MULTICOLLINEARITY, AUTOCORRELATION, AND RIDGE REGRESSION

by

JACKIE JEN-CHY HSU

B.A. in Econ., The National Taiwan University, 1977

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE (BUSINESS ADMINISTRATION)
IN
THE FACULTY OF GRADUATE STUDIES*

THE FACULTY OF COMMERCE AND BUSINESS ADMINISTRATION

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

February 1980

c Jackie Jen-Chy Hsu, 1980
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 representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Commerce & Business Admin.

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date Feb. 8, 1980
ABSTRACT

The presence of multicollinearity can induce large variances in the ordinary Least-squares estimates of regression coefficients. It has been shown that ridge regression can reduce this adverse effect on estimation. The presence of serially correlated error terms can also cause serious estimation problems. Various two-stage methods have been proposed to obtain good estimates of the regression coefficients in this case. Although the multicollinearity and autocorrelation problems have long been recognized in regression analysis, they are usually dealt with separately. This thesis explores the joint effects of these two conditions on the mean square error properties of the ordinary ridge estimator as well as the ordinary least-squares estimator. We show that ridge regression is doubly advantageous when multicollinearity is accompanied by autocorrelation in both the errors and the principal components. We then derive a new ridge type estimator that is adjusted for autocorrelation.

Finally, using simulation experiments with different degrees of multicollinearity and autocorrelation, we compare the mean square error properties of various estimators.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2.</td>
<td>NOTATION AND PRELIMINARIES</td>
<td>3</td>
</tr>
<tr>
<td>3.</td>
<td>MULTICOLLINEARITY</td>
<td>7</td>
</tr>
<tr>
<td>3.1</td>
<td>Sources</td>
<td>7</td>
</tr>
<tr>
<td>3.2</td>
<td>Effects</td>
<td>8</td>
</tr>
<tr>
<td>3.3</td>
<td>Detection</td>
<td>9</td>
</tr>
<tr>
<td>4.</td>
<td>AUTOCORRELATION</td>
<td>13</td>
</tr>
<tr>
<td>4.1</td>
<td>Sources</td>
<td>13</td>
</tr>
<tr>
<td>4.2</td>
<td>Effects</td>
<td>14</td>
</tr>
<tr>
<td>4.3</td>
<td>Detection</td>
<td>15</td>
</tr>
<tr>
<td>5.</td>
<td>JOINT EFFECTS OF MULTICOLLINEARITY AND AUTOCORRELATION</td>
<td>18</td>
</tr>
<tr>
<td>5.1</td>
<td>Mean Square Error of the OLS Estimates of $\hat{\beta}$</td>
<td>18</td>
</tr>
<tr>
<td>5.2</td>
<td>Mean Square Error of the Ridge Estimates of $\hat{\beta}$</td>
<td>21</td>
</tr>
<tr>
<td>5.3</td>
<td>When will Ridge estimates be better than the OLS estimates?</td>
<td>24</td>
</tr>
<tr>
<td>5.4</td>
<td>Use of the &quot;Ridge Trace&quot;</td>
<td>27</td>
</tr>
<tr>
<td>6.</td>
<td>RIDGE REGRESSION: ESTIMATES, MEAN SQUARE ERROR AND PREDICTION</td>
<td>29</td>
</tr>
<tr>
<td>6.1</td>
<td>Derivation of Ridge Estimator for a CLR Model</td>
<td>29</td>
</tr>
<tr>
<td>6.2</td>
<td>Derivation of Ridge Estimator for an ALR Model</td>
<td>32</td>
</tr>
<tr>
<td>6.3</td>
<td>Mean Square Error of the &quot;Generalized Estimates&quot;</td>
<td>34</td>
</tr>
<tr>
<td>6.4</td>
<td>Estimation</td>
<td>35</td>
</tr>
<tr>
<td>6.5</td>
<td>Prediction</td>
<td>37</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS (cont'd)

7. THE MONTE CARLO STUDY  
   7.1 Design of the Experiments  38  
   7.2 Sampling Results  43  
      7.2a. Results assuming $\rho_\varepsilon$ is known  43  
      7.2b. Results assuming $\rho_\varepsilon$ is unknown  45  
      7.2c. Forecasting  53  

8. CONCLUSIONS  56  

REFERENCES  59
INTRODUCTION

Multicollinearity and Autocorrelation are two very common problems in regression analysis. As is well-known, the presence of some degree of multicollinearity results in estimation, instability and model misspecification while the presence of serially correlated errors leads to underestimation of the variances of parameter estimates and inefficient prediction. Because these two conditions have adverse effects on estimation and prediction, a wide range of tests have been developed to reduce their impact. Invariably, the multicollinearity and autocorrelation problems are dealt with separately in most if not all the proceedings.

In this thesis we address the question "What are the joint effects of multicollinearity and autocorrelation on estimation and prediction?" Thereafter we shall study analytically the possible changes in the effectiveness of various estimation methods in the joint presence of these two conditions. As a result of these new findings, a new ridge estimator adjusted for autocorrelation is then proposed and its properties are investigated by conducting a simulation study.

We briefly outline this thesis. Section 2 provides the setting for our analysis. Sections 3 and 4 give a general discussion of the problems of multicollinearity and autocorrelation. In addition, we comment on the validity of various existing diagnostic tests. The analytical study of the joint effects of multicollinearity and autocorrelation is presented in Section 5. In Section 6, a new ridge estimator adjusted for autocorrelation is derived and its mean square error properties are analyzed. Also, we discuss how these new estimates can be obtained in practice. The methodology and the results of sampling experiments appear in Section
7. The thesis concludes with the presentation of several two-stage methods that can be used with the new ridge rule that hopefully will achieve better estimates and predictions.
2. NOTATION AND PRELIMINARIES

The Classical Linear Regression (CLR) model can be represented by the equation

\[(2.1) \quad Y = X\beta + \varepsilon\]

where \( Y \) is a \( nx1 \) vector of observations on the dependent variable, \( X \) is a \( nxp \) matrix of observations on the explanatory variables, \( \beta \) is a \( px1 \) vector of regression coefficients to be estimated and \( \varepsilon \) is a \( nx1 \) vector of true error terms. The standard assumptions of the linear regression model are:

\begin{enumerate}
  \item \( \mathbb{E}(\varepsilon) = 0 \), where 0 is the zero vector
  \item \( \mathbb{E}(\varepsilon\varepsilon^T) = \sigma^2 I \), where I is the identity matrix.
  \item The explanatory variables are non-stochastic, hence they are independent of the error terms.
  \item \( \text{Rank } (X) = p \leq n \).
\end{enumerate}

The Ordinary Least-squares (OLS) estimator of \( \beta \) is given

\[(2.2) \quad \hat{\beta}_{OLS} = (X^TX)^{-1}X^TY\]

with variance-covariance matrix

\[(2.3) \quad \text{Var}(\hat{\beta}_{OLS}) = \sigma^2 (X^TX)^{-1} \]

For simplicity, we will assume that \( (X^TX) \) is in correlation form. Let \( P \) be the \( pxp \) orthogonal matrix such that \( PX^TP = \Lambda \) where \( \Lambda \) is a diagonal matrix with the eigenvalues of \( (X^TX) \), \( \lambda_1, \ldots, \lambda_p \), displayed on the diagonal of \( \Lambda \). We assume further that

\[ \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p. \]
After applying an orthogonal rotation, $P$, it follows from (2.1) that

$$E(Y) = X^T P^T P \hat{\beta} = X* \alpha$$

where $X^* = X P^T$ is the data matrix represented in the rotated coordinates and the columns of $X^*$ are linearly independent. The vector $\alpha = P \hat{\beta}$ is the vector of regression coefficients of the principal components. It follows that the OLS estimator of $\alpha$ is given by

$$\hat{\alpha}_{OLS} = (X^* X^*)^{-1} X^* Y = P \hat{\beta}_{OLS}.$$ 

We will consider ridge estimators for $\beta$ of the form

$$\hat{\beta}_R(k) = (X^T X + kI)^{-1} X^T Y \quad 0 \leq k \leq 1.$$ 

Where $k$ is independent of $\hat{\beta}_{OLS}$. When $k$ is a function of $\hat{\beta}_{OLS}$, $\hat{\beta}_R(k)$ is called an "adaptive ridge estimator" [12]. If $kI$ is replaced by a symmetric nonnegative definite matrix $k$, then the estimator is said to be a "generalized ridge estimator" [11:p.63].

Expressed in the rotated coordinates, the ridge estimator for $\alpha$ is given by

$$\hat{\alpha}_R(k) = Z \hat{\alpha}_{OLS}$$

where $Z = (A + kI)^{-1} A$.

By substituting $X*P = X$ in (2.6), I found

$$\hat{\beta}_R(k) = (P^T A P + kI)^{-1} P^T A \hat{\beta}_{OLS}$$

$$= P^T A \hat{\beta}_{OLS}.$$ 

It follows from (2.7) that

$$\hat{\alpha}_R(k) = P \hat{\beta}_R(k).$$
For the CLR model, assumption (2) that the errors are uncorrelated is often violated in practice. This leads to the formulation of an Autoregressive Linear Regression (ALR) model. Mathematically the ALR model is given by replacing assumptions (2) and (3) by assumptions (2') and (3') below.

\[(2') \quad E(\xi^T) = \sigma_u^2 \Omega \quad \text{where} \quad \Omega \text{ is a nondiagonal positive definite matrix.}\]

\[(3') \quad X_t = [x_{t1}, x_{t2}, \ldots, x_{tp}] \quad \text{the observation on the explanatory variables, is independent of the contemporaneous and succeeding errors,} \]
\[\epsilon_t, \epsilon_{t+1}, \ldots, \epsilon_n.\]

We assume that the error term \(\epsilon_t\) follows a first-order autoregressive scheme, that is

\[(2.9) \quad \epsilon_t = \rho \epsilon_{t-1} + U_t\]

where \(\rho\) is the autocorrelation coefficient. We require that \(|\rho| < 1\) and that \(U_t\) satisfies the following for all \(t\)

\[(2.10) \quad E(U_t) = 0\]
\[E(U_t U_{t+s}) = \sigma_u^2 \quad s = 0\]
\[E(U_t U_{t+s}) = 0 \quad s \neq 0\]

and

\[(2.11) \quad E(\xi^T) = \sigma_u^2 \Phi\]

where

\[\Phi = \begin{pmatrix}
1 & \rho \epsilon & \rho^2 \epsilon & \ldots & \rho^{n-1} \\
\rho \epsilon & 1 & \rho \epsilon & \ldots & \rho^{n-2} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\rho^{n-1} \epsilon & \rho^{n-2} \epsilon & \ldots & 1
\end{pmatrix}\]
For an ALR model, the "Generalized Least-squares" (GLS) will give the "Best Linear Unbiased Estimator" (BLUE) of $\beta$, denoted as $\hat{\beta}_{GLS}$. The matrix $\Omega$ can be written

$$\Omega = \Omega^T$$

where $\Omega$ is nonsingular. Hence

$$\Omega^{-1} \Omega (\Omega^{-1})^T = I$$

and

$$\Omega^{-1} \Omega^{-1} = \Omega^{-1}$$

$\hat{\beta}_{GLS}$ is obtained by making the following substitution in the ALR model;

$$Y_* = \Omega^{-1} Y$$
$$X_* = \Omega^{-1} X$$
$$\xi_* = \Omega^{-1} \xi$$

Then it follows that

$$Y_* = X_* \hat{\beta} + \xi_*$$

$$E(\xi_* \xi_*^T) = \sigma^2_{ulY}$$

Since (2.12) satisfies all the assumptions of a CLR model, OLS will give the BLUE of $\beta$. Hence it follows that

$$\hat{\beta}_{GLS} = (X_*^T X_*)^{-1} X_*^T Y_*$$

$$= (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} Y$$

For prediction, formula (2.15) gives the "Best Linear Unbiased Predictor" (BLUP) in a first-order ALR model

$$Y_{t+1} = X_{t+1} \hat{\beta}_{GLS} + \rho e_t$$

where $e_t$ is the $t^{th}$ GLS residual.
3. MULTICOLLINEARITY

In applying multiple regression models, some degree of interdependence among explanatory variables can be expected. As this interdependence grows and the correlation matrix \( (X^T X) \) approaches singularity, multicollinearity constitutes a problem. Therefore it is preferable to think of multicollinearity in terms of its "severity" rather than its "existence" or "nonexistence".

3.1 Sources

In general, multicollinearity can be considered to be a symptom of poor experimental design. The sources of severe multicollinearity may be classified as follows [20:p.99-101].

(i) Not enough data or too many variables

In many cases large data sets only contain a few basic factors. As the number of variables extracted from the data increases, each variable tends to measure the different nuances of the same basic factors and each highly collinear variable only has little information content. In this case, deleting some variables or collecting more data can usually solve the problem.

(ii) Physical or structural singularity

Sometimes highly collinear variables, due to mathematical or physical constraints, are inadvertently included in the model.

(iii) Sampling singularity

Due to expense, accident or mistake, sampling was only conducted in a small region of the design space.
3.2 Effects

The major effects of serious multicollinearity are

(i) Estimation instability

As the correlation matrix \((X^T X)\) becomes ill-conditioned, the elements of the inverse matrix \((X^T X)^{-1}\) explode. (2.3) shows that the variances of the OLS estimates for \(\hat{\beta}\) are affected by the diagonal elements of the inverse matrix \((X^T X)^{-1}\). As a result of the instability of the inverse matrix \((X^T X)^{-1}\), the OLS estimates of the regression coefficients of the collinear variables might be numerically impossible to obtain. In any case they have large variances and are quite sensitive to small changes in the data set.

(ii) Structure misspecification

The increase in the size of the variable set of \(X\) decreases the information content of each explanatory variable, thereby decreasing the sample significance of each variable's contribution to the explained variance of \(Y\). Therefore, even though \(Y\) really depends on each member of a relatively large variable set of \(X\), erroneous deletion of variables may happen. As asserted by many authors [6:p.94][13:p.160][15], in the process of model-building, data limitation rather than the theoretical limitation is responsible for the tendency to underspecify the models.

(iii) Forecast inaccuracy

If an important variable is omitted because it is highly collinear, but later in the prediction period, this omitted variable changes its behavior and moves independently of other variables, then any forecasting under this oversimplified model will be very inaccurate.
(iv) **Numerical problems**

The correlation matrix \( (XX^T) \) is not invertible if the columns of \( X \) are linearly dependent. With the matrix \( (XX^T) \) being singular, the OLS estimates of \( \beta \), represented by (2.2), are completely indeterminate. In case of an almost singular set of variables, the numerical instability in calculating inverse matrix \( (XX^T)^{-1} \) still remains.

### 3.3 Detection

Tests for the presence and location of serious multicollinearity are briefly outlined and followed by comments.

(i) **Tests based on various correlation coefficients**

Here, harmful multicollinearity is generally recognized by rules of thumb. For instance, an admitted rule of thumb requires simple pair-wise correlation coefficients of explanatory variables to be less than 0.8. Certainly, those more extended and sophisticated rules of thumb with prudent use of various correlation coefficients will give more satisfactory results. The following rule of thumb is generally considered to be superior to other rules: a variable \( X_i \) is said to be highly multicollinear if its coefficient of multiple correlation, \( R_i^2 \), with the remaining \( (p-1) \) variables is greater than the coefficient of multiple correlation, \( R_y^2 \), with all the explanatory variables \([14:p.101]\). The variance of the estimate of \( \beta_i \) can be expressed as follows \([9]\)

\[
(3.1) \quad \text{Var}(\hat{\beta}_i) = \frac{1}{n-p-1} \frac{\sigma_y^2}{\sigma_{X_i}^2} \frac{1}{1 - R_i^2} \frac{1}{1 - R_y^2}
\]

where \( \sigma_y^2 \) is the variance of the dependent variable \( Y \) and \( \sigma_{X_i}^2 \) is the
variance of the explanatory variable $X_i$. From (3.1), it is obvious that multicollinearity constitutes a problem only when $R_i^2$ is relatively high to $R_y^2$. Unfortunately the geometric interpretation of this rule of thumb is apparent only when there are two explanatory variables [6:p.98].

(ii) **Three-stage hierarchy test**

This is proposed by Farrar and Glauber [6]. At the first stage, if the null hypothesis $H_0: |X^TX| = 1$ is rejected based on the Wilks-Bartlett's test, we may assert that multicollinearity is severe and move toward the second stage. The F statistic is then computed for each $R_i^2$

$$F_i = \frac{R_i^2/(p-1)}{(1-R_i^2)/(n-p)} \quad i = 1, \ldots, p$$

Statistical significant $F_i$ implies $X_i$ is collinear. At the third stage, inspection of the partial correlation coefficients between $X_i$ and the remaining $(p-1)$ variables and the associated t-ratios can show the pattern of interdependency among the explanatory variables. Farrar and Glauber claimed that detecting, localizing severe multicollinearity and learning the pattern of interdependence among explanatory variables can be respectively achieved at three different stages of their test.

(iii) **Haitovsky Chi Square test**

In 1969, Haitovsky [9] proposed a heuristic statistic for the hypothesis test of severe multicollinearity. This heuristic statistic is a function of the determinant of the correlation matrix $(X^TX)$, and approximately distributed as Chi Square. Applications
to Farrar and Glauber's data show that this test gives more satisfactory results than the Wilks-Bartlett's test that is adopted at the first stage of the Farrar and Glauber three-stage test. Therefore Haitovsky claimed the superiority of his test and suggested a replacement of Wilk-Bartlett's test by his test in the Farrar and Glauber three-stage test. However, any test based on the determinant of correlation matrix has some built-in deficiencies. As will be shown later, the mean square error properties depend only on the eigenvalues of the matrix $(X^TX)$. Only when the $(X^TX)$ has a broad eigenvalue spectrum, that is to say the ratio of the largest eigenvalue to the smallest one, $\lambda_1/\lambda_p$, is large, the performance of the OLS estimates may deteriorate. Since the determinant of the correlation matrix is equal to the product of all the eigenvalues, this test will treat the matrix having broad eigenvalue spectrum equivalently to those having relatively narrow eigenvalue spectra, so long as they have the same or nearly the same determinants. The relative magnitude of the eigenvalues is difficult if not impossible to infer from the results of any test that is based on the determinant of the correlation matrix. However, Haitovsky test gives a fairly good indication in the presence of severe multicollinearity in our simulation study.

(iv) Examining the spectrum of matrix $(X^TX)$

If the matrix $(X^TX)$ has a broad eigenvalue spectrum, that is, $\lambda_1/\lambda_p$ is large, then the mean square error of the OLS estimates of $\beta$ becomes very large. Since the trace of the correlation matrix $(X^TX)$ is equal to the number of explanatory variables $p$, an arbitrary
rule of thumb may consider $\frac{\lambda_1}{\lambda_p}$ is large if $\frac{\lambda_1}{\lambda_p} > p$. Besides, the minimax index \( \text{MMI} = \sum_{i=1}^{p} \frac{\lambda_i^{-2}}{\lambda_p^{-2}} \) is a useful indicator too. Small MMI, say, less than two implies the presence of multicollinearity [21:p.13-14].

Among all these tests and methods proposed, examining the eigenvalue spectrum of the matrix \((X^T X)\) provides not only a sound theoretical basis but also the lightest computation burden.
4. AUTOCORRELATION

One of the basic assumptions of the CLR model is that the error terms are independent of each other. However, when regression analysis is applied to time series data, the residuals are often found to be serially correlated. Like multicollinearity, autocorrelation is another widespread problem in applying regression models. For simplicity, first-order autocorrelation is assumed in our study.

4.1 Sources

The sources are mainly the following:

(i) Omission of variables

The time-ordered effects of the omitted variables will be included in the error terms. This prevents the errors from displaying random behavior. In this case, finding the missing variables and identifying the correct relationship can solve the problem.

(ii) Systematic measurement error in the dependent variable

Again, the error terms absorb the systematic measurement error in the dependent variable and then display non-random behavior.

(iii) Error structure is time dependent

The great impacts of some random events or shocks, such as war, strikes, flood, etc., are spread over several periods of time, causing the error terms to be serially correlated. This is so-called "true-autocorrelation".
4.2 Effects

When the OLS technique is still used for estimation, the major effects are:

(i) Unbiased but inefficient estimator of $\beta$

GLS provides the BLUE of $\beta$ when the dispersion matrix of $\epsilon$, $\sigma^2 \Omega$, is nondiagonal. That is to say on the average the sampling variances of GLS estimates of $\beta$ are less than that of OLS estimates of $\beta$, hence OLS is inefficient compared with GLS.

(ii) Underestimation of the variances of the estimates of $\beta$

As an illustration, consider the very simple model

$$y_t = \beta x_t + \epsilon_t$$

$$\epsilon_t = \rho \epsilon_{t-1} + u_t$$

where $u_t$ satisfies assumptions (2.10). It has been shown that the variance of OLS estimate of $\beta$ is [13:p.247]

$$\text{Var}(\hat{\beta}_{OLS}) = \frac{\sigma_u^2}{n} \left[ 1 + 2\rho \sum_{i=1}^{n-1} x_i^2 + \frac{\sum_{i=1}^{n-2} x_i^2 x_{i+1}}{2} + 2\rho \sum_{i=1}^{n-2} x_i^2 x_{i+2} + \ldots \right]$$

The OLS formula (2.3) ignores the term in parentheses in (4.1) and gives the variances of the estimates of $\beta$ as $\sigma^2 / \sum_{i=1}^{n} x_i^2$. If both $\epsilon$ and $x$ are positively autocorrelated the expression in parentheses
is almost certainly greater than unity, therefore the OLS formula will underestimate the true variance of $\hat{\beta}_{OLS}$.

(iii) Inefficient predictor of $Y$

When autocorrelation is present, error made at one point in time gives information about the error made at a subsequent point in time. The OLS predictor fails to take this information into account, hence it is not the BLUP of $Y$ [13:p.265-266].

4.3 Detection

The tests which are commonly used to recognize the existence of first-order autocorrelation are the following.

(i) Eye-ball tests

The plot of OLS residuals $e_t$ against time $t$ can be informative. Any nonrandom behavior of $e_t$ can be considered as an indication of autocorrelation. We may also plot the OLS residual $e_t$ against its lagged value $e_{t-1}$. If the observations are not evenly spread over the four quadrants, we may conclude the first-order autocorrelation is present. These eye-ball tests are quite effective, however they are imprecise and do not lend themselves to classical inferential methods.

(ii) von-Neumann ratio

In 1941, the ratio of the mean square successive difference to the variance was proposed by von-Neumann as a test statistic for the existence of first-order autocorrelation [22]. Though various applications have proven the usefulness of the von-Neumann ratio, we emphasize that this test is applicable only when $e$ values are independently distributed and the sample size is large. In practice,
the OLS residuals used to compute the von-Neumann ratio usually are not independently distributed even when the true error terms are.

(iii) Durbin-Watson test

This test, named after its originators Durbin and Watson, is widely used for small sample sizes [4][5]. There are some shortcomings of the Durbin-Watson test. First, there exist two regions of indeterminancy. Though an exact test was suggested by Henshaw in 1966, its heavy computational burden prevents the test from wide applications [10]. Secondly, the Durbin-Watson test is derived for non-stochastic explanatory variables only. It has been shown that if the lagged dependent variables are present either in single regression equation models or in systems of simultaneous regression equations, the Durbin-Watson test is biased towards the value for a random error, that is, \( d \) is biased towards 2, thereby giving very misleading information [17]. It is as necessary as important to test for serial correlation for models containing lagged dependent variables since autocorrelated models are usually repaired by inserting lagged Y values into the right-hand side of the regression equation. To this end, Durbin developed a test based on the h statistic in 1970[3].

"h" is defined as the following,

\[
h = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sqrt{\frac{n}{\sum_{t=1}^{n} e_t^2} \left(1-n \text{ var}(b')\right)}}
\]

where \( b' \) is the coefficient of \( Y_{t-1} \).
This test is computational cheap but only applicable for large sample sizes. The small sample properties of the "h" statistic are still unknown.
5. JOINT EFFECTS OF MULTICOLLINEARITY AND AUTOCORRELATION

In statistical analysis, a point estimate is usually of little use unless accompanied by an estimate of its accuracy. In this connection, Mean Square Error (MSE) is widely used as a measure of accuracy. Since it is true that accurate parameter estimates constitute an effective model. MSE can be used to determine the model's effectiveness when the underlying objective is simply to obtain good parameter estimates. In 1970, Hoerl and Kennard presented the MSE properties for the OLS and Ridge estimates of $\beta$ [11]. Thereafter, the results of various studies have confirmed that ridge regression will improve the MSE of estimation and predictions in the presence of severe multicollinearity.

In this section, we will present expressions for the MSE of $\hat{\beta}_{OLS}$ and $\hat{\beta}_{R}(k)$ when the error terms follow a first-order autocorrelated pattern (2.9). These expressions will enable us to examine the effect of these two conditions on the ridge and the OLS estimates. Our analysis can be reduced to that of Hoerl and Kennard by setting $\rho = 0$.

5.1 Mean Square Error of the OLS Estimates of $\hat{\beta}$

We begin with the analysis for the OLS estimates for a first-order ALR model. Let $L_1 = \text{Distance from } \hat{\beta}_{OLS} \text{ to } \beta$

$$L_1^2 = (\hat{\beta}_{OLS} - \beta)^T (\hat{\beta}_{OLS} - \beta).$$

We define the MSE of $\hat{\beta}_{OLS}$ to be $E(L_1^2)$.

**Proposition 5.1**

$$E(L_1^2) = \sum_{j=1}^{n} \sum_{k=1}^{n} D_{jk} \rho |j-k|$$

(5.1)
where \( D = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-2} \mathbf{X}^T \mathbf{X} \)

**Proof:** From (2.1) (2.2)

\[
(5.2) \quad \hat{\beta}_{\text{OLS}} - \beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} - \beta
\]

\[
= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X} \beta + \varepsilon) - \beta
\]

\[
= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon
\]

By definition and (5.2) it follows that

\[
E(L_1^2) = E[(\hat{\beta}_{\text{OLS}} - \beta)^T (\hat{\beta}_{\text{OLS}} - \beta)]
\]

\[
= E[\mathbf{X}^T (\mathbf{X}^T \mathbf{X})^{-2} \mathbf{X}^T \varepsilon]^2.
\]

Noting that \( E(\varepsilon) = 0 \) it follows from Theorem 4.6.1 Graybill [7:p.139] that

\[
(5.3) \quad E(L_1^2) = \sigma_u^2 \mathbf{X}^T (\mathbf{X}^T \mathbf{X})^{-2} \mathbf{X}^T \varepsilon
\]

From the definition of \( V \) and \( D \), (5.1) follows.

(5.1) does not give much insight into the effect of multicollinearity and autocorrelation on the MSE of \( \hat{\beta}_{\text{OLS}} \). By rotating axes (using principal components) the effect can be more clearly demonstrated.

**Proposition 5.2**

\[
(5.4) \quad E(L_1^2) = \sigma_u^2 \sum_{i=1}^{p} \frac{1}{\lambda_i^2} + 2\sigma_u^2 \sum_{j > \ell} \sum_{i=1}^{n} \left( \sum_{i=1}^{p} \frac{x_{ji}^* x_{\ell i}^*}{\lambda_i^2} \right) \rho_{j \ell}^{j-\ell}
\]

where \( x_{ji}^* \) is the \( j \)th observation on the \( i \)th principal component.

---

For a matrix \( \mathbf{A} \) we use the notation \( t_\mathbf{r}(\mathbf{A}) \) to denote the trace of \( \mathbf{A}. \)
Proof: From (2.4) and (2.5)

\[ (5.5) \quad \hat{\alpha}_{\text{OLS}} - \alpha = (X^T X)\hat{\alpha} - \alpha \]

\[ = (X^T X)\hat{\alpha} - \alpha \]

By definition and (5.5) it follows that

\[ E(L_1^2) = E[(\hat{\alpha}_{\text{OLS}} - \beta)^T P^T (\hat{\beta}_{\text{OLS}} - \beta)] \]

\[ = E[(\hat{\alpha}_{\text{OLS}} - \alpha)^T (\hat{\alpha}_{\text{OLS}} - \alpha)] \]

\[ = E[\hat{\beta}_{\text{OLS}}^T (X^T X)^{-2} \hat{\alpha}^T \hat{\epsilon}] \]

Hence by the same argument used in proving Proposition 5.1

\[ E(L_1^2) = \sigma_u^2 \frac{\hat{\alpha}}{X^T V} \]

\[ = \sigma_u^2 \left( \begin{array}{cccc}
\frac{p}{\lambda_1^2} & \frac{p}{\lambda_1^2} & \cdots & \frac{p}{\lambda_1^2} \\
\frac{p}{\lambda_1} & \frac{p}{\lambda_1} & \cdots & \frac{p}{\lambda_1} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{p}{\lambda_1} & \frac{p}{\lambda_1} & \cdots & \frac{p}{\lambda_1} \\
\frac{p}{\lambda_1^2} & \frac{p}{\lambda_1^2} & \cdots & \frac{p}{\lambda_1^2} \\
\frac{p}{\lambda_1^2} & \frac{p}{\lambda_1^2} & \cdots & \frac{p}{\lambda_1^2} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{p}{\lambda_1^2} & \frac{p}{\lambda_1^2} & \cdots & \frac{p}{\lambda_1^2} \\
\end{array} \right) \]

By definition of \( \hat{\alpha} \) and \( V \), (5.4) follows.

After orthogonal rotation, the effect of multicollinearity and autocorrelation becomes apparent from (5.4). First, if \( \rho_\epsilon \) is positive and most of the principal components are also positively autocorrelated, almost certainly the second term in (5.4) will be positive. That is to
say that the MSE of \( \hat{\beta}_{OLS} \) will be larger than when these effects are not present; moreover the difference will be in proportion to the magnitude of \( \rho_\varepsilon \). Secondly, we obtain a cross term of eigenvalues, \( \lambda_i \) and autocorrelation coefficient, \( \rho_\varepsilon \). If the matrix \((X^T X)\) is ill-conditioned, that is, \( \lambda_p \) is close to zero and there is a high degree of positive autocorrelation both in the \( p^{th} \) component and the error terms, then the second term in (5.4) dominates and the MSE of \( \hat{\beta}_{OLS} \) can be very large. It is then extremely dangerous to apply OLS to data with the above characteristics. However, the problem will not be that serious if \( \rho_\varepsilon \) is negative or the principal components, especially those weak components, are not autocorrelated. Finally, from (5.4) we are able to tell by how much the MSE of \( \hat{\beta}_{OLS} \) changes because of the existence of first-order autocorrelated errors in general regression models containing \( p \) explanatory variables. Note that when \( \rho_\varepsilon = 0 \), (5.4) reduces to

\[
E(L^2_1) = \sigma^2 \text{tr}(X^T X)
= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}
\]

5.2 Mean Square Error of the Ridge Estimates of \( \hat{\beta} \)

In parallel with 5.1, we define

\[
L_2(k) = \text{Distance from } \hat{\beta}_{R}(k) \text{ to } \hat{\beta}
\]

The MSE of \( \hat{\beta}_{R}(k) \) is given by

\[
E[L^2_2(k)] = E[\hat{\beta}_{R}(k) - \beta^T (\hat{\beta}_{R}(k) - \beta)].
\]

Proposition 5.3

\[
(5.6) \quad E[L^2_2(k)] = \gamma_1(k) + \gamma_2(k) + \gamma_3(k)
\]

where

\[
\gamma_1(k) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2}
\]

\[
\gamma_2(k) = k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2}
\]
\[-22-\]

\[
\gamma_3(k) = 2\sigma_u^2 \sum_{j} \sum_{i} \frac{n-1}{i} \frac{\sum_{j \geq i} x_j^* x_i^*}{\sum_{i=1} \left( \lambda_i + k \right)^2} \rho_{\varepsilon}^{j-l}
\]

**Proof:** From (2.7) and (2.8), the MSE of \( \hat{\beta}_R(k) \) can be written as

\[
E[L_2^2(k)] = E[(\hat{\beta}_R(k) - \beta)^T P T (\hat{\beta}_R(k) - \beta)]
\]

\[
= E[(\hat{\beta}_R - \beta)^T (\hat{\beta}_R - \beta)]
\]

\[
(5.7)
\]

Since the first term in (5.7) is a scalar, from (2.7) and Proposition 5.2, it follows

\[
E[(\hat{\beta}_R - \beta)^T (\hat{\beta}_R - \beta)] = \sigma_u^2 \sum_{j} \sum_{i} \frac{n-1}{i} \frac{\sum_{j \geq i} x_j^* x_i^*}{\sum_{i=1} \left( \lambda_i + k \right)^2} \rho_{\varepsilon}^{j-l}
\]

(5.8) \[
E[(\hat{\beta}_R - \beta)^T (\hat{\beta}_R - \beta)] = \sigma_u^2 \sum_{j} \sum_{i} \frac{n-1}{i} \frac{\sum_{j \geq i} x_j^* x_i^*}{\sum_{i=1} \left( \lambda_i + k \right)^2} \rho_{\varepsilon}^{j-l}
\]

Since the matrix \((Z-I)\) can be written as

\[
(5.9) \quad Z - I = Z(I-Z^{-1})
\]

\[
= Z(-kA^{-1})
\]

\[
= -k(A+kI)^{-1}
\]

From (5.9), the second term in (5.7) can be expressed as follows.
\[(Z \alpha - a)^T (Z \alpha - a) = a^T (Z - I) \alpha^2 = k^2 \frac{\alpha^2}{(\lambda_k + k)^2} = \gamma_2(k)\]

Completing the proof.

The MSE of \(\hat{\beta}_R(k)\) consists of three parts, \(\gamma_1(k)\), \(\gamma_2(k)\) and \(\gamma_3(k)\). \(\gamma_1(k)\) can be considered to be the total variance of the parameter estimates and is a monotonically decreasing function of \(k\), \(\gamma_2(k)\) is the square of the bias brought by the augmented matrix \(kI\) and is monotonically increasing function of \(k\) while \(\gamma_3(k)\) is related to the autocorrelation in the error terms. Hoerl and Kennard claim that in the presence of severe multicollinearity, it is possible to reduce MSE substantially by taking a little bias, that is, choosing \(k > 0\). This is because in the neighborhood of origin, \(\gamma_1(k)\) will drop sharply while \(\gamma_2(k)\) will only increase slightly as \(k\) increases [11:p.60-61]. After incorporating autocorrelation in the context of ridge regression analysis, their assertion will still be true only if certain conditions are satisfied. From (5.6) we see that the effects of multicollinearity and autocorrelation are the following.

(i) If \(\rho_{\epsilon}\) is positive and the principal components, especially the weak components, are also positively autocorrelated, then ridge method will be even more desirable than OLS. This is because substantial decrease in both \(\gamma_1(k)\) and \(\gamma_3(k)\) can be achieved by choosing \(k > 0\) while the increase in \(\gamma_2(k)\) is relatively small as moving to \(k > 0\).

(ii) If \(\rho_{\epsilon}\) is negative or almost all of the principal components are
not autocorrelated, then on the average $\gamma_3(k)$ is close to zero, hence the ridge and the OLS estimates will perform relatively the same as in the uncorrelated case.

(iii) Since ridge regression is similar to shrinking the model by dropping the least important component [21:p.24-28] (5.6) gives a theoretical justification to shrink the model if both the last component and the error terms are positively autocorrelated. From the point of view of estimation stability, the ridge method will be helpful when severe multicollinearity is accompanied by high degree of positive autocorrelation both in the weakest component and the error terms.

5.3 When will Ridge estimates be better than the OLS estimates?

Taking the derivatives of $\gamma_1(k)$ and $\gamma_2(k)$, Hoerl and Kennard found a condition on $k$ such that ridge regression gives better parameter estimates than OLS in terms of MSE. That is when $k$ is smaller than $\sigma_u^2 / \alpha_{\max}^2$, where $\alpha_{\max}$ is the largest regression coefficient in magnitude, the MSE of $\hat{\beta}_R(k)$ will be less than that of $\hat{\beta}_{OLS}$ [11]. When autocorrelation is present, the condition on $k$ such that ridge regression will perform better than OLS regression are described below. Consider the derivatives of $\gamma_1(k)$, $\gamma_2(k)$ and $\gamma_3(k)$.
$$\frac{d\gamma_1}{dk} = -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{u_i (\lambda_i+k)^3}$$

$$\frac{d\gamma_2}{dk} = 2k \sum_{i=1}^{p} \frac{\lambda_i}{u_i (\lambda_i+k)^3}$$

(5.10) $$\frac{d\gamma_3}{dk} = -2\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i (\lambda_i+k)^3}$$

$$\sum_{j=1}^{n-1} \sum_{k=1}^{n-j} X^* \ell_{2+j} X^* \ell_{2+j} X^* \ell_{2+j}$$

When \((X^T X)\) approaches singularity which implies that \(\lambda_p \to 0\), the values of the first two derivatives in the neighborhood of origin are given by

(5.11) $$\lim_{\lambda_p \to 0} \lim_{k \to 0^+} (d\gamma_1/dk) = -\infty$$

(5.12) $$\lim_{\lambda_p \to 0} \lim_{k \to 0^+} (d\gamma_2/dk) = 0$$

As \(k\) increases, a huge drop in \(\gamma_1\) with slight increase in \(\gamma_2\) may be expected. However (5.10) shows that the behavior of \(\gamma_3\) depends on the degree of autocorrelation both in the principal components and the error terms. Therefore \(\gamma_3\) may increase or decrease at various rates as \(k\) increases. The use of ridge regression is most favourable when there is a high degree of positive autocorrelation both in the components and the error terms. We now formalize these arguments and present a condition on \(k\) such that ridge regression will be better than OLS regression in MSE criterion.
Let \( F(k) = E(L_1^2) - E[L_2^2(k)] \)

\[
= \sigma_u^2 \sum_{i=1}^{p} \left[ \frac{1}{\lambda_i^2} - \frac{1}{(\lambda_i + k)^2} \right] (\lambda_i + k)^j \lambda_i \sum_{j=1}^{n} \sum_{i=1}^{n-j} X_{i1}^* X_{i+1}^* \rho^{j-i}_e - \frac{2}{\lambda_i} \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3}
\]

Then

\[
(5.13) \quad \frac{dF}{dk} = 2\sigma_u^2 \sum_{i=1}^{p} \frac{1}{(\lambda_i + k)^3} [\lambda_i + k] \lambda_i \sum_{j=1}^{n} \sum_{i=1}^{n-j} X_{i1}^* X_{i+1}^* \rho^{j-i}_e - 2k \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3}
\]

Assume that \( \gamma_3(k) \) is a non-increasing function of \( k \) in the neighborhood of origin. From (5.11) and (5.12) we may expect \( F(k) \) to increase as moving towards \( k > 0 \), i.e. \( \lim_{k \to 0^+} (dF/dk) > 0 \). In other words, there exists \( k > 0 \) such that the OLS estimates have higher MSE than the ridge estimates.

\[
\lim_{k \to 0^+} (dF/dk) > 0 \implies \text{Theorem 5.1}
\]

If

\[
(5.14) \quad \frac{\sigma_u^2}{\alpha_{\max}^2} + \frac{\sigma_u^2}{\alpha_{\max}^2} \left[ \frac{1}{\lambda_1} \sum_{j=1}^{n} \sum_{i=1}^{n-j} X_{i1}^* X_{i+1}^* \rho^{j-i}_e \right] > k
\]

then \( E(L_1^2) - E[L_2^2(k)] > 0 \).
Again (5.14) will reduce to Hoerl and Kennard's result if $\rho_c = 0$. When positive autocorrelation exists in the error terms and the principal components, the second term in (5.14) may well be positive, hence the range of $k$ for ridge estimates to be better than OLS estimates in MSE criterion will be larger than what Hoerl and Kennard asserted in uncorrelated case. (5.14) shows that the extension in the range of $k$ is positively related with the magnitude of $\rho_c$. However, (5.14) is just a necessary condition on $k$ for $E(L_1^2)$ to be greater than $E[L_2^2(k)]$ since $F(k)$ is increasing in $k$ over the range shown by (5.14). It is possible that for some values of $k$, $F(k)$ is decreasing in $k$ while the function value is still positive, that is $E(L_1^2)$ is still greater than $E[L_2^2(k)]$. Therefore, we may consider (5.14) as a stringent condition on $k$ for ridge estimates to be better than OLS estimates in MSE criterion.

If either $\rho_c$ is negative or the principal components, especially those weak ones, are not autocorrelated, the behavior of $\gamma_3(k)$ and thereby $F(k)$ as $k$ increases will be hard to predict. The effect of autocorrelation on the range of $k$ depends on the data set we gathered.

In practice, the true parameters are unknown, the range of $k$ shown by (5.14) can be approximated by conducting a principal component analysis and substituting the estimates for the parameters.

5.4 Use of the "Ridge Trace"

In ridge regression the augmented matrix $(kI)$ is used to cause the system to have the general characteristics of an orthogonal system.
Hoerl and Kennard claimed that at certain value of \( k \), the system will stabilize [11:p.65]. They proposed the usage of a "Ridge Trace" as a diagnostic tool to select a single value of \( k \) and a unique ridge estimate of \( \beta \) in practice. The "Ridge Trace" will portray the behavior of all the parameter estimates as \( k \) varies. Therefore instead of suppressing dimensions either by deleting collinear variables or dropping the principal components of small importance, the Ridge Trace will show how singularity is causing instability, over/under-estimations and incorrect signs. In connection with autocorrelation where ridge regression is even more desirable, certainly the "Ridge Trace" will be of great help in getting better point estimates and thereby better predictions. Even when \( \rho \varepsilon \) is negative or the principal components are not autocorrelated, the merits and usefulness of the "Ridge Trace" and the "Ridge Regression" are still preserved in dealing with the problem of multi-collinearity.
RIDGE REGRESSION: ESTIMATES, MEAN SQUARE ERROR AND PREDICTION

The MSE of OLS estimates of \( \hat{\beta} \) can be written as the difference in the length between two vectors, \( \hat{\beta}_{OLS} \) and \( \hat{\beta} \) [11:p.56]

\[
E(I_1^2) = E(\hat{\beta}_{OLS}^T \hat{\beta}_{OLS}) - \beta^T \beta
\]

(6.1) shows that in the presence of severe multicollinearity, the MSE can be improved by shortening the OLS estimates of \( \hat{\beta} \). In this section we will show that this reasoning appears to be compatible with the derivation of ridge estimator of \( \hat{\beta} \). Hence ridge regression can be expected to be better in terms of MSE.

6.1 Derivation of Ridge Estimator for a CLR Model

Let \( \tilde{\beta} \) be any estimate of \( \hat{\beta} \). Its residual sum of squares, \( \phi \), can be written as the value of minimum sum of squares, \( \phi_{min} \), plus the distance from \( \tilde{\beta} \) to \( \hat{\beta}_{OLS} \) weighted through \( (X^T X) \).

\[
\phi = (Y-X\hat{\beta})^T (Y-X\hat{\beta})
\]

(6.2) \[
= (Y-X\hat{\beta}_{OLS})^T (Y-X\hat{\beta}_{OLS}) + (\hat{\beta}_{OLS} - \hat{\beta}_{OLS})^T X^T X (\hat{\beta}_{OLS} - \hat{\beta}_{OLS})
\]

\[
= \phi_{min} + \phi(\tilde{\beta})
\]

For a specific value of \( \phi(\tilde{\beta}) \), \( \phi_{o} \), the ridge estimator is founded by choosing a \( \tilde{\beta} \) to

Minimize \( \tilde{\beta}^T \tilde{\beta} \)

(6.3) Subject to \( (\hat{\beta}_{OLS} - \hat{\beta}_{OLS})^T X^T X (\hat{\beta}_{OLS} - \hat{\beta}_{OLS}) = \phi_{o} \)

This problem can be solved by use of Lagrange multiplier techniques.
where \((1/k)\) is the multiplier corresponding to the constraint (6.3).

The problem is to minimize

\[
F = B^T B + (1/k)[(B - \hat{B}_{OLS})^T X^T X (B - \hat{B}_{OLS}) - \phi_0]
\]

A necessary condition for \(B\) to minimize (6.4) is that

\[
\frac{\partial F}{\partial B} = 2B + \frac{1}{k} [2(X^T X)B - 2(X^T X)\hat{B}_{OLS}] = 0
\]

Hence

\[
[I + \frac{1}{k}(X^T X)]B = \frac{1}{k}(X^T X)\hat{B}_{OLS}
\]

and

\[
B^* = \hat{B}_R(k) = (X^T X + kI)^{-1} X^T Y
\]

where \(k\) is chosen to satisfy constraint (6.3). In practice, we usually work the other way round since it is easier to choose a \(k > 0\) and then compute the additional residual sum of squares, \(\phi_o\).

It is clear that for a fixed increment \(\phi_o\), there is a continuum of values of \(B_0\) that will satisfy the relationship \(\phi = \phi_{\text{min}} + \phi_o\), and the ridge estimate so derived is the one with the minimum length.

Therefore, we may well expect the ridge estimates to yield less MSE in the presence of multicollinearity since they are originally derived by minimizing the length of the regression vector. It is true to a certain extent that minimizing the length of the regression vector is equivalent to reducing the MSE of parameter estimates. In addition, (6.2) shows that it is possible to move further away from \(\hat{B}_{OLS}\) without an appreciable increase in the residual sum of squares as \((X^T X)\) approaching singularity. That is to say, ridge regression may
achieve large reduction in MSE at virtually no cost in terms of the residual sum of squares if the conditioning of $(X^TX)$ is poor enough. In 1971, Newhouse and Oman [18] used MSE as evaluation criterion in their Monte Carlo studies of ridge regression. Since then it has become the standard way to evaluate proposals for ridge estimators. From the above derivation of the ridge estimator, obviously we realize that ridge estimates are designed to be better in MSE criterion.

Now we would like to study the implications of the constraint in deriving the ridge estimator. Since orthogonalization can ease interpretation, we represent (6.2) in the rotated axes.

Let $A = PB$

Then

$$\phi = (Y - X^T\hat{\beta}_{OLS})^T (Y - X^T\hat{\beta}_{OLS}) + \hat{\beta}^T X^* X^* (\hat{\beta} - \hat{\beta}_{OLS})$$

$$= \phi_{\min} + \sum_{i=1}^{p} (A_i - \hat{\beta}_{OLS_i})^2 \lambda_i$$

Where $A_i$ is the estimate of regression coefficient for the $i$th component and $\hat{\beta}_{OLS_i}$ is the OLS estimate of regression coefficient for the $i$th component.

The problem is

Minimize $A^T A$

(6.5) Subject to $(A - \hat{\beta}_{OLS})^T A (A - \hat{\beta}_{OLS}) = \phi_o$

or equivalently

(6.6) Subject to $\sum_{i=1}^{p} (A_i - \hat{\beta}_{OLS_i})^2 \lambda_i = \phi_o$.

(6.5) Shows that the vector $[A - \hat{\beta}_{OLS}]$ is normed through $A$ to have the
length equal to $\phi_0$. Since the eigenvalue $\lambda_1$ can be considered as an indicator of the information content and explanatory power of the $i$th principal component, we may well conclude that the derivation of the ridge estimator has already taken the relative information content and explaining power of the explanatory variables into account. (6.6) shows that the constraint has incorporated the concept of square-error-loss function as well. It increases the length of $A$ the most when the parameter estimates of the important components deviate from OLS estimates since it is found by taking the square of the deviations multiplied by their corresponding eigenvalues. This implies that it is best to shrink the estimate of $\beta$ for those components that have small eigenvalues, i.e. the ones most subject to instability.

6.2 Derivation of Ridge Estimator for an ALR Model

In the presence of autocorrelated error terms, the OLS estimator of $\beta$ will no longer have the minimum-variance property; the GLS type estimator will be the BLUE of $\beta$. Our derivation of a new ridge estimator adjusted for autocorrelation, $\hat{\beta}_{CR}(k)$, will parallel with the derivation of $\hat{\beta}_{R}(k)$ in the previous section. Again let $B$ be any estimate of $\beta$. Its residual sum of squares, $\phi$, can be written as the value of the minimum sum of squares, $\phi_{min}$, plus the distance from $B$ to $\hat{\beta}_{GLS}$ weighted through $(X^T X)^{-1}$. (See (2.10) for notation). We have

$$\phi = (Y^* - X^* B)^T (Y^* - X^* B)$$

$$= (Y^* - X^* \hat{\beta}_{GLS})^T (Y^* - X^* \hat{\beta}_{GLS}) + (B - \hat{\beta}_{GLS})^T X^T X (B - \hat{\beta}_{GLS})$$

$$= \phi_{min} + \phi(B).$$
For a specific value $\phi(B) = \phi_0$, $\hat{\beta}_{GR}$ is derived to minimize $B^T B$

(6.7) subject to $\left( B - \hat{\beta}_{GLS} \right)^T X^* X^* \left( B - \hat{\beta}_{GLS} \right) = \phi_0$.

The Lagrangian is given by

$$F = B^T B + \frac{1}{k} \left[ \left( B - \hat{\beta}_{GLS} \right)^T X^* X^* \left( B - \hat{\beta}_{GLS} \right) - \phi_0 \right].$$

A necessary condition for a minimum is that

$$\frac{\partial F}{\partial B} = 2B + \frac{1}{k} \left[ 2X^* X^* B - 2 \left( X^* X^* \right) \hat{\beta}_{GLS} \right] = 0.$$

This reduces to

(6.8) $B^* = \hat{\beta}_{GR} (k) = \left( X^* X^* + kI \right)^{-1} X^* Y^*$

$$= \left( X^T \Omega^{-1} X + kI \right)^{-1} X^T \Omega^{-1} Y.$$

Where $k$ is chosen to satisfy (6.7). The characterization of the ridge trace of $\hat{\beta}_{GR} (k)$ will be essentially the same as that of $\hat{\beta}_R (k)$. That is, for a specific increment $\phi_0$, the $\hat{\beta}_{GR}$ so derived is the regression vector with the minimum length among a continuum of values of $B_0$ that will satisfy the relationship $\phi = \phi_{\text{min}} + \phi_0$. However multicollinearity may no longer be a substantial problem after transforming $X$ into $X^*$ in some rare cases. For instance, this may happen in time series studies where multicollinearity is a result of the explanatory variables increasing together over time. It is then possible that the transformed variables are not close to being collinear with each other. If that is the case, the reduction in MSE can not be obtained with only a slight increase in the residual sum of squares. This is because of the low MSE of $\hat{\beta}_{GLS}$ already achieved and the non-singularity of the $(X^* X^*)^{-1}$. 
In most cases, if not all, the matrix $(X^T T^{-1} X)$ is very likely to have a broad eigenvalue spectrum if $(X^T X)$ does. Then the previous discussion on the motivation of minimizing the length of the regression vector, the interpretation and implications of the constraint in the derivation of ridge estimator of $\beta$ for a CLR model will be applicable to that in the derivation of $\hat{\beta}_{GR}$.

6.3 Mean Square Error of the "Generalized Estimators"

The MSE of $\hat{\beta}_{GLS}$ and $\hat{\beta}_{GR}(k)$ are readily established. Since

$Y_k = X_k \beta + \xi_k$ satisfies all the assumptions of a CLR model, (5.3) with $\rho_\epsilon = 0$ gives the MSE of $\hat{\beta}_{GLS}$ as follows.

Let $L_3 = \text{Distance from } \hat{\beta}_{GLS} \text{ to } \beta$

$E[L_3^2] = \sigma_u^2 \text{ tr } (X_k^T X_k)^{-1}$

(6.9) \[ \sigma_u^2 \text{ tr } (X_k^T \Omega_k^{-1} X_k)^{-1} \]

Setting $\rho_\epsilon = 0$ in (5.8), (6.10) gives the MSE of $\hat{\beta}_{GR}(k)$.

Let $L_4(k) = \text{Distance from } \hat{\beta}_{GR}(k) \text{ to } \beta$

(6.10) \[ E[L_4^2(k)] = \sigma_u^2 \text{ tr } [(X_k^T \Omega_k^{-1} X_k + kI)^{-2} (X_k^T \Omega_k^{-1} X_k)] + k^2 \beta^T (X_k^T \Omega_k^{-1} X_k + kI)^{-2} \beta \]

The effect of autocorrelation is difficult to infer from (6.9) and (6.10), since $\Omega_k$ is not a diagonal matrix, however, normally we may expect $E[L_3^2]$ and $E[L_4^2(k)]$ to be less than $E[L_1^2]$ and $E[L_2^2(k)]$ respectively.
6.4 Estimation

Theoretically, the GLS gives the BLUE of $\beta$ for an ALR model. But usually in practice, neither the order of autocorrelation structure nor the value of the parameter $\rho$ is known. Hence the GLS or GR estimates can not be computed directly. Many two-stage methods have been proposed to approximate the GLS estimates and have proven to be quite effective. These include the Cochrane-Ocutt iterative process [1] and Durbin's two-step method [2].

In the joint presence of multicollinearity and autocorrelation, it is actually quite straightforward to combine ridge regression with one of the two-stage regression methods in the hope of achieving better estimates of $\beta$. We illustrate how ridge regression can be incorporated in Durbin's two-step method for a simple model with only two collinear explanatory variables:

\begin{equation}
Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \epsilon_t \quad t = 1, 2, \ldots, n
\end{equation}

\begin{equation}
\epsilon_t = \rho \epsilon_{t-1} + u_t \quad |\rho| < 1
\end{equation}

and for all $t$

\begin{align*}
E(u_t) &= 0 \\
E(u_t, u_{t+s}) &= \sigma_u^2 \quad \text{for } s = 0 \\
&= 0 \quad \text{for } s \neq 0
\end{align*}

The transformed relation is given by

\begin{equation}
Y_t - \rho \epsilon Y_{t-1} = \beta_0 (1-\rho) + \beta_1 (X_{t1} - \rho \epsilon X_{t-1,1}) + \beta_2 (X_{t2} - \rho \epsilon X_{t-1,2}) + u_t
\end{equation}
Combining (6.11) and (6.12) gives

\[(6.13) Y_t = \beta_0 (1-\rho_e) + \beta_1 X_{t1} - \beta_1 \rho_e X_{t-1,1} + \beta_2 X_{t2} - \beta_2 \rho_e X_{t-1,2} + \rho Y_{t-1} + u_t \]

The first step is to estimate the parameters of (6.13) using OLS. Then use the estimated coefficient of \(Y_{t-1}\) to compute the transformed variables \(\hat{Y}_{t-1} - \rho \hat{Y}_{t-1}\), \(\hat{X}_{t1} - \rho \hat{X}_{t-1,1}\) and \(\hat{X}_{t2} - \rho \hat{X}_{t-1,2}\). At the second step, ridge regression is highly recommended to be used in place of OLS and applied to relationship (6.12) containing those transformed variables. The coefficient estimate of \(\hat{X}_{t1} - \rho \hat{X}_{t-1,1}\) is our approximation of \(\hat{\beta}_{GR1}\) and the intercept term divided by \((1-\rho_e)\) is our approximation of \(\hat{\beta}_{GR0}\).

It might seem reasonable to apply ridge regression at the first step of Durbin's method since \(X_{t1}\) and \(X_{t2}\) are collinear. As stated earlier in Section 3, high pair-wise correlation coefficient of explanatory variables does not necessarily result in estimation instability. Besides, the lagged values of \(X_{t1}\) and \(X_{t2}\) are inserted into the explanatory variable set. If \(X_{t1}\) and \(X_{t2}\) are not autocorrelated, the conditioning of the enlarged \((X'X)\) may be satisfactory. Moreover \(u_t\) in (6.12) has a scalar dispersion matrix, therefore OLS gives consistent estimates of regression coefficients. Also among these estimates, only the coefficient estimate of \(Y_{t-1}\) will be used to compute the transformed variables. Hence OLS technique is recommended to be used at the first step even when \(X_{t1}\) and \(X_{t2}\) are collinear.

This combination of ridge regression and Durbin's two-step method can easily be extended to a p-variable model with higher order of autocorrelation.
6.5 Prediction

Consider a first-order ALR model, (2.15) gives the minimum variance predictor (BLUP). In practice, both $\hat{\beta}_{\text{GLS}}$ and $\rho_\varepsilon$ are replaced by their estimated values.

If ridge regression is used in conjunction with some other methods to cope with the joint problem of multicollinearity and autocorrelation, the prediction is given by

\begin{equation}
\hat{Y}_{t+1} = X_{t-1} B_{GR}(k) + \hat{\rho}_\varepsilon e_t
\end{equation}

where $X_{t+1}$ is a 1 x p vector of the $(t+1)^{st}$ observation on the explanatory variables, $B_{GR}(k)$ is a p x 1 vector of approximated regression coefficients, $\hat{\rho}_\varepsilon$ is an estimate of autocorrelation coefficient and $e_t$ is the ridge residual at time $t$. 
7. THE MONTE CARLO STUDY

Consider a first-order ALR model with two explanatory variables, the error terms in the transformed relation, as shown by (6.12), have a scalar dispersion. The residual sum of squares from (6.12) is given by

\[ \sum_{t=1}^{n} u_t^2 = \sum_{t=1}^{n} [(Y_t - \hat{\beta}_0 Y_{t-1} - \hat{\beta}_1 (1 - \rho_e) X_{t-1} - \hat{\beta}_2 (X_{t-1} - \rho_e X_{t-1}) - \hat{\rho}_e X_{t-1})]^2 \]

If \( X_{o1}, X_{o2} \), and \( Y_0 \) are given, the summation can run from 1 to \( n \), otherwise it can only run from 2 to \( n \). The direct minimization of (7.1) with respect to \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \) and \( \rho_e \) leads to non-linear equations, therefore, the analytic expressions for \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \) and \( \rho_e \) cannot be obtained.

As mentioned before, many two-stage methods have been proposed to approximate these parameters. Usually in practice, not only the parameters and the true error terms but also the order of autocorrelation structure is unknown. As indicated previously, joint presence of autocorrelation and multicollinearity will further complicate the situation. Under this circumstance, the relative effectiveness of those two-stage methods can best be studied by the Monte Carlo experiments [19].

7.1 Design of the Experiments

The main purpose of the experiment is to give an empirical support to the inference drawn from our analytic studies. The sampling experiments are conducted in the following manner. Basically, the sampling experiments comprise nine different experiments with different
degree of multicollinearity and autocorrelation. They are summarized in Table 1.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>$r_{12}$</th>
<th>$\rho_\varepsilon$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>2</td>
<td>.05</td>
<td>.50</td>
</tr>
<tr>
<td>3</td>
<td>.05</td>
<td>.90</td>
</tr>
<tr>
<td>4</td>
<td>.50</td>
<td>.05</td>
</tr>
<tr>
<td>5</td>
<td>.50</td>
<td>.50</td>
</tr>
<tr>
<td>6</td>
<td>.50</td>
<td>.90</td>
</tr>
<tr>
<td>7</td>
<td>.95</td>
<td>.05</td>
</tr>
<tr>
<td>8</td>
<td>.95</td>
<td>.50</td>
</tr>
<tr>
<td>9</td>
<td>.95</td>
<td>.90</td>
</tr>
</tbody>
</table>

In our experiments, $\gamma_{12}$ and $\rho_\varepsilon$ are used to indicate the severity of multicollinearity and autocorrelation respectively. Usually in practice, multicollinearity constitutes a problem only when $\gamma_{12}$ is as high as 0.8 or 0.9. In addition, the error terms are normally considered to be independent, moderately and highly autocorrelated when $\rho_\varepsilon = .05$, .50 and .90 respectively. As shown by Table 1, the experiments are set up to have different characteristics. Through this design, we can study the effects of autocorrelation on estimation and prediction for a given degree of multicollinearity. Moreover, we can observe how these effects of autocorrelation change as the degree of multicollinearity varies.

The data is generated as follows: first, values are assigned to $\beta_0$, $\beta_1$, $\beta_2$ and the probability characteristics of error terms $u_t$ in
Three series of $\varepsilon_t$ in (2.9) are subsequently generated, given the values of $\varepsilon_0$ and different value of $\rho_\varepsilon$. The probability characteristics of the joint distribution $X_{t1}$ and $X_{t2}$ are chosen to generate the series of $X_{t1}$ and $X_{t2}$ that are suitable for the first three experiments. By varying the correlation coefficient of $X_{t1}$ and $X_{t2}$, another two series of $X_{t1}$ and $X_{t2}$ are generated for the remaining six experiments. We have also assured that there is no significant first-order autocorrelation in $X_{t1}$ and $X_{t2}$ so that the error structure is first-order. Solving for $Y_t$ based on the data, nine different sets can be generated for the experiments. For each experiment, ten samples of forty observations are generated. In each sample, thirty observations on the $Y_t$'s from $t = 1$ to $t = 30$ are employed to estimate the equation by appropriate methods. Observations 31 to 40 are used to study the prediction properties of estimators. The BLUP is used in the presence of significant autocorrelation in the error terms.

Special care has to be exercised in controlling the serial correlation properties of the error terms. In this connection, OLS regression has to be run on

$$e_{jt} = \rho_\varepsilon e_{jt-1} + u_{jt} \quad j = 1, 2, \ldots, 10$$

$$t = 1, 2, \ldots, 40$$

(7.2) to determine whether the estimated regression coefficient $\hat{\rho}_\varepsilon$ is consistent with the $\rho_\varepsilon$ which is used to generate them. However, as is well-known, the OLS estimates of parameters for small samples may be badly biased if some of the regressors are lagged dependent variables [23]. This is because the error terms, $u_{jt}$ will no longer be independent of the regressors, $\varepsilon_{jt-1}, \varepsilon_{jt}, \ldots, \varepsilon_{j40}$. In (7.2),
\( E(u_{jt}, \epsilon_{jt+s}) \neq 0 \) for \( s \neq 0 \) and all \( t \), hence the OLS estimate for the coefficient of \( \epsilon_{jt-1} \) is biased. The usual \( t \) test on the estimate of regression coefficient may be quite misleading, therefore we can only ascertain that the desired serial correlation properties are obtained by assuring that \( u_{jt} \) are randomly distributed. For each sample, we first test whether the series of \( u_t \) is consistent with the probability characteristics chosen to generate them, then we use run test to determine whether \( u_t \) is randomly distributed. Only those series of \( u_t \) passed all the tests are adopted in our simulation study. We are now ready to estimate the regression equation.

First, for each experiment, the OLS principle is applied to estimate the parameters. The Durbin-Watson statistic is used as a filter to test the existence of autocorrelation. Whenever the Durbin-Watson statistic computed from the fitted model is less than the corresponding upper critical value \( d_u (\alpha=0.05) \), autocorrelation is assumed to be present in the error terms, then Durbin's two-step method is used in conjunction with Ridge regression for estimation as described in Section 6.4; otherwise, autocorrelation is assumed to be absent, and only OLS and Ridge regressions techniques are employed for estimation purposes. In addition, whenever the existence of autocorrelation is recognized, \( \hat{\beta}_{GLS} \) and \( \hat{\beta}_{GR}(k) \) are computed for comparison purposes. Since the true value of the autocorrelation coefficient is known for each experiment, calculations of \( \hat{\beta}_{GLS} \) and \( \hat{\beta}_{GR}(k) \) will simply be the straightforward multiplication of matrices as shown by (2.13) and (6.8). The methods adopted for estimation in each experiment are recorded in Table 2.
Table 2

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OLS, RR</td>
</tr>
<tr>
<td>2</td>
<td>Durb., Durb.+RR, GLS, GR</td>
</tr>
<tr>
<td>3</td>
<td>Durb., Durb.+RR, GLS, GR</td>
</tr>
<tr>
<td>4</td>
<td>OLS, RR</td>
</tr>
<tr>
<td>5</td>
<td>Durb., Durb.+RR, GLS, GR</td>
</tr>
<tr>
<td>6</td>
<td>Durb., Durb.+RR, GLS, GR</td>
</tr>
<tr>
<td>7</td>
<td>OLS, RR</td>
</tr>
<tr>
<td>8</td>
<td>Durb., Durb.+RR, GLS, GR</td>
</tr>
<tr>
<td>9</td>
<td>Durb., Durb.+RR, GLS, GR</td>
</tr>
</tbody>
</table>

OLS: Ordinary Least-squares Regression
Durb.: Durbin's Two-step Method
Durb.+RR: Durbin's Two-step in conjunction with Ridge Regression
GLS: Centralized Least-squares Regression
GR: Ridge Regression adjusted for Autocorrelation

As is expected, no correction for autocorrelation is necessary for experiments 1, 4 and 7. Whenever the ridge method is applied seven or eight values of k have been used in our study. In order to minimize the effects of subjectivity resulting from selecting the value of k in ridge regressions, we compute the mean $\hat{\beta}_R$ of the samples for every specific value of k in each experiment. Then the value of k is selected based on a "Mean Ridge Trace". That is to say, a unique value of k selected will generally be the best for all then samples. Obviously, the value of k so-selected may well not to be the best for every individual sample. Therefore, the minimum of the MSE of ridge estimates of $\hat{\beta}$ achieved for each experiment is slightly upward biased.
Certainly this way of selecting the value of k cannot be used in practice.

7.2 Sampling Results

For each method for each experiment, the MSE of the estimates of $\hat{\beta}$, the adjusted $R^2$, the residual sum of squares, the MSE forecast and the Durbin-Watson statistic are averaged over ten samples. In addition, the mean estimate of $\rho_\epsilon$ and the mean Haitovsky heuristic statistic are also computed. $u_t$ is assumed to follow a normal distribution with mean zero and variance equal to 6 i.e. $u_t \sim N(0, 6)$. The true mean of $X_{t1}$ is 10 and that of $X_{t2}$ is 8. The respective variance of $X_{t1}$ and $X_{t2}$ are 18 and 15. $\epsilon_0$ is chosen to be 3 for each sample. The true value of $\beta_0$ is 5, $\beta_1$ is 1.1 and $\beta_2$ is 1.

7.2a Results assuming $\rho_\epsilon$ is Known

First we assume $\rho_\epsilon$ is known. The results here will indicate whether the methods described in Section 6.4 can show promise in the best of situations. Table 3 contains the average MSE of $\hat{\beta}_{GLS}$ and $\hat{\beta}_{GR}$ for experiments 2, 3, 5, 6, 8 and 9.
Table 3

<table>
<thead>
<tr>
<th>Experiment</th>
<th>2 ((\gamma_{12}=.05))</th>
<th>3 ((\gamma_{12}=.05))</th>
<th>5 ((\gamma_{12}=.50))</th>
<th>6 ((\gamma_{12}=.50))</th>
<th>8 ((\gamma_{12}=.95))</th>
<th>9 ((\gamma_{12}=.95))</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>((\rho_{\varepsilon}=.50))</td>
<td>((\rho_{\varepsilon}=.90))</td>
<td>((\rho_{\varepsilon}=.50))</td>
<td>((\rho_{\varepsilon}=.90))</td>
<td>((\rho_{\varepsilon}=.50))</td>
<td>((\rho_{\varepsilon}=.90))</td>
</tr>
<tr>
<td>0.0*</td>
<td>0.4824</td>
<td>3.2913</td>
<td>0.0834</td>
<td>2.3940</td>
<td>0.1011</td>
<td>2.1681</td>
</tr>
<tr>
<td>0.025</td>
<td>0.0591</td>
<td>1.8084</td>
<td>0.0018</td>
<td>1.4991</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>0.05</td>
<td>0.0363</td>
<td>0.8073</td>
<td>0.1215</td>
<td>0.8337</td>
<td>0.0561</td>
<td>0.9405</td>
</tr>
<tr>
<td>0.075</td>
<td>0.3603</td>
<td>0.2274</td>
<td>0.3033</td>
<td>0.4263</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>0.1</td>
<td>0.9849</td>
<td>0.0156</td>
<td>0.8871</td>
<td>0.1092</td>
<td>0.4929</td>
<td>0.2901</td>
</tr>
<tr>
<td>0.2</td>
<td>5.7786</td>
<td>2.0100</td>
<td>4.0851</td>
<td>0.6153</td>
<td>2.4426</td>
<td>0.4242</td>
</tr>
<tr>
<td>0.3</td>
<td>2.9771</td>
<td>7.0896</td>
<td>9.3743</td>
<td>4.0521</td>
<td>5.4924</td>
<td>1.0113</td>
</tr>
<tr>
<td>0.5</td>
<td>30.1494</td>
<td>21.7887</td>
<td>21.1581</td>
<td>11.4165</td>
<td>13.7628</td>
<td>4.8285</td>
</tr>
<tr>
<td>0.7</td>
<td>47.5367</td>
<td>36.6241</td>
<td>35.2063</td>
<td>22.2489</td>
<td>23.7003</td>
<td>10.9524</td>
</tr>
<tr>
<td>1.0</td>
<td>75.6732</td>
<td>63.4542</td>
<td>56.5869</td>
<td>39.9762</td>
<td>38.5830</td>
<td>-----</td>
</tr>
</tbody>
</table>

* GLS regressions can be considered as a special case of Ridge regressions, adjusted for autocorrelation, with k = 0.

In Section 5, it has been shown that the MSE of \( \hat{\beta}_{GLS} \) will increase rapidly if significantly positive autocorrelation exists both in the disturbances and in the principal components. Correction for autocorrelation will then be necessary in estimating the regression equation. Though the GLS regression yields the BLUE of \( \hat{\beta} \), the behavior of the MSE of \( \hat{\beta}_{GLS} \) is very difficult to infer from (6.10). From the MSE of experiments 3, 6 and 9 when k = 0, we observe that the MSE of \( \hat{\beta}_{GLS} \) decreases as the degree of multicollinearity increases for sufficiently high degree of autocorrelation. On the other hand, for a given degree of multicollinearity, the MSE of \( \hat{\beta}_{GLS} \) will increase as the degree of autocorrelation increases. But the magnitude of the increase in the MSE
of $\hat{\beta}_{\text{GLS}}$ decreases as the relation among explanatory variables increases. For instance, the difference in MSE of $\hat{\beta}_{\text{GLS}}$ between experiments 8 and 9 is less than that between experiments 5 and 6. Moreover, Table 3 shows that there exists at least one value of $k$ for each experiment such that the MSE of $\hat{\beta}_{\text{GR}}$ is less than that of $\hat{\beta}_{\text{GLS}}$. Note that when $\rho_{\varepsilon} = 9$, $k = .1$ obtains the minimum MSE of the estimates of $\hat{\beta}$ in experiment 9. This also implies that the transformed matrix $(X^T X)^{-1}$ is still ill-conditioned. In Section 5.3 we have shown that the range of $k$ such that the MSE of $\hat{\beta}_{\text{GR}}$ is less than that of $\hat{\beta}_{\text{OLS}}$ will be larger if multicollinearity is accompanied by high degree of autocorrelation. Now with parameter estimates fully adjusted for autocorrelation (since $\rho_{\varepsilon}$ is known) experiment 9 still has the largest admissible range of $k$. That is, the range of $k$ such that the MSE of $\hat{\beta}_{\text{GR}}$ is less than that of $\hat{\beta}_{\text{GLS}}$ is still larger if multicollinearity is accompanied by high degree of autocorrelation and autocorrelation has been fully adjusted. We also observe that as the degree of autocorrelation increases, a larger reduction in the MSE of the estimates of $\hat{\beta}$ can be obtained by replacing $\hat{\beta}_{\text{GLS}}$ with $\hat{\beta}_{\text{GR}}$. For instance, the difference in the MSE of $\hat{\beta}_{\text{GLS}}$ and $\hat{\beta}_{\text{GR}} (.05)$ in experiment 8 is less than that in experiment 9. However, the behavior of the MSE of $\hat{\beta}_{\text{GR}}$ is very difficult, if not impossible, to predict. (6.10) shows that the MSE of $\hat{\beta}_{\text{GR}}$ is comprised of two terms. How each term behaves will depend not only on the data matrix $X$ and the degree of autocorrelation but also on the way the matrix $X$ is linked with the matrix $\Omega^{-1}$.

7.2b Results assuming $\rho_{\varepsilon}$ is unknown

In practice $\rho_{\varepsilon}$ is unknown. We assume that the true autocorrelation coefficient is unknown and we try to fit the equation using heuristic
techniques akin to the Durbin's two-step method in which has been shown by Griliches and Rao [8] to perform well when there is autocorrelation. We apply these techniques as described in Section 7.1.

Tables 4, 5 and 6 report the mean adjusted $R^2$ and the mean Durbin-Watson statistic for each experiment. $(du(\alpha = 5\%) = 1.57$ for experiments 1, 4 and 7; $du(\alpha = 5\%) = 1.56$ for the remaining six experiments).

Table 4 ($\gamma_{12} = .05$)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>1 $\rho_\varepsilon = .05$</th>
<th>2 $\rho_\varepsilon = .50$</th>
<th>3 $\rho_\varepsilon = .90$</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>$2^*$ $R_a$</td>
<td>$d^{**}$</td>
<td>$2^*$ $R_a$</td>
</tr>
<tr>
<td>0.0</td>
<td>0.8640  2.0791</td>
<td></td>
<td>0.8979  1.8766</td>
</tr>
<tr>
<td>0.025</td>
<td>0.8640  2.0903</td>
<td></td>
<td>0.8884  1.8713</td>
</tr>
<tr>
<td>0.05</td>
<td>0.8620  2.1001</td>
<td></td>
<td>0.8868  1.8728</td>
</tr>
<tr>
<td>0.075</td>
<td>0.8579  2.1083</td>
<td></td>
<td>0.8844  1.8805</td>
</tr>
<tr>
<td>0.1</td>
<td>0.8567  2.1154</td>
<td></td>
<td>0.8813  1.8916</td>
</tr>
<tr>
<td>0.2</td>
<td>0.8401  2.1353</td>
<td></td>
<td>0.8619  1.9619</td>
</tr>
<tr>
<td>0.3</td>
<td>0.8167  2.1463</td>
<td></td>
<td>0.8399  2.0383</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7651  2.1536</td>
<td></td>
<td>0.7865  2.1563</td>
</tr>
<tr>
<td>1.0</td>
<td>0.6394  2.1527</td>
<td></td>
<td>0.6571  2.2960</td>
</tr>
</tbody>
</table>

* $R_a^2$: the mean adjusted $R^2$

** $d$: the mean Durbin-Watson statistic
Table 5 \((y_{12} = .50)\)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>4 (\rho_{\varepsilon} = .05)</th>
<th>5 (\rho_{\varepsilon} = .50)</th>
<th>6 (\rho_{\varepsilon} = .90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>(R^2_a) (d)</td>
<td>(R^2_a) (d)</td>
<td>(R^2_a) (d)</td>
</tr>
<tr>
<td>0.0</td>
<td>.8973 2.0984</td>
<td>.9178 1.8958</td>
<td>.9475 2.0754</td>
</tr>
<tr>
<td>.025</td>
<td>.8970 2.1040</td>
<td>.9175 1.9054</td>
<td>.9472 2.0865</td>
</tr>
<tr>
<td>.05</td>
<td>.8962 2.1095</td>
<td>.9168 1.9194</td>
<td>.9464 2.1059</td>
</tr>
<tr>
<td>.075</td>
<td>.8950 2.1147</td>
<td>.9155 1.9371</td>
<td>.9451 2.1317</td>
</tr>
<tr>
<td>.1</td>
<td>.8933 2.1198</td>
<td>.9138 1.9576</td>
<td>.9434 2.1622</td>
</tr>
<tr>
<td>.2</td>
<td>.8832 2.1381</td>
<td>.9032 2.0540</td>
<td>.9336 2.3024</td>
</tr>
<tr>
<td>.3</td>
<td>.8702 2.1592</td>
<td>.8895 2.1504</td>
<td>.9184 2.4532</td>
</tr>
<tr>
<td>.5</td>
<td>.8351 2.1721</td>
<td>.8546 2.2988</td>
<td>.8834 2.6086</td>
</tr>
<tr>
<td>.7</td>
<td>.7970 2.1826</td>
<td>.8187 2.4203</td>
<td>.8440 2.7084</td>
</tr>
<tr>
<td>1.0</td>
<td>.7412 2.1900</td>
<td>.7572 2.4732</td>
<td>.7846 2.7887</td>
</tr>
</tbody>
</table>

Table 6 \((y_{12} = .95)\)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>7 (\rho_{\varepsilon} = .05)</th>
<th>8 (\rho_{\varepsilon} = .50)</th>
<th>9 (\rho_{\varepsilon} = .90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>(R^2_a) (d)</td>
<td>(R^2_a) (d)</td>
<td>(R^2_a) (d)</td>
</tr>
<tr>
<td>0.0</td>
<td>.9208 2.0511</td>
<td>.9391 1.8785</td>
<td>.9575 1.8707</td>
</tr>
<tr>
<td>.05</td>
<td>.9201 2.0873</td>
<td>.9380 1.9058</td>
<td>.9565 1.8883</td>
</tr>
<tr>
<td>.1</td>
<td>.9181 2.1128</td>
<td>.9359 1.9437</td>
<td>.9544 1.9335</td>
</tr>
<tr>
<td>.2</td>
<td>.9116 2.1473</td>
<td>.9293 2.0280</td>
<td>.9677 2.0562</td>
</tr>
<tr>
<td>.3</td>
<td>.9024 2.1659</td>
<td>.9199 2.1084</td>
<td>.9382 2.1660</td>
</tr>
<tr>
<td>.5</td>
<td>.8786 2.1768</td>
<td>.8957 2.2312</td>
<td>.9136 2.8381</td>
</tr>
<tr>
<td>.7</td>
<td>.8505 2.1725</td>
<td>.8673 2.3082</td>
<td>.8847 2.4417</td>
</tr>
<tr>
<td>1.0</td>
<td>.8056 2.1594</td>
<td>.8217 2.3739</td>
<td>.8399 2.5259</td>
</tr>
</tbody>
</table>
From Tables 4-6, we observe that the adjusted $R^2$ increases as the degree of autocorrelation increases for a given value of $k$ and a given degree of multicollinearity. This is intuitively plausible since autocorrelation can account for part of the variation in the errors, thereby decreasing the residual sum of squares and increasing the adjusted $R^2$. For each experiment, the adjusted $R^2$ decreases as $k$ increases. The reason is obvious from the derivation of ridge estimators. Besides, the best $R^2_a$ achieved for each experiment is pretty high, that is, the estimated model can explain most of the variation in $Y_t$. This also implies that the estimation methods adopted in our experiments are fairly efficient and powerful. The mean Durbin-Watson statistic computed for each method for each experiment is high enough to ascertain that the fitted model has successfully removed the problem of autocorrelation. Since the model is reasonably well fitted, simulation comparisons of the experimental results should be meaningful as well as informative.

The average MSE of the estimates of $\hat{\theta}$ is computed for each method for each experiment and recorded in Table 7.
<table>
<thead>
<tr>
<th>Experiment</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>k = 0.0</td>
<td>0.1101</td>
<td>0.4824</td>
<td>0.9594</td>
<td>0.0030</td>
<td>0.0342</td>
<td>0.3996</td>
<td>0.0180</td>
<td>0.0720</td>
<td>0.6951</td>
</tr>
<tr>
<td>k = 0.025</td>
<td>0.0192</td>
<td>0.0570</td>
<td>0.2691</td>
<td>0.1104</td>
<td>0.0210</td>
<td>0.0945</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>k = 0.05</td>
<td>0.2865</td>
<td>0.0390</td>
<td>0.0087</td>
<td>0.4158</td>
<td>0.1965</td>
<td>0.0024</td>
<td>0.1833</td>
<td>0.0719</td>
<td>0.0939</td>
</tr>
<tr>
<td>k = 0.075</td>
<td>0.8820</td>
<td>0.3744</td>
<td>0.1200</td>
<td>0.8973</td>
<td>0.5778</td>
<td>0.1026</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>k = 0.1</td>
<td>1.7430</td>
<td>1.0167</td>
<td>0.5559</td>
<td>1.5366</td>
<td>1.1097</td>
<td>0.3765</td>
<td>0.8307</td>
<td>0.5643</td>
<td>0.0041</td>
</tr>
<tr>
<td>k = 0.2</td>
<td>7.3539</td>
<td>5.9115</td>
<td>4.7694</td>
<td>5.3823</td>
<td>4.5690</td>
<td>2.7540</td>
<td>3.2001</td>
<td>1.2549</td>
<td>0.5822</td>
</tr>
<tr>
<td>k = 0.3</td>
<td>15.1383</td>
<td>13.2456</td>
<td>11.5962</td>
<td>10.8432</td>
<td>9.5949</td>
<td>6.8219</td>
<td>6.6531</td>
<td>5.7624</td>
<td>3.3012</td>
</tr>
<tr>
<td>k = 0.5</td>
<td>33.5067</td>
<td>31.1559</td>
<td>28.8369</td>
<td>23.9694</td>
<td>22.3449</td>
<td>18.6255</td>
<td>15.6804</td>
<td>14.2377</td>
<td>10.0251</td>
</tr>
<tr>
<td>k = 0.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>38.6334</td>
<td>36.9396</td>
<td>32.0214</td>
<td>26.3049</td>
<td>24.3789</td>
<td>18.5115</td>
</tr>
<tr>
<td>k = 1.0</td>
<td>79.3134</td>
<td>76.8708</td>
<td>73.8630</td>
<td>60.7620</td>
<td>50.3176</td>
<td>52.8276</td>
<td>43.3464</td>
<td>40.2351</td>
<td>32.3991</td>
</tr>
</tbody>
</table>
As the known autocorrelation case, when \( k = 0 \) the MSE of estimates of \( \hat{\beta} \) increases as the degree of autocorrelation increases, given the degree of multicollinearity. However, being different from the known autocorrelation case, the MSE of estimates of \( \hat{\beta} \) first decreases then increases as the degree of multicollinearity increases for \( k = 0 \) and a given degree of autocorrelation. Table 7 shows that except for experiments 4 and 7, better estimates of \( \beta \) in MSE criterion can be obtained if Durbin's two-step method is combined with ridge regression for estimation. Besides, amazingly we have found that we are able to obtain better estimates of \( \beta \) in terms of MSE if the true autocorrelation coefficient \( \rho_\epsilon \) is unknown. For clarity, we shall compare only the minimum of the average MSE of the estimates of \( \hat{\beta} \) achieved for each experiment in the \( \rho_\epsilon \) unknown case with that in the \( \rho_\epsilon \) known case. Table 8 reports the minima of the average MSE of the estimates of \( \hat{\beta} \) achieved for each experiment in both the \( \rho_\epsilon \) known and unknown cases. In addition, the estimation method and the characteristics of each experiment are also tabulated.
A couple of interesting observations can be made from Table 8. First, if the degree of multicollinearity is held constant, the minimum of the average MSE of parameter estimates will first increase then decrease as the degree of autocorrelation increases. On the other hand, given the degree of autocorrelation, the minimum of the average MSE of the parameter estimates will first decrease then increase as the degree of multicollinearity increases. These are intuitively plausible since sufficient high degree of autocorrelation should lead to more stable parameter estimates while sufficient high degree of multicollinearity usually results in very unstable parameter estimates. Secondly, we observe that the value of ridge parameter $k$, used to achieve the minimum $\text{MSE}$ of the estimates of $\hat{\beta}$, increases with the degree of multicollinearity and autocorrelation. This is consistent with our analytic findings.
shown in Section 5.3. Moreover, we have found that knowing \( \rho_e \) does not give better estimates of \( \beta \) for sufficient high degree of autocorrelation. This may result from sample sizes being small.

Table 9 contains the mean estimates of \( \rho_e \) obtained in the first step of Durbin's method and the mean Haitovsky heuristic statistic for each experiment.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\rho}_e )</td>
<td>---</td>
<td>.3581</td>
<td>.7182</td>
<td>---</td>
<td>.3586</td>
<td>.7231</td>
<td>---</td>
<td>.3849</td>
<td>.7498</td>
</tr>
<tr>
<td>Bias in ( \hat{\rho}_e )</td>
<td>---</td>
<td>.1419</td>
<td>.1818</td>
<td>---</td>
<td>.1414</td>
<td>.1769</td>
<td>---</td>
<td>.1151</td>
<td>.1502</td>
</tr>
<tr>
<td>( H_{x^2} )</td>
<td>125.7</td>
<td>123.1</td>
<td>111.7</td>
<td>39.1</td>
<td>38.7</td>
<td>37.4</td>
<td>2.78</td>
<td>2.53</td>
<td>2.40</td>
</tr>
<tr>
<td>df = 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In all cases, Durbin's two-step method tends to underestimate the true autocorrelation coefficient. This results from the presence of the lagged \( Y \) values among the explanatory variables [16]. If the degree of multicollinearity is held constant, the bias of estimate of \( \rho_e \) increases as the degree of autocorrelation increases; while given the degree of autocorrelation, the bias decreases as the degree of multicollinearity increases. In our simulation study, Haitovsky heuristic statistic can recognize the existence of severe multicollinearity in experiments 7, 8 and 9. However, it does not give any warning when there exists a fairly high degree of multicollinearity, i.e. based on the Haitovsky test, multicollinearity is insignificant in experiments 4, 5 and 6. Since the Haitovsky test is based on the determinant of correlation matrix, it has some built-in deficiencies (see Section 3.3 for details). Our experiments have
disclosed these deficiencies to a certain extent, hence we suggest that special care has to be exercised in applying this test.

7.2.c. Forecasting

Tables 10, 11 and 12 report the average residual sums of squares and the mean square error of prediction from the given values for the forecast period of each experiment, under the assumption that \( \rho_c \) is unknown.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\sigma}_u^2 )</td>
<td>MSE}_{F/C}^{**}</td>
<td>( \hat{\sigma}_u^2 )</td>
</tr>
<tr>
<td>k</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>5.9700</td>
<td>8.1132</td>
<td>5.7055</td>
</tr>
<tr>
<td>.025</td>
<td>5.9924</td>
<td>8.0343</td>
<td>5.7332</td>
</tr>
<tr>
<td>.05</td>
<td>6.0554</td>
<td>7.9961</td>
<td>5.8113</td>
</tr>
<tr>
<td>.075</td>
<td>6.1536</td>
<td>7.9986</td>
<td>5.9328</td>
</tr>
<tr>
<td>.2</td>
<td>7.0250</td>
<td>8.4293</td>
<td>7.0074</td>
</tr>
<tr>
<td>.3</td>
<td>8.0022</td>
<td>9.0877</td>
<td>8.2087</td>
</tr>
<tr>
<td>.5</td>
<td>10.245</td>
<td>10.755</td>
<td>10.957</td>
</tr>
<tr>
<td>1.0</td>
<td>15.733</td>
<td>15.075</td>
<td>17.656</td>
</tr>
</tbody>
</table>

* \( \hat{\sigma}_u^2 \): the average of the residual sums of squares over ten samples.

** MSE}\_{F/C}^{**}: the average MSE of predictions from the given values for the forecast period.
Table 11 \((r_{12} = .50, \sigma_u^2 = 6)\)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>4</th>
<th>(\hat{\sigma}_u^2)</th>
<th>(\text{MSE}_{F/C})</th>
<th>5</th>
<th>(\hat{\sigma}_u^2)</th>
<th>(\text{MSE}_{F/C})</th>
<th>6</th>
<th>(\hat{\sigma}_u^2)</th>
<th>(\text{MSE}_{F/C})</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td></td>
<td>6.0169</td>
<td>8.2093</td>
<td>5.8625</td>
<td>9.3838</td>
<td>5.7757</td>
<td>10.733</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.05</td>
<td></td>
<td>6.0797</td>
<td>8.1690</td>
<td>5.9409</td>
<td>9.3874</td>
<td>5.8838</td>
<td>10.672</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.3</td>
<td></td>
<td>7.2030</td>
<td>9.2476</td>
<td>7.9231</td>
<td>10.783</td>
<td>8.6992</td>
<td>12.482</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td></td>
<td>9.7190</td>
<td>10.851</td>
<td>10.470</td>
<td>12.735</td>
<td>12.128</td>
<td>15.327</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.7</td>
<td></td>
<td>11.933</td>
<td>12.726</td>
<td>13.070</td>
<td>14.851</td>
<td>15.759</td>
<td>18.018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>15.429</td>
<td>15.620</td>
<td>17.566</td>
<td>18.301</td>
<td>21.929</td>
<td>22.680</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12 \((r_{12} = .95, \sigma_u^2 = 6)\)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>7</th>
<th>(\hat{\sigma}_u^2)</th>
<th>(\text{MSE}_{F/C})</th>
<th>8</th>
<th>(\hat{\sigma}_u^2)</th>
<th>(\text{MSE}_{F/C})</th>
<th>9</th>
<th>(\hat{\sigma}_u^2)</th>
<th>(\text{MSE}_{F/C})</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td></td>
<td>6.0443</td>
<td>8.3699</td>
<td>5.6725</td>
<td>9.8589</td>
<td>5.4033</td>
<td>11.335</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.05</td>
<td></td>
<td>6.1186</td>
<td>8.1165</td>
<td>5.7759</td>
<td>9.4141</td>
<td>5.5390</td>
<td>10.758</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.3</td>
<td></td>
<td>7.5180</td>
<td>8.9408</td>
<td>7.5161</td>
<td>10.116</td>
<td>7.9187</td>
<td>12.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td></td>
<td>9.4101</td>
<td>10.445</td>
<td>9.8466</td>
<td>11.683</td>
<td>11.120</td>
<td>14.301</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>15.198</td>
<td>15.343</td>
<td>16.973</td>
<td>16.918</td>
<td>20.713</td>
<td>21.591</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Though the BLUP is adopted for forecast purposes, the MSE of prediction will still increase as the degree of autocorrelation increases. However, the main point is that the presence of multi-collinearity will adversely affect the predictive performance if the disturbances are highly serially correlated. The commonly held belief that the predictive power of the model is not affected by existence of multi-collinearity is only true if the problem of autocorrelation is not serious. In the 9th experiment, the model fitted by Durbin's method gives satisfactory results on various diagnostic tests, still the prediction of the BLUP leaves much to be desired. Fortunately, with Durbin's method combined with Ridge regression, we are able to determine the model which will yield less MSE of prediction and perform well on various criteria and tests in the joint presence of multi-collinearity and autocorrelation. We also observed that the value of $k$ that yields the best estimate of the residual sum of squares will not usually give the minimum MSE of predictions. However, the value of $k$ giving the best estimates of $\beta$ in MSE criterion tends to yield less MSE of prediction for each experiment. Hence, we may conclude the validity of the MSE criterion in the evaluation of parameter estimates.

To avoid confusion, we have not reported the MSE of prediction based on $\hat{\beta}_{GLS}$. However, we found that the MSE of prediction based on $\hat{\beta}_{GLS}$ and true $\rho_\varepsilon$ is less than that based on the parameter estimates obtained by Durbin's two-step method in conjunction with ridge regression. Though Durbin's two-step method combined with ridge regression gives better estimates of $\beta$, but in general it underestimates $\rho_\varepsilon$. Therefore, the BLUP based on $\hat{\beta}_{GLS}$ and true $\rho_\varepsilon$
gives the minimal MSE of prediction in each of the experiments 2, 3, 5, 6, 8, and 9.

8. CONCLUSIONS

It has been shown that in the presence of multicollinearity with sufficient high degrees of autocorrelation. The OLS estimates of regression coefficients can be highly inaccurate. Improving the estimation procedure is obviously necessary. Combining GLS and Ridge regression, we derived a new estimator.

\[ \hat{\beta}_{GR}^{(k)} = (X^T\Omega^{-1}X + kI)^{-1}X^T\Omega^{-1}y \]

where \( 0 < k < 1 \) and \( \Omega \) is defined in (2'). \( \hat{\beta}_{GR}^{(k)} \), though biased, is expected to perform well in the joint presence of multicollinearity and autocorrelation. However, since \( \Omega \) is unknown, parameter estimates based on the biased estimator \( \hat{\beta}_{GR}^{(k)} \) cannot be obtained in practice. Therefore, we combined Durbin's two-step method with ordinary Ridge regression to approximate those parameters. The effectiveness of our approximation can then best be examined by the Monte Carlo simulation.

Our study has confirmed that, for a given degree of multicollinearity, the MSE of the GLS estimates of \( \hat{\beta} \) is directly proportioned to the degree of autocorrelation. This agrees with conventional wisdom. Unexpectedly, we found that the MSE of the GLS estimates of \( \hat{\beta} \) is inversely proportional to the degree of multicollinearity for a sufficiently high degree of autocorrelation. This implies that in the application of the GLS technique, the symptom of the existence of multicollinearity may be disguised. However, since in practice neither
the true error terms nor the autocorrelation coefficient is known, no
GLS estimates can possibly be obtained. We were pleased to find that
in the joint presence of multicollinearity and autocorrelation;
whatever the degree is, Durbin's two-step method in conjunction with
Ridge regression (p known) yields even better estimates of \( \hat{\beta} \) than
the GLS technique (p known) does in MSE criterion. Though the value
of k giving better estimates of \( \hat{\beta} \) tends to yield less MSE of
prediction, still the GLS gives the minimal MSE of prediction in all
the cases. Besides, our experimental results have shown that the
Durbin-Watson test for detecting the existence of first-order auto-
correlation remains powerful in the presence of multicollinearity while
the Haitovsky heuristic statistic gives relatively limited information
about the existence of multicollinearity either with or without the
presence of autocorrelated error terms.

Our results also suggest that it might be possible to find an
"optimal estimation package" that deals with the joint problem of
multicollinearity and autocorrelation. Empirical research has hitherto
been confined to the search for optimal estimation techniques in
dealing with multicollinearity and autocorrelated errors as separate
and independent phenomena. The ordinary ridge regression, i.e.,
adding a constant k on the diagonal of correlation matrix \( (X^TX) \) and
Durbin's two-step method have been shown to be very powerful
techniques in handling multicollinearity and autocorrelation problems.
Even though satisfactory estimation and prediction are obtained by
the combination of Durbin's method and ordinary ridge regression,
there may still exist some other even more efficient approaches to
the joint problem of multicollinearity and autocorrelation. For
instance, the combination of the Cochrane-Orcutt procedure with Generalized Ridge regression is a more flexible estimation technique and thereby should lead to better estimation and prediction. Allowing for higher order and mixed order autocorrelation will be a good direction to pursue as well.
BIBLIOGRAPHY


