MODEL SELECTION METHODS IN THE LINEAR MIXED MODEL FOR LONGITUDINAL DATA

by
Anita A. Abraham

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics, School of Public Health.

Chapel Hill
2008

Approved by:

Advisor: Lloyd J. Edwards

Reader: Peggye Dilworth-Anderson

Reader: John S. Preisser

Reader: Pranab K. Sen

Reader: Paul W. Stewart

ABSTRACT

ANITA A. ABRAHAM: Model Selection Methods in the Linear Mixed Model for Longitudinal Data.
(Under the direction of Dr. Lloyd J. Edwards)

The increased use of repeated measures for longitudinal studies has resulted in the necessity for more research in the modeling of this type of data. In this dissertation, we extend three candidate model selection methods from the univariate linear model to the linear mixed model, and investigate their behavior.

Mallows' C_p statistic was developed for the univariate linear model in 1964. Here we propose a C_p statistic for the linear mixed model and show that it can be a promising method for fixed effects selection. Of all the methods investigated in this dissertation, the C_p statistic gave the most favorable results in terms of fixed effects selection and is the least computationally demanding of all the candidate methods.

The KIC statistic, a symmetric divergence information criteria, explored here appears to be promising as a model selection method for both fixed effects and covariance structure. In the selection of the correct covariance structure, the KIC tended to hold middle ground between the AIC and the BIC. In terms of fixed effects, the KIC appears to perform significantly better than either the AIC or BIC in the selection of fixed effects when there is no interaction effect present.

The predicted sum of squares (PRESS) statistic has been developed for the linear mixed model and is available in the SAS statistical software, but its abilities as a model selection method lacked sufficient evaluation. From our study, it appears that the PRESS

statistic does not add much as a fixed effect selection method compared to the C_p or the KIC while being more computationally intensive.

All three criteria are investigated using simulation studies and a large example dataset evaluating health outcomes in the elderly to determine their reliability. As a by-product of this research, the reliability of standard selection criteria in the linear mixed model, namely the AIC and BIC, are also evaluated. Numerous areas of future research within the context of model selection methods in the linear mixed model, are identified.

To my parents, Abraham and Grace

and my brother, Akash

ACKNOWLEDGEMENTS

It is evident after working on this dissertation that not only does "it take a village to raise a child", it takes a village to complete a doctorate. This task would not have been completed without years of support from an extended village of family, friends, professors and co-workers.

First and most importantly, I would like to thank my family. I am blessed with amazing parents that have not only provided me with a boundless amount of love and support but have also taught me, amongst a lifetime of valuable lessons, that there are no limits to my abilities. Beyond that, I am also blessed with a wonderful brother who was my teacher when I was young and now serves as a constant advocate, friend and protector. It is because of their support that I have been able to accomplish all that I have, not the least of which is this dissertation.

Many thanks to my dissertation advisor, Lloyd Edwards. His words of advice and encouragement have been priceless through the years. His enthusiasm and love for learning and research has created the same dedication in me. His sense of humor and jovial personality has aided in making this journey a bit easier. It is because of him that I am confident in my abilities as a statistician and a researcher.

I would like to thank Paul Stewart, who provided me with an assistantship for the past four years which has given me invaluable work experience as a biostatistician, and has been both an encouraging supervisor and a supportive member of my committee. I must also thank the other members of my committee, P.K. Sen, John Preisser and Peggye Dilworth-Anderson for their time and encouragement with regards to my research, and to

C.M. Suchindran for being my academic advisor at the beginning of my career at Carolina and for the financial support he provided through his grant #T32-HD 007237 , "Research Training in Population Statistics," from the National Institute of Child Health and Human Development .

Special thanks to Matt Gurka at the University of Virginia for the use of his simulation code as a starting point for my research, and for the time and feedback he has provided for my work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Objectives

The linear mixed model is an important tool in the modeling of continuous outcome longitudinal data. The linear mixed model extends the univariate linear model with independent and identically distributed (i.i.d.) Gaussian errors in a way that accommodates for the correlation of measurements within the same subject. While many different types of frequentist selection methods have been developed for the univariate linear model, the quantity and quality of frequentist model selection methods that have been developed for the linear mixed model leave much room for improvement. The linear mixed model requires selecting both a mean and a covariance model, and each must be considered separately. Unfortunately, only limited and ambiguous results have been published which evaluate the candidate methods including ones considered standard practice such as the Akaike Information Criterion (AIC, Akaike, 1974) and the Bayesian Information Criterion (BIC, Schwarz, 1978).

In the univariate model, Mallows' C_p criterion requires a pool of candidate models which are each separately nested within a single full model. It compares the mean square error (*MSE*) of each candidate model to the *MSE* of the full model, which then allows comparing one candidate to another. Presently, there is no C_p statistic for the linear mixed model.

Information theoretic criteria have played a prominent role in model selection and is probably the most active area of current research in model selection for the linear mixed model. Most practitioners use the Akaike Information Criterion (AIC, Akaike, 1974) and the Bayesian Information Criterion (BIC, Schwarz, 1978). Cavanaugh (1999) developed a nondirectional divergence criterion, based on Kullback's symmetric divergence, for the linear univariate model (KIC). However, no investigation of the KIC has been done in the linear mixed model. In addition, while the AIC and BIC are accepted as appropriate model selection criteria in the linear mixed model, little research has been done to look at how well they actually perform as selection criteria.

For modeling repeated measures data with correlated errors, Liu et al. (1999) generalized a cross-validation model selection method, the Predicted Residual Sum of Squares (PRESS). Allen (1974) originally suggested PRESS as a model selection criterion in the univariate linear model. While the PRESS statistic has been developed in the linear mixed model, no investigation of its performance has been done in the literature.

1.2 Motivation

As the use of repeated measures data for longitudinal studies has grown, the necessity for more research in the modeling of this type of data has also increased. The linear mixed model serves the same role in longitudinal data analysis as the linear univariate model does in cross-sectional analysis. However, in comparison to standard univariate linear models, fewer frequentist model selection methods have been developed or evaluated for the linear mixed model.

The literature shows a utility in the development of a C_p statistic (Mallows, 1973) for the linear mixed model as well as a need for investigation into the ability of the KIC, developed by Cavanaugh (1999), and the PRESS (Allen, 1974) to select correct model

structures in the linear mixed model. In addition, little evaluation has been done regarding the information criteria that are considered to be standard model selection tools in the linear mixed model, namely the AIC and BIC. In addition to the primary consideration of the C_p, the KIC and the PRESS, this thesis will also serve as an evaluation of these standard tools. The C_p and PRESS statistics will be used to select the correct mean structure using a predictive approach, while the KIC will be evaluated in its selection of both the correct mean and covariance structure (though not simultaneously).

1.3 Literature Review

In this section, a brief overview of the available frequentist model selection techniques in linear and non-linear studies involving the criteria to be studied and their analogs are summarized. The first section presents and defines the univariate linear model and reviews the studies available regarding the analogs of the criteria that will be presented in this thesis. The second section defines the linear mixed model and presents the notation that will be used throughout and contains a review of the literature available regarding the criteria that are the primary focus of this thesis as to their use in the linear mixed model and in other non-linear modeling techniques.

1.3.1 The Univariate Linear Model

Since the model selection methods described in this dissertation were first developed for the univariate linear model, it is important to familiarize ourselves with this model. The univariate linear model, also known as the general linear model for a single response, can be represented as follows for n individuals and p regression parameters (Muller and Stewart, 2006, p. 40-41):

$$\underset{(n \times 1)}{\boldsymbol{y}} = \underset{(n \times p)}{\boldsymbol{X}} \underset{(p \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{e}} \tag{1.1}$$

$$\boldsymbol{e} \sim N_n\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_n\right)$$

This model assumes that the elements of response vector $\boldsymbol{y} = \{y_i\}$ are mutually independent. $\boldsymbol{X}$ is a known constant design matrix of covariate values, and $\boldsymbol{\beta}$ is an unknown vector of regression parameters. The $n \times 1$ constant vector $\boldsymbol{X\beta}$ describes the mean of the responses and the $n \times 1$ random vector $\boldsymbol{e}$ describes the variance of the responses.

1.3.2 Model Selection: General Overview

Model selection provides only one step of many in assessing model adequacy. Other steps include, but are not limited to: reducing the largest model using hypothesis tests and scientific questions of interest, assessing the fit of the model using appropriate diagnostic tools, and assessing the strength of association between the response variable and the predictors using appropriate statistics. While it is important to remember that model selection criteria are not the sole method of deciding what variables should be in a model, it is still a necessary step in selecting the appropriate model for the data available. In the linear mixed model, there are only a limited number of methods that have been developed thus far in comparison to the linear univariate model, and since the structure of the linear mixed model is more complicated than the univariate linear model, it is essential that more methods be developed so that there are better possibilities to choose the correct model.

In general, a selection criterion scores every fitted model in a candidate class by how effectively the model conforms to the data based on its size. Ideally, unwanted scores will be assigned not only to models that omit essential variables, but also to models that adequately accommodate the data yet involve extraneous or irrelevant

variables (Cavanaugh, 2004). In other words, the ideal selection criterion will select the model with the most parsimonious set of variables to describe the data available.

A different method of model selection is cross-validation. This method changes the goal of model selection from explaining a given set of data to predicting a new set of data which comes from the same background as the given set. This is the method targeted by the C_p and PRESS statistics.

1.3.2.1 Mallows' C_p

The following conceptual predictive (C_p) criterion for use in the univariate linear model was first proposed by C. L. Mallows (1964) and first published by Gorman and Toman (1966):

$$C_p = \frac{SSE(p)}{\widehat{\sigma}^2} - [N - 2(p+1)] \tag{1.2}$$

where $\widehat{\sigma}^2$ is an estimate of the variance of the data. In linear regression, the formula used most often is:

$$C_p = \frac{SSE(p)}{MSE(p_{max})} - [n - 2(p+1)] \tag{1.3}$$

where n is the number of observations, p is the number of parameters in the candidate model, p_{max} is the number of parameters in the saturated (maximum) model, $MSE(p_{max})$ represents the mean square error for the saturated (maximum) model, and $SSE(p)$ represents the sums of square error for the candidate model. This criterion requires a pool of candidate models which are each separately nested within a single full model. It is an estimate of the measure of adequacy for prediction given by the scaled sum of errors (Ronchetti and Staudte, 1994).

Mallows suggested that a value of C_p too large or too far above $p + 1$ indicates an inaccurate model. In 1976, Hocking suggested requiring $C_p \leq p + 1$ in choosing a model

for prediction, and requiring $C_p \leq 2(p+1) - p_{max}$ in choosing a model for parameter estimation.

The F_p statistic and C_p are related in that they both are tests of comparisons within nested models. The criterion F_p when comparing to the saturated model is defined as:

$$F_p = \frac{[SSE(p) - SSE(p_{max})]/(p_{max} - p)}{SSE(p_{max})/(n - p_{max} - 1)} \tag{1.4}$$

The F_p statistic compares to an F distribution with $p_{max} - p$ and $N - p_{max} - 1$ numerator and denominator degrees of freedom, respectively. Since the F statistic corresponds to a test of two models, then F_p corresponds to a test, using the saturated model, of whether the $p_{max} - p$ regression coefficients not in the candidate model are simultaneously zero. If this criterion is significant, the saturated model includes variables that significantly improve upon the predictive ability of the model when compared to the model with p variables. We can express the C_p statistic as a simple function of the F_p statistic as follows:

$$C_p = (p_{max} - p)F_p + (2p - p_{max} + 1) \tag{1.5}$$

Mallows (1973) expounded on the statistic saying that he "feels that the greatest value of the device is that it helps the statistician to examine some aspects of the structure of his data and helps him to recognize the ambiguities that confront him." He further stated that "this device cannot be expected to provide a single 'best' equation when the data are intrinsically inadequate to support such a strong inference." Mallows found that the ambiguous cases where the "minimum C_p" rule will give bad results are where there are a large number of subsets that have C_p's that are close to each other.

Ronchetti and Staudte (1994) noticed that since the C_p is based on least squares estimation, it is very sensitive to outliers and other departures from the normality assumption of the error distribution. To correct this issue, they proposed a robust

criterion for prediction, which they called RC_p, which can be used to choose the best models that fit the majority of the data by taking into account the presence of outliers and possible departures from the normality assumption. The robust version of C_p is defined as follows:

$$RC_p = \frac{W_p}{\widehat{\sigma}^2} - (U_p - V_p) \tag{1.6}$$

where W_p is a weighted residual sum of squares that is centered and rescaled based on the data and where U_p and V_p are variance terms that are defined as:

$$V_p = \sum_i var(\widehat{w}_i \delta_i) \tag{1.7}$$

$$U_p = \sum_i var(\widehat{w}_i r_i) \tag{1.8}$$

where $\delta_i = \widehat{y}_i - \mathrm{E}(y_i)$ (the prediction error) and $r_i = y_i - \widehat{y}_i$ (the residual). Plotting RC_p vs. V_p is the analog of the C_p vs. p plot. Note that when the weights are identically 1, W_p becomes the residual sum of squares of a least squares fit, $V_p = p$, $U_p = n - p$ and RC_p reduces to Mallows' C_p. Through the use of simulated data and some classic data examples, Ronchetti and Staudte (1994) were able to show that their RC_p is more efficient than the classical C_p in selecting the correct model when outliers are present.

Gilmour (1996) looked at the problem that occurs when there are a large number of models with $C_p < p$. This can be observed in most data sets that have a reasonably large number of unimportant regressors. By looking at the expectation of C_p, he found that if there are a relatively large number of regressors, the model with the lowest C_p will tend to overfit, that is, to suggest the inclusion of at least one unimportant regressor. To correct this problem, Gilmour (1996) suggested the following adjustment:

$$\overline{C}_p = C_p - \frac{2(p_{max} - p + 1)}{n - p_{max} - 3} \tag{1.9}$$

In addition to comparing this value to p, Gilmour (1996) suggested that the plots of $\overline{C}_p$ vs. p be interpreted more conservatively. He believed an interactive approach is needed where hypothesis tests are looked at to investigate the importance of specific variables in the models that are chosen with the smallest $\overline{C}_p$. Gilmour (1996) also commented on the RC_p statistic devised by Ronchetti and Staudte (1994), stating that since its numerator is a weighted residual sum of squares, where the weights are calculated from the data, the distributional results cannot be worked out, therefore its expectation cannot be found in a generalized form.

1.3.2.2 Information Criteria

Information theoretic criteria have played a prominent role in model selection for the linear mixed model. Information theoretic criteria are defined as an estimate of the measure of fit of a model to the data. The most common criteria used in mixed models are the Akaike Information Criterion (AIC, Akaike, 1974) and the Bayesian Information Criterion (BIC, Schwarz, 1978). The AIC is the directed divergence between the true model and candidate model with respect to the true model. The BIC, on the other hand, technically is not a divergence criterion since it does not assume a true model exists. However, it is generally used as an approximation to a measure of directed divergence. The AIC assumes models are nested whereas the BIC does not make the assumption. Under the same conditions as the AIC, Cavanaugh (1999) defined the nondirectional divergence criterion, the KIC.

1.3.2.2.1 Directed Divergences: AIC and BIC

Kullback and Leibler (1951) defined Kullback's directed divergence as a measure of the disparity between the true model $f(Y|\theta_0)$ and an approximating model $f(Y|\theta_k)$. For two arbitrary parametric densities $f(Y|\theta)$ and $f(Y|\theta^*)$, Kullback's directed divergence between $f(Y|\theta)$ and $f(Y|\theta^*)$ is defined as:

$$I(\theta, \theta^*) = E_\theta\left\{\log\frac{f(Y|\theta)}{f(Y|\theta^*)}\right\} \tag{1.10}$$

where E_θ denotes the expectation under $f(Y|\theta)$ (Kullback and Leibler, 1951). Therefore, $I(\theta_0, \theta_k) = E_{\theta_0}\left\{\log\frac{f(Y|\theta_0)}{f(Y|\theta_k)}\right\}$ defines the directed divergence between $f(Y|\theta_0)$ and $f(Y|\theta_k)$ with respect to $f(Y|\theta_0)$.

For $f(Y|\theta)$ and $f(Y|\theta^*)$, define

$$d(\theta, \theta^*) = E_\theta\{-2\log f(Y|\theta^*)\} \tag{1.11}$$

Therefore,

$$2I(\theta_0, \theta_k) = d(\theta_0, \theta_k) - d(\theta_0, \theta_0) \tag{1.12}$$

For the purpose of discriminating between various candidate models, we can now say,

$$d(\theta_0, \theta_k) = E_{\theta_0}\{-2\log f(Y|\theta_k)\} \tag{1.13}$$

serves as a valid substitute for $I(\theta_0, \theta_k)$.

This suggests that

$$d(\theta_0, \hat{\theta}_k) = E_{\theta_0}\{-2\log f(Y|\theta_k)\}|_{\theta_k=\widehat{\theta}_k} \tag{1.14}$$

would provide a suitable measure of the separation between the generating model $f(Y|\theta_0)$ and a fitted candidate model $f(Y|\widehat{\theta}_k)$.

Akaike (1974) suggested that $-2\log f(Y|\theta_k)$ serves as an unbiased estimator of $d(\theta_0, \hat{\theta}_k)$ and that the bias adjustment

$$E_{\theta_0}\left\{d(\theta_0, \hat{\theta}_k)\right\} - E_{\theta_0}\left\{-2\log f(Y|\widehat{\theta}_k)\right\} \tag{1.15}$$

can often be asymptotically estimated by twice the dimension of $\hat{\theta}_k$. Thus, since k denotes the dimension of $\hat{\theta}_k$, under appropriate conditions, the expected value of

$$AIC = -2\log f(Y|\theta_k) + 2k \tag{1.16}$$

should asymptotically approach the expected value of

$$d(\theta_0, \hat{\theta}_k) = E_{\theta_0}\{-2\log f(Y|\theta_k)\}|_{\theta_k=\widehat{\theta}_k} \tag{1.17}$$

and is therefore asymptotically unbiased.

Schwarz (1978) presented a Bayesian alternative to the AIC. In a model of given dimension, ML estimators can be obtained as large sample limits of the Bayes estimators for arbitrary nowhere vanishing a priori distributions. Therefore, by studying the asymptotic behavior of Bayes estimators under a special class of priors, Schwarz arrived at the procedure where you choose the model for which $\log f(Y|\widehat{\theta}_k) - \frac{1}{2}k\log n$ is largest. Thus

$$BIC = \log f(Y|\widehat{\theta}_k) - \frac{1}{2}k\log n \tag{1.18}$$

The BIC, though technically not a directed divergence, is generally used as an approximation to the Kullback-Leibler directed divergence.

Since the BIC differs from the AIC only in that the dimension is multiplied by $\frac{1}{2}\log n$, the BIC leans more than Akaike's toward lower-dimensional models. Both the AIC and BIC are essentially log-likelihood values with a penalty (or adjustment) for the number of parameters estimated.

1.3.2.2.2 Symmetric Divergence: KIC

Kullback's symmetric divergence is defined as:

$$J(\theta_0, \theta_k) = I(\theta_0, \theta_k) + I(\theta_k, \theta_0) = E_{\theta_0}\left\{\log\frac{f(Y|\theta_0)}{f(Y|\theta_k)}\right\} + E_{\theta_k}\left\{\log\frac{f(Y|\theta_k)}{f(Y|\theta_0)}\right\} \tag{1.19}$$

This divergence is symmetric in that $J(\theta_0, \theta_k) = J(\theta_k, \theta_0)$. This symmetric divergence measures the average combined measure of fit of a sample Y generated under the true model $f(Y|\theta_0)$ and a sample Z generated under the true model $f(Z|\theta_k)$. Using

arguments similar to Akaike (1974) and assuming nested models, Cavanaugh (1999) proposed the following large sample model selection criterion:

$$KIC = -2\log f(Y|\widehat{\theta}_k) + 3k \tag{1.20}$$

This criterion serves as an asymptotically unbiased estimator of a variant of the symmetric divergence between the true model and a fitted approximating model.

1.3.2.2.3 Information Criteria Performance in Various Model Types

The articles discussed in this section look at the performance of many types of information criteria in a variety of model settings. Cavanaugh (1999) looked at the performance of different criteria in an autoregressive model setting, Cavanaugh (2004) used a univariate linear model setting, Kim and Cavanaugh (2005) used a nonlinear model setting, and Hafidi and Mkhadri (2006) used a multivariate linear model setting.

Cavanaugh (1999) looked at the performance of the KIC in comparison to a wide spectrum of information criteria, namely, the AIC, the corrected AIC (AIC_c) (Suguira, 1978), the Final Prediction Error (FPE) (Akaike, 1969), the HQ (Hannan and Quinn, 1979), the ABIC (Akaike, 1978), and the BIC (Schwarz, 1978), in a univariate autoregressive (AR) process of order p, in a setting in which the criteria are used to select p. This univariate AR process of order p is defined as follows:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \text{iid } N\left(0, \sigma^2\right).$$

where $1 \leq p \leq P$ in P candidate families with a set of observations $Y = \{y_1, y_2, \ldots, y_n\}$ from such a process. Also, note that $\widehat{\theta}_k$=$(\sigma^2, \phi_1, \phi_2, \ldots, \phi_p)^{'}$, $k = p + 1$ and $t = p + 1, ..., n$.

The AIC is defined above (1.16), the other criteria used in this study are defined, within this setting, as:

$$AIC_c = \left(n \log \widehat{\sigma}^2 + n\right) + \frac{2n(p+1)}{n-p-2}$$

$$FPE = n\left(\frac{n+p}{n-p}\right)\widehat{\sigma}^2$$

$$HQ = n \log \widehat{\sigma}^2 + 2p \log \log n$$

$$ABIC = (n-p)\log\left(\frac{n\widehat{\sigma}^2}{n-p}\right) + p\log\left\{\frac{\left(\sum_{t=1}^{n} y_t^2\right) - n\widehat{\sigma}^2}{p}\right\}$$

$$BIC = -2\log f\left(Y|\widehat{\theta}_k\right) + k \log n$$

We note that Cavanaugh (1999) uses two different Bayesian information criteria. The BIC, that is defined in (1.18), is multiplied by a factor of -2, and the $ABIC$, as defined above, is from Akaike's 1978 paper.

Cavanaugh found that, for autoregressive modeling, the BIC, HQ and ABIC are consistent whereas AIC, AIC_c and FPE are asymptotically efficient. The KIC is also asymptotically efficient for a broad class of generating models. However, the AIC, AIC_c, and FPE are asymptotically efficient in an even broader class. For Gaussian white noise processes of mean 0 and variance 1 with sample size of 40 or 60 and maximum model of order $p = 8$, he found that the KIC obtains substantially more correct order selections than any of the asymptotically efficient criteria, consistently outperforms HQ in terms of correct order selection, and outperforms ABIC in half of the simulations. KIC is generally outperformed by BIC. However, KIC does not exhibit as strong a tendency as BIC to choose underparameterized models.

In conclusion, Cavanaugh found that the results from the simulation study suggested that KIC should function as an effective model selection criteria in large-sample applications. The results also suggested that the symmetric divergence may

provide a foundation for the development of model selection criteria which is preferable to that provided by the directed divergence.

Cavanaugh (2004) derived the KIC_c and the MKIC as analogs of the AIC_c and the modified AIC, (MAIC) (Fujikoshi and Satoh, 1997) using Kullback's symmetric divergence (as opposed to the directed divergence used to develop the AIC family of criteria). Cavanaugh describes the motivation of this method by saying that "the directed divergence which serves as the basis for the AIC is more sensitive towards detecting overfitted models, whereas its counterpart is more sensitive towards detecting underfitted models. Since the symmetric divergence reflects the sensitivities of both directed divergences, it functions as a discrepancy measure which is arguably more balanced than either of its individual components." Cavanaugh (2004) also conducts a simulation study (in the linear univariate model) to look at the effectiveness of these new criteria compared to their AIC-based counterparts. In this setting, the AIC and KIC are defined as in equations 1.16 and 1.20 respectively. Since this study used the linear univariate model, the parameter p here refers to the number of predictors in the model and $k = p + 1$.

The other criteria used in this study are defined as:

$$AIC_c = -2\log f\left(Y|\widehat{\theta}_k\right) + \frac{2nk}{n-k-1}$$

$$MAIC = -2\log f\left(Y|\widehat{\theta}_k\right) + \frac{2nk}{n-k-1} + 2p\left(\frac{(n-p)\widehat{\sigma}_P^2}{(n-P)\widehat{\sigma}_p^2} - 1\right) - 2\left(\frac{(n-p)\widehat{\sigma}_P^2}{(n-P)\widehat{\sigma}_p^2} - 1\right)^2$$

$$KIC_c = -2\log f\left(Y|\widehat{\theta}_k\right) + n\log\left(\frac{n}{n-p}\right) + \frac{n((n-p)(2p+3)-2)}{(n-k-1)(n-p)}$$

$$MKIC = 2\left(\frac{(n-P-2)\widehat{\sigma}_p^2}{\widehat{\sigma}_P^2} + p - (n-2)\right) + \frac{2nk}{n-k-1}$$

The modified criteria (MAIC and MKIC) are based on the assumption that the true model is a member of the largest family in the candidate class. In this scenario, $\mathcal{F}(P)$ denotes this largest family, and its associated maximum likelihood estimates are defined as $\widehat{\theta}_{\mathrm{P}} = \left(\widehat{\sigma}_P^2, \widehat{\beta}_P\right)$.

In Cavanaugh's (2004) simulation study he used two different frameworks: a nested model (NM) framework and an "all possible regressions" (APR) framework. In the NM framework, it is assumed that the candidate models are nested (i.e. each successive design matrix contains all of the regressors of its predecessors). In the APR framework, it is assumed that the candidate models correspond to all possible subsets of the regressor variables.

In the NM framework, Cavanaugh (2004) found that the MKIC outperformed the MAIC for smaller sample sizes. However, as sample size increased, MAIC overtook MKIC. For the 'corrected' criteria, AIC_c initially outperformed KIC_c, but was overtaken as sample size increased. With the original criteria, KIC outperforms AIC over all sets. In the APR framework, MKIC is consistently outperformed by MAIC (though it is marginal). However, with the 'corrected' criteria KIC_c always obtains higher selection rates than AIC_c. And again, KIC markedly outperforms AIC over all sets.

From these results, Cavanaugh concludes that MKIC shows promise as a small-sample selection criteria, whereas KIC_c and KIC show promise as large-sample selection criteria.

Kim and Cavanaugh (2005) looked at modified versions of the AIC (the "corrected" AIC_c and the "improved" AIC_I) and the KIC (the "corrected" KIC_c and the "improved" KIC_I), in the nonlinear regression framework. The AIC_c was originally proposed by Suguira (1978) and was found to be useful even in relatively small samples in linear regression models by Hurvich and Tsai (1989). The AIC_I was originally proposed by Hurvich et al. (1990) for use in autoregressive models. However, Kim and Cavanaugh (2005) modify the AIC_I in this paper to adjust for bias. From this background, Kim and Cavanaugh derived the KIC_I using Kullback's symmetric divergence.

The AIC_I and the KIC_I are defined in this paper as follows:

$$AIC_I = n\big(\log\widehat{\sigma}^2 + 1\big) + \frac{1}{R}\sum_{j=1}^{R}\left[\frac{n}{\widehat{\sigma}^2(j)} + \frac{\left\{h\Big(\widehat{\delta}(j)\Big)' h\Big(\widehat{\delta}(j)\Big)\right\}}{\widehat{\sigma}^2(j)} - n\right]$$

$$KIC_I = n\big(\log\widehat{\sigma}^2 + 1\big) + \frac{1}{R}\sum_{j=1}\Bigg[- n\log\widehat{\sigma}^2(j) + \frac{n}{\widehat{\sigma}^2(j)} + n\widehat{\sigma}^2(j) + \left\{\frac{1}{\widehat{\sigma}^2(j)} + 1\right\}\left\{h\Big(\widehat{\delta}(j)\Big)' h\Big(\widehat{\delta}(j)\Big)\right\} - 2n\Bigg]$$

where $h\Big(\widehat{\delta}(j)\Big)$ is the mean vector for the candidate model and R is the number of samples generated.

From the simulation studies, Cavanaugh and Kim (2005) found that generally the "improved" criteria outperformed the "corrected" criteria, which in turn outperformed the non-adjusted criteria. They also found that the KIC family performed favorably against the AIC family.

Hafidi and Mkhadri (2003) derived a different version of the "corrected" KIC (KIC_c) and compared it to the AIC_c derived by Hurvich and Tsai (1989) (which was used in Kim and Cavanaugh (2005) and Cavanaugh (2004) as well). Hafidi and Mkhadri (2003) studied the behavior of their criteria in three different settings: multiple regression (i.e. univariate linear regression with multiple regression parameters), multivariate regression, and univariate autoregressive models. Multiple regression was described in section 1.3.1 of this paper, and the univariate autoregressive models used here are identical to those of the Cavanaugh (1999) paper. The multivariate regression model used here is described as:

$$\underset{(n\times q)}{\boldsymbol{Y}} = \underset{(n\times p)}{\boldsymbol{X}}\underset{(p\times q)}{\boldsymbol{\beta}} + \underset{(n\times q)}{\boldsymbol{E}}$$

$$\boldsymbol{E} \sim N_q(\mathbf{0}, \Sigma)$$

with the rows of $\boldsymbol{Y}(n \times q)$ correspond to q response variables on each of n individuals, $\boldsymbol{\beta}$ $(p \times q)$ is a matrix of unknown regression parameters and $\boldsymbol{X}(n \times p)$ is a constant design matrix of covariate values.

The derivation of the KIC$_c$ in this paper used the methodology of McQuarrie and Tsai (1998, p. 131-132) and resulted in the following definition of the KIC$_c$ (again, $k = p + 1$) :

$$KIC_c = KIC + \frac{2k(k+1)}{n-k-1} \tag{1.21}$$

Using simulations with various types of data structures, Hafidi and Mkhadri found that among the efficient criteria studied (i.e., AIC, KIC, AIC$_c$ and KIC$_c$, HQ and BIC), the KIC performed the best in multiple and multivariate linear regression. For univariate autoregression, the KIC$_c$ was slightly outperformed by the consistent criterion BIC. The signal-to-noise study found that when the sample size is small, KIC$_c$ has a greater signal-to-noise ratio, and its probability of overfitting is almost zero. In contrast, KIC has a tendency to overfit, because it has a low signal-to-noise ratio, which leads to a higher probability of overfitting.

1.3.2.3 PRESS Statistic

The predicted residual sum of squares (PRESS) statistic was proposed by Allen (1974). The PRESS criterion is obtained by deleting the ith case from a data set, estimating the regression function for the subset model from the remaining $n - 1$ cases, and then using the fitted regression function to obtain the predicted value $\widehat{y}_{(i)}$ (Neter et al., 1996). The PRESS residuals are defined as:

$$e_{(i)} = y_i - \widehat{y}_{(i)} \tag{1.22}$$

This process is repeated for all n observations and the PRESS statistic is then computed as:

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 \tag{1.23}$$

As the PRESS residuals are a measure of how well the fitted model is able to predict the response, the smaller the PRESS statistic the better the model is for prediction. In other words, the PRESS statistic selects the model with the smallest mean square error of prediction.

1.3.3 The Linear Mixed Model

1.3.3.1 Notation

The linear mixed model may be thought of as a two-stage hierarchy: an individual stage and a population stage (Davidian, 2001). For independent sampling units (or subjects) $i = 1, ..., m$:

Individual Stage:

$$\boldsymbol{y}_i = \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{e}_i, \qquad \boldsymbol{e}_i \sim \boldsymbol{\mathcal{N}}_{n_i}(\mathbf{0},\, \boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e)) \tag{1.24}$$

This models each subject's measurements as an individual regression line where $\boldsymbol{y}_i$ and $\boldsymbol{e}_i$ are $n_i \times 1$ vectors with the "design matrix" $\boldsymbol{Z}_i$ which is $n_i \times q$ and the regression parameter vector $\boldsymbol{\beta}_i$ which is $q \times 1$.

Population Stage:

$$\boldsymbol{\beta}_i = A_i\boldsymbol{\beta}_0 + \boldsymbol{d}_i, \qquad \boldsymbol{d}_i \sim \boldsymbol{\mathcal{N}}_q(\mathbf{0},\, \boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d)) \tag{1.25}$$

This describes the common features between individuals through a regression line where $\boldsymbol{\beta}_i$ and $\boldsymbol{d_i}$ are $q \times 1$ vectors. The matrix A_i is $q \times p$ and summarizes information like group membership, which allows the mean of $\boldsymbol{\beta}_i$ to be different for different groups, and

$\boldsymbol{\beta}_0$ is a $p \times 1$ vector of unknown, constant population parameters. From this set up, we can see that within-unit variation is described by the covariance matrix $\boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e)$, while between-unit (or among-unit) variation is described by the covariance matrix $\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d)$.

Combining these two stages, we arrive at the standard format for the general linear mixed model, where $\boldsymbol{X}_i = \boldsymbol{Z}_i \boldsymbol{A}_i$ (Laird and Ware, 1982; Muller and Stewart, 2006):

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_0 + \boldsymbol{Z}_i \boldsymbol{d}_i + \boldsymbol{e}_i \tag{1.26}$$

Here, $\boldsymbol{y_i}$ is an $n_i \times 1$ vector of observations on subject i; $\boldsymbol{X}_i$ is an $n_i \times p$ known, constant design matrix for subject i, with full column rank p; $\boldsymbol{\beta}_0$ is a $p \times 1$ vector of unknown, constant, population parameters; $\boldsymbol{Z}_i$ is an $n_i \times q$ known, constant design matrix for subject i corresponding to the random effects $\boldsymbol{d}_i$, with rank q; $\boldsymbol{d}_i$ is a $q \times 1$ vector of unknown, random individual parameters; $\boldsymbol{e}_i$ is an $n_i \times 1$ vector of random errors. Also, $n = \sum_{i=1}^{m} n_i$.

Throughout, $\boldsymbol{d}_i$ is Gaussian with mean $\mathbf{0}$ $(q \times 1)$ and covariance $\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d)$ $(q \times q)$, independently of Gaussian $\boldsymbol{e}_i$ $(n_i \times 1)$ with mean $\mathbf{0}$ $(n_i \times 1)$ and covariance $\boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e)$ $(n_i \times n_i)$, so that

$$\boldsymbol{\mathcal{V}}\left(\begin{bmatrix} \boldsymbol{d}_i \\ \boldsymbol{e}_i \end{bmatrix}\right) = \begin{pmatrix} \boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e) \end{pmatrix} \tag{1.27}$$

Here $\boldsymbol{\mathcal{V}}(\,\cdot\,)$ is the covariance operator, $\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d)$ is a positive-definite, symmetric covariance matrix of the random effects, and $\boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e)$ is an unknown, constant positive-definite matrix of the fixed effects. Under the assumptions, $\boldsymbol{\mathcal{V}}(\boldsymbol{y}_i)$ can be expressed as $\boldsymbol{\Sigma}_i(\boldsymbol{\tau}) = \boldsymbol{Z}_i \boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d) \boldsymbol{Z}_i' + \boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e)$. Generally, it is assumed that the covariance $\boldsymbol{\Sigma_i}$ can be characterized by a finite set of parameters represented by an $r\,x\,1$ vector $\boldsymbol{\tau}$, which consists of the unique parameters in $\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d)$ and $\boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e)$. Throughout, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\tau}')'$ will be the s $x\,1$ vector of parameters for model (1), where $s = p + r$.

1.3.3.2 Estimation and Inference Techniques

We consider two of the estimation techniques employed in the mixed model: maximum likelihood (ML) and restricted-maximum likelihood (REML). It is important to consider both estimation techniques because the choice affects the availability of model selection criteria. The log-likelihood function, $L(\boldsymbol{\beta}, \boldsymbol{\tau}) = log[f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\tau})]$, allows finding ML estimates for model (1.26):

$$-2L_{ML}(\boldsymbol{\beta}, \boldsymbol{\tau}) = nlog(2\pi) + \sum_{i=1}^{m} log|\boldsymbol{\Sigma}_i(\boldsymbol{\tau})| + \sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i(\boldsymbol{\tau})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \qquad (1.28)$$

The corresponding function for REML estimates is given by:

$$-2\mathrm{L}_{REML}(\boldsymbol{\tau}) = (n-p)log(2\pi) + log\left|\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{X}_i\right| + \sum_{i=1}^{m} log|\boldsymbol{\Sigma}_i(\boldsymbol{\tau})|$$

$$+ log\left|\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right| + \sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \qquad (1.29)$$

where $\boldsymbol{y} = (\boldsymbol{y}_1', \boldsymbol{y}_2', \ldots, \boldsymbol{y}_m')'$ is $n\,x\,1$. From here on, we will mostly use $L(\boldsymbol{\theta})$ to denote the log-likelihood for either ML or REML. Since the second term in (1.29), $log\left|\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{X}_i\right|$, is a constant not depending on model parameters, it is typically omitted from the REML likelihood.

1.3.4 Model Selection Methods in the Linear Mixed Model

1.3.4.1 Background

Comparisons in the linear mixed model can be grouped into three types: 1) comparing mean models using the same covariance structure, 2) comparing covariance models using the same mean structure (where the covariance structure is either nested or nonnested), and 3) comparing linear mixed models with different mean and covariance structures. There are model selection techniques in all three of these areas that warrant

further exploration. Amongst these techniques are the KIC, the C_p statistic, and the PRESS statistic.

1.3.4.2 Mallows' C_p

While the C_p statistic has not been developed in the linear mixed model, Cantoni, et al. (2005) did derive the generalized C_p, which they called the GC_p, for use with parametric and nonparametric models. The simulation study performed used a marginal longitudinal model with a logit link, where the linear predictor was $\beta_0 + \boldsymbol{x}_{it}^T\boldsymbol{\beta}$ with $\boldsymbol{x}_{it}^T$ and $\boldsymbol{\beta}$ of dimension 5. In other words, they used a population-averaged model, i.e. a model with no random effects. The formula for the GC_p is given by:

$$\begin{aligned} GC_p = WSR - \sum_{i=1}^{K} E\{w_i^2(\epsilon_i)\epsilon_i^2\} + 2\sum_{i=1}^{K} E[\{w_i(\epsilon_i)w_i'(\epsilon_i)\epsilon_i^2 + w_i^2(\epsilon_i)\epsilon_i\}\delta_i] \\ - \sum_{i=1}^{K} E[\{w_i(\epsilon_i)w_i''(\epsilon_i)\epsilon_i^2 + (w_i'(\epsilon_i))^2\epsilon_i^2 \\ + 4w_i(\epsilon_i)w_i'(\epsilon_i)\epsilon_i\}\delta_i^2] \end{aligned} \tag{1.30}$$

where w_i is a weight function that may be chosen to achieve a number of objectives including heteroscedasticity or robustness, WSR is the weighted sum of residuals given by $\sum_{i=1}^{K} w_i^2(r_i)r_i^2$, $\epsilon_i = \frac{y_i - E(y_i)}{\sigma v_i^{1/2}}$, $\delta_i = \frac{\widehat{y}_i - E(y_i)}{\sigma v_i^{1/2}}$, $r_i = \frac{y_i - \widehat{y}_i}{\sigma v_i^{1/2}}$, and $V(y_i) = \sigma^2 v_i$.

They explored the use of this statistic in GEE modeling of longitudinal data and compared it to variable selection technique based on Wald-type tests and score-type tests. They found that the GC_p performed as well as stepwise selection and much better than the z-test and score-test in identifying good models in the GEE framework.

1.3.4.3 Information Criteria

For the linear mixed model given by (1.26), the AIC for both ML and REML is defined as (Vonesh and Chinchilli 1997, pp 262-263)

$$AIC = L\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\tau}}\right) - s, \tag{1.31}$$

The BIC for ML is given by

$$BIC = L\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\tau}}\right) - 2^{-1} s \log(n), \tag{1.32}$$

and the BIC for REML is given by

$$BIC = L\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\tau}}\right) - 2^{-1} s \log(n - p) \tag{1.33}$$

Both the AIC and BIC involve differences in log-likelihood values and numbers of parameters. For purposes of interpretation, all forms have the unfortunate characteristic of being data scale dependent. However, the formula for the BIC includes the number of observations in the data, which seeks to give a finite sample correction. In both the AIC and BIC the model with the larger value is considered a better fit. Equivalently, AIC and BIC can also be defined by using a multiplicative factor of -2, where then the smaller value is considered a better fit.

The penalty terms used in the AIC and BIC, based on formulas for longitudinal data familiar to readers of Vonesh and Chinchilli (1997), may not always be what is actually implemented in software. Gurka (2006) showed that for some software, but not all, $s = r$, the number of unique covariance parameters only. The unsuspecting user of such software will be unaware that the associated form of BIC is only for covariance model selection under REML estimation. In addition, Gurka (2006) showed that some software replaces the number of observations, n, with the number of independent sampling units, m, under ML and REML. As a result, many users of such software may

have been misled when using AIC or BIC to select either the mean or covariance structures in the linear mixed model.

Gomez (2005) looked at the performance of the Kenward-Roger method in the linear mixed model when the covariance structure is selected using AIC and BIC. He investigated Type I error rates for tests of fixed effects in mixed linear models using Wald F-statistics with the Kenward-Roger adjustment, and he generated data using 15 covariance structures. Correct covariance structures as well as those selected using AIC and BIC were examined. Type I error rates for the correct models were often adequate depending on the sample size and complexity of covariance structure. Type I error rates for the best AIC and BIC models were always higher than target values, but those obtained using BIC were closer to the target value than those obtained using AIC. For unbalanced data , Type I error rates for the between-subjects effects were closer for negative pairing. Success of AIC and BIC in selecting the correct covariance structure was low.

- The success rate (proportion of times AIC or BIC chose the correct covariance structure) depended greatly on the sample size and covariance structure. The highest success rate was 73.91%, for the largest sample size and a simple covariance structure (ARRE). Success rates were higher for larger sample sizes and simpler covariance structures. AIC had a higher success rate than BIC for complicated structures, especially for those with heterogeneity between treatments. BIC had higher success than AIC for simpler structures.
- Type I error rates for Kenward-Roger method hypothesis tests were always higher than the target values for best AIC and best BIC models. The best BIC models usually produced closer Type I error rates to target values than the best AIC models. Tests for within-subject effect generally produce closer Type I error rates to target values.

- Unless sample sizes are large, if AIC and BIC are used, users should be aware that Type I error rates are higher than target values.
- Even if the correct covariance structure is known, Type I error rates are inflated for complex structures and small sample sizes.

1.3.4.4 PRESS Statistic

Liu et al. (1999) generalized the PRESS statistic to multivariate linear models with correlated errors in repeated measures data. Liu et al. (1999) uses the PRESS statistic to select the linear predictor in linear models with correlated errors. They define the PRESS statistic in linear models with correlated errors as:

$$PRESS = \sum_{i=1}^{n} e'_{(i)} e_{(i)} \tag{1.34}$$

where $e_{(i)} = y_i - X_i \widehat{\boldsymbol{\beta}}_{(i)}$ is the deleted residual with $\widehat{\boldsymbol{\beta}}_{(i)}$ defined as the regression parameter estimate when the ith person is deleted from the analysis. This definition of PRESS is applicable to both balanced and unbalanced data where each person has a different number of measurements, and can only be used to select the linear predictor (the covariance structure is treated as a nuisance parameter). This definition is very close to the form of the original PRESS statistic in traditional linear regression.

Liu et al. (1999) proposed a new efficient computing method based on pivoting and then proposed to apply the PRESS model selection method to real data. They compared the top 10 models that were selected by the PRESS to the top 10 models selected by AIC, BIC and likelihood ratio tests using a linear mixed effects model. They found that there were 5 models that were selected by both AIC and PRESS in the top 10, but while the AIC values were very close together for the top 10 models, the PRESS statistic had a small difference between the first two models and a large difference between those and the other 8 models. The BIC and the PRESS statistic had 6 of the

same models in their top 10 and the model chosen as best by the BIC was the same as that chosen by the forward selection likelihood ratio test with Type I error of $\alpha = 0.01$.

1.4 Aspects of Design and Summary Plan

From this literature review we can see that while a tremendous amount of exploration into model selection methods has occurred in the linear univariate model (and, some of these methods have been extended to include more general models), fewer extensions have been done to include the linear mixed model.

Exploration of the behavior of the KIC, a proposed C_p, and PRESS statistics in the linear mixed model will allow for a wider variety of model selection methods to be available when modeling repeated measures data. This will thereby further increase the probability that the correct model structure is being chosen when finding the appropriate model for the data.

In order to explore the behavior of these statistics in the linear mixed model empirically, and to evaluate their performance compared to the widely used AIC and BIC, we used both simulation studies and real data examples. Simulation studies are useful in that when the data are simulated, we have prior knowledge of the underlying fixed and random effects, and therefore can determine the accuracy of our statistics in identifying the underlying models correctly. Our real data examples are useful in that they can be an important gauge of how well our criteria behave when missing data are present or when looking at other aspects of experimental design.

1.5 Summary

In this dissertation, we attempt to both evaluate the performance of the widely-used information criteria, AIC and BIC, and compare their performance to that of the developed but less familiar KIC, our C_p, and a developed, but untested PRESS statistic.

In chapter 2, we propose and evaluate a C_p statistic for the linear mixed model. We developed the C_p statistic using the F_p statistic defined previously, and evaluate its behavior in a variety of simulations testing selection of fixed effects, in a small-sample, complete real data set, and in a large data set containing missing data. In all simulations and examples, we compare the performance of the C_p statistic with that of the AIC and BIC.

In chapter 3, we evaluate the performance of the KIC in comparison to that of the AIC and BIC. This evaluation is done by looking at the performance of all three criteria in simulation studies where it is necessary to select the correct covariance structure when the fixed effect structure is known, and in selecting the correct fixed effects structure when the correct covariance structure is known. In addition, the performance of the KIC in selecting the correct set of fixed effects is evaluated using the same large data set with missing data as was used in chapter 2.

In chapter 4, we evaluate the performance of the PRESS statistic, as developed and defined in SAS version 9.1.3 (SAS Institute, Cary, NC 2007) and compare its performance to the $C_{p,}$ the KIC and the well-known AIC and BIC. The simulation study performed here is identical to that explored in chapter 3 for selection of fixed effects when the covariance structure is known. In addition, the real data example used in chapter 3 is again used in this chapter to evaluate the performance of the PRESS statistic and to compare it to the results received in chapters 2 and 3. We conclude in chapter 5 with an overall discussion and analysis of model selection criteria in the linear mixed model and look into areas of future research.

This dissertation format will follow the "manuscript" style of dissertations, where Chapters 2, 3 and 4 will represent individual papers that are publishable in various statistical journals. Each of these chapters will include more detail than would normally be seen in a journal article, however, as this is a dissertation. A version of Chapter 2 has

been submitted for publication, with plans to submit versions of the other two chapters in the near future. Each chapter contains a brief literature review with reference sections specific to the topic.

CHAPTER 2

A C_p STATISTIC FOR FIXED EFFECTS VARIABLE SELECTION IN THE LINEAR MIXED MODEL

2.1 Introduction

Mallows' C_p statistic (Mallows 1964, Gorman and Toman 1966) has been of great use in the univariate linear model when selecting from a pool of models which are each separately nested within a single full model. It compares the mean square error (*MSE*) of each candidate model to the *MSE* of the full model, which then allows comparing one candidate to another.

In the longitudinal data setting, data analysts often use the linear mixed model with Gaussian deviations (Demidenko, 2004; Fitzmaurice et al., 2008; Laird and Ware, 1982). The linear mixed model serves the same role in longitudinal data analysis as the linear univariate model does in cross-sectional analysis. The linear mixed model extends the univariate linear model with independently and identically distributed (i.i.d.) Gaussian errors to a wide variety of correlated and commensurate Gaussian data (all responses measured in the same units). Both the univariate linear model and the linear mixed model provide basic foundations for developing model selection procedures. However, in contrast to standard univariate linear models, the quantity and quality of model selection methods for the linear mixed model leave much room for improvement.

Unlike the linear univariate model, there is no C_p statistic for the linear mixed model. In this paper we propose a C_p statistic for fixed effects variable selection in the linear mixed

model with a focus on the analysis of longitudinal data. This paper is an empirical evaluation of this newly formed statistic. The distributional aspects have not been evaluated as of yet. In what follows, section 2 provides background on the C_p statistic in the univariate linear model. In section 3 we propose a C_p statistic for the linear mixed model. The AIC and BIC are also discussed. To assess the performance of the proposed C_p statistic, we present results of simulation studies in section 4. In section 5 we apply the proposed C_p statistic to actual data from two different studies, one a small sample study and another a large sample study. A discussion of the proposed C_p statistic is provided in section 6.

2.2 Notation and Basic Concepts of C_p for the Univariate Linear Model

The univariate linear model is given in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}\,.$$

Here $\boldsymbol{y}$ $(N \times 1)$ is a vector of responses corresponding to N persons (independent sampling units). The design matrix, $\boldsymbol{X}$ $(N \times (p+1))$, consists of p independent predictors and an intercept. The vector $\boldsymbol{\beta}$ $((p+1) \times 1)$ is an unknown vector of regression parameters and $\boldsymbol{e}$ $(N \times 1)$ is a vector of independent random errors. Typically, the rows of $\boldsymbol{e}$ are assumed to be normally distributed with mean 0 and common variance σ^2; i.e., $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{\mathcal{I}})$. For Mallows C_p statistic, the case of no intercept is handled similarly.

The following conceptual predictive (C_p) criterion for use in the univariate linear model was first proposed by C. L. Mallows (1964) and first published by Gorman and Toman (1966):

$$C_p = \frac{SSE(p)}{\widehat{\sigma}^2} - [N - 2(p+1)] \tag{2.1}$$

where $SSE(p)$ represents the sums of square error from a nested sub-model and $\widehat{\sigma}^2$ is an estimate of the variance of the random error from the saturated (maximum) model. In linear regression, the formula used most often is:

$$C_p = \frac{SSE(p)}{MSE(p_{max})} - [N - 2(p+1)] \tag{2.2}$$

where N is the number of independent sampling units, p is the number of independent predictors in the candidate model, p_{max} is the number of independent predictors in the maximum model, $MSE(p_{max})$ represents the mean square error for the maximum model, and $SSE(p)$ represents the sums of square error for the candidate model. This criterion requires a pool of candidate models which are each separately nested within a single full model. It is an estimate of the measure of adequacy for prediction given by the scaled sum of errors (Ronchetti and Staudte 1994).

Mallows suggested that a value of C_p too large or too far above $p+1$ indicates an inaccurate model. Hocking (1976) suggested requiring $C_p \leq p+1$ in choosing a model for prediction, and requiring $C_p \leq 2(p+1) - p_{max}$ in choosing a model for parameter estimation.

In the univariate linear model, C_p can also be defined using the F statistic for comparing a candidate model to the maximum model, denoted F_p,

$$F_p = \frac{[SSE(p) - SSE(p_{max})]/(p_{max} - p)}{SSE(p_{max})/(N - p_{max} - 1)} \tag{2.3}$$

The F_p statistic compares to an F distribution with $p_{max} - p$ and $N - p_{max} - 1$ degrees of freedom. Since the F statistic corresponds to a test of two models, then F_p corresponds to a test, using the saturated model, of whether the $p_{max} - p$ regression coefficients not in the candidate model are simultaneously zero. If this criterion is significant, the saturated model includes variables that significantly improve upon the predictive ability of the model, when compared to the model with p variables. We can express the C_p statistic as a simple function of the F_p statistic as follows:

$$C_p = (p_{max} - p)F_p + (2p - p_{max} + 1) \tag{2.4}$$

Mallows (1973) expounded on the statistic saying that he "feels that the greatest value of the device is that it helps the statistician to examine some aspects of the structure of his data and helps him to recognize the ambiguities that confront him." He further stated that "this device cannot be expected to provide a single 'best' equation when the data are intrinsically inadequate to support such a strong inference." Mallows found that the ambiguous cases where the "minimum C_p" rule will give bad results are where there are a large number of subsets that have C_p's that are close to each other.

2.3 Proposed C_p for the Linear Mixed Model

With N independent sampling units (often *persons* in practice), the linear mixed model for person i may be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{d}_i + \boldsymbol{e}_i\,.$$

Here, $\boldsymbol{y}_i$ is a $p_i \times 1$ vector of observations on person i; $\boldsymbol{X}_i$ is a $p_i \times q$ known, constant design matrix for person i, with full column rank q while $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown, constant, population parameters. Also $\boldsymbol{Z}_i$ is a $p_i \times m$ known, constant design matrix with rank m for person i corresponding to the $m \times 1$ vector of unknown random effects $\boldsymbol{d}_i$, while $\boldsymbol{e}_i$ is a $p_i \times 1$ vector of unknown random errors. Gaussian $\boldsymbol{d}_i$ and $\boldsymbol{e}_i$ are independent with mean $\mathbf{0}$ and

$$\mathcal{V}\left(\begin{bmatrix}\boldsymbol{d}_i \\ \boldsymbol{e}_i\end{bmatrix}\right) = \begin{bmatrix}\boldsymbol{\Sigma}_{di}(\boldsymbol{\tau}_d) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e)\end{bmatrix}.$$

Here $\boldsymbol{\mathcal{V}}(\,\cdot\,)$ is the covariance operator, while both $\boldsymbol{\Sigma}_{di}(\boldsymbol{\tau}_d)$ and $\boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e)$ are positive-definite, symmetric covariance matrices. Therefore $\boldsymbol{\mathcal{V}}(\boldsymbol{y}_i)$ may be written $\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{\Sigma}_{di}(\boldsymbol{\tau}_d)\boldsymbol{Z}_i' + \boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e)$. We assume that $\boldsymbol{\Sigma}_i$ can be characterized by a finite set of parameters represented by an $r \times 1$ vector $\boldsymbol{\tau}$ which consists of the unique parameters in $\boldsymbol{\tau}_d$ and $\boldsymbol{\tau}_e$. Throughout $n = \sum_{i=1}^{N} n_i$.

While there have been many developments in the C_p statistic since it was proposed by Mallows (1964), none of these developments have extended the C_p to the linear mixed

model. Cantoni, et al. (2005) did suggest the GC_p for use outside the linear model, but due to its focus on marginal longitudinal generalized linear models, we were motivated to seek a solution specific to the linear mixed model. Also, while many articles differ on whether to use the lowest value of C_p as the model selected by the statistic or to use the C_p closest to $p+1$, we decided it was best to follow Mallows' decision to use the lowest value of C_p to decide what model is chosen by this statistic.

As shown in the previous section, the F_p statistic and C_p are related in that they both are tests of comparisons within nested models. In the case of the linear mixed model, we focus our discussion on the fixed effects portion of the model. Sum of squares error or mean squared error in the linear mixed model cannot be expressed the same as in the univariate linear model, due to the dual components of fixed and random effects involved in the model. However, we are able to test the fixed effects in the linear mixed model via an F statistic. The global F statistic test$s$ whether the fixed effects included in the linear mixed model are equal to zero. Hence, we can manipulate this statistic to form an F_p statistic that corresponds to a test, using the saturated model, of whether the regression coefficients not included in the candidate model are simultaneously zero. We can then use this F_p statistic to derive our C_p statistic for the linear mixed model in the same manner as was discussed in section 2. The equation for our C_p statistic is:

$$C_p = (p_{max} - p)F_p + (2p - p_{max} + 1) \tag{2.5}$$

where, p is the number of parameters in the fixed effects portion of the candidate model and p_{max} is the number of parameters in the fixed effects portion of the saturated model.

The linear mixed model explicitly specifies not only the mean structure, but also the covariance structure. Therefore, three types of model comparisons can occur. I) Compare mean models with the same covariance structure. Nested mean models are the most common. II) Compare covariance models with the same mean structure. Two linear mixed models may be nested or nonnested in the covariance models. III) Compare linear mixed

models with different mean and different covariance structures. Here, the proposed C_p statistic is in relation to item I, i.e., comparing nested mean models with the same covariance structure.

In order to assess the proposed C_p, we will compare its performance against commonly used information criteria. Information theoretic criteria have played a prominent role in mixed model selection. Most practitioners use the Akaike Information Criterion (AIC, Akaike 1974) and the Bayesian Information Criterion (BIC, Schwarz 1978). Both criteria are essentially log-likelihood values with a penalty for the number parameters estimated. For the linear mixed model we use,

$$AIC = -2\log f(Y|\widehat{\theta}_k) + 2k \tag{2.6}$$

$$BIC = -2\log f\left(Y|\widehat{\theta}_k\right) + k\log n, \tag{2.7}$$

where $\log f(Y|\widehat{\theta}_k)$ is the log-likehood evaluated at $\widehat{\theta}_k$, a vector of model parameter estimates.

When comparing models, the smaller the value of the information criterion, the better the fit. Both the AIC and BIC involve differences in log-likelihood values and numbers of parameters. For purposes of interpretation, both criteria have the unfortunate characteristic of being data scale dependent. The BIC includes the number of observations in the data, which seeks to give a finite-sample correction.

In this paper, we will compare the ability of our newly formulated C_p statistic to select the correct set of fixed effects in the linear mixed model, to that of the AIC and BIC. For our simulation studies and our real data analysis we will use REML for estimation and Kenward and Roger (1997) F statistic for inference.

2.4 Simulation Study

In the following simulation study we looked at the ability of the C_p statistic to determine the correct set of fixed effects for simulated data in three different scenarios. All three scenarios base their random effects structure on a classic data example from Pothoff and Roy (1964) that is examined in greater depth later in this paper. In addition, the first and second scenarios are similar to the fixed effects in that paper. In the first scenario, the fixed effects take into account time, a group effect and an interaction effect, where the magnitude of the group effect and interaction effect are varied. The second scenario has fixed effects that include only a time and group effect, and the magnitude of each are varied. The third scenario looks at the ability of the statistic to determine the correct set of fixed effects when there are multiple categorical predictors.

2.4.1 First Scenario: Two Groups, Time and Interaction Effect

2.4.1.1 Introduction

For the first scenario, the data simulated have random effects loosely based on the Potthoff and Roy (1964) data example which was an orthodontic study with 27 children, 16 boys and 11 girls. For each child, the distance (mm) from the center of the pituitary to the pterygomaxillary fissure was measured at ages 8, 10, 12, and 14 years. However, in our simulated data, we considered a sample size of 50 children with about half belonging to each gender classification. Fixed effects for the simulation include a shared intercept, covariates for time and group effect and an interaction effect between group and time. This scenario looks at how the C_p, AIC and BIC statistics behave when the group effect and the interaction effect are varied.

2.4.1.2 Methods

For this set of simulations, we used an i.i.d. within-subject covariance structure i.e.,

$$\mathbf{\Sigma}_{e_i}(\boldsymbol{\tau}_e) = \sigma_0^2 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

with an unstructured between-subject covariance structure containing a random intercept and slope, i.e.,

$$\mathbf{\Sigma}_{d_i}(\boldsymbol{\tau}_d) = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix}$$

where $\sigma_0^2 = 2,\ \sigma_1^2 = 7.823$, $\sigma_2^2 = 0.051$, and $\rho_d = -0.77$ (i.e., $\sigma_{12}^2 = -0.486$).

For the fixed effects, where β_1 is the common intercept, β_2 is the coefficient for time, β_3 is the coefficient for the group membership, and β_4 is the coefficient for the interaction between group and time, we set $\beta_1 = 3$ and $\beta_2 = 1$ for all the simulations, varied the value of β_3 from 0.5 to 10, and varied the value of β_4 from 0 to -2.0.

For each set of simulations, three different models of fixed effect were assessed using SAS Proc Mixed (2007). These three models were: 1) Model with time only, 2) Model with time and group effect only, and 3) Model with time, group and interaction. For this approach, model (3) is considered the saturated model.

The ability of the C_p, AIC and BIC to select the correct model (i.e. model 3 when $\beta_4 \neq 0$) is determined over 10,000 simulations, and a percentage for correctly selecting the true linear mixed model is determined.

2.4.1.3 Results

Table 2.1 gives the results for the C_p, AIC and BIC statistics from the simulation study where the data simulated had a within-subject covariance structure that was i.i.d., unstructured random effects covariance with random intercept and random slope, and fixed effects given by:

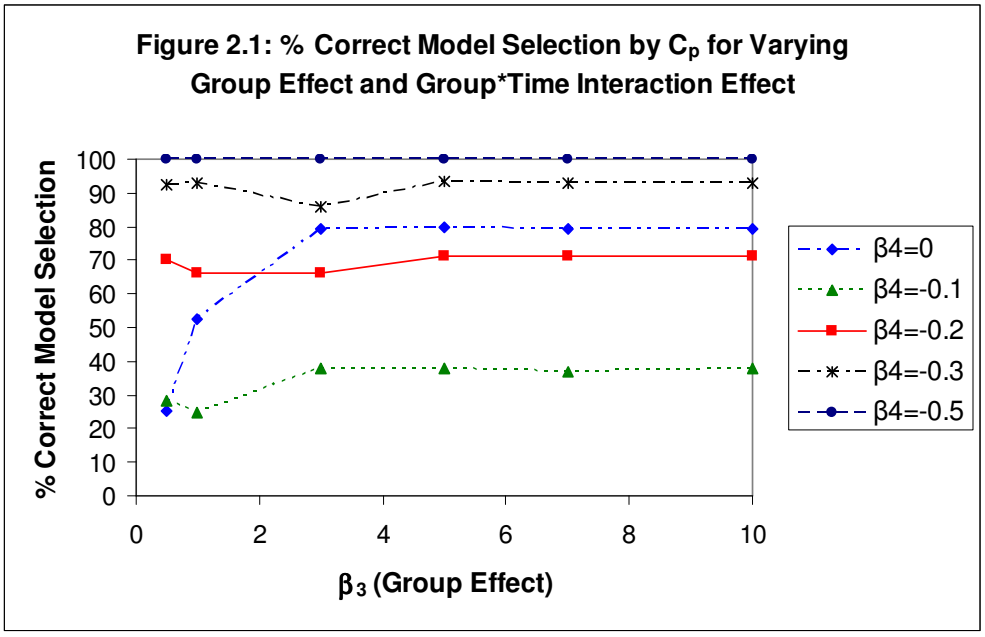
$$E(y) = 3 + t + \beta_3 x + \beta_4 (x{*}t)$$

where β_3 varies between 0.5 and 10 and β_4 varies between 0 and -2.0.

Table 2.1: Correct Model Selection by the C_p Statistic for Varying Group Effect (β_3) and Group * Time Interaction Effect (β_4)

		% Correct Model Selection									
		β_4=									
β_3	Criterion	0	-0.1	-0.2	-0.3	-0.5	-0.7	-1.0	-1.2	-1.5	-2.0
0.5	C_p	25.1	28.1	70.3	92.6	99.9	100	100	100	100	100
	AIC	86.3	28.3	62.4	88.5	99.9	100	100	100	100	100
	BIC	86.1	28.3	62.3	88.4	99.9	100	100	100	100	100
1	C_p	52.4	25.0	66.0	92.9	100	100	100	100	100	100
	AIC	86.7	28.2	61.3	88.9	99.9	100	100	100	100	100
	BIC	86.4	28.1	61.1	88.9	99.9	100	100	100	100	100
3	C_p	79.3	37.7	66.0	86.1	100	100	100	100	100	100
	AIC	86.6	28.4	62.1	88.6	99.9	100	100	100	100	100
	BIC	86.5	28.4	61.7	88.3	99.9	100	100	100	100	100
5	C_p	79.6	38.1	71.2	93.4	99.8	100	100	100	100	100
	AIC	86.9	29.2	62.3	89.9	99.9	100	100	100	100	100
	BIC	86.9	29.2	62.2	89.8	99.9	100	100	100	100	100
7	C_p	79.1	37.0	71.1	93.1	99.9	100	100	100	100	100
	AIC	86.5	27.8	62.7	89.3	99.8	100	100	100	100	100
	BIC	86.5	27.9	62.6	89.2	99.8	100	100	100	100	100
10	C_p	78.9	37.8	71.7	93.1	99.9	100	100	100	100	100
	AIC	86.5	28.0	62.7	89.2	99.9	100	100	100	100	100
	BIC	86.5	28.0	62.6	89.1	99.8	100	100	100	100	100

Figure 2.1 graphically presents the results in Table 2.2 for the C_p statistic but does not include the results where β_4 is less than -0.5 (as these results are all the same).



2.4.1.4 Discussion

From this scenario, we see that, for all of our criteria, if the interaction effect is fairly strong (i.e. $\beta_4 \geq$ -0.5), the correct fixed effect structure is selected 100% of the time. However, when the group effect is small and the interaction effect is non-existent, the C_p appears to struggle, in comparison to the AIC and BIC, in correctly identifying the fixed effects. In all other situations, the C_p is comparable to the AIC and BIC in its ability to identify the correct set of fixed effects.

2.4.2 Second Scenario: Two Groups and Time

2.4.2.1 Introduction

In this scenario, the random effects are the same as in scenario 1, and the fixed effects are comprised of a shared intercept, a time effect and a group effect. This scenario looks at how the C_p, AIC and BIC statistics perform when the time effect and group effect are varied.

2.4.2.2 Methods

For this set of simulations, the same covariance structure is used as in the previous section, and time is defined as in the Pothoff and Roy example (i.e. age 8, 10, 12 and 14 years). For the fixed effects, where β_1 is the common intercept, β_2 is the coefficient for the group membership, and β_3 is the coefficient for time, we set $\beta_1 = 3$, varied the value of β_2 from 0 to 10 and varied the value of β_3 from 0 to 2.

For each set of simulations, two different models of fixed effects were looked at using SAS Proc Mixed. These three models were: 1) Model with time only and 2) Model with time and group effect only. For this approach, model (2) is considered the saturated model. The ability of the C_p, AIC and BIC to select the correct model (i.e. model 2 when $\beta_4 \neq 0$) is determined over 10,000 simulations, and a percentage for correctly selecting the true linear mixed model is determined.

2.4.2.3 Results

Table 2.2 gives the results from the simulation study where the data simulated had a covariance structure that was i.i.d. with random intercept and random slope and fixed effects given by:

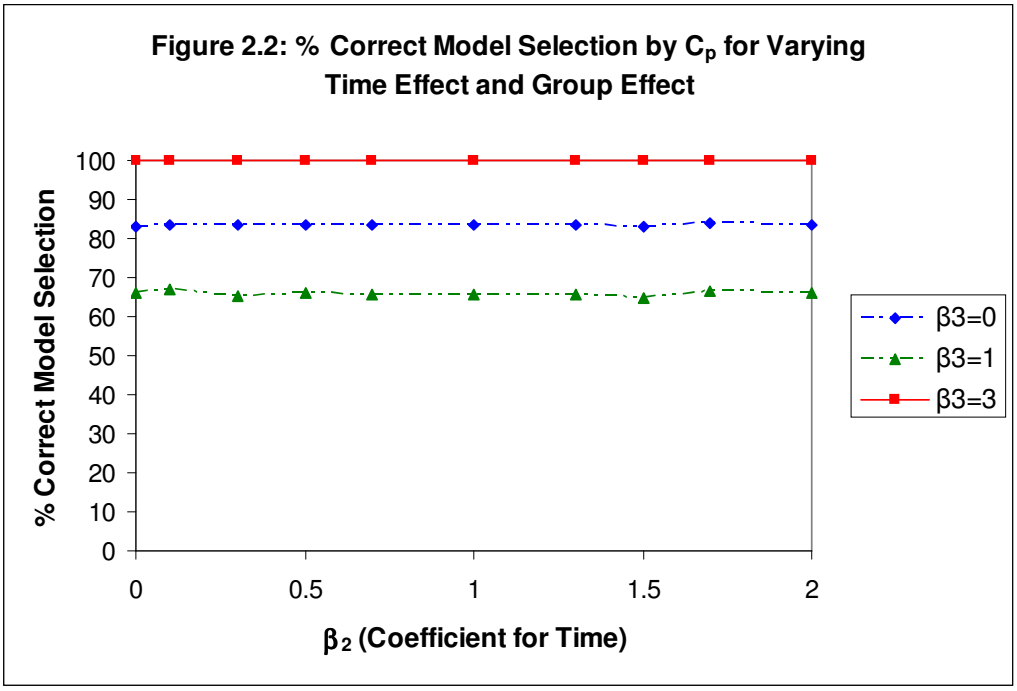
$$E(y) = 3 + \beta_2 t + \beta_3 x$$

where β_2 varies between 0.05 and 2 and β_3 varies between 0 and 10.

Table 2.2: Correct Model Selection by the C_p Statistic for Varying Time Effect (β_2) and Group Effect (β_3)

		% Correct Model Selection									
		(Coefficient for Time) β_2=									
β_3	**Criterion**	**0.05**	**0.1**	**0.3**	**0.5**	**0.7**	**1.0**	**1.3**	**1.5**	**1.7**	**-2.0**
0	C_p	83.2	83.5	83.5	83.5	83.6	83.4	83.6	83.1	84.0	83.6
	AIC	0.7	0.7	0.9	0.8	0.6	0.7	0.7	0.8	0.8	0.7
	BIC	1.0	1.0	1.1	1.4	0.8	1.0	0.9	1.1	1.1	1.0
1	C_p	65.9	66.9	65.1	65.9	65.8	65.6	65.5	64.9	66.6	66.2
	AIC	99.5	99.7	99.5	99.6	99.7	99.8	99.7	99.7	99.8	99.6
	BIC	99.1	99.1	99.0	99.2	99.2	99.2	99.1	99.2	99.3	99.1
3	C_p	100	100	100	100	100	100	100	100	100	100
	AIC	100	100	100	100	100	100	100	100	100	100
	BIC	100	100	100	100	100	100	100	100	100	100
5	C_p	100	100	100	100	100	100	100	100	100	100
	AIC	100	100	100	100	100	100	100	100	100	100
	BIC	100	100	100	100	100	100	100	100	100	100
7	C_p	100	100	100	100	100	100	100	100	100	100
	AIC	100	100	100	100	100	100	100	100	100	100
	BIC	100	100	100	100	100	100	100	100	100	100
10	C_p	100	100	100	100	100	100	100	100	100	100
	AIC	100	100	100	100	100	100	100	100	100	100
	BIC	100	100	100	100	100	100	100	100	100	100

Figure 2.2 graphically presents the results for C_p in Table 4.3 but does not include the results where β_3 is greater than 3 (as these results are all the same).



2.4.2.4 Discussion

From this scenario, we see that, for all of our criteria, if the group effect is fairly strong (i.e. $\beta_3 \geq 3$), the correct fixed effect structure is selected 100% of the time. However, when the group effect is non-existent, the C_p works fairly well but the AIC and BIC are not effective at all in correctly identifying the fixed effects. When the coefficient for the group effect is 1, i.e. $\beta_3 = 1$, the AIC and BIC work better than the C_p in correctly identifying the fixed effects, but the performance of the C_p is still fairly good.

2.4.3 Third Scenario: Multiple Categorical Predictors

2.4.3.1 Introduction

The multiple categorical predictors that are used to generate the data simulate the case where there are various factors (in addition to the time effect), such as race, gender etc., that have an impact on the outcome variable. To determine which denominator degrees of freedom to use for our statistic, both Kenward-Roger F statistic and the F statistic using the

residual ($n - \text{rank}(\boldsymbol{X})$) denominator degrees of freedom (ddf) were used when evaluating the simulated data

2.4.3.2 Methods

For this set of simulations, again, we used a sample size of 50 subjects, an i.i.d. within-subject covariance structure with an unstructured between-subject covariance structure containing a random intercept and slope. In this simulation, $\sigma_0^2 = 2$, $\sigma_1^2 = 2$, $\sigma_2^2 = 1$, and $\rho_d = 0.25$ (i.e. $\sigma_{12}^2 = 0.35$).

For the fixed effects, β_1 is the common intercept, β_2 is the coefficient for time and four different categorical predictors, $x_1,\ x_2,\ x_3$ and x_4, are considered with the corresponding coefficients β_3, β_4, β_5, and β_6. And time is again defined as in the two previous scenarios (i.e. age 8, 10, 12 and 14 years) The vector for $\beta' = (3, 1, 4, 3, 5, 5)$, which are coefficients that were chosen at random, and 16 different variations are considered where 0,1, 2, 3, or 4 of the coefficients for the categorical predictors (i.e. $x_1,\ x_2,\ x_3$ and x_4) are set to 0. Therefore, the saturated model, which includes all coefficients for the categorical predictors and time, is given by:

$$E(y) = \ 3 + t + 4x_1 + 3x_2 + 5x_3 + 5x_4$$

For each set of simulations, 16 different sets of fixed effects were assessed using SAS Proc Mixed. In order to calculate the C_p statistic for the 16 different models, the saturated model is fit and tests of hypotheses about the fixed effects (contrast statements) are conducted that test whether the covariate(s) excluded from the model of interest is equal to zero. This method is equivalent to the creation of an F_p statistic testing whether the covariate(s) equal zero. This F_p statistic is then used to formulate our C_p statistic, using formula 2.1. The model selected by the C_p statistic is the set of fixed effects that has the smallest C_p value.

The ability of the C_p to select the correct model is determined over 1,000 simulations using REML estimation and Kenward-Roger F statistic and associated ddf. A percentage of correct model selection is calculated.

2.4.3.3 Results

Table 2.3 gives the results from the simulation study where the data simulated had a within-subject covariance structure that was i.i.d., unstructured random effects covariance with random intercept and random slope, and fixed effects given by:

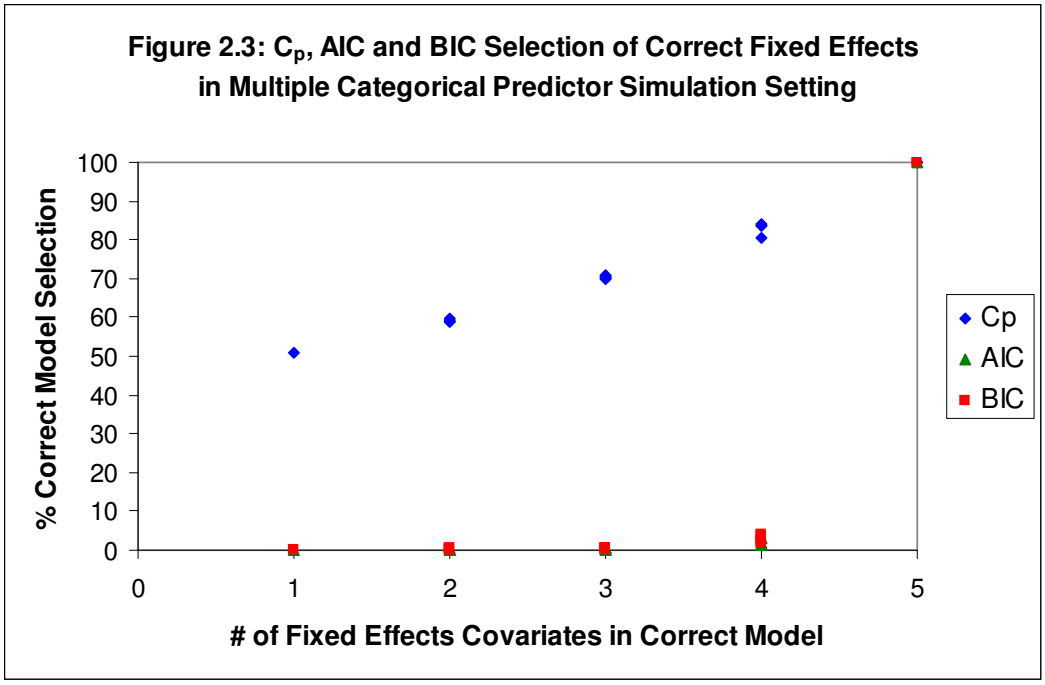
$$E(y) = \beta_1 + \beta_2 t + \beta_3 x_1 + \beta_4 x_2 + \beta_5 x_3 + \beta_6 x_4$$

Where β_1 is the common intercept, β_2 is the coefficient for time and four different categorical predictors, x_1, x_2, x_3 and x_4, are considered with the corresponding coefficients β_3, β_4, β_5, and β_6. The vector for $\beta' = (3, 1, 4, 3, 5, 5)$, and 16 different variations are considered where 0,1, 2, 3, or 4 of the coefficients for the categorical predictors (i.e. x_1, x_2, x_3 and x_4) are set to 0. The table shows the correct model for the fixed effects (i.e. the model used to simulate the data), the number of covariates in the model used, and the ability of the C_p statistic and the AIC and BIC statistics to correctly identify the model using SAS Proc Mixed.

Table 2.3: % Correct Model Selection by C_p Statistic for Multiple Categorical Predictors

Fixed Effects for Simulated Data	# of Covariates	% Correct Selection by		
		C_p	AIC	BIC
$3 + t + 4x_1 + 3x_2 + 5x_3 + 5x_4$	5	100	99.9	99.9
$3 + t + 4x_1 + 3x_2 + 5x_3$	4	84.0	1.3	1.8
$3 + t + 4x_1 + 3x_2 + 5x_4$	4	83.6	3.1	3.8
$3 + t + 4x_1 + 5x_3 + 5x_4$	4	80.7	1.7	2.4
$3 + t + 3x_2 + 5x_3 + 5x_4$	4	83.8	3.1	3.5
$3 + t + 4x_1 + 3x_2$	3	70.0	0.1	0.1
$3 + t + 4x_1 + 5x_3$	3	70.9	0.4	0.4
$3 + t + 4x_1 + 5x_4$	3	70.2	0.5	0.6
$3 + t + 3x_2 + 5x_3$	3	70.6	0.1	0.3
$3 + t + 3x_2 + 5x_4$	3	69.9	0.4	0.4
$3 + t + 5x_3 + 5x_4$	3	70.5	0.5	0.6
$3 + t + 4x_1$	2	59.6	0	0
$3 + t + 3x_2$	2	58.9	0.1	0.3
$3 + t + 5x_3$	2	58.9	0.1	0.1
$3 + t + 5x_4$	2	59.3	0.3	0.5
$3 + t$	1	50.7	0	0

Figure 2.3 graphically represents the data in Table 2.3 by plotting the percentage of correct model selection by the C_p statistic vs. the number of covariates in the correct model.



2.4.3.4 Discussion

From the results of this part of the simulation study, we can clearly see that in this situation, our C_p statistic far exceeds the AIC and BIC statistics in its ability to identify the correct set of fixed effects. The AIC and BIC are only able to correctly identify the fixed effects when the saturated model is the correct model. On the other hand, our C_p statistic is able to correctly identify the fixed effects in more than half of the simulated data sets regardless of the identity of the correct model.

2.4.4 Conclusions

From the three different scenarios covered in this simulation study, we see that while there are moments where the AIC and BIC outperform the C_p statistic, the C_p statistic performs well overall. In comparison, while the AIC and BIC have moments where their

performance is near perfection, when these criteria do not perform well, they perform abysmally.

From the scenario where multiple categorical predictors are simulated, we see that while the AIC and BIC are not capable of correctly identifying the fixed effects in any setting other than when the saturated model is simulated, the C_p statistic is able to correctly identify the fixed effects over 50% of the time in all settings.

2.5 Example Data

To further investigate the performance of the C_p statistic for fixed effects variable selection in the linear mixed model, we looked at two different example datasets. The first set of example data is well-known and come from a Potthoff and Roy (1964) dental data study. The second set of example data come from a larger study of blood pressure measurements containing missing data from the North Carolina Established Populations for the Epidemiologic Studies of the Elderly (EPESE).

2.5.1 Example Data: Dental Study

2.5.1.1 Background

We used a well known example from Potthoff and Roy (1964) for the first investigation into the utility of the C_p statistic. The data come from an orthodontic study that involved 27 children, 16 boys and 11 girls. For each child, the distance (mm) from the center of the pituitary to the pterygomaxillary fissure was measured at ages 8, 10, 12, and 14 years with complete data for each child. The objectives of the dental study were to determine whether, on the average over time, distances are larger for boys than for girls and whether, on the average over time, the rate of change of the distance is similar for boys and girls.

2.5.1.2 Methods

As the only explanatory variables for the outcome are age and gender, we looked at the values of the C_p statistic for the linear mixed model with four different fixed effects

structures 1) model with continuous age effect alone, 2) model with classification gender effect alone, 3) model with both continuous age effect and classification gender effect, and 4) model with continuous age, classification gender and their interaction (maximum model). The covariance structure we used had random intercepts and slopes with unstructured covariance for the random effects, $\mathbf{\Sigma}_{d_i}(\boldsymbol{\tau}_d)(2 \times 2)$, and $\mathbf{\Sigma}_{e_i}(\boldsymbol{\tau}_e) = \sigma_0^2 \boldsymbol{I}_{n_i}$ (i.e. i.i.d.) for the covariance structure of the within-subject error. The covariance structure of the random effects can be represented as:

$$\mathbf{\Sigma}_{d_i}(\boldsymbol{\tau}_d) = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix}$$

where $\sigma_{12}^2 = \rho_d * \sigma_1 * \sigma_2$.

In order to calculate the C_p statistic for each set of fixed effects, we fit the saturated model (i.e. the model that included age, gender and the interaction term), and then set up contrast statements that test whether the excluded covariate in the models of interest are equal to zero, thus formulating the appropriate F_p statistics, and then calculating the C_p for the model using equation 3.1.

For example, to find the C_p for model (1), we use a contrast statement that includes both "gender" and "age*gender", this creates an F_p statistic testing whether these two terms are equal to zero. To then calculate the C_p for this model, where $p = 1$ and $p_{max} = 3$, we get:

$$C_p(\text{model 1}) = 2 * F_p$$

2.5.1.3 Results

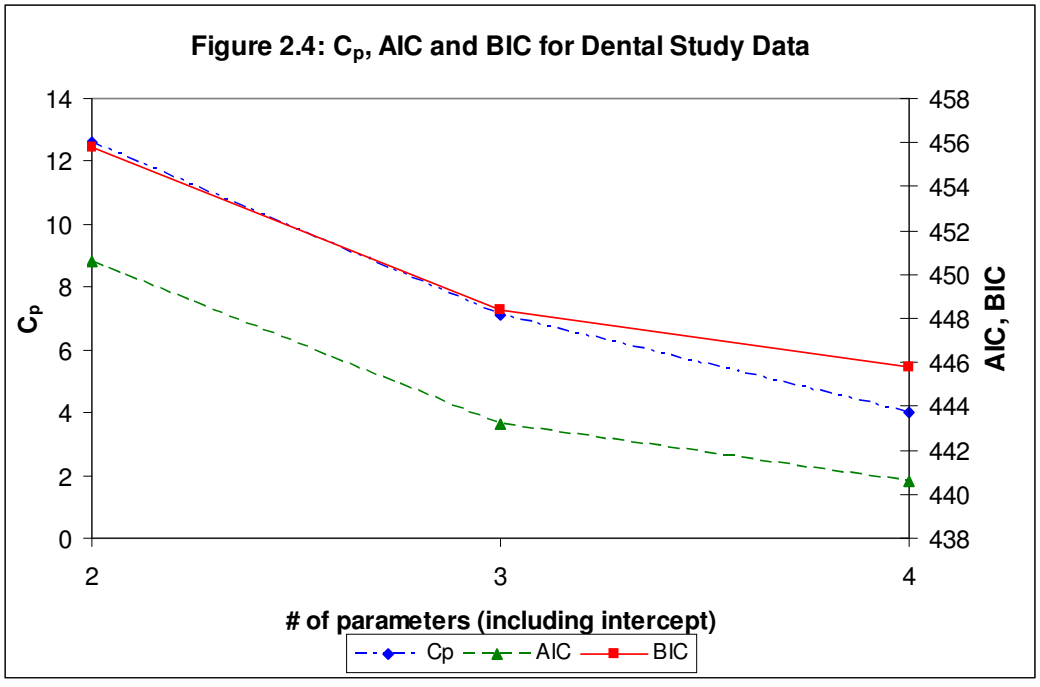
Table 2.4 provides the estimates, standard errors, and *p*-values for fixed effects in Models 1, 2, 3 and 4 described above and the estimates of covariance parameters. Table 2.5 gives the calculated C_p statistic and the AIC and BIC statistics for the models. Figure 2.4 give a graphical representation of the data in table 2.4 (excluding Model 2).

Table 2.4: Mixed Model Fixed Effect Estimates, Standard Errors (SE), p-values and Covariance Estimates for Dental Data

					Covariance Estimates	
Model	**Fixed Effect**	**Estimate**	**SE**	**p-value**	**Random effects**	**Error**
1	Intercept	16.76	0.775	<0.001	$\hat{\sigma}_1^2 = 5.415$	$\hat{\sigma}_0^2 = 1.716$
	Age	0.66	0.071	<0.001	$\hat{\sigma}_2^2 = 0.051$	
					$\hat{\rho}_d = -0.61$	
2	Intercept	22.37	0.583	<0.001	$\hat{\sigma}_2^2 = 54.63$	$\hat{\sigma}_0^2 = 1.716$
	Gender	2.15	0.757	0.0055	$\hat{\sigma}_1^2 = 0.482$	
					$\hat{\rho}_d = -52.96$	
3	Intercept	15.49	0.944	<0.001	$\hat{\sigma}_1^2 = 7.823$	$\hat{\sigma}_0^2 = 1.716$
	Age	0.66	0.071	<0.001	$\hat{\sigma}_2^2 = 0.051$	
	Gender	2.15	0.758	0.0055	$\hat{\rho}_d = -0.77$	
4	Intercept	17.37	1.2284	<0.001	$\hat{\sigma}_1^2 = 5.786$	$\hat{\sigma}_0^2 = 1.716$
	Age	0.4795	0.1037	<0.001	$\hat{\sigma}_2^2 = 0.033$	
	Gender	−1.0321	1.5957	0.519	$\hat{\rho}_d = -0.66$	
	Age*Gender	0.3048	0.1347	0.026		

Table 2.5: C_p, AIC and BIC Results for Dental Data

Model	**Fixed Effects**	F_p	C_p	AIC	BIC
1	Age	6.31	12.62	450.6	455.8
2	Gender	52.28	104.56	479.2	484.3
3	Age, Gender	5.12	7.12	443.2	448.4
4	Age, Gender, Age*Gender	–	4.00	440.6	445.8



2.5.1.4 Discussion

When looking at the results, we see that the lowest C_p, AIC and BIC is associated with the full model which includes the interaction term. By definition, the C_p of the full model is always equal to $p + 1$, but in our example, it is also the lowest C_p value overall. This assertion is sustained is further supported when looking at the p-values for the parameter estimates for Model 4, where we see that the interaction term is significant at a level of $\alpha = 0.05$.

2.5.2 Example Data: Elderly Blood Pressure Study

2.5.2.1 Background

This data comes from a retrospective longitudinal cohort study from the North Carolina Established Populations for the Epidemiologic Studies of the Elderly (EPESE) (Blazer and George, 2004). The goals of the EPESE project were to describe and identify predictors of mortality, hospitalization, and placement in long-term care facilities and to investigate risk factors for chronic diseases and of functioning among the elderly. The North Carolina cohort was established in a 1986-1987 baseline survey, and was a sample of persons 65 or older residing in households in Durham, Warren, Franklin, Granville, and Vance counties (one urban, four rural) in the Central Piedmont area of North Carolina. The site was over 50% black, and the geographic area selected was diverse, allowing both racial and urban/rural comparisons to be made regarding the distribution of certain risk factors and disease. Of the 4162 subjects selected on the basis of a four-stage, race-stratified sampling design, 48% (including similar proportions of Blacks and Whites) lived in the urban community. Participants were surveyed at four time periods: Wave 1(1986); Wave 2 (1990); Wave 3 (1994); and Wave 4 (1998).

2.5.2.2 Methods

To investigate the behavior of the C_p statistic using this data, we used an all-possible regressions approach (where time in years is in every model) for the outcome variable using

mixed models. We looked at average diastolic blood pressure as the response variable and used only main effects as predictors. There were 12 separate main effects chosen for study from a set of predictors numbering larger than 50: including time in years, 4 self-reported illness indices, race, marital status, gender, weight, diagnosis of diabetes, diagnosis of heart disease and whether the subject lived in a rural area). All predictors were binary categorical variables, except for time (which was labeled as 0, 4, 8, and 12 years). We calculated the C_p statistic for 2048 sets of fixed effects for the outcome variable. We used an unstructured random effects covariance structure, $\mathbf{\Sigma}_{d_i}(\boldsymbol{\tau}_d)$ (2×2), and i.i.d. within-subject error covariance, $\mathbf{\Sigma}_{e_i}(\boldsymbol{\tau}_e) = \sigma_0^2 \boldsymbol{I}_{n_i}$.

2.5.2.3 Results

When the outcome was diastolic blood pressure, there were 4 models that had a C_p statistic less than or equal to the number of parameters including the intercept, i.e., $p + 1$. Figure 2.5 is a graph of the C_p of these 4 models vs. the number of parameters. Table 2.6 lists the models with the lowest C_p statistic for p number of parameters. Table 2.7 lists the models with the 3 lowest AIC and BIC values. Figure 2.6 plots C_p vs. $p + 1$ for these models. Table 2.8 lists the models and parameter estimates of the 3 models that had the lowest C_p values.

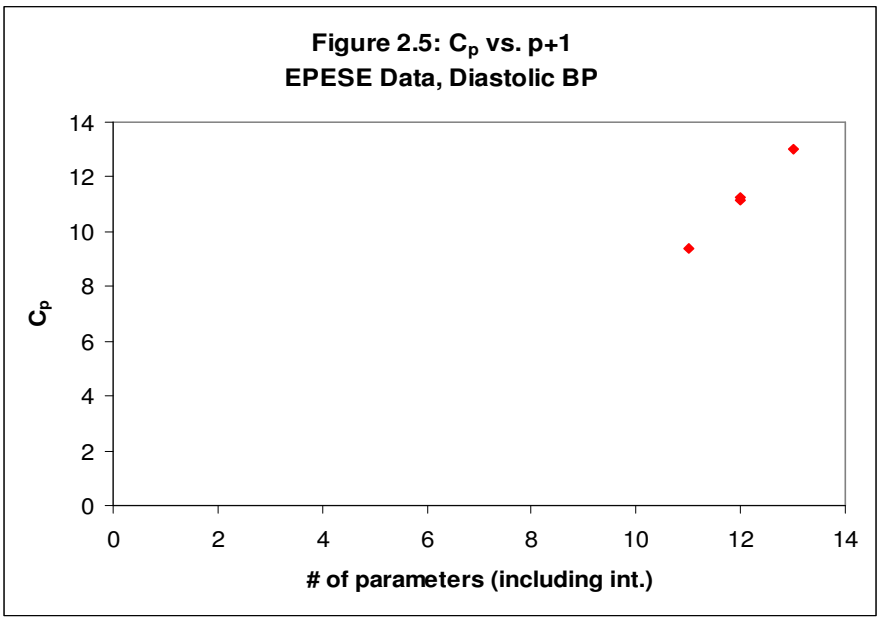


Table 2.6: C_p Results for Elderly Diastolic Blood Pressure Data
Models with the lowest C_p statistic for p number of covariates

p	Fixed Effects	F_p	C_p
1	year	89.60	976.59
2	year, weight	83.23	825.28
3	year, fair_ill, poor_ill	53.52	476.72
4	year, fair_ill, poor_ill, heart	42.72	338.77
5	year, weight, fair_ill, poor_ill, heart	29.86	208.06
6	year, weight, fair_ill, poor_ill, heart, diabet	19.28	116.70
7	year, weight, fair_ill, poor_ill, heart, diabet, blackpat	7.65	41.25
8	year, weight, fair_ill, poor_ill, heart, diabet, blackpat, rural	4.26	22.04
9	year, weight, fair_ill, poor_ill, heart, diabet, blackpat, rural, male	2.76	15.29
10	year, weight, fair_ill, poor_ill, heart, diabet, blackpat, rural, male, poor_hlth	0.20	9.40
11	year, weight, fair_ill, poor_ill, heart, diabet, blackpat, rural, male, poor_hlth, married	0.17	11.17
12	year, weight, fair_ill, poor_ill, heart, diabet, blackpat, rural, male, poor_hlth, married, fair_hlth	–	13.00

Table 2.7: AIC and BIC Results for Elderly Diastolic Blood Pressure Data
Models with the 3 lowest AIC and BIC values

Fixed Effects	AIC	BIC
year, weight, fair_ill, poor_ill, heart, diabet, blackpat, rural, male, poor_hlth, married	72397.35	72422.41
year, weight, fair_ill, poor_ill, heart, diabet, blackpat, rural, male, poor_hlth, married, fair_hlth	72398.00	72423.05
year, weight, fair_ill, poor_ill, heart, diabet, blackpat, rural, poor_hlth, married	72405.48	72430.54

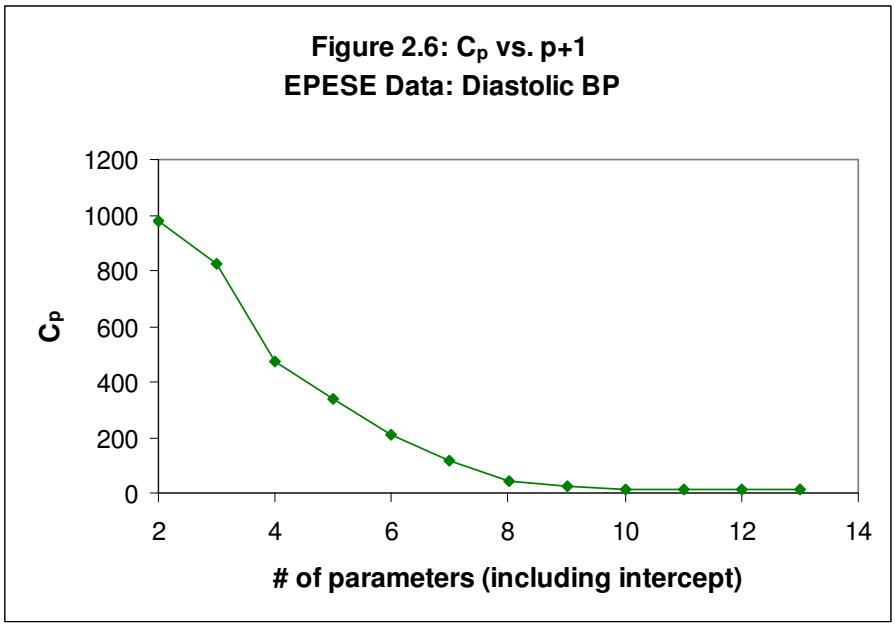


Table 2.8: Mixed Model Fixed Effect Estimates, Standard Errors (SE), p-values, and Covariance Estimates for the 3 Models with Lowest C_p Values (Outcome = Diastolic BP)

					Covariance Estimates	
C_p	Fixed Effect	Estimate	SE	p-value	Random effects	Error
9.40	Intercept	65.48	0.78	<0.001	$\widehat{\sigma}_1^2 = 47.11$	$\widehat{\sigma}_0^2 = 82.55$
	year	− 0.65	0.03	<0.001	$\widehat{\sigma}_2^2 = 0.26$	
	weight	0.04	0.004	<0.001		
	fair_ill	7.03	0.30	<0.001	$\widehat{\rho}_d = -0.16$	
	poor_ill	9.96	0.41	<0.001		
	heart	− 4.83	0.41	<0.001		
	diabet	− 3.95	0.38	<0.001		
	blackpat	2.65	0.31	<0.001		
	poor_hlth	− 1.17	0.38	0.002		
	rural	1.41	0.30	<0.001		
	male	0.96	0.34	0.004		
11.17	Intercept	65.61	0.78	<0.001	$\widehat{\sigma}_1^2 = 47.19$	$\widehat{\sigma}_0^2 = 82.46$
	year	− 0.66	0.03	<0.001	$\widehat{\sigma}_2^2 = 0.26$	
	weight	0.04	0.005	<0.001		
	fair_ill	6.98	0.30	<0.001	$\widehat{\rho}_d = -0.16$	
	poor_ill	9.94	0.41	<0.001		
	heart	− 4.87	0.40	<0.001		
	diabet	− 3.93	0.38	<0.001		
	blackpat	2.62	0.31	<0.001		
	poor_hlth	− 1.08	0.38	0.005		
	rural	1.42	0.30	<0.001		
	male	1.05	0.36	0.004		
	married	− 0.16	0.32	0.632		
11.23	Intercept	65.51	0.78	<0.001	$\widehat{\sigma}_1^2 = 47.08$	$\widehat{\sigma}_0^2 = 82.58$
	year	− 0.65	0.03	<0.001	$\widehat{\sigma}_2^2 = 0.26$	
	weight	0.04	0.004	<0.001		
	fair_ill	7.02	0.30	<0.001	$\widehat{\rho}_d = -0.16$	
	poor_ill	9.96	0.41	<0.001		
	heart	− 4.82	0.40	<0.001		
	diabet	− 3.94	0.38	<0.001		
	blackpat	2.66	0.31	<0.001		
	poor_hlth	− 1.22	0.40	<0.001		
	rural	1.42	0.30	<0.001		
	male	0.96	0.34	0.004		
	fair_hlth	− 0.11	0.26	0.663		

2.5.2.4 Discussion

From the results we can see that for the outcome of diastolic blood pressure, we have that the model with the smallest C_p statistic was the model whose fixed effects were all significant. The second best model, according to the C_p statistic, adds "married" as a fixed effect, but it is not significant and the same occurs in the third best model. Therefore, it can be said that the C_p statistic successfully chose the most parsimonious set of fixed effects for this outcome variable.

From the results using the information criteria, we see that both the AIC and the BIC selected the model that was selected as "second-best" by the C_p. When we look at the estimates and p-values for this model, we see that the "married" covariate is not significant at the α=0.05 level, and therefore should be excluded. Therefore the AIC and BIC's "best model" included an extraneous covariate.

2.5.3 Conclusions

From the results of both real data applications of the proposed C_p statistic, we can see that this statistic may be valuable as a model selection tool for mean structure in the linear mixed model. In both real data examples, it appears that using the lowest C_p statistic as a gauge for the best model resulted in finding the model which was most parsimonious in describing the amount of variation in the data due to the fixed effects and which had the most significant fixed effects. On the other hand, we see that the AIC and BIC agreed with the proposed C_p statistic only when using the dental data set.

2.6 Discussion

Using empirical studies, we find that our C_p statistic tends to do a good job of selecting fixed effects in most cases. It significantly outperforms the AIC and BIC in the simulation study for multiple categorical predictor setting, and in almost all cases, selects the correct set of fixed effects with at least 50% accuracy.

Looking at the results from the Potthoff and Roy data, we can see that our C_p statistic agrees with the AIC and BIC and selects the model with the interaction term. We also find that this interaction term is significant, and therefore the saturated model is best in this case.

This statistic, from both example data sets used here, appears to accurately identify the best set of fixed effects available to describe the data provided. In both the small data set with complete data and the large data set with missing data, the model with the smallest C_p statistic was also the model which had fixed effects that were all significant at the level of α=0.05. Going to the model with the next smallest C_p statistic would cause the user to either omit a statistically important covariate or to add a covariate that was not statistically significant.

In comparison to the AIC and BIC, the C_p statistic appears to work as well or better than the information criteria in most cases. In addition, the information criteria are more computationally intensive in that they require that every candidate model be fit in order to determine which one is best. For our C_p statistic, we only need to fit the saturated model, and look at different tests of hypotheses about the fixed effects in order to calculate the statistic for each set of fixed effects in our candidate models. The ease of these calculations in combination with the improved performance leads us to conclude that this statistic can be a valuable tool in the linear mixed model.

We note that this investigation was purely an empirical study of this new C_p statistic. The distributional properties of this statistic have yet to be explored.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**, 716.

Blazer, D.G., and George, L.K. (2004). Established Populations for Epidemiologic Studies of the Elderly, 1996-1997: Piedmont Health Survey of the Elderly, Fourth In-Person Survey [Durham, Warren, Vance, Granville, and Franklin Counties, North Carolina] [Computer file]. ICPSR version. Bethesda, MD: United States Department of Health and Human Services. National Institutes of Health. National Institute on Aging [producer], 1999. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004.

Cantoni, E., Flemming, J. M., & Ronchetti, E. (2005). Variable selection for marginal longitudinal generalized linear models. *Biometrics*, **61**(2), 507-514.

Demidenko E. (2004) Mixed Models—Theory and Application. Wiley: New York.

Fitzmaurice G, Davidian M, Molenberghs G, Verbeke G. (eds) (2008). Longitudinal Data Analysis: A Handbook of Modern Statistical Methods. Chapman & Hall/CRC: Boca Raton, Florida.

Gorman, J. W., & Toman, R. J. (1966). Selection of variables for fitting equations to data. Technometrics, 8(1), 27-51.

Hocking, R. R. (1976). A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, **32**(1), 1-49.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983-997.

Laird, NM, and Ware, JH. (1982). Random-effects models for longitudinal data. *Biometrics*; **38**: 963-974.

Mallows, C.L. (1964). "Choosing Variables in a Linear Regression: A Graphical Aid," unpublished paper presented at the Central Regional Meeting of the Institute of Mathematical Statistics. Manhattan, KS, May 7-9.

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

Potthoff, R. F., Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313-326.

Ronchetti, E., & Staudte, R. G. (1994). A robust version of Mallows's C_p. *Journal of the American Statistical Association*, **89**, 550-559.

SAS Institute Inc. (2007), SAS (release 9.1.3), Cary, NC: SAS Institute Inc.

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

CHAPTER 3

SELECTING THE BEST LINEAR MIXED MODEL USING THE KIC

3.1 Introduction

In general, a selection criterion scores every fitted model in a candidate class by how effectively the model conforms to the data based on its size. Ideally, unwanted scores will be assigned not only to models that omit essential variables, but also to models that adequately accommodate the data yet involve extraneous or irrelevant variables (Cavanaugh, 2004). In other words, the ideal selection criterion will select the model with the most parsimonious set of variables to describe the data available.

Information theoretic criteria have played a prominent role in model selection and is probably the most active area of current research in model selection for the linear mixed model. These criteria are defined as an estimate of the measure of fit of a model to the data. The most common criteria used in mixed models are the Akaike Information Criterion (AIC, Akaike 1974) and the Bayesian Information Criterion (BIC, Schwarz 1978) which are both directed divergences. In 1999, Cavanaugh defined the nondirectional divergence criterion, the KIC. However, no investigation of the KIC has been done in the linear mixed model.

In the linear univariate model, the sample squared multiple correlation coefficient, R^2, measures the maximum overall linear association of a single dependent variable with several independent variables. In the univariate model, R^2 corresponds to comparing two

models (Muller and Fetterman, 2002, Chapter 6, Sections 6.9–6.11): 1. a *full* model that consists of p independent predictors and an intercept; 2. a null model that has only the intercept. It also estimates the proportionate reduction of total variation in the dependent variable associated with the set of p independent variables. Most linear regression and ANOVA software packages provide the model (overall) R^2. However, until recently, little attention has been given to developing an R^2 statistic for the linear mixed model.

With N independent sampling units (often *persons* in practice), the linear mixed model for person i may be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{d}_i + \boldsymbol{e}_i\,.$$

Here, $\boldsymbol{y}_i$ is a $n_i \times 1$ vector of observations on person i; $\boldsymbol{X}_i$ is a $n_i \times q$ known, constant design matrix for person i, with full column rank q while $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown, constant, population parameters. Also $\boldsymbol{Z}_i$ is a $n_i \times m$ known, constant design matrix with rank m for person i corresponding to the $m \times 1$ vector of unknown random effects $\boldsymbol{d}_i$, while $\boldsymbol{e}_i$ is a $p_i \times 1$ vector of unknown random errors. Gaussian $\boldsymbol{d}_i$ and $\boldsymbol{e}_i$ are independent with mean $\mathbf{0}$ and

$$\mathcal{V}\left(\begin{bmatrix}\boldsymbol{d}_i \\ \boldsymbol{e}_i\end{bmatrix}\right) = \begin{bmatrix}\boldsymbol{\Sigma}_{di}(\boldsymbol{\tau}_d) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e)\end{bmatrix}.$$

Here $\boldsymbol{\mathcal{V}}(\,\cdot\,)$ is the covariance operator, while both $\boldsymbol{\Sigma}_{di}(\boldsymbol{\tau}_d)$ and $\boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e)$ are positive-definite, symmetric covariance matrices. Therefore $\boldsymbol{\mathcal{V}}(\boldsymbol{y}_i)$ may be written $\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{\Sigma}_{di}(\boldsymbol{\tau}_d)\boldsymbol{Z}_i' + \boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e)$. We assume that $\boldsymbol{\Sigma}_i$ can be characterized by a finite set of parameters represented by an $r \times 1$ vector $\boldsymbol{\tau}$ which consists of the unique parameters in $\boldsymbol{\tau}_d$ and $\boldsymbol{\tau}_e$. Throughout $n = \sum_{i=1}^{N} n_i$.

Edwards et al. (2008) have derived an R^2 statistic for the fixed effects in the linear mixed model. The proposed R^2 statistic is interpreted as a multivariate measure of association of the response $\boldsymbol{y}$ and the fixed effects given by $\boldsymbol{X}$. While this paper can not be considered a formal examination of this statistic, we have included its calculation in

the simulation studies as a method of observing its behavior for further, more in-depth studies.

3.1.1 Directed Divergences: AIC and BIC

In 1951, Kullback and Leibler defined Kullback's directed divergence as a measure of the disparity between the generating or true model $f(Y|\theta_0)$ and an approximating model $f(Y|\theta_k)$. For two arbitrary parametric densities $f(Y|\theta)$ and $f(Y|\theta^*)$, Kullback's directed divergence between $f(Y|\theta)$ and $f(Y|\theta^*)$ is defined as:

$$I(\theta, \theta^*) = E_\theta \left\{ \log \frac{f(Y|\theta)}{f(Y|\theta^*)} \right\} \tag{3.1}$$

where E_θ denotes the expectation under $f(Y|\theta)$ (Kullback and Leibler, 1951). Therefore, $I(\theta_0, \theta_k) = E_\theta \left\{ \log \frac{f(Y|\theta_0)}{f(Y|\theta_k)} \right\}$ defines the directed divergence between $f(Y|\theta_0)$ and $f(Y|\theta_k)$ with respect to $f(Y|\theta_0)$.

For $f(Y|\theta)$ and $f(Y|\theta^*)$, define

$$d(\theta, \theta^*) = E_\theta \{ -2\log f(Y|\theta^*) \} \tag{3.2}$$

Therefore,

$$2I(\theta_0, \theta_k) = d(\theta_0, \theta_k) - d(\theta_0, \theta_0) \tag{3.3}$$

For the purpose of discriminating between various candidate models, we can now say,

$$d(\theta_0, \theta_k) = E_{\theta_0} \{ -2\log f(Y|\theta_k) \} \tag{3.4}$$

serves as a valid substitute for $I(\theta_0, \theta_k)$.

This suggests that

$$d(\theta_0, \hat{\theta}_k) = E_{\theta_0} \{ -2\log f(Y|\theta_k) \}|_{\theta_k = \widehat{\theta}_k} \tag{3.5}$$

would provide a suitable measure of the separation between the generating model $f(Y|\theta_0)$ and a fitted candidate model $f(Y|\widehat{\theta}_k)$. However, we cannot evaluate $d(\theta_0, \theta_k)$ because θ_0 is unknown.

In 1974, Akaike suggested that $-2\log f(Y|\theta_k)$ serves as an unbiased estimator of $d(\theta_0, \hat{\theta}_k)$ and that the bias adjustment

$$E_{\theta_0}\Big\{d(\theta_0, \hat{\theta}_k)\Big\} - E_{\theta_0}\Big\{-2\log f(Y|\widehat{\theta}_k)\Big\} \tag{3.6}$$

can often be asymptotically estimated by twice the dimension of $\hat{\theta}_k$. Thus, since k denotes the dimension of $\hat{\theta}_k$, under appropriate conditions, the expected value of

$$AIC = -2\log f(Y|\theta_k) + 2k \tag{3.7}$$

should asymptotically approach the expected value of

$$d(\theta_0, \hat{\theta}_k) = E_{\theta_0}\{-2\log f(Y|\theta_k)\}|_{\theta_k=\widehat{\theta}_k} \tag{3.8}$$

and is therefore asymptotically unbiased.

In 1989, Hurvich and Tsai developed the corrected AIC (AIC_c), as a correction of use, in particular, when the sample size is small. The formula for this statistic is as follows:

$$AIC_c = -2\log f(Y|\theta_k) + \frac{2n(k+1)}{n-k-2} \tag{3.9}$$

Note: In this paper we are focused primarily on large sample situations, therefore, we did not include the AIC_c in our investigation.

In 1978, Schwarz presented a Bayesian alternative to the AIC. In a model of given dimension, ML estimators can be obtained as large sample limits of the Bayes estimators for arbitrary nowhere vanishing a priori distributions. Therefore, by studying the asymptotic behavior of Bayes estimators under a special class of priors, Schwarz arrived at the procedure where you choose the model for which $\log f(Y|\widehat{\theta}_k) - \frac{1}{2}k\log n$ is

largest.

Thus

$$BIC = \log f(Y|\widehat{\theta}_k) - \frac{1}{2}k\log n \tag{3.10}$$

The BIC, though technically not a directed divergence, is treated as another directed divergence.

Both the AIC and BIC are essentially log-likelihood values with a penalty (or adjustment) for the number of parameters estimated. Since the BIC differs from Akaike's only in that the dimension is multiplied by $\frac{1}{2}\log n$, the BIC leans more than Akaike's toward lower-dimensional models.

3.1.2 Symmetric Divergence: KIC

Kullback's symmetric divergence is defined as:

$$J(\theta_0, \theta_k) = I(\theta_0, \theta_k) + I(\theta_k, \theta_0) = E_{\theta_0}\left\{\log\frac{f(Y|\theta_0)}{f(Y|\theta_k)}\right\} + E_{\theta_k}\left\{\log\frac{f(Y|\theta_k)}{f(Y|\theta_0)}\right\} \tag{3.11}$$

This divergence is symmetric in that $J(\theta_0, \theta_k) = J(\theta_k, \theta_0)$. This symmetric divergence measures the average combined measure of fit of a sample Y generated under the generating or true model $f(Y|\theta_0)$ and a sample Z generated under the candidate model $f(Z|\theta_k)$. Using arguments similar to Akaike and assuming nested models, Cavanaugh (1999) proposed the following large sample model selection criterion:

$$KIC = -2\log f(Y|\widehat{\theta}_k) + 3k \tag{3.12}$$

This criterion serves as an asymptotically unbiased estimator of a variant of the symmetric divergence between the true model and a fitted approximating model.

3.1.3 Previous Model Selection Studies Using Information Criteria

Cavanaugh (1999) looked at the performance of the KIC in comparison to a wide spectrum of information criteria, including the AIC, the corrected AIC (AIC_c) and the BIC, in a univariate autoregressive (AR) process of order p, in a setting in which the criteria are used to select p. This univariate AR process of order p is defined as follows:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim iid\, N\left(0, \sigma^2\right).$$

where $1 \leq p \leq P$ in P candidate families with a set of observations $Y = \{y_1, y_2, \ldots, y_n\}$ from such a process. Also, note that $\widehat{\theta}_k$=$(\sigma^2, \phi_1, \phi_2, \ldots, \phi_p)^{'}$, $k = p + 1$ and $t = p + 1, ..., n$.

Cavanaugh found that, for autoregressive modeling BIC is consistent whereas the AIC and AIC_c are asymptotically efficient. The KIC is also asymptotically efficient for a broad class of generating models. However, the AIC, AIC_c are asymptotically efficient in an even broader class.

For Gaussian white noise processes of mean 0 and variance 1 with sample size of 40 or 60 and maximum model of order p = 8, KIC obtains substantially more correct order selections than any of the asymptotically efficient criteria. KIC is generally outperformed by BIC, however, KIC does not exhibit as strong a tendency as BIC to choose underparameterized models.

In conclusion, Cavanaugh found that the results from the simulation study suggested that KIC should function as an effective model selection criteria in large-sample applications. The results also suggested that the symmetric divergence may provide a foundation for the development of model selection criteria which is preferable to that provided by the directed divergence.

Gomez et al. (2005) looked at the performance of the Kenward-Roger method in the linear mixed model when the covariance structure is selected using AIC and BIC. They investigated Type I error rates for tests of fixed effects in mixed linear models using

Wald F-statistics with the Kenward-Roger adjustment, and he generated data using 15 covariance structures. Correct covariance structures as well as those selected using AIC and BIC were examined. Type I error rates for the correct models were often adequate depending on the sample size and complexity of covariance structure. Type I error rates for the best AIC and BIC models were always higher than target values, but those obtained using BIC were closer to the target value than those obtained using AIC. For unbalanced data , Type I error rates for the between-subjects effects were closer for negative pairing. Success of AIC and BIC in selecting the correct covariance structure was low.

The success rate (proportion of times AIC or BIC chose the correct covariance structure) depended greatly on the sample size and covariance structure. The highest success rate was 73.91%, for the largest sample size and a simple covariance structure (ARRE). Success rates were higher for larger sample sizes and simpler covariance structures. AIC had a higher success rate than BIC for complicated structures, especially for those with heterogeneity between treatments. BIC had higher success than AIC for simpler structures.

3.1.4 Investigation of the KIC and R^2 in This Chapter

As information criteria can be used as a model selection method in the linear mixed model for selection of the correct covariance structure and the correct fixed effect structure, in our simulation studies, we investigated the abilities of the KIC for both of these situations. First, we looked at the ability of the KIC to select the correct covariance structure when the fixed effect structure was correctly identified and different parameters of the random effects and fixed effects were varied. Secondly, we looked at the KIC's ability to select the correct fixed effect structure when the covariance structure and random effects structure was correctly identified. In both cases, we compared the ability of the KIC to that of the AIC and BIC as provided in the standard SAS output. We also

looked at the average value of the R^2 statistic as defined by Edwards, et al. (2008) to see if there was any correlation between the behaviors of this statistic and the various information criteria.

In this paper, section 3.2 looks at the results of the simulation study examining the selection of the correct covariance structure given the correct mean structure is identified, section 3.3 looks at the results examining the correct selection of the mean structure when the covariance structure is identified, section 3.4 applies the KIC to the EPESE data example that was introduced in Chapter 2 and section 3.4 discusses the results from both types of simulation studies and the data example.

3.2 Simulation Study: Covariance Structure Selection

3.2.1 Introduction

This simulation study looked at the ability of the KIC to select the correct covariance structure when the correct fixed effects structure was specified. The performance of the KIC was compared to that of the AIC and BIC. We considered only the large sample performance with complete data in all our scenarios, and used Kenward-Roger F and associated denominator degrees of freedom and REML likelihood estimation.

3.2.2 Methods

Data were simulated from a true linear mixed model consisting of the following fixed effects: an intercept, a dummy variable indicating membership in one of two groups, and a continuous covariate. To look at the impact of a varying R^2 statistic, we looked at two different fixed effects structures: one where the true linear mixed model has $\boldsymbol{\beta} = (1,1,1)'$ corresponding to an intercept, group and slope and one where $\boldsymbol{\beta} = (1,1,2)'$ which will result in a higher value for R^2, due to the stronger signal. Two different

covariance structures were simulated for the within-unit error: an independent and identically distributed (i.i.d.) structure and an autoregressive structure of order 1 (i.e. AR(1)).

An i.i.d. covariance matrix is represented as:

$$\boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e) = \sigma_0^2 \boldsymbol{I}_{n_i},$$

where $\boldsymbol{I}_{n_i}$ is an $n_i \times n_i$ identity matrix. An AR(1) structure is often used when it is expected that the correlation between observations "tails off" as the observations grow further apart in time. Formulaically, this is represented as:

$$\boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e) = \sigma_0^2 \begin{pmatrix} 1 & \rho_R & \rho_R^2 & \cdots & \rho_R^{n_i-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_R^{n_i-1} & \rho_R^{n_i-2} & \cdots & \rho_R & 1 \end{pmatrix}.$$

Random effect structures took three different forms: (1) no random effect; (2) random intercept only; and (3) a random intercept and a random slope. When there are no random effects and the within-unit error is independent and identically distributed (i.i.d.), the structure reduces to a univariate linear model with variance σ_0^2. When there is a random intercept only, the variance of the random effects, i.e. $\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d)$, is given by σ_1^2. When there is a random intercept and random slope, the variance of the random effects is given by:

$$\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d) = \begin{pmatrix} \sigma_1^2 & \rho_d \sigma_1 \sigma_2 \\ \rho_d \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}$$

The random intercepts, random slopes (where applicable) and within-unit error terms were generated as independent normal random variables with means zero and variances σ_1^2, σ_2^2 and σ_0^2, respectively. Datasets were simulated using varying values of variance: for random effect structure (1) $\sigma_0^2 = 1$ and 4; for (2) $\sigma_1^2 = \sigma_0^2 = 1$ and 4; and for (3) $\sigma_1^2 = \sigma_2^2 = \sigma_0^2 = 1$ and 4. To look at the impact of varying parameters of the AR(1) structure on the performance of the criteria, data were simulated using $\rho_{\mathrm{R}} = 0.10$,

0.25, 0.50 and 0.75 for the within-subject correlation. When a random intercept and slope were present, we looked at cases where there is small or no correlation between these effects, $\rho_d = 0$, 0.1 and 0.2. Large sample performance was assessed using simulated datasets which consisted of 200 subjects and 5 observations each in the interval [0,1] and involved 1,000 realizations.

In order to assess the performance of the criteria in choosing the true covariance structure, eight candidate models were fit for each generated dataset, and the number of times the criteria chose the correct model from this set of 1,000 was tallied. The set of candidate models consisted of eight models each having the same correct set of fixed effects but random effects covariance structures corresponding to: (1) no random effects; (2) a random intercept only; (3) a random intercept and a random slope with unstructured covariance; and (4) a random intercept and a random slope with variance components (VC) covariance. Within-unit error covariance structures corresponded to: (1) independent and identically distributed (i.i.d.); and (2) autoregressive structure of order 1 (i.e. AR(1)). The number of times out of the 1,000 possibilities that the criterion in question chose the correct covariance model (based on the model by which the data were simulated) as the best model was recorded. All simulations in both scenarios were run using REML estimation, Kenward-Roger$F$, and associated denominator degrees of freedom (ddf).

3.2.3 Results

Tables 3.1, 3.2 and 3.3 display the results of the simulations for iid models with no random effects, a random intercept only, and both a random intercept and random slope, respectively. Figure 3.1 and 3.2 graphically illustrate the ability of the information criteria to select the correct model for varying values of R^2 when the true within-unit error covariance structure was i.i.d. and the model had, respectively, no random effects or a random intercept alone.

Table 3.1: Monte Carlo Assessment of Covariance Model Selection:
iid Within-Unit Error Covariance, No Random Effects
1,000 datasets, 200 subjects each, 5 observations per subject

	Percentage Correct Model Selection					
	$\beta_1 = 1$			$\beta_1 = 2$		
$\sigma_0^2 =$	**0.5**	**1**	**2**	**0.5**	**1**	**2**
Mean R^2	0.494	0.346	0.226	0.602	0.430	0.274
KIC	88.1	87.4	88.1	88.5	89.4	89.2
AIC	75.6	76.4	75.6	76.1	80.4	79.0
BIC	96.6	97.0	96.6	96.2	97.4	96.9

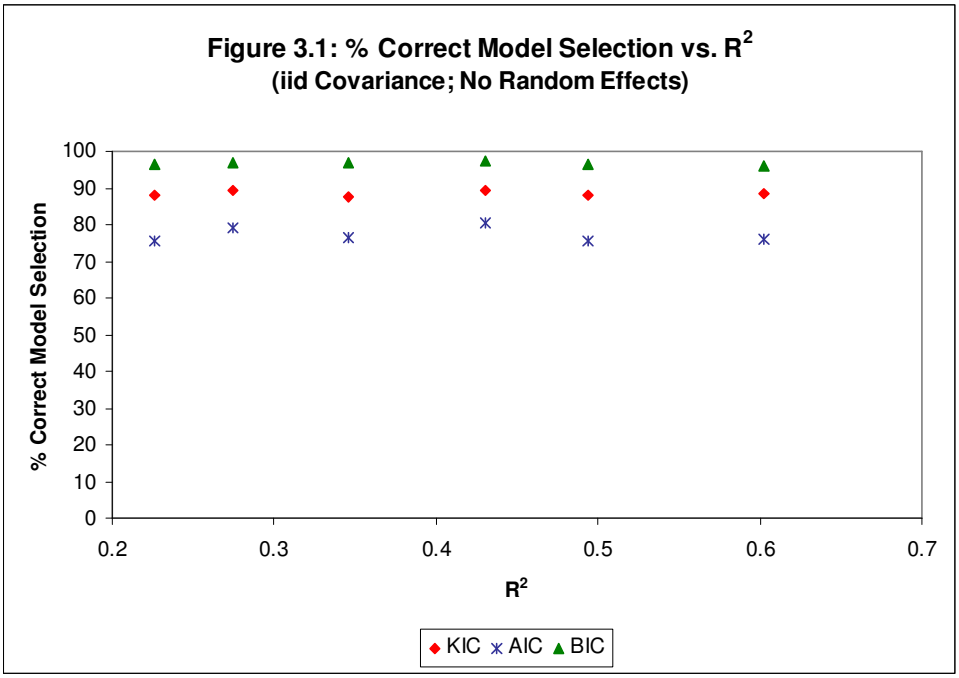


Table 3.2: Monte Carlo Assessment of Covariance Model Selection:
iid Within-Unit Error Covariance, Random Intercept Only
1,000 datasets, 200 subjects each, 5 observations per subject

	Percentage Correct Model Selection					
	$\beta_1 = 1$			$\beta_1 = 2$		
$\sigma_0^2, \sigma_1^2 =$	**0.5**	**1**	**2**	**0.5**	**1**	**2**
Mean R^2	0.251	0.144	0.079	0.520	0.352	0.215
KIC	86.9	87.5	89.4	87.0	87.9	87.0
AIC	73.5	76.8	77.6	74.6	80.1	73.6
BIC	96.4	96.2	97.8	97.3	96.9	96.3

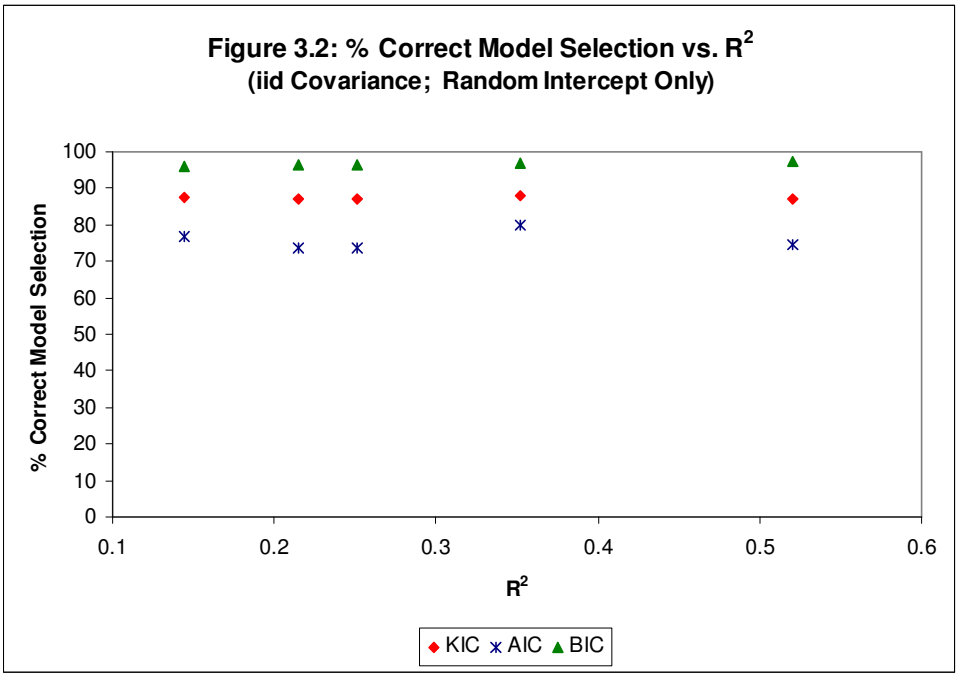


Table 3.3: Monte Carlo Assessment of Covariance Model Selection: iid Within-Unit Error Covariance, Random Intercept and Random Slope 1,000 datasets, 200 subjects each, 5 observations per subject

	Percentage Correct Model Selection					
	$\beta_1 = 1; \rho_d = 0$			$\beta_1 = 2; \rho_d = 0$		
$\sigma_0^2, \sigma_1^2, \sigma_2^2 =$	**0.5**	**1**	**2**	**0.5**	**1**	**2**
Mean R^2	0.251	0.144	0.079	0.520	0.352	0.215
KIC	86.9	87.5	89.4	87.0	87.9	87.0
AIC	73.5	76.8	77.6	74.6	80.1	73.6
BIC	96.4	96.2	97.8	97.3	96.9	96.3
	$\beta_1 = 1; \rho_d = 0.1$			$\beta_1 = 2; \rho_d = 0.1$		
$\sigma_0^2, \sigma_1^2, \sigma_2^2 =$	**0.5**	**1**	**2**	**0.5**	**1**	**2**
Mean R^2	0.186	0.104	0.055	0.408	0.257	0.149
KIC	92.1	92.2	92.0	91.0	90.3	92.6
AIC	85.6	85.8	85.5	82.6	82.0	83.3
BIC	93.6	98.4	97.3	97.0	97.1	97.6
	$\beta_1 = 1; \rho_d = 0.2$			$\beta_1 = 2; \rho_d = 0.2$		
$\sigma_0^2, \sigma_1^2, \sigma_2^2 =$	**0.5**	**1**	**2**	**0.5**	**1**	**2**
Mean R^2	0.262	0.102	0.055	0.407	0.339	0.149
KIC	90.6	89.6	90.4	89.7	92.4	92.4
AIC	83.1	83.8	83.7	81.2	84.3	84.3
BIC	98.0	96.8	97.0	95.9	98.0	97.7

Tables 3.4, 3.5, 3.6 and 3.7 display the results of the simulations for AR(1) models with no random effects, a random intercept only, and both a random intercept and

random slope, respectively, Tables 3.6 and 3.7 differ in that the variance parameters in Table 3.7 is twice that of the variance parameters in Table 3.6. Varying parameters for the AR(1) structure are described by the within-unit correlation ρ_R . Figure 3.3 graphically illustrates the data for an AR(1) covariance structure with random intercept and slope when $\beta_1 = 2; \rho_d = 0.2$ and σ_0^2, σ_1^2, $\sigma_2^2 = 2$ (i.e. the last quadrant of Table 3.7).

Table 3.4: Monte Carlo Assessment of Covariance Model Selection: AR(1) Within-Unit Error Covariance, No Random Effects 1,000 datasets, 200 subjects each, 5 observations per subject

	Percentage Correct Model Selection							
	$\beta_1 = 1; \sigma_0^2 = 1$				$\beta_1 = 2; \sigma_0^2 = 1$			
$\rho_R =$	**0.10**	**0.25**	**0.50**	**0.75**	**0.10**	**0.25**	**0.50**	**0.75**
Mean R^2	0.249	0.219	0.185	0.185	0.406	0.377	0.356	0.400
KIC	76.2	95.3	94.0	94.5	76.6	96.3	94.4	93.8
AIC	76.5	89.6	87.7	86.5	77.7	91.2	89.1	87.8
BIC	66.1	99.1	98.9	99.2	65.3	98.9	98.5	98.9
	$\beta_1 = 1; \sigma_0^2 = 2$				$\beta_1 = 2; \sigma_0^2 = 2$			
$\rho_R =$	**0.10**	**0.25**	**0.50**	**0.75**	**0.10**	**0.25**	**0.50**	**0.75**
Mean R^2	0.145	0.124	0.103	0.103	0.254	0.233	0.217	0.251
KIC	75.8	94.5	94.7	94.3	78.4	95.1	94.6	94.8
AIC	76.7	88.1	86.8	86.5	79.4	89.1	89.2	88.4
BIC	62.8	98.9	98.7	88.4	68.3	98.9	99.2	98.8

Table 3.5: Monte Carlo Assessment of Covariance Model Selection: AR(1) Within-Unit Error Covariance, Random Intercept Only 1,000 datasets, 200 subjects each, 5 observations per subject

	Percentage Correct Model Selection							
	$\beta_1 = 1; \sigma_0^2, \sigma_1^2 = 1$				$\beta_1 = 2; \sigma_0^2, \sigma_1^2 = 1$			
$\rho_R =$	**0.10**	**0.25**	**0.50**	**0.75**	**0.10**	**0.25**	**0.50**	**0.75**
Mean R^2	0.137	0.131	0.128	0.146	0.337	0.405	0.319	0.371
KIC	54.1	94.2	94.7	63.8	59.6	95.2	95.8	65.5
AIC	59.8	87.4	88.5	71.0	65.7	89.8	90.0	72.2
BIC	36.6	98.3	97.8	42.9	38.8	97.3	98.4	44.3
	$\beta_1 = 1; \sigma_0^2, \sigma_1^2 = 2$				$\beta_1 = 2; \sigma_0^2, \sigma_1^2 = 2$			
$\rho_R =$	**0.10**	**0.25**	**0.50**	**0.75**	**0.10**	**0.25**	**0.50**	**0.75**
Mean R^2	0.074	0.069	0.068	0.081	0.123	0.127	0.191	0.229
KIC	55.2	93.3	94.0	65.2	54.3	94.6	94.1	65.0
AIC	63.1	88.6	88.8	72.8	62.5	90.0	87.8	73.3
BIC	36.2	97.8	97.7	48.2	36.2	98.2	98.2	45.4

Table 3.6: Monte Carlo Assessment of Covariance Model Selection: AR(1) Within-Unit Error Covariance, Random Intercept and Random Slope 1,000 datasets, 200 subjects each, 5 observations per subject

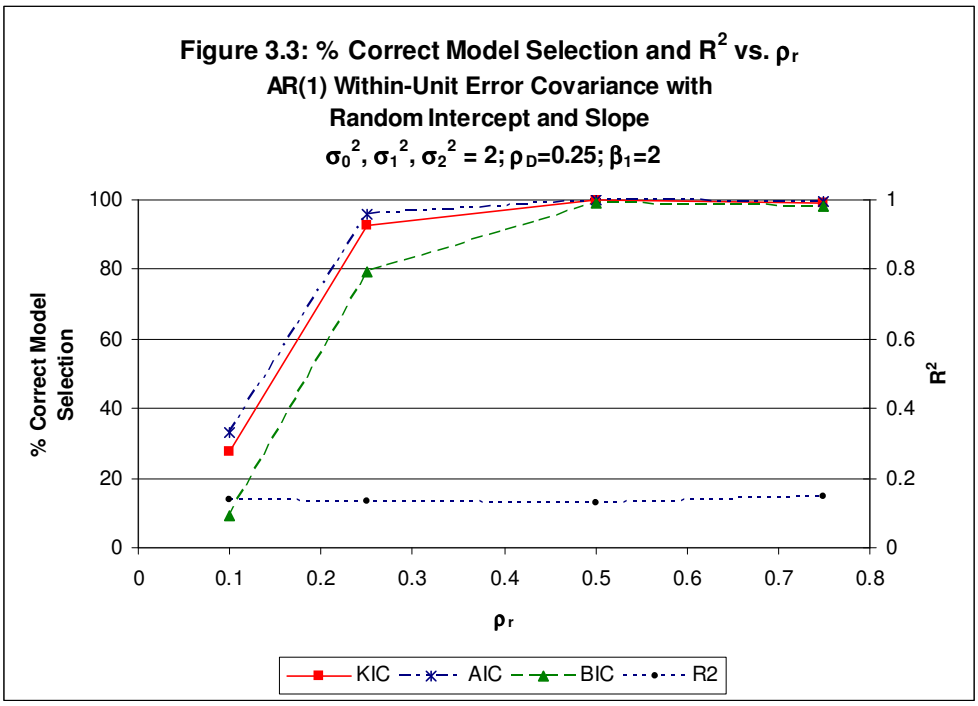
$\boldsymbol{\sigma_0^2}, \sigma_1^2, \sigma_2^2 = 1$

	Percentage Correct Model Selection							
	$\beta_1 = 1; \rho_d = 0$				$\beta_1 = 2; \rho_d = 0$			
$\rho_R =$	**0.10**	**0.25**	**0.50**	**0.75**	**0.10**	**0.25**	**0.50**	**0.75**
Mean R^2	0.100	0.096	0.093	0.098	0.251	0.240	0.240	0.256
KIC	44.8	93.9	95.7	93.4	42.4	93.9	96.3	91.6
AIC	54.5	96.6	98.0	95.0	54.0	88.1	98.7	93.1
BIC	22.6	84.9	82.2	83.1	21.5	89.9	83.9	83.1
	$\beta_1 = 1; \rho_d = 0.1$				$\beta_1 = 2; \rho_d = 0.1$			
$\rho_R =$	**0.10**	**0.25**	**0.50**	**0.75**	**0.10**	**0.25**	**0.50**	**0.75**
Mean R^2	0.098	0.094	0.093	0.098	0.244	0.236	0.237	0.256
KIC	33.9	95.3	98.8	95.9	34.6	95.7	98.7	97.1
AIC	48.3	97.0	99.8	96.3	47.3	97.6	99.5	97.8
BIC	17.4	87.3	92.5	91.0	14.2	87.1	92.0	92.6
	$\beta_1 = 1; \rho_d = 0.2$				$\beta_1 = 2; \rho_d = 0.2$			
$\rho_R =$	**0.10**	**0.25**	**0.50**	**0.75**	**0.10**	**0.25**	**0.50**	**0.75**
Mean R^2	0.095	0.092	0.090	0.097	0.239	0.232	0.229	0.257
KIC	25.6	94.0	99.8	98.8	25.9	92.5	99.8	99.6
AIC	40.9	97.1	99.9	98.8	43.8	96.3	100	99.7
BIC	8.8	79.7	99.5	97.1	8.3	78.7	98.1	97.9

Table 3.7: Monte Carlo Assessment of Covariance Model Selection: AR(1) Within-Unit Error Covariance, Random Intercept and Random Slope 1,000 datasets, 200 subjects each, 5 observations per subject

$\sigma_0^2, \sigma_1^2, \sigma_2^2 = 2$

	Percentage Correct Model Selection							
	$\beta_1 = 1; \rho_d = 0$				$\beta_1 = 2; \rho_d = 0$			
$\rho_R =$	**0.10**	**0.25**	**0.50**	**0.75**	**0.10**	**0.25**	**0.50**	**0.75**
Mean R^2	0.054	0.052	0.050	.053	0.143	0.137	0.136	0.147
KIC	42.9	94.6	95.2	92.3	41.9	94.8	96.5	92.2
AIC	55.3	97.2	98.7	93.9	54.2	96.5	98.7	94.2
BIC	21.5	87.3	80.7	83.7	22.9	86.7	83.4	81.6
	$\beta_1 = 1; \rho_d = 0.1$				$\beta_1 = 2; \rho_d = 0.1$			
$\rho_R =$	**0.10**	**0.25**	**0.50**	**0.75**	**0.10**	**0.25**	**0.50**	**0.75**
Mean R^2	0.052	0.051	0.049	0.052	0.141	0.135	0.133	0.147
KIC	36.2	93.8	98.1	92.0	34.6	94.5	98.8	97.8
AIC	48.7	96.4	99.6	96.8	48.2	96.9	99.8	98.5
BIC	14.7	84.7	92.5	91.6	16.7	86.0	92.8	93.9
	$\beta_1 = 1; \rho_d = 0.2$				$\beta_1 = 2; \rho_d = 0.2$			
$\rho_R =$	**0.10**	**0.25**	**0.50**	**0.75**	**0.10**	**0.25**	**0.50**	**0.75**
Mean R^2	0.051	0.048	0.048	0.052	0.138	0.133	0.131	0.148
KIC	26.7	92.5	99.5	99.2	27.5	92.6	99.8	99.3
AIC	41.1	96.0	99.9	99.5	33.3	95.8	99.8	99.5
BIC	8.5	79.1	97.9	98.5	9.0	79.3	99.2	98.0



3.2.4 Conclusions

From the results of this simulation study, we see that the BIC is the information criteria with the best performance when the correct within-unit error covariance structure is i.i.d and the random effects covariance varied. However, when the correct within-unit error covariance structure is AR(1) and the random effects covariance varied, the AIC is the best performer of the information criteria. In all scenarios, the KIC tends to hold the middle ground.

When the time effect was doubled in magnitude, the R^2 statistic, as expected, increased due to the increase in the signal-to-noise ratio. However, as can be seen in Figures 3.1 and 3.2, this increase did not significantly affect the ability of any of the information criteria as far as correct model selection was concerned. Also, as seen in Figure 3.3, while the change in ρ_R has a great impact on the performance of the information criteria when the within-unit error covariance structure is AR(1), this variation has little impact on the magnitude of the R^2 statistic. On the other hand, an

increase in the variance parameters, σ_0^2, σ_1^2 and σ_2^2 does result in a decrease in the R^2 statistic in both the i.i.d. and AR(1) scenarios.

While the KIC was not the best performer in any of the scenarios, its ability to hold the middle ground in all scenarios asserts that it may be the best criterion to use when one is truly unaware of what the underlying covariance structure may be.

3.3 Simulation Study: Mean Structure Selection

3.3.1 Introduction

This simulation study investigated the ability of the KIC to select the correct mean structure when the correct covariance structure was identified, and we compared its ability to that of the AIC and BIC. We also observed the behavior of the R^2 statistic in these scenarios. For both scenarios investigated, we used an i.i.d. covariance structure, REML estimation, and Kenward Roger F statistic and associated ddf for inference with complete data.

3.3.2 Methods

3.3.2.1 Methods: Scenario One

For the first mean structure selection scenario, a repeated measures study consisting of 3 measurement occasions was simulated that assumes compound symmetry of the data; that is, the data generated were based on a model with only a random intercept and an independent and identically distributed (i.i.d.) within-unit error term i.e., $\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d) = \sigma_1^2$ and

$$\boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e) = \sigma_0^2 \boldsymbol{I}_{n_i},$$

where $\boldsymbol{I}_{n_i}$ is an $n_i \times n_i$ identity matrix. Therefore, the compound symmetry structure was simulated such that $\sigma^2 = \sigma_1^2 + \sigma_0^2$ where σ^2 is the total variance, σ_1^2 is the variance of the

random intercept, and σ_0^2 is the variance of the within-unit error term. Datasets were simulated using varying values of the total variance, namely $\sigma^2 = 1$, 4 and 8, to get a sense of the impact of the true variance on the performance of the criteria. Corresponding to the standard model assumptions, the random intercepts and within-unit error terms were generated as independent normal random variables with means zero and variances σ_1^2 and σ_0^2, respectively. To understand the impact of the within-unit correlations on the performance of the criteria, data were simulated using correlations between observations $\rho = 0.25$, 0.50 and 0.75, where $\rho = \sigma_1^2/\sigma^2$.

As for the mean structure, data were simulated from a true linear mixed model consisting of the following fixed effects: an intercept, a dummy variable indicating membership in one of two groups, and a continuous covariate. Nine different variations of models for these fixed effects were considered to see how the information criteria performed in their ability to select the correct set of fixed effects. The fixed effects models simulated took on three variations: 1) model with intercept and time only, 2) model with intercept, time and group effect, and 3) model with intercept, time, group and interaction between group and time. In each variation, three values for the coefficient associated with the time effect, i.e. β_1, were considered: $\beta_1 = 0.5,\ 1,$ and 2. Thus, with three models for fixed effects and three values for β_1, we have nine different variations of models for the fixed effects. In addition, when the correct model included an interaction term, two values for the coefficient associated with the interaction, i.e. β_3, were considered: $\beta_3 = 0.25,\ 0.5$. Large sample performance was assessed; simulated datasets consisted of 200 subjects with 5 observations each in the interval [0,1]. Each simulation for the varying sample sizes, variances and correlation values consisted of 1,000 realizations.

In order to assess the performance of the criteria in choosing the proper set of fixed effects based on the simulation study, a set of candidate models was fit for each generated dataset, and the number of times the criteria chose the correct model from this

set of 1,000 was tallied. The set of candidate models consisted of three models each having the same covariance structure and fixed effects corresponding to: (1) a model with common intercept and common slope; (2) a model with common intercept, a slope, and an additional group covariate; and (3) a model with intercept, slope, group, and group x slope interaction. For example, if the true model was one with a common intercept and common slope, we assessed the performance of the criteria in choosing the true model as a candidate model or the other two candidate models. The number of times out of the 1,000 possibilities that the criterion in question chose the correct model as the best model and the average R^2 for the correct model was recorded.

3.3.2.2 Methods: Scenario Two

The multiple categorical predictors that are used to generate the data simulate the case where there are various factors (in addition to the time effect), such as race, gender etc., that have an impact on the outcome variable. The REML estimation, Kenward-Roger$F$, and associated ddf were used when evaluating the simulated data.

For this set of simulations, we used an i.i.d within-subject covariance structure i.e.,

$$\boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e) = \sigma_0^2 \boldsymbol{I}_5,$$

with a between-subject covariance structure that contained either a random intercept alone, or a random intercept and slope, i.e., $\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d) = \sigma_1^2$ or

$$\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d) = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix}$$

In this simulation, $\sigma_0^2 = 2$, $\sigma_1^2 = 2$, $\sigma_2^2 = 0$ or 2, and $\rho_d = 0.25$ when $\sigma_2^2 = 2$ ($\sigma_{12}^2 = \rho_d \sigma_1 \sigma_2$).

For the fixed effects, β_1 is the common intercept, β_2 is the coefficient for time and four different categorical predictors, x_1, x_2, x_3 and x_4, are considered with the corresponding coefficients β_3, β_4, β_5, and β_6. The vector for $\beta' = (3, 1, 4, 3, 5, 5)$, and 16 different variations are considered where 0,1, 2, 3, or 4 of the coefficients for the

categorical predictors (i.e. x_1, x_2, x_3 and x_4) are set to 0. Therefore, the saturated mean model, which includes all coefficients for the categorical predictors and time, is given by:

$$E(y) = 3 + t + 4x_1 + 3x_2 + 5x_3 + 5x_4$$

For each set of simulations, 16 different sets of fixed effects were assessed. We looked at the ability of the KIC compared to the AIC and BIC to select the correct fixed effect structure among the 16 possible outcomes over the 1,000 simulations, and also recorded the average R^2 statistic for the correct fixed effect model. All simulations were done using Kenward-Roger F, associated ddf, and REML likelihood.

3.3.3 Results

3.3.3.1 Results: Scenario One

Tables 3.8, 3.9, 3.10 and 3.11 display the results for the fixed effects model selection when the correct model included time alone, time and group alone, and time, group and their interaction. Tables 3.10 and 3.11 differ in that in Table 3.11 the interaction effect has double the magnitude of the interaction effect in Table 3.10. In all tables the variation of the variance, covariance and time effect parameters are illustrated. Figures 3.4, 3.5 and 3.6 are a graphic representation of the middle panels in Table 3.9 to illustrate the impact of the increase in the variance parameter on the R^2 and the performance of the information criteria. As the AIC and BIC have identical values in these tables, only the AIC is graphed (in addition to the KIC and R^2).

Table 3.8: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 200 subjects each, 5 observations per subject

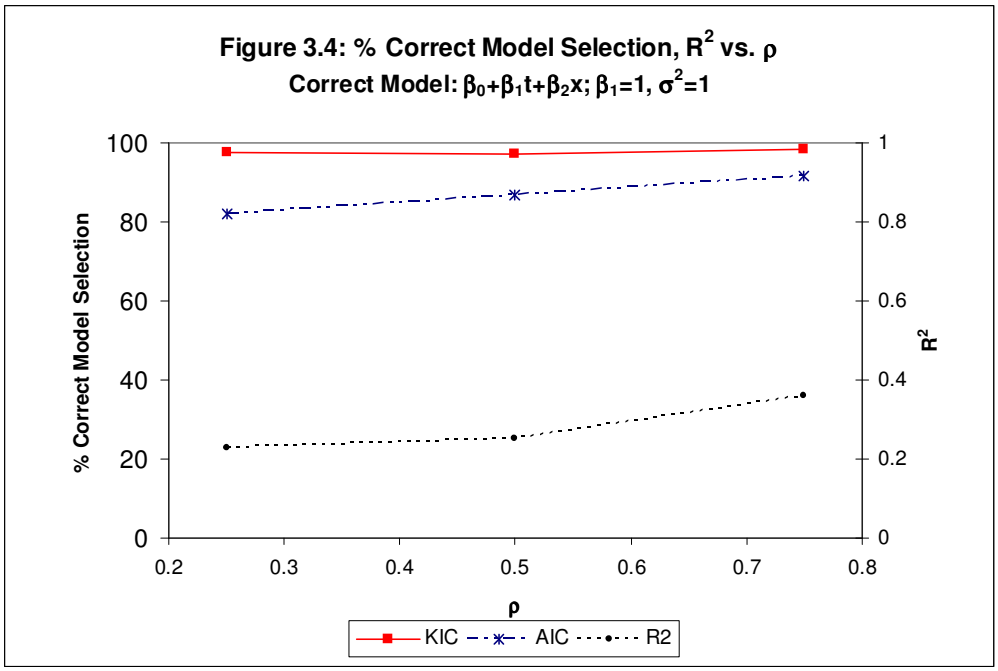
Correct Model: $\beta_0 + \beta_1 t$

	Percentage Correct Model Selection								
	$\sigma^2 = 1$, $\beta_1 = 0.5$			$\sigma^2 = 1$, $\beta_1 = 1$			$\sigma^2 = 1$, $\beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
Mean R^2	0.041	0.060	0.112	0.145	0.201	0.334	0.401	0.501	0.667
KIC	98.1	98.0	97.9	98.7	98.2	97.5	98.0	98.2	97.4
AIC	87.2	84.3	84.2	87.4	84.7	85.1	87.7	85.1	84.7
BIC	87.2	84.3	84.2	87.4	84.7	85.1	87.7	85.1	84.7
	$\sigma^2 = 4$, $\beta_1 = 0.5$			$\sigma^2 = 4$, $\beta_1 = 1$			$\sigma^2 = 4$, $\beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
Mean R^2	0.012	0.016	0.031	0.042	0.060	0.112	0.146	0.202	0.333
KIC	95.8	93.5	94.9	96.0	95.2	93.6	96.9	95.2	95.9
AIC	63.4	57.8	56.9	64.8	60.8	58.5	61.9	55.6	59.8
BIC	63.4	57.8	56.9	64.8	60.8	58.5	61.9	55.6	59.8
	$\sigma^2 = 8$, $\beta_1 = 0.5$			$\sigma^2 = 8$, $\beta_1 = 1$			$\sigma^2 = 8$, $\beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
Mean R^2	0.006	0.009	0.016	0.021	0.032	0.060	0.079	0.114	0.202
KIC	93.8	92.2	91.8	94.4	91.5	91.6	95.7	91.8	90.7
AIC	29.5	29.2	26.2	29.2	27.2	23.5	30.8	25.3	20.9
BIC	29.5	29.2	26.2	29.2	27.2	23.5	30.8	25.3	20.9

Table 3.9: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 200 subjects each, 5 observations per subject

Correct Model: $\beta_0 + \beta_1 t + \beta_2 x$

	Percentage Correct Model Selection								
	$\sigma^2 = 1, \beta_1 = 0.5$			$\sigma^2 = 1, \beta_1 = 1$			$\sigma^2 = 1, \beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
Mean R^2	0.144	0.130	0.160	0.227	0.251	0.361	0.442	0.522	0.673
KIC	96.9	98.4	98.6	97.7	97.4	98.4	96.4	97.8	97.6
AIC	82.7	86.8	90.9	82.2	86.8	91.8	83.3	87.2	90.9
BIC	82.7	86.8	90.9	82.2	86.8	91.8	83.3	87.2	90.9
	$\sigma^2 = 4, \beta_1 = 0.5$			$\sigma^2 = 4, \beta_1 = 1$			$\sigma^2 = 4, \beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
Mean R^2	0.042	0.037	0.047	0.070	0.078	0.124	0.167	0.214	0.342
KIC	92.9	93.6	95.2	94.6	94.5	93.5	93.1	94.9	94.4
AIC	53.7	65.6	77.0	51.3	67.5	77.6	52.2	64.8	78.8
BIC	53.7	65.6	77.0	51.3	67.5	77.6	52.2	64.8	78.8
	$\sigma^2 = 8, \beta_1 = 0.5$			$\sigma^2 = 8, \beta_1 = 1$			$\sigma^2 = 8, \beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
Mean R^2	0.022	0.020	0.025	0.037	0.042	0.068	0.091	0.121	0.208
KIC	87.4	85.4	78.8	88.6	82.7	80.7	88.7	85.6	80.0
AIC	0	35.7	66.0	0	34.6	65.9	0	34.7	64.7
BIC	0	35.7	66.0	0	34.6	65.9	0	34.7	64.7



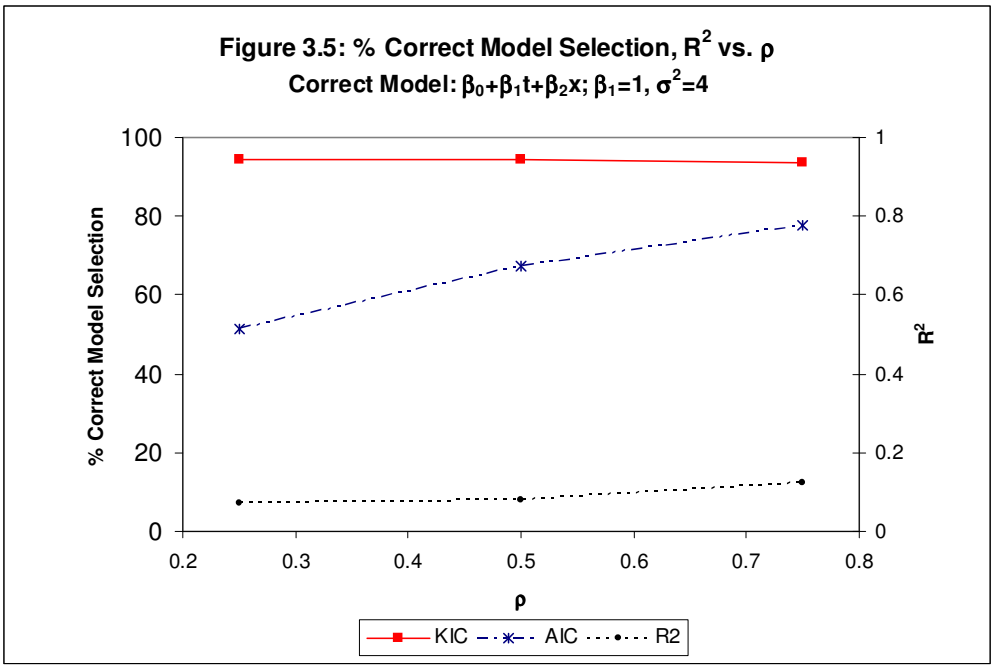


Figure 3.5: % Correct Model Selection, R^2 vs. ρ
Correct Model: $\beta_0+\beta_1 t+\beta_2 x$; $\beta_1=1$, $\sigma^2=4$
% Correct Model Selection
R^2
ρ
KIC
AIC
R2

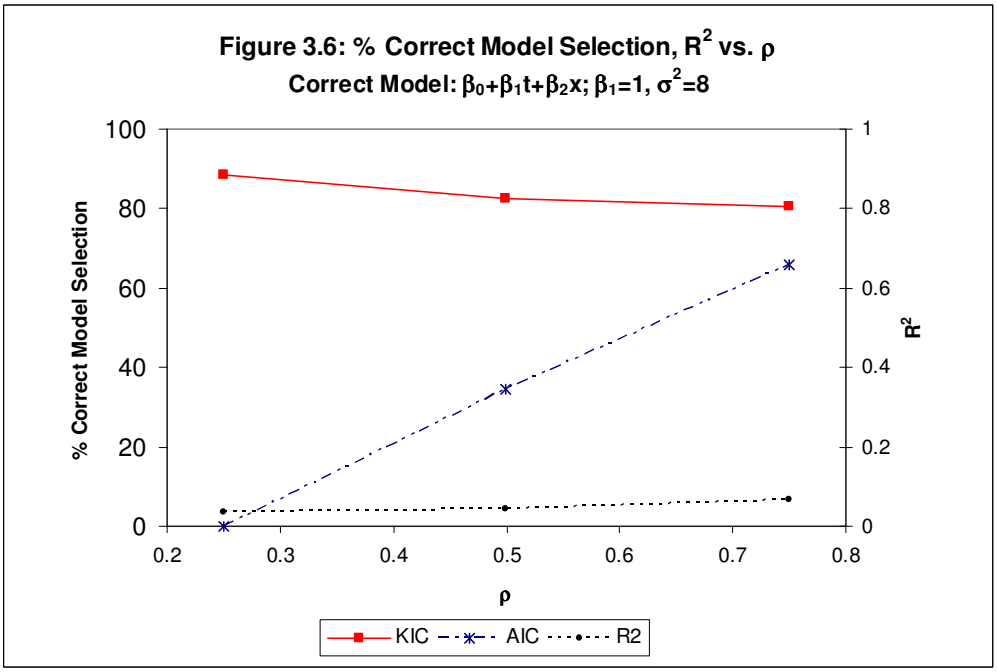


Figure 3.6: % Correct Model Selection, R^2 vs. ρ
Correct Model: $\beta_0+\beta_1 t+\beta_2 x$; $\beta_1=1$, $\sigma^2=8$
% Correct Model Selection
R^2
ρ
KIC
AIC
R2

Table 3.10: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 200 subjects each, 5 observations per subject

Correct Model: $\beta_0 + \beta_1 t + \beta_2 x + \beta_3 x*t$; β_3=0.25

	Percentage Correct Model Selection								
	$\sigma^2 = 1$, $\beta_1 = 0.5$			$\sigma^2 = 1$, $\beta_1 = 1$			$\sigma^2 = 1$, $\beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
Mean R^2	0.186	0.174	0.222	0.273	0.300	0.420	0.480	0.555	0.702
KIC	25.7	37.2	64.5	25.7	37.5	61.5	27.5	37.8	62.8
AIC	56.0	68.6	86.5	56.9	70.1	85.8	59.5	67.6	85.8
BIC	56.0	68.6	86.5	56.9	70.1	85.8	59.5	67.6	85.8
	$\sigma^2 = 4$, $\beta_1 = 0.5$			$\sigma^2 = 4$, $\beta_1 = 1$			$\sigma^2 = 4$, $\beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
Mean R^2	0.056	0.052	0.068	0.089	0.099	0.154	0.187	0.238	0.371
KIC	13.2	16.1	21.4	12.9	16.5	21.7	13.7	14.5	21.6
AIC	60.9	57.6	53.1	59.1	55.8	54.8	58.0	51.7	55.7
BIC	60.9	57.6	53.1	59.1	55.8	54.8	58.0	51.7	55.7
	$\sigma^2 = 8$, $\beta_1 = 0.5$			$\sigma^2 = 8$, $\beta_1 = 1$			$\sigma^2 = 8$, $\beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
Mean R^2	0.030	0.028	0.037	0.047	0.054	0.085	0.106	0.135	0.228
KIC	14.4	14.5	13.8	16.3	14.4	16.5	16.3	14.4	15.1
AIC	100	72.4	53.9	100	73.3	55.2	100	72.0	54.4
BIC	100	72.4	53.9	100	73.3	55.2	100	72.0	54.4

Table 3.11: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 200 subjects each, 5 observations per subject

Correct Model: $\beta_0 + \beta_1 t + \beta_2 x + \beta_3 x*t$; β_3=0.50

	Percentage Correct Model Selection								
	$\sigma^2 = 1$, $\beta_1 = 0.5$			$\sigma^2 = 1$, $\beta_1 = 1$			$\sigma^2 = 1$, $\beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
Mean R^2	0.233	0.227	0.291	0.319	0.351	0.478	0.512	0.586	0.728
KIC	82.7	95.4	100	85.9	94.5	99.8	84.8	95.9	99.9
AIC	96.8	99.3	100	96.3	99.3	99.9	96.4	99.4	100
BIC	96.8	99.3	100	96.3	99.3	99.9	96.4	99.4	100
	$\sigma^2 = 4$, $\beta_1 = 0.5$			$\sigma^2 = 4$, $\beta_1 = 1$			$\sigma^2 = 4$, $\beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
Mean R^2	0.073	0.071	0.096	0.107	0.121	0.187	0.211	0.263	0.401
KIC	39.2	49.3	72.8	40.9	49.8	73.3	37.3	50.2	72.1
AIC	83.0	84.0	94.3	82.6	86.6	93.2	81.0	83.9	93.7
BIC	83.0	84.0	94.3	82.6	86.6	93.2	81.0	83.9	93.7
	$\sigma^2 = 8$, $\beta_1 = 0.5$			$\sigma^2 = 8$, $\beta_1 = 1$			$\sigma^2 = 8$, $\beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
Mean R^2	0.039	0.037	0.051	0.058	0.065	0.105	0.118	0.151	0.252
KIC	28.7	34.9	47.3	30.3	34.3	52.2	28.6	35.7	50.1
AIC	100	85.1	83.2	100	86.9	87.7	100	85.5	83.3
BIC	100	85.1	83.2	100	86.9	87.7	100	85.5	83.3

3.3.3.2 Results: Scenario Two

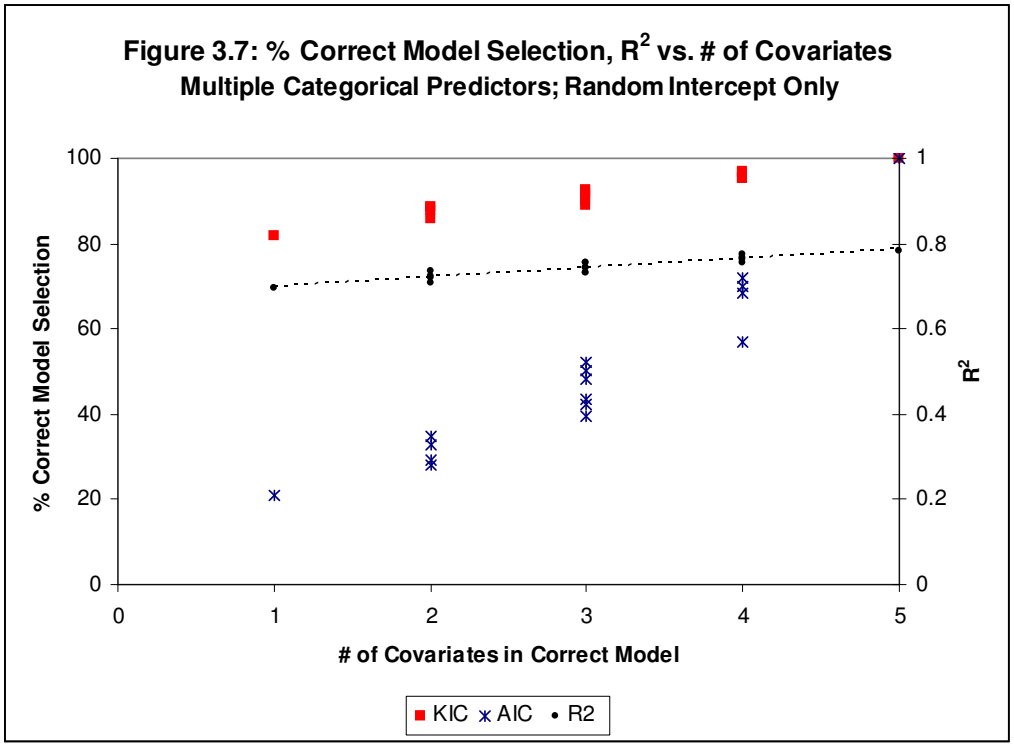
Tables 3.12, 3.13, 3.14 and 3.15 illustrate the results of the simulation study where there are multiple categorical predictors present. Tables 3.12 and 3.13 contain a random intercept only covariance structure while Tables 3.13 and 3.15 have both random intercept and random slope. Tables 3.12 and 3.13 have large sample sizes, and Tables 3.14 and 3.15 have smaller sample sizes. Figures 3.7 and 3.8 illustrate the data in Tables 3.12 and 3.13 respectively, and include a linear trendline for the R^2 data (dashed line in figure). Again, since the results for AIC and BIC were identical, only AIC was plotted (in addition to the KIC and the R^2).

Table 3.12: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 200 subjects, 5 observations per subject Scalar Covariance for Random Intercept Only

Fixed Effects for Simulated Data	# of Covariates	Mean R^2	% Correct Selection by KIC	AIC	BIC
$3+t+4x_1+3x_2+5x_3+5x_4$	5	.781	100	100	100
$3+t+4x_1+3x_2+5x_3$	4	.765	95.2	56.9	56.9
$3+t+4x_1+3x_2+5x_4$	4	.755	96.7	68.5	68.5
$3+t+4x_1+5x_3+5x_4$	4	.775	96.5	70.0	70.0
$3+t+3x_2+5x_3+5x_4$	4	.764	95.3	72.0	72.0
$3+t+4x_1+3x_2$	3	.733	90.1	42.1	42.1
$3+t+4x_1+5x_3$	3	.755	92.4	43.3	43.3
$3+t+4x_1+5x_4$	3	.744	91.4	48.2	48.2
$3+t+3x_2+5x_3$	3	.744	88.9	39.4	39.4
$3+t+3x_2+5x_4$	3	.733	92.5	52.1	52.1
$3+t+5x_3+5x_4$	3	.754	91.8	50.2	50.2
$3+t+4x_1$	2	.722	88.5	29.2	29.2
$3+t+3x_2$	2	.707	87.7	32.8	32.8
$3+t+5x_3$	2	.734	85.7	28.0	28.0
$3+t+5x_4$	2	.719	88.4	34.6	34.6
$3+t$	1	.694	82.0	20.8	20.8

Table 3.13: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 200 subjects, 5 observations per subject Unstructured Covariance for Random Intercept and Slope

Fixed Effects for Simulated Data	# of Covariates	Mean R^2	% Correct Selection by KIC	AIC	BIC
$3+t+4x_1+3x_2+5x_3+5x_4$	5	.259	100	100	100
$3+t+4x_1+3x_2+5x_3$	4	.222	83.7	0.5	0.5
$3+t+4x_1+3x_2+5x_4$	4	.203	88.3	0.2	0.2
$3+t+4x_1+5x_3+5x_4$	4	.240	88.8	0.2	0.2
$3+t+3x_2+5x_3+5x_4$	4	.219	88.4	0.3	0.3
$3+t+4x_1+3x_2$	3	.163	74.4	0	0
$3+t+4x_1+5x_3$	3	.203	73.5	0.1	0.2
$3+t+4x_1+5x_4$	3	.181	75.5	0	0
$3+t+3x_2+5x_3$	3	.181	75.8	0	0.1
$3+t+3x_2+5x_4$	3	.158	78.1	0.2	0.2
$3+t+5x_3+5x_4$	3	.201	76.5	0.2	0.2
$3+t+4x_1$	2	.139	65.2	0.1	0.2
$3+t+3x_2$	2	.114	65.0	0	0.1
$3+t+5x_3$	2	.161	66.4	0	0.1
$3+t+5x_4$	2	.135	69.3	0.1	0.1
$3+t$	1	.089	62.1	0	0.1



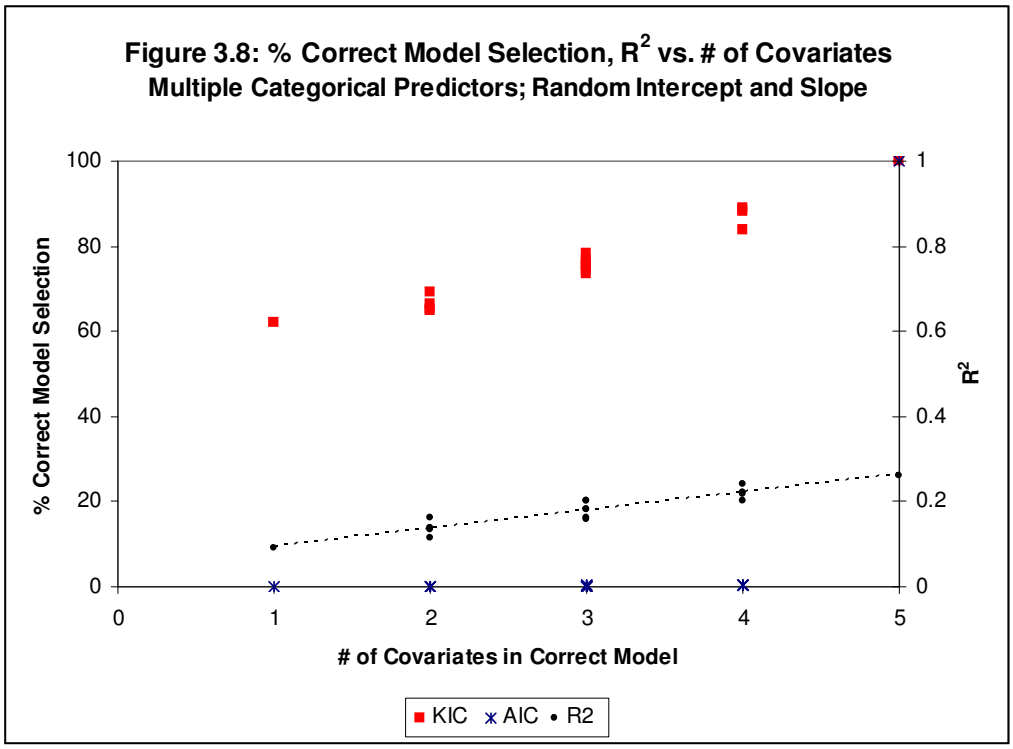


Table 3.14: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 50 subjects, 4 observations per subject Scalar Covariance for Random Intercept Only

Fixed Effects for Simulated Data	# of Covariates	Mean R^2	% Correct Selection by KIC	AIC	BIC
$3+t+4x_1+3x_2+5x_3+5x_4$	5	.807	100	100	100
$3+t+4x_1+3x_2+5x_3$	4	.791	86.5	0.1	0.1
$3+t+4x_1+3x_2+5x_4$	4	.780	88.4	1.6	1.6
$3+t+4x_1+5x_3+5x_4$	4	.799	88.7	1.1	1.1
$3+t+3x_2+5x_3+5x_4$	4	.789	90.6	2.1	2.1
$3+t+4x_1+3x_2$	3	.759	76.2	0	0
$3+t+4x_1+5x_3$	3	.782	76.2	0	0
$3+t+4x_1+5x_4$	3	.770	81.0	0.5	0.5
$3+t+3x_2+5x_3$	3	.770	75.4	0	0
$3+t+3x_2+5x_4$	3	.759	85.7	0.3	0.3
$3+t+5x_3+5x_4$	3	.779	78.8	0.1	0.1
$3+t+4x_1$	2	.746	66.7	0	0
$3+t+3x_2$	2	.732	69.4	0	0
$3+t+5x_3$	2	.758	68.0	0	0
$3+t+5x_4$	2	.745	72.8	0.1	0.1
$3+t$	1	.714	61.3	0	0

Table 3.15: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 50 subjects, 4 observations per subject Unstructured Covariance for Random Intercept and Slope

Fixed Effects for Simulated Data	# of Covariates	Mean R^2	% Correct Selection by KIC	AIC	BIC
$3 + t + 4x_1 + 3x_2 + 5x_3 + 5x_4$	5	.298	92.7	100	99.9
$3 + t + 4x_1 + 3x_2 + 5x_3$	4	.262	41.5	0	2.4
$3 + t + 4x_1 + 3x_2 + 5x_4$	4	.238	61.9	0.5	1.5
$3 + t + 4x_1 + 5x_3 + 5x_4$	4	.275	61.7	0.2	1.8
$3 + t + 3x_2 + 5x_3 + 5x_4$	4	.257	64.4	0.2	0.7
$3 + t + 4x_1 + 3x_2$	3	.195	28.6	0	0
$3 + t + 4x_1 + 5x_3$	3	.238	28.9	0	0
$3 + t + 4x_1 + 5x_4$	3	.217	41.1	0	0.1
$3 + t + 3x_2 + 5x_3$	3	.184	29.3	0	0
$3 + t + 3x_2 + 5x_4$	3	.194	43.0	0	0
$3 + t + 5x_3 + 5x_4$	3	.236	42.3	0	0.1
$3 + t + 4x_1$	2	.116	19.1	0	0
$3 + t + 3x_2$	2	.144	18.6	0	0
$3 + t + 5x_3$	2	.191	19.2	0	0
$3 + t + 5x_4$	2	.163	30.7	0	0
$3 + t$	1	.115	13.4	0	0

3.3.4 Conclusions

From both scenarios, we see that the AIC and BIC are nearly identical in their ability to select the correct mean structure when the correct covariance structure is specified, in all the variations explored. The KIC outperforms both criteria in every situation except when the correct mean structure includes an interaction term between time and group effect.

From the first scenario, when there is only one group effect present, we see that the KIC performs better than the other information criteria when there are no interaction effects present. When there are only main effects, the performance of the KIC does not seem to be as affected by the increase in total variance as the AIC and BIC, whose performance decrease in these two scenarios. Again, as in the selection of the covariance structure, we see that the increase in the R^2 statistic, due to the increased signal-to-noise ratio from the increase in the coefficient associated with the time effect does not affect the ability of any of the information criteria to select the correct fixed effect structure.

When the interaction effect is present, the AIC and BIC outperform the KIC in every variation. In this case, since the correct model is the model with all the available covariates included, it seems as though the fact that the AIC and BIC tend to included extraneous covariates serves as an advantage. This is most evident in the high variance case, i.e. where $\sigma^2 = 8$. We also see that when the strength of the signal of the interaction term is increased, the ability of all the criteria to correctly identify the model increases significantly. Also, we see that as the variance increased, the KIC was less likely to select the correct model, while the AIC and BIC was more likely to select the correct model. In this situation, the KIC appears to be more sensitive to the change in variance than the AIC and BIC.

From all the cases in this scenario, we see that as the correlation between observations were increased, the R^2 increased and the ability of the criteria to select the correct model also tended to increase.

In the second scenario, we see that the KIC performs best in all scenarios. It is the least likely of all the information criteria to add extraneous covariates to the model. We also see that the AIC and BIC only succeed in selecting the correct fixed effect structure when the covariance structure is simple, i.e., composed of a random intercept alone. While the ability of the KIC to select the correct fixed effect structure does decrease when the covariance structure is more complex, its ability to select the correct the fixed effect far exceeds that of the other criteria in the large sample size case. When the smaller sample size is used, we see that AIC and BIC lose all ability to identify the fixed effects regardless of covariance structure. Again, for the KIC, while its ability decreases, it still succeeds over 50% of the time in both cases when the covariance structure is composed of only a random intercept and in both larger sample size cases.

3.4 Example Data: Elderly Blood Pressure Study

3.4.1 Background

In order to investigate how the KIC performs in comparison to the AIC and BIC in a real world setting, we applied these criteria to the same example dataset used in Chapter 2 with the C_p statistic. This data comes from a retrospective longitudinal cohort study from the North Carolina Established Populations for the Epidemiologic Studies of the Elderly (EPESE). The goals of the EPESE project were to describe and identify predictors of mortality, hospitalization, and placement in long-term care facilities and to investigate risk factors for chronic diseases and of functioning among the elderly. The study followed 4162 subjects, aged 65 years and older, over a period of 12 years. The more intricate details of the study population can be found in Chapter 2. It should be noted that due to the subject matter (i.e. the elderly) and timeline of this project, this large dataset contains a great deal of missing data. Participants were surveyed at four time periods: Wave 1(1986); Wave 2 (1990); Wave 3 (1994); and Wave 4 (1998).

3.4.2 Methods

As was done in Chapter 2, to investigate the performance of the information criteria on this dataset, we used an all-possible regressions approach (where time in years is included in all models) using linear mixed models. We looked at average diastolic blood pressure as the response variable and used only main effects as predictors. There were 12 separate main effects chosen for study from a set of predictors numbering larger than 50: including time in years, 4 self-reported illness indices, race, marital status, gender, weight, diagnosis of diabetes, diagnosis of heart disease and whether the subject lived in a rural area). All predictors were binary categorical variables, except for time (which was labeled as 0, 4, 8, and 12 years).

We calculated the KIC, AIC and BIC for 2048 sets of fixed effects for the outcome variable. We used a random intercept, $\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d) = \sigma_1^2$, and i.i.d. within-subject error covariance, $\boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e) = \sigma_0^2 \boldsymbol{I}_{n_i}$. The 3 models with the lowest KIC, AIC and/or BIC values are further investigated by comparing what fixed effects were chosen and excluded by these models and by looking at the corresponding p-values for these fixed effects.

3.4.3 Results

Table 3.16 lists the fixed effects and information criteria values for the three models that were selected as 'best' (i.e. the smallest information criterion value) by the KIC, AIC and BIC. The three information criteria selected the same three models.

Table 3.17 lists the models and parameter estimates of these 3 models that had the lowest KIC, AIC and BIC values.

Table 3.16: KIC, AIC and BIC Results for Elderly Diastolic Blood Pressure Data
Models with the 3 lowest KIC, AIC and BIC values

Fixed Effects	KIC	AIC	BIC
year, weight, fair_ill, poor_ill, heart, diabet, blackpat, rural, male, poor_hlth, married	72469.33	72397.35	72422.41
year, weight, fair_ill, poor_ill, heart, diabet, blackpat, rural, male, poor_hlth, married, fair_hlth	72472.99	72398.00	72423.05
year, weight, fair_ill, poor_ill, heart, diabet, blackpat, rural, poor_hlth, married	72474.40	72405.48	72430.54

Table 3.17: Mixed Model Fixed Effect Estimates, Standard Errors (SE), p-values, and Covariance Estimates for the 3 Models with Lowest KIC Values (Outcome = Diastolic BP)

					Covariance Estimates	
KIC	**Fixed Effect**	**Estimate**	**SE**	**p-value**	**Random effects**	**Error**
72469.33	Intercept	65.60	0.78	<0.001	$\widehat{\sigma}_1^2 = 45.99$	$\widehat{\sigma}_0^2 = 87.92$
	year	− 0.65	0.02	<0.001		
	weight	0.04	0.005	<0.001		
	fair_ill	6.96	0.30	<0.001		
	poor_ill	9.88	0.41	<0.001		
	heart	− 4.88	0.40	<0.001		
	diabet	− 3.96	0.38	<0.001		
	blackpat	2.63	0.31	<0.001		
	poor_hlth	− 1.10	0.38	0.004		
	rural	1.40	0.30	<0.001		
	male	1.05	0.36	0.004		
	married	− 0.18	0.32	0.578		
72472.99	Intercept	65.61	0.78	<0.001	$\widehat{\sigma}_1^2 = 45.97$	$\widehat{\sigma}_0^2 = 87.94$
	year	− 0.66	0.02	<0.001		
	weight	0.04	0.005	<0.001		
	fair_ill	6.96	0.30	<0.001		
	poor_ill	9.89	0.41	<0.001		
	heart	− 4.87	0.40	<0.001		
	diabet	− 3.96	0.38	<0.001		
	blackpat	2.63	0.31	<0.001		
	poor_hlth	− 1.15	0.40	0.004		
	rural	1.40	0.30	<0.001		
	male	1.05	0.36	0.004		
	married	− 0.18	0.32	0.576		
	fair_hlth	− 0.10	0.26	0.697		
72474.40	Intercept	65.18	0.77	<0.001	$\widehat{\sigma}_1^2 = 46.29$	$\widehat{\sigma}_0^2 = 87.86$
	year	− 0.65	0.02	<0.001		
	weight	0.05	0.004	<0.001		
	fair_ill	6.96	0.30	<0.001		
	poor_ill	9.90	0.41	<0.001		
	heart	− 4.81	0.40	<0.001		
	diabet	− 4.03	0.38	<0.001		
	blackpat	2.62	0.31	<0.001		
	poor_hlth	− 1.11	0.38	0.004		
	rural	1.41	0.30	<0.001		
	married	0.18	0.30	0.560		

3.4.4 Conclusions

From this investigation, we see that the models chosen as best by the KIC are the same as those chosen by the AIC and BIC. In Chapter 2, where the same models were chosen by the AIC and BIC, we found that these models appeared to include extraneous variables. This conclusion again appears to be confirmed here, as we can see in Table 3.17. In this table we see that in the model that is chosen as 'best' by the three criteria, the covariate "married" is included, but from its associated p-value, it appears to not be a significant predictor of diastolic blood pressure in this dataset.

3.5 Discussion

The exploration of the ability of the KIC to select the correct covariance structure when the correct mean structure is specified, and the correct mean structure when the covariance structure is specified, has given us a great deal of information regarding the behavior of this criterion in comparison to other information criteria. In addition this study has provided us with more information with regards to the behavior of information criteria in general. While the AIC_c was recorded in all aspects of this study, the results were not posted here as they were identical to that of the AIC (due to the large sample nature of all the scenarios explored here).

From the exploration of the selection of the correct covariance structure when the mean structure is specified, we see that the strength of the BIC lies in being able to detect an i.i.d. covariance structure when it is the true structure, while the AIC excels in identifying the AR(1) structure when it is true. This confirmed the conclusions of Gomez (2005), who stated that the BIC worked best with simple covariance structures while the AIC worked best in more complex structures.

When looking at the selection of the correct mean structure when the covariance structure is specified, we see that the behavior of the AIC and BIC are nearly identical in all circumstances, and that the KIC far exceeds the other criteria in performance in all

situations except when an interaction term between group and time effect is present. We also see that the AIC and BIC are more sensitive to increases in variability of the data as opposed to the KIC which seems to work well (when the main effects models are correct) in all of the different variance scenarios.

Where the interaction effect is present, we see that the KIC does not work as well in detecting the presence of the interaction effect compared to the AIC and BIC. Also, in contrast to the situations where only main effects are present, the KIC is much more sensitive to changes in variance than the AIC and BIC when there is an interaction effect. It seems that the AIC and BIC are more apt than the KIC to include extraneous variables when selecting the mean structure, and this inclination may explain the better selection rates by these criteria when the interaction term is present.

From the observation of the R^2 statistic, we see that it is a good measure of the signal-to-noise ratio with regards to the mean structure in the linear mixed model. We also notice that the increase in the signal-to-noise ratio does not translate into an increase in the ability of any of the information criteria to correctly select the right set of fixed effects or covariance structures.

From our examination of the KIC using the real data from the EPESE project, we see that the KIC gives the same results as the AIC and BIC in this scenario. The tendency of these information criteria to choose the model that contains an extraneous variable may be due to the high variability in the data or the presence of a large amount of missing data in the dataset. Regardless, we can see that there are shortcomings in all of the information criteria examined here.

The use of the KIC in the realm of information criteria needs to be explored further in the linear mixed model as this simulation study and example data study show that it works as well as the criteria available in the situations tested here. From these explorations, we see that the KIC can be extremely valuable as a method of selection of

fixed effects, and can be as good as the standard measures as a selection method for covariance structure.

In addition to being an examination of the abilities of the KIC, this chapter has also served as an examination of the AIC and BIC. From our simulation studies, we can see that while these criteria are often blindly accepted as being accurate methods of model selection in the linear mixed model their error rates are extremely variable, and the user should be cautious with them and should not use them as the sole method of model selection.

Further studies could include an exploration of the penalty term of the KIC to accommodate for its use in the linear mixed model, as well as explorations into situations involving missing data and unbalanced designs.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**, 716.

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters*, **42**, 333-343.

Cavanaugh, J. E. (2004). Criteria for linear model selection based on Kullback's symmetric divergence. *Australian & New Zealand Journal of Statistics*, **46**, 257-274.

Edwards, L.J., Muller, K.E., Wolfinger, R.D., Qaqish, B.F., Schabenberger, O. (2008). An R^2 statistic for fixed effects in the linear mixed model. *In Press: Statistics in Medicine*.

Gomez, E., Schaalje, G., and Fellingham, G. E. -. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics: Simulation and Computation*, **34**, 377-392.

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

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

Muller, K. and Fetterman, B. (2002). Regression and ANOVA: An integrated approach using SAS software. Cary, NC, USA: SAS Institute Inc.

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

CHAPTER 4

SELECTION OF FIXED EFFECTS IN THE LINEAR MIXED MODEL USING THE PRESS STATISTIC

4.1 Introduction

The linear mixed model is an important tool in the modeling of continuous outcome longitudinal data. The linear mixed model extends the univariate linear model with independent and identically distributed (i.i.d.) Gaussian errors in a way that accommodates for the correlation of measurements within the same subject. While many different types of frequentist selection methods have been developed for the univariate linear model, the quantity and quality of frequentist model selection methods that have been developed for the linear mixed model leave much room for improvement. The linear mixed model requires selecting both a mean and a covariance model, and each must be considered separately.

With N independent sampling units (often *persons* in practice), the linear mixed model for person i may be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{d}_i + \boldsymbol{e}_i\,.$$

Here, $\boldsymbol{y}_i$ is a $n_i \times 1$ vector of observations on person i; $\boldsymbol{X}_i$ is a $n_i \times q$ known, constant design matrix for person i, with full column rank q while $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown, constant, population parameters. Also $\boldsymbol{Z}_i$ is a $n_i \times m$ known, constant design matrix with rank m for person i corresponding to the $m \times 1$ vector of unknown random effects $\boldsymbol{d}_i$,

while $\boldsymbol{e}_i$ is a $p_i \times 1$ vector of unknown random errors. Gaussian $\boldsymbol{d}_i$ and $\boldsymbol{e}_i$ are independent with mean $\mathbf{0}$ and

$$\mathcal{V}\left(\begin{bmatrix}\boldsymbol{d}_i \\ \boldsymbol{e}_i\end{bmatrix}\right) = \begin{bmatrix}\boldsymbol{\Sigma}_{di}(\boldsymbol{\tau}_d) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e)\end{bmatrix}.$$

Here $\boldsymbol{\mathcal{V}}(\,\cdot\,)$ is the covariance operator, while both $\boldsymbol{\Sigma}_{di}(\boldsymbol{\tau}_d)$ and $\boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e)$ are positive-definite, symmetric covariance matrices. Therefore $\boldsymbol{\mathcal{V}}(\boldsymbol{y}_i)$ may be written $\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{\Sigma}_{di}(\boldsymbol{\tau}_d)\boldsymbol{Z}_i' + \boldsymbol{\Sigma}_{ei}(\boldsymbol{\tau}_e)$. We assume that $\boldsymbol{\Sigma}_i$ can be characterized by a finite set of parameters represented by an $r \times 1$ vector $\boldsymbol{\tau}$ which consists of the unique parameters in $\boldsymbol{\tau}_d$ and $\boldsymbol{\tau}_e$. Throughout $n = \sum_{i=1}^{N} n_i$.

As discussed previously in this dissertation, model selection in the linear mixed model is complicated as both the correct mean structure and covariance structure need to be correctly identified and they cannot be selected simultaneously. Previously, we have looked at the ability of the C_p criterion as a method of mean structure selection and at the ability of the KIC as a method of mean structure and covariance structure selection. Recently, SAS Proc Mixed (SAS Institute, Cary, NC, 2007) has included a predicted residual sum of squares (PRESS) computation that can be output when modeling the linear mixed model. However, very little is known its performance since no broad studies have been conducted. In this chapter we will look at how this PRESS statistic performs as a method of fixed effect selection in the linear mixed model and we will compare its performance to that of the C_p statistic and the KIC, in addition to the standard information criteria used in the linear mixed model, the AIC and BIC.

4.1.1 Overview of the PRESS Statistic

The PRESS statistic uses a method of model selection called cross-validation. This method changes the goal of model selection from explaining a given set of data to predicting a new set of data which comes from the same background as the given set. Conceptually, it is the same "predictive" method that is used with the C_p statistic.

Proposed by Allen (1974), the PRESS criterion is obtained by deleting the ith case from a data set, estimating the regression function for the subset model from the remaining $n-1$ cases, and then using the fitted regression function to obtain the predicted value $\widehat{y}_{(i)}$ (Neter et al., 1996). The PRESS residuals are defined as:

$$e_{(i)} = y_i - \widehat{y}_{(i)} \tag{4.1}$$

This process is repeated for all n observations and the PRESS statistic is then computed as:

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 \tag{4.2}$$

As the PRESS residuals are a measure of how well the fitted model is able to predict the response, the smaller the PRESS statistic the better the model is for prediction. In other words, the PRESS statistic selects the model with the smallest mean square error of prediction.

Liu et al. (1999) generalized the PRESS statistic to multivariate linear models with correlated errors in repeated measures data. Liu et al. (1999) use the PRESS statistic to select the linear predictor in linear models with correlated errors. They define the PRESS statistic in linear models with correlated errors as:

$$PRESS = \sum_{i=1}^{n} e'_{(i)} e_{(i)} \tag{4.3}$$

where $e_{(i)} = y_i - X_i \widehat{\boldsymbol{\beta}}_{(i)}$ is the deleted residual with $\widehat{\boldsymbol{\beta}}_{(i)}$ defined as the regression parameter estimate when the ith person is deleted from the analysis. This definition of PRESS is applicable to both balanced and unbalanced data where each person has a different number of measurements, and can only be used to select the linear predictor (the covariance structure is treated as a nuisance parameter). This definition is very close to the form of the original PRESS statistic in traditional linear regression.

4.1.2 Studies Using the PRESS Statistic in the Linear Mixed Model

Liu et al. (1999) proposed a new efficient computing method based on pivoting and then proposed to apply the PRESS model selection method to real data. They compared the top 10 models that were selected by the PRESS to the top 10 models selected by AIC, BIC and likelihood ratio tests using a linear mixed effects model. They found that there were 5 models that were selected by both AIC and PRESS in the top 10, but while the AIC values were very close together for the top 10 models, the PRESS statistic had a small difference between the first two models and a large difference between those and the other 8 models. The BIC and the PRESS statistic had 6 of the same models in their top 10 and the model chosen as best by the BIC was the same as that chosen by the forward selection likelihood ratio test with Type I error of $\alpha = 0.01$.

4.1.3 Calculation of the PRESS Statistic in SAS

The most recent version of SAS (version 9.1.3, Cary, NC) includes the calculation of PRESS residuals as part of its experimental "influence" option (Schabenberger 2005). The definition of the PRESS residuals used here is given as:

$$\widehat{e}_{i(U)} = y_i - x_i'\widehat{\beta}_{(U)} \tag{4.4}$$

where the subscript (U) denote quantities obtained without observations in the set U.The PRESS statistic is then computed as the sum of these squared residuals. The "influence" option in SAS allows for influence diagnostics to be calculated iteratively or non-iteratively. A noniterative influence analysis relies on closed-form update formulas for the fixed effects without updating the covariance parameters, while an iterative influence analysis involves iterative reestimation of the covariance parameters. In this paper, an iterative influence analysis is performed to obtain the PRESS statistics where the covariance parameters are updated up to five times for each deletion set.

4.1.4 Investigation of PRESS and Comparison to KIC and C_p in This Chapter

In this chapter, we will compare the performance of the PRESS statistic to that of the C_p and the KIC, using the same method of simulation study for fixed effects and the same data example as were used in Chapter 3. From this investigation, we will be able to determine the strengths and limitations of all three criteria explored in this dissertation and compare them to the standard criteria used in the linear mixed model, the AIC and BIC.

4.2 Summary of Past Results

In the last two chapters we have investigated two model selection methods in the linear mixed model: the C_p statistic and the KIC. The C_p statistic is a method of selection for the fixed effects and requires that the models be nested. The KIC is an information criterion that requires comparison to another model that is similar to it, but can be used for selection of both fixed and random effects. In both chapters, we compared the model selection abilities of these two criteria to those provided in the standard SAS output, namely the AIC and the BIC.

In Chapter 2, we found that the C_p appeared to work as well or better than the AIC and BIC, in most cases, as a selection criterion for the fixed effects in linear mixed model. In addition, information criteria are more computationally intensive in that they require that every candidate model be fit individually in order to determine which one is best. In contrast, in order to calculate the C_p statistic, the full (or saturated) model is fit, and the C_p is calculated via F statistics generated through contrasts within the same model. The ease of these calculations in combination with the improved performance led us to conclude that the C_p statistic can be a valuable tool for fixed effects selection in the linear mixed model.

In Chapter 3, when looking at the selection of the correct mean structure when the covariance structure is specified, we saw that the KIC far exceeded the AIC and BIC in

performance in all situations except when an interaction term between group and time effect were present. It seemed that the AIC and BIC are more apt than the KIC to include extraneous variables when selecting the mean structure, while the KIC was more adept at selecting the most parsimonious model, and this trait may explain why the AIC and BIC performed better in our study involving the interaction term.

4.3 Simulation Study: Mean Structure Selection

4.3.1 Introduction

This simulation study investigates the ability of the PRESS to select the correct mean structure when the correct covariance structure is identified, and we compare its ability to that of the C_p, KIC, AIC and BIC. Again, since this is a large sample study, we found that the behavior of the AIC_c was identical to that of the AIC, so it is omitted here. The scenarios used are identical to those in Chapter 3 when looking at KIC performance in the selection of fixed effects. For both scenarios investigated, we used an i.i.d. within-unit error covariance structure and REML estimation with complete data. For the PRESS statistic we used residual denominator degrees of freedom, due to the computational intensity involved with calculating the PRESS. Kenward-Roger denominator degrees of freedom were used for the other criteria.

4.3.2 Methods

4.3.2.1 Methods: Scenario One

For the first mean structure selection scenario, a repeated measures study consisting of five measurement occasions was simulated that assumes compound symmetry of the data; that is, the data generated were based on a model with only a random intercept and an independent and identically distributed (i.i.d.) within-unit error term i.e., $\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d) = \sigma_1^2$ and

$$\boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e) = \sigma_0^2 \boldsymbol{I}_5,$$

where $\boldsymbol{I}_5$ is a 5×5 identity matrix. Therefore, the compound symmetry structure was simulated such that $\sigma^2 = \sigma_1^2 + \sigma_0^2$ where σ^2 is the total variance, σ_1^2 is the variance of the random intercept, and σ_0^2 is the variance of the within-unit error term. Datasets were simulated using varying values of the total variance, namely $\sigma^2 = 1$, 4 and 8, to get a sense of the impact of the true variance on the performance of the criteria. Corresponding to the standard model assumptions, the random intercepts and within-unit error terms were generated as independent normal random variables with means zero and variances σ_1^2 and σ_0^2, respectively. To understand the impact of the within-unit correlations on the performance of the criteria, data were simulated using correlations between observations $\rho = 0.25$, 0.50 and 0.75, where $\rho = \sigma_1^2/\sigma^2$.

As for the mean structure, data were simulated from a true linear mixed model consisting of the following fixed effects: an intercept, a dummy variable indicating membership in one of two groups, and a continuous covariate. Nine different variations of models for these fixed effects were considered to see how the information criteria performed in their ability to select the correct set of fixed effects. The fixed effects models simulated took on three variations: 1) model with intercept and time only, 2) model with intercept, time and group effect, and 3) model with intercept, time, group and interaction between group and time. In each variation, three values for the coefficient associated with the time effect, i.e. β_1, were considered: $\beta_1 = 0.5,\ 1,$ and 2. Thus, with three models for fixed effects and three values for β_1, we have nine different variations of models for the fixed effects. In addition, when the correct model included an interaction term, two values for the coefficient associated with the interaction, i.e. β_3, were considered: $\beta_3 = 0.25,\ 0.5$. Large sample performance was assessed; simulated datasets consisted of 200 subjects with 5 observations each in the time interval [0,1]. Each

simulation for the varying sample sizes, variances and correlation values consisted of 1,000 realizations.

In order to assess the performance of the criteria in choosing the proper set of fixed effects based on the simulation study, a set of candidate models was fit for each generated dataset, and the number of times the criteria chose the correct model from this set of 1,000 was tallied. The set of candidate models consisted of three models each having the same covariance structure and fixed effects corresponding to: (1) a model with common intercept and common slope; (2) a model with common intercept, a slope, and an additional group covariate; and (3) a model with intercept, slope, group, and group x slope interaction. For example, if the true model was one with a common intercept and common slope, we assessed the performance of the criteria in choosing the true model as a candidate model or the other two candidate models. The number of times out of the 1,000 realizations that the criterion in question chose the correct model as the best model was recorded.

4.3.2.2 Methods: Scenario Two

The multiple categorical predictors that are used to generate the data simulate the case where there are various factors (in addition to the time effect), such as race, gender etc., that have an impact on the outcome variable.

For this set of simulations, we again used an i.i.d within-subject error covariance structure i.e.,

$$\boldsymbol{\Sigma}_{e_i}(\boldsymbol{\tau}_e) = \sigma_0^2 \boldsymbol{I}_5,$$

with a between-subject covariance structure that contained either a random intercept alone, or a random intercept and slope, i.e., $\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d) = \sigma_1^2$ or

$$\boldsymbol{\Sigma}_{d_i}(\boldsymbol{\tau}_d) = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix}$$

In this simulation, $\sigma_0^2 = 2$, $\sigma_1^2 = 2$, $\sigma_2^2 = 0$ or 2, and $\rho_d = 0.25$ when $\sigma_2^2 = 2$ ($\sigma_{12}^2 = \rho_d \sigma_1 \sigma_2$).

For the fixed effects, β_1 is the common intercept, β_2 is the coefficient for time and four different categorical predictors, x_1, x_2, x_3 and x_4, are considered with the corresponding coefficients β_3, β_4, β_5, and β_6. The vector for $\beta' = (3, 1, 4, 3, 5, 5)$, and 16 different variations are considered where 0,1, 2, 3, or 4 of the coefficients for the categorical predictors (i.e. x_1, x_2, x_3 and x_4) are set to 0. Therefore, the saturated mean model, which includes all coefficients for the categorical predictors and time, is given by:

$$E(y) = 3 + t + 4x_1 + 3x_2 + 5x_3 + 5x_4$$

For each set of simulations, 16 different sets of fixed effects were assessed. We looked at the ability of the PRESS compared to the C_p, KIC, AIC and BIC to select the correct fixed effect structure among the 16 possible outcomes over the 1,000 simulations.

4.3.3 Results

4.3.3.1 Results: Scenario One

Table 4.1 displays the results for the fixed effects model selection when the correct model is time alone. Figures 4.1, 4.2 and 4.3 illustrate the results from the middle panel of this table (i.e. when $\beta_1 = 1$). Table 4.2 displays the results when the correct model is time and group, and Figures 4.4, 4.5 and 4.6 illustrate the results from the middle panel of this table (again when $\beta_1 = 1$). Table 4.3 and 4.4 display the results when the correct model is time, group and their interaction. Tables 4.3 and 4.4 differ in that in Table 4.4 the interaction effect has double the magnitude of the interaction effect in Table 4.3. Again, Figures 4.7, 4.8, 4.9 and 4.10, 4.11, 4.12 illustrate the results from the middle panel of these two tables, respectively (when $\beta_1 = 1$). In all tables the changes of the variance, covariance and time effect parameters are illustrated. As the AIC and BIC have identical values in all of these tables, only the AIC is graphed.

Table 4.1: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 200 subjects each, 5 observations per subject

Correct Model: $\beta_0 + \beta_1 t$

	Percentage Correct Model Selection								
	$\sigma^2 = 1$, $\beta_1 = 0.5$			$\sigma^2 = 1$, $\beta_1 = 1$			$\sigma^2 = 1$, $\beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
PRESS	77.6	80.4	80.4	77.6	81.0	81.2	79.2	79.6	79.1
C_p	77.6	77.1	79.8	77.0	79.1	79.9	79.2	78.5	78.9
KIC	98.1	98.0	97.9	98.7	98.2	97.5	98.0	98.2	97.4
AIC	87.2	84.3	84.2	87.4	84.7	85.1	87.7	85.1	84.7
BIC	87.2	84.3	84.2	87.4	84.7	85.1	87.7	85.1	84.7
	$\sigma^2 = 4$, $\beta_1 = 0.5$			$\sigma^2 = 4$, $\beta_1 = 1$			$\sigma^2 = 4$, $\beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
PRESS	79.1	78.1	79.8	79.3	80.4	79.7	78.9	79.4	81.0
C_p	78.6	76.9	79.1	80.9	80.7	77.2	78.4	79.3	81.7
KIC	95.8	93.5	94.9	96.0	95.2	93.6	96.9	95.2	95.9
AIC	63.4	57.8	56.9	64.8	60.8	58.5	61.9	55.6	59.8
BIC	63.4	57.8	56.9	64.8	60.8	58.5	61.9	55.6	59.8
	$\sigma^2 = 8$, $\beta_1 = 0.5$			$\sigma^2 = 8$, $\beta_1 = 1$			$\sigma^2 = 8$, $\beta_1 = 2$		
ρ =	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**
PRESS	80.3	78.5	80.3	78.9	79.1	78.9	80.9	79.5	78.6
C_p	80.1	77.5	79.1	78.2	76.5	78.8	79.9	76.2	76.1
KIC	93.1	92.2	91.8	94.4	91.5	91.6	95.7	91.8	90.7
AIC	29.5	29.2	26.2	29.2	27.2	23.9	30.8	25.3	20.9
BIC	29.5	29.2	26.2	29.2	27.2	23.9	30.8	25.3	20.9

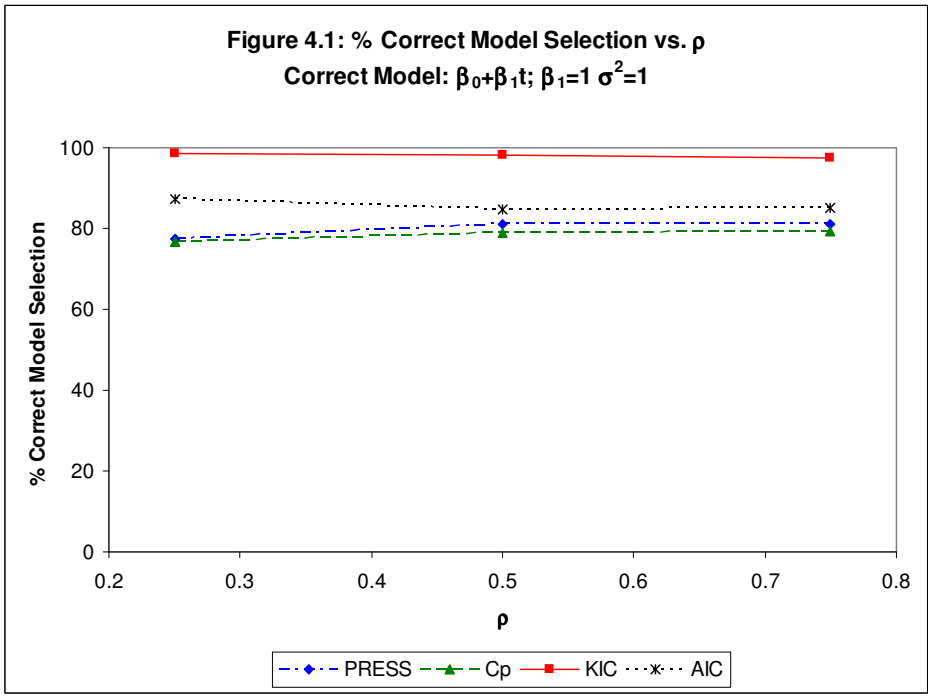


Figure 4.1: % Correct Model Selection vs. ρ
Correct Model: β0+β1t; β1=1 σ2=1
% Correct Model Selection
100
80
60
40
20
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ρ
PRESS
Cp
KIC
AIC

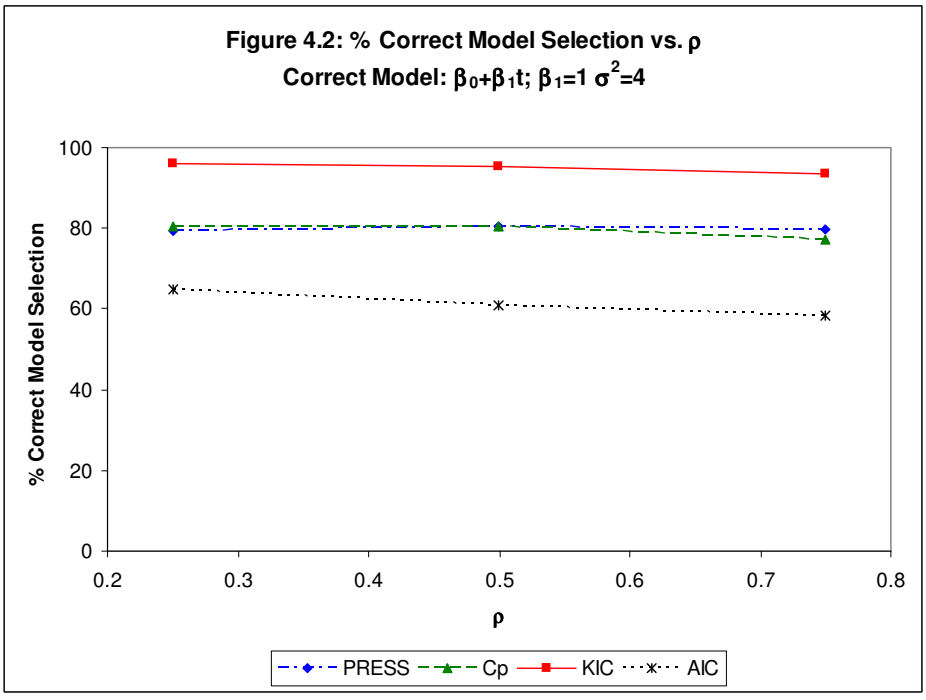


Figure 4.2: % Correct Model Selection vs. ρ
Correct Model: β0+β1t; β1=1 σ2=4
% Correct Model Selection
100
80
60
40
20
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ρ
PRESS
Cp
KIC
AIC

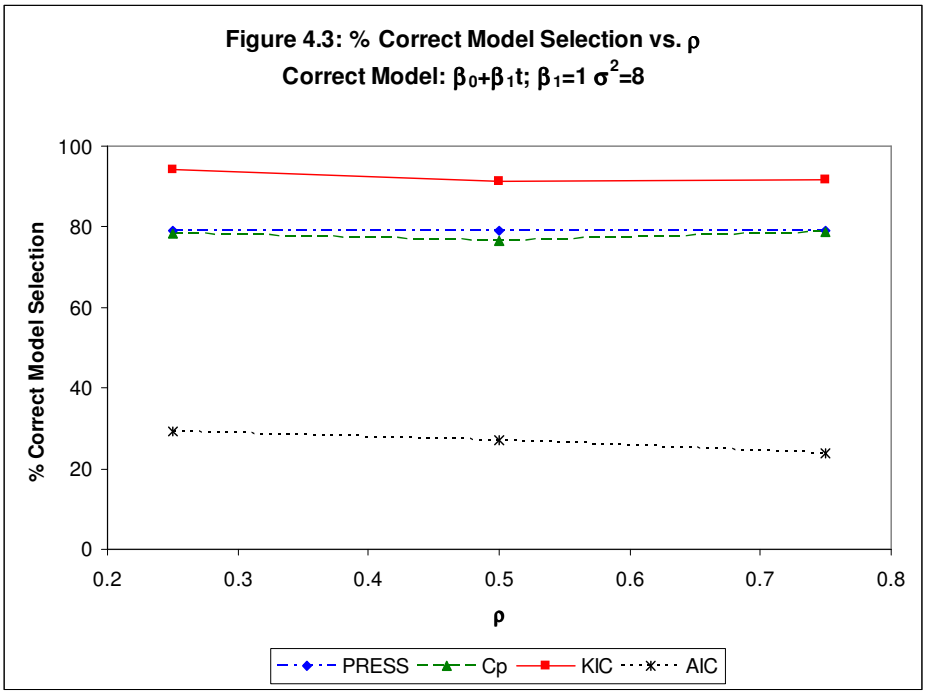


Table 4.2: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 200 subjects each, 5 observations per subject

Correct Model: $\beta_0 + \beta_1 t + \beta_2 x$

	Percentage Correct Model Selection								
	$\sigma^2 = 1, \beta_1 = 0.5$			$\sigma^2 = 1, \beta_1 = 1$			$\sigma^2 = 1, \beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
PRESS	80.6	81.0	77.6	82.5	77.8	78.8	82.4	79.3	77.1
C_p	84.2	84.9	85.0	83.9	85.8	85.7	84.9	84.5	84.4
KIC	96.9	98.4	98.6	97.7	97.4	98.4	96.4	97.8	97.6
AIC	82.7	86.8	90.9	82.2	86.8	91.8	83.3	87.2	90.9
BIC	82.7	86.8	90.9	82.2	86.8	91.8	83.3	87.2	90.9
	$\sigma^2 = 4, \beta_1 = 0.5$			$\sigma^2 = 4, \beta_1 = 1$			$\sigma^2 = 4, \beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
PRESS	81.6	77.8	75.7	81.8	79.0	73.7	79.6	79.9	74.0
C_p	85.3	83.9	82.4	85.8	84.9	82.2	83.4	83.3	83.9
KIC	92.9	93.6	95.2	94.6	94.5	93.5	93.1	94.9	94.4
AIC	53.7	65.6	77.0	51.3	67.5	77.6	52.2	64.8	78.8
BIC	53.7	65.6	77.0	51.3	67.5	77.6	52.2	64.8	78.8
	$\sigma^2 = 8, \beta_1 = 0.5$			$\sigma^2 = 8, \beta_1 = 1$			$\sigma^2 = 8, \beta_1 = 2$		
ρ =	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**
PRESS	79.8	75.1	66.5	79.4	70.1	66.7	79.3	75.5	66.5
C_p	82.9	82.3	77.3	84.4	78.6	79.4	85.2	82.8	78.7
KIC	87.4	85.4	78.8	88.6	82.7	80.7	88.7	85.6	80.0
AIC	0	35.7	66.0	0	34.6	65.9	0	34.7	64.7
BIC	0	35.7	66.0	0	34.6	65.9	0	34.7	64.7

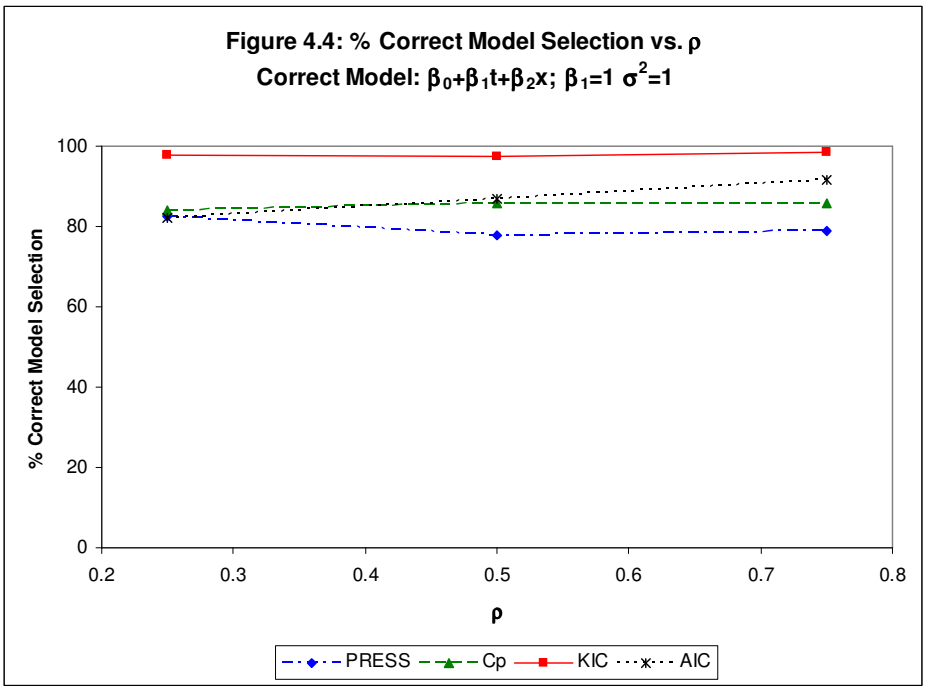


Figure 4.4: % Correct Model Selection vs. ρ
Correct Model: β0+β1t+β2x; β1=1 σ2=1
% Correct Model Selection
100
80
60
40
20
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ρ
PRESS
Cp
KIC
AIC

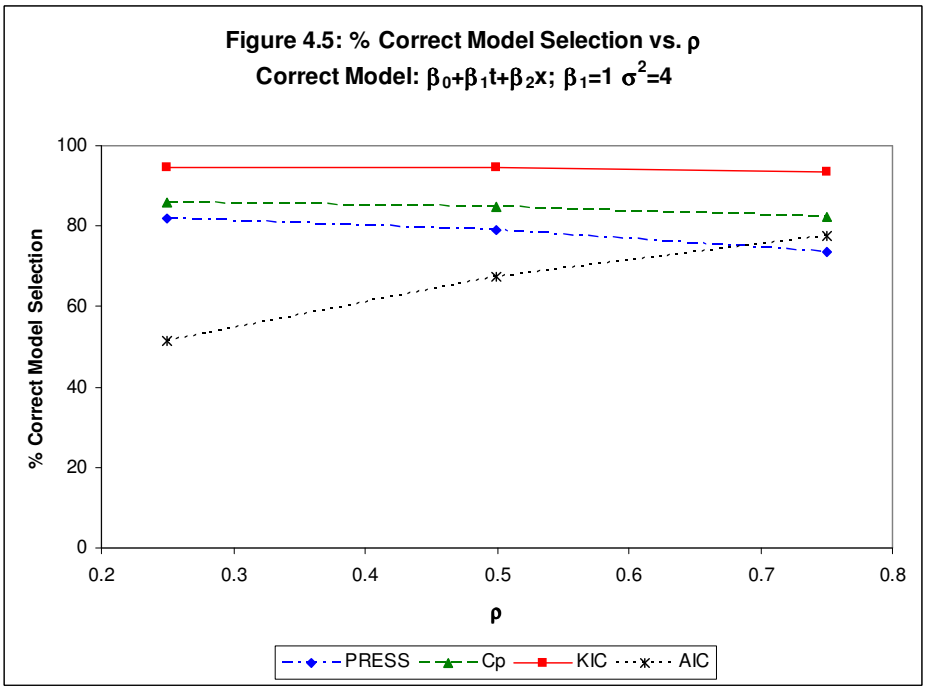


Figure 4.5: % Correct Model Selection vs. ρ
Correct Model: β0+β1t+β2x; β1=1 σ2=4
% Correct Model Selection
100
80
60
40
20
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ρ
PRESS
Cp
KIC
AIC

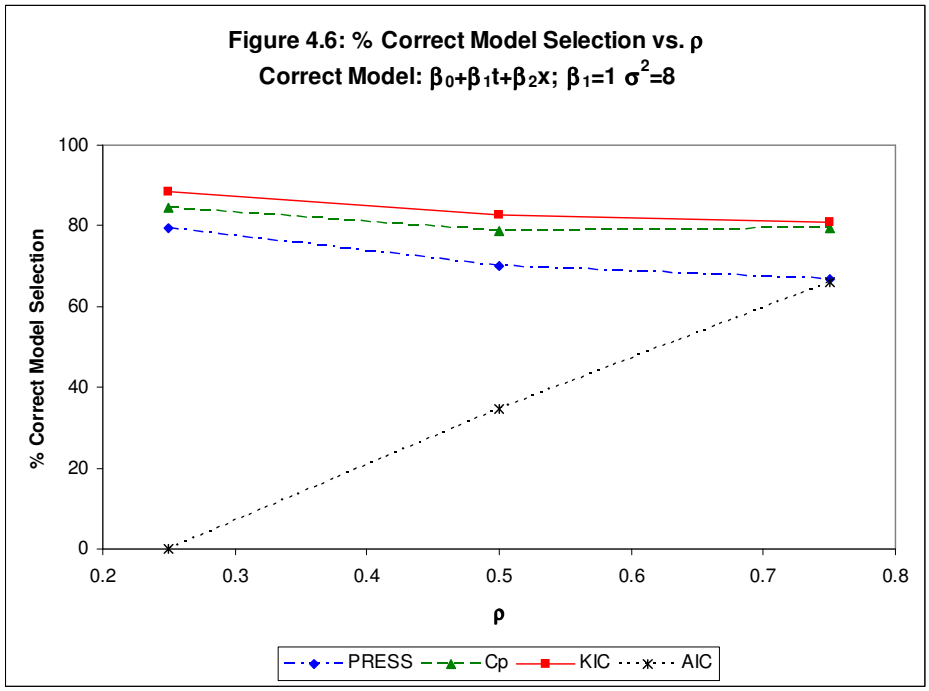


Table 4.3: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 200 subjects each, 5 observations per subject

Correct Model: $\beta_0 + \beta_1 t + \beta_2 x + \beta_3 x*t$; β_3=0.25

	Percentage Correct Model Selection								
	$\sigma^2 = 1, \beta_1 = 0.5$			$\sigma^2 = 1, \beta_1 = 1$			$\sigma^2 = 1, \beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
PRESS	50.9	63.9	79.2	52.0	67.2	79.1	54.4	63.3	77.5
C_p	54.2	71.9	93.7	54.9	73.2	92.0	58.0	70.5	91.4
KIC	25.7	37.2	64.5	25.7	37.5	61.5	27.5	37.8	62.8
AIC	56.0	68.6	86.5	56.9	70.1	85.8	59.5	67.6	85.8
BIC	56.0	68.6	86.5	56.9	70.1	85.8	59.5	67.6	85.8
	$\sigma^2 = 4, \beta_1 = 0.5$			$\sigma^2 = 4, \beta_1 = 1$			$\sigma^2 = 4, \beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
PRESS	28.5	36.6	43.0	28.1	35.7	41.7	29.1	34.0	44.4
C_p	28.7	37.6	48.0	27.5	34.8	48.1	26.5	31.9	49.4
KIC	13.2	16.1	21.4	12.9	16.5	21.7	13.7	14.5	21.6
AIC	60.9	57.6	53.1	59.1	55.8	54.8	58.0	51.7	55.7
BIC	60.9	57.6	53.1	59.1	55.8	54.8	58.0	51.7	55.7
	$\sigma^2 = 8, \beta_1 = 0.5$			$\sigma^2 = 8, \beta_1 = 1$			$\sigma^2 = 8, \beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.5**	**0.75**
PRESS	22.9	26.9	31.7	27.4	26.2	33.5	26.3	24.1	30.5
C_p	21.9	26.1	34.3	23.6	24.0	34.4	24.5	23.1	33.1
KIC	14.4	14.5	13.8	16.3	14.4	16.5	16.3	14.4	15.1
AIC	100	72.4	53.9	100	73.3	55.2	100	72.0	54.4
BIC	100	72.4	53.9	100	73.3	55.2	100	72.0	54.4

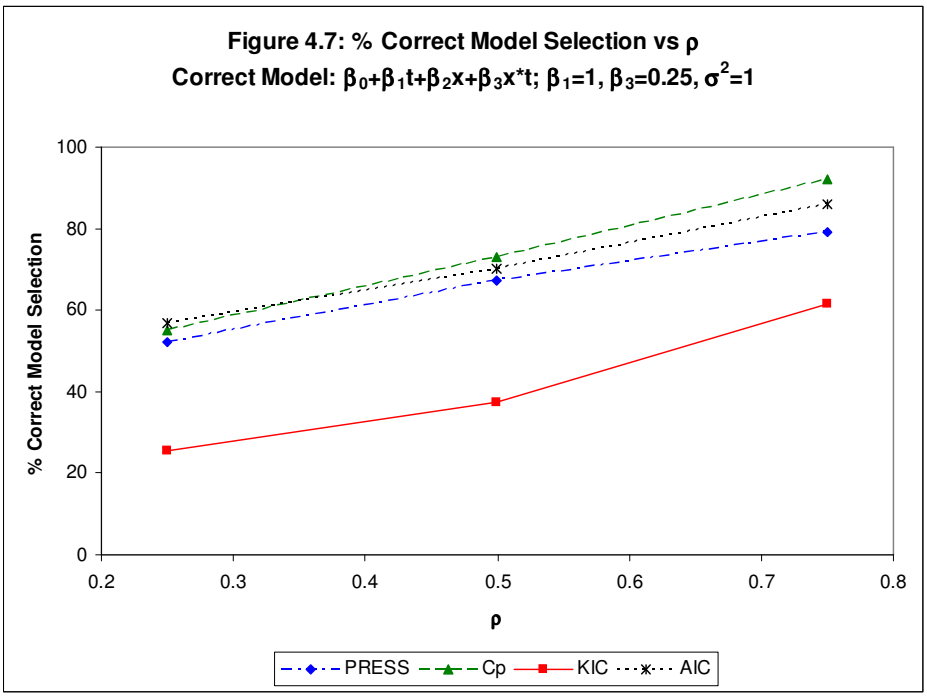


Figure 4.7: % Correct Model Selection vs ρ
Correct Model: β0+β1t+β2x+β3x*t; β1=1, β3=0.25, σ2=1
% Correct Model Selection
100
80
60
40
20
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ρ
PRESS
Cp
KIC
AIC

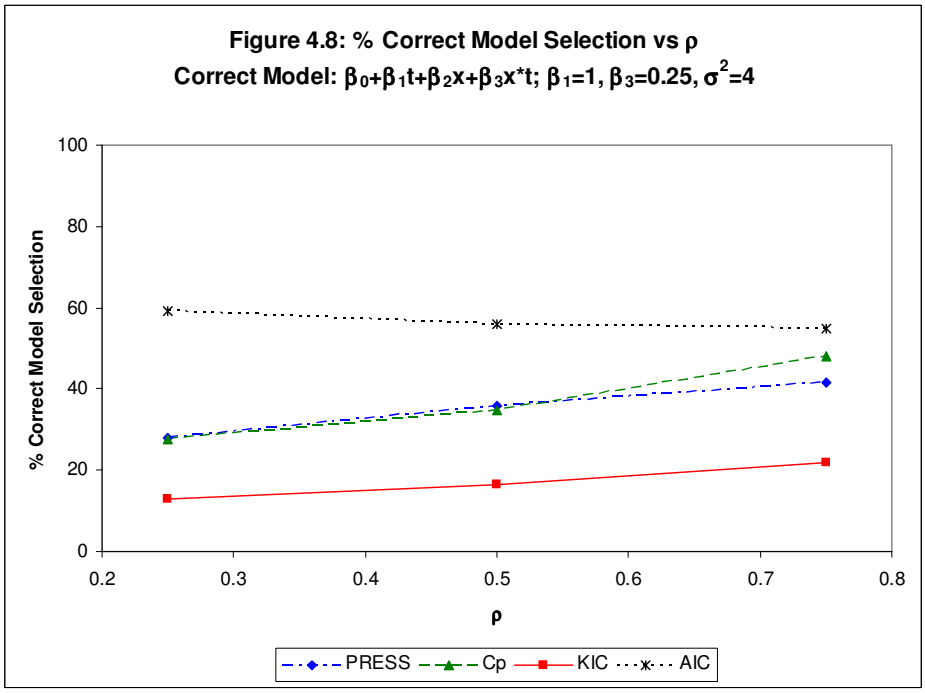


Figure 4.8: % Correct Model Selection vs ρ
Correct Model: β0+β1t+β2x+β3x*t; β1=1, β3=0.25, σ2=4
% Correct Model Selection
100
80
60
40
20
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ρ
PRESS
Cp
KIC
AIC

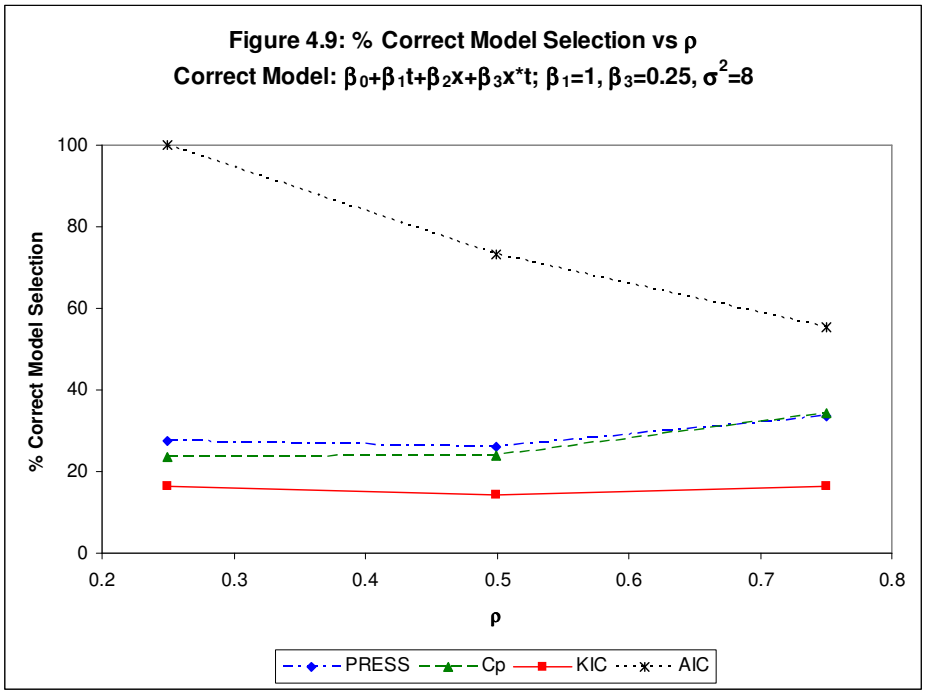


Table 4.4: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 200 subjects each, 5 observations per subject

Correct Model: $\beta_0 + \beta_1 t + \beta_2 x + \beta_3 x*t$; β_3=0.50

	Percentage Correct Model Selection								
	$\sigma^2 = 1, \beta_1 = 0.5$			$\sigma^2 = 1, \beta_1 = 1$			$\sigma^2 = 1, \beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
PRESS	91.9	96.0	98.6	92.5	95.8	99.1	92.0	95.5	98.5
C_p	96.8	99.6	100	96.3	99.7	100	96.4	99.4	100
KIC	82.7	95.4	100	85.9	94.5	99.8	84.8	95.9	99.9
AIC	96.8	99.3	100	96.3	99.3	99.3	96.4	99.4	100
BIC	96.8	99.3	100	96.3	99.3	99.3	96.4	99.4	100
	$\sigma^2 = 4, \beta_1 = 0.5$			$\sigma^2 = 4, \beta_1 = 1$			$\sigma^2 = 4, \beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
PRESS	55.0	63.3	75.7	56.4	62.0	79.3	51.1	64.4	78.1
C_p	57.4	70.8	92.6	57.6	72.7	91.2	55.9	70.7	90.7
KIC	39.2	49.3	72.8	40.9	49.8	73.3	37.3	50.2	72.1
AIC	83.0	84.0	94.3	82.6	86.6	93.2	81.0	83.9	93.7
BIC	83.0	84.0	94.3	82.6	86.6	93.2	81.0	83.9	93.7
	$\sigma^2 = 8, \beta_1 = 0.5$			$\sigma^2 = 8, \beta_1 = 1$			$\sigma^2 = 8, \beta_1 = 2$		
ρ =	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
PRESS	39.3	45.8	55.9	39.4	46.2	58.7	39.0	46.0	56.1
C_p	37.9	49.8	70.4	40.9	49.0	73.3	39.7	49.6	71.5
KIC	28.7	34.9	47.3	30.3	34.3	52.2	28.6	35.7	50.1
AIC	100	85.1	83.2	100	86.9	87.7	100	85.5	83.3
BIC	100	85.1	83.2	100	86.9	87.7	100	85.5	83.3

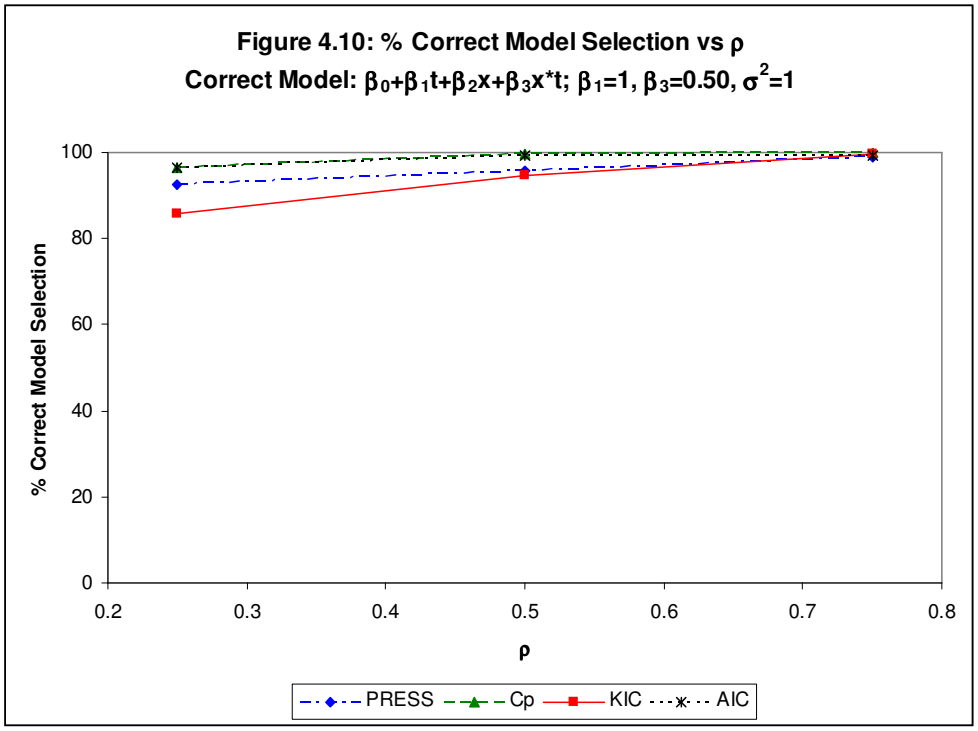


Figure 4.10: % Correct Model Selection vs ρ
Correct Model: β0+β1t+β2x+β3x*t; β1=1, β3=0.50, σ2=1
% Correct Model Selection
100
80
60
40
20
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ρ
PRESS
Cp
KIC
AIC

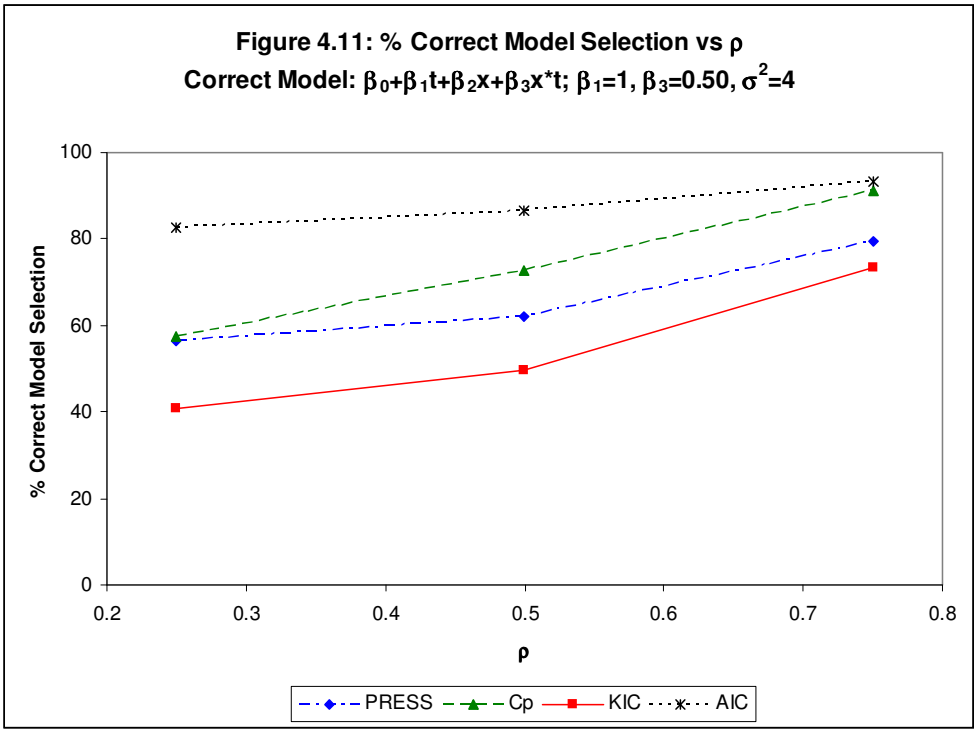
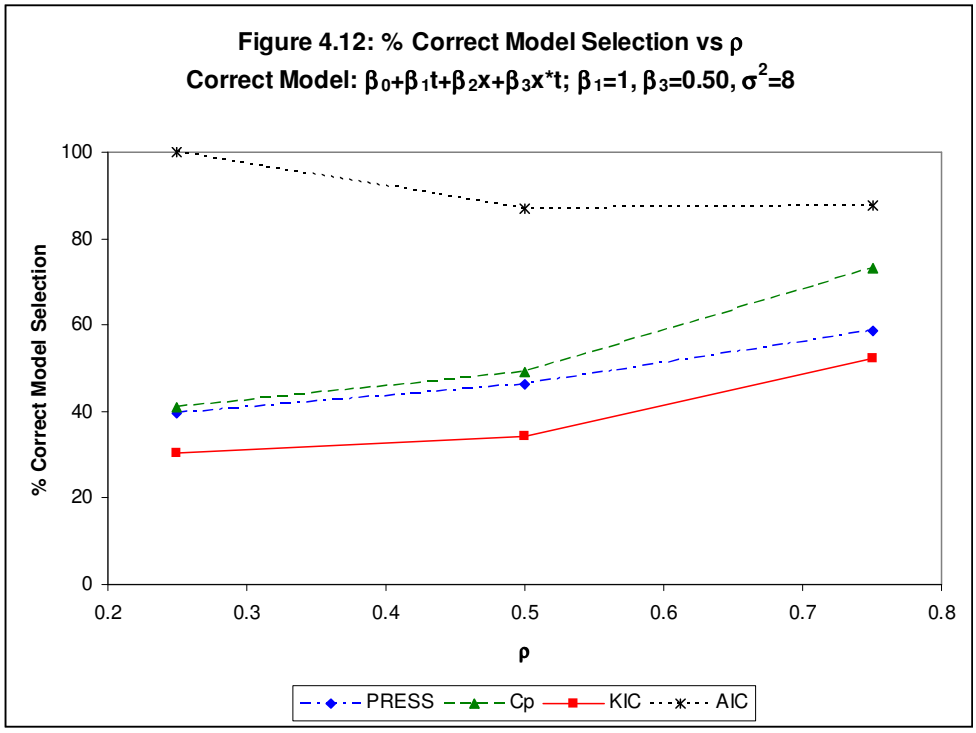


Figure 4.11: % Correct Model Selection vs ρ
Correct Model: β0+β1t+β2x+β3x*t; β1=1, β3=0.50, σ2=4
% Correct Model Selection
100
80
60
40
20
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ρ
PRESS
Cp
KIC
AIC



4.3.3.2 Results: Scenario Two

Tables 4.5, 4.6, 4.7 and 4.8 illustrate the results of the simulation study where there are multiple categorical predictors present. Tables 4.5 and 4.7 contain a random intercept only while Tables 4.6 and 4.8 have both random intercept and random slope. Tables 4.5 and 4.6 have large sample sizes, and Tables 4.7 and 4.8 have smaller sample sizes. Figures 4.13-16 illustrate the data present in Tables 4.5-4.8 respectively. Again, since the results for AIC and BIC were nearly identical, only AIC was plotted.

Table 4.5: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 200 subjects, 5 observations per subject Scalar Covariance for Random Intercept Only

Fixed Effects for Simulated Data	# of Covariates	% Correct Selection by PRESS	C_p	KIC	AIC	BIC
$3+t+4x_1+3x_2+5x_3+5x_4$	5	100	100	100	100	100
$3+t+4x_1+3x_2+5x_3$	4	81.0	84.6	95.2	56.9	56.9
$3+t+4x_1+3x_2+5x_4$	4	80.3	83.1	96.7	68.5	68.5
$3+t+4x_1+5x_3+5x_4$	4	81.9	85.3	96.5	70.0	70.0
$3+t+3x_2+5x_3+5x_4$	4	80.7	82.7	95.3	72.0	72.0
$3+t+4x_1+3x_2$	3	65.7	72.7	90.1	42.1	42.1
$3+t+4x_1+5x_3$	3	65.5	72.2	92.4	43.3	43.3
$3+t+4x_1+5x_4$	3	63.3	69.7	91.4	48.2	48.2
$3+t+3x_2+5x_3$	3	60.8	69.1	88.9	39.4	39.4
$3+t+3x_2+5x_4$	3	65.1	71.7	92.5	52.1	52.1
$3+t+5x_3+5x_4$	3	65.3	70.1	91.8	50.2	50.2
$3+t+4x_1$	2	52.0	62.0	88.5	29.2	29.2
$3+t+3x_2$	2	55.7	60.9	87.7	32.8	32.8
$3+t+5x_3$	2	49.9	56.7	85.7	28.0	28.0
$3+t+5x_4$	2	51.0	60.2	88.4	34.6	34.6
$3+t$	1	41.3	48.2	82.0	20.8	20.8

Figure 4.13: % Correct Model Selection vs. # of Covariates
Multiple Categorical Predictors
(Random Intercept Only; Large Sample Size)

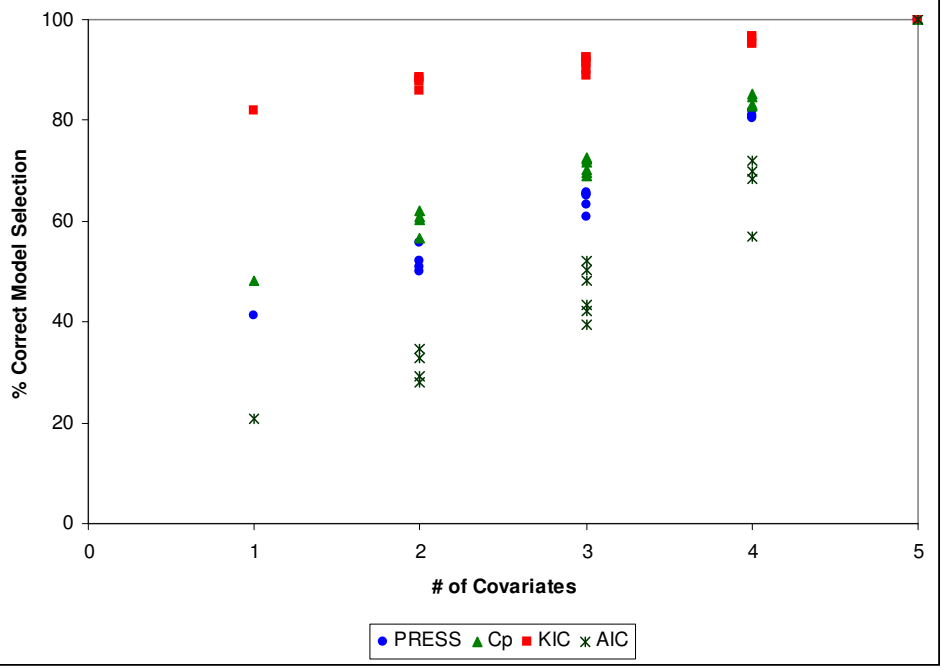


Table 4.6: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 200 subjects, 5 observations per subject Unstructured Covariance for Random Intercept and Slope

Fixed Effects for Simulated Data	# of Covariates	% Correct Selection by PRESS	C_p	KIC	AIC	BIC
$3+t+4x_1+3x_2+5x_3+5x_4$	5	30.1	100	100	100	100
$3+t+4x_1+3x_2+5x_3$	4	23.2	83.6	83.7	0.5	0.5
$3+t+4x_1+3x_2+5x_4$	4	22.0	83.9	88.3	0.2	0.2
$3+t+4x_1+5x_3+5x_4$	4	28.7	84.7	88.8	0.2	0.2
$3+t+3x_2+5x_3+5x_4$	4	24.5	83.7	88.4	0.3	0.3
$3+t+4x_1+3x_2$	3	20.9	70.5	74.4	0	0
$3+t+4x_1+5x_3$	3	22.8	70.2	73.5	0.1	0.2
$3+t+4x_1+5x_4$	3	21.2	71.3	75.5	0	0
$3+t+3x_2+5x_3$	3	22.3	71.9	75.8	0	0.1
$3+t+3x_2+5x_4$	3	20.6	70.6	78.1	0.2	0.2
$3+t+5x_3+5x_4$	3	24.2	68.9	76.5	0.2	0.2
$3+t+4x_1$	2	15.5	60.9	65.2	0.1	0.2
$3+t+3x_2$	2	15.0	59.3	65.0	0	0.1
$3+t+5x_3$	2	18.7	60.7	66.4	0	0.1
$3+t+5x_4$	2	17.9	59.9	69.3	0.1	0.1
$3+t$	1	15.6	53.0	62.1	0	0.1

Figure 4.14: % Correct Model Selection vs. # of Covariates Multiple Categorical Predictors (Random Intercept and Slope; Large Sample Size)

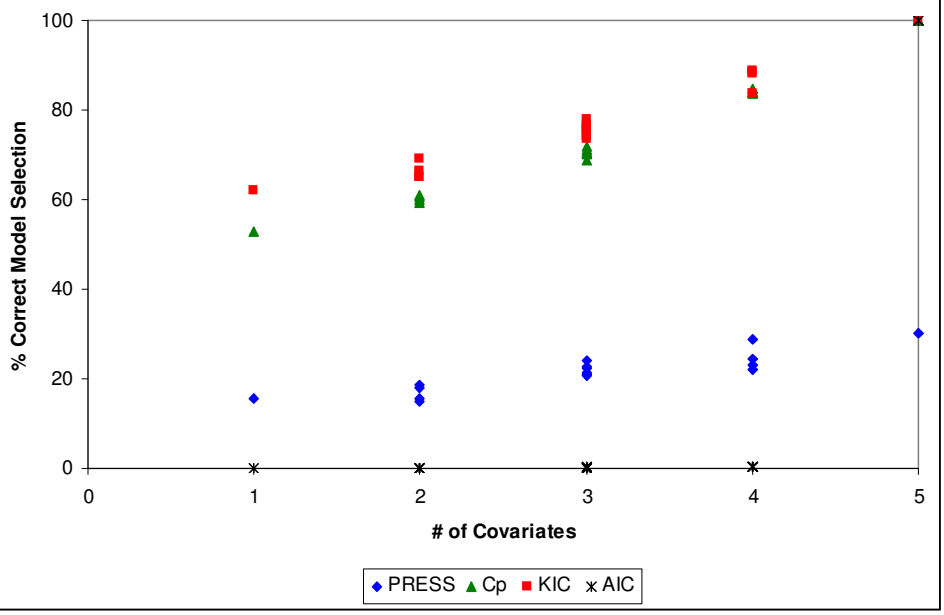


Table 4.7: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 50 subjects, 4 observations per subject Scalar Covariance for Random Intercept Only

Fixed Effects for Simulated Data	# of Covariates	% Correct Selection by PRESS	C_p	KIC	AIC	BIC
$3+t+4x_1+3x_2+5x_3+5x_4$	5	100	100	100	100	100
$3+t+4x_1+3x_2+5x_3$	4	80.8	86.0	86.5	0.1	0.1
$3+t+4x_1+3x_2+5x_4$	4	79.5	83.2	88.4	1.6	1.6
$3+t+4x_1+5x_3+5x_4$	4	79.4	84.0	88.7	1.1	1.1
$3+t+3x_2+5x_3+5x_4$	4	80.7	84.7	90.6	2.1	2.1
$3+t+4x_1+3x_2$	3	62.5	71.4	76.2	0	0
$3+t+4x_1+5x_3$	3	63.3	72.6	76.2	0	0
$3+t+4x_1+5x_4$	3	64.3	71.4	81.0	0.5	0.5
$3+t+3x_2+5x_3$	3	61.1	70.6	75.4	0	0
$3+t+3x_2+5x_4$	3	62.1	69.5	85.7	0.3	0.3
$3+t+5x_3+5x_4$	3	65.3	70.5	78.8	0.1	0.1
$3+t+4x_1$	2	50.1	59.0	66.7	0	0
$3+t+3x_2$	2	48.8	59.4	69.4	0	0
$3+t+5x_3$	2	49.9	59.8	68.0	0	0
$3+t+5x_4$	2	50.9	58.1	72.8	0.1	0.1
$3+t$	1	39.1	51.5	61.3	0	0

Figure 4.15: % Correct Model Selection vs. # of Covariates Multiple Categorical Predictors (Random Intercept Only; Small Sample Size)

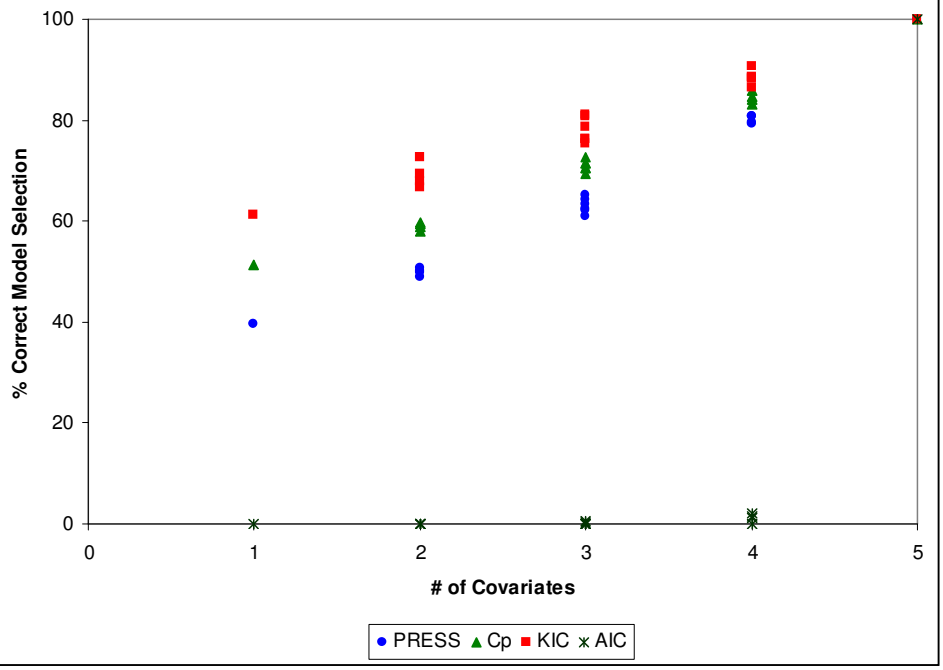
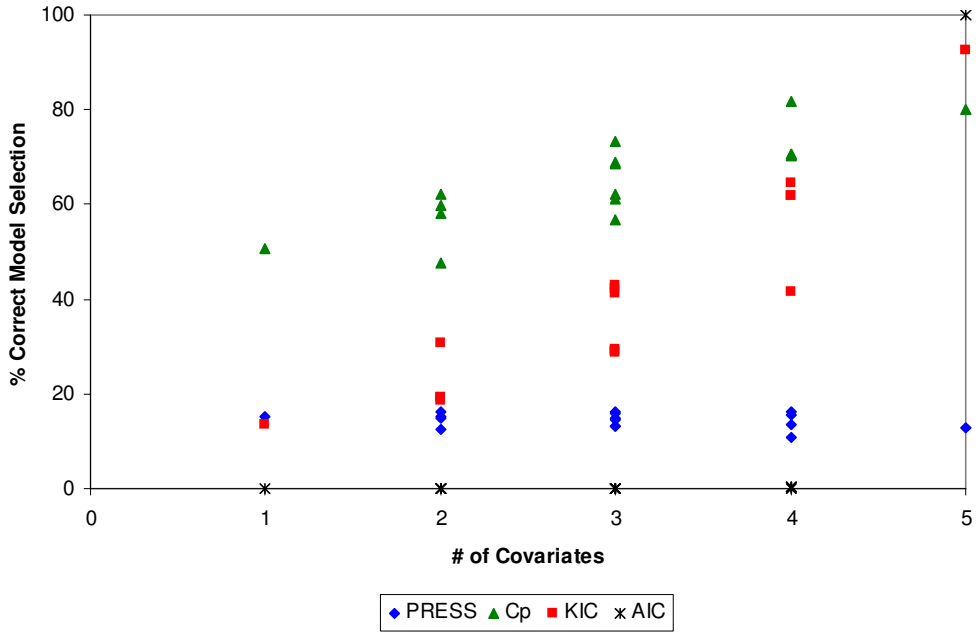


Table 4.8: Monte Carlo Assessment of Fixed Effects Model Selection: 1,000 datasets, 50 subjects, 4 observations per subject Unstructured Covariance for Random Intercept and Slope

Fixed Effects for Simulated Data	# of Covariates	% Correct Selection by PRESS	C_p	KIC	AIC	BIC
$3+t+4x_1+3x_2+5x_3+5x_4$	5	12.9	80.2	92.7	100	100
$3+t+4x_1+3x_2+5x_3$	4	10.9	70.4	41.5	0	0.1
$3+t+4x_1+3x_2+5x_4$	4	13.5	70.7	61.9	0.5	0.8
$3+t+4x_1+5x_3+5x_4$	4	16.2	81.8	61.7	0.2	0.4
$3+t+3x_2+5x_3+5x_4$	4	15.6	70.6	64.4	0.2	0.4
$3+t+4x_1+3x_2$	3	13.1	61.3	28.6	0	0
$3+t+4x_1+5x_3$	3	16.1	73.2	28.9	0	0
$3+t+4x_1+5x_4$	3	14.5	68.6	41.1	0	0
$3+t+3x_2+5x_3$	3	13.3	62.0	29.3	0	0
$3+t+3x_2+5x_4$	3	14.7	56.7	43.0	0	0
$3+t+5x_3+5x_4$	3	16.0	68.8	42.3	0	0
$3+t+4x_1$	2	15.1	59.8	19.1	0	0
$3+t+3x_2$	2	12.4	47.8	18.6	0	0
$3+t+5x_3$	2	14.9	62.0	19.2	0	0
$3+t+5x_4$	2	16.3	58.0	30.7	0	0
$3+t$	1	15.2	50.6	13.4	0	0

Figure 4.16: % Correct Model Selection vs. # of Covariates Multiple Categorical Predictors (Random Intercept and Slope; Small Sample Size)



4.3.4 Conclusions

In the scenarios explored here, we see that all three criteria investigated in this dissertation, namely the C_p statistic, the KIC and the PRESS statistic, outperform the standard AIC and BIC in almost every variation. The only time where the AIC and BIC outperform the other criteria is when there is an interaction term present.

In the first scenario, when the interaction term is not included in generating the data, the KIC is best in all situations. However, the C_p and the PRESS do fairly well in all of these situations as well. The performance of the KIC, C_p and PRESS do not seem to be affected by the changes in the variance in any of the situations investigated. On the other hand, the performance of the AIC and BIC decreases significantly as the total variance in the generating data increases.

When there is an interaction term present, the performance of the KIC, C_p and PRESS decreases as the total variance of the generating data increases. The opposite is true of the AIC and BIC, their performance increases as the variance increases. Another difference can be seen in the pattern of behavior as the correlation between measurements increase. For the KIC, C_p and PRESS, as the correlation increases, the ability of these criteria to identify the correct model increases. However, for the AIC and BIC, when the total variation is high and the correlation increases, the ability of the AIC and BIC to correctly identify the correct model decreases. As mentioned in Chapter 3, the AIC and BIC may have an advantage in the situation where the interaction term is present, as they tend to include extraneous variables, and in this situation, it is the model with the most variables that is correct. The difference in the patterns of behavior of the AIC and BIC compared to the other criteria seems to support this theory.

When there are multiple categorical predictors present, the KIC performs best in both of the large sample cases and in the small sample case when there is only a random intercept. However, both the C_p and PRESS statistic tend to do as well when only a random intercept is present. In fact, the C_p statistic appears to fare the best in this

scenario as it does as well or better than all the criteria in all four situations. On the other hand, the PRESS statistic is only able to detect the correct mean structure when there is only a random intercept, and the AIC and BIC can perform satisfactorily only when the sample size is large and there is only a random intercept.

From all simulation scenarios, it appears as though the PRESS statistic tends to behave similar to the C_p statistic in most of the variations investigated. However, the PRESS statistic's performance is poor compared to that of the C_p statistic in the second scenario, when multiple categorical predictors are present and the random effects consist of a random intercept and slope. In general, the PRESS statistic outperforms the AIC and BIC in all situations, except where there is an interaction effect present.

4.4 Example Data: Elderly Blood Pressure Data

4.4.1 Background

As in Chapters 2 and 3, we decided to explore a "real world" application of the PRESS statistic by applying it to the large data set from the North Carolina Established Populations for the Epidemiologic Studies of the Elderly (EPESE). As stated previously, the goals of the EPESE project were to describe and identify predictors of mortality, hospitalization, and placement in long-term care facilities and to investigate risk factors for chronic diseases and of functioning among the elderly. The study followed 4162 subjects, aged 65 years and older, over a period of 12 years. The more intricate details of the study population can be found in Chapter 2. It should be noted that due to the subject matter (i.e. the elderly) and timeline of this project, this large dataset contains a great deal of missing data. Participants were surveyed at four time periods: Wave 1(1986); Wave 2 (1990); Wave 3 (1994); and Wave 4 (1998). Here, we will focus on the outcome of diastolic blood pressure.

4.4.2 Methods

Originally, when planning to explore the ability of the PRESS statistic on this data, we expected to use the same all-possible regression approach that we used in Chapters 2 and 3. However, we found the calculation of the PRESS statistic by the SAS system to be so computationally intensive that performing this calculation for all 2048 possible models was not possible. Therefore, in order to evaluate the PRESS statistic in comparison to the C_p, KIC, AIC and BIC, we looked at the three best models as selected by the C_p statistic and the three best models selected by the KIC, AIC and BIC (all three criteria selected the same models). As both the C_p and the information criteria selected the saturated model as one of its three best models, there were a total of five models evaluated using the PRESS statistic. As was done in Chapter 3, the models were investigated using an i.i.d. within-unit error covariance with a random intercept (compound symmetry for the response). However, since the PRESS is computationally intensive, residual denominator degrees of freedom were used as opposed to the Kenward Roger denominator degrees of freedom.

4.4.3 Results

Table 4.9 shows the PRESS, C_p, KIC, AIC and BIC results for the 5 models that for which the PRESS statistic was calculated. Table 4.10 gives the fixed effect estimates, standard errors, p-values and covariance estimates for the models selected as best by either the PRESS, KIC, AIC and BIC or the C_p statistic.

Table 4.9: PRESS, $C_{p,}$ KIC, AIC and BIC Results for Elderly Diastolic Blood Pressure Data
Models with the lowest KIC and/or C_p values

(Model) Fixed Effects	PRESS	C_p	KIC	AIC	BIC
(1) year, weight, fair_ill, poor_ill, heart, diabet, black, rural, male, poor_hlth, married	$151.91x10^7$	11.17	72469.33	72397.35	72422.41
(2) year, weight, fair_ill, poor_ill, heart, diabet, black, rural, male, poor_hlth, married, fair_hlth	$151.96x10^7$	13.00	72472.99	72398.00	72423.05
(3) year, weight, fair_ill, poor_ill, heart, diabet, black, rural, poor_hlth, married	$152.90x10^7$	17.50	72474.40	72405.48	72430.54
(4) year, weight, fair_ill, poor_ill, heart, diabet, black, rural, male, poor_hlth	$154.31x10^7$	9.40	72643.04	72430.33	72449.12
(5) year, weight, fair_ill, poor_ill, heart, diabet, black, rural, male, poor_hlth, fair_hlth	$154.35x10^7$	11.23	73646.69	73607.69	73626.48

Table 4.10: Mixed Model Fixed Effect Estimates, Standard Errors (SE), p-values, and Covariance Estimates for the Models with the Lowest KIC and PRESS Value (Model 1) and the Model with the Lowest C_p Statistic Value (Model 4) (Outcome = Diastolic BP)

					Covariance Estimates	
	Fixed Effect	**Estimate**	**SE**	**p-value**	**Random effects**	**Error**
Model 1	Intercept	65.60	0.78	<0.001	$\widehat{\sigma}_1^2 = 45.99$	$\widehat{\sigma}_0^2 = 87.92$
	year	– 0.65	0.02	<0.001		
KIC	weight	0.04	0.005	<0.001		
72469.33	fair_ill	6.96	0.30	<0.001		
	poor_ill	9.88	0.41	<0.001		
PRESS	heart	– 4.88	0.40	<0.001		
151.91x10^7	diabet	– 3.96	0.38	<0.001		
	blackpat	2.63	0.31	<0.001		
C_p	poor_hlth	– 1.10	0.38	0.004		
11.17	rural	1.40	0.30	<0.001		
	male	1.05	0.36	0.004		
	married	– 0.18	0.32	0.578		
Model 4	Intercept	65.48	0.78	<0.001	$\widehat{\sigma}_1^2 = 45.72$	$\widehat{\sigma}_0^2 = 88.10$
	year	– 0.65	0.02	<0.001		
KIC	weight	0.04	0.005	<0.001		
72643.04	fair_ill	7.00	0.30	<0.001		
	poor_ill	9.90	0.40	<0.001		
PRESS	heart	– 4.82	0.40	<0.001		
$154.31x10^7$	diabet	– 3.97	0.38	<0.001		
	blackpat	2.67	0.31	<0.001		
C_p	poor_hlth	– 1.26	0.40	0.002		
9.40	rural	1.40	0.30	<0.001		
	male	0.94	0.34	<0.001		

4.4.4 Conclusions

From Table 4.9, we see that the PRESS statistic tends to order the models in the same way as the information criteria. The PRESS, KIC, AIC and BIC all considered Model 1, which includes all variables except the "fair_health" covariate, as best, while the C_p statistic chose Model 4, which excludes both the "fair_health" covariate and the marital status covariate (i.e. "married") as best.

From Table 4.10, we see that in Model 1, (the model chosen as 'best' by the PRESS, KIC, AIC and BIC) the "married" covariate appears to be extraneous to the outcome of diastolic blood pressure, as it is not significant at α = 0.05. On the other hand, Model 4, (the model chosen by C_p) does not appear to include any covariates that are not significant at α = 0.05.

While our exploration of the EPESE data using the PRESS statistic is not ideal, in that an all-possible regression approach could not be used, it appears as though the PRESS statistic arrives at the same conclusion as the KIC, AIC and BIC. This conclusion apears to include an extraneous covariate, and it seems as though of the criteria explored here, the C_p statistic may be best when our dataset is large, includes many covariates, and missing data, if our goal is to be as parsimonious as possible when choosing the correct model.

4.5 Discussion

From our exploration of the PRESS statistic in this chapter via simulation studies and a data example, we have discovered a great deal about this statistic's abilities and limitations in comparison to that of the C_p statistic and the information criteria available for the linear mixed model.

From our simulation studies, we see that the PRESS works about as well as the C_p statistic when the random effects are composed of only a random intercept. In general, as long as there is no interaction term in the fixed effects, the PRESS statistic performs

better than the current standard model selection criteria, the AIC and the BIC. The performance of the PRESS statistic was weakest in the multiple categorical predictor case when the random effects were composed of a random intercept and slope.

In the real data example, however, the PRESS seemed to behave more like the information criteria than the C_p statistic. From the limited information we have, the PRESS seemed to favor the same models that the KIC, AIC and BIC favored, rather than choosing the more parsimonious model chosen by the C_p statistic. This difference in behavior may have been due to a number of factors, including the presence of missing data, the larger sample size, or the large number of covariates being investigated.

The greatest limitation of the PRESS statistic is its computational intensity. From simple observation, we noticed that the PRESS statistic took up to five times longer than the information criteria when trying to determine the correct model from the same scenarios and this problem prevented us from being able to fully explore the statistic when using the EPESE study. This issue serves as a great hindrance in using the PRESS statistic as a model selection criterion in the linear mixed model.

References

Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics,* **16,** 125-127.

Liu, H., Weiss, R. E., Jennrich, R. I., & Wenger, N. S. (1999). PRESS model selection in repeated measures data. *Computational Statistics & Data Analysis*, **30**, 169-184.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W. (1996). Applied Linear Regression Models: Third Edition. USA: The McGraw-Hill Companies.

SAS Institute Inc. (2007), SAS (release 9.1.3), Cary, NC: SAS Institute Inc.

Schabenberger, O. (2005). Mixed model influence diagnostics. In: SUGI 29. SAS Institute, Paper 189–29.

CHAPTER 5

DISCUSSION AND FUTURE RESEARCH

5.1 Discussion

The increased use of repeated measures for longitudinal studies has resulted in the necessity for more research in the modeling of this type of data. While the linear mixed model is an extension of the linear univariate model that accounts for longitudinal data, few model selection methods have been developed or evaluated for the linear mixed model. In this dissertation, we investigated the behavior of three candidate model selection methods that have all been extended from the univariate linear model to the linear mixed model.

Mallows' C_p statistic was developed for the univariate linear model in 1964 and our extension of its use into the linear mixed model has shown that it can be a promising method for fixed effects selection. Of all the methods investigated in this dissertation, the C_p statistic we used gave us the most favorable results in terms of fixed effects selection. In addition to that, this statistic is the least computationally demanding of all the candidate methods in that it only requires one model to be fit in order to determine which subset of fixed effects is best for modeling the response.

The C_p statistic is limited in that it is a model selection method that requires nested models. In other words, the determination of how well a specific model fits the data using the C_p statistic can only be seen relative to the C_p statistic of another model

nested within the original saturated model. Also, when the number of potential covariates is large, the number of subsets of covariates that need to be considered can be overwhelming. This shortcoming is illustrated by the 2047 subsets of covariates that were evaluated in order to investigate the EPESE data set discussed in Chapter 2.

The KIC statistic explored in Chapter 3 appears to be promising as a model selection method for both fixed effects and covariance structure. In the selection of the correct covariance structure, the KIC tended to hold middle ground between the AIC and the BIC. While it was never the best method in terms of selection of covariance structure, it seems that it can be a useful method when one is completely unsure of what type of covariance structure is best. Unlike the AIC, which worked best when the covariance structure was complex, or the BIC, which excelled with a simple underlying covariance structure, the consistency of the KIC makes it appealing when aspects of the covariance structure are unknown.

In terms of fixed effects, the KIC appears to perform significantly better than either the AIC or BIC in the selection of fixed effects when there is no interaction effect present. It appears from our study of the selection of fixed effects that the KIC is far more conservative than the AIC and the BIC when it comes to allowing additional covariates into the model. In addition, the KIC was able to maintain its performance level better than the AIC and BIC when parameters in the covariance structure or random effects structure changed.

A shortcoming of the KIC, unlike the C_p, is that in order to determine which subset of fixed effects is best, each individual model must be fit. In addition, the value of the KIC, as with all information criteria, is only meaningful when compared to another model in the same family. For example, when looking for the correct set of fixed effects, KIC values must be compared with those of other models with the same covariance

structure. And if the covariance is structure is being evaluated, KIC values of models with the same mean structure must be compared.

The PRESS statistic has been developed for the linear mixed model in a widely available statistical software package (SAS), but its abilities as a model selection method were never evaluated. From our study, it appears that the PRESS statistic does not add much as a fixed effect selection method compared to the C_p or the KIC. The selection of fixed effects in the simulation studies appeared to mirror the abilities of the C_p statistic, except that the PRESS statistic appeared to be more susceptible to changes in covariance and random effects structures. However, when looking at the EPESE study data, the PRESS ranked models in a similar fashion as the information criteria, which selected a model that appears to have an extraneous covariate.

The greatest drawback of the PRESS statistic is its computational requirements. When initially trying out this statistic using the computations provided by SAS v9.1, I experimented with both the non-iterative and iterative analyses. I found that both analyses required the same amount of time and computational intensity, and therefore proceeded with the iterative analysis, hoping that it would provide the best results. The computation of the statistic required at least five times more time than the information criteria, which took significantly longer to calculate compared to the $C_{p.}$ In fact, this drawback prevented us from conducting the same level of research with the PRESS as was done with the C_p and the KIC. The fact that the PRESS requires this much computational power and does not provide results that far surpass the performance of the other candidate methods evaluated here leads us to believe that the PRESS is probably best used as an influence diagnostic and not as a more general model selection criterion.

When comparing the results of Chapter 2 to those of Chapter 4 involving the C_p statistic and the AIC and BIC, we notice that in Chapter 2, it appeared that the C_p statistic was better than the AIC and BIC at correctly identifying the fixed effects in all scenarios,

including when the correct model included an interaction term, while in Chapter 4, it was the AIC and BIC that performed better in this scenario. This difference in outcomes is probably due to the fact that in Chapter 2, the simulations involved smaller sample sizes of 50 subjects as opposed to the 200 subject sample size used in the similar simulation studies in Chapter 4. As we saw in Chapters 3 and 4, when looking at the multiple categorical predictor setting, the AIC and BIC work best when there is a larger sample size. The decrease in the sample size resulted in a dramatic decrease in these information criteria's abilities to detect the correct fixed effect structure.

With regards to the EPESE data set, it should be noted that the distribution of the variables investigated in the models were not evaluated beforehand to detect if they were normally distributed. As this is the case, transformations on this data to ensure normal distrbution may result in different results. In addition, using a more complex covariance structure with this data could allow for a closer fit.

Perhaps the most interesting development of this research is the revelation of how variable the information criteria that are available as standard output in most packages for linear mixed model computing (i.e. the AIC and BIC) performed in many of the simulation scenarios explored here. While investigators tend to accept these criteria as good measures of the fit of linear mixed models to their data, we see now that the results received when using these criteria should be taken with a grain of salt.

5.2 Future Research

The results of this dissertation shine a light on many different areas of further research for model selection in the linear mixed model. In addition to distributional aspects of some of the candidate methods explored here, further computational studies will also lead to important results.

The C_p statistic that was used in Chapter 2 was developed as a pure extension from the one developed by Mallows for the univariate linear model. Our exploration of this suggested statistic was purely empirical in nature and none of the distributional aspects were explored. Further research of the C_p statistic should include development of the distributional aspects of this statistic for the linear mixed model and correction terms that account for the complexity involved in using the linear mixed model.

The KIC statistic used in Chapter 3 was directly taken from the univariate linear model. Further research for this statistic would look at refining the criterion to take into account the number of parameters estimated when using the linear mixed model.

The PRESS statistic used in Chapter 4 could use a great deal of refinement. The development of a technique that would not be as computationally intensive, yet would yield the same or better results in terms of model selection would be a great service to this field. Christiansen et al. (1992), suggest a method of providing case-deletion diagnostics in the linear mixed model via one-step estimates of diagnostics for variance components. This technique may lead the way to the development of a less computationally intensive method of finding the PRESS statistic in the linear mixed model.

For all three criteria, further computational studies could include an investigation of model selection performance when the sample size is small, when missing data are present, and when there are multiple interaction terms involved in generating the data. It may also be beneficial to design further simulation studies based on the parameters of classic data sets, similar to what was done in the simple mean structure simulation studies performed in Chapter 2, where the simulations were based on the underlying fixed effect and covariance parameters of the Pothoff and Roy data. By looking at simulated variations based on real data, we may be able to get a better picture of how our criteria will perform in the real world. In addition, it would be useful to know how robust our statistics are when the data investigated are not normally distributed, and a further

investigation of their performance in a wider variety of covariance structures would also be useful.

As the use of longitudinal studies continue to grow in popularity, it is increasingly important that we refine and discover additional tools that can be useful in confirming the models we choose accurately describe the data we receive from these studies.

REFERENCES

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, **21**, 243.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**, 716.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, **30**, 9.

Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics,* **16,** 125-127.

Blazer, D.G., and George, L.K. (2004). Established Populations for Epidemiologic Studies of the Elderly, 1996-1997: Piedmont Health Survey of the Elderly, Fourth In-Person Survey [Durham, Warren, Vance, Granville, and Franklin Counties, North Carolina] [Computer file]. ICPSR version. Bethesda, MD: United States Department of Health and Human Services. National Institutes of Health. National Institute on Aging [producer], 1999. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004.

Cantoni, E., Flemming, J. M., & Ronchetti, E. (2005). Variable selection for marginal longitudinal generalized linear models. *Biometrics*, **61**, 507-514.

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters*, **42**, 333-343.

Cavanaugh, J. E. (2004). Criteria for linear model selection based on Kullback's symmetric divergence. *Australian & New Zealand Journal of Statistics*, **46**, 257-274.

Christiansen, R., Pearson L.M., Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics,***34**, 38-45.

Davidian, M (2001). ST 732: Applied Longitudinal Data Analysis, Lecture Notes. North Carolina State University.

Demidenko E. (2004) Mixed Models—Theory and Application. Wiley: New York.

Fujikoshi, Y., & Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707-716.

Edwards, L.J., Muller, K.E., Wolfinger, R.D., Qaqish, B.F., Schabenberger, O. (2008). An R^2 statistic for fixed effects in the linear mixed model. *In Press: Statistics in Medicine*.

Fitzmaurice G, Davidian M, Molenberghs G, Verbeke G. (eds) (2008). Longitudinal Data Analysis: A Handbook of Modern Statistical Methods. Chapman & Hall/CRC: Boca Raton, Florida.

Gilmour, S. G. (1996). The interpretation of Mallows's C_p-statistic. *The Statistician,* **45**, 49-56.

Gomez, E., Schaalje, G., & Fellingham, G. E. -. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics: Simulation and Computation*, **34**, 377-392.

Gorman, J. W., & Toman, R. J. (1966). Selection of variables for fitting equations to data. *Technometrics*, 8(1), 27-51.

Gurka, M. J. (2006), "Selecting the Best Linear Mixed Model under REML," *The American Statistician*, **60**, 19-26.

Hafidi, B., & Mkhadri, A. (2006). A corrected Akaike criterion based on Kullback's symmetric divergence: Applications in time series, multiple and multivariate regression. *Computational Statistics & Data Analysis*, **50**, 1524-1550.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, **41**, 190-195.

Hocking, R. R. (1976). A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, **32**, 1-49.

Hurvich, C. M., Shumway, R., & Tsai, C. (1990). Improved estimators of kullback-leibler information for autoregressive model selection in small samples. *Biometrika*, **77**, 709-719.

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

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983-997.

Kim, H., and Cavanaugh, J. E. (2005). Model selection criteria based on Kullback information measures for nonlinear regression. *Journal of Statistical Planning and Inference*, **134**, 332-349.

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

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Liu, H., Weiss, R. E., Jennrich, R. I., & Wenger, N. S. (1999). PRESS model selection in repeated measures data. *Computational Statistics & Data Analysis*, **30**, 169-184.

Mallows, C.L. (1964). "Choosing Variables in a Linear Regression: A Graphical Aid," unpublished paper presented at the Central Regional Meeting of the Institute of Mathematical Statistics. Manhattan, KS, May 7-9.

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

McQuarrie, A. D. R., and Tsai, C.-L. (1998). Regression and Time Series Model Selection. World Scientific.

Muller, K., and Fetterman, B. (2002). Regression and ANOVA: An integrated approach using SAS software. Cary, NC, USA: SAS Institute Inc.

Muller, K., & Stewart, P. (2006). Linear Model Theory: Univariate, Multivariate, and Mixed Models. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W. (1996). Applied Linear Regression Models: Third Edition. USA: The McGraw-Hill Companies.

Potthoff, R. F., Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313-326.

Ronchetti, E., & Staudte, R. G. (1994). A robust version of Mallows's C_p. *Journal of the American Statistical Association*, **89**, 550-559.

SAS Institute Inc. (2007), SAS (release 9.1.3), Cary, NC: SAS Institute Inc.

Schabenberger, O. (2005). Mixed model influence diagnostics. In: SUGI 29. SAS Institute, Paper 189–29.

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

Sugiura, N. (1978). Further analysts of the data by akaike' s information criterion and the finite corrections. *Communications in statistics. Theory and methods*, **7**, 13.

Vonesh, E.F., and Chinchilli, V.M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker.