

Fall 2011

Criteria for generalized linear model selection based on Kullback's symmetric divergence

Cristina Laura Acion
University of Iowa

Copyright 2011 Laura Acion

This dissertation is available at Iowa Research Online: <https://ir.uiowa.edu/etd/2665>

Recommended Citation

Acion, Cristina Laura. "Criteria for generalized linear model selection based on Kullback's symmetric divergence." PhD (Doctor of Philosophy) thesis, University of Iowa, 2011.
<https://doi.org/10.17077/etd.xibxpnc3>

Follow this and additional works at: <https://ir.uiowa.edu/etd>

Part of the [Biostatistics Commons](#)

CRITERIA FOR GENERALIZED LINEAR MODEL SELECTION BASED ON
KULLBACK'S SYMMETRIC DIVERGENCE

by

Cristina Laura Acion

An Abstract

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

December 2011

Thesis Supervisor: Professor Joseph Cavanaugh

ABSTRACT

Model selection criteria frequently arise from constructing estimators of discrepancy measures used to assess the disparity between the data generating model and a fitted approximating model. The widely known Akaike information criterion (AIC) results from utilizing Kullback’s directed divergence (KDD) as the targeted discrepancy. Under appropriate conditions, AIC serves as an asymptotically unbiased estimator of KDD. The directed divergence is an asymmetric measure of separation between two statistical models, meaning that an alternate directed divergence may be obtained by reversing the roles of the two models in the definition of the measure. The sum of the two directed divergences is Kullback’s symmetric divergence (KSD).

A comparison of the two directed divergences indicates an important distinction between the measures. When used to evaluate fitted approximating models that are improperly specified, the directed divergence which serves as the basis for AIC is more sensitive towards detecting overfitted models, whereas its counterpart is more sensitive towards detecting underfitted models. Since KSD combines the information in both measures, it functions as a gauge of model disparity which is arguably more balanced than either of its individual components. With this motivation, we propose three estimators of KSD for use as model selection criteria in the setting of generalized linear models: KIC_o , KIC_u , and QKIC. These statistics function as asymptotically unbiased estimators of KSD under different assumptions and frameworks.

As with AIC, KIC_o and KIC_u are both justified for large-sample maximum likelihood settings; however, asymptotic unbiasedness holds under more general assumptions for KIC_o and KIC_u than for AIC. KIC_o serves as an asymptotically unbiased estimator of KSD in settings where the distribution of the response is

misspecified. The asymptotic unbiasedness of KIC_u holds when the candidate model set includes underfitted models.

QKIC is a modification of KIC_o . In the development of QKIC, the likelihood is replaced by the quasi-likelihood. QKIC can be used as a model selection tool when generalized estimating equations, a quasi-likelihood-based method, are used for parameter estimation.

We examine the performance of KIC_o , KIC_u , and QKIC relative to other relevant criteria in simulation experiments. We also apply QKIC in a model selection problem for a randomized clinical trial investigating the effect of antidepressants on the temporal course of disability after stroke.

Abstract Approved: _____
Thesis Supervisor

Title and Department

Date

CRITERIA FOR GENERALIZED LINEAR MODEL SELECTION BASED ON
KULLBACK'S SYMMETRIC DIVERGENCE

by

Cristina Laura Acion

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

December 2011

Thesis Supervisor: Professor Joseph Cavanaugh

Copyright by
CRISTINA LAURA ACION
2011
All Rights Reserved

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

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

Cristina Laura Acion

has been approved by the Examining Committee for the
thesis requirement for the Doctor of Philosophy degree in
Biostatistics at the December 2011 graduation.

Thesis Committee: _____

Joseph Cavanaugh, Thesis Supervisor

Stephan Arndt

Ricardo Jorge

Joseph Lang

Jane Pendergast

To My Father.

ACKNOWLEDGMENTS

This work would not have been possible without the abundant and generous help of many people that will stay in my heart for life. Each of the persons named or implied below lent me their hand repeatedly during the years that took to complete this challenging part of my life.

First of all I want to thank the patience, understanding, generosity, and inspirational advice from Prof. Joseph Cavanaugh. He guided me through every single step that I took to see the light at the end of this dissertation.

I would not have achieved this point in my education without the help of Prof. Stephan Arndt. He started it all by inviting me into the Biostatistics program back in 2002 and, since then, never stopped helping me advance my career.

My thanks go also to Prof. Ricardo Jorge and his wife Ms. Marcela Casiraghi for their unconditional support these years.

I want to thank as well the constant aid of all the people at the Departments of Biostatistics and Statistics, in particular, that of Prof. Kathryn Chaloner, Prof. Jane Pendergast, Prof. Joseph Lang, and Ms. Terry Kirk.

This journey was pleasant because at different times I shared it with my colleagues and friends at the Departments of Biostatistics and Statistics Ms. Kelly Bach, Dr. Emine Bayman, Ms. Jaysri Buttler, Dr. Yu-Hui Chang, Ms. Mijin Jang, Mr. Eric Schaefer, and Prof. Qian Cicci Shi.

I would not have got here without my dear friends in Argentina: Ms. Inés Abelló, Mr. Pablo Alí, Dr. Marcela Borinsky, Prof. Débora Burin, Ms. Paola Lefer, Mr. Maximiliano Sacco, and Mr. Alejandro Weil, as well as those in the United States: Ms. Erin Arndt, Ms. Mary Hansman, Ms. Kyla Kennedy, Dr. Blanca Márquez de Prado, Ms. Elizabeth Smothers, and Dr. Luis Tecedor. I want to thank them all for being there for me regardless if it was good, bad or ugly.

These years I learned that there is no career without parents that raised their children wholly. I am proud of being Mr. Jorge Ación's and Ms. Esther Gatti's daughter. They made me who I am. They taught me through example the values of sacrifice, perseverance and hard work among many others. ¡Muchas gracias, Papá! ¡Muchas gracias, Mamá!

Last, but not least, I would like to mention Mr. Daniel Xifra, my husband. He has been right next to me during almost all my brightest and darkest times these years. I want to thank him for making life wonderful in spite of any Ph.D.-related or unrelated hardship.

ABSTRACT

Model selection criteria frequently arise from constructing estimators of discrepancy measures used to assess the disparity between the data generating model and a fitted approximating model. The widely known Akaike information criterion (AIC) results from utilizing Kullback’s directed divergence (KDD) as the targeted discrepancy. Under appropriate conditions, AIC serves as an asymptotically unbiased estimator of KDD. The directed divergence is an asymmetric measure of separation between two statistical models, meaning that an alternate directed divergence may be obtained by reversing the roles of the two models in the definition of the measure. The sum of the two directed divergences is Kullback’s symmetric divergence (KSD).

A comparison of the two directed divergences indicates an important distinction between the measures. When used to evaluate fitted approximating models that are improperly specified, the directed divergence which serves as the basis for AIC is more sensitive towards detecting overfitted models, whereas its counterpart is more sensitive towards detecting underfitted models. Since KSD combines the information in both measures, it functions as a gauge of model disparity which is arguably more balanced than either of its individual components. With this motivation, we propose three estimators of KSD for use as model selection criteria in the setting of generalized linear models: KIC_o , KIC_u , and QKIC. These statistics function as asymptotically unbiased estimators of KSD under different assumptions and frameworks.

As with AIC, KIC_o and KIC_u are both justified for large-sample maximum likelihood settings; however, asymptotic unbiasedness holds under more general assumptions for KIC_o and KIC_u than for AIC. KIC_o serves as an asymptotically unbiased estimator of KSD in settings where the distribution of the response is

misspecified. The asymptotic unbiasedness of KIC_u holds when the candidate model set includes underfitted models.

QKIC is a modification of KIC_o . In the development of QKIC, the likelihood is replaced by the quasi-likelihood. QKIC can be used as a model selection tool when generalized estimating equations, a quasi-likelihood-based method, are used for parameter estimation.

We examine the performance of KIC_o , KIC_u , and QKIC relative to other relevant criteria in simulation experiments. We also apply QKIC in a model selection problem for a randomized clinical trial investigating the effect of antidepressants on the temporal course of disability after stroke.

TABLE OF CONTENTS

LIST OF TABLES	ix
LIST OF FIGURES	xii
CHAPTER	
1 INTRODUCTION	1
1.1 Model Selection Principles	2
1.1.1 Generalizability	2
1.1.2 Goodness-of-Fit versus Parsimony	3
1.1.3 Over and Under Specification	4
1.1.4 Variability and Bias	5
1.2 Introduction to Model Selection Criteria	6
1.3 Introduction to Kullback's Divergences	8
1.4 Research Goals	8
1.5 Relevant Literature Review	9
1.6 Outline	14
2 PRELIMINARY CONCEPTS	16
2.1 Relevant Parameter Estimation and GLM Concepts	16
2.1.1 Likelihood-Based Estimation for Independent Responses	17
2.1.2 Notation and Basic GLM Concepts	20
2.1.3 Quasi-Likelihood-Based Estimation for Correlated Responses	24
2.2 Model Selection Concepts	27
2.2.1 Kullback's Directed and Symmetric Divergences	27
2.2.2 Kullback's Discrepancies	33
2.2.3 Likelihood-Based Model Selection Criteria: AIC, TIC and KIC	34
2.2.4 Quasi-Likelihood-Based Model Selection Criterion: QIC	38
3 LIKELIHOOD-BASED MODEL SELECTION CRITERIA DERIVED FROM KULLBACK'S SYMMETRIC DIVERGENCE	41
3.1 KIC_o Derivation	41
3.2 KIC_u Derivation	46
4 SIMULATION STUDIES FOR KIC_o AND KIC_u	52
4.1 Selection of a Model with Correctly Specified Mean Structure	54
4.1.1 Linear Regression with Correctly Specified Error Distribution and with Incorrectly Specified Error Distribution	55
4.1.2 Linear Regression with Collinear Regressors	62

4.1.3 Logistic Regression without Overdispersion and with Ignored Overdispersion	69
4.1.4 Poisson Regression without Overdispersion and with Ignored Overdispersion	76
4.2 Selection of a Model with Optimal Predictive Properties.	83
4.2.1 Linear Regression.	84
4.3 General Conclusions	85
 5 QUASI-LIKELIHOOD-BASED MODEL SELECTION CRITERIA FOR CORRELATED RESPONSE DATA DERIVED FROM KULLBACK'S SYMMETRIC DIVERGENCE	87
5.1 QKIC Derivation.	87
5.2 Simulation Studies for QKIC	93
5.2.1 Selection of Working Correlation Structure for Correlated Binary Response Data	94
5.2.2 Selection of Mean Structure	95
5.3 Application.	102
5.3.1 Study Description	103
5.3.2 Comparison of Model Selection Criteria	103
 6 CONCLUSIONS AND FUTURE DIRECTIONS	107
6.1 Conclusions	107
6.2 Limitations and Future Directions	108
 REFERENCES	110

LIST OF TABLES

Table

2.1	Average values for $\hat{\sigma}^2$, \hat{Q} , KDD and KSD, and KDD and KSD number of model selections for 10,000 samples; results for the generating model are bolded.	31
4.1	Settings 1 and 2: NM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the linear regression framework; results for the generating model are bolded.	57
4.2	Settings 3 and 4: NM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the linear regression framework; results for the generating model are bolded.	58
4.3	Settings 5 and 6: NM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the linear regression framework; results for the generating model are bolded.	59
4.4	Settings 1 and 2: APM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the linear regression framework. . .	60
4.5	Settings 3 and 4: APM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the linear regression framework. . .	60
4.6	Settings 5 and 6: APM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the linear regression framework. . .	61
4.7	Settings 7 and 8: NM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the logistic regression framework; results for the generating model are bolded.	72
4.8	Settings 9 and 10: NM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the logistic regression framework; results for the generating model are bolded.	73
4.9	Settings 11 and 12: NM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the logistic regression framework; results for the generating model are bolded.	74
4.10	Settings 7 and 8: APM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the logistic regression framework. . .	74

4.11	Settings 9 and 10: APM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the logistic regression framework. . .	75
4.12	Settings 11 and 12: APM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the logistic regression framework. . .	76
4.13	Settings 13 and 14: NM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the Poisson regression framework; results for the generating model are bolded.	79
4.14	Settings 15 and 16: NM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the Poisson regression framework; results for the generating model are bolded.	80
4.15	Settings 17 and 18: NM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the Poisson regression framework; results for the generating model are bolded.	81
4.16	Settings 13 and 14: APM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the Poisson regression framework. .	82
4.17	Settings 15 and 16: APM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the Poisson regression framework. .	82
4.18	Settings 17 and 18: APM order selections for AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the Poisson regression framework; results for the generating model are bolded.	83
4.19	Settings 19 to 22: Average MSEP for models selected by AIC, TIC, KIC, KIC_o , KIC_u and their oracles.	85
5.1	Settings 23 and 24: Working correlation structure selections for QIC^R and $QKIC^R$ in the binary correlated response data framework; results for the generating model are bolded.	95
5.2	Settings 25 and 26: NM order selections for QIC^I , $QKIC^I$, QIC^R , $QKIC^R$, QIC^U , and $QKIC^U$ in the binary correlated response data framework; results for the generating model are bolded.	97
5.3	Setting 27: NM order selections for QIC^I , $QKIC^I$, QIC^R , $QKIC^R$, QIC^U , and $QKIC^U$ in the binary correlated response data framework; results for the generating model are bolded.	98
5.4	Setting 28: NM order selections for QIC^I , $QKIC^I$, QIC^R , $QKIC^R$, QIC^U , and $QKIC^U$ in the binary correlated response data framework; results for the generating model are bolded.	99

5.5	Settings 29 and 30: NM order selections for QIC^I , QKIC^I , QIC^R , QKIC^R , QIC^U , and QKIC^U in the Poisson correlated response data framework; results for the generating model are bolded.	101
5.6	Settings 31 and 32. NM order selections for QIC^I , QKIC^I , QIC^R , QKIC^R , QIC^U , and QKIC^U in the Poisson correlated response data framework; results for the generating model are bolded.	102
5.7	QIC^R and QKIC^R rankings for working correlation structure candidate models.	104
5.8	QIC^I , QKIC^I , QIC^R , QKIC^R , QIC^U , and QKIC^U rankings for mean structure candidate models.	105

LIST OF FIGURES

Figure

2.1	Three-dimensional plots comparing KDD and KSD	30
4.1	Model selection criteria comparison with correlated regressors, $n = 30$, no misspecified errors	65
4.2	Model selection criteria comparison with correlated regressors, $n = 30$, with ignored misspecified errors	66
4.3	Model selection criteria comparison with correlated regressors, $n = 60$, no misspecified errors	67
4.4	Model selection criteria comparison with correlated regressors, $n = 60$, with ignored misspecified errors	68

CHAPTER 1

INTRODUCTION

A statistical model can be defined as a simplified or idealized description of a phenomenon, generally cast in probabilistic terms. Models help us understand phenomena, extract information, validate, predict and draw inferences. Modeling plays a critical role in scientific discovery. Hence, statistical modeling is one of the main activities for applied statisticians.

One of the fundamental problems in modeling data is that of selecting an appropriate model from a potentially large set of candidates. Choosing the model that best characterizes the data requires the determination of an appropriate and explicit structural form for the model. Improper model specification may substantially affect both the model estimators and predictors, not to mention scientific interpretations.

Consider a collection of data that has been generated according to an unknown parametric model. The model that gave rise to the data is called the *generating model*.

In traditional modeling, outcome data are represented using random variables. A parametric model implies a probability distribution for these random variables, where the parameters of the distribution may be defined as functions of explanatory variables. A model where the parameters are estimated from the observed data is referred to as a *fitted model*.

The goal of statistical modeling is to find a fitted model that provides the “best” approximation to the generating model. To this end, a parametric set of *candidate models* is often proposed that contains a collection of models with various structures. A candidate model is a model that could potentially be used to describe the data.

Model selection criteria are statistical tools that help identify the “best” fitted candidate model among the set of candidates. Inspired by the seminal work by Hirotugu Akaike (1927-2009), investigators have designed model selection criteria for different types of data and statistical frameworks and under different sets of assumptions. Many of the criteria developed are well justified when modeling continuous, independent, normally distributed outcomes. The problem of adequately modeling a phenomenon is more intricate for outcomes without these characteristics.

1.1 Model Selection Principles

An optimal statistical model is characterized by the principles of generalizability, goodness-of-fit and parsimony. Failure to conform with these principles can lead to improper model specification, affecting both the estimators of the model parameters and the predictors of the response variable. In this section, we discuss the fundamental modeling principles that determine model quality.

1.1.1 Generalizability

The principle of generalizability refers to the ability of a model to explain new data. A generalizable model is capable of describing or predicting future observations as accurately as possible. Akaike (1974, 1985) believed that the purpose of statistical modeling should be that of predicting new data as opposed to precisely characterizing the true model that generated the data. With an infinite amount of data or with data that are noiseless, Konishi and Kitagawa (2008) explain that a model designed to predict new data does not differ from one intended to mirror the structure of the data generating model. However, in usual practice, with finite sample sizes and noisy data, the differences between these two types of models can be substantial.

In some instances, the existence of a generating model that is accessible and amenable to estimation is often questioned (Burham and Anderson, 2002; Konishi

and Kitagawa, 2008). For example, in the life sciences, the phenomenon of interest can be so complex that any statistical model is necessarily a major simplification of reality. In spite of this limitation, striving for generalizability is one of the main model selection objectives.

1.1.2 Goodness-of-Fit versus Parsimony

Goodness-of-fit refers to the extent to which a fitted candidate model will conform to the data used in the model construction. With many model selection criteria, a goodness-of-fit term includes a measure that reflects the discrepancy between the observed outcomes employed to construct the model and their expected values under the fitted model. It may be possible to fit a model that conforms to the data very well, but does so because the model is excessively complicated and possibly difficult to interpret. For this reason, the principle of parsimony must also be considered.

The idea of parsimony relates directly to Occam’s razor, which is a principle credited to the medieval English philosopher William of Ockham (1285-1349). Occam’s Razor may be stated as follows: given two or more competing explanations for a phenomenon, none of which can be discounted, the simplest explanation is to be preferred. Occam’s razor recommends that we “shave off” extraneous ideas to better reveal the truth. Relating this principle to model selection, within a candidate collection of fitted models, the simplest model that adequately fits the data should be preferred. A key objective in model selection is achieving a balance between goodness-of-fit and parsimony. As Albert Einstein stated, “Everything should be made as simple as possible, but not simpler.”

1.1.3 Over and Under Specification

The concepts of under and over specification are also pertinent in determining the quality of a model. Both concepts are defined in terms of the generating model. Suppose the generating model belongs to the set of candidate models.

A candidate model that has the same structure as the generating model is called *correctly specified*. The resulting fitted candidate model would be *correctly fit*.

A candidate model that provides an incomplete representation of the generating model, perhaps because it does not include necessary variables, is called *underspecified*. The resulting fitted candidate model would be *underfitted*. Choosing such a model is referred to as *underfitting*.

A candidate model that is more complex than the generating model (e.g., one that contains extraneous variables) is called *overspecified*. The resulting fitted candidate model would be *overfitted*. Choosing such a model is referred to as *overfitting*.

In a practical setting, where researchers do not have access to the generating model, overfitting and underfitting can be thought of in terms of the best fitted candidate model. An underfitted model will fail to include the important variables, while an overfitted model will contain all the important variables as well as some spurious ones.

Both underfitting and overfitting can lead to problems in statistical modeling. Underfitting may lead to results that are biased while overfitting may lead to results with unnecessarily high variability. Burnham and Anderson (2002, p 17), in reference to Shibata (1989), state “While one must worry about errors due to both underfitting and overfitting, it seems that modest overfitting is less damaging than underfitting.” This may be conceptualized by realizing that it may be less damaging to additionally include an irrelevant variable in a correctly specified model

(i.e., overfitting), whereas the failure to include an important variable in a model (i.e., underfitting) could be more problematic. In the next section we develop these notions further.

1.1.4 Variability and Bias

As mentioned in the previous section, a key objective in model selection is achieving a balance between goodness-of-fit and parsimony. This objective is similar to balancing bias and variability. Technically speaking, a parameter estimate is biased when its expected value differs from the true parameter value. For instance, if a candidate model fails to include all of the relevant variables in the generating model, the estimates for the included parameters will be biased. Intuitively speaking, the parameters included in the model will “absorb” in an erratic way the missing effects, and will not correctly represent the corresponding effects in the generating model.

On the other hand, if a candidate model includes more variables than that which are required to explain the data, extraneous parameters will be estimated. This will not result in biased parameter estimates, but rather in parameter estimates that are excessively variable. That is, if we collected many different samples to study the same phenomenon, the parameter estimates resulting from each sample will vary more when the model includes parameters that are not necessary. In such a setting, we will be using a finite amount of data to estimate more parameters than that which are actually needed, and as a result, we will have less evidence per parameter than the evidence we would have for the adequate number of parameters. The larger the variability, the less precise the parameter estimates.

In summary, from a theoretical standpoint, a goal in model selection is to

choose a candidate model that “best” approximates the generating model. The concept of “best” implies properly achieving the objectives of generalizability, goodness-of-fit and parsimony. In many statistical modeling applications, especially those in the biomedical and health sciences, the notion of having access to the generating model is somewhat difficult to defend. However, in theoretical frameworks, it must be acknowledged that there is a probabilistic mechanism that generated the data. From a practical point of view, a more realistic goal in model selection is to attempt to capture the most salient features of the generating model using the best fitted candidate model. As George Box famously stated, “All models are wrong, some are useful.” Model selection criteria are the statistical tools designed to choose a useful model.

1.2 Introduction to Model Selection Criteria

Model selection criteria are statistical instruments that serve the purpose of choosing a suitable statistical model from a candidate class. A researcher studying a phenomenon often postulates different mechanisms that could explain the phenomenon. The different mechanisms hypothesized usually generate a group of candidate models that could serve as viable characterizations of the data. Model selection criteria are used to assign scores to each of the fitted candidate models in order to assist the data analyst in selecting a good model; that is, a model that conforms to the principles explained in the previous section.

Different criteria were developed under different assumptions and for different statistical frameworks and data types. Akaike pioneered the work in this area. Indeed, the Akaike information criterion (AIC) (Akaike, 1973) remains the most widely known and used model selection criterion. AIC is applicable in a broad array of modeling frameworks, since its justification primarily relies upon conventional large-sample properties of maximum likelihood estimators. A short list of

other popular model selection criteria includes the Bayesian information criterion (BIC), the corrected Akaike information criterion (AIC_c) and Mallows' conceptual predictive statistic (C_p).

BIC, also known as Schwarz information criterion, was introduced by Schwarz (1978) as a competitor to AIC. BIC was justified for the case of independent, identically distributed observations, and linear models, under the assumption that the likelihood is from the regular exponential family. The use of BIC seems justifiable for model screening in large-sample Bayesian analyses. However, BIC is often employed in frequentist analyses. Some frequentist practitioners prefer BIC to AIC because BIC tends to choose fitted models that are more parsimonious than those favored by AIC.

AIC_c was first suggested for normal linear regression by Sugiura (1978). Hurvich and Tsai (1989) demonstrated the small-sample superiority of AIC_c over AIC, and justified the use of AIC_c in the frameworks of nonlinear regression and autoregressive models. In the last 20 years, AIC_c has been extended to a number of additional modeling frameworks (e.g., autoregressive moving-average models, multivariate linear regression models, models for longitudinal data analysis under the assumption of a known covariance structure, etc.).

C_p was introduced by Mallows (1973) as a screening tool in multiple linear regression analyses. AIC and C_p are asymptotically equivalent: in large-sample settings, the two criteria will select the same fitted model from a candidate family.

Model selection criteria are often developed by constructing estimators of oracle measures that quantify the separation between the generating model and a fitted model. These measures are considered oracles because they have access to the “truth” (i.e., the generating model). C_p is derived as an estimator of an oracle called the Gauss discrepancy, while AIC and AIC_c serve as estimators of another oracle based on Kullback's directed divergence (KDD).

1.3 Introduction to Kullback’s Divergences

Kullback’s directed divergence (Kullback and Leibler, 1951), also known as the Kullback-Leibler information, the I -divergence, or the relative entropy, is one of many possible oracles used to design model selection criteria. Other oracles include the Kolmogorov discrepancy, the Cramer-von Mises discrepancy, the Pearson chi-squared discrepancy, the Neyman chi-squared discrepancy and the Gauss discrepancy (Linhart and Zucchini, 1986).

KDD is an asymmetric disparity measure, meaning that an alternative directed divergence can be obtained by reversing the roles of the two models in the definition of the measure. The sum of the two directed divergences is Kullback’s symmetric divergence (KSD), also known as the J -divergence.

In the framework of linear models, a comparison of the two directed divergences indicates an important distinction between the measures. When used to evaluate fitted approximating models that are improperly specified, the directed divergence which serves as a basis for AIC is more sensitive towards detecting overfitted models, whereas its counterpart is more sensitive towards detecting underfitted models (Cavanaugh, 2004). Since KSD reflects the sensitivities of both directed divergences, it functions as a discrepancy measure which is arguably more balanced than either of its individual components. With this motivation, KSD is the oracle we choose as a basis for the development of the criteria presented in this dissertation.

1.4 Research Goals

Some model selection criteria were developed under very general assumptions and can be used in a wide range of statistical settings (e.g., AIC, BIC, etc.). Other criteria (e.g., C_p , AIC_c , etc.) were initially developed only in a particular modeling framework. The applicability of such criteria in other settings requires further justification.

We are interested in general model selection tools. Thus, we will focus on generalized linear models (GLMs) using both likelihood-based analytical frameworks (e.g., normal, logistic, and Poisson regressions, etc.) and frameworks based on quasi-likelihood and generalized estimating equations (GEEs), suitable for modeling correlated binary or count response data.

Our goal is to develop and investigate model selection criteria based on KSD for use in GLM frameworks with independent or correlated responses. We expect that the criteria we propose will improve upon model selection criteria based on KDD by reflecting the increased sensitivity of KSD over KDD as a disparity measure.

1.5 Relevant Literature Review

In this section, we present the literature on existing criteria for modeling independent and correlated data under some of the modeling options available in the GLM framework. We also introduce previously developed model selection criteria based on using KSD as an oracle.

A review of the model selection tools available within the GLM framework is daunting, as this is a broad framework for which a variety of model selection tools has been designed from various perspectives. Common model selection tools developed for this framework include modifications of AIC, Bayesian-inspired methodologies, and AIC-like approaches combined with resampling techniques such as cross validation and bootstrapping. Since the criteria proposed in this dissertation are based on estimators of KSD and KDD and do not entail resampling techniques, we only review criteria relevant to the work presented in the following chapters.

Likelihood-based GLMs can be used to model independent continuous, binary, count or nominal data. In this framework, any model selection criterion can be used that is primarily justified by assuming the traditional large-sample properties of maximum likelihood estimation. AIC (Akaike, 1973) is the most popular

criterion for this type of modeling. However, when the sample size is small relative to the larger model dimensions represented within the candidate set, AIC is less protective against overfitting than many other likelihood-based criteria (Cavanaugh and Shumway, 1997; McQuarrie and Tsai, 1998; Rao and Wu, 2001).

Another likelihood-based criterion is the Takeuchi information criterion (TIC) (Takeuchi, 1976). TIC is justified for use in the GLM framework. TIC is a more general criterion than AIC; in fact, AIC may be viewed as a simplification of TIC that results under more restrictive conditions. However, TIC is considerably less known among practitioners than AIC. TIC has not become widely accepted because it was published in a difficult-to-find Japanese paper, and because the criterion requires the evaluation of more likelihood-based constructs than AIC. Shibata (1989) noted that the error incurred by this additional estimation can cause instability of the model selection results yielded by TIC. Therefore, TIC is not universally recommended (Burnham and Anderson, 2002).

AIC and TIC are two criteria that serve as estimators of KDD. Specifically, both may be viewed as asymptotically unbiased estimators of KDD, derived under different sets of assumptions. Both can be used with independent data in the GLM framework. Other criteria that can be used in this framework include those presented in Konishi and Kitagawa (1996), Ishiguro et al. (1997), Goutis and Robert (1998), Claeskens and Hjort (2003), Claeskens and Hjort (2008), Muller and Welsh (2009), Nott and Leng (2010), and many others. We do not discuss these contributions as they are not directly related to the criteria presented in this dissertation.

When data are correlated, one can model them using conditional or marginal

models. This choice depends on the investigator's question and the type of interpretation that she or he seeks to make using the data. When a subject-specific interpretation is sought, then conditional models are indicated and pseudo-likelihood-based generalized linear mixed models are employed. Alternatively, if a population-averaged interpretation is of interest, then marginal models are adequate. For marginal model parameter estimation, it is common to use quasi-likelihood-based GEEs.

An investigator seeking to model correlated data using a GLM will need to choose not only a suitable set of covariates for the mean structure, but also the working correlation structure, the variance function and the link function. Model selection criteria can inform all these choices. In this literature review we focus only on model selection tools primarily developed for choosing the mean and the working correlation structures when GEEs are used for parameter estimation. Model selection tools designed for additional GLM frameworks, and approaches to modeling correlated data, can be found in Liu et al. (1999), Vaida and Blanchard (2005), Yafune et al. (2005), Azari et al. (2006), Pu and Niu (2006), Kinney and Dunson (2007), Lavergne et al. (2008), Shang and Cavanaugh (2008) and Jiang et al. (2009) among others.

In the GEE approach for parameter estimation a likelihood is not specified; thus, AIC is not available as a model selection tool in this setting. Instead of a likelihood, unbiased estimating equations are employed. These estimating equations were derived as an extension of the quasi-likelihood equations introduced by Wedderburn (1974). Pan (2001) considered the problem of model selection in GEE applications and proposed the quasi-likelihood information criterion (QIC), a modification of AIC for use with GEEs and correlated data. QIC is widely known and used in this framework; since its introduction, over one hundred published applications have appeared where QIC is used for model selection. The success of QIC

stems from its ease of use, its similarity to AIC, and the fact that it has been implemented in popular statistical softwares such as Stata (Cui and Qian, 2007), R (Cui and Qian, 2007) and SAS (SAS Institute, 2007). QIC was also included in the guidelines by Hardin and Hilbe (2002, pp 139-142) for choosing an appropriate marginal model. A clear advantage of QIC is that it can be used to select suitable mean and working correlation structures.

Comparisons of the performance of QIC to other criteria are rare in the literature. Hin and colleagues (2007) use simulations to compare QIC and the Rotnisky-Jewell criterion (RJC) for the selection of working correlation structures. RJC is based on a heuristic approach that assesses the adequacy of the correlation structure when using GEEs by comparing the fitted model covariance structure to the empirical (a.k.a. robust) covariance estimate (Rotnitzky and Jewel, 1990). Hin et al. (2007) contrast the criteria performance to identify the true correlation structure for Gaussian or binomial data, covariates varying at the cluster or observation level, and exchangeable or autoregressive of order 1 (AR-1) intraclass correlation structures. The results indicate that QIC outperforms RJC for AR-1 structures, while RJC is better than QIC for exchangeable correlation structures.

Based partly on their previous results, Hin and Wang (2009) propose the correlation information criterion (CIC) as a complement to QIC for correlation structure selection when marginal models are used. CIC is a refinement of QIC that works appreciably better than QIC for correlation structure selection. However, CIC cannot be used for mean structure selection. Other work (e.g., Pan and Connett, 2002; Cantoni et al., 2005; Shults et al., 2009; Wang and Qu, 2009; Wang and Hin, 2010; etc.) present potentially competing alternatives to QIC; however, none of these publications compare the proposed criteria to QIC.

The first criterion developed using KSD as an oracle is the Kullback information criterion (KIC) (Cavanaugh, 1999). KIC functions as an asymptotically

unbiased estimator of KSD, under the same assumptions for which AIC serves as an asymptotically unbiased estimator of KDD. Similar to AIC, KIC is an effective model selection criterion in large-sample applications. KIC can be used in the GLM framework with independent data, and selects overspecified candidate models less often than AIC. Cavanaugh’s development is supported by Broersen and Wensink (1996), who empirically find that the penalty term for KIC provides the best protection against underfitting and overfitting for autoregressive order selection in finite samples.

Nevertheless, in settings where the sample size is small and the candidate set consists of models which are excessively over parameterized, both KIC and AIC may exhibit a tendency to choose overfitted models. As mentioned in section 1.2, small-sample refinements of AIC have been derived assuming a particular modeling framework, giving rise to AIC_c (e.g., Sugiura, 1978; Hurvich and Tsai, 1989; and 1993; Bedrick and Tsai, 1994; Azari et al., 2006; etc.). AIC_c serves as an exactly unbiased estimator of KDD in the framework of linear regression with normal errors. AIC_c also functions as an approximately unbiased estimator in many other modeling frameworks. Analogously, Cavanaugh (2004) proposes corrected KIC (KIC_c) as an exactly unbiased estimator of KSD for traditional linear regression models. In this setting, KIC_c outperforms AIC, AIC_c and KIC in both small and large sample settings when all possible combinations of covariates are considered for the class of candidate models. These results further document the advantage of using KSD instead of KDD as an oracle.

Other contributions based on KIC include an improved KIC for nonlinear regression models (Kim and Cavanaugh, 2005), corrected KICs for time series and multivariate regression (Hafidi and Mkhadri, 2006), and a corrected KIC for vector autoregressive models (Seghouane, 2006). Other criteria designed using KSD as oracle include a criterion for the simultaneous determination of the number of

components and predictors in finite mixture regression models (Hafidi and Mkhadri, 2010).

Cavanaugh (2004) concludes his work by suggesting that different estimators of KSD are possible, both in the setting of linear models and in other modeling frameworks. He specifically mentions that “criteria based on more sophisticated estimators of KSD have the potential to effectively guard against both under and over fitting over a wide array of different applications” (Cavanaugh, 2004, p 272). This notion provides the impetus for the developments presented in this dissertation.

1.6 Outline

The remainder of this dissertation is organized as it follows:

Chapter 2 includes an overview of parameter estimation and GLM concepts, the definition and characterization of KDD and KSD, and a simulation example to illustrate the efficacy of KSD as a discrepancy measure for linear model selection. We also include technical details of model selection criteria directly related to the criteria proposed in the following chapters.

In chapter 3 we derive KIC_o and KIC_u , two new likelihood-based model selection criteria using KSD as the oracle. KIC_o is designed to excel in settings where overspecification prevails. KIC_u results from modifying the assumptions adopted for developing KIC_o , and unlike KIC_o , is suitable for model selection in scenarios prone to underspecification.

Chapter 4 compares the performance of KIC_o and KIC_u using a simulation study based on factorial experimental designs. In this study, we also evaluate the performance of model selection criteria presented in chapter 2.

Chapter 5 consists of the development of QKIC, a quasi-likelihood-based model selection criterion for correlated response data. The oracle for the development of QKIC is KSD. We characterize the performance of QKIC through a

simulation study. The chapter concludes with a modeling application where QKIC is used for model selection.

In chapter 6 we discuss the model selection criteria presented in chapters 3 through 5 and describe future directions suggested by this work.

CHAPTER 2

PRELIMINARY CONCEPTS

This chapter presents and develops notions that provide a technical and conceptual foundation for the remainder of the dissertation. We include an overview of likelihood-based and quasi-likelihood-based parameter estimation and the GLM framework. We also formally define KDD and KSD and use a simulation example to illustrate the potential superiority of KSD over KDD for linear model selection. Finally, we provide an overview of those model selection criteria that are directly pertinent to the tools proposed in the following chapters.

2.1 Relevant Parameter Estimation and GLM Concepts

Nelder and Wedderburn (1972) introduce the term generalized linear model to unify a wide array of modeling frameworks for continuous, discrete, and categorical outcome data (e.g., linear, logistic and Poisson regression; ANOVA; ANCOVA; multinomial regression; etc.). GLMs can be used to model outcomes which have distributions within the exponential family of probability distributions. When outcomes are independent, parameter estimation is usually performed by maximizing the likelihood function.

However, data are generally correlated when measurements are collected to describe or explain clustered phenomena. For instance, with data that represent the evolution of patients over time, the geographical disposition of biological specimens, or the heritability of a disease within a family, clustering would occur naturally among the measured outcomes. In such cases, the practitioner should use modeling approaches that accommodate the sources of correlation within the data.

Marginal models are among the different avenues available for modeling correlated data. Unlike GLM approaches for independent data, the specification of a

likelihood is often not possible for parameter estimation in marginal models. Instead, a quasi-likelihood may be employed. The quasi-likelihood is formulated based upon the postulated mean and variance of the individual outcomes. The outcomes are treated as independent, which leads to straightforward estimating equations. Liang and Zeger (1986) propose an extension of quasi-likelihood-based estimation for correlated data, utilizing GEEs within the GLM framework. The GEE approach allows one to incorporate a proposed correlation structure in the parameter estimation process.

We start by introducing notation and basic results regarding likelihood-based estimation. We then proceed to introduce GLM notation and some of the fundamental concepts of the GLM framework. We conclude this section with a discussion of results pertaining to quasi-likelihood-based estimation and GEEs.

2.1.1 Likelihood-Based Estimation for Independent Responses

Let y_i ($i = 1, \dots, n$) be a sequence of independent observations taken on n experimental units. Assume that each y_i has a marginal probability density or mass function $f(y_i|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)'$. Given a collection of observations $\mathbf{y} = (y_1, y_2, \dots, y_n)'$, the function of $\boldsymbol{\theta}$ given by

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^n f(y_i|\boldsymbol{\theta}) \quad (2.1)$$

defines the *joint likelihood function*. Maximizing $L(\boldsymbol{\theta}|\mathbf{y})$ is the most popular technique for deriving parameter estimators. The parameter value at which $L(\boldsymbol{\theta}|\mathbf{y})$ attains its maximum as a function of $\boldsymbol{\theta}$, with \mathbf{y} held fixed, is the *maximum likelihood estimator* (MLE) of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$. One can maximize any monotonically increasing function of $L(\boldsymbol{\theta}|\mathbf{y})$ to find the MLE. The natural logarithm is the most

convenient among such functions. Let

$$\ell_i(\boldsymbol{\theta}|y_i) = \ln f(y_i|\boldsymbol{\theta}).$$

The *log-likelihood*,

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}|y_i),$$

is then maximized by solving the *likelihood equations*:

$$\begin{aligned} \mathbf{u}(\boldsymbol{\theta}|\mathbf{y}) &\equiv \sum_{i=1}^n \mathbf{u}_i(\boldsymbol{\theta}|y_i) = \mathbf{0}, \text{ where} \\ \mathbf{u}_i(\boldsymbol{\theta}|y_i) &= \frac{\partial \ell_i(\boldsymbol{\theta}|y_i)}{\partial \boldsymbol{\theta}}. \end{aligned}$$

The function $\mathbf{u}(\boldsymbol{\theta}|\mathbf{y})$ is the *score* of $\boldsymbol{\theta}$. To determine whether the solution to the likelihood equations represents a maximum, one must check that the matrix of second derivatives of the log likelihood is negative definitive. The negative of the matrix of second derivatives provides the amount of information the data have available for $\boldsymbol{\theta}$ and is known as the *observed information*,

$$\mathcal{I}(\boldsymbol{\theta}|\mathbf{y}) = -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}. \quad (2.2)$$

The expected value of (2.2) is known as the *expected information*; we denote it as $\mathcal{I}(\boldsymbol{\theta})$. The matrix $\mathcal{I}(\boldsymbol{\theta})$ is often also referred to as the *Fisher information*.

Let

$$\mathcal{J}(\boldsymbol{\theta}) = E\{\mathbf{u}(\boldsymbol{\theta}|\mathbf{y})\mathbf{u}(\boldsymbol{\theta}|\mathbf{y})'\}. \quad (2.3)$$

When the model is correctly specified, $\mathcal{J}(\boldsymbol{\theta}) = \mathcal{I}(\boldsymbol{\theta})$. The matrix $\mathcal{J}(\boldsymbol{\theta})$ can be estimated using $\mathcal{J}(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^n \mathbf{u}_i(\boldsymbol{\theta}|y_i)\mathbf{u}_i(\boldsymbol{\theta}|y_i)'$.

Maximum likelihood estimation is widely used because $\hat{\boldsymbol{\theta}}$ has desirable statistical properties. Assume that the model is properly specified and let $\boldsymbol{\theta}_0$ represent the data generating model parameter vector. Given suitable regularity conditions

(e.g., Casella and Berger, 2002, p 516), the MLE is consistent for $\boldsymbol{\theta}_0$: $\hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_0$ as $n \rightarrow \infty$. The MLE is also asymptotically unbiased for $\boldsymbol{\theta}_0$: $E(\hat{\boldsymbol{\theta}}) \rightarrow \boldsymbol{\theta}_0$ as $n \rightarrow \infty$. The large-sample variance/covariance matrix of the MLE is given by the inverse of the Fisher information:

$$\Sigma(\boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\theta}_0)^{-1}.$$

Under certain conditions, $\mathcal{I}(\boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\theta}_0|\mathbf{y})$. In general, $\text{Var}(\hat{\boldsymbol{\theta}})$ can be estimated by using either the observed or expected information matrix evaluated at $\hat{\boldsymbol{\theta}}$.

The MLE of $\boldsymbol{\theta}$ is also asymptotically efficient and asymptotically normally distributed:

$$\hat{\boldsymbol{\theta}} \sim N_k[\boldsymbol{\theta}_0, \Sigma(\boldsymbol{\theta}_0)].$$

The specification of the model chosen for parameter estimation may affect the consistency, bias, and variability of $\hat{\boldsymbol{\theta}}$. Misspecification may arise due to an improper choice for the distribution of the response, an improper formulation for the variance/covariance structure, or an improper formulation for the mean structure. Other sources of misspecification may arise, yet the preceding are the most relevant for our purposes.

If the distribution of the response is misspecified but the model structure based on $\boldsymbol{\theta}$ is correctly specified, according to White (1982), $\hat{\boldsymbol{\theta}}$ is still asymptotically normally distributed with mean $\boldsymbol{\theta}_0$, but the large-sample variance/covariance matrix of $\hat{\boldsymbol{\theta}}$ becomes

$$\Sigma(\boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\theta}_0)^{-1} \mathcal{J}(\boldsymbol{\theta}_0) \mathcal{I}(\boldsymbol{\theta}_0)^{-1}. \quad (2.4)$$

The parameter (2.4) is known as the *sandwich*, *robust* or *empirical variance*. The sandwich variance can be estimated by replacing $\mathcal{I}(\boldsymbol{\theta}_0)$ and $\mathcal{J}(\boldsymbol{\theta}_0)$ with $\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})$ and $\mathcal{J}(\hat{\boldsymbol{\theta}}|\mathbf{y})$, respectively. Note that, from a practical perspective, $\mathcal{I}(\hat{\boldsymbol{\theta}})$ and $\mathcal{J}(\hat{\boldsymbol{\theta}})$ are

not accessible due to the misspecification of the distribution needed to evaluate the expectations of $\mathcal{I}(\boldsymbol{\theta}|\mathbf{y})$ and $\mathcal{J}(\boldsymbol{\theta}|\mathbf{y})$.

If the model structure based on $\boldsymbol{\theta}$ is misspecified, $\hat{\boldsymbol{\theta}}$ will not necessarily converge to $\boldsymbol{\theta}_0$ as n increases. If the model is overspecified, $\hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_0$ assuming $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ are configured to have the same dimension. If the model is underspecified, $\hat{\boldsymbol{\theta}}$ converges to the *pseudo-true parameter*, denoted by $\bar{\boldsymbol{\theta}}$. For the postulated model, the pseudo-true parameter is the parameter that is closest to $\boldsymbol{\theta}_0$, where proximity is measured by the Kullback-Leibler information. Under appropriate regularity conditions (e.g., Ljung and Caines, 1979), $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically unbiased for $\bar{\boldsymbol{\theta}}$: $\hat{\boldsymbol{\theta}} \rightarrow \bar{\boldsymbol{\theta}}$ and $E(\hat{\boldsymbol{\theta}}) \rightarrow \bar{\boldsymbol{\theta}}$ as $n \rightarrow \infty$. Also,

$$\begin{aligned}\hat{\boldsymbol{\theta}} &\sim N_k[\bar{\boldsymbol{\theta}}, \Sigma(\bar{\boldsymbol{\theta}})], \text{ where} \\ \Sigma(\bar{\boldsymbol{\theta}}) &= \mathcal{I}(\bar{\boldsymbol{\theta}})^{-1} \mathcal{J}(\bar{\boldsymbol{\theta}}) \mathcal{I}(\bar{\boldsymbol{\theta}})^{-1}.\end{aligned}\tag{2.5}$$

For estimating (2.5), one may use $\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})$ to approximate $\mathcal{I}(\bar{\boldsymbol{\theta}})$ and $\mathcal{J}(\hat{\boldsymbol{\theta}}|\mathbf{y})$ to approximate $\mathcal{J}(\bar{\boldsymbol{\theta}})$.

2.1.2 Notation and Basic GLM Concepts

Let y_i ($i = 1, \dots, n$) be a sequence of independent observations taken on n experimental units. Assume that each y_i has a marginal probability density or mass function of the form

$$f(y_i|\gamma_i, \phi) = \exp\left(\frac{y_i\gamma_i - b(\gamma_i)}{a(\phi)} + c(y_i, \phi)\right),\tag{2.6}$$

which depends on the unknown parameters γ_i and ϕ .

A density of the form (2.6) belongs to the *exponential dispersion family*. When ϕ is known, this form simplifies to the *one-parameter exponential family*. Distributions such as the Bernoulli, binomial, multinomial, Poisson, negative binomial, normal, geometric, gamma and inverse Gaussian are members of the exponential

family. In (2.6), γ_i corresponds to the *natural parameter* and ϕ is the *dispersion parameter*.

The GLM framework is formulated based on the assumption that y_i has a density in the exponential family. Suppose that we seek to explain each y_i with a set of explanatory variables. Let x_{ij} ($j = 1, \dots, p$) be the measurement of the j^{th} covariate for the i^{th} experimental unit. For convenience, $x_{i1} = 1$. The measurements x_{ij} can be arranged in a *design matrix* as follows:

$$X = \begin{pmatrix} 1 & x_{12} & x_{13} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & x_{n3} & \cdots & x_{np} \end{pmatrix}. \quad (2.7)$$

GLMs are comprised of three fundamental elements: a *random component*, a *systematic component* and a *link function*. The random component consists of a collection of random variables y_i and their postulated exponential family distribution. The systematic component or *mean structure* is defined as a collection of linear forms in the covariates: $\eta_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$, with $\mathbf{x}'_i = (1, x_{i2}, x_{i3}, \dots, x_{ip})$. The systematic component is considered fixed. With $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$, we can write $\eta_i = \mathbf{x}'_i \boldsymbol{\beta} = \sum_{j=1}^p x_{ij} \beta_j$.

Let $\mu_i = E(y_i)$ and $v(\mu_i) = Var(y_i)$ denote the mean and variance of y_i , respectively. The link function provides a functional relation between μ_i and η_i which maps the mean of y_i to the linear form $\mathbf{x}'_i \boldsymbol{\beta}$; that is, $g(\mu_i) = \eta_i$. The link function is a strictly monotonic, differentiable function. When $g(\mu_i) = \mu_i$, $g(\cdot)$ is called the *identity link*; when $g(\mu_i) = \gamma_i$, $g(\cdot)$ is the *canonical link*.

The introduction of the link function allows one to express the likelihood (and hence, the log-likelihood) as a function of $\boldsymbol{\beta}$, $L(\boldsymbol{\beta}|\mathbf{y})$. In this context the likelihood

equations for $\boldsymbol{\beta}$ can be written as follows:

$$\mathbf{u}(\boldsymbol{\beta}|\mathbf{y}) \equiv X'DV^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}, \quad \text{where} \quad (2.8)$$

$$V = \begin{pmatrix} v(\mu_1) & 0 & \cdots & 0 \\ 0 & v(\mu_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v(\mu_n) \end{pmatrix} \equiv \text{Diag}(v(\mu_i)) \quad \text{and}$$

$$D = \begin{pmatrix} \frac{\partial \mu_1}{\partial \eta_1} & 0 & \cdots & 0 \\ 0 & \frac{\partial \mu_2}{\partial \eta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\partial \mu_n}{\partial \eta_n} \end{pmatrix} \equiv \text{Diag} \left(\frac{\partial \mu_i}{\partial \eta_i} \right).$$

Solving (2.8), one can find $\hat{\boldsymbol{\beta}}$, the MLE of $\boldsymbol{\beta}$. The Fisher information for $\boldsymbol{\beta}$ can be expressed as

$$\mathcal{I}(\boldsymbol{\beta}) = X'DV^{-1}DX.$$

Thus, provided the model is correctly specified, the large-sample variance/covariance matrix for $\hat{\boldsymbol{\beta}}$, $\text{Var}(\hat{\boldsymbol{\beta}})$, may be estimated by $\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}$.

With reference to (2.3), one can also define

$$\mathcal{J}(\boldsymbol{\beta}) = E\{\mathbf{u}(\boldsymbol{\beta}|\mathbf{y})\mathbf{u}(\boldsymbol{\beta}|\mathbf{y})'\}.$$

The natural estimate of this matrix is

$$\begin{aligned} \mathcal{J}(\boldsymbol{\beta}|\mathbf{y}) &= \sum_{i=1}^n \mathbf{u}_i(\boldsymbol{\beta}|y_i)\mathbf{u}_i(\boldsymbol{\beta}|y_i)' \quad \text{where} \\ \mathbf{u}_i(\boldsymbol{\beta}|y_i) &= \frac{y_i - \mu_i}{v(\mu_i)} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\eta}_i} \right) \mathbf{x}_i. \end{aligned}$$

If the canonical link is employed, it can be shown that $\mathcal{I}(\boldsymbol{\beta}) = \mathcal{I}(\boldsymbol{\beta}|\mathbf{y})$. Also, if the model is correctly specified, $\mathcal{I}(\boldsymbol{\beta}) = \mathcal{J}(\boldsymbol{\beta})$. When the distribution of the

random component is misspecified, the large-sample variance/covariance matrix for $\hat{\boldsymbol{\beta}}$, $Var(\hat{\boldsymbol{\beta}})$, can be estimated using

$$\Sigma(\hat{\boldsymbol{\beta}}) = \mathcal{I}(\hat{\boldsymbol{\beta}}|\mathbf{y})^{-1} \mathcal{J}(\hat{\boldsymbol{\beta}}|\mathbf{y}) \mathcal{I}(\hat{\boldsymbol{\beta}}|\mathbf{y})^{-1}. \quad (2.9)$$

Thus far, we have ignored ϕ . Before considering this parameter, it is helpful to introduce the concept of *overdispersion*. Overdispersion arises when the empirical variance in the data exceeds the variance under the fitted model. This problem does not arise for members of the exponential dispersion family such as the normal distribution, where all of the residual dispersion is accommodated through a separate dispersion parameter ϕ . (In the case of the normal distribution, this parameter is simply the variance, σ^2 .) However, for members of the one-parameter exponential family such as the Poisson and binomial distributions, the moments are completely specified when the mean is determined. Thus, the dispersion is fixed by the mean.

If y_i follows a distribution from the exponential dispersion family, ϕ is viewed as an unknown parameter that must be estimated. If y_i follows a distribution from the one-parameter exponential family and overdispersion is not present, we assume that $a(\phi) = \phi = 1$. If overdispersion is present, it is common to model the variance as $\phi v(\mu_i)$ and to estimate ϕ . Here, $v(\mu_i)$ is the variance based on the one-parameter distribution.

If overdispersion is ignored, the standard error of $\hat{\boldsymbol{\beta}}$ can be underestimated, possibly leading to incorrect inferences. There are various approaches to correctly estimate the standard error of $\hat{\boldsymbol{\beta}}$ in the presence of overdispersion. When the variance is modeled by introducing a dispersion parameter ϕ , an estimate of ϕ can be computed based on the Pearson chi-square statistic or the deviance, and the variances of the components of $\hat{\boldsymbol{\beta}}$ can be multiplied by this estimate. An alternative approach is to change the data distribution to one designed to accommodate

overdispersion: e.g., use a negative binomial distribution as opposed to a Poisson distribution. Yet another approach is estimate $Var(\hat{\beta})$ using (2.9).

2.1.3 Quasi-Likelihood-Based Estimation for Correlated Responses

In the case of correlated responses, it is not always possible to formulate a likelihood for parameter estimation. Wedderburn (1974) proposes a method for estimating the parameters of interest by treating the responses as if they were independent. This method preserves many of the appealing large-sample properties of maximum likelihood estimators.

Suppose that we are interested in modeling the evolution of n patients over time, and measure the response of each patient and the associated covariates t times. Then, we have t correlated responses for each subject; that is, for subject i , $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{it})'$. Let $\mu_{il} = E(y_{il})$ and $v(\mu_{il}) = Var(y_{il})$ with $l = 1, 2, \dots, t$. Then, $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{it})'$. Also, let x_{ijl} be the measurement for the j^{th} covariate ($j = 1, 2, \dots, p$) corresponding to the i^{th} subject at time period l . For convenience, $x_{i1l} = 1$. In this case the mean structure becomes $\eta_{il} = \beta_1 + \beta_2 x_{i2l} + \dots + \beta_p x_{ipl}$.

We can define a $t \times p$ subject-specific design matrix X_i with the same layout as matrix (2.7). In X_i , the x_{ijl} s for the i^{th} subject take the place of the x_{ijs} in X .

One can define

$$Q(\boldsymbol{\mu}|\mathbf{y}) = \sum_{i=1}^n Q_i(\boldsymbol{\mu}_i|\mathbf{y}_i), \quad \text{where} \quad (2.10)$$

$$Q_i(\boldsymbol{\mu}_i|\mathbf{y}_i) = \sum_{l=1}^t \int_{y_{il}}^{\mu_{il}} \frac{y_{il} - z}{v(z)} dz.$$

The expression in (2.10) is the quasi-likelihood and plays a role analogous to $\ell(\boldsymbol{\mu}|\mathbf{y})$.

Differentiating (2.10) with respect to $\boldsymbol{\beta}$ yields

$$\mathbf{q}(\boldsymbol{\beta}|\mathbf{y}) \equiv \frac{\partial Q(\boldsymbol{\mu}|\mathbf{y})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n X_i' D_i V_i^{-1} \{\mathbf{y}_i - \boldsymbol{\mu}_i\}, \quad (2.11)$$

where $D_i = \text{Diag}(\partial\mu_{il}/\partial\eta_{il})$ and V_i is the modeled variance/covariance matrix for \mathbf{y}_i : i.e., $V_i = \text{Diag}(v(\mu_{il}))$. Solving $\sum_{i=1}^n X_i' D_i V_i^{-1} \{\mathbf{y}_i - \boldsymbol{\mu}_i\} = \mathbf{0}$ yields the quasi-likelihood-based estimator $\hat{\boldsymbol{\beta}}$. Assuming that the mean structure is correctly specified, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta}$ as $n \rightarrow \infty$.

Note that expression (2.8) can be solved for any choice of link and variance functions. For modeling overdispersed data, it suffices to replace $v(\mu_{il})$ by $\phi v(\mu_{il})$.

Let

$$\mathcal{I}(\boldsymbol{\beta}|\mathbf{y}) = -\frac{\partial^2 Q(\boldsymbol{\mu}|\mathbf{y})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}$$

The preceding functions as the observed information matrix in the quasi-likelihood framework.

The sandwich variance estimator can be used to estimate $\text{Var}(\hat{\boldsymbol{\beta}})$. Let

$$\begin{aligned}\Sigma(\boldsymbol{\beta}) &= \mathcal{I}(\boldsymbol{\beta})^{-1} \mathcal{J}(\boldsymbol{\beta}) \mathcal{I}(\boldsymbol{\beta})^{-1}, \text{ where} \\ \mathcal{I}(\boldsymbol{\beta}) &= \sum_{i=1}^n X_i' D_i V_i^{-1} D_i X_i, \\ \mathcal{J}(\boldsymbol{\beta}) &= E\{\mathcal{J}(\boldsymbol{\beta}|\mathbf{y})\}, \text{ and} \\ \mathcal{J}(\boldsymbol{\beta}|\mathbf{y}) &= \sum_{i=1}^n X_i' D_i V_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)' V_i^{-1} D_i X_i.\end{aligned}$$

The large-sample variance/covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by $\Sigma(\boldsymbol{\beta})$. The sandwich estimator which is of the form of (2.9) is given by

$$\Sigma(\hat{\boldsymbol{\beta}}) = \mathcal{I}(\hat{\boldsymbol{\beta}}|\mathbf{y})^{-1} \mathcal{J}(\hat{\boldsymbol{\beta}}|\mathbf{y}) \mathcal{I}(\hat{\boldsymbol{\beta}}|\mathbf{y})^{-1}. \quad (2.12)$$

However, the conventional sandwich estimator is based on

$$\Sigma(\hat{\boldsymbol{\beta}}) = \mathcal{I}(\hat{\boldsymbol{\beta}})^{-1} \mathcal{J}(\hat{\boldsymbol{\beta}}|\mathbf{y}) \mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}.$$

(See Liang and Zeger, 1986).

When responses are correlated, the true $\text{Var}(\mathbf{y}_i)$ has off-diagonal values that are different from 0 and usually unknown. In quasi-likelihood-based estimation, a

diagonal V_i is used as if the data were independent, although they are not. This approach is known as the *working independence model*.

The simplicity of the working independence model carries the cost of a loss of efficiency for the β estimates. Nevertheless, assuming a correctly specified mean structure, the β estimates are consistent and asymptotically unbiased. The estimates are also asymptotically normally distributed (Wedderburn, 1974).

Liang and Zeger (1986) extend the quasi-likelihood approach and suggest that improved efficiency can be obtained by simultaneously estimating β and the parameters in $Var(\mathbf{y}_i)$. This corresponds to employing (2.11), yet instead of modeling $Var(\mathbf{y}_i)$ using a diagonal V_i , using a more general structure:

$$V_i(\boldsymbol{\alpha}) = \Delta_i^{1/2} R_i(\boldsymbol{\alpha}) \Delta_i^{1/2}. \quad (2.13)$$

Here, $\Delta_i = \text{Diag}(v(\mu_{il}))$ for $l = 1, 2, \dots, t$ and $R_i(\boldsymbol{\alpha})$ is a $t \times t$ *working correlation matrix* with parameter vector $\boldsymbol{\alpha}$. Liang and Zeger (1986) call equations (2.11) in conjunction with (2.13) *generalized estimating equations*. In fact, if $R_i(\boldsymbol{\alpha})$ is the identity matrix, $V_i(\boldsymbol{\alpha})$ is diagonal and this approach reduces to the aforementioned working independence model.

Other common choices for $R_i(\boldsymbol{\alpha})$ are an *exchangeable* correlation matrix with a single parameter (i.e., $\text{corr}(y_{ij}, y_{il}) = \alpha$ for $j \neq l$), an *AR-1* correlation matrix with a single parameter (i.e., $\text{corr}(y_{ij}, y_{il}) = \alpha^{|l-j|}$ for $j \neq l$), and an *unstructured* correlation matrix with $t(t-1)/2$ parameters (i.e., $\text{corr}(y_{ij}, y_{il}) = \alpha_{jl}$).

With a correctly specified mean structure, the β estimates resulting from this extension of quasi-likelihood are consistent and asymptotically unbiased. The estimates are also asymptotically normal. An appropriate specification of $V_i(\boldsymbol{\alpha})$ improves asymptotic efficiency. If $V_i(\boldsymbol{\alpha})$ is assumed to be correct, $Var(\hat{\beta})$ can be estimated by $\mathcal{I}(\hat{\beta})^{-1}$. Otherwise, $Var(\hat{\beta})$ can be estimated using (2.12).

2.2 Model Selection Concepts

We now review model selection concepts relevant to the developments presented in chapters 3 through 5. We start by establishing the technical notions that facilitate the development of estimators for KDD and KSD. We then illustrate the performance of KDD and KSD using a simulation example. We also introduce Kullback discrepancies based on variants of KDD or KSD. Finally, we introduce relevant likelihood-based and quasi-likelihood-based model selection criteria, against which we compare the performance of the new criteria we propose in chapters 3 and 5.

2.2.1 Kullback's Directed and Symmetric Divergences

Let the generating model be denoted by $g(\mathbf{y}|\boldsymbol{\theta}_0)$. To determine which of the proposed fitted models best resembles $g(\mathbf{y}|\boldsymbol{\theta}_0)$, we need a measure that reflects the disparity between $g(\mathbf{y}|\boldsymbol{\theta}_0)$ and a candidate model $f(\mathbf{y}|\boldsymbol{\theta})$. The oracles KDD and KSD both fulfill this objective. KDD between $g(\mathbf{y}|\boldsymbol{\theta}_0)$ and $f(\mathbf{y}|\boldsymbol{\theta})$ with respect to $g(\mathbf{y}|\boldsymbol{\theta}_0)$ is defined as

$$\text{KDD} \equiv I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = E_g \left(\ln \frac{g(\mathbf{y}|\boldsymbol{\theta}_0)}{f(\mathbf{y}|\boldsymbol{\theta})} \right). \quad (2.14)$$

Here, E_g denotes the expectation under $g(\mathbf{y}|\boldsymbol{\theta}_0)$. Note that $I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ is not symmetric in terms of the densities that define the measure. Thus, an alternative directed divergence, $I_{fg}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$, may be obtained by switching the roles of $g(\mathbf{y}|\boldsymbol{\theta}_0)$ and $f(\mathbf{y}|\boldsymbol{\theta})$ in (2.14). The sum of the two directed divergences yields KSD:

$$\begin{aligned} \text{KSD} \equiv J_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) &= I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) + I_{fg}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \\ &= E_g \left(\ln \frac{g(\mathbf{y}|\boldsymbol{\theta}_0)}{f(\mathbf{y}|\boldsymbol{\theta})} \right) + E_f \left(\ln \frac{f(\mathbf{y}|\boldsymbol{\theta})}{g(\mathbf{y}|\boldsymbol{\theta}_0)} \right), \end{aligned} \quad (2.15)$$

where E_f denotes the expectation under $f(\mathbf{y}|\boldsymbol{\theta})$. It is well known that $I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \geq 0$ with equality if and only if $f(\mathbf{y}|\boldsymbol{\theta}) = g(\mathbf{y}|\boldsymbol{\theta}_0)$ (Kullback, 1968, p 14); the same property follows for $I_{fg}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ and $J_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$.

To assess the proximity between a certain fitted candidate model $f(\mathbf{y}|\hat{\boldsymbol{\theta}})$ and $g(\mathbf{y}|\boldsymbol{\theta}_0)$, we consider the measures $J_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = J_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ and $I_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$. In practical settings $I_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $J_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ are oracles; that is, they can be estimated but neither can be directly employed for model selection purposes, because both measures depend upon the unknown generating model $g(\mathbf{y}|\boldsymbol{\theta}_0)$.

At this point, one might raise the following question: which of KSD or KDD is a better measure of disparity between $g(\mathbf{y}|\boldsymbol{\theta}_0)$ and $f(\mathbf{y}|\hat{\boldsymbol{\theta}})$? We address this question using an example taken from Cavanaugh (2004).

The objective of the example is to examine the performance of KSD and KDD as oracles for finding a fitted linear model that provides a suitable approximation to the normal linear model

$$\mathbf{y} = X_0\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}, \quad (2.16)$$

where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma_0^2 I)$ and X_0 is $n \times p_0$ of rank p_0 , $\boldsymbol{\beta}_0$ is $p_0 \times 1$, and $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \sigma_0^2)$.

As a candidate model, consider the normal linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (2.17)$$

where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 I)$ and X is $n \times p$ of rank p , $\boldsymbol{\beta}$ is $p \times 1$, and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$. The MLE of $\boldsymbol{\theta}$ is denoted as $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$, and $f(\mathbf{y}|\hat{\boldsymbol{\theta}})$ represents the resulting empirical likelihood.

For models (2.16) and (2.17), one can show

$$I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \frac{1}{2}n \left(\ln \left(\frac{\sigma^2}{\sigma_0^2} \right) + \frac{\sigma_0^2}{\sigma^2} \right) + \frac{1}{2\sigma^2} (X_0\boldsymbol{\beta}_0 - X\boldsymbol{\beta})' (X_0\boldsymbol{\beta}_0 - X\boldsymbol{\beta}) - \frac{1}{2}n, \quad (2.18)$$

$$I_{fg}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{1}{2}n \left(\ln \left(\frac{\sigma_0^2}{\sigma^2} \right) + \frac{\sigma^2}{\sigma_0^2} \right) + \frac{1}{2\sigma_0^2} (X_0\boldsymbol{\beta}_0 - X\boldsymbol{\beta})' (X_0\boldsymbol{\beta}_0 - X\boldsymbol{\beta}) - \frac{1}{2}n. \quad (2.19)$$

For the interpretations to follow, it is useful to recall that $f(x) = \ln(1/x) + x$ is positive and increasing in x for $x > 1$. Considering $I_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $I_{fg}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ as

functions of $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$, $I_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ is often large when $(X_0\boldsymbol{\beta}_0 - X\hat{\boldsymbol{\beta}})'(X_0\boldsymbol{\beta}_0 - X\hat{\boldsymbol{\beta}})$ is large and $\hat{\sigma}^2$ is small. For models that are correctly specified or overfitted, it follows that

$$E_g\{(X_0\boldsymbol{\beta}_0 - X\hat{\boldsymbol{\beta}})'(X_0\boldsymbol{\beta}_0 - X\hat{\boldsymbol{\beta}})\} = p\sigma_0^2 \quad \text{and} \quad E_g(\hat{\sigma}^2) = \left(1 - \frac{p}{n}\right)\sigma_0^2.$$

Thus, the form of $I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ implies that $I_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ is often sensitive towards overfitting, assuming large values when p is large and $\hat{\sigma}^2$ is substantially deflated.

On the other hand, $I_{fg}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ is often large when $(X_0\boldsymbol{\beta}_0 - X\hat{\boldsymbol{\beta}})'(X_0\boldsymbol{\beta}_0 - X\hat{\boldsymbol{\beta}})$ and $\hat{\sigma}^2$ are both large. For models that are underfitted,

$$E_g\{(X_0\boldsymbol{\beta}_0 - X\hat{\boldsymbol{\beta}})'(X_0\boldsymbol{\beta}_0 - X\hat{\boldsymbol{\beta}})\} = p\sigma_0^2 + (X_0\boldsymbol{\beta}_0)'(I - H)(X_0\boldsymbol{\beta}_0) \quad \text{and}$$

$$E_g(\hat{\sigma}^2) = \left(1 - \frac{p}{n}\right)\sigma_0^2 + \frac{1}{n}(X_0\boldsymbol{\beta}_0)'(I - H)(X_0\boldsymbol{\beta}_0),$$

where I is the identity matrix and H is the projection matrix onto the column space of X . The size of the quadratic form $(X_0\boldsymbol{\beta}_0)'(I - H)(X_0\boldsymbol{\beta}_0)$ is dictated by the extent to which X is underspecified. Thus, the form of $I_{fg}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ implies that $I_{fg}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ is often sensitive towards underfitting, assuming large values when $(X_0\boldsymbol{\beta}_0)'(I - H)(X_0\boldsymbol{\beta}_0)$ is large and $\hat{\sigma}^2$ is substantially inflated.

If $X_0\boldsymbol{\beta}_0$, n , and σ_0^2 are fixed, the two directed divergences in (2.18) and (2.19) may be regarded as a function of the candidate model error variance (σ^2 or Variance) and the squared difference between the mean vectors under the true and candidate models (Q). Figure 2.1 features three-dimensional plots of KDD and KSD as functions of Q and the Variance. The minimum value of each function occurs at the point $(Q, \text{Variance}) = (\mu_0, \sigma_0^2) = (0, 36)$. Moving in any direction from this point, KSD increases more than KDD. The differences in curvature are particularly evident in the back upper-left corners of the plots.

To further illustrate the effectiveness of KSD and KDD as measures for model selection in the linear regression framework, one can simulate data using $n = 26$,

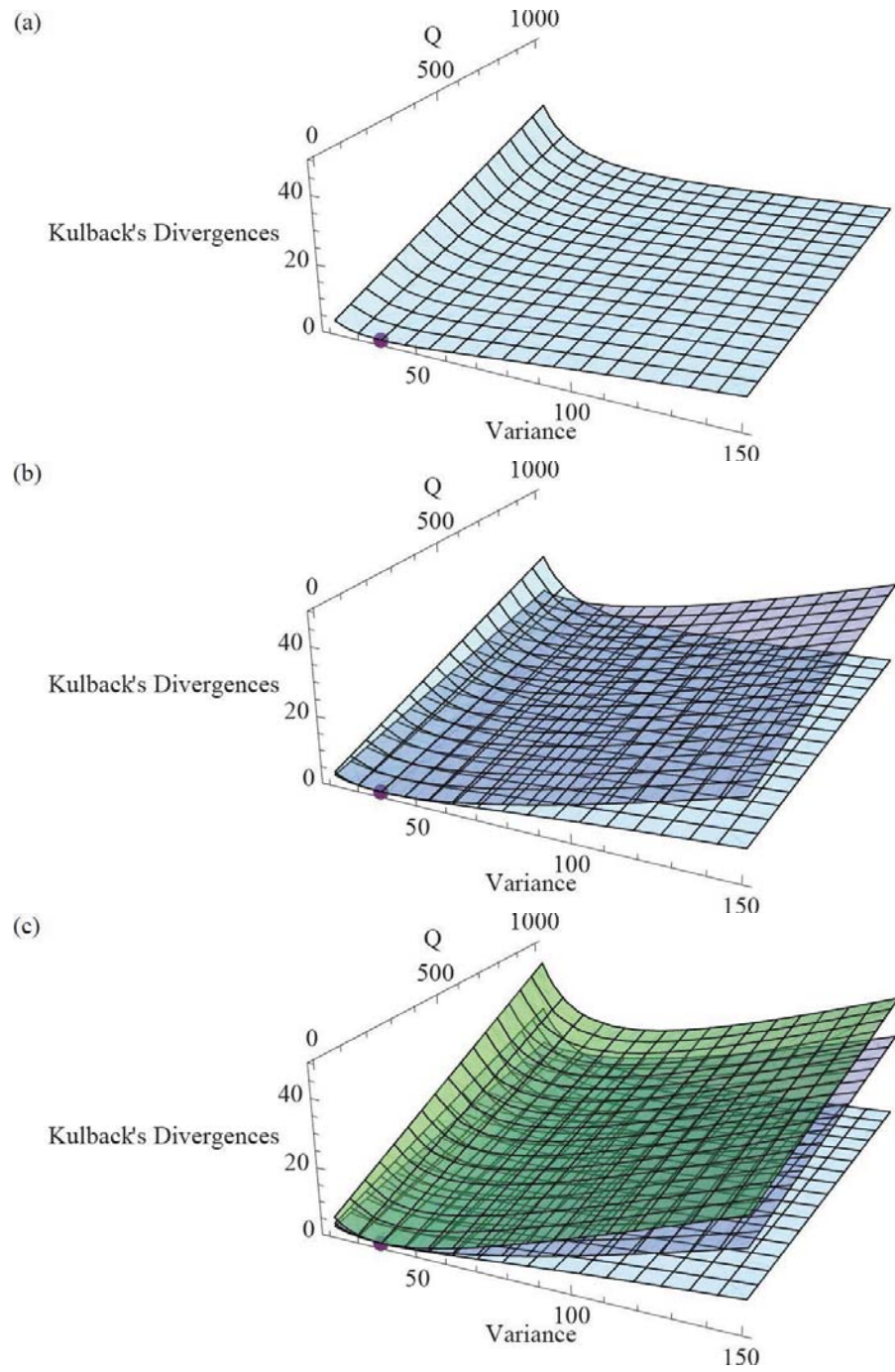


Figure 2.1: Three-dimensional plots comparing KDD (light blue surface), KDD counterpart (light purple surface) and KSD (green surface); the dot represents the minimum of the surfaces, at (0,36).

$\mu_0 = 0$, $\sigma_0^2 = 36$ and the linear model

$$y_i = 1 + x_{i2} + x_{i3} + x_{i4} + \epsilon_i, \quad (2.20)$$

where $\epsilon_i \sim N(0, 36)$.

Consider using KSD and KDD to determine which of the following fitted models best describes the data:

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\epsilon}_i, \quad (2.21)$$

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} + \hat{\epsilon}_i, \quad \text{or} \quad (2.22)$$

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_5 x_{i5} + \hat{\beta}_6 x_{i6} + \hat{\epsilon}_i. \quad (2.23)$$

A total of 10,000 samples are generated using model (2.20). For every sample, the fitted models (2.21), (2.22) and (2.23) are obtained. KSD and KDD are then evaluated for all three models, and the fitted model corresponding to the minimum value of each measure is recorded. The regressors for all models are generated from a $U(0, 10)$ distribution. Table 2.1 shows the results, along with the average value of $\hat{\sigma}^2$, \hat{Q} , KDD and KSD for each candidate model.

					KDD	KSD
$f(\mathbf{y} \hat{\boldsymbol{\theta}})$	$\hat{E}_g(\hat{Q})$	$\hat{E}_g(\hat{\sigma}^2)$	$\hat{E}_g(\text{KDD})$	$\hat{E}_g(\text{KSD})$	Selections	Selections
(2.23)	354.4	35.9	6.15	11.69	738	576
(2.22)	143.4	30.5	3.68	6.43	8408	9026
(2.21)	472.4	48.7	6.03	13.78	854	398

Table 2.1: Average values for $\hat{\sigma}^2$, \hat{Q} , KDD and KSD, and KDD and KSD number of model selections for 10,000 samples; results for the generating model are bolded.

The correctly specified model, (2.22), is chosen more frequently by KSD than

by KDD. The reason for this can be seen in Figure 2.1 by comparing the plots in the neighborhoods of the points $(\hat{E}_g(\hat{Q}), \hat{E}_g(\hat{\sigma}^2))$ corresponding to each of the candidate models. For model (2.22), the point $(\hat{E}_g(\hat{Q}), \hat{E}_g(\hat{\sigma}^2)) = (143.4, 30.5)$ lies to the back right of the point at which the divergence surfaces attain their minimum, $(0, 36)$. Moving from the point $(143.4, 30.5)$ towards either of the points $(472.4, 48.7)$ (for model (2.21)) or $(354.4, 35.9)$ (for model (2.23)), the curvature of KSD is more pronounced than that of KDD, especially in the direction of $(472.4, 48.7)$. For a particular sample, it is more likely for KSD than KDD to be minimized at the coordinate $(\hat{Q}, \hat{\sigma}^2)$ corresponding to the correctly specified model.

A plausible explanation for the preceding results is that KSD combines the information in $I_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $I_{fg}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$, two measures which are related and yet distinct. Over the 10,000 samples generated for this example, the correlations between $I_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $I_{fg}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ for models (2.21), (2.22) and (2.23) are 0.483, 0.909 and 0.716, respectively. This reinforces the notion that $I_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $I_{fg}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ are not redundant and advances the premise that KSD improves upon KDD by incorporating the additional information in $I_{fg}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$.

In evaluating the adequacy of a fitted candidate model, Cavanaugh (2004) notes that $I_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ may better reflect the error due to estimation variability, whereas $I_{fg}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ may better reflect the error due to estimation bias. This suggests that when $I_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $I_{fg}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ are combined, KSD might provide a more balanced gauge of model disparity. In settings where the set of fitted candidate models consists of both underfitted and overfitted models, KSD may better indicate those models that are improperly specified than KDD. As a consequence, an estimator of KSD may be preferable to an estimator of KDD as a model selection criterion.

2.2.2 Kullback's Discrepancies

Let \mathcal{M} denote the collection of probability densities or mass functions for the data vector \mathbf{y} . Then, $g(\mathbf{y}|\boldsymbol{\theta}_0) \in \mathcal{M}$ and $f(\mathbf{y}|\boldsymbol{\theta}) \in \mathcal{M}$. A *discrepancy*, d , is a mapping from $\mathcal{M} \times \mathcal{M}$ to \mathcal{R} such that $d(g(\mathbf{y}|\boldsymbol{\theta}_0), f(\mathbf{y}|\boldsymbol{\theta})) \geq d(g(\mathbf{y}|\boldsymbol{\theta}_0), g(\mathbf{y}|\boldsymbol{\theta}_0))$ (Linhart and Zucchini, 1986, p 11).

A discrepancy is not a proper distance function, as it is not always the case that $d(a, a) = 0$, $d(a, b) = d(b, a)$, and $d(a, c) \leq d(a, b) + d(b, c)$ for all a , b , and c . For example, KDD and KSD are discrepancies that satisfy $d(a, a) = 0$, but only KSD satisfies $d(a, b) = d(b, a)$. However, even though a discrepancy is not a formal metric, it shares the same spirit as a distance, in that it is designed to measure the separation between two entities. Thus, ideally, as the separation between the generating and the candidate model increases, $d(g(\mathbf{y}|\boldsymbol{\theta}_0), f(\mathbf{y}|\boldsymbol{\theta}))$ also increases. For brevity of notation, we use $d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ instead of $d(g(\mathbf{y}|\boldsymbol{\theta}_0), f(\mathbf{y}|\boldsymbol{\theta}))$ for the remainder of this manuscript.

In order to develop KDD and KSD estimators we consider discrepancies of the following form:

$$d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = E_g\{-2 \ln f(\mathbf{y}|\boldsymbol{\theta})\} \text{ and} \quad (2.24)$$

$$d_{fg}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = E_f\{-2 \ln g(\mathbf{y}|\boldsymbol{\theta}_0)\}. \quad (2.25)$$

Evaluating the second argument of (2.24) at $g(\mathbf{y}|\boldsymbol{\theta}_0)$ yields

$$d_{gg}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = E_g\{-2 \ln g(\mathbf{y}|\boldsymbol{\theta}_0)\}, \quad (2.26)$$

the minimum of $d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$. Also, evaluating the second argument of $d_{fg}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ at $f(\mathbf{y}|\boldsymbol{\theta})$ yields

$$d_{ff}(\boldsymbol{\theta}, \boldsymbol{\theta}) = E_f\{-2 \ln f(\mathbf{y}|\boldsymbol{\theta})\}, \quad (2.27)$$

the minimum of $d_{fg}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$. Combining (2.14), (2.24) and (2.26), we can write

$$2I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) - d_{gg}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0).$$

Since $d_{gg}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ does not depend on $\boldsymbol{\theta}$, any ranking of candidate models based on $I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ would be identical to a ranking based on $d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$. Thus, for discriminating among various candidate models, the measure $d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ is a valid substitute for $I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$.

From (2.15) and (2.24) through (2.27), it follows that

$$2J_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) - d_{gg}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) + d_{fg}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - d_{ff}(\boldsymbol{\theta}, \boldsymbol{\theta}).$$

As with $I_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, $d_{gg}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ can be discarded as it will not affect the ranking of candidate models, and thus,

$$K_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) + d_{fg}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - d_{ff}(\boldsymbol{\theta}, \boldsymbol{\theta}) \quad (2.28)$$

is a valid substitute for $J_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$.

2.2.3 Likelihood-Based Model Selection Criteria: AIC, TIC and KIC

The *overall discrepancy*

$$d_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = E_g\{-2 \ln f(\mathbf{y}|\boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \quad (2.29)$$

reflects the separation between $g(\mathbf{y}|\boldsymbol{\theta}_0)$ and $f(\mathbf{y}|\hat{\boldsymbol{\theta}})$. The expectation with respect to $g(\mathbf{y}|\boldsymbol{\theta}_0)$ of (2.29) is the *expected discrepancy*, $E_g\{d_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\}$. The evaluation of both discrepancies requires knowledge of $g(\mathbf{y}|\boldsymbol{\theta}_0)$. Akaike (1973) suggests that the *estimated discrepancy*,

$$-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}),$$

serves as a biased estimator of $E_g\{d_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\}$.

The expected discrepancy reflects how well, on average, the fitted candidate model predicts new data generated under the true model. On the other hand, the estimated discrepancy reflects how well the fitted model predicts the data at hand. By evaluating the adequacy of the fitted model based on its ability to predict the data used in its own construction, the estimated discrepancy yields an overly optimistic assessment of how effectively the fitted model performs. Thus, the estimated discrepancy serves as a negatively biased estimator of the expected discrepancy. Correcting for this bias leads to the *penalty term* or *bias correction term* of a KDD-based model selection criterion.

Let $\mathcal{F}(k) = \{f(\mathbf{y}|\boldsymbol{\theta})|\boldsymbol{\theta} \in \Theta(k)\}$ denote a k -dimensional parametric *candidate family* of probability densities or mass functions; that is, a family in which the parameter space $\Theta(k)$ consists of k -dimensional vectors whose components are functionally independent. Akaike (1974) shows that if $g(\mathbf{y}|\boldsymbol{\theta}_0) \in \mathcal{F}(k)$, under a set of regularity conditions ensuring the large-sample properties of $\hat{\boldsymbol{\theta}}$, the statistic

$$\text{AIC} = -2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + 2k$$

serves as an asymptotically unbiased estimator of $E_g\{d_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\}$. The first term of AIC is usually referred to as the *goodness-of-fit term* and $2k$ is the penalty term.

AIC provides an approximately unbiased estimator of the expected discrepancy in settings where n is large and k is comparatively small. In settings where n is small and k is comparatively large (e.g., $k \approx n/2$), $2k$ is often much smaller than the bias adjustment, making AIC substantially negatively biased as an estimator of the expected discrepancy. If AIC severely underestimates the expected discrepancy for higher dimensional fitted models in the candidate set, the criterion may favor the higher dimensional models even when the expected discrepancy between these models and the generating model is rather large. Examples illustrating this phenomenon appear in Linhart and Zucchini (1986, p 86), who comment (p 78) that “in

some cases the criterion simply continues to decrease as the number of parameters in the approximating model is increased.”

AIC is applicable in a broad array of modeling frameworks, since its justification only requires conventional large-sample properties of MLEs. AIC can be used to compare non-nested models and models based on different probability distributions. In a model selection application, the optimal fitted model is identified by the minimum value of AIC.

Takeuchi (1976) shows that when $g(\mathbf{y}|\boldsymbol{\theta}_0)$ is not necessarily included in $\mathcal{F}(k)$, the statistic

$$\text{TIC} = -2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + 2\text{tr}\{\mathcal{J}(\hat{\boldsymbol{\theta}}|\mathbf{y})\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})^{-1}\}$$

functions as an asymptotically unbiased estimator of $E_g\{d_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\}$.

If $g(\mathbf{y}|\boldsymbol{\theta}_0) \in \mathcal{F}(k)$, $\mathcal{I}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})$; hence, $\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y}) \approx \mathcal{J}(\hat{\boldsymbol{\theta}}|\mathbf{y})$, and the bias correction term of TIC is close to that of AIC. However, if $g(\mathbf{y}|\boldsymbol{\theta}_0)$ is not well approximated by $f(\mathbf{y}|\hat{\boldsymbol{\theta}})$, then $\text{tr}\{\mathcal{J}(\hat{\boldsymbol{\theta}}|\mathbf{y})\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})^{-1}\}$ could be considerably different from k . The data-dependent estimator of the bias correction, $2\text{tr}\{\mathcal{J}(\hat{\boldsymbol{\theta}}|\mathbf{y})\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})^{-1}\}$, might be substantially less biased than $2k$. However, a data-dependent estimator might also be highly variable. This issue discourages some authors to recommend the use of TIC (Burnham and Anderson, 2002).

Critics of model selection criteria with data-dependent penalty terms sometimes argue that such estimators of the bias adjustment are highly inaccurate in settings conducive to underfitting. In simulation studies to evaluate the performance of model selection criteria, two common conditions that tend to promote a high frequency of underfitted selections are a small sample size and a low signal-to-noise ratio (SNR).

SNR is usually defined as $\text{Var}(\eta_i)$ over $\text{Var}(y_i)$. In traditional regression applications, the GLM systematic component is regarded as deterministic and thereby

has a variance of zero. However, in simulation studies, the preceding SNR definition is sensible since the systematic components are randomly generated. For normal linear regression, the SNR definition is amenable to a familiar interpretation: if a correctly specified model is fitted to data generated under a true model with a given SNR, the coefficient of determination for the fit will be approximately $\text{SNR}/(1+\text{SNR})$.

In settings where SNR is low or the sample size is small, we have seen that TIC tends to choose underfitted models more frequently than AIC. However, for a fixed sample size, as SNR increases, the propensity of TIC to select underfitted models is attenuated and TIC usually outperforms AIC. For a fixed SNR, as the sample size grows, the probability of the criteria choosing an underfitted model converges to zero, and TIC and AIC exhibit the same selection properties.

TIC may outperform AIC in settings where the data distribution is misspecified. For instance, with a normal linear likelihood, if sample size is small (i.e., $n = 25$ to 100) and the error is distributed with thicker tails than those of a normal distribution (e.g., $\epsilon_i \sim t(\text{df})$ with $\text{df} \leq 6$), Kitagawa (1987) shows that the penalty term of TIC is markedly different from that in AIC.

The last likelihood-based criterion of relevance to the present work is based on KSD. Cavanaugh (1999) shows that if $g(\mathbf{y}|\boldsymbol{\theta}_0) \in \mathcal{F}(k)$, under a set of regularity conditions ensuring the large-sample properties of $\hat{\boldsymbol{\theta}}$, the statistic

$$\text{KIC} = -2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + 3k$$

serves as an asymptotically unbiased estimator of

$$\Omega_{gf}(\boldsymbol{\theta}_0) \equiv E_g\{K_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\} = E_g\{d_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) + d_{fg}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})\}. \quad (2.30)$$

In the linear regression framework, various simulations presented in Cavanaugh (2004) show that KIC outperforms AIC by obtaining more correct model selections

than AIC. However, in small-sample applications where excessively overparameterized families are entertained, KIC tends to underestimate $E_g\{K_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\}$ in the same manner that AIC tends to underestimate $E_g\{d_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\}$. Thus, KIC may be an improvement over AIC in terms of guarding against overfitting, but in small-sample settings, where underfitting may be likely, there is room for criteria based on more sophisticated estimators of $\Omega_{gf}(\boldsymbol{\theta}_0)$. Ideally, such estimators have the potential to effectively guard against both underfitting and overfitting.

More extensive comparisons of AIC, TIC and KIC in the GLM framework are included in the simulations presented in chapter 4.

2.2.4 Quasi-Likelihood-Based Model Selection Criterion: QIC

We conclude this chapter by defining QIC. This criterion is widely used for the selection of a GLM for correlated data, where the model parameters are estimated using either a quasi-likelihood or GEEs. QIC is suitable for selecting the working correlation structure and the covariates for the mean structure. QIC is introduced by Pan (2001) as a modification of AIC that is developed by replacing $f(\boldsymbol{\beta}|\mathbf{y})$ in $d_{gf}(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ with the quasi-likelihood.

Let $\hat{\boldsymbol{\beta}}^I$ denote the estimator of $\boldsymbol{\beta}$ under the working independence model. Then the overall discrepancy can be expressed as

$$d_{gq}(\boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}}^I) = E_g\{-2Q(\boldsymbol{\beta}|\mathbf{y})\}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^I},$$

and the expected and estimated discrepancies become

$$E_g\{d_{gq}(\boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}}^I)\} \text{ and} \tag{2.31}$$

$$-2Q(\hat{\boldsymbol{\beta}}^I|\mathbf{y}), \tag{2.32}$$

respectively. Correcting for the bias in (2.32) introduced by estimating (2.31) with

(2.32) yields

$$\text{QIC}^I = -2Q(\hat{\beta}^I|\mathbf{y}) + 2\text{tr}\{\mathcal{I}(\hat{\beta}^I|\mathbf{y})\Sigma(\hat{\beta}^I)\}. \quad (2.33)$$

Here, the sandwich variance estimator $\Sigma(\hat{\beta}^I)$ is evaluated using the estimate of $\text{Var}(\mathbf{y}_i)$ under the working independence model: i.e., using a diagonal \hat{V}_i . QIC^I is an asymptotically unbiased estimator of $E_g\{d_{gq}(\beta_0, \hat{\beta}^I)\}$ (Pan, 2001).

In the GEE framework, let $\hat{\beta}^R$ be the estimator of β obtained using the working correlation matrix $R_i(\alpha)$. A variant of QIC^I can be proposed that takes into account this postulated correlation structure:

$$\text{QIC}^R = -2Q(\hat{\beta}^R|\mathbf{y}) + 2\text{tr}\{\mathcal{I}(\hat{\beta}^R|\mathbf{y})\Sigma(\hat{\beta}^R)\}. \quad (2.34)$$

Here, the sandwich variance estimator $\Sigma(\hat{\beta}^R)$ is evaluated using the estimate of $\text{Var}(\mathbf{y}_i)$ based on the postulated variance/covariance structure: i.e., using $\hat{V}_i(\hat{\alpha})$.

The criterion in (2.34) differs from (2.33) in that it is calculated using $\hat{\beta}^R$. QIC^R is a biased estimator of $E_g\{d_{gq}(\beta_0, \hat{\beta}^I)\}$. Although this does not seem to appreciably affect its performance in terms of choosing a proper mean structure, Pan (2001) recommends the use of QIC^I .

Note that ϕ may need to be estimated. According to Pan (2001), ϕ can be estimated using any method of choice, but the estimate should be based on the largest candidate model; that is, the one including all the available covariates.

When all modeling specifications are correct, $\mathcal{I}(\hat{\beta}^R|\mathbf{y})^{-1}$ and $\Sigma(\hat{\beta}^R)$ are asymptotically equivalent. Thus, $\text{tr}\{\mathcal{I}(\hat{\beta}^R|\mathbf{y})\Sigma(\hat{\beta}^R)\} \approx k$. In that case, QIC^R reduces to a form that is similar to that of AIC,

$$\text{QIC}^U = -2Q(\hat{\beta}^R|\mathbf{y}) + 2k.$$

QIC^U is an approximation to (2.34) and potentially useful in mean structure selection. However, QIC^U cannot be applied to select the working correlation structure

because the postulated correlation structure is not sufficiently represented in either the goodness-of-fit term or the penalty term.

As with AIC, the smaller the value of QIC^I or any of its variants, the better the fitted model explains the data. Hardin and Hilbe (2002) propose using QIC^R to choose among competing correlation structures that cannot be discerned using scientific reasoning. Once the correlation structure is determined, the authors recommend estimating β using the chosen working correlation (i.e., computing $\hat{\beta}^R$), and then utilizing QIC^U to determine the best subset of covariates. We compare the performance of the different forms of QIC in chapter 5.

CHAPTER 3

LIKELIHOOD-BASED MODEL SELECTION CRITERIA DERIVED FROM KULLBACK'S SYMMETRIC DIVERGENCE

In this chapter we present KIC_o and KIC_u , two new model selection criteria based on KSD. KIC_o and KIC_u are developed, in part, by relaxing the assumptions under which AIC and KIC are derived.

In the large-sample justifications of AIC and KIC, it is assumed that $g(\mathbf{y}|\boldsymbol{\theta}_0) \in \mathcal{F}(k)$. This assumption implies that (1) the structure of the fitted model is either correctly specified or overspecified (i.e., the mean and variance/covariance structures of the candidate model include at least all the necessary parameters); and (2) the distribution of \mathbf{y} is correctly specified. We derive KIC_o and KIC_u under a set of less stringent assumptions. We relax (1) in the development of KIC_u and (2) in the development of KIC_o . Thus, in both developments, $g(\mathbf{y}|\boldsymbol{\theta}_0)$ is not necessarily included in the candidate family $\mathcal{F}(k)$. In this respect, KIC_o and KIC_u resemble TIC. Not surprisingly, the penalty terms of KIC_o and KIC_u both include the trace statistic that serves as the basis for the penalty term of TIC.

3.1 KIC_o Derivation

For deriving KIC_o , we assume that the fitted model is correctly specified or overspecified, but that the distribution of \mathbf{y} is potentially misspecified. Distributional misspecification can occur with the error distribution; for example, when a linear model is fitted assuming normally distributed errors, yet the true errors arise from a t distribution. Another form of distributional misspecification arises when the outcomes are overdispersed but the fitted model does not accommodate overdispersion. For instance, when a Poisson random component is fitted but the outcomes follow a negative binomial process, or when a binomial random component is fitted

but the outcomes follow a beta-binomial process. Relaxing the assumption of a correctly specified distribution introduces the notion of the *best approximating model*, $f(\mathbf{y}|\boldsymbol{\theta}_0)$. In the candidate family $\mathcal{F}(k)$, the best approximating model is the model parameterized by $\boldsymbol{\theta}_0$; i.e., the model in $\mathcal{F}(k)$ where the parameters of interest are fixed at values that are identical to those of the data generating model $g(\mathbf{y}|\boldsymbol{\theta}_0)$. However, since the form of the postulated distribution of \mathbf{y} possibly differs from the data generating distribution, $f(\mathbf{y}|\boldsymbol{\theta}_0) \neq g(\mathbf{y}|\boldsymbol{\theta}_0)$.

To further clarify the concept of the best approximating model, consider the following scenarios. Suppose that $g(\mathbf{y}|\boldsymbol{\theta}_0)$ represents a negative binomial distribution, but a Poisson distribution is chosen for $f(\mathbf{y}|\boldsymbol{\theta})$, or that $g(\mathbf{y}|\boldsymbol{\theta}_0)$ represents a beta binomial distribution, but a binomial distribution is chosen for $f(\mathbf{y}|\boldsymbol{\theta})$. In each case, assume that the parameters of interest relate to the mean structure. Although the data generating distribution features an additional parameter to model dispersion, this parameter would not be included among the parameters represented in $\boldsymbol{\theta}_0$. Rather, the vector $\boldsymbol{\theta}_0$ only includes the regression parameters that characterize the mean. Note that the MLE for $\boldsymbol{\theta}_0$ should be consistent under the candidate model $f(\mathbf{y}|\boldsymbol{\theta})$ even though this model is misspecified in terms of the distribution. The best approximating model $f(\mathbf{y}|\boldsymbol{\theta}_0)$ is based on the true parameter vector (i.e., the true mean structure) as applied to the postulated distribution.

In the development of KIC_o , we also assume a set of regularity conditions to ensure that $\hat{\boldsymbol{\theta}}$ satisfies the large-sample properties of MLEs under distributional misspecification, as described in section 2.1.1. Recall that when $L(\boldsymbol{\theta}|\mathbf{y})$ is misspecified, the sandwich variance estimator should be used to calculate $\text{Var}(\hat{\boldsymbol{\theta}})$ because $\mathcal{I}(\boldsymbol{\theta}) \neq \mathcal{J}(\boldsymbol{\theta})$.

For the purpose of the derivations that follow, we define the following information matrices:

$$\begin{aligned}\mathcal{I}_f(\boldsymbol{\theta}) &= E_f(\mathcal{I}(\boldsymbol{\theta}|\mathbf{y})) \text{ and} \\ \mathcal{I}_g(\boldsymbol{\theta}) &= E_g(\mathcal{I}(\boldsymbol{\theta}|\mathbf{y})).\end{aligned}$$

In section 2.2.2, we present $K_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ as a valid substitute for $J_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$. For developing KIC_o , we will use yet another form of KSD. Let

$$2J_{ff}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = d_{ff}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) - d_{ff}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) + d_{ff}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - d_{ff}(\boldsymbol{\theta}, \boldsymbol{\theta}). \quad (3.1)$$

The preceding expression is another possible KSD. It only differs from $J_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ in that the data generating model $g(\mathbf{y}|\boldsymbol{\theta}_0)$ is replaced by the best approximating model $f(\mathbf{y}|\boldsymbol{\theta}_0)$.

Adding the constant $d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ to (3.1) does not alter the ranking of competing models. Thus,

$$L_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = d_{ff}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) - d_{ff}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) + d_{ff}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - d_{ff}(\boldsymbol{\theta}, \boldsymbol{\theta}) + d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \quad (3.2)$$

is also an appealing measure for creating KSD-based model selection criteria. This measure will serve as the basis for the development of KIC_o .

If we evaluate (3.2) at $\hat{\boldsymbol{\theta}}$ and take the expected value with respect to $g(\mathbf{y}|\boldsymbol{\theta}_0)$, we obtain

$$\begin{aligned}\Phi_{gf}(\boldsymbol{\theta}_0) &\equiv E_g\{L_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\} \\ &= E_g\{d_{ff}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) - d_{ff}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) + d_{ff}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}) + d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\}.\end{aligned}$$

We may then decompose $\Phi_{gf}(\boldsymbol{\theta}_0)$ as follows:

$$\begin{aligned}\Phi_{gf}(\boldsymbol{\theta}_0) &= E_g\{-2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} \\ &\quad + d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) - E_g\{-2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\}\end{aligned}\tag{3.3}$$

$$+ E_g\{d_{ff}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\} - d_{ff}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\tag{3.4}$$

$$+ E_g\{d_{ff}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})\}.\tag{3.5}$$

The purpose of the preceding decomposition is to introduce $-2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})$ as a platform for estimating $\Phi_{gf}(\boldsymbol{\theta}_0)$. According to the developments that follow, (3.3), (3.4), and (3.5) are positive; hence, $-2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})$ is negatively biased. If we can obtain estimates of these terms, we can correct for the negative bias.

Consider taking a second-order Taylor series expansion of $-2\ln f(\mathbf{y}|\boldsymbol{\theta}_0)$ about $\hat{\boldsymbol{\theta}}$. Since $\ln f(\mathbf{y}|\boldsymbol{\theta}_0)$ is maximized at $\hat{\boldsymbol{\theta}}$, one can establish

$$-2\ln f(\mathbf{y}|\boldsymbol{\theta}_0) = -2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + r_1(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}).$$

Here, as n increases, $r_1(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ is $o_p(1)$; $E_g\{r_1(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\}$ is therefore $o(1)$. Taking the expectation of both sides of this expansion with respect to $g(\mathbf{y}|\boldsymbol{\theta}_0)$ yields

$$d_{gf}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) - E_g\{-2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} = E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} + o(1).\tag{3.6}$$

Next, consider taking a second-order Taylor series expansion in the second argument of $d_{ff}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ about $\boldsymbol{\theta}_0$, and a second-order expansion in the second argument of $d_{ff}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ about $\hat{\boldsymbol{\theta}}$. By definition, $d_{ff}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ is minimized when $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$ and $d_{ff}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ is minimized when $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}$. Thus,

$$d_{ff}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = d_{ff}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}_f(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + r_2(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) \text{ and}$$

$$d_{ff}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) = d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}_f(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + r_3(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}).$$

Here, $r_2(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $r_3(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ are both $o_p(1)$; $E_g\{r_2(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\}$ and $E_g\{r_3(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\}$ are therefore $o(1)$. Taking the expectation of both sides of these two last expansions

with respect to $g(\mathbf{y}|\boldsymbol{\theta}_0)$ yields

$$E_g\{d_{ff}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\} - d_{ff}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}_f(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} + o(1) \text{ and} \quad (3.7)$$

$$E_g\{d_{ff}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})\} = E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}_f(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} + o(1). \quad (3.8)$$

Thus, replacing (3.3) by (3.6), (3.4) by (3.7), and (3.5) by (3.8), $\Phi_{gf}(\boldsymbol{\theta}_0)$ can be written as follows:

$$\begin{aligned} \Phi_{gf}(\boldsymbol{\theta}_0) &= E_g\{-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} \\ &\quad + E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} \\ &\quad + E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}_f(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} \\ &\quad + E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}_f(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} + o(1). \end{aligned}$$

Since the structure of the fitted model is either correctly specified or overspecified, $\hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_0$ as $n \rightarrow \infty$. Thus, we can approximate the first quadratic form with $E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}_g(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\}$. Also, we can approximate the third quadratic form with $E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}_f(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\}$. Hence, $\Phi_{gf}(\boldsymbol{\theta}_0)$ becomes

$$\begin{aligned} \Phi_{gf}(\boldsymbol{\theta}_0) &= E_g\{-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} \\ &\quad + E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}_g(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} \\ &\quad + 2E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}_f(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} + o(1). \end{aligned}$$

Given that $E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}_g(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\}$ and $E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathcal{I}_f(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\}$ are scalars, we have

$$\begin{aligned} \Phi_{gf}(\boldsymbol{\theta}_0) &= E_g\{-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} + \text{tr}\{\mathcal{I}_g(\boldsymbol{\theta}_0) E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\}\} \\ &\quad + 2[\text{tr}\{\mathcal{I}_f(\boldsymbol{\theta}_0) E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\}\}] + o(1) \\ &= E_g\{-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} + \text{tr}\{\mathcal{I}_g(\boldsymbol{\theta}_0) \Sigma(\boldsymbol{\theta}_0)\} \\ &\quad + 2[\text{tr}\{\mathcal{I}_f(\boldsymbol{\theta}_0) \Sigma(\boldsymbol{\theta}_0)\}] + o(1). \end{aligned}$$

Let $\mathcal{J}_g(\boldsymbol{\theta}_0) = E_g(\mathcal{J}(\boldsymbol{\theta}_0|\mathbf{y}))$. Replacing $\Sigma(\boldsymbol{\theta}_0)$ by result (2.4), we obtain

$$\begin{aligned}\Phi_{gf}(\boldsymbol{\theta}_0) &= E_g\{-2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} + \text{tr}\{\mathcal{J}_g(\boldsymbol{\theta}_0)\mathcal{I}_g(\boldsymbol{\theta}_0)^{-1}\} \\ &\quad + 2[\text{tr}\{\mathcal{I}_f(\boldsymbol{\theta}_0)\mathcal{I}_g(\boldsymbol{\theta}_0)^{-1}\mathcal{J}_g(\boldsymbol{\theta}_0)\mathcal{I}_g(\boldsymbol{\theta}_0)^{-1}\}] + o(1).\end{aligned}$$

Now, define the statistic KIC_o as follows:

$$\begin{aligned}\text{KIC}_o &= -2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + \text{tr}\{\mathcal{J}(\hat{\boldsymbol{\theta}}|\mathbf{y})\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})^{-1}\} \\ &\quad + 2[\text{tr}\{\mathcal{I}_f(\hat{\boldsymbol{\theta}})\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})^{-1}\mathcal{J}(\hat{\boldsymbol{\theta}}|\mathbf{y})\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})^{-1}\}].\end{aligned}\tag{3.9}$$

Based on the preceding development, one can conclude that

$$E_g\{\text{KIC}_o\} + o(1) = \Phi_{gf}(\boldsymbol{\theta}_0).$$

In the GLM framework, when the canonical link is used, the expected information equals the observed information because the Hessian matrix does not depend on the data. That is, $\mathcal{I}_f(\boldsymbol{\theta}_0) = \mathcal{I}_g(\boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\theta}|\mathbf{y})$. Hence, expression (3.9) becomes

$$\begin{aligned}\text{KIC}_o &= -2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + \text{tr}\{\mathcal{J}(\hat{\boldsymbol{\theta}}|\mathbf{y})\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})^{-1}\} \\ &\quad + 2[\text{tr}\{\mathcal{J}(\hat{\boldsymbol{\theta}}|\mathbf{y})\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})^{-1}\}] \\ &= -2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + 3[\text{tr}\{\mathcal{J}(\hat{\boldsymbol{\theta}}|\mathbf{y})\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})^{-1}\}].\end{aligned}$$

In summary, KIC_o is an asymptotically unbiased estimator of $\Phi_{gf}(\boldsymbol{\theta}_0)$ for models with correctly or overspecified mean and variance/covariance structures. Note that even though we are assuming that the fitted model is correctly specified or overspecified, the trace terms in KIC_o do not reduce to k because we allow for misspecification of the likelihood.

3.2 KIC_u Derivation

We propose KIC_u as another asymptotically unbiased estimator of KSD. Let $\mathcal{F} = \{\mathcal{F}(k_1), \mathcal{F}(k_2), \dots, \mathcal{F}(k_L)\}$ represent the collection of candidate families. We

refer to \mathcal{F} as the *candidate set*. Assume that the largest candidate family in \mathcal{F} is $\mathcal{F}(K)$: i.e., $K = \max\{k_1, k_2, \dots, k_L\}$. The *largest candidate model* refers to the model defined by the structure of the candidate family $\mathcal{F}(K)$. In a linear regression setting, this model would include all available covariates that have been collected to explain the outcome. In a linear mixed modeling framework, this model would not only include all available covariates, but its variance/covariance structure would subsume all other structures represented in the candidate set.

KIC_u is based on the assumption that $g(\mathbf{y}|\boldsymbol{\theta}_0) \in \mathcal{F}(K)$. This assumption implies that (1) the data distribution is correctly specified, and (2) the generating model is subsumed by the largest, K -dimensional, candidate model (i.e., the largest candidate model is either correctly specified or overspecified). Thus, in terms of the model structures, various models in the candidate set \mathcal{F} may be either correctly specified, underspecified or overspecified.

In the development of KIC_u, we also assume a set of regularity conditions to ensure that $\hat{\boldsymbol{\theta}}$ satisfies the large-sample properties of MLEs under structural misspecification, as described in section 2.1.1. Thus, the sandwich variance estimator should be used to estimate $Var(\hat{\boldsymbol{\theta}})$.

Let $\boldsymbol{\theta}_0$ represent the data generating model parameter. Let $\boldsymbol{\theta}_*$ be the K -dimensional parameter for a candidate model in $\mathcal{F}(K)$ and let $\hat{\boldsymbol{\theta}}_*$ be its MLE. For the largest candidate model, $\hat{\boldsymbol{\theta}}_* \rightarrow \boldsymbol{\theta}_0$ as $n \rightarrow \infty$. For other candidate models, we may only assume that $\hat{\boldsymbol{\theta}} \rightarrow \bar{\boldsymbol{\theta}}$ as $n \rightarrow \infty$, since the candidate family $\mathcal{F}(k)$ may be underspecified. Underspecified candidate models will be based on mean and/or variance/covariance structures that are insufficient to accommodate the data.

In section 2.2.3, we present $\Omega_{gf}(\boldsymbol{\theta}_0)$ as an appealing measure for creating

KSD-based model selection criteria. For deriving KIC_u , we use the following decomposition:

$$\begin{aligned}\Omega_{gf}(\boldsymbol{\theta}_0) &= E_g\{-2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} \\ &\quad + d_{gf}(\boldsymbol{\theta}_0, \bar{\boldsymbol{\theta}}) - E_g\{-2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\}\end{aligned}\tag{3.10}$$

$$+ E_g\{d_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\} - d_{gf}(\boldsymbol{\theta}_0, \bar{\boldsymbol{\theta}})\tag{3.11}$$

$$+ E_g\{d_{gf}(\boldsymbol{\theta}_0, \bar{\boldsymbol{\theta}}) - d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})\}\tag{3.12}$$

$$+ E_g\{d_{fg}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - d_{gf}(\boldsymbol{\theta}_0, \bar{\boldsymbol{\theta}})\}.\tag{3.13}$$

Again, the preceding decomposition allows us to introduce $-2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})$ as a platform for estimating $\Omega_{gf}(\boldsymbol{\theta}_0)$. We can correct for the negative bias of $-2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})$ as an estimator for $\Omega_{gf}(\boldsymbol{\theta}_0)$ by obtaining estimators for (3.10), (3.11), (3.12), and (3.13).

Consider taking a second-order Taylor series expansion of $-2\ln f(\mathbf{y}|\bar{\boldsymbol{\theta}})$ about $\hat{\boldsymbol{\theta}}$, and a second-order expansion in the second argument of $d_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ about $\bar{\boldsymbol{\theta}}$. The log likelihood $\ln f(\mathbf{y}|\bar{\boldsymbol{\theta}})$ is maximized at $\hat{\boldsymbol{\theta}}$. Also, $d_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ is minimized when $\hat{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}$ (as opposed to $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$), because $g(\mathbf{y}|\boldsymbol{\theta}_0)$ is not necessarily included in the candidate model family. Hence, one can establish

$$\begin{aligned}-2\ln f(\mathbf{y}|\bar{\boldsymbol{\theta}}) &= -2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})' \mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) + r_4(\bar{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}) \text{ and} \\ d_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) &= d_{gf}(\boldsymbol{\theta}_0, \bar{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})' \mathcal{I}_g(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) + r_5(\bar{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}).\end{aligned}$$

Here, as n increases, $r_4(\bar{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})$ and $r_5(\bar{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})$ are $o_p(1)$; $E_g\{r_4(\bar{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})\}$ and $E_g\{r_5(\bar{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})\}$ are therefore $o(1)$. Taking the expectation of both sides of these two last expansions with respect to $g(\mathbf{y}|\boldsymbol{\theta}_0)$ yields

$$d_{gf}(\boldsymbol{\theta}_0, \bar{\boldsymbol{\theta}}) - E_g\{-2\ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} = E_g\{(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})' \mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\} + o(1) \text{ and}\tag{3.14}$$

$$E_g\{d_{gf}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\} - d_{gf}(\boldsymbol{\theta}_0, \bar{\boldsymbol{\theta}}) = E_g\{(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})' \mathcal{I}_g(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\} + o(1).\tag{3.15}$$

To obtain an approximation for (3.12), we use the following result:

$$d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}) = -2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}). \quad (3.16)$$

For a normal linear model, it is straightforward to demonstrate (3.16). However, it is not trivial to generalize (3.16) to other distributions in the exponential family. We have found empirically that (3.16) holds for the Bernoulli and for the parameter-dependent part of the Binomial and Poisson distributions: i.e., $p^y(1-p)^{n-y}$ and $e^{-\lambda}\lambda^y$, respectively. However, it remains to be shown theoretically that (3.16) is a general result that holds for all distributions in the exponential family.

Using result (3.16), (3.12) can be expressed as

$$\begin{aligned} E_g\{d_{gf}(\boldsymbol{\theta}_0, \bar{\boldsymbol{\theta}}) - (-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}))\} &= d_{gf}(\boldsymbol{\theta}_0, \bar{\boldsymbol{\theta}}) - E_g\{-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} \\ &= E_g\{-2 \ln f(\mathbf{y}|\bar{\boldsymbol{\theta}})\} - E_g\{-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\}. \end{aligned} \quad (3.17)$$

Note that the left-hand side of (3.17) is the same as the left-hand side of (3.14).

Next, we approximate (3.13) using a result that has been established empirically yet not theoretically. Given that $\hat{\boldsymbol{\theta}}_* \rightarrow \boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}} \rightarrow \bar{\boldsymbol{\theta}}$ as $n \rightarrow \infty$, and that the data distribution is correctly specified, the statistic

$$d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_*) - d_{ff}(\hat{\boldsymbol{\theta}}_*, \hat{\boldsymbol{\theta}}) \quad (3.18)$$

appears to be a reasonable approximation for (3.13). However, simulation results show that (3.18) is negatively biased by a factor of $K - k$, where K and k are the dimensions of $\boldsymbol{\theta}_*$ and $\boldsymbol{\theta}$, respectively. Hence, we propose

$$d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_*) - d_{ff}(\hat{\boldsymbol{\theta}}_*, \hat{\boldsymbol{\theta}}) - K + k \quad (3.19)$$

to approximate (3.13). It remains to be shown theoretically that (3.19) is an asymptotically unbiased estimator of (3.13).

Replacing (3.10) and (3.12) by (3.14), (3.11) by (3.15), and (3.13) by

$$E_g\{d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_*) - d_{ff}(\hat{\boldsymbol{\theta}}_*, \hat{\boldsymbol{\theta}}) - K + k\},$$

$\Omega_{gf}(\boldsymbol{\theta}_0)$ can be represented as follows:

$$\begin{aligned}\Omega_{gf}(\boldsymbol{\theta}_0) &= E_g\{-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} \\ &\quad + E_g\{(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})' \mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\} \\ &\quad + E_g\{(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})' \mathcal{I}_g(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\} \\ &\quad + E_g\{(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})' \mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\} \\ &\quad + E_g\{d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_*) - d_{ff}(\hat{\boldsymbol{\theta}}_*, \hat{\boldsymbol{\theta}}) - K + k\} + o(1).\end{aligned}$$

Since $\hat{\boldsymbol{\theta}} \rightarrow \bar{\boldsymbol{\theta}}$ as $n \rightarrow \infty$, the first and third quadratic forms in the preceding expression can be approximated by $E_g\{(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})' \mathcal{I}_g(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\}$. Hence, $\Omega_{gf}(\boldsymbol{\theta}_0)$ becomes

$$\begin{aligned}\Omega_{gf}(\boldsymbol{\theta}_0) &= E_g\{-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} + 3E_g\{(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})' \mathcal{I}_g(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\} \\ &\quad + E_g\{d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_*) - d_{ff}(\hat{\boldsymbol{\theta}}_*, \hat{\boldsymbol{\theta}}) - K + k\} + o(1).\end{aligned}$$

Given that $E_g\{(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})' \mathcal{I}_g(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\}$ is a scalar, we have

$$\begin{aligned}\Omega_{gf}(\boldsymbol{\theta}_0) &= E_g\{-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} + 3[\text{tr}\{\mathcal{I}_g(\bar{\boldsymbol{\theta}})E_g\{(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})'(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\}\}] \\ &\quad + E_g\{d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_*) - d_{ff}(\hat{\boldsymbol{\theta}}_*, \hat{\boldsymbol{\theta}}) - K + k\} + o(1) \\ &= E_g\{-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} + 3[\text{tr}\{\mathcal{I}_g(\bar{\boldsymbol{\theta}})\Sigma(\bar{\boldsymbol{\theta}})\}] \\ &\quad + E_g\{d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_*) - d_{ff}(\hat{\boldsymbol{\theta}}_*, \hat{\boldsymbol{\theta}}) - K + k\} + o(1).\end{aligned}$$

Replacing $\Sigma(\bar{\boldsymbol{\theta}})$ by result (2.5), we obtain

$$\begin{aligned}\Omega_{gf}(\boldsymbol{\theta}_0) &= E_g\{-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} + 3[\text{tr}\{\mathcal{I}_g(\bar{\boldsymbol{\theta}})\mathcal{I}_g(\bar{\boldsymbol{\theta}})^{-1}\}] \\ &\quad + E_g\{d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_*) - d_{ff}(\hat{\boldsymbol{\theta}}_*, \hat{\boldsymbol{\theta}}) - K + k\} + o(1).\end{aligned}$$

Now, define the statistic KIC_u as follows:

$$\text{KIC}_u = -2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + 3\text{tr}\{\mathcal{J}(\hat{\boldsymbol{\theta}}|\mathbf{y})\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})^{-1}\} + d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_*) - d_{ff}(\hat{\boldsymbol{\theta}}_*, \hat{\boldsymbol{\theta}}) - K + k.$$

In light of the preceding development, one can conclude that

$$E_g\{\text{KIC}_u\} + o(1) = \Omega_{gf}(\boldsymbol{\theta}_0).$$

Thus, KIC_u is an asymptotically unbiased estimator of $\Omega_{gf}(\boldsymbol{\theta}_0)$ for models with correctly specified, overspecified or underspecified mean and variance/covariance structures.

Note that KIC_u has the following form:

$$\text{KIC}_u = \text{KIC}_o + \text{underfitting penalty term}.$$

For correctly specified or overspecified models, the expectation of the underfitting penalty term is approximately zero. For underfitted models, the size of the term reflects the degree of underspecification.

CHAPTER 4

SIMULATION STUDIES FOR KIC_O AND KIC_U

The assumptions employed in the development of KIC_o and KIC_u allow us to apply these criteria in the selection of a wide range of models. Importantly, this includes models grouped under the umbrella of GLMs for uncorrelated responses. The goal of the simulation studies presented in this chapter is to characterize the performances of KIC_o and KIC_u , illustrating their behaviors in scenarios where, given the assumptions used in their development, each criterion is expected to excel. Our simulation studies are based on factorial designs. We consider settings based on two modeling objectives (i.e., selection of models with a correctly specified mean structure and of models with optimal predictive properties), two types of candidate sets (i.e., nested models and all possible models), and three types of GLMs (i.e., linear, logistic and Poisson regressions) both without misspecification and with ignored misspecification. We also explore the effect of regressor collinearity on the performance of the criteria.

The two modeling objectives considered are to select models with a correctly specified mean structure and to choose a model with optimal predictive properties. Selecting a model with a correctly specified mean structure is reflective of the usual situation that arises in the practice of clinical health sciences, where the researcher has various candidate covariates that could explain an outcome, but he or she is unsure of which covariates should be included in the model. Simulation experiments can imitate this problem by including in the set of candidate models the data generating model, together with other candidate models including fewer or more of the candidate covariates. In this scenario, the model selection criterion that chooses the generating model most often is considered best.

Selecting a model with optimal predictive properties is directly related to a

typical situation in the basic sciences where the model explaining a natural phenomenon is well characterized by different effects of various magnitudes, but the sample size is limited and will not allow all of the effects to be accurately estimated. In this situation, the goal is to decide which of the explanatory variables should be included in the final model to optimize prediction of the outcome given a new sample. To simulate this type of setting, the highest order model in the candidate set is the generating model, and this model contains strong, moderate, and weak effects. Thus, all candidate models, except for the one with the highest order, are underspecified. The best model selection criterion in this type of modeling problem is one that chooses models which minimize some measure of prediction error. This modeling objective is more aligned with the assumptions made in the development of KIC_u .

Each simulation experiment presented is based on 1,000 replications. In these experiments, *order* refers to the number of parameters in the linear predictor, referred to as p in chapter 2. When the candidate set \mathcal{F} contains nested candidate families or nested models (NM), the model selection problem is reduced to choosing an appropriate order. Even though NM simulations settings do not mirror the manner in which regression models are selected in practice, simulation experiments with NM allow a better conceptualization of underfitting and overfitting than when the candidate set contains all possible models (APM). A candidate model is underfitted unless the generating model is subsumed by the candidate model. In the APM setting, models of the same order as the generating model but with a different mean structure are underfitted. In fact, candidate models can have a higher order than the generating model but be underfitted. Thus, in APM settings, the number of underfitted candidate models grows quickly with the number of covariates considered. APM candidate sets are more reflective of real life model selection situations than NM sets.

We illustrate the performances of KIC_o and KIC_u using simulations based on linear, logistic and Poisson regression frameworks. We compare their performances to three other criteria that can be used in the GLM framework: AIC, TIC, and KIC. We also compare the criteria to suitable oracles. Oracles serve as gold standards for the selection criteria and allow us to evaluate the best possible criterion performance in each experiment.

For computational tractability, the two oracles we use in the simulation experiments are

$$d_{ff}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = E_f\{-2 \ln f(\mathbf{y}|\hat{\boldsymbol{\theta}})\} \text{ and} \quad (4.1)$$

$$K_{ff}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = d_{ff}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) + d_{ff}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - d_{ff}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}). \quad (4.2)$$

Expression (4.1) can be used as an oracle for KDD-based criteria and (4.2) can be used as an oracle for KSD-based criteria. For brevity of notation, we will use $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ instead of $d_{ff}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K_{ff}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$, respectively.

4.1 Selection of a Model with Correctly Specified Mean Structure

All of the simulation studies in this section are structured as factorial experiments with two levels of sample size and different types of misspecification. Different levels of distributional misspecification are introduced in order to reflect settings that follow the assumptions used in the development of KIC_o . In settings where distributional misspecification is present, the models are fitted ignoring such misspecification.

Data are generated following the configurations detailed in each section. For each replication, AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$, and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ are evaluated for each fitted candidate model. The fitted model with the minimum value for each statistic or oracle is recorded. The results are then summarized in tables and figures.

4.1.1 Linear Regression with Correctly Specified Error Distribution and with Incorrectly Specified Error Distribution

For continuous outcomes, one can often assume $y_i \sim N(\mu_i, \sigma^2)$ for $i = 1, 2, \dots, n$, with $E(y_i) = \mu_i$. If the goal is to explain a sample \mathbf{y} with effects represented by a linear combination of explanatory variables (a.k.a, covariates or regressors), then the GLM of choice is linear regression. The canonical link function for linear regression is the identity link:

$$\mu_i = \sum_{j=1}^p x_{ij}\beta_j,$$

where the β_j s are the parameters we seek to estimate.

For this framework, we configure a simulation experiment based on six settings with the characteristics that follow. The latter four of these six settings represent a 2x2 factorial design. Data are generated using the third-order model

$$y_i = 1 + x_{i2} + x_{i3} + \epsilon_i,$$

where $\epsilon_i \sim N(0, 3)$ and $n = 30$ (setting 1) or $n = 60$ (setting 2), $\epsilon_i \sim t(2)$ and $n = 60$ (setting 3) or $n = 120$ (setting 4), and $\epsilon_i \sim I * N(0, 1) + (1 - I) * N(0, 10)$ ($I \sim \text{Bernoulli}(0.95)$) and $n = 60$ (setting 5) or $n = 120$ (setting 6). Settings 1 and 2 do not present error misspecification, since the normal distribution is assumed in fitting candidate models to the data. Settings 3 to 6 present two forms of error misspecification, with error distributions having thicker tails than the normal distribution.

For the simulations featuring NM, the model selection criteria choose from the

following candidate set:

$$\begin{aligned}
y_i &= \beta_1 + \epsilon_i, \\
y_i &= \beta_1 + \beta_2 x_{i2} + \epsilon_i, \\
&\vdots \\
y_i &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i,
\end{aligned}$$

where x_{ij} ($j = 2, \dots, 5$) are independent and identically distributed (iid) as $N(0, 1)$, and ϵ_i are iid $N(0, \sigma^2)$.

For APM simulations, a total of 15 models are included in the candidate set (i.e., all combinations of the covariates, except for x_{i1} , excluding the intercept-only model). We choose to exclude the intercept-only model from the APM selections with the intent of making this simulation experiment as close as possible to a real-world setting. However, the intercept-only model is included in the NM simulations.

The results for the six settings are summarized in Tables 4.1 to 4.6. In all cases, $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ chooses the data generating model at least as often as $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$. The only instances where the two oracles exhibit a very similar number of correct model selections are those where the correct selection rate for $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ is 97% or higher, leaving almost no room for improvement.

For NM order selections (Tables 4.1 to 4.3), although all criteria perform well, KIC_o is the criterion that chooses the data generating model most often. The most remarkable difference between KIC_o and the rest of the criteria is that KIC_o chooses overspecified models less often. However, KIC_o selects underspecified models more often than any of the other criteria. Note that KIC_u protects against underfitting, as does AIC; however, KIC_u chooses overfitted models less often than AIC.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
30	5	117	77	49	20	88	0	0
30	4	154	138	94	61	150	0	0
30	3	677	721	753	781	708	952	973
30	2	49	59	85	113	51	43	27
30	1	3	5	19	25	3	5	0
60	5	89	77	46	28	81	0	0
60	4	130	111	78	68	116	0	0
60	3	779	810	870	897	801	997	999
60	2	2	2	6	7	2	3	1
60	1	0	0	0	0	0	0	0

Table 4.1: Settings 1 and 2: NM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the linear regression framework; results for the generating model are bolded.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
60	5	84	65	24	12	70	0	0
60	4	141	134	77	59	138	0	0
60	3	690	710	757	773	707	983	986
60	2	55	58	88	98	55	15	13
60	1	30	33	54	58	30	2	1
120	5	71	65	18	12	67	0	0
120	4	130	121	74	69	126	0	0
120	3	771	784	857	866	780	996	996
120	2	15	17	28	30	15	3	3
120	1	13	13	23	23	12	1	1

Table 4.2: Settings 3 and 4: NM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the linear regression framework; results for the generating model are bolded.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
60	5	80	63	20	9	71	0	0
60	4	140	124	74	63	129	0	0
60	3	707	736	775	789	730	838	860
60	2	55	59	81	87	55	114	132
60	1	18	18	50	52	15	48	8
120	5	77	65	33	26	65	0	0
120	4	143	135	81	71	139	0	0
120	3	773	793	874	891	789	979	983
120	2	6	6	10	10	6	18	17
120	1	1	1	2	2	1	3	0

Table 4.3: Settings 5 and 6: NM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the linear regression framework; results for the generating model are bolded.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
30	overfitted	925	815	996	941	723	78	47
30	correct	51	141	4	51	213	922	953
30	underfitted	24	44	0	8	64	0	0
60	overfitted	115	127	29	39	154	0	0
60	correct	413	520	153	325	569	996	999
60	underfitted	472	353	818	636	277	4	1

Table 4.4: Settings 1 and 2: APM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the linear regression framework.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
60	overfitted	56	60	17	21	69	0	0
60	correct	227	304	88	172	334	973	978
60	underfitted	717	636	895	807	597	27	22
120	overfitted	116	125	33	35	127	0	0
120	correct	494	530	336	405	552	992	992
120	underfitted	390	345	631	560	321	8	8

Table 4.5: Settings 3 and 4: APM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the linear regression framework.

The tables featuring results for APM simulations (Tables 4.4 to 4.6) show that none of the criteria exceed a 60% rate of selection for the correct model. In these scenarios, KIC_u selects the correct model more often than the rest of the criteria,

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
60	overfitted	83	97	29	29	99	0	0
60	correct	330	382	221	308	414	821	841
60	underfitted	587	521	750	663	487	179	159
120	overfitted	128	133	48	48	144	0	0
120	correct	519	554	368	438	569	975	980
120	underfitted	353	313	584	514	287	25	20

Table 4.6: Settings 5 and 6: APM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the linear regression framework.

and it is the criterion that exhibits the best protection against underfitting when $n \geq 60$. KIC_u and TIC are the only criteria that sometimes exceed a 55% rate of correct selections.

As mentioned in the introduction to this chapter, in APM settings, the number of underfitted candidate models grows quickly with the number of covariates considered. Thus, not only is the candidate model set larger than in the NM simulations (i.e., 15 versus 5 candidate models), but the number of underspecified models is much larger (i.e., 11 underspecified candidate models in the APM setting versus 2 in the NM setting). Given the assumptions under which KIC_u was developed, the prevalence of underspecified models is a plausible explanation of why KIC_u works better in APM than in NM simulations.

Fitting a model that ignores error misspecification appears to slightly affect the performance of all criteria and their oracles (Tables 4.2, 4.3, 4.5, and 4.6). The two forms of error misspecification considered yield similar results. Even in the presence of ignored error misspecification, KIC_o outperforms the rest of the criteria

in NM selections (Tables 4.2 and 4.3), while KIC_u does so in APM selections (Tables 4.5 and 4.6).

4.1.2 Linear Regression with Collinear Regressors

We also compile simulation experiments for linear regression based on collinear regressors. In this setting, we configure a $2 \times 2 \times 4 \times 11$ factorial experiment with two levels of sample size (i.e., $n = 30$ and $n = 60$), two levels of error misspecification, four types of collinearity among the candidate covariates, and 11 levels of covariate correlation. We generate data using the third-order model

$$y_i = 1 + x_{i2} + 2x_{i3} + \epsilon_i,$$

where $\epsilon_i \sim N(0, 3)$ (no error misspecification) or $\epsilon_i \sim t(2)$ (ignored error misspecification). The rationale for settings with $\epsilon_i \sim t(2)$ is as described for settings 3 and 4 in the previous section.

We report only NM selections, where the model selection criteria choose from the following candidate set:

$$\begin{aligned} y_i &= \beta_1 + \epsilon_i, \\ y_i &= \beta_1 + \beta_2 x_{i2} + \epsilon_i, \\ &\vdots \\ y_i &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i, \end{aligned}$$

where $(x_{i2}, x_{i3}, x_{i4}, x_{i5})' \sim N_4(\mathbf{0}, \Sigma)$ and $\epsilon_i \sim N(0, \sigma^2)$. Σ corresponds to one of the following structures:

$$\Sigma_1 = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\Sigma_3 = \begin{pmatrix} 1 & 0 & \rho & 0 \\ 0 & 1 & 0 & 0 \\ \rho & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \text{ or } \Sigma_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \rho & 0 \\ 0 & \rho & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Here, ρ is chosen among the following values: 0, 0.10, 0.25, 0.40, 0.55, 0.70, 0.75, 0.80, 0.85, 0.90, or 0.95. Σ_1 (henceforth referred to as Type 1) corresponds to all pairwise correlations among x_{i2} to x_{i5} , Σ_2 (Type 2) corresponds to x_{i2} correlated with x_{i3} , Σ_3 (Type 3) corresponds to x_{i2} correlated with x_{i4} , and Σ_4 (Type 4) corresponds to x_{i3} correlated with x_{i4} .

We are interested in evaluating how the different degrees of correlation affect the performance of the model selection criteria, and how the performance varies depending on whether the correlation is among the regressors in the data generating model or not. This interest provides the rationale for the four preceding correlation structures. We also seek to determine if the results depend on whether a spurious regressor (x_{i4}) is correlated with a regressor corresponding to a weaker effect (x_{i2}) or a stronger effect (x_{i3}) in the data generating model. This issue is investigated using the Type 3 and 4 correlation structures. The results are summarized in Figures 4.1 to 4.4.

Figure 4.1 shows the effect of the four different regressor correlation types on the performance of the five criteria and their oracles. The performance ranking for the criteria is the same as that observed for the NM selections in setting 1 in the

previous section. Based on correct model selections, KIC_o outperforms the rest of the criteria, except when the covariates in the data generating model are highly correlated (i.e., $\rho > 0.80$). When the regressors in the data generating model are correlated (Type 1 and Type 2 plots), the performance of the criteria falls abruptly when ρ is 0.8 or higher. The performance is not affected by regressor correlation if the collinearity is between a data generating model covariate and a spurious covariate (Type 3 and Type 4 plots). This holds true even if the correlation occurs between a data generating model covariate corresponding to a stronger effect (Type 3 versus Type 4 plots).

Figure 4.2 shows that the presence of ignored error misspecification degrades the performance of all the criteria and their oracles. KIC_o exhibits the best performance, as is the case with the results in Table 4.2. In Figure 4.2, the effect of collinearity between the regressors is evident only with Type 1 and Type 2 correlation structures, as in Figure 4.1. For Type 1 and Type 2 correlation structures, the performance of the criteria and oracles starts decreasing abruptly at $\rho = 0.6$.

Increasing the sample size from $n = 30$ to $n = 60$ (Figures 4.3 and 4.4) improves the performance of all criteria and alleviates the negative effect of the correlation between the regressors on the proportion of correct selections. This is evident in the Type 1 and Type 2 plots with no error misspecification (Figure 4.3). The performance of the criteria begins to decline at $\rho = 0.95$ compared to $\rho = 0.80$ for $n = 30$ in Figure 4.1. This shift is, however, less marked when ignored error misspecification is present (Figures 4.4 versus 4.2).

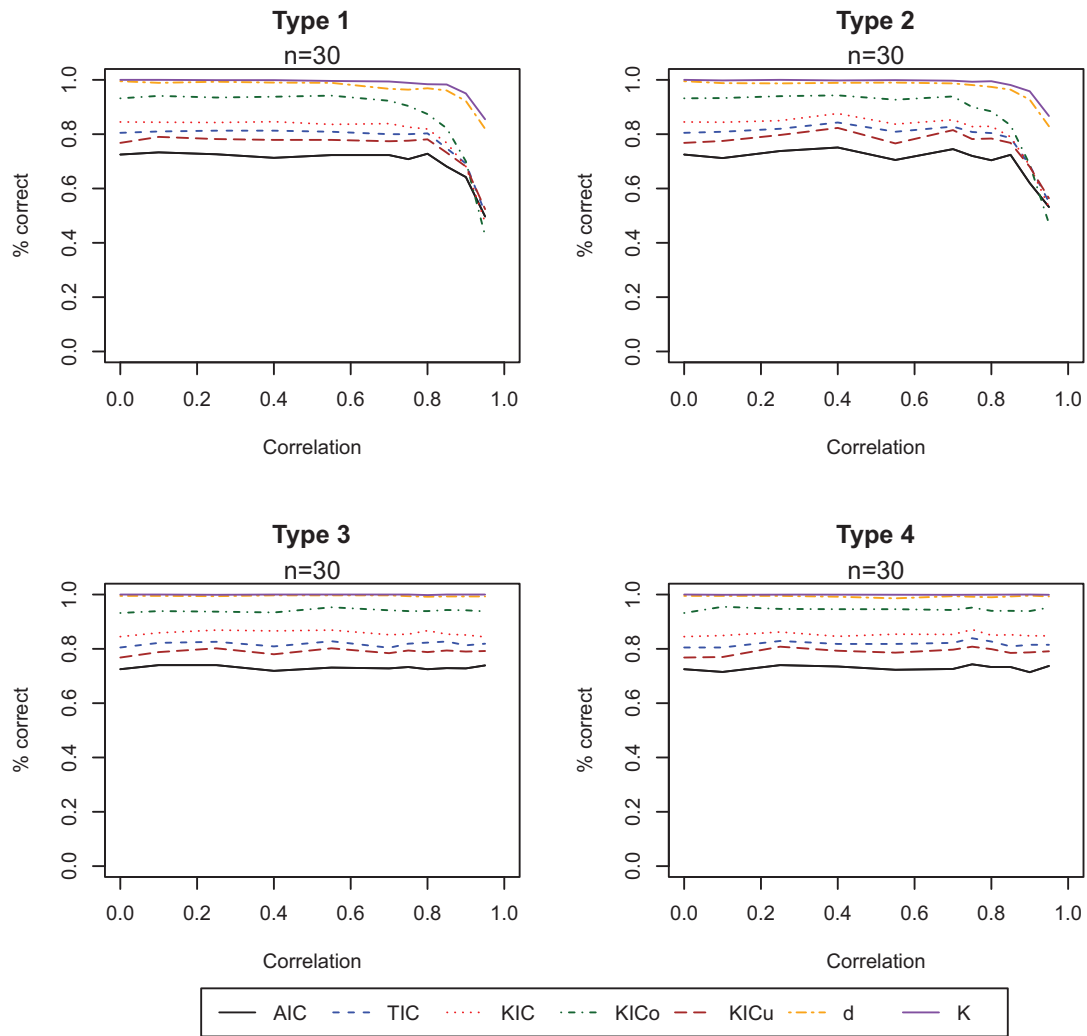


Figure 4.1: Model selection criteria comparison with correlated regressors, $n = 30$, no misspecified errors.

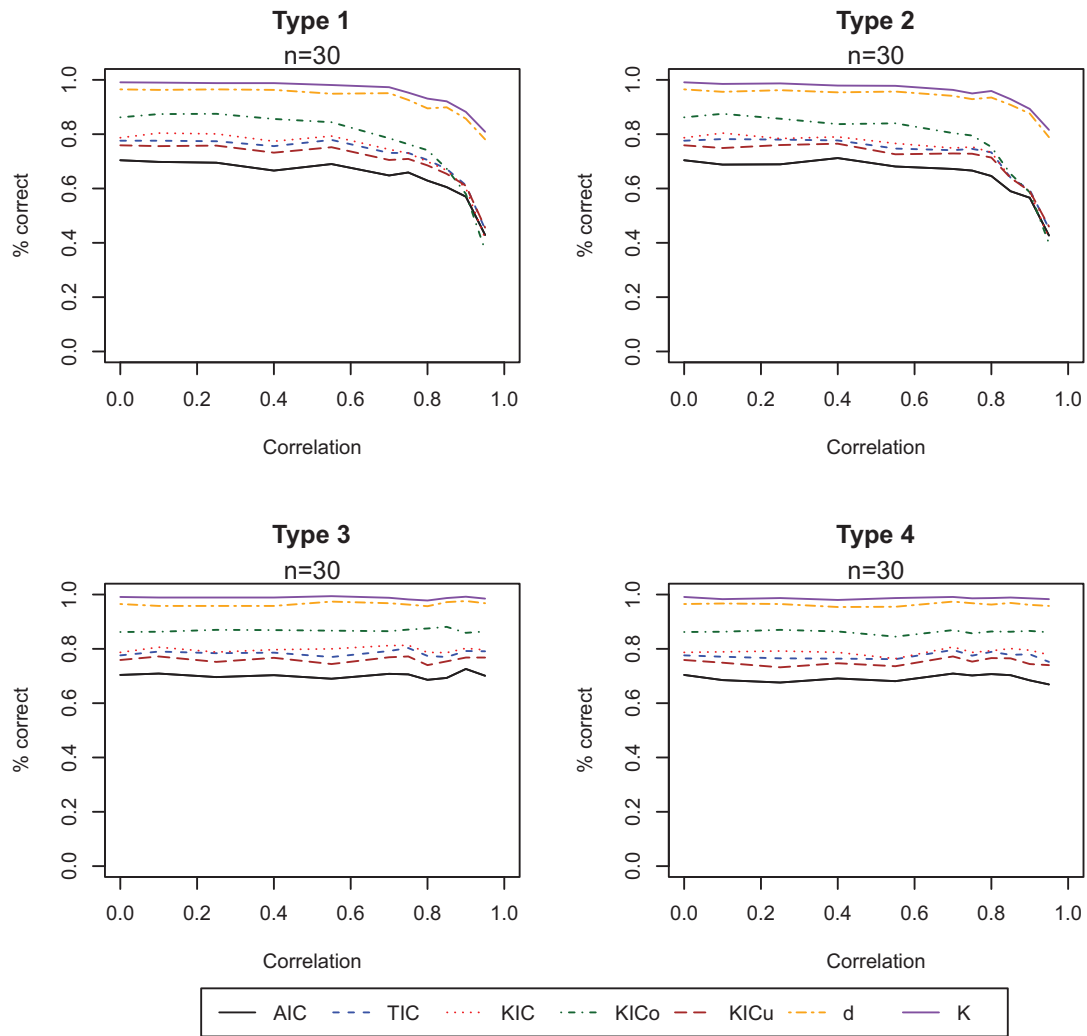


Figure 4.2: Model selection criteria comparison with correlated regressors, $n = 30$, with ignored misspecified errors.

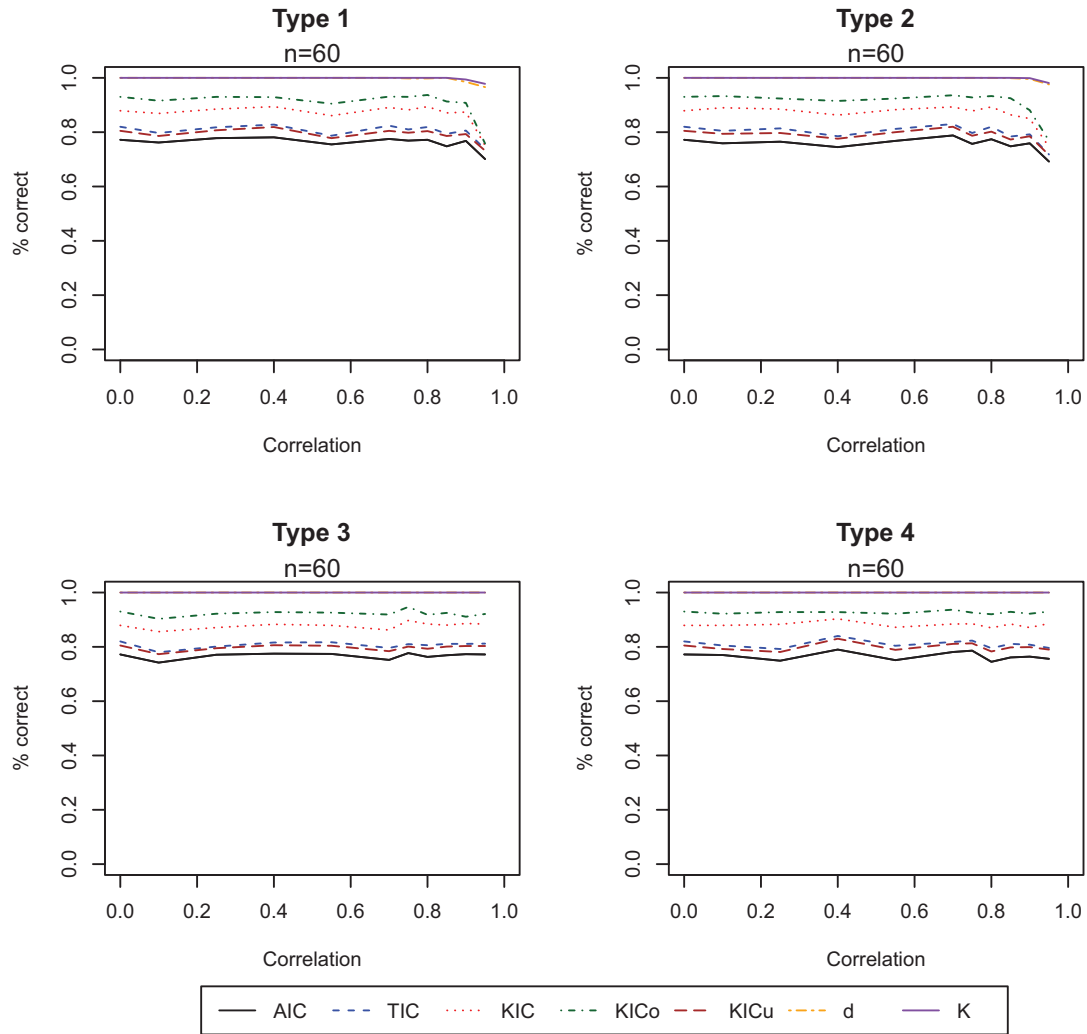


Figure 4.3: Model selection criteria comparison with correlated regressors, $n = 60$, no misspecified errors.

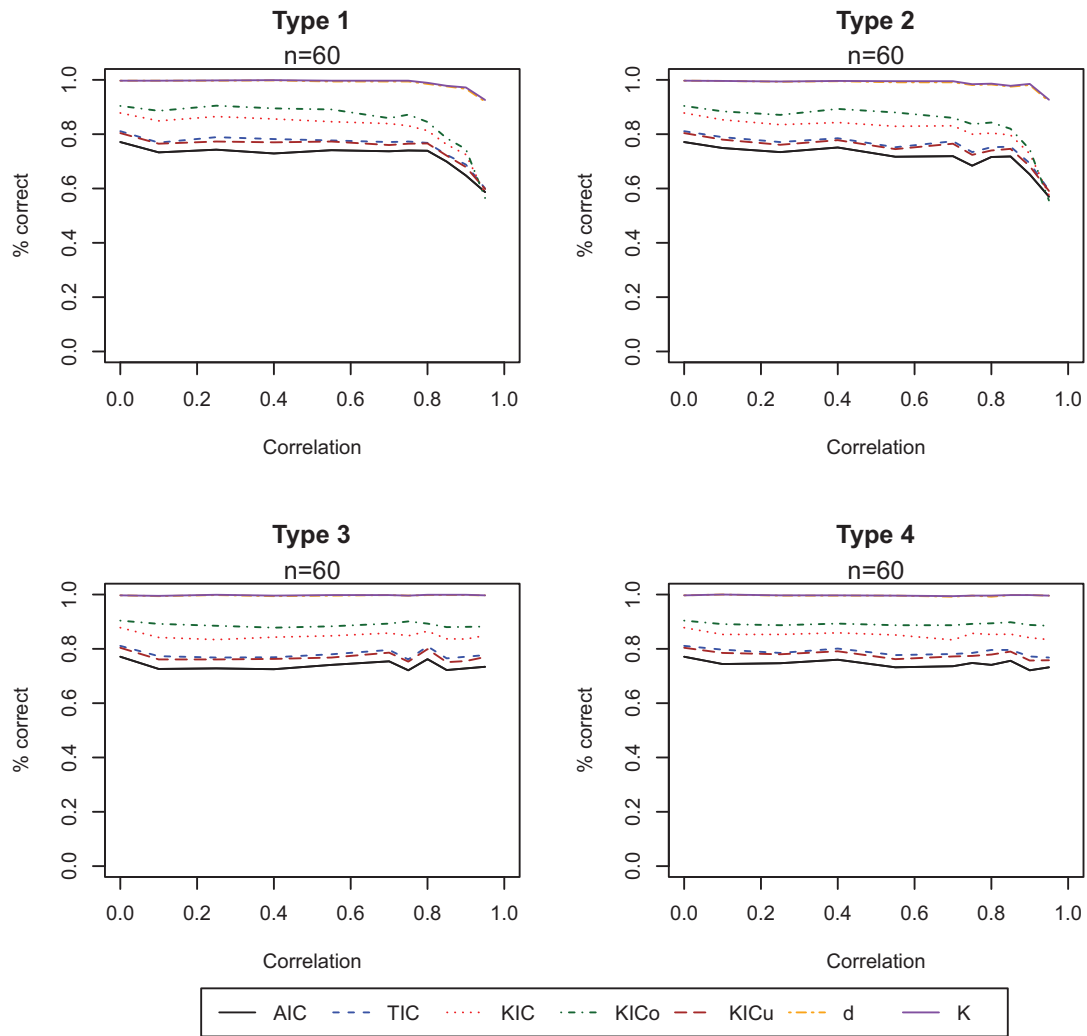


Figure 4.4: Model selection criteria comparison with correlated regressors, $n = 60$, with ignored misspecified errors.

Overall, we found that correlation among regressors affects the performance of the criteria when the correlation between data generating model regressors is high (i.e., $\rho \geq 0.60$). For moderate and low correlations among data generating regressors, the criteria perform in the same manner as when the regressors are independent. Interestingly, as the correlation between regressors increases, KIC_o chooses second-order models more often than the rest of the criteria (data not shown). In this particular setting, choosing models featuring x_{i2} only is a desirable alternative, given that a high collinearity between x_{i2} and x_{i3} indicates that there is little information that is unique to either one of the two covariates. Thus, one can argue that models including either covariate but not both would be a better choice. In general, avoiding highly collinear regressors is good practice in any GLM framework. Even if a high degree of collinearity does not affect the goodness of fit of the model, it interferes with the model interpretation and induces higher variability in the parameter estimates, ultimately affecting inference.

These experiments indicate that the criteria are robust to the presence of low to moderate collinearity. Given the robustness of the criteria to regressor collinearity, we limit our simulation experiment to the linear regression framework and the NM setting. In subsequent simulation experiments, we revert to the use of independent regressors.

4.1.3 Logistic Regression without Overdispersion and with Ignored Overdispersion

For dichotomous outcomes, one often can assume $y_i \sim \text{Bernoulli}(\pi_i)$ for $i = 1, 2, \dots, n$, with $E(y_i) = \pi_i$. If the objective is to explain a sample \mathbf{y} with effects represented by a linear combination of explanatory variables, then the GLM of choice is logistic regression. The canonical link function for logistic regression is the

logit link:

$$\text{logit}(\pi_i) = \ln \left(\frac{\pi_i}{1 - \pi_i} \right) = \sum_{j=1}^p x_{ij} \beta_j,$$

where the β_j s are the parameters we seek to estimate.

For logistic regression, we present a simulation study based on six settings. The latter four of these six settings represent a 2x2 factorial design. Data are generated as $y_i \sim \text{Bernoulli}(\pi_i)$, where π_i is determined using the third-order model

$$\text{logit}(\pi_i) = 1 + x_{i2} - x_{i3}, \quad (4.3)$$

with $i = 1, 2, \dots, n$. Settings 7 ($n = 95$) and 8 ($n = 145$) do not present overdispersion, as π_i is solely based on the systematic component, defined by the x_{ij} s and $\beta = (1, 1, -1)'$.

Settings for experiments 9 to 12 introduce overdispersion in the data generating model. For these settings, the model is the same as (4.3); however, $y_i \sim \text{Bernoulli}(\pi_i^*)$ with $\pi_i^* \sim \text{Beta}(\alpha, \beta = (\alpha - \alpha\pi_i)/\pi_i)$. Note that $E(\pi_i^*) = \alpha/(\alpha + \beta) = \pi_i$; hence, it is straightforward to obtain β by assigning a value to α .

Settings 9 to 12 feature two levels of sample size and two levels of overdispersion: $n = 95$ and $\pi_i^* \sim \text{Beta}(2, (2 - 2\pi_i)/\pi_i)$ (setting 9), $n = 145$ and $\pi_i^* \sim \text{Beta}(2, (2 - 2\pi_i)/\pi_i)$ (setting 10), $n = 95$ and $\pi_i^* \sim \text{Beta}(10, \beta = (10 - 10\pi_i)/\pi_i)$ (setting 11), and $n = 145$ and $\pi_i^* \sim \text{Beta}(10, \beta = (10 - 10\pi_i)/\pi_i)$ (setting 12).

For the simulations featuring NM, the model selection criteria choose from the following candidate set:

$$\begin{aligned} \text{logit}(\pi_i) &= \beta_1, \\ \text{logit}(\pi_i) &= \beta_1 + \beta_2 x_{i2}, \\ &\vdots \\ \text{logit}(\pi_i) &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}, \end{aligned}$$

where $x_{i2} \sim N(0, 0.5)$, $x_{i3} \sim \text{Bernoulli}(0.4)$, $x_{i4} \sim \text{Poisson}(2)$, and $x_{i5} \sim N(5, 5)$. In fitting the models, we assume the data follow a Bernoulli distribution even if the data are generated from an overdispersed Bernoulli distribution (i.e., beta-Bernoulli).

For the APM simulations, the candidate set includes a total of 15 models (i.e., all combinations of the covariates, except for x_{i1} , excluding the intercept-only model).

In the logistic and Poisson frameworks, we include simulation sets where the data are generated using overdispersed models but the candidate models are fitted using a distribution that does not accommodate this overdispersion, so as to simulate a setting that reflects the assumptions under which KIC_o was developed.

The results for the six settings are summarized in Tables 4.7 to 4.12. Again, $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ consistently outperforms $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$, albeit marginally in instances where both oracles obtain a very high rate of correct model selections.

In the NM framework (Tables 4.7 to 4.9) with a moderate sample size (i.e., settings 7, 9, and 11), all criteria choose the correct model barely over 55% of the time. In spite of the superiority of $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ over $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$, it is difficult to advocate the superiority of any of the model selection criteria presented. For settings with a larger sample size (i.e., settings 8, 10 and 12), all criteria perform better than in the moderate sample size settings, with KIC and KIC_o performing best.

Note that KIC_o (together with KIC) chooses overfitted models less often than the rest of the criteria, again demonstrating its advantage in protecting against overfitting. Also, note that KIC_u exhibits less protection against overfitting in the NM order selections, yet together with AIC and TIC, selects underfitted models less frequently and outperforms KIC_o in this respect. These behaviors are consistent with the assumptions under which KIC_o and KIC_u were developed.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
95	5	81	85	31	28	90	0	0
95	4	124	127	62	66	129	7	2
95	3	598	591	573	569	588	939	953
95	2	136	136	196	195	138	42	38
95	1	61	61	138	142	55	12	7
145	5	98	94	27	30	99	0	0
145	4	138	140	78	79	144	3	2
145	3	685	685	738	727	679	985	988
145	2	66	68	115	122	65	10	10
145	1	13	13	42	42	13	2	0

Table 4.7: Settings 7 and 8: NM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the logistic regression framework; results for the generating model are bolded.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
95	5	67	73	25	24	79	0	0
95	4	120	128	67	64	132	6	4
95	3	619	603	563	566	601	941	952
95	2	142	142	206	205	137	40	38
95	1	52	54	139	141	51	13	6
145	5	88	87	24	26	89	0	0
145	4	138	143	74	80	147	4	2
145	3	685	683	726	723	678	983	989
145	2	74	74	134	129	73	11	7
145	1	15	13	42	42	13	2	2

Table 4.8: Settings 9 and 10: NM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the logistic regression framework; results for the generating model are bolded.

Tables 4.10 to 4.12 feature results for the APM simulations. Underfitting is a serious problem for most of the criteria in these settings, especially for AIC and KIC. Only with a larger sample size ($n = 145$) do criteria with data-dependent penalty terms (i.e., TIC, KIC_o and KIC_u) exhibit success at selecting the data generating model the majority of the time. Among the criteria that perform adequately for APM selections, KIC_o chooses overfitted models less often, while KIC_u exhibits better protection against underfitting.

The presence of ignored overdispersion does not appear to appreciably modify the results for any of the NM or APM settings.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
95	5	79	85	27	32	88	0	0
95	4	125	128	68	73	134	6	3
95	3	584	579	556	551	574	949	958
95	2	132	132	191	190	132	36	32
95	1	80	76	158	154	72	9	7
145	5	78	82	28	27	84	0	0
145	4	151	155	87	89	156	1	0
145	3	692	684	728	724	682	990	994
145	2	67	67	118	122	66	9	6
145	1	12	12	39	38	12	0	0

Table 4.9: Settings 11 and 12: NM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the logistic regression framework; results for the generating model are bolded.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
95	overfitted	1	183	0	74	188	12	3
95	correct	10	433	0	394	435	898	917
95	underfitted	989	384	1000	532	377	90	80
145	overfitted	16	258	1	140	262	8	5
145	correct	29	561	1	569	560	973	979
145	underfitted	955	181	998	291	178	19	16

Table 4.10: Settings 7 and 8: APM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the logistic regression framework.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
95	overfitted	2	191	0	90	195	12	6
95	correct	7	421	0	384	423	904	917
95	underfitted	991	388	1000	526	382	84	77
145	overfitted	11	247	0	114	250	9	5
145	correct	25	564	2	585	561	970	981
145	underfitted	964	189	998	301	189	21	14

Table 4.11: Settings 9 and 10: APM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the logistic regression framework.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
95	overfitted	0	197	0	87	202	13	6
95	correct	7	411	0	357	408	912	927
95	underfitted	993	392	1000	556	390	75	67
145	overfitted	12	270	0	133	273	5	2
145	correct	31	562	1	596	563	975	983
145	underfitted	957	168	999	271	164	20	15

Table 4.12: Settings 11 and 12: APM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the logistic regression framework.

4.1.4 Poisson Regression without Overdispersion and with Ignored Overdispersion

When the outcome of interest is a discrete count variable, one can often assume $y_i \sim \text{Poisson}(\lambda_i)$ for $i = 1, 2, \dots, n$, with $E(y_i) = \lambda_i$. If the objective is to explain a sample \mathbf{y} with effects represented by a linear combination of explanatory variables, then the GLM of choice is Poisson regression. The canonical link function for Poisson regression is the *log link*:

$$\log(\lambda_i) = \sum_{j=1}^p x_{ij}\beta_j,$$

where the β_j s are the parameters we seek to estimate.

For Poisson regression, we present a 2x3 factorial simulation study. Data are generated as $y_i \sim \text{Poisson}(\lambda_i)$, where λ_i is determined using the third-order model

$$\log(\lambda_i) = 1 + x_{i2} + x_{i3} \quad (4.4)$$

with $i = 1, 2, \dots, n$. Settings 13 ($n = 60$) and 14 ($n = 120$) do not present overdispersion, as λ_i is solely based on the systematic component, defined by the x_{ij} s and $\beta = (1, 1, 1)'$.

Settings for experiments 15 to 18 introduce overdispersion in the data generating model. For these settings, the model is the same as (4.4); however, $y_i \sim \text{Poisson}(\lambda_i^*)$ with $\lambda_i^* \sim \text{Gamma}(\alpha = \lambda_i/\beta, \beta)$. Note that $E(\lambda_i^*) = \alpha\beta = \lambda_i$; hence, it is straightforward to obtain α by assigning a value to β .

Settings 15 to 18 feature two levels of sample size and two levels of overdispersion: $n = 60$ and $\lambda_i^* \sim \text{Gamma}(\lambda_i/50, 50)$ (setting 15), $n = 120$ and $\lambda_i^* \sim \text{Gamma}(\lambda_i/50, 50)$ (setting 16), $n = 60$ and $\lambda_i^* \sim \text{Gamma}(\lambda_i/100, 100)$ (setting 17), $n = 120$ and $\lambda_i^* \sim \text{Gamma}(\lambda_i/100, 100)$ (setting 18).

For the simulations featuring NM, the model selection criteria choose from the following candidate set:

$$\begin{aligned} \log(\lambda_i) &= \beta_1, \\ \log(\lambda_i) &= \beta_1 + \beta_2 x_{i2}, \\ &\vdots \\ \log(\lambda_i) &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}, \end{aligned}$$

where $x_{i2} \sim N(0, 0.5)$, $x_{i3} \sim \text{Binomial}(2, 0.5)$, $x_{i4} \sim \text{Poisson}(10)$, $x_{i5} \sim N(5, 1)$, and $x_{i6} \sim \text{Binomial}(3, 0.5)$. In fitting the models, we assume the data follow a Poisson distribution even if the data are generated from an overdispersed Poisson distribution (i.e., gamma-Poisson).

For the APM simulations, the candidate set includes a total of 31 models (i.e., all combinations of the covariates, except for x_{i1} , excluding the intercept-only model).

The results for the six settings are summarized in Tables 4.13 to 4.18. Once more, $K(\theta_0, \hat{\theta})$ performs at least as well as $d(\theta_0, \hat{\theta})$ in all settings.

For the NM simulations in settings 13 and 14, Table 4.13 shows that all criteria perform well. KIC is the criterion that chooses the data generating model most often. This table illustrates a phenomenon that is not evident in previous simulation sets: as sample size increases, the performance of AIC and KIC appears to degrade. This is seemingly at odds with what is expected for asymptotic statistics. However, this behavior may be reflecting that all criteria used in these studies are *asymptotically efficient* but are not *consistent*. With an asymptotically efficient criterion, as sample size increases, the criterion chooses models prone to minimizing prediction error. With a consistent criterion, as sample size increases, the probability of choosing the data generating model converges to one. For an asymptotically efficient criterion, the asymptotic probability of selecting an overfitted model is always positive. Thus, in settings 13 and 14, the performance of AIC and KIC might be at a plateau for choosing the correct model. Table 4.13 shows, once more, the enhanced overfitting protection of KIC_o when compared to KIC_u .

The introduction of ignored overdispersion (Tables 4.14 and 4.15) affects the performance of all criteria, in particular that of AIC and KIC. In these settings, the enhanced protection of KIC_o against overfitting is a clear advantage, and the criterion exhibits the highest rate of correct model selections. The different protection patterns of KIC_o and KIC_u are again noticeable in these tables. KIC_u chooses underfitted models less often than KIC_o while KIC_o selects overfitted models less often than KIC_u .

In the APM settings (Tables 4.16 to 4.18), the criteria behave in the same manner as in the NM settings. It is evident that the rate of correct model selections markedly degrades when overdispersion is present in the generating model but is ignored in the candidate set.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
60	6	34	113	9	56	113	0	0
60	5	60	122	17	72	121	0	0
60	4	123	157	67	130	157	8	5
60	3	783	608	907	742	609	992	995
60	2	0	0	0	0	0	0	0
60	1	0	0	0	0	0	0	0
120	6	52	88	16	34	87	0	0
120	5	59	93	21	49	93	0	0
120	4	135	154	87	108	153	0	0
120	3	754	665	876	809	667	1000	1000
120	2	0	0	0	0	0	0	0
120	1	0	0	0	0	0	0	0

Table 4.13: Settings 13 and 14: NM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the Poisson regression framework; results for the generating model are bolded.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
60	6	818	203	772	125	205	0	0
60	5	141	166	162	118	166	2	0
60	4	28	152	43	137	155	35	14
60	3	13	398	23	471	396	909	943
60	2	0	40	0	65	42	28	19
60	1	0	41	0	84	36	26	24
120	6	838	140	804	74	140	0	0
120	5	120	97	141	58	97	2	1
120	4	28	144	33	109	145	13	4
120	3	14	582	22	688	581	977	989
120	2	0	31	0	48	31	6	4
120	1	0	6	0	23	6	2	2

Table 4.14: Settings 15 and 16: NM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the Poisson regression framework; results for the generating model are bolded.

Overall, the comparison of the criteria in the Poisson regression framework confirms the selection patterns of KIC_o and KIC_u exhibited for the linear and logistic regression settings. However, in the Poisson setting, the effect of the presence of ignored misspecification in the form of overdispersion is more evident.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
60	6	849	253	819	157	264	0	0
60	5	133	139	148	102	143	1	0
60	4	11	141	21	120	136	27	19
60	3	7	298	12	318	299	821	855
60	2	0	75	0	109	78	80	64
60	1	0	94	0	194	80	71	62
120	6	856	148	831	95	151	0	0
120	5	125	108	141	73	108	0	0
120	4	17	147	23	117	150	22	11
120	3	2	488	5	528	489	942	959
120	2	0	63	0	89	59	21	15
120	1	0	46	0	98	43	15	15

Table 4.15: Settings 17 and 18: NM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the Poisson regression framework; results for the generating model are bolded.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
60	overfitted	183	547	90	414	548	15	9
60	correct	817	453	905	586	452	985	991
60	underfitted	0	0	5	0	0	0	0
120	overfitted	204	501	91	344	501	9	4
120	correct	796	499	909	656	499	991	996
120	underfitted	0	0	0	0	0	0	0

Table 4.16: Settings 13 and 14: APM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the Poisson regression framework.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
60	overfitted	784	471	696	351	477	64	29
60	correct	13	207	23	250	205	812	859
60	underfitted	203	322	281	399	318	124	112
120	overfitted	917	452	875	314	455	34	17
120	correct	17	387	31	461	387	932	951
120	underfitted	66	161	94	225	158	34	32

Table 4.17: Settings 15 and 16: APM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the Poisson regression framework.

n	order	AIC	TIC	KIC	KIC _o	KIC _u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
60	overfitted	682	381	609	265	379	51	32
60	correct	7	127	11	134	128	671	709
60	underfitted	311	492	380	601	493	278	259
120	overfitted	874	407	837	288	412	50	25
120	correct	4	267	9	296	266	855	888
120	underfitted	122	326	154	416	322	95	87

Table 4.18: Settings 17 and 18: APM order selections for AIC, TIC, KIC, KIC_o, KIC_u, $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ in the Poisson regression framework; results for the generating model are bolded.

4.2 Selection of a Model with Optimal Predictive Properties

Simulating the problem of choosing a model with optimal predictive properties requires modifications to the simulation settings presented in section 4.1. In this situation, the data generating model is the highest order model. Hence, all candidate models except for the model of the highest order model are underspecified. In an application where the data generating model contains strong, moderate, and weak effects, the researcher might not be able to accurately estimate all the effects given the sample size at hand, yet might search for a good model to predict new outcomes. Such a model would invariably dismiss some of the weaker effects, based on the notion that the bias incurred by omitting such effects is less damaging than the variability induced by including them.

For simulating this problem, all candidate models are fitted using a training data set produced via the generating model. All the fitted models are then evaluated

using some measure of prediction error based on a new data set: i.e., a testing data set. We present simulation studies in the linear regression framework, and consider the *mean squared error of prediction*,

$$\text{MSEP} = \frac{\sum_{i=n+1}^{n+n^*} (y_i - \hat{y}_i)^2}{n^*},$$

as a suitable prediction error measure for this framework. In the MSEP definition, n^* represents the sample size of the testing data set, and \hat{y}_i denotes the predicted value under the fitted model for each observation. In these experiments, the criteria that choose models which yield a small average MSEP are considered optimal.

Each simulation set is based on 1,000 replications. For every sample, AIC, TIC, KIC, KIC_o , KIC_u , $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$, and $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ are evaluated for each fitted candidate model. For each criterion, the MSEP value is recorded for the selected model; the average MSEP is then computed over the 1,000 replications. For this simulation study, we present results in the APM setting.

4.2.1 Linear Regression

We present a 2x2 factorial simulation study based on two levels of sample size and two levels of error variance. Both the training and testing data sets are generated using the sixth-order model

$$y_i = 5 + 2.5x_{i2} + x_{i3} + 0.75x_{i4} + 0.5x_{i5} + 0.25x_{i6} + \epsilon_i,$$

where x_{ij} ($j = 2, \dots, 6$) are iid $N(0,10)$. Following the notation used in chapter 3, $K = 7$ (after counting one parameter for σ^2). All training sets have a sample size of $n = 30$. We consider four settings: $\epsilon_i \sim N(0,10)$ and $n^* = 15$ (setting 19) or $n^* = 30$ (setting 20), and $\epsilon_i \sim N(0,30)$ and $n^* = 15$ (setting 21) or $n^* = 30$ (setting 22). Note that the model is comprised of effects of different strengths, corresponding to the magnitudes of the components of $\boldsymbol{\beta}$.

The candidate model set includes 31 models (i.e., all combinations of the covariates, except for x_{i1} , excluding the intercept-only model).

n^*	$\text{Var}(\epsilon_i)$	AIC	TIC	KIC	KIC_o	KIC_u	$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$	$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$
15	10	102.56	29.17	109.10	32.92	25.21	20.17	17.90
30	10	26.55	16.79	82.59	20.32	15.52	13.08	12.79
15	30	134.77	77.27	132.85	83.33	70.90	49.94	46.80
30	30	96.21	51.08	117.51	55.79	48.69	37.59	37.15

Table 4.19: Settings 19 to 22: Average MSEP for models selected by AIC, TIC, KIC, KIC_o , KIC_u and their oracles.

Based on the results in Table 4.19, $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ is the oracle that tends to choose models with the smallest MSEP. Among the criteria, KIC_u clearly performs best in this scenario. This is as expected, since this simulation study reflects the assumptions under which KIC_u was developed. TIC also performs well, followed by KIC_o . As the testing sample size increases, the average MSEP is lower. Also, a higher error variance increases the average prediction error.

4.3 General Conclusions

Probably the most common question posed by practitioners to researchers in the field of model selection criteria is “What criterion should I use?” Tailored to this chapter, an analogous question would be “What criterion do you recommend, KIC_o or KIC_u ?” Unfortunately, as is evident from the results of the featured simulation studies, the answer is not straightforward. The use of KIC_o or KIC_u depends on what modeling objective is pursued, and whether the practitioner anticipates a modeling setting where overfitting or underfitting prevails.

The simulation results show that there is no optimal criterion for all modeling situations. There are instances in the preceding experiments where any of the considered criteria succeed at selecting the best model. Yet there are also instances where none of the criteria perform well. Among model selection criteria, the perfect balance between underfitting and overfitting is elusive, and we know of no criterion that performs uniformly best throughout all possible settings. At most, one criterion can be tailored to succeed in a particular framework.

Our general advice is to base model selection decisions on scientific guidance, to the greatest extent that the science will allow. In principle, this should constrain the candidate model set to the smallest feasible size. If at this point, the modeling objective is to predict new data, KIC_u is particularly tailored to this setting and performs well in selecting a model with optimal predictive properties (section 4.2.1). Based on the form of the candidate set, if the practitioner expects underspecification to be a problem, then KIC_u can be also be recommended. In applications where overfitting is expected, or where model misspecification is suspected, KIC_o should be the criterion of choice.

CHAPTER 5

QUASI-LIKELIHOOD-BASED MODEL SELECTION CRITERIA FOR CORRELATED RESPONSE DATA DERIVED FROM KULLBACK'S SYMMETRIC DIVERGENCE

In this chapter we present three forms of QKIC, a quasi-likelihood-based model selection criterion for correlated response data. QKIC is developed using a variant of KSD similar to that used for deriving KIC_o . We also characterize the performance of QKIC in simulation studies, and illustrate the use of QKIC with a real-world data example.

5.1 QKIC Derivation

Let β_0 denote the parameter in the data generating model $g(\mathbf{y}|\beta_0)$, and let $\hat{\beta}$ represent its estimator. For deriving the different forms of QKIC, we employ a similar set of assumptions to that used for KIC_o . As mentioned in chapter 2, in the GLM framework with correlated data, we do not always have access to the distribution of the response. In such settings, we often use the quasi-likelihood or GEEs for estimation, which only require the specification of the first and second moments of the response. Thus, for the development of the different variants of QKIC, we cannot assume that $g(\mathbf{y}|\beta_0) \in \mathcal{F}(k)$. In this setting, misspecification is present in the distribution of the response. However, we assume that the mean structure is correctly specified or overspecified. This assumption is also imposed in the theoretical development of quasi-likelihood and GEE estimation.

For deriving QKIC, we assume a set of regularity conditions required to ensure that $\hat{\beta}$ satisfies the large-sample properties described in section 2.1.3. In the quasi-likelihood and GEE frameworks, regardless of the working correlation used to estimate $\hat{\beta}$, $\hat{\beta}$ is consistent for β_0 : $\hat{\beta} \rightarrow \beta_0$ as $n \rightarrow \infty$.

We denote $\hat{\beta}$ obtained under the working independence model as $\hat{\beta}^I$, and $\hat{\beta}$ obtained using the working correlation matrix $R(\alpha)$ as $\hat{\beta}^R$.

For the purpose of the derivation that follows, we implicitly assume the working independence model. We define the following information matrices:

$$\mathcal{I}_q(\boldsymbol{\beta}) = E_q\{\mathcal{I}(\boldsymbol{\beta}|\mathbf{y})\} \text{ and}$$

$$\mathcal{I}_g(\boldsymbol{\beta}) = E_g\{\mathcal{I}(\boldsymbol{\beta}|\mathbf{y})\}.$$

Here, $E_q\{\cdot\}$ denotes the expected value under the quasi-likelihood model $Q(\boldsymbol{\beta}|\mathbf{y})$ (i.e., the working independence model). For our purposes, this expected value is only applied to obtain expressions involving the first and second moments of the response, which are well defined under the postulated model.

Consider the discrepancy

$$d_{qq}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = E_q\{-2Q(\boldsymbol{\beta}^*|\mathbf{y})\},$$

where $\boldsymbol{\beta}$ refers to the parameters of the quasi-likelihood model $Q(\boldsymbol{\beta}|\mathbf{y})$ under which the expectation $E_q\{\cdot\}$ is calculated, and where $\boldsymbol{\beta}^*$ refers to the term $-2Q(\boldsymbol{\beta}^*|\mathbf{y})$ in the expectation. We define the following variant of KSD based on this measure:

$$2J_{qq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = d_{qq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) - d_{qq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0) + d_{qq}(\boldsymbol{\beta}, \boldsymbol{\beta}_0) - d_{qq}(\boldsymbol{\beta}, \boldsymbol{\beta}). \quad (5.1)$$

The expression (5.1) is similar to the expression (3.1) used in the development of KIC_o . Here, $Q(\boldsymbol{\beta}|\mathbf{y})$ and $Q(\boldsymbol{\beta}_0|\mathbf{y})$ function as replacements for $\ln f(\mathbf{y}|\boldsymbol{\beta})$ and $\ln f(\mathbf{y}|\boldsymbol{\beta}_0)$, respectively. When the quasi-likelihood is viewed as the equivalent of a log-likelihood in the exponential family of distributions, the Kullback discrepancies in this section are of the traditional form. Establishing the general properties of Kullback discrepancies for quasi-likelihoods is a potentially worthwhile problem, yet one that is outside the scope of this dissertation.

Now consider the discrepancy

$$d_{gq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = E_g\{-2Q(\boldsymbol{\beta}|\mathbf{y})\},$$

where the first argument refers to the parameters of the data generating model

$g(\mathbf{y}|\boldsymbol{\beta}_0)$ under which the expectation $E_g\{\cdot\}$ is calculated, and where the second argument refers to the term $-2Q(\boldsymbol{\beta}|\mathbf{y})$ in the expectation. We add the constant $d_{gq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0)$ to (5.1), which does not alter the ranking of competing models. With this change, we have

$$L_{gq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = d_{qq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) - d_{qq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0) + d_{qq}(\boldsymbol{\beta}, \boldsymbol{\beta}_0) - d_{qq}(\boldsymbol{\beta}, \boldsymbol{\beta}) + d_{gq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0). \quad (5.2)$$

The preceding is an appealing measure for creating KSD-based model selection criteria when a quasi-likelihood or GEEs are used for parameter estimation. This measure will serve as the basis for the development of QKIC.

If we evaluate (5.2) at $\hat{\boldsymbol{\beta}}^I$ and take the expected value with respect to $g(\mathbf{y}|\boldsymbol{\beta}_0)$, we obtain

$$\begin{aligned} \Psi_{gq}(\boldsymbol{\beta}_0) &\equiv E_g\{L_{gq}(\boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}}^I)\} \\ &= E_g\{d_{qq}(\boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}}^I) - d_{qq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0) + d_{qq}(\hat{\boldsymbol{\beta}}^I, \boldsymbol{\beta}_0) - d_{qq}(\hat{\boldsymbol{\beta}}^I, \hat{\boldsymbol{\beta}}^I) + d_{gq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0)\}. \end{aligned} \quad (5.3)$$

We may then decompose $\Psi_{gq}(\boldsymbol{\beta}_0)$ as follows:

$$\begin{aligned} \Psi_{gq}(\boldsymbol{\beta}_0) &= E_g\{-2Q(\hat{\boldsymbol{\beta}}^I|\mathbf{y})\} \\ &\quad + d_{gq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0) - E_g\{-2Q(\hat{\boldsymbol{\beta}}^I|\mathbf{y})\} \end{aligned} \quad (5.4)$$

$$+ E_g\{d_{qq}(\boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}}^I)\} - d_{qq}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0) \quad (5.5)$$

$$+ E_g\{d_{qq}(\hat{\boldsymbol{\beta}}^I, \boldsymbol{\beta}_0) - d_{qq}(\hat{\boldsymbol{\beta}}^I, \hat{\boldsymbol{\beta}}^I)\}. \quad (5.6)$$

The purpose of the preceding decomposition is to introduce $-2Q(\hat{\boldsymbol{\beta}}^I|\mathbf{y})$ as a platform for estimating $\Psi_{gq}(\boldsymbol{\beta}_0)$. Since (5.4), (5.5) and (5.6) are positive, $-2Q(\hat{\boldsymbol{\beta}}^I|\mathbf{y})$ is negatively biased. If we can obtain estimates of these terms, we can correct for the negative bias.

Consider taking a second-order Taylor series expansion of $-2Q(\boldsymbol{\beta}_0|\mathbf{y})$ about

$\hat{\beta}^I$. Since $-2Q(\beta_0|\mathbf{y})$ is maximized at $\hat{\beta}^I$, one can establish

$$-2Q(\beta_0|\mathbf{y}) = -2Q(\hat{\beta}^I|\mathbf{y}) + (\hat{\beta}^I - \beta_0)' \mathcal{I}(\hat{\beta}^I|\mathbf{y})(\hat{\beta}^I - \beta_0) + r_1(\beta_0, \hat{\beta}^I).$$

Here, as n increases, $r_1(\beta_0, \hat{\beta}^I)$ is $o_p(1)$; $E_g\{r_1(\beta_0, \hat{\beta}^I)\}$ is therefore $o(1)$. Taking the expectation of both sides of this expansion with respect to $g(\mathbf{y}|\beta_0)$ yields

$$E_g\{-2Q(\beta_0|\mathbf{y})\} - E_g\{-2Q(\hat{\beta}^I|\mathbf{y})\} = E_g\{(\hat{\beta}^I - \beta_0)' \mathcal{I}(\hat{\beta}^I|\mathbf{y})(\hat{\beta}^I - \beta_0)\} + o(1). \quad (5.7)$$

Next, consider taking second-order expansions in the second argument of $d_{qq}(\beta_0, \hat{\beta}^I)$ about β_0 and in the second argument of $d_{qq}(\hat{\beta}^I, \beta_0)$ about $\hat{\beta}^I$. By definition, $d_{qq}(\beta_0, \hat{\beta}^I)$ is minimized when $\hat{\beta}^I = \beta_0$ and $d_{qq}(\hat{\beta}^I, \beta_0)$ is minimized when $\beta_0 = \hat{\beta}^I$. We have

$$d_{qq}(\beta_0, \hat{\beta}^I) = d_{qq}(\beta_0, \beta_0) + (\hat{\beta}^I - \beta_0)' \mathcal{I}_q(\beta_0)(\hat{\beta}^I - \beta_0) + r_2(\beta_0, \hat{\beta}^I) \text{ and} \quad (5.8)$$

$$d_{qq}(\hat{\beta}^I, \beta_0) = d_{qq}(\hat{\beta}^I, \hat{\beta}^I) + (\hat{\beta}^I - \beta_0)' \mathcal{I}_q(\hat{\beta}^I)(\hat{\beta}^I - \beta_0) + r_3(\beta_0, \hat{\beta}^I). \quad (5.9)$$

Here, $r_2(\beta_0, \hat{\beta}^I)$ and $r_3(\beta_0, \hat{\beta}^I)$ are both $o_p(1)$; $E_g\{r_2(\beta_0, \hat{\beta}^I)\}$ and $E_g\{r_3(\beta_0, \hat{\beta}^I)\}$ are therefore $o(1)$. Taking the expectation of both sides of (5.8) and (5.9) with respect to $g(\mathbf{y}|\beta_0)$ yields

$$E_g\{d_{qq}(\beta_0, \hat{\beta}^I) - d_{qq}(\beta_0, \beta_0)\} = E_g\{(\hat{\beta}^I - \beta_0)' \mathcal{I}_q(\beta_0)(\hat{\beta}^I - \beta_0)\} + o(1) \text{ and} \quad (5.10)$$

$$E_g\{d_{qq}(\hat{\beta}^I, \beta_0) - d_{qq}(\hat{\beta}^I, \hat{\beta}^I)\} = E_g\{(\hat{\beta}^I - \beta_0)' \mathcal{I}_q(\hat{\beta}^I)(\hat{\beta}^I - \beta_0)\} + o(1). \quad (5.11)$$

Thus, replacing (5.4) by (5.7), (5.5) by (5.10) and (5.6) by (5.11), $\Psi_{gq}(\beta_0)$ can be represented as follows:

$$\begin{aligned}
\Psi_{gq}(\beta_0) &= E_g\{-2Q(\hat{\beta}^I|\mathbf{y})\} \\
&+ E_g\{(\hat{\beta}^I - \beta_0)'\mathcal{I}(\hat{\beta}^I|\mathbf{y})(\hat{\beta}^I - \beta_0)\} \\
&+ E_g\{(\hat{\beta}^I - \beta_0)'\mathcal{I}_q(\beta_0)(\hat{\beta}^I - \beta_0)\} \\
&+ E_g\{(\hat{\beta}^I - \beta_0)'\mathcal{I}_q(\hat{\beta}^I)(\hat{\beta}^I - \beta_0)\} + o(1).
\end{aligned}$$

Since $\hat{\beta}^I \rightarrow \beta_0$ as $n \rightarrow \infty$, we can approximate the first quadratic form with $E_g\{(\hat{\beta}^I - \beta_0)'\mathcal{I}_g(\beta_0)(\hat{\beta}^I - \beta_0)\}$. Also, we can approximate the third quadratic form with $E_g\{(\hat{\beta}^I - \beta_0)'\mathcal{I}_q(\beta_0)(\hat{\beta}^I - \beta_0)\}$. Hence, $\Psi_{gq}(\beta_0)$ becomes

$$\begin{aligned}
\Psi_{gq}(\beta_0) &= E_g\{-2Q(\hat{\beta}^I|\mathbf{y})\} \\
&+ E_g\{(\hat{\beta}^I - \beta_0)'\mathcal{I}_g(\beta_0)(\hat{\beta}^I - \beta_0)\} \\
&+ 2E_g\{(\hat{\beta}^I - \beta_0)'\mathcal{I}_q(\beta_0)(\hat{\beta}^I - \beta_0)\} + o(1).
\end{aligned}$$

Given that $E_g\{(\hat{\beta}^I - \beta_0)'\mathcal{I}_g(\beta_0)(\hat{\beta}^I - \beta_0)\}$ and $E_g\{(\hat{\beta}^I - \beta_0)'\mathcal{I}_q(\beta_0)(\hat{\beta}^I - \beta_0)\}$ are scalars, we have

$$\begin{aligned}
\Psi_{gq}(\beta_0) &= E_g\{-2Q(\hat{\beta}^I|\mathbf{y})\} + \text{tr}\{E_g\{\mathcal{I}_g(\beta_0)(\hat{\beta}^I - \beta_0)(\hat{\beta}^I - \beta_0)'\}\} \\
&+ 2[\text{tr}\{E_g\{\mathcal{I}_q(\beta_0)(\hat{\beta}^I - \beta_0)(\hat{\beta}^I - \beta_0)'\}\}] + o(1) \\
&= E_g\{-2Q(\hat{\beta}^I|\mathbf{y})\} + \text{tr}\{\mathcal{I}_g(\beta_0)\Sigma(\beta_0)\} \\
&+ 2[\text{tr}\{\mathcal{I}_q(\beta_0)\Sigma(\beta_0)\}] + o(1).
\end{aligned}$$

Let $\mathcal{J}_g(\beta_0) = E_g(\mathcal{J}(\beta_0|\mathbf{y}))$. Replacing $\Sigma(\beta_0)$ by result (2.12), we obtain

$$\begin{aligned}
\Psi_{gq}(\beta_0) &= E_g\{-2Q(\hat{\beta}^I|\mathbf{y})\} + \text{tr}\{\mathcal{J}_g(\beta_0)\mathcal{I}_g(\beta_0)^{-1}\} \\
&+ 2[\text{tr}\{\mathcal{I}_q(\beta_0)\mathcal{I}_g(\beta_0)^{-1}\mathcal{J}_g(\beta_0)\mathcal{I}_g(\beta_0)^{-1}\}] + o(1).
\end{aligned}$$

Now, define the statistic QKIC^I as follows:

$$\begin{aligned}\text{QKIC}^I &= -2Q(\hat{\beta}^I|\mathbf{y}) + \text{tr}\{\mathcal{J}(\hat{\beta}^I|\mathbf{y})\mathcal{I}(\hat{\beta}^I|\mathbf{y})^{-1}\} \\ &\quad + 2[\text{tr}\{\mathcal{I}_q(\hat{\beta}^I)\mathcal{I}(\hat{\beta}^I|\mathbf{y})^{-1}\mathcal{J}(\hat{\beta}^I|\mathbf{y})\mathcal{I}(\hat{\beta}^I|\mathbf{y})^{-1}\}].\end{aligned}\quad (5.12)$$

Based on the preceding development, one can conclude that

$$E_g\{\text{QKIC}^I\} + o(1) = \Psi_{gq}(\beta_0).$$

In the GLM framework, when the canonical link is used, the expected information equals the observed information because the Hessian matrix based on $Q(\beta|\mathbf{y})$ does not depend on the data. That is, $\mathcal{I}_q(\beta_0) = \mathcal{I}_g(\beta_0) = \mathcal{I}(\beta|\mathbf{y})$. Hence, expression (5.12) becomes

$$\begin{aligned}\text{QKIC}^I &= -2Q(\hat{\beta}^I|\mathbf{y}) + \text{tr}\{\mathcal{J}(\hat{\beta}^I|\mathbf{y})\mathcal{I}(\hat{\beta}^I|\mathbf{y})^{-1}\} \\ &\quad + 2[\text{tr}\{\mathcal{J}(\hat{\beta}^I|\mathbf{y})\mathcal{I}(\hat{\beta}^I|\mathbf{y})^{-1}\}] \\ &= -2Q(\hat{\beta}^I|\mathbf{y}) + 3[\text{tr}\{\mathcal{J}(\hat{\beta}^I|\mathbf{y})\mathcal{I}(\hat{\beta}^I|\mathbf{y})^{-1}\}].\end{aligned}\quad (5.13)$$

In summary, QKIC^I is an asymptotically unbiased estimator of $\Psi_{gq}(\beta_0)$ for models with correctly specified or overspecified mean structures.

If $\hat{\beta}^I$ is replaced by $\hat{\beta}^R$ in (5.13), the following variant of QKIC arises:

$$\text{QKIC}^R = -2Q(\hat{\beta}^R|\mathbf{y}) + 3[\text{tr}\{\mathcal{J}(\hat{\beta}^R|\mathbf{y})\mathcal{I}(\hat{\beta}^R|\mathbf{y})^{-1}\}].\quad (5.14)$$

This replacement is not based on solid theoretical principles. In fact, QKIC^R is not an asymptotically unbiased estimator of $\Psi_{gq}(\beta_0)$, because the first-order terms in the Taylor series expansions used in the derivation of QKIC^I are not zero and do not converge to zero as n increases. We present this criterion merely as a parallel to the criterion QIC^R , introduced by Pan (2001) without theoretical justification. Both QKIC^R and QIC^R warrant a rigorous theoretical development; however, such developments are outside of the scope of this dissertation.

Using KSD as the target oracle, another criterion is suggested based on the work of Pan (2001). When all modeling specifications are correct, $\mathcal{J}(\hat{\beta}^R|\mathbf{y})$ and $\mathcal{I}(\hat{\beta}^R|\mathbf{y})$ are asymptotically equivalent. Thus, $\text{tr}\{\mathcal{J}(\hat{\beta}^R|\mathbf{y})\mathcal{I}(\hat{\beta}^R|\mathbf{y})^{-1}\} \approx k$. In that case, QKIC^R reduces to a form that is similar to that of KIC:

$$\text{QKIC}^U = -2Q(\hat{\beta}^R|\mathbf{y}) + 3k.$$

QKIC^U is an approximation of (5.14) and potentially useful in mean structure selection. However, QKIC^U cannot be applied to select the working correlation structure, because the postulated correlation structure is not sufficiently represented in either the goodness-of-fit term or the penalty term.

5.2 Simulation Studies for QKIC

The assumptions employed in the development of QKIC allow us to apply these criteria for selecting models grouped under the umbrella of GLMs for correlated responses, provided the parameters are estimated using a quasi-likelihood or GEEs. The goal of the simulation studies presented in this chapter is to characterize the performance of QKIC. We compare the performance of QKIC and QIC in these experiments.

We first illustrate the selection of the working correlation structure and then the selection of covariates for the mean structure. For the latter, we consider correlated binary and count response data, using experiments with different sample sizes, different generating models, and different candidate model sets. Each simulation experiment presented is based on 1,000 replications.

5.2.1 Selection of Working Correlation Structure for Correlated Binary Response Data

We begin our simulation studies by evaluating the performance of QKIC in the setting of simulation experiments that are already published. Specifically, we illustrate the selection of a working correlation structure for correlated binary response data by reproducing the simulation study presented in Pan (2001).

Correlated binary data are generated using the methods suggested by Leisch et al. (1998). In these experiments, any two measurements in the same cluster have the same correlation ρ ; that is, repeated measures within clusters follow an exchangeable correlation structure. For each of the 1,000 replications, QIC^R and QKIC^R are evaluated for each candidate model. The fitted model with the minimum value for each statistic is recorded. The results are then summarized in Table 5.1.

Data for settings 23 and 24 are generated using the following generating model:

$$\text{logit}(\pi_{it}) = 0.25 - 0.25x_{i2t} - 0.25t,$$

where $x_{i2t} \sim \text{iid Bernoulli}(0.5)$, $t = 0, 1, 2$ (i.e., three measurements per cluster), $\rho = 0.5$ (exchangeable correlation structure), $n = 50$ (setting 23) and $n = 100$ (setting 24). For these two settings, following the simulation experiment in Pan (2001), we limit the candidate model set to the data generating model and two other candidate models: one assuming the responses are independent and the other assuming an AR-1 correlation structure.

Table 5.1 shows that for both $n = 50$ and $n = 100$, QKIC^R chooses the generating model more often than QIC^R . Also, in agreement with the asymptotic characteristics of the criteria, as the sample size increases, the performance of the criteria improves.

n	model	QIC^R	QKIC^R
50	AR-1	172	183
50	Exchangeable	668	712
50	Independence	160	105
100	AR-1	123	128
100	Exchangeable	727	795
100	Independence	150	77

Table 5.1: Settings 23 and 24: Working correlation structure selections for QIC^R and QKIC^R in the binary correlated response data framework; results for the generating model are bolded.

5.2.2 Selection of Mean Structure

To investigate the performance of the criteria for mean structure selection, data are generated following the configurations detailed in the two following subsections. For each of the 1,000 replications, QIC^I , QKIC^I , QIC^R , QKIC^R , QIC^U , and QKIC^U are evaluated for each fitted model. The fitted model with the minimum value for each statistic is recorded. The results are then summarized in tables. The term order again refers to the number of parameters in the linear predictor. We present results in the NM setting, where the model selection problem is reduced to choosing an appropriate order.

5.2.2.1 Correlated Binary Response Data

For selecting the mean structure, when the response data are correlated and binary, we generate the data using the methods suggested by Leisch et al. (1998). Any two measurements in the same cluster have the same correlation ρ ; that is, repeated measures within clusters follow an exchangeable correlation structure. We

configure a simulation experiment based on two data generating models and two sample sizes.

Data for settings 25 and 26 are generated using the following third-order model:

$$\text{logit}(\pi_{it}) = 0.25 - 0.25x_{i2t} - 0.25t$$

where $x_{i2t} \sim \text{iid Bernoulli}(0.5)$, $t = 0, 1, 2$ (i.e., three measurements per cluster), $\rho = 0.5$ (exchangeable correlation structure), $n = 50$ (setting 25) and $n = 100$ (setting 26).

In settings 25 and 26, we apply the model selection criteria to the following candidate set:

$$\text{logit}(\pi_{it}) = \beta_1 + \beta_2x_{i2t} \tag{5.15}$$

$$\text{logit}(\pi_{it}) = \beta_1 + \beta_2x_{i2t} + \beta_3t \tag{5.16}$$

$$\text{logit}(\pi_{it}) = \beta_1 + \beta_2x_{i2t} + \beta_4x_{i4t} \tag{5.17}$$

$$\text{logit}(\pi_{it}) = \beta_1 + \beta_2x_{i2t} + \beta_3t + \beta_4x_{i4t} \tag{5.18}$$

$$\text{logit}(\pi_{it}) = \beta_1 + \beta_2x_{i2t} + \beta_3t + \beta_4x_{i4t} + \beta_5x_{i5t}, \tag{5.19}$$

where x_{i2t} and t are as in the previous paragraph, and x_{i4t} and $x_{i5t} \sim \text{iid } U(-1, 1)$. Settings 25 and 26 recreate the simulation experiment for mean structure selection presented in Pan (2001).

Data for settings 27 and 28 are generated using the following fifth-order model:

$$\text{logit}(\pi_{it}) = 0.25 + 0.25x_{i2t} - 0.25t + 0.25x_{i4t} - 0.25x_{i5t},$$

where x_{i2t} , x_{i4t} and $x_{i5t} \sim \text{iid Binomial}(7, 0.5)$, $t = 0, 1, 2$ (i.e., three measurements per cluster), $\rho = 0.1$ (exchangeable correlation structure), $n = 100$ (setting 27) and $n = 150$ (setting 28).

The candidate model set for settings 27 and 28 is as follows:

$$\text{logit}(\pi_{it}) = \beta_1 + \beta_2 x_{i2t}$$

$$\text{logit}(\pi_{it}) = \beta_1 + \beta_2 x_{i2t} + \beta_3 t$$

$$\text{logit}(\pi_{it}) = \beta_1 + \beta_2 x_{i2t} + \beta_3 t + \beta_4 x_{i4t}$$

$$\vdots$$

$$\text{logit}(\pi_{it}) = \beta_1 + \beta_2 x_{i2t} + \beta_3 t + \beta_4 x_{i4t} + \dots + \beta_1 1 x_{i11t},$$

where x_{i2t} , x_{i3t} , t and x_{i5t} are as in the previous paragraph, and x_{i6t} to $x_{i11t} \sim \text{iid } N(0, 1)$.

n	model	QIC^I	QKIC^I	QIC^R	QKIC^R	QIC^U	QKIC^U
50	(5.19)	50	45	122	51	43	9
50	(5.18)	85	29	108	68	52	19
50	(5.17)	101	49	60	61	60	42
50	(5.16)	492	445	458	415	323	209
50	(5.15)	272	432	252	405	522	721
100	(5.19)	18	19	135	59	12	48
100	(5.18)	92	38	123	82	33	66
100	(5.17)	115	66	21	30	43	21
100	(5.16)	671	665	619	633	606	426
100	(5.15)	104	212	102	196	306	439

Table 5.2: Settings 25 and 26: NM order selections for QIC^I , QKIC^I , QIC^R , QKIC^R , QIC^U , and QKIC^U in the binary correlated response data framework; results for the generating model are bolded.

Table 5.2 shows that in settings 25 and 26, the variants of QIC outperform their QKIC counterparts by choosing more often the data generating model. This pattern holds except for QIC^R when $n = 100$. QIC^I is the criterion that performs best. Note that the variants of QIC tend to choose overfitted models more often than the variants of QKIC. The latter tend to choose underfitted models more often than the former.

n	order	QIC^I	QKIC^I	QIC^R	QKIC^R	QIC^U	QKIC^U
100	11	22	2	24	2	21	1
100	10	21	6	21	6	22	5
100	9	25	4	25	4	19	2
100	8	37	12	37	12	31	11
100	7	63	27	64	31	65	25
100	6	116	63	118	61	111	59
100	5	579	610	581	614	595	619
100	4	103	169	99	166	101	166
100	3	15	29	13	27	13	24
100	2	19	78	18	77	22	88

Table 5.3: Setting 27: NM order selections for QIC^I , QKIC^I , QIC^R , QKIC^R , QIC^U , and QKIC^U in the binary correlated response data framework; results for the generating model are bolded.

For settings 27 and 28 (Tables 5.3 and 5.4), the variants of QKIC exhibit a better performance than the variants of QIC. In these cases, QKIC^U chooses the generating model most often; however, it is difficult to argue that any of the variants of QKIC is best, because the number of correct selections for each of the variants

n	order	QIC^I	QKIC^I	QIC^R	QKIC^R	QIC^U	QKIC^U
150	11	21	1	21	1	16	1
150	10	18	1	20	1	18	1
150	9	24	3	25	3	23	2
150	8	46	15	44	15	41	13
150	7	57	26	56	26	53	24
150	6	131	82	133	81	129	77
150	5	663	771	665	774	681	776
150	4	33	75	31	74	32	76
150	3	2	7	1	6	1	9
150	2	5	19	4	19	6	21

Table 5.4: Setting 28: NM order selections for QIC^I , QKIC^I , QIC^R , QKIC^R , QIC^U , and QKIC^U in the binary correlated response data framework; results for the generating model are bolded.

is similar. The tendencies of QIC to select overfitted models and QKIC to select underfitted models are quite evident.

5.2.2.2 Correlated Count Response Data

For correlated count response data, we present a 2x2 factorial simulation study based on two data generating models and two samples sizes. Correlated count response data are generated using the methodology suggested by Yahav and Shmueli (2008) for multivariate Poisson random variables. For these experiments, any two measurements in the same cluster have the same correlation ρ ; that is, repeated measures within clusters follow an exchangeable correlation structure.

Data for settings 29 and 30 are generated using the following third-order model:

$$\log(\lambda_{it}) = 0.35 + 0.35x_{i2t} + 0.35t,$$

where $x_{i2t} \sim \text{iid Bernoulli}(0.5)$, $t = 0, 1, 2$ (i.e., three measurements per cluster), $\rho = 0.5$ (exchangeable correlation structure), $n = 50$ (setting 29) and $n = 100$ (setting 30).

In settings 29 and 30, model selection criteria choose from the following candidate set:

$$\log(\lambda_{it}) = \beta_1 + \beta_2x_{i2t}$$

$$\log(\lambda_{it}) = \beta_1 + \beta_2x_{i2t} + \beta_3t$$

$$\log(\lambda_{it}) = \beta_1 + \beta_2x_{i2t} + \beta_3t + \beta_4x_{i4t}$$

$$\vdots$$

$$\log(\lambda_{it}) = \beta_1 + \beta_2x_{i2t} + \beta_3t + \beta_4x_{i4t} + \beta_5x_{i5t} + \beta_6x_{i6t} + \beta_7x_{i7t} + \beta_8x_{i8t},$$

where x_{i2t} and t are as in the preceding paragraph, and x_{i4t} to $x_{i8t} \sim \text{iid } N(0, 1)$.

Data for settings 31 and 32 are generated using the following fifth-order generating model:

$$\log(\lambda_{it}) = 0.30 + 0.30x_{i2t} + 0.30t + 0.30x_{i4t} + 0.30x_{i5t},$$

where $x_{i2t} \sim \text{iid Bernoulli}(0.5)$, $x_{i4t} \sim \text{iid } N(0, 1)$, $x_{i5t} \sim \text{iid } N(0, 1)$, $t = 0, 1, 2$ (i.e., three measurements per cluster), $\rho = 0.5$ (exchangeable correlation structure), $n = 50$ (setting 31) and $n = 150$ (setting 32). The candidate set for settings 31 and 32 is the same as that for settings 29 and 30.

Tables 5.5 and 5.6 illustrate a similar pattern of results as the pattern observed for settings 29 and 30 in the previous section: the variants of QKIC select the correct model more often than the corresponding variants of QIC. When the variants of

n	order	QIC^I	QKIC^I	QIC^R	QKIC^R	QIC^U	QKIC^U
50	7	48	8	81	22	19	3
50	6	71	27	85	44	34	11
50	5	68	38	83	53	43	20
50	4	123	87	131	104	97	65
50	3	683	830	613	768	798	890
50	2	7	10	7	9	9	11
100	7	33	12	64	18	16	3
100	6	48	15	55	25	21	4
100	5	73	30	93	49	48	13
100	4	132	81	140	108	99	57
100	3	714	862	648	800	816	923
100	2	0	0	0	0	0	0

Table 5.5: Settings 29 and 30: NM order selections for QIC^I , QKIC^I , QIC^R , QKIC^R , QIC^U , and QKIC^U in the Poisson correlated response data framework; results for the generating model are bolded.

QKIC and QIC are compared, it can be observed that the variants of QIC are more prone to overfitting than those of QKIC . On the other hand, the latter are more prone to underfitting than the former. These experiments also reflect the asymptotic nature of the criteria: as sample size increases, the criteria choose the correct model more often.

In all the settings in this section, QKIC^U performs better than the rest of the criteria. QKIC^U is usually followed by QKIC^I .

n	order	QIC^I	QKIC^I	QIC^R	QKIC^R	QIC^U	QKIC^U
50	7	165	103	152	106	102	55
50	6	172	137	167	137	123	88
50	5	641	716	644	707	717	773
50	4	21	38	33	44	51	68
50	3	1	6	4	6	7	16
50	2	0	0	0	0	0	0
150	7	144	90	134	99	101	67
150	6	179	143	175	143	145	105
150	5	677	767	690	757	753	827
150	4	0	0	1	1	1	1
150	3	0	0	0	0	0	0
150	2	0	0	0	0	0	0

Table 5.6: Settings 31 and 32. NM order selections for QIC^I , QKIC^I , QIC^R , QKIC^R , QIC^U , and QKIC^U in the Poisson correlated response data framework; results for the generating model are bolded.

5.3 Application

We illustrate the use of QIC and QKIC by analyzing part of the data recently published by Mikami et al. (2011). These authors study the effect of antidepressants on the temporal course of disability in stroke patients.

5.3.1 Study Description

A total of 83 patients entered a double-blind randomized clinical trial to investigate the efficacy of antidepressants to treat depressive disorders and reduce disability after stroke. Patients were assigned to either active antidepressant treatment (i.e., 32 participants received fluoxetine and 22 nortriptyline) or placebo ($n = 29$). The modified Rankin Scale (mRS) (van Swieten et al., 1988) was used to evaluate the temporal course of disability following stroke. The mRS is a scale that ranges from 0 to 6. Lower scores mean the patient is less disabled. Patients within 6 months of sustaining a stroke were examined at the time of entry to the study and at three-month follow-up visits throughout one year, for a total of 5 evenly-spaced times (i.e., $t = 1, 2, \dots, 5$). In addition to the effect of the antidepressant treatment, the effects of age and the intensity of rehabilitation care are variables of interest in predicting the longitudinal course of mRS scores. The intensity of rehabilitation care is assessed by the total hours of physical rehabilitation received during the study.

We consider the data for the 56 patients that completed all evaluations during one year. Their age and total rehabilitation hours were also available. In this group with complete data, a total of 21 participants received placebo and 35 participants received antidepressants. Given the correlated and count-like nature of the data, we use a candidate set based on the Poisson distribution, similar to the set presented in section 5.2.2.2.

5.3.2 Comparison of Model Selection Criteria

We first use QIC^R and QKIC^R to choose the best working correlation structure among independent, exchangeable or AR-1. We include the following effects of interest in the mean structure: age (henceforth referred to as x_{i2}), total of physical rehabilitation hours (x_{i3}), antidepressant treatment (x_{i4}), evaluation time (t), and

the interaction between evaluation time and antidepressant treatment (tx_{i4}). After choosing a working correlation structure, we compare all variants of QIC and QKIC to choose a suitable model from the following candidate set:

$$\log(\lambda_{it}) = \beta_1 + \beta_5 t \quad (5.20)$$

$$\log(\lambda_{it}) = \beta_1 + \beta_2 x_{i2} + \beta_5 t \quad (5.21)$$

$$\log(\lambda_{it}) = \beta_1 + \beta_3 x_{i3} + \beta_5 t \quad (5.22)$$

$$\log(\lambda_{it}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_5 t \quad (5.23)$$

$$\log(\lambda_{it}) = \beta_1 + \beta_4 x_{i4} + \beta_5 t + \beta_6 tx_{i4} \quad (5.24)$$

$$\log(\lambda_{it}) = \beta_1 + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 t + \beta_6 tx_{i4} \quad (5.25)$$

$$\log(\lambda_{it}) = \beta_1 + \beta_2 x_{i2} + \beta_4 x_{i4} + \beta_5 t + \beta_6 tx_{i4} \quad (5.26)$$

$$\log(\lambda_{it}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 t + \beta_6 tx_{i4}. \quad (5.27)$$

Tables 5.7 and 5.8 show how the model selection criteria ranked the different candidate models.

model	QIC ^R	QKIC ^R
AR-1	1	1
Exchangeable	2	2
Independence	3	3

Table 5.7: QIC^R and QKIC^R rankings for working correlation structure candidate models.

As can be seen in Table 5.7, both QIC^R and QKIC^R indicate that an AR-1 working correlation structure is better than working independence or an exchangeable working correlation structure. Both criteria rank candidate models in the same way.

model	QIC^I	QKIC^I	QIC^R	QKIC^R	QIC^U	QKIC^U
(5.27)	2	3	2	3	1	1
(5.26)	5	7	5	6	5	5
(5.25)	1	1	1	1	2	2
(5.23)	4	4	3	4	3	3
(5.24)	7	6	7	7	7	7
(5.22)	3	2	4	2	4	4
(5.21)	6	5	6	5	6	6
(5.20)	8	8	8	8	8	8

Table 5.8: QIC^I , QKIC^I , QIC^R , QKIC^R , QIC^U , and QKIC^U rankings for mean structure candidate models.

Table 5.8 shows that the model including age, antidepressant treatment, evaluation time and the interaction between antidepressant treatment and evaluation time is the choice of QIC^I , QKIC^I , QIC^R and QKIC^R . This model is the second choice for QIC^U and QKIC^U , both of which select the largest model in the candidate set. The smallest model in the set, which includes only evaluation time, is ranked last by all of the model selection criteria. Also, for this application, QIC^U and QKIC^U agree in all the model rankings. With the exception of two models, QIC^I and QIC^R also agree on their rankings. The same is true for QKIC^I and QKIC^R . Except for those models where the rankings of QKIC^I and QKIC^R agree with those of their QIC analogues, QKIC^I and QKIC^R rank models with fewer parameters higher than their QIC analogues.

If researchers cannot discern based on scientific reasoning between models (5.25) and (5.27), most of the simulations in the previous section indicate that

model (5.27), the choice of QKIC^U , should be adopted for these data. As a side note, all the effects included in model (5.27) are statistically significant, meaning that model (5.25) lacks what appears to be an important covariate.

CHAPTER 6

CONCLUSIONS AND FUTURE DIRECTIONS

We close this manuscript with general conclusions based on the work presented in chapters 3 to 5. We also mention the limitations of our work and propose some important future directions.

6.1 Conclusions

We have developed and investigated three asymptotically unbiased estimators of different variants of KSD that serve as selection criteria in different modeling frameworks and under different sets of assumptions.

We have characterized the performance of our proposed criteria in an extensive collection of simulation studies. Our simulation experiments provide preliminary indications that KIC_o is a selection criterion well suited for traditional GLM frameworks where the candidate model set is prone to overfitting. The experiments also suggest that KIC_u performs best in traditional GLM frameworks where the candidate set is prone to underfitting, or where the objective is to choose a model for predicting new data. QKIC^I appears to be a suitable model selection criterion for correlated data in the GLM framework, when estimation is performed using either a quasi-likelihood or GEEs. We illustrate the use of QKIC variants with a real-world data example.

A noteworthy conclusion of the simulation studies for KIC_o and KIC_u is that KSD performs systematically at least as well as KDD. The only instances where the two oracles exhibit a similar performance are those where there is almost no room for improvement (e.g., where the oracle success rates at choosing the data generating model are over 95%). To our knowledge, this is the most extensive comparison of both oracles published to date.

6.2 Limitations and Future Directions

Our work has some limitations, both in the theoretical development of the proposed criteria, and in the scope of coverage of some of the topics developed in this dissertation.

The derivation of KIC_u has two steps that were derived heuristically. Specifically, the relations (3.16) and (3.19) were used in the derivation of KIC_u even though they were not rigorously justified. $QKIC^R$ and $QKIC^U$ are also presented without a solid theoretical foundation. In addition, KIC_o and $QKIC^I$ could be further explored in the case where the canonical link is not used.

The characterization of the behaviors of the proposed criteria through simulation is somewhat limited. We present more extensive simulation studies for KIC_o and KIC_u than for the $QKIC$ variants. However, these experiments are insufficient to uncover all of the settings where each of these criteria might perform optimally. In particular, the study in the prediction setting is very limited. In order to conduct further experiments, prediction measures more suitable for the GLM framework than the MSEP should be explored for binary and count outcomes (e.g., area under the ROC curve for binary data).

In the case of $QKIC^I$, the literature shows that criteria developed using only the bias correction term of QIC are better than QIC at identifying suitable working correlation structures. Our simulation studies for choosing working correlation structures are very restricted. We anticipate that further experiments comparing $QKIC$ to other alternatives proposed in the literature (e.g., CIC) would show that $QKIC$ is better suited for mean structure selection than for the selection of a working correlation structure. For all the criteria presented in this work, a more complete behavioral characterization through simulation studies would enrich their scope of usage.

Shibata (1997) shows that the bootstrap estimates of KDD proposed by Efron

(1983, 1986) and Cavanaugh and Shumway (1997) are asymptotically equivalent to AIC when $g(\mathbf{y}|\boldsymbol{\theta}_0) \in \mathcal{F}(k)$ and asymptotically equivalent to TIC when $g(\mathbf{y}|\boldsymbol{\theta}_0)$ is not necessarily included in $\mathcal{F}(k)$. This indicates that bootstrap estimates of KSD could also be pursued, possibly leading to the development of bootstrap analogues of KIC, KIC_o , and KIC_u .

All the aforementioned limitations and avenues warrant further work and investigation, and provide possible future directions based on the results presented in this dissertation.

REFERENCES

- Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in: B. N. Petrov and F. Csaki, eds., *2nd International Symposium on Information Theory* (Akademia Kiado, Budapest), 267–281.
- Akaike, H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **AC-19**, 716–723.
- Akaike, H. (1985), Prediction and entropy, in: A. Atkinson and E. Fienberg, eds., *A Celebration of Statistics* (Springer-Verlag, New York), 1–24.
- Azari, R., Li, L. and Tsai, C. L. (2006), Longitudinal data model selection, *Computational Statistics and Data Analysis* **50**, 3053–3066.
- Bedrick, E. J. and Tsai, C. L. (1994), Model selection for multivariate regression in small samples, *Biometrics* **50**, 226–231.
- Broersen P. M. T. and Wensink H. E. (1996), On the penalty factor for autoregressive order selection in finite samples, *IEEE Transactions on Signal Processing* **44**, 748–752.
- Burnham, K. P. and Anderson, D. R. (2002), *Model Selection and Multimodel Inference* (Springer-Verlag, New York).
- Cantoni, E., Flemming, J. M. and Ronchetti, E. (2005), Variable selection for marginal longitudinal generalized linear models, *Biometrics* **61**, 507–514.
- Casella, G. and Berger, R (2002), *Statistical Inference* (Duxbury Press).
- Cavanaugh, J. E. (1999), A large-sample model selection criterion based on Kullback’s symmetric divergence, *Statistics and Probability Letters* **44**, 333–344.
- Cavanaugh, J. E. (2004), Criteria for linear model selection based on Kullback’s symmetric divergence, *Australian and New Zealand Journal of Statistics* **46**, 257–274.
- Cavanaugh, J. E. and Shumway, R. H. (1997), A bootstrap variant of AIC for state-space model selection, *Statistica Sinica* **7**, 473–496.
- Claeskens, G. and Hjort, N. (2003), The focused information criterion, *Journal of the American Statistical Association* **98**, 900–916.
- Claeskens, G. and Hjort, N. (2008), *Model Selection and Model Averaging* (Cambridge University Press, Cambridge).

- Cui, J. and Qian, G. (2007), Selection of working correlation structure and best model in GEE analyses of longitudinal data, *Communications in Statistics – Simulation and Computation* **36**, 987–996.
- Efron, B. (1983), Estimating the error rate of a prediction rule: Improvement on cross-validation, *Journal of the American Statistical Association* **78**, 316–331.
- Efron, B. (1986), How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81**, 461–470.
- Goutis, C. and Robert, C. P. (1998), Model choice in generalized linear models: A Bayesian approach via Kullback-Leibler projections, *Biometrika* **85**, 29–37.
- Hafidi, B. and Mkhadri, A. (2006), A corrected Akaike criterion based on Kullback’s symmetric divergence: applications in time series, multiple and multivariate regression, *Computational Statistics and Data Analysis* **50**, 1524–1550.
- Hafidi, B. and Mkhadri, A. (2010), The Kullback information criterion for mixture regression models, *Statistics and Probability Letters* **80**, 807–815.
- Hardin, J. W. and Hilbe, J. M. (2002), *Generalized Estimating Equations* (Chapman and Hill/CRC, Boca Raton, Florida).
- Hin, L. Y., Carey, V. J. and Wang, Y. G. (2007), Criterion for working-correlation-structure selection in GEE: Assessment via simulation, *The American Statistician* **61**, 360–364.
- Hin, L. Y. and Wang, Y. G. (2009), Working-correlation-structure identification in generalized estimating equations, *Statistics in Medicine* **28**, 642–658.
- Hurvich, C. M. and Tsai, C. L. (1989), Regression and time series model selection in small samples, *Biometrika* **76**, 297–307.
- Hurvich, C. M. and Tsai, C. L. (1993), A corrected Akaike information criterion for vector autoregressive model selection, *Journal of Time Series Analysis* **14**, 271–279.
- Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997), Bootstrapping log likelihood and EIC, an extension of AIC, *Annals of the Institute of Statistical Mathematics* **49**, 411–434.
- Jiang, J., Nguyen, T. and Rao, J. (2009), A simplified adaptive fence procedure, *Statistics and Probability Letters* **79**, 625–629.
- Kim, H. J. 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.

- Kinney, S. K. and Dunson, D. B. (2007), Fixed and random effects selection in linear and logistic models, *Biometrics* **63**, 690–698.
- Kitagawa, G. (1987), Non-Gaussian state-space modeling of nonstationary time series: Rejoinder, *Journal of the American Statistical Association* **82**, 1060–1063.
- Konishi, S. and Kitagawa, G. (1996), Generalized information criteria in model selection, *Biometrika* **83**, 875–890.
- Konishi, S. and Kitagawa, G. (2008), *Information Criteria and Statistical Modeling* (Springer, New York).
- Kullback, S. (1968), *Information Theory and Statistics* (Dover, New York).
- Kullback, S. and Leibler, R. (1951), On information and sufficiency, *Annals of Mathematical Statistics* **22**, 79–86.
- Lavergne, C., Martinez, M. J. and Trottier, C. (2008), Empirical model selection in generalized linear mixed effects models, *Computational Statistics* **23**, 99–109.
- Liang, K. Y. and Zeger, S. L. (1986), Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.
- Leisch, F., Weingessel, A. and Hornik, K. (1998), On the generation of correlated artificial binary data, *Adaptive Information Systems and Modelling in Economics and Management Science – Working Paper Series* **13** – <http://epub.wu.ac.at/286/> Technical Report Vienna University of Economics and Business, last accessed August 4th, 2011.
- Linhart, H. and Zucchini, W. (1986), *Model Selection* (Wiley, New York).
- Liu, H., Weiss, R. E., Jennrich, R. I. and Wenger, N. S. (1999), PRESS model selection in repeated measures data, *Computational Statistics and Data Analysis* **30**, 169–184.
- Ljung, L. and Caines, P. E. (1979), Asymptotic normality of prediction error estimators for approximate system models, *Stochastics* **3**, 29–46.
- 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, Singapore).
- Mikami, K., Jorge, R., Adams, H., Davis, P., Leira, E., Jang, M. and Robinson, R. (2011), Effect of antidepressants on the course of disability following stroke, *American Journal of Geriatric Psychiatry* **in press**.

- Muller, S. and Welsh, A. H. (2009), Robust model selection in generalized linear models, *Statistica Sinica* **19**, 1155–1170.
- Nelder, J. and Wedderburn, R. (1972), Generalized linear models, *Journal of the Royal Statistical Society Series A – General* **135**, 370–384.
- Nott, D. J. and Leng, C. (2010), Bayesian projection approaches to variable selection in generalized linear models, *Computational Statistics and Data Analysis* **54**, 3227–3241.
- Pan, W. (2001), Akaike’s information criterion in generalized estimating equations, *Biometrics* **57**, 120–125.
- Pan, W. and Connett, J. E. (2002), Selecting the working correlation structure in generalized estimating equations with application to the lung health study, *Statistica Sinica* **12**, 475–490.
- Pu, W. and Niu, X. F. (2006), Selecting mixed-effects models based on a generalized information criterion, *Journal of Multivariate Analysis* **97**, 733–758.
- Rao, C. R. and Wu, Y. (2001), On model selection, in: P. Lahiri, ed., *Model Selection, Lecture Notes – Monograph Series* (Institute of Mathematical Statistics) **38**, 1–57.
- Rotnitzky, A. and Jewell, N. P. (1990), Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data, *Biometrika* **77**, 485–497.
- SAS Institute (2007), Sample 26100: QIC goodness of fit statistic for GEE models, <http://support.sas.com/kb/26/100.html#ref>, last accessed on June 7th, 2011.
- Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics* **6**, 461–464.
- Seghouane, A. (2006), Vector autoregressive model-order selection from finite samples using Kullback’s symmetric divergence, *IEEE Transactions Circuits and Systems* **53**, 2327–2335.
- Shang, J. and Cavanaugh, J. E. (2008), Bootstrap variants of the Akaike information criterion for mixed model selection, *Computational Statistics and Data Analysis* **52**, 2004–2021.
- Shibata, R. (1989), Statistical aspects of model selection, in: J. C. Willemsa, ed., *From Data to Model*, Springer-Verlag, London, 215–240.
- Shibata, R. (1997), Bootstrap estimate of Kullback-Leibler information for model selection, *Statistica Sinica* **7**, 375–394.

- Shults, J., Sun, W., Tu, X., Kim, H., Amsterdam, J., Hilbe, J. and Ten-Have, T (2009), A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data, *Statistics in Medicine* **28**, 2338–2355.
- Sugiura, N. (1978), Further analysis of the data by Akaike’s information criterion and the finite corrections, *Communications in Statistics* **A7**, 13–26.
- Takeuchi, K. (1976), Distributions of information statistics and criteria for adequacy of models, *Mathematical Science* **153**, 12–18.
- Vaida, F. and Blanchard, S. (2005), Conditional Akaike information for mixed-effects models, *Biometrika* **92**, 351–370.
- van Swieten, J., Koudstaal, P., Visser, M., Schouten, H. and van Gijn, J. (1988), Interobserver agreement for the assessment of handicap in stroke patients, *Stroke* **19**, 604–607.
- Wang, Y. and Hin, L. (2010), Modeling strategies in longitudinal data analysis: Covariate, variance function and correlation structure selection, *Computational Statistics and Data Analysis* **54**, 3359–3370.
- Wang, L. and Qu, A. (2009), Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach, *Journal of the Royal Statistical Society Series B – Statistical Methodology* **71**, 177–190.
- Wedderburn, R. (1974), Quasi-likelihood functions, generalized linear-models, and Gauss-Newton method, *Biometrika* **61**, 439–447.
- White, H. (1982), Maximum likelihood estimation of misspecified models, *Econometrica* **50**, 1–25.
- Yafune, A., Funatogawa, T. and Ishiguro, M. (2005), Extended information criterion (EIC) approach for linear mixed effects models under restricted maximum likelihood (REML) estimation, *Statistics in Medicine* **24**, 3417–3429.
- Yahav, I. and Shmueli, G. (2008), An elegant method for generating multivariate Poisson random variables, *arXiv:0710.5670v2 [stat.CO]* – <http://arxiv.org/abs/0710.5670> Technical Report Cornell University Library, last accessed August 4th, 2011.