Information Matrices in Estimating Function Approach: Tests for Model Misspecification and Model Selection

by

Qian Zhou

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2009

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Estimating functions have been widely used for parameter estimation in various statistical problems. Regular estimating functions produce parameter estimators which have desirable properties, such as consistency and asymptotic normality. In quasi-likelihood inference, an important example of estimating functions, correct specification of the first two moments of the underlying distribution leads to the information unbiasedness, which states that two forms of the information matrix: the negative sensitivity matrix (negative expectation of the first order derivative of an estimating function) and the variability matrix (variance of an estimating function) are equal, or in other words, the analogue of the Fisher information is equivalent to the Godambe information. Consequently, the information unbiasedness indicates that the model-based covariance matrix estimator and sandwich covariance matrix estimator are equivalent. By comparing the model-based and sandwich variance estimators, we propose information ratio (IR) statistics for testing model misspecification of variance/covariance structure under correctly specified mean structure, in the context of linear regression models, generalized linear regression models and generalized estimating equations. Asymptotic properties of the IR statistics are discussed. In addition, through intensive simulation studies, we show that the IR statistics are powerful in various applications: test for heteroscedasticity in linear regression models, test for overdispersion in count data, and test for misspecified variance function and/or misspecified working correlation structure. Moreover, the IR statistics appear more powerful than the classical information matrix test proposed by White (1982).

In the literature, model selection criteria have been intensively discussed, but almost all of them target choosing the optimal mean structure. In this thesis, two model selection procedures are proposed for selecting the optimal variance/covariance structure among a collection of candidate structures. One is based on a sequence of the IR tests for all the competing variance/covariance

structures. The other is based on an "information discrepancy criterion" (IDC), which provides a measurement of discrepancy between the negative sensitivity matrix and the variability matrix. In fact, this IDC characterizes the relative efficiency loss when using a certain candidate variance/covariance structure, compared with the true but unknown structure. Through simulation studies and analyses of two data sets, it is shown that the two proposed model selection methods both have a high rate of detecting the true/optimal variance/covariance structure. In particular, since the IDC magnifies the difference among the competing structures, it is highly sensitive to detect the most appropriate variance/covariance structure.

Acknowledgements

First and foremost, I would like to express my deep and sincere gratitude to my supervisors, Professor Mary E. Thompson and Professor Peter X.-K. Song. They have supported me throughout my thesis with their wide knowledge, great patience, constant encouragement and personal guidance.

I also wish to express my special appreciation to Professor Richard J. Cook, Professor Grace Y. Yi, Professor Anindya Sen, and Professor Charmaine Dean for their assistance and valuable advice and for serving as thesis committee members.

Many thanks go to all faculty members, administrative staff and my fellow graduate students for their help rendered to me during my studies.

Finally, I am forever indebted to my parents for their understanding, endless encouragement and unconditional love.

Dedication

This is dedicated to my dear father.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Estimating functions have been widely used for parameter estimation in various statistical models. Estimating function theory can be regarded as a generalization of maximum likelihood theory, which was originally introduced by [28] and [29]. An estimating function takes a form $\Psi(y, \boldsymbol{\theta})$, where y represents the data, and $\boldsymbol{\theta}$ is a set of parameters of interest. An estimate of the parameter $\boldsymbol{\theta}$ is obtained as a solution to an estimating equation of the form

$$\Psi(y, \boldsymbol{\theta}) = 0.$$

Estimating equations may be derived from a fully specified parametric model. For example, a score equation derived from a log linear regression model for count data is an estimating equation, where $\boldsymbol{\theta}$ is the vector of regression coefficients. This equation produces the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$, known as being a fully efficient estimator. However, often statistical models cannot be fully specified, due to the lack of enough information or knowledge about the underlying probabilistic mechanism from which the data are generated. In other cases, from the previous analysis of similar data, it may be suspected that some of the distributional assumptions are violated. As a result, investigators are only able to impose assumptions on some aspects of the probability distribution. For instance, in least squares estimation for linear regression models (LM), only

the first two moments of the data distribution are assumed, instead of a complete parametric distribution.

1.1 Literature Review

The term "equation of estimation" was first used in [30]. [51] presented a non-trivial example of an estimating function, where estimating equations are proposed to construct confidence regions for the parameters in Gumbel distributions. Later, [60] generalized Kimball's idea of stable estimating equations to establish a theory of sufficiency and ancillarity for estimating functions.

The theory of optimal estimating functions was first studied by [35]. He introduced a measure, now known as *Godambe information*, as a criterion to define an optimal estimating function among the class of regular estimating functions. In addition, he pointed out that the score function is optimal in the sense that it has the maximum Godambe information, and this maximum Godambe information is equal to the Fisher information (Also see [36]). Moreover, the asymptotic covariance matrix of the parameter estimator is equal to the inverse of the Godambe information. An estimator of this covariance matrix was called the *heteroscedasticity-consistent covariance matrix estimator* in the LM (see [45] and [88]), and was later referred to as the *sandwich variance estimator*, which has been widely used in longitudinal data analysis; see [23], [53] and [54]. Even when the distributional assumption cannot be fully specified or fails to hold, the sandwich estimator is still able to provide a consistent estimate of the asymptotic covariance matrix for parameter estimators. Because of this property, the sandwich covariance matrix estimator is also called the *robust covariance matrix estimator*.

In the theory of maximum likelihood estimation, if the distributional assumption for the data analysis is correct and regularity conditions are satisfied, the Bartlett identity holds. That is, the Fisher information matrix can be expressed

as a negative sensitivity or variability matrix form, namely

$$E\left\{-\ddot{l}(\theta)\right\} = Var\left\{\dot{l}(\theta)\right\},$$

where $l(\theta)$ is the log-likelihood function, and $\dot{l}$ and $\ddot{l}$ stand for the first order and second order derivatives of the function l with respect to (w.r.t.) θ, respectively. This implies that when the Bartlett identity fails, a certain distributional assumption is misspecified. By comparing the negative sensitivity and variability matrices, [87] introduced an information matrix (IM) test for model misspecification. The IM test can detect the inconsistency of the usual maximum likelihood covariance matrix estimator at the very least, as well as possible inconsistency of the MLE for parameters of interest. This IM test will be reviewed in this thesis.

As an important example of estimating functions, quasi-likelihood inference in the context of generalized linear models (GLMs) (see [86]) only imposes the assumptions of the first two moments, instead of fully specifying the parametric distribution. If the quasi-score function is unbiased, which usually results from the correct formulation of the mean structure of the responses, the resultant estimator of the parameter of interest is consistent. Moreover, if the assumption of the variance structure is correctly specified, the resultant estimator will achieve the same estimation efficiency as that of the most efficient estimator. In the statistical literature, three test statistics were most commonly considered for testing the adequacy of the mean structure under the likelihood and quasi-likelihood theory: the Wald test based on comparison of estimated coefficients with their standard errors ([82]); the (quasi) likelihood ratio test based on comparison of deviances among nested models ([59] and [6]); and the (quasi) likelihood score test ([59], [83] and [69]). In longitudinal data analysis, [68] proposed a method of quadratic inference functions (QIF) to test for regression misspecification. On the other hand, it is also of importance to assess the validity of the second moment assumption for the sake of estimation efficiency.

In the LM setting, it is conventional to assume that the error terms all have equal variances, which is referred sometimes to as a homogeneity assumption.

Methods for checking on this assumption have been well investigated in the literature. [88] obtained a direct test for heteroscedasticity, which is a special case dealt with by the IM test of [87]. In the literature, many other tests were proposed either on the basis of a specific alternative model for heteroscedasticity (see for example [5], [7] and [14]), or on the basis of a certain non-parametric or semi-parametric variance function model (see for example [62] and [22]), or on the basis of plausible, but *ad hoc* grounds (see for example [37]; [34]; [42]). In addition, some robust tests were proposed by [7]; see also [39] and [11].

In the context of GLM, overdispersion is a common case contributing to violation of the mean-variance relation. It prohibits investigators from using a specific parametric distribution, for example, a Poisson regression model, for the count data. [20] proposed score tests against arbitrary mixed Poisson alternatives, which are generalizations of tests of [27] and [13]. See also [21]. However, for other types of misspecification of the mean-variation relation, there are few proposals available to develop and assess statistical methods for testing the variance structure. For longitudinal data analysis, in the context of generalized estimating equations (GEE), there is a lack of a thorough investigation on tests for misspecification of covariance structure, including variance function and/or working correlation structure.

Model selection is the task of choosing a statistical model, from a set of potential models, which is the best approximation of reality manifested in the observed data. In the statistical literature, numerous model selection criteria have been intensively discussed. Some model selection procedures can be constructed based on a sequence of hypothesis testing. For example, forward selection, backward selection and stepwise regression are popular model selection methods in LM (See [55]). In their applications, controlled by one or two thresholds, the model is selected based on statistical hypothesis testing. See [52] and [61]. To avoid the difficulty in the choice of thresholds, some alternative model selection procedures were suggested based on the prediction errors. [1] defined a final prediction error (FPE), which is the mean squared prediction error when a model

fitted to the current data is applied to another independent future observation, or to make a one step prediction. The FPE is asymptotically equivalent to the C_p criterion proposed by [56].

In LM, a large number of predictors are usually introduced at the initial stage of modeling. [80] proposed a new approach, called least absolute shrinkage and selection operator (LASSO), which simultaneously selects variables and estimates parameters. The LASSO is closely related to the penalized likelihood with the L_1 penalty. Cross-validation was discussed as a common method for model selection in terms of the predictive ability of model. See [78], [24], [4] and [33]. Some model selection criteria were constructed based upon the Kullback-Leibler distance (information). One of the most important and popular criteria is Akaike's information criterion (AIC) defined by [2], based on the concept of minimizing the expected Kullback-Leibler distance. The AIC takes the form:

$$\text{AIC} = -2\log(\text{maximum likelihood}) + 2k,$$

where k is the number of parameters, which provides a balance of goodness of fit and simplicity of the model. To address the inconsistency of AIC, [3] and [73] introduced two equivalent consistent model selection criteria, SIC and BIC respectively, from a Bayesian perspective.

In the context of GEE, if we use a more general working correlation matrix, there is no guarantee that a corresponding quasi-likelihood exists unless certain conditions are satisfied ([59]). Furthermore, even if it exists, in general it is difficult to construct. [65] proposed a modification to AIC, named the "quasi-log-likelihood under the independence working correlation information criterion" (QIC). The QIC involves using the quasi-likelihood constructed under the working independence model, and the penalty term involves both the model-based and sandwich covariance matrix estimates of estimated regression coefficients. However, [43] conducted an investigation about the performance of the QIC concerning the selection of working correlation structure. They found that the performance of the QIC is impaired by the fact that the key term of goodness

of fit, $-2\log L(\widehat{\boldsymbol{\theta}})$, is theoretically irrelevant to the parameters in any correlation structure, but has to be estimated with an error. [43] proposed a correlation information criterion (CIC), which is defined by only the utility of the penalty term of the QIC. Extensive simulation studies in this paper have shown that the CIC has remarkable improvements in selecting the true correlation structure. Almost all the model selection criteria above aim at selecting the optimal mean structure. However, there is lack of powerful methods providing systematic criteria for selecting the optimal variance/covariance structure, even though the CIC is used to select a working correlation structure.

1.2 Objectives

In this thesis, we focus on the quasi-likelihood inference for independent data in the context of GLM and the GEE method for longitudinal data analysis. Both the quasi-score equations and GEE require only the assumptions of the first two moments. Under the correct formulation of the mean structure, if the variance structure is correctly specified, two forms of the information matrix, the negative sensitivity matrix and the variability matrix, will be equivalent. As a result, the asymptotic covariance matrix of the regression coefficient estimator can be estimated by either the model-based covariance matrix estimator, or the sandwich covariance matrix estimator. On the other hand, certain model misspecifications of the second moment assumption will lead to a discrepancy between these two covariance matrix estimators.

Thus, the main objective of the thesis is, through comparison between the model-based and sandwich covariance matrix estimators, to construct a systematic test, called the information ratio (IR) test. This IR test targets testing for model misspecifications of the variance/covariance structure, assuming that the mean structure is correctly specified. In addition, the p-values from the proposed IR tests may be used to select the optimal variance/covariance structure

among a collection of candidates. Moreover, for the variance/covariance selection, we propose an "information discrepancy criterion" (IDC), which measures the discrepancy between two forms of information matrices under a general variance/covariance structure.

1.3 Main Results

(i) We propose a new estimator of the dispersion parameter in the context of GLM and GEE. In model-based variance estimators, the dispersion parameter is usually estimated by the method of moments, if its true value is unknown. Analogously, we show that the sandwich variance estimator of each individual regression coefficient estimator provides a Godambian estimator of the dispersion parameter in the form of a weighted sum of the squared Pearson residuals for GLM, or a weighted sum of the squared transformed residuals for GEE. The weights in these Godambian estimators are differences between the leverage from two hat matrices, which characterize the influence from the corresponding covariate variables.

(ii) We propose information ratio statistics by taking ratios of the Godambian estimators to the true value, if known, or the moment estimator, otherwise, of the dispersion parameter. It shows that the information ratio statistics are asymptotically distributed as $N(0,1)$ random variables. The finite-sample distribution of the proposed IR statistic is found to be heavily right skewed, but a normalized χ^2_ν approximation can improve the performance. Through several simulation studies, we apply the information ratio statistics to test for heteroscedasticity in LM, test for overdispersion in count data, and test for misspecified variance function and/or working correlation structure in GEE. The simulation experiments have shown that the information ratio statistics are robustly powerful to reject the null hypothesis under different scenarios of alternative hypotheses.

(iii) We propose a new criterion for selecting the optimal variance/covariance structure. This criterion is constructed on the discrepancy between two forms of information matrices, so it is called the "Information Discrepancy Criterion" (IDC). We show via simulation studies and data analyses that the proposed IDC has a high rate of detecting the true/optimal variance/covariance structure.

1.4 Organization of Thesis

Chapter 2 gives an introduction to the theory of estimating functions. In this chapter, properties of regular estimating functions are discussed. Quasi-likelihood inference for GLM and GEE method are studied in detail. In addition, several simulation studies illustrate the model robustness of the sandwich variance estimators. At the end of the chapter, we review the information matrix test proposed by [87], which is one of the important contributions to tests for model misspecification.

Chapter 3 focuses on the formulation of the Godambian estimators of the dispersion parameter in the sandwich variance estimators. We show that the Godambian estimator takes the form of a weighted sum of the squared Pearson residuals in GLM or the squared transformed residuals in GEE. Properties of the weights in the Godambian estimators have been investigated in the special case of LM.

In Chapter 4, we propose information ratio statistics that take ratios of the Godambian estimators to the true value or the moment estimator of the dispersion parameter. Several simulation studies are carried out to investigate the asymptotic distributions of these test statistics under the null hypothesis. Moreover, we assess and compare the power of the proposed IR tests under different alternative hypotheses with the information matrix test proposed by [87].

We propose two model selection procedures in Chapter 5 for selecting the optimal variance/covariance structure from a collection of candidates. One is based on the information ratio tests proposed in Chapter 4, and the other is based on an "information discrepancy criterion" which measures the discrepancy between two forms of information matrices under a general variance/covariance structure. Two simulation experiments are conducted to assess the detection rate of these two model selection methods. In addition, two data sets are analyzed using these two model selection methods to choose the most appropriate variance/covariance structure.

Chapter 6 gives a summary and discussion of future work.

Chapter 2

Estimating Functions

2.1 Preliminaries

Let y be a point of the sample space $\mathcal{Y}$, on which a measure μ is defined. Furthermore, let $p(y, \boldsymbol{\theta})$ denote the probability density function w.r.t. μ from a family of parametric statistical models, which is completely specified for all $y \in \mathcal{Y}$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. An estimating function is a function of the form $\Psi(y, \boldsymbol{\theta})$, which contains at least p independent components. When the dimension of the estimating function is larger than p, according to [40], the parameter $\boldsymbol{\theta}$ is said to be over-identified. In this thesis, we consider only the regular case of non-over-identification. The following definitions are given in [77].

Definition 2.1 (Estimating functions) *A function $\Psi : \mathcal{Y} \times \Theta \to \mathbb{R}^p$ is called an estimating function (or inference function) if $\Psi(\cdot; \boldsymbol{\theta})$ is measurable for any $\boldsymbol{\theta} \in \Theta$ and $\Psi(y; \cdot)$ is continuous in a compact subspace of Θ containing the true parameter $\boldsymbol{\theta}_0$ for any sample $y \in \mathcal{Y}$.*

Given an estimating function Ψ and a single observation $y \in \mathcal{Y}$, an estimating equation can be defined by

$$\Psi(y; \boldsymbol{\theta}) = 0,$$

and as a result, an estimate $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(y)$ of parameter $\boldsymbol{\theta}$ is obtained as an solution to this equation.

Definition 2.2 (Equivalent estimating functions) *Let Ψ_1 and Ψ_2 be any two estimating functions. If they give the same estimate of $\boldsymbol{\theta}$, or equivalently, they lead to the same solution to the equations $\Psi_1(y;\boldsymbol{\theta}) = 0$ and $\Psi_2(y;\boldsymbol{\theta}) = 0$ for any given sample $y \in \mathcal{Y}$, Ψ_1 and Ψ_2 are said to be equivalent, denoted by $\Psi_1 \sim \Psi_2$.*

It turns out that with any estimating function Ψ_0, we can construct a class of equivalent estimating functions $\{\Psi : \Psi(y;\boldsymbol{\theta}) = K(\boldsymbol{\theta})\Psi_0(y;\boldsymbol{\theta})\}$, where $K(\boldsymbol{\theta})$ is a $p\times p$ matrix of full rank and independent of observation y.

Definition 2.3 (Unbiased estimating functions) *An estimating function Ψ is said to be unbiased if it has expectation zero in the sense that*

$$E_{\boldsymbol{\theta}}\{\Psi(Y;\boldsymbol{\theta})\} = \mathbf{0},\ \forall\ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p.$$

Consider a sample of observations $\mathbf{y} = (y_1, \cdots, y_n)^T$, independently drawn from a parametric statistical model with the probability density function $p(y;\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is the true value of the parameter $\boldsymbol{\theta}$. An additive estimating function is defined by

$$\Psi_n(\mathbf{y};\boldsymbol{\theta}) = \sum_{i=1}^{n} \Psi\left(y_i;\boldsymbol{\theta}\right).$$

Note that if Ψ is unbiased, the additive estimating function Ψ_n is also unbiased. Consequently, an estimate $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n(\mathbf{y})$ of the parameter $\boldsymbol{\theta}$ is obtained by solving the equation

$$\Psi_n(\mathbf{y};\boldsymbol{\theta}) = \sum_{i=1}^{n} \Psi\left(y_i;\boldsymbol{\theta}\right) = \mathbf{0}. \tag{2.1}$$

We are interested in defining a class of unbiased estimating functions, within which the resultant estimators $\widehat{\boldsymbol{\theta}}_n$, of parameter $\boldsymbol{\theta}$ have desirable properties, such as consistency and asymptotic normality. Let us begin with a simple case where the parameter θ is one-dimensional (p=1).

Definition 2.4 *An estimating function Ψ is said to be regular, if it satisfies the following conditions:*

(i) $E_\theta\{\Psi(Y;\theta)\} = 0$, $\forall\ \theta \in \Theta \subset \mathbb{R}$;

(ii) $\partial\Psi(y;\theta)/\partial\theta$ *exists,* $\forall\ y \in \mathcal{Y}$;

(iii) The order of integration and differentiation may be interchanged

$$\frac{\partial}{\partial\theta}\int_{\mathcal{Y}} f(y)\Psi(y;\theta)p(y;\theta)dy = \int_{\mathcal{Y}} f(y)\frac{\partial}{\partial\theta}\{\Psi(y;\theta)p(y;\theta)\}\,dy$$

for any bounded measurable function $f(y)$ that is independent of θ;

(iv) $0 < E_\theta\left\{\Psi^2(Y;\theta)\right\} < \infty$;

(v) $0 < \{E_\theta|\partial\Psi(Y;\theta)/\partial\theta|\}^2 = \left\{E_\theta\left|\dot{\Psi}(Y;\theta)\right|\right\}^2 < \infty$, *where $\dot{\Psi}$ denotes the first-order derivative of the estimating function Ψ w.r.t. θ, that is, $\dot{\Psi}(y;\theta) = \partial\Psi(y;\theta)/\partial\theta$.*

Let $\mathfrak{G}$ denote the class of all the regular estimating functions. Given an estimating function $\Psi \in \mathfrak{G}$, let $\Psi'(y;\theta) = k(\theta)\Psi(y;\theta)$, where $k(\theta) \neq 0$ is a differentiable function of θ in Θ. By the definition 2.2, Ψ' is an equivalent estimating function to Ψ. However, the variance of the estimating function Ψ', given by

$$Var_\theta\{\Psi'(Y;\theta)\} = k^2(\theta)Var_\theta\{\Psi(Y;\theta)\},$$

could be substantially smaller or larger than that of Ψ, depending on the choice of $k(\theta)$. Thus, it is not reasonable to compare two estimating functions on the basis of variance alone. [35] proposed the standardization of estimating functions before comparing their variances. For any estimating function $\Psi \in \mathfrak{G}$, the standardized version of Ψ is defined by

$$\Psi_s = \Psi/E_\theta\left\{\dot{\Psi}(Y;\theta)\right\}.$$

Then, the standardized versions, Ψ_s and Ψ'_s, of equivalent estimating functions Ψ and Ψ' are equivalent, and consequently, they have the same variance, i.e., $Var_\theta\{\Psi'_s(Y;\theta)\} = Var_\theta\{\Psi_s(Y;\theta)\}$, or equivalently,

$$\frac{Var_\theta\{\Psi'(Y;\theta)\}}{\left[E_\theta\left\{\dot{\Psi}'(Y;\theta)\right\}\right]^2} = \frac{Var_\theta\{\Psi(Y;\theta)\}}{\left[E_\theta\left\{\dot{\Psi}(Y;\theta)\right\}\right]^2}.$$

As a result, the variance of the standardized estimating functions is unique among equivalent estimating functions, and its inverse is referred to as *the Godambe information* in the following definition.

Definition 2.5 (Godambe information) *For a regular estimating function $\Psi \in \mathfrak{G}$ and a single observation $Y \in \mathcal{Y}$,*

(i) the sensitivity*, denoted by $\boldsymbol{S}_\Psi$, of Ψ is defined as*

$$\boldsymbol{S}_\Psi(\theta) = E_\theta\left\{\dot{\Psi}(Y;\theta)\right\} = E_\theta\left\{\frac{\partial\Psi(Y;\theta)}{\partial\theta}\right\}, \theta \in \Theta;$$

(ii) the variability*, denoted by $\boldsymbol{V}_\Psi$, of Ψ is defined as*

$$\boldsymbol{V}_\Psi(\theta) = E_\theta\left\{\Psi^2(Y;\theta)\right\} = Var_\theta\{\Psi(Y;\theta)\}, \theta \in \Theta;$$

(iii) the Godambe information*, denoted by $\boldsymbol{J}_\Psi$, of Ψ is defined as*

$$\boldsymbol{J}_\Psi(\theta) = \frac{\boldsymbol{S}^2_\Psi(\theta)}{\boldsymbol{V}_\Psi(\theta)}, \quad \theta \in \Theta. \tag{2.2}$$

A regular estimating function Ψ_1 is said to be at least as good as the regular estimating function Ψ_2 if

$$\frac{Var_\theta(\Psi_1)}{\left[E_\theta\left(\dot{\Psi}_1\right)\right]^2} \leq \frac{Var_\theta(\Psi_2)}{\left[E_\theta\left(\dot{\Psi}_2\right)\right]^2},$$

or equivalently,

$$\mathbf{J}_{\Psi_1}(\theta) \geq \mathbf{J}_{\Psi_2}(\theta).$$

Moreover, from Theorem 2.2 in Section 2.2, we note that, for large samples, the asymptotic mean square error of the resultant estimator $\widehat{\theta_n}$ is equal to $\{n\mathbf{J}_\Psi(\theta_0)\}^{-1}$, where θ_0 is the true value of the parameter θ. Then, the larger the Godambe information is, the more efficient the resultant estimator is. In other words, a desirable estimating function should have large sensitivity ($\Psi(y; \theta_0 + \delta\theta_0)$ should be as far away from zero as possible) and small variability ($\Psi(y; \theta_0)$ should be as close to zero as possible).

Definition 2.6 *A statistical model is said to be regular if its score function*

$$u\,(y; \theta) = \partial \log p(y; \theta)/\partial\theta$$

is a regular estimating function; that is, $u(y; \theta) \in \mathfrak{G}, \theta \in \Theta \subset \mathbb{R}$. Moreover, for a regular model and a single observation $Y \in \mathcal{Y}$, the Fisher information *is defined by*

$$\boldsymbol{I}(\theta) = -E_\theta \left\{ \frac{\partial^2 \log p(Y; \theta)}{\partial\theta^2} \right\} = -E_\theta \left\{ \frac{\partial u\,(Y; \theta)}{\partial\theta} \right\}.$$

For a regular score function $u(y, \theta)$, the Bartlett identity implies that

$$E_\theta \left(\frac{\partial u}{\partial\theta} \right) + E_\theta(u^2) = 0,$$

or equivalently,

$$\mathbf{S}_u(\theta) + \mathbf{V}_u(\theta) = 0.$$

In other words, the equality of Godambe information and Fisher information $\mathbf{J}_u(\theta) = \mathbf{I}(\theta)$ holds for the regular score function.

Now let us consider the case where the parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T$ is a p-dimensional vector. Similarly to univariate estimating functions, a regular multi-dimensional estimating function is defined as follows. A p-element estimating

function

$$\Psi(y;\boldsymbol{\theta}) = \Big(\Psi_1(y;\boldsymbol{\theta}), \cdots, \Psi_p(y;\boldsymbol{\theta})\Big)^T$$

is said to be *regular* if it satisfies the following conditions:

(i) $E_{\boldsymbol{\theta}}\{\Psi(Y;\boldsymbol{\theta})\} = \mathbf{0}$, $\forall\ \boldsymbol{\theta} \in \Theta$;

(ii) $\frac{\partial}{\partial\theta_k}\Psi_j(y;\boldsymbol{\theta})$ exists, $\forall\ y \in \mathcal{Y}$, and $j,k = 1,\cdots,p$;

(iii) The order of integration and differentiation may be interchanged

$$\frac{\partial}{\partial\theta_k}\int_{\mathcal{Y}} f(y)\Psi_j(y;\boldsymbol{\theta})p(y;\boldsymbol{\theta})dy = \int_{\mathcal{Y}} f(y)\frac{\partial}{\partial\theta_k}\Big\{\Psi_j(y;\boldsymbol{\theta})p(y;\boldsymbol{\theta})\Big\}dy$$

for $j,k = 1,\cdots,p$, and any bounded measurable function $f(y)$ that is independent of $\boldsymbol{\theta}$;

(iv) $E_{\boldsymbol{\theta}}\Big\{\Psi_j(Y;\boldsymbol{\theta})\Psi_k(Y;\boldsymbol{\theta})\Big\}$ exists for $j,k = 1,\cdots,p$, and the $p\times p$ matrix

$$\mathbf{V}_\Psi(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\Big\{\Psi(Y;\boldsymbol{\theta})\Psi^T(Y;\boldsymbol{\theta})\Big\}$$

is positive-definite. $\mathbf{V}_\Psi(\boldsymbol{\theta})$ is called the *variability matrix*.

(v) $E_{\boldsymbol{\theta}}\Big\{\frac{\partial}{\partial\theta_k}\Psi_j(Y;\boldsymbol{\theta})\Big\}$ exists for $j,k = 1,\cdots,p$, and the $p\times p$ matrix

$$\mathbf{S}_\Psi(\boldsymbol{\theta}) = E_\theta\Big\{\nabla_{\boldsymbol{\theta}}\Psi(Y;\boldsymbol{\theta})\Big\}$$

is non-singular. $\mathbf{S}_\Psi(\boldsymbol{\theta})$ is called the *sensitivity matrix*.

Here the $\nabla_{\boldsymbol{\theta}}$ denotes the gradient operator on real function $f(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$, defined by

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \left(\frac{\partial f(\boldsymbol{\theta})}{\partial\theta_1},\cdots,\frac{\partial f(\boldsymbol{\theta})}{\partial\theta_p}\right)^T.$$

Let $\mathfrak{G}$ denote the class of all p-dimensional regular estimating functions. For a given regular estimating function $\Psi \in \mathfrak{G}$, the standardized version of Ψ is defined by

$$\Psi_s(y;\boldsymbol{\theta}) = \mathbf{S}_\Psi^{-1}(\boldsymbol{\theta})\Psi(y;\boldsymbol{\theta}),$$

and the Godambe information matrix then takes the form

$$\mathbf{J}_{\Psi}(\boldsymbol{\theta}) = \mathbf{S}_{\Psi}^{T}(\boldsymbol{\theta})\mathbf{V}_{\Psi}^{-1}(\boldsymbol{\theta})\mathbf{S}_{\Psi}(\boldsymbol{\theta}). \tag{2.3}$$

Note that the inverse of the Godambe information also equals to the covariance matrix of the standardized estimating function Ψ_s. Moreover, similarly to comparing univariate estimating functions, the regular p-element estimating function $\Psi_1 \in \mathfrak{G}$ is said to be at least as good as the p-element regular estimating function $\Psi_2 \in \mathfrak{G}$, if $\mathbf{J}_{\Psi_1} - \mathbf{J}_{\Psi_2}$ is non-negative definite, denoted by $\mathbf{J}_{\Psi_1} \geq \mathbf{J}_{\Psi_2}$.

[12] showed that the following three inequalities are equivalent:

(i) matrix inequality: $\mathbf{J}_{\Psi_1} - \mathbf{J}_{\Psi_2}$ is non-negative definite;

(ii) trace inequality: $\mathrm{tr}\left(\mathbf{J}_{\Psi_1}\right) \geq \mathrm{tr}\left(\mathbf{J}_{\Psi_2}\right)$;

(iii) determinant inequality: $\left|\mathbf{J}_{\Psi_1}\right| \geq \left|\mathbf{J}_{\Psi_2}\right|$.

2.2 Properties

Consider a regular statistical model $p(y, \boldsymbol{\theta})$, and suppose that the true value of the parameter $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_0$. Let Ψ be a regular estimating function in $\mathfrak{G}$, and let $\left\{\widehat{\boldsymbol{\theta}}_n\right\}$ be a sequence of roots to the sequence of additive estimating equations (2.1):

$$\Psi_n(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \Psi\left(y_i; \boldsymbol{\theta}\right) = \mathbf{0}, \quad n = 1, 2, \cdots,$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ is a sample of independent observations from the statistical model $p(y, \boldsymbol{\theta})$. The properties of regular estimating functions have been discussed and summarized in [77]. Here we only list the main properties, and omit all the proofs of the following theorems shown in [77].

Theorem 2.1 (Consistency) *Let*

$$\lambda(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0}\{\Psi(X;\boldsymbol{\theta})\} = \int \Psi(y;\boldsymbol{\theta})p(y;\boldsymbol{\theta}_0)dy.$$

If $\lambda(\boldsymbol{\theta})$ has a unique zero at $\boldsymbol{\theta}_0$, then

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0, \quad \textit{under} \quad P_{\boldsymbol{\theta}_0}.$$

Let us first discuss other properties of regular estimating functions in $\mathfrak{G}$ for the case where the parameter θ is a scalar ($p = 1$).

Theorem 2.2 (Asymptotic Normality) *Suppose that the estimator, $\widehat{\theta}_n$, of parameter θ is consistent, namely*

$$\widehat{\theta}_n \xrightarrow{p} \theta_0, \quad \textit{under} \quad P_{\theta_0}.$$

Moreover, suppose that the second derivative of Ψ w.r.t. θ is bounded in the sense that there exist a constant c and a P_θ-measurable function $M(y)$ with finite expectation, i.e. $E_\theta\{M(Y)\} < \infty$, such that

$$\left|\ddot{\Psi}(y;\theta)\right| < M(y), \quad \textit{for} \quad \theta \in (\theta_0 - c, \theta_0 + c).$$

Then,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \boldsymbol{J}_\Psi^{-1}(\theta_0)\right), \quad \textit{under} \quad P_{\theta_0},$$

where $\boldsymbol{J}_\Psi^{-1}(\theta)$ is the Godambe information of Ψ given by (2.2).

Therefore, as mentioned in Section 2.1, the motivation of defining the class of all regular estimating functions $\mathfrak{G}$ is that the resultant estimator of parameter θ is consistent and asymptotically normally distributed. Among these regular estimating functions in class $\mathfrak{G}$, [35] defined an *optimal* estimating function as the one which maximizes the Godambe information. Thus the optimal estimating function produces an estimator with the smallest asymptotic variance and hence has the highest asymptotic efficiency.

Definition 2.7 (Optimal estimating function) *A regular estimating function* $\Psi^* \in \mathfrak{G}$ *is said to be an* optimal estimating function *if*

$$\boldsymbol{J}_{\Psi^*}(\theta) \geq \boldsymbol{J}_{\Psi}(\theta), \text{ for all } \Psi \in \mathfrak{G}, \text{ and } \theta \in \Theta \subset \mathbb{R}.$$

Moreover, [35] also showed that if the class $\mathfrak{G}$ includes the score function, the optimal estimating function is equivalent to the score function, which leads to maximum likelihood estimation.

Theorem 2.3 (Godambe Inequality) *Assume an estimating function* $\Psi \in \mathfrak{G}$. *Then*

$$\boldsymbol{J}_{\Psi}(\theta) \leq \boldsymbol{I}(\theta), \quad \forall \theta \in \Theta \subset \mathbb{R},$$

where the equality holds if and only if $\Psi \sim u$, *namely* Ψ *is equivalent to the score function.*

In general, let $\{\Psi_i(y_i; \theta); i = 1, \cdots, n\}$ be a set of *elementary* estimating functions belonging to the class $\mathfrak{G}$. Now define a special subclass $\mathfrak{G}_c$, of regular estimating functions in the following form of a linear combination of elementary estimating functions,

$$\Psi_c(\theta) = \sum_{i=1}^{n} c_i(\theta)\Psi_i(y_i; \theta), \quad \theta \in \Theta \subset \mathbb{R},$$

where $c_i(\theta)$ is a non-random function of θ. The resultant estimator of θ, obtained from solving the equation $\Psi_c(\theta) = 0$, is consistent. The subclass $\mathfrak{G}_c$ is referred to as the *Crowder class* of regular estimating functions. Furthermore, [17] obtained the optimal estimating function in class $\mathfrak{G}_c$.

Theorem 2.4 (Crowder Optimality) *Consider regular estimating functions* $\Psi_c \in \mathfrak{G}_c$. *Then, the optimal estimating function in the class* $\mathfrak{G}_c$, *the one which has the largest Godambe information, is the one with the* $c_i(\cdot)$ *functions being a ratio of the sensitivity over the variability, namely*

$$c_i(\theta) = \frac{E_\theta\left\{\dot{\Psi}_i(Y_i; \theta)\right\}}{Var_\theta\left\{\Psi_i(Y_i; \theta)\right\}} = \frac{\boldsymbol{S}_{\Psi_i}(\theta)}{\boldsymbol{V}_{\Psi_i}(\theta)}, \quad \theta \in \Theta \subset \mathbb{R}.$$

These properties of univariate regular estimating functions can also be generalized to the case with the p-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T$.

Theorem 2.5 (Multivariate Asymptotic Normality) *If $\widehat{\boldsymbol{\theta}}_n$ is consistent, and in a small neighborhood, $\mathcal{N}(\boldsymbol{\theta}_0)$, centered at the true value $\boldsymbol{\theta}_0$,*

$$\|\ddot{\Psi}(y; \boldsymbol{\theta})\| < M(y), \quad \boldsymbol{\theta} \in \mathcal{N}(\boldsymbol{\theta}_0),$$

with a $P_{\boldsymbol{\theta}}$-measurable function $M(y)$ such that $E_{\boldsymbol{\theta}}\{M(Y)\} < \infty$, then

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} MVN_p\left(\mathbf{0}, \boldsymbol{J}_\Psi^{-1}(\boldsymbol{\theta}_0)\right), \tag{2.4}$$

where $\boldsymbol{J}_\Psi^{-1}(\boldsymbol{\theta})$ is the Godambe information of Ψ given by (2.3).

Definition 2.8 (Multivariate optimal estimating function) *A regular estimating function $\Psi^* \in \mathfrak{G}$ is said to be an* optimal estimating function *if*

$$\boldsymbol{J}_{\Psi^*}(\boldsymbol{\theta}) \geq \boldsymbol{J}_\Psi(\boldsymbol{\theta}), \text{ for all } \Psi \in \mathcal{G}, \text{ and } \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p.$$

Theorem 2.6 (Multivariate Godambe Inequality) *Consider a regular estimating function $\Psi \in \mathfrak{G}$. Then*

$$\boldsymbol{J}_\Psi(\boldsymbol{\theta}) \preceq \boldsymbol{I}(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p,$$

where the equality holds if and only if $\Psi \sim \boldsymbol{u}$, the score function.

Define a Crowder class of regular estimating functions $\mathfrak{G}_c$ by

$$\Psi_c(\boldsymbol{y}; \boldsymbol{\theta}) = \sum_{i=1}^{n} C_i(\boldsymbol{\theta}) \Psi_i(y_i; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p,$$

where $C_i(\boldsymbol{\theta})$ is a non-random $p \times p$ matrix of $\boldsymbol{\theta}$ such that the sequence of roots, $\left\{\widehat{\boldsymbol{\theta}}_n, n \geq 1\right\}$, to the estimating equation $\Psi_c(\boldsymbol{y}; \boldsymbol{\theta}) = 0$ is consistent.

Theorem 2.7 (Multivariate Crowder Optimality) *Consider regular estimating functions $\Psi_c \in \mathfrak{G}_c$. Then, the optimal estimating function in the Crowder class $\mathfrak{G}_c$, the one which has the largest Godambe information, is the one with the matrix $C_i(\cdot)$ functions given by*

$$C_i(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}^T \left\{ \nabla_{\boldsymbol{\theta}} \Psi_i(Y_i; \boldsymbol{\theta}) \right\} Var_{\boldsymbol{\theta}}^{-1} \left\{ \Psi_i(Y_i; \boldsymbol{\theta}) \right\}, \quad \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p. \tag{2.5}$$

2.3 Quasi-likelihood Inference

The score function is the optimal estimating function among all the regular estimating functions. However, in practice, because the underlying mechanism is not fully understood or there is lack of previous informative experience of analyzing similar data, it is common that the probability density function from which the data are generated cannot be fully specified. In addition, in some cases, from some preliminary analysis, practitioners found that the parametric models proposed for the data analysis were violated, for example, due to overdispersion.

Usually, the main interest of data analyses attaches to how the response variables are affected by one or multiple explanatory variables. Often, it is natural for investigators to propose assumptions on some aspects of the probability mechanisms, such as the first two moments, instead of the full parametric distributions. [86] proposed an idea of quasi-likelihood estimation for regression coefficients in the setting of GLM. Also see [59].

2.3.1 Quasi-score equations in GLM for independent data

Consider a set of observations $\left\{ (y_i, \boldsymbol{x}_i^T), i = 1, \cdots, n \right\}$, independently drawn from a regular statistical model, where y_i is the response variable, and $\boldsymbol{x}_i$ is a $p \times 1$ vector of covariates. When the parametric model cannot be fully specified, part of the objectives in data analysis can be addressed by the following regression

model, specified only by the first two moments:

$$\mu_i = E(Y_i) = h(\boldsymbol{x}_i^T \boldsymbol{\beta}), \tag{2.6}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $h(\cdot)$ is the *link function*, which is assumed to be known and continuous; and

$$Var(Y_i) = \sigma^2 V(\mu_i), \tag{2.7}$$

where σ^2 is called the *dispersion parameter*, which is usually unknown in practice, and $V(\cdot)$ is called the *unit variance function*. Usually, the intercept term is included in the covariate vector $\boldsymbol{x}_i$'s. In the rest of the thesis, for each i, the covariate vector is a $p\times1$ vector with the first element 1, i.e., $\boldsymbol{x}_i = (1, x_{i1}, \cdots, x_{i,p-1})^T$, and the coefficient vector is denoted by $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})^T$.

Let $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ be an $n \times 1$ vector of response variables. To estimate the coefficient vector $\boldsymbol{\beta}$, it is suggested to solve a p-element additive estimating equation,

$$\Psi_n(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T \frac{y_i - \mu_i(\boldsymbol{\beta})}{\sigma^2 V(\mu_i)} = 0, \tag{2.8}$$

where $\Psi_n(\boldsymbol{\beta})$ is referred to as a *quasi-score function*. Correspondingly, it is possible to yield a function similar to the likelihood function in the MLE setting, called *quasi-likelihood*, by taking integration w.r.t. μ, that is,

$$l_q(y; \mu) = \int_y^{\mu} \frac{y - t}{V(t)} dt.$$

For $i = 1, \cdots, n$, define $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$, and then, rewrite $\partial \mu_i(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ as

$$\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \dot{\mu}_i \boldsymbol{x}_i,$$

where $\dot{\mu}_i$ stands for the first order derivative of μ_i w.r.t. η_i. For the additive quasi-score function $\Psi_n(\boldsymbol{\beta}; \boldsymbol{y})$ (2.8), the aggregated sensitivity matrix is given by

$$\mathbf{S}_{\Psi_n}(\boldsymbol{\beta}) = E_{\boldsymbol{\beta}} \left\{ \frac{\partial \Psi_n(\boldsymbol{\beta}; \boldsymbol{Y})}{\partial \boldsymbol{\beta}} \right\} = - \sum_{i=1}^{n} \frac{\dot{\mu}_i^2}{\sigma^2 V(\mu_i)} \boldsymbol{x}_i \boldsymbol{x}_i^T, \tag{2.9}$$

and the aggregated variability matrix is given by

$$\mathbf{V}_{\Psi_n}(\boldsymbol{\beta}) = E_{\boldsymbol{\beta}}\left\{\Psi_n(\boldsymbol{\beta};\boldsymbol{Y})\Psi(\boldsymbol{\beta};\boldsymbol{Y})^T\right\} = \sum_{i=1}^{n} \frac{\dot{\mu}_i^2 Var(Y_i)}{(\sigma^2)^2\, V^2(\mu_i)} \boldsymbol{x}_i \boldsymbol{x}_i^T. \tag{2.10}$$

The aggregated Godambe information matrix is given by

$$\mathbf{J}_{\Psi_n}(\boldsymbol{\beta}) = \mathbf{S}_{\Psi_n}^T(\boldsymbol{\beta})\mathbf{V}_{\Psi_n}^{-1}(\boldsymbol{\beta})\mathbf{S}_{\Psi_n}(\boldsymbol{\beta}). \tag{2.11}$$

Let $\widehat{\boldsymbol{\beta}}_n$ denote the estimator of the parameter $\boldsymbol{\beta}$ obtained by solving the quasi-score equation $\Psi_n(\boldsymbol{\beta};\boldsymbol{y}) = 0$. Under some mild regularity conditions in Theorem 2.1 and Theorem 2.5, the estimator $\widehat{\boldsymbol{\beta}}_n$ is consistent, and $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_*)$ is asymptotically multivariate Gaussian distributed with zero mean and covariance matrix of the form $\lim_n n\mathbf{J}_{\Psi_n}^{-1}(\boldsymbol{\beta}_*)$, where $\boldsymbol{\beta}_*$ is the true value of the parameter $\boldsymbol{\beta}$.

In order to use this result to make inference about the parameter $\boldsymbol{\beta}$, for example, constructing confidence intervals, an estimated Godambe information matrix with $\boldsymbol{\beta}_*$ replaced with $\widehat{\boldsymbol{\beta}}_n$ is usually used to obtain an estimate of the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_n$.

Let us first give some necessary notation which will be used in the rest of the thesis. Let X denote an $n \times p$ matrix, referred to as the *design matrix*, with the i-th row given by $\boldsymbol{x}_i^T$. For each i, with the estimator $\widehat{\boldsymbol{\beta}}_n$, the fitted values are defined by $\widehat{\mu}_i = \mu_i(\widehat{\boldsymbol{\beta}}_n) = h(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_n)$. Then, the raw residuals are defined by $r_i = y_i - \widehat{\mu}_i$, and the Pearson residuals are given by

$$r_{P,i} = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}, \quad i = 1, \cdots, n, \tag{2.12}$$

the standardized residuals. Let

$$\widehat{\Delta} = \operatorname{diag}\left\{\widehat{\dot{\mu}}_1, \cdots, \widehat{\dot{\mu}}_n\right\} \tag{2.13}$$

be an $n \times n$ diagonal matrix, where $\widehat{\dot{\mu}}_i = \dot{\mu}_i(\widehat{\boldsymbol{\beta}}_n)$, and let

$$\widehat{\mathcal{V}} = \operatorname{diag}\left\{V(\widehat{\mu}_1), \cdots, V(\widehat{\mu}_n)\right\} \tag{2.14}$$

be an $n \times n$ diagonal matrix, and

$$\mathcal{R} = diag\left\{r_1^2, \cdots, r_n^2\right\}. \tag{2.15}$$

Then, based on the data $\{(y_i, \boldsymbol{x}_i); i = 1, \cdots, n\}$, the sensitivity matrix (2.9) is estimated by

$$\begin{aligned} \widehat{\mathbf{S}}_{\Psi_n}(\widehat{\boldsymbol{\beta}}_n) &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{\widehat{\dot{\mu}}_i^2}{V(\widehat{\mu}_i)}\boldsymbol{x}_i\boldsymbol{x}_i^T \\ &= -\frac{1}{\sigma^2}X^T\widehat{\Delta}\widehat{\mathcal{V}}^{-1}\widehat{\Delta}X, \end{aligned} \tag{2.16}$$

when the true value of the dispersion parameter σ^2 is known. However, in practice, the true value is rarely known, so the dispersion parameter σ^2 is estimated by, for example, a moment estimator

$$\widehat{\sigma}_m^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_{P,i}^2. \tag{2.17}$$

In addition, the variability matrix (2.10) is estimated by

$$\begin{aligned} \widehat{\mathbf{V}}_{\Psi_n}(\widehat{\boldsymbol{\beta}}_n) &= \frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}\frac{\widehat{\dot{\mu}}_i^2}{V^2(\widehat{\mu}_i)}(y_i - \widehat{\mu}_i)^2\boldsymbol{x}_i\boldsymbol{x}_i^T \\ &= \frac{1}{(\sigma^2)^2}X^T\widehat{\Delta}\widehat{\mathcal{V}}^{-1}\mathcal{R}\widehat{\mathcal{V}}^{-1}\widehat{\Delta}X. \end{aligned} \tag{2.18}$$

Consequently, the Godambian information matrix is estimated by

$$\widehat{\mathbf{J}}_{\Psi_n}(\widehat{\boldsymbol{\beta}}_n) = \left\{-\widehat{\mathbf{S}}_{\Psi_n}(\widehat{\boldsymbol{\beta}}_n)\right\}^T\left\{\widehat{\mathbf{V}}_{\Psi_n}(\widehat{\boldsymbol{\beta}}_n)\right\}^{-1}\left\{-\widehat{\mathbf{S}}_{\Psi_n}(\widehat{\boldsymbol{\beta}}_n)\right\}.$$

Its inverse can be an estimator of the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_n$, which is given by

$$\left\{\widehat{\mathbf{J}}_{\Psi_n}(\widehat{\boldsymbol{\beta}}_n)\right\}^{-1} = \left\{X^T\widehat{\Delta}\widehat{\mathcal{V}}^{-1}\widehat{\Delta}X\right\}^{-1}\left\{X^T\widehat{\Delta}\widehat{\mathcal{V}}^{-1}\mathcal{R}\widehat{\mathcal{V}}^{-1}\widehat{\Delta}X\right\}\left\{X^T\widehat{\Delta}\widehat{\mathcal{V}}^{-1}\widehat{\Delta}X\right\}^{-T}. \tag{2.19}$$

Because of its unique sandwich structure, this covariance matrix estimator has been referred to as the "*sandwich covariance matrix estimator*" in [53]'s paper on generalized estimating equations. Moreover, its diagonal elements can be used to estimate the asymptotic variances of individual regression coefficient estimator $\widehat{\beta}_j$'s. We will call these estimators the *sandwich covariance matrix estimator*, denoted by $ASCOV_s(\widehat{\boldsymbol{\beta}}_n)$, or *sandwich variance estimators*, denoted by $ASVAR_s(\widehat{\beta}_j)$, in the rest of the thesis.

There are several appealing features associated with the utility of the quasi-score equation for the estimator $\widehat{\boldsymbol{\beta}}_n$. First of all, the quasi-score function only requires the assumptions about the first two moments without specifying the explicit form of the underlying parametric models. Secondly, it does not need to estimate the dispersion parameter σ^2, because this parameter is a constant which can be canceled out in the equation $\Psi_n(\boldsymbol{\beta}; \boldsymbol{y}) = 0$, as well as in the calculation of the sandwich covariance matrix estimator (2.19).

Furthermore, if the two assumptions (2.6) and (2.7) correctly specify the true mean and variance structures of the responses, the quasi-score function preserves the two key properties of the real score function. They are:

(1) The quasi-score function is unbiased in the sense that $E_{\boldsymbol{\beta}}\{\Psi(\boldsymbol{\beta}; \boldsymbol{Y})\} = 0$ if the mean structure is correctly specified. The unbiasedness of the quasi-score function guarantees the consistency of the resultant estimator $\widehat{\boldsymbol{\beta}}_n$;

(2) Consider a vector of elementary estimating functions $\Psi_i(\boldsymbol{\beta}; y_i) \in \mathfrak{G}$, defined by

$$\Psi_i(\boldsymbol{\beta}; y_i) = y_i - \mu_i(\boldsymbol{\beta}). \tag{2.20}$$

By the Crowder Optimality Theorem 2.7, in the Crowder class $\mathfrak{G}_c$ defined by

$$\Psi_c(\boldsymbol{\beta}) = \sum_{i=1}^{n} C_i(\boldsymbol{\beta})\Psi_i(\boldsymbol{\beta}; y_i),$$

the optimal estimating function is given by

$$\begin{aligned}\Psi^*(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{S}_{\Psi_i}^T(\boldsymbol{\beta})\mathbf{V}_{\Psi_i}^{-1}(\boldsymbol{\beta})\left(y_i - \mu_i(\boldsymbol{\beta})\right) \\ &= -\sum_{i=1}^{n}\left(\frac{\partial\mu_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)^T \frac{y_i - \mu_i(\boldsymbol{\beta})}{Var(Y_i)}.\end{aligned}$$

Thus, if the assumption of the second moment (2.7) is correct, the quasi-score function (2.8) is actually equivalent to the optimal estimating function among the Crowder class of the elementary estimating functions (2.20). Moreover, the negative sensitivity matrix (2.9) and the variability matrix (2.10) are equivalent. In other words, the following identity

$$E_{\boldsymbol{\beta}}\left\{-\frac{\partial\Psi_n(\boldsymbol{\beta};Y)}{\partial\boldsymbol{\beta}}\right\} = Var_{\boldsymbol{\beta}}\left\{\Psi_n(\boldsymbol{\beta};Y)\right\}$$

holds if the unit variance function $V(\cdot)$ correctly specifies the true variance structure of the response variables. This equality ensures that the asymptotic covariance matrix of the resultant estimator $\widehat{\boldsymbol{\beta}}_n$ equals to that of the most efficient estimator, i.e., the GLM analogue of the inverse of Fisher information matrix. If the underlying distribution belongs to the exponential family, even though explicit parametric models are not assumed, the estimator $\widehat{\boldsymbol{\beta}}_n$ can achieve the same estimation efficiency as that of the MLE, which is the fully efficient estimator.

As a result, the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_n$ can be estimated by

$$\left\{-\widehat{\mathbf{S}}_{\Psi_n}(\widehat{\boldsymbol{\beta}}_n)\right\}^{-1} = \sigma^2\left(X^T\widehat{\Delta}\widehat{\mathcal{V}}^{-1}\widehat{\Delta}X\right)^{-1}, \tag{2.21}$$

if the true value of the dispersion parameter σ^2 is known; otherwise,

$$\left\{-\widehat{\mathbf{S}}_{\Psi_n}(\widehat{\boldsymbol{\beta}}_n)\right\}^{-1} = \widehat{\sigma}_m^2\left(X^T\widehat{\Delta}\widehat{\mathcal{V}}^{-1}\widehat{\Delta}X\right)^{-1}, \tag{2.22}$$

where $\widehat{\sigma}_m^2$ is the moment estimator of σ^2. This covariance matrix estimator has been referred to as the "*model-based covariance matrix estimator*", denoted by $ASCOV_m(\widehat{\boldsymbol{\beta}}_n)$. Consequently, its diagonal elements are regarded as the asymptotic variance estimators of individual regression coefficient estimators. Here, they are called the *model-based variance estimators*, denoted by $ASVAR_m(\widehat{\beta}_j)$. Therefore, confidence intervals constructed from the sandwich and model-based variance estimators are called the sandwich confidence intervals and the model-based confidence intervals, respectively.

Example 2.1 (Linear Regression Models) LM for continuous data is regarded as a special case of GLM. The first moment assumption is

$$\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad i = 1, \cdots, n,$$

where the link function $h(\cdot)$ is an identity function, i.e., $h(\eta_i) = \eta_i$. The error terms are usually assumed to be independent random variables with mean 0 and constant variance σ^2. The second moment assumption is

$$Var(Y_i) = \sigma^2, \quad i = 1, \cdots, n,$$

where the unit variance function $V(\mu_i) = 1$.

Least squares (LS) estimation provides an estimator of the regression coefficient vector $\boldsymbol{\beta}$, given by

$$\widehat{\boldsymbol{\beta}}_n = \left(X^T X\right)^{-1} X^T \boldsymbol{y}, \tag{2.23}$$

by solving the estimating equation

$$\Psi_n(\boldsymbol{\beta}; \boldsymbol{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) = 0.$$

Note that the quasi-score function $\Psi_n(\boldsymbol{\beta})$ is equivalent to the real score function if the responses $\boldsymbol{y} = (y_1, \cdots, y_n)$ are independent observations from a normal distribution with the probability density function, given by

$$p(y_i, \boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2}\right\}, \quad i = 1, \cdots, n.$$

In this case, the LS estimator $\widehat{\boldsymbol{\beta}}_n$ (2.23) is the MLE.

With the estimator $\widehat{\boldsymbol{\beta}}_n$, the sandwich covariance matrix estimator is given by

$$ASCOV_s(\widehat{\boldsymbol{\beta}}_n) = \left(X^T X\right)^{-1} \left(X^T \mathcal{R} X\right) \left(X^T X\right)^{-1}, \tag{2.24}$$

where the matrix $\mathcal{R}$ is given in (2.15). In addition, the model-based covariance matrix estimator is given by

$$ASCOV_m(\widehat{\boldsymbol{\beta}}_n) = \widehat{\sigma}^2_m \left(X^T X\right)^{-1}, \tag{2.25}$$

where $\widehat{\sigma}^2_m$ is a moment estimator of the dispersion parameter σ^2, if the true value is unknown, given by

$$\widehat{\sigma}^2_m = \frac{1}{n-p} \sum_{i=1}^{n} r_i^2. \tag{2.26}$$

Note that in LM, the Pearson residuals reduce to the raw residuals. That is, $r_{P,i} = r_i$, for $i = 1, \cdots, n$.

2.3.2 Generalized estimating equations (GEE) in longitudinal data analysis

Longitudinal data is a data type frequently encountered in many subject-matter areas such as biology, medical and public health sciences and social science. Since the defining feature of longitudinal data is that the measurements of the same individuals are taken repeatedly over a period of time, the primary interest of longitudinal data analysis lies in the mechanism of change over time, including growth, aging, time profiles or effects of covariates.

In most cases, the maximum likelihood inference is either unavailable or numerically too intricate to be implemented. One of the popular methods is the generalized estimating equations (GEE) approach proposed by [53], which does not require us to specify a complete probability model. In fact, the GEE method

can be viewed as a multivariate extension of the quasi-likelihood method proposed by [86], which only requires us to correctly specify the first two moments of the underlying data distribution.

Consider a longitudinal data set denoted by

$$(y_{ij}, \boldsymbol{x}_{ij}^T, t_{ij}),\ j = 1, \cdots, n_i, i = 1, \cdots, K,$$

where y_{ij} is the response variable and $\boldsymbol{x}_{ij}$ is a set of covariate variables observed at the j-th time point t_{ij} for subject i, and n_i is the number of measurements for the subject i. In total, there are K subjects in the data set. Let $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{i,n_i})^T$, and $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{i,n_i})^T$, for $i = 1, \cdots, K$. Assume that the first two moments of the response vector $\boldsymbol{y}_i$ are given by

$$E(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i, \quad \text{with} \quad \mu_{ij} = \mu_{ij}(\boldsymbol{\beta}) = h(\boldsymbol{x}_{ij}^T \boldsymbol{\beta})$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients and $h(\cdot)$ is the link function; and

$$Cov(\boldsymbol{Y}_i) = \sigma^2 \Sigma_i = \sigma^2 G_i^{1/2} R_i(\boldsymbol{\rho}) G_i^{1/2},$$

where σ^2 is the dispersion parameter, G_i is an $n_i \times n_i$ diagonal matrix given by

$$G_i = \text{diag}\left\{V(\mu_{i1}), \cdots, V(\mu_{i,n_i})\right\},$$

with $V(\cdot)$ being the unit variance function, and $R_i(\boldsymbol{\rho})$ is an $n_i \times n_i$ correlation matrix that is fully characterized by a q-dimensional correlation parameter vector $\boldsymbol{\rho}$. This $R_i(\boldsymbol{\rho})$ is referred to as a *working correlation matrix*. A p-element estimating function is given by

$$\Psi_K(\boldsymbol{\beta}, \boldsymbol{\rho}) = \frac{1}{\sigma^2} \sum_{i=1}^{K} D_i^T(\boldsymbol{\beta}) \Sigma_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\rho}) \left(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right) = \boldsymbol{0}, \tag{2.27}$$

where $D_i^T(\boldsymbol{\beta}) = X_i^T \text{diag}\left\{\dot{h}(\mu_{i1}), \cdots, \dot{h}(\mu_{i,n_i})\right\}$, with X_i being an $n_i \times p$ design matrix for the subject i. The equation $\Psi_K(\boldsymbol{\beta}) = 0$ is termed as *the generalized estimating*

equation (GEE) by [53], where the nuisance parameter $\boldsymbol{\rho}$ is involved, and the dispersion parameter σ^2 is factorized out of the equation (2.27). Consequently, the estimator, $\widehat{\boldsymbol{\beta}}_K$, of parameter $\boldsymbol{\beta}$ is obtained as the solution to the GEE (2.27). [53] showed that the estimators of these coefficients are consistent and asymptotically normal even when the correlation structure is incorrectly specified.

For the estimating function $\Psi_K(\boldsymbol{\beta})$, the sensitivity and variability matrices are obtained, respectively, as

$$\mathbf{S}_{\Psi_K}(\boldsymbol{\beta}) = -\frac{1}{\sigma^2}\sum_{i=1}^{K} D_i^T \Sigma_i^{-1} D_i$$

and

$$\mathbf{V}_{\Psi_K}(\boldsymbol{\beta}) = \frac{1}{(\sigma^2)^2}\sum_{i=1}^{K} D_i^T \Sigma_i^{-1} Cov(\boldsymbol{Y}_i)\Sigma_i^{-1} D_i.$$

Hence the Godambe information matrix is given by

$$\begin{aligned}\mathbf{J}_{\Psi_K}(\boldsymbol{\beta}) &= \left\{-\mathbf{S}_{\Psi_K}(\boldsymbol{\beta})\right\}^{-1}\left\{\mathbf{V}_{\Psi_K}(\boldsymbol{\beta})\right\}\left\{-\mathbf{S}_{\Psi_K}(\boldsymbol{\beta})\right\}^{-1} \\ &= \left\{\sum_{i=1}^{K} D_i^T \Sigma_i^{-1} D_i\right\}\left\{\sum_{i=1}^{K} D_i^T \Sigma_i^{-1} Cov(\boldsymbol{Y}_i)\Sigma_i^{-1} D_i\right\}^{-1}\left\{\sum_{i=1}^{K} D_i^T \Sigma_i^{-1} D_i\right\}.\end{aligned}$$

When the covariance structure, including the variance function involved in G_i and the working correlation structure $R_i(\boldsymbol{\rho})$, is correctly specified, the following information matrix equality holds:

$$\mathbf{S}_{\Psi_K}(\boldsymbol{\beta}) + \mathbf{V}_{\Psi_K}(\boldsymbol{\beta}) = 0, \quad \text{or} \quad \mathbf{J}_{\Psi_K}(\boldsymbol{\beta}) = -\mathbf{S}_{\Psi_K}(\boldsymbol{\beta}).$$

Let $\widehat{D}_i$ and $\widehat{\Sigma}_i$ be the matrices which use the estimates $\widehat{\boldsymbol{\beta}}_K$ and $\widehat{\boldsymbol{\rho}}_K$ in the matrices D_i and Σ_i. The residuals are defined by

$$r_{ij} = y_{ij} - \widehat{\mu}_{ij} = y_{ij} - \mu_{ij}(\widehat{\boldsymbol{\beta}}_K), \quad j = 1, \cdots, n_i, i = 1, \cdots, K.$$

Let $\boldsymbol{r}_i = (r_{i1}, \cdots, r_{i,n_i})^T$ be an $n_i \times 1$ vector of the residuals for subject i, for $i = 1, \cdots, K$. Then, the sandwich covariance matrix estimator of the parameter

estimator $\widehat{\boldsymbol{\beta}}_K$ is given by

$$ASCOV_s(\widehat{\boldsymbol{\beta}}_K) = \left\{\sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}_i^{-1} \widehat{D}_i\right\}^{-1} \left\{\sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}_i^{-1} \boldsymbol{r}_i \boldsymbol{r}_i^T \widehat{\Sigma}_i^{-1} \widehat{D}_i\right\} \left\{\sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}_i^{-1} \widehat{D}_i\right\}^{-1}. \tag{2.28}$$

Moreover, the model-based covariance matrix estimator is given by

$$ASCOV_m(\widehat{\boldsymbol{\beta}}_K) = \sigma^2 \left\{\sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}_i^{-1} \widehat{D}_i\right\}^{-1}, \tag{2.29}$$

if the true value of the dispersion parameter σ^2 is known; otherwise,

$$ASCOV_m(\widehat{\boldsymbol{\beta}}_K) = \widehat{\sigma}_m^2 \left\{\sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}_i^{-1} \widehat{D}_i\right\}^{-1}, \tag{2.30}$$

where $\widehat{\sigma}_m^2$ is a moment estimator of the dispersion parameter σ^2, suggested by [53],

$$\widehat{\sigma}_m^2 = \frac{1}{N-p} \sum_{i=1}^{K} \boldsymbol{r}_{P,i}^T \boldsymbol{r}_{P,i} = \frac{1}{N-p} \sum_{i=1}^{K} \sum_{j=1}^{n_i} r_{P,ij}^2, \tag{2.31}$$

where $N = \sum_{i=1}^{K} n_i$.

2.4 Robustness of Sandwich Variance Estimators

The covariance matrix estimator in LM (2.24) has been referred to as the *heteroscedasticity-consistent covariance matrix estimator*, originally introduced by [45] and [88]. In the context of LM, this method provides a consistent covariance matrix estimator even when the errors of the LM are heteroscedastic. In the context of GEE for correlated data, efficient estimation for parameters of interest

requires correct specification of the correlation structure among the observations, which is, however, typically unknown. Therefore, a so-called working correlation structure is employed in point estimation. The sandwich estimator yields a consistent estimate of the covariance matrix under a misspecified working correlation matrix as well as under heteroscedastic errors. In both of these two settings, the sandwich method provides asymptotically consistent estimates of the covariance matrix for parameter estimators when the distributional assumptions fail to hold or are not specified. Due to these two desirable model-robustness properties, the sandwich covariance matrix estimator is also called the *robust covariance matrix estimator* or the *empirical covariance matrix estimator*.

[50] commented that the argument in favor of the sandwich estimate is that asymptotic normality and asymptotic coverage of confidence intervals require only a consistent variance estimate, so there is no direct need to construct a highly accurate covariance matrix estimate. But the consistency of the sandwich variance estimate has its price in increased variability; that is, sandwich variance estimators generally have a larger variance than model-based classical variance estimates. In addition, the authors pointed out that under certain conditions when the model assumptions are correct, the sandwich estimator is often far more variable than the usual parametric variance estimates. The additional variability directly affects the coverage probability of confidence intervals constructed from the sandwich variance estimates, which is the price one pays to obtain consistency even when the parametric model fails. More discussions are given in [91], [9] and among others.

In the quasi-likelihood estimation for GLM, the sandwich covariance matrix estimators, such as (2.19) and (2.24), lead to a consistent estimation of the covariance matrix even when the variance structure in the second moment assumption (2.6) is incorrect. Note that under the misspecification of the first moment assumption (2.7), confidence intervals constructed from the sandwich estimators do not achieve the nominal coverage probability, because the estimators of the parameters of interest are inconsistent.

2.4.1 Simulation experiments

In this section, we describe three simulation studies to investigate the coverage probabilities of confidence intervals (CIs) from the sandwich covariance matrix estimators and the model-based covariance matrix estimators under misspecifications of the variance structure or the mean structure.

Simulation 2.1 (LM - Heteroscedastic errors) A sample of observations

$$\{(y_i, x_i); i = 1, \cdots, n\}$$

is generated from the following simplest linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \cdots, n.$$

Here the covariate x_i's are i.i.d. observations from a Gaussian distribution with mean 0 and variance 1. In practice, it is common that the sampling errors show large variability when the observations drawn from the population are far from the center. Therefore, in this experiment, the error terms e_i are independently generated from a Gaussian distribution with mean 0 and heteroscedastic variance $Var(Y_i) = \sigma^2 h_{ii}$, where h_{ii} is the i-th diagonal element of the hat matrix H, defined as $H = X\left(X^T X\right)^{-1} X^T$. The true values of the two regression coefficients are $\beta_0 = 1$ and $\beta_1 = 2$, and the parameter σ^2 is set to be 100. We generate the data with different sample sizes $n = 20, 50, 100, 200, 500$. Note that the average variance $Var(Y_i)$ would be 10, 4, 2, 1, and 0.25, corresponding to each of the sample sizes.

Based on 1000 replicates, we show, in Table 2.1, the empirical coverage probabilities of the 95% sandwich CIs (95% CIs obtained from the sandwich variance estimates), and the 95% model-based CIs (95% CIs obtained from the model-based variance estimates), with different sample sizes $n = 20, 50, 100, 200, 500$. Table 2.2 reports the results of the empirical standard deviations of the coefficient estimates, denoted by SD_e, the average squared root of the sandwich variance

estimates, i.e. the average sandwich standard deviations, denoted by SD_s^a, and the average squared root of the model-based variance estimates, i.e. the average model-based standard deviations, denoted by SD_m^a.

Conclusion

As shown in Table 2.1, as the sample size increases, the empirical coverage probabilities of the sandwich CIs for each regression coefficient, β_0 and β_1, approach the nominal value 0.95. Compared to the sandwich CIs, the model-based CIs perform poorly in the sense that they cannot attain the nominal coverage probability, especially for the slope parameter β_1. Moreover, in Table 2.2, we find that the average sandwich standard deviations, SD_s^a, are closer to the empirical standard deviations SD_e than the average model-based standard deviations, SD_m^a.

Table 2.1: Empirical coverage probabilities of the 95% sandwich CIs and model-based CIs with different sample sizes.

	β_0		β_1	
sample size	model-based	sandwich	model-based	sandwich
20	0.925	0.911	0.820	0.852
50	0.938	0.937	0.819	0.903
100	0.951	0.948	0.820	0.926
200	0.953	0.949	0.830	0.939
500	0.943	0.942	0.836	0.954

Simulation 2.2 (GLM - Misspecified variance structure) A set of observations

$$\{(y_i, x_i); i = 1, \cdots, n\}$$

is generated from the following model:

$$y_i|p_i \sim \text{Negative Binomial}(k, p_i),$$
$$p_i = \frac{1}{1 + \exp(\eta_i)/k}, \quad \text{where } \eta_i = \beta_0 + \beta_1 x_i,$$

Table 2.2: The empirical standard deviations of the coefficient estimates, denoted by SD_e, the average sandwich standard deviations, denoted by SD_s^a, and the average model-based standard deviations, denoted by SD_m^a, based on 1000 replicates.

	β_0			β_1		
sample size	SD_e	SD_m^a	SD_s^a	SD_e	SD_m^a	SD_s^a
20	0.752	0.694	0.723	1.038	0.663	0.819
50	0.291	0.281	0.285	0.412	0.278	0.367
100	0.141	0.141	0.142	0.204	0.14	0.19
200	0.07	0.071	0.071	0.101	0.07	0.098
500	0.029	0.028	0.028	0.04	0.028	0.039

for $i = 1, \cdots, n$. Note that the mean and variance of y_i from this model are given by $E(Y_i) = \mu_i = \exp(\beta_0 + \beta_1 x_i)$ and $Var(Y_i) = \mu_i(1 + \mu_i/k)$, for $i = 1, \cdots, n$. Here the covariate x_i's are independently sampled from a Gaussian distribution with mean 0 and variance 0.1. The true values of the regression coefficients are $\beta_0 = 1$ and $\beta_1 = 2$. The sample size is set to be $n = 200$. In addition, the value of k is set to be 1, 5, or 9.

We fit the data with a negative binomial log-linear regression model and a Poisson log-linear regression model, respectively. That is, for the negative binomial regression model, the estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is obtained by solving the equation

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\mu_i + \mu_i^2/k} \mu_i \boldsymbol{x}_i = 0, \tag{2.32}$$

where $\boldsymbol{x}_i = (1, x_i)^T$ and $\mu_i = \exp\{\eta_i\}$, from the log likelihood function

$$l_n(\boldsymbol{\beta}) \propto \sum_{i=1}^{n} k \log p_i + y_i \log(1 - p_i).$$

For the Poisson regression model, the estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is obtained by

solving the equation

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\mu_i} \mu_i \boldsymbol{x}_i = 0, \tag{2.33}$$

from the log likelihood function

$$l_n(\boldsymbol{\beta}) \propto \sum_{i=1}^{n} y_i \log \mu_i - \mu_i.$$

Note that in both of these two models, the mean structure is correctly specified. The variance structure, given by $Var(Y_i) = \mu_i + \mu_i^2/k$, is correctly specified in the negative binomial regression model, but is misspecified in the Poisson regression model, with the variance function $V(\mu_i) = \mu_i$.

We calculate the empirical coverage probabilities of the 95% sandwich CIs and the 95% model-based CIs under the negative binomial regression model and the Poisson regression model, based on 5000 replicates. The results are shown in Table 2.3. Table 2.4 reports the empirical standard deviations of the coefficient estimates, denoted by SD_e, the average sandwich standard deviations, denoted by SD_s^a, and the average model-based standard deviations, denoted by SD_m^a, under both of the negative binomial regression model and the Poisson regression model.

Conclusion

As shown in Table 2.3, under the negative binomial regression model which correctly specifies the variance structure, the model-based CIs and sandwich CIs give approximately the same empirical coverage probabilities. In addition, for the slope parameter β_1, the model-based CIs have slightly more stable performance than the sandwich CIs. This conclusion agrees with the discussion about the variability of sandwich variance estimators by [50]. Under the Poisson regression model, for each value of k, the sandwich CIs perform better than the model-based CIs in the sense that the empirical coverage probabilities of the sandwich CIs are closer to the nominal value 95% than those of the model-based

CIs. Moreover, from Table 2.4, we find that under the negative binomial regression model, the average sandwich standard deviations SD_s^a and the average model-based standard deviations SD_m^a are approximately equally closer to the empirical standard deviations SD_e. However, under the Poisson regression model, the average sandwich standard deviations SD_s^a are closer to the empirical standard deviations SD_e than the average model-based standard deviations SD_m^a. In addition, as the value of k increases, the difference between these two average standard deviations shrinks.

Table 2.3: Empirical coverage probabilities of the 95% sandwich CIs and model-based CIs under the negative binomial regression model and the Poisson regression model with different values of k.

	Negative Binomial			
k	β_0		β_1	
	model-based	sandwich	model-based	sandwich
1	0.949	0.944	0.952	0.938
5	0.952	0.948	0.951	0.944
9	0.941	0.939	0.952	0.942
	Poisson			
k	β_0		β_1	
	model-based	sandwich	model-based	sandwich
1	0.698	0.943	0.689	0.943
5	0.888	0.947	0.880	0.943
9	0.900	0.939	0.912	0.942

Simulation 2.3 (LM - outliers) A sample of observations

$$\{(y_i, x_i); i = 1, \cdots, n\}$$

is generated from the following simplest linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \cdots, n.$$

Here the covariate variables x_i's are i.i.d. observations from a Gaussian distribution with mean 0 and variance 1. The error terms e_i are independently generated

Table 2.4: The empirical standard deviations of the coefficient estimates, denoted by SD_e, the average sandwich standard deviations, denoted by SD_s^a, and the average model-based standard deviations, denoted by SD_m^a, under both of the negative binomial regression model and the Poisson regression model, based on 5000 replicates.

			Negative Binomial			
		β_0			β_1	
k	SD_e	SD_m^a	SD_s^a	SD_e	SD_m^a	SD_s^a
1	0.0835	0.0831	0.0822	0.7764	0.7881	0.7633
5	0.0536	0.0538	0.0535	0.487	0.4864	0.4771
9	0.051	0.0495	0.0491	0.4897	0.4879	0.4807
			Poisson			
		β_0			β_1	
k	SD_e	SD_m^a	SD_s^a	SD_e	SD_m^a	SD_s^a
1	0.0836	0.0436	0.0822	0.7902	0.4069	0.7797
5	0.0536	0.0435	0.0535	0.4902	0.3865	0.4785
9	0.051	0.0435	0.0491	0.4900	0.4255	0.4814

from a Gaussian distribution with mean 0 and variance 1. The true values of the two regression coefficients are $\beta_0 = 1$ and $\beta_1 = 2$. The sample size n is set to be 200. In order to create a certain proportion of outliers, we select the observations with the smallest values of the covariate x_i (usually negative), and then drag them horizontally around the reflect point of zero as they were likely to be recorded with a mistake of sign, shown in Figure 2.1. The proportion of outliers is set to be 0.5%, 1%, 1.5% or 2%.

Based on 500 replicates, we show, in Table 2.5, the empirical coverage probabilities of the 95% sandwich CIs and the 95% model-based CIs. In addition, Table 2.6 reports the average biases of the LS estimates of the regression coefficients, based on 500 replicates, with different proportions of outliers for the given sample size 200. Table 2.7 reports the average biases of the LS estimates of the regression coefficients, based on 500 replicates, with different sample sizes for a given proportion 2% of outliers.

Conclusion

As shown in Table 2.6 and Table 2.7, the biases of the LS estimates of regression coefficients increase as higher proportions of outliers are included in the data. In addition, in the presence of a certain proportion of outliers, these biases will not vanish even when the sample size gets larger. Due to the biases, the sandwich CIs are not able to attain the nominal coverage probabilities except for a substantially small proportion of outliers, say 0.5%. The model-based CIs have even poorer performance. Especially for the slope parameter β_1, the model-based CIs have zero tolerance, namely, they are tolerant at most for 0% percentage of outliers, but the sandwich CIs are tolerant for a lower percentage of outliers, say lower than 1%.

Table 2.5: Empirical coverage probabilities of the 95% model-based CIs and the 95% sandwich CIs with different proportions of outliers. The sample size of the data is 200.

	β_0		β_1	
proportion of outliers	model-based	sandwich	model-based	sandwich
0.5%	0.950	0.958	0.604	0.976
1.0%	0.926	0.902	0.250	0.922
1.5%	0.870	0.822	0.030	0.752
2.0%	0.838	0.756	0.012	0.374

Table 2.6: Average biases of the LS estimates of regression coefficients, based on 500 replicates, with different proportions of outliers. The sample size is 200.

proportion of outliers	0.5%	1.0%	1.5%	2.0%
Bias($\widehat{\beta_0}$)	-0.050	-0.084	-0.123	-0.149
Bias($\widehat{\beta_1}$)	-0.153	-0.263	-0.369	-0.453

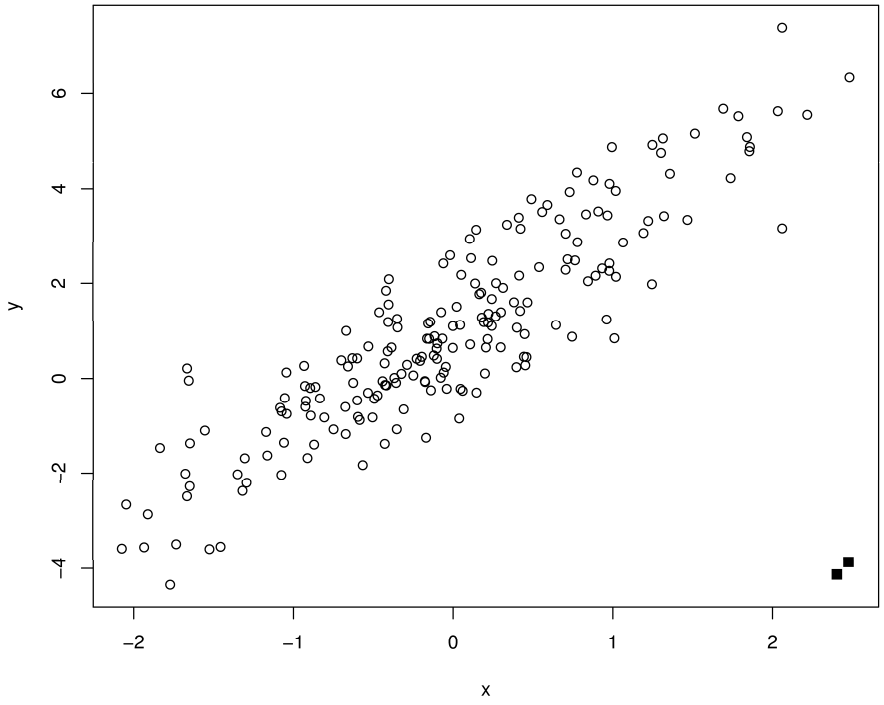

Figure 2.1: A sample of 200 observations from a simple linear regression model with 1% of outliers (solid square).

Table 2.7: Average biases of the LS estimates of regression coefficients, based on 500 replicates, with different sample sizes. The proportion of the outliers is 2%.

sample size	50	100	200	500
Bias($\widehat{\beta_0}$)	-0.153	-0.150	-0.149	-0.151
Bias($\widehat{\beta_1}$)	-0.432	-0.441	-0.445	-0.457

Summary

(1) The first two simulation studies have illustrated the consistency of the sandwich variance estimates even when the variance structure is misspecified.

(2) The validity of the consistency property requires the correct specification of the mean structure. In Simulation 2.3, since outliers violate the mean structure, neither the sandwich CIs or model-based CIs can attain the nominal coverage probabilities due to the presence of biases. However, the sandwich CIs are more tolerant of a low proportion of outliers than the model-based CIs.

(3) In Simulation 2.2, as $k \to \infty$, the variance structure assumed in a negative binomial regression model $V(\mu) \to \mu$. The sandwich and model-based variance estimators become closer, as the unit variance function approaches the true variance structure of the data distribution. Due to this property, a test statistic is proposed in Chapter 4 by comparing the two types of variance estimators.

2.5 Information Matrix Test (IM) for Model Misspecification Proposed by Hulbert White

As shown from the three simulation studies in Section 2.4, if the distributional model is misspecified, the coverage probabilities differ substantially between the

sandwich and model-based confidence intervals. This is due largely to the discrepancy between the sandwich variance estimates and the model-based variance estimates, and furthermore, essentially to the difference between the negative sensitivity matrix and the variability matrix. According to [88], a direct test for heteroscedasticity in LM can be constructed by comparing the elements of the difference between the consistent estimates of the negative sensitivity matrix and the variability matrix. With the absence of heteroscedasticity, these two matrix estimates will be approximately equal, but will generally differ otherwise. [87] extended this test further to more general situations.

Suppose that $f(y, \boldsymbol{\theta})$ is the density function of the parametric distribution imposed for the data analysis, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$ is the parameter of interest. Define two $p \times p$ matrices

$$A(\boldsymbol{\theta}) = \left\{ E\left(\frac{\partial^2 \log f(Y; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) \right\} = \left\{ A_{ij}(\boldsymbol{\theta}) \right\},$$

$$B(\boldsymbol{\theta}) = \left\{ E\left(\frac{\partial \log f(Y; \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(Y; \boldsymbol{\theta})}{\partial \theta_j} \right) \right\} = \left\{ B_{ij}(\boldsymbol{\theta}) \right\}.$$

Note that these two matrices are essentially the negative sensitivity and variability matrices of the p-element estimating function

$$\nabla_{\boldsymbol{\theta}} \log f(y; \boldsymbol{\theta}) = \left(\partial \log f(y; \boldsymbol{\theta}) / \partial \theta_1, \cdots, \partial \log f(y; \boldsymbol{\theta}) / \partial \theta_p \right)^T.$$

If the model is correctly specified, according to [87], the Bartlett identity holds at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, namely

$$-A_{ij}(\boldsymbol{\theta}_0) = B_{ij}(\boldsymbol{\theta}_0), \quad \text{or} \quad A_{ij}(\boldsymbol{\theta}_0) + B_{ij}(\boldsymbol{\theta}_0) = 0, \quad \text{for } i, j = 1, \cdots, p,$$

which is called the information matrix equivalence [87], also referred to as the "information unbiasedness" ([75]). On the other hand, the failure of the information matrix equivalence indicates that the model is misspecified. [87] proposed an information matrix (IM) test for model misspecification based on the

argument above. First, define another two $p \times p$ matrices:

$$A_n(\boldsymbol{\theta}) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(y_i; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right\},$$

$$B_n(\boldsymbol{\theta}) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f(y_i; \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(y_i; \boldsymbol{\theta})}{\partial \theta_j} \right\}.$$

The matrices $A_n(\widehat{\boldsymbol{\theta}}_n)$ and $B_n(\widehat{\boldsymbol{\theta}}_n)$ are consistent estimators of $A(\boldsymbol{\theta})$ and $B(\boldsymbol{\theta})$, respectively, where $\widehat{\boldsymbol{\theta}}_n$ is the so-called *quasi-maximum likelihood estimator* (QMLE), which maximizes $L_n(\boldsymbol{y}; \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta})$, the quasi-log-likelihood of the sample $\boldsymbol{y} = (y_1, \cdots, y_n)^T$. [87] derived the asymptotic distribution of the elements of $\sqrt{n}\left(A_n(\widehat{\boldsymbol{\theta}}_n) + B_n(\widehat{\boldsymbol{\theta}}_n)\right)$. With a consistent estimator for the asymptotic covariance matrix, denoted by $V_n(\widehat{\boldsymbol{\theta}}_n)$, he constructed an asymptotic χ^2 statistic of the [82] type. The IM test by White is discussed in detail in Section 2.5.1.

2.5.1 Information matrix test statistics

Consider the upper triangle elements of $A(\boldsymbol{\theta}) + B(\boldsymbol{\theta})$,

$$d_l(y, \boldsymbol{\theta}) = \partial \log f(y, \boldsymbol{\theta}) / \partial \theta_j \cdot \partial \log f(y, \boldsymbol{\theta}) / \partial \theta_t + \partial^2 \log f(y, \theta) / \partial \theta_j \partial \theta_t,$$

where $l = 1, \cdots, p(p+1)/2$; $j = 1, \cdots, p$; $t = j, \cdots, p$. It is useful to consider a test on certain linear combinations of these elements, or simply a subset of these elements. This is because, firstly, some may be identically zero, and secondly, $d_l(y, \boldsymbol{\theta})$ may consist of the set of linear combinations of the others.
Define the $q \times 1$ vector $d(y, \boldsymbol{\theta})$, $q \leq p(p+1)/2$. Then

$$D_n(\widehat{\boldsymbol{\theta}}_n) = n^{-1} \sum_{i=1}^{n} d(y_i, \widehat{\boldsymbol{\theta}}_n) \quad \text{and} \quad D(\boldsymbol{\theta}) = E\{d(Y, \boldsymbol{\theta})\}$$

are both $q \times 1$ vectors. Also define the $q \times p$ Jacobian matrices

$$\nabla D_n(\boldsymbol{\theta}) = \left\{ n^{-1} \sum_{i=1}^{n} \partial d_l(y_i, \boldsymbol{\theta}) / \partial \theta_k \right\} \quad \text{and} \quad \nabla D(\boldsymbol{\theta}) = \left\{ E\left(\partial d_l(Y, \boldsymbol{\theta}) / \partial \theta_k \right) \right\}.$$

Define

$$\begin{aligned} V(\boldsymbol{\theta}) &= E\left\{ \left[d(Y, \boldsymbol{\theta}) - \nabla D(\boldsymbol{\theta}) A(\boldsymbol{\theta})^{-1} \nabla \log f(Y, \boldsymbol{\theta}) \right] \right. \\ &\qquad \left. \left[d(Y, \boldsymbol{\theta}) - \nabla D(\boldsymbol{\theta}) A(\boldsymbol{\theta})^{-1} \nabla \log f(Y, \boldsymbol{\theta}) \right]^T \right\}. \end{aligned}$$

Note that $V(\boldsymbol{\theta}_0)$ is the asymptotic covariance matrix of $\sqrt{n} D_n(\widehat{\boldsymbol{\theta}}_n)$ under the null hypothesis that the model is correctly specified, and it can be consistently estimated by

$$\begin{aligned} V_n(\widehat{\boldsymbol{\theta}}_n) &= n^{-1} \sum_{i=1}^{n} \left[d(y_i, \widehat{\boldsymbol{\theta}}_n) - \nabla D_n(\widehat{\boldsymbol{\theta}}_n) A_n(\widehat{\boldsymbol{\theta}}_n)^{-1} \nabla \log f(y_i, \widehat{\boldsymbol{\theta}}_n) \right] \\ &\qquad \left[d(y_i, \widehat{\boldsymbol{\theta}}_n) - \nabla D_n(\widehat{\boldsymbol{\theta}}_n) A_n(\widehat{\boldsymbol{\theta}}_n)^{-1} \nabla \log f(y_i, \widehat{\boldsymbol{\theta}}_n) \right]^T. \end{aligned}$$

Theorem 2.8 (White's Information Matrix Test) *Under mild regularity conditions, if the model $f(y; \boldsymbol{\theta})$ is correctly specified,*

(i) $\sqrt{n} D_n(\widehat{\boldsymbol{\theta}}_n) \underrightarrow{d} MVN(\mathbf{0}, V(\boldsymbol{\theta}_0))$*;*

(ii) $V_n(\widehat{\boldsymbol{\theta}}_n) \underrightarrow{a.s.} V(\boldsymbol{\theta}_0)$*, and* $V_n(\widehat{\boldsymbol{\theta}}_n)$ *is nonsingular almost surely for all n sufficiently large;*

(iii) the information matrix test statistic

$$n D_n(\widehat{\boldsymbol{\theta}}_n)^T V_n(\widehat{\boldsymbol{\theta}}_n)^{-1} D_n(\widehat{\boldsymbol{\theta}}_n) \sim \chi^2_q, \quad \textit{asymptotically.}$$

More details can be found in [87]. In the rest of the thesis, we call this test the *White's IM test*. Rejection of the null hypothesis that the model has been correctly specified implies that, at least, the model-based covariance matrix estimator $-A_n(\widehat{\boldsymbol{\theta}}_n)$ is inconsistent, and possibly, the QMLE $\widehat{\boldsymbol{\theta}}_n$ for the parameters of interest is inconsistent. However, in practice, the calculation of the estimator $V_n(\widehat{\boldsymbol{\theta}}_n)$ can be cumbersome due to the requirement of third derivatives. Often it can be shown under the null hypothesis that $\nabla D(\boldsymbol{\theta}_0)$ vanishes, so that $V(\boldsymbol{\theta}_0)$ is consistently estimated by $n^{-1}\sum_{i=1}^{n} d(y_i, \widehat{\boldsymbol{\theta}}_n) d(y_i, \widehat{\boldsymbol{\theta}}_n)^T$. When $\nabla D(\boldsymbol{\theta}_0)$ does not vanish, $V(\boldsymbol{\theta}_0)$ can be consistently estimated by

$$n^{-1}\sum_{i=1}^{n} d(y_i, \widehat{\boldsymbol{\theta}}_n) d(y_i, \widehat{\boldsymbol{\theta}}_n)^T - \nabla D_n(\widehat{\boldsymbol{\theta}}_n) C_n(\widehat{\boldsymbol{\theta}}_n) \nabla D_n(\widehat{\boldsymbol{\theta}}_n),$$

where $C_n(\boldsymbol{\theta}) = A_n(\boldsymbol{\theta})^{-1} B_n(\boldsymbol{\theta}) A_n(\boldsymbol{\theta})$. However, even though these alternative estimators can be employed for simplification, they are neither consistent nor necessarily positive semi-definite when the null hypothesis fails.

Moreover, if the model misspecifications amount only to the loss in efficiency associated with quasi-maximum likelihood estimation, rather than inconsistency of the resultant parameter estimator or covariance matrix estimator, the information matrix test will lose power. To overcome this shortcoming, in Chapter 4, we propose an information ratio test, targeting model misspecification of the variance/covariance structure, but with the correct specification of the mean structure. Several simulation studies show that the proposed information ratio test is more powerful than the White's IM test when the model misspecification leads to consistent but inefficient estimators of parameters of interest, such as regression coefficients.

Chapter 3

Godambian Estimator of Dispersion Parameter

Model misspecification leads to a large discrepancy between the model-based and sandwich variance estimators. Usually, the dispersion parameter σ^2, in the model-based variance estimators, is estimated by a moment estimator, if its true value is unknown. The moment estimators, for example, (2.17) and (2.31), can be regarded as an equally weighted sum of the squared Pearson residuals. In this Chapter, it can be shown that in the sandwich variance estimators, compared with the model-based variance estimators, the dispersion parameter σ^2 is analogously estimated by a weighted sum of the squared Pearson residuals in GLM or the squared transformed residuals in GEE. This estimator is called the Godambian estimator of the dispersion parameter. In addition, we show that, for each individual regression coefficient, the weights in the Godambian estimator take a form of the difference between the diagonal elements of two hat matrices: one is obtained from the full "weighted" design matrix, and the other is obtained from the sub-matrix, with the corresponding covariate column deleted from this full matrix. Moreover, in LM, it can be shown that the weights in the Godambian estimator, related to a certain individual regression coefficient, characterize the influence from the corresponding covariate.

We will start with the GLM for independent data, and then extend the result to the context of GEE for correlated data.

3.1 Generalized Linear Regression Models

In the context of GLM, the model-based and sandwich covariance matrix estimators, (2.22) and (2.19), of the regression coefficient estimator $\widehat{\boldsymbol{\beta}}_n$ are given by, respectively,

$$ASCOV_m(\widehat{\boldsymbol{\beta}}_n) = \widehat{\sigma}^2_m \left(X^T \widehat{\Delta} \widehat{\mathcal{V}}^{-1} \widehat{\Delta} X\right)^{-1},$$

where the matrix $\widehat{\Delta}$ is given in (2.13), the matrix $\widehat{\mathcal{V}}$ is given in (2.14), and the dispersion parameter σ^2 is estimated by a moment estimator $\widehat{\sigma}^2_m$, if the true value is unknown, and

$$ASCOV_s(\widehat{\boldsymbol{\beta}}_n) = \left(X^T \widehat{\Delta} \widehat{\mathcal{V}}^{-1} \widehat{\Delta} X\right)^{-1} \left(X^T \widehat{\Delta} \widehat{\mathcal{V}}^{-1} \mathcal{R} \widehat{\mathcal{V}}^{-1} \widehat{\Delta} X\right) \left(X^T \widehat{\Delta} \widehat{\mathcal{V}}^{-1} \widehat{\Delta} X\right)^{-1}.$$

Let $\widehat{\mathcal{V}}^{1/2}$ be an $n \times n$ diagonal matrix with the i-th diagonal element $\sqrt{V(\widehat{\mu}_i)}$. Let $\mathcal{R}_P$ be an $n \times n$ diagonal matrix, defined by

$$\mathcal{R}_P = \text{diag}\left\{r^2_{P,1}, \cdots, r^2_{P,n}\right\}, \tag{3.1}$$

where $r_{P,i}$ is the Pearson residual, given by

$$r_{P,i} = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}, \quad i = 1, \cdots, n.$$

Define $\widehat{U} = \widehat{\mathcal{V}}^{-1/2} \widehat{\Delta} X$. Note that the matrix $\widehat{U}$ consists of the weighted covariates with the weights $\widehat{\mu}_i / \sqrt{V(\widehat{\mu}_i)}$ for each subject, so it can be regarded as a "weighted" design matrix. Then, the model-based and sandwich covariance matrix estimators, (2.22) and (2.19), can be written as

$$ASCOV_m(\widehat{\boldsymbol{\beta}}_n) = \widehat{\sigma}^2_m \left(\widehat{U}^T \widehat{U}\right)^{-1}, \tag{3.2}$$

and

$$ASCOV_s(\widehat{\boldsymbol{\beta}}_n) = \left(\widehat{U}^T\widehat{U}\right)^{-1}\left(\widehat{U}^T\mathcal{R}_p\widehat{U}\right)\left(\widehat{U}^T\widehat{U}\right)^{-1}. \tag{3.3}$$

Theorem 3.1 *Suppose that, for the regression coefficient β_{j-1}, $j = 1, \cdots, p$, the sandwich variance estimator of the coefficient estimator $\widehat{\beta}_{j-1}$ can be written as*

$$ASVAR_s(\widehat{\beta}_{j-1}) = \widetilde{\sigma}^2_{j-1}a_j, \tag{3.4}$$

where a_j is the j-th diagonal element of the matrix $\left(\widehat{U}^T\widehat{U}\right)^{-1}$, given in the model-based covariance matrix estimator (3.2).

Here $\widetilde{\sigma}^2_{j-1}$ can be regarded as an estimator of the dispersion parameter σ^2, written as a weighted sum of the squared Pearson residuals, that is,

$$\widetilde{\sigma}^2_{j-1} = \sum_{i=1}^{n}\widehat{w}_i^{(j-1)}r^2_{P,i}, \quad j = 1, \cdots, p. \tag{3.5}$$

The weights $\widehat{w}_i^{(j-1)}$ are given by

$$\widehat{w}_i^{(j-1)} = \widehat{h}_{ii} - \widehat{h}_{ii}^{(-j)}, \tag{3.6}$$

where $\widehat{h}_{ii}$ is the i-th diagonal element of the hat matrix $\widehat{H}$, defined as

$$\widehat{H} = \widehat{U}\left(\widehat{U}^T\widehat{U}\right)^{-1}\widehat{U}^T, \tag{3.7}$$

and $\widehat{h}_{ii}^{(-j)}$ is the i-th diagonal element of the hat matrix $\widehat{H}^{(-j)}$, defined as

$$\widehat{H}^{(-j)} = \widehat{U}_{(-j)}\left(\widehat{U}^T_{(-j)}\widehat{U}_{(-j)}\right)^{-1}\widehat{U}^T_{(-j)},$$

obtained from the sub-matrix $\widehat{U}_{(-j)}$ with the j-th column deleted from $\widehat{U}$, for $j = 1, \cdots, p$.

*The estimator $\widetilde{\sigma}^2_{j-1}$ is called the β_{j-1}-**specific Godambian estimator** of the dispersion parameter σ^2.*

Note that the hat matrix $\widehat{H}$ (3.7) is also referred to the leverage matrix for Poisson regression models ([32] and [89]).

Proof. First of all, suppose that the matrix $\widehat{U}$ is an orthogonal matrix, i.e., $\widehat{U}^T\widehat{U} = \boldsymbol{I}_p$, where $\boldsymbol{I}_p$ is a p-dimensional identity matrix. Its diagonal elements are $a_j = 1$, for $j = 1, \cdots, p$. In addition, the sandwich variance estimator of $\widehat{\beta}_{j-1}$ is given by

$$ASVAR_s(\widehat{\beta}_{j-1}) = \sum_{i=1}^{n} u_{i,j-1}^2 r_{P,i}^2, \quad j = 1, \cdots, p,$$

where $u_{i,j-1}$ is the (i, j)-th element of the matrix $\widehat{U}$. It follows from (3.4) that

$$\widetilde{\sigma}_{j-1}^2 = \sum_{i=1}^{n} \widehat{w}_i^{(j-1)} r_{P,i}^2 = \sum_{i=1}^{n} u_{i,j-1}^2 r_{P,i}^2,$$

where the weights are $\widehat{w}_i^{(j-1)} = u_{i,j-1}^2$ with $\sum_{i=1}^n \widehat{w}_i^{(j-1)} = 1$.

The hat matrix $\widehat{H}$ (3.7) has the i-th diagonal element $\widehat{h}_{ii} = \sum_{k=1}^{p} u_{i,k-1}^2$. If we delete the j-th column from the matrix $\widehat{U}$, the resulting matrix $\widehat{U}_{(-j)}$ is also an orthogonal matrix, i.e., $\widehat{U}_{(-j)}^T\widehat{U}_{(-j)} = \boldsymbol{I}_{p-1}$. Then, the corresponding hat matrix $\widehat{H}^{(-j)}$ has the i-th diagonal element $\widehat{h}_{ii}^{(-j)} = \sum_{k\neq j} u_{i,k-1}^2$. Thus, the weights

$$\widehat{w}_i^{(j-1)} = u_{i,j-1}^2 = \widehat{h}_{ii} - \widehat{h}_{ii}^{(-j)}.$$

Secondly, let us consider the case that the matrix $\widehat{U}$ is an arbitrary matrix of full rank. By the QR factorization [38], the matrix $\widehat{U}$ can be decomposed as follows:

$$\Pi = \widehat{U}Q, \tag{3.8}$$

where Q is a $p \times p$ upper triangular matrix, and Π is an orthogonal matrix, i.e., $\Pi^T\Pi = \boldsymbol{I}_p$. Under this decomposition, $(\widehat{U}^T\widehat{U})^{-1} = QQ^T$, and the sandwich covariance matrix estimator (3.3) can be re-written as

$$ASCOV_s(\widehat{\boldsymbol{\beta}}_n) = Q\left(\Pi^T\mathcal{R}_P\Pi\right)Q^T. \tag{3.9}$$

Let us start with the parameter β_{p-1}, the last element of the coefficient vector $\boldsymbol{\beta}$. Let $\widehat{U}_{(-p)}$ denote the sub-matrix which deletes the p-th column from the matrix $\widehat{U}$. Re-express the matrix Π as

$$\Pi = \begin{pmatrix} \Pi_1 & \boldsymbol{\pi} \end{pmatrix},$$

where Π_1 is an $n \times (p-1)$ matrix containing the first $p-1$ columns of the matrix Π, and $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_n)^T$ is the last column vector of Π. We can also re-express the matrix Q as

$$Q = \begin{pmatrix} Q_{11} & \boldsymbol{q}_{12} \\ \mathbf{0} & q_{22} \end{pmatrix},$$

where Q_{11} is a $(p-1)\times(p-1)$ upper triangular matrix, $\boldsymbol{q}_{12}$ is a $(p-1)\times 1$ vector, $\mathbf{0}$ is a $1 \times (p-1)$ zero vector, and q_{22} is a scalar.

The p-th diagonal element of the matrix $(\widehat{U}^T\widehat{U})^{-1}$ is $a_p = q_{22}^2$. The sandwich variance estimator of $\widehat{\beta}_{p-1}$ can be written as

$$ASVAR_s(\widehat{\beta}_{p-1}) = q_{22}^2 \boldsymbol{\pi}^T \mathcal{R}_P \boldsymbol{\pi} = q_{22}^2 \sum_{i=1}^{n} \pi_i^2 r_{P,i}^2. \tag{3.10}$$

Then, we can obtain

$$\widetilde{\sigma}_{p-1}^2 = \sum_{i=1}^{n} \widehat{w}_i^{(p-1)} r_{P,i}^2 = \sum_{i=1}^{n} \pi_i^2 r_{P,i}^2, \tag{3.11}$$

where the weights are $\widehat{w}_i^{(p-1)} = \pi_i^2$.

The hat matrices are invariant under orthogonalization. That is, the hat matrix $\widehat{H}$ (3.7) w.r.t. $\widehat{U}$ is equivalent to the hat matrix H_{Π} w.r.t. Π, i.e.,

$$\widehat{H} = \widehat{U}\left(\widehat{U}^T\widehat{U}\right)^{-1}\widehat{U}^T = \Pi\left(\Pi^T\Pi\right)^{-1}\Pi^T = H_{\Pi}.$$

Since $H_{\Pi} = \Pi_1\Pi_1^T + \boldsymbol{\pi}\boldsymbol{\pi}^T$, the i-th diagonal element, $\widehat{h}_{ii}$, of $\widehat{H}$ equals to the sum of the i-th diagonal elements of the matrices $\Pi_1\Pi_1^T$ and $\boldsymbol{\pi}\boldsymbol{\pi}^T$, for $i = 1, \cdots, p$.

Moreover, the QR decomposition of $\widehat{U}_{(-p)}$ is obtained by

$$\Pi_1 = \widehat{U}_{(-p)} Q_{11},$$

from the QR decomposition (3.8). Then, the i-th diagonal element, $\widehat{h}_{ii}^{(-p)}$, of $\widehat{H}^{(-p)}$ equals to the i-th diagonal element of $\Pi_1 \Pi_1^T$. Because the i-th diagonal element of $\boldsymbol{\pi}\boldsymbol{\pi}^T$ is π_i^2, we can obtain $\widehat{h}_{ii} = \widehat{h}_{ii}^{(-p)} + \pi_i^2$. Therefore, the weights $\widehat{w}_i^{(p-1)}$ can be written as

$$\widehat{w}_i^{(p-1)} = \pi_i^2 = \widehat{h}_{ii} - \widehat{h}_{ii}^{(-p)}, \quad \text{for } i = 1, \cdots, p.$$

Next consider other elements of the coefficient vector $\boldsymbol{\beta}, \beta_0, \cdots, \beta_{p-2}$. Given an index $j \in \{1, \cdots, p-2\}$, let $\widetilde{X}$ be the resulting matrix that swaps the j-th and last columns of the matrix X, that is,

$$\widetilde{X} = X\mathcal{S},$$

where the matrix $\mathcal{S}$ is a $p \times p$ matrix given by

$$\begin{array}{c} \\ \\ \\ \\ j \\ \\ \\ p \end{array} \begin{array}{c} \begin{array}{cccccc} & & & j & & p \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & 1 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & \cdots & 1 & \cdots & 0 & 0 \end{pmatrix} \end{array}.$$

Let $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_0, \cdots, \widetilde{\beta}_{p-1})$ denote the vector which switches the j-th and p-th element of the regression coefficients $\boldsymbol{\beta}$, i.e., $\widetilde{\boldsymbol{\beta}} = \mathcal{S}\boldsymbol{\beta}$. Let $\widehat{\widetilde{\boldsymbol{\beta}}}_n = (\widehat{\widetilde{\beta}}_0, \cdots, \widehat{\widetilde{\beta}}_{p-1})$ denote the corresponding estimator of $\widetilde{\boldsymbol{\beta}}$. It is easy to show that $\widehat{\widetilde{\boldsymbol{\beta}}}_n = \mathcal{S}\widehat{\boldsymbol{\beta}}_n$, which implies that the estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ switches the positions of $\widehat{\beta}_{j-1}$ and $\widehat{\beta}_{p-1}$ in

the estimator $\widehat{\boldsymbol{\beta}}_n$. That is, $\widehat{\beta}_{j-1} = \widetilde{\widehat{\beta}}_{p-1}$. Let $\widetilde{\widehat{U}} = \widehat{\mathcal{V}}^{-1/2}\widetilde{\widehat{\Delta \boldsymbol{X}}} = \widehat{U}\mathcal{S}$, which swaps the j-th and p-th columns of the matrix $\widehat{U}$. Note that the fitted values, variance function and Pearson residuals remain the same under the swap because they depend on the regression coefficient estimator only through the linear predictors $\widetilde{\widehat{X\boldsymbol{\beta}}} = X\widehat{\boldsymbol{\beta}}$. The sandwich covariance matrix estimator of $\widetilde{\widehat{\boldsymbol{\beta}}}_n$ can be written as

$$ASCOV_s(\widetilde{\widehat{\boldsymbol{\beta}}}_n) \;=\; \mathcal{S}\left\{ ASCOV_s(\widehat{\boldsymbol{\beta}}_n)\right\}\mathcal{S},$$

which swaps the j-th and p-th diagonal elements of $ASCOV_s(\widehat{\boldsymbol{\beta}}_n)$. It implies that the sandwich variance estimator of $\widehat{\beta}_{j-1}$, given by

$$ASVAR_s(\widehat{\beta}_{j-1}) = \widetilde{\sigma}^2_{j-1}a_j,$$

is equivalent to that of $\widetilde{\widehat{\beta}}_{p-1}$, given by

$$ASVAR_s(\widetilde{\widehat{\beta}}_{p-1}) = \widetilde{\widetilde{\sigma}}^2_{p-1}\widetilde{a}_p,$$

where a_j is the j-th diagonal element of $\left(\widehat{U}^T\widehat{U}\right)^{-1}$ and $\widetilde{a}_p$ is the p-th diagonal element of $\left(\widetilde{\widehat{U}}^T\widetilde{\widehat{U}}\right)^{-1}$. Under the swap, $\left(\widetilde{\widehat{U}}^T\widetilde{\widehat{U}}\right)^{-1} = \mathcal{S}\left(\widehat{U}^T\widehat{U}\right)^{-1}\mathcal{S}$, which indicates that $\widetilde{a}_p = a_j$. It follows that $\widetilde{\widetilde{\sigma}}^2_{p-1} = \widetilde{\sigma}^2_{j-1}$.

According to the results (3.11), $\widetilde{\widetilde{\sigma}}^2_{p-1} = \sum_{i=1}^{n} \widetilde{\widetilde{w}}_i^{(p-1)} r^2_{P,i}$, where the weights are

$$\widetilde{\widetilde{w}}_i^{(p-1)} = \widetilde{\widehat{h}}_{ii} - \widetilde{\widetilde{h}}_{ii}^{(-p)}, \tag{3.12}$$

where $\widetilde{\widehat{h}}_{ii}$ is the i-th diagonal element of the hat matrix $\widetilde{\widehat{H}} = \widetilde{\widehat{U}}\left(\widetilde{\widehat{U}}^T\widetilde{\widehat{U}}\right)^{-1}\widetilde{\widehat{U}}^T$, and $\widetilde{\widetilde{h}}_{ii}^{(-p)}$ is the i-th diagonal element of the hat matrix

$$\widetilde{\widetilde{H}}^{(-p)} = \widetilde{\widehat{U}}_{(-p)}\left(\widetilde{\widehat{U}}^T_{(-p)}\widetilde{\widehat{U}}_{(-p)}\right)^{-1}\widetilde{\widehat{U}}^T_{(-p)}.$$

Here, $\widehat{\widetilde{U}}_{(-p)}$ is the sub-matrix with the last column deleted from $\widehat{\widetilde{U}}$, which is the same as the sub-matrix $\widehat{U}_{(-j)}$ with the j-th column deleted from $\widehat{U}$. Consequently, $\widehat{\widetilde{h}}_{ii}^{(-p)} = \widehat{h}_{ii}^{(-j)}$. In addition, since the hat matrix is invariant w.r.t. swap, $\widehat{\widetilde{h}}_{ii} = \widehat{h}_{ii}$. Therefore, $\widetilde{\sigma}^2_{j-1}$, given in the sandwich variance estimator $ASVAR_s(\widehat{\beta}_{j-1})$, can be written as

$$\widetilde{\sigma}^2_{j-1} = \sum_{i=1}^{n} \widehat{w}_i^{(j-1)} r^2_{P,i}, \quad j = 1, \cdots, p,$$

where the weights $\widehat{w}_i^{(j-1)} = \widehat{h}_{ii} - \widehat{h}_{ii}^{(-j)}$. This completes the proof of Theorem 3.1.

3.1.1 Special Case: Godambian estimators of the variance parameter σ^2 in LM

In the context of LM, under the assumptions

$$\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad \text{and} \quad V(\mu_i) = 1,$$

the matrix $\widehat{U}$ reduces to the original design matrix X, and the Pearson residuals $r_{P,i}$ reduce to the raw residuals r_i.

Corollary 3.1 *Suppose that, for the regression coefficient β_{j-1}, $j = 1, \cdots, p$, the sandwich variance estimator of the coefficient estimator $\widehat{\beta}_{j-1}$ can be written as*

$$ASVAR_s(\widehat{\beta}_{j-1}) = \widetilde{\sigma}^2_{j-1} a_j, \tag{3.13}$$

where a_j is the j-th diagonal element of the matrix $\left(X^T X\right)^{-1}$, given in the model-based covariance matrix estimator (2.25).

Here $\widetilde{\sigma}^2_{j-1}$ can be regarded as an estimator of the variance parameter σ^2, written as a weighted sum of the squared residuals, that is,

$$\widetilde{\sigma}^2_{j-1} = \sum_{i=1}^{n} w_i^{(j-1)} r_i^2, \quad j = 1, \cdots, p. \tag{3.14}$$

The weights $w_i^{(j-1)}$ *are given by*

$$w_i^{(j-1)} = h_{ii} - h_{ii}^{(-j)}, \tag{3.15}$$

where h_{ii} *is the* i*-th diagonal element of the hat matrix* H*, defined as*

$$H = X\left(X^T X\right)^{-1} X^T, \tag{3.16}$$

and $h_{ii}^{(-j)}$ *is the* i*-th diagonal element of the hat matrix* $H^{(-j)}$ *obtained from the sub-matrix* $X_{(-j)}$*, with the* j*-th column deleted from the full design matrix* X*, defined as*

$$H^{(-j)} = X_{(-j)}\left\{X_{(-j)}^T X_{(-j)}\right\}^{-1} X_{(-j)}^T. \tag{3.17}$$

The estimator $\widetilde{\sigma}_{j-1}^2$ *is called the* β_{j-1}**-specific Godambian estimator** *of the variance parameter* σ^2*.*

Here, the weights take differences between the diagonal elements, also called leverages, from two hat matrices. [44] discussed that the diagonal element, h_{ii}, of the hat matrix can be interpreted as the amount of leverage or influence of the i-th observation exerted on the fitted value based on the full model. Specifically, a large value of h_{ii} represents high influence if the i-th observation $(x_{i0}, \cdots, x_{i,p-1})$ is far away from the sample center in the space $\mathbb{R}^p$. This observation is called a "high leverage point". On the other hand, when the observation is closer to the center, it has lower influence on the fitted value, and consequently, the value of h_{ii} becomes smaller.

Similarly, $h_{ii}^{(-j)}$ characterizes the amount of influence on the fitted value by the i-th observation $(x_{i0}, \cdots, x_{i,j-2}, x_{i,j}, \cdots, x_{i,p-1})$ in the space $\mathbb{R}^{p-1}$ based on a sub-model without the j-th covariate. Therefore, the weights $w_i^{(j-1)} = h_{ii} - h_{ii}^{(-j)}$ can be interpreted as the amount of influence of the i-th observation on the fitted value contributed from the j-th covariate. It is worth pointing out that

(1) $\sum_{i=1}^{n} w_i^{(j-1)} = 1$, for $j = 1, \cdots, p$;

(2) it is expected that when the observation $x_{i,j-1}$ approaches to the sample center of the j-th covariate variable, the value of $w_i^{(j-1)}$ tends to get smaller; on the other hand, the value of $w_i^{(j-1)}$ is likely to get larger if the observation deviates from the center.

Simulation

Through a simulation study, we will investigate the properties of these Godambe weights in the LM setting with independent covariates. Consider a sample of observations $\{x_{11}, \cdots, x_{n,1}\}$ independent drawn from a Gaussian distribution $N(1, 1)$, and another sample $\{x_{12}, \cdots, x_{n,2}\}$ from a Gaussian distribution $N(2, 0.5)$. The sample size n is 200. The full design matrix X is given by an $n \times 3$ matrix with the i-th row $(1, x_{i1}, x_{i2})$, for $i = 1, \cdots, n$. For each $j = 1, 2, 3$, the weights $w_i^{(j-1)}$, calculated from (3.15), are regressed non-parametrically on all the three covariate variables $\{x_{1,k-1}, \cdots, x_{n,k-1}\}$, $k = 1, 2, 3$, using a kernel smoothing technique (see [64] and [74]). That is, for each i, the kernel regression estimate of the weight $w_i^{(k-1)}$ on the covariate $x_{i,j-1}$'s is given by

$$\widetilde{w}_i^{(k-1)} = \frac{\sum_{l=1}^{n} \kappa_h(x_{l,j-1} - x_{i,j-1}) w_i^{(k-1)}}{\sum_{l=1}^{n} \kappa_h(x_{l,j-1} - x_{i,j-1})}, \quad j, k = 1, 2, 3,$$

where $\kappa_h(\cdot)$ is the kernel function, and h is the bandwidth. Here we use a Gaussian kernel with bandwidth 2. For $j = 1$, $x_{i,0} = 1$, for $i = 1, \cdots, n$. Then, the kernel estimates of the weights $w_i^{(0)}$, regressing on the vector of 1's, is in fact its sample mean $n^{-1} \sum_{i=1}^{n} w_i^{(0)} = 1/n = 0.05$.

Figure 3.1 displays the kernel estimates of the weights $w_i^{(1)}$ regressing on x_{i1} and x_{i2}, respectively. As expected, in the left panel of the figure, the pattern of the kernel estimates of the weights $w_i^{(1)}$ strongly associated with the covariate x_{i1} agrees with what we have discussed above. That is, the estimates get smaller when the covariate x_{i1} gets closer to the sample center. In addition, the right panel shows that the kernel estimates are almost constant, without any obvious relation between the weights $w_i^{(1)}$ and the covariate x_{i2}. Similar results are found in Figure 3.2, which shows the kernel estimates of the weights $w_i^{(2)}$ regressing on x_{i1} and x_{i2}, respectively.

Summary. This simulation experiment shows that the weights $w_i^{(j-1)}$ are strongly associated with the covariate $x_{i,j-1}$, but rarely have relation with other covariates, for $j = 1, \cdots, p$. In LM, the error terms are usually assumed to be independent random variables with mean 0, and constant variance σ^2. Checking on these assumptions is often carried out by visual examination of appropriate residual plots. For instance, plotting the residuals versus certain covariate variables is able to show certain kinds of heteroscedasticity, which violate the assumption of constant error variance. Let r_i denote the residual from the i-th data point. It can be shown that under the homoscedasticity assumption,

$$E(r_i^2) = (1 - h_{ii})\sigma^2, \quad i = 1, \cdots, n.$$

Suppose that a certain covariate, say $x_{i,j-1}$, accounts for the heteroscedasticity in the model; that is, the squared residuals can be modelled in the following form:

$$r_i^2 = (1 - h_{ii})\sigma^2 + g(x_{i,j-1}) + \xi_i, \quad i = 1, \cdots, n,$$

where $g(\cdot)$ is a certain function defined on only the covariate $x_{i,j-1}$, and ξ_i is an error term, assumed to have expectation 0. Consider a weighted sum of the squared residuals with a certain standardization procedure,

$$\sum_{i=1}^{n} \frac{w_i^{(k-1)}}{\sum_{i=1}^{n}(1 - h_{ii})w_i^{(k-1)}\sigma^2} r_i^2,$$

for $k = 1, \cdots, p$. If $k \neq j$, its expectation is

$$E\left\{\sum_{i=1}^{n} \frac{w_i^{(k-1)}}{\sum_{i=1}^{n}(1-h_{ii})w_i^{(k-1)}\sigma^2} r_i^2\right\} = 1,$$

if we assume that the covariate $(x_{1,k-1}, \cdots, x_{n,k-1})$ is sampled independent of the covariate $\left(x_{1,j-1}, \cdots, x_{n,j-1}\right)$. If $k = j$, its expectation is

$$E\left\{\sum_{i=1}^{n} \frac{w_i^{(k-1)}}{\sum_{i=1}^{n}(1-h_{ii})w_i^{(k-1)}\sigma^2} r_i^2\right\} = 1 + \frac{\sum_{i=1}^{n} w_i^{(j-1)} g(x_{i,j-1})}{\sum_{i=1}^{n} w_i^{(j-1)}(1-h_{ii})\sigma^2},$$

which deviates from 1. As a result, this property is helpful to identify the covariate which is responsible for the heteroscedasticity by the information ratio test, proposed in Chapter 4, when the null hypothesis is rejected.

3.2 Generalized estimating equations

In this section, all the results derived from the quasi-likelihood inference in GLM for independent data will be extended to the GEE in longitudinal data analysis. The model-based and sandwich covariance matrix estimators of $\widehat{\boldsymbol{\beta}}_K$ are given by (2.30) and (2.28)

$$ASCOV_m(\widehat{\boldsymbol{\beta}}_K) = \widehat{\sigma}_m^2 \left\{\sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}_i^{-1} \widehat{D}_i\right\}^{-1},$$

and

$$ASCOV_s(\widehat{\boldsymbol{\beta}}_K) = \left\{\sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}_i^{-1} \widehat{D}_i\right\}^{-1} \left\{\sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}_i^{-1} \boldsymbol{r}_i \boldsymbol{r}_i^T \widehat{\Sigma}_i^{-1} \widehat{D}_i\right\}^{-1} \left\{\sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}_i^{-1} \widehat{D}_i\right\}.$$

Let $N = \sum_{i=1}^{K} n_i$. Let $\boldsymbol{y}$ be an $N \times 1$ vector, defined as

$$\boldsymbol{y}^T = (\boldsymbol{y}_1^T, \cdots, \boldsymbol{y}_K^T),$$

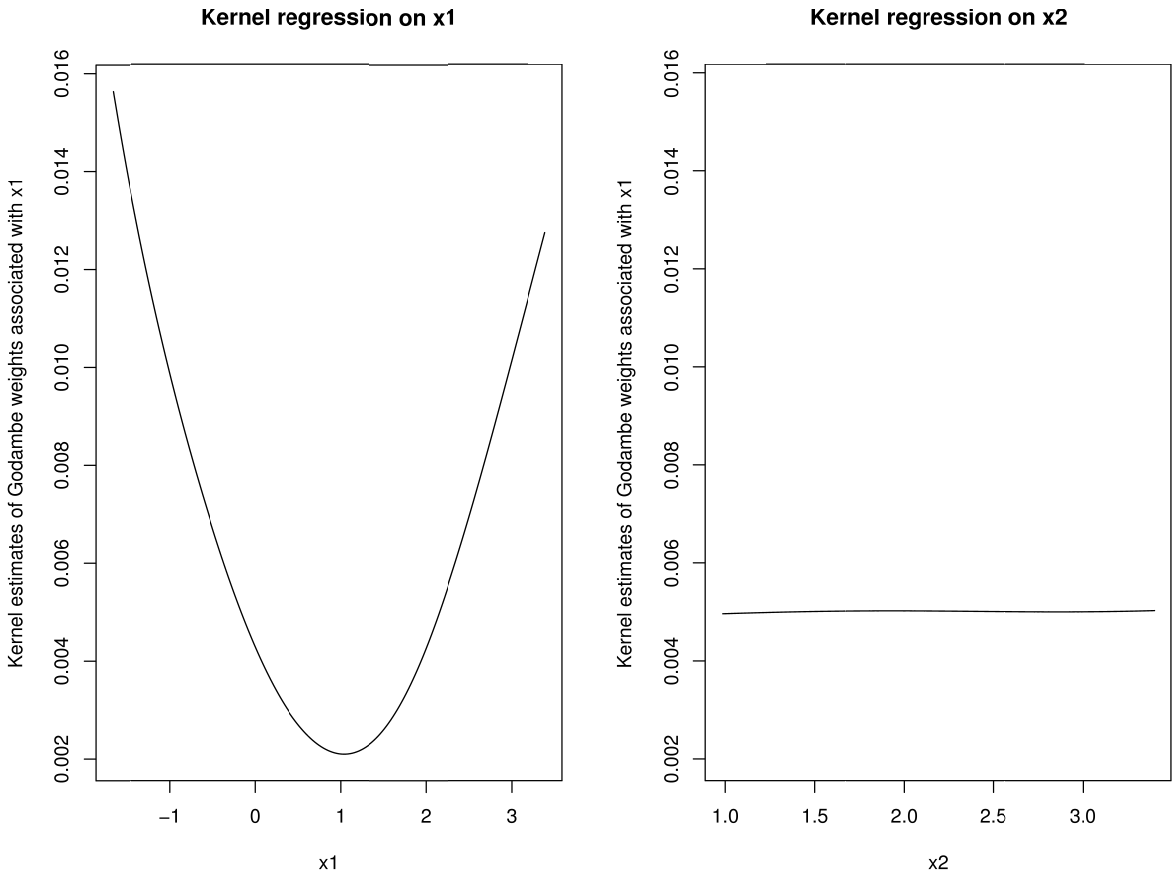

Figure 3.1: The kernel estimates of the weights $w_i^{(1)}$ associated with the covariate x_{i1}, regressing on both x_{i1} and x_{i2}, respectively, with the bandwidth 2. The left panel shows the kernel estimates regressing on x_{i1}, and the right panel shows the kernel estimates regressing on x_{i2}.

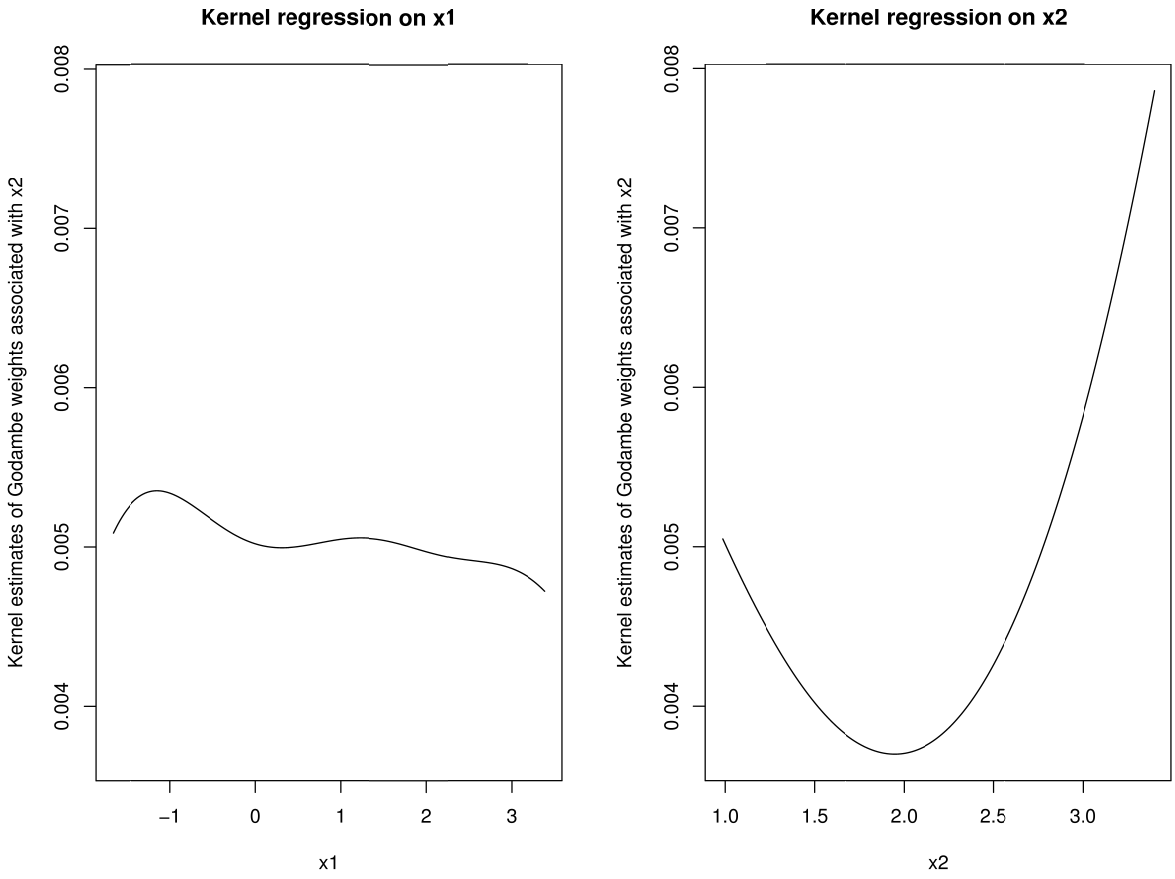

Figure 3.2: The kernel estimates of the weights $w_i^{(2)}$ associated with the covariate x_{i2}, regressing on both x_{i1} and x_{i2}, respectively, with the bandwidth 2. The left panel shows the kernel estimates regressing on x_{i1}, and the right panel shows the kernel estimates regressing on x_{i2}.

and $\boldsymbol{\mu}(\boldsymbol{\beta})$ be an $N \times 1$ vector, defined as

$$\boldsymbol{\mu}(\boldsymbol{\beta})^T = (\boldsymbol{\mu}_1^T, \cdots, \boldsymbol{\mu}_K^T),$$

and $\mathcal{D}(\boldsymbol{\beta})$ be an $N \times p$ matrix, defined as

$$\mathcal{D}(\boldsymbol{\beta})^T = \left(D_1^T, \cdots, D_K^T\right),$$

and $\boldsymbol{\Sigma}(\boldsymbol{\beta}, \boldsymbol{\rho})$ be an $K \times K$ block diagonal matrix, defined as

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}, \boldsymbol{\rho}) = \text{diag}\left\{\Sigma_1, \cdots, \Sigma_K\right\}.$$

Then, the GEE (2.27) can be written as

$$\Psi_K(\boldsymbol{\beta}, \boldsymbol{\rho}) = \frac{1}{\sigma^2}\mathcal{D}(\boldsymbol{\beta})^T\boldsymbol{\Sigma}(\boldsymbol{\beta}, \boldsymbol{\rho})^{-1}\left\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right\} = \mathbf{0}. \tag{3.18}$$

Let $\widehat{\boldsymbol{\beta}}_K$ and $\widehat{\boldsymbol{\rho}}_K$ denote $K^{1/2}$-consistent estimators of the regression coefficient $\boldsymbol{\beta}$ and correlation parameter $\boldsymbol{\rho}$. Let $\widehat{\boldsymbol{\mu}}$, $\widehat{\mathcal{D}}$ and $\widehat{\boldsymbol{\Sigma}}$ denote the vector and matrices $\boldsymbol{\mu}$, $\mathcal{D}$ and $\boldsymbol{\Sigma}$ evaluated at the estimates $\widehat{\boldsymbol{\beta}}_K$ and $\widehat{\boldsymbol{\rho}}_K$. Define the residual vector $\boldsymbol{r}$ by

$$\boldsymbol{r} = \boldsymbol{y} - \widehat{\boldsymbol{\mu}}. \tag{3.19}$$

Note that the residual vector can be written as $\boldsymbol{r}^T = (\boldsymbol{r}_1^T, \cdots, \boldsymbol{r}_K^T)$, where $\boldsymbol{r}_i = \boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i$, for $i = 1, \cdots, K$. The sandwich covariance matrix estimator of $\widehat{\boldsymbol{\beta}}_K$ can be written as

$$ASCOV_s(\widehat{\boldsymbol{\beta}}_K) = \left\{\widehat{\mathcal{D}}^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathcal{D}}\right\}^{-1}\left\{\widehat{\mathcal{D}}^T\widehat{\boldsymbol{\Sigma}}^{-1}\mathcal{R}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathcal{D}}\right\}\left\{\widehat{\mathcal{D}}^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathcal{D}}\right\}^{-1}, \tag{3.20}$$

where $\mathcal{R}$ is a $K \times K$ diagonal block matrix, defined as

$$\mathcal{R} = \text{diag}\left\{\boldsymbol{r}_1\boldsymbol{r}_1^T, \cdots, \boldsymbol{r}_K\boldsymbol{r}_K^T\right\}. \tag{3.21}$$

For $i = 1, \cdots, K$, suppose that the matrix $R_i(\widehat{\boldsymbol{\rho}}_K)$ is positive definite. By the Cholesky decomposition ([31]), this matrix can be decomposed as follows

$$R_i(\widehat{\boldsymbol{\rho}}_K) = \widehat{L}_i\widehat{L}_i^T, \tag{3.22}$$

where $\widehat{L_i}$ is a lower triangular matrix. Define the Pearson residual vector for subject i, which can be written as

$$\boldsymbol{r}_{P,i} = (r_{P,i,1}, \cdots, r_{P,i,n_i})^T, \tag{3.23}$$

where $r_{P,i,j}$ is the j-th Pearson residual for subject i, given by

$$r_{P,ij} = \frac{y_{ij} - \widehat{\mu}_{ij}}{V(\widehat{\mu}_{ij})}, \quad i = 1, \cdots, K, j = 1, \cdots, n_i. \tag{3.24}$$

The Pearson residual vector $\boldsymbol{r}_{P,i}$ for subject i can also be written as $\boldsymbol{r}_{P,i} = \widehat{G}_i^{-1/2}\boldsymbol{r}_i$. Define a transformed residual vector for the i-th subject by

$$\widetilde{\boldsymbol{r}}_i = \widehat{L}_i^{-1}\widehat{G}_i^{-1/2}\boldsymbol{r}_i = \widehat{L}_i^{-1}\boldsymbol{r}_{P,i}, \quad i = 1, \cdots, K, \tag{3.25}$$

where the matrix $\widehat{L_i}$ is given in the decomposition (3.22). In this way, the residuals are so-called "de-correlated" so that they mimic residuals from a standard linear regression, which have constant variance and zero correlation. [31] gave an interesting interpretation of the transformed residuals. For example, the first element of $\widetilde{\boldsymbol{r}}_i$ is the standardized residuals for the first repeated observation (often the baseline measurement). The subsequent residuals represent standardized deviations from the conditional mean of the response given all previous observations. Specifically, the k-th transformed residual $\widetilde{r}_{ik}$ estimates

$$\frac{y_{ik} - E(Y_{ik}|Y_{i1}, \cdots, Y_{i,k-1})}{\sqrt{V(Y_{ik}|Y_{i1}, \cdots, Y_{i,k-1})}}, \quad k = 2, \cdots, n_i.$$

Let $\widetilde{\mathcal{R}}$ be a $K \times K$ diagonal block matrix, with the i-th diagonal matrix $\widetilde{\boldsymbol{r}}_i\widetilde{\boldsymbol{r}}_i^T$. Let $\widehat{U_i}$ be an $n_i \times p$ matrix, defined by

$$\widehat{U}_i = \widehat{L}_i^{-1}\widehat{G}_i^{-1/2}\widehat{D}_i,$$

and let $\widehat{\mathcal{U}}$ be an $N \times p$ matrix, defined by

$$\widehat{\mathcal{U}}^T = \left(\widehat{U}_1^T, \cdots, \widehat{U}_K^T\right).$$

Then the model-based and sandwich covariance matrix estimators, (2.30) and (2.28), can be rewritten as

$$ASCOV_m(\widehat{\boldsymbol{\beta}}_K) = \widetilde{\sigma}^2_m \left(\widehat{\mathcal{U}}^T \widehat{\mathcal{U}}\right)^{-1}, \tag{3.26}$$

and

$$ASCOV_s(\widehat{\boldsymbol{\beta}}_K) = \left(\widehat{\mathcal{U}}^T \widehat{\mathcal{U}}\right)^{-1} \left(\widehat{\mathcal{U}}^T \widetilde{\mathcal{R}} \widehat{\mathcal{U}}\right) \left(\widehat{\mathcal{U}}^T \widehat{\mathcal{U}}\right)^{-1}.$$

Theorem 3.2 *Suppose that, for the regression coefficient β_{j-1}, $j = 1, \cdots, p$, the sandwich variance estimator of the coefficient estimator $\widehat{\beta}_{j-1}$ can be written as*

$$ASVAR_s(\widehat{\beta}_{j-1}) = \widetilde{\sigma}^2_{j-1} a_j, \tag{3.27}$$

where a_j is the j-th diagonal element of the matrix $\left(\widehat{\mathcal{U}}^T \widehat{\mathcal{U}}\right)^{-1}$, given in the model-based covariance matrix estimator (3.26).

Here $\widetilde{\sigma}^2_{j-1}$ can be regarded as an estimator of the dispersion parameter σ^2, written as a sum of quadratic forms in the transformed residuals (3.25), that is,

$$\widetilde{\sigma}^2_{j-1} = \sum_{i=1}^{K} \widetilde{\boldsymbol{r}}_i^T \widehat{W}_i^{(j-1)} \widetilde{\boldsymbol{r}}_i, \quad j = 1, \cdots, p. \tag{3.28}$$

The weight matrices $\widehat{W}_i^{(j-1)}$ can be written as

$$\widehat{W}_i^{(j-1)} = \widehat{H}_{ii} - \widehat{H}_{ii}^{(-j)}, \tag{3.29}$$

where $\widehat{H}_{ii}$ is the i-th diagonal matrix of the $K \times K$ block hat matrix $\widehat{\mathcal{H}}$, defined as

$$\widehat{\mathcal{H}} = \widehat{\mathcal{U}} \left\{\widehat{\mathcal{U}}^T \widehat{\mathcal{U}}\right\}^{-1} \widehat{\mathcal{U}}^T, \tag{3.30}$$

and $\widehat{H}_{ii}^{(-j)}$ is the i-th diagonal matrix of the $K \times K$ block hat matrix $\widehat{\mathcal{H}}^{(-j)}$, defined as

$$\widehat{\mathcal{H}}^{(-j)} = \widehat{\mathcal{U}}_{(-j)} \left(\widehat{\mathcal{U}}_{(-j)}^T \widehat{\mathcal{U}}_{(-j)}\right)^{-1} \widehat{\mathcal{U}}_{(-j)}^T, \tag{3.31}$$

obtained from the sub-matrix $\widehat{\mathcal{U}}_{(-j)}$ *with the* j*-th column deleted from* $\widehat{\mathcal{U}}$*, for* $j = 1, \cdots, p$.

The estimator $\widetilde{\sigma}^2_{j-1}$ *is called the* β_{j-1}**-specific Godambian estimator** *of the dispersion parameter* σ^2.

Compared to the proof of Theorem 3.1, in the context of GEE, the matrix $\widehat{\mathcal{U}}$ plays the same role as the matrix $\widehat{U}$ in GLM. The i-th diagonal element of the hat matrix is an $n_i \times n_i$ matrix, instead of a scalar. Consequently, the weights are matrices. Analogous to the form of a weighted sum of the squared residuals in LM or the squared Pearson residuals in GLM, the Godambian estimator of σ^2 can be written as a sum of quadratic forms in the transformed residuals.

Under certain model misspecification of variance/covariance structure, discrepancy between the sandwich and model-based variance estimators will lead to the discrepancy between the Godambian estimator of the dispersion parameter and its true value or the moment estimator. In the next section, assuming that the mean structure of the response is correctly specified, we construct a test statistic by taking a ratio of the Godambian estimator to its true value or the moment estimator.

Chapter 4

Information Ratio Test

Testing for model misspecification of the mean structure has drawn much attention in the literature. A large class of test statistics has been proposed, including the quasi-likelihood ratio tests, Rao's score tests and Wald's statistics. In addition, [68] developed a chi-squared inference function for testing nested models and a chi-squared regression misspecification test using quadratic inference functions (QIF) for longitudinal data analysis. More details can be found in Section 1.1. However, assessing the adequacy of the variance/covariance assumption is also important. For example, it is a common assumption of LM that the error terms all have equal variances. When this assumption is not met, the loss in efficiency when using ordinary least squares estimation may be substantial. Moreover, the incorrect estimation of standard errors may lead to invalid inference. For regression analysis of count data, overdispersion is often encountered. Although the excess variation has little effect on estimation of the regression coefficients of primary interest, standard errors, tests and confidence intervals may be seriously in error unless it is appropriately taken into account. The GEE method for regression modeling of clustered outcomes allows for specification of covariance structure, including variance function and working correlation matrix. Much work has been done on investigating the impact of misspecifying the correlation structure. [84] pointed out that an inappropriate choice will lead

to inefficient parameter estimation. Also see [79]. Moreover, [85] studied the effects of the variance-function misspecification on estimation of the mean parameters for quantitative responses. Their numerical studies showed that even if the variance function is misspecified, correct choice of the correlation structure may not necessarily improve estimation efficiency. Even though some methods have been considered to test for model misspecification of variance/covariance structure, e.g. heteroscedasticity and overdispersion, there is no systematic statistical test available in the framework of regression analysis using GLM for independent data and GEE for correlated data.

In this chapter, we will develop a statistical test for misspecification of variance/covariance structures. For independent data, let $\sigma^2 V^*(\mu_i)$, $i = 1, \cdots, n$, denote the *true* variance structure of the underlying distribution. Let $V(\cdot)$ denote the *working* unit variance function which is actually used in the quasi-score equation (2.8) in GLM. Analogously, for correlated data, the *true* covariance structure is denoted by $\sigma^2\Sigma_i^* = \sigma^2 G_i^{*1/2} R_i^* G_i^{*1/2}$, where $G_i^* = \text{diag}\{V^*(\mu_i)\}$. Let $\Sigma_i = G_i^{1/2} R_i G_i^{1/2}$ denote the *working* covariance structure, including the *working* unit variance function and the *working* correlation matrix, used in the GEE (2.27). We consider to test the null hypothesis

$$H_0 : V(\cdot) = V^*(\cdot), \tag{4.1}$$

in GLM, or

$$H_0 : \Sigma_i = \Sigma_i^*, \quad i = 1, \cdots K, \tag{4.2}$$

in GEE.

By taking a ratio of the Godambian estimator of the dispersion parameter σ^2, given in the sandwich variance estimators, to its true value or the moment estimator, given in the model-based variance estimators, we propose a statistic, called the *information ratio (IR) statistic*, to test for model misspecification of the variance/covariance structure. When the mean structure is misspecified, testing for misspecifying the second moment is meaningless because the residuals

would be distorted by the incorrect mean function. Therefore, the test proposed in this chapter and model selection of the optimal variance/covariance structure suggested in Chapter 5 are constructed based on the assumption that the mean structure has been chosen and correctly specified.

4.1 Asymptotic Distributions of the Dispersion Parameter Estimators

Both of the Godambian estimators and moment estimators of the dispersion parameter are functions of the Pearson residuals $r_{P,i}$ (2.12) in GLM, or the transformed residuals $\widetilde{\boldsymbol{r}}_i$ (3.25) in GEE.

4.1.1 Generalized Linear Models

Pearson residuals are based on the Pearson goodness-of-fit statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

In the literature of GLM, the Pearson residuals have been widely studied. See [16], [57], [58], [66] and [15].

Let $\boldsymbol{\beta}_*$ denote the limiting value of the sequence of the estimators

$$\left\{\widehat{\boldsymbol{\beta}}_n, n = 1, 2, \cdots\right\},$$

namely, $\widehat{\boldsymbol{\beta}}_n \to_p \boldsymbol{\beta}_*$, as $n \to \infty$.

Note that since the estimator $\widehat{\boldsymbol{\beta}}_n$ is consistent, the limiting value $\boldsymbol{\beta}_*$ equals to the true value of the parameter $\boldsymbol{\beta}$. Define

$$e_i = y_i - \mu_i^* = y_i - h(\boldsymbol{x}_i^T \boldsymbol{\beta}_*), \quad i = 1, \cdots, n.$$

Note that e_i's are independent random variables with mean 0 and variance $V^*(\mu_i^*)$. Let $\boldsymbol{e} = (e_1, \cdots, e_n)^T$. Under mild regularity conditions, straightforward asymptotic expansion of the residual vector $\boldsymbol{r} = \boldsymbol{y} - \widehat{\boldsymbol{\mu}}$ yields

$$\boldsymbol{r} = \left\{ \boldsymbol{I}_n - \Delta_* X \left(X^T \Delta_* \mathcal{V}_*^{-1} \Delta_* X \right)^{-1} X^T \Delta_* \mathcal{V}_*^{-1} \right\} \boldsymbol{e} + o_p(1),$$

where Δ_* and $\mathcal{V}_*$ are the resulting matrices which substitute the estimate $\widehat{\boldsymbol{\beta}}_n$ in the matrices $\widehat{\Delta}$ and $\widehat{\mathcal{V}}$, (2.13) and (2.14), with its limiting value $\boldsymbol{\beta}_*$. See [20]. Then,

(i) the expectation of the Pearson residual $r_{P,i} = (y_i - \widehat{\mu}_i)/\sqrt{V(\widehat{\mu}_i)}$ for the i-th observation is approximately 0, and

(ii) from [25], it can be shown that, for large sample size, the covariance matrix of the Pearson residual vector $\boldsymbol{r}_P = (r_{P,1}, \cdots, r_{P,n})^T$ may be approximated by $\sigma^2(\boldsymbol{I}_n - H_*)\Omega_*(\boldsymbol{I}_n - H_*)$, where the matrix H_* is given by

$$H_* = \mathcal{V}_*^{-1/2} \Delta_* X \left(X^T \Delta_* \mathcal{V}_*^{-1} \Delta_* X \right)^{-1} X^T \Delta_* \mathcal{V}_*^{-1/2}, \tag{4.3}$$

and the matrix Ω_* is an $n \times n$ diagonal matrix with the i-th diagonal element

$$\omega_i^* = V^*(\mu_i^*)/V(\mu_i^*),$$

for $i = 1, \cdots, n$. Note that under the null hypothesis H_0 (4.1), $V^*(\mu_i^*) = V(\mu_i^*)$, i.e., $\omega_i^* = 1$, and consequently the matrix $\Omega_* = \boldsymbol{I}_n$.

Suppose that $\boldsymbol{\epsilon}_P$ is an n-variate random vector with mean $\mathbf{0}$ and covariance matrix $\sigma^2\Omega_*$. Then, for large sample size, the Pearson residual vector can be approximated by

$$\boldsymbol{r}_P \simeq (\boldsymbol{I}_n - H_*)\, \boldsymbol{\epsilon}_P. \tag{4.4}$$

For Gaussian responses, the Pearson residual vector $\boldsymbol{r}_P$ is asymptotically multivariate normal distributed. However, this statement is not always true for non-Gaussian responses, e.g., binary data, count data and etc. [47] discussed so-called "small dispersion asymptotics". In general, when the dispersion parameter is small, the Pearson residuals are asymptotically normally distributed. Also

see [48] and [77]. But for the Poisson and binomial cases, this small-dispersion asymptotic normality needs to be modified because the dispersion parameter σ^2 is assumed to be 1 when no over-dispersion or under-dispersion is concerned. For the Poisson case, the Pearson residuals

$$r_{P,i} = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}} \to^d N(0,1), \quad \text{as } \mu \to \infty.$$

For the binomial case, the Pearson residuals $r_{P,i}$ are asymptotically normal as the number of Bernoulli trials m approaches infinity. In this sense, $1/m$ sometimes is regarded as the "dispersion" parameter. Note that because of this, the Pearson residuals in the binary case would not have asymptotic normal distribution.

Godambian estimators

The Godambian estimators of σ^2 can be re-written as quadratic forms in the Pearson residuals $r_{P,i}$ (2.12). Specifically, for $j = 1, \cdots, p$, the β_{j-1}-specific Godambian estimator of σ^2 can be written as

$$\widetilde{\sigma}^2_{j-1} = \boldsymbol{r}_P^T \widehat{\boldsymbol{W}}^{(j-1)} \boldsymbol{r}_P, \tag{4.5}$$

where the weight matrix $\widehat{\boldsymbol{W}}^{(j-1)}$ is an $n \times n$ diagonal matrix with the i-th diagonal element $\widehat{w}_i^{(j-1)}$, given in (3.6).

Let $\boldsymbol{W}_*^{(j-1)}$ be an $n \times n$ diagonal matrix with the i-th diagonal element $w_{i,*}^{(j-1)}$, which substitutes the estimate $\widehat{\boldsymbol{\beta}}_n$, in the weights $\widehat{w}_i^{(j-1)}$ (3.6), with $\boldsymbol{\beta}_*$.

Lemma 4.1 *Under the null hypothesis H_0 (4.1), for $j = 1, \cdots, p$, the β_{j-1}-specific Godambian estimator (4.5) has expectation*

$$E\left(\widetilde{\sigma}^2_{j-1}\right) \simeq \left(\sum_{k=1}^{n} \lambda_{k,*}^{(j-1)}\right)\sigma^2,$$

where $\lambda_{k,*}^{(j-1)}$, $k = 1, \cdots, n$, *are the eigenvalues of the matrix*

$$(\boldsymbol{I}_n - H_*)\, \boldsymbol{W}_*^{(j-1)}\, (\boldsymbol{I}_n - H_*)\,.$$

Lemma 4.1 is proved in the appendix. Note that

$$\sum_{k=1}^{n} \lambda_{k,*}^{(j-1)} = tr\left\{(\boldsymbol{I}_n - H_*)\, \boldsymbol{W}_*^{(j-1)}\, (\boldsymbol{I}_n - H_*)\right\},$$

for $j = 1, \cdots, p$, where $tr\{\cdot\}$ denotes the trace of a matrix. Under the null hypothesis H_0 (4.1),

$$E(\widetilde{\sigma}_{j-1}^2/\sigma^2) \simeq \sum_{i=1}^{n} w_{i,*}^{(j-1)} - \sum_{i=1}^{n} h_{ii}^* w_{i,*}^{(j-1)} = 1 + O(1/n),$$

because $h_{ii}^* = O(1/n)$ and $w_{i,*}^{(j-1)} = O(1/n)$, following from $\sum_{i=1}^{n} h_{ii}^* = p$ and $\sum_{i=1}^{n} w_{i,*}^{(j-1)} = 1$, where h_{ii}^* is the i-th diagonal element of the matrix H_*. Then, with the order of $O(1/n)$, the Godambian estimator $\widetilde{\sigma}_{j-1}^2$ is an asymptotically unbiased estimator of the dispersion parameter σ^2, under the null hypothesis H_0.

For finite sample size, with a bias correction, an unbiased version of the β_{j-1}-specific Godambian estimator of σ^2 is proposed as

$$\widetilde{\sigma}_{j-1,u}^2 = \sum_{i=1}^{n} \frac{\widehat{w}_i^{(j-1)}}{1 - \sum_{k=1}^{n} \widehat{w}_k^{(j-1)} \widehat{h}_{kk}} r_{P,i}^2 = \boldsymbol{r}_P^T \widehat{\boldsymbol{W}}_u^{(j-1)} \boldsymbol{r}_P, \tag{4.6}$$

where $\widehat{\boldsymbol{W}}_u^{(j-1)}$ is an $n \times n$ diagonal matrix with the i-th diagonal element $\widehat{w}_i^{(j-1)}/(1 - \sum_{k=1}^{n} \widehat{w}_k^{(j-1)} \widehat{h}_{kk})$, for $j = 1, \cdots, p$.

In Section 3.1.1, we have discussed, in LM, the fact that the weights in the β_{j-1}-specific Godambian estimator characterize the influence from the corresponding covariate $x_{i,j-1}$'s. To incorporate the overall impact from all the covariates, a new Godambian estimator is defined by

$$\widetilde{\sigma}_{pool}^2 = \sum_{i=1}^{n} \widehat{w}_i^{pool} r_{P,i}^2 = \boldsymbol{r}_P^T \widehat{\boldsymbol{W}}^{pool} \boldsymbol{r}_P, \tag{4.7}$$

where $\widehat{\boldsymbol{W}}^{pool}$ is an $n \times n$ diagonal matrix with the i-th diagonal element $\widehat{w}_i^{pool} = \widehat{h}_{ii}/p$. Here, since the influences from all the covariates are pooled in the weights $\widehat{w}_i^{pool}$, this estimator is called the *pooled Godambian estimator* of the dispersion parameter σ^2. We can also show that the pooled Godambian estimator $\widetilde{\sigma}^2_{pool}$ is asymptotically unbiased with the order of $O(1/n)$, under the null hypothesis H_0. For finite sample size, an unbiased pooled Godambian estimator of σ^2 is defined by

$$\widetilde{\sigma}^2_{pool,u} = \sum_{i=1}^{n} \frac{\widehat{w}_i^{pool}}{1 - \sum_{k=1}^{n} \widehat{h}_{kk}^2/p} r_{P,i}^2 = \boldsymbol{r}_P^T \widehat{\boldsymbol{W}}_u^{pool} \boldsymbol{r}_P, \tag{4.8}$$

where $\widehat{\boldsymbol{W}}_u^{pool}$ is an $n \times n$ diagonal matrix with the i-th diagonal element $\widehat{h}_{ii}/(p - \sum_{k=1}^{n} \widehat{h}_{kk}^2)$.

Moment estimator

The moment estimator (2.17) of the dispersion parameter σ^2 can be written as a quadratic form in Pearson residuals

$$\widehat{\sigma}^2_m = \boldsymbol{r}_P^T \left(\frac{1}{n-p} \boldsymbol{I}_n \right) \boldsymbol{r}_P. \tag{4.9}$$

It can be shown that under the null hypothesis H_0 (4.1), the moment estimator $\widehat{\sigma}^2_m$ is approximately a Pearson χ^2 statistic, i.e.,

$$\frac{(n-p)\widehat{\sigma}^2_m}{\sigma^2} \sim \chi^2_{n-p}.$$

Also see [59]. Moreover, it is approximately an unbiased estimator of the dispersion parameter σ^2.

4.1.2 Generalized Estimating Equations

In the context of GEE, let $\boldsymbol{\beta}_*$ be the limiting value of the sequence of the estimators $\left\{\widehat{\boldsymbol{\beta}}_K, K = 1, 2, \cdots\right\}$, that is,

$$\widehat{\boldsymbol{\beta}}_K \to_p \boldsymbol{\beta}_* \quad \text{as} \quad K \to \infty.$$

The limiting value $\boldsymbol{\beta}_*$ equals to the true value due to the consistency of the estimator $\widehat{\boldsymbol{\beta}}_K$.

It is rarely encountered that an arbitrary working correlation matrix with given values of the correlation parameters is assumed, so we only consider the cases where the values of the correlation parameters in the working correlation matrix are unknown. For the sequence of the estimators $\{\widehat{\boldsymbol{\rho}}_K, K = 1, 2, \cdots\}$, suppose that there exists a limiting value $\boldsymbol{\rho}_*$, namely,

$$\widehat{\boldsymbol{\rho}}_K \to_p \boldsymbol{\rho}_* \quad \text{as} \quad K \to \infty.$$

Note that if the working correlation matrix correctly specifies the true correlation structure, the limiting value $\boldsymbol{\rho}_*$ is equal to the true value of the correlation parameter involved in the true correlation structure. On the other hand, if the working correlation matrix departs from the true correlation structure, the estimator of the "working" correlation parameter converges to a certain value, which is unlikely to be equal to the true value of the "true" correlation parameter.

Similarly to GLM, suppose that $\widetilde{\boldsymbol{\epsilon}}$ is an N-variate random vector with mean $\mathbf{0}$ and covariance matrix $\sigma^2\boldsymbol{\Omega}_*$, where $\boldsymbol{\Omega}_*$ is a $K \times K$ diagonal block matrix with i-th diagonal matrix

$$\Omega_i^* = L_{i,*}^{-1} G_{i,*}^{-1/2} G_i^{*1/2} R_i^* G_i^{*1/2} G_{i,*}^{-1/2} L_{i,*}^{-T}. \tag{4.10}$$

The matrix $G_{i,*}$ substitutes the estimate $\widehat{\boldsymbol{\beta}}_K$ in the matrix $\widehat{G}_i$ with its true value $\boldsymbol{\beta}_*$, and under the null hypothesis H_0 (4.2), $G_{i,*} = G_i^*$. The matrix $L_{i,*}$ is an $n_i \times n_i$ lower triangular matrix such that $L_{i,*}L_{i,*}^T = R_i(\boldsymbol{\rho}_*)$, and under the null

hypothesis H_0 (4.2), $L_{i,*}L_{i,*}^T = R_i(\boldsymbol{\rho}_*) = R_i^*$. Thus, the matrix $\boldsymbol{\Omega}_*$ is equivalent to an identity matrix $\boldsymbol{I}_N$ under the null hypothesis H_0 (4.2). For large sample size, the transformed residual vector $\widetilde{\boldsymbol{r}}$ can be approximated by

$$\widetilde{\boldsymbol{r}} \simeq (\boldsymbol{I}_N - \mathcal{H}_*)\widetilde{\boldsymbol{\epsilon}}, \tag{4.11}$$

where $\mathcal{H}_*$ is the matrix substituting the estimate $\widehat{\boldsymbol{\beta}}_K$ and $\widehat{\boldsymbol{\rho}}_K$ with its limiting values in the hat matrix $\widehat{\mathcal{H}}$ (3.30). Similarly to GLM, in the case of Gaussian responses, or non-Gaussian responses for small dispersion, the transformed residual vector $\widetilde{\boldsymbol{r}}$ is asymptotically normal distributed.

Godambian estimator

For $j = 1, \cdots, p$, the β_{j-1}-specific Godambian estimator of the dispersion parameter σ^2 can be written as a quadratic form in the transformed residuals $\widetilde{\boldsymbol{r}}_i$, i.e.,

$$\widetilde{\sigma}_{j-1}^2 = \widetilde{\boldsymbol{r}}^T \widehat{\mathcal{W}}^{(j-1)} \widetilde{\boldsymbol{r}}, \tag{4.12}$$

where $\widehat{\mathcal{W}}^{(j-1)}$ is a $K \times K$ diagonal block matrix with the i-th diagonal matrix $\widehat{W}_i^{(j-1)}$ given in (3.29).

Lemma 4.2 *Under the null hypothesis H_0 (4.2), for $j = 1, \cdots, p$, the β_{j-1}-specific Godambian estimator (4.12) has expectation*

$$E\left(\widetilde{\sigma}_{j-1}^2\right) \simeq \left(\sum_{k=1}^{N} \lambda_{k,*}^{(j-1)}\right)\sigma^2,$$

where $\lambda_{k,*}^{(j-1)}$, $k = 1, \cdots, N$, are the eigenvalues of the matrix

$$(\boldsymbol{I}_N - \mathcal{H}_*)\, \mathcal{W}_*^{(j-1)}\, (\boldsymbol{I}_N - \mathcal{H}_*)\,.$$

Lemma 4.2 is proved in the appendix. Similarly to GLM, the β_{j-1}-specific Godambian estimator is an asymptotically unbiased estimator of σ^2 to the order of $O(1/K)$, under the null hypothesis H_0 (4.2). Let

$$b_{j-1} = tr\left\{\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\widehat{\mathcal{W}}^{(j-1)}\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\right\}.$$

For finite sample size, an unbiased β_{j-1}-specific Godambian estimator is defined by

$$\widetilde{\sigma}^2_{j-1,u} = \widetilde{\boldsymbol{r}}^T\widehat{\mathcal{W}}^{(j-1)}_u\widetilde{\boldsymbol{r}} = \sum_{i=1}^{K}\widetilde{\boldsymbol{r}}_i^T\,\widehat{W}^{(j-1)}_{i,u}\widetilde{\boldsymbol{r}}_i, \tag{4.13}$$

where $\widehat{\mathcal{W}}^{(j-1)}_u$ is a $K \times K$ diagonal block matrix with the i-th diagonal matrix $\widehat{W}^{(j-1)}_{i,u} = \widehat{W}^{(j-1)}_i/b_{j-1}$, for $j = 1, \cdots, p$.

Moreover, the pooled Godambian estimator of σ^2 is defined by

$$\widetilde{\sigma}^2_{pool} = \widetilde{\boldsymbol{r}}^T\widehat{\mathcal{W}}^{pool}\widetilde{\boldsymbol{r}} = \sum_{i=1}^{K}\widetilde{\boldsymbol{r}}_i^T\,\widehat{W}^{pool}_i\widetilde{\boldsymbol{r}}_i, \tag{4.14}$$

where $\widehat{\mathcal{W}}^{pool}$ is a $K \times K$ diagonal block matrix with the i-th diagonal matrix $\widehat{W}^{pool}_i = \widehat{H}_{ii}/p$. Let

$$b = tr\left\{\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\widehat{\mathcal{W}}^{pool}\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\right\}.$$

The unbiased pooled Godambian estimator of σ^2 is defined by

$$\widetilde{\sigma}^2_{pool,u} = \widetilde{\boldsymbol{r}}^T\widehat{\mathcal{W}}^{pool}_u\widetilde{\boldsymbol{r}} = \sum_{i=1}^{K}\widetilde{\boldsymbol{r}}_i^T\,\widehat{W}^{pool}_{i,u}\widetilde{\boldsymbol{r}}_i, \tag{4.15}$$

where $\widehat{\mathcal{W}}^{pool}_u$ is a $K \times K$ diagonal block matrix with the i-th diagonal matrix $\widehat{W}^{pool}_{i,u} = \widehat{W}_i/b$.

Moment estimators

In the context of GEE, the moment estimator (2.31) of the dispersion parameter σ^2 can also be written as a quadratic form in the transformed residuals

$$\widehat{\sigma}^2_m = \boldsymbol{r}_P^T \left(\frac{1}{N-p} \boldsymbol{I}_N \right) \boldsymbol{r}_P = \widetilde{\boldsymbol{r}}^T \left(\frac{1}{N-p} \widehat{\mathcal{L}}^T \widehat{\mathcal{L}} \right) \widetilde{\boldsymbol{r}} = \widetilde{\boldsymbol{r}}^T \widehat{\mathcal{W}}_P \widetilde{\boldsymbol{r}}, \tag{4.16}$$

where $\boldsymbol{r}_P = (\boldsymbol{r}_{P,1}^T, \cdots, \boldsymbol{r}_{P,K}^T)^T$ is a $K \times 1$ Pearson vector with the i-th element $\boldsymbol{r}_{P,i}$, given in (3.23), and $\widetilde{\boldsymbol{r}} = (\widetilde{\boldsymbol{r}}_1^T, \cdots, \widetilde{\boldsymbol{r}}_K^T)^T$ is a $K \times 1$ transformed residual vector with the i-th element $\widetilde{\boldsymbol{r}}_i$, given in (3.25), and $\widehat{\mathcal{W}}_P$ is a $K \times K$ diagonal block matrix with the i-th diagonal matrix $\widehat{L}_i^T \widehat{L}_i/(N-p)$, with the matrix $\widehat{L}_i$ given in (3.22). Under the null hypothesis H_0 (4.2), this estimator is approximately a Pearson χ^2_{N-p} statistic. However, under the null hypothesis H_0 (4.2), the expectation of this "Pearson" moment estimator is given by

$$E\left(\widehat{\sigma}^2_m\right) \simeq \sigma^2 tr\left\{ (\boldsymbol{I}_N - \mathcal{H}_*) \mathcal{W}_{P,*} (\boldsymbol{I}_N - \mathcal{H}_*) \right\} = \sigma^2 + O(1/K).$$

Then, with the order of $O(1/K)$, the "Pearson" moment estimator is asymptotically unbiased, under the null hypothesis H_0. For finite sample size, with a bias correction, the unbiased "Pearson" moment estimator of σ^2 is defined by

$$\widehat{\sigma}^2_{m,u} = \widetilde{\boldsymbol{r}}^T \left(\widehat{\mathcal{W}}_P / m_P \right) \widetilde{\boldsymbol{r}}, \tag{4.17}$$

where $m_P = tr\left\{ \left(\boldsymbol{I}_N - \widehat{\mathcal{H}} \right) \widehat{\mathcal{W}}_P \left(\boldsymbol{I}_N - \widehat{\mathcal{H}} \right) \right\}$.

However, Liang and Zeger pointed out that any consistent estimate of σ^2 is admissible. Due to the "de-correlation" property of the transformed residuals, a new moment estimator of the dispersion parameter σ^2 in GEE is defined by

$$\widehat{\sigma}^2_{tr} = \widetilde{\boldsymbol{r}}^T \widetilde{\boldsymbol{r}}/(N-p) = \frac{1}{N-p} \sum_{i=1}^{K} \widetilde{\boldsymbol{r}}_i^T \widetilde{\boldsymbol{r}}_i. \tag{4.18}$$

Under the null hypothesis H_0 (4.2), this new moment estimator $\widehat{\sigma}^2_{tr}$ is approximately an unbiased estimator of σ^2, and moreover, is approximately distributed as a χ^2_{N-p} statistic.

4.2 Information Ratio Statistics

As shown in Chapter 3, the discrepancy between the sandwich variance estimators and model-based variance estimators reduces to a discrepancy between the Godambian estimator of the dispersion parameter (in the sandwich variance estimators) and the true value, if known, or the moment estimator, otherwise (in the model-based variance estimators). Thus, a large difference between the Godambian estimator and its true value or the moment estimator indicates a certain model misspecification of the variance/covariance structure. Here, we will propose a statistic by taking a ratio of the Godambian estimator of σ^2 to its true value, if known, or the moment estimator, otherwise. Due to the ratio construction, the statistics proposed are called the *information ratio (IR) statistics*. In addition, asymptotic distributions of the statistics are discussed.

4.2.1 If the true value of σ^2 is known

In some cases, the true value of the dispersion parameter is assumed to be known. For example, the variance structure of binary data is usually assumed to be $\mu(1-\mu)$, which implies that the dispersion parameter σ^2 is known to be 1. Generally, if the true value of σ^2 is known, define the β_{j-1}***-specific information ratio statistic*** by taking a ratio of the unbiased β_{j-1}-specific Godambian estimator of σ^2 to its true value

$$IR_{j-1} = \frac{\widetilde{\sigma}^2_{j-1,u}}{\sigma^2}, \quad j = 1, \cdots, p, \tag{4.19}$$

where $\widetilde{\sigma}^2_{j-1,u}$ is the unbiased β_{j-1}-specific Godambian estimator, (4.6) or (4.13), of σ^2. Analogously, the ***pooled information ratio statistic*** is defined by

$$IR_{pool} = \frac{\widetilde{\sigma}^2_{pool,u}}{\sigma^2}, \tag{4.20}$$

where $\widetilde{\sigma}^2_{pool,u}$ is the unbiased pooled Godambian estimator, (4.8) or (4.15), of σ^2.

By Lemma 4.1 and 4.2, these information ratio statistics IR_{j-1} and IR_{pool} can be approximated by quadratic forms in random variables. [46] discussed some central limit theorems for quadratic forms. By normalization, pivotal statistics, standardized information ratio statistics, are proposed as follows.

Theorem 4.1 *In the context of GLM, under the null hypothesis H_0 (4.1),*

(i) the standardized β_{j-1}-specific information ratio statistic

$$IR^s_{j-1} = \frac{IR_{j-1} - 1}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\lambda}_k^{(j-1)}\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad as\ n \to \infty, \tag{4.21}$$

where $\widehat{\lambda}_k^{(j-1)}$ are the eigenvalues of the matrix $\left(\boldsymbol{I}_n - \widehat{H}\right)\widehat{\boldsymbol{W}}_u^{(j-1)}\left(\boldsymbol{I}_n - \widehat{H}\right)$, and $\widehat{\boldsymbol{W}}_u^{(j-1)}$ is given in (4.6), for $j = 1, \cdots, p$;

(ii) the standardized pooled information ratio statistic

$$IR^s_{pool} = \frac{IR_{pool} - 1}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\lambda}_k^{pool}\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad as\ n \to \infty, \tag{4.22}$$

where $\widehat{\lambda}_k^{pool}$ are the eigenvalues of the matrix $\left(\boldsymbol{I}_n - \widehat{H}\right)\widehat{\boldsymbol{W}}_u^{pool}\left(\boldsymbol{I}_n - \widehat{H}\right)$, and $\widehat{\boldsymbol{W}}^{pool}$ is given in (4.8).

In the context of GEE, under the null hypothesis H_0 (4.2),

(iii) the standardized β_{j-1}-specific information ratio statistic

$$IR^s_{j-1} = \frac{IR_{j-1} - 1}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\lambda}_k^{(j-1)}\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad as\ n \to \infty, \tag{4.23}$$

where $\widehat{\lambda}_k^{(j-1)}$ are the eigenvalues of the matrix $\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\widehat{\mathcal{W}}_u^{(j-1)}\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)$, and $\widehat{\mathcal{W}}_u^{(j-1)}$ is given in (4.13), for $j = 1, \cdots, p$;

(iv) the standardized pooled information ratio statistic

$$IR^s_{pool} = \frac{IR_{pool} - 1}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\lambda}_k^{pool}\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad as\ n \to \infty, \tag{4.24}$$

where $\widehat{\lambda}_k^{pool}$ are the eigenvalues of the matrix $\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\widehat{\mathcal{W}}_u^{pool}\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)$, and $\widehat{\mathcal{W}}_u^{pool}$ is given in (4.15).

The proof of Theorem 4.1 is provided in the appendix.

4.2.2 If the true value of σ^2 is unknown

If the true value of σ^2 is unknown, the dispersion parameter σ^2 is estimated by a moment estimator. Then, information ratio statistics can be defined by taking ratios of the Godambian estimators to the moment estimators of σ^2.

Theorem 4.2 *In the context of GLM, under the null hypothesis H_0 (4.1),*

(i) the standardized β_{j-1}-specific information ratio statistic

$$IR^s_{j-1} = \frac{\widetilde{\sigma}^2_{j-1,u}/\widehat{\sigma}^2_m - 1}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\tau}_k^{(j-1)}\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad as\ n \to \infty, \tag{4.25}$$

where $\widetilde{\sigma}^2_{j-1,u}$ is the unbiased β_{j-1}-specific Godambian estimator (4.6), $\widehat{\sigma}^2_m$ is the moment estimator (4.9), and $\widehat{\tau}_k^{(j-1)}$ are the eigenvalues of the matrix

$$\left(\boldsymbol{I}_n - \widehat{H}\right)\left(\widehat{\boldsymbol{W}}_u^{(j-1)} - \frac{1}{n-p}\boldsymbol{I}_n\right)\left(\boldsymbol{I}_n - \widehat{H}\right),$$

for $j = 1, \cdots, p$;

(ii) the standardized pooled information ratio statistic

$$IR^s_{pool} = \frac{\widetilde{\sigma}^2_{pool,u}/\widehat{\sigma}^2_m - 1}{\sqrt{2\sum_{k=1}^n \left[\widehat{\tau}_k^{pool}\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad as\ n \to \infty, \tag{4.26}$$

where $\widetilde{\sigma}^2_{pool,u}$ is the unbiased pooled Godambian estimator (4.8), $\widehat{\sigma}^2_m$ is the moment estimator, and $\widehat{\tau}_k^{pool}$ are the eigenvalues of the matrix

$$\left(\boldsymbol{I}_n - \widehat{H}\right)\left(\widehat{\boldsymbol{W}}_u^{pool} - \frac{1}{n-p}\boldsymbol{I}_n\right)\left(\boldsymbol{I}_n - \widehat{H}\right).$$

Theorem 4.2 is proved in the appendix.

In the context of GEE, the dispersion parameter σ^2 can be estimated by either the "Pearson" moment estimator $\widehat{\sigma}^2_{m,u}$ (4.16) or the "transformed" moment estimator $\widehat{\sigma}^2_{tr}$ (4.18). Correspondingly, the information ratio statistics are defined by taking ratios of the Godambian estimators to the "Pearson" moment estimator or "transformed" moment estimator of σ^2.

Theorem 4.3 *In the context of GEE, under the null hypothesis H_0 (4.2),*

(i) the standardized β_{j-1}-specific information ratio statistic

$$IR^s_{j-1} = \frac{\widetilde{\sigma}^2_{j-1,u}/\widehat{\sigma}^2_{m,u} - 1}{\sqrt{2\sum_{k=1}^n \left[\widehat{\tau}_k^{(j-1)}\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad as\ n \to \infty, \tag{4.27}$$

where $\widetilde{\sigma}^2_{j-1,u}$ is the unbiased β_{j-1}-specific Godambian estimator (4.13), $\widehat{\sigma}^2_{m,u}$ is the unbiased "Pearson" moment estimator (4.17), and $\widehat{\tau}_k^{(j-1)}$ are the eigenvalues of the matrix

$$\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\left(\widehat{\mathcal{W}}_u^{(j-1)} - \widehat{\mathcal{W}}_P/m_P\right)\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right),$$

and $\widehat{\mathcal{W}}_P$ is given in (4.17), for $j = 1, \cdots, p$;

(ii) the standardized pooled information ratio statistic

$$IR^s_{pool} = \frac{\widetilde{\sigma}^2_{pool,u}/\widehat{\sigma}^2_{m,u} - 1}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\tau}_k^{pool}\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad as\ n \to \infty, \tag{4.28}$$

where $\widetilde{\sigma}^2_{pool,u}$ is the unbiased pooled Godambian estimator (4.15), $\widehat{\sigma}^2_{m,u}$ is the unbiased "Pearson" moment estimator, and $\widehat{\tau}_k^{pool}$ are the eigenvalues of the matrix

$$\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\left(\widehat{\mathcal{W}}_u^{pool} - \widehat{\mathcal{W}}_{_P}/m_{_P}\right)\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right).$$

Theorem 4.4 *In the context of GEE, under the null hypothesis H_0 (4.2),*

(i) the standardized β_{j-1}-specific information ratio statistic

$$IR^s_{j-1} = \frac{\widetilde{\sigma}^2_{j-1,u}/\widehat{\sigma}^2_{tr} - 1}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\tau}_k^{(j-1)}\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad as\ n \to \infty, \tag{4.29}$$

where $\widehat{\sigma}^2_{tr}$ is the "transformed" moment estimator (4.18), and $\widehat{\tau}_k^{(j-1)}$ are the eigenvalues of the matrix

$$\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\left(\widehat{\mathcal{W}}_u^{(j-1)} - \frac{1}{N-p}\boldsymbol{I}_N\right)\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right),$$

for $j = 1, \cdots, p$;

(ii) the standardized pooled information ratio statistic

$$IR^s_{pool} = \frac{\widetilde{\sigma}^2_{pool,u}/\widehat{\sigma}^2_{tr} - 1}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\tau}_k^{pool}\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad as\ n \to \infty, \tag{4.30}$$

where $\widehat{\sigma}^2_{tr}$ is the "transformed" moment estimator, and $\widehat{\tau}_k^{pool}$ are the eigenvalues of the matrix

$$\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\left(\widehat{\mathcal{W}}_u^{pool} - \frac{1}{N-p}\boldsymbol{I}_N\right)\left(\boldsymbol{I}_N - \widehat{H}\right).$$

The proof of Theorem 4.3 and 4.4 is similar to the proof of Theorem 4.2.

4.3 Application of Information Ratio Statistics

4.3.1 Test for heteroscedasticity in linear regression models

As discussed in the beginning of this chapter, in LM, it is common to assume that all the error terms have equal variance. Consider testing the null hypothesis of homoscedasticity,

$$H_0 : Var(Y_1) = \cdots = Var(Y_n) = \sigma^2. \tag{4.31}$$

In Simulation 4.1, we investigate the asymptotic distributions of the proposed information ratio statistics under the null hypothesis H_0 (4.31). In addition, in Simulations 4.2 - 4.7, we compare the power of the proposed IR statistics with that of the White's IM test, under different scenarios of heteroscedasticity.

A data sample $\{(y_i; x_{i1}, x_{i2}), i = 1, \cdots, n\}$ is generated from the following centered linear regression model:

$$y_i = \beta_0 + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + e_i \quad i = 1, \cdots, n,$$

where x_{i1} and x_{i2} are both generated from the Gaussian distributions $N(0, 1)$. Here, $\bar{x}_1 = n^{-1}\sum_i x_{i1}$ and $\bar{x}_2 = n^{-1}\sum_i x_{i2}$. The true values of the regression coefficients are $\beta_0 = 1$, $\beta_1 = 2$ and $\beta_2 = 2$. The sample size n is set to be 20, 100, 200 and 400. For each sample size, we generate 5000 replicates. To stress that the covariates are here fixed, for a given sample size n, the same covariates values were used for each replicate.

Note that because the covariate variables are centered, the unbiased β_0-specific Godambian estimator of the variance parameter σ^2 is identical to the moment estimator, i.e., $\widetilde{\sigma}^2_{0,u} = \widehat{\sigma}^2_m$. As a result, if we assume that the true value of σ^2 is unknown, the β_0-specific IR statistic $IR_0 = 1$, so it is not included in the comparison.

Simulation 4.1 To evaluate the null distributions, the error terms e_i's are generated from a Gaussian distribution with mean 0 and constant variance 0.25. First, we consider the case that the true value of the variance parameter $\sigma^2 = 0.25$ is known. Figures 4.1, 4.2, 4.3 and 4.4 display the kernel density estimates of the standardized IR statistics IR^s_{pool}, IR^s_0, IR^s_1 and IR^s_2, over different sample sizes. In addition, we also assume that the true value of σ^2 is unknown. Figures 4.5, 4.6 and 4.7 display the kernel density estimates of the standardized IR statistics, IR^s_{pool}, IR^s_1 and IR^s_2. Table 4.1 reports their empirical type I errors, the proportions of rejecting the null hypothesis (4.31), among the 5000 replicates, at the significance level 5%, when the true value of σ^2 is either known or unknown.

Conclusion.

As shown in these figures, under the null hypothesis H_0 (4.31), the IR statistics are heavily right skewed with small sample sizes, but their performance improves in terms of approaching limiting $N(0, 1)$ distribution as the sample size increases. Compared to the coefficient-specific statistics, the distributions of the standardized pooled IR statistics are closer to the limiting $N(0, 1)$.

Similar results are found in Table 4.1. The empirical type I errors of all the IR statistics approach the nominal level as the sample size gets larger. In addition, the type I errors of the standardized pooled IR statistics are closer to the nominal level than those of the coefficient-specific statistics. Compared to these proposed IR statistics, the empirical type I error of the White's IM test differs from the nominal level substantially. Due to the poor performance of the IR statistics with small sample size, a proper approximation or bootstrap method is required to obtain the correct upper 5% quartiles.

Finite-sample Approximation

For small sample size, the distributions of the IR statistics are heavily right skewed. In addition, as shown in Table 4.1, with small sample size, say $n = 20$,

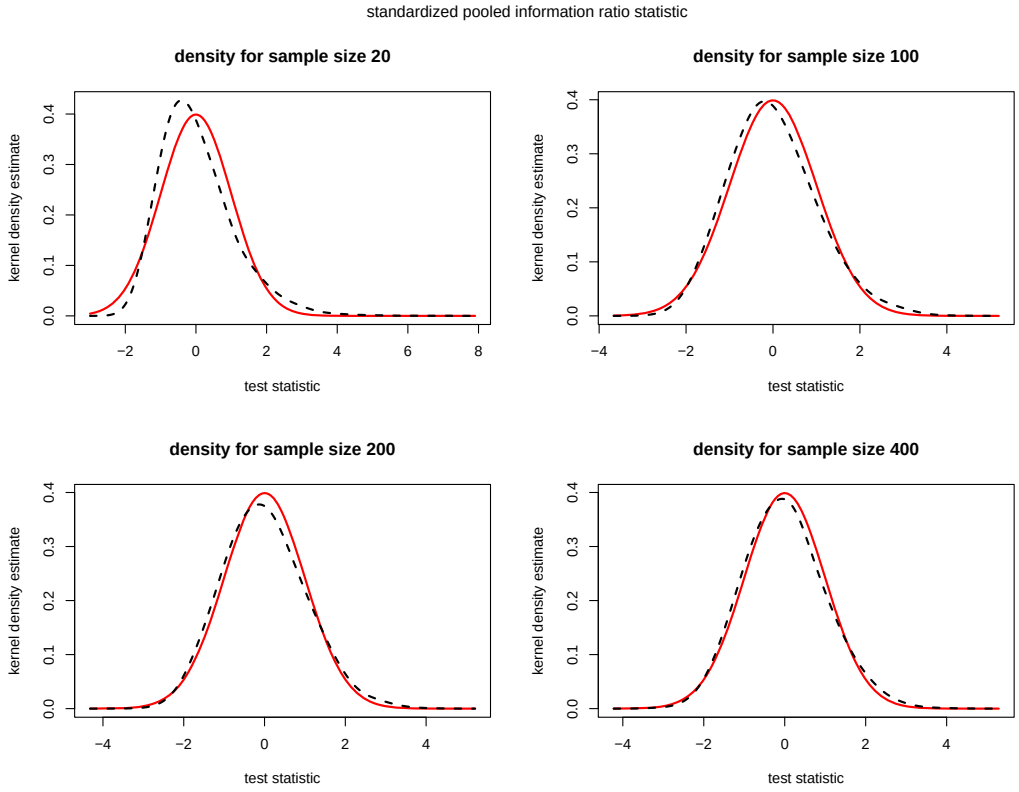

Figure 4.1: The kernel density estimates of the standardized pooled IR statistic (4.22) over different sample sizes, under the assumption that the true value of the variance parameter $\sigma^2 = 0.25$ is known. The solid line is the density function of the limiting normal distribution $N(0,1)$ under the null hypothesis. The dashed line represents the kernel density estimates of the standardized pooled IR statistic IR^s_{pool}.

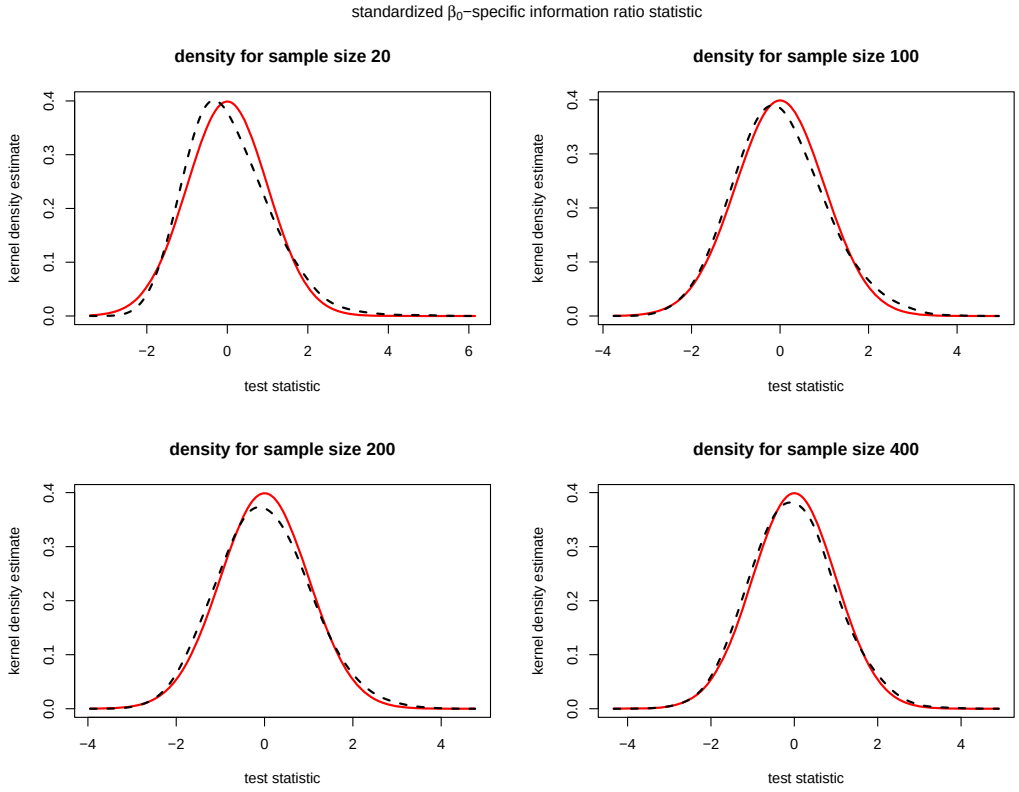

Figure 4.2: The kernel density estimates of the standardized β_0-specific IR statistic (4.21) over different sample sizes, under the assumption that the true value of the variance parameter $\sigma^2 = 0.25$ is known. The solid line is the density function of the limiting normal distribution $N(0, 1)$ under the null hypothesis. The dashed line represents the kernel density estimates of the standardized β_0-specific IR statistic IR_0^s.

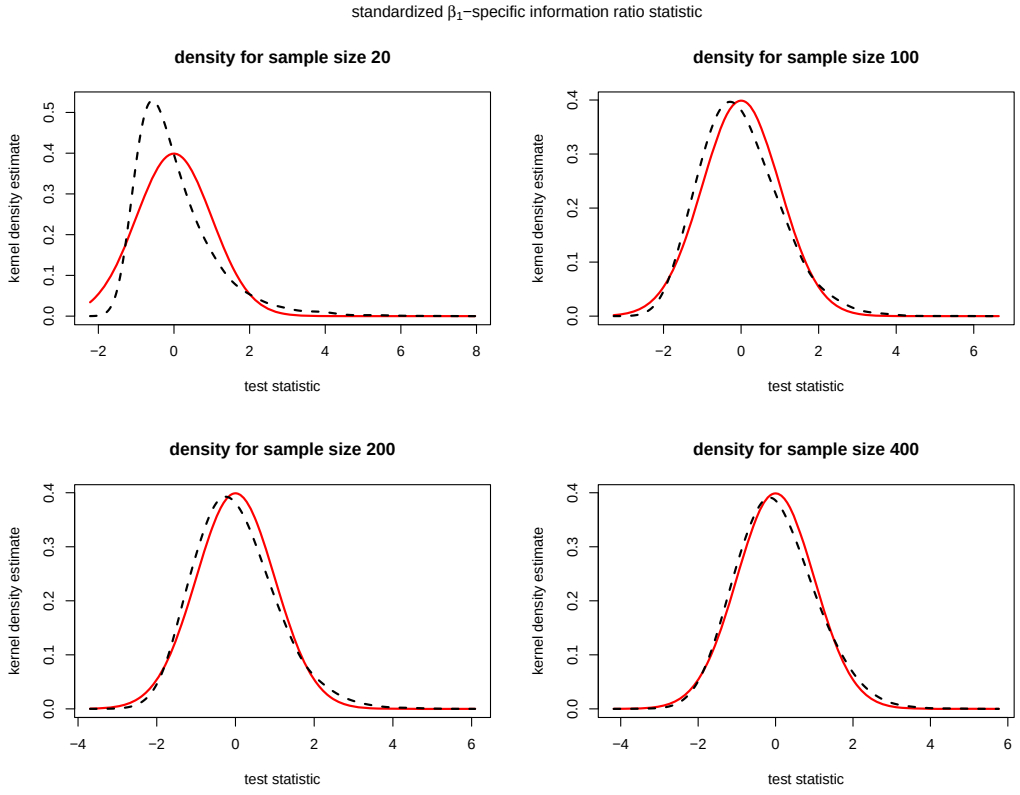

Figure 4.3: The kernel density estimates of the standardized β_1-specific IR statistic (4.21) over different sample sizes, under the assumption that the true value of the variance parameter $\sigma^2 = 0.25$ is known. The solid line is the density function of the limiting normal distribution $N(0, 1)$ under the null hypothesis. The dashed line represents the kernel density estimates of the standardized β_1-specific IR statistic IR_1^s.

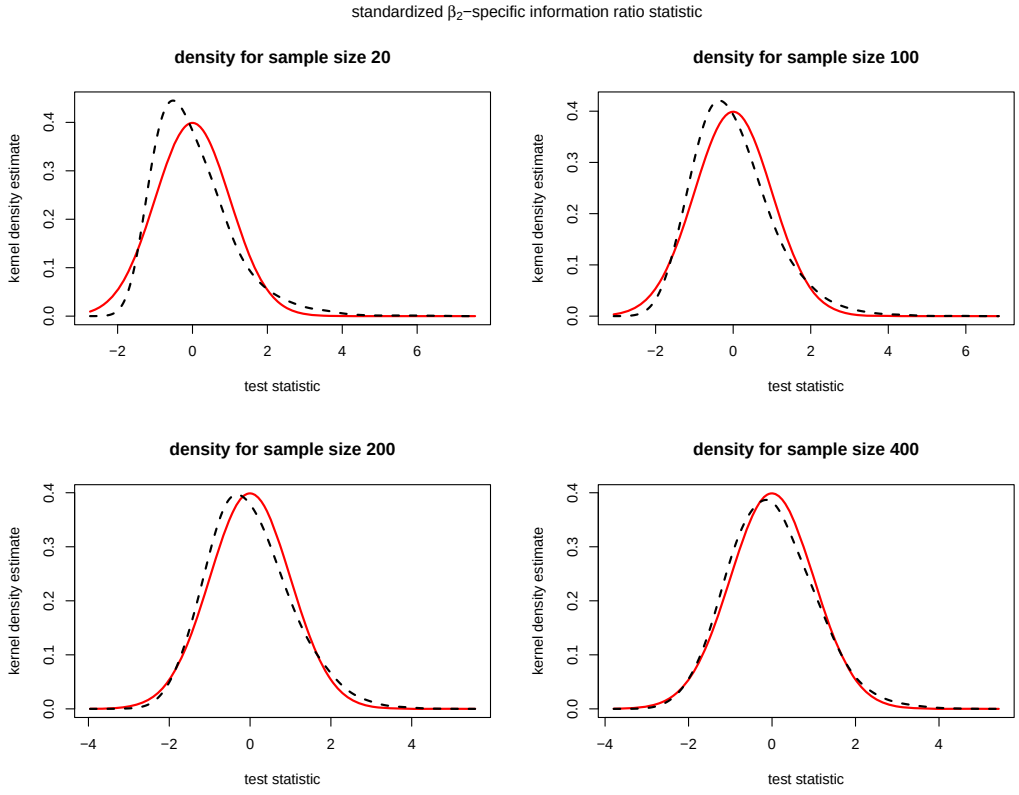

Figure 4.4: The kernel density estimates of the standardized β_2-specific IR statistic (4.21) over different sample sizes, under the assumption that the true value of the variance parameter $\sigma^2 = 0.25$ is known. The solid line is the density function of the limiting normal distribution $N(0, 1)$ under the null hypothesis. The dashed line represents the kernel density estimates of the standardized β_2-specific IR statistic IR_2^s.

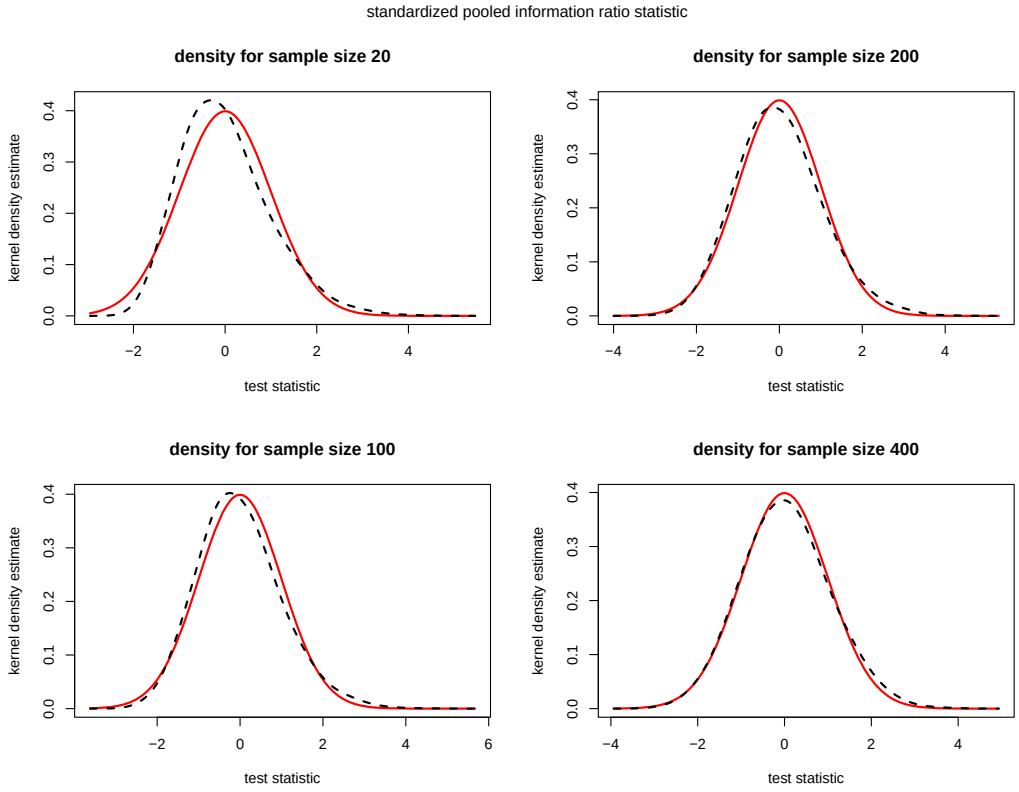

Figure 4.5: The kernel density estimates of the standardized pooled IR statistic (4.26) over different sample sizes, under the assumption that the true value of the variance parameter σ^2 is unknown. The solid line is the density function of the limiting normal distribution $N(0,1)$ under the null hypothesis. The dashed line represents the kernel density estimates of the standardized pooled IR statistic IR^s_{pool}.

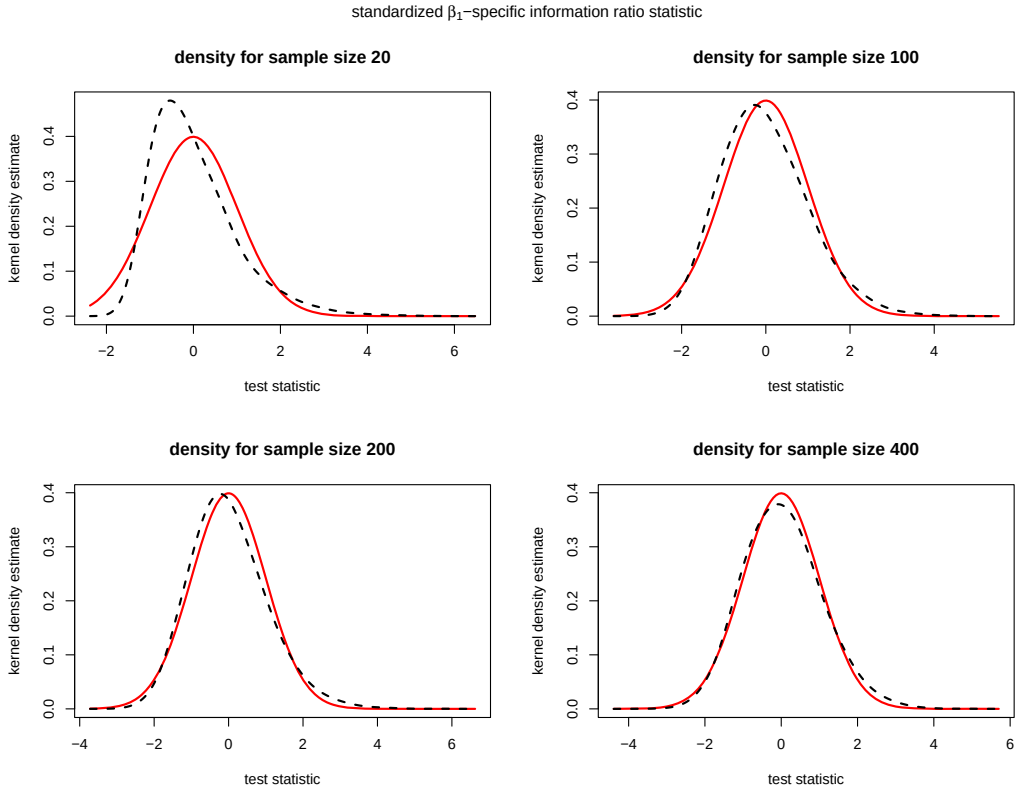

Figure 4.6: The kernel density estimates of the standardized β_1-specific IR statistic (4.25) over different sample sizes, under the assumption that the true value of the variance parameter σ^2 is unknown. The solid line is the density function of the limiting normal distribution $N(0,1)$ under the null hypothesis. The dashed line represents the kernel density estimates of the standardized β_1-specific IR statistic IR_1^s.

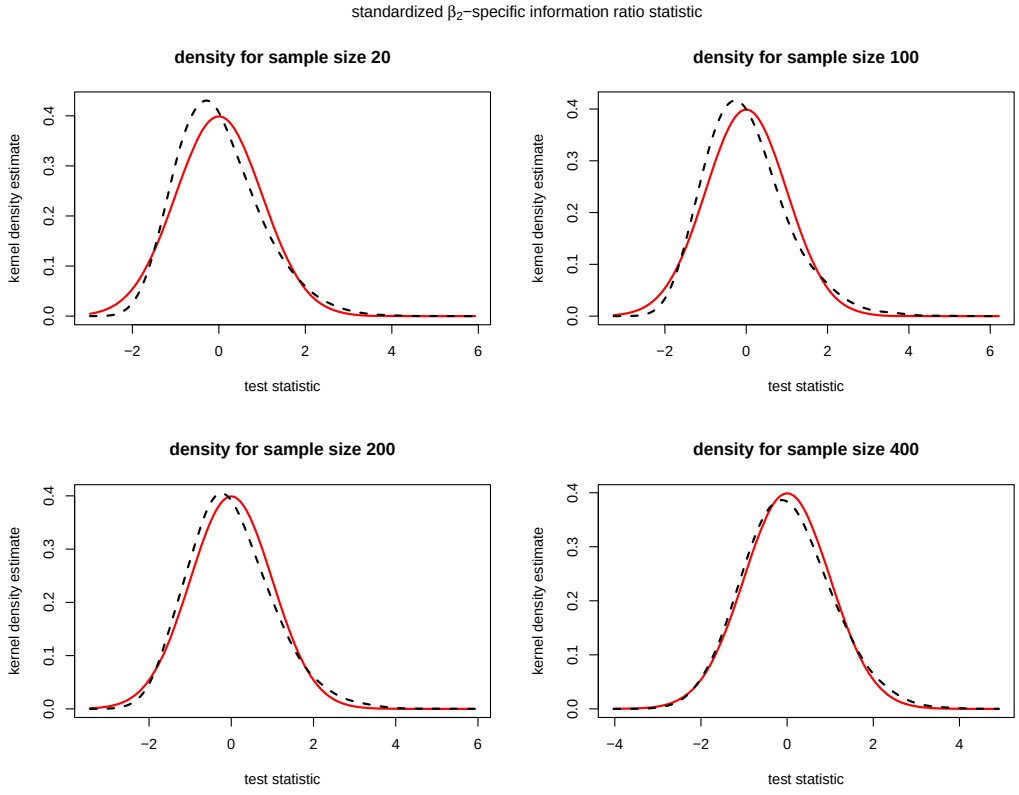

Figure 4.7: The kernel density estimates of the standardized β_2-specific IR statistic (4.25) over different sample sizes, under the assumption that the true value of the variance parameter σ^2 is unknown. The solid line is the density function of the limiting normal distribution $N(0,1)$ under the null hypothesis. The dashed line represents the kernel density estimates of the standardized β_2-specific IR statistic IR_2^s.

Table 4.1: The empirical type I errors of the standardized IR statistics, IR^s_{pool}, IR^s_0, IR^s_1, and IR^s_2, as well as the White's IM test statistic T_w over different sample sizes at the significance level 5%, under the assumption that the true value of the variance parameter is either known or unknown

sample size	IR^s_{pool}	IR^s_0	IR^s_1	IR^s_2	T_w
	$\sigma^2 = 0.25$ is known				
20	0.0500	0.0434	0.0534	0.0514	0.0032
100	0.0456	0.0540	0.0434	0.0436	0.0308
200	0.0462	0.0532	0.0498	0.0468	0.0718
400	0.0478	0.0478	0.0464	0.0460	0.0736
	σ^2 is unknown				
20	0.0358	-	0.0464	0.0362	0.0032
100	0.0466	-	0.0492	0.0432	0.0308
200	0.0532	-	0.0496	0.0458	0.0718
400	0.0508	-	0.0518	0.0510	0.0736

the type I errors of the proposed IR statistics are considerably smaller than the nominal level 0.05. Thus, the limiting distribution $N(0, 1)$ is not appropriate in order to obtain the critical value for the purpose of testing.

As shown in the proof of Theorem 4.1 and Theorem 4.2, the IR statistics can be approximated by quadratic forms in the vector ϵ_p, which has mean $\mathbf{0}$ and covariance matrix $\sigma^2 \boldsymbol{I}_n$ under the null hypothesis H_0 (4.31). For example, if the true value of the dispersion parameter σ^2 is known, the pooled IR statistic can be approximated by

$$IR_{pool} = \boldsymbol{\epsilon}_p^T (\boldsymbol{I}_n - \widehat{H}) \widehat{\boldsymbol{W}}_u^{pool} (\boldsymbol{I}_n - \widehat{H}) \boldsymbol{\epsilon}_p.$$

In the case of Gaussian responses or non-Gaussian responses for small dispersion, since the Pearson residuals have asymptotically normal distribution, the quadratic forms are approximately distributed as weighted sums of χ^2_1 distributions. For example, the pooled IR statistic

$$IR_{pool} = \sum_{k=1}^{n} \widehat{\lambda}_k^{pool} \chi^2_1,$$

where $\widehat{\lambda}_k^{pool}$, $k = 1, 2, \cdots, n$, are the eigenvalues of the matrix

$$(\boldsymbol{I}_n - \widehat{H})\widehat{\boldsymbol{W}}_u^{pool}(\boldsymbol{I}_n - \widehat{H}).$$

In this case, a certain proper approximation is required. In the literature, some approximation methods have been developed. [72] suggested to approximate the distribution of a quadratic form by that of $a\chi_\nu^2$, with a and ν chosen to equate the first two moments of the two distributions. Also see [26]. This approximation seems to perform well but no theoretical justification has been derived. Another approximation proposed by [76] is to use $a(\chi_\nu^2)^b$, with the parameters being chosen to equate the first three moments. However, this approximation is difficult to use in practice as the equations defining a, b and ν require an iterative solution. Later, [10] proposed to use a normalized χ_ν^2 to approximated the normalized quadratic forms. The value of ν is chosen by equating the first three moments of the quadratic form with those of $a\chi_\nu^2 + b$. Here, we used the method of this normalized χ^2 approximation to obtain the upper 5% quartile. Suppose that we approximate $\sum_{k=1}^n \lambda_k\chi_1^2$ by a normalized χ_ν^2, i.e., $(\chi_\nu^2 - \nu)/\sqrt{2\nu}$. The value of ν is chosen to be

$$\nu = \frac{\left(\sum_{k=1}^n \lambda_k^2\right)^3}{\left(\sum_{k=1}^n \lambda_k^3\right)^2}.$$

Therefore, $(\chi_\nu^2(0.95) - \nu)/\sqrt{2\nu}$ can be obtained as the approximate 95% quartile, where $\chi_\nu^2(0.95)$ is the 95% quartile of the distribution χ_ν^2.

Figure 4.8 shows the Q-Q plots of the standardized IR statistics versus the limiting $N(0, 1)$ distribution, and the normalized χ_ν^2 distribution, respectively, for small sample size $n = 20$, under the assumption that the true value $\sigma^2 = 0.25$ is known. Figure 4.9 shows the Q-Q plots under the assumption that the true value of σ^2 is unknown. Table 4.2 reports the empirical type I errors of the standardized IR statistics IR_{pool}^s, IR_0^s, IR_1^s, and IR_2^s using the critical values at the significance level 5% obtained from the normalized χ_ν^2 approximation for small sample size $n = 20$. Both of the Q-Q plots and numerical results show that with

small sample size, the normalized χ^2_ν distribution gives a better approximation to the distribution of the standardized IR statistics than the limiting $N(0,1)$ distribution, especially in the cases where the true value of the dispersion parameter σ^2 is unknown. In addition, the empirical type I errors are closer to the nominal level 5% when the normalized χ^2_ν approximation is used, compared to the limiting $N(0,1)$ distribution.

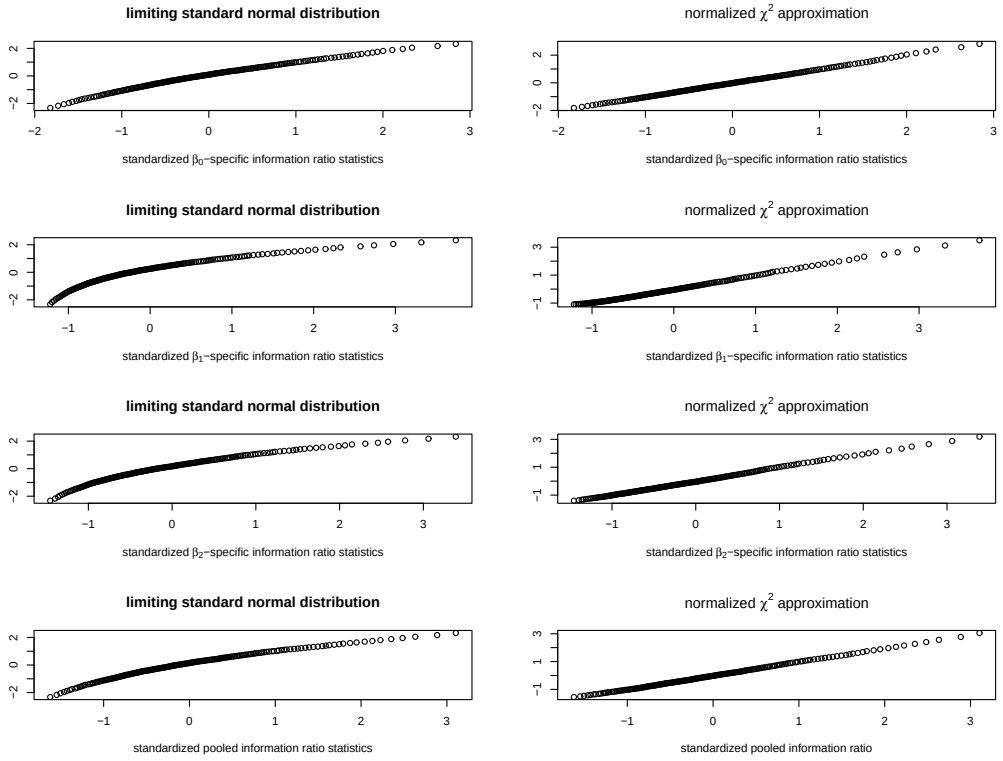

Figure 4.8: Q-Q plots of the standardized IR statistics, IR^s_0, IR^s_1, IR^s_2 and IR^s_{pool}, from top to bottom, for small sample size $n = 20$ under the assumption that the true value of $\sigma^2 = 0.25$ is known. The left panels plot the quartiles of the standard $N(0,1)$ distribution versus the quartiles of the standardized IR statistics. The right panels plot the quartiles of the normalized χ^2_ν distribution versus the quartiles of the standardized IR statistics.

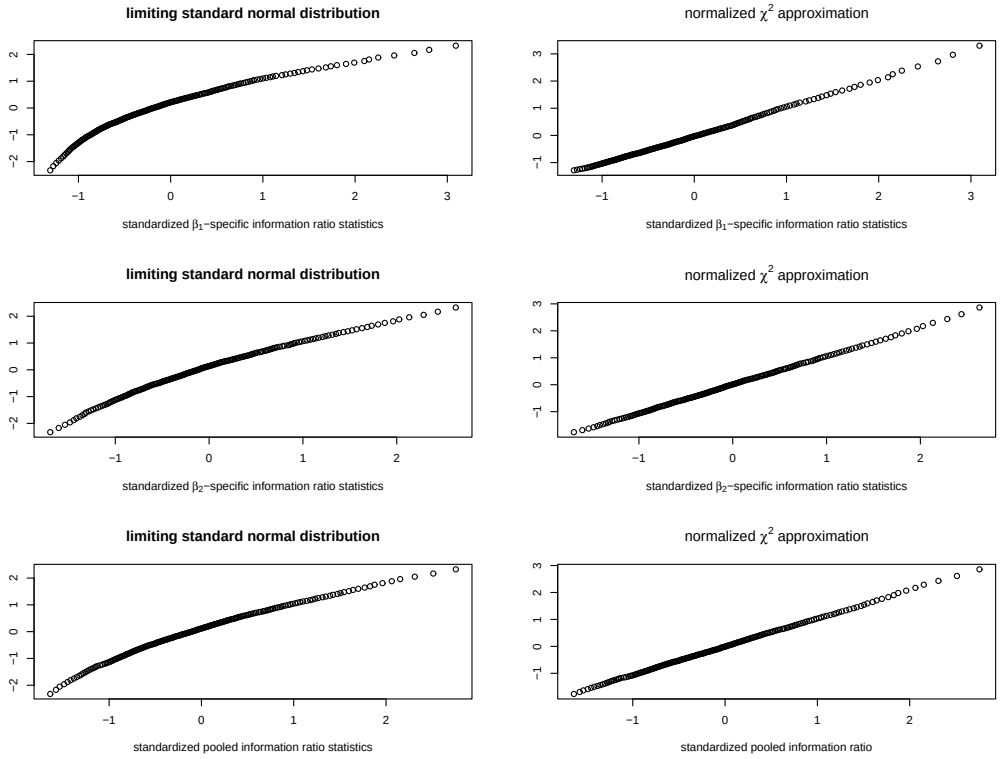

Figure 4.9: Q-Q plots of the standardized IR statistics, IR_1^s, IR_2^s and IR_{pool}^s, from top to bottom, for small sample size $n = 20$ under the assumption that the true value of σ^2 is unknown. The left panels plot the quartiles of the standard $N(0, 1)$ distribution versus the quartiles of the standardized IR statistics. The right panels plot the quartiles of the normalized χ_ν^2 distribution versus the quartiles of the standardized IR statistics.

Table 4.2: The empirical type I errors of the standardized IR statistics, IR^s_{pool}, IR^s_0, IR^s_1, and IR^s_2 using the normalized χ^2_ν approximation for small sample size $n = 20$ at the significance level 5%.

	IR^s_{pool}	IR^s_0	IR^s_1	IR^s_2
	$\sigma^2 = 0.25$ is known			
$N(0,1)$	0.0500	0.0434	0.0534	0.0514
$(\chi^2_\nu - \nu)/\sqrt{2\nu}$	0.0540	0.0514	0.0524	0.0528
	σ^2 is unknown			
$N(0,1)$	0.0358	-	0.0464	0.0362
$(\chi^2_\nu - \nu)/\sqrt{2\nu}$	0.0450	-	0.0476	0.0428

In LM, the true value of the variance parameter σ^2 is usually unknown in practice. Then, in the following simulation studies, we focus on only the power of the proposed IR statistics, which take ratios of the Godambian estimators to the moment estimators, under the assumption that the true value is unknown. In the following simulation studies, we will carry out power comparison among the proposed IR statistics and the White's IM test for small sample size 20 and large sample size 200. Because the White's IM test cannot attain the nominal level 5%, the critical values used for comparing power, 7.878 for sample size 20 and 11.758 for sample size 200, are obtained from the empirical 95% quartiles of the 5000 replicates under the null distribution in Simulation 4.1.

In many applications, when the observations drawn from the population are far from the center, the sampling error variability tends to be larger. Thus, the error variance could be a function of the leverage h_{ii}, the diagonal elements of the hat matrix $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$.

Simulation 4.2 In this experiment, the error terms e_i's are generated from a Gaussian distribution with mean 0 and variance

$$Var(Y_i) = 0.2h_{ii},$$

for $i = 1, \cdots, n$. Here, we compare the power of the proposed IR statistics with the White's IM test to reject the null hypothesis H_0 (4.31) under the alternative

hypothesis,

$$H_A : Var(Y_i) = \sigma^2 h_{ii}, \quad i = 1, \cdots, n,$$

with small sample size $n = 20$ and large sample size $n = 200$.

Table 4.3 reports the empirical power, which is in fact the proportion of rejecting the null hypothesis (4.31) among the 5000 replicates, of the standardized IR statistics, IR^s_{pool}, IR^s_1, and IR^s_2, as well as the White's IM test statistic T_w, at the significance level 5%. For small sample size $n = 20$, the results obtained from the normalized χ^2_ν approximation are also included in this table.

Conclusion.

The power of all the test statistics improves as the sample size increases. For small sample size $n = 20$, the normalized χ^2_ν approximation improves the power of the IR tests. Compared with the IR tests, the White's IM test is much less powerful to reject the null hypothesis. Moreover, the standardized pooled IR statistic outperforms the coefficient-specific statistics.

Table 4.3: The empirical power of the standardized IR statistics, IR^s_{pool}, IR^s_1, and IR^s_2, as well as the White's IM test statistic T_w, over different sample sizes 20 and 200, at the significance level 5% to reject the null hypothesis H_0 (4.31) under the heteroscedasticity $H_A : Var(Y_i) = \sigma^2 h_{ii}$.

	$n = 20$		$n = 200$
	$N(0,1)$	$(\chi^2_\nu - \nu)/\sqrt{2\nu}$	
IR^s_{pool}	0.2500	0.2728	0.9964
IR^s_1	0.2224	0.2300	0.8894
IR^s_2	0.1082	0.1680	0.8804
T_w	0.0992		0.4838

Simulation 4.3 In some applications, the heteroscedasticity in the model is associated with a certain covariate. For example, when the covariate x_{i1} is sampled far from the corresponding population center, the error variance gets large. In this case, the error variance is a function of the covariate x_{i1}. In this experiment,

the error terms e_i's are generated from a Gaussian distribution with mean 0 and variance

$$Var(Y_i) = 0.2\mathbb{h}_{ii}^{(1)}, \quad \mathbb{h}_{ii}^{(1)} = (x_{i1} - \bar{x}_1)^2 / \sum_k (x_{k1} - \bar{x}_1)^2,$$

for $i = 1, \cdots, n$. Then, the alternative hypothesis in this simulation is given by

$$H_A : Var(Y_i) = \sigma^2 \mathbb{h}_{ii}^{(1)}, \quad i = 1, \cdots, n.$$

Table 4.4 reports the empirical power of the standardized IR statistics, IR^s_{pool}, IR^s_1, and IR^s_2, as well as the White's IM test statistic T_w, at the significance level 5% with the sample size 20 and 200. In addition, this table lists the results by using the normalized χ^2_ν approximation for small sample size $n = 20$.

Conclusion.

Similarly to the results in Simulation 4.2, the power of all the test statistics improves as the sample size increases. The normalized χ^2_ν approximation improves the power of the IR statistics for small sample size. In addition, the White's IM test is generally less powerful to reject the null hypothesis H_0 (4.31) than the proposed IR statistics. Among all the IR statistics, the standardized pooled IR statistic and the β_1-specific IR statistic have the top two highest empirical power. However, at the significance level 5%, there is no evidence to reject the null hypothesis when using the standardized β_2-specific IR statistic.

Simulation 4.4 The error terms e_i's are generated from a Gaussian distribution with mean 0 and variance

$$Var(Y_i) = 0.2\mathbb{h}_{ii}^{(2)}, \quad \mathbb{h}_{ii}^{(2)} = (x_{i2} - \bar{x}_2)^2 / \sum_k (x_{k2} - \bar{x}_2)^2,$$

for $i = 1, \cdots, n$. Then, the alternative hypothesis in this simulation is given by

$$H_A : Var(Y_i) = \sigma^2 \mathbb{h}_{ii}^{(2)}, \quad i = 1, \cdots, n.$$

Table 4.4: The empirical power of the standardized IR statistics, IR^s_{pool}, IR^s_1 and IR^s_2, as well as the White's IM test statistic T_w, over different sample sizes at the significance level 5% to reject the null hypothesis H_0 (4.31) under the heteroscedasticity $H_A : Var(Y_i) = \sigma^2 \mathbb{h}^{(1)}_{ii}$.

	$n = 20$		$n = 200$
	$N(0,1)$	$(\chi^2_\nu - \nu)/\sqrt{2\nu}$	
IR^s_{pool}	0.1840	0.2396	1.0000
IR^s_1	0.6016	0.6922	1.0000
IR^s_2	0.0126	0.0108	0.2528
T_w	0.1626		0.7650

Table 4.5 reports the empirical power of the standardized IR statistics, IR^s_{pool}, IR^s_1, and IR^s_2, as well as the White's IM test statistic T_w, at the significance level 5% with the sample size 20 and 200. In addition, this table lists the results by using the normalized χ^2_ν approximation for small sample size $n = 20$.

Conclusion.

The results are similar to those of Simulation 4.2. Among all the proposed IR statistics, the pooled IR statistic and the β_2-specific IR statistic are the most powerful to reject the null hypothesis H_0 (4.31). However, under the significance level 5%, there is no strong evidence to reject the null hypothesis when using the standardized β_1-specific IR statistic.

In some applications, the error variance is associated with the mean values. For example, the sampling error variability gets larger when the mean values increase. In this case, the error variance is a function of the mean values. In the following three simulation experiments, the error variances are modelled as exponential functions of the covariates. In addition, we will also investigate the effect of the covariate variability on the performance of the IR statistics. Thus, in the following simulation studies, the covariate x_{i1} is generated from $N(0,1)$,

Table 4.5: The empirical power of the standardized IR statistics, IR^s_{pool}, IR^s_1 and IR^s_2, as well as the White's IM test statistic T_w, over different sample sizes at the significance level 5% to reject the null hypothesis H_0 (4.31) under the heteroscedasticity $H_A : Var(Y_i) = \sigma^2 \mathbb{h}^{(2)}_{ii}$.

	$n = 20$		$n = 200$
	$N(0,1)$	$(\chi^2_\nu - \nu)/\sqrt{2\nu}$	
IR^s_{pool}	0.3834	0.4236	1.0000
IR^s_1	0.1204	0.1266	0.2522
IR^s_2	0.5954	0.6370	1.0000
T_w	0.1620		0.6638

and the covariate x_{i2} is generated from $N(0, 0.1)$. Since x_{i1} and x_{i2} are generated independently, their centered versions are approximately orthogonal.

Simulation 4.5 In this experiment, the error terms e_i's are generated from a Gaussian distribution with mean 0 and variance

$$Var(Y_i) = 0.5 \exp\{\beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2)\},$$

for $i = 1, \cdots, n$. Then, the alternative hypothesis here is given by

$$H_A : Var(Y_i) = \sigma^2 \exp\{\beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2)\}, \quad i = 1, \cdots, n.$$

Simulation 4.6 In this experiment, only the covariate x_{i1} accounts for the heteroscedasticity in an exponential form. Therefore, the error terms e_i's are generated from a Gaussian distribution with mean 0 and variance

$$Var(Y_i) = 0.5 \exp\{\beta_1(x_{i1} - \bar{x}_1)\},$$

for $i = 1, \cdots, n$. The alternative hypothesis is given by

$$H_A : Var(Y_i) = \sigma^2 \exp\{\beta_1(x_{i1} - \bar{x}_1)\}, \quad i = 1, \cdots, n.$$

Simulation 4.7 The error terms e_i's are generated from a Gaussian distribution with mean 0 and variance

$$Var(Y_i) = 0.5 \exp\{\beta_2(x_{i2} - \bar{x}_2)\},$$

for $i = 1, \cdots, n$. The alternative hypothesis is given by

$$H_A : Var(Y_i) = \sigma^2 \exp\{\beta_2(x_{i2} - \bar{x}_2)\}, \quad i = 1, \cdots, n.$$

Table 4.6 reports the power of the standardized IR statistics, IR^s_{pool}, IR^s_1 and IR^s_2, at the significance level 5% among 5000 replications under the three different alternative models in Simulation 4.5, 4.6 and 4.7. For the small sample size $n = 20$, the results are obtained from the normalized χ^2_ν approximation for small sample size $n = 20$.

Conclusion.

The numerical results in Table 4.6 appear similar to the results in Simulations 4.2, 4.3 and 4.4. Since the variability in the covariate x_{i2} is considerably smaller than that in the covariate x_{i1}, the heteroscedasticity results from mainly the covariate x_{i1} even though the error variances are functions of both of the covariates x_{i1} and x_{i2}. Thus, it shows that the IR^s_2 statistic is less powerful than the IR^s_1 statistic as well as the IR^s_{pool} statistic. Especially, in Simulation 4.7, the error variances are less heteroscedastic due to its form as an exponential of the value of x_{i2}, so all the IR statistics have a lower frequency of rejecting the null hypothesis.

Summary

(1) Under the null hypothesis (4.31), the standardized IR statistics are heavily right skewed for small sample size, but their performance improves as the sample size increases. In addition, among all the proposed IR statistics, the pooled IR statistic outperforms the coefficient-specific statistics.

Table 4.6: The empirical power of the standardized IR statistics, IR^s_{pool}, IR^s_1, and IR^s_2, as well as the White's IM test statistic T_w, over different sample sizes at the significance level 5% to reject the null hypothesis H_0 (4.31) under the alternative hypotheses $H_A : Var(Y_i) = \sigma^2 \exp\{\beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2)\}$, $H_A : Var(Y_i) = \sigma^2 \exp\{\beta_1(x_{i1} - \bar{x}_1)\}$ and $H_A : Var(Y_i) = \sigma^2 \exp\{\beta_1(x_{i2} - \bar{x}_2)\}$.

H_A	$\exp\{\beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2)\}$		$\exp\{\beta_1(x_{i1} - \bar{x}_1)\}$		$\exp\{\beta_2(x_{i2} - \bar{x}_2)\}$	
	$n = 20$	$n = 200$	$n = 20$	$n = 200$	$n = 20$	$n = 200$
IR^s_{pool}	0.3884	0.9988	0.4776	0.9998	0.0392	0.0692
IR^s_1	0.4170	1.0000	0.4842	1.0000	0.0446	0.0468
IR^s_2	0.2564	0.3948	0.2064	0.3902	0.0446	0.1084

(2) Compared with the IR statistics, the White's IM test has poorer performance under the null hypothesis. Its type I error differs from the nominal level substantially.

(3) Compared with the White's IM test, all the IR tests are more powerful to reject the null hypothesis under certain alternative hypotheses of heteroscedasticity. In addition, if the heteroscedasticity arises from a certain covariate, say the j-th covariate $x_{i,j-1}$, orthogonal to the rest, the corresponding β_{j-1}-specific IR statistic is more powerful than the other regression coefficients. The result illustrates the discussion about the properties of the weights incorporated in the Godambian estimator in linear regression models in Section 3.1.1. Since the weights $w_i^{(j-1)}$ characterize the influence from the j-th covariate, the β_{j-1}-specific IR statistic, IR_{j-1}, is expected to be far from 1, compared to the other regression coefficients. Therefore, the statistic IR_{j-1} is more powerful to reject the null hypothesis. In addition, since the weights in the standardized pooled IR statistic IR_{pool} reflect the overall influence from all the covariates, the statistic IR_{pool} is also powerful to reject the null hypothesis.

In the literature of LM, residuals play an important role in graphical diagnostics, such as plotting residuals versus a certain covariate. However, few statistical methods are available to carry out a formal statistical test for

dependence of the error variance on a certain variable. The Godambian estimator of the variance parameter σ^2 is able to identify the responsible covariate variable for the heteroscedasticity in LM, due to the unique form of a weighted sum of squared residuals, as well as the properties of these weights. But the IR test is not able to provide any information about the explicit form of the error variance associated with the responsible covariates. In addition, the test will be less sensitive to capture the dependence if the heteroscedasticity is caused by complex interactions among multiple covariates. Thus, the IR test is recommended as part of the exploratory analysis in LM.

(4) In the literature of LM, some tests for heteroscedasticity are based on specific alternative hypotheses; see [5]; [7] and [14]. However, these tests may not be powerful under other types of heteroscedasticity. The simulation studies have shown that the strong power of the proposed IR statistics are consistent among various scenarios of heteroscedasticity.

(5) A two stage-wise procedure of testing heteroscedasticity can be suggested. First, if the p value obtained from the standardized pooled IR statistic is less than 0.05, the null hypothesis of homoscedasticity (4.31) is rejected at the significance level 5%. Secondly, each regression coefficient specific test will be carried out. If a certain β_{j-1}-specific IR statistic gives a p value smaller than 0.05, we may conclude that the error variance is a function of the j-th covariate.

(6) The proposed IR statistics have poor performance for small sample size using the limiting $N(0, 1)$ distribution. The simulation studies have shown that normalized χ^2_ν approximation improves the performance for small sample size.

4.3.2 Test for overdispersion in count data

Poisson regression models are widely used in analyzing count data. In these models, given independent responses Y_i and associated $p \times 1$ covariate vectors $\boldsymbol{x}_i$, the distribution of Y_i, given $\boldsymbol{x}_i$ is assumed to be Poisson with mean $\mu_i = \mu_i(\boldsymbol{x}_i; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients. Here, the mean structure of the responses is given by

$$\mu_i = E(Y_i) = h(\boldsymbol{x}_i^T \boldsymbol{\beta}), \quad i = 1, \cdots, n,$$

where $h(.)$ is the link function associating the mean of the responses with the covariates. The log-linear regression model, where $\mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$, is most commonly considered. The variance structure of the responses is given by

$$Var(Y_i) = \mu_i, \quad i = 1, \cdots, n,$$

which indicates that the unit variance function $V(\mu_i) = \mu_i$, for $i = 1, \cdots, n$, and the dispersion parameter is 1.

In this section, we apply the proposed IR statistics to test for overdispersion in Poisson regression models. Then, the null hypothesis is defined as

$$H_0 : V(\mu_i) = \mu_i, \quad i = 1, \cdots, n. \tag{4.32}$$

Because the dispersion parameter σ^2 in Poisson regression models is assumed to be known as 1, we focus on the IR statistics which take ratios of the Godambian estimators to the true value, for example, (4.21) and (4.22).

Simulation 4.8 In this simulation study, we investigate the asymptotic distributions of the proposed IR statistics under the null hypothesis H_0 (4.32). A data set $\{(y_i; x_i), i = 1, \cdots, n\}$ is generated from a Poisson distribution as follows:

$$Y_i \mid x_i \sim \text{Poisson}(\mu_i)$$

where $\mu_i = \exp(\beta_0 + \beta_1 x_i)$, for $i = 1, \cdots, n$. The covariate x_i's are independently generated from a uniform distribution $UNIF(0, 1)$. We consider two models where the range of μ_i's is moderate or large.

Model 1: The true values of the regression coefficients $\beta_0 = 1$, and $\beta_1 = 1$. The range of μ_i is 2.72 to 7.39.

Model 2: The true values of the regression coefficients $\beta_0 = 1$ and $\beta_1 = 4$. The range of μ_i is 2.72 to 148.41.

The sample size is set to be 10, 50, 100 and 200. For each sample size, 5000 replicates are generated. In addition, the same values of the covariates are used for each replicate to stress that the covariates are fixed. Table 4.7 reports the empirical type I errors, the proportions of rejecting the null hypothesis H_0 (4.32), of the standardized IR statistics (4.21) and (4.22), IR^s_{pool}, IR^s_0 and IR^s_1, over different sample sizes at the significance level 10%, 5% and 1% for Model 1 (moderate range of μ_i's) and Model 2 (large range of μ_i's). The numerical results show that for larger sample size, the type I errors of the standardized IR statistics are closer to the nominal levels. Thus, as the sample size increases, the normal asymptotic distribution is more accurate.

To incorporate possible extra-Poisson variation, we consider alternative mixed Poisson models. Let $\zeta_1, \cdots, \zeta_n$ be continuous positive-valued independent random variables from a certain distribution with finite first and second moments. For each $i = 1, \cdots, n$, given $\boldsymbol{x}_i$ and ζ_i, Y_i is Poisson distributed with mean $\zeta_i \mu_i$, where $\mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$. Without loss of generality, we assume that $E(\zeta_i) = 1$ and $Var(\zeta_i) = \varpi > 0$. If the ζ_i's follow a gamma distribution, then Y_i has a negative binomial distribution with the mean

$$E(Y_i) = \mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}),$$

and the variance

$$Var(Y_i) = \mu_i + \varpi \mu_i^2.$$

Table 4.7: The empirical type I errors of the standardized IR statistics, IR^s_{pool}, IR^s_0, and IR^s_1, from the limiting $N(0,1)$ distribution, over different sample sizes at the significance level 10%, 5% and 1%, under the null hypothesis H_0 (4.32) for Model 1 (moderate range of μ_i's) and Model 2 (large range of μ_i's).

	Model 1			Model 2		
n	.10	.05	.01	.10	.05	.01
			IR^s_{pool}			
10	0.0650	0.0458	0.0232	0.0628	0.0430	0.0188
50	0.0920	0.0466	0.0132	0.0846	0.0472	0.0186
100	0.0938	0.0504	0.0108	0.0914	0.0472	0.0142
200	0.0966	0.0496	0.0110	0.0956	0.0470	0.0118
			IR^s_0			
10	0.0590	0.0438	0.0222	0.0644	0.0414	0.0180
50	0.0796	0.0450	0.0176	0.0858	0.0440	0.0112
100	0.0998	0.0482	0.0150	0.0984	0.0500	0.0114
200	0.0964	0.0504	0.0118	0.0994	0.0506	0.0120
			IR^s_1			
10	0.0700	0.0506	0.0298	0.0628	0.0440	0.0196
50	0.0890	0.0458	0.0150	0.0798	0.0430	0.0148
100	0.0970	0.0518	0.0138	0.0924	0.0492	0.0124
200	0.0994	0.0524	0.0134	0.0980	0.0500	0.0114

We will use the simulation studies included in the paper [20] to carry out power comparisons for the proposed IR statistics; T_a proposed by [20], given by

$$T_a = \frac{\sum_{i=1}^{n}\{(Y_i - \widehat{\mu}_i)^2 - Y_i + \widehat{h}_{ii}\widehat{\mu}_i\}}{(2\sum_{i=1}^{n}\widehat{\mu}_i^2)^{1/2}},$$

where $\widehat{h}_{ii}$ is the diagonal element of the hat matrix $\widehat{H}$ given in (3.7); Pearson χ^2 statistic P, given by

$$P = \sum_{i=1}^{n}\frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i};$$

and the deviance statistic D, given by

$$D = 2\sum_{i=1}^{n}[y_i \log(y_i/\widehat{\mu}_i) - (y_i - \widehat{\mu}_i)].$$

Simulation 4.9 A data set $\{(y_i, x_i), i = 1, \cdots, n\}$ with sample size $n = 15$ is generated from the mixed Poisson model described above. One-third of the x_i's are equal to each of 0, 0.5 and 1. For each i, given μ_i and ϖ, generate ζ_i from a gamma distribution with the shape parameter $1/\varpi$, and the scale parameter ϖ. Then, generate the response Y_i from a Poisson distribution with mean $\zeta_i\mu_i$. We will investigate two alternative models with different ranges of μ_i.

Model 3: $\mu_i = \exp(2.6 + 2x_i)$ for $i = 1, \cdots, n$. The μ_i's range in value from roughly 13.5 to 99.

Model 4: $\mu_i = \exp(2.6 + 3x_i)$ for $i = 1, \cdots, n$. The μ_i's range from 13.5 to 270.

Table 4.8 reports the empirical power, the proportions of rejecting the null hypothesis H_0, of the standardized IR statistics, IR^s_{pool}, IR^s_0, and IR^s_1; the score test T_a proposed by [20]; the Pearson χ^2 statistic P; and the deviance statistic D over different values of ϖ, among 5000 replicates, for Model 3 and Model 4 at the significance level 5%. For the standardized IR statistics, the results obtained

from the limiting $N(0, 1)$ distribution as well as the normalized χ^2_ν approximation are listed in the table. For the statistic T_a proposed by [20], the critical value used at $\varpi = 0$ is obtained from the standard normal upper 5% point of 1.645, and for the Pearson χ^2 statistic P and the deviance statistic D, the critical value used at $\varpi = 0$ is the χ^2_{13} upper 5% point of 22.4. These critical values lead to larger type I errors, shown with the $*$ in Table 4.8. As a result, at other values of ϖ, the critical values used are obtained from the empirical 95% quartiles from 5000 replicates under the null hypothesis ($\varpi = 0$). For Model 3, the critical values are 1.723 for T_a, 22.494 for P, and 22.732 for D, and for Model 4, the critical values are 1.779 for T_a, 22.213 for P, and 22.427 for D.

Conclusion.

The type I errors of the standardized IR statistics under the null hypothesis ($\varpi = 0$) in both of the models are much smaller than the nominal level 0.05 for small sample size. However, the normalized χ^2_ν distribution gives a better approximation. Among all the IR statistics, the standardized pooled IR statistic is substantially more powerful to reject the null hypothesis under the mixed Poisson model. Moreover, the performance of the pooled IR statistic using the normalized χ^2_ν approximation is comparable to that of the score statistic T_a proposed by [20], and even better in some cases. In addition, the pooled IR statistic is generally more powerful than the Pearson χ^2 statistic P and the deviance statistic D. The improvement of the pooled IR statistic over other statistics is greater with larger values of ϖ, which indicate greater levels of overdispersion. Moreover, the difference in power between IR^s_{pool} and T_a (or P, D) is larger when the covariates effect is larger and the μ_i's vary more widely. Note that one of the appealing feature of the IR statistics is that only the Poisson model needs to be fit, and there is no requirement of modeling alternative hypotheses.

Table 4.8: The empirical power of the standardized IR statistics, IR^s_{pool}, IR^s_0, and IR^s_1, score test T_a proposed by [20], Pearson χ^2 statistic P, and deviance statistic D, among 5000 replicates, for Model 3 and Model 4 at the significance level 5% to reject the null hypothesis H_0 (4.32) under the mixed Poisson model.

	IR^s_{pool}		IR^s_0		IR^s_1		T_a	P	D
ϖ	$N(0,1)$	$\frac{\chi^2_\nu-\nu}{\sqrt{2\nu}}$	$N(0,1)$	$\frac{\chi^2_\nu-\nu}{\sqrt{2\nu}}$	$N(0,1)$	$\frac{\chi^2_\nu-\nu}{\sqrt{2\nu}}$			
Model 3: moderate range									
0	0.044	0.050	0.046	0.050	0.042	0.048	0.054*	0.052*	0.057*
0.005	0.164	0.177	0.071	0.078	0.109	0.120	0.188	0.154	0.152
0.015	0.456	0.476	0.166	0.177	0.329	0.347	0.481	0.443	0.441
0.025	0.655	0.677	0.280	0.295	0.533	0.556	0.672	0.657	0.651
0.04	0.833	0.843	0.464	0.479	0.752	0.766	0.836	0.837	0.837
Model 4: large range									
0	0.043	0.051	0.041	0.048	0.038	0.044	0.058*	0.048*	0.052*
0.002	0.164	0.178	0.080	0.087	0.096	0.106	0.180	0.156	0.156
0.005	0.372	0.392	0.112	0.123	0.194	0.214	0.395	0.335	0.333
0.01	0.640	0.654	0.206	0.220	0.403	0.426	0.646	0.595	0.593
0.02	0.870	0.878	0.423	0.440	0.733	0.751	0.858	0.857	0.855

4.3.3 Test for misspecified variance function and/or working correlation matrix in GEE

In the context of GEE, misspecification of the covariance structure may result from misspecification of the variance function and/or misspecification of the working correlation structure. [85] carried out intensive studies of the impacts of misspecifying the variance function on the mean parameter estimators for quantitative responses. Their numerical results have shown that (1) correct specification of the variance function can improve the estimation efficiency even if the correlation structure is misspecified; (2) misspecification of the variance function impacts much more on estimators for within-cluster covariates than for cluster-level covariates; and (3) if the variance function is misspecified, correct choice of the correlation structure may not necessarily improve estimation efficiency. Moreover, [84] have shown that the choice of working correlation structure has a substantial impact on estimation efficiency of regression coefficients. In this section, we apply the IR statistics to test for misspecification of variance function and/or working correlation structure in GEE. Since [85] have shown that misspecified covariance structures have stronger impact for within-cluster covariates, in the following simulation, the covariates are generated from a time-dependent distribution. (See [85]).

Suppose that a longitudinal data set

$$\Big\{(y_{ij}, x_{ij}), j = 1, \cdots, n_i, i = 1, \cdots, K\Big\}$$

is generated as follows: for each subject i,

- generate a time-dependent covariate x_{ij} from a uniform distribution $UNIF(j-1, j)$, and let $\mu_{ij} = \exp(\beta_0 + \beta_1 x_{ij})$, for $j = 1, \cdots, n_i$;
- given $\boldsymbol{\mu}_i^T = (\mu_{i1}, \cdots, \mu_{i,n_i})$ and an $n_i \times n_i$ correlation matrix $R_*^{(i)}(\rho) = \left(r_{j,k}^{(i)}\right)$ with a certain value of correlation parameter ρ, generate a random vector $\boldsymbol{y}_i$ from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_i$ and covariance

matrix $\sigma^2\Sigma_i^*$ with the (j,k)-th element $\sigma^2\sqrt{\mu_{ij}}\sqrt{\mu_{ik}}r_{jk}^{(i)}$, where σ^2 is the dispersion parameter.

The underlying mean and covariance structures of the responses are

$$\mu_{ij} = E(Y_{ij}) = \exp(\beta_0 + \beta_1 x_{ij}), \quad j = 1, \cdots, n_i, i = 1, \cdots, K,$$

and

$$Cov(\boldsymbol{Y}_i) = \sigma^2\Sigma_i^* = \sigma^2 G^*(\mu_i)^{1/2} R_i^*(\rho_0) G^*(\mu_i)^{1/2}, \quad i = 1, \cdots, K,$$

where $G^*(\mu_i)$ is an $n_i \times n_i$ diagonal matrix with the j-th diagonal element μ_{ij}, for $j = 1, \cdots, n_i$. Here, $R^*(\rho_0)$ is the true correlation matrix with a certain value ρ_0 of the correlation parameter, and $V^*(\mu_{ij}) = \mu_{ij}$ is the true unit variance function.

Fit the data by a p-element estimating equation

$$\sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T \left\{G^{1/2}(\boldsymbol{\mu}_i) R_i(\boldsymbol{\rho}) G^{1/2}(\boldsymbol{\mu}_i)\right\}^{-1} \{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} = 0, \tag{4.33}$$

where $\mu_i(\boldsymbol{\beta})$ is an n_i-dimensional vector with j-th element $\mu_{ij} = \exp(\beta_0 + \beta_1 x_{ij})$, and $G(\mu_i) = \text{diag}\left\{V(\mu_{ij})\right\}$ with $V(\cdot)$ the "working" variance function, and $R_i(\boldsymbol{\rho})$ is an $n_i \times n_i$ working correlation matrix, fully specified by the correlation parameter $\boldsymbol{\rho}$.

In the following simulation studies, the true values of the regression coefficients are $\beta_0 = 1$ and $\beta_1 = 2$. The true value of the dispersion parameter $\sigma^2 = 0.01$.

Simulation 4.10 Generate 2000 replicates of the data set

$$\left\{(y_{ij}, x_{ij}), j = 1, \cdots, n_i, i = 1, \cdots, K\right\}$$

from an exchangeable correlation structure with 5 repeated measurements for each subject, and different numbers of subjects $K = 20, 50, 100$. In addition, the true value of the correlation parameter is $\rho = 0.5$. Fit the data by GEE (4.33)

using the variance function $V(\mu) = \mu$, and the true correlation structure, i.e., exchangeable structure. In this simulation, we will consider IR tests for the null hypothesis

$$H_0 : V(\mu_{ij}) = \mu_{ij}, \quad j = 1, \cdots, n_i, i = 1, \cdots, K. \tag{4.34}$$

The previous applications of the IR tests have shown that the pooled IR statistic usually has the best performance. We consider only the pooled IR statistics (4.28) or (4.30), taking ratios of the Godambian estimators (4.15) to the Pearson moment estimator (4.17) or transformed moment estimator (4.18).

Table 4.9 reports the empirical type I errors of the standardized pooled IR statistics, which take ratios of the unbiased pooled Godambian estimator to the unbiased Pearson moment estimator or the transformed moment estimator over different sample sizes. The type I errors are obtained at different significance levels 10%, 5% and 1%. The numerical results have shown that the IR statistics which use the "transformed" moment estimator have better performance, with the type I errors closer to the nominal levels, than those with the unbiased "Pearson" moment estimator.

Table 4.9: The empirical type I errors of the pooled IR statistic IR^s_{pool}, which take ratios of the unbiased pooled Godambian estimator to the unbiased Pearson moment estimator or the transformed moment estimator among 2000 replicates. The type I errors are obtained from the limiting $N(0, 1)$ distribution, over different sample sizes at the significance levels 10%, 5% and 1%, under the null hypothesis H_0 (4.34).

	Pearson moment			Transformed moment		
n	.10	.05	.01	.10	.05	.01
20	0.0530	0.0290	0.0075	0.0815	0.0425	0.0160
50	0.0525	0.0215	0.0045	0.0930	0.0510	0.0110
100	0.0505	0.0205	0.0025	0.0985	0.0515	0.0090

Simulation 4.11 Generate 5000 replicates of the data with 5 observations for each subject and the number of subjects $K = 20$, and 2000 replicates with sample

size $K = 50$ from either exchangeable or AR(1) correlation structure with the true value of the correlation parameter 0.5. Fit the data by the GEE (4.33) using three different working variance functions

$$V_1(\mu) = 1, \quad V_2(\mu) = \mu, \quad V_3(\mu) = \mu^2,$$

combining with three different working correlation structures: independence (IND), exchangeable(EXCH), and AR(1). We will assess the power of the pooled IR statistic to reject the three null hypotheses

$$H_{01} : V_1(\mu) = V^*(\mu); \quad H_{02} : V_2(\mu) = V^*(\mu); \quad H_{03} : V_3(\mu) = V^*(\mu).$$

Note that the true variance function is $V^*(\mu) = \mu$, so only the null hypothesis H_{02} is true. Table 4.10 and Table 4.11 show the empirical power of the pooled IR statistic to reject three different hypotheses H_{01}, H_{02} and H_{03} under three different working correlation structures. They show that the IR statistics are powerful to detect misspecification of variance function, but they are less powerful to test for misspecification of working correlation structure. In addition, as shown in previous simulation studies, the test power becomes stronger as the sample size increases.

Remark 4.1 If the true value of the dispersion parameter σ^2 is assumed to be known, we can still obtain the moment estimator of σ^2. Thus, the IR statistics can be constructed by taking ratios of the Godambian estimator of σ^2 to the moment estimator. However, when using the moment estimator, asymptotic distributions of the IR statistics are based on the first order Taylor approximation of the IR statistics. Thus, the IR statistics taking ratios of the Godambian estimator to the moment estimator are likely to perform more poorly than those taking ratios of the Godambian estimator to its true value. In addition, as shown in Figures 4.1, 4.2 4.3, 4.4, 4.5, 4.6 and 4.7, as well as Table 4.1, Simulation 4.1 indicated that the distributions of the IR statistics using the true value of σ^2 are closer to the limiting $N(0, 1)$ than those using the moment estimator, especially for

Table 4.10: The empirical power of the pooled IR statistic IR^s_{pool}, which takes a ratio of the unbiased pooled Godambian estimator to the unbiased Pearson moment estimator (4.17) or the transformed moment estimator (4.18). The results are obtained from the limiting $N(0,1)$ distribution among 5000 replicates for small sample size $K = 20$, and 2000 replicates for the sample size $K = 50$ from the true correlation structure: exchangeable with the correlation parameter 0.5, at the significance level 5% to reject the null hypotheses H_{01}, H_{02} and H_{03}.

		Godambian versus Pearson moment		
		$H_{01} : V_1(\mu) = \mu$	$H_{02} : V_2(\mu) = \mu$	$H_{03} : V_3(\mu) = \mu$
	IND	1.0000	**0.6338**	0.9996
$K = 20$	EXCH	1.0000	**0.0276**	0.9996
	AR(1)	1.0000	**0.1530**	0.9998
	IND	1.0000	**0.9370**	1.0000
$K = 50$	EXCH	1.0000	**0.0265**	1.0000
	AR(1)	1.0000	**0.2435**	1.0000
		Godambian versus Transformed moment		
		$H_{01} : V_1(\mu) = \mu$	$H_{02} : V_2(\mu) = \mu$	$H_{03} : V_3(\mu) = \mu$
	IND	1.0000	**0.6338**	0.9996
$K = 20$	EXCH	1.0000	**0.0434**	0.9996
	AR(1)	1.0000	**0.0614**	0.9998
	IND	1.0000	**0.9370**	1.0000
$K = 50$	EXCH	1.0000	**0.0540**	1.0000
	AR(1)	1.0000	**0.1735**	1.0000

Table 4.11: The empirical power of the pooled IR statistic IR^s_{pool}, which takes a ratio of the unbiased pooled Godambian estimator to the unbiased Pearson moment estimator (4.17) or the transformed moment estimator (4.18). The results are obtained from the limiting $N(0,1)$ distribution among 5000 replicates for small sample size $K = 20$, and 2000 replicates for the sample size $K = 50$ from the true correlation structure: AR(1) with the correlation parameter 0.5, at the significance level 5% to reject the null hypotheses H_{01}, H_{02} and H_{03}.

		Godambian versus Pearson moment		
		$H_{01} : V_1(\mu) = \mu$	$H_{02} : V_2(\mu) = \mu$	$H_{03} : V_3(\mu) = \mu$
	IND	1.0000	**0.4616**	0.9940
$K = 20$	EXCH	1.0000	**0.1714**	0.9942
	AR(1)	1.0000	**0.0520**	0.9918
	IND	1.0000	**0.8165**	1.0000
$K = 50$	EXCH	1.0000	**0.3155**	1.0000
	AR(1)	1.0000	**0.0400**	1.0000
		Godambian versus Transformed moment		
		$H_{01} : V_1(\mu) = \mu$	$H_{02} : V_2(\mu) = \mu$	$H_{03} : V_3(\mu) = \mu$
	IND	1.0000	**0.4616**	0.9940
$K = 20$	EXCH	1.0000	**0.1892**	0.9944
	AR(1)	1.0000	**0.0522**	0.9902
	IND	1.0000	**0.8165**	1.0000
$K = 50$	EXCH	1.0000	**0.3420**	1.0000
	AR(1)	1.0000	**0.0430**	1.0000

small sample size. Therefore, in some applications where the true value of the dispersion parameter σ^2 is assumed to be known from external information or previous studies, it is suggested that we use the IR statistics which take ratios of the Godambian estimator to its true value.

Remark 4.2 In the literature of GEE, several methods of estimating the correlation parameter in the working correlation structure have been suggested. See [53], [67] and [84]. No matter which method is used to estimate the correlation parameter, the only necessary assumption required in Theorem 4.1, Theorem 4.3 and Theorem 4.4 is that the sequence of the estimators $\{\widehat{\boldsymbol{\rho}}_K, K = 1, 2, \cdots, \}$ converges to a certain value $\boldsymbol{\rho}_*$. Moreover, under the null hypothesis H_0 (4.2), the limiting value $\boldsymbol{\rho}_*$ is equal to the true value of the correlation parameter involved in the true correlation structure.

Chapter 5

Model Selection

Model selection problems are encountered almost everywhere. In high dimensional data analysis, for example, gene expression analysis, the number of covariates is considerably larger than the sample size. Variable selection is an important method to increase the power of statistical conclusions and to facilitate the biological interpretation. In the analysis of time series, it is essential to know the true order of an ARMA model. In the analysis of clustering, it is important to determine the number of clusters. In the statistical literature, numerous model selection procedures have been intensively discussed. They are developed based on hypothesis testing, prediction errors, cross-validation and information measurement and so on. [70] surveyed several model selection methods with reference to regression, categorical data and time series analysis. However, almost all of the criteria are for the selection of the optimal mean structure. There is a lack of a systematic criterion for selecting the variance/covariance structure.

In this section, we will develop two different model selection procedures: one is based on the information ratio tests proposed in Chapter 4, and the other is based on a discrepancy measurement related with information matrices. In addition, these two model selection methods will be illustrated by two simulation studies: selecting the optimal variance function in compound Poisson models and selecting the true working correlation structure in GEE (2.27).

5.1 Model Selection Criterion

Let $\mathfrak{V} = \left\{\mathcal{V}_\gamma, \gamma \in \Gamma\right\}$ be a class of candidate variance or covariance structures for the given data, where Γ is an index set. For instance, in the analysis of insurance claim data, Tweedie's compound Poisson models are commonly considered. (See [49] and [90]). In these models, the variance function takes a form of power function as follows:

$$V(\mu) = \mu^\kappa, \quad 1 < \kappa < 2.$$

However, the value of κ is rarely known in practice. Given a list of candidate values of κ, for example, $\Gamma = \{\kappa = 1.1, 1.2, 1.3, \cdots, 1.9\}$, a class of variance functions is given by

$$\mathfrak{V} = \{V_\kappa : V_\kappa(\mu) = \mu^\kappa, \kappa \in \Gamma\},$$

where the shape parameter κ is the index. In another example of selecting the optimal working correlation structure in GEE (2.27), we may define a class of damped exponential correlation structures proposed by [63]. The correlation between two observations, taken on the same subject, separated by s-units of time was modelled as ρ^{s^θ}, where ρ is the correlation between elements separated by one s-unit, and θ is a damping parameter which permits attenuation or acceleration of the exponential decay of the autocorrelation function defining an AR(1). Note that $\rho = 0$ corresponds to the independence correlation, $\theta = 0$ corresponds to the exchangeable (or compound symmetric) correlation, and $\theta = 1$ corresponds to the AR(1) correlation. Then, a class of candidate correlation structures is given by

$$\mathfrak{V} = \left\{R_{\rho,\theta}; 0 \le \rho \le 1, \theta \ge 0\right\},$$

where (ρ, θ) is the index.

We should bear in mind that it is possible that the true variance/covariance structure is not included in $\mathfrak{V} = \left\{\mathcal{V}_\gamma\right\}$. Then, based on the data, we need to select

the optimal one among those given in $\mathfrak{V}$, which is not necessarily the true one, but the closest approximation, by using a suitable model selection criterion.

5.1.1 Selection of Variance/Covariance Structure based on Hypothesis Testing

In Chapter 4, IR tests were proposed to test for misspecification of variance/covariance structure. Several theorems, for example, Theorem 4.1 and Theorem 4.2, stated that the proposed standardized IR statistics are asymptotically $N(0,1)$ distributed, under the null hypothesis that the variacne/covariance structure is correctly specified. In this section, we will propose a model selection procedure based on a sequence of IR tests. Specifically, for each candidate variance/covariance structure $\mathcal{V}_\gamma \in \mathfrak{V}$, we will test for the null hypothesis

$$H_{0,\gamma} : \mathcal{V}_\gamma = \mathcal{V}^*,$$

using the IR statistic, denoted by IR_γ, obtained from the estimating equation using $\mathcal{V}_\gamma$ as the working variance/covariance structure, where $\mathcal{V}^*$ represents the true variance/covariance structure.

Selection of the optimal variance function

Consider a class of candidate variance functions $\mathfrak{V} = \left\{V_\gamma(\cdot), \gamma \in \Gamma\right\}$. First of all, we need to define equivalency of variance functions.

Definition 5.1 *Two variance functions V_1 and V_2 are said to be equivalent up to a constant, if there exists a constant c such that $V_1(\mu) = cV_2(\mu)$ for any value of μ.*

Two assumptions should be made in the following investigation.

Assumption 5.1 *For the given class of candidate variance functions*

$$\mathfrak{V} = \left\{ V_\gamma(\cdot), \gamma \in \Gamma \right\},$$

(i) any variance function in $\mathfrak{V}$ is not equivalent to all the others; and

(ii) if the true variance function $V^*(\cdot)$ is not included in $\mathfrak{V}$, it is not equivalent to any function $V_\gamma(\cdot) \in \mathfrak{V}$.

Assumption 5.2 *Let $\mathfrak{Y}$ be the domain of the expectation of the responses. Assuming that $\mathfrak{Y}$ is compact, for any candidate variance function $V_\gamma(\cdot) \in \mathfrak{V}$ and the true unit variance function $V^*(\cdot)$, there exists a constant M such that*

$$\left| \frac{V^*(\mu)}{V_\gamma(\mu)} \right| \leq M, \quad \mu \in \mathfrak{Y},$$

where $|\cdot|$ denotes the absolute value of a real number.

In the context of GLM, given an observed sample of size n and a candidate variance function $V_\gamma \in \mathfrak{V}$ used in the quasi-score equation (2.8), we consider an IR statistic, which takes a ratio of an unbiased Godambian estimator of the dispersion parameter σ^2 to its true value, if given. That is,

$$IR_{\gamma,n} = \frac{\widetilde{\sigma}^2}{\sigma^2} = \left(\frac{\boldsymbol{r}_P}{\sigma} \right)^T \widehat{\boldsymbol{W}} \left(\frac{\boldsymbol{r}_P}{\sigma} \right),$$

where $\widetilde{\sigma}^2$ could be any unbiased individual coefficient-specific Godambian estimator $\widetilde{\sigma}^2_{j-1,u}$ (4.6), or the unbiased pooled Godambian estimator $\widetilde{\sigma}^2_{pool,u}$ (4.8), and correspondingly, the matrix $\widehat{\boldsymbol{W}}$ could be $\widehat{\boldsymbol{W}}_u^{(j-1)}$ or $\widehat{\boldsymbol{W}}_u^{pool}$. Shown in the proof of Theorem 4.1, in the context of GLM, the statistic $IR_{\gamma,n}$ can be approximated by

$$IR_{\gamma,n} \simeq \left(\frac{\boldsymbol{\epsilon}_P}{\sigma} \right)^T \left\{ (\boldsymbol{I}_n - H_*) \boldsymbol{W}_* (\boldsymbol{I}_n - H_*) \right\} \left(\frac{\boldsymbol{\epsilon}_P}{\sigma} \right).$$

Note that $tr\{(\boldsymbol{I}_n - H_*)\boldsymbol{W}_*(\boldsymbol{I}_n - H_*)\} = 1$. The vector $\boldsymbol{\epsilon}_P/\sigma$ is multivariate distributed with mean $\mathbf{0}$ and covariance matrix Ω, an $n \times n$ diagonal matrix with the i-th diagonal element $\omega_i = V^*(\mu_{i,*})/V_\gamma(\mu_{i,*})$. The expectation of the statistic $IR_{\gamma,n}$ is given by

$$\varsigma_{\gamma,n} = E\left(IR_{\gamma,n}\right) \simeq tr\left\{\Omega^{1/2}(\boldsymbol{I}_n - H_*)\boldsymbol{W}_*(\boldsymbol{I}_n - H_*)\Omega^{1/2}\right\}.$$

If the null hypothesis $H_{0,\gamma} : V_\gamma(\cdot) = V^*(\cdot)$ is true, so that the matrix $\Omega = \boldsymbol{I}_n$, then $\varsigma_{\gamma,n} = E\left(IR_{\gamma,n}\right) \simeq 1$. On the other hand, if the null hypothesis $H_{0,\gamma}$ is not true, the matrix Ω is no longer an identity matrix. In this case, the expectation can be expressed as

$$\varsigma_{\gamma,n} \simeq \sum_{i=1}^{n} \omega_i \vartheta_{ii} = 1 + \sum_{i=1}^{n} (\omega_i - 1)\vartheta_{ii}, \tag{5.1}$$

where ϑ_{ii} are the diagonal elements of the matrix $(\boldsymbol{I}_n - H_*)\boldsymbol{W}_*(\boldsymbol{I}_n - H_*)$. Note that $\sum_{i=1}^{n} \vartheta_{ii} = 1$.

If an IR statistic takes a ratio of an unbiased Godambian estimator to the moment estimator of the dispersion parameter, it can be approximated by

$$IR_{\gamma,n} \simeq 1 + \left(\frac{\boldsymbol{\epsilon}_P}{\sigma}\right)^T \left\{(\boldsymbol{I}_n - H_*)\left(\boldsymbol{W}_* - \frac{1}{n-p}\boldsymbol{I}_n\right)(\boldsymbol{I}_n - H_*)\right\}\left(\frac{\boldsymbol{\epsilon}_P}{\sigma}\right).$$

If the null hypothesis $H_{0,\gamma}$ is true, the expectation, denoted by $\varsigma_{\gamma,n}$, of $IR_{\gamma,n}$ is approximately 1; otherwise, the expectation is given by

$$\varsigma_{\gamma,n} \simeq 1 + \sum_{i=1}^{n} (\omega_i - 1)\vartheta_{ii} - \sum_{i=1}^{n} (\omega_i - 1)\frac{1 - h_{ii}^*}{n-p}, \tag{5.2}$$

where h_{ii}^* is the i-th diagonal element of the matrix H_*. Note that under the Assumption 5.1, $\varsigma_{\gamma,n} - 1$ is a nonzero constant.

Equations (5.1) and (5.2) indicate that the expectation $\varsigma_{\gamma,n}$ is a function of ω_i. It can show that, under the Assumption 5.2, $|\omega_i| = \left|V^*(\mu_{i,*})/V_\gamma(\mu_{i,*})\right|$ is bounded,

and consequently, $|\omega_i - 1|$ is also bounded. Since $\sum_{i=1}^{n} \vartheta_{ii} = 1$, from the equation (5.1),

$$\left|\varsigma_{\gamma,n} - 1\right| = \left|\sum_{i=1}^{n} (\omega_i - 1)\vartheta_{ii}\right| = O(1).$$

Similarly, from the equation (5.2),

$$\left|\varsigma_{\gamma,n} - 1\right| = \left|\sum_{i=1}^{n} (\omega_i - 1)\vartheta_{ii} - \sum_{i=1}^{n} (\omega_i - 1)\frac{1 - h_{ii}^*}{n - p}\right| = O(1),$$

Moreover, the magnitude of $|\varsigma_{\gamma,n} - 1|$ depends on the departure of the candidate variance function $V_\gamma(\cdot)$ from the true one $V^*(\cdot)$. That is, the larger the relative discrepancy, i.e. $V^*/V_\gamma - 1$ is, the further the expectation $\varsigma_{\gamma,n}$ differs from 1. Let $\widehat{\lambda}_k$, $k = 1, \cdots, n$, be the eigenvalues of the matrix

$$\left(\boldsymbol{I}_n - \widehat{H}\right)\widehat{\boldsymbol{W}}\left(\boldsymbol{I}_n - \widehat{H}\right) \quad \text{or} \quad \left(\boldsymbol{I}_n - \widehat{H}\right)\left(\widehat{\boldsymbol{W}} - \frac{1}{n-p}\boldsymbol{I}_n\right)\left(\boldsymbol{I}_n - \widehat{H}\right).$$

The standardized IR statistic

$$IR_{\gamma,n}^s = \frac{IR_{\gamma,n} - 1}{\sqrt{2\sum_{k=1}^{n} \widehat{\lambda}_k^2}} \quad \xrightarrow{d} \quad N(0,1), \quad \text{as} \quad n \to \infty,$$

if the null hypothesis $H_{0,\gamma} : V_\gamma(\cdot) = V^*(\cdot)$ is true. Then, given a significance level α, the p-value

$$p_\gamma(\alpha) = 2P\left(Z \geq |IR_{\gamma,obs}^s|\right)$$

is equal to or larger than α, where Z is a $N(0,1)$ random variable, and $IR_{\gamma,obs}^s$ is the observed value of the statistic $IR_{\gamma,n}^s$ obtained from the given data.

If the null hypothesis $H_{0,\gamma}$ is false,

$$\frac{IR_{\gamma,n} - \varsigma_{\gamma,n}}{\sqrt{2\sum_{k=1}^{n} \upsilon_k^2}} \quad \xrightarrow{d} \quad N(0,1), \quad \text{as} \quad n \to \infty,$$

where $\varsigma_{\gamma,n}$ is given in (5.1) or (5.2), and υ_k, $k = 1, \cdots, n$ are the eigenvalues of the matrix

$$\Omega^{1/2}\left(\boldsymbol{I}_n - H_*\right)\boldsymbol{W}_*\left(\boldsymbol{I}_n - H_*\right)\Omega^{1/2},$$

or

$$\Omega^{1/2}\left(\boldsymbol{I}_n - H_*\right)\left(\boldsymbol{W}_* - \frac{1}{n-p}\boldsymbol{I}_n\right)\left(\boldsymbol{I}_n - H_*\right)\Omega^{1/2}.$$

In this case, for large sample size n,

$$IR^s_{\gamma,n} = \frac{IR_{\gamma,n} - 1}{\sqrt{2\sum_{k=1}^n \widehat{\lambda}_k^2}} \sim N\left(\frac{\varsigma_{\gamma,n} - 1}{\sqrt{2\sum_{k=1}^n \widehat{\lambda}_k^2}}, \sqrt{\frac{\sum_{k=1}^n \upsilon_k^2}{\sum_{k=1}^n \widehat{\lambda}_k^2}}\right),$$

approximately. Since $\sum_{k=1}^n \widehat{\lambda}_k = 1$, then $\sqrt{2\sum_{k=1}^n \widehat{\lambda}_k^2} = o(1)$, and consequently, the magnitude of $(\varsigma_{\gamma,n} - 1)/\sqrt{2\sum_{k=1}^n \widehat{\lambda}_k^2}$ is considerably large valued. In addition, under the Assumption 5.2, $\sum_{k=1}^n \upsilon_k^2 / \sum_{k=1}^n \widehat{\lambda}_k^2$ is bounded. Therefore, the p-value

$$p_\gamma(\alpha) = 2P\left(Z \geq |IR^s_{\gamma,obs}|\right)$$

is much smaller than α. Thus, a model selection procedure can be suggested as follows. Given a class of variance functions $\mathfrak{V} = \left\{V_\gamma(\cdot), \gamma \in \Gamma\right\}$, for each candidate variance function V_γ, a p-value, p_γ, is obtained from the test for $H_{0,\gamma}$: $V_\gamma = V^*$, using a standardized IR statistic at a significance level α. The optimal variance function is the selected one with the maximum p-value, i.e.,

$$V_{opt} = \arg\max_{V_\gamma \in \mathfrak{V}} \left\{p_\gamma(\alpha), \gamma \in \Gamma\right\},$$

for a given significance level α. A same model selection procedure may be obtained for the selection of the optimal variance function in GEE.

Selection of the optimal working correlation structure in GEE

Suppose that the variance function is correctly specified in the context of GEE, that is, $G_i = G_i^*$, for $i = 1, \cdots, K$. In this section, we want to select the optimal working correlation structure from a given class of working correlation matrices $\mathfrak{V} = \left\{R_\gamma, \gamma \in \Gamma\right\}$.

Given a candidate working correlation matrix R_γ, an IR statistic can be approximated by

$$IR_{\gamma,K} \simeq \left(\frac{\widetilde{\boldsymbol{\epsilon}}}{\sigma}\right)^T \left\{\left(\boldsymbol{I}_N - \mathcal{H}_*\right) \mathcal{W}_* \left(\boldsymbol{I}_N - \mathcal{H}_*\right)\right\} \left(\frac{\widetilde{\boldsymbol{\epsilon}}}{\sigma}\right),$$

which is a ratio of an unbiased Godambian estimator of σ^2 to its true value, or

$$IR_{\gamma,K} \simeq 1 + \left(\frac{\widetilde{\boldsymbol{\epsilon}}}{\sigma}\right)^T \left\{\left(\boldsymbol{I}_N - \mathcal{H}_*\right) \left(\mathcal{W}_* - \mathcal{W}_m\right) \left(\boldsymbol{I}_N - \mathcal{H}_*\right)\right\} \left(\frac{\widetilde{\boldsymbol{\epsilon}}}{\sigma}\right),$$

which is a ratio of an unbiased Godambian estimator to the moment estimator. For example, the matrix $\mathcal{W}_*$ could be $\mathcal{W}_{u,*}^{(j-1)}$ for the unbiased β_{j-1}-specific Godambian estimator (4.13), or $\mathcal{W}_{u,*}^{pool}$ for the unbiased pooled Godambian estimator (4.15), and the matrix $\mathcal{W}_m$ could be $\mathcal{W}_{_P}^*/m_{_P}^*$ for the unbiased Pearson moment estimator (4.17), or $\frac{1}{N-p}\boldsymbol{I}_N$ for the transformed moment estimator (4.18). The vector $\widetilde{\boldsymbol{\epsilon}}/\sigma$ is multivariate distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is a $K \times K$ diagonal block matrix with the i-th diagonal matrix $\Omega_i = L_{\gamma,i,*}^{-1} R_i^* L_{\gamma,i,*}^{-T}$, because the variance function is correctly specified, i.e., $G_i = G_i^*$. If the null hypothesis $H_{0,\gamma} : R_\gamma = R^*$ is true, the matrix $\boldsymbol{\Omega} = \boldsymbol{I}_N$. Then, the expectation, $\varsigma_{\gamma,K}$, of the IR statistic $IR_{\gamma,K}$ is approximately 1.

By the Cholesky decomposition, the matrix R_i^* can be decomposed as

$$R_i^* = L_i^* L_i^{*T}, \quad i = 1, \cdots, K,$$

where L_i^* is a lower triangular matrix. If the null hypothesis $H_{0,\gamma}$ is false, the

expectation of the IR statistic is given by

$$\varsigma_{\gamma,K} \simeq \sum_{i=1}^{K} tr\left\{L_{\gamma,i,*}^{-1} L_i^* \digamma_{ii} L_i^{*T} L_{\gamma,i,*}^{-T}\right\} = 1 + \sum_{i=1}^{K} tr\left\{\left(L_i^{*T} R_{\gamma,i,*}^{-1} L_i^* - \boldsymbol{I}_{n_i}\right) \digamma_{ii}\right\},$$

where $\digamma_{ii}$ is the diagonal block matrix of

$$\left(\boldsymbol{I}_N - \mathcal{H}_*\right) \mathcal{W}_* \left(\boldsymbol{I}_N - \mathcal{H}_*\right),$$

or

$$\varsigma_{\gamma,K} \simeq 1 + \sum_{i=1}^{K} tr\left\{L_{\gamma,i,*}^{-1} L_i^* \digamma_{ii} L_i^{*T} L_{\gamma,i,*}^{-T}\right\} = 1 + \sum_{i=1}^{K} tr\left\{L_i^{*T} R_{\gamma,i,*}^{-1} L_i^* \digamma_{ii}\right\},$$

where $\digamma_{ii}$ is the diagonal block matrix of

$$\left(\boldsymbol{I}_N - \mathcal{H}_*\right) \left(\mathcal{W}_* - \mathcal{W}_m\right) \left(\boldsymbol{I}_N - \mathcal{H}_*\right).$$

Note that $\digamma_{ii}$ has the same role as ϑ_{ii} given in (5.1) or (5.2). Similarly to the selection of the optimal variance function, the magnitude of $|\varsigma_{\gamma,K} - 1|$ depends on the eigenvalues of the matrix $L_i^{*T} R_{\gamma,i,*}^{-1} L_i^*$, which characterizes the departure of the candidate working correlation structure from the true one. If the null hypothesis $H_{0,\gamma}$ is true, the p-value obtained from the standardized IR statistic is equal to or larger than α, for a given significance level α. If the null hypothesis is false, the p-value is considerably smaller than α. Thus, among the class of working correlation structures $\mathfrak{V} = \left\{R_\gamma, \gamma \in \Gamma\right\}$, the optimal one R_{opt} leads to the maximum p-value.

However, from Section 4.3.3, we have noticed that the IR statistics are relatively less sensitive to the discrepancy of the working correlation structures than that of variance functions. Thus, the model selection procedure based on the IR tests might not be powerful to detect the optimal correlation structure.

5.1.2 Information Discrepancy Criterion

Several model selection criteria have been built upon the use of Kullback-Leibler distance (or information) between the true underlying distribution and the distributional model imposed for parameter estimation. For example, Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) take a form of $-2\log L(\widehat{\boldsymbol{\theta}})$ plus a certain penalty term, where $L(\widehat{\boldsymbol{\theta}})$ is the maximum likelihood, which is the likelihood function evaluated at the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$. In the context of estimating equations, because no parametric density function is assumed, the likelihood function is unavailable. Thus, these popular information criteria cannot be directly used. [65] proposed a modification to AIC, named the "quasi-log-likelihood under the independence working correlation information criteria" (QIC). Later, [43] suggested to use only the penalty term in the QIC for selecting a working correlation structure. The penalty term takes a ratio of two information matrices: one is the model-based covariance matrix estimator using the independence working correlation, and the other is the sandwich covariance matrix estimator using a general correlation structure. It indicates that the departure of the candidate working correlation structure from the true one can be reflected from the ratio of the model-based and sandwich covariance matrix estimators.

Let $\mathfrak{V} = \left\{\mathcal{V}_\gamma, \gamma \in \Gamma\right\}$ be a class of variance or covariance structures for the data, where Γ is an index set. Corresponding to the candidate variance or covariance structures, we define a class of candidate estimating equations $\mathfrak{G} = \left\{\Psi_\gamma(\boldsymbol{\beta}), \gamma \in \Gamma\right\}$.

Selection of the optimal variance function in GLM

In the context of GLM, given a candidate variance function $V_\gamma \in \mathfrak{V}$, a quasi-likelihood method requires us to specify an additive estimating function $\Psi_{\gamma,n}(\boldsymbol{\beta})$,

given by (2.8)

$$\Psi_{\gamma,n}(\boldsymbol{\beta}) = \frac{1}{\sigma^2} X^T \Delta \mathcal{V}_\gamma^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}) \right),$$

which is used to estimate the regression coefficients $\boldsymbol{\beta}$, based on the data

$$\{(y_i, \boldsymbol{x}_i), i = 1, \cdots, n\}.$$

Given the estimating function $\Psi_{\gamma,n}(\boldsymbol{\beta})$, the aggregated sensitivity and variability matrices are given by, respectively, (2.9) and (2.10)

$$\mathbf{S}_{\Psi_{\gamma,n}}(\boldsymbol{\beta}) = E\left\{ \frac{\partial \Psi_{\gamma,n}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right\} = -\frac{1}{\sigma^2} X^T \Delta \mathcal{V}_\gamma^{-1} \Delta X,$$

and

$$\mathbf{V}_{\Psi_{\gamma,n}}(\boldsymbol{\beta}) = E\left\{ \Psi_{\gamma,n}(\boldsymbol{\beta}) \Psi_{\gamma,n}^T(\boldsymbol{\beta}) \right\} = \frac{1}{\sigma^2} X^T \Delta \mathcal{V}_\gamma^{-1} \mathcal{V}^* \mathcal{V}_\gamma^{-1} \Delta X,$$

where $Cov(\boldsymbol{Y}) = \sigma^2 \mathcal{V}^*$. Given any value of $\boldsymbol{\beta}$, define two limiting information matrices by

$$\mathbf{S}_\gamma(\boldsymbol{\beta}) = \lim_{n \to \infty} \frac{1}{n} \mathbf{S}_{\Psi_{\gamma,n}}(\boldsymbol{\beta}),$$

and

$$\mathbf{V}_\gamma(\boldsymbol{\beta}) = \lim_{n \to \infty} \frac{1}{n} \mathbf{V}_{\Psi_{\gamma,n}}(\boldsymbol{\beta}).$$

To measure the discrepancy between these two information matrices, under a certain candidate variance function, we define an information matrix ratio by

$$\mathcal{IMR}_\gamma(\boldsymbol{\beta}) = \left\{ -\mathbf{S}_\gamma(\boldsymbol{\beta}) \right\}^{-1} \mathbf{V}_\gamma(\boldsymbol{\beta}). \tag{5.3}$$

If the variance function is correctly specified, i.e., $V_\gamma(\cdot) = V^*(\cdot)$, the information unbiasedness holds at the true value $\boldsymbol{\beta}_*$ of the parameter $\boldsymbol{\beta}$, that is,

$$-\mathbf{S}_\gamma(\boldsymbol{\beta}_*) = \mathbf{V}_\gamma(\boldsymbol{\beta}_*).$$

Thus, in this case, the information matrix ratio is an identity matrix under the true variance function, that is, $\mathcal{IMR}_*(\boldsymbol{\beta}_*) = \boldsymbol{I}_p$.

Let $\mathcal{D}(V_\gamma, V^*)$ be a function characterizing the discrepancy between the candidate variance function and the true one. It can be defined by a measurement of the discrepancy between these two information matrices, which is the distance between $\mathcal{IMR}_\gamma$, the information matrix ratio under the candidate variance function V_γ, and $\mathcal{IMR}_* = \boldsymbol{I}_p$ under the true variance function V^*. (See [81]). Define

$$\mathcal{D}(V_\gamma, V^*) = tr\left\{[\mathcal{IMR}_\gamma(\boldsymbol{\beta}_*) - \mathcal{IMR}_*(\boldsymbol{\beta}_*)]^2\right\} = tr\left\{[\mathcal{IMR}_\gamma(\boldsymbol{\beta}_*) - \boldsymbol{I}_p]^2\right\}. \quad (5.4)$$

However, in practice, the true value of $\boldsymbol{\beta}$ is unknown, and the explicit forms of the limiting matrices $\mathbf{S}_\gamma$ and $\mathbf{V}_\gamma$ are not available. Based on the data, the sensitivity and variability matrices are estimated by (2.16) and (2.18)

$$\widehat{\mathbf{S}}_{\gamma,n}(\widehat{\boldsymbol{\beta}}) = -\frac{1}{\sigma^2} X^T \widehat{\Delta} \widehat{\mathcal{V}}_\gamma^{-1} \widehat{\Delta} X,$$

and

$$\widehat{\mathbf{V}}_{\gamma,n}(\widehat{\boldsymbol{\beta}}) = \frac{1}{(\sigma^2)^2} X^T \widehat{\Delta} \widehat{\mathcal{V}}_\gamma^{-1} \mathcal{R} \widehat{\mathcal{V}}_\gamma^{-1} \widehat{\Delta} X,$$

if the true value of σ^2 is known; otherwise, it is replaced by a moment estimator. Then, the information matrix ratio can be estimated by

$$\mathcal{IMR}_{\gamma,n}(\widehat{\boldsymbol{\beta}}) = \left\{-\frac{1}{n}\widehat{\mathbf{S}}_{\gamma,n}(\widehat{\boldsymbol{\beta}})\right\}^{-1} \left\{\frac{1}{n}\widehat{\mathbf{V}}_{\gamma,n}(\widehat{\boldsymbol{\beta}})\right\} = \left\{-\widehat{\mathbf{S}}_{\gamma,n}(\widehat{\boldsymbol{\beta}})\right\}^{-1} \widehat{\mathbf{V}}_{\gamma,n}(\widehat{\boldsymbol{\beta}}),$$

and consequently, the estimated information discrepancy function can be obtained by

$$d(V_\gamma, V^*) = tr\left\{[\mathcal{IMR}_{\gamma,n}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{I}_p]^2\right\}. \quad (5.5)$$

Let $\lambda_\gamma = (\lambda_{\gamma,1}, \cdots, \lambda_{\gamma,p})$, be the eigenvalues of the matrix $\mathcal{IMR}_{\gamma,n}(\widehat{\boldsymbol{\beta}})$. Then, the estimated information discrepancy function can be re-written as

$$d(V_\gamma, V^*) = \sum_{j=1}^{p} (\lambda_{\gamma,j} - 1)^2 = \| \boldsymbol{\lambda}_\gamma - \mathbf{1}_p \|^2,$$

where $\mathbf{1}_p$ is a $p \times 1$ vector of 1's, and $\|\cdot\|$ denotes the Euclidean distance. Let $\widehat{U} = \widehat{\mathcal{V}}_\gamma^{-1/2}\widehat{\Delta X}$, then the estimated sensitivity and variability matrices can be re-written as, respectively,

$$-\widehat{\mathbf{S}}_{\gamma,n}(\widehat{\boldsymbol{\beta}}) = \frac{1}{\sigma^2}\widehat{U}^T\widehat{U},$$

$$\widehat{\mathbf{V}}_{\gamma,n}(\widehat{\boldsymbol{\beta}}) = \frac{1}{(\sigma^2)^2}\widehat{U}^T\mathcal{R}_{P}\widehat{U},$$

where $\mathcal{R}_{P} = \text{diag}\left\{r_{P,i}^2\right\}$. By the QR decomposition, the matrix $\widehat{U}$ can be decomposed by

$$\widehat{U} = \Pi Q,$$

where Π is an $n \times p$ orthogonal matrix, and Q is a $p \times p$ matrix. Then the estimated information matrix ratio $\mathcal{IMR}_{\gamma,n}(\widehat{\boldsymbol{\beta}})$ can be written as

$$\mathcal{IMR}_{\gamma,n}(\widehat{\boldsymbol{\beta}}) = \frac{1}{\sigma^2}Q^{-1}\Pi^T\mathcal{R}_{P}\Pi Q.$$

By the eigen-decomposition, there exists a $p \times p$ matrix $\mathcal{E}$ such that

$$\mathcal{IMR}_{\gamma,n}(\widehat{\boldsymbol{\beta}}) = \mathcal{E}\Lambda\mathcal{E}^{-1},$$

where $\Lambda = \text{diag}\left\{\lambda_{\gamma,1}, \cdots, \lambda_{\gamma,p}\right\}$. Thus,

$$\Lambda = \frac{1}{\sigma^2}\mathcal{E}^{-1}Q^{-1}\Pi^T\mathcal{R}_{P}\Pi Q\mathcal{E} = \frac{1}{\sigma^2}\mathcal{P}\mathcal{R}_{P}\mathcal{Q},$$

where $\mathcal{P} = \mathcal{E}^{-1}Q^{-1}\Pi^T$ is an $p \times n$ matrix with (i, j)-th element p_{ij}, and $\mathcal{Q} = \Pi Q\mathcal{E}$ is an $n \times p$ matrix with (i, j)-th element q_{ij}. It can be shown that $\lambda_{\gamma,j}$ can be written as a quadratic form in Pearson residuals. That is,

$$\lambda_j = \sum_{i=1}^{n} p_{ji}q_{ij}r_{P,i}^2/\sigma^2 = \left(\frac{\boldsymbol{r}_P}{\sigma}\right)^T \mathcal{K}_j\left(\frac{\boldsymbol{r}_P}{\sigma}\right), \quad j = 1, \cdots, p,$$

where $\mathcal{K}_j$ is an $n \times n$ diagonal matrix with the i-th diagonal element $\kappa_i^{(j)} = p_{ji}q_{ij}$, for $i = 1, \cdots, n$. Note that $\sum_{i=1}^n \kappa_i^{(j)} = 1$ for $j = 1, \cdots, p$, from $\mathcal{P}\mathcal{Q} = \boldsymbol{I}_p$. From the approximation (4.4), the eigenvalue λ_j can be approximated by

$$\lambda_j \simeq \left(\frac{\boldsymbol{\epsilon}_p}{\sigma}\right)^T (\boldsymbol{I}_n - H_*)\,\mathcal{K}_j^*\,(\boldsymbol{I}_n - H_*)\left(\frac{\boldsymbol{\epsilon}_p}{\sigma}\right),$$

where $\boldsymbol{\epsilon}_p/\sigma$ is a multivariate random vector with mean $\mathbf{0}$, and covariance matrix Ω, which is an $n \times n$ diagonal matrix with the i-th diagonal element $\omega_i = V^*(\mu_{i,*})/V_\gamma(\mu_{i,*})$.

Let $\eta_1, \cdots, \eta_n$ be the eigenvalues of the matrix

$$\Omega^{1/2}\,(\boldsymbol{I}_n - H_*)\,\mathcal{K}_j^*\,(\boldsymbol{I}_n - H_*)\,\Omega^{1/2}.$$

Then the expectation of λ_j is $\sum_{k=1}^n \eta_k$ and its variance is $2\sum_{k=1}^n \eta_k^2$. If $V_\gamma = V^*$, the covariance matrix of $\boldsymbol{\epsilon}_p/\sigma$ is $\boldsymbol{I}_n$, then the bias of λ_j, $\left|E(\lambda_j) - 1\right|$, is of the order $o(1)$, and the variance of λ_j is of the order $o(1)$. Consequently, $E(\lambda_j - 1)^2 = o(1)$, and then,

$$E\left[d(V_\gamma, V^*)\right] = \sum_{j=1}^p E(\lambda_j - 1)^2 = o(1).$$

If $V_\gamma \neq V^*$, the bias $\left|E(\lambda_j) - 1\right|$ is of the order $O(1)$, and the variance is still of the order $o(1)$. Consequently, $E\left[d(V_\gamma, V^*)\right] = O(1)$. The optimal variance function is defined as the one with the minimum estimated information discrepancy, that is,

$$V_{opt} = \arg\min_{V_\gamma \in \mathfrak{V}} \left\{d(V_\gamma, V^*), \gamma \in \Gamma\right\}.$$

We call $d(V_\gamma, V^*)$ the *information discrepancy criterion (IDC)*.

Remark 5.1 *The model-based and sandwich covariance matrix estimators of* $\widehat{\boldsymbol{\beta}}$ *are given by*

$$ASCOV_m(\widehat{\boldsymbol{\beta}}) = \left\{-\widehat{\boldsymbol{S}}_{\gamma,n}(\widehat{\boldsymbol{\beta}})\right\}^{-1},$$

and

$$ASCOV_s(\widehat{\boldsymbol{\beta}}) = \left\{-\widehat{\boldsymbol{S}}_{\gamma,n}(\widehat{\boldsymbol{\beta}})\right\}^{-1} \widehat{\boldsymbol{V}}_{\gamma,n}(\widehat{\boldsymbol{\beta}}) \left\{-\widehat{\boldsymbol{S}}_{\gamma,n}(\widehat{\boldsymbol{\beta}})\right\}^{-1}.$$

If the variance structure is misspecified, there is loss of efficiency when using the model-based covariance matrix estimator. Thus, since $ASCOV_s(\widehat{\boldsymbol{\beta}})$ is a consistent covariance estimator, then

$$\mathcal{IMR}_{\gamma,n}(\widehat{\boldsymbol{\beta}}) = ASCOV_s(\widehat{\boldsymbol{\beta}}) \left\{ASCOV_m(\widehat{\boldsymbol{\beta}})\right\}^{-1} \geq \boldsymbol{I}_p.$$

In this case, the IDC characterizes the loss in relative efficiency under the candidate variance structure.

Selection of the optimal working correlation structure in GEE

In the context of GEE, let $\mathfrak{V}$ be a class of candidate working correlation matrices

$$\mathfrak{V} = \left\{R_\gamma, \gamma \in \Gamma\right\}.$$

Given a longitudinal data $\left\{(y_{ij}, \boldsymbol{x}_{ij}^T), j = 1, \cdots, n_i, i = 1, \cdots, K\right\}$, the GEE is an additive estimating function $\Psi_{\gamma,K}(\boldsymbol{\beta})$, given by (2.27)

$$\Psi_{\gamma,K}(\boldsymbol{\beta}) = \frac{1}{\sigma^2} \sum_{i=1}^{K} D_i^T(\boldsymbol{\beta}) \Sigma_{\gamma,i}^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right),$$

which is used to estimate the regression coefficients $\boldsymbol{\beta}$, where $\Sigma_{\gamma,i} = G_i^{1/2} R_\gamma G_i^{1/2}$. Given the estimating function $\Psi_{\gamma,K}(\boldsymbol{\beta})$, the aggregated sensitivity and variability matrices are given by

$$\mathbf{S}_{\Psi_{\gamma,K}}(\boldsymbol{\beta}) = -\frac{1}{\sigma^2} \sum_{i=1}^{K} D_i^T(\boldsymbol{\beta}) \Sigma_{\gamma,i}^{-1} D_i(\boldsymbol{\beta}),$$

and

$$\mathbf{V}_{\Psi_{\gamma,K}}(\boldsymbol{\beta}) = \frac{1}{\sigma^2} \sum_{i=1}^{K} D_i^T(\boldsymbol{\beta}) G_i^{-1/2} R_\gamma^{-1} R^* R_\gamma^{-1} G_i^{-1/2} D_i(\boldsymbol{\beta}),$$

where $Cov(\boldsymbol{Y}_i) = \sigma^2 G_i^{1/2} R^* G_i^{1/2}$, when the variance function is correctly specified. Given any value of $\boldsymbol{\beta}$, define two limiting information matrices by

$$\mathbf{S}_\gamma(\boldsymbol{\beta}) = \lim_{K\to\infty} \frac{1}{K} \mathbf{S}_{\Psi_{\gamma,K}}(\boldsymbol{\beta}),$$

and

$$\mathbf{V}_\gamma(\boldsymbol{\beta}) = \lim_{K\to\infty} \frac{1}{K} \mathbf{V}_{\Psi_{\gamma,K}}(\boldsymbol{\beta}).$$

Based on the data, the sensitivity and variability matrices are estimated by

$$\widehat{\mathbf{S}}_{\Psi_{\gamma,K}}(\boldsymbol{\beta}) = -\frac{1}{\sigma^2} \sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}_{\gamma,i}^{-1} \widehat{D}_i,$$

and

$$\widehat{\mathbf{V}}_{\Psi_{\gamma,K}}(\boldsymbol{\beta}) = \frac{1}{(\sigma^2)^2} \sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}_{\gamma,i}^{-1} \boldsymbol{r}_i \boldsymbol{r}_i^T \widehat{\Sigma}_{\gamma,i}^{-1} \widehat{D}_i$$

where the true value of σ^2 is known; otherwise, it is replaced by a moment estimator. Then, the estimated information matrix ratio is

$$\mathcal{IMR}_{\gamma,K}(\widehat{\boldsymbol{\beta}}) = \left\{-\widehat{\mathbf{S}}_{\gamma,K}(\widehat{\boldsymbol{\beta}})\right\}^{-1} \widehat{\mathbf{V}}_{\gamma,K}(\widehat{\boldsymbol{\beta}}),$$

and consequently, the IDC is given by

$$d(V_\gamma, V^*) = tr\left\{[\mathcal{IMR}_{\gamma,K}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{I}_p]^2\right\}.$$

The optimal working correlation is defined as the one with the minimum IDC, that is,

$$V_{opt} = \arg\min_{V_\gamma \in \mathfrak{V}} \left\{d(V_\gamma, V^*), \gamma \in \Gamma\right\}.$$

Remark 5.2 *The correlation information criterion (CIC) proposed by [43] is defined by*

$$CIC(R_\gamma) = tr\left\{\widehat{\Omega}_I \widehat{\mathbb{V}}_\gamma\right\},$$

where

$$\widehat{\Omega}_I = -\widehat{\boldsymbol{S}}_{I,K}(\widehat{\boldsymbol{\beta}}_{R_\gamma}, \widehat{\sigma}^2_{R_\gamma}) = \frac{1}{\widehat{\sigma}^2} \sum_{i=1}^{K} \widehat{D}_i^T \widehat{G}_i^{-1} \widehat{D}_i$$

evaluated at the estimators of $\boldsymbol{\beta}$ and σ^2 from the GEE with the candidate working correlation matrix R_γ, and

$$\begin{aligned}
\widehat{\mathbb{V}}_\gamma &= \left\{-\boldsymbol{S}_{\gamma,K}(\widehat{\boldsymbol{\beta}}_{R_\gamma})\right\}^{-1} \boldsymbol{V}_{\gamma,K}(\widehat{\boldsymbol{\beta}}_{R_\gamma}) \left\{-\boldsymbol{S}_{\gamma,K}(\widehat{\boldsymbol{\beta}}_{R_\gamma})\right\}^{-1} \\
&= \left(\sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}^{-1}_{\gamma,i} \widehat{D}_i\right)^{-1} \left(\sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}^{-1}_{\gamma,i} \boldsymbol{r}_i \boldsymbol{r}_i^T \widehat{\Sigma}^{-1}_{\gamma,i} \widehat{D}_i\right) \left(\sum_{i=1}^{K} \widehat{D}_i^T \widehat{\Sigma}^{-1}_{\gamma,i} \widehat{D}_i\right)^{-1}
\end{aligned}$$

evaluated at the estimator of $\boldsymbol{\beta}$ from the GEE with the candidate working correlation matrix R_γ.

The optimal working correlation structure is the one with the minimum CIC, that is,

$$R_{opt} = \arg\min_{R_\gamma \in \Re} \left\{CIC(R_\gamma), \gamma \in \Gamma\right\}.$$

Remark 5.3 *Both of these two model selection procedures can work for selecting the optimal covariance structure from a general class of candidate covariance structures. This class of candidates could be a closed collection of arbitrary covariance structures, or the combination of a class of variance structures and a class of correlation structures.*

5.2 Numerical Illustration

5.2.1 Selection of Variance Function

Through the following simulation studies, we will evaluate the two model selection procedures based on the IR tests and the IDC. A data set

$$\{(y_i; x_{i1}); i = 1, \cdots, n\}$$

is generated from a compound Poisson model through the following steps: for each $i = 1, \cdots, n$,

(i) generate x_{i1} from a uniform distribution $UNIF(0, 1)$;

(ii) let $\eta_i = \beta_0 + \beta_1 x_{i1}$, where the true values of the regression coefficients are $\beta_0 = 10, \beta_1 = 5$, and let $\mu_i = \exp(\eta_i)$;

(iii) let

$$\lambda_i = \frac{\mu_i^{2-\kappa_0}}{\sigma^2(2-\kappa_0)}, \quad 1 < \kappa_0 < 2,$$

where the true value of σ^2 is 1.5; and then generate a random number N_i from a Poisson distribution with mean λ_i;

(iv) generate N_i random numbers $\{z_{i1}, \cdots, z_{i,N_i}\}$ from a gamma distribution with the shape parameter $\alpha = \frac{2-\kappa_0}{\kappa_0-1}$, and the scale parameter $\zeta = \sigma^2(\kappa_0 - 1)\mu_i^{\kappa_0-1}$;

(v) let $y_i = \sum_{j=1}^{N_i} z_{i,j}$.

According to [47], the true mean and variance of the responses in the underlying distribution are given by

$$E(Y_i) = \mu_i = \exp\{\beta_0 + \beta_1 x_{i1}\}, \quad \text{and} \quad Var(Y_i) = \sigma^2 \mu_i^{\kappa_0},$$

for $i = 1, \cdots, n$.

The task of this simulation study is to select the optimal variance function from a collection of candidate variance functions

$$\mathcal{V} = \{V(\mu; \kappa) = \mu^\kappa, \kappa = 1, 1.2, 1.5, 1.8, 2\}.$$

For each candidate variance function, the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ are estimated from solving the quasi-score equation

$$\sum_{i=1}^{n} \boldsymbol{x}_i \mu_i \frac{y_i - \mu_i}{\sigma^2 \mu_i^\kappa} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \boldsymbol{x}_i \mu_i^{1-\kappa}(y_i - \mu_i) = 0,$$

where $\boldsymbol{x}_i = (1, x_{i1})^T$ and $\mu_i = \exp\{\beta_0 + \beta_1 x_{i1}\}$.

Given a value of κ_0, we generate 1000 replicates for each of the sample sizes $n = 20, 50, 100, 200$. Table 5.1 reports the empirical selection frequencies of each candidate variance function among 1000 replicates using the two model selection procedures based on the IR tests and the IDC, over different sample sizes and different values of κ_0. The numerical results suggest that

(i) both of these two model procedures have higher detection rate with larger sample size;

(ii) the detection rate of the model selection procedure based on the IR tests increases as κ_0 increases. Figure 5.1 illustrates the relative discrepancy between the true variance function and candidate variance structure

$$\frac{V^*(\mu)}{V_\gamma(\mu)} - 1 = \frac{\mu^{\kappa_0}}{\mu^{\kappa}} - 1$$

evaluated at smaller value $\mu = 10$ and larger value $\mu = 100$. The plots show that for larger value of κ_0, the relative discrepancy is larger, and consequently, the IR tests are more sensitive to detect the departure of the candidate variance function from the true one in this case. Correspondingly, the model selection procedure based on the IR tests is more powerful to detect the true variance function for large value of κ_0;

(iii) the IDC has higher detection rate than the model selection based on the IR tests for all the values of κ_0.

5.2.2 Selection of Working Correlation Structure

In this section, we will assess and compare the performance of the IDC, QIC and CIC in detecting the true working correlation structure among independence, exchangeable and AR(1), for continuous and discrete responses.

Table 5.1: The empirical frequencies of selecting each candidate variance function in $\mathfrak{V}$ among 1000 replicates using the two model selection procedures based on the IR tests and the IDC over different sample sizes and different true values of κ_0. The numbers in the parentheses are the ratios of the detection rate, obtained from the model selection procedure based on the IDC, to the detection rate, obtained from the model selection based on the IR tests, of the true variance function.

	$IR^s_{pool,u}$					$d(V_\gamma, V^*)$				
	1	1.2	1.5	1.8	2	1	1.2	1.5	1.8	2
						$\kappa_0 = 1.2$				
20	245	**157**	319	201	78	502	**233**(1.48)	179	40	46
50	234	**220**	169	264	113	391	**455**(2.07)	153	1	0
100	171	**310**	78	334	107	308	**591**(1.91)	101	0	0
200	139	**500**	138	222	1	164	**803**(1.61)	33	0	0
						$\kappa_0 = 1.5$				
20	113	184	**273**	262	168	172	207	**389**(1.42)	197	35
50	47	156	**329**	234	234	37	221	**616**(1.87)	125	1
100	16	108	**375**	407	94	9	125	**824**(2.20)	42	0
200	5	61	**350**	331	253	3	76	**886**(2.53)	35	0
						$\kappa_0 = 1.8$				
20	47	109	207	**293**	344	64	76	227	**334**(1.14)	299
50	4	47	231	**463**	255	5	15	159	**595**(1.29)	226
100	0	11	165	**444**	380	0	2	95	**756**(1.70)	147
200	0	1	130	**539**	330	0	0	31	**875**(1.62)	94

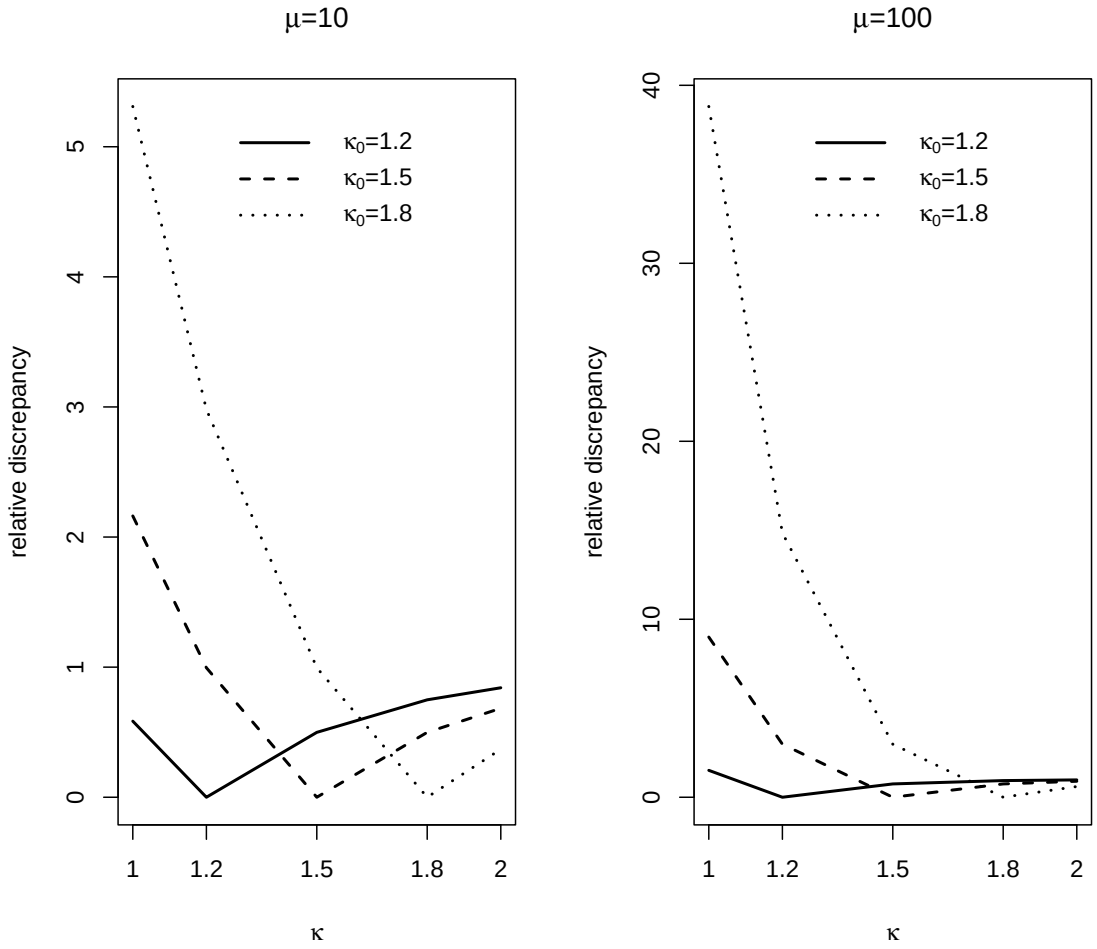

Figure 5.1: The relative discrepancy between two variance functions, measured by $|V^*(\mu)/V_\gamma(\mu) - 1| = |\mu^{\kappa_0}/\mu^{\kappa} - 1|$ evaluated at small and large values of μ. The solid line represents $\kappa_0 = 1.2$. The dashed line represents $\kappa_0 = 1.5$. The dotted line represents $\kappa_0 = 1.8$.

Simulation 5.1 We conduct a simulation study with 1000 replicates for each of the following settings:

Model 1: Gaussian response with the mean structure $\mu_{ij} = 3 + 5x_{ij}$. The true correlation structure is either exchangeable or AR(1). The true value of the dispersion parameter is 1. For this model, three studies will be investigated.

Study 1: For each replicate, we generate a set of balanced longitudinal data of size $K = 30$ with 5 repeated measurements taken for each subject. We will generate two types of covariate: time-dependent and time-independent but individual level. Time-dependent covariate x_{ij} is generated from a uniform distribution $UNIF(j, j + 1)$, and individual level covariate x_{ij}'s are independently generated from $UNIF(0, 1)$. The true value of the correlation parameter is 0.5.

Study 2: For each replicate, we generate a set of balanced longitudinal data of size $K = 30$ with different number of measurements 5 and 10 for each subject. The covariate x_{ij} is time-dependent, generated from a uniform distribution $UNIF(j, j + 1)$. The true value of the correlation parameter is 0.5.

Study 3: For each replicate, we generate a set of balanced longitudinal data with 5 observations for each subject. The covariate x_{ij} is time-dependent, generated from a uniform distribution $UNIF(j, j + 1)$. The sample size $K = 20, 100$, and the true value of the correlation parameter is $\rho = 0.1, 0.5, 0.9$.

Model 2: Binary response with the mean structure with logistic link

$$\mu_{ij} = \frac{\exp(-1 + 1/6x_{ij})}{1 + \exp(-1 + 1/6x_{ij})}.$$

The true correlation structure is either exchangeable or AR(1). The true value of the dispersion parameter is 1. In this setting, we generate 1000 replicates of balanced longitudinal data with sample size $K = 30$ and 5 observations for each subject. The covariate x_{ij} is time-dependent, generated from a uniform distribution $UNIF(j, j+1)$. Binary responses in Model 2 are generated using the BINDATA library in R, which requires some constraints on the correlation parameter. In the

case that the true correlation structure is exchangeable, the true value of the correlation parameter is set to be 0.2, 0.5; for AR(1), the true value of the correlation parameter is 0.2, 0.5, 0.7.

Table 5.2, 5.3, 5.4 and 5.5 report the empirical selection frequencies of each candidate working correlation structure: independence, exchangeable and AR(1) for the three studies of Model 1. These empirical frequencies are obtained from the criteria QIC, CIC and IDC using the true value, the Pearson moment estimator (4.17) and the transformed moment estimator (4.18) of the dispersion parameter σ^2. The numerical results have shown that:

(i) In general, the CIC has higher detection rate than the QIC, which agrees with the conclusions in [43]. Moreover, the performance of the IDC is better than both of the QIC and CIC, except in the case when the true correlation is AR(1) and the covariate is time-independent. The advantage of the IDC over the QIC or CIC is stronger when the covariate is time-dependent.

(ii) If the true value of the dispersion parameter is assumed unknown, it can estimated by the Pearson residuals or the transformed residuals. The IDC has a higher detection rate when when using the transformed moment estimator than using the Pearson moment estimator. However, when the true correlation structure is exchangeable, both of the QIC and CIC, when using the transformed moment estimator, lead to an incorrect conclusion, that is, AR(1) is selected as the optimal working correlation. When the true correlation structure is AR(1), both the QIC and CIC have a higher detection rate when using the Pearson moment estimator than when using the transformed moment estimator.

(iii) The QIC, CIC and IDC all have better performance as the sample size increases. However, more repeated measurements do not necessarily improve the performance of these three criteria.

Table 5.6 reports the empirical selection frequencies of each candidate work-

ing correlation structure: independence, exchangeable and AR(1) for the Model 2. For binary data, the dispersion parameter is usually assumed to be known as 1. Then the empirical frequencies are obtained from the QIC, CIC and IDC criteria using the true value 1. The numerical results are consistent with the results obtained from Model 1. The proposed criterion IDC performs better than the QIC and CIC.

5.3 Data Analysis

We will analyze two data sets to illustrate the two model selection methods of the most appropriate variance/covariance structure, proposed in Chapter 5.

5.3.1 Vehicle Insurance Data

This data set records one-year vehicle insurance policies taken out in the year 2004 or 2005, provided in [19]. Among 67856 policies, 4624 (6.8%) had at least one claim. There are two response variables: the number of claims, and the claim size. Several explanatory variables were also recorded, shown in the following Table 5.7. Given these variables, we define the covariates, shown in Table 5.8, which will be used in the analysis. Note that $x_{v.age1}$, $x_{v.age2}$, $x_{v.age4}$ are the dummy variables for the vehicle age categories, with category 3 as the baseline; x_{gender} is the dummy variable for the gender of policy holder, with male as the baseline; x_{age1}, x_{age2}, x_{age4}, x_{age5} and x_{age6} are the dummy variables for the age categories of policy holders, with the category 3 as the baseline; x_{areaA}, x_{areaB}, x_{ageD}, x_{ageE} and x_{ageF} are the dummy variables for the residence area, with area C as the baseline; x_{bus}, x_{convt}, x_{coupe}, x_{hback}, x_{hdtop}, x_{mcara}, x_{mibus}, x_{panvn}, x_{rdstr}, x_{stnwg}, x_{truck}, and x_{ute} are the dummy variables for the vehicle body type, with "sedan" as the baseline.

Table 5.2: The empirical frequencies of selecting each of the independence (IND), exchangeable(EXCH) and AR(1) correlation structures, among 1000 replicates using the QIC, CIC and IDC for the Study 1 of Model 1. QIC_*, CIC_* and IDC_* represent the QIC, CIC and IDC using the true value of σ^2. QIC_P, CIC_P and IDC_P represent the QIC, CIC and IDC using the Pearson moment estimator (4.17) of σ^2. QIC_{tr}, CIC_{tr} and IDC_{tr} represent the QIC, CIC and IDC using the transformed moment estimator (4.18) of σ^2. For each replicate, the data is of size $K = 30$, with 5 observations for each subject. The true correlation structure is either exchangeable or AR(1), and the true value of the correlation parameter is 0.5. Two types of covariate are generated: time-dependent and time-independent but individual-level.

	True correlation: Exchangeable with $\rho = 0.5$					
	x_{ij}: time-dependent			x_{ij}: individual-level		
	IND	EXCH	AR(1)	IND	EXCH	AR(1)
QIC_*	416	**519**	65	267	**638**	95
CIC_*	310	**546**	144	45	**795**	160
IDC_*	1	**742**	257	0	**758**	242
QIC_P	303	**549**	148	37	**796**	167
CIC_P	303	**549**	148	37	**796**	167
IDC_P	0	**713**	287	0	**839**	161
QIC_{tr}	0	**0**	1000	0	**0**	1000
CIC_{tr}	4	**3**	993	4	**11**	985
IDC_{tr}	0	**926**	74	0	**766**	234
	True correlation: AR(1) with $\rho = 0.5$					
	x_{ij}: time-dependent			x_{ij}: individual-level		
	IND	EXCH	AR(1)	IND	EXCH	AR(1)
QIC_*	208	147	**645**	169	210	**621**
CIC_*	127	99	**744**	36	122	**842**
IDC_*	13	146	**841**	24	524	**252**
QIC_P	125	97	**778**	32	115	**853**
CIC_P	125	97	**778**	32	115	**853**
IDC_P	0	82	**918**	0	553	**447**
QIC_{tr}	242	239	**519**	104	429	**467**
CIC_{tr}	138	106	**756**	42	105	**853**
IDC_{tr}	0	108	**892**	0	527	**473**

Table 5.3: The empirical frequencies of selecting each of the independence (IND), exchangeable(EXCH) and AR(1) correlation structures, among 1000 replicates using the QIC, CIC and IDC for the Study 2 of Model 1. QIC_*, QIC_P, QIC_{tr}, CIC_*, CIC_P, CIC_{tr}, IDC_*, IDC_P, and IDC_{tr} are the same as Table 5.2. For each replicate, the sample size is $K = 30$, and 5 or 10 repeated measurements were taken for each subject. The true correlation structure is either exchangeable or AR(1), and the true value of the correlation parameter is 0.5. The covariate is time-dependent.

	True correlation: Exchangeable with $\rho = 0.5$					
	5 repeated measurements			10 repeated measurements		
	IND	EXCH	AR(1)	IND	EXCH	AR(1)
QIC_*	391	**538**	71	466	**517**	17
CIC_*	280	**585**	135	411	**538**	51
IDC_*	1	**747**	252	0	**695**	305
QIC_P	272	**589**	139	410	**538**	52
CIC_P	272	**589**	139	410	**538**	52
IDC_P	0	**749**	251	0	**697**	303
QIC_{tr}	0	**0**	1000	0	**0**	1000
CIC_{tr}	2	**3**	995	1	**0**	999
IDC_{tr}	0	**929**	71	0	**995**	5
	True correlation: AR(1) with $\rho = 0.5$					
	5 repeated measurements			10 repeated measurements		
	IND	EXCH	AR(1)	IND	EXCH	AR(1)
QIC_*	200	158	**642**	172	160	**668**
CIC_*	111	112	**777**	87	91	**822**
IDC_*	15	111	**874**	0	43	**957**
QIC_P	110	108	**782**	86	89	**825**
CIC_P	110	108	**782**	86	89	**825**
IDC_P	0	72	**928**	0	26	**974**
QIC_{tr}	244	264	**492**	272	251	**477**
CIC_{tr}	124	128	**748**	102	110	**788**
IDC_{tr}	0	84	**916**	0	28	**972**

Table 5.4: The empirical frequencies of selecting each of the independence (IND), exchangeable(EXCH) and AR(1) correlation structures, among 1000 replicates using the QIC, CIC and IDC for the Study 3 of Model 1. QIC_*, QIC_P, QIC_{tr}, CIC_*, CIC_P, CIC_{tr}, IDC_*, IDC_P, and IDC_{tr} are the same as Table 5.2. For each replicate, 5 repeated measurements were taken for each subject. The sample size is 20 or 100. The true correlation structure is exchangeable, and the true value of the correlation parameter is 0.1, 0.5 or 0.9. The covariate is time-dependent.

	$\rho = 0.1$			$\rho = 0.5$			$\rho = 0.9$		
	IND	EXCH	AR(1)	IND	EXCH	AR(1)	IND	EXCH	AR(1)
	sample size $K = 20$								
QIC_*	260	**329**	411	403	**477**	120	273	**505**	222
CIC_*	229	**311**	460	297	**492**	211	113	**569**	318
IDC_*	92	**616**	292	9	**674**	317	3	**566**	431
QIC_P	226	**312**	462	287	**485**	228	102	**572**	326
CIC_P	226	**312**	462	287	**485**	228	102	**572**	326
IDC_P	32	**702**	266	0	**666**	334	0	**589**	411
QIC_{tr}	141	**177**	682	2	**1**	997	0	**0**	1000
CIC_{tr}	193	**261**	546	23	**16**	961	0	**0**	1000
IDC_{tr}	32	**701**	267	0	**846**	154	0	**955**	45
	sample size $K = 100$								
QIC_*	376	**408**	216	342	**656**	2	206	**752**	42
CIC_*	329	**403**	268	189	**798**	13	0	**895**	105
IDC_*	14	**851**	135	0	**870**	130	0	**562**	438
QIC_P	329	**403**	268	186	**799**	15	0	**894**	106
CIC_P	329	**403**	268	186	**799**	15	0	**894**	106
IDC_P	0	**862**	138	0	**877**	113	0	**574**	426
QIC_{tr}	26	**58**	916	0	**0**	1000	0	**0**	1000
CIC_{tr}	249	**315**	436	0	**0**	1000	0	**0**	1000
IDC_{tr}	0	**859**	141	0	**991**	9	0	**1000**	0

Table 5.5: The empirical frequencies of selecting each of the independence (IND), exchangeable(EXCH) and AR(1) correlation structures, among 1000 replicates using the QIC, CIC and IDC for the Study 3 of Model 1. QIC_*, QIC_P, QIC_{tr}, CIC_*, CIC_P, CIC_{tr}, IDC_*, IDC_P, and IDC_{tr} are the same as Table 5.2. For each replicate, 5 repeated measurements were taken for each subject. The sample size is 20 or 100. The true correlation structure is AR(1), and the true value of the correlation parameter is 0.1, 0.5 or 0.9. The covariate is time-dependent.

	$\rho = 0.1$			$\rho = 0.5$			$\rho = 0.9$		
	IND	EXCH	AR(1)	IND	EXCH	AR(1)	IND	EXCH	AR(1)
	sample size $K = 20$								
QIC_*	211	282	**507**	199	168	**633**	198	215	**587**
CIC_*	191	273	**536**	112	129	**759**	130	172	**698**
IDC_*	102	527	**371**	42	201	**757**	1	101	**898**
QIC_P	189	273	**538**	108	127	**765**	124	166	**710**
CIC_P	189	273	**538**	108	127	**765**	124	166	**710**
IDC_P	50	577	**373**	1	150	**849**	0	11	**989**
QIC_{tr}	230	328	**442**	258	218	**524**	176	298	**526**
CIC_{tr}	184	265	**551**	123	124	**753**	128	220	**652**
IDC_{tr}	50	577	**373**	1	165	**834**	0	23	**977**
	sample size $K = 100$								
QIC_*	206	243	**551**	142	103	**755**	110	191	**699**
CIC_*	170	215	**615**	35	29	**936**	17	97	**886**
IDC_*	81	361	**558**	0	18	**982**	0	0	**1000**
QIC_P	170	215	**615**	33	29	**938**	17	97	**886**
CIC_P	170	215	**615**	33	29	**938**	17	97	**886**
IDC_P	18	392	**590**	0	3	**997**	65	394	**541**
QIC_{tr}	271	277	**452**	222	277	**501**	32	184	**784**
CIC_{tr}	166	215	**619**	42	39	**929**	0	0	**1000**
IDC_{tr}	18	391	**591**	0	8	**992**	0	0	**1000**

Table 5.6: The empirical frequencies of selecting each of the independence (IND), exchangeable(EXCH) and AR(1) correlation structures, among 1000 replicates using the QIC, CIC and IDC for the Model 2. *QIC*, *CIC* and *IDC* represent the QIC, CIC and IDC using the true value of σ^2. For each replicate, the sample size is $K = 30$, and 5 repeated measurements were taken for each subject. The true correlation structure is either exchangeable with the correlation 0.2 and 0.5, or AR(1) with the correlation 0.2, 0.5 and 0.7. The covariate is time-dependent.

	True: Exchangeable								
	$\rho = 0.2$			$\rho = 0.5$					
	IND	EXCH	AR(1)	IND	EXCH	AR(1)			
QIC	231	**487**	282	212	**658**	130			
CIC	234	**539**	227	184	**756**	60			
IDC	8	**863**	129	0	**712**	288			
	True: AR(1)								
	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.7$		
	IND	EXCH	AR(1)	IND	EXCH	AR(1)	IND	EXCH	AR(1)
QIC	130	207	**663**	76	108	**816**	50	98	**852**
CIC	97	167	**736**	30	74	**896**	14	42	**944**
IDC	12	345	**643**	0	81	**919**	0	19	**981**

We are interested in investigating significant covariate effects on the claim size. The observations with nonzero claim sizes are used in the analysis, where the response variable is the claim size, and the others are covariates. We fit the GLMs, which include only the main effects, with the log link function

$$\begin{aligned} log(\mu_i) &= \beta_0 + \beta_{value}x_{i,value} + \beta_{exp}x_{i,exposure} + \beta_{gender}x_{i,gender} \\ &+\beta_{v.age1}x_{i,v.age1} + \beta_{v.age2}x_{i,v.age2} + \beta_{v.age4}x_{i,v.age4} \\ &+\beta_{age1}x_{i,age1} + \beta_{age2}x_{i,age2} + \beta_{age4}x_{i,age4} + \beta_{age5}x_{i,age5} + \beta_{age6}x_{i,age6} \\ &+\beta_{areaA}x_{i,areaA} + \beta_{areaB}x_{i,areaB} + \beta_{areaD}x_{i,areaD} + \beta_{areaE}x_{i,areaE} \\ &+\beta_{areaF}x_{i,areaF} + \beta_{bus}x_{i,bus} + \beta_{convt}x_{i,convt} + \beta_{coupe}x_{i,coupe} + \beta_{hback}x_{i,hback} \\ &+\beta_{hdtop}x_{i,hdtop} + \beta_{mcara}x_{i,mcara} + \beta_{mibus}x_{i,mibus} + \beta_{panvn}x_{i,panvn} \\ &+\beta_{rdstr}x_{i,rdstr} + \beta_{stnwg}x_{i,stnwg} + \beta_{truck}x_{i,truck} + \beta_{ute}x_{i,ute} \end{aligned} \tag{5.6}$$

under different working unit variance functions

$$V(\mu_i, \kappa) = \mu_i^{\kappa}, \quad \kappa = 1.2, 1.5, 1.8, 2, 3, \tag{5.7}$$

where $\mu_i = E(Y_i)$, for $i = 1, \cdots, n$. The models with $\kappa = 1.2, 1.5, 1.8$ correspond to Tweedie's distributions ([49]), the one with $\kappa = 2$ corresponds to a gamma regression model, and the one with $\kappa = 3$ corresponds to an inverse gaussian regression model. The parameter estimates as well as the corresponding sandwich standard errors obtained from these five models were reported in Table 5.9. This analysis found strong evidence that exposure, gender, age category of policy holder, and residence area have significant effects on the claim sizes under all the variance functions. Specifically, larger degree of exposure is significant in reducing the claim sizes. Compared to men, women are likely to have smaller claim sizes. In addition, younger policy holders tend to have larger claim sizes. Compared to the area C, the claim sizes in area F are shown to be larger. However, the point estimates, as well as the standard errors, differ remarkably under different working variance functions. In particular, under the variance functions $V(\mu) = \mu^{\kappa}$ for $\kappa = 1.8, 2, 3$, the vehicle body type is shown to have a significant

effect on the claim sizes, and specifically, compared to "sedan", the "motorized caravan/combi" vehicles tend to have insurance claims with significantly smaller sizes. However, under the variance function $V(\mu) = \mu^{\kappa}$ with $\kappa = 1.2, 1.5$, the vehicle body type does not show as a significant factor.

For each variance function, we evaluate and compare the model selection criteria: IDC, p-values from the IR test using the pooled IR statistic IR_{pool}, AIC and BIC. The values of these criteria are listed in Table 5.10. AIC and BIC, the most popular model selection methods, select the variance function $V(\mu) = \mu^3$, which corresponds to an inverse gaussian regression model, as the optimal variance function. However, IDC selects $V(\mu) = \mu^2$ as the most appropriate variance function, and p-values obtained from the IR statistic IR_{pool} select $V(\mu) = \mu^{1.8}$.

Given the same mean structure, the penalty terms in the AIC and BIC are the same under different variance functions. The quasi-likelihood function is given by $Q(\mu, y) = \mu^{-\kappa}\left(\frac{\mu y}{1-\kappa} - \frac{\mu^2}{2-\kappa}\right)$ for Tweedie's distribution with the variance function $V(\mu) = \mu^{\kappa}$, $1 < \kappa < 2$, and $Q(\mu, y) = -y/\mu - log(\mu)$ for the gamma regression model with the variance function $V(\mu) = \mu^2$, $Q(\mu, y) = -y/2\mu^2 + 1/\mu$ for the inverse gaussian regression model with the variance function $V(\mu) = \mu^3$. Even though the models with different variance functions produce different point estimates to a certain extent, the values of $\widehat{\mu_i}$ are quite similar. The differences in the magnitude of AIC and BIC mainly depend on the value of κ especially for large values of μ and y. We found that both the AIC and BIC monotonically decrease when κ gets larger. Therefore, neither of these two criteria is appropriate to choose the optimal variance function because they tend to select larger values of κ. On the other hand, the IDC and p-value obtained from the IR test did not show any monotonic pattern associated with the value of κ. Even though different variance functions were chosen by the model selection methods which are based on the IDC and the p-values of the IR test ($\kappa = 2$ for the IDC and $\kappa = 1.8$ for the p-values), they lead to similar point estimates as well as standard errors, and consequently, the same statistical conclusion. Therefore, either the variance function $V(\mu) = \mu^{1.8}$ or $V(\mu) = \mu^2$ can be regarded as the selected best-fitting

Table 5.7: The variables given in vehicle insurance data set.

Variables	Type	Range
Age category of policy holder	categorical	1(youngest),2,3,4,5,6
Gender	categorical	male,female
Area of residence	categorical	A,B,C,D,E,F
Vehicle value	numerical	$0 - $350,000
Vehicle age	categorical	1(new),2,3,4
vehicle body type	categorical	bus, convertible,coupe,hatchback, hardtop,motorized caravan/combi, minibus, panel van, roaster, sedan,station wagon truck, utility
exposure	numerical	(0,1)
the number of claims	categorical	1,2,3,4
claim size	numerical	$0 - $55922.13

variance function.

5.3.2 Madras Longitudinal Schizophrenia Study

This data set was used as an example in [23]. Schizophrenia is a psychiatric disorder with symptoms of thought disorders. This study tracked positive and negative psychiatric symptoms over the first year after initial hospitalization for schizophrenia. To investigate the time pattern of the symptoms as well as age and gender effects, we fit the GEE models. The presence of thought disorder, which is a binary variable, is used as the response variable in the analysis. In addition, three variables are used as the covariates:

(i) x_{month}: duration since hospitalization (in months)

(ii) x_{age}: age category of patient at the onset of symptom, defined as

$$x_{age} = \begin{cases} 1, & \text{if age} < 20; \\ 0, & \text{otherwise.} \end{cases}$$

Table 5.8: The covariates used in the model fitting.

covariates	description
x_{value}	vehicle value: \$0 - \$350,000
$x_{v.age1}$	= 1, if the vehicle age category is 1; = 0, otherwise
$x_{v.age2}$	= 1, if the vehicle age category is 2; = 0, otherwise
$x_{v.age4}$	= 1, if the vehicle age category is 4; = 0, otherwise
$x_{exposure}$	exposure: (0,1)
x_{gender}	gender: = 1, if female; = 0, otherwise
x_{age1}	= 1, if the holder age category is 1; = 0, otherwise
x_{age2}	= 1, if the holder age category is 2; = 0, otherwise
x_{age4}	= 1, if the holder age category is 4; = 0, otherwise
x_{age5}	= 1, if the holder age category is 5; = 0, otherwise
x_{age6}	= 1, if the holder age category is 6; = 0, otherwise
x_{areaA}	= 1, if the residence area is A; = 0, otherwise
x_{areaB}	= 1, if the residence area is B; = 0, otherwise
x_{areaD}	= 1, if the residence area is D; = 0, otherwise
x_{areaE}	= 1, if the residence area is E; = 0, otherwise
x_{areaF}	= 1, if the residence area is F; = 0, otherwise
x_{bus}	= 1, if the vehicle body type is bus; = 0, otherwise
x_{convt}	= 1, if the vehicle body type is convertible; = 0, otherwise
x_{coupe}	= 1, if the vehicle body type is coupe; = 0, otherwise
x_{hback}	= 1, if the vehicle body type is hatchback; = 0, otherwise
x_{hdtop}	= 1, if the vehicle body type is hardtop; = 0, otherwise
x_{mcara}	= 1, if the vehicle body type is motorized caravan/combi; = 0, otherwise
x_{mibus}	= 1, if the vehicle body type is minibus; = 0, otherwise
x_{panvn}	= 1, if the vehicle body type is panel van; = 0, otherwise
x_{rdstr}	= 1, if the vehicle body type is roster; = 0, otherwise
x_{stnwg}	= 1, if the vehicle body type is station wagon; = 0, otherwise
x_{truck}	= 1, if the vehicle body type is truck; = 0, otherwise
x_{ute}	= 1, if the vehicle body type is utility; = 0, otherwise

Table 5.9: Parameter estimates of regression coefficients with the sandwich standard errors (in the parentheses) obtained from the GLMs, under different working variance functions, for the vehicle insurance data with at least one claim. The $*$ represents rejection of the null hypothesis $H_0 : \beta = 0$ using the Wald test at the significance level 0.05.

	$\kappa = 1.2$	$\kappa = 1.5$	$\kappa = 1.8$	$\kappa = 2$	$\kappa = 3$
β_0	8.027(0.108)	8.011(0.109)$*$	7.996(0.110)$*$	7.986(0.110)$*$	7.942(0.113)$*$
β_{value}	0.037(0.032)	0.035(0.032)	0.034(0.033)	0.033(0.033)	0.032(0.033)
β_{exp}	-0.826(0.089)$*$	-0.794(0.090)$*$	-0.766(0.090)$*$	-0.749(0.091)$*$	-0.681(0.094)$*$
β_{gender}	-0.173(0.050)$*$	-0.162(0.050)$*$	-0.152(0.050)$*$	-0.147(0.050)$*$	-0.127(0.050)$*$
$\beta_{v.age1}$	-0.104(0.082)	-0.105(0.081)	-0.106(0.081)	-0.106(0.081)	-0.108(0.079)
$\beta_{v.age2}$	-0.031(0.068)	-0.035(0.067)	-0.038(0.067)	-0.040(0.067)	-0.040(0.066)
$\beta_{v.age4}$	0.078(0.068)	0.078(0.068)	0.078(0.069)	0.078(0.069)	0.083(0.070)
β_{age1}	0.261(0.083)$*$	0.249(0.085)$*$	0.239(0.086)$*$	0.234(0.088)$*$	0.213(0.093)$*$
β_{age2}	0.093(0.072)	0.093(0.072)	0.092(0.072)	0.092(0.072)	0.093(0.072)
β_{age4}	0.024(0.070)	0.026(0.069)	0.029(0.069)	0.031(0.069)	0.043(0.068)
β_{age5}	-0.078(0.086)	-0.08(0.084)	-0.081(0.082)	-0.082(0.081)	-0.093(0.076)
β_{age6}	0.029(0.101)	0.021(0.100)	0.013(0.099)	0.008(0.098)	-0.009(0.094)
β_{areaA}	-0.049(0.066)	-0.048(0.066)	-0.046(0.065)	-0.045(0.065)	-0.043(0.065)
β_{areaB}	-0.094(0.069)	-0.094(0.068)	-0.095(0.068)	-0.097(0.067)	-0.111(0.066)
β_{areaD}	-0.092(0.088)	-0.097(0.087)	-0.103(0.085)	-0.108(0.085)	-0.136(0.081)
β_{areaE}	0.075(0.091)	0.072(0.092)	0.071(0.093)	0.070(0.094)	0.070(0.098)
β_{areaF}	0.328(0.099)$*$	0.315(0.103)$*$	0.303(0.106)$*$	0.296(0.109)$*$	0.270(0.122)$*$
β_{bus}	-0.362(0.609)	-0.351(0.580)	-0.343(0.555)	-0.339(0.539)	-0.330(0.471)
β_{convt}	0.136(0.900)	0.105(0.918)	0.071(0.931)	0.048(0.937)	-0.077(0.917)
β_{coupe}	0.300(0.181)	0.290(0.188)	0.281(0.196)	0.275(0.201)	0.249(0.227)
β_{hback}	0.125(0.065)	0.125(0.064)	0.124(0.064)	0.123(0.064)	0.119(0.063)
β_{hdtop}	0.107(0.146)	0.095(0.148)	0.087(0.149)	0.082(0.150)	0.067(0.152)
β_{mcara}	-1.045(0.638)	-1.023(0.550)	-1.004(0.478)$*$	-0.992(0.436)$*$	-0.952(0.286)$*$
β_{mibus}	0.149(0.206)	0.139(0.208)	0.130(0.209)	0.125(0.209)	0.108(0.213)
β_{panvn}	-0.831(1.765)	-0.857(1.502)	-0.884(1.275)	-0.902(1.140)	-1.010(0.626)
β_{rdstr}	-0.023(0.077)	-0.023(0.077)	-0.023(0.076)	-0.023(0.076)	-0.023(0.074)
β_{stnwg}	0.245(0.144)	0.250(0.149)	0.256(0.154)	0.259(0.157)	0.283(0.178)
β_{truck}	0.083(0.110)	0.088(0.111)	0.092(0.112)	0.094(0.113)	0.110(0.117)
β_{ute}	0.359(0.227)	0.365(0.235)	0.375(0.245)	0.383(0.252)	0.445(0.307)

Table 5.10: The IDC, p-values of the IR test (p_{IR}), AIC and BIC obtained from the GLM models with different working variance functions.

	$\kappa = 1.2$	$\kappa = 1.5$	$\kappa = 1.8$	$\kappa = 2$	$\kappa = 3$
IDC	9.483	7.181	6.232	**6.012**$*$	7.437
p_{IR}	0.042	0.526	**0.879**$*$	0.679	0.554
AIC	2.523E07	1.642E06	2.630E05	7.927E04	**55.521**$*$
BIC	2.523E07	1.642E06	2.631E05	7.946E04	**242.252**$*$

(iii) x_{gender}: gender of patient, defined as

$$x_{gender} = \begin{cases} 1, & \text{if female;} \\ 0, & \text{if male.} \end{cases}$$

Besides these three main effects, the interaction between variables x_{month} and x_{age}, and the interaction between variables x_{month} and x_{gender} are also included in the models.

We fit three GEE models with the same logistic link

$$\begin{aligned} logit(\mu_{ij}) &= \beta_0 + \beta_{month} x_{month,ij} + \beta_{age} x_{age,ij} + \beta_{gender} x_{gender,ij} \\ &\quad + \beta_{m.a} x_{month,ij} * x_{age,ij} + \beta_{m.g} x_{month,ij} * x_{gender}, \end{aligned}$$

where $\mu_{ij} = E(Y_{ij})$, under three different working correlation structures: independence, exchangeable, and AR(1). The results are reported in Table 5.11. The point estimates of the regression coefficients as well as the sandwich standard errors vary substantially over these three working correlation structures. In addition, the model with AR(1) working correlation structure leads to the smallest standard errors for all the coefficient estimates. Moreover, the baseline log odds ratio of the presence of symptom is significantly larger than zero for the patients who were male and over 20 years old at the onset of symptom, under the "independence" and "exchangeable" working correlation structures, but it is not significantly different from 0 under the AR(1) working correlation structure.

For each model, we calculate and compare the IDC, p-value obtained from the IR statistic IR_{pool} as well as the QIC and CIC, listed in Table 5.12. The results show that the IDC, p-value of the IR_{pool} and CIC all select AR(1) as the optimal working correlation structure, which is consistent with the result in [43]. In addition, the QIC selects the independence working correlation structure, also the same as their result. Note that GEE fitting is performed here using the *geepack* library in R, but [43] used the *yags* library. Thus, the values of the QIC and CIC here are slightly different from those in [43].

It is interesting to see that the differences in the magnitudes of the QIC and CIC are substantially smaller than those of the IDC. In other words, the IDC magnifies the discrepancy among different working correlation structures. In addition, the p-values obtained from the IR_{pool} are also sensitive to the difference among different correlation structures.

Remark 5.4 In this data analysis, we did not check the adequacy of the mean structure, in which other interactions (for example, three way interaction of age, gender and month) may be also a significant factor. It may require further tests to verify the mean structure, for example, using the QIF proposed by [68]. However, in order to compare our results with those in [43], we keep the same mean structure used in their paper [43].

Among 186 subjects in this data set, 17 have missing observations, only partial follow-up that ranges from 1 to 11 months. Regression analysis of drop-out suggests that subjects whose current disease status is $Y_{ij} = 1$ are at increased risk to drop-out at time $j+1$ with odds ratio 1.76 and p-value 0.345. The potential association between drop-out and the observed outcome data implies consideration of an analysis that is valid if the drop-out mechanism is missing at random. See [23]. However, GEE is valid only if the missing mechanism is missing completely at random. When missing data are missing at random, GEE is unable to produce consistent estimates of the mean parameters because its estimating equations are not unbiased. [71] proposed inverse probability weighted GEE

Table 5.11: Parameter estimates of regression coefficients with the sandwich standard errors (in the parentheses) obtained from the GEE models under different working correlation structures, for the Madras Longitudinal Schizophrenia data. The $*$ represents rejection of the null hypothesis $H_0 : \beta = 0$ using the Wald test at the significance level 0.05.

coefficients	independence	exchangeable	AR(1)
β_0	0.643(0.304)$*$	0.620(0.315)$*$	0.542(0.292)
β_{month}	-0.254(0.060)$*$	-0.272(0.065)$*$	-0.233(0.055)$*$
β_{age}	0.811(0.493)	1.059(0.547)	0.619(0.459)
β_{gender}	-0.388(0.449)	-0.593(0.525)	-0.130(0.420)
$\beta_{m.a}$	-0.137(0.094)	-0.087(0.093)	-0.096(0.084)
$\beta_{m.g}$	-0.113(0.096)	-0.140(0.097)	-0.157(0.088)

Table 5.12: The IDC, QIC, CIC and p-values of the IR test (p_{IR}) obtained from the GEE models under different working correlation structures.

	independence	exchangeable	AR(1)
IDC	35.943	20.706	**0.317**
QIC	**955.322**	964.745	955.965
CIC	19.011	19.005	**18.540**
p_{IR}	0.000	0.000	**0.398**

which yields unbiased estimating equations, and hence produce consistent parameter estimates. However, for the purpose of comparison with [43], we do not consider the issue of missing data.

Chapter 6

Summary and Future Work

6.1 Summary

As an important example of estimating equations, quasi-likelihood inference is widely used to estimate parameters of interest in various statistical problems where the investigators are uncertain about the complete probabilistic mechanism by which the data are generated. Based on the assumptions on certain aspects of the underlying probability distribution, typically on the first two moments, quasi-score equations for independent data or GEEs for correlated data can be constructed. These estimating equations can provide consistent estimators of the regression coefficients, and can obtain the same estimation efficiency as the most efficient estimator if the mean and variance/covariance structures are correctly specified. Thus, it is of importance to assess the adequacy of the assumptions on the first two moments. Numerous tests have been suggested in the literature to test for misspecification of the mean structure as well as to select covariates that have significant effects on responses. But, so far in the literature there have been no systematic methods available to assess the validity of the variance/covariance assumption.

In this thesis, we focus on the circumstances where the mean structure is correctly specified. It shows that misspecified variance/covariance structures lead

to some discrepancy between the two forms of information matrix, the negative sensitivity matrix and the variability matrix. This discrepancy is equivalent to that between the model-based and sandwich covariance matrix estimators. Contrasting the two types of covariance matrix estimators, we construct the information ratio (IR) statistics that enable us to test for misspecification of variance/covariance structures, as well as to select the optimal variance/covariance structure. Also, we propose the "information discrepancy criterion" (IDC) for selecting the optimal variance/covariance structure, which gives a better performance than the model selection procedure based on the IR test statistics. The IDC essentially measures the loss in relative estimation efficiency in using a candidate variance/covariance structure compared to the true one.

For the IR tests, we have derived related asymptotic distributions, and carried out intensive simulation experiments for a test for heteroscedasticity, a test for overdispersion, and a test for misspecified variance functions and/or working correlation structure in GEE. The numerical results have shown that the proposed IR statistics give adequate performance under the null hypothesis. When the statistics have poor performance due to a heavy right tail for small sample size, a normalized χ^2_ν approximation can make an improvement. The IR statistics are considerably powerful to reject the null hypothesis, and more powerful than the classic information matrix test proposed by [87]. Furthermore, the performance of the IR tests is very consistent among different scenarios of alternative hypotheses, because there is no need to model alternative hypotheses. The pooled IR statistics usually perform the best, because their weights incorporate the overall influence from all the covariates. Moreover, in linear regression models, the IR statistics corresponding to individual regression coefficients can provide a powerful tool to detect responsible variables for heteroscedasticity.

We proposed two model selection procedures. One is based on a sequence of IR tests. The simulation studies have shown that in the context of GEE, the IR tests based criterion is more useful to select the true/optimal variance function than to select the true/optimal correlation structure. The other selection criterion

is an "information discrepancy criterion" (IDC), that measures the information loss (difference) of a candidate variance/covariance structure relative to the true one. In the two simulation studies considered, the IDC has a high detection rate of the true variance function in compound Poisson models, and the true correlation structure in GEE. Moreover, the IDC has better performance than the criterion QIC proposed by [65] and the CIC proposed by [43] for selecting the optimal correlation structure, especially for time-dependent covariates.

6.2 Future Work

Both of the IR tests and model selection approach based on the IDC have been proposed under the assumption that the first moment of the underlying distribution is correctly specified. However, in some applications, the assumption of the mean structure may be approximately correct, but with mild departure from the true structure, for example, excluding an uninformative covariate, or misspecification of the link function. It is necessary to conduct further investigation on the robustness of the IR tests and the IDC against minor misspecification of the mean structure.

In the context of GEE, the regression coefficients are regarded as the parameter of main interest, and both the dispersion parameter and the correlation parameters are treated as nuisance parameters. They are called GEE-1 in the literature. [67] formalized the estimation of the parameters related to the variance/covariance structure, which leads to simultaneous inferences about the regression and association parameters. See also [54]. They are referred to as GEE-2 in [41]. In GEE-2, the asymptotic covariance matrix of the parameter estimators is available in an expanded form involving mean, dispersion and correlation parameters. It would be straightforward to extend the proposed IR statistics to the GEE-2 setting where we can test for any postulated variance/covariance structures.

Many selection information criteria describe the tradeoff between the complexity and precision of the model. Among competing correlation structures in GEE, exchangeable and AR(1) correlation structures both have only one correlation parameter, but unstructured correlation matrices have $(n^2 - n)/2$ correlation parameters, where n is the number of repeated observations for a subject. It is certain that an unstructured correlation structure provides a more complex and flexible model of correlation, but on the other hand, it may cause overfitting that affects the estimation efficiency. In the proposed criterion IDC, as well as the CIC, no penalty on the number of correlation parameters estimated has been accounted for. It is unfair to directly compare correlation structures with different numbers of correlation parameters. A future work is to make a modification to the IDC by penalizing the complexity of competing correlation structures.

GEE has been regarded as the most popular estimation method in the marginal model for longitudinal data analysis. GEE may run into some difficulty occasionally. For example, the estimates of the correlation parameter do not exist in some cases of misspecification ([18]). [68] proposed a method of quadratic inference functions (QIF) which do not involve direct estimation of the correlation parameter, and that remains the optimal even if the working correlation structure is misspecified. They suggested that the inverse of the working correlation matrix can be represented by a linear combination of basis matrices. Moreover, this approach provides a χ^2 inference function for testing nested models and a χ^2 test for the mean structure misspecification. However, like the GEE, QIF does not provide a test for misspecified covariance structure. Tests for the second-moment misspecification will help to improve the estimation efficiency of the QIF. Thus, in the context of QIF, it is possible to establish procedures of testing for both mean and covariance misspecification. In addition, the complexity of the candidate correlation structures can be characterized by the minimum number of the basis matrices in the decomposition of the inverse of the working correlation matrix. Then, a possible penalty term for the IDC may be related to the number of the basis matrices. Some additional exploration is needed.

APPENDICES

A.1. Proof of Lemma 4.1 and 4.2

In the context of GLM, since the Pearson residual vector can be approximated by (4.4)

$$\boldsymbol{r}_{_P} \simeq (\boldsymbol{I}_n - H_*)\, \boldsymbol{\epsilon}_{_P},$$

the β_{j-1}-specific Godambian estimator of the dispersion parameter σ^2, for $j = 1, \cdots, p$, can be approximated by

$$\widetilde{\sigma}^2_{j-1} \simeq \boldsymbol{\epsilon}_{_P}^T\, (\boldsymbol{I}_n - H_*)\, \boldsymbol{W}_*^{(j-1)}\, (\boldsymbol{I}_n - H_*)\, \boldsymbol{\epsilon}_{_P},$$

following from the quadratic form (4.5) of the Godambian estimator. Here we use the fact that $\widehat{w}_i^{(j-1)}$ are consistent estimates of $w_{i,*}^{(j-1)}$ due to the consistency of $\widehat{\boldsymbol{\beta}}_n$.

Under the null hypothesis H_0 (4.1), the vector $\boldsymbol{\epsilon}_{_P}$ has mean $\mathbf{0}$ and covariance matrix $\sigma^2 \boldsymbol{I}_n$. Consider a general quadratic form $\boldsymbol{Q}(\boldsymbol{\epsilon}_{_P}) = \boldsymbol{\epsilon}_{_P}^T \mathcal{A} \boldsymbol{\epsilon}_{_P}$, where $\mathcal{A}$ is an $n \times n$ symmetric matrix. By the eigen-decomposition, the matrix $\mathcal{A}$ can be decomposed as

$$\mathcal{A} = \mathcal{E} \Lambda \mathcal{E}^T,$$

where $\mathcal{E}$ is an orthogonal matrix and

$$\Lambda = \mathrm{diag}\,\{\lambda_1, \cdots, \lambda_n\},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the eigenvalues of the matrix $\mathcal{A}$. Let $\boldsymbol{\varepsilon} = \mathcal{E}^T \boldsymbol{\epsilon}_p = (\varepsilon_1, \cdots, \varepsilon_n)^T$ be an $n \times 1$ vector, which also has mean $\mathbf{0}$ and covariance matrix $\sigma^2 \boldsymbol{I}_n$, because $\mathcal{E}$ is an orthogonal matrix. The quadratic form $Q(\boldsymbol{\epsilon}_p)$ can be written as

$$Q(\boldsymbol{\epsilon}_p) = \boldsymbol{\varepsilon}^T \Lambda \boldsymbol{\varepsilon} = \sum_{k=1}^{n} \lambda_k \varepsilon_k^2.$$

Then, the expectation of the quadratic form $Q(\boldsymbol{\epsilon}_p)$ is given by

$$E\left(Q(\boldsymbol{\epsilon}_p)\right) = \sigma^2 \left(\sum_{k=1}^{n} \lambda_k \right).$$

Therefore, the β_{j-1}-specific Godambian estimator (4.5) has expectation

$$E\left(\widetilde{\sigma}_{j-1}^2\right) \simeq \sigma^2 \left(\sum_{k=1}^{n} \lambda_{k,*}^{(j-1)} \right), \quad j = 1, \cdots, p,$$

where $\lambda_{k,*}^{(j-1)}$ are the eigenvalues of the matrix $(\boldsymbol{I}_n - H_*)\, \boldsymbol{W}_*^{(j-1)}\, (\boldsymbol{I}_n - H_*)$. This completes the proof of Lemma 4.1.

Similarly, in the context of GEE, the β_{j-1}-specific Godambian estimator (4.12) of the dispersion parameter σ^2 can be approximated by

$$\widetilde{\sigma}_{j-1}^2 \simeq \widetilde{\boldsymbol{\epsilon}}^T \left\{ (\boldsymbol{I}_N - \mathcal{H}_*)\, \mathcal{W}_*^{(j-1)}\, (\boldsymbol{I}_N - \mathcal{H}_*) \right\} \widetilde{\boldsymbol{\epsilon}},$$

from the approximation of the transformed residual vector (4.11) and the quadratic form of the Godambian estimator (4.12). Under the null hypothesis H_0 (4.2), the vector $\widetilde{\boldsymbol{\epsilon}}$ is multivariate distributed with mean $\mathbf{0}$ and covariance matrix $\sigma^2 \boldsymbol{I}_N$. Then, the estimator $\widetilde{\sigma}_{j-1}^2$ has expectation as

$$E\left(\widetilde{\sigma}_{j-1}^2\right) \simeq \left(\sum_{k=1}^{N} \lambda_{k,*}^{(j-1)} \right) \sigma^2,$$

where $\lambda_{k,*}^{(j-1)}$ are the eigenvalues of the matrix $(\boldsymbol{I}_N - \mathcal{H}_*)\, \mathcal{W}_*^{(j-1)}\, (\boldsymbol{I}_N - \mathcal{H}_*)$. This completes the proof of Lemma 4.2.

A.2. Proof of Theorem 4.1

In the context of GLM, the unbiased β_{j-1}-specific Godambian estimator (4.6) of σ^2 can be rewritten as a quadratic form given by

$$\widetilde{\sigma}^2_{j-1,u} = \boldsymbol{r}_P^T \widehat{\boldsymbol{W}}_u^{(j-1)} \boldsymbol{r}_P.$$

The Pearson residual vector $\boldsymbol{r}_P$ can be approximated by $\boldsymbol{r}_P \simeq (\boldsymbol{I}_n - H_*)\,\boldsymbol{\epsilon}_P$ given by (4.4) for large sample size, where $\boldsymbol{\epsilon}_P$ is an n-variate random vector with mean $\mathbf{0}$ and covariance matrix $\sigma^2\Omega_*$. Thus, for large sample size, the Godambian estimator $\widetilde{\sigma}^2_{j-1,u}$ can be approximated by a quadratic form,

$$\widetilde{\sigma}^2_{j-1,u} \simeq Q(\boldsymbol{\epsilon}_P) = \boldsymbol{\epsilon}_P^T \left\{ (\boldsymbol{I}_n - H_*)\, \boldsymbol{W}_{u,*}^{(j-1)}\, (\boldsymbol{I}_n - H_*) \right\} \boldsymbol{\epsilon}_P, \tag{6.1}$$

where $\boldsymbol{W}_{u,*}^{(j-1)}$ is the matrix which substitutes the estimate $\widehat{\boldsymbol{\beta}}_n$ in the matrix $\widehat{\boldsymbol{W}}_u^{(j-1)}$ with the true value $\boldsymbol{\beta}_*$, for $j = 1, \cdots, p$. Note that because $\widehat{\boldsymbol{\beta}}_n$ is a consistent estimator, the matrix $\widehat{\boldsymbol{W}}_u^{(j-1)}$ may be approximated by $\boldsymbol{W}_{u,*}^{(j-1)}$ for large sample size. Let

$$C_* = (\boldsymbol{I}_n - H_*)\, \boldsymbol{W}_{u,*}^{(j-1)}\, (\boldsymbol{I}_n - H_*)\,,$$

which is an $n \times n$ symmetric matrix. Let $\lambda_{k,*}^{(j-1)}$, $k = 1, \cdots, n$, be the eigenvalues of the matrix C_*. Note that $\left[\lambda_{k,*}^{(j-1)}\right]^h$, $k = 1, \cdots, n$, are the eigenvalues of the matrix C_*^h, for $h = 1, 2, \cdots$.

As discussed in Section 4.1.1, under the null hypothesis H_0 (4.1), the matrix $\Omega_* = \boldsymbol{I}_n$, and consequently, the vector $\boldsymbol{\epsilon}_P$ has mean $\mathbf{0}$ and covariance matrix $\sigma^2\boldsymbol{I}_n$. In addition, under the null hypothesis H_0, the expectation and variance of the quadratic form $Q(\boldsymbol{\epsilon}_P)/\sigma^2$ are

$$E\left\{Q(\boldsymbol{\epsilon}_P)/\sigma^2\right\} = \sum_{k=1}^{n} \lambda_{k,*}^{(j-1)} = 1, \tag{6.2}$$

and

$$Var\left\{Q(\boldsymbol{\epsilon}_P)/\sigma^2\right\} = 2\sum_{k=1}^{n}\left[\lambda_{k,*}^{(j-1)}\right]^2. \tag{6.3}$$

By the central limit theorem for quadratic forms discussed in [46], we have

$$\frac{Q(\boldsymbol{\epsilon}_P)/\sigma^2 - E\left\{Q(\boldsymbol{\epsilon}_P)/\sigma^2\right\}}{\sqrt{Var\left\{Q(\boldsymbol{\epsilon}_P)/\sigma^2\right\}}} = \frac{Q(\boldsymbol{\epsilon}_P)/\sigma^2 - 1}{\sqrt{2\sum_{k=1}^{n}\left[\lambda_{k,*}^{(j-1)}\right]^2}} \to^d N(0,1) \quad \text{as } n \to \infty,$$

under the null hypothesis H_0. A necessary condition for the central limit theorem is

$$\frac{\max\left\{\left[\lambda_{k,*}^{(j-1)}\right]^2\right\}}{\sum_{k=1}^{n}\left[\lambda_{k,*}^{(j-1)}\right]^2} \to 0 \quad \text{as } n \to \infty,$$

because $\lambda_{k,*}^{(j-1)} = O(1/n)$, obtained from $\sum_{k=1}^{n}\lambda_{k,*}^{(j-1)} = 1$.

Let

$$\widehat{C} = \left(\boldsymbol{I}_n - \widehat{H}\right)\widehat{\boldsymbol{W}}_u^{(j-1)}\left(\boldsymbol{I}_n - \widehat{H}\right),$$

and let $\widehat{\lambda}_k^{(j-1)}$, $k = 1, \cdots, n$, be the eigenvalues of the matrix $\widehat{C}$. Note that this matrix $\widehat{C}$ can be regarded as a function of $\widehat{\boldsymbol{\beta}}_n$, and because the estimator $\widehat{\boldsymbol{\beta}}_n$ is a consistent estimator, it can be shown that

$$\widehat{C} \to_p C_*, \quad \text{as } n \to \infty,$$

and consequently,

$$\sum_{k=1}^{n}\left[\widehat{\lambda}_k^{(j-1)}\right]^2 = tr\left\{\widehat{C}^2\right\} \to_p \sum_{k=1}^{n}\left[\lambda_{k,*}^{(j-1)}\right]^2 = tr\left\{C_*^2\right\}, \quad \text{as } n \to \infty,$$

i.e.,

$$\frac{\sqrt{2\sum_{k=1}^{n}\left[\lambda_{k,*}^{(j-1)}\right]^2}}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\lambda}_k^{(j-1)}\right]^2}} \to_p 1, \quad \text{as } n \to \infty.$$

Thus, by the Slutsky Theorem ([8]),

$$\frac{Q(\boldsymbol{\epsilon}_P)/\sigma^2 - 1}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\lambda}_k^{(j-1)}\right]^2}} = \frac{Q(\boldsymbol{\epsilon}_P)/\sigma^2 - 1}{\sqrt{2\sum_{k=1}^{n}\left[\lambda_{k,*}^{(j-1)}\right]^2}} \frac{\sqrt{2\sum_{k=1}^{n}\left[\lambda_{k,*}^{(j-1)}\right]^2}}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\lambda}_k^{(j-1)}\right]^2}} \to^d N(0,1) \quad \text{as } n \to \infty,$$

under the null hypothesis H_0. Therefore, under the null hypothesis H_0, as $n \to \infty$, the standardized β_{j-1}-specific information ratio statistic (4.21)

$$\frac{\widetilde{\sigma}^2_{j-1,u}/\sigma^2 - 1}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\lambda}_k^{(j-1)}\right]^2}} \to^d N(0,1), \quad \text{as } n \to \infty,$$

for $j = 1, \cdots, p$.

Similarly, the unbiased pooled Godambian estimator (4.15) of the dispersion parameter can be approximated, for large sample size, by

$$\widetilde{\sigma}^2_{pool,u} \simeq \boldsymbol{\epsilon}_P^T \left\{ (\boldsymbol{I}_n - H_*)\, \boldsymbol{W}_{u,*}^{pool} (\boldsymbol{I}_n - H_*) \right\} \boldsymbol{\epsilon}_P, \tag{6.4}$$

where $\boldsymbol{W}_{u,*}^{pool}$ is the matrix which substitutes the estimate $\widehat{\boldsymbol{\beta}}_n$ in the matrix $\widehat{\boldsymbol{W}}_u^{pool}$ with the true value $\boldsymbol{\beta}_*$. Let $\widehat{\lambda}_k^{pool}$ be the eigenvalues of the matrix

$$\left(\boldsymbol{I}_n - \widehat{H}\right) \widehat{\boldsymbol{W}}_u^{pool} \left(\boldsymbol{I}_n - \widehat{H}\right).$$

Thus, the standardized pooled information ratio statistic (4.22)

$$\frac{\widetilde{\sigma}^2_{pool,u}/\sigma^2 - 1}{\sqrt{2\sum_{k=1}^{n}\left[\widehat{\lambda}_k^{pool}\right]^2}} \to^d N(0,1), \quad \text{as } n \to \infty,$$

under the null hypothesis H_0 (4.1).

In the context of GEE, the unbiased β_{j-1}-specific Godambian estimator of the dispersion parameter can be written as a quadratic form in the transformed residuals given in (4.13)

$$\widetilde{\sigma}^2_{j-1,u} = \widetilde{\boldsymbol{r}}^T \widehat{\boldsymbol{W}}_u^{(j-1)} \widetilde{\boldsymbol{r}}.$$

By the large-sample approximation of the transformed residuals given in (4.11)

$$\widetilde{\boldsymbol{r}} \simeq (\boldsymbol{I}_N - \mathcal{H}_*)\widetilde{\boldsymbol{\epsilon}},$$

the Godambian estimator $\widetilde{\sigma}^2_{j-1,u}$ can be approximated, for large sample size, by

$$\widetilde{\sigma}^2_{j-1,u} \simeq \widetilde{\boldsymbol{\epsilon}}^T(\boldsymbol{I}_N - \mathcal{H}_*)\mathcal{W}^{(j-1)}_{u,*}(\boldsymbol{I}_N - \mathcal{H}_*)\widetilde{\boldsymbol{\epsilon}}, \tag{6.5}$$

where $\widetilde{\boldsymbol{\epsilon}}$ is an N-variate random vector with mean $\mathbf{0}$ and covariance matrix $\sigma^2\boldsymbol{\Omega}_*$. Under the null hypothesis H_0 (4.2), the matrix $\boldsymbol{\Omega}_* = \boldsymbol{I}_N$. Let $\widehat{\lambda}^{(j-1)}_k$ be the eigenvalues of the matrix $(\boldsymbol{I}_N - \widehat{\mathcal{H}})\widehat{\mathcal{W}}^{(j-1)}_u(\boldsymbol{I}_N - \widehat{\mathcal{H}})$. Then, the standardized β_{j-1}-specific information ratio statistic (4.23)

$$\frac{\widetilde{\sigma}^2_{j-1,u}/\sigma^2 - 1}{\sqrt{2\sum_{k=1}^n \left[\widehat{\lambda}^{(j-1)}_k\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad \text{as } n \to \infty,$$

under the null hypothesis H_0.

Similarly, the unbiased pooled Godambian estimator (4.15) of the dispersion parameter can be approximated, for large sample size, by

$$\widetilde{\sigma}^2_{pool,u} \simeq \widetilde{\boldsymbol{\epsilon}}^T(\boldsymbol{I}_N - \mathcal{H}_*)\mathcal{W}^{pool}_{u,*}(\boldsymbol{I}_N - \mathcal{H}_*)\widetilde{\boldsymbol{\epsilon}}. \tag{6.6}$$

Let $\widehat{\lambda}^{pool}_k$ be the eigenvalues of the matrix $(\boldsymbol{I}_N - \widehat{\mathcal{H}})\widehat{\mathcal{W}}^{pool}_u(\boldsymbol{I}_N - \widehat{\mathcal{H}})$. Then, the standardized pooled information ratio statistic (4.24)

$$\frac{\widetilde{\sigma}^2_{pool,u}/\sigma^2 - 1}{\sqrt{2\sum_{k=1}^n \left[\widehat{\lambda}^{pool}_k\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad \text{as } n \to \infty,$$

under the null hypothesis H_0.

This completes the proof of Theorem 4.1.

A.3. Proof of Theorem 4.2, 4.3 and 4.4

Let us firstly consider a general form of a ratio of two quadratic forms in random variables. Suppose that a random vector $\boldsymbol{\xi}$ has mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}_n$. Let Q_j, $j = 1, 2$, denote quadratic forms in random variables $\boldsymbol{\xi}$, given by

$$Q_j = \boldsymbol{\xi}^T \mathcal{A}_j \boldsymbol{\xi}, \quad j = 1, 2.$$

Let μ_{Q_j} be the expectation of the quadratic form Q_j, for $j = 1, 2$. Assume that Q_j, $j = 1, 2$ have approximately the same expectations, say

$$\mu_{Q_j} \simeq \mu_Q, \quad j = 1, 2.$$

A Taylor expansion of the ratio Q_1/Q_2 about the point (μ_{Q_1}, μ_{Q_2}) is given by

$$\begin{aligned}
\frac{Q_1}{Q_2} &= \frac{\mu_{Q_1}}{\mu_{Q_2}} + \frac{\mu_{Q_1}}{\mu_{Q_2}} \left(\frac{Q_1}{\mu_{Q_1}} - \frac{Q_2}{\mu_{Q_2}} \right) + o_p(\|(Q_1 - \mu_{Q_1}, Q_2 - \mu_{Q_2})\|) \\
&\simeq 1 + \frac{1}{\mu_Q} (Q_1 - Q_2) + o_p(\|(Q_1 - \mu_{Q_1}, Q_2 - \mu_{Q_2})\|) \\
&= 1 + \frac{1}{\mu_Q} \boldsymbol{\xi}^T (\mathcal{A}_1 - \mathcal{A}_2) \boldsymbol{\xi} + o_p(\|(Q_1 - \mu_{Q_1}, Q_2 - \mu_{Q_2})\|).
\end{aligned}$$

Let τ_k, $k = 1, \cdots, n$ be the eigenvalues of the matrix $\mathcal{A}_1 - \mathcal{A}_2$. By the central limit theorem in [46], we get

$$\frac{Q_1/Q_2 - 1}{\left(\sqrt{2 \sum_{k=1}^n \tau_k^2} \right) / \mu_Q} \to^d N(0, 1), \quad \text{as } n \to \infty.$$

In the context of GLM, for the β_{j-1}-specific information ratio statistic, the unbiased β_{j-1}-specific Godambian estimator $\widetilde{\sigma}^2_{j-1,u}$ can be approximated by a quadratic form, for large sample size, given in (6.1)

$$\frac{\widetilde{\sigma}^2_{j-1,u}}{\sigma^2} \simeq Q_1 = \left(\frac{\boldsymbol{\epsilon}_P}{\sigma} \right)^T C_{j-1} \left(\frac{\boldsymbol{\epsilon}_P}{\sigma} \right),$$

where $C_{j-1} = (\boldsymbol{I}_n - H_*)\,\boldsymbol{W}_{u,*}^{(j-1)}\,(\boldsymbol{I}_n - H_*)$. And, the moment estimator of σ^2 can also be approximated, for large sample size, by

$$\frac{\widetilde{\sigma}_m^2}{\sigma^2} \simeq Q_2 = \left(\frac{\boldsymbol{\epsilon}_P}{\sigma}\right)^T C_m \left(\frac{\boldsymbol{\epsilon}_P}{\sigma}\right),$$

where $C_m = \frac{1}{n-p}\,(\boldsymbol{I}_n - H_*)$. Under the null hypothesis H_0 (4.1), the vector $\boldsymbol{\epsilon}_P/\sigma$ has mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}_n$ for large sample size. Then, the expectations, μ_{Q_1} and μ_{Q_2}, of these two quadratic forms Q_1 and Q_2 are approximately the same, equal to 1. Let $\tau_{k,*}^{(j-1)}$, $k = 1, \cdots, n$, be the eigenvalues of the matrix $C_{j-1} - C_m$. Then, under the null hypothesis H_0,

$$\frac{\widetilde{\sigma}_{j-1,u}^2/\widehat{\sigma}_m^2 - 1}{\sqrt{2\sum_{k=1}^n \left[\tau_{k,*}^{(j-1)}\right]^2}} \to^d N(0,1), \quad \text{as } n \to \infty.$$

Let $\widehat{C}_{j-1} = \left(\boldsymbol{I}_n - \widehat{H}\right)\widehat{\boldsymbol{W}}_u^{(j-1)}\left(\boldsymbol{I}_n - \widehat{H}\right)$ and $\widehat{C}_m = \frac{1}{n-p}\left(\boldsymbol{I}_n - \widehat{H}\right)$. Let $\widehat{\tau}_k^{(j-1)}$ be the eigenvalues of the matrix

$$\widehat{C}_{j-1} - \widehat{C}_m = \left(\boldsymbol{I}_n - \widehat{H}\right)\left(\widehat{\boldsymbol{W}}_u^{(j-1)} - \frac{1}{n-p}\boldsymbol{I}_n\right)\left(\boldsymbol{I}_n - \widehat{H}\right).$$

Note that the matrices $\widehat{C}_{j-1}$ and $\widehat{C}_m$ are functions of $\widehat{\boldsymbol{\beta}}_n$. Since $\widehat{\boldsymbol{\beta}}_n$ is a consistent estimator, then

$$\widehat{C}_{j-1} \to_p C_{j-1} \quad \text{and} \quad \widehat{C}_m \to_p C_m, \quad \text{as } n \to \infty.$$

Consequently, as $n \to \infty$,

$$\sum_{k=1}^n \left[\widehat{\tau}_k^{(j-1)}\right]^2 = tr\left\{\left(\widehat{C}_{j-1} - \widehat{C}_m\right)^2\right\} \to_p \sum_{k=1}^n \left[\tau_{k,*}^{(j-1)}\right]^2 = tr\left\{\left(C_{j-1} - C_m\right)^2\right\},$$

i.e.,

$$\frac{\sqrt{2\sum_{k=1}^n \left[\tau_{k,*}^{(j-1)}\right]^2}}{\sqrt{2\sum_{k=1}^n \left[\widehat{\tau}_k^{(j-1)}\right]^2}} \to_p 1, \quad \text{as } n \to \infty.$$

Thus, by the Slutsky Theorem, the standardized β_{j-1}-specific information ratio statistic (4.25)

$$\frac{\widetilde{\sigma}^2_{j-1,u}/\widehat{\sigma}^2_m - 1}{\sqrt{2\sum_{k=1}^n \left[\widehat{\tau}_k^{(j-1)}\right]^2}} = \frac{\frac{\widetilde{\sigma}^2_{j-1,u}}{\sigma^2}/\frac{\widehat{\sigma}^2_m}{\sigma^2} - 1}{\sqrt{2\sum_{k=1}^n \left[\tau_{k,*}^{(j-1)}\right]^2}} \frac{\sqrt{2\sum_{k=1}^n \left[\tau_{k,*}^{(j-1)}\right]^2}}{\sqrt{2\sum_{k=1}^n \left[\widehat{\tau}_k^{(j-1)}\right]^2}} \to^d N(0,1), \quad \text{as } n \to \infty,$$

under the null hypothesis H_0.

Similarly, the unbiased pooled Godambian estimator can be approximated, for large sample size, as given in (6.4), by

$$\frac{\widetilde{\sigma}^2_{pool,u}}{\sigma^2} \simeq Q_1 = \left(\frac{\boldsymbol{\epsilon}_P}{\sigma}\right)^T C_{pool} \left(\frac{\boldsymbol{\epsilon}_P}{\sigma}\right),$$

where $C_{pool} = (\boldsymbol{I}_n - H_*)\, \boldsymbol{W}^{pool}_{u,*} (\boldsymbol{I}_n - H_*)$. Let $\widehat{C}_{pool} = \left(\boldsymbol{I}_n - \widehat{H}\right) \widehat{\boldsymbol{W}}^{pool}_u \left(\boldsymbol{I}_n - \widehat{H}\right)$, and let $\widehat{\tau}^{pool}_k$, $k = 1, \cdots, n$, be the eigenvalues of the matrix

$$\widehat{C}_{pool} - \widehat{C}_m = \left(\boldsymbol{I}_n - \widehat{H}\right)\left(\widehat{\boldsymbol{W}}^{pool}_u - \frac{1}{n-p}\boldsymbol{I}_n\right)\left(\boldsymbol{I}_n - \widehat{H}\right).$$

Then, the standardized pooled information ratio statistic (4.26)

$$\frac{\widetilde{\sigma}^2_{pool,u}/\widehat{\sigma}^2_m - 1}{\sqrt{2\sum_{k=1}^n \left[\widehat{\tau}_k^{pool}\right]^2}} \to^d N(0,1), \quad \text{as } n \to \infty,$$

under the null hypothesis H_0. This completes the proof of Theorem 4.2.

In the context of GEE, for $j = 1, \cdots, p$, the unbiased β_{j-1}-specific Godambian estimator can be approximated, for large sample size, as given in (6.5), by

$$\frac{\widetilde{\sigma}^2_{j-1,u}}{\sigma^2} \simeq \left(\frac{\widetilde{\boldsymbol{\epsilon}}}{\sigma}\right)^T C_{j-1} \left(\frac{\widetilde{\boldsymbol{\epsilon}}}{\sigma}\right),$$

where $C_{j-1} = (\boldsymbol{I}_N - \mathcal{H}_*)\,\mathcal{W}_{u,*}^{(j-1)}\,(\boldsymbol{I}_N - \mathcal{H}_*)$. The unbiased"Pearson" moment estimator (4.17) can also be approximated, for large sample size, by

$$\frac{\widetilde{\sigma}^2_{m,u}}{\sigma^2} \simeq \left(\frac{\widetilde{\boldsymbol{\epsilon}}}{\sigma}\right)^T C_{_P} \left(\frac{\widetilde{\boldsymbol{\epsilon}}}{\sigma}\right),$$

where $C_{_P} = (\boldsymbol{I}_N - \mathcal{H}_*)\left(\mathcal{W}_{_{P,*}}/m_{_{P,*}}\right)(\boldsymbol{I}_N - \mathcal{H}_*)$, with $\mathcal{W}_{_{P,*}}$ and $m_{_{P,*}}$ substituting the estimators $\widehat{\boldsymbol{\beta}}_K$ and $\widehat{\boldsymbol{\rho}}_K$ in $\widehat{\mathcal{W}}_{_P}$ and $m_{_P}$ with the limiting value $\boldsymbol{\beta}_*$ and $\boldsymbol{\rho}_*$.

Let $\widehat{C}_{j-1} = \left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\widehat{\mathcal{W}}_u^{(j-1)}\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)$ and $\widehat{C}_{_P} = \left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\left(\widehat{\mathcal{W}}_{_P}/m_{_P}\right)\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)$. Let $\widehat{\tau}_k^{(j-1)}$, $k = 1, \cdots, n$, be the eigenvalues of the matrix

$$\widehat{C}_{j-1} - \widehat{C}_{_P} = \left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\left(\widehat{\mathcal{W}}_u^{(j-1)} - \widehat{\mathcal{W}}_{_P}/m_{_P}\right)\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right).$$

Then, under the null hypothesis H_0 (4.2), the standardized β_{j-1}-specific information ratio statistic (4.27)

$$\frac{\widetilde{\sigma}^2_{j-1,u}/\widehat{\sigma}^2_{m,u} - 1}{\sqrt{2\sum_{k=1}^n \left[\widehat{\tau}_k^{(j-1)}\right]^2}} \to^d N(0,1), \qquad \text{as } n \to \infty.$$

The unbiased pooled Godambian estimator can be approximated, for large sample size, as given in (6.5), by

$$\frac{\widetilde{\sigma}^2_{pool,u}}{\sigma^2} \simeq \left(\frac{\widetilde{\boldsymbol{\epsilon}}}{\sigma}\right)^T C_{pool} \left(\frac{\widetilde{\boldsymbol{\epsilon}}}{\sigma}\right),$$

where $C_{pool} = (\boldsymbol{I}_N - \mathcal{H}_*)\,\mathcal{W}_{u,*}^{pool}\,(\boldsymbol{I}_N - \mathcal{H}_*)$. Let $\widehat{C}_{pool} = \left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\widehat{\mathcal{W}}_u^{pool}\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)$, and let $\widehat{\tau}_k^{pool}$, $k = 1, \cdots, n$, be the eigenvalues of the matrix

$$\widehat{C}_{pool} - \widehat{C}_{_P} = \left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\left(\widehat{\mathcal{W}}_u^{pool} - \widehat{\mathcal{W}}_{_P}/m_{_P}\right)\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right).$$

Then, under the null hypothesis H_0 (4.2), the standardized pooled information ratio statistics (4.28)

$$\frac{\widetilde{\sigma}^2_{pool,u}/\sigma^2 - 1}{\sqrt{2\sum_{k=1}^n \left[\widehat{\tau}_k^{pool}\right]^2}} \to^d N(0,1), \qquad \text{as } n \to \infty.$$

This completes the proof of Theorem 4.3.

The "transformed" moment estimator (4.18) of the dispersion parameter can be approximated, for large sample size, by

$$\widehat{\sigma}^2_{tr} \simeq \left(\frac{\widetilde{\boldsymbol{\epsilon}}}{\sigma}\right)^T C_{tr} \left(\frac{\widetilde{\boldsymbol{\epsilon}}}{\sigma}\right),$$

where $C_{tr} = \frac{1}{N-p}(\boldsymbol{I}_N - \mathcal{H}_*)$. Let $\widehat{C}_{tr} = \frac{1}{N-p}\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)$, and let $\widehat{\tau}_k^{(j-1)}$, $j = 1, \cdots, p$, be the eigenvalues of the matrix

$$\widehat{C}_{j-1} - \widehat{C}_{tr} = \left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\left(\widehat{\mathcal{W}}_u^{(j-1)} - \frac{1}{N-p}\boldsymbol{I}_N\right)\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right).$$

Under the null hypothesis H_0 (4.2), the standardized β_{j-1}-specific information ratio statistic (4.29)

$$\frac{\widetilde{\sigma}^2_{j-1,u}/\widehat{\sigma}^2_{tr} - 1}{\sqrt{2\sum_{k=1}^n \left[\widehat{\tau}_k^{(j-1)}\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad \text{as } n \to \infty.$$

Let $\widehat{\tau}_k^{pool}$ be the eigenvalues of the matrix

$$\widehat{C}_{pool} - \widehat{C}_{tr} = \left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right)\left(\widehat{\mathcal{W}}_u^{pool} - \frac{1}{N-p}\boldsymbol{I}_N\right)\left(\boldsymbol{I}_N - \widehat{\mathcal{H}}\right).$$

Under the null hypothesis H_0 (4.2), the standardized pooled information ratio statistic (4.30)

$$\frac{\widetilde{\sigma}^2_{pool,u}/\widehat{\sigma}^2_{tr} - 1}{\sqrt{2\sum_{k=1}^n \left[\widehat{\tau}_k^{pool}\right]^2}} \quad \xrightarrow{d} \quad N(0,1), \quad \text{as } n \to \infty.$$

This completes the proof of Theorem 4.4.

Bibliography

[1] H. Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21:243–247, 1969. 4

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281. Budapest: Akademiai Kiado, 1973. 5

[3] H. Akaike. A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, 30:9–14, 1978. 5

[4] Davis M. Allen. The relationship between variable selection and data augumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974. 5

[5] F. Anscombe. Examination of residuals. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, 1:1–3, 1961. 4, 99

[6] O.E. Barndorff-Nielsen. Quasi profile and directed likelihoods from estimating functions. *Annals of the Institute of Statistical Mathematics*, 47:461–464, 1995. 3

[7] P.J. Bickel. Using residuals robustly I: Tests for heteroscedasticity, nonlinearity. *The Annals of Statistics*, 6(2):266–291, 1978. 4, 99

[8] Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2nd edition, 1999. 158

[9] N. Breslow. Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association*, 85:565–571, 1990. 31

[10] M.J. Buckley and G.K. Eagleson. An approximation to the distribution of quadratic forms in normal random variables. *Australian & New Zealand Journal of Statistics*, 30:150–159, 1988. 89

[11] Raymond J. Carroll and David Ruppert. On robust tests for heteroscedasticity. *The Annals of Statistics*, 9(1):206–210, 1981. 4

[12] B. Chandrasekar and B.K. Kale. Unbiased statistical estimation functions in the presence of nuisance parameters. *Journal of Statistical Planning and Inference*, 9:45–54, 1984. 16

[13] B. J. Collings and B. H. Margolin. Testing goodness of fit for the Poisson assumption when observations are not identically distributed. *Journal of the American Statistical Association*, 80:411–418, 1985. 4

[14] R. Dennis Cook and Sanford Weisberg. Diagnostics for heteroscedasticity in regression. *Biometrika*, 70(1):1–10, 1983. 4, 99

[15] Gauss M. Cordeiro. On Pearson's residuals in generalized linear models. *Statistics & Probability Letters*, 66:213–219, 2004. 65

[16] D.R. Cox and E.J. Snell. A general definition of residuals. *Journal of the Royal Statistical Society*, 30:277–299, 1968. 65

[17] M. Crowder. On linear and quadratic estimating functions. *Biometrika*, 74:591–597, 1987. 18

[18] M. Crowder. On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, 82:407–410, 1995. 153

[19] Piet De Jong and Gillian Heller. *Generalized Linear Models for Insurance Data*. Cambridge University Press, 2008. 136

[20] C. Dean and Jerry F. Lawless. Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association*, 84(46):467–472, 1989. xii, 4, 66, 103, 104, 105

[21] C. B. Dean. Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*, 87(418):451–457, 1992. 4

[22] Angela Diblasi and Adrian Bowman. Testing for constant variance in a linear model. *Statistics and Probability Letters*, 33:95–103, 1997. 4

[23] Peter J. Diggle, Kung-Yee Liang, and Scott L. Zeger. *Analysis of Longitudinal Data*. Clarendon Press, Oxford, 1994. 2, 144, 148

[24] B. Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331, 1983. 5

[25] Ludwig Fahrmier, Gerhard Tutz, and Wolfgang Hennevogl. *Multivariate Statistical Modelling based on Generalized Linear Models*. Springer, 2nd edition, 2001. 66

[26] H. Fairfield-Smith. The problem of comparing the results of two experiments with unequal errors. *Journal of the Council for Scientific & Industrial Research*, 9:211–212, 1936. 89

[27] R. A. Fisher. The significance of deviations from expectation in a Poisson series. *Biometrics*, 6:17–24, 1950. 4

[28] Ronald Aylmer Fisher. The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85:597–612, 1922. 1

[29] Ronald Aylmer Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, 222:309–368, 1922. 1

[30] Ronald Aylmer Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6:391–398, 1935. 2

[31] Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics, 2004. 59, 60

[32] E.L. Frome. The analysis of rates using Poisson regression models. *Biometrics*, 39:665–674, 1983. 48

[33] Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:320–328, 1975. 5

[34] H. Glejser. A new test for heteroscedasticity. *Journal of the American Statistical Association*, 64:316–323, 1969. 4

[35] V.P. Godambe. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211, 1960. 2, 12, 17, 18

[36] V.P. Godambe. *Estimating Functions*. Clarendon Press, Oxford, 1991. 2

[37] S. M. Goldpeld and R. E. Quandt. Some tests for heteroscedasticity. *Journal of the American Statistical Association*, 60:539–547, 1965. 4

[38] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins, 3rd edition, 1996. 48

[39] Thomas Hammerstrom. Asymptotically optimal tests for heteroscedasticity in the general linear model. *The Annals of Statistics*, 9(2):368–380, 1981. 4

[40] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982. 10

[41] James William Hardin and Joseph Hilbe. *Generalized Estimating Equations*. Chapman and Hall, 2002. 152

[42] M.J. Harrison and B.P.M. McCabe. A test for heteroscedasticity based on ordinary least squares residuals. *Journal of the American Statistical Association*, 74(366):494–499, 1979. 4

[43] Lin-Yee Hin and You-Gan Wang. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine*, 28:642–658, 2009. 5, 6, 122, 128, 135, 148, 149, 152

[44] David C. Hoaglin and Roy E. Welsch. The hat matrix in regression and ANOVA. *The American Statistician*, 32(1):17 – 22, 1978. 53

[45] Peter J. Huber. The behavior of maximum likelihood estimation under nonstandard conditions. In L.M. LeCam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233, Berkeley, 1967. University of California Press. 2, 30

[46] Jiming Jiang. REML estimation: Asymptotic behavior and related topics. *The Annals of Statistics*, 24(1):255–286, 1996. 75, 157, 160

[47] B. Jørgensen. Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society, Series B*, 49:127–162, 1987. 66, 130

[48] B. Jørgensen. *The Theory of Dispersion Models*. Chapman and Hall, London, 1997. 67

[49] R. Kaas. Compound Poisson distribution and GLM's - Tweedie's distribution. In *Proceedings of the Contact Forum "3rd Actuarial and Financial Mathematics Day"*, pages 3–612. Royal Flemish Academy of Belgium for Science and the Arts, 2005. 114, 142

[50] Göran Kauermann and Raymond J. Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396, 2001. 31, 35

[51] B.F. Kimball. Sufficient statistical estimation functions for the parameters of the distribution of maximum values. *Annals of Mathematical Statistics*, 17:299–309, 1946. 2

[52] P.R. Krishnaiah. Selection of variables under univariate regression models. In P.R. Krishnaiah, editor, *Handbook of Statistics II*, pages 805–820, North-Holland,Amsterdam, 1982. 4

[53] Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986. 2, 24, 27, 29, 30, 112

[54] Kung-Yee Liang, Scott L. Zeger, and Bahajt Qaqish. Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, Series B*, 54:3–40, 1992. 2, 152

[55] H. Linhart and W. Zucchini. *Model Selection*. Wiley, 1986. 4

[56] C.L. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973. 5

[57] Peter McCullagh. Quasi-likelihood functions. *The Annals of Statistics*, 11(1):59–67, 1983. 65

[58] Peter McCullagh. On the asymptotic distribution of Pearson's statistics in linear exponential-family models. *International Statistical Review*, 53(1):61–67, 1985. 65

[59] Peter McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, 2nd edition, 1989. 3, 5, 20, 69

[60] Don L. McLeish and C.G. Small. The theory and applications of statistical inference functions. Lecture Notes in Statistics 44, New York, 1988. 2

[61] A.J. Miller. *Subset Selection in Regression*. Chapman and Hall, London, 1990. 4

[62] Hans-Georg Müller and Peng-Liang Zhao. On a semiparametric variance function model and a test for heteroscedasticity. *The Annals of Statistics*, 23(3):946–967, 1995. 4

[63] Alvaro Munoz, Vincent Carey, Jan P. Schouten, Mark Segal, and Bernard Rosner. A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics*, 48(3):733–742, 1992. 114

[64] E.A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964. 54

[65] Wei Pan. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57:120–125, 2001. 5, 122, 152

[66] Donald A. Pierce and Daniel W. Schafer. Residuals in generalized linear models. *Journal of the American Statistical Association*, 81(396):977–986, 1986. 65

[67] Ross L. Prentice. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44:1033–1048, 1988. 112, 152

[68] Annie Qu, Bruce Lindsay, and Bing Li. Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87:823–836, 2000. 3, 63, 148, 153

[69] C.R. Rao. *Linear Statistical Inference and its Applications*. Wiley, New York, 2nd edition, 1973. 3

[70] C.R. Rao and Y. Wu. On model selection (with discussion). In P. Lahiri, editor, *Lecture Notes-Monograph series, Model Selection*, volume 38, pages 1–64, Beachwood OH, 2001. Institute of Mathematical Statistics. 113

[71] J. M. Robins, A. Rotnitzky, and L.P. Zhao. Analysis of semiparametric regression models of repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90:106–121, 1995. 148

[72] F.E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics*, 2:110–114, 1946. 89

[73] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. 5

[74] Jeffrey S. Simonoff. *Smoothing Methods in Statistics*. Springer, 1996. 54

[75] Christopher G. Small and Jinfang Wang. *Numerical Methods for Nonlinear Estimating Equations*. Oxford, 2003. 41

[76] H. Solomon and M.A. Stephens. Distribution of a weighted sum of chi-square variables. *Journal of the American Statistical Association*, 72:881–885, 1977. 89

[77] Peter X. K. Song. *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer, 2007. 10, 16, 67

[78] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36(2):111–147, 1974. 5

[79] Brajendra C. Sutradhar and Kalyan Das. On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika*, 86:459–465, 1999. 64

[80] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996. 5

[81] Marcel D.T. Vieira and Christ J. Skinner. Estimating models for panel survey data under complex sampling. *Journal of Official Statistics*, 24:343–364, 2008. 124

[82] Abraham Wald. Tests of statistical hypothesis concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54:426–482, 1943. 3, 42

[83] Abraham Wald. Score tests in GLIM with applications. In R. Gilchrist, editor, *GLIM'82: Proceedings of the International Conference on Generalised Linear Models*, pages 87–97, Berlin, 1982. Springer-Verlag. 3

[84] You-Gan Wang and Vincent Carey. Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. *Biometrika*, 90(1):29–41, 2003. 63, 106, 112

[85] You-Gan Wang and Xu Lin. Effects of variance-function misspecification in analysis of longitudinal data. *Biometrics*, 61:413–421, 2005. 64, 106

[86] R.W.M. Wedderburn. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, 61:699–704, 1974. 3, 20, 28

[87] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. 3, 4, 8, 41, 42, 44, 151

[88] Hulbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980. 2, 4, 30, 41

[89] D.A. Williams. Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics*, 36:181–191, 1987. 48

[90] G.E. Willmot. Mixed compound Poisson distribution. *ASTIN Bulletin Supplement*, 16:59–79, 1986. 114

[91] C.F.J. Wu. Jackknife, bootstrap and other resampling methods in statistics. *Annals of Statistics*, 14:1261–1350, 1986. 31