Effect of Misspecified Response Correlation in Regression Analysis

by

Grace Shung-Lai Chiu

B.Sc. University of British Columbia, 1994

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June 1996

©Grace S. Chiu, 1996

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of STATISTICS

The University of British Columbia
Vancouver, Canada

Date JUNE 25, 1996

Abstract

One can imagine a possible loss of parameter estimation efficiency when response correlation is ignored or misspecified in modeling a response-covariate relationship. Under what conditions is efficiency lost? How much is lost?

Whether the responses are correlated or independent, standard theory for the distribution of least squares parameter estimates in linear models (Gaussian responses) can be readily determined. We find that the linear regression analysis assuming independent responses is (theoretically) never more efficient than that incorporating response dependence. The "difference" in efficiencies between these two analyses — measured by how much more readily the latter detects a non-zero regression coefficient — generally increases as the coefficient-to-noise ratio increases.

To incorporate response correlation in GLM parameter estimation, Liang & Zeger (1986) extended the quasi-likelihood theory and developed the *generalized estimating equations* (GEE) approach. Despite being a popular method, the effects of misspecifying response correlation (e.g. assuming independence when responses are correlated) on parameter estimation efficiency using GEE are not obvious. To investigate such effects, we use simulation studies in which we generate count data and use the GEE approach to estimate the model parameters, using both the correct and misspecified correlation structures. The generated counts, the number of correlated responses in each cluster/replicate, and the total number of replicates are all small to imitate health impact studies (in which hospital admission counts are often the responses). Despite possible loss of parameter estimation efficiency due to such "obstacles" intrinsic in the model, simulation results indicate that the GEE approach produces

1. regression parameter estimates with relatively small empirical biases using either a correct or misspecified response correlation;

2. a good estimate of the response correlation matrix if its structure is correctly specified;

3. *naive* and *robust* variance estimates both of which estimate the true variance well when the response correlation structure is correctly specified; and

4. good *robust* variance estimates even when the response correlation is misspecified.

Furthermore, in a GLM with exchangeably correlated Poisson data and no covariates, specifying independence or exchangeable dependence yields the same intercept estimate and estimation efficiency, provided that inference is based on the *robust* variance estimate. The *naive* variance estimate can significantly underestimate the true variance if the responses are assumed independent when analyzing such a GLM.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my most sincere thanks to my two supervisors, Dr. Nhu Le and Dr. James Zidek, for their patience and guidance throughout the past twelve months. It has been an honor and a wonderful experience working for them. I would also like to thank Dr. John Petkau for being my mentor for the past four years. His guidance, support, and numerous advices will be eternally appreciated. Many thanks to the rest of the faculty members for being kind and helpful throughout my Master's program; to the Statistics Department office staff for always being helpful and patient; to Paige Eveson for her friendship, patience, help with my writing, MVN r.v. generator, and being the best officemate one can ever have; and to the rest of the Statistics graduate students for sharing both tough and wonderful times with me.

Chapter 1

Introduction

Researchers are often interested in the relationship between a response variable and explanatory variables (covariates). In many health care studies, for instance, the response may be some measure of health status of a population, and the covariates may be indices of different pollution types to which the population is exposed.

The relationship between response and covariates can be determined through regression analysis. Which regression method to use depends on the nature of the response variable. For example, *linear regression* via least squares estimation can be applied to response data which can be assumed to follow a normal distribution. When we can no longer assume Gaussian response data, we may sometimes use a *generalized linear model* (GLM) for the response-covariate relationship. Such a model may be appropriate, for instance, when the response data are discrete and consist of small counts, such as hospital admission counts as a measure of health status of a population.

Regardless of the regression method, researchers should be very careful in fitting the model when any correlation possibly exists among the response data. For example, people living in neighboring geographical regions may tend to be at similar risk of developing respiratory ailments due to similar environmental factors shared by these regions. Efficient statistical inference is essential to, say, ensure appropriate allocation of government funding for controlling pollutant levels. Thus, introducing response correlation into the statistical model when studying the health impacts of pollution may be necessary. (This is an example of *spatial correlation* among the responses.) However, some researchers in the past would not incorporate response correlation when modeling a response-covariate relationship. This was sometimes due to the belief that the only source of variability came from independent and identically distributed (i.i.d.) measurement errors.[1] It might also be due to the complexity of incorporating response correlation into some non-linear statistical models such as a GLM.

Using data which resemble those in health care studies, in this thesis we investigate the effect of misspecified response correlation on the efficiency of parameter estimation in both linear and generalized linear regression analyses. For example, ignoring response correlation is a form of correlation misspecification. The effect of such misspecification in linear regression is closely investigated in Chapter 2. We then turn our attention to GLM's in the following chapters. Our GLM's involve small Poisson count responses which resemble hospital admission counts observed from neighboring regions in studies of rare diseases. We use the *generalized estimating equations* (GEE) approach for fitting our GLM's, which allows the specification of a *working* response correlation structure.[2] The effect of ignoring response correlation in a GLM is studied in Chapter 5, while Chapter 4 examines the effect of specifying some non-independence structure for independent responses.

Simulation experiments are used in our investigation. Each experiment involves generation of many datasets (responses and covariates), regression analyses using the true and incorrect response correlation structures respectively for each dataset, and the (empirical) efficiency comparison between the two analyses. We also include analytical arguments to support some of our simulation results.

We measure (empirical) *efficiency* by how often a regression analysis rejects the null hypothesis of a zero regression coefficient in a simulation experiment.[3] Rejecting such a

[1] See Cressie, 1991, page 25.

[2] See page 80.

[3] That is, the sample probability of hypothesis rejection by the analysis in the simulation experiment.

hypothesis is equivalent to detecting a non-zero coefficient in the regression model. Naturally, we say that the analysis that detects an existing non-zero coefficient more readily is *more efficient.* However, efficiency in this context is directly influenced by the variability of the coefficient estimate produced by the analysis. If the estimate has a very large standard error, then the test statistic for the above hypothesis test will be very small, the null hypothesis will be retained and hence the analysis will be inefficient (not very powerful) for detecting this coefficient. For example, we find that in a linear regression, the *generalized least squares* estimator (see Section 2.3) of a coefficient often has much smaller variability than the *ordinary least squares* estimator. Thus, the analysis based on the former estimator (which incorporates the response correlation) is more efficient/powerful than that based on the latter (which assumes response independence). We also find that an analysis generally increases in its power to detect a non-null regression coefficient as the coefficient-to-noise[4] ratio increases.

The GLM coefficient estimates produced by the GEE approach have a known asymptotic Gaussian distribution. To assess efficiency of a GLM analysis, we can compute the *naive* and *robust* variance estimates[5] for the GEE coefficient estimates based on this asymptotic distribution. However, each simulation experiment involves a finite number of data replicates.[6] Thus, the asymptotic variance estimates may not estimate the true variances well. The assessment of efficiency is then affected by how good (close to the true variances) these variance estimates are. In particular, an analysis that produces a smaller but severely biased variance estimate may not be more efficient than another that produces a larger but unbiased variance estimate. Yet how do we know if an estimate is unbiased?

Unlike the least squares estimates in a linear model, the exact distribution of a GEE

[4] *Noise* is the standard deviation of the random errors (with zero mean) in the model.

[5] See Section 3.3.

[6] See Section 3.3.

coefficient estimate with finite number of replicates is typically unavailable. However, given enough datasets in a simulation experiment, we may use the sample variance of the GEE estimates[7] as the "gold standard" for the true variance. In a simulation experiment, for each analysis (using correct/incorrect response correlation), we first assess how good the *naive* and *robust* variance estimates are by comparing them to this gold standard. Then we compare the efficiencies of the analyses based on their *naive* and *robust* estimates respectively. Finally, considering how well the true variance is estimated, we can conclude which analysis based on which variance estimate is more efficient. For example, we find that for the GLM in Chapter 4, the analyses using the correct and misspecified response correlations respectively yield good coefficient estimates[8] and good *robust* variance estimates. Thus, both analyses based on their *robust* variance estimates have similar efficiencies. (We do not examine analyses based on the *naive* variance estimates in this chapter.)

For the GLM in Chapter 5, assuming[9] either independent responses (incorrect) or exchangeably correlated responses (correct) yields the same GEE coefficient estimate and *robust* estimate of its variance. Both are good estimates of the true parameters. However, the *naive* estimate can considerably underestimate the true variance when assuming response independence. On the other hand, the *naive* estimate is good if the response correlation structure is correctly specified. In other words, the analysis assuming either independent or exchangeable responses based on the *robust* variance estimate yields the same efficiency. The latter analysis based on the *naive* variance estimate yields similar efficiency to that of these two analyses based on the *robust* estimate. However, assuming

[7] Each dataset produces one vector of coefficient estimates.

[8] That is, the estimates have small empirical biases.

[9] Technically, specifying a *working* correlation structure is not an assumption in model fitting. However, we will use the word "assume" to reflect the belief that the response data have the correlation structure as specified by the working correlation. See Section 3.4 for more details.

independent responses and using the *naive* variance estimate yields an incorrect[10] and thus inefficient analysis.

[10]... because the variance of the coefficient estimate is *incorrectly*/underestimated.

Chapter 2

Multiple Linear Regression

2.1 Introduction

Suppose we are interested in the relationship between some response variable and g explanatory variables, or *covariates.* The objective is to estimate the *regression parameters* associated with each covariate. Suppose we believe that the response-covariate relationship is linear. In other words, we believe that the response variable can be expressed as a linear combination of the covariates, with the coefficients in the linear combination being the regression parameters/coefficients. Then we may estimate these parameters via *least squares linear regression.* If we measure r sets of the response variable and the covariates, $r > g + 1$,[1] least squares parameter estimates in such a regression can then be explicitly expressed in terms of the (r) responses, their covariances (if they are correlated), and the observed covariates.

In many real life situations, the measurements of the response variable of interest may not be independent. For example, in a clinical trial to determine the effects of a drug, we measure some response on each patient several times over the period of the trial. Thus, repeated measurements on the same subject may be correlated over time.

In this thesis, we are more interested in *spatially correlated* response variables. Suppose we take a measurement of a single response from each of several sites within the same geographical domain of interest. The measurements (of the same response) across the different sites may be correlated when they come from proximate sites. Correlation

[1] See Section 2.3.

among the measurements can be induced by similar environmental factors acting across the geographical domain. Such a factor would be air pollution level. Our scenario might well be realized in health care studies, where the impact of air pollution on human health is measured through some response variable on a human subject from each site, together with pollution and meteorological covariates.

Researchers may overlook the correlation in the responses. In particular, they often believe that i.i.d. measurement errors constitute the only source of response variability.[2] Furthermore, it may be tempting to assume that the geographical sites are so far apart that correlation among the responses may be ignored without affecting the inferences about regression parameters. However, the strength of the correlations may depend not only on the distance between the sites, but also on the similarity of their shared environmental factors.

In this chapter, we concentrate on potential dangers which may arise in multiple linear regression if one overlooks the correlation among the response measurements. We will assume that the covariances of the responses are known. That assumption will not usually be realistic, in which case the covariance matrix must be estimated from the observed data. (See Johnson & Wichern, 1982 for details on how regression inference changes when the covariance matrix is estimated.) However, when accounting for correlated responses in regression, we expect that the conclusions based on estimated parameters will not change significantly if the correlation is well estimated. We only compare conclusions based on (1) assumed response independence to (2) admitted response dependence in parameter estimation. We are not interested in the possible efficiency loss in parameter estimation due to a poorly estimated response covariance matrix.

Our investigation uses simulation studies. In each simulation, we generate realizations of model (2.1) (see Section 2.3). With each realization/dataset, the regression parameters

[2] See Cressie, 1991, page 25.

are estimated with and without taking response dependence into account. Note that in all our simulation experiments, we fit the correct regression model (i.e. the model with the correct mean response structure) with each dataset. Thus, if the realizations of model (2.1) are generated, say, with an intercept and two slope coefficients, then the fitted model will contain an intercept and two slope coefficients. On the other hand, if the realizations are generated with no intercept and only one slope coefficient in model (2.1), then the fitted model will contain solely one slope regression parameter. We do not fit models that do not coincide with the true models to investigate the impact of factors such as incorrectly specified mean response structure. We only consider the impact on inference due to response dependence.

2.2 Literature Review

Numerous publications address multiple linear regression. Many of them, such as Weisberg, 1985, Johnson & Wichern, 1982, and Marascuilo & Levin, 1983, are undergraduate text books. Most discuss how to account for correlated responses in regression parameter estimation. However, they usually do not investigate the impact of ignoring such correlation.

Advanced books and articles have been written since the 1960's on the loss of efficiency in linear regression parameter estimation when response dependence is ignored. For example, Plackett, 1960, Rao, 1965, Anderson, 1971, and Martin, 1982 all mention that, when the responses are dependent, the matrix $\boldsymbol{A} - \boldsymbol{B}$ is always non-negative definite, $\boldsymbol{A}$ being the covariance matrix of the regression parameter estimate vector under the independent response model, $\boldsymbol{B}$ that under the model with a correctly specified response correlation matrix. (Also mentioned is the unbiasedness of the parameter estimates despite ignoring response dependence.) In other words, admitting the known covariance of the responses in regression analysis yields parameter estimates which are *at least as efficient* as those produced by ignoring response dependence.

A natural question arises: "Under what conditions can response dependence be ignored without loss of efficiency?" Some of these conditions are stated in Plackett, 1960 (as an end-of-chapter exercise); Rao, 1965; Zyskind, 1967; McElroy, 1967; Anderson, 1971; Seber, 1977; Martin, 1982; and Cressie, 1991. Some of these articles supply necessary and sufficient conditions for the equivalence of $\boldsymbol{A}$ and $\boldsymbol{B}$. One of these conditions, less technical than the others, is when the responses have an *exchangeable* correlation structure. This is explained in Section 2.5.

However, none of these articles mentions the fact that, when an investigator overlooks the response dependence, s/he naturally does statistical inference based on parameter estimates whose standard errors are calculated as if the responses were truly independent. We then have three covariance matrices for two parameter estimate vectors:

1. the (true) covariance matrix of the *correct* parameter estimates (considering the known response dependence);

2. the *true* covariance matrix of the *incorrect* parameter estimates (obtained while ignoring response dependence); and

3. the *incorrect* covariance matrix of the *incorrect* parameter estimates (both obtained assuming the responses are uncorrelated).

We can think of the first covariance matrix as that computed by the "enlightened" statistician, who knows the true covariance matrix of the responses and obtains his/her parameter estimates and their covariances taking account of this knowledge. The third covariance matrix is that computed by the "naive" statistician, who does not know the responses are correlated, and thus obtains his/her parameter estimates, as well as their covariances, without accounting for response dependence. Finally, the second covariance matrix is that computed by Enlightened *for* Naive: Enlightened knows Naive's covariance matrix is incorrect, and does Naive a favor by computing a correct one for him/her.

For reasons given in Section 2.3, we refer to the least squares parameter vector estimate assuming response dependence as the **GLS estimator**. We call the one obtained ignoring response dependence the **OLS estimator**. The first covariance in the above list is thus the covariance matrix of the GLS estimator. Let us call it the **GLS covariance**. We call the true covariance matrix of the OLS estimator the **true OLS covariance**. This is the one given in the second item of the above list. Finally, we call the third covariance matrix in the list the **naive OLS covariance** for obvious reasons.

In this chapter, we will investigate the loss of efficiency resulting from using the OLS instead of the GLS estimator, by essentially comparing the GLS covariance to the true OLS covariance, as well as the GLS covariance to the naive OLS covariance, with different covariate (design) matrices and regression parameter values.[3] In this way, we hope to understand the loss of efficiency in parameter estimation when ignoring response dependence. At the same time, we also try to understand what the effects on parameter estimation are when Naive bases his/her inference on his/her naive parameter estimates and covariances. We will base our investigation on computer simulations implemented in the statistical software S-Plus.

Later in the chapter, we address the question of how to model spatial correlation. A few such models are presented in Cressie, 1991, and Basu and Reinsel, 1994.

[3] For reasons given in Section 2.7, we calculate p-values (for hypothesis tests) based on these covariances. We then compare the p-values (hence, functions of the covariances) instead of directly comparing the covariance matrices.

2.3 Notation

Suppose the spatial domain of interest consists of r regions. In an investigation, we observe in region i the response (dependent) variable Y_i, $i = 1, 2, \ldots, r$. We then decide to include in our regression model, say, g covariates (explanatory variables) $x_{i1}, \ldots, x_{ig}$ for each region i, together with an intercept β_0 and a slope β_j for the jth covariate, $j = 1, 2, \ldots, g$. In matrix notation, we may then write our model as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_r \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1g} \\ 1 & x_{21} & x_{22} & \ldots & x_{2g} \\ & & \vdots & & \\ 1 & x_{r1} & x_{r2} & \ldots & x_{rg} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_g \end{pmatrix}, \text{and } \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_r \end{pmatrix}.$$

(The intercept β_0 and the column of 1's in the design matrix, $\boldsymbol{X}$, are of course omitted if we decide not to fit an intercept in the above model.) It is often convenient to assume that $\boldsymbol{X}$ is of full rank, and that $\text{rank}(\boldsymbol{X}) = g + 1 < r$. We also assume that the expected value and the correlation matrix of the error vector $\boldsymbol{\epsilon}$ are respectively $\text{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\text{Cor}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_\epsilon$, where $\boldsymbol{\Sigma}_\epsilon$ is symmetric and positive definite. Assuming that the variance of each ϵ_i is σ^2, the covariance matrix of $\boldsymbol{\epsilon}$ is then $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{\Sigma}_\epsilon$.

A simple case of (2.1) obtains when $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$, the identity matrix. In this case, the *best linear unbiased estimator* (BLUE) of the parameter $\boldsymbol{\beta}$ is simply the *ordinary least squares* (OLS) estimator

$$\widehat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{2.2}$$

with covariance matrix

$$\text{Cov}(\widehat{\boldsymbol{\beta}}^{OLS}) = \sigma^2\ (\boldsymbol{X}^T\boldsymbol{X})^{-1}. \tag{2.3}$$

With a known $\boldsymbol{\Sigma}_\epsilon \neq \boldsymbol{I}$, however, the BLUE for $\boldsymbol{\beta}$ is the *generalized least squares* (GLS) estimator

$$\widehat{\boldsymbol{\beta}}^{GLS} = (\boldsymbol{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{Y} \tag{2.4}$$

with covariance matrix

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}^{GLS}) = \sigma^2 \, (\boldsymbol{X}^T \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{X})^{-1}. \tag{2.5}$$

(See Seber, 1977, page 49 for derivation of the BLUE's.) The covariance formulas for these BLUE's are derived in standard textbooks on regression theory.

Recall the notion of the naive and enlightened statisticians. In the case $\boldsymbol{\Sigma}_\epsilon \neq \boldsymbol{I}$, Enlightened uses the estimator in (2.4) to estimate the regression parameters. S/he uses the formula in (2.5) to calculate $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}^{GLS})$. On the other hand, since Naive assumes $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$ in all situations, s/he uses the estimator in (2.2) and the covariance formula in (2.3) in his/her analyses.

However, if the responses are truly *not* independent, i.e. $\boldsymbol{\Sigma}_\epsilon \neq \boldsymbol{I}$, then the true OLS covariance is actually

$$\begin{aligned} \mathrm{Cov}(\widehat{\boldsymbol{\beta}}^{OLS}) &= \mathrm{Cov}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \, \mathrm{Cov}(\boldsymbol{Y}) \, [(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \end{aligned} \tag{2.6}$$

Thus, we see that, in general, the covariance matrices in (2.3), (2.5), and (2.6) will differ.

2.4 Which estimator is appropriate?

In principle, when the responses are known to be correlated, we should always account for their correlation in fitting the model. However, sometimes $\boldsymbol{\Sigma}_\epsilon$ may be difficult to determine (or estimate). Other times, it is believed that the correlation among the responses is too weak to make much difference in the inference. Also, we can easily see from (2.2) and (2.4) that both $\widehat{\boldsymbol{\beta}}^{OLS}$ and $\widehat{\boldsymbol{\beta}}^{GLS}$ are unbiased for $\boldsymbol{\beta}$ regardless of whether $\boldsymbol{\Sigma}_\epsilon = \text{Cor}(\boldsymbol{\epsilon}) = \text{Cor}(\boldsymbol{Y})$ is the identity matrix. In fact, the expected value of $\widehat{\boldsymbol{\beta}}^{OLS}$ does not depend on the value of $\boldsymbol{\Sigma}_\epsilon$. For reasons such as these, some investigators may ignore the presence of response correlation.

However, the following theorem indicates a potential danger in using $\widehat{\boldsymbol{\beta}}^{OLS}$ to estimate $\boldsymbol{\beta}$ when $\boldsymbol{\Sigma}_\epsilon \neq \boldsymbol{I}$.

Theorem 1 *Cov($\widehat{\boldsymbol{\beta}}^{OLS}$)− Cov($\widehat{\boldsymbol{\beta}}^{GLS}$) is always non-negative definite.*

Rao, 1965 presents a proof of this theorem.

Here, we are referring to the true OLS covariance as Cov($\widehat{\boldsymbol{\beta}}^{OLS}$). In other words, although both $\widehat{\boldsymbol{\beta}}^{OLS}$ and $\widehat{\boldsymbol{\beta}}^{GLS}$ are unbiased for $\boldsymbol{\beta}$, Theorem 1 states that the former is never "better" than the latter in that there is at least as much variability in $\widehat{\boldsymbol{\beta}}^{OLS}$ as in $\widehat{\boldsymbol{\beta}}^{GLS}$. When the matrix (in Theorem 1) is positive definite, the parameter estimation based on $\widehat{\boldsymbol{\beta}}^{OLS}$ is lower in efficiency (with the true OLS covariance) than that based on $\widehat{\boldsymbol{\beta}}^{GLS}$. We will show some simulation results later in the chapter to further illustrate and quantify this point.

When is the matrix in Theorem 1 a zero matrix, then? That is, when are these two covariances[4] equal?

Two obvious conditions for the two covariances to be equal are:

1. $\widehat{\boldsymbol{\beta}}^{OLS} = \widehat{\boldsymbol{\beta}}^{GLS}$, and in particular,

2. $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$.

The second condition is not of much interest to us. We are interested in situations where $\widehat{\boldsymbol{\beta}}^{OLS}$ is as efficient as $\widehat{\boldsymbol{\beta}}^{GLS}$ while the responses Y_i's are *not* independent.

[4] GLS covariance and true OLS covariance.

Necessary and sufficient conditions for $\widehat{\boldsymbol{\beta}}^{OLS} = \widehat{\boldsymbol{\beta}}^{GLS}$ are presented in Rao, 1967; Zyskind, 1967; and McElroy, 1967. However, most of these conditions are very technical and difficult to verify. Thus, we believe investigators should always use $\widehat{\boldsymbol{\beta}}^{GLS}$ to estimate $\boldsymbol{\beta}$ whenever the responses are known to be correlated.

However, one realistic exception to the need to use $\widehat{\boldsymbol{\beta}}^{GLS}$[5] stems from the *exchangeable structure*, also known as the *intraclass correlation structure*, often met in analyses of repeated measurements, and sometimes in spatially correlated data. With such structure, fitting an intercept in model (2.1) will entail the actual equality of $\widehat{\boldsymbol{\beta}}^{OLS}$ and $\widehat{\boldsymbol{\beta}}^{GLS}$!

2.5 The Exchangeable Model

An exchangeable model for correlation structure assumes that the correlation matrix of the errors is

$$\mathrm{Cor}(\epsilon) = \boldsymbol{\Sigma}_\epsilon = \begin{pmatrix} 1 & \rho & \rho & \cdots & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \cdots & \rho \\ \rho & \rho & 1 & \rho & \cdots & \rho \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \rho & \cdots & \cdots & \rho & 1 & \rho \\ \rho & \rho & \cdots & \cdots & \rho & 1 \end{pmatrix} = (1-\rho)\boldsymbol{I} + \rho\boldsymbol{J} \qquad (2.7)$$

where $\boldsymbol{J}$ is a matrix with all elements equal to 1, and $0 \leq \rho < 1$. This last inequality guarantees that $\boldsymbol{\Sigma}_\epsilon$ is positive definite for any r. For a given r, we can replace this inequality by $[-1/(r-1)] < \rho < 1$.[6]

Theorem 2 *Assume that we fit an intercept in the regression model (2.1). Then*

$$\widehat{\boldsymbol{\beta}}^{OLS} = \widehat{\boldsymbol{\beta}}^{GLS} \textit{ for all } \boldsymbol{Y} \iff \boldsymbol{\Sigma}_\epsilon \textit{ has the form}$$

[5] However, this does not mean that we can use the naive OLS covariance in place of the GLS covariance.

[6] See Graybill, 1983, page 215.

$$\boldsymbol{\Sigma}_\epsilon = (1-\rho)\boldsymbol{I} + \rho\boldsymbol{J}.$$

To prove Theorem 2, we need the following lemma.

Lemma 1 *Let the matrix* $\boldsymbol{W}$ *have the form*

$$\boldsymbol{W} = a\boldsymbol{I} + b\boldsymbol{P}$$

where $\boldsymbol{P}$ *is an* $r \times r$ *matrix such that* $\boldsymbol{P}^2 = t\boldsymbol{P}$, *with* $t > 0$. *Then*

$$\boldsymbol{W}^{-1} = \tfrac{1}{a}\,\boldsymbol{I} - \tfrac{b}{a(a+bt)}\,\boldsymbol{P}\ .$$

Proof

$$\begin{aligned}\boldsymbol{W}^{-1}\boldsymbol{W} &= (\tfrac{1}{a}\,\boldsymbol{I} - \tfrac{b}{a(a+bt)}\,\boldsymbol{P})\,(a\boldsymbol{I} + b\boldsymbol{P}) \\ &= \boldsymbol{I} + \tfrac{b}{a}\boldsymbol{P} - \tfrac{b}{a+bt}\boldsymbol{P} - \tfrac{b^2t}{a(a+bt)}\boldsymbol{P} \\ &= \boldsymbol{I} + \tfrac{b}{a(a+bt)}[(a+bt) - a - bt]\boldsymbol{P} \\ &= \boldsymbol{I}\end{aligned}$$

Similarly, $\boldsymbol{W}\boldsymbol{W}^{-1} = \boldsymbol{I}$, and hence the result.

Proof of Theorem 2[7]

Note that

$$\begin{aligned} & \widehat{\boldsymbol{\beta}}^{OLS} &= \widehat{\boldsymbol{\beta}}^{GLS} \\ \Longleftrightarrow\quad & (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} &= (\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{Y} \\ \Longleftrightarrow\quad & (\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} &= \boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{Y}. \qquad (*)\end{aligned}$$

[7] This proof is an elaboration of the one by McElroy (1967).

Therefore, proving Theorem 2 is equivalent to proving $(*)$ holds for all $\boldsymbol{Y}$ if and only if $\boldsymbol{\Sigma}_\epsilon = (1-\rho)\boldsymbol{I} + \rho\boldsymbol{J}$.

Now, $(*)$ automatically holds for all $\boldsymbol{\Sigma}_\epsilon^{-1}$ if $\boldsymbol{Y}$ is in the range space of $\boldsymbol{X}$, i.e. if $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{Z}$ for some $\boldsymbol{Z}$. This is so because then, in $(*)$,

$$\begin{aligned} \text{LHS} &= (\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{Z} \\ &= \boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{X}\boldsymbol{Z}) \\ &= \boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{Y} \\ &= \text{RHS}. \end{aligned}$$

Therefore, we now need only to consider those $\boldsymbol{Y}$ in the null space of $\boldsymbol{X}$, i.e. those $\boldsymbol{Y}$ such that $\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{0}$.

$(\Longleftarrow)$ Assume $\boldsymbol{\Sigma}_\epsilon = (1-\rho)\boldsymbol{I} + \rho\boldsymbol{J}$. Denote the column vector of 1's as $\boldsymbol{1}$. Then

$$\begin{aligned} \boldsymbol{J} &= \boldsymbol{1}\boldsymbol{1}^T \\ \Longrightarrow \boldsymbol{J}^2 &= \boldsymbol{1}(\boldsymbol{1}^T\boldsymbol{1})\boldsymbol{1}^T \\ &= r\boldsymbol{1}\boldsymbol{1}^T \\ &= r\boldsymbol{J}. \end{aligned}$$

Since $r > 0$, by Lemma 1, we have

$$\begin{aligned} \boldsymbol{\Sigma}_\epsilon^{-1} &= \tfrac{1}{1-\rho}\boldsymbol{I} - \tfrac{\rho}{(1-\rho)(1-\rho+\rho r)}\boldsymbol{J} \\ &= \tfrac{1}{1-\rho}\boldsymbol{I} - \tfrac{\rho}{(1-\rho)[1+(r-1)\rho]}\boldsymbol{J} \\ &= \tfrac{1}{1-\rho}\,[\boldsymbol{I} - \tfrac{\rho}{1+(r-1)\rho}\boldsymbol{1}\boldsymbol{1}^T]. \end{aligned}$$

Now, let $\boldsymbol{y}$ be in the null space of $\boldsymbol{X}$, i.e.

$$\boldsymbol{0} = \boldsymbol{X}^T\boldsymbol{y} = \begin{pmatrix} \boldsymbol{1}^T \\ \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_g^T \end{pmatrix} \boldsymbol{y}$$

where $\boldsymbol{x}_i^T = (x_{1i},\ x_{2i},\ \ldots,\ x_{ri})$. Then, $\mathbf{1}^T\boldsymbol{y} = 0$.

With $\boldsymbol{y}$ in place of $\boldsymbol{Y}$ in $(*)$, we now have

$$\begin{aligned} \text{LHS} &= (\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \mathbf{0} \qquad\qquad (\text{since } \boldsymbol{X}^T\boldsymbol{y} = \mathbf{0}) \end{aligned}$$

and

$$\begin{aligned} \text{RHS} &= \boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{y} \\ &= \tfrac{1}{1-\rho}\boldsymbol{X}^T\boldsymbol{y} - \tfrac{\rho}{1+(r-1)\rho}\boldsymbol{X}^T\mathbf{1}\mathbf{1}^T\boldsymbol{y} \\ &= \mathbf{0} - \mathbf{0} \qquad\qquad (\text{since } \boldsymbol{X}^T\boldsymbol{y} = \mathbf{0} \text{ and } \mathbf{1}^T\boldsymbol{y} = 0) \\ &= \text{LHS}. \end{aligned}$$

Thus, $(*)$ holds for all $\boldsymbol{y}$ such that $\boldsymbol{X}^T\boldsymbol{y} = \mathbf{0}$.
Hence, $\widehat{\boldsymbol{\beta}}^{OLS} = \widehat{\boldsymbol{\beta}}^{GLS}$.

$(\Longrightarrow)$ Suppose $\widehat{\boldsymbol{\beta}}^{OLS} = \widehat{\boldsymbol{\beta}}^{GLS}$. Therefore, $(*)$ holds for all $\boldsymbol{Y}$.
Let $\boldsymbol{y}$ be such that $\boldsymbol{X}^T\boldsymbol{y} = \mathbf{0}$. Hence, by $(*)$,

$$\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{y} = (\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \mathbf{0}. \tag{2.8}$$

Now, fix i, j, and k such that $1 \le i, j, k \le r$, and $i \ne j, k$.
Let $\boldsymbol{e}_i$ denote the elementary vector all of whose co-ordinates are 0's except the ith being 1. In turn, let $\boldsymbol{U}$ be an $r \times (g+1)$ matrix of the form

$$\boldsymbol{U} = (\mathbf{1},\ \boldsymbol{e}_i,\ \boldsymbol{v}_2,\ \ldots,\ \boldsymbol{v}_g)$$

where the set of distinct elementary vectors $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_g$ is chosen such that $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_g \ne \boldsymbol{e}_i$ or $\boldsymbol{e}_j$ or $\boldsymbol{e}_k$. (That is, the vectors $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_g$, together with any of $\boldsymbol{e}_i$, $\boldsymbol{e}_j$, or $\boldsymbol{e}_k$, form a basis.) Then, $\boldsymbol{U}$ has the specified form of the design matrix $\boldsymbol{X}$ in model (2.1), and we can substitute $\boldsymbol{U}$ for $\boldsymbol{X}$ in (2.8).

Now,

$$\begin{aligned} \boldsymbol{U}^T(\boldsymbol{e}_j - \boldsymbol{e}_k) &= \begin{pmatrix} \mathbf{1}^T \\ \boldsymbol{e}_i^T \\ \boldsymbol{v}_2^T \\ \vdots \\ \boldsymbol{v}_g^T \end{pmatrix} (\boldsymbol{e}_j - \boldsymbol{e}_k) \\ &= \boldsymbol{e}_1 - \boldsymbol{e}_1 \\ &= \mathbf{0}. \end{aligned}$$

This implies that the vector $(\boldsymbol{e}_j - \boldsymbol{e}_k)$ is in the null space of $\boldsymbol{U}$. Thus, by (2.8), we have

$$\boldsymbol{U}^T \boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{e}_j - \boldsymbol{e}_k) = \mathbf{0}. \tag{2.9}$$

Write the (i, j)th element of $\boldsymbol{\Sigma}_\epsilon^{-1}$ as $(\boldsymbol{\Sigma}_\epsilon^{-1})_{ij} = \sigma_{ij}^*$. Now, we can write

$$\boldsymbol{\Sigma}_\epsilon^{-1} = (\boldsymbol{\sigma}_1^*,\ \boldsymbol{\sigma}_2^*, \dots, \boldsymbol{\sigma}_r^*) \text{ where } \boldsymbol{\sigma}_l^* = \begin{pmatrix} \sigma_{1l}^* \\ \sigma_{2l}^* \\ \vdots \\ \sigma_{rl}^* \end{pmatrix}.$$

Then, the matrix

$$\boldsymbol{U}^T \boldsymbol{\Sigma}_\epsilon^{-1} = \begin{pmatrix} \mathbf{1}^T \\ \boldsymbol{e}_i^T \\ \boldsymbol{v}_2^T \\ \vdots \\ \boldsymbol{v}_g^T \end{pmatrix} (\boldsymbol{\sigma}_1^*\ \boldsymbol{\sigma}_2^* \dots \boldsymbol{\sigma}_r^*) = \begin{pmatrix} \mathbf{1}^T\boldsymbol{\sigma}_1^* & \dots & \mathbf{1}^T\boldsymbol{\sigma}_r^* \\ \boldsymbol{e}_i^T\boldsymbol{\sigma}_1^* & \dots & \boldsymbol{e}_i^T\boldsymbol{\sigma}_r^* \\ \boldsymbol{v}_2^T\boldsymbol{\sigma}_1^* & \dots & \boldsymbol{v}_2^T\boldsymbol{\sigma}_r^* \\ & \vdots & \\ \boldsymbol{v}_g^T\boldsymbol{\sigma}_1^* & \dots & \boldsymbol{v}_g^T\boldsymbol{\sigma}_r^* \end{pmatrix}$$

has its second row equal to

$$\begin{aligned}(\boldsymbol{e}_i^T\boldsymbol{\sigma}_1^*,\ \boldsymbol{e}_i^T\boldsymbol{\sigma}_2^*, \dots, \boldsymbol{e}_i^T\boldsymbol{\sigma}_r^*) &= (\sigma_{i1}^*,\ \sigma_{i2}^*,\ \dots,\ \sigma_{ir}^*)\\ &= i\text{th row of } \boldsymbol{\Sigma}_\epsilon^{-1}.\end{aligned}$$

Hence, equation (2.9) implies

$$\begin{aligned}(\sigma_{i1}^*,\ \sigma_{i2}^*,\ \dots,\ \sigma_{ir}^*)(\boldsymbol{e}_j - \boldsymbol{e}_k) &= 0\\ \Longrightarrow \qquad \sigma_{ij}^* - \sigma_{ik}^* &= 0.\end{aligned} \tag{2.10}$$

This is true for all $1 \le i, j, k \le r$ such that $i \ne j$ or k.

Thus, $\boldsymbol{\Sigma}_\epsilon^{-1}$ has all off-diagonal elements equal.

Equation (2.9) also implies

$$\begin{aligned}0 &= (\boldsymbol{1}^T\boldsymbol{\sigma}_1^*, \dots, \boldsymbol{1}^T\boldsymbol{\sigma}_r^*)(\boldsymbol{e}_j - \boldsymbol{e}_k)\\ &= \left(\textstyle\sum_{l=1}^r \sigma_{l1}^*,\ \dots,\ \sum_{l=1}^r \sigma_{lr}^*\right)(\boldsymbol{e}_j - \boldsymbol{e}_k)\\ &= \textstyle\sum_{l=1}^r \sigma_{lj}^* - \sum_{l=1}^r \sigma_{lk}^*\\ &= \textstyle\sum_{l=1}^r (\sigma_{lj}^* - \sigma_{lk}^*).\end{aligned} \tag{2.11}$$

Now, (2.10) and (2.11) together imply

$$\begin{aligned}0 &= \textstyle\sum_{l=1}^r (\sigma_{lj}^* - \sigma_{lk}^*)\\ &= (\sigma_{jj}^* - \sigma_{jk}^*) + (\sigma_{kj}^* - \sigma_{kk}^*)\\ &= \sigma_{jj}^* - \sigma_{kk}^*\end{aligned}$$

(since $\boldsymbol{\Sigma}_\epsilon$, and hence $\boldsymbol{\Sigma}_\epsilon^{-1}$, is symmetric).

Finally, we have

$$\sigma_{jj}^* = \sigma_{kk}^* \ \text{ for all } \ 1 \le j, k \le r.$$

Hence, $\boldsymbol{\Sigma}_\epsilon^{-1}$ has all diagonal elements equal.

Note that Lemma 1 implies that $\boldsymbol{\Sigma}_\epsilon$ must also have its off-diagonal elements equal, and its diagonal elements equal. But $\boldsymbol{\Sigma}_\epsilon$ is positive definite, and thus it has positive diagonal elements. In addition, a correlation matrix has the property

$$\text{trace}\ (\boldsymbol{\Sigma}_\epsilon) = r.$$

Therefore, the diagonal elements of $\boldsymbol{\Sigma}_\epsilon$ must all equal 1.

Hence, $\boldsymbol{\Sigma}_\epsilon$ is of the form $\boldsymbol{\Sigma}_\epsilon = (1-\rho)\boldsymbol{I} + \rho\boldsymbol{J}$ for some $\rho \in [0,1)$, as was to be proved.

Q.E.D.

Although $\widehat{\boldsymbol{\beta}}^{OLS} = \widehat{\boldsymbol{\beta}}^{GLS}$ when the responses have an exchangeable correlation structure and when an intercept exists in model (2.1), we need to be very cautious in computing its covariance in such situations.

Recall that we have the GLS covariance, the true OLS covariance, and the naive OLS covariance of the corresponding $\boldsymbol{\beta}$ estimates. Theorem 2 says that, since $\widehat{\boldsymbol{\beta}}^{OLS} = \widehat{\boldsymbol{\beta}}^{GLS}$, their covariance matrices are the same. This is to say that the *true* OLS covariance equals the GLS covariance. This is *not* to say that the *naive* OLS covariance and the GLS covariance are the same, *unless* $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$. When $\boldsymbol{\Sigma}_\epsilon \neq \boldsymbol{I}$, the naive OLS covariance can be considerably different from the true OLS covariance (which is the same as the GLS covariance). The loss of efficiency due to the use of the incorrect covariance formula will be illustrated in Section 2.7.3.

Thus, although the $\boldsymbol{\beta}$ estimates obtained by both Naive and Enlightened may be the same, the covariances that they compute for the estimates may be very different, regardless of what the response correlation structure may be. In other words, Theorem 2 only allows Enlightened to use $\widehat{\boldsymbol{\beta}}^{OLS}$ to estimate $\boldsymbol{\beta}$ for simplicity, but does not allow either to ignore $\boldsymbol{\Sigma}_\epsilon$ in the covariance formula for their estimates.

2.6 Common Spatial Correlation Structures

We have now seen that with an exchangeable response correlation structure, we may use either the OLS or the GLS estimator to estimate the regression coefficients in model (2.1) without affecting estimation efficiency, provided that (1) the model includes an intercept and (2) $\mathrm{Cov}(\widehat{\beta}^{OLS})$ is taken to be the true OLS covariance in (2.6). However, when we have a different response correlation structure, using the OLS estimator may result in significant loss of efficiency in parameter estimation, as will be shown later in the chapter.

So what response correlation structures are seen in spatial statistics?

Let us first assume that the regions of interest have a "linear spatial arrangement," as do the major cities in the Pacific Northwest. Then the spatial correlation in the responses may resemble the common autocorrelation structures in repeated measurement or time series analyses.[8] Some such structures are the *exchangeable* and the *AR(1) (autoregressive process with lag 1)* structures. The former is useful when we believe that observations from regions either nearby or far apart are correlated to the same extent. The latter is used in cases where the strength of correlation declines with increasing geographical separation.

However, regions of interest are usually spread out in a two-dimensional geographical domain. The natural measure of how far apart two regions are is then the physical distance between them. This distance is often measured from the centroid of one region to that of the other.[9] For each centroid, we record its *bearing*, i.e. its latitude and longitude. The distance between two centroids can then be calculated from their bearings (see equation (2.12)).

It is often believed that correlation between the observations from two distinct regions

[8] See Cressie, 1991, page 340.

[9] See Rosychuk, 1994.

declines with increasing separation.[10] This belief suggests that the *variogram*,[11] $\boldsymbol{\Sigma}_\epsilon$, be modeled to have a form like:[12]

$$(\boldsymbol{\Sigma}_\epsilon)_{ij} = \text{the } (i,j)\text{th element of } \boldsymbol{\Sigma}_\epsilon = \exp\{-||\boldsymbol{z}_i - \boldsymbol{z}_j||^2\} \tag{2.12}$$

where

$$\boldsymbol{z}_i = \begin{pmatrix} \text{latitude of region } i \\ \text{longitude of region } i \end{pmatrix}.$$

Note that $||\boldsymbol{z}_i - \boldsymbol{z}_j||^2$ is a measure of the distance between regions i and j. In other words, the correlation between two regions decreases exponentially as their separation increases. This structure is widely used in spatial statistical models.

Another way to model spatial correlation in a two-dimensional layout is presented in Basu & Reinsel, 1994. This type of model requires that the regions be arranged in a two-dimensional grid. The observation from each region is thus assigned a co-ordinate, (i, j). The correlation is introduced by a *unilateral first order ARMA* error process that depends on the co-ordinates. The drawback of using such a model is the difficulty in estimating the extra parameters (in the ARMA process) intrinsic in the regression model.

In our computer simulations of linear regressions, we used both standard and non-standard correlation structures. The standard ones include the exchangeable, AR(1), and common variogram structures. We did not use the model presented by Basu & Reinsel (1994) for the reasons mentioned above. The non-standard correlation structures include some which resemble the AR(1) structure. All of these will be introduced in the next section.

[10] See Cressie, 1991, page 24.

[11] That is, the covariance function of spatially correlated observations. See Cressie, 1991, page 40 for a detailed definition.

[12] See Cressie, 1991.

2.7 The Simulation Experiments

Our objective is mainly to investigate how Naive's inferences compare with those of Enlightened.

First, we specified the **model parameters:** $\boldsymbol{\beta}$ (the vector of **regression parameters**), $\boldsymbol{\Sigma}_\epsilon$ (the correlation matrix of the errors, and hence, of the response data), and σ^2 (the noise level, or the "scale" of the covariances of the errors/responses). Next, we fixed an arbitrary covariate (design) matrix, $\boldsymbol{X}$. We then generate error data, ϵ_i's, from a multivariate normal (MVN) distribution with zero mean vector and covariance matrix equal to $\sigma^2\boldsymbol{\Sigma}_\epsilon$ (σ^2 and $\boldsymbol{\Sigma}_\epsilon$ specified separately). Finally, $\boldsymbol{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ were "put together" to give correlated response data, $\boldsymbol{Y}$, according to model (2.1).

With this dataset $\boldsymbol{Y}$, Naive and Enlightened fitted the linear regression model. Knowing the true regression parameters $\boldsymbol{\beta}$, we can judge how well these two statisticians have estimated $\boldsymbol{\beta}$. We are concerned in particular with how much efficiency in parameter estimation is lost in Naive's analysis.

Recall that Naive believes the response data to be independent, and hence computes a naive covariance matrix of $\widehat{\boldsymbol{\beta}}^{OLS}$, according to formula (2.3). Let us denote it by $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}^{OLS})_{\mathbf{naive}}$. Enlightened knows the true $\boldsymbol{\Sigma}_\epsilon$, and hence computes the correct covariance matrix of $\widehat{\boldsymbol{\beta}}^{GLS}$, according to (2.5). We denote this by $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}^{GLS})$. Finally, Enlightened realizes Naive's mistake and computes the true covariance matrix of $\widehat{\boldsymbol{\beta}}^{OLS}$ for Naive, according to (2.6). We denote it by $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}^{OLS})_{\mathbf{true}}$.

In each simulation experiment, we fixed one set of model parameters with a covariate matrix, and carried out n repetitions of the process of:

1. generating $\boldsymbol{\epsilon}$ and calculating $\boldsymbol{Y}$;

2. fitting Naive's regression model to obtain $\widehat{\boldsymbol{\beta}}^{OLS}$, then estimating σ^2 from this

naively fitted model, obtaining $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}^{OLS})_{\mathrm{naive}}$ and $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}^{OLS})_{\mathrm{true}}$ using the estimated (OLS) σ^2, but using the true $\boldsymbol{\Sigma}_\epsilon$;[13]

3. fitting Enlightened's regression model to obtain $\widehat{\boldsymbol{\beta}}^{GLS}$, then estimating σ^2 from this fitted model, obtaining $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}^{GLS})$ using the estimated (GLS) σ^2, but using the true $\boldsymbol{\Sigma}_\epsilon$;

4. obtaining t-test statistics based on each of these three covariance matrices for testing the hypotheses

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0$$

at significance level 0.05, for all $j = 1, 2, \ldots, g$; and

5. calculating the corresponding p-values of the above t-test statistics.

The number of covariates g was usually set to anywhere from 1 to 30. After n such sets of p-values were obtained, we compared two pairs of analyses. We first compared Enlightened's analysis to that of Naive. We then compared Enlightened's analysis to that s/he did for Naive. The comparisons were done via some *p-value comparisons.* For each regression coefficient, we put the corresponding p-values of the two pairs of analyses into a contingency table such as Table 2.1, to compare the **consistency** of hypothesis rejection (number of times both methods rejected H_0) and the **empirical power** (number of times H_0 was rejected) of the two methods. "$\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$" in the second column of Table 2.1 means that the p-values were computed based on $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}^{OLS})_{\mathrm{naive}}$, "$\boldsymbol{\Sigma}_\epsilon$ known" that the p-values were computed based on $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}^{OLS})_{\mathrm{true}}$.

[13] That is, when fitting a regression model, we assume the response correlation matrix, $\boldsymbol{\Sigma}_\epsilon$, is given, but the "scale" σ is not. We adopted our assumption from Weisberg, 1985, page 40.

			GLS	
			β_1	
			$p < 0.05$	$p \geq 0.05$
OLS	$\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$	$p < 0.05$		
		$p \geq 0.05$		
	$\boldsymbol{\Sigma}_\epsilon$ known	$p < 0.05$		
		$p \geq 0.05$		

Table 2.1: An example of a p-value contingency table for β_1.

2.7.1 Generating Correlated Responses

The (model) parameters in model (2.1) include σ^2, $\boldsymbol{\Sigma}_\epsilon$, and $\boldsymbol{\beta}$. In a simulation experiment, we need to specify all of these parameters, together with the covariate matrix $\boldsymbol{X}$, in order to obtain the response vector $\boldsymbol{Y}$.

In one simulation study of n realizations of model (2.1), we would first specify $\boldsymbol{X}$, the design/covariate matrix, and the regression coefficient vector, $\boldsymbol{\beta}$. Instead of arbitrarily "inventing" these quantities everytime we ran a simulation experiment, we decided to have a computer implemented procedure which could be used to generate both $\boldsymbol{X}$ and $\boldsymbol{\beta}$. In this way, we would only need to specify a limited number of program arguments instead of each of the $r \times g$ covariate values and $g + 1$ coefficient values.

The procedure for obtaining the design matrix and the regression coefficients is presented in Section 2.10, the CHAPTER APPENDIX. This procedure requires the specification of $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$, the common expectation and covariance matrix of the g covariate vectors. These were truly arbitrarily chosen. For example, we sometimes took $\boldsymbol{\mu}_X = \boldsymbol{0}$, and other times $\boldsymbol{\mu}_X = (1, 2, 3)^T$ for a model with an intercept and two slope coefficients. $\boldsymbol{\Sigma}_X$ was taken to be the identity in some cases. We also tried $\boldsymbol{\Sigma}_X$'s which have exchangeable structures.

Once $\boldsymbol{X}$ and $\boldsymbol{\beta}$ were generated, they were used to generate all the n $\boldsymbol{Y}$'s from model

(2.1). In fact, we did not generate a new $\boldsymbol{X}$ and $\boldsymbol{\beta}$ in every simulation. Once generated, the same $\boldsymbol{X}$ and $\boldsymbol{\beta}$ were sometimes re-used in other simulations, varying only $\boldsymbol{\Sigma}_\epsilon$ and σ^2 of the distribution of the random errors in the model.

In other simulations, we fixed the same 30 × 2 covariate matrix for those cases in which an intercept along with one slope coefficient was present in model (2.1). The first column of this $\boldsymbol{X}$ is simply the $\mathbf{1}$ vector. The second column was generated from a normal distribution with mean 10 and standard deviation 1/2.[14] The first five elements of this second column are 9.73, 9.61, 9.33, 10.62, and 10.57.

Next, we needed to generate the random errors in the model by specifying σ^2 and $\boldsymbol{\Sigma}_\epsilon$. The matrix $\boldsymbol{\Sigma}_\epsilon$ was chosen from the following:

1. an AR(1) correlation matrix with a lag one correlation of 0.6 or 0.9;

2. a correlation matrix (which resembles the AR(1) structure) that has its (i, j)th element equal to

   (a) 1 if $i = j$,

   (b) 0.5 if $|i - j| = 1$,

   (c) 0.2 if $|i - j| = 2$,

   (d) 0.1 if $|i - j| = 3$, and

   (e) 0 if $|i - j| > 3$,

   for $i, j = 1, 2, \ldots, r$;

3. a correlation matrix, similar to the above, with its (i, j)th element equal to

   (a) 1 if $i = j$,

[14] Again, the choice of the moments was arbitrary.

(b) 0.8 if $|i-j|=1$,

(c) 0.45 if $|i-j|=2$,

(d) 0.15 if $|i-j|=3$, and

(e) 0 if $|i-j|>3$,

for $i,j=1,2,\ldots,r$;

4. an exchangeable correlation matrix with an intraclass correlation coefficient of 0.6 or 0.9; and

5. the variogram computed according to equation (2.12) with

$$\boldsymbol{Z}=\begin{pmatrix}\boldsymbol{z}_1^T\\ \boldsymbol{z}_2^T\\ \vdots\\ \boldsymbol{z}_r^T\end{pmatrix}$$

whose columns were generated independently from a Uniform(25, 50) and a Uniform(100, 110) distribution respectively.[15]

Define $\boldsymbol{\epsilon}^{(s)}$ to be the sth sample of errors, $s=1,2,\ldots,n$, generated from

$$\boldsymbol{\epsilon}^{(s)}\sim \text{MVN}(\mathbf{0},\sigma^2\boldsymbol{\Sigma}_\epsilon). \tag{2.13}$$

After generating these n (independent) samples of $\boldsymbol{\epsilon}^{(s)}$'s, we would use the $\boldsymbol{X}$ matrix and the $\boldsymbol{\beta}$ vector (same for all s) obtained using our algorithm mentioned earlier, and compute

$$\boldsymbol{Y}^{(s)}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}^{(s)}. \tag{2.14}$$

[15] Once generated, the $\boldsymbol{Z}$ matrix (and hence the variogram) was fixed throughout those simulations in which the variogram was used.

Thus, these $\boldsymbol{Y}^{(s)}$'s are i.i.d. MVN, with mean $\boldsymbol{X\beta}$ and covariance $\boldsymbol{\Sigma}_\epsilon$.

Finally, we provided Naive and Enlightened with the n observed response vectors and the set of covariates. (Again, Naive believes that the responses are independent, whereas Enlightened knows the value of $\boldsymbol{\Sigma}_\epsilon$.) They returned their n sets of $\boldsymbol{\beta}$ estimates with their corresponding p-values. We then obtained the sample mean and variance of each regression coefficient estimate as summaries of the simulation experiment. Their p-values were then tabulated into a contingency table such as Table 2.1 for comparison of their estimation efficiencies.

2.7.2 Where are Naive and Enlightened?

We brought these two statisticians to life (figuratively speaking) by a module in a computer program which we implemented in the statistical software S-Plus. Other modules in the program include one for generating the n $\boldsymbol{\epsilon}^{(s)}$'s, and another for tabulating the p-values that the statisticians have computed. An independent module was implemented to generate covariates and regression coefficients. The program user has to specify $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$ in this module if s/he wishes to generate the covariates and regression coefficients. These values can be stored and re-used in other simulations to save computational time.

The user then has to specify a covariate matrix and a vector of regression coefficients, one or both of which may be generated by the above module, or fixed by the user. In addition, s/he has to specify σ and $\boldsymbol{\Sigma}_\epsilon$ in the main program. This program will first compute the responses, $\boldsymbol{Y}$, and then "call on" the two statisticians to obtain p-values from them. Next, the main program will hand these p-values over to the "output department" to produce the p-value contingency tables.

Terminology

In the rest of this chapter, we will refer to Enlightened's analysis as the **GLS analysis**, and to Naive's as the **naive OLS analysis**. Finally, the analysis done by Enlightened by using the OLS estimator and its true OLS covariance will be referred to as the **true OLS analysis**.

When examining the p-value contingency table for a particular regression coefficient, we will compare the off-diagonal values. For example, we will refer to the ratio of the (2,1) element to the (1,2) element in the contingency table for the GLS analysis to the naive OLS analysis as the **consistency ratio with respect to (wrt) the naive OLS analysis**. That is, a 10-to-1 consistency ratio wrt the OLS analysis means that there are 10 datasets of a simulation in which the GLS analysis detected a non-zero regression coefficient (at significance level 0.05) when the naive OLS analysis did not; whereas there is only one dataset in which the naive OLS analysis detected the coefficient, but the GLS analysis did not. In other words, the larger this consistency ratio, the more efficient the GLS analysis relative to the naive OLS analysis (provided that the detected coefficient is truly non-zero). Similarly, the corresponding consistency ratio with respect to the true OLS analysis will be the **consistency ratio wrt the true OLS analysis**.

By "**detecting a non-zero regression coefficient**," we mean that such a coefficient exists in the regression model and is detected (by the particular analysis).

2.7.3 Simulation Results

(i) Simulations 1 and 2

Simulations 1 and 2 were somewhat small scale studies with few correlated responses ($r = 4$) and only $n = 200$ realizations of model (2.1). The small r and n allowed us to

Program input variables:

r	g	n	$\boldsymbol{\mu}_X$	$\boldsymbol{\Sigma}_X$	σ^2	$\boldsymbol{\Sigma}_\epsilon$	fit intercept?
4	2	200	$(1,2,3)^T$	$\boldsymbol{I}$	1	AR(1) lag 1 correlation = 0.6	yes

Program output:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1.36 & 3.33 \\ 1 & 1.66 & 4.82 \\ 1 & 1.67 & 4.86 \\ 1 & 0.81 & 4.94 \end{pmatrix}$$

$\boldsymbol{\beta} = (1, 0, 0)^T$ (fixed; see page 32)

	OLS estimates			GLS estimates		
	β_0	β_1	β_2	β_0	β_1	β_2
sample mean of n estimates	0.67	0.05	0.06	0.69	0.06	0.05
sample variance of n estimates	8.54	1.10	0.41	7.85	1.06	0.35
true variance of each estimate	8.02	1.06	0.38	7.59	0.98	0.33

		GLS					
		β_0		β_1		β_2	
		$p < 0.05$	$p \geq 0.05$	$p < 0.05$	$p \geq 0.05$	$p < 0.05$	$p \geq 0.05$
OLS $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$	$p < 0.05$	11	0	12	0	11	0
	$p \geq 0.05$	2	187	3	185	1	188
OLS $\boldsymbol{\Sigma}_\epsilon$ known	$p < 0.05$	13	1	15	0	12	0
	$p \geq 0.05$	0	186	0	185	0	188

Table 2.2: Results of Simulation 1, with AR(1) response dependence

Program input variables:

r	g	n	$\boldsymbol{\mu}_X$	$\boldsymbol{\Sigma}_X$	σ^2	$\boldsymbol{\Sigma}_\epsilon$	fit intercept?
4	2	200	$(1,2,3)^T$	exchangeable, intraclass $\rho = 0.6$	1	AR(1), lag 1 correlation = 0.6	yes

Program output:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1.44 & 2.80 \\ 1 & 1.52 & 3.09 \\ 1 & 3.86 & 4.59 \\ 1 & 2.14 & 2.82 \end{pmatrix}, \boldsymbol{\beta} = (-0.875, 0.375, 0.375)^T$$

	OLS estimates			GLS estimates		
	β_0	β_1	β_2	β_0	β_1	β_2
sample mean of n estimates	-0.544	0.563	0.137	-0.602	0.494	0.197
sample variance of n estimates	6.02	1.70	2.49	5.73	1.47	2.27
true variance of each estimate	6.36	1.57	2.33	6.24	1.40	2.20

		GLS					
		β_0		β_1		β_2	
		$p < 0.05$	$p \geq 0.05$	$p < 0.05$	$p \geq 0.05$	$p < 0.05$	$p \geq 0.05$
OLS $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$	$p < 0.05$	12	0	14	0	12	0
	$p \geq 0.05$	2	186	0	186	3	185
OLS $\boldsymbol{\Sigma}_\epsilon$ known	$p < 0.05$	14	3	14	2	15	2
	$p \geq 0.05$	0	183	0	184	0	183

Table 2.3: Results of Simulation 2, with AR(1) response dependence

determine the approximate computational time needed for larger scale computer simulations. However, technicalities such as the computational efficiency of our simulation program and the capacity of the computer system on which the simulation experiments were run are beyond the scope of this thesis.

Simulations 1 and 2 were run mostly with the same program input variables. One difference lay in Σ_X, the covariance matrix for generating the covariates and the regression coefficients. However, only in Simulation 1 did we discard the regression coefficients generated by the corresponding module and replace it with the vector $(1, 0, 0)^T$. We did this to examine how often the GLS or OLS methods falsely detected non-zero coefficients.

Results of Simulation 1 are shown in Table 2.2. Those of Simulation 2 are shown in Table 2.3.

First, let us examine the results of Simulation 1.

We notice that the true variance of each OLS regression coefficient estimate is larger than that of the corresponding GLS estimate. This illustrates the result presented in Theorem 1. However, the covariate matrix generated from the program yielded very similar OLS and GLS estimates, and hence, similar (true and sample) variances.

Notice that both the OLS and the GLS estimates for the intercept are substantially smaller than the true intercept (=1). The empirical biases may be due to chance, as suggested by the large variances of the intercept estimates. Nevertheless, in most regression analyses, the analyst's prime interest lies not in the intercept of the model, but the slope coefficients. S/he would be concerned about how much each unit change in a covariate would affect the response. This change is measured by the corresponding slope estimate and its variability.

The estimates for the slope coefficients seem quite accurate.

Now, to assess the *consistency* (see page 24) of the methods of Enlightened and Naive, we examine the off-diagonal values of each 2×2 p-value contingency table. First, we see that the *consistency ratio wrt the naive OLS analysis* for the intercept is 2:0. That is, Naive (in the naive OLS analysis) twice (out of 200) failed to detect the non-zero intercept at the 0.05 significance level when Enlightened succeeded (in the GLS analysis). On the other hand, this ratio wrt the true OLS analysis is 0:1. That is, Enlightened (in his/her GLS analysis) failed to detect it once when Naive succeeded (in his/her naive OLS analysis). However, the small number of inconsistent conclusions made by the two statisticians could well be due to chance. Similar reasoning obtains for the estimates of the two slopes.

To compare the empirical power of the three analyses, we compare the sample probabilities of detecting a non-zero regression coefficient at significance level 0.05. For example, Enlightened's power for detecting a non-zero intercept was $(11+2)/200 = 0.065 = 6.5\%$. This power was slightly lower at $11/200 = 5.5\%$ for Naive. In the true OLS analysis, Enlightened's power for the OLS estimator was $(13+1)/200 = 7.0\%$. The power is slightly lower (6.5% as above) for his/her GLS analysis. These powers are all very low, possibly due to the large variances in the intercept estimates. Also, the powers are similar because of the particular choice of covariate matrix which gave similar OLS and GLS estimators.

We do not see striking power differences either between those associated with the OLS and GLS slope estimates. In fact, if we compare the GLS power and that of either OLS analysis, the Pearson's χ^2 test of independence of analysis (GLS/OLS) and power ($p <$ or ≥ 0.05)[16] of detecting each of β_0, β_1, and β_2 retains the null hypothesis (of independence). For example, the p-value[17] for the χ^2 test is 0.83 for the GLS and naive OLS analyses in detecting β_0.[18] Thus, the power differences in this simulation experiment are very likely

[16] In such a χ^2 test, the two categories of analysis are "GLS" and "OLS," each with two power levels, namely, "$p < 0.05$" and "$p \geq 0.05$." We test the independence of analysis and power (category and level).

[17] ... obtained using the S-Plus function `chisq.test`.

[18] In this χ^2 test, the data are (naive OLS $p < 0.05$, GLS $p < 0.05$) = (11, 13), and similarly (189,

to be results of chance[19] according to the χ^2 test.

Finally, the fact that the two powers were low in detecting non-zero slopes in this simulation is desirable, because the true slopes *were* zero!

In Simulation 2, much the same sort of results were obtained.

All the (true) OLS estimate variances are noticeably larger than their GLS counterparts. We again see large variability in the slope coefficient estimate for both the OLS and the GLS methods. In fact, this is true for every co-ordinate of the $\boldsymbol{\beta}$ vector. The sample mean values of the parameter estimates indicate that overall, the coefficients were not well estimated, possibly due to chance as a result of their large variabilities. The large variances were mainly due to the choice of $\boldsymbol{X}$ and $\boldsymbol{\Sigma}_\epsilon$. However, the GLS estimates appear to be slightly more accurate with smaller empirical biases than the OLS estimates.

Large variability in the coefficient estimates also entails low power in detecting non-zero true coefficients. This power for all three regression coefficients is under 10% based on any of the naive OLS, true OLS, and GLS covariances.

Finally, minimal inconsistency is found between the methods in the detection of non-zero coefficients. However, notice that in a few cases (3 for β_0 and 2 for both β_1 and β_2), the true OLS analysis actually out-performed the GLS analysis in detecting non-zero coefficients. Of course, the small number of such cases does not strongly suggest much about the overall power of this true OLS analysis. In fact, this out-performance of the true OLS analysis is likely due to chance: the χ^2 tests[20] for this simulation experiment all yielded large p-values. Also, theoretically, the true OLS analysis is never more efficient than the GLS analysis, as indicated by Theorem 1.

187) for the "$p \geq 0.05$" level.

[19] ... in the terminology of Freedman, Pisani, Purves & Adhikari, 1991, Chapter 26.

[20] See footnotes on page 33.

Program input variables:

r	g	n	$\boldsymbol{X}$	$\boldsymbol{\beta}$	σ^2	$\boldsymbol{\Sigma}_\epsilon$	fit intercept?
30	1	500	see below	$(3, 1.5)^T$	1	item 2 on page 26	yes

Program output:

	OLS estimates		GLS estimates	
	β_0	β_1	β_0	β_1
sample mean of n estimates	3.11	1.49	3.05	1.49
sample variance of n estimates	11.54	0.11	8.60	0.08
true variance of each estimate	13.19	0.13	9.93	0.10

		GLS			
		β_0		β_1	
		$p < 0.05$	$p \geq 0.05$	$p < 0.05$	$p \geq 0.05$
OLS $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$	$p < 0.05$	35	8	482	0
	$p \geq 0.05$	46	411	15	3
OLS $\boldsymbol{\Sigma}_\epsilon$ known	$p < 0.05$	46	24	491	1
	$p \geq 0.05$	35	395	6	2

Table 2.4: Results of Simulation 3

(ii) Simulations 3 and 4

The next two simulations were of larger scale than the previous two. Instead of $n = 200$, we had 500 realizations of model (2.1) in Simulations 3 and 4. Such larger simulations were naturally more time consuming. To save computational time, we fixed the covariate matrix and the regression coefficient vector. $\boldsymbol{X}$ was set to be the 30 × 2 covariate matrix described on page 26, and $\boldsymbol{\beta} = (3, 1.5)^T$. The results are presented in Tables 2.4 and 2.5.

The two simulations were done with the same program input variables, except the response dependence correlation. The responses have a covariance matrix as described in item 2 on page 26 for Simulation 3, and item 3 on page 26 for Simulation 4.

Program input variables:

r	g	n	$\boldsymbol{X}$	$\boldsymbol{\beta}$	σ^2	$\boldsymbol{\Sigma}_\epsilon$	fit intercept?
30	1	500	see page 35	$(3, 1.5)^T$	1	item 3 on page 26	yes

Program output:

	OLS estimates		GLS estimates	
	β_0	β_1	β_0	β_1
sample mean of n estimates	3.14	1.49	3.07	1.50
sample variance of n estimates	8.96	0.089	1.25	0.011
true variance of each estimate	9.41	0.093	1.24	0.011

		GLS			
		β_0		β_1	
		$p < 0.05$	$p \geq 0.05$	$p < 0.05$	$p \geq 0.05$
OLS $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$	$p < 0.05$	43	1	487	0
	$p \geq 0.05$	348	108	13	0
OLS $\boldsymbol{\Sigma}_\epsilon$ known	$p < 0.05$	91	2	498	0
	$p \geq 0.05$	300	107	2	0

Table 2.5: Results of Simulation 4

In Simulation 3, because of the choice of $\boldsymbol{X}$ and $\boldsymbol{\Sigma}_\epsilon$, the variance is very large for the intercept estimate, but it is very small for the slope estimate. Using a slightly different $\boldsymbol{\Sigma}_\epsilon$ in Simulation 4, the variances for the GLS estimates decrease considerably (by about eight times). In both simulations, both regression coefficients in the model are well estimated by even the OLS estimators, although the GLS estimators seem to have slightly smaller empirical biases.

However, due to the large difference in the variances of the GLS intercept estimate, the p-value tables of the two simulations for β_0 are very different. In particular, the GLS analysis obviously out-performed the OLS analysis (naive or true) in Simulation 4 (in which the variance of the GLS intercept estimate is much smaller). In terms of consistency, the consistency ratio wrt the naive OLS analysis is 348:1. The true OLS analysis is slightly better than the naive in terms of consistency. Instead of a 348:1 ratio, it is a 300:2 ratio wrt the true OLS analysis. As for power, the empirical powers of the GLS, true OLS, and naive OLS analyses are respectively 78%, 19%, and 9%. The χ^2 test of independence (of analysis and power) yielded a virtually zero p-value for comparing the GLS analysis to either OLS analysis. (In other words, the power of detecting the intercept is significantly associated with the analysis (GLS/OLS).) Here, the difference in power is striking. Although the intercept is typically not the main interest of a regression analysis, we can nonetheless see "the power of the enlightened" in detecting β_0 in this simulation experiment.

This contrast in power, with respect to the intercept estimates, is not as striking in Simulation 3 as in 4. Nonetheless, the GLS analysis seems more efficient than the two OLS analyses. The consistency ratios wrt the naive and the true OLS analyses are respectively 46:8 and 35:24. Thus, the GLS analysis was the most efficient in this simulation, followed by the true OLS analysis, then the naive OLS analysis. This efficiency ranking is heralded by their empirical powers, which are 16%, 14%, and 9% respectively. In fact, the χ^2 test of independence for the GLS and naive OLS analyses for β_0 yielded a near-zero p-value. Thus, the difference in power between the the analyses was extremely unlikely to be a result of chance. We may say that Enlightened is significantly more powerful than Naive in detecting the intercept, according to the χ^2 test. However, the χ^2 test for the GLS and true OLS analyses indicated insignificant association between the analyses and their powers.

Although the difference in efficiency is not as extreme as in the intercept analysis, the GLS analysis seems more efficient in both simulations than either of the OLS analyses in detecting the slope coefficient. The difference is slightly more noticeable in Simulation 3 than Simulation 4. We did not apply the χ^2 test of independence for the slope estimate of either simulation due to the presence of zero counts.

When we compare the results of these two simulations, we notice that with stronger response dependence (in Simulation 4), more efficiency is lost in the OLS analyses (wrt the GLS analysis).

(iii) Simulations 5 and 6

Simulations 5 and 6 were run with an exchangeable $\boldsymbol{\Sigma}_\epsilon$. Recall that with such a response dependence structure and a non-zero intercept in model (2.1), the OLS and the GLS estimators are identical. Thus, the true OLS covariance is no different from the GLS covariance, as stated in Theorem 2. However, the naive OLS covariance is not the same as the GLS covariance, unless $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$. With $\boldsymbol{\Sigma}_\epsilon \neq \boldsymbol{I}$, we expect to see loss of efficiency in parameter estimation.

In these two simulations, we used the same program input variables as in Simulations 3 and 4, except that $\boldsymbol{\Sigma}_\epsilon$ has the exchangeable structure. The intraclass correlation coefficient in $\boldsymbol{\Sigma}_\epsilon$ is 0.6 in Simulation 5 and 0.9 in Simulation 6. The results of these two simulations are shown in Tables 2.6 and 2.7. Their bottom panels tabulate the p-value

Program input variables:
exchangeable $\boldsymbol{\Sigma}_\epsilon$, intraclass $\rho = 0.6$ (see Table 2.4 for other program input variables)
Program output:

	GLS = OLS estimates	
	β_0	β_1
sample mean of n estimates	3.14	1.49
sample variance of n estimates	9.76	0.077
true variance of each estimate	7.19	0.066

		GLS			
		β_0		β_1	
		$p < 0.05$	$p \geq 0.05$	$p < 0.05$	$p \geq 0.05$
OLS $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$	$p < 0.05$	109	0	500	0
	$p \geq 0.05$	122	269	0	0
GLS power (= true OLS power)		46%		100%	
naive OLS power		22%		100%	

Table 2.6: Results of Simulation 5

Program input variables:
exchangeable $\boldsymbol{\Sigma}_\epsilon$, intraclass $\rho = 0.9$ (see Table 2.4 for other program input variables)
Program output:

	GLS = OLS estimates	
	β_0	β_1
sample mean of n estimates	2.96	1.50
sample variance of n estimates	2.53	0.018
true variance of each estimate	2.55	0.017

		GLS			
		β_0		β_1	
		$p < 0.05$	$p \geq 0.05$	$p < 0.05$	$p \geq 0.05$
OLS	$p < 0.05$	297	0	500	0
$\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$	$p \geq 0.05$	155	48	0	0
GLS power (= true OLS power)		90%		100%	
naive OLS power		59%		100%	

Table 2.7: Results of Simulation 6

of the GLS analysis versus the naive analysis. We will not show the actual contingency table corresponding to the true OLS analysis because this analysis is equivalent to the GLS analysis. Thus, the off-diagonal values of this contingency table will be the same. (Hence, the consistency ratio wrt the true OLS analysis must be $k : k$ for some non-negative k.) Also for this reason, we will only quote the empirical power of the GLS analysis and the naive OLS analysis.

Tables 2.6 and 2.7 show the extent of efficiency loss in estimating β_0 in the naive OLS analysis. (The χ^2 test of independence of analysis and power of detecting β_0 yielded a virtually zero p-value in either simulation experiment.) Comparing the two tables, we see that stronger response dependence heralds greater efficiency loss in the naive analysis. This relationship between the response dependence and efficiency loss was also seen in the previous two simulations. As we will see in some other simulations (see the next subsection), this relationship is generally true.

(iv) Other Simulation Experiments

We ran about fifteen more simulations in addition to the six described above. They were mainly intended to investigate under what conditions the GLS analysis would significantly out-perform either of the OLS analyses when the responses are positively correlated.[21] Tables 2.2 to 2.7 were shown in full to illustrate how we interpret the program output. For brevity, we will merely highlight the more notable results, summarized as follows:

1. In general, the stronger the response correlation, the larger is the consistency ratio wrt either OLS analysis.[22]

[21] Positively correlated responses are often encountered in studies of repeated measurements or in spatial statistics.

[22] In the case of a $k : 0$ ratio for some non-negative k, a "large" ratio means a large k.

2. While the last observation is generally true, in some specific cases, the two OLS analyses were empirically more powerful than the GLS analysis. This happened mainly in the analysis for the intercept, which usually had a very large variance. A large variance could have led to a slightly higher empirical power of an OLS analysis, by chance. (Note that the covariance of a $\boldsymbol{\beta}$ estimate depends on the choice of $\boldsymbol{X}$, $\boldsymbol{\Sigma}_{\epsilon}$, and σ^2.) In the trials without an intercept, an empirically more powerful OLS analysis for a slope could also be due to a large variance of the slope estimate.

3. In some cases, with relatively weak response correlation, the GLS analysis proved to be more powerful than the OLS analyses. For example, a simulation experiment was done with the response dependence variogram described on page 27. Our choice of the region bearings (longitudes and latitudes) gave us a variogram which was extremely close to the identity matrix. However, for one of the two slope coefficients, the GLS analysis turned out to have 89% power, compared to 77% for the true OLS analysis, and 75% for the naive OLS analysis. The χ^2 test of independence of analysis and power of detecting this slope yielded a virtually zero p-value for each of the "GLS/true OLS" and "GLS/naive OLS" combinations. Such a small p-value indicates that the difference in power was extremely unlikely to be a result of chance.

4. In general, the loss in efficiency (wrt to the GLS analysis) of the true OLS analysis is less than that of the naive OLS analysis. The difference in efficiency lost between the two OLS analyses can sometimes be extreme. For example, in one simulation, the consistency ratio wrt the true OLS analysis was 4:2, but it was 147:0 wrt the naive OLS analysis.

5. As expected, in cases for which the OLS (naive or true) variances are much larger than the GLS variances, the GLS analysis is much more powerful than the OLS analysis.

6. In most cases, both the OLS and the GLS estimators accurately estimate the true regression parameters, in that they both have small empirical biases. However, the empirical bias for both estimators can be quite large if their variances are large.

7. In general, the OLS and the GLS estimates are very close to each other. (They are the same if the response dependence is exchangeable, with an intercept in the regression model.) Sometimes, the GLS estimate has less empirical bias, as shown in Simulations 1 to 4.

2.8 How bad is Naive's analysis?

In the previous section, we have seen that in general, ignoring response dependence in linear regression can be dangerous. A considerable loss in efficiency can result. There has been some established theory on the relative efficiency in parameter estimation of the true OLS analysis to the GLS analysis. As Theorem 1 indicates, the true OLS analysis can never be more efficient than the GLS analysis.[23] Our simulations also support this difference in efficiency, in terms of consistency and power, in detecting non-zero regression coefficients.

However, we are not aware of theory which extensively compares the naive OLS analysis with the GLS analysis. From our simulation studies, we do know that in general, the naive OLS analysis is the most inefficient in parameter estimation, followed by the true

[23] In practice, the empirical relative efficiency of the true OLS analysis to the GLS analysis may be greater than unity. This is the case in, for example, Simulation 2 in Section 2.7.3. This is mainly due to chance. If we had more data sets (i.e. more than 500) in this simulation, the GLS analysis could have been empirically more efficient than the true OLS analysis.

OLS analysis,[24] then by the GLS analysis. Now, we direct our focus onto understanding the extent of efficiency loss using the naive OLS analysis relative to the GLS analysis.

2.8.1 Empirical Power Curves

First, let us examine how the power in detecting a non-zero regression coefficient of the naive OLS analysis compares to that of the GLS analysis. We compare them under varying conditions. For example, we may fix a regression coefficient and $\boldsymbol{\Sigma}_\epsilon$, letting σ^2 vary. Similarly, we may fix the covariances of the responses, and let the regression coefficient vary. We will once again use simulation experiments as the basis of our investigation.

We ran several such experiments for each set of varying conditions. For each of these simulations, we recorded the empirical powers of both the GLS and the naive OLS analyses. Each simulation consisted of 500 realizations of model (2.1), with no intercept and one slope coefficient. Having only one slope in the model provides us with enough insight into the question at hand while reducing the computational burden. We used a model with no intercept term because, as explained before, the slope is the primary determinant of the relationship between the response and the covariate. We chose a 30×1 covariate matrix $\boldsymbol{X}$ in this case, assuming there were 30 regions in the geographical domain of interest. The covariates in this $\boldsymbol{X}$ matrix (vector) were arbitrarily generated, independently, from a normal distribution with mean 0 and standard deviation 0.2. Once generated, the $\boldsymbol{X}$ matrix was fixed throughout the 500 simulation trials. $\boldsymbol{\Sigma}_\epsilon$ was taken to be the one described in item 3 on page 26 in all simulations. Such a $\boldsymbol{\Sigma}_\epsilon$ was chosen because from the previous simulations with this $\boldsymbol{\Sigma}_\epsilon$, the GLS analysis was often noticeably more powerful than the naive OLS analysis. We wish to exploit to a greater extent this

[24] The naive OLS analysis may appear to be more efficient than the true OLS analysis when the matrix $(\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{X})^{-1} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is positive definite (i.e. when the true OLS covariance is "larger than" the naive). However, technically the naive analysis is *not* more efficient in this case because $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is not the true covariance of $\widehat{\beta}^{OLS}$.

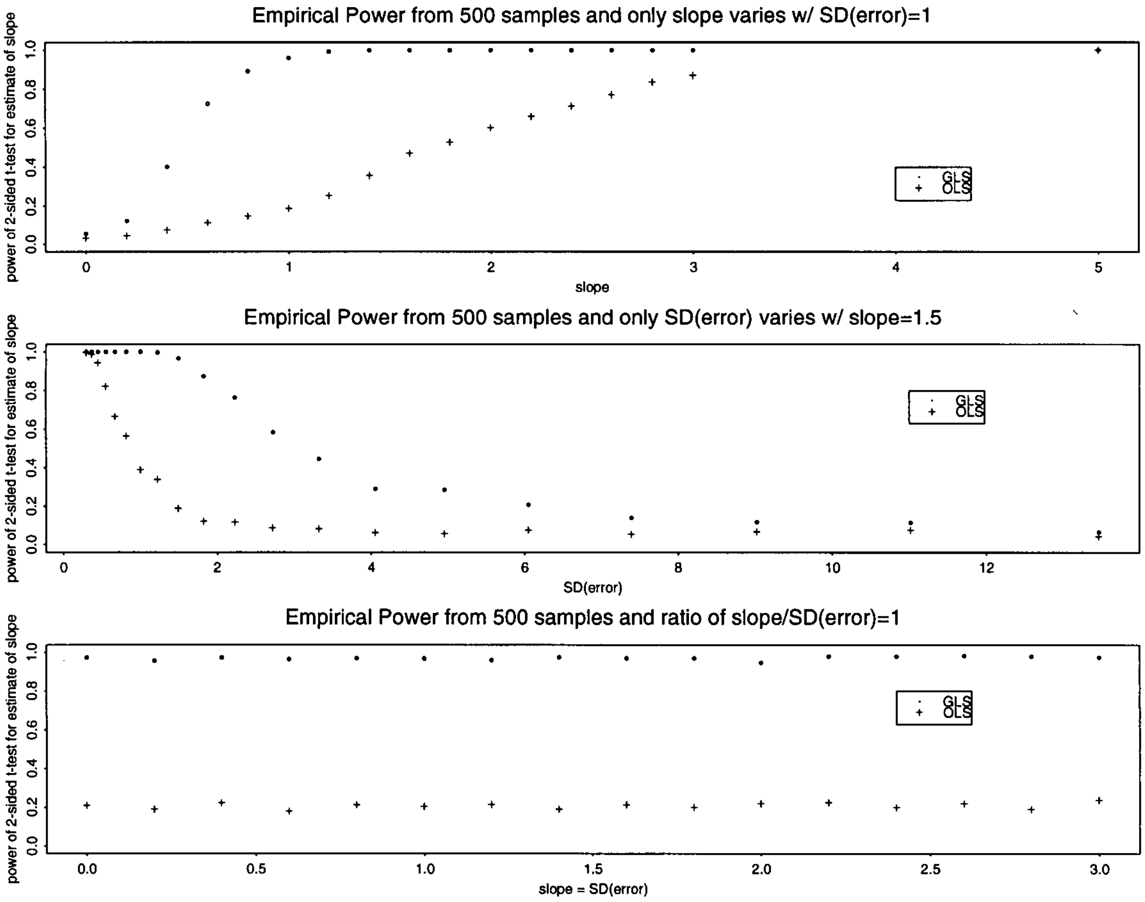


Figure 2.1: Empirical Power Curves for the GLS and naive OLS analyses.

difference between the two analyses to gain increased clarity in the result. A graph of all the recorded empirical powers from the several simulations were then produced.

The above process was repeated under each set of varying conditions. We have provided the graphs of the empirical powers in these sets of simulations in Figures 2.1 and 2.2.

In the two figures, the graph titles designate the parameters chosen to vary. For example, the first panel of Figure 2.1 has $\boldsymbol{X}$ and $\boldsymbol{\Sigma}_\epsilon$ described as above, with $\sigma = 1$

fixed, but a varying β_1. The third panel of the same figure has the ratio β_1/σ fixed, although both β_1 and σ vary together.

Let us examine Figure 2.1 closely.

In the first panel, we let the slope β_1 increase from 0 to 5 in the simulations. With $\sigma = 1$ and every other parameter fixed, both the GLS and the naive OLS analyses increase in power as β_1 increases. Of course, this is not surprising, since the larger the true coefficient, the easier it is to detect, and hence, the higher the power of either test. However, the power of the GLS analysis increases at a much higher rate than the naive OLS analysis, as β_1 increases. The GLS power reaches almost 1 when $\beta_1 \approx 1$, whereas the naive OLS power increases very gradually, and becomes close to 1 only when β_1 is beyond 4. The loss of efficiency in using the OLS analysis is especially big in the interval (0.8, 1.2).

In the second panel of Figure 2.1, we let σ increase from near-zero to beyond 12. $\beta_1 = 1.5$ is fixed, as are other parameters in the model. As σ increases, the power of either analysis should decrease, because detecting a non-zero coefficient becomes more difficult as the variation of the data increases. However, the GLS power remains very close to 1 until $\sigma > 1.5$, then decreases very gradually and stabilizes as σ grows beyond 8. Meanwhile, the naive OLS power starts decreasing as soon as σ grows beyond 0. It decreases much faster than the GLS power, stabilizing when $\sigma \approx 4$.

Finally, in the third panel of Figure 2.1, we let the ratio $\beta_1/\sigma = 1$, i.e. $\beta_1 = \sigma$, although letting both parameters vary together. The other program input arguments are fixed. The graph indicates that both the GLS and the naive OLS power functions remain roughly constant, but the latter is much lower. The naive OLS power seems to remain at around 0.2, whereas the GLS power remains close to 1 at all values of β_1 $(= \sigma)$.

Indeed, the fact that the GLS power remains relatively constant agrees with theoretical results. The following argument shows that the theoretical GLS power function is constant if the ratio β_1/σ is constant.

Let $\nu = \beta_1/\sigma$. Let $\pi(\nu)$ be the power function of the GLS analysis evaluated at ν, i.e. the probability of failing to detect a non-zero β_1, given ν. The t-test statistic is

$$\frac{\widehat{\beta_1}^{GLS}}{\widehat{\mathrm{SE}}(\widehat{\beta_1}^{GLS})}$$

where

$$\widehat{\mathrm{SE}}(\widehat{\beta_1}^{GLS}) = \hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\epsilon}^{-1}\boldsymbol{X})^{-1}}$$

estimates

$$\mathrm{SE}(\widehat{\beta_1}^{GLS}) = \sigma\sqrt{(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\epsilon}^{-1}\boldsymbol{X})^{-1}}.$$

The estimate $\hat{\sigma}$ is the square-root of the MSE obtained from the regression analysis of variance. That is,

$$\hat{\sigma}^2 = \frac{\widehat{\boldsymbol{\epsilon}}^T\widehat{\boldsymbol{\epsilon}}}{r-g}$$

where

$$\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

Recall that the covariate matrix $\boldsymbol{X}$ has dimension $r \times g$, where $r = 30$ and $g = 1$. Now, from classical theory, we know that $\widehat{\beta_1}^{GLS}/\widehat{\mathrm{SE}}(\widehat{\beta_1}^{GLS})$ follows a non-central t distribution with $r - g$ degrees of freedom, and non-centrality parameter equal to[25]

$$\frac{\beta_1}{\mathrm{SE}(\widehat{\beta_1}^{GLS})} = \frac{\beta_1}{\sigma\sqrt{(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\epsilon}^{-1}\boldsymbol{X})^{-1}}} = \frac{\nu}{\sqrt{(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\epsilon}^{-1}\boldsymbol{X})^{-1}}}$$

which only varies in ν, because the values in the denominator of the last expression are all fixed.

Now, let t_c be the critical value of this non-central t distribution at significance level 0.05. It is not difficult to see that

$$\pi(\nu) = \mathrm{P}\left(\left|\frac{\widehat{\beta_1}}{\widehat{\mathrm{SE}}(\widehat{\beta_1}^{GLS})}\right| > t_c\right)$$

[25] See Encyclopedia of Statistical Sciences, Wiley & Sons, 1985, vol. 6, pp. 286 – 288.

is constant if ν is constant.

From the same empirical power graph, we also see that the naive OLS analysis has constant power. In fact, we can deduce that theoretically, the naive OLS analysis should also have constant power. This is true despite the fact that the true standard error of $\widehat{\beta_1}^{OLS}$ is not equal to the naive S.E. (= $\text{SE}(\widehat{\beta_1}^{OLS})_{\text{naive}}$) so that $\widehat{\beta_1}^{OLS}/\text{SE}(\widehat{\beta_1}^{OLS})_{\text{naive}}$ does not have a non-central t distribution.

Recall that $\widehat{\beta_1}^{OLS}$, in the case of $\boldsymbol{\Sigma}_\epsilon \neq \boldsymbol{I}$, has true S.E. and naive S.E. equal to $\sigma\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}}$ and $\sigma\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}}$ respectively. That is,

$$\text{SE}(\widehat{\beta_1}^{OLS})_{\text{true}} = \sqrt{(\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}}\ \text{SE}(\widehat{\beta_1}^{OLS})_{\text{naive}}.$$

Now, notice that $\widehat{\beta_1}^{OLS}/\widehat{\text{SE}}(\widehat{\beta_1}^{OLS})_{\text{true}}$, the t-statistic for the true OLS analysis, has non-central t distribution with non-centrality parameter equal to $\beta_1/\text{SE}(\widehat{\beta_1}^{OLS})_{\text{true}}$. Hence, the power function of the true OLS analysis should be constant if β_1/σ is constant (like that of the GLS analysis).

Now, the naive t-statistic is

$$\frac{\widehat{\beta_1}^{OLS}}{\widehat{\text{SE}}(\widehat{\beta_1}^{OLS})_{\text{naive}}} = \frac{\widehat{\beta_1}^{OLS}}{\widehat{\text{SE}}(\widehat{\beta_1}^{OLS})_{\text{true}}}\sqrt{(\boldsymbol{X}^T\boldsymbol{\Sigma}_\epsilon\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}},$$

which is simply a multiple of the t-statistic of the true OLS analysis. Therefore, the power function of the naive OLS analysis should also be constant if β_1/σ is constant.

Finally, the fact that the empirical power curve of the naive OLS analysis is below that of the GLS analysis in all three panels of Figure 2.1 illustrates that ignoring response dependence in regression analysis can lead to significant loss in parameter estimation efficiency.

Next, let us examine Figure 2.2.

The two panels of this figure show empirical power curves of the two statisticians' analyses with all program input arguments fixed as before, except that β_1 and σ both

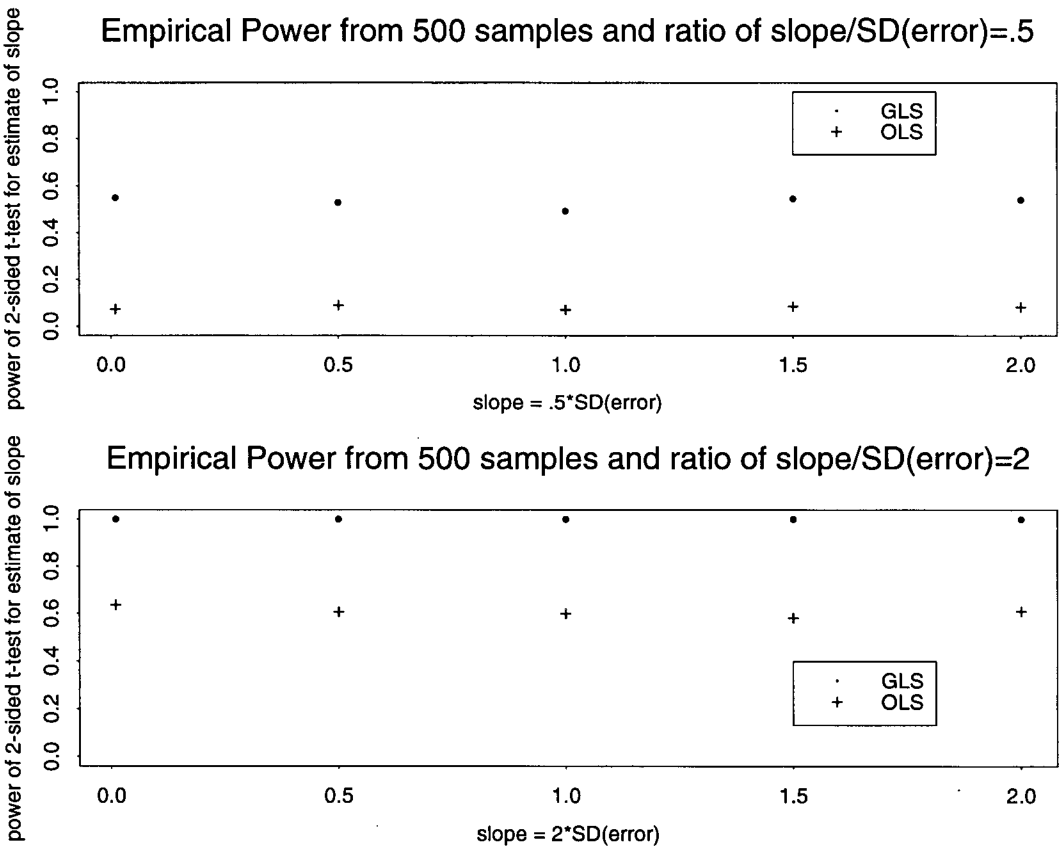


Figure 2.2: More Empirical Power Curves for the GLS and naive OLS analyses.

vary with the ratio $\nu = \beta_1/\sigma$ constant. This ratio was fixed as 1/2 and 2 respectively in the two graphs. Again, we see that the naive OLS analysis is clearly less powerful than the GLS analysis.

Comparing the two panels in Figure 2.2 with the last one in Figure 2.1, the gap between the two power curves seems to decrease as the ratio of β_1 to σ moves away from unity in either direction. Although the GLS power curve is everywhere above the naive OLS power curve, the gap between them is the largest when the ratio is 1, a little smaller when the ratio is 1/2, and even smaller when the ratio is 2. Therefore, the power difference between the two statisticians is most apparent when the slope and the noise level (i.e. σ) are the same.

Furthermore, from the relative positions of the power curves as the ratio of the slope to the noise level changes, we again see the following: the noisier the data with respect to the size of the slope coefficient, the less sensitive either analysis, in terms of the detection of the slope. The GLS analysis has almost 100% power when the slope is at least as large as the noise (σ). On the other hand, the OLS analysis has very low power when the noise is at least at large as the slope.

These observations seem reasonable, by the following argument. When $\beta_1/\sigma = 1$, the two empirical curves are almost as far apart as they can be. That is, the GLS power is almost 1, and the OLS power is close to 0.05, the size of the hypothesis test. As the ratio increases from 1, the size of the slope relative to the noise level increases. Then both analyses improve in their power in detecting the slope, but the OLS analysis has more room to improve in its power (from nearly 0.05 to anywhere below 1). Afterall, how much more can the GLS power improve when it has almost reached 1 at $\beta_1/\sigma = 1$? Similarly, as the ratio decreases from 1, the size of the slope relative to the noise level decreases. Then the powers of both analyses decrease, but that of the GLS analysis has more room to decline.

2.8.2 Histograms of p-value Ratios and p-value Scatter Plots

We have compared the efficiency of parameter estimation using GLS analysis and that using naive OLS analysis by examining their empirical power curves. Another way of comparing their efficiencies is to compare their actual p-values in each simulation. In regression parameter estimation, a method is more efficient than another if its p-value is smaller. This is because a smaller p-value gives stronger evidence against a zero regression coefficient.

In a simulation experiment with, say, 500 data sets, there are 500 pairs of (GLS p, naive OLS p) recorded for each coefficient estimate. Suppose we have only one slope coefficient and no intercept in the regression model. How then should we compare all the 500 pairs of p-values of this slope estimate?

We decided to compute, for each pair of p-values, the ratio of naive OLS p to the GLS p. Since we expected the naive OLS analysis to be less efficient, the ratio should be larger than unity most of the time.[26] We then log-transformed the square of each of these 500 p-value ratios. Log transformations are often used to map ratios in the interval $(0, \infty)$ to $(-\infty, \infty)$. Moreover, our choice of squaring the p-value ratio before log-transformation was inspired by Wilks' Theorem: twice the log-likelihood ratio statistic has an asymptotic Chi-square distribution. We expected that such a transformation of the p-value ratios would allow nicer graphical displays of the simulation results.

Histograms of these log-transformed p-value ratios were then plotted to examine their distribution. For example, a histogram with most of its mass at, say, 10, would indicate that in most of the 500 cases in the simulation, the naive OLS p-values were $\exp\{10/2\}(\approx 148)$ times larger than its GLS counterpart. This in turn would indicate that the GLS analysis was much more powerful than the naive OLS analysis. On the other hand, a

[26] The ratios which are smaller than 1 are expected to be due to chance.

histogram centered at 0 would indicate that the naive OLS p-value was just as likely to be larger than its corresponding GLS p-value as it was smaller. In other words, such a histogram would indicate that the two analyses were equally powerful, empirically.

Before presenting some detailed observations, let us introduce an analogy of the power comparisons between Enlightened and Naive. A rabbit, under normal conditions, usually runs much faster than a tortoise. If both the rabbit and the tortoise wear good running shoes (large β_1 relative to σ), it may be more apparent that the rabbit is a faster runner than the tortoise. Afterall, how much improvement can running shoes do to an animal which simply cannot run? On the other hand, if both animals have a broken leg (large σ relative to β_1), the hindrance to the rabbit's running may be a lot more severe than that to the tortoise's, because the rabbit's speed has much more room to decline. Hence, if we measure the difference between the two animals' running speeds, this difference should be more apparent when both wear good running shoes, but less so when they both have a broken leg. We will leave for the reader the task of linking this analogy between these two animals in a race and the two statisticians in an efficiency competition of regression parameter estimation.

Now, using this analogy, we wish to understand how much faster the rabbit out-runs the tortoise under different conditions. Thus, we need the two animals to compete in several races (simulation experiments). In each race, they wear the same kind of running shoes *assigned to that particular race.* Running shoes of a different quality are assigned to a different race. Over several races, we hope to recognize which quality of running shoes makes the rabbit's superiority the most or the least apparent. To do this, we can display histograms of logged p-value ratios from several simulation experiments in the same graph. This way, the change in spread and location of the histogram due to the change in the "quality of the running shoes" can be spotted easily.

Some technical difficulties might arise in plotting some of the histograms. For example, the log of zero is undefined. To get around the problem of taking the log of a virtually zero p-value, we used *started logs*[27] and added 10^{-10} to each of the p-values, regardless of its size,[28] before taking logs. Another difficulty emerges in displaying the histogram itself. A regular histogram resembling a bar chart may not provide an effective display. Since our main interest in these histograms concerns location and spread, we will present their smoothed sketches instead of the actual histograms.

Plotting the histogram of the logged p-value ratios is only one way to compare the p-values, and hence, the power of the two analyses. In addition, we also produced a scatter plot of all the logged OLS p-values versus their GLS counterparts for each simulation. For example, if the scatter cloud is concentrated around a point, this corresponds to a histogram which has a tall spike.

In the previous section, empirical power curves were given for different sets of varying conditions. For example, in one graph, we fixed all program input arguments but let β_1 vary. A set of varying conditions is analogous to "different models/qualities of running shoes of the same brand." That is, the brand could be β_1, and the different models could be $\beta_1 = 0.1$, $\beta_1 = 0.2$, etc. The noise level σ could be considered a different brand of running shoes, when it varies but all other parameters (and program input arguments) are fixed. For a certain brand, several races are run between the two animals, and each race is run with a different shoe model. Let us now relate this back to our p-value comparisons. For each brand for which, say, five different races (simulations) were run, five sets of race results in terms of the histogram sketches and the corresponding scatter plots were displayed. Hence, we could compare the five histograms and/or the five scatter plots to examine how the varying values (shoe models) of this parameter (brand) affected

[27]... in the terminology of Mosteller & Tukey, 1977, page 91.

[28]For the sake of consistency, we added the infinitesimal amount to *all* p-values.

the power comparison of the naive OLS and the GLS analyses (speed comparison of the two animals).

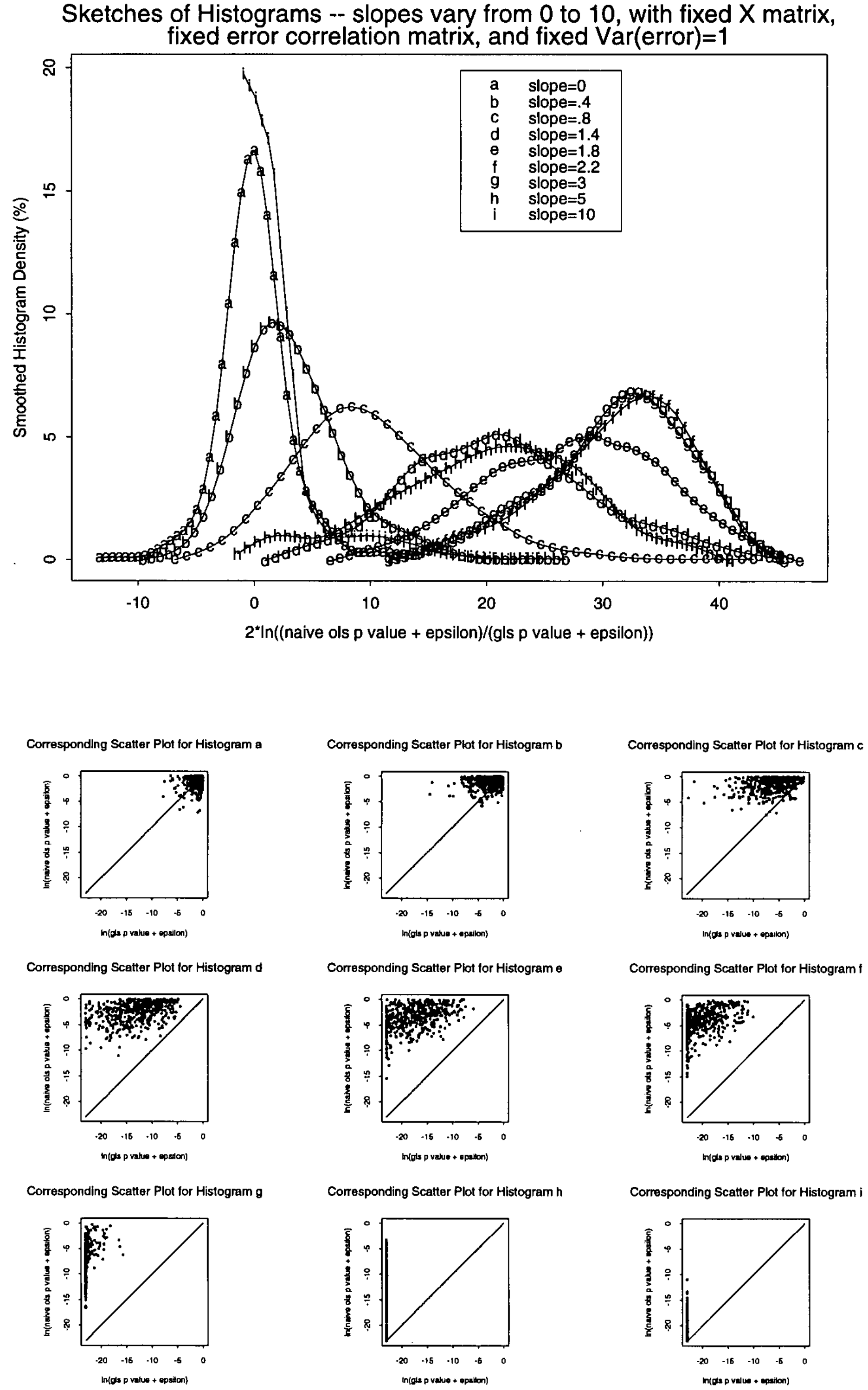
The Graphs

Several sets of varying conditions were investigated. Here, we used the same $\boldsymbol{X}$ and $\boldsymbol{\Sigma}_\epsilon$ matrices as those used in the previous section where simulations were run to produce empirical power curves. Thus, we have assumed 30 regions, and no intercept but only one slope coefficient, β_1, in regression model (2.1). Again, each simulation had 500 realizations (datasets) of the model. The varying parameter is indicated in the title of each histogram graph. Notice also that the title of each scatter plot indicates its corresponding histogram sketch. The x-axis of a scatter plot is always "log (GLS $p + 10^{-10}$)" and the y-axis is always "log (naive OLS $p + 10^{-10}$)." A 45-degree line is drawn on each scatter plot to indicate clearly the relative position of the scatter cloud.

Figure 2.3 shows the histogram sketches and scatter plots for nine simulations under one set of varying conditions. $\boldsymbol{X}$ and $\boldsymbol{\Sigma}_\epsilon$ were fixed as described above, σ was fixed as 1, but β_1 varied from 0 to 10. Each histogram sketch is labeled with a letter from the alphabet for two reasons: (1) by comparing the labeling letter to those in the legend, we can match the histogram to the β_1 value with which the simulation was run; (2) as β_1 increases, the shape and the location of the histogram changes, and we can trace the changes of the histogram by following the *alphabetically ordered* sketches. That is, increasing β_1 to the next value corresponds to changing the labeling letter to the following letter in the alphabet. With this in mind, we may trace how the histogram "moves" as β_1 increases.

From the histogram sketches in this figure, we notice a "tidal effect" in the location of the histograms as β_1 increases. That is, with an increasing slope coefficient, the center of the histogram of the logged p-value ratios "moves" from small values (≤ 0) up to

Figure 2.3: Histogram Sketches of logged p-value Ratios and their Scatter Plots: β_1 varies.



larger values, then returns down to smaller values as the tide does. Note that a slope of moderate size, such as 0.8 (histogram "c"), gives a histogram mode of about 9. That is, for such a slope (and $\sigma = 1$), the OLS p-value is typically $\exp\{4.5\} \approx 90$ times as large as the GLS p-value!

Another interesting observation is the tidal effect in the spread of the histogram which accompanies that of the location. Apparently, extreme (very small and very large) values of the slope are associated with smaller spread in the distribution of the log-transformed p-value ratios. Moderate values of the slope are associated with wider spread in this distribution.

The tidal effect in the location and spread of the histogram can also be seen from the logged p-value scatter plots.

First, we see the changing location of the scatter cloud. As β_1 increases, the scatter cloud moves from being around the top portion of the 45-degree line to being away from the line to finally being solely on the left of the line. Thus, when the slope is small (difficult to detect with either the OLS or the GLS method), both analyses have big p-values (so that the logged p-value is around 0). A small slope also means that empirically, the naive OLS analysis has more probability of out-performing the GLS analysis than with a larger slope. (Recall the rabbit and the tortoise racing against each other when both have a broken leg.) This explains the portion of the scatter cloud below the 45-degree line in the first few scatter plots. With a slope that is easier to detect using either method, the GLS analysis' superiority in power becomes more apparent. Having a large slope is analogous to having good quality running shoes in the rabbit-tortoise race. Recall how we explained that the rabbit's superiority in speed is most apparent when the two animals are wearing the best running shoes. Thus, as the slope increases, more and more GLS p-values are smaller than their naive OLS counterparts, and hence, the cloud moves further and further to the left of the 45-degree line. Finally, when $\beta_1 = 10$, an extremely large slope, the scatter cloud converges along a vertical line, almost solely above the 45-degree line, at a highly negative value on the x-axis. Such a vertical line indicates that almost all GLS p-values were at least as small as their naive OLS counterparts in the simulation. In other words, the GLS analysis was at least as empirically powerful in almost all 500 cases in the simulation.

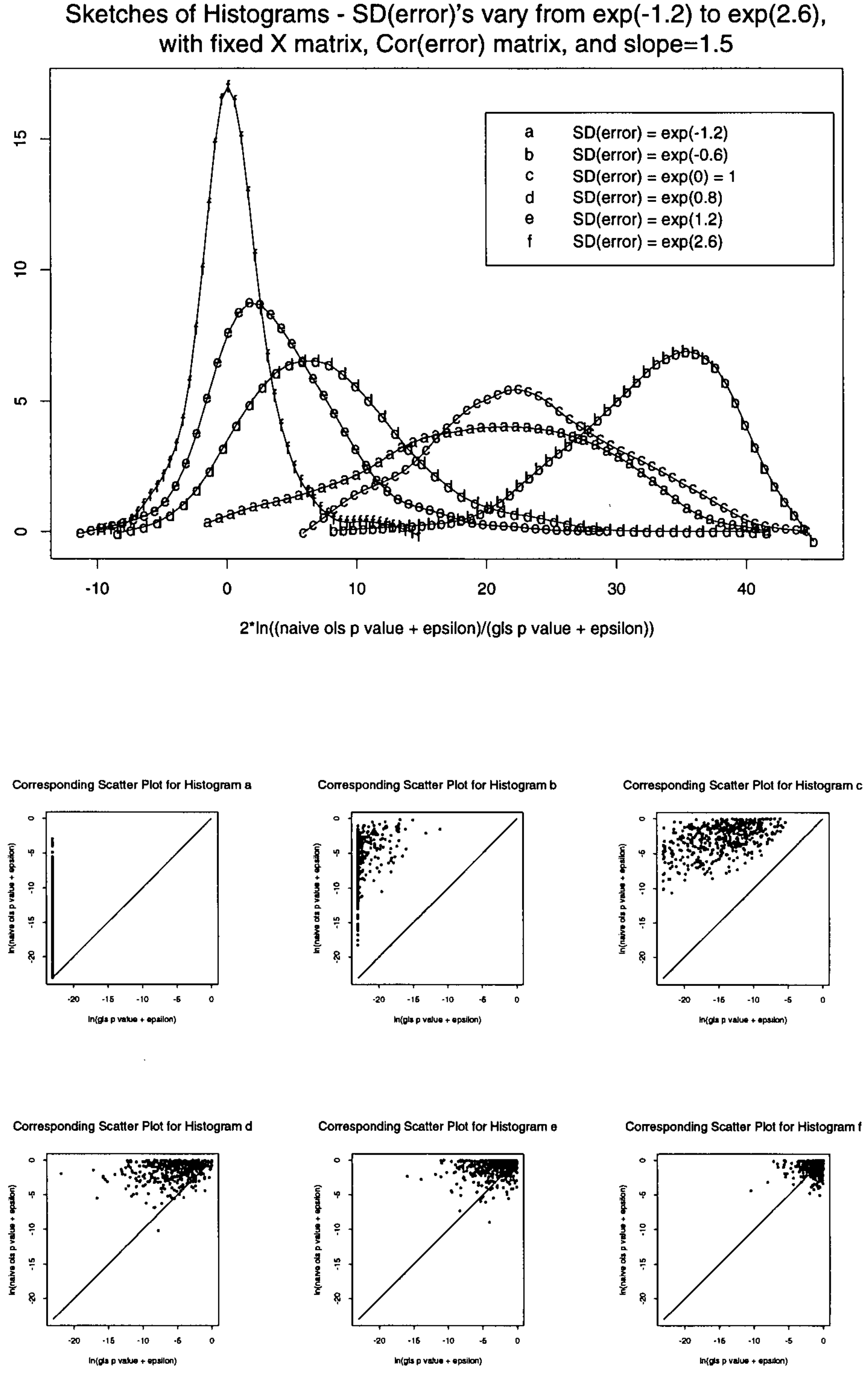
Notice that this last scatter plot corresponds to histogram sketch "i," which shows that there are almost no log-transformed p-value ratios that are negative, out of the 500 cases in the simulation. (A negative value corresponds to a naive OLS p-value smaller than the GLS p-value.) This is simply the exact translation of the observation of the last scatter plot.

In addition to the moving location, the spread of the scatter cloud also changes as β_1 increases. From being very concentrated around one point, it becomes somewhat sparsely scattered, then becomes concentrated again. Although the cloud appears concentrated along a vertical line at large values of the slope coefficient, the last plot actually shows some tendency of the cloud converging to a point near (-25, -25). This change in the spread of the cloud simply translates back to the changing spread of the histograms.

Now, putting back together the observations of both the histogram sketches and the scatter plots in Figure 2.3, it is not difficult to understand the tidal effect in both the location and the spread of either the histogram or the scatter cloud. Basically, it depends on the relationship between the varying quality of the animals' running shoes and the varying extent of superiority of the rabbit over the tortoise in a race. Here, the larger β_1 is, the better the quality of the shoes, and hence the larger the rabbit's "margin of victory" over the tortoise.

Figure 2.4 shows some similar tidal effects in the histogram and scatter cloud, but in the opposite direction of that in Figure 2.3. The simulation experiments were run with $\boldsymbol{X}$ and $\boldsymbol{\Sigma}_\epsilon$ fixed as described above, β_1 fixed as 1.5, but σ increased from $\exp\{-1.2\}$ to

Figure 2.4: Histogram Sketches of logged p-value Ratios and their Scatter Plots: σ varies.


Sketches of Histograms - SD(error)'s vary from exp(-1.2) to exp(2.6), with fixed X matrix, Cor(error) matrix, and slope=1.5

Corresponding Scatter Plot for Histogram a

Corresponding Scatter Plot for Histogram b

Corresponding Scatter Plot for Histogram c

Corresponding Scatter Plot for Histogram d

Corresponding Scatter Plot for Histogram e

Corresponding Scatter Plot for Histogram f

$\exp\{2.6\}$. In fact, the opposite direction of the tide is not surprising if we understand why the histogram and the scatter cloud "travel" the way they do in Figure 2.3. In Figure 2.3, a large β_1 corresponds to high quality running shoes in the rabbit-tortoise race. On the other hand, a large σ (noise level in data) in Figure 2.4 corresponds to a broken leg, whereas having a small σ is analogous to having good running shoes.

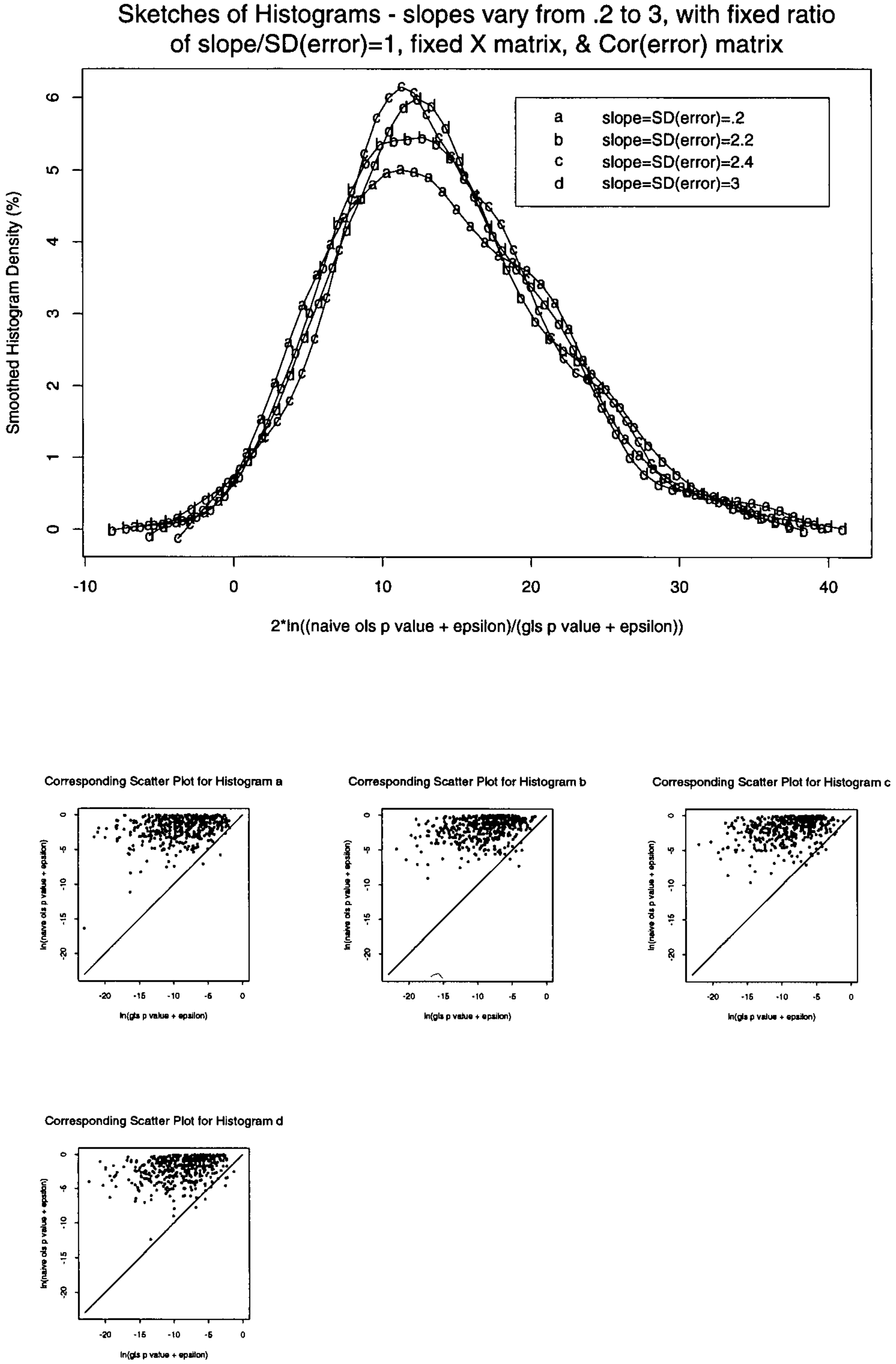
This brings us back to the relative size of the slope coefficient with respect to the noise level in the data. In other words, if both parameters vary together, it is not so clear what shoe is considered to have good quality. In the simulation experiments for Figure 2.5, the ratio $\nu = \beta_1/\sigma$ was fixed as unity, and β_1 and σ increased together from 0.2 to 3. Similarly, ν was fixed as 1/2 and 2 respectively in the sets of simulation experiments for Figures 2.6 and 2.7. The slope β_1 ranged from 0.01 to 2 in both of these figures.

From these three figures, we find that their histograms (and scatter clouds) remain at a relatively constant location and have almost the same shapes as β_1 and σ vary together at a fixed rate. This observation agrees with the constant power (of detecting β_1) of the GLS and OLS analyses seen in Figures 2.1 and 2.2. The logged p-value ratio is similar to the ratio of the powers of the two analyses. If the powers are constant for all values of β_1 and σ (while β_1/σ is fixed), then their ratio must also be constant for all values of the two parameters. Hence, the locations and shapes of the histogram sketches almost coincide in each of Figures 2.5 to 2.7.

However, if we compare these three figures, the tide moves from left to right as the ratio of β_1/σ increases. When $\nu = \beta_1/\sigma = 1/2$, the OLS p-value is typically larger than the GLS p-value by a factor of $\exp\{1.5\} \approx 4.5$. When $\nu = 1$, this factor is $\exp\{6\} \approx 400$. When $\nu = 2$, it is $\exp\{16\} \approx 8.9$ million! The increasing factor indicates that, when the ratio of the slope to the noise level is fixed, the size of this ratio is essentially what determines the relative location of the tide.

In other words, both β_1 and σ increasing together is like having two opposing forces

Figure 2.5: Histogram Sketches of logged p-value Ratios and their Scatter Plots: $\beta_1/\sigma = 1$.

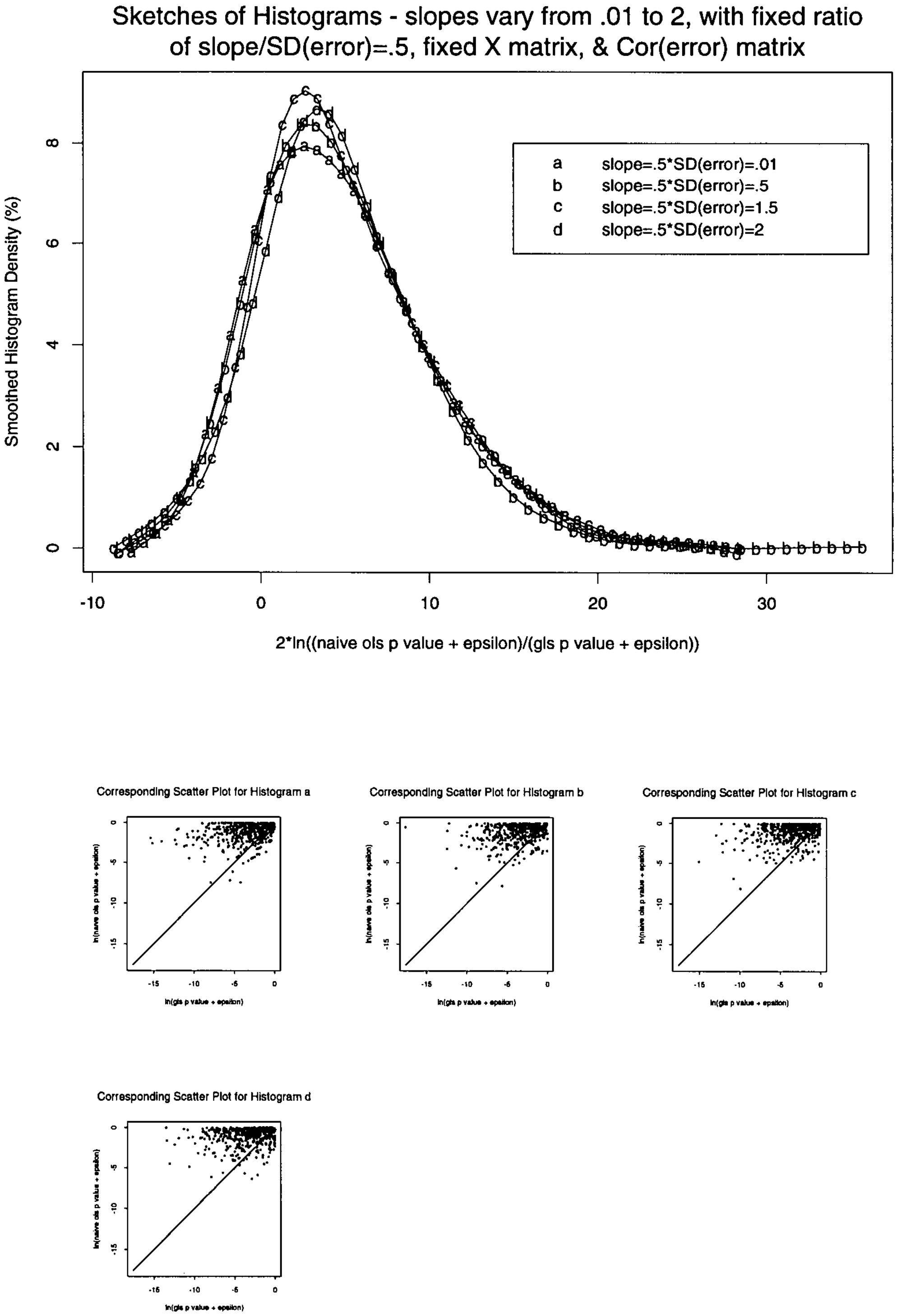


"pulling the tide." A large β_1 makes more apparent the GLS analysis' victory over the naive OLS analysis. On the other hand, a large σ makes this less noticeable. Again, with σ fixed, a large β_1 corresponds to good running shoes. With β_1 fixed, a large σ corresponds to a broken leg. When only the ratio of the two parameters is fixed, but the parameters themselves vary together, the efficiency competition between the two analyses is analogous to the two animals running a race in which they both have a broken leg but wear good running shoes. From the locations of the histogram sketches in Figures 2.5 to 2.7, we deduce that the rabbit's superiority in speed over the tortoise is constant as the parameters increase together with their ratio fixed, but increases if the ratio increases. (This can also be seen from the location of the scatter clouds — the clouds move farther towards the left of the 45-degree line as the ratio increases.) That is, the superiority of the GLS analysis over the naive OLS analysis is more apparent as the relative size of β_1 to the noise level increases.

One may argue that this trend seems to disagree with what we observed from the empirical power curves. They indicate that the superiority of the GLS analysis was most apparent when $\beta_1 = \sigma$, but less so if $\beta_1 \neq \sigma$. From the histogram sketches, we see that this superiority increases as the ratio β_1/σ increases.

However, the empirical power graphs were obtained based on a fixed significance level of the t-tests for the slope estimates. On the other hand, the histogram sketches were obtained using the actual p-values computed from these t-tests. Then, a larger OLS p-value would indicate a lower OLS power than the GLS power, regardless of how small both p-values are. Thus, if we measure the superiority of the GLS analysis based on how much smaller its p-values are, then we see the trend of increasing superiority as the ratio β_1/σ increases. If we measure superiority based on how often the null of $\beta_1 = 0$ is rejected, then this superiority decreases as β_1/σ moves farther away from unity. This is because both p-values might be smaller than the significance level so that both analyses

Figure 2.6: Histogram Sketches of logged p-value Ratios and their Scatter Plots: $\beta_1/\sigma = 1/2$.



Sketches of Histograms - slopes vary from .01 to 2, with fixed ratio of slope/SD(error)=.5, fixed X matrix, & Cor(error) matrix

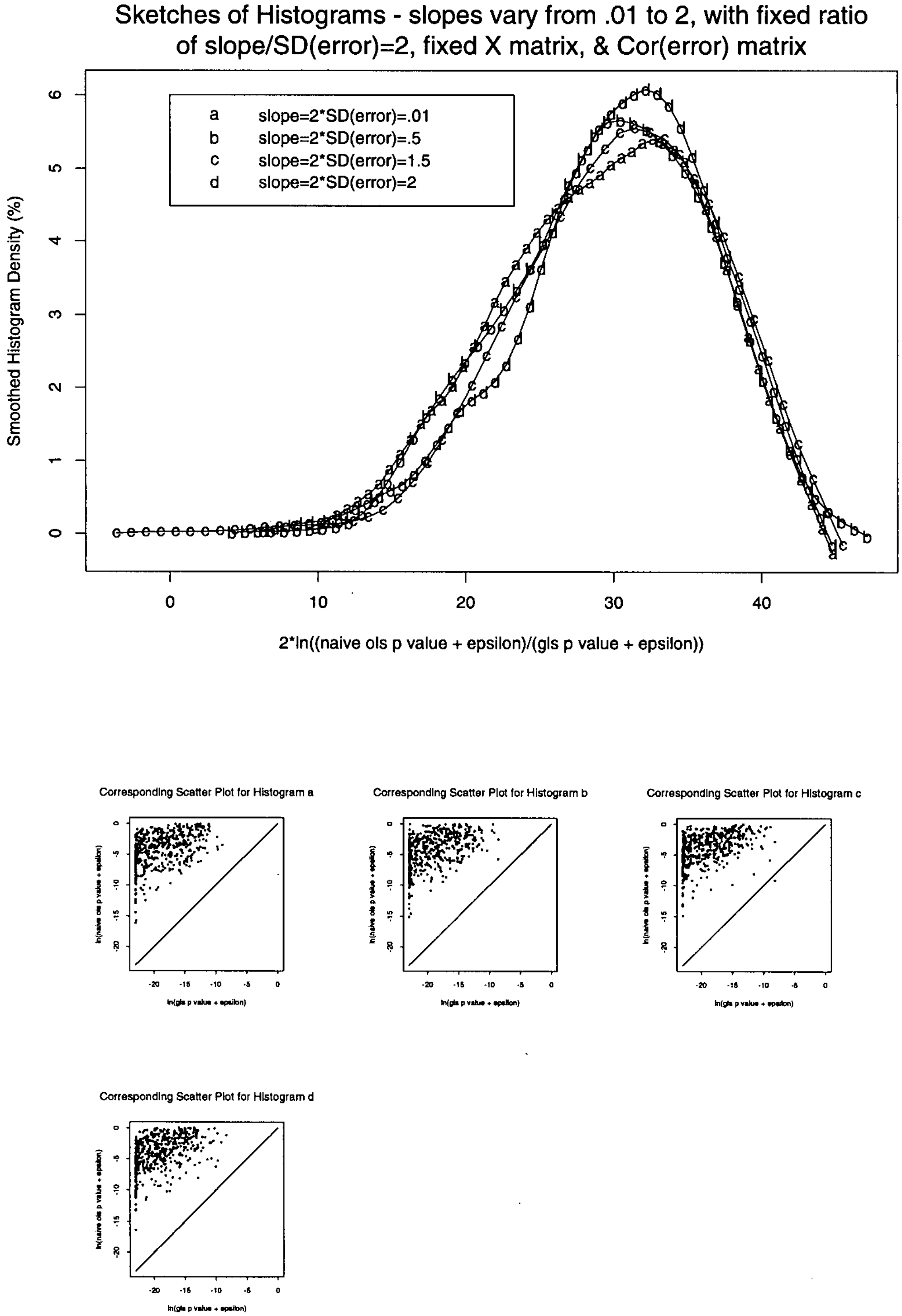
Corresponding Scatter Plot for Histogram a

Corresponding Scatter Plot for Histogram b

Corresponding Scatter Plot for Histogram c

Corresponding Scatter Plot for Histogram d

Figure 2.7: Histogram Sketches of logged p-value Ratios and their Scatter Plots: $\beta_1/\sigma = 2$.



appear powerful, despite the larger OLS p-value .

2.8.3 The case of a chosen $\boldsymbol{\Sigma}_\epsilon \approx \boldsymbol{I}$

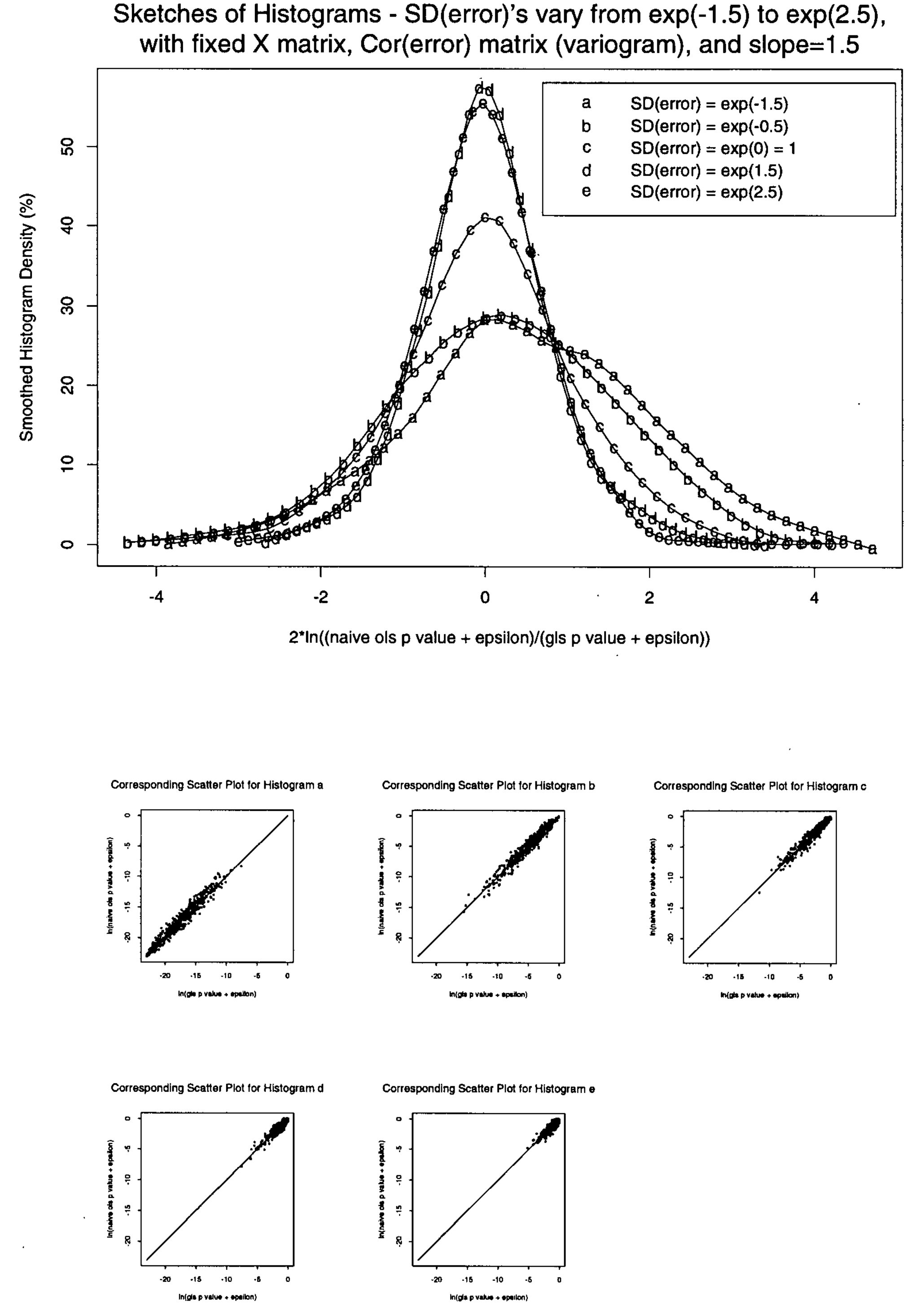
So far in this section, the results are based on simulation experiments with $\boldsymbol{\Sigma}_\epsilon$ that might produce a potentially noticeable difference between the two statisticians' analyses. What happens if $\boldsymbol{\Sigma}_\epsilon$ is chosen to be very close to the identity?

Naturally, we do not expect the GLS analysis to out-perform the naive OLS analysis. Afterall, the two analyses differ only in what the two statisticians believe $\boldsymbol{\Sigma}_\epsilon$ to be. Out of curiosity, we ran a few sets of simulations using $\boldsymbol{\Sigma}_\epsilon$ as the variogram structure described in item 5 of page 27. Recall that our choice of the region bearings produced a $\boldsymbol{\Sigma}_\epsilon$ which was very close to $\boldsymbol{I}$. $\boldsymbol{X}$ was again the 30×1 covariate vector that we used in the previous few sets of simulations. As expected, the histograms for these sets of simulations all locate at around 0, even under different sets of varying conditions. This indicates that in general, with such a $\boldsymbol{\Sigma}_\epsilon$, the GLS analysis is not noticeably more powerful than the naive OLS analysis.

Nonetheless, we have included the histograms and scatter plots for one of these sets of simulations in Figure 2.8. In these simulations, all program input arguments were fixed ($\beta_1 = 1.5$), except that σ varied from $\exp\{-1.5\}$ to $\exp\{2.5\}$. The other sets of simulations produced very similar graphs as Figure 2.8. Perhaps the only interesting observation of these simulations lies in the movement of the scatter cloud as a parameter varies.

In Figure 2.8, the scatter cloud moves along the 45-degree line as σ increases. It starts out at around (-20, -20), then spreads upwards along the line, until it concentrates at around (0, 0). The fact that the cloud is scattered around the 45-degree is because with $\boldsymbol{\Sigma}_\epsilon \approx \boldsymbol{I}$, both the GLS and the naive OLS analyses are about as powerful as each other. At small values of σ, the running shoes are of very good quality, and hence, both

Figure 2.8: Histogram and Scatter Plots with Variogram Σ_ϵ: σ varies.



analyses are powerful. Thus, the cloud is concentrated at highly negative co-ordinates (extremely small p-values). As σ increases, the quality of the running shoes worsens, and hence, the cloud moves up to near-zero co-ordinates (p-values close to 1).

2.9 Discussion

Recall that our main goal in the chapter was to investigate the loss of parameter estimation efficiency in a linear regression model if one ignores the response dependence. We have used the "storyline" of the two analysts to clarify the distinctions among three different approaches to a regression problem:

1. the naive OLS analysis: Naive simply assumes that the data are uncorrelated, estimates the regression parameters with $\widehat{\boldsymbol{\beta}}^{OLS}$ (OLS estimator), and ignores the response dependence in computing the covariance of this $\widehat{\boldsymbol{\beta}}^{OLS}$;

2. the GLS analysis: Enlightened knows the true response dependence structure, $\boldsymbol{\Sigma}_\epsilon$, and hence obtains $\widehat{\boldsymbol{\beta}}^{GLS}$ (GLS estimator) as the regression parameter vector estimate and bases his/her regression analysis on the knowledge of $\boldsymbol{\Sigma}_\epsilon$; and

3. the true OLS analysis: Enlightened realizes Naive's mistake in computing the covariance of $\widehat{\boldsymbol{\beta}}^{OLS}$, and carries out Naive's regression analysis, using the true OLS covariance, a function of $\boldsymbol{\Sigma}_\epsilon$.

Before our investigation, theory has been established on the loss of efficiency using the true OLS analysis compared to the GLS analysis. When $\boldsymbol{\Sigma}_\epsilon \neq \boldsymbol{I}$, that is, when the responses are correlated, the GLS estimator is the BLUE for $\boldsymbol{\beta}$ in model (2.1). When $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$, the GLS estimator reduces to the OLS estimator, and thus both are the BLUE's for $\boldsymbol{\beta}$. Theorem 1 indicates that for all $\boldsymbol{\Sigma}_\epsilon$, the GLS estimator is always at least as efficient as the OLS estimator, *based on the true covariance of the OLS estimator.* In other words, the true OLS analysis (regression analysis based on the OLS estimator and its true covariance) can never be more efficient than the GLS analysis. Of course, the two analyses are equivalent (as efficient as each other) when the two estimators are the same. However, we would like to know the conditions under which the two estimators are equivalent.

First, $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{I}$ is one such condition. It is not an interesting condition because we are trying to investigate cases where $\boldsymbol{\Sigma}_\epsilon \neq \boldsymbol{I}$. Some research has been done in this area since the 1960's. (See Section 2.2.) Among the complicated algebraic conditions which guarantee that $\widehat{\boldsymbol{\beta}}^{OLS} = \widehat{\boldsymbol{\beta}}^{GLS}$, one condition is readily verifiable. We have stated this condition in Theorem 2. It says that $\widehat{\boldsymbol{\beta}}^{OLS} = \widehat{\boldsymbol{\beta}}^{GLS}$ if and only if the model in (2.1) contains an intercept and that $\boldsymbol{\Sigma}_\epsilon$ is exchangeable.

However, Theorem 2 does not imply that under these conditions, the naive OLS analysis is equivalent to the GLS analysis. This brought us to our simulation studies.

In each simulation experiment, we generated data according to model (2.1) with some $\boldsymbol{\Sigma}_\epsilon \neq \boldsymbol{I}$. The three analyses (naive OLS, true OLS, and GLS) were compared. The analyses consisted of independent hypothesis tests of zero regression coefficients, each at 5% significance level. We compared the analyses by comparing their p-values of these hypothesis tests. Empirical power (sample probability of rejection of the null hypothesis) was calculated for each regression coefficient estimate for each analysis.

We found that in general, the stronger the dependence of the responses, the more apparent the GLS analysis' superiority in power. The naive OLS analysis is the least powerful, followed by the true OLS analysis, then by the GLS analysis. Even when the true OLS and the GLS analyses are equivalent, the naive OLS analysis can have a considerably lower power than the other two. We then decided to only compare the naive OLS analysis and the GLS analysis in the remaining part of our investigation.

We ran more sets of simulations. Empirical power curves of the GLS and naive OLS analyses were produced to display their powers graphically. They show that the GLS empirical power curve is almost[29] everywhere above that of the naive OLS analysis. As the size of the slope relative to the noise level (σ) either increases or decreases, the gap between the power curves always increases and then decreases.

In addition, instead of explicitly calculating the empirical power of each analysis, we graphically displayed their p-values using "logged p-value ratio histograms" and "logged p-value scatter plots." We could deduce which analysis was more powerful from these graphs. (An analysis is more powerful if its p-value is smaller.) We found that with all other program input arguments ($\boldsymbol{X}$, $\boldsymbol{\Sigma}_\epsilon$, and σ) fixed, the larger the regression coefficient in the model, the larger the "margin of Enlightened's victory" over Naive in their efficiency competition. On the other hand, with the regression coefficient and the other program input arguments fixed, a large σ (high noise level in the data) corresponds to a smaller margin of victory.

What if the ratio of the regression coefficient to the noise level in the data is fixed?

We ran sets of simulation experiments with this ratio fixed, and let the two parameters vary together, keeping their ratio constant. In these simulations, we found that the powers of the two analysts remain roughly constant as both parameters increase together at a fixed ratio. (We have seen this from both the empirical power graphs and the histogram sketches.) However, Enlightened's margin of victory varies according to the size of the coefficient-to-noise ratio, as can be seen by comparing the locations of the histogram sketches in Figures 2.5 to 2.7.[30] In general, when the coefficient is smaller than the noise level (σ), Enlightened barely wins over Naive. When the coefficient and the noise level are equal, Enlightened wins by a larger margin. Finally, if the coefficient is bigger than

[29] A higher naive OLS empirical power is due to chance.

[30] Figures 2.3 and 2.4 also show the effect of a varying slope-to-noise ratio.

the noise level, s/he wins "by a landslide."

2.10 CHAPTER APPENDIX

The following is the algorithm for the computer implemented procedure for generating the design matrix and regression coefficients in a simulation experiment of linear regression.

1. Generate the $(g+1) \times 1$ vectors $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\ldots$, $\boldsymbol{X}_r$ ($\boldsymbol{X}_i = (X_{i0}, X_{i1}, \ldots, X_{ig})^T$ for all i) independently from an MVN distribution with mean $\boldsymbol{\mu}_X$ and covariance $\boldsymbol{\Sigma}_X$ as specified by the program user. (r = number of regions.) Thus, all r of these $(g \times 1)$-vectors are i.i.d.

2. Write $\boldsymbol{\mu}_X = (\mu_0, \mu_1, \ldots, \mu_g)^T$, and

$$\boldsymbol{\Sigma}_X = \begin{pmatrix} \sigma_{00} & \boldsymbol{\sigma}^T_{\text{o},-\text{o}} \\ \boldsymbol{\sigma}_{\text{o},-\text{o}} & \boldsymbol{\sigma}_{-\text{o},-\text{o}} \end{pmatrix}$$

where

$$\begin{aligned} \sigma_{00} &= (\boldsymbol{\Sigma}_X)_{\text{oo}} \\ &= \text{Var}(X_{i0}) \\ & \text{(i.e. variance of the first element of any of the } (g+1)\text{-vectors),} \\ \boldsymbol{\sigma}_{\text{o},-\text{o}} &= \text{Cov}(X_{i0}, \boldsymbol{X}_{i,-\text{o}}) \\ &= [\text{Cov}(X_{i0}, X_{i1}),\ \text{Cov}(X_{i0}, X_{i2}),\ \ldots,\ \text{Cov}(X_{i0}, X_{ig})]^T \end{aligned}$$

where $\boldsymbol{X}_{i,-\text{o}} = (X_{i1}, X_{i2}, \ldots, X_{ig})^T$ for all $i = 1, 2, \ldots, r$, and

$$\boldsymbol{\sigma}_{-\text{o},-\text{o}} = \text{Cov}(\boldsymbol{X}_{i,-\text{o}}) = \text{the covariance matrix of } \boldsymbol{X}_{i,-\text{o}}.$$

3. According to the above, we thus have

$$[X_{i0} | \boldsymbol{X}_{i,-\text{o}} = \boldsymbol{x}] \sim \text{N}(\nu, \psi^2) \tag{2.15}$$

where

$$\begin{aligned}\boldsymbol{x} &= (x_1, x_2, \ldots, x_g)^T \\ \nu &= \mu_0 + (\boldsymbol{\sigma}_{\text{o},-\text{o}})^T(\boldsymbol{\sigma}^{-1}_{-\text{o},-\text{o}})(\boldsymbol{x} - \boldsymbol{\mu}_{-\text{o}}) \qquad (2.16) \\ \psi^2 &= \sigma_{00} - \ (\boldsymbol{\sigma}_{\text{o},-\text{o}})^T(\boldsymbol{\sigma}^{-1}_{-\text{o},-\text{o}})(\boldsymbol{\sigma}_{\text{o},-\text{o}}).\end{aligned}$$

(See Johnson & Wichern, 1982, Result 4.6 for the derivation of ν and ψ^2.)

4. If we let $\boldsymbol{\theta}_{-\text{o}} = (\theta_1, \ldots, \theta_g)^T$ be $(\boldsymbol{\sigma}_{\text{o},-\text{o}})^T(\boldsymbol{\sigma}^{-1}_{-\text{o},-\text{o}})$, then we see that (2.15) itself specifies a linear regression model with X_{i0} as the response, x_j's as the covariates, and θ_j's as the slopes. From equation (2.16), we derive the intercept θ_0 in this regression model as
$$\theta_0 = \mu_0 - (\boldsymbol{\theta}_{-\text{o}})^T \boldsymbol{\mu}_{-\text{o}}.$$

5. Now, we extract from the regression model specified by (2.15) its covariates and regression coefficients, and specify these as the corresponding program input arguments in (2.1). That is, in the case that model (2.1) has no intercepts, we let the $p \times g$ covariate matrix and the g-vector of slope coefficients to be $\boldsymbol{X} = (\boldsymbol{X}_{1,-\text{o}}, \boldsymbol{X}_{2,-\text{o}}, \ldots, \boldsymbol{X}_{g,-\text{o}})^T$ and $\boldsymbol{\beta} = \boldsymbol{\theta}_{-\text{o}}$. If we fit an intercept in model (2.1), we simply add a column of 1's to this $\boldsymbol{X}$. The $(g+1)$-vector of regression coefficients (including the intercept) in model (2.1) then becomes $\boldsymbol{\beta} = \boldsymbol{\theta} = (\theta_0, \boldsymbol{\theta}_{-\text{o}})^T$.

Chapter 3

Introduction to Multiple Regression of Count Data

In the previous chapter, we have seen that ignoring response dependence may result in serious reduction of efficiency in parameter estimation in a linear regression model with normally distributed random errors. In real life, however, the data we need to analyze may not be normally distributed. For example, in health care studies, hospital admission counts are often used to indicate the health of a population. These counts are usually small (less than 20) for rare diseases, and thus should not be assumed to follow a normal distribution. In such cases, linear models may no longer be efficient in modeling the relationship between counts and possible covariates. In fact, small admission counts are usually assumed to follow a Poisson distribution in health care studies. As a result, these counts are often analyzed using Poisson regression. However, theoretical results are generally not widely available for models with non-Gaussian responses, especially when the responses are correlated. Perhaps this is why health care studies in the past generally ignored correlation among responses.

In 1986, Liang & Zeger proposed the *generalized estimating equations* (GEE) approach to incorporate response correlation for estimating parameters in generalized linear models (GLM's). A Poisson regression model, for example, is a GLM. The generalized estimating equations are an extension of the estimating equations in quasi-likelihood theory, implemented in the widely used GLIM software. Response independence is assumed in the original quasi-likelihood approach, whereas the extension (GEE) allows the incorporation of any response correlation structure in the GLM. GEE has become a popular model fitting method in the past decade for dealing with correlated responses in GLM's, ever since its computer implementation became readily available. The approach is widely used in longitudinal data analyses in which the repeated responses over time are correlated within each independent subject (given the covariates).

When the response correlation matrix or structure is known, it is natural to incorporate this known matrix/structure into the regression model when using GEE to fit the model. However, misspecification of response correlation results when an investigator incorporates an incorrect matrix/structure into the model while believing it to be true. For example, ignoring response correlation (i.e. believing that the responses are independent) is a misspecification of the true correlation structure.

In the next two chapters, we examine the impact of misspecification of the response correlation structure on parameter estimation in multiple regression models for count data using GEE. Our investigation is based mainly on computer simulations.

The first part of our simulations is presented in Chapter 4. In these simulations, we generate count data through Poisson-lognormal distributions. Conditioned on the covariates, the data are independent Poisson counts. Marginally, the counts have a Poisson-lognormal distribution, and are correlated through the correlation among the covariates. The GEE approach is used to fit the data using correct and incorrect correlation structures. We then examine the performance of the approach with the different structures, through its empirical power (of detecting non-zero regression coefficients). We can use this empirical power as a measure of parameter estimation efficiency.

Recall from the previous chapter the notion of the model assuming independent responses (while the responses are correlated) and that assuming correlated responses with a known (correlation) structure. In the next two chapters, the *independence model* refers to the GLM in which the responses are assumed independent. The *dependence model*, on the other hand, is the GLM in which the responses are assumed to have certain correlation structure. (See Section 3.4 for detailed explanations of terminology.) These two *models* are then fitted to the data using GEE, and their efficiencies (empirical powers) are compared to determine their relative efficiency.

We present the second part of our simulations in Chapter 5. In this chapter, we concentrate on studying the impact of ignoring response correlation on parameter estimation in a GLM. We generate exchangeably correlated Poisson data. Generating correlated Poisson data with non-linear relationship to the covariates is generally a difficult task. Instead, the Poisson data that we generate are linearly related to the intercept — the only regression coefficient — without true covariates in the model.

The following section is a summary of literature references for Chapters 4 and 5. In the remainder of this chapter, we describe the software we use for our simulations in Section 3.2, the theory of the GEE approach in Section 3.3, and the terminology we use throughout the next two chapters in Section 3.4.

3.1 Literature Review

Since the development of the GEE theory by Liang & Zeger (1986), many longitudinal data studies involving non-Gaussian data have been carried out through the GEE approach for model fitting. For example, Zeger & Liang (1986) applied their theory on a dichotomous dataset (response and covariates) intended to describe the association of a mother's stress and her child's morbidity. Apparently, the GEE approach for model fitting with such data types is simple and easy to use.

However, the application of the GEE approach requires *data replication.* For example, the time series measurements on each individual subject in a longitudinal data analysis constitute one set of data replication, or a *cluster.* When there is only one correlated series in the study, such as the case in the previous chapter, the GEE approach is not appropriate for model fitting.

In the case of count data for a single time series, Zeger (1988) developed a "parameter-driven regression model for time series of counts." In his paper, he mentions both a *parameter-driven* model, in which autocorrelation of the counts (Poisson) is introduced through a latent process $\{\epsilon_t\}$, and an *observation-driven* model, in which the distribution of the current count observation is a function of all the past counts. He concentrates on the parameter-driven model in this paper, and develops an estimating equation approach for parameter estimation. Simulations are done to compare the efficiency of the parameter-driven model with two common log-linear models. Results show that the parameter-driven model leads to "more valid inferences."

The effect of ignoring correlation among response variables on the efficiency of regression parameter estimation in non-linear models has been studied in different contexts, mostly involving longitudinal data analyses using the GEE approach. In some of these "GEE studies," the dependence model is shown to be more efficient than the independence model in parameter estimation. It is hence believed that the model that takes into account the correlation among the responses should be more efficient than the model which does not. However, the effect of ignoring response correlation on the efficiency of parameter estimation is still not clearly understood because such effect varies across different contexts. Fitzmaurice (1995) mentioned that different research papers concluded different findings about the relative efficiency of the independence model to the dependence model. Some conclude that the latter is clearly more efficient than the former, while others conclude that the former is nearly as efficient as the latter. Fitzmaurice illustrates in his paper that, with binary data, this relative efficiency depends on how strongly the responses are correlated, and on the covariate design. In particular, with a fixed strength of response correlation, the GEE independence model (the model fitted using GEE assuming independent responses) is noticeably less efficient than the GEE dependence model (the model fitted using GEE with a correctly specified response correlation structure) when a covariate varies within the cluster. On the other hand, ignoring response dependence does not lead to loss of efficiency in parameter estimation using GEE if "group is a cluster-level covariate."

Inspired by health care studies using hospital admission counts as responses, here we study the impact of misspecifying the *spatial correlation* on parameter estimation. In this context, the correlated "series" is relatively short compared to those in a longitudinal study. This is because the spatially correlated responses are collected from the regions of interest, which are often of a limited number. Since the implementation of the GEE algorithm is readily available, we will not try to implement the algorithm for fitting a parameter-driven model as developed by Zeger (1988). Instead, we use the S-Plus function gee in our regression simulations with count data.

Data replication is present in our simulation studies so that the GEE approach can be applied. Each spatially correlated series is analogous to the correlated repeated measurements (on each subject) over time in a longitudinal data analysis. The analog of the independent subjects in a longitudinal study is the *independent replicates*, i.e. the repeated observations (of the response and explanatory variables), within each region. Thus, instead of treating the subjects as the means of data replication in a longitudinal study, we can independently observe replicates of the response and covariates in each region in the context of spatially correlated data. For example, datasets collected by two independent investigators can be viewed as two sets of data replicates. Datasets observed at different time points far apart (such as a year) from each other may also be thought of as independent data replicates.

3.2 GEE in S-Plus

In using the GEE method to fit a model to correlated responses with fixed covariates, we need to specify the response correlation structure. The S-Plus function `gee` asks for either a correlation structure (e.g. exchangeable, AR(1)) or a fixed correlation matrix for these responses as one of the input arguments. If we specify the correlation structure, then `gee` will estimate the parameters in the response correlation matrix with the specified structure. (For example, the intraclass correlation coefficient is the parameter for the exchangeable structure.) If we specify a fixed correlation matrix, then `gee` will treat it as known and fixed throughout the parameter estimation iterations.

In addition to the response correlation structure, using `gee` also requires the specification of the link function and the response distribution family (given the covariates) as a means to specify the relationship between the mean and variance of each response. `gee` then outputs the parameter estimates and their standard error estimates, with the final estimate of the response correlation matrix if it was not fixed as an input argument.

This S-Plus function `gee` version 3.5 that we use in our simulations was obtained from the share-ware organization, *StatLib*, and then installed into S-Plus version 3.1. This function allows the user to specify the following response correlation structures:

1. `independence` (in which case the `gee` estimates of the regression coefficients are equivalent to those produced by the GLIM software),

2. `fixed` (user-specified/known correlation matrix),

3. `exchangeable`,

4. `AR-M` (with a user-specified M),

5. `stat_M_dep` (stationary M-dependence, with a user-specified M),

6. `non_stat_M_dep` (non-stationary M-dependence), and

7. `unstructured` (gee estimates the structure and the values in the working correlation matrix).

(For convenience, we do not apply items 5 and 6 in the above list to our simulation studies.) Notice that if we specify the "independence structure" in gee, the GEE coefficient estimates are equivalent to those produced by the usual GLIM software. Also, the independence structure is a special case of both the exchangeable and the AR(1) structures. For the exchangeable structure, a 0 intraclass correlation coefficient gives the independence structure. Similarly, a 0 lag one correlation reduces the AR(1) structure to the independence structure. If the responses are independent (given the covariates), one would expect the independence model to provide very similar results as those of the dependence model with an exchangeable or AR(1) structure. On the other hand, fixing a correlation matrix that is very different from the identity matrix in the dependence model may yield very different results from those of the independence model. In particular, one would expect the estimation efficiency (in terms of bias and power) of the dependence model to be lower than that of the independence model in this case, simply because the structure is incorrectly specified in the former model. This will be closely investigated in the next chapter. Similarly, we will closely investigate in Chapter 5 the efficiencies of the two models when the responses are exchangeably correlated.

3.3 Some Notation and Theory in GEE

In this section, we present the applicable results in Liang & Zeger, 1986 in the context of spatially correlated responses.

First, assume that the geographical domain of interest consists of r regions. Denote the response vector by $\boldsymbol{Y} = (Y_1, \ldots, Y_r)^T$. For each of the r regions, we observe the same

g covariates. In the ith region, denote the g-dimensional covariate vector for Y_i by $\boldsymbol{X}_{i.} = (X_{i1}, X_{i2}, \ldots, X_{ig})^T$. Now, we can write the $r \times g$ covariate/design matrix $\boldsymbol{X}$ as

$$\boldsymbol{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1g} \\ X_{21} & X_{22} & \cdots & X_{2g} \\ \vdots & \cdots & \cdots & \vdots \\ X_{r1} & X_{r2} & \cdots & X_{rg} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_{1.}^T \\ \boldsymbol{X}_{2.}^T \\ \vdots \\ \boldsymbol{X}_{r.}^T \end{pmatrix} = (\boldsymbol{X}_{.1}, \boldsymbol{X}_{.2}, \ldots, \boldsymbol{X}_{.g}).$$

The relationship between the responses and the covariates is expressed in terms of the generalized linear model

$$\mathcal{G}\left[\mathrm{E}(Y_i|\boldsymbol{X}_{i.})\right] = \boldsymbol{X}_{i.}^T\boldsymbol{\beta} \qquad \text{for all } i = 1, \ldots, r$$

for some *link function* $\mathcal{G}$ and fixed parameters $\boldsymbol{\beta}$.

To apply the GEE approach in estimating $\boldsymbol{\beta}$, we need data replication.

Let $\{\boldsymbol{Y}^{(s)} : s = 1, 2, \ldots, K\}$ be a random sample of $r \times 1$ response vectors from the spatial domain of interest. Associated with this sample is the set of covariate matrices, $\{\boldsymbol{X}^{(s)} : s = 1, 2, \ldots, K\}$. That is, we have K independent sets of replicates of the response vector and the covariate matrix. In this set up, it is natural to assume that the K replicates of the responses given the covariates, $[\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)}]$'s, all have a common correlation matrix. Denote this matrix by $\mathrm{Cor}(\boldsymbol{Y}|\boldsymbol{X})$. In other words, $\mathrm{Cor}(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)})$ is the same $\mathrm{Cor}(\boldsymbol{Y}|\boldsymbol{X})$ matrix for all s.

Now, define the $r \times 1$ vector $\boldsymbol{\eta}^{(s)}$ to be

$$\boldsymbol{\eta}^{(s)} = \begin{pmatrix} \eta_1^{(s)} \\ \eta_2^{(s)} \\ \vdots \\ \eta_r^{(s)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_{1.}^{(s)T}\boldsymbol{\beta} \\ \boldsymbol{X}_{2.}^{(s)T}\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{X}_{r.}^{(s)T}\boldsymbol{\beta} \end{pmatrix} = \boldsymbol{X}^{(s)}\boldsymbol{\beta}.$$

If, for all s, the conditional density of the ith response given the covariates, $[Y_i^{(s)}|\ \boldsymbol{X}_{i.}^{(s)}]$, can be written as

$$f\left(y_i^{(s)}|\boldsymbol{x}_{i.}^{(s)}\right) = \exp\left\{y_i^{(s)}\theta_i^{(s)} - a(\theta_i^{(s)}) + b(y_i^{(s)})\phi\right\} \tag{3.17}$$

where $\theta_i^{(s)} = h(\eta_i^{(s)})$, for some functions a, b, and h, then

$$\begin{aligned} \mathrm{E}\left(Y_i^{(s)}|\boldsymbol{X}_{i.}^{(s)}\right) &= a'(\theta_i^{(s)}) \\ \mathrm{Var}\left(Y_i^{(s)}|\boldsymbol{X}_{i.}^{(s)}\right) &= \tfrac{a''(\theta_i^{(s)})}{\phi} \end{aligned} \tag{3.18}$$

where a' and a'' are the first and second derivatives of a respectively. ϕ is called the *scale* or the *dispersion* parameter. Note that the link function $\mathcal{G}$ is related to h in the following manner:

$$\begin{aligned} \mathcal{G}\left[\mathrm{E}(Y_i^{(s)}|\boldsymbol{X}_{i.}^{(s)})\right] = \mathcal{G}\left(\frac{\mathrm{d}a}{\mathrm{d}\theta_i^{(s)}}\right) &= \boldsymbol{X}_{i.}^{(s)T}\boldsymbol{\beta} \\ &= \eta_i^{(s)} = h^{-1}\left(\theta_i^{(s)}\right) \end{aligned}$$

The $r \times r$ matrices $\boldsymbol{A}^{(s)}$, $\boldsymbol{\Delta}^{(s)}$, and $\boldsymbol{V}^{(s)}$ are defined as

$$\boldsymbol{A}^{(s)} = \begin{pmatrix} a''(\theta_1^{(s)}) & 0 & \cdots & \cdots & 0 \\ 0 & a''(\theta_2^{(s)}) & \ddots & & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a''(\theta_{r-1}^{(s)}) & 0 \\ 0 & \cdots & \cdots & 0 & a''(\theta_r^{(s)}) \end{pmatrix}$$

$$\boldsymbol{\Delta}^{(s)} = \begin{pmatrix} \frac{\mathrm{d}\theta_1^{(s)}}{\mathrm{d}\eta_1^{(s)}} & 0 & \cdots & \cdots & 0 \\ 0 & \frac{\mathrm{d}\theta_2^{(s)}}{\mathrm{d}\eta_2^{(s)}} & \ddots & & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \frac{\mathrm{d}\theta_{r-1}^{(s)}}{\mathrm{d}\eta_{r-1}^{(s)}} & 0 \\ 0 & \cdots & \cdots & 0 & \frac{\mathrm{d}\theta_r^{(s)}}{\mathrm{d}\eta_r^{(s)}} \end{pmatrix}$$

$$\boldsymbol{V}^{(s)} = \frac{[\boldsymbol{A}^{(s)}]^{\frac{1}{2}}\boldsymbol{R}[\boldsymbol{A}^{(s)}]^{\frac{1}{2}}}{\phi} \tag{3.19}$$

where $\boldsymbol{R}$ is any $r \times r$ matrix that satisfies the requirements of being a correlation matrix. This $\boldsymbol{R}$ is the *working correlation matrix.* Similarly, $\boldsymbol{V}^{(s)}$ is the *working covariance matrix* for the sth replicate. Note that $\boldsymbol{V}^{(s)}$ is the true covariance matrix of $[\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)}]$ (i.e. $\boldsymbol{V}^{(s)} = \text{Cov}\left(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)}\right)$) if $\boldsymbol{R}$ is the true correlation matrix of each $[\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)}]$. Also note that since all $[\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)}]$'s have the common correlation matrix Cor($\boldsymbol{Y}|\boldsymbol{X}$), the working correlation matrix $\boldsymbol{R}$ (which may be used to estimate Cor($\boldsymbol{Y}|\boldsymbol{X}$)) does not depend on s.

Next, define the $r \times g$ matrix

$$\boldsymbol{D}^{(s)} = \boldsymbol{A}^{(s)}\boldsymbol{\Delta}^{(s)}\boldsymbol{X}^{(s)}.$$

Then the *generalized estimating equations* are defined as

$$\sum_{s=1}^{K} \boldsymbol{D}^{(s)T}[\boldsymbol{V}^{(s)}]^{-1}[\boldsymbol{Y}^{(s)} - \boldsymbol{a}'(\boldsymbol{\theta}^{(s)})] = \boldsymbol{0} \tag{3.20}$$

where $\boldsymbol{a}'(\boldsymbol{\theta}^{(s)}) = (a'(\theta_1^{(s)}), \cdots, a'(\theta_r^{(s)}))^T$. The estimate of $\boldsymbol{\beta}$ is then obtained by solving (3.20) for $\boldsymbol{\beta}$.

Since the quantities $\boldsymbol{D}^{(s)}$, $\boldsymbol{V}^{(s)}$, and $\boldsymbol{a}'(\boldsymbol{\theta}^{(s)})$ in (3.20) are all functions of some unknown parameters, we need an iterative numerical method for *updating* (at each iteration) each unknown parameter estimate so that these quantities in (3.20) are consequently *updated.* Upon convergence, we obtain the final estimates of these unknown parameters such as $\boldsymbol{\beta}$, and those in the working correlation matrix $\boldsymbol{R}$. If the working correlation structure (i.e. the structure of $\boldsymbol{R}$) is specified according to the true response correlation structure, then this final "estimate" of $\boldsymbol{R}$ is the estimate of the conditional correlation matrix of the response $\boldsymbol{Y}$, i.e. Cor($\boldsymbol{Y}|\boldsymbol{X}$).[1]

[1] In our investigation, we assume that the initial specification of $\boldsymbol{R}$ as the working correlation matrix is based on the believed structure of Cor($\boldsymbol{Y}|\boldsymbol{X}$). The final estimate of $\boldsymbol{R}$ may not be a good estimate of Cor($\boldsymbol{Y}|\boldsymbol{X}$) if its specified structure does not coincide with that of Cor($\boldsymbol{Y}|\boldsymbol{X}$).

More advanced statistical packages such as S-Plus provide standard procedures for fitting GLM's using the GEE approach. Notice that if $\boldsymbol{R} = \boldsymbol{I}$, i.e. assuming[2] the responses to be independent (given the covariates), the estimating equations (3.20) reduce to those derived from quasi-likelihood theory. In other words, the GEE parameter estimates reduce to those of the ordinary GLIM software approach when $\boldsymbol{R} = \boldsymbol{I}$.

The following theorem is a re-statement of that presented in Liang & Zeger (1986):

Theorem 3 *Suppose regularity conditions hold, and that there exist K-consistent estimators for ϕ and the parameters in $\boldsymbol{R}$. Define $\widehat{\boldsymbol{\beta}}^G$ to be the solution of $\boldsymbol{\beta}$ in (3.20), with ϕ and the parameters in $\boldsymbol{R}$ being replaced by their estimators, if they were not fully specified. Then $\sqrt{K}\ (\widehat{\boldsymbol{\beta}}^G - \boldsymbol{\beta})$ has an asymptotic MVN distribution with mean $\mathbf{0}$ and covariance*

$\lim_{K\to\infty} \mathrm{Cov}[\sqrt{K}\ (\widehat{\boldsymbol{\beta}}^G - \boldsymbol{\beta})] =$
$\lim_{K\to\infty} K \left(\sum_{s=1}^K \boldsymbol{D}^{(s)T}[\boldsymbol{V}^{(s)}]^{-1}\boldsymbol{D}^{(s)}\right)^{-1} \left\{\sum_{s=1}^K \boldsymbol{D}^{(s)T}[\boldsymbol{V}^{(s)}]^{-1}\mathrm{Cov}(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)})[\boldsymbol{V}^{(s)}]^{-1}\boldsymbol{D}^{(s)}\right\}$
$\left(\sum_{s=1}^K \boldsymbol{D}^{(s)T}[\boldsymbol{V}^{(s)}]^{-1}\boldsymbol{D}^{(s)}\right)^{-1}.$

In a GEE analysis, the above theorem is used for estimating the covariance of $\widehat{\boldsymbol{\beta}}^G$, in the following manner:

$$\begin{aligned}
\mathrm{Cov}(\widehat{\boldsymbol{\beta}}^G) &= \frac{K}{K}\mathrm{Cov}(\widehat{\boldsymbol{\beta}}^G - \boldsymbol{\beta}) \\
&= \frac{1}{K}\mathrm{Cov}[\sqrt{K}(\widehat{\boldsymbol{\beta}}^G - \boldsymbol{\beta})] \\
&\approx \left(\sum_{s=1}^K \boldsymbol{D}^{(s)T}[\boldsymbol{V}^{(s)}]^{-1}\boldsymbol{D}^{(s)}\right)^{-1} \left\{\sum_{s=1}^K \boldsymbol{D}^{(s)T}[\boldsymbol{V}^{(s)}]^{-1}\mathrm{Cov}(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)})[\boldsymbol{V}^{(s)}]^{-1}\boldsymbol{D}^{(s)}\right\} \\
&\quad \left(\sum_{s=1}^K \boldsymbol{D}^{(s)T}[\boldsymbol{V}^{(s)}]^{-1}\boldsymbol{D}^{(s)}\right)^{-1} \qquad (3.21)
\end{aligned}$$

Recall that $\boldsymbol{R}$ and $\boldsymbol{V}^{(s)}$ are respectively the working correlation matrix and the working covariance matrix for the responses $\boldsymbol{Y}^{(s)}$ given the covariates $\boldsymbol{X}^{(s)}$. If we assume $\boldsymbol{V}^{(s)}$

[2] See Section 3.4.

to be the true covariance matrix $\text{Cov}(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)})$, then equation (3.21) reduces to

$$\text{Cov}(\widehat{\boldsymbol{\beta}}^G) \approx \left(\sum_{s=1}^{K} \boldsymbol{D}^{(s)T}[\boldsymbol{V}^{(s)}]^{-1}\boldsymbol{D}^{(s)}\right)^{-1}. \qquad (3.22)$$

The GEE estimate of the covariance of $\widehat{\boldsymbol{\beta}}^G$ using equation (3.22) is called the *naive* covariance estimate. The one using equation (3.21) is known as the *robust* covariance estimate of $\widehat{\boldsymbol{\beta}}^G$. In practice, we usually do not know the true correlation structure of the responses. The working correlation, $\boldsymbol{R}$, which we specify in running the GEE analysis, is only based on the belief of what the true structure is. The *naive* covariance estimate might then be an inappropriate estimate of the true covariance of $\widehat{\boldsymbol{\beta}}^G$ if $\boldsymbol{R}$ is specified as very different from the true $\text{Cor}(\boldsymbol{Y}|\boldsymbol{X})$.[3] On the other hand, the *robust* covariance estimate accounts for the true covariance of the responses, $\text{Cov}(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)})$, and thus is a "corrected version" of the *naive* covariance estimate.

Of course, $\text{Cov}(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)})$ is usually unknown. In particular, misspecification of the response correlation structure is usually a result of not knowing the true $\text{Cor}(\boldsymbol{Y}|\boldsymbol{X})$. In fact, $\text{Cov}(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)})$ in the *robust* variance estimate formula is estimated from the data using $[\boldsymbol{Y}^{(s)} - \boldsymbol{a}'(\widehat{\boldsymbol{\theta}}^{(s)})][\boldsymbol{Y}^{(s)} - \boldsymbol{a}'(\widehat{\boldsymbol{\theta}}^{(s)})]^T$. Here, $\boldsymbol{a}'(\widehat{\boldsymbol{\theta}}^{(s)})$ is computed using $\boldsymbol{X}^{(s)}$ and $\widehat{\boldsymbol{\beta}}^G$, for each s. Since the observed data should be representative of the population, the estimate of $\text{Cov}(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)})$ based on the data should usually resemble the true covariance more than $\boldsymbol{V}^{(s)}$ does (when $\text{Cor}(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)})$ is misspecified). Hence, the *robust* estimate is often a better estimate of the covariance of $\widehat{\boldsymbol{\beta}}^G$.

[3] That is, the asymptotic $\text{Cov}(\widehat{\boldsymbol{\beta}}^G)$ does not equal the *naive* covariance estimate if the structure of $\boldsymbol{R}$ is different from that of $\text{Cor}(\boldsymbol{Y}|\boldsymbol{X})$.

3.4 Terminology

In the next two chapters, we will assume the scenario of an investigator specifying the working correlation matrix/structure in GEE according to what s/he believes the response correlation matrix/structure to be. **Misspecification** of the response correlation refers to the case when the investigator's belief/assumption (and thus specification) of the response correlation is incorrect when using GEE for model fitting.

We will refer to the GLM in which the responses are assumed to be independent as the **independence model**. Similarly, we will refer to the GLM in which a correlation structure of the response dependence is taken into account as the **dependence model**. In the simulations, we fit both models to a given dataset using the GEE approach. Notice that either model may be incorrect. For example, suppose the true response correlation is exchangeable. Then the independence model is incorrect, but the dependence model that assumes exchangeably correlated responses is correct. On the other hand, the dependence model is incorrect if it assumes the response correlation structure to be anything other than exchangeable. In other words, the model is only correct if the specified correlation structure is the true response correlation structure.

We say that one model is more **efficient** than another if its parameter estimates produced by the GEE approach are closer to the true parameters, and if these estimates have smaller variability. One way to measure unbiasedness and variability of a parameter estimate is to observe many realizations of the estimate. For each simulation, we repeat the GEE model fitting process to n datasets. Summary statistics of these n parameter estimates are then obtained. Unbiasedness can be measured by how close the sample mean of these n estimates is to the true parameter. Similarly, variability can be measured by the sample variance of the n estimates.

However, small variability in a parameter estimate also means higher power in detecting a non-zero value of the parameter. This leads us to the notion of **empirical power** of the independence and dependence models. The power of detecting a particular non-zero regression parameter can be measured by the proportion of the n datasets in which the parameter is statistically significant at a given level. A model is then also more efficient than the other if it is empirically more powerful than the other in detecting a given parameter from the same dataset.

As in the previous chapter, we may wish to determine if the difference in power between the independence and dependence models is statistically significant. We will again use the Pearson's χ^2 test of independence to assess the association between model (independence/dependence) and power ($\hat{\beta}_j$ significant or not) for each regression coefficient in the model. (See page 33 for examples of the χ^2 test.) If the association is significant, we conclude that **the difference in** (empirical) **power is significant.** Otherwise, **the difference in power is likely due to chance.**[4]

[4]... in the terminology of Freedman, Pisani, Purves & Adhikari, 1991, Chapter 26.

Chapter 4

Simulations of Multiple Regression of Count Data — Part One

In this part of our simulations, we generate count data which are independently Poisson distributed when given the covariates. (Marginally, the responses are correlated with a Poisson-lognormal distribution.) We are interested in the effect of misspecifying a response correlation structure on parameter estimation using GEE. Thus, any non-independence structure is a misspecification.

4.1 Generating Count Data and their (Random) Covariates

Recall from Section 3.4 the notation for the response vector $\boldsymbol{Y}^{(s)}$ and the covariate matrix $\boldsymbol{X}^{(s)}$ for the sth replicate. Fixing s, the equations in this section hold for all $s = 1, \ldots, K$. In other words, each set of replicates $\{\boldsymbol{Y}^{(s)}, \boldsymbol{X}^{(s)}\}$ are independent and identically distributed (i.i.d.). Thus, (4.1) to (4.7) hold for all s.

We can generate $Y_i^{(s)}$'s through generating covariate vectors $\boldsymbol{X}_{i.}^{(s)}$'s. In a given region i, the g-dimensional covariate vector is $\boldsymbol{X}_{i.}^{(s)}$, and we assume that the jth covariate is distributed as a normal random variable with mean μ_{ij} and variance σ_{ij}. Also, within the ith region, the jth and the lth covariates are assumed to be uncorrelated, i.e. $\text{Cov}(X_{ij}^{(s)}, X_{il}^{(s)}) = 0$, for all $j \neq l$. Fixing j, we assume that the jth covariate in the ith region and that in the kth region have covariance σ_{ik}. In matrix notation, we have the $r \times 1$ covariate vector

$$\boldsymbol{X}_{.j}^{(s)} = \begin{pmatrix} X_{1j}^{(s)} \\ X_{2j}^{(s)} \\ \vdots \\ X_{rj}^{(s)} \end{pmatrix} \sim \mathrm{MVN}(\boldsymbol{\mu}_{.j}, \boldsymbol{\Sigma}) \text{ for all } j \tag{4.1}$$

where

$$\boldsymbol{\mu}_{.j} = \begin{pmatrix} \mu_{1j} \\ \mu_{2j} \\ \vdots \\ \mu_{rj} \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1r} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2r} \\ \vdots & & \ddots & & \vdots \\ \sigma_{1r} & \cdots & \cdots & \cdots & \sigma_{rr} \end{pmatrix},$$

and that $\boldsymbol{X}_{.j}^{(s)}$ and $\boldsymbol{X}_{.l}^{(s)}$ are independent vectors for all $j \neq l$.

In health impact studies of air pollution, for instance, this distribution for the covariate vectors $\boldsymbol{X}_{.j}^{(s)}$'s (for a given s) is reasonable where the log-scale pollutant concentration often follows a normal distribution. Moreover, given an explanatory variable, observed across all regions, we expect the readings in the different regions to be spatially correlated. To reduce the complexity of our simulation study to computationally realistic levels, we assume that the readings across all regions are spatially correlated in the same way for each explanatory variable. Hence, we have the same covariance matrix $\boldsymbol{\Sigma}$ for all j in (4.1). For the same reason, we assume the readings of the different explanatory variables within a region to be independent to avoid singularity problems in model fitting due to multi-collinearity in the explanatory variables.

Now, we can define the expectation of $\boldsymbol{X}^{(s)}$ in the matrix form

$$\boldsymbol{\mu} = \mathrm{E}\left(\boldsymbol{X}^{(s)}\right) = \begin{pmatrix} \boldsymbol{\mu}_{1.}^{T} \\ \boldsymbol{\mu}_{2.}^{T} \\ \vdots \\ \boldsymbol{\mu}_{r.}^{T} \end{pmatrix} = (\boldsymbol{\mu}_{.1}, \boldsymbol{\mu}_{.2}, \ldots, \boldsymbol{\mu}_{.g}).$$

Again, let the $r \times 1$ response vector be $\boldsymbol{Y}^{(s)} = (Y_1^{(s)}, \ldots, Y_r^{(s)})^T$, where for the ith region, the response (e.g. hospital admission count) has the conditional Poisson distribution

$$[Y_i^{(s)} | \boldsymbol{X}_{i.}^{(s)}] \overset{\text{ind}}{\sim} \text{Poisson}\left(\exp\{\boldsymbol{X}_{i.}^{(s)T}\boldsymbol{\beta}\}\right) \tag{4.2}$$

for some $g \times 1$ coefficient vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_g)^T$, for $i = 1, 2, \ldots, r$. That is, conditional on the g covariates, the response in the ith region has a Poisson distribution with mean $\exp(\boldsymbol{X}_{i.}^{(s)T}\boldsymbol{\beta})$.

Note that we generate the $Y_i^{(s)}$'s so that $[Y_i^{(s)} \; | \boldsymbol{X}_{i.}^{(s)}]$ is independent of $[Y_k^{(s)} | \boldsymbol{X}_{k.}^{(s)}]$ for $i \neq k$.

Also note that $\boldsymbol{X}_{i.}^{(s)T}\boldsymbol{\beta} = \sum_{j=1}^{g} \beta_j X_{ij}^{(s)}$ is a normal random variable with mean $\boldsymbol{\mu}_{i.}^T\boldsymbol{\beta}$ and variance $\sum_{j=1}^{g} \beta_j^2 \sigma_{ii} = \sigma_{ii} \, \boldsymbol{\beta}^T\boldsymbol{\beta}$, and that

$$\begin{aligned} \text{Cov}(\boldsymbol{X}_{i.}^{(s)T}\boldsymbol{\beta}, \boldsymbol{X}_{k.}^{(s)T}\boldsymbol{\beta}) &= \text{Cov}(\sum_{j=1}^{g} X_{ij}^{(s)}\beta_j, \sum_{l=1}^{g} X_{kl}^{(s)}\beta_l) \\ &= \sum_{j=1}^{g} \beta_j^2 \text{Cov}(X_{ij}^{(s)}, X_{kj}^{(s)}) \\ &= \sum_{j=1}^{g} \beta_j^2 \sigma_{ik} \\ &= \sigma_{ik}\boldsymbol{\beta}^T\boldsymbol{\beta} \end{aligned}$$

for all $i \neq k$. Thus, letting $\boldsymbol{\nu} = \boldsymbol{\mu}\boldsymbol{\beta}$ and $\boldsymbol{W} = \boldsymbol{\beta}^T\boldsymbol{\beta}\boldsymbol{\Sigma}$, we have

$$\boldsymbol{X}^{(s)}\boldsymbol{\beta} \sim \text{MVN}(\boldsymbol{\nu}, \boldsymbol{W}). \tag{4.3}$$

Now, let

$$Z_i^{(s)} = \text{E}(Y_i^{(s)} | \boldsymbol{X}_{i.}^{(s)}) = \exp(\boldsymbol{X}_{i.}^{(s)T}\boldsymbol{\beta}).$$

Therefore, $\boldsymbol{Z}^{(s)} = \exp(\boldsymbol{X}^{(s)}\boldsymbol{\beta})$ is an r-dimensional lognormal random vector with parameters $\boldsymbol{\nu}$ and $\boldsymbol{W}$. Hence, marginally (unconditional on $\boldsymbol{X}^{(s)}$), $\boldsymbol{Y}^{(s)}$ has a multivariate

Poisson-lognormal distribution (Aitchison & Ho, 1989) with

$$\begin{aligned} \mathrm{E}(Y_i^{(s)}) &= \exp(\nu_i + \tfrac{1}{2}W_{ii}) \\ &= \exp(\boldsymbol{\mu}_i^T\boldsymbol{\beta} + \tfrac{1}{2}\sigma_{ii}\boldsymbol{\beta}^T\boldsymbol{\beta}) \end{aligned} \tag{4.4}$$

$$\begin{aligned} \mathrm{Var}(Y_i^{(s)}) &= \mathrm{E}(Y_i^{(s)}) + [\mathrm{E}(Y_i^{(s)})]^2[\exp(W_{ii}) - 1] \\ &= \mathrm{E}(Y_i^{(s)}) + [\mathrm{E}(Y_i^{(s)})]^2[\exp(\sigma_{ii}\boldsymbol{\beta}^T\boldsymbol{\beta}) - 1] \end{aligned} \tag{4.5}$$

$$\begin{aligned} \mathrm{Cov}(Y_i^{(s)}, Y_k^{(s)}) &= \mathrm{E}(Y_i^{(s)})\mathrm{E}(Y_k^{(s)})[\exp(W_{ik}) - 1] \\ &= \mathrm{E}(Y_i^{(s)})\mathrm{E}(Y_k^{(s)})[\exp(\sigma_{ik}\boldsymbol{\beta}^T\boldsymbol{\beta}) - 1] \end{aligned} \tag{4.6}$$

$$\begin{aligned} \mathrm{Cor}(Y_i^{(s)}, Y_k^{(s)}) &= \frac{\exp(W_{ik}) - 1}{\sqrt{\left[\exp(W_{ii}) - 1 + \frac{1}{\mathrm{E}(Y_i^{(s)})}\right]\left[\exp(W_{kk}) - 1 + \frac{1}{\mathrm{E}(Y_k^{(s)})}\right]}} \\ &= \frac{\exp(\sigma_{ik}\boldsymbol{\beta}^T\boldsymbol{\beta}) - 1}{\sqrt{\left[\exp(\sigma_{ii}\boldsymbol{\beta}^T\boldsymbol{\beta}) - 1 + \frac{1}{\mathrm{E}(Y_i^{(s)})}\right]\left[\exp(\sigma_{kk}\boldsymbol{\beta}^T\boldsymbol{\beta}) - 1 + \frac{1}{\mathrm{E}(Y_k^{(s)})}\right]}} \end{aligned} \tag{4.7}$$

for all i and k where $i \neq k$.

Finally, we can generate the covariate matrix and the response vector according to (4.1) and (4.2) by specifying the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ matrices, and the coefficient vector $\boldsymbol{\beta}$. We thus know the exact marginal distribution of the response variables according to equations (4.3) to (4.7).

Notice that so far, we have assumed that we do not have an intercept in the GLM. Some of the equations presented above should be slightly modified if an intercept is to be fitted. Please refer to Section 4.4 in the CHAPTER APPENDIX for details.

Technically, in our simulations, applying a Poisson regression to these $\boldsymbol{Y}^{(s)}$'s to obtain an estimate of $\boldsymbol{\beta}$ and the subsequent data analysis does not require the knowledge of the above marginal distribution. However, as mentioned in the previous chapter, some response correlation structure needs to be specified in the dependence model when using GEE for parameter estimation. When deliberately misspecifying the response correlation, we may make use of the marginal response correlation structure, which is generally very different from the independence structure. In fact, the responses are often strongly correlated, marginally, due to the correlation among the covariates.

4.2 Multiple Regression via Generalized Estimating Equations

In the dependence model, we assume that the responses are correlated, given the covariates, with certain correlation structure. In this part of our simulations, since the responses are uncorrelated given the covariates, we may examine if specifying a non-independence structure for the responses in the dependence model would decrease the efficiency of the regression parameter estimation. We do so by comparing the efficiency of such a model to that of the independence model (both models are fitted using the GEE approach).

4.2.1 GEE in the Conditional Poisson Model

The conditional density of $[Y_i^{(s)}| \boldsymbol{X}_{i.}^{(s)}]$ in (4.2) is

$$\begin{aligned} f\left(y_i^{(s)}|\boldsymbol{x}_{i.}^{(s)}\right) &= \frac{\exp\{-e^{\eta_i^{(s)}}\}(e^{\eta_i^{(s)}})^{y_i^{(s)}}}{y_i^{(s)}!} \\ &= \exp\{-e^{\eta_i^{(s)}} + y_i^{(s)} \log e^{\eta_i^{(s)}} - \log y_i^{(s)}!\} \\ &= \exp\{y_i^{(s)}\eta_i^{(s)} - e^{\eta_i^{(s)}} - \log y_i^{(s)}!\} \end{aligned} \tag{4.8}$$

where $\eta_i^{(s)} = \boldsymbol{X}_{i.}^{(s)T}\boldsymbol{\beta}$ for $i = 1, 2, \ldots, r$ and $s = 1, 2, \ldots, K$.

Comparing (3.17) and (4.8), we have $\theta_i^{(s)} = \eta_i^{(s)}$, $a(\theta_i^{(s)}) = e^{\theta_i^{(s)}}$, $b(y_i^{(s)}) = -\log y_i^{(s)}!$, and $\phi = 1$. Hence, (3.18) in the conditional Poisson model becomes

$$\begin{aligned} \mathrm{E}\left(Y_i^{(s)} | \boldsymbol{X}_{i.}^{(s)}\right) &= a'(\theta_i^{(s)}) &= e^{\theta_i^{(s)}} \\ \mathrm{Var}\left(Y_i^{(s)} | \boldsymbol{X}_{i.}^{(s)}\right) &= \frac{a''(\theta_i^{(s)})}{\phi} &= e^{\theta_i^{(s)}}. \end{aligned} \tag{4.9}$$

Note that the *link function* in a generalized linear model is defined to be the function that maps the mean (the conditional mean in our case) of the response to $\eta_i^{(s)}$. Thus, we have the log-*link* in the conditional Poisson model, i.e. $\log \mathrm{E}(Y_i^{(s)} | \boldsymbol{X}_{i.}^{(s)}) = \eta_i^{(s)}$.

Now, the matrices $\boldsymbol{A}^{(s)}$, $\boldsymbol{\Delta}^{(s)}$, $\boldsymbol{V}^{(s)}$, and $\boldsymbol{D}^{(s)}$ simply become

$$\boldsymbol{A}^{(s)} = \begin{pmatrix} e^{\theta_1^{(s)}} & 0 & \cdots & \cdots & 0 \\ 0 & e^{\theta_2^{(s)}} & \ddots & & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & e^{\theta_{r-1}^{(s)}} & 0 \\ 0 & \cdots & \cdots & 0 & e^{\theta_r^{(s)}} \end{pmatrix}$$

$$\boldsymbol{\Delta}^{(s)} = \boldsymbol{I}$$

$$\boldsymbol{V}^{(s)} = [\boldsymbol{A}^{(s)}]^{\frac{1}{2}} \boldsymbol{R} [\boldsymbol{A}^{(s)}]^{\frac{1}{2}}$$

$$\boldsymbol{D}^{(s)} = \boldsymbol{A}^{(s)} \boldsymbol{X}^{(s)}.$$

By specifying an initial working correlation matrix $\boldsymbol{R}$ with a certain structure (e.g. exchangeable), we may now substitute the above matrices into equations (3.20) and solve for the coefficient vector $\boldsymbol{\beta}$. Hence, we obtain a GEE estimate $\widehat{\boldsymbol{\beta}}^G$ for $\boldsymbol{\beta}$. If the specified structure of $\boldsymbol{R}$ coincides with that of $\mathrm{Cor}(\boldsymbol{Y} | \boldsymbol{X})$, then the final $\boldsymbol{R}$[1] is the estimate of $\mathrm{Cor}(\boldsymbol{Y} | \boldsymbol{X})$, the (spatial) correlation matrix of the conditional Poisson responses.

To estimate the covariance of $\widehat{\boldsymbol{\beta}}^G$, we need to know $\mathrm{Cov}(\boldsymbol{Y}^{(s)} | \boldsymbol{X}^{(s)})$. Recall that in this part of our simulations, we generate count data which are independent given the covariates. Thus, we can replace $\mathrm{Cov}(\boldsymbol{Y}^{(s)} | \boldsymbol{X}^{(s)})$ in equation (3.21) by

[1] See page 80.

1. $\boldsymbol{I}$, or

2. $[\boldsymbol{Y}^{(s)} - \boldsymbol{a}'(\widehat{\boldsymbol{\theta}}^{(s)})][\boldsymbol{Y}^{(s)} - \boldsymbol{a}'(\widehat{\boldsymbol{\theta}}^{(s)})]^T$ (see page 82).

In practice, the true correlation structure of the responses is not known, in which case item 1 above may not be used. Moreover, although we may sometimes believe that the data are so weakly correlated that they are close to independent, we should base our inference on the *robust* S.E. produced by gee. This is because the computation of the *robust* S.E. is a "corrected version" of the *naive* S.E., accounting for the true covariance structure of the responses which is estimated by item 2 above.

4.3 Computer Simulations of Multiple Regression of Count Data

Our goal is to examine the effect of incorrectly specifying the response correlation on the efficiency of parameter estimation in a generalized linear model. (The response correlation is specified through the structure of the working correlation matrix, $\boldsymbol{R}$, or fixing the matrix itself, when running gee.) In particular, count data are of special interest because many health impact studies use hospital admission counts to indicate health conditions of a population.

Throughout our simulations presented in this chapter, our specification of the incorrect response correlation was based on the marginal correlation of $\boldsymbol{Y}$, not conditional on $\boldsymbol{X}$. See Section 4.4.2 in the CHAPTER APPENDIX for details of how we specified $\boldsymbol{R}$. We used the statistical software S-Plus for all our simulations. Each simulation included three main parts, namely,

1. generation of the count responses and the covariates[2] according to equations (4.1) and (4.2), with $r = 5$ and $K = 20$;

[2] We would need to specify some $\boldsymbol{\Sigma}$ in (4.1) to generate the covariates.

2. regression (GLM fitting) of the generated count responses on the covariates using gee (function in S-Plus) assuming
   (a) independent responses ($\boldsymbol{R} = \boldsymbol{I}$), and
   (b) dependent responses ($\boldsymbol{R} \neq \boldsymbol{I}$), where
      i. only the structure of $\boldsymbol{R}$ was specified as either exchangeable or AR(1); or
      ii. $\boldsymbol{R}$ was fully specified as a fixed matrix based on the *marginal* distribution (Poisson-lognormal) of $\boldsymbol{Y}$ (as a misspecification of the correlation structure);

   and

3. comparison of empirical relative efficiency of the independence and the dependence models, and other analyses.

Note that estimating Cor($\boldsymbol{Y}|\boldsymbol{X}$) (by the final $\boldsymbol{R}$) is not the primary concern of a GEE analysis. However, a measure of the variability of $\widehat{\boldsymbol{\beta}}^G$ is crucial. In S-Plus, the function gee estimates the standard error (S.E.) of each regression coefficient estimate. It gives both the *naive* and the *robust* S.E. estimates. As mentioned before, in computing the *robust* S.E., the fact that the final estimate of $\boldsymbol{R}$ is only an estimate of Cor($\boldsymbol{Y}|\boldsymbol{X}$) is taken into consideration. In this part of our simulations, since the responses are in fact conditionally independent, $\boldsymbol{R} = \boldsymbol{I}$ (fixed) as specified in the independence model is actually the true Cor($\boldsymbol{Y}|\boldsymbol{X}$) (so that the working covariance matrix $\boldsymbol{V}$ is the true Cov($\boldsymbol{Y}|\boldsymbol{X}$) in the case of $\boldsymbol{R} = \boldsymbol{I}$). Hence, the *naive* S.E. in the fit for the independence model should be a better estimate of the true S.E. of a regression parameter estimate than the *robust* S.E.

However, in our simulations, we only compared the *robust* S.E.'s of the parameter estimates produced by the independence and dependence models. This is because with real data, one is never certain that they are truly independent or not. Analyses with real data should naturally be based on the *robust* S.E. estimates. In addition, we also computed the sample covariance of the $\widehat{\boldsymbol{\beta}}^G$'s in each simulation study[3] for comparison to the *robust* S.E.'s produced by the two models. With enough observations of $\widehat{\boldsymbol{\beta}}^G$, the sample covariance should resemble most the true covariance of $\widehat{\boldsymbol{\beta}}^G$, which can only be roughly estimated using equation (3.21) in a regression of real data.

This brings us back to our objective of the simulations — to determine how misspecification of the response correlation structure affects the performance of GEE parameter estimation in a GLM. A model *performs well* if the parameter estimates have small empirical biases and variability. Hence, a good model should be more effective in detecting departures from the null model of $\boldsymbol{\beta} = \mathbf{0}$. Intuitively, as supported by the previous chapter, the model that accounts for the true response correlation should perform no worse than that with a misspecified correlation structure. However, the effect of such misspecification on the efficiency of parameter estimation is not so clear in the current non-Gaussian setting. Our simulations are intended to help us understand at least empirically what this effect is on our non-Gaussian GLM.

4.3.1 Simulations with Exchangeable and AR(1) Correlation Structures of the Covariates

The most convenient correlation structures to use in the simulations are the exchangeable and the AR(1) correlation structures. These structures only consist of one unknown parameter. Hence, if the responses approximately have either of these correlation structures, then a small sample size, K, and a small number of regions, r, should provide sufficient information for GEE to estimate in a few iterations all the unknown parameters in the

[3] Each sample of K response vectors produced one estimate $\widehat{\boldsymbol{\beta}}^G$. In each simulation study, we generated n such samples, and then obtained the sample covariance of the n $\widehat{\boldsymbol{\beta}}^G$'s.

GLM. Indeed, conditioned on the covariates, the responses generated according to equations (4.1) and (4.2) have the independence correlation structure, which is a special case of both the exchangeable and the AR(1) structures.

At the same time, the exchangeable and AR(1) correlation structures are also convenient to specify in generating the covariates. We ran numerous simulations in which the covariates had either of these two correlation structures. Let ρ_x denote the intraclass correlation coefficient in $\boldsymbol{\Sigma}$ for the exchangeable covariates, and the lag one correlation in $\boldsymbol{\Sigma}$ for the AR(1) covariates. We specified different values of ρ_x, together with different $\boldsymbol{\mu}$'s and $\boldsymbol{\beta}$'s to generate different sets of count data, each set being marginally correlated, but independent conditional on the covariates. With each generated dataset, we used the S-Plus function gee to separately fit the independence and dependence models. We will denote by $\widehat{\boldsymbol{\beta}}^I$ the $\widehat{\boldsymbol{\beta}}^G$ produced by the independence model, and by $\widehat{\boldsymbol{\beta}}^D$ that produced by the dependence model. We would like to compare the empirical relative efficiency of the two models as reflected by how frequently the models reject the null hypotheses (for each j) of $\beta_j = 0$, each at significance level $\alpha = 0.05$. The z-test-statistics were computed based on normal approximations for the distributions of the estimates $\widehat{\boldsymbol{\beta}}^I$ and $\widehat{\boldsymbol{\beta}}^D$. The S.E.'s of the β_j's in these z-statistics were taken to be the *robust* S.E.'s produced by the S-Plus function gee.

We used two ways to specify a response correlation structure in the dependence model: (1) by specifying the structure for the working correlation, $\boldsymbol{R}$ (i.e. exchangeable or AR(1)) as a functional argument of gee; and (2) by specifying a fixed $\boldsymbol{R}$. In our experiments, the fixed $\boldsymbol{R}$ matrices were often very different from the identity matrix. We determined if the dependence model — either having an extra unknown parameter in $\boldsymbol{R}$ (i.e. case (1)), or a misspecified response correlation structure (i.e. case (2)) — would be less efficient than the independence model.

However, in all the simulations, both the independence and the dependence models generally produced estimates of $\boldsymbol{\beta}$ with small empirical biases. This result may not be surprising at first glance, because the consistency part of Theorem 3 applies to a model with any working correlation $\boldsymbol{R}$. However, it is merely an asymptotic result.[4] In fact, our study indicates that the GEE parameter estimates for our GLM is reasonably robust to misspecification of response correlation for small sample sizes; afterall, $K = 20$ is a small sample size. Slight bias was apparent only in isolated cases, particularly in estimating the intercept. This bias could merely be due to chance. As revealed in Table 4.9, the intercept estimate was typically the most variable (had the largest **relative S.E.**[5]) among all the regression coefficient estimates.

As for the empirical relative efficiency of $\widehat{\boldsymbol{\beta}}^I$ to $\widehat{\boldsymbol{\beta}}^D$, we examine the p-value comparisons in the simulations. We tabulate the p-values as in Table 2.1 from Chapter 2. For each regression coefficient in one such comparison table, we are interested in comparing the following:

1. the margins of the 2×2 p-value table, and
2. the off-diagonals of the 2×2 p-value table.

In item 1, we are essentially comparing the the empirical powers of the independence and the dependence GLM's. Hence, we can compute the empirical relative efficiency of the two models from the p-value tables as we have done in Chapter 2.

We will not present the results of all of our simulations. However, the results we will present in Tables 4.8 to 4.14 are representative of all of the simulations.

In these selected simulations, there were $g = 2$ covariates, and the $\boldsymbol{\mu}$ matrix that we

[4] The misspecification of the correlation structure of the responses does not affect the (K-)consistency of the GEE estimate for the regression coefficients. However, this is only true provided that, among other criteria, the link function is correctly specified. See Liang & Zeger, 1986 for other criteria for the consistency of $\widehat{\boldsymbol{\beta}}^G$.

[5] We define the relative S.E. of an estimator as the ratio of the S.E. of the estimator to the estimator itself. This is actually the inverse of the standardized estimator.

used in the "no-intercept" GLM's was

$$\boldsymbol{\mu} = \begin{pmatrix} 0 & 3 \\ 3 & 1 \\ 0 & 2 \\ 1 & 1 \\ 3 & 2 \end{pmatrix}.$$

We used the same $\boldsymbol{\mu}$ for those GLM's with an intercept, where we simply added a column of 1's to the covariate matrix generated with the above expectation.

Before examining more closely the simulation results, let us first explain the meaning of the items in Table 4.8.

In running the S-Plus function `gee`, we specify `corstr="ex"` as an argument to specify an exchangeable response correlation structure. Similarly, `corstr="fixed"` specifies a fixed response correlation matrix, in which case such a matrix needs to be completely supplied to `gee`. In simulations with `corstr="fixed"`, we would supply that fixed matrix computed in the way described in Section 4.4.2 (i.e. the response correlation unconditioned on the covariates), with $\widehat{\boldsymbol{\beta}}^I$ (obtained from the independence model) in place of $\boldsymbol{\beta}$.

Again, ρ_x represents the intraclass correlation coefficient or alternately the lag 1 correlation in $\boldsymbol{\Sigma}$, the correlation matrix of the covariates. In our simulations, we noticed that the marginal correlation structure of the responses were often approximately exchangeable if the covariates were either exchangeable or AR(1). (See Example 1 in Section 4.4.2.) We use $\hat{\rho}_y$ to denote the approximate value of the marginal correlation of each pair $(Y_i^{(s)}, Y_k^{(s)})$, $i \neq k$, for each s. (That is, unconditioned on the covariates, the responses are approximately exchangeable, with an intraclass correlation coefficient approximately equal to $\hat{\rho}_y$.) Since this marginal correlation matrix was what we would fix as $\boldsymbol{R}$ for the dependence model in simulations with `corstr="fixed"`, $\hat{\rho}_y$ indicates how

Sim	`cor`
	`str`
	`=`
		ρ_x	$\hat{\rho}_y$	$\boldsymbol{\beta}$	Fit inter-cept?	n	sample mean of n $\widehat{\boldsymbol{\beta}}^{I}$'s	sample mean of n $\widehat{\boldsymbol{\beta}}^{D}$'s

3^a	`"fixed"`	0.9	0.8	$\begin{pmatrix} 0.1 \\ 0.01 \\ 1 \end{pmatrix}$	yes	500	$\begin{pmatrix} 0.100 \\ 0.010 \\ 1.001 \end{pmatrix}$	$\begin{pmatrix} 0.101 \\ 0.010 \\ 1.000 \end{pmatrix}$
4^a	`"fixed"`	0.9	0.8	$\begin{pmatrix} 0.01 \\ 1 \end{pmatrix}$	no	500	$\begin{pmatrix} 0.011 \\ 1.000 \end{pmatrix}$	$\begin{pmatrix} 0.010 \\ 1.000 \end{pmatrix}$
7	`"ex"`	0.9	0.03 to 0.04	$\begin{pmatrix} 0.15 \\ 0.1 \\ 0.12 \end{pmatrix}$	yes	100	$\begin{pmatrix} 0.138 \\ 0.098 \\ 0.123 \end{pmatrix}$	$\begin{pmatrix} 0.139 \\ 0.097 \\ 0.123 \end{pmatrix}$
8	`"ex"`	0.9	0.03	$\begin{pmatrix} 0.1 \\ 0.12 \end{pmatrix}$	no	100	$\begin{pmatrix} 0.100 \\ 0.117 \end{pmatrix}$	$\begin{pmatrix} 0.100 \\ 0.116 \end{pmatrix}$

a: In these simulations, when fitting the dependence GLM, we assumed ρ_x to be known, and then substituted it and the final $\widehat{\boldsymbol{\beta}}^{I}$ into equations (4.4) and (4.7). We then specified in `gee` a correlation structure `fixed` as that in equation (4.7).

Table 4.8: Results of selected simulations.

badly the response correlation structure was misspecified in the model. In fact, most of these marginal response correlation matrices turned out to be quite different from the identity. However, this led to only slight loss of efficiency of the dependence model in some isolated cases. In the next section, we will see the simulation results which illustrate this point.

As for the regression parameter vector $\boldsymbol{\beta}$, we wanted its components to be small enough such that badly misspecifying the response correlation might deter the dependence model from detecting them,[6] making any possible loss of estimation efficiency noticeable. However, the size of a coefficient of course depends on the variability of its (GEE) estimate. Without any *a priori* information about the standard error of the estimates before the actual simulation experiment was done, we specified "near-zero" regression coefficients in the model. We tried values in the order of between 10^{-2} and 10^{1} in all the simulations.

[6] The larger the regression coefficients, the easier they are detected in any regression model.

Next, denote in each simulation by n the number of datasets generated, i.e. the number of GLM's fitted. For each dataset, the sample size K (= the number of replicates) was fixed as 20 per region, and r (= the number of regions) as 5.

The items "sample mean of n $\widehat{\boldsymbol{\beta}}^I$'s" and "sample mean of n $\widehat{\boldsymbol{\beta}}^D$'s" are self-explanatory.

4.3.2 Detailed Results for Simulations with Exchangeable and AR(1) Covariates

Among the simulations with exchangeable or AR(1) covariates, Simulations 3 and 4 are representative of those which were run with `corstr="fixed"`, and are the more interesting to examine. Similarly, Simulations 7 and 8 are representative of those which were run with `corstr="ex"`. Notice that the two pairs of simulations are paired by their regression coefficient values. For example, the latter pair have $\beta_1 = 0.1$ and $\beta_2 = 0.12$. The only difference between the two simulations is the intercept, $\beta_0 = 0.15$, in the GLM of Simulation 7, whereas in Simulation 8 the GLM had no intercept.

As mentioned before, both $\widehat{\boldsymbol{\beta}}^I$ and $\widehat{\boldsymbol{\beta}}^D$ seem to have small empirical biases in most of our simulations. Table 4.8 shows the typical amount of empirical biases of these two estimates. The β_j's in most cases are extremely well estimated by the $\widehat{\beta}_j^I$'s and $\widehat{\beta}_j^D$'s. The intercept estimates are usually those $\widehat{\beta}_j$'s that have slightly larger empirical biases. In fact, the intercepts usually have larger relative S.E.'s compared to the other β_j's. Table 4.9 shows the sample S.D. of the n $\widehat{\beta}_j^I$'s produced by the GEE analysis for our selected simulations, together with the approximate relative S.E. of $\widehat{\beta}_j^I$, computed by

$$\frac{\text{sample S.D. of } n\ \widehat{\beta}_j^I\text{'s}}{\text{sample mean of } n\ \widehat{\beta}_j^I\text{'s}}$$

for each j.

For example, β_0 in Simulation 7 is not particulary well estimated. Not coincidentally, its relative S.E. is more than twice the relative S.E. of each of the two slope coefficient

Simu-	sample S.D. for			approximate relative S.E. for		
lation	$\widehat{\beta}_0^I$	$\widehat{\beta}_1^I$	$\widehat{\beta}_2^I$	$\widehat{\beta}_0^I$	$\widehat{\beta}_1^I$	$\widehat{\beta}_2^I$
3	0.093	0.019	0.026	0.930	1.90	0.026
4	—	0.017	0.0096	—	1.55	0.0096
7	0.16	0.041	0.064	1.16	0.418	0.520
8	—	0.043	0.048	—	0.430	0.410

Table 4.9: Sample S.D.'s and approximate Relative S.E.'s of coefficient estimates in the independence model in selected simulations.

estimates.

We will not include a similar table for the $\widehat{\boldsymbol{\beta}}^D$'s (by the dependence model), because their sample S.D.'s and sample means are very similar to those of the $\widehat{\boldsymbol{\beta}}^I$'s (by the independence model).

This brings us to the efficiency/power comparisons of the independence and the dependence models. The sample S.D.'s should be the closest (compared to the GEE *naive* and *robust* S.E. estimates) to the true S.E.'s of the coefficient estimates, because n was large. However, in a real life investigation, we only have one set of data, on which we fit a GLM using GEE. In that case, we can only estimate the true S.E.'s of the regression coefficient estimates, usually by the *robust* S.E. estimates from the GEE analysis.

To compare the efficiency of the two models, we examine their p-value comparison tables. Tables 4.10 to 4.13 are the p-value comparison tables for Simulations 3, 4, 7, and 8 respectively. The following is an example of how one can examine a p-value comparison table.

Example

Table 4.12 shows the p-value comparisons for each of β_0, β_1, and β_2 in Simultion 7. The 2×2 p-value comparison for β_1 (middle row) shows that the empirical power[7] of the independence model ($\boldsymbol{R} = \boldsymbol{I}$) is $58/100 = 0.58$. The empirical power of the dependence model ($\boldsymbol{R} \neq \boldsymbol{I}$) is slightly higher at $60/100 = 0.60$. We say that the dependence model is empirically 2% more powerful. This is obtained from $(60 - 58)/100 = 0.02 = 2\%$. The off-diagonal values in this 2×2 table show the *discordance* of the two models in detecting a non-zero β_1. For example, among the 100 cases, there are $4+2=6$ cases where the two models did not agree in the rejection of the null hypothesis. In considering the overall p-value comparison for the three regression coefficients in this simulation, the dependence model seems to be slightly more powerful than the independence model. However, the χ^2 test of independence of model and power yields p-values[8] of 0.66, 0.89, and 0.67 for β_0, β_1, and β_2 respectively. Thus, the difference in power of the two models (in detecting any of the regression coefficients) is likely due to chance. Both models are powerful in detecting β_1 and β_2, but not powerful in detecting β_0. This is because $\widehat{\beta}_0^G$ has the largest relative S.E. (i.e. $\mathrm{SE}(\widehat{\beta}_0^G)/\widehat{\beta}_0^G$), as is shown in Table 4.9.

□

From our simulation results, we see that generally, both the dependence and the independence models would decrease in power of detecting a non-zero regression coefficient as the coefficient became smaller. Also, both models had very similar empirical power in a given simulation. For example, in Table 4.10, both models only had approximately

[7] By power, we mean the probability of rejecting the null hypothesis in testing $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ for that j, at significance level $\alpha = 0.05$, when H_1 is true. We consider hypothesis tests of β_j's individually for convenience. One can easily modify this into a test of $H_0 : \boldsymbol{\beta} = \mathbf{0}$, at a certain significance level.

[8] ... obtained by the S-Plus function `chisq.test`.

			$\boldsymbol{R} \neq \boldsymbol{I}$		
			$p < 0.05$	$p \geq 0.05$	total
β_0	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	58	76	134
		$p \geq 0.05$	47	319	366
		total	105	395	500
β_1	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	38	35	73
		$p \geq 0.05$	31	396	427
		total	69	431	500
β_2	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	500	0	500
		$p \geq 0.05$	0	0	0
		total	500	0	500

Table 4.10: p-value comparisons of Simulation 3 in Table 4.8.

			$\boldsymbol{R} \neq \boldsymbol{I}$		
			$p < 0.05$	$p \geq 0.05$	total
β_1	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	34	43	77
		$p \geq 0.05$	33	390	423
		total	67	433	500
β_2	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	500	0	500
		$p \geq 0.05$	0	0	0
		total	500	0	500

Table 4.11: p-value comparisons of Simulation 4 in Table 4.8.

70/500 = 14% empirical power in detecting $\beta_1 = 0.01$, but had 100% empirical power in detecting $\beta_2 = 1$.

We also noticed that both models tended to decrease in power if the number of parameters increased. That is, in those simulations where the GLM did not include an intercept, the two models were sometimes somewhat more powerful (empirically) than if the GLM included an intercept. This is illustrated by Simulations 7 and 8. We may compare Tables 4.12 and 4.13. In Simulation 7, the independence and dependence models both had only about 60% and 50% empirical power in detecting the two slope coefficients ($\beta_1 = 0.1$, $\beta_2 = 0.12$) respectively, but had about 70% and 75% empirical power for the respective slopes in Simulation 8. This trend was not always apparent, mainly because of other factors which would also affect power, such as the actual size of the regression coefficients. For example, in both Simulations 3 and 4, the two models had similar empirical power in detecting the two slopes. $\beta_1 = 0.01$ (whose estimate has an approximate S.E. of 0.02) is considered small, whether the GLM had many or few coefficients. Thus, the two models had low but similar empirical power (about 14%) in detecting it in both simulations. On the other hand, $\beta_2 = 1$ (whose estimate has an approximate S.E. of 0.03) is considered very large, with or without many extra coefficients in the GLM. Therefore, the two models both had 100% empirical power in detecting it in both simulations.

In some simulations, such as Simulations 3 and 4, we specified a fixed correlation matrix — the marginal $\text{Cor}(\boldsymbol{Y})$ — which was often very different from the identity, as can be seen from the $\hat{\rho}_y$ column in Table 4.8. In these simulations, the dependence model often had slightly less empirical power than the independence model. For example, in Simulation 3, the independence model is

$$\frac{134 - 105}{500} \times 100\% = 5.8\%$$

			$\boldsymbol{R} \neq \boldsymbol{I}$		
			$p < 0.05$	$p \geq 0.05$	total
β_0	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	10	0	10
		$p \geq 0.05$	3	87	90
		total	13	87	100
β_1	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	56	2	58
		$p \geq 0.05$	4	38	42
		total	60	40	100
β_2	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	46	1	47
		$p \geq 0.05$	5	48	53
		total	51	49	100

Table 4.12: p-value comparisons of Simulation 7 in Table 4.8.

			$\boldsymbol{R} \neq \boldsymbol{I}$		
			$p < 0.05$	$p \geq 0.05$	total
β_1	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	66	2	68
		$p \geq 0.05$	3	29	32
		total	69	31	100
β_2	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	74	0	74
		$p \geq 0.05$	1	25	26
		total	75	25	100

Table 4.13: p-value comparisons of Simulation 8 in Table 4.8.

more powerful than the dependence model in detecting β_0. In fact, the χ^2 test of independence of model and power yields a p-value of 0.038, implying that this difference in power between the models is significant (at level 0.05, say). This difference in empirical power in detecting β_1 is 0.8% in Simulation 3, and is 2.0% in Simulation 4, although the differences in both cases seem to be results of chance according the χ^2 tests. The p-value discordance also seems to be more severe in those simulations where `corstr="fixed"`.

We re-ran Simulation 3 in which we saw a noticeable difference in empirical power of the two models. In the re-run, we let the GEE analysis update the values in the working correlation matrix $\boldsymbol{R}$. In running `gee` here, we specified the exchangeable response correlation structure. To save computational time, we set $n = 100$, but K remained 20. Table 4.14 shows the p-value comparison for this simulation.

From Table 4.14, we can see that the dependence model indeed had a better empirical power in this re-run than the original simulation. In fact, in the re-run, the empirical powers of the independence and dependence models did not differ in testing $\beta_0 = 0$, as opposed to the dependence model being 5.8% weaker in power in the original Simulation 3. In addition, the dependence model in the re-run was actually 3.0% more powerful than the independence model in the re-run in testing $\beta_1 = 0$, as opposed to 0.8% less powerful in the original simulation. (However, χ^2 tests indicate that these differences were likely due to chance in both the original simulation and its re-run.) The p-value discordance was also much less severe in the re-run than in the original simulation.

The smaller power of the dependence model in those cases such as Simulations 3 and 4 might be due to the inappropriately specified marginal response correlation matrix for the dependence model, when data were in fact independent. However, the dependence model did not have substantially lower power than the independence model in these cases. The power difference was the most "extreme" in Simulation 3, but the dependence model turned out to be only 5.8% weaker than the independence model in detecting $\beta_0 = 0.1$.

			$\boldsymbol{R} \neq \boldsymbol{I}$		
			$p < 0.05$	$p \geq 0.05$	total
β_0	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	22	1	23
		$p \geq 0.05$	1	76	77
		total	23	77	100
β_1	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	21	1	22
		$p \geq 0.05$	4	74	78
		total	25	75	100
β_2	$\boldsymbol{R} = \boldsymbol{I}$	$p < 0.05$	100	0	100
		$p \geq 0.05$	0	0	0
		total	100	0	100

Table 4.14: p-value comparisons of a re-run of Simulation 3 in Table 4.8, this time without fixing $\boldsymbol{R}$.

The small difference in power of the two models is particularly surprising if we consider how different the marginal response correlation matrix was from the the identity in most of these cases. For example, in Simulations 3 and 4, we specified `corstr="fixed"`, and this fixed correlation matrix had most non-diagonal values of about 0.8. This in fact illustrates the robustness of the GEE approach to the misspecification of the response correlation structure.

There is another worth-noting observation from our simulation results. Recall that we specified `corstr="ex"` in the dependence model when we ran `gee` in some of the simulations. Since the responses were truly independent given the covariates, the independence model should be the correct model. However, since independence is a special case of exchangeability, the final $\boldsymbol{R}$ matrix produced by `gee` should be close to the identity matrix, because the GEE approach is supposed to estimate the true response correlation reasonably well when the structure is correctly specified. Therefore, one would expect both the dependence and independence models to have the same efficiencies here. In fact, if we consider the extra parameter (response intraclass correlation coefficient) that GEE needs to estimate in the dependence model, one might even expect the independence model to be more efficient than the dependence model.

Indeed, in some of our simulations with `corstr="ex"` in the dependence model, this model was slightly less powerful than the independence model, possibly due to the extra unknown parameter in $\mathrm{Cor}(\boldsymbol{Y}|\boldsymbol{X})$. However, some other simulations with `corstr="ex"` in the dependence model indicate that this model could also be slightly more powerful than the independence model. For example, in Simulation 7, as presented in Table 4.12, the dependence model is empirically 3%, 2%, and 4% more powerful than the independence model in detecting the three regression coefficients respectively. However, such small differences in power seem to have resulted from chance according to χ^2 tests.

4.3.3 Additional observations

We have so far discussed how the power of independence and dependence models compare in terms of their p-values. We may also compare their power by comparing their actual *robust* S.E.'s for the regression coefficient estimates.

Among the simulations for which we have a retained record of these S.E.'s,[9] the S.E.'s for both the independence and the dependence models were very similar. This explains why both models had virtually the same power in detecting a non-zero regression coefficient.

The GEE *robust* S.E.'s seemed to be relatively close to the sample S.D.'s of the coefficient estimates. A few *robust* S.E.'s were smaller than the corresponding sample S.D. by a factor of about 2, making the regression analyses appear more powerful than they really were. Only a few other *robust* S.E.'s differed from the sample S.D.'s by a slightly larger factor. However, many of them were extremely close to the sample S.D.'s. Although the difference would well have been due to chance, the reason behind a smaller

[9] These include simulations yet to be discussed.

robust S.E. than the sample S.D. in some cases is not obvious. Further investigation may be needed.

4.3.4 Other Simulations

In addition to those simulations with exchangeable and AR(1) covariates, we also ran several other simulations with very different covariate correlation structures. Some were "unstructured," while others had the variogram structure as in Chapter 2. Some of these extra simulations were also run with different $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X})$ matrices. Their results were very similar to those of the previous simulations. That is, we found in these simulations that

1. both the independence and dependence models produced regression coefficient estimates that have relatively small empirical biases;

2. the independence model tended to be slightly more powerful than the dependence model in those simulations with fixed response correlation matrix;

3. the dependence model was slightly more powerful than the independence model in some of the simulations where only the response correlation structure was specified;

4. the difference in power between the two models mentioned in the previous two items was not always significant according to the χ^2 test;

5. both models had very similar overall power in detecting the non-zero regression coefficients in all simulations, suggesting the robustness of GEE to misspecification of response correlation structure;

6. this power tended to decrease as the coefficient size decreased, and/or as the number of parameters in the GLM increased; and

K	n	summary statistics of n intraclass correlation coefficient estimates		
		mean	median	sample S.D.
10	30	-0.028	-0.029	0.080
20	20	-0.012	-0.020	0.056
50	15	0.0039	0.0042	0.048
100	15	-0.0011	-0.0029	0.024
200	15	-0.011	-0.013	0.022
500	10	-0.000084	-0.00059	0.012
800	10	-0.00093	-0.0023	0.010

Table 4.15: Final GEE $\boldsymbol{R}$ as K increases.

7. the *robust* S.E.'s from the GEE analysis were often very close to the sample S.D.'s of the coefficient estimates, even with such small r $(= 5)$ and K $(= 20)$.

In addition, we also investigated how well the GEE approach estimates Cor($\boldsymbol{Y}\,|\,\boldsymbol{X}$). We ran small scale simulations ($n \leq 30$) with increasing K but r was held at 5 (as in the previous simulations). Each was run with $\boldsymbol{\mu} = (3, 1, 2, 1, 2)^T$, a common $\boldsymbol{\Sigma}$ matrix, `corstr="ex"`, and a GLM that consisted of only one slope coefficient, $\beta_1 = -1$, but no intercept. Since the final $\boldsymbol{R}$ matrix estimates Cor($\boldsymbol{Y}\,|\,\boldsymbol{X}$) in our assumed scenario (see Section 3.4), its intraclass correlation coefficient should estimate that of Cor($\boldsymbol{Y}\,|\,\boldsymbol{X}$) which is zero. Intuitively, the larger the sample size K (= number of replicates per dataset), the better the response correlation matrix is estimated. We found that with K ranging from 10 to 800, the estimated intraclass correlation coefficient (in the final $\boldsymbol{R}$) produced by `gee` was always close to zero. The simulation results are shown in Table 4.15. Boxplots of the estimates are shown in Figure 4.9.

The second column of Table 4.15 shows the number of regressions in the simulation experiment. As K increases, the computational time for the simulation experiment

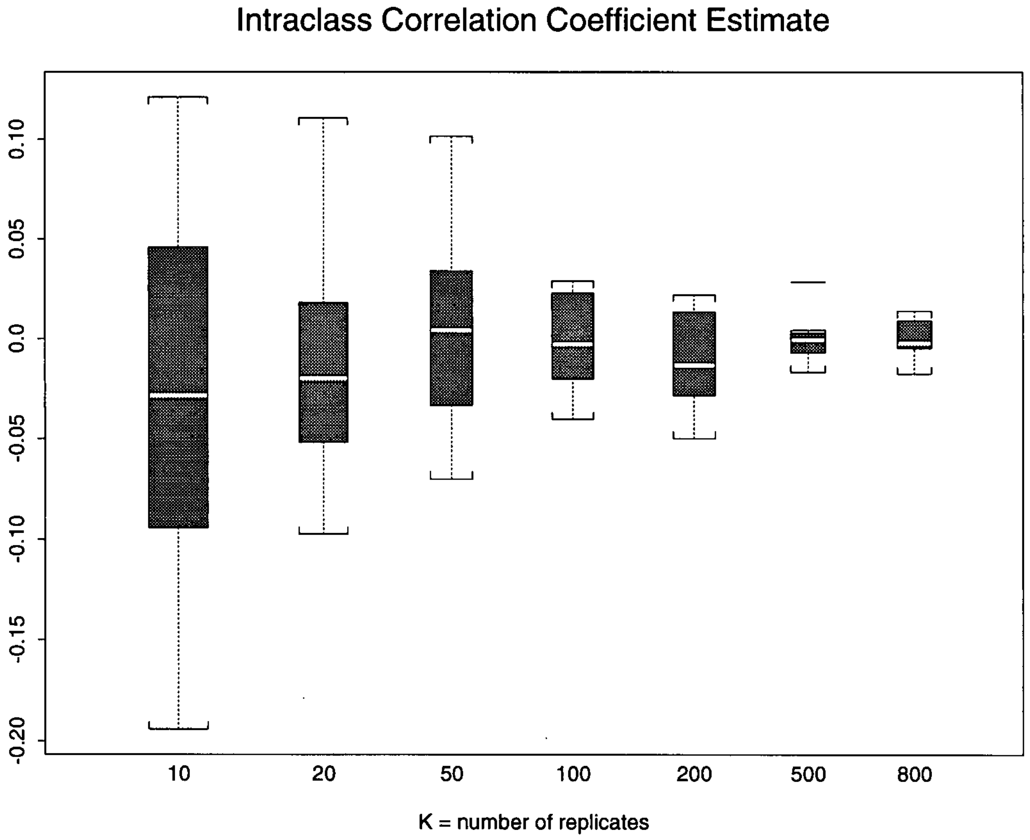


Figure 4.9: Boxplots of estimates of intraclass correlation coefficient (in $\boldsymbol{R}$) with various sample sizes, K.

increases.[10] This explains our choice of small n's for the larger K's. The sample median of the estimates may be more appropriate than the sample mean for assessing how well the intraclass correlation coefficient is estimated in the cases with small n's. In fact, the boxplots in Figure 4.9 show asymmetric empirical distributions of the estimates with most K values, suggesting the comparisons of the sample medians instead of means.

From Table 4.15, we see that the median of the estimates appears to approach closer to zero (the true intraclass correlation) as K increases, except for $K = 200$ and $K = 800$. The exceptions could be due to chance, but the real reason is not obvious since n was small for both cases. However, the variability (as reflected by the sample S.D.) of the estimate seems to decrease steadily with an increasing K. Figure 4.9 also shows that as K increases, the sample median of the estimates seems to stabilize around 0, and the estimates become less variable.

Although the intraclass correlation coefficient estimate seems to become better (smaller bias and variability) as K increases, we are quite satisfied with the GEE approach in that with such a short correlated series ($r = 5$), even a small sample size of 10 seems to yield a reasonable estimate of $\text{Cor}(\boldsymbol{Y}|\boldsymbol{X})$.

4.3.5 Concluding Remarks

We have thus far examined the effect of misspecifying response correlation, when the data are independent, on parameter estimation using the GEE approach. Another type of misspecification of response correlation is using the independence model to estimate regression parameters, when the data are in fact correlated. In the next chapter, we will investigate the relative efficiency of the independence and dependence models in such a setting. This time, we will specify the correct correlation structure in the dependence model when running gee.

[10] ... with our implementation of the simulation experiments in S-Plus.

4.4 CHAPTER APPENDIX

4.4.1 Fitting an intercept β_{o} in the GLM

As in Chapter 2, when fitting an intercept in the model, we have to add a column of 1's to the covariate matrix $\boldsymbol{X}$. Let us refer to this column as $\boldsymbol{X}_{.\mathrm{o}}$. Then $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X})$ also has a column of 1's, denoted by $\boldsymbol{\mu}_{.\mathrm{o}}$. The distribution in (4.1) remains unchanged, except that j only runs from 1 to r (i.e. $j \neq 0$). The coefficient vector $\boldsymbol{\beta}$ now becomes $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_g)^T$. Now, the row vectors $\boldsymbol{X}_{i.}$ and $\boldsymbol{\mu}_{i.}$, $i = 1, 2, \ldots, r$, all include an extra first element of 1. Nevertheless, the conditional distribution in (4.2) is unchanged with the new definitions of $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\beta}$.

However, now $\boldsymbol{X}_{i.}^T\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{g} \beta_j X_{ij}$, and hence, $\mathrm{Cov}(\boldsymbol{X}_{i.}^T\boldsymbol{\beta}, \boldsymbol{X}_{k.}^T\boldsymbol{\beta}) = \sigma_{ik}\, (\boldsymbol{\beta}_{-\mathrm{o}})^T\boldsymbol{\beta}_{-\mathrm{o}}$ for all i and k from 1 to r, where $\boldsymbol{\beta}_{-\mathrm{o}} = (\beta_1, \beta_2, \ldots, \beta_g)^T$. Note that $X_{i0}\beta_0 = \beta_0$ is non-random, and it does not have a variance. Thus, the distribution in (4.3) should now be written as $\boldsymbol{X\beta} \sim \mathrm{MVN}(\boldsymbol{\nu}, \boldsymbol{W})$ where $\boldsymbol{\nu} = \boldsymbol{\mu\beta}$ ($\boldsymbol{\mu}$ with column of 1's and $\boldsymbol{\beta}$ with β_0) and $\boldsymbol{W} = (\boldsymbol{\beta}_{-\mathrm{o}})^T\boldsymbol{\beta}_{-\mathrm{o}}\boldsymbol{\Sigma}$.

Equations (4.4) to (4.7) basically remain the same, except that $\boldsymbol{\nu} = \boldsymbol{\mu\beta}$ is to be computed using the new $\boldsymbol{\mu}$ (with the column of 1's), the new $\boldsymbol{\beta}$ (with β_0), and $\boldsymbol{W} = (\boldsymbol{\beta}_{-\mathrm{o}})^T\boldsymbol{\beta}_{-\mathrm{o}}\boldsymbol{\Sigma}$.

The rest of the equations in the chapter remain unchanged.

4.4.2 Examples of calculating (marginal) Cor($\boldsymbol{Y}$) with a specified $\boldsymbol{\Sigma}$

First, assume that we do not fit an intercept in the GLM. Let the $r \times g$ covariate matrix $\boldsymbol{X}$ have expected value $\boldsymbol{\mu}$, and let each of its column have covariance $\boldsymbol{\Sigma}$.

Example 1

Suppose $r = 5$ and $g = 1$, and

$$\boldsymbol{\mu} = \boldsymbol{\mu}_{.1} = \begin{pmatrix} 50 \\ 35 \\ 43 \\ 26 \\ 32 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.9 & 0.9 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 & 0.9 & 0.9 \\ 0.9 & 0.9 & 1 & 0.9 & 0.9 \\ 0.9 & 0.9 & 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 0.9 & 0.9 & 1 \end{pmatrix}.$$

That is, the covariate vector $\boldsymbol{X} = \boldsymbol{X}_{.1}$ has an exchangeable correlation structure, with an intraclass correlation coefficient of 0.9. Also, let $\boldsymbol{\beta} = (\beta_1) = (0.1)$.

Substituting these values into equations (4.4) and (4.7), we have

$$\mathrm{E}(\boldsymbol{Y}) = \begin{pmatrix} 149.16 \\ 33.28 \\ 74.07 \\ 13.53 \\ 24.66 \end{pmatrix} \text{ and } \mathrm{Cor}(\boldsymbol{Y}) = \begin{pmatrix} 1 & 0.35 & 0.46 & 0.24 & 0.31 \\ 0.35 & 1 & 0.29 & 0.16 & 0.20 \\ 0.46 & 0.29 & 1 & 0.20 & 0.26 \\ 0.24 & 0.16 & 0.20 & 1 & 0.14 \\ 0.31 & 0.20 & 0.26 & 0.14 & 1 \end{pmatrix}.$$

Here, $\mathrm{Cor}(\boldsymbol{Y})$ does not have a standard structure, but the structure is roughly exchangeable with an approximate intraclass correlation coefficient of 0.2 to 0.3. As an entry in Table 4.8, this approximate coefficient is denoted by $\hat{\rho}_y$ to indicate how badly the response correlation is misspecified through the dependence model with `corstr="fixed"` as this marginal $\mathrm{Cor}(\boldsymbol{Y})$.

In the next example, we will fit an intercept β_0 in the GLM. With an extra column of 1's, the matrices $\boldsymbol{X}$ and $\boldsymbol{\mu}$ now have dimension $r \times (g+1)$. $\boldsymbol{\beta}$ is now a $(g+1)$-vector.

Example 2

Suppose $r = 3$ and $g = 1$, and

$$\boldsymbol{\mu} = \begin{pmatrix} 1 & 10 \\ 1 & 8 \\ 1 & 25 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.9 & 0.81 \\ 0.9 & 1 & 0.9 \\ 0.81 & 0.9 & 1 \end{pmatrix}.$$

That is, the covariate vector $\boldsymbol{X}_{.1}$ has an AR(1) correlation structure, with a lag 1 correlation of 0.9. Now, let $\boldsymbol{\beta} = (\beta_0, \beta_1)^T = (5, -0.15)^T$.

Substituting these values into the modified[11] equations (4.4) and (4.7), we have

$$\mathrm{E}(\boldsymbol{Y}) = \begin{pmatrix} 33.49 \\ 45.21 \\ 3.53 \end{pmatrix} \text{ and } \mathrm{Cor}(\boldsymbol{Y}) = \begin{pmatrix} 1 & 0.421 & 0.145 \\ 0.421 & 1 & 0.175 \\ 0.145 & 0.175 & 1 \end{pmatrix}.$$

We see that the response vector $\boldsymbol{Y}$ has a correlation structure that is approximately AR(1) with an approximate lag 1 correlation of 0.4. As an entry in Table 4.8, this approximate lag 1 correlation is denoted by $\hat{\rho}_y$ to indicate how badly the response correlation is misspecified through the dependence model with `corstr="fixed"` as this marginal Cor($\boldsymbol{Y}$).

[11] See Section 4.4.1.

Chapter 5

Simulations of Multiple Regression of Count Data — Part Two

In this part of our simulations, we generate Poisson responses that are correlated (conditional on the covariates). We apply the GEE approach to the same correlated Poisson data for estimating the regression parameters in each of the independence and dependence models. When the correct response correlation structure is specified for the dependence model in running GEE, we expect it to be more efficient in parameter estimation than the independence model (which assumes the data to be uncorrelated).

However, generating correlated Poisson responses with covariates is a difficult task. Instead, we use a GLM which only has an intercept, with no covariates. In other words, the covariate matrix is simply a column of 1's, and the regression coefficient vector consists of only one element — the intercept value β_0. As explained in the following section, the resulting Poisson counts have an exchangeable correlation structure.

5.1 Generating Correlated Poisson Responses

As before, we assume that there are r regions. From each region, we sample/generate K independent replicates. Each time we sample, i.e. for each fixed s for $s = 1, 2, \ldots, K$, we would like the r observations $Y_1^{(s)}, Y_2^{(s)}, \ldots, Y_r^{(s)}$ to be spatially correlated.

Now, let us fix an s.

We generate $\{W_1^{(s)}, \ldots, W_r^{(s)}\}$ and $Z^{(s)}$, independently, according to

$$W_i^{(s)} \overset{\text{iid}}{\sim} \text{Poisson}(\lambda_1)$$

and

$$Z^{(s)} \sim \text{Poisson}(\lambda_2)$$

for all $i = 1, 2, \dots, r$, for some positive λ_1 and λ_2.

We then let the response in region i to be $Y_i^{(s)} = W_i^{(s)} + Z^{(s)}$, and hence

$$\begin{aligned}
Y_i^{(s)} &\sim \text{Poisson}(\lambda_1 + \lambda_2) && (5.1)\\
\text{E}(Y_i^{(s)}) &= \text{Var}(Y_i^{(s)}) = \lambda_1 + \lambda_2 && (5.2)\\
\text{Cov}(Y_i^{(s)}, Y_k^{(s)}) &= \text{Cov}(W_i^{(s)} + Z^{(s)}, W_k^{(s)} + Z^{(s)}) &&\\
&= \text{Var}(Z^{(s)}) &&\\
&= \lambda_2 \qquad \text{for all } i \neq k && (5.3)\\
\text{Cor}(Y_i^{(s)}, Y_k^{(s)}) &= \frac{\text{Cov}(Y_i^{(s)}, Y_k^{(s)})}{\sqrt{\text{Var}(Y_i^{(s)})\text{Var}(Y_k^{(s)})}} &&\\
&= \frac{\lambda_2}{\lambda_1 + \lambda_2} \qquad \text{for all } i \neq k. && (5.4)
\end{aligned}$$

Let $\lambda = \lambda_1 + \lambda_2$. Hence, these responses $Y_1, \dots, Y_r$ are exchangeably correlated, with intraclass correlation coefficient $\rho = \lambda_2/\lambda$.

Let us now put the above into the GLM format.

The only observations here are the r-dimensional response vector, $\boldsymbol{Y}^{(s)} = (Y_1^{(s)}, \dots, Y_r^{(s)})^T$. In other words, we do not include any covariates in the GLM (a regression model). However, an intercept, β_0, does exist in the regression model. In other words, the GLM is

$$\mathcal{G}\left[\text{E}\left(Y_i^{(s)} | X_i^{(s)}\right)\right] = X_i^{(s)}\beta_0 \qquad \text{where } X_i^{(s)} = 1 \text{ for all } i = 1, \dots, r$$

for some link function $\mathcal{G}$. In vector notation, the covariate (design) matrix $\boldsymbol{X}^{(s)}$ is thus simply a vector of 1's. The GLM in vector notation then becomes

$$\mathcal{G}\left[\text{E}\left(\boldsymbol{Y}^{(s)} | \boldsymbol{X}^{(s)}\right)\right] = \boldsymbol{X}^{(s)}\beta_0 \qquad \text{where } \boldsymbol{X}^{(s)} = \mathbf{1}. \tag{5.5}$$

The goal is to estimate the unknown regression parameter, β_0, in the above GLM. The use of the GEE approach in estimating β_0 allows the incorporation of the correlation among $Y_i^{(s)}$'s (for each s).

5.1.1 The Link Function

To be consistent with the GEE notation as before, we define $\boldsymbol{\eta}^{(s)} = \boldsymbol{X}^{(s)}\beta_0$. In other words, $\boldsymbol{\eta}^{(s)}$ is simply the vector of β_0's.

As explained before, the link function $\mathcal{G}$ in a GLM maps $\mathrm{E}(Y_i^{(s)}|X_i^{(s)})$ to $\eta_i^{(s)}$ for all $i = 1, \ldots, r$. An easy parametrization obtains by letting

$$\lambda = \lambda_1 + \lambda_2 = \beta_0$$

so that $\mathcal{G}$ is the *identity link*, since

$$\mathcal{G}\left[\mathrm{E}\left(Y_i^{(s)}|X_i^{(s)}\right)\right] = \mathcal{G}(\lambda) = \mathcal{G}(\beta_0) = \beta_0 = \eta_i^{(s)}$$

for all i.

In other words, this GLM is a true linear regression model because of the linear response-intercept relationship. However, there is a difference between this GLM and the linear model in Chapter 2. Here, the individual responses have a Poisson distribution, whereas before, they were normally distributed.

After recognizing the link function, we may proceed to the actual model fitting for the simulated data.

5.2 Model Fitting using the GEE approach

Once again, we fit two different GLM's to these simulated data, namely, the independence and the dependence models.

In fitting the independence model, we ignore the response correlation. We use the S-Plus function **gee** to fit this model, by specifying (5.5) as the linear model, Poisson as the distribution of $[Y_i^{(s)}|X_i^{(s)}]$,[1] the identity link function, and the independence response correlation structure. **gee** then produces an estimate of β_0 and the *naive* and *robust* estimates of its S.E.

We may also use the same **gee** function for incorporating the response correlation structure in parameter estimation. Recall that the GEE approach allows the estimation of the true response correlation matrix, $\text{Cor}(\boldsymbol{Y}|\boldsymbol{X})$,[2] provided that the structure of $\text{Cor}(\boldsymbol{Y}|\boldsymbol{X})$ is correctly specified. To use the S-Plus **gee** function in such a case, we specify the function arguments as in the independence model, except for some non-independence correlation structure for $[Y_i^{(s)}|X_i^{(s)}]$ instead of the independence structure. Since the responses we generate are exchangeably correlated, we may specify `"exchangeable"` for the `corstr` function argument. **gee** would then use an initial working correlation matrix, $\boldsymbol{R}$, which has the exchangeable correlation structure, in the estimation process. Thus, in addition to estimating the regression parameter β_0 and the *naive* and *robust* S.E.'s of its estimate, **gee** also updates the intraclass correlation coefficient in $\boldsymbol{R}$ throughout the estimating process. The final $\boldsymbol{R}$ would then be the estimate of $\text{Cor}(\boldsymbol{Y}|\boldsymbol{X})$ (= $\text{Cor}(\boldsymbol{Y})$ here). In other words, the non-diagonal value in the final $\boldsymbol{R}$ would be an estimate of ρ.

Before fitting the independence and the dependence models, we notice that with an identity link in the GLM (5.5), and no true covariates (except the **1** vector as the design matrix), the GEE estimate of β_0 is simply the sample mean of the responses. This is true for both the independence and dependence models, and will be proved in the proof of Theorem 4 in Section 5.4. In fact, we can see that the sample mean is the best linear unbiased estimator (BLUE) of the intercept for either model (as in ordinary and

[1] The function **gee** in S-Plus asks for the response distribution as a means of specifying the relationship between the mean and the variance of each response.

[2] See Section 3.3 page 78.

generalized least squares estimation in Chapter 2) due to the linearity of the response-intercept relationship and the exchangeability of the responses. Denoting $\widehat{\beta}_0^I$ and $\widehat{\beta}_0^D$ as the β_0 estimates produced by the independence and dependence models respectively, we have,

$$\widehat{\beta}_0^I = \widehat{\beta}_0^D = \overline{Y} = \frac{1}{Kr}\sum_{s=1}^{K}\sum_{i=1}^{r} Y_i^{(s)}. \tag{5.6}$$

However, the *naive* and *robust* S.E. estimates for $\widehat{\beta}_0^I$ and $\widehat{\beta}_0^D$ may be different, since their variance formulae differ. In particular, for the independence model, $\boldsymbol{R} = \boldsymbol{I}$ (throughout the GEE iterations) in equation (3.19); hence, $\boldsymbol{V}^{(s)} = \boldsymbol{A}^{(s)}$ from equation (3.19)[3] and thus in (3.21) and (3.22). For the dependence model, $\boldsymbol{R}$ has an exchangeable correlation structure. Thus, the $\boldsymbol{R}$ matrix in the final iteration of GEE has some $\hat{\rho}$ as its non-diagonal value, possibly non-zero. This is because we expect $\hat{\rho} \approx \rho = \lambda_2/\lambda \neq 0$. In the case of $\hat{\rho} \neq 0$, $\boldsymbol{V}^{(s)}$ in equation (3.19) (computed using the final $\boldsymbol{R}$ matrix) is no longer equal to $\boldsymbol{A}^{(s)}$. This $\boldsymbol{V}^{(s)}$ will possibly produce a different value in each of equations (3.21) and (3.22) from the one for the independence model (when $\boldsymbol{V}^{(s)} = \boldsymbol{A}^{(s)}$). In other words, the *naive* and *robust* Var($\widehat{\beta}_0^D$) estimates are possibly different from the *naive* and *robust* Var($\widehat{\beta}_0^I$) estimates.

Since the GEE approach allows the consideration of response correlation when fitting the model, our first thought was that the two variance estimates for $\widehat{\beta}_0^D$ (the dependence model (correct) parameter estimate) would be no bigger than their independence model (incorrect) counterparts. However, this is not exactly what we saw in our simulations, as explained in the next section.

[3] The scale parameter ϕ in our Poisson model is simply 1 (assumed known when using gee).

5.3 The Simulations

In our first simulation, we tried to fit $n = 100$ GLM regressions. For each regression, we generated $K = 20$ independent samples from each of $r = 5$ regions according to Section 5.1. We let $\lambda_1 = 0.05$ and $\lambda_2 = 0.5$, so that $\rho = 0.5/0.55 \approx 0.91$ would give very highly correlated Poisson responses. The two λ's were deliberately chosen to have such minute size because we would like the generated responses to resemble hospital admission counts (for rare diseases) — in other words, very small.

However, we overlooked the possibility of non-convergence of the GEE estimation process due to high number of zero counts. This was in fact what happened in fitting the dependence model in some of the regressions. The GEE analysis converged for the independence model despite the zero counts because $\widehat{\beta}_0^I$ could be readily computed as the sample mean, and no other unknown parameter needed to be estimated. On the other hand, for the dependence model, the estimation of ρ might become difficult with too many zero observations.

Despite the non-convergence of some of the regressions, we did observe some very interesting results.

Firstly, as we expected, $\widehat{\beta}_0^I$ and $\widehat{\beta}_0^D$ were the exact same value for all the regressions that converged. However, contrary to what we believed, the *naive* $\text{Var}(\widehat{\beta}_0^I)$ estimate was always smaller than the *naive* $\text{Var}(\widehat{\beta}_0^D)$ estimate! Another curious observation is that the *robust* variance estimate of $\widehat{\beta}_0$ was exactly the same for both the independence and the dependence regression models. We can summarize the observations of these simulations as follows:

$$\begin{aligned} \widehat{\beta}_0^I &= \widehat{\beta}_0^D && (5.7) \\ \widehat{\text{Var}}(\widehat{\beta}_0^I)_{\text{robust}} &= \widehat{\text{Var}}(\widehat{\beta}_0^D)_{\text{robust}} && (5.8) \\ \widehat{\text{Var}}(\widehat{\beta}_0^I)_{\text{naive}} &< \widehat{\text{Var}}(\widehat{\beta}_0^D)_{\text{naive}}. && (5.9) \end{aligned}$$

To double-check that the result in (5.8) and (5.9) were not by chance, we ran a second set of simulations. However, this time, we used a much larger λ_1 and λ_2 to avoid the non-convergence of the GEE method.

In this simulation, we used the same n, r, and K as before, but $\lambda_1 = 5$ and $\lambda_2 = 50$. Thus, $\lambda = 5 + 50 = 55$, and $\rho = 50/55 \approx 0.91$ is the same large intraclass correlation coefficient as in the first simulation.

In this simulation, the GEE analysis successfully converged for both the independence and dependence models for each of the 100 regressions. From the results here, we confirmed that (5.8) and (5.9) were indeed not due to chance.

The result in equation (5.8) actually seems familiar to us. Recall that in the linear regression setting from Chapter 2, both the dependence and the independence models produce the exact same regression parameter estimate (and thus the same covariance matrix for the estimate) if the responses are exchangeably correlated, provided that an intercept is fitted in the model. However, the *naive* covariance for the independence model estimate is never smaller than that for the dependence model estimate in linear Gaussian regression. Why, then, is (5.9) true? That is, why is the *naive* variance estimate always smaller in the independence model than the one in the dependence model for a linear Poisson regression?

Note that the word *naive* in both Chapter 2 and this chapter has the same meaning. In Chapter 2, the *naive* covariance is the one calculated believing that the responses were truly independent. Here, we are *naive* in calculating the variance estimate for $\widehat{\beta}_0$ if we *naively* take the working covariance matrix $\boldsymbol{V}$ as the true covariance matrix of the responses. In other words, we believe that $\boldsymbol{V}$ is the true response covariance matrix, and thus the covariance estimate for $\widehat{\beta}_0$ obtained according to such a belief is *naive.*

In the next section, we present the theoretical reasoning for the results in (5.6) to (5.9). We also further compare the notion of *naiveté* in the contexts of Gaussian and

Poisson regression.

5.4 The Naive and Robust Variances for $\widehat{\beta}_0^I$ and $\widehat{\beta}_0^D$

We now present (5.6) to (5.9), with a slight variation to (5.9), as the next theorem, with a proof following it.

Theorem 4 *Suppose we have responses $Y_1^{(s)}, \ldots, Y_r^{(s)}$ sampled from a Poisson distribution with positive parameter $\lambda = \lambda_1 + \lambda_2$, such that they follow equations (5.1) to (5.4). In particular, for each s, $Y_i^{(s)}$'s are exchangeably correlated with intraclass correlation coefficient $\rho = \lambda_2/\lambda$. Note that $\lambda_2 > 0$, and hence $\rho > 0$.*

Now, suppose we fit a Poisson regression model (a GLM) according to (5.5), respectively assuming the responses to be (1) uncorrelated (the independence model) and (2) exchangeably correlated (the dependence model). We use the identity link in such a GLM. For the dependence model, we specify in GEE an exchangeable response working correlation structure without specifying the value of the intraclass correlation coefficient.

Then, equations (5.6) to (5.8) are true. Furthermore, $\widehat{\mathrm{Var}(\widehat{\beta}_0^I)}_{\mathrm{naive}} \neq \widehat{\mathrm{Var}(\widehat{\beta}_0^D)}_{\mathrm{naive}}$ in general.

To prove this theorem, we first need to "translate" our notations into the GEE notations. In particular, we need to express equations (3.17) to (3.22) in terms of λ_1, λ_2, and ρ.

Let us fix s. Since $\boldsymbol{Y}^{(s)}$'s are K i.i.d. vectors, the following is true for all $s = 1, \ldots, K$.

Recall that the design matrix $\boldsymbol{X}^{(s)}$ is simply the r-vector of 1's. Denote such a vector by $\mathbf{1}$. Also, we let $\eta_i^{(s)} = \lambda = \beta_0$ for all i so that we have the identity link. Hence, the conditional density for each $Y_i^{(s)}$ given $X_i^{(s)}$ is

$$f\left(y_i^{(s)} | x_{i.}^{(s)}\right) = \frac{e^{-\eta_i^{(s)}} (\eta_i^{(s)})^{y_i^{(s)}}}{y_i^{(s)}!}$$

$$\begin{aligned} &= \exp\{y_i^{(s)} \log \eta_i^{(s)} - \eta_i^{(s)} - \log y_i^{(s)}!\} \\ &= \exp\{y_i^{(s)} \log \lambda - \lambda - \log y_i^{(s)}!\}. \qquad (5.10) \end{aligned}$$

Comparing (5.10) to (3.17), we have $\theta_i^{(s)} = \log \eta_i^{(s)} = \log \lambda$ and $a(\theta_i^{(s)}) = \exp\{\theta_i^{(s)}\} = \lambda$. Also,

$$\begin{aligned} \mathrm{E}\left(Y_i^{(s)} | X_i^{(s)}\right) &= a'(\theta_i^{(s)}) = \exp\{\theta_i^{(s)}\} = \lambda \\ \mathrm{Var}\left(Y_i^{(s)} | X_i^{(s)}\right) &= a''(\theta_i^{(s)}) = \lambda. \end{aligned}$$

Thus,

$$\begin{aligned} \boldsymbol{A}^{(s)} &= \mathrm{diag}\left\{a'(\theta_i^{(s)})\right\} = \mathrm{diag}(\lambda, \ldots, \lambda) \\ \boldsymbol{\Delta}^{(s)} &= \mathrm{diag}\left\{\frac{\mathrm{d}\theta_i^{(s)}}{\mathrm{d}\eta_i^{(s)}}\right\} = \mathrm{diag}\left\{\frac{1}{\eta_i^{(s)}}\right\} = \mathrm{diag}\left(\frac{1}{\lambda}, \ldots, \frac{1}{\lambda}\right) \\ \Longrightarrow \boldsymbol{D}^{(s)} &= \boldsymbol{A}^{(s)} \boldsymbol{\Delta}^{(s)} \boldsymbol{X}^{(s)} = \boldsymbol{I1} = \boldsymbol{1} \qquad (5.11) \\ \boldsymbol{V}^{(s)} &= [\boldsymbol{A}^{(s)}]^{\frac{1}{2}} \boldsymbol{R} [\boldsymbol{A}^{(s)}]^{\frac{1}{2}} = \lambda \boldsymbol{R} \qquad (5.12) \\ \Longrightarrow [\boldsymbol{V}^{(s)}]^{-1} &= \frac{1}{\lambda} \boldsymbol{R}^{-1} \qquad (5.13) \end{aligned}$$

where $\boldsymbol{R}$ is the working correlation matrix that GEE uses in its estimation process. Since we specify an exchangeable response correlation structure for the dependence model, $\boldsymbol{R}$ has all non-diagonal values equal to some α. In each iteration of the estimation process, the values of both α and β_0 are updated. Upon convergence, the final value of α is the estimate of ρ. That is, $\hat{\rho} = \alpha$. Also, upon convergence of the GEE estimation process, equations (3.21) and (3.22) are used to estimate $\mathrm{Var}(\widehat{\beta}_0^I)$ and $\mathrm{Var}(\widehat{\beta}_0^D)$ (using the final values of $\widehat{\beta}_0$ and α).

Note that in the GEE estimation process, we do not expect the matrix $\boldsymbol{D}^{(s)}$ to be computed based on the data. This is because when we specify the Poisson distribution of the responses, and the identity link function, with a design matrix which is simply the **1** vector, the matrix product $\boldsymbol{A}^{(s)}\boldsymbol{\Delta}^{(s)} = \boldsymbol{I}$ is readily obtained, without the need to estimate any unknown parameter in this product. However, the computation of $\boldsymbol{A}^{(s)}$ and $\boldsymbol{\Delta}^{(s)}$ (and hence, of $\boldsymbol{V}^{(s)}$) individually should be data based, because they both involve the unknown parameter λ. By the identity link, we have $\lambda = \beta_0$, and hence, $\hat{\lambda} = \widehat{\beta}_0$. Finally, as mentioned on page 82, the covariance matrix $\text{Cov}(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)})$ in equation (3.21) is estimated by $\boldsymbol{S}^{(s)}\boldsymbol{S}^{(s)T}$ in the GEE estimation process, where $\boldsymbol{S}^{(s)} = \boldsymbol{Y}^{(s)} - \boldsymbol{a}'(\widehat{\boldsymbol{\theta}}^{(s)})$.

Now, we may substitute equations (5.11) to (5.13) into equations (3.20) to (3.22) to prove Theorem 4.

Proof of Theorem 4

First, let us prove equation (5.6).

For either the independence or dependence model, substituting equations (5.11) to (5.13) into estimating equation (3.20) gives

$$\begin{aligned}
\mathbf{0} &= \sum_{s=1}^{K} \mathbf{1}^T \frac{1}{\lambda} \boldsymbol{R}^{-1} \left[\mathbf{Y}^{(s)} - \lambda \mathbf{1}\right] \\
&\quad \text{(since } a'(\theta_i^{(s)}) = \lambda \text{ for all } i \text{ for all } s\text{)} \\
\iff \quad \mathbf{1}^T \boldsymbol{R}^{-1} \sum_{s=1}^{K} \mathbf{Y}^{(s)} &= \beta_0 K \mathbf{1}^T \boldsymbol{R}^{-1} \mathbf{1} \qquad (5.14) \\
&\quad \text{(since } \lambda = \beta_0 \text{ by the identity link).}
\end{aligned}$$

For the independence model, $\boldsymbol{R} = \boldsymbol{R}^{-1} = \boldsymbol{I}$, so that equation (5.14) becomes

$$\begin{aligned}
\sum_{s=1}^{K} \mathbf{1}^T \mathbf{Y}^{(s)} &= \beta_0 K \mathbf{1}^T \mathbf{1} \qquad (5.15) \\
\iff \quad \sum_{s=1}^{K} \sum_{i=1}^{r} Y_i^{(s)} &= \beta_0 K r \\
\iff \quad \beta_0 &= \frac{1}{Kr} \sum_{s=1}^{K} \sum_{i=1}^{r} Y_i^{(s)} = \overline{Y}.
\end{aligned}$$

Hence, $\widehat{\beta}_0^I = \overline{Y}$.

Now, recall from Chapter 2 that for the dependence model, we can write $\boldsymbol{R}(\alpha) = (1-\alpha)\boldsymbol{I}+\alpha\mathbf{1}\mathbf{1}^T$. (Again, α is the estimate of ρ, the intraclass correlation coefficient of the exchangeably correlated responses.) Also, from Lemma 1 and the proof of Theorem 2, we know that

$$\boldsymbol{R}^{-1} = \tfrac{1}{1-\alpha}\,[\boldsymbol{I} - \tfrac{\alpha}{1+(r-1)\alpha}\mathbf{1}\mathbf{1}^T]. \tag{5.16}$$

Thus, we have

$$\begin{aligned}
\mathbf{1}^T\boldsymbol{R}^{-1} &= \mathbf{1}^T\frac{1}{1-\alpha}\left[\boldsymbol{I} - \frac{\alpha}{1+(r-1)\alpha}\mathbf{1}\mathbf{1}^T\right] \\
&= \frac{1}{1-\alpha}\left[\mathbf{1}^T - \frac{\alpha}{1+(r-1)\alpha}r\mathbf{1}^T\right] \\
&= \frac{1}{1-\alpha}\left[\frac{1-\alpha}{1+(r-1)\alpha}\right]\mathbf{1}^T \\
&= \frac{1}{1+(r-1)\alpha}\mathbf{1}^T.
\end{aligned}\tag{5.17}$$

Substituting equation (5.17) into estimating equation (5.14), we have

$$\begin{aligned}
\frac{1}{1+(r-1)\alpha}\sum_{s=1}^{K}\mathbf{1}^T\boldsymbol{Y}^{(s)} &= \beta_0 K\left(\frac{1}{1+(r-1)\alpha}\right)\mathbf{1}^T\mathbf{1} \\
\Longleftrightarrow \qquad \sum_{s=1}^{K}\mathbf{1}^T\boldsymbol{Y}^{(s)} &= \beta_0 K\mathbf{1}^T\mathbf{1}
\end{aligned}$$

which is the same as estimating equation (5.15) whose solution is $\beta_0 = \overline{Y}$. In other words, $\widehat{\beta}_0^D = \overline{Y} = \widehat{\beta}_0^I$. This proves equation (5.6) (and thus (5.7)).

Next, we will prove equation (5.8).

For either the independence or dependence model, substituting equations (5.11) to (5.13) into equation (3.21) gives

$$\begin{aligned}
\widehat{\mathrm{Var}(\widehat{\beta}_0)}_{\text{robust}} = &\left(\sum_{s=1}^{K}\mathbf{1}^T\left[\frac{1}{\hat{\lambda}}\boldsymbol{R}^{-1}\right]\mathbf{1}\right)^{-1}\left\{\sum_{s=1}^{K}\mathbf{1}^T\left[\frac{1}{\hat{\lambda}}\boldsymbol{R}^{-1}\right]\boldsymbol{S}^{(s)}\boldsymbol{S}^{(s)T}\left[\frac{1}{\hat{\lambda}}\boldsymbol{R}^{-1}\right]\mathbf{1}\right\} \\
&\left(\sum_{s=1}^{K}\mathbf{1}^T\left[\frac{1}{\hat{\lambda}}\boldsymbol{R}^{-1}\right]\mathbf{1}\right)^{-1}
\end{aligned}$$

$$\begin{aligned} &= \frac{\frac{1}{\hat{\lambda}^2}\sum_{s=1}^{K} \mathbf{1}^T \boldsymbol{R}^{-1} \boldsymbol{S}^{(s)} \boldsymbol{S}^{(s)T} \boldsymbol{R}^{-1} \mathbf{1}}{\left(K\frac{1}{\hat{\lambda}}\mathbf{1}^T \boldsymbol{R}^{-1}\mathbf{1}\right)^2} \\ &= \frac{\mathbf{1}^T \boldsymbol{R}^{-1}\left(\sum_{s=1}^{K} \boldsymbol{S}^{(s)} \boldsymbol{S}^{(s)T}\right) \boldsymbol{R}^{-1}\mathbf{1}}{K^2\left(\mathbf{1}^T \boldsymbol{R}^{-1}\mathbf{1}\right)^2}. \end{aligned} \tag{5.18}$$

For the independence model, $\boldsymbol{R} = \boldsymbol{I}$, and equation (5.18) becomes

$$\begin{aligned} \widehat{\mathrm{Var}}(\hat{\beta}_0^I)_{\text{robust}} &= \frac{\mathbf{1}^T\left(\sum_{s=1}^{K} \boldsymbol{S}^{(s)} \boldsymbol{S}^{(s)T}\right)\mathbf{1}}{K^2(\mathbf{1}^T\mathbf{1})^2} \\ &= \frac{\mathbf{1}^T\left(\sum_{s=1}^{K} \boldsymbol{S}^{(s)} \boldsymbol{S}^{(s)T}\right)\mathbf{1}}{K^2 r^2}. \end{aligned}$$

For the dependence model, first notice that $\boldsymbol{R}^{-1} = (\boldsymbol{R}^{-1})^T$ because $\boldsymbol{R}$ (and thus $\boldsymbol{R}^{-1}$) is symmetric. Thus, by equation (5.17), we have

$$\begin{aligned} \boldsymbol{R}^{-1}\mathbf{1} &= (\mathbf{1}^T \boldsymbol{R}^{-1})^T \\ &= \frac{1}{1+(r-1)\alpha}\mathbf{1}. \end{aligned} \tag{5.19}$$

Now, we substitute equations (5.17) and (5.19) into equation (5.18) and have

$$\begin{aligned} \widehat{\mathrm{Var}}(\hat{\beta}_0^D)_{\text{robust}} &= \frac{\sum_{s=1}^{K} \frac{\mathbf{1}^T \boldsymbol{S}^{(s)} \boldsymbol{S}^{(s)T}\mathbf{1}}{[1+(r-1)\alpha]^2}}{K^2\left(\frac{r}{1+(r-1)\alpha}\right)^2} \\ &= \frac{\mathbf{1}^T\left(\sum_{s=1}^{K} \boldsymbol{S}^{(s)} \boldsymbol{S}^{(s)T}\right)\mathbf{1}}{K^2 r^2} \\ &= \widehat{\mathrm{Var}}(\hat{\beta}_0^I)_{\text{robust}}. \end{aligned}$$

This proves equation (5.8).

Now, to prove that $\widehat{\mathrm{Var}}(\hat{\beta}_0^I)_{\text{naive}} \neq \widehat{\mathrm{Var}}(\hat{\beta}_0^D)_{\text{naive}}$, we substitute equations (5.11) to (5.13) into equation (3.22). Then, for either the dependence or independence model,

$$\begin{aligned} \widehat{\mathrm{Var}}(\hat{\beta}_0)_{\text{naive}} &= \frac{1}{\sum_{s=1}^{K} \mathbf{1}^T \frac{1}{\hat{\lambda}} \boldsymbol{R}^{-1}\mathbf{1}} \\ &= \frac{\hat{\beta}_0}{K\mathbf{1}^T \boldsymbol{R}^{-1}\mathbf{1}} \\ &\quad \text{(since } \lambda = \beta_0 \text{ by the identity link).} \end{aligned} \tag{5.20}$$

First, for the dependence model, the working correlation matrix $\boldsymbol{R}$ has inverse as in equation (5.16). Thus, we can substitute equation (5.17) into equation (5.20) and obtain

$$\widehat{\text{Var}}(\widehat{\beta}_0^D)_{\text{naive}} = \frac{\widehat{\beta}_{0.}^D}{K}\left(\frac{1+(r-1)\alpha}{r}\right).$$

By equation (5.7), we may denote both $\widehat{\beta}_0^D$ and $\widehat{\beta}_0^I$ by a common $\widehat{\beta}_0$. The above equation then becomes

$$\widehat{\text{Var}}(\widehat{\beta}_0^D)_{\text{naive}} = \frac{\widehat{\beta}_0}{K}\left(\frac{1+(r-1)\alpha}{r}\right). \tag{5.21}$$

Again, for the independence model, $\boldsymbol{R} = \boldsymbol{R}^{-1} = \boldsymbol{I}$. Thus, equation (5.20) simply becomes

$$\begin{aligned}\widehat{\text{Var}}(\widehat{\beta}_0^I)_{\text{naive}} &= \frac{\widehat{\beta}_0^I}{K\mathbf{1}^T\mathbf{1}} \\ &= \frac{\widehat{\beta}_0}{Kr} \\ &= \frac{\widehat{\text{Var}}(\widehat{\beta}_0^D)_{\text{naive}}}{1+(r-1)\alpha}.\end{aligned}$$

Therefore, in general, unless $\alpha = 0$, we will have $\widehat{\text{Var}}(\widehat{\beta}_0^I)_{\text{naive}} \neq \widehat{\text{Var}}(\widehat{\beta}_0^D)_{\text{naive}}$.

Q.E.D.

Notice that, in our simulations, we generated the Poisson responses with an extremely high intraclass correlation ρ (> 0.9). Since α estimates ρ, we naturally observed all positive α values in our simulations. In fact, in both sets of simulations, α turned out to be an extremely good estimate of ρ (α between 0.8 and 1) in all of the n observations which converged. This large positive α explains the result in (5.9).

In principal, α can take on negative values. In particular, if we generated the responses with a near-zero (or zero) ρ, then it is possible to have negative α values. We can easily see that $\widehat{\text{Var}}(\widehat{\beta}_0^I)_{\text{naive}} > \widehat{\text{Var}}(\widehat{\beta}_0^D)_{\text{naive}}$ if $\alpha < -1/(r-1)$.

Also, notice that $\widehat{\beta}_0 = \overline{Y}$ is the MLE of β_0 *when assuming response independence.* Recall from Chapter 2 the notion of the naive statistician who blindly believes that the responses are truly independent. Thus, the theoretical variance of $\widehat{\beta}_0^I$ that he computes is

$$\mathrm{Var}(\widehat{\beta}_0) = \mathrm{Var}(\overline{Y}) = \frac{\lambda}{Kr} = \frac{\beta_0}{Kr}.$$

The estimate of this variance is naturally computed by replacing β_0 with $\widehat{\beta}_0^I$. It, then, is indeed the *naive* variance estimate, $\widehat{\mathrm{Var}(\widehat{\beta}_0^I)}_{\text{naive}}$. From this set up of a GLM, it is clear that the two notions of *naiveté* — that of the naive statistician and that of the naive variance estimate — coincide.

5.4.1 More on the simulation results

Let us now return to the question of why the *naive* variance estimate is always smaller than the *robust* variance estimate in our simulations. Such a phenomenon creates an illusion that Naive, using his/her *naive* variance estimate, has better efficiency in his/her inference than that of Enlightened who bases it on his/her *naive* variance estimate. However, a closer look reveals that the former estimate is simply the MLE variance estimate in the case of (assumed) independent responses coming from a single population, forming one big i.i.d. sample (of size Kr). Naturally, such a variance estimate must be smaller than the one computed assuming the r responses per replicate to be (positively) correlated, and hence more *between-group* response variability among the K replicates.

As for the *robust* variance estimates, we have the results from Theorem 4 for both the independence and dependence models. Such variance estimates are derived from the asymptotic distribution of $\widehat{\beta}_0$ presented in Theorem 3. For convenience, instead of determining the true variance of $\widehat{\beta}_0$, we use the sample S.D. of the n $\widehat{\beta}_0$'s to estimate the true SE($\widehat{\beta}_0$). With enough datasets, n, we may assume this sample S.D. to be reasonably

β_0	sample mean of 50 $\widehat{\beta}_0^I$'s (= $\widehat{\beta}_0^D$'s)	$\widehat{\mathrm{SE}}(\widehat{\beta}_0^I)_{\text{sample}}$ $= \widehat{\mathrm{SE}}(\widehat{\beta}_0^D)_{\text{sample}}$	$\widehat{\mathrm{SE}}(\widehat{\beta}_0^D)_{\text{naive}}$	$\widehat{\mathrm{SE}}(\widehat{\beta}_0^I)_{\text{robust}}$ $= \widehat{\mathrm{SE}}(\widehat{\beta}_0^D)_{\text{robust}}$	$\widehat{\mathrm{SE}}(\widehat{\beta}_0^I)_{\text{naive}}$
55	59.90	1.69	1.66	1.53	0.77

Table 5.16: S.E. estimate comparison based on the first 50 simulated datasets.

close to the true S.E. On the other hand, our goal is only to obtain some idea of the relative variabilities of the intercept estimates. Thus, in this investigation, we decide not to use large scale simulation experiments although they would allow better understanding of the relative variabilities of the estimates.

Let us revisit our latter simulation where the GEE method converged for all its $n = 100$ datasets. Our results, based on the first 50 datasets in this simulation,[4] are summarized in Table 5.16. For the independence model, $\widehat{\mathrm{SE}}(\widehat{\beta}_0^I)_{\text{sample}}$ denotes the square root of the sample variance of the 50 $\widehat{\beta}_0^I$'s. $\widehat{\mathrm{SE}}(\widehat{\beta}_0^I)_{\text{naive}}$ and $\widehat{\mathrm{SE}}(\widehat{\beta}_0^I)_{\text{robust}}$ respectively denote the sample means of the 50 *naive* and the 50 *robust* estimates of $\mathrm{SE}(\widehat{\beta}_0^I)$. We denote those S.E. estimates for $\widehat{\beta}_0^D$ (using the dependence model) in a similar way. Of course, $\widehat{\mathrm{SE}}(\widehat{\beta}_0^I)_{\text{sample}} = \widehat{\mathrm{SE}}(\widehat{\beta}_0^D)_{\text{sample}}$ by equation (5.7), and $\widehat{\mathrm{SE}}(\widehat{\beta}_0^I)_{\text{robust}} = \widehat{\mathrm{SE}}(\widehat{\beta}_0^D)_{\text{robust}}$ by equation (5.8).

The results in Table 5.16 show that $\overline{Y}$, the sample mean of the responses, is a reasonbly good estimate of the true β_0. Using $\widehat{\mathrm{SE}}(\widehat{\beta}_0^I)_{\text{sample}}(= \widehat{\mathrm{SE}}(\widehat{\beta}_0^D)_{\text{sample}})$ as the "gold standard" (g.d.) of the true $\mathrm{SE}(\widehat{\beta}_0)$ for both models, $\widehat{\mathrm{SE}}(\widehat{\beta}_0^D)_{\text{naive}}$ is thus closest to the g.d. among the *naive* and *robust* estimates. It is followed by $\widehat{\mathrm{SE}}(\widehat{\beta}_0^I)_{\text{robust}} = \widehat{\mathrm{SE}}(\widehat{\beta}_0^D)_{\text{robust}}$, and $\widehat{\mathrm{SE}}(\widehat{\beta}_0^I)_{\text{naive}}$ is the farthest away from the g.d. In fact, $\widehat{\mathrm{SE}}(\widehat{\beta}_0^I)_{\text{naive}}$ is the only S.E. estimate which gives a 95% confidence interval that does not contain the true β_0. This is an indication that, in

[4] The way we recorded the simulation output made it difficult to base our results here on all 100 datasets.

general,[5] the *naive* variance estimate of the independence model is a poor estimate which substantially underestimates the true Var($\widehat{\beta}_0$). The underestimation can be explained by the between-group variability among the K replicates being ignored in the computation of this variance estimate. In other words, if response correlation is ignored when fitting the GLM, then inference based on the *naive* variance estimate may be far off from the truth. Fortunately, inference done by ignoring response correlation but using the *robust* variance estimate does not affect estimation efficiency, in the sense that this variance estimate is the same as the *robust* variance estimate considering response correlation. Of course, the advantage of the dependence model over the independence model is the possibility of estimation of the parameters in the response correlation structure.

5.5 Concluding Remarks

The above results of Theorem 4 and the size comparison of the different S.E. estimates of $\widehat{\beta}_0$ are all based on the particular GLM we are dealing with here. Therefore, we may not generalize our observations to cover such conditions as GLM's with covariates and/or non-linear link functions.

Simulations of non-linear Poisson regressions may be difficult, mainly due to the complexity of generating correlated (conditional on the covariates) Poisson count data with non-linear link to the covariates.[6] However, based on our results, we recommend the use of the *robust* variance estimates, regardless of the assumption of correlated or independent responses, for such non-linear regressions of real data. This is because we have seen from simulation results here and in Chapter 4 that the data analyses based on the *robust* variance estimates for both the independence and dependence models are generally very similar. We have also seen that in our set up of a GLM in this

[5] ... for the conditions as in Theorem 4.

[6] A method for this task is developed in the Ph.D. thesis by Jianmeng Xu (1996).

chapter, the *naive* variance estimate of the dependence model estimates the true variance overwhelmingly better than that of the independence model.

However, assuming the dependence model allows the estimation of the true response correlation matrix if its structure is known or can be approximated. In this case, one should base inference on the dependence model instead of the independence model.

An extra remark is that the GEE method seems to estimate reasonably well all the model parameters, i.e. the regression parameters, and those in the response correlation structure, even with relatively small r and K (5 and 20 in our simulations).

Chapter 6

Discussion

In this thesis, we have presented some results on the effect of misspecified response correlation in regression analysis. We focussed on regression analyses applied to health impact studies, in which the responses are often obtained from neighboring geographical sites and thus are spatially correlated. (See Chapter 1.) In Chapter 2, we made use of the storyline of the naive and enlightened statisticians to illustrate the severe loss of parameter estimation efficiency that might result from ignoring response correlation when fitting a linear response-covariate relationship with Gaussian responses. In particular, both theory and our simulation results indicate that Naive's regression analysis is almost always[1] less powerful/efficient than Enlightened's. Such and other forms of response correlation misspecification in GLM's were examined in Chapters 4 and 5. The GLM's were fitted via the GEE approach, which allows the incorporation of response correlation in the parameter estimation process. Our results suggest reasonable robustness[2] of the GEE approach to misspecification of response correlation with a small number (about 20) of independent replicates/clusters in the sample if the analysis is based on the *robust* variance estimates. In other words, the GLM analyses based on such variance estimates using the correct and incorrect response correlations respectively both have similar efficiencies. The analyses based on the *naive* variance estimates are similarly efficient if the response correlation structure is correctly specified. Otherwise, the true variance estimates may

[1] Theoretically, Naive's analysis is never more powerful. However, it can be more powerful empirically due to chance.

[2] ... with those GLM's that we investigated.

be considerably underestimated, thus falsely revealing high efficiencies of the analyses. Moreover, when the response correlation structure is correctly specified, the parameters in the correlation matrix are relatively well estimated[3] even with very few replicates (10 to 20) and correlated measurements per series[4] (about 5).

However, we have not examined the GLM setting in which the clusters are *not* independent. For example, if the replicates are repeated measurements obtained over a short period of time, they may be temporally correlated. Thus, both spatial and temporal correlations may exist in the responses: the measurements within a replicate may be spatially correlated, and the replicates may be temporally correlated.

Studying the effect of ignoring correlation among replicates in a GLM analysis may be very complex although important to health impact investigations. The complexity lies in the incorporation of such correlation into the GLM and hence the parameter estimation process. In fact, the GEE approach only allows incorporation of correlation along a series, but strictly assumes the replicates to be independent. This assumption is often met in longitudinal data analyses in which the measurements are correlated temporally, but the measurement replicates are obtained from separate individuals and thus independent. On the other hand, the assumption of independent replicates may fail in some health impact studies for reasons mentioned earlier. Unfortunately, an alternative method to the GEE approach which accounts for response correlation along both spatial and temporal "axes" in parameter estimation is apparently not available.

[3] ... provided that the GEE estimation process converges. Non-convergence of the process may result due to insufficient information in the data (e.g. too many unknown parameters or too many zero counts in a Poisson regression).

[4] See page 110.

Bibliography

[1] Aitchison, J., Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76**, 643–653.

[2] Anderson, Theodore W. (1971). *The Statistical Analysis of Time Series.* John Wiley & Sons.

[3] Basu, Sabyasachi, Reinsel, Gregory C. (1994). Regression Models With Spatially Correlated Errors. *American Statistical Association Journal* **89**, 88–99.

[4] Cressie, Noel A. C. (1991). *Statistics For Spatial Data.* John Wiley & Sons.

[5] Fitzmaurice, Garrett M. (1995). A Caveat Concerning Independence Estimating Equations With Multivariate Binary Data. *Biometrics* **51**, 309–317.

[6] Freedman, D., Pisani, R., Purves, R., Adhikari, A. (1991). *Statistics,* 2nd Ed. W. W. Norton & Co.

[7] Graybill, Franklin A. (1983). *Matrices with Applications in Statistics,* 2nd Ed. Wadsworth.

[8] Herzberg, Agnes M. (1988). Some further results for the equivalence of ordinary least squares and weighted least squares estimators: an advantage for rotable designs. *SIAM Institute for Mathematics and Society Technical Report* #116.

[9] Johnson, Richard A., Wichern, Dean W. (1982). *Applied Multivariate Statistical Analysis.* Prentice Hall.

[10] Kotz, Samuel, Johnson, Norman L., Read, Campbell B. (1985). *Encyclopedia of Statistical Sciences.* John Wiley & Sons.

[11] Liang, Kung-Yee, Zeger, Scott L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

[12] McElroy, F. W. (1967). A necessary and sufficient condition that ordinary least squares estimators be best linear unbiased. *American Statistical Association Journal* **62**, 1302–1304.

[13] Marascuilo, Leonard A., Levin, Joel R. (1983). *Multivariate Statistics in the Social Sciences.* Brooks/Cole.

[14] Martin, R. J. (1982). Some aspects of experimental design and analysis when errors are correlated. *Biometrika* **69**, 597–612.

[15] Mosteller, Frederick, Tukey, John W. (1977). *Data Analysis and Regression.* Addison-Wesley.

[16] Plackett, Roger L. (1960). *Principles of Regression Analysis.* Oxford @ The Clarendon Press.

[17] Rao, C. Radhakrishna (1965). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. *Fifth Berkeley Symposium* **1**, 355–372.

[18] Rosychuk, Rhonda J. (1994). Cancer cluster detection in British Columbia school districts 1983–1989. *Master of Science Thesis.* University of British Columbia.

[19] Seber, George A. F. (1977). *Linear Regression Analysis.* John Wiley & Sons.

[20] Weisberg, Sanford (1985). *Applied Linear Regression,* 2nd Ed. John Wiley & Sons.

[21] Zeger, Scott L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621–629.

[22] Zeger, Scott L., Liang, Kung-Yee (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics* **42**, 121–130.

[23] Zyskind, George (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Annals of Mathematical Statistics* **38**, 1092–1109.