression model is more flexible in missing data settings than both the multivariate linear regression model and the baseline-category logit model, since more data is utilized.

6.2 Missing Data in the Outcome

In the following subsections, (6.2.1), (6.2.2), and (6.2.3) we consider a situation in which there is potentially missing data in each outcome vector y_i from the candidate model (6.2). The elements of the design matrix X_i are assumed to contain *no* missing values.

The observed and missing elements for each y_i can be delineated as $y_i =$

$(y'_{obs,i}, y'_{mis,i})'$. Let $t_{obs,i}$ denote the number of observations in $y_{obs,i}$. The design matrix X_i can also be partitioned as $X_i = (X'_{obs,i}, X'_{mis,i})'$, which corresponds to $y_i = (y'_{obs,i}, y'_{mis,i})'$. Similarly, the design matrix $X_{o,i}$ in the true model (6.1) can be partitioned as $X_{o,i} = (X'_{obs,o,i}, X'_{mis,o,i})'$. Let the observed data for the outcome measurements of all cases be represented as the vector $Y_{obs} = (y'_{obs,1}, y'_{obs,2}, \ldots, y'_{obs,n})'$, and similarly, the missing data for the outcome measurements can be depicted as the vector $Y_{mis} = (y'_{mis,1}, y'_{mis,2}, \ldots, y'_{mis,n})'$.

6.2.1 Criteria Using the Observed Data

A candidate model analogous to (6.2) can be postulated by using $y_{obs,i}$ and $X_{obs,i}$ in place of y_i and X_i. The observed data log-likelihood for this candidate model, ignoring the constant terms, is given as

$$\log L(\beta, \Sigma | Y_{obs}) \propto \sum_{i=1}^{n} \Big(-\frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (y_{obs,i} - X_{obs,i}\beta)' \Sigma_i^{-1} (y_{obs,i} - X_{obs,i}\beta) \Big), \quad (6.5)$$

where Σ_i is the $t_{obs,i} \times t_{obs,i}$ covariance matrix for $y_{obs,i}$ corresponding to a partition of Σ from the candidate model (6.2).

In this setting, the *observed data KL discrepancy* is defined as

$$\begin{aligned} & E_o\{-2 \log L(\beta, \Sigma | Y_{obs})\} \quad (6.6) \\ = & \sum_{i=1}^{n} \Big(\log |\Sigma_i| + \mathrm{tr}(\Sigma_i^{-1} \Sigma_{o,i}) + (X_{obs,o,i}\beta_o - X_{obs,i}\beta)' \Sigma_i^{-1} (X_{obs,o,i}\beta_o - X_{obs,i}\beta) \Big), \end{aligned}$$

where E_o denotes the expectation under the true model (6.1), and $\Sigma_{o,i}$ is the $t_{obs,i} \times t_{obs,i}$ covariance matrix for $y_{obs,i}$ corresponding to a partition of Σ_o from the true model (6.1). We emphasize that we are working with the observed data KL discrepancy based on the observed data Y_{obs}, which differs from the fully observed data KL discrepancy of Chapters 3, 4, 5, and 6 based on the fully observed data Y_{fobs}.

Model selection criteria can be developed based on the observed data KL

discrepancy (6.6), or the complete data KL discrepancy which will be given in equation (6.7). Cavanaugh and Shumway (1998) argue why it may be preferable to base model selection on the latter as opposed to the former. They show that the complete data KL discrepancy is potentially more sensitive than the observed data KL discrepancy for accessing the separation between the candidate model and the true model, thus, it may be desirable to develop model selection criteria based on the complete data KL discrepancy.

The maximum likelihood estimators of β and Σ corresponding to (6.5) can be found by maximizing the observed data log-likelihood via the EM algorithm. Let $\hat{\beta}$ and $\hat{\Sigma}$ denote such estimators of β and Σ.

To define the *overall observed data KL discrepancy*, which is estimated by model selection criteria developed using the observed data, we substitute $\hat{\beta}$ and $\hat{\Sigma}$ as estimators of β and Σ in (6.6). The *expected overall observed data KL discrepancy* is found by averaging the preceding over the sampling distribution of the estimators base on Y_{obs}.

We now give four model selection criteria based on the observed data in a normal longitudinal regression setting. Let k denote the number of functionally independent parameters in the candidate model. If the covariance matrix Σ is unstructured, we will have $k = p + .5t(t+1)$. For other covariance structures, k is equal to the sum of the number of functionally independent parameters in Σ and p. The observed data version of AIC is given as

$$\mathrm{AIC}_{obs} = \sum_{i=1}^{n} \Big(\log|\hat{\Sigma}_i| + (y_{obs,i} - X_{obs,i}\hat{\beta})'\hat{\Sigma}_i^{-1}(y_{obs,i} - X_{obs,i}\hat{\beta}) \Big) + 2k.$$

Azari, Li, and Tsai (2006) justify an estimator of the expected optimism that can be used in the construction of AICc in a normal longitudinal regression model setting. As with the justification of AIC and the derivation of AICc in other frameworks, their derivation assumes that the candidate model is correctly specified

or overfit. They also assume the same covariance structure Σ for all models under consideration, which is consistent with the setting under consideration. Let the total number of observed responses be denoted as $N_{obs} = \sum_{i=1}^{n} t_{obs,i}$. The evaluation of AICc based on the observed data is

$$\text{AICc}_{obs} = \sum_{i=1}^{n} \Big(\log|\hat{\Sigma}_i| + (y_{obs,i} - X_{obs,i}\hat{\beta})' \hat{\Sigma}_i^{-1} (y_{obs,i} - X_{obs,i}\hat{\beta}) \Big) + 2k(N_{obs}/(N_{obs} - k - 1)).$$

To construct the estimators of the expected optimism for the observed data bootstrap criteria, bootstrap samples of Y_{obs} must be obtained. The bootstrap samples could be obtained using either a parametric, semi-parametric, or non-parametric bootstrap. Again, we will outline the generation of the bootstrap samples using a parametric bootstrap, since this is how the bootstrap samples were obtained in the simulation study to be presented in section (6.2.3).

The bootstrap responses $y^*_{obs,i}$ in a bootstrap sample Y^*_{obs} are generated case-by-case by taking a draw from the distribution of the fitted candidate model, depicted as

$$N(X_{obs,i}\hat{\beta}, \hat{\Sigma}), \quad \text{for } i = 1, 2, \ldots, n.$$

Let $\{Y^*_{obs}(b) | b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y_{obs}, generated as described. Let $\{\hat{\beta}^*_{obs}(b) | b = 1, 2, \ldots, B\}$ and $\{\hat{\Sigma}^*_{obs}(b) | b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap replicates of MLEs of β and Σ, respectively, corresponding to the B bootstrap samples of Y_{obs}. Note that for the b^{th} bootstrap sample, $\hat{\Sigma}^*_{obs,i}(b)$ represents the partitioning of $\hat{\Sigma}^*_{obs}(b)$ based on $y_{obs,i}$.

The observed data version of EIC is given as

$$\begin{aligned}\mathrm{EIC}_{obs} &= \sum_{i=1}^{n}\Big(\log|\hat{\Sigma}_i| + (y_{obs,i} - X_{obs,i}\hat{\beta})'\hat{\Sigma}_i^{-1}(y_{obs,i} - X_{obs,i}\hat{\beta})\Big) \\ &+ \frac{1}{B}\sum_{b=1}^{B}\Bigg\{\sum_{i=1}^{n}\Big(\log|\hat{\Sigma}^*_{obs,i}(b)| + (y_{obs,i} - X_{obs,i}\hat{\beta}^*_{obs}(b))'\hat{\Sigma}^*_{obs,i}(b)^{-1}(y_{obs,i} - X_{obs,i}\hat{\beta}^*_{obs}(b))\Big) \\ &- \sum_{i=1}^{n}\Big(\log|\hat{\Sigma}^*_{obs,i}(b)| + (y^*_{obs,i}(b) - X_{obs,i}\hat{\beta}^*_{obs}(b))'\hat{\Sigma}^*_{obs,i}(b)^{-1}(y^*_{obs,i}(b) - X_{obs,i}\hat{\beta}^*_{obs}(b))\Big)\Bigg\}.\end{aligned}$$

The observed data version of AICb is

$$\begin{aligned}\mathrm{AICb}_{obs} &= \sum_{i=1}^{n}\Big(\log|\hat{\Sigma}_i| + (y_{obs,i} - X_{obs,i}\hat{\beta})'\hat{\Sigma}_i^{-1}(y_{obs,i} - X_{obs,i}\hat{\beta})\Big) \\ &+ \frac{2}{B}\sum_{b=1}^{B}\Bigg\{\sum_{i=1}^{n}\Big(\log|\hat{\Sigma}^*_{obs,i}(b)| + (y_{obs,i} - X_{obs,i}\hat{\beta}^*_{obs}(b))'\hat{\Sigma}^*_{obs,i}(b)^{-1}(y_{obs,i} - X_{obs,i}\hat{\beta}^*_{obs}(b))\Big) \\ &\qquad - \sum_{i=1}^{n}\Big(\log|\hat{\Sigma}_i| + (y_{obs,i} - X_{obs,i}\hat{\beta})'\hat{\Sigma}_i^{-1}(y_{obs,i} - X_{obs,i}\hat{\beta})\Big)\Bigg\}.\end{aligned}$$

6.2.2 Criteria Using the Complete Data

In the normal longitudinal regression setting, we now develop the complete data analogues of AIC, AICc, EIC, and AICb. In this modeling framework, the complete data KL discrepancy is given as

$$\begin{aligned}&E_o\{-2\log L(\beta,\Sigma|Y)\} \\ &= \sum_{i=1}^{n}\Big(\log|\Sigma| + \ \mathrm{tr}(\Sigma^{-1}\Sigma_o) + (X_{o,i}\beta_o - X_i\beta)'\Sigma^{-1}(X_{o,i}\beta_o - X_i\beta)\Big). \qquad (6.7)\end{aligned}$$

To define the overall compete data KL discrepancy, which is estimated by the complete data criteria, we substitute the maximum likelihood estimators based on Y for β and Σ in (6.7). The expected overall complete data KL discrepancy is found by averaging the preceding over the sampling distribution of the estimators based on Y.

Following the proposed methodology of section (3.4), the complete data goodness-of-fit term can now be calculated. Recall that for case i, the response vector for the missing and observed elements can be delineated as $y_i = (y'_{obs,i}, y'_{mis,i})'$. Bootstrap samples of each $y_{mis,i}$ that is contained in Y_{mis} must be collected in order to construct the complete data goodness-of-fit term. For a particular fitted candidate model, the distribution of y_i is estimated as $N(X_i\hat{\beta}, \hat{\Sigma})$, where the estimators $\hat{\beta}$ and $\hat{\Sigma}$ are obtained via the EM algorithm. This distribution can be partitioned corresponding to the observed and missing elements of y_i, and is given by

$$(y'_{obs,i}, y'_{mis,i})' \sim N\left((X'_{obs,i}, X'_{mis,i})'\hat{\beta}, \begin{pmatrix} \hat{\Sigma}_{obs,i} & \hat{\Sigma}_{obsmis,i} \\ \hat{\Sigma}_{misobs,i} & \hat{\Sigma}_{mis,i} \end{pmatrix}\right), \text{ for } i = 1, 2, \ldots, n. \tag{6.8}$$

The conditional distribution of $y_{mis,i}|y_{obs,i}$ can now be obtained by using (6.8). The bootstrap vectors $y^*_{mis,i}$ in a bootstrap sample Y^*_{mis} are generated for the cases that have at least one missing element in y_i by taking a draw from the conditional distribution of $y_{mis,i}|y_{obs,i}$, depicted as

$$\begin{aligned} &N(\hat{\mu}_{y_{mis,i}|y_{obs,i}}, \hat{\Sigma}_{y_{mis,i}|y_{obs,i}}), \quad \text{where} \\ &\hat{\mu}_{y_{mis,i}|y_{obs,i}} = X_{mis,i}\hat{\beta} + \hat{\Sigma}_{misobs,i}\hat{\Sigma}^{-1}_{obs,i}(y_{obs,i} - X_{obs,i}\hat{\beta}), \quad \text{and} \\ &\hat{\Sigma}_{y_{mis,i}|y_{obs,i}} = \hat{\Sigma}_{mis,i} - \hat{\Sigma}_{misobs,i}\hat{\Sigma}^{-1}_{obs,i}\hat{\Sigma}_{obsmis,i}, \text{ for } i = 1, 2, \ldots, n. \end{aligned}$$

Let $\{Y^*_{mis}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y_{mis} generated as described. Let $\{Y^*_{par}(b) = (Y_{obs}, Y^*_{mis}(b))|b = 1, 2, \ldots, B\}$ represent a collection of B partially bootstrapped samples of Y. Let $\{\hat{\beta}^*_{par}(b)|b = 1, 2, \ldots, B\}$ and $\{\hat{\Sigma}^*_{par}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B partial bootstrap replicates of MLEs of β and Σ, corresponding to the B partially bootstrapped samples of Y.

The complete data goodness-of-fit term reduces to

$$\begin{aligned}\text{GOF}_{comp} &= \frac{1}{B}\sum_{b=1}^{B} -2\log L(\hat{\beta}^*_{par}(b), \hat{\Sigma}^*_{par}(b)|Y^*_{par}(b)) \\ &= \frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}^*_{par}(b)| + N\Big\}. \end{aligned} \tag{6.9}$$

The four complete data criteria proposed in section (3.4) are now given, which should have expectations that are approximately equal to the expected complete data KL discrepancy. Again, let k denote the number of functionally independent parameters in the candidate model. Again, assuming that Σ is an unstructured covariance matrix, we have $k = p + .5t(t+1)$. The complete data analogues of AIC and AICc are

$$\begin{aligned}\text{AIC}_{comp} &= \frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}^*_{par}(b)| + N\Big\} + 2k, \quad \text{and} \\ \text{AICc}_{comp} &= \frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}^*_{par}(b)| + N\Big\} + 2k(N/(N-k-1)).\end{aligned}$$

In order to construct the estimators of the expected optimism in the complete data bootstrap criteria, bootstrap samples of both the observed and missing elements in Y must be obtained. The bootstrap vectors y^*_i in a bootstrap sample Y^* are generated case-by-case by taking a draw from the distribution of the fitted candidate model as given in (6.8). Let $\{Y^*(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y generated as described. Let $\{\hat{\beta}^*(b)|b = 1, 2, \ldots, B\}$ and $\{\hat{\Sigma}^*(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap replicates of MLEs of β and Σ, corresponding to the B bootstrap samples of Y.

The complete data analogue of EIC simplifies to

$$\begin{aligned}\text{EIC}_{comp} &= \frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}^*_{par}(b)| + N\Big\} \\ &+ \frac{1}{B}\sum_{b=1}^{B}\left\{\sum_{i=1}^{n}\Big((y^*_{par,i}(b) - X_i\hat{\beta}^*(b))'\hat{\Sigma}^*(b)^{-1}(y^*_{par,i}(b) - X_i\hat{\beta}^*(b))\Big) - N\right\}.\end{aligned}$$

The complete data analogue of AICb simplifies to

$$\text{AICb}_{comp} = \frac{1}{B}\sum_{b=1}^{B}\Big\{n\log|\hat{\Sigma}^*_{par}(b)| + N\Big\}$$
$$+ 2\Bigg[\frac{1}{B}\sum_{b=1}^{B}\Bigg\{n\log|\hat{\Sigma}^*(b)| + \sum_{i=1}^{n}\Big((y^*_{par,i}(b) - X_i\hat{\beta}^*(b))'\hat{\Sigma}^*(b)^{-1}(y^*_{par,i}(b) - X_i\hat{\beta}^*(b))\Big)$$
$$- (n\log|\hat{\Sigma}^*_{par}(b)| + N)\Bigg\}\Bigg].$$

6.2.3 Simulation Study

Consider a setting where a sample of n outcome vectors are generated according to the true model (6.1). Suppose our objective is to search among a class of nested candidate models for the fitted candidate model that best approximates the true model (6.1). The candidate models are assumed to be of the form (6.2).

In the simulation study, we consider a setting where each of the n outcome vectors has a length of $t = 2$; thus, Y is a $2n \times 1$ vector. For case i, let the first and second elements of y_i be represented as $y_i = (y_{i1}, y_{i2})$.

From model (6.2), we can determine that X_i is a $2 \times p$ known design matrix of covariates for case i. Consider a class of five nested candidate models where the number of covariates range from 1 to 5, corresponding to design matrices having from $p = 2$ to $p = 6$ columns. We will refer to p as the order of the candidate model. The first column in each X_i is taken to be a vector of ones. The other covariate values of each X_i were generated independently from a $N(0, 1)$ distribution.

Simulation sets with 200 samples of data of $n = 60$ were generated from a model of the form (6.1) with one of two types of covariance structures:

$$y_i = X_{o,i}\beta_o + e_{o,i}, \text{ where } \Sigma_o = \begin{bmatrix} 25 & 0 \\ 0 & 25 \end{bmatrix} \text{ or } \Sigma_o = \begin{bmatrix} 25 & 20 \\ 20 & 25 \end{bmatrix}, \text{for } i = 1, 2, \ldots, n.$$

Here, β_o is a 5×1 vector consisting of all ones. In the first covariance structure,

$\text{corr}(y_{i1}, y_{i2}) = 0$, and in the second covariance structure, $\text{corr}(y_{i1}, y_{i2}) = .8$. Note that the total number of generated observations for each sample is $N = 120$.

The order of the true model, or true order p_o, is five. In the nested model setting that we are considering, the true model includes the first four covariates. An underspecified model would include the first covariate, the first two covariates, or the first three covariates. An overspecified model would include all five covariates. All candidate models include an intercept.

For some of the cases, y_{i1} or y_{i2} were randomly discarded so as to create a situation where the missing data in Y are MAR. Let $\Pr(y_{i1} \text{ mis})$ denote the probability that for the i^{th} case, y_{i1} is discarded and y_{i2} is retained, and let $\Pr(y_{i2} \text{ mis})$ denote the probability that for the i^{th} case, y_{i2} is discarded and y_{i1} is retained. We considered discard probabilities of ($\Pr(y_{i1} \text{ mis})$, $\Pr(y_{i2} \text{ mis})$) set at (0.0, 0.0), (0.15, 0.15), and (0.30, 0.30).

In summary, simulation results from a total of six different settings will be presented. We consider all possible combinations of the following factors: two different correlation levels of $\text{corr}(y_{i1}, y_{i2}) = 0$ and $\text{corr}(y_{i1}, y_{i2}) = .8$, and three levels of missingness of $(\Pr(y_{i1} \text{ mis}), \Pr(y_{i2} \text{ mis}))$ set at (0.0, 0.0), (0.15, 0.15), and (0.30, 0.30).

For each of the six settings, the five candidate models were fit for each sample assuming an unstructured covariance matrix for Σ. The complete data and observed data analogues of AIC, AICc, EIC, and AICb were then calculated for every set of fitted candidate models. The overall complete data KL discrepancy, which is found by substituting the maximum likelihood estimators based on Y for β and Σ in (6.7), along with the overall observed data KL discrepancy, which is obtained by substituting the maximum likelihood estimators based on Y_{obs} for β and Σ in (6.6), were also computed under the six settings for each sample and candidate model. The model selected by each of the observed data criteria and complete data criteria,

along with both discrepancies, was determined. For every candidate model, a total of 250 bootstrap samples were generated for EIC and AICb, along with 250 partially bootstrapped samples for the complete data criteria.

To characterize the effectiveness of the complete and observed data criteria, we use the two performance measures given in section (4.2.3): the average criterion values and distribution of model selections. The results from the six simulation settings are summarized in figures (6.1), (6.2), and (6.3), and in tables (6.1) and (6.2).

The figures (6.1), (6.2), and (6.3) demonstrate how effectively the criteria estimate their target discrepancies. With no missing data (see figure 6.1), all criteria estimate the KL target reasonably well at both correlation levels. For the correctly specified or underspecified models, the slopes of the criterion curves become more steep with $\text{corr}(y_{i1}, y_{i2}) = .8$, which will consequently aid in selecting more models of the true order.

The complete data criterion curves are quite similar when level of missingness are moderate (see figure 6.2) or high (see figure 6.3). Once again, the slopes of the curves are more steep at the high level of correlation. The criterion AICb_{comp} estimates the complete data target very well. The other criteria, AIC_{comp}, AICc_{comp}, and EIC_{comp}, tend to have mean values that are below the complete data target, although the slopes of their curves generally parallel the complete data target. Similar trends hold for the observed data criteria.

Tables (6.1) and (6.2) demonstrate the model selections for each of the criteria. Under a setting of no missing data, all four criteria show similar distributions of model selections. For the setting with no correlation (see table 6.1), the complete data criteria and the observed data criteria have a very similar number of correct model selections at the moderate level of missingness. At the highest level of missingness, the complete data criteria select more correctly specified models than their

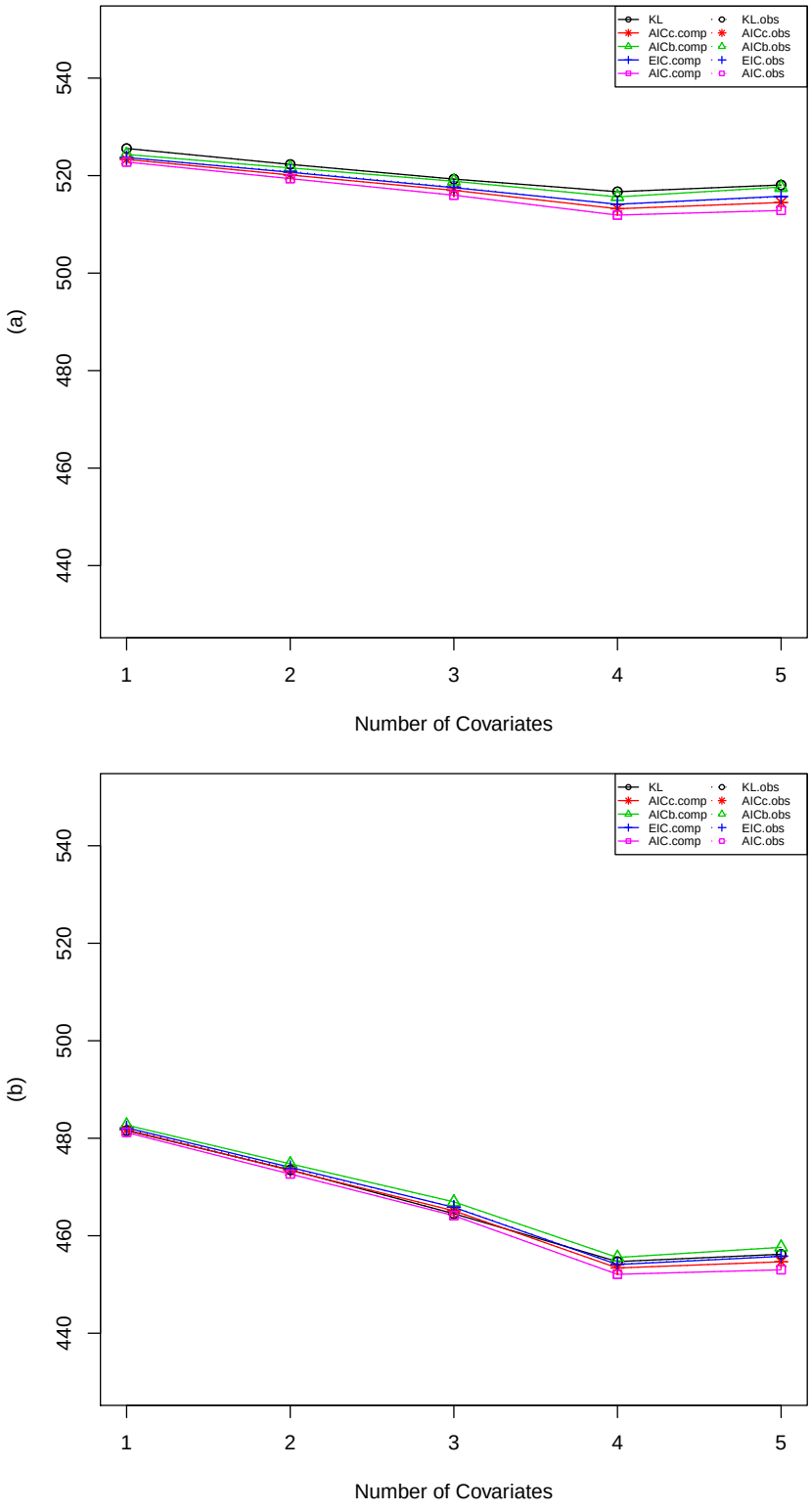

Figure 6.1: Average criterion values under longitudinal regression setting with missing data in the outcome: $(\Pr(y_{i1} \text{ mis}), \Pr(y_{i2} \text{ mis})) = (0, 0)$. (a) $\text{corr}(y_{i1}, y_{i2}) = 0$; (b) $\text{corr}(y_{i1}, y_{i2}) = .8$.

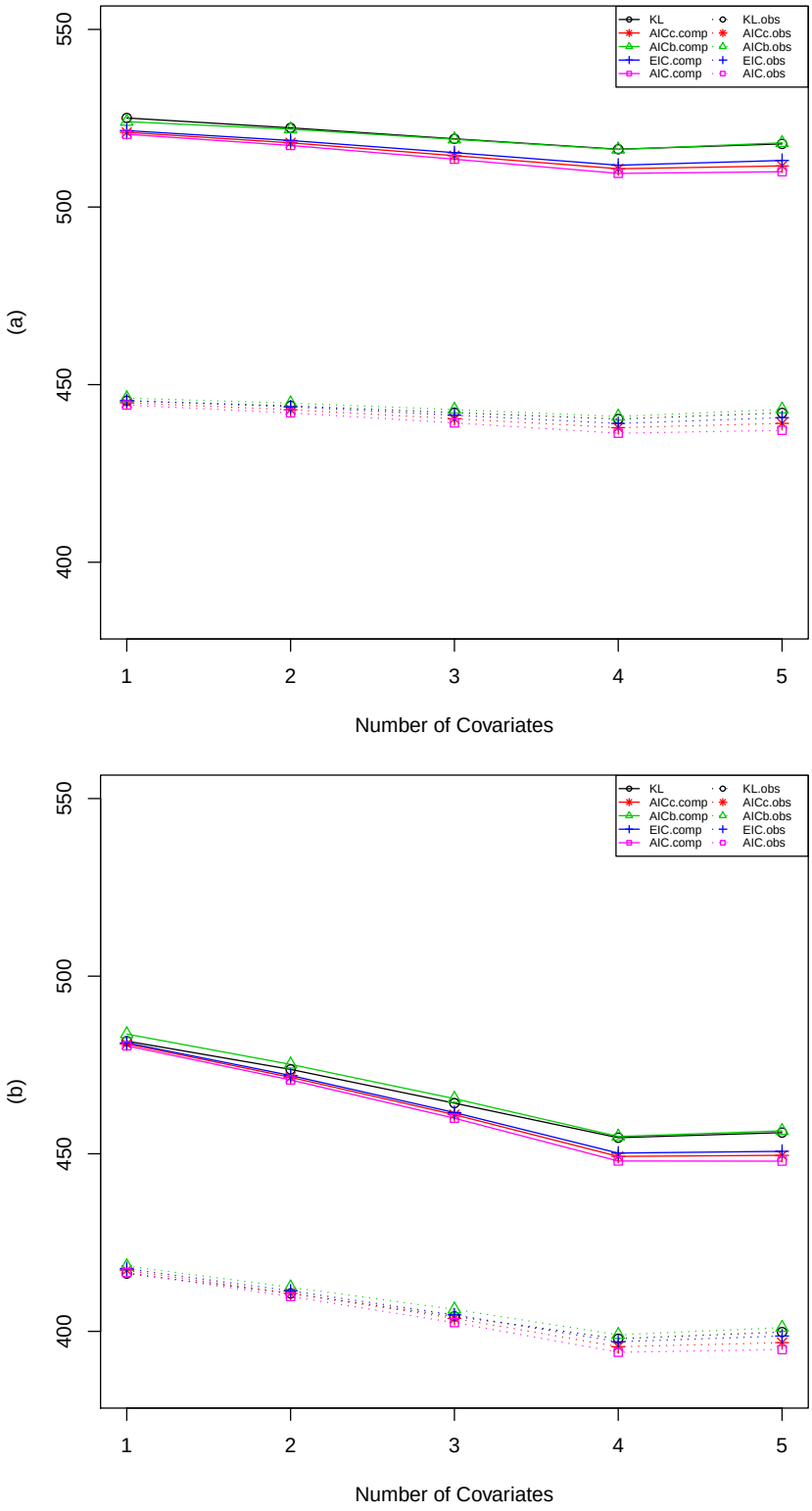

Figure 6.2: Average criterion values under longitudinal regression setting with missing data in the outcome: $(\Pr(y_{i1} \text{ mis}), \Pr(y_{i2} \text{ mis})) = (.15, .15)$. (a) $\text{corr}(y_{i1}, y_{i2}) = 0$; (b) $\text{corr}(y_{i1}, y_{i2}) = .8$.

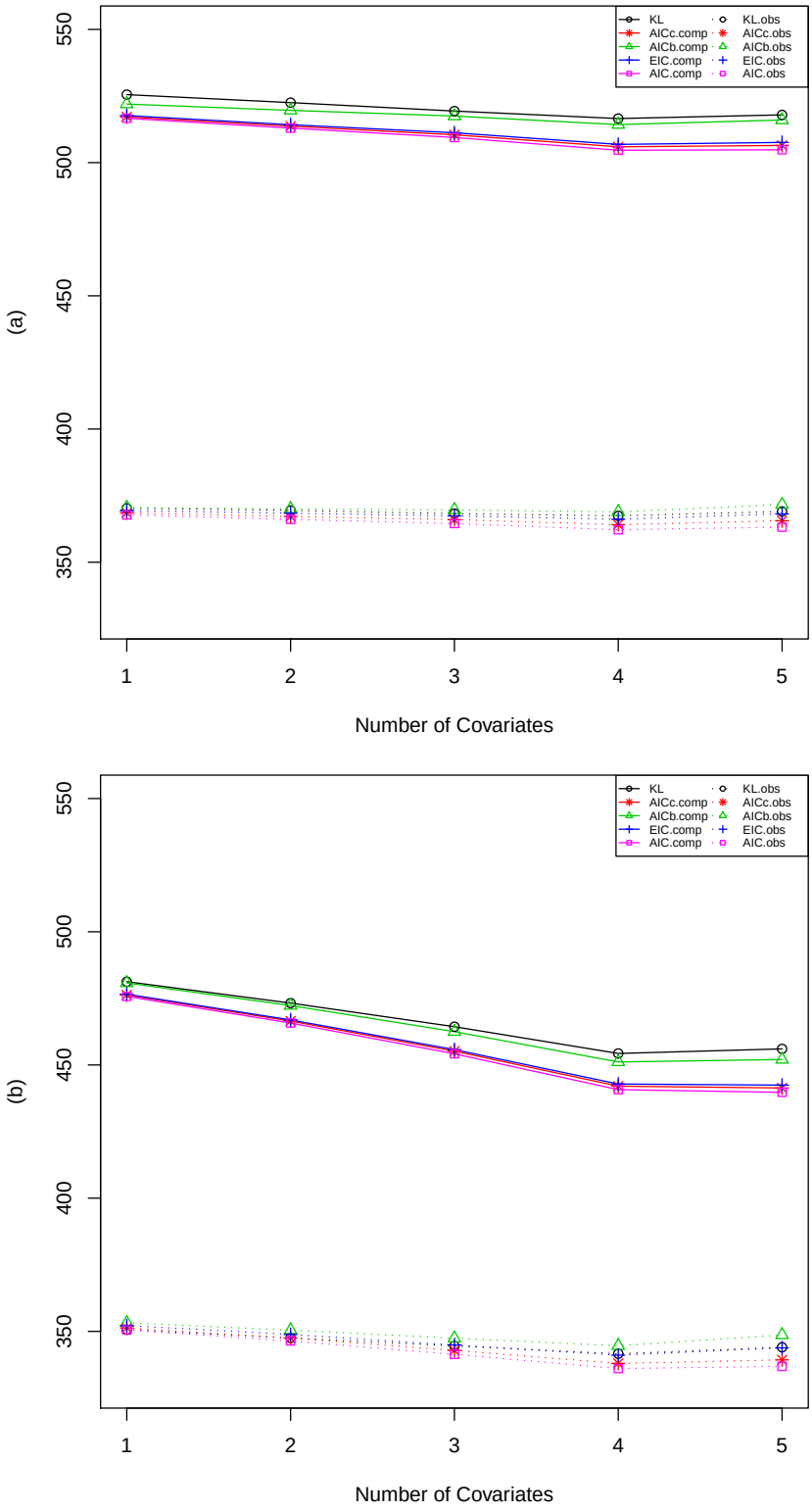

Figure 6.3: Average criterion values under longitudinal regression setting with missing data in the outcome: $(\Pr(y_{i1} \text{ mis}), \Pr(y_{i2} \text{ mis})) = (.30, .30)$. (a) $\text{corr}(y_{i1}, y_{i2}) = 0$; (b) $\text{corr}(y_{i1}, y_{i2}) = .8$.

Table 6.1: Frequency of criterion selections under longitudinal regression setting with missing data in the outcome: $\text{corr}(y_{i1}, y_{i2}) = 0, N = 120$.

$\Pr(y_{i1}$ mis),		Complete Data					Observed Data				
$\Pr(y_{i2}$ mis)	Covariates	KL	AICc	AIC	EIC	AICb	KL	AICc	AIC	EIC	AICb
(0, 0)	1	3	11	10	12	17	3	11	10	12	15
	2	3	18	14	21	22	3	18	14	10	21
	3	18	30	27	38	36	18	30	27	46	44
	4	153	115	117	102	110	153	115	117	103	103
	5	23	26	32	27	15	23	26	32	29	17
(.15, .15)	1	5	10	8	9	20	16	17	11	22	31
	2	5	14	10	19	20	7	20	14	26	25
	3	14	32	31	33	35	27	42	37	43	40
	4	154	105	110	95	100	134	98	105	91	89
	5	22	39	41	44	25	16	23	33	18	15
(.30, .30)	1	3	11	8	15	30	19	37	29	44	60
	2	2	14	11	9	26	14	30	24	36	40
	3	22	33	29	37	42	26	47	43	47	42
	4	161	95	95	78	73	127	75	79	65	51
	5	12	47	57	61	29	14	11	25	8	7

Table 6.2: Frequency of criterion selections under longitudinal regression setting with missing data in the outcome: $\text{corr}(y_{i1}, y_{i2}) = .8, N = 120$.

$\Pr(y_{i1}$ mis),		Complete Data					Observed Data				
$\Pr(y_{i2}$ mis)	Covariates	KL	AICc	AIC	EIC	AICb	KL	AICc	AIC	EIC	AICb
(0, 0)	1	1	0	0	0	0	1	0	0	0	0
	2	0	0	0	0	1	0	0	0	1	1
	3	3	2	1	6	5	3	2	1	0	4
	4	179	168	166	155	169	179	168	166	161	170
	5	17	30	33	39	25	17	30	33	38	25
(.15, .15)	1	0	0	0	0	1	0	1	1	0	2
	2	0	0	0	0	2	3	1	1	3	3
	3	3	7	4	9	11	13	10	8	12	11
	4	181	132	125	112	144	173	149	142	143	157
	5	16	61	71	79	42	11	39	48	42	27
(.30, .30)	1	1	0	0	0	0	7	0	0	5	15
	2	0	2	2	2	4	15	12	9	18	22
	3	1	19	15	22	30	23	31	29	42	50
	4	177	107	101	99	113	138	132	129	116	103
	5	21	72	82	77	53	17	25	33	19	10

observed data counterparts. Furthermore, as the amount of missingness increases, the tendency of the observed data criteria is to select underspecified models, yet the complete data criteria protect against this tendency. This tendency is due to the fact that the overall selection patterns of the observed data target worsen as the missingness increases; consequently, the criteria that estimate this target tend to under perform in selecting the correct model.

At the high correlation level (see table 6.2), the observed data criteria outperform the complete data criteria with a moderate level of missingness. At the highest level of missingness, the observed data criteria select more correctly specified models than the observed data criteria; however, we see again that the complete data criteria select fewer underspecified models than the observed data criteria.

Overall, both the complete data and the observed data criteria have better selection performance at the high correlation setting than compared to the setting with no correlation. In the simulation study presented in subsection (4.2.3) under a normal bivariate linear regression model framework, we saw an opposite result. In that framework, the complete data and fully observed data criteria exhibited better selection performance at the setting with no correlation than compared to the high correlation setting.

Often, model selection criteria are used to determine a model with optimal properties for prediction of future observations. In section (2.1), we introduced the concepts of overfitting and underfitting, which directly impact prediction. As stated in Burnham and Anderson (2002, p.32), the use of an underfitted model to predict will often result in severely biased predictions. Conversely, an overfitted model may lead to predictions that are unbiased yet are potentially highly variable. Burnham and Anderson, in reference to Shibata (1989), state "Shibata argues that underfitted models are a more serious issue in data analysis and inference than overfitted models." One may conclude from the results of the simulation study that

in smaller sample-size settings and in other settings conducive to underfitting, an underfit model is often selected if the observed data criteria are utilized. Thus, using a model selected from the observed data criteria may lead to less accurate predictions since the fitted model is more likely to be underfit than if the complete data criteria were used to select a model. The greatest strength of the complete data criteria is the protection against underfitting, which could potentially lead to better predictions.

6.3 Missing Data in the Outcome and/or Covariates

We have previously demonstrated in subsection (6.2.2) how to calculate the complete data criteria when there is potentially missing data in each y_i from candidate model (6.2). In research settings, it may be very likely to also have missing elements in the covariate matrix X_i. Recall, that most statistical packages do not use a row in a longitudinal data file if the row contains at least one missing element. Thus, additional missing data in the covariates can possibly be even more problematic since the total number of observations with no missing data that are used to select a model can quickly dwindle.

In order to construct the complete data criteria in settings where there potentially is missing data in y_i and/or X_i, bootstrap samples of the missing elements in both the outcome and/or covariates must be obtained. The bootstrap samples are generated by taking a draw from the conditional distribution of the missing data given the observed data. To postulate the conditional distribution, one must first assume a joint distribution for the outcome and covariates, or a distribution for the covariates given the outcome. Secondly, the parameters must be estimated for the conditional distribution of the missing data given the observed data. Both of these tasks may be difficult to accomplish in a longitudinal data setting.

Currently, the majority of the literature for missing data in a normal longitudinal regression model setting pertains strictly to missing data in the outcome, and the covariates are assumed to be observed. A few EM algorithm type approaches exist for missing data in the outcome and/or covariates in a longitudinal regression model setting, and are found in Roy and Lin (2002), Stubbendick and Ibrahim (2003), and Roy and Lin (2005). Ideally, parameter estimates for the posited conditional distribution of the missing data given the observed data can be provided using the given approaches, or other approaches. Bootstrap samples of the missing data can then be generated. Once the distributions and the bootstrapping approach are established, one can follow the techniques of subsection (4.3.3) in calculating the complete data criteria. Further investigation of these ideas is left as an area of possible future research.

CHAPTER 7
GENERAL DISCREPANCY-BASED MODEL SELECTION CRITERIA

In this chapter, we present an overview of general discrepancy-based model selection criteria under the assumption of no missing data. We then propose analogous criteria that can be implemented in incomplete data settings. Under both settings, the criteria can be formulated using a wide variety of discrepancies.

7.1 General Discrepancy-Based Criteria With No Missing Data

Suppose that we have a collection of data $Y = \{y_1, y_2, \ldots, y_n\}$, where the y_is may be scalars or vectors. Let $L(\theta_o|Y)$ (or equivalently, $f(Y|\theta_o)$) denote an unknown parametric model that presumably generated Y. Let θ denote a k-dimensional parameter vector that is an element of the parameter space Θ ($\theta \in \Theta$), and let $L(\theta|Y)$ (or equivalently, $f(Y|\theta)$) denote the candidate model. For now, assume that Y contains *no* missing values.

As discussed in section (2.1), a family of parametric candidate models is often postulated, one that is comprised of a collection of models that could potentially describe the data. A useful way of selecting a model from the candidate family is to assign scores to the fitted candidate models that can be used to assess the propriety of each model. A discrepancy is a measure that reflects the disparity between two statistical models, and can therefore be used to develop criteria that provide such scores.

Many discrepancies exist that can serve as the basis for the development of model selection criteria. In section (3.1), we defined the KL discrepancy, which is applicable in nearly all modeling frameworks and serves as the basis for AIC and its variants. Another widely used discrepancy is the Gauss discrepancy, which is a measure that reflects the mean squared error of prediction. The Gauss discrepancy

serves as the basis for Mallows' (1973) C_p and Akaike's (1970) final prediction error (FPE). Other discrepancies include the Pearson chi-squared discrepancy and the Neyman chi-squared discrepancy, which are often utilized in discrete data settings, the Kullback symmetric divergence (Cavanaugh, 1999), and a Gauss-type discrepancy based on the median squared error of prediction (Neath and Cavanaugh, 2000).

For our purpose, we consider discrepancies that reflect the separation between the true model $L(\theta_o|Y)$ and the candidate model $L(\theta|Y)$. We will assume that these discrepancies are of the following form:

$$\Delta(\theta) = E_o\{\delta(Y,\theta)\},$$

where E_o denotes the expectation under the true model. Here, $\delta(Y,\theta)$ represents a function that gauges the accuracy with which Y is predicted under the candidate model (parameterized by θ), where smaller values of $\delta(Y,\theta)$ imply greater predictive accuracy. The discrepancy $\Delta(\theta)$ shares the first of the following properties of a distance function, $f(a,b)$.

1. $f(a,b) \geq f(a,a)$ for all a, b.
2. $f(a,a) = 0$.
3. Symmetry: $f(a,b) = f(b,a)$.
4. Triangle Inequality: $f(a,c) \leq f(a,b) + f(b,c)$.

Let $\hat{\theta}$ denote an estimator of θ, which is obtained as follows:

$$\hat{\theta} = \text{argmin}_{\theta\in\Theta}\delta(Y,\theta). \tag{7.1}$$

Such an estimator $\hat{\theta}$ is called the *minimum discrepancy estimator* (MDE).

The *overall discrepancy* results from evaluating the discrepancy at $\theta = \hat{\theta}$, and is depicted as

$$\Delta(\hat{\theta}) = E_o\{\delta(Y,\theta)\}|_{\theta=\hat{\theta}}.$$

The overall discrepancy reflects the disparity between the true model $L(\theta_o|Y)$ and the fitted candidate model $L(\hat{\theta}|Y)$.

The *expected (overall) discrepancy* is obtained by averaging the overall discrepancy over the sampling distribution of $\hat{\theta}$, and is therefore given by

$$E_o\{\Delta(\hat{\theta})\} = E_o\{E_o\{\delta(Y,\theta)\}|_{\theta=\hat{\theta}}\}.$$

The expected discrepancy reflects how well, on average, the fitted candidate model predicts "new" data generated under the true model. A model selection criterion is often formulated by constructing a statistic that has an expectation equal to (or approximately equal to) $E_o\{\Delta(\hat{\theta})\}$.

The *estimated discrepancy* is given by

$$\hat{\Delta}(\hat{\theta}) = \delta(Y,\hat{\theta}),$$

and is a natural estimator of the expected discrepancy. The estimated discrepancy reflects how well the fitted model predicts the data at hand. By evaluating the adequacy of the fitted model based on its ability to recover the data used in its own construction, the estimated discrepancy yields an overly optimistic assessment of how effectively the fitted model predicts new data. Thus, the estimated discrepancy serves as a negatively biased estimator of the expected discrepancy.

We will employ the estimated discrepancy $\hat{\Delta}(\hat{\theta})$ as a platform for estimating the expected discrepancy $E_o\{\Delta(\hat{\theta})\}$. Consider writing $E_o\{\Delta(\hat{\theta})\}$ as follows:

$$E_o\{\Delta(\hat{\theta})\} = E_o\{\hat{\Delta}(\hat{\theta})\} + \left[E_o\{\Delta(\hat{\theta}) - \hat{\Delta}(\hat{\theta})\}\right]. \tag{7.2}$$

The bracketed quantity on the right is often referred to as the *expected optimism* in judging the fit of a model using the same data as that which was used to construct the fit (Efron, 1983, 1986). The expected optimism is positive, again implying that $\hat{\Delta}(\hat{\theta})$ is a negatively biased estimator of $E_o\{\Delta(\hat{\theta})\}$. In order to correct for the negative bias, we must evaluate or approximate the bias adjustment represented by

the expected optimism. The estimator of the expected optimism, say $\widehat{eo}$, provides a bias correction which is then added to $\hat{\Delta}(\hat{\theta})$ to produce an approximately unbiased estimator of the expected discrepancy $E_o\{\Delta(\hat{\theta})\}$:

$$\hat{\Delta}(\hat{\theta}) + \widehat{eo}.$$

The preceding is a conventional form of a model selection criterion. The estimated discrepancy $\hat{\Delta}(\hat{\theta})$ functions as a goodness-of-fit term, and $\widehat{eo}$ serves as a penalty term for model complexity.

There exist numerous approaches to estimating the expected optimism. These include deriving an exact expression (Suguira, 1978; Hurvich and Tsai, 1989; Bedrick and Tsai, 1994), or using an approximation based on Monte Carlo simulation (Hurvich, Shumway and Tsai, 1990; Bengtsson and Cavanaugh, 2006). We will now outline two other methods: developing an estimator based on large-sample arguments, and formulating an estimator based on the bootstrap.

Linhart and Zucchini (1986) derive an estimator of the expected optimism based on large-sample arguments. They develop this estimator by employing second-order Taylor series expansions. This approach leads to a bias correction based on the expectation of the Hessian matrix of $\delta(Y,\theta)$ evaluated at $\hat{\theta}$, given as

$$\Omega(\hat{\theta}) = E_o\left[\frac{\partial^2\delta(Y,\theta)}{\partial\theta\partial\theta'}\right]|_{\theta=\hat{\theta}},$$

and the expectation of the cross products of the first-order partial derivatives of $\delta(Y,\theta)$ evaluated at $\hat{\theta}$, depicted as

$$\Sigma(\hat{\theta}) = E_o\left[\left\{\frac{\partial\delta(Y,\theta)}{\partial\theta}\right\}\left\{\frac{\partial\delta(Y,\theta)}{\partial\theta}\right\}'\right]|_{\theta=\hat{\theta}}.$$

In general, $\Omega(\hat{\theta})$ and $\Sigma(\hat{\theta})$ will depend on θ_o, and will not be directly accessible. The matrix

$$\hat{\Omega}(\hat{\theta}) = \frac{\partial^2\delta(Y,\hat{\theta})}{\partial\theta\partial\theta'}$$

is often used to estimate $\Omega(\hat{\theta})$. Similarly, the matrix

$$\hat{\Sigma}(\hat{\theta}) = \left\{ \frac{\partial \delta(Y, \hat{\theta})}{\partial \theta} \right\} \left\{ \frac{\partial \delta(Y, \hat{\theta})}{\partial \theta} \right\}'$$

if often used to estimate $\Sigma(\hat{\theta})$.

Using these estimators, an approximately unbiased estimator of the expected discrepancy $E_o\{\Delta(\hat{\theta})\}$ can be given as

$$\hat{\Delta}(\hat{\theta}) + \text{tr}\left\{ [\hat{\Omega}(\hat{\theta})]^{-1} \hat{\Sigma}(\hat{\theta}) \right\}. \tag{7.3}$$

Here, the penalty term $\text{tr}\{[\hat{\Omega}(\hat{\theta})]^{-1}\hat{\Sigma}(\hat{\theta})\}$ serves as an estimator of the bias adjustment $\text{tr}\{[\Omega(\hat{\theta})]^{-1}\Sigma(\hat{\theta})\}$. For the KL discrepancy, the criterion (7.3) is equivalent to a criterion developed by Takeuchi (1976) known as the Takeuchi information criterion (TIC).

If one assumes that the candidate model is correctly specified or overfit, then under some discrepancies, the bias adjustment $\text{tr}\{[\Omega(\hat{\theta})]^{-1}\Sigma(\hat{\theta})\}$ asymptotically reduces to a function of k, the dimension of the parameter vector θ. For example, under the KL discrepancy, as well as both the Pearson chi-squared and Neyman chi-squared discrepancies, $\text{tr}\{[\Omega(\hat{\theta})]^{-1}\Sigma(\hat{\theta})\}$ asymptotically reduces to $2k$. Hence, in large-sample settings, under the assumption that the candidate model is correctly specified or overfit, the criteria TIC and AIC are equivalent. Even when the assumption is violated, estimators of the expected optimism based on the function of k may still be used and may perform well, unless the true and fitted candidate models are very different (Linhart and Zucchini, 1986, p.22). The estimator of the expected optimism based on the function of k can also be used if the estimators $\hat{\Omega}(\hat{\theta})$ and $\hat{\Sigma}(\hat{\theta})$ are potentially inaccurate, such as in settings when the sample size is small or the data is highly variable.

We now present an estimator of the expected optimism based on techniques relating to the bootstrap. Following Efron (1983, 1986) and Konishi and Kitagawa

(2008), assume that the bootstrap distribution corresponds to the empirical distribution $\hat{F}(Y)$ of the original sample, which assigns probability $1/n$ to each of the cases comprising $Y = \{y_1, y_2, \ldots, y_n\}$. The bootstrap sample of Y is obtained by generating a random sample of size n drawn with replacement from $\hat{F}(Y)$ and is denoted as Y^*. Let $\hat{\theta}^*$ denote the bootstrap replicate of the MLE of θ based on Y^*. The bootstrap analogue of the expected optimism is

$$\left[E_*\{\hat{\Delta}(\hat{\theta}^*) - \hat{\Delta}^*(\hat{\theta}^*)\}\right],$$

where * corresponds to the bootstrap sample, E_* represents the expectation taken with respect to the distribution of the bootstrap sample, $\hat{\Delta}(\hat{\theta}^*) = \delta(Y, \hat{\theta}^*)$, and $\hat{\Delta}^*(\hat{\theta}^*) = \delta(Y^*, \hat{\theta}^*)$.

Now, let $\{Y^*(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples. Let $\{\hat{\theta}^*(b)|b = 1, 2, \ldots, B\}$ represent a set of B bootstrap replicates of MLEs of θ corresponding to the B bootstrap samples. The preceding bootstrap analogue of the expected optimism can be estimated by

$$\frac{1}{B}\sum_{b=1}^{B}\left\{\hat{\Delta}(\hat{\theta}^*(b)) - \hat{\Delta}^*(\hat{\theta}^*(b))\right\}. \tag{7.4}$$

An approximately unbiased estimator of the expected discrepancy $E_o\{\Delta(\hat{\theta})\}$ is thereby given as

$$\hat{\Delta}(\hat{\theta}) + \frac{1}{B}\sum_{b=1}^{B}\left\{\hat{\Delta}(\hat{\theta}^*(b)) - \hat{\Delta}^*(\hat{\theta}^*(b))\right\}. \tag{7.5}$$

Under the KL discrepancy, (7.5) is equivalent to EIC, which was outlined in section (3.1).

7.2 General Discrepancy-Based Criteria Using the Complete Data

In this section, we will assume that some of the elements in Y are missing. We propose analogues of the discrepancy-based criteria, (7.3) and (7.5), that are

based on the complete data $Y = (Y_{obs}, Y_{mis})$. The criteria can be utilized in incomplete data settings. The goal is to develop criteria that have expectations equal to (or approximately equal to) the expected discrepancy $E_o\{\Delta(\hat{\theta})\}$. Recall the representation of the expected discrepancy in (7.2). We first propose a complete data term which approximates $E_o\{\hat{\Delta}(\hat{\theta})\}$, and then two complete data estimators of the expected optimism, which approximate $[E_o\{\Delta(\hat{\theta}) - \hat{\Delta}(\hat{\theta})\}]$.

In settings with no missing data, the MDE was defined as an estimator of θ in equation (7.1). The MDE may be difficult to obtain in incomplete data settings; thus, other estimators may be more practical. For the remainder of this chapter, we propose using the maximum likelihood estimator (MLE) obtained via the EM algorithm in developing the complete data criteria. Note that the large-sample based model selection criterion (7.3) is only justified if the estimator of θ is the MDE. The bootstrap based model selection criterion (7.5) is also outlined using the MDE. However, in some instances, such as when the KL discrepancy is used, the MDE and the MLE are equivalent. Also, with the Gauss discrepancy, the MDE and MLE for the regression parameters in the normal linear regression setting are equivalent. In other instances, the MDE and MLE may be asymptotically equivalent. Let $\hat{\theta}_{obs}$ denote the MLE for a particular candidate model that is found by maximizing $L(\theta|Y_{obs})$ via the EM algorithm.

When Y contains no missing data, a natural estimator of $E_o\{\hat{\Delta}(\hat{\theta})\}$ is simply the estimated discrepancy $\hat{\Delta}(\hat{\theta}) = \delta(Y, \hat{\theta})$. However, in a setting with missing data, the estimated discrepancy is inaccessible since some of the elements of Y are unobserved. We propose to use $\hat{\theta}_{obs}$ obtained from the EM algorithm, along with techniques related to the bootstrap as discussed in section (3.4) in developing a complete data analogue of $\delta(Y, \hat{\theta})$.

Let $\{Y^*_{mis}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples of Y_{mis}, generated by taking draws from the conditional distribution $f(Y_{mis}|Y_{obs}, \hat{\theta}_{obs})$.

Let $\{Y^*_{par}(b) = (Y_{obs}, Y^*_{mis}(b))|b = 1, 2, \ldots, B\}$ represent a collection of B partially bootstrapped samples. Let $\{\hat{\theta}^*_{par}(b)|b = 1, 2, \ldots, B\}$ represent a collection of B partial bootstrap replicates of MLEs of θ corresponding to the B partially bootstrapped samples. For the b^{th} bootstrap sample, let $\hat{\Delta}^*_{par}(b)(\hat{\theta}^*_{par}(b)) = \delta(Y^*_{par}(b), \hat{\theta}^*_{par}(b))$. The complete data analogue of the estimated discrepancy using partial bootstrapping is now proposed as

$$\frac{1}{B}\sum_{b=1}^{B}\left\{\hat{\Delta}^*_{par}(b)(\hat{\theta}^*_{par}(b))\right\} = \frac{1}{B}\sum_{b=1}^{B}\left\{\delta(Y^*_{par}(b), \hat{\theta}^*_{par}(b))\right\}. \tag{7.6}$$

We now develop two estimators of the expected optimism based on the complete data.

In settings where there is no missing data, recall that Linhart and Zucchini use large-sample arguments to show that $\text{tr}\{[\hat{\Omega}(\hat{\theta})]^{-1}\hat{\Sigma}(\hat{\theta})\}$ serves as an estimator of the expected optimism used in (7.3). If the degree of missingness of Y is not too extreme, and the MLE and MDE are at least asymptotically equivalent, then the large-sample result should still hold. The estimators $\hat{\Omega}(\hat{\theta})$ and $\hat{\Sigma}(\hat{\theta})$ both depend on Y and are unattainable if some of the elements of Y are missing.

In the complete data analogue of the estimated discrepancy (7.6), the missing elements of Y are replaced with bootstrap samples obtained by taking draws from $f(Y_{mis}|Y_{obs}, \hat{\theta}_{obs})$. In conjunction with the complete data analogue of the estimated discrepancy, Y can be replaced with the collection of B partially bootstrapped samples $\{Y^*_{par}(b) = (Y_{obs}, Y^*_{mis}(b))|b = 1, 2, \ldots, B\}$. Furthermore, the partial bootstrap replicates $\{\hat{\theta}^*_{par}(b)|b = 1, 2, \ldots, B\}$ corresponding to the B partially bootstrapped samples can also be used in the estimators. A complete data analogue of $\hat{\Omega}(\hat{\theta})$ based on the b^{th} partially bootstrapped sample is

$$\hat{\Omega}^*_{par}(b)(\hat{\theta}^*_{par}(b)) = \frac{\partial^2\delta(Y^*_{par}(b), \hat{\theta}^*_{par}(b))}{\partial\theta\partial\theta'}.$$

Similarly, a complete data estimate of $\hat{\Sigma}(\hat{\theta})$ based on the b^{th} partially bootstrapped

sample is

$$\hat{\Sigma}^*_{par}(b)(\hat{\theta}^*_{par}(b)) = \left\{ \frac{\partial \delta(Y^*_{par}(b), \hat{\theta}^*_{par}(b))}{\partial \theta} \right\} \left\{ \frac{\partial \delta(Y^*_{par}(b), \hat{\theta}^*_{par}(b))}{\partial \theta} \right\}' .$$

A complete data analogue of the large-sample based criterion (7.3) is

$$\frac{1}{B}\sum_{b=1}^{B} \left\{ \hat{\Delta}^*_{par}(b)(\hat{\theta}^*_{par}(b)) \right\} + \frac{1}{B}\sum_{b=1}^{B} \left\{ \text{tr} \left\{ [\hat{\Omega}^*_{par}(b)(\hat{\theta}^*_{par}(b))]^{-1} \hat{\Sigma}^*_{par}(b)(\hat{\theta}^*_{par}(b)) \right\} \right\} .$$

In settings where there is no missing data, a bootstrap based estimator of the expected optimism is given by (7.4). The estimator depends on Y and is unattainable if some of the elements of Y are missing. Again, in conjunction with the complete data analogue of the estimated discrepancy, Y can be replaced with the collection of B partially bootstrapped samples $\{Y^*_{par}(b) = (Y_{obs}, Y^*_{mis}(b)) | b = 1, 2, \ldots, B\}$ as needed in the bootstrap based estimator.

The maximum likelihood estimates obtained using the EM algorithm can also be used to create parametric bootstrap samples of the entire sample $Y = (Y_{obs}, Y_{mis})$, which are needed for the bootstrap based estimator of the expected optimism. Let $\{Y^*(b) | b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap samples generated by taking draws from the fitted candidate model $f(Y|\hat{\theta}_{obs})$. Let $\{\hat{\theta}^*(b) | b = 1, 2, \ldots, B\}$ represent a collection of B bootstrap replicates of MLEs of θ corresponding to the B bootstrap samples. For the b^{th} bootstrap sample, let $\hat{\Delta}^*_{par}(b)(\hat{\theta}^*(b)) = \delta(Y^*_{par}(b), \hat{\theta}^*(b))$, and let $\hat{\Delta}^*(b)(\hat{\theta}^*(b)) = \delta(Y^*(b), \hat{\theta}^*(b))$. A complete data analogue of the bootstrap based criterion (7.5) is then given by

$$\frac{1}{B}\sum_{b=1}^{B} \left\{ \hat{\Delta}^*_{par}(b)(\hat{\theta}^*_{par}(b)) \right\} + \frac{1}{B}\sum_{b=1}^{B} \left\{ \hat{\Delta}^*_{par}(b)(\hat{\theta}^*(b)) - \hat{\Delta}^*(b)(\hat{\theta}^*(b)) \right\}$$

In closing, we have proposed two complete data criteria that can be utilized in missing data settings. The criteria can be implemented under a variety of different discrepancies.

CHAPTER 8
APPLICATION

In order to evaluate the performance of the proposed methodology in a real world application, we obtained a data set from the Visual Function and SIREN Laboratories at the University of Iowa Department of Neurology. Uc, Rizzo, Anderson, Dastrup, Sparks and Dawson (2009) compared the driving performance of individuals with Parkinson's disease (PD) under low contrast (fog) conditions in a driving simulator to that of healthy elderly controls. From this study, we considered a data set of 61 PD cases.

The goal of our analysis was to develop a model for predicting the driving performance of individuals with Parkinson's disease. Often times, it is difficult for family members and medical professionals to determine when an individual should stop driving or limit driving (e.g., only drive in daylight) due to the effects of Parkinson's disease. If a battery of laboratory tests could effectively predict driving performance in a simulator, the results could possibly be extended to real world driving applications to ensure the driving safety of individuals with Parkinson's disease and others who share the road.

The outcome for the data set was the standard deviation of lateral position (SDLP) during a 60 second straight segment in foggy conditions in the driving simulator. The outcome vector can be represented as $Y = (y_1, y_2, \ldots, y_{61})'$. SDLP measures an individual's ability to maintain vehicular control, where lower values of SDLP suggest more vehicular control. In constructing a model, we consider the following four covariates that measure general mobility and cognitive ability.

- Reach: the distance (in inches) one can lean forward and reach while standing (higher is better).

- Tap: the average number of right and left hand finger taps per 20 seconds

Table 8.1: Summary statistics for 61 Parkinson's disease cases.

Variable	Mean (s.d.)	Median	Minimum	Maximum
SDLP	0.36 (0.16)	0.32	0.12	0.82
Tap	35.3 (7.41)	34.0	19.0	52.5
TMT-D	87.7 (74.6)	59.3	20.2	323.3
Reach	10.9 (3.60)	10.8	1.50	20.0
Walk	14.7 (4.70)	14.1	4.86	30.4

(higher is better).

- TMT-D: the difference in times (in seconds) between two connect-the-dots type tests administered by paper. The Trails Making Test A (TMT-A) requires an individual to draw lines to connect numbers in sequential order (e.g. 1, 2, 3, etc.), while the Trails Making Test B (TMT-B) requires an individual to draw lines to connect numbers and letters that alternate in a sequential order (e.g. 1, A, 2, B, 3, etc.). The value for TMT-D is calculated as TMT-D = TMT-B − TMT-A (lower is better).

- Walk: the time (in seconds) required to walk 7 meters at a safe and comfortable pace (lower is better).

The summary statistics for the data set are included in table (8.1).

The data set of the 61 cases used to generate table (8.1) contains no missing values. In order to illustrate our proposed methodology, a missing data setting must be introduced. We deleted a total of 16 data points (7 from Reach and 9 from Walk) from 16 different cases, to create a missing data mechanism that was missing at random (MAR). Starting with a data set that contains no missing values, and then randomly deleting some data points, is identical to the process that was used

for the simulations in chapters 4, 5, and 6, and also follows the framework of other simulation studies (e.g. Cavanaugh and Shumway, 1998; Claeskens and Consentino, 2008).

Generally, in practical settings, the data set containing the 61 cases with no missing data would not be available. We investigate this setting since the model selected using this data set could serve as the "gold standard" for subsequent models chosen under missingness. Hypothetically, the desired remedy for missingness would be to simply recover and use the missing data values. Thus, any model selected based on the incomplete data set would ideally include a structure that was identical to, or similar to, the gold standard model.

As mentioned throughout this dissertation, a common approach to selecting and fitting a model based on a data set containing missing values is to simply use a complete case analysis. In this situation, a complete case analysis would be based on the 45 fully observed cases. In section (4.3), we proposed model selection criteria based on the complete data (45 fully observed cases and 16 partially observed cases) that can be used in a normal linear models framework with missing data in the covariates. We consider model selection under the three aforementioned settings: using a complete case analysis with the fully observed data, using our proposed methodology based on the complete data, and using standard methods on the sample with no missing data.

Linear regression models were fit to all 15 possible covariate subsets under the three settings, and AIC, AICc, EIC, and AICb, were calculated for each model. Table (8.2) contains the AICc values for the 15 candidate models. The criterion values of AIC, EIC, and AICb are not included in table (8.2) for simplification of presentation. However, the criteria have similar values and select the exact same model as AICc under each of the three settings. The bolded values in table (8.2) indicate the model with the lowest criterion value under each setting, which would

Table 8.2: AICc values for the 3 settings: Fully Observed Data (45 fully observed cases), Complete Data (45 fully observed and 16 partially observed cases), and No Missing Data (61 cases with no missing data).

Covariates	Fully Observed	Complete Data	No Missing Data
Reach, Tap, Trails, Walk	-143.15	**-176.74**	**-176.23**
Tap, Trails, Walk	-139.82	-173.61	-175.25
Reach, Trails, Walk	-145.73	-175.73	-174.16
Tap, Trails	-137.79	-172.09	-172.09
Reach, Tap, Walk	-145.53	-175.81	-171.83
Trails, Walk	-142.13	-171.13	-171.37
Reach, Tap, Trails	-139.86	-172.69	-171.15
Reach, Trails	-142.38	-171.62	-170.20
Trails	-140.10	-169.82	-169.82
Reach, Tap	-142.20	-172.81	-168.79
Tap, Walk	-138.80	-167.68	-168.71
Tap	-137.45	-168.04	-168.04
Reach, Walk	**-147.91**	-173.72	-167.53
Reach	-144.55	-170.68	-165.93
Walk	-140.35	-160.93	-160.97

generally be selected for inference.

The criterion values from the column labeled "No Missing Data" in table (8.2) indicate that the model with the lowest AICc value using the 61 cases with no missing data includes all four covariates. The fitted candidate model corresponding to the selected model for cases $i = 1, 2, \ldots, 61$, is

$$\hat{y}_i = .7471 - .0097 * \text{Reach}_i - .0052 * \text{Tap}_i + .0007 * \text{Trl-D}_i - .0106 * \text{Walk}_i. \quad (8.1)$$

The column labeled "Fully Observed" in table (8.2) can be used to select a model based on the 45 fully observed cases. The model with the lowest AICc value includes only the covariates Reach and Walk. The fitted candidate model corresponding to the selected model for the fully observed cases $i = 1, 2, \ldots, 45$, is

$$\hat{y}_i = .6166 - .0160 * \text{Reach}_i - .0085 * \text{Walk}_i. \tag{8.2}$$

The criterion values from the column labeled "Complete Data" in table (8.2) indicate that the selected model would include all 4 covariates if the complete data analogue of AICc was used for model selection. Using the parameter estimates from the EM algorithm, the fitted candidate model corresponding to the selected model for cases $i = 1, 2, \ldots, 61$, is

$$\hat{y}_i = .7511 - .0119 * \text{Reach}_i - .0048 * \text{Tap}_i + .0003 * \text{Trl-D}_i - .0089 * \text{Walk}_i. \tag{8.3}$$

The fitted model selected using the 61 cases with no missing data (8.1), and the fitted model selected using the 45 fully observed cases (8.2), differ by the inclusion (deletion) of the covariates TMT-D and Tap. (We emphasize that although the regression parameter estimates for TMT-D and Tap are small, the covariate values are relatively large as indicated in table (8.1).) This is a rather substantial difference in selected models. The model selected using the fully observed data is also the third to last model that would be selected if there was no missing data. Furthermore, although examining univariable results in a multivariable framework can be dangerous, the univariable models that include Reach or Walk are the two worst models if model selection was based on the data set with no missing data. The drawbacks and limitations of using only the fully observed data in selecting a model are illustrated with this data set. One could argue that under the fully observed data setting, less data is being used, and subsequently the model selection criteria are protecting against overfitting. However, the data from the 16 partially

observed cases could certainly be used in selecting a better model under a more desirable paradigm. The fitted model selected using the fully observed data is underfitted, which is consistent with what would be expected based on the results of the simulation studies in Chapters 4, 5, and 6.

The model selected using the complete data (8.2), and the model selected using the 61 cases with no missing data (8.1), both include all four covariates. The fitted candidate models (8.1) and (8.3) can also be examined in order to evaluate the effectiveness of using the parameter estimates from the EM algorithm once a candidate model is selected. The parameter estimates are quite similar for both models, which is what one would desire. The proposed complete data criteria outperform the fully observed data criteria in selecting a model, if the model selected based on the original sample with no missing data is assumed to be the best model. This application illustrates that the complete data criteria effectively determine a desirable structural form of a model when a real data set is used.

REFERENCES

Agresti, A. (2002), *Categorical Data Analysis* (Wiley, New Jersey).

Akaike, H. (1970), Statistical predictor identification, *Annals of the Institute of Statistical Mathematics* **22**, 203–217.

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.

Azari, R., Li, L. and Tsai, C. L. (2006), Longitudinal data model selection, *Computational Statistics and Data Analysis* **50**, 3053–3066.

Beaton, A. E. (1964), The use of special matrix operations in statistical calculus, Educational Testing Service Research Bulletin, RB–64–51.

Bedrick, E. J. and Tsai, C. L. (1994), Model selection for multivariate regression in small samples, *Biometrics* **50**, 226–231.

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.

Bueso, M. C., Qian, G. and Angulo, J. M. (1999), Stochastic complexity and model selection from incomplete data, *Journal of Statistical Planning and Inference* **76**, 273–284.

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

Cavanaugh, J. E. (1997), Unifying the derivations of the Akaike and corrected Akaike information criteria, *Statistics & Probability Letters* **33**, 201–208.

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.

Cavanaugh, J. E. and Shumway, R. H. (1998), An Akaike information criterion for model selection in the presence of incomplete data, *Journal of Statistical Planning and Inference* **67**, 45–65.

Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006), Deviance information criteria for missing data models, *Bayesian Analysis* **1**, 651–674.

Claeskens, G. and Consentino, F. (2008), Variable selection with incomplete covariate data, *Biometrics* **64**, 1062–1069.

Dempster, A. P. (1969), *Elements of Continuous Multivariate Analysis* (Addison-Wesley, Massachusetts).

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Efron, B. (1979), Bootstrap methods: another look at the jackknife, *American Statistician* **7**, 1–26.

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.

Efron, B. and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap* (Chapman and Hall, New York).

Gelfand, A. E. and Smith, A. F. M. (1990), Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.

Geman, D. and Geman, S. (1984), Stochastic relaxation, Gibbs distributions, and the Bayesian reconstruction of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Hens, N., Aerts, M. and Molenberghs, G. (2006), Model selection for incomplete and design-based samples, *Statistics in Medicine* **25**, 2502–2520.

Heymans, M. W., Buuren, S., Knol, D. L., van Mechelen, W. and de Vet, H. C. W. (2007), Variable selection under multiple imputation using the bootstrap in a prognostic study, *BMC Medical Research Methodology* **7**.

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.

Hurvich, C. M. and Tsai, C. L. (1991), Bias of the corrected AIC criterion for underfitted regression and time series models, *Biometrika* **78**, 499–509.

Hurvich, C. M. and Tsai, C. L. (1993), A corrected Akaike information criterion for vector autoregressive model selection, *Journal of Time Series Analysis* **14**,

271–279.

Hurvich, C. M. and Tsai, C. L. (1995), Model selection for extended quasi-likelihood models in small samples, *Biometrics* **51**, 1077–1084.

Ibrahim, J. G. (1990), Incomplete data in generalized linear models, *Journal of the American Statistical Association* **85**, 765–769.

Ibrahim, J. G., Zhu, H. and Tang, N. (2008), Model selection criteria for missing data problems via the EM algorithm, *Journal of the American Statistical Association* **103**, 1648–1658.

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.

Konishi, S. and Kitagawa, G. (1996), Generalized information criteria in model selection, *Biometrika* **83**, 875–890.

Konishi, S. and Kitagawa, G. (2008), *Information Criteria and Statistical Modeling* (Springer, New York).

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

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

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data* (Wiley, New Jersey).

Mallows, C. L. (1973), Some comments on C_p, *Technometrics* **15**, 661–675.

McKnight, P. E., McKnight, K. M., Sidani, S. and Figueredo, A. J. (2007), *Missing Data: A Gentle Introduction* (The Guilford Press, New York).

Neath, A. A. and Cavanaugh, J. E. (2000), A regression model selection criterion based on bootstrap bumping for use with resistant fitting, *Computational Statistics and Data Analysis* **35**, 155–169.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. and Solenberger, P. (2001), A multivariate technique for multiply imputing missing values using a sequence of regression models, *Survey Methodology* **27**, 85–95.

Roy, J. and Lin, X. (2002), Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: changes in Methadone treatment practices, *Journal of the American Statistical Association* **97**, 40–52.

Roy, J. and Lin, X. (2005), Missing covariates in longitudinal data with informative dropouts: bias analysis and variance, *Biometrics* **61**, 837–846.

Rubin, D. B. (1976), Inference and missing data, *Biometrika* **63**, 581–592.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys* (Wiley, New York).

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data* (Chapman and Hall, New York).

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

Seghouane, A., Bekara, M. and Fleury, G. (2005), A criterion for model selection in the presence of incomplete data based on Kullback's symmetric divergence, *Signal Processing* **85**, 1405–1417.

Shang, J. and Cavanaugh, J. E. (2008), Bootstrap variants of the Akaike information criterion for mixed model selection, *Computational Statistics and Data Analysis* **52**, 2004–2021.

Shibata, R. (1989), Statistical aspects of model selection, in: J. C. Willemsa, ed., *From Data to Model*, Springer-Verlag, London, 215–240.

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

Shimodaira, H. (1994), A new criterion for selecting models from partially observed data, in: P. Cheeseman and R. W. Oldford, eds., *Selecting Models from Data: Artificial Intelligence and Statistics IV, Lecture Notes in Statistics* **89**, Springer-Verlag, New York, 21–29.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002), Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society, Series B* **64**, 1–34.

Stubbendick, A. L. and Ibrahim, J. G. (2003), Maximum likelihood methods for nonignorable missing responses and covariates in random effects models, *Biometrics* **59**, 1140–1150.

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), Distributions of information statistics and criteria for adequacy of models, *Mathematical Science* **153**, 12–18.

Uc, E. Y., Rizzo, M., Anderson, S. W., Dastrup, E., Sparks, J. and Dawson, J. D. (2009), Driving under low contrast visibility conditions in Parkinson's disease, *Neurology* **6**, 1103–1110.

Verbeke, G. and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal*

Data (Springer-Verlag, New York).

Wood, A. M., White, I. R. and Royston, P. (2008), How should variable selection be performed with multiply imputed data? *Statistics in Medicine* **27**, 3227–3246.

Yang, X., Belin, T. R. and Boscardin W. J. (2005), Imputation and Variable Selection in Linear Regression Models with Missing Covariates, *Biometrics* **61**, 498–506.