5-1978

Multicollinearity and the Estimation of Regression Coefficients

John Charles Teed
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/etd

Part of the Applied Statistics Commons

Recommended Citation
https://digitalcommons.usu.edu/etd/6947

This Thesis is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Theses and Dissertations by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.
MULTICOLLINEARITY AND THE ESTIMATION OF REgression COEFFICIENTs

by

John Charles Teed

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Applied Statistics

APPROVED:

UTAH STATE UNIVERSITY
Logan, Utah
1978
ACKNOWLEDGEMENTS

I would like to give a very special thanks to my Major Professor, Dr. Rex L. Hurst, for his guidance and help throughout this assignment.

A wish to thank Dr. David L. Turner for his unending help and encouragement and Dr. Bartell Jensen for his time and help on my committee.

Finally, to my wife, Kayleen, for her patience and support in fulfilling this assignment, I extend a husband’s gratitude.

John Charles Teed
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>iv</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>v</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Multiple Regression Model</td>
<td>1</td>
</tr>
<tr>
<td>Definition of Multicollinearity</td>
<td>3</td>
</tr>
<tr>
<td>REVIEW OF LITERATURE</td>
<td>5</td>
</tr>
<tr>
<td>Causes of Multicollinearity</td>
<td>5</td>
</tr>
<tr>
<td>Consequences of Multicollinearity</td>
<td>7</td>
</tr>
<tr>
<td>Detection of Multicollinearity</td>
<td>8</td>
</tr>
<tr>
<td>Solutions to Multicollinearity</td>
<td>13</td>
</tr>
<tr>
<td>METHODOLOGY</td>
<td>17</td>
</tr>
<tr>
<td>Scope of Study</td>
<td>17</td>
</tr>
<tr>
<td>Variables Measured</td>
<td>20</td>
</tr>
<tr>
<td>Analysis of Data</td>
<td>21</td>
</tr>
<tr>
<td>RESULTS</td>
<td>22</td>
</tr>
<tr>
<td>Analysis of Variance Tests</td>
<td>22</td>
</tr>
<tr>
<td>Interpretation of results</td>
<td>23</td>
</tr>
<tr>
<td>Effect of correlation structure</td>
<td>36</td>
</tr>
<tr>
<td>Effect of sample size</td>
<td>38</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Analysis of variance for model (3.1)</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>Analysis of variance for model (3.2)</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>Treatment averages for $(S_{\beta_1})$</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>Effect of standard error of y and sample size on the precision of estimation $(S_{\beta_1})$ for model (3.1)</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>Effect of standard error of y and the standard deviation of $X$ on the precision of estimation $(S_{\beta_1})$ for model (3.1)</td>
<td>32</td>
</tr>
<tr>
<td>6</td>
<td>Effect of standard deviation of $X$ and sample size on the precision of estimation $(S_{\beta_1})$ for model (3.1)</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>Effect of standard deviation of $X$ and correlation structure on precision of estimation $(S_{\beta_1})$ for model (3.1)</td>
<td>33</td>
</tr>
<tr>
<td>8</td>
<td>Effect of standard error of y and sample size on precision of estimation $(S_{\beta_1})$ for model (3.2)</td>
<td>34</td>
</tr>
<tr>
<td>9</td>
<td>Effect of standard error of y and standard deviation of $X$ on the precision of estimation $(S_{\beta_1})$ for model (3.2)</td>
<td>34</td>
</tr>
<tr>
<td>10</td>
<td>Effect of standard deviation of $X$ and sample size on the precision of estimation for model (3.2)</td>
<td>35</td>
</tr>
<tr>
<td>11</td>
<td>Coefficient of variation on $\beta_1$ with bad multicollinearity.</td>
<td>39</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>Effect of standard error of y and sample size on precision of estimation ( S_{\beta_1} )</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>Effect of standard error of y and standard deviation of X on the precision of estimation ( S_{\beta_1} )</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>Effect of standard deviation of X and sample size on the precision of estimation ( S_{\beta_1} )</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>Effect of standard deviation of X and correlation structure of precision of estimation ( S_{\beta_1} )</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>Effect of sample size on coefficient of variation of ( \beta_1 )</td>
<td>40</td>
</tr>
</tbody>
</table>
ABSTRACT
Multicollinearity and the Estimation of Regression Coefficients

by

John C. Teed, Master of Science
Utah State University, 1977

Major Professor: Dr. Rex L. Hurst
Department: Applied Statistics

The precision of the estimates of the regression coefficients in a regression analysis is affected by multicollinearity. The effect of certain factors on multicollinearity and the estimates was studied. The response variables were the standard error of the regression coefficients and a standardized statistic that measures the deviation of the regression coefficient from the population parameter.

The estimates are not influenced by any one factor in particular, but rather some combination of factors. The larger the sample size, the better the precision of the estimates no matter how "bad" the other factors may be.

The standard error of the regression coefficients proved to be the best indication of estimation problems.
CHAPTER I
INTRODUCTION

Multiple regression is one of the most frequently used tools of modern statistics. One of the serious concerns in using multiple regression is how well the estimated regression coefficients agree with the population parameters. It is well known that when the various independent variables are highly correlated, problems in estimation arise. The problem of highly correlated independent variables falls under the general term of multicollinearity. Collinearity means that strong linear relationships exist among the independent variables. There is some degree of collinearity in nearly all variable sets. The only truly non-collinear situation results when the independent variables have zero correlation, which is extremely unusual.

It has long been suspected that many factors, in addition to the correlation structure among the independent variables, affect the estimation problems. This paper intends to explore how several factors, including the correlation structure, affect the estimation problems in multiple regression.

Multiple Regression Model

The usual multiple regression model may be written in matrix notation as:

\[ Y = X\beta + \varepsilon \]  \hspace{1cm} (1.1)
where $Y$ is a $n \times 1$ vector of observable random variables known as the response or dependent variables. $X$ is an $n \times k$ matrix of constants known as the explanatory or independent variables.

$$X = \begin{bmatrix}
1 & X_{12} & \cdots & X_{1k} \\
1 & X_{22} & \cdots & X_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
1 & X_{n2} & \cdots & X_{nk}
\end{bmatrix}$$

where $n$ denotes the $n$th observation on the $k$th independent variable. $\hat{\beta}$ is an $n \times 1$ vector of unknown constants that correspond to the regression coefficients. $\epsilon$ is an $n \times 1$ vector of unobservable random variables denoting the error variables.

$$\epsilon \sim \text{NID}(0, \sigma^2).$$

Least squares regression analysis leads to estimates of the regression coefficient.

$$\hat{\beta} = (X'X)^{-1}X'Y$$

with variance-covariance matrix

$$\text{VAR}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

where $\sigma^2$ is the population variance of $\epsilon$.

The purpose of a multiple regression analysis is to estimate parameters of a dependency, not an interdependency relationship. To use the multiple regression techniques, estimation of the unknown parameters is required. Many of the applications of the multiple regression model rely on the estimation of the regression coefficients, $\hat{\beta}$. Many decisions
made by a researcher are based on the individual estimates of the coefficients or on some statistical test associated with the estimates. There are times when relationships between independent and dependent variables are inferred from the coefficient estimates. When multicollinearity is present in the data, it is viewed as an interdependency condition and the decisions made may be misleading or totally wrong.

One of the basic assumptions of the general multiple regression model is that the data matrix $X$, which is of order $n \times k$, has rank of $k$. This is to say that there is no linear dependence among the independent variables. The reason for this assumption is that the least squares estimators, $\hat{\beta}$, require $(X'X)^{-1}$ to exist. This is impossible if the rank of $X$ is less than $k$. If the rank of $X$ is less than $k$, then there is a linear dependency among some or all of the independent variables. These affected variables are perfectly collinear, that is the correlation between these variables is equal to one, and perfect multicollinearity is said to exist.

If no interpretation is being made from the estimates of the coefficients, a generalized inverse could be used when the model is not of full rank. The problem with this is that the solution is not unique and therefore interpretation is a problem.

Definition of Multicollinearity

In this paper, the definition of multicollinearity that will be used is that given by Webster, Gunst and Mason (1973). It is given in terms of the linear dependent column vectors of the data matrix $X$. Column vectors $X_1, X_2, ..., X_j$ are perfectly collinear if there exists a non-zero constant $a_j$ such that
When this holds for any subset of column vectors of $X$, the independent variables associated with these column vectors are perfectly correlated and extreme multicollinearity exists. However, in most sets of data, (1.2) does not hold exactly, but it is approximately true. Farrar and Glauber (1967) point out that if (1.2) is approximately true, that is if some of the independent variables are highly correlated, a less extreme but still very serious case of multicollinearity exists. This can be seen if (1.2) is approximately true then the determinant of $(X'X)$, denoted by $|X'X|$, is close to zero and $|(X'X)^{-1}|$ approaches infinity. As $|(X'X)^{-1}|$ approaches infinity, $\text{var} \left( \hat{\beta} \right) = \sigma^2 |(X'X)^{-1}|$ will approach infinity and therefore $\text{var} \left( \hat{\beta} \right) = s^2 |(X'X)^{-1}|$ approaches infinity. This is an undesirable condition since the large variances on the regression coefficients produced by the multicollinearity indicate the low information content of the observed data and therefore the low quality of the resulting parameter estimates. This points out the inability to distinguish the independent contribution to the explained variance by an independent variable that has little or no independent variation.

It can be seen therefore that multicollinearity should be considered in terms of severity rather than existence or non-existence. In this paper, "Multicollinearity" will be used when (1.2) is approximately true and "Perfect Multicollinearity" used when (1.2) is exactly true.
Causes of Multicollinearity

Mason, Gunst and Webster (1975) suggest that there are three main causes of multicollinearity. The first of these is the over-defined model. The cause of this is having more independent variables in the model than observations or by having duplicate independent variables in the model. This situation can arise in biological studies where multiple measurements are made on a small number of subjects. This problem is solved by redefining the model with some of the independent variables deleted. To decide which of the variables to delete, either preliminary investigations must be performed using subsets of the independent variables, or some sort of analysis is performed to decide which variables are similar. Ways of analyzing the data to make this decision are principle components analysis or cluster analysis. The principle components analysis techniques are preferred over other ways of choosing the variables to delete by Massy (1965), Webster, Gunst, and Mason (1973, 1974) and Lott (1973) because it is felt that they are more objective in their approach to the solution.

The second cause of multicollinearity arises when a subspace of the space of independent variables is sampled. This sampled subspace is approximately a hyperplane that is defined by one or more of the
relationships of the form defined in (1.2). Usually the researcher is unaware of this problem because the relationships are not exact and difficult to visualize. An example of this cause is presented by Mason, Gunst, and Webster (1973). They suggest this might occur in a situation when prediction of profits is obtained from variables such as income and labor costs. The analysis might show a linear relationship between income and labor costs; that is, higher labor costs results in higher prices which result in increased income. This type of multicollinearity is not inherent since the data could be collected during a period when labor costs are increasing yet prices are held constant or are decreasing. Therefore, the multicollinearity is due to sampling technique.

The third cause of multicollinearity is that of physical constraints on the model or in the population. This cause is similar to those of the sampling techniques, but will exist regardless of the sampling scheme used. An example given by Mason et. al. is one of a chemical analysis where the sum of certain constituents in a solution must always be constant although the values of the individual constituents may vary. It is often hard to determine whether there are unknown physical constraints on the model or the sampling techniques used to collect the data are the cause of multicollinearity. It is important to try and decide which is the cause so that the correct steps are taken to analyze and interpret the data.

Koursoyiannis (1973) has indicated two types of data which cause multicollinearity; the first of which is time series. There is a tendency for variables to move together over time with growth and trend factors
in time series to be the most serious cause of multicollinearity. The second is the use of lagged values of some of the independent variables as separate independent factors.

**Consequences of Multicollinearity**

When multicollinearity is present in a set of data, there are consequences which are not desirable. One consequence is that the precision of the estimation diminishes so that it becomes difficult, if not impossible, to tell which of the $X$ variables has an influence on what. This loss of precision causes three things to happen: specific estimates may have infinitely large standard errors; these standard errors may be highly correlated with each other; and the sampling variances of the coefficients may be very large. In model building, a result of multicollinearity could be that a variable may be dropped which is meaningful but not significantly different. The problem here is that multicollinearity has influenced the sample data so as to not allow the data to pick up the true situation. If relevant variables have been incorrectly omitted, there is no indication as to what bias has been introduced into the remaining coefficient estimates. Another result of multicollinearity is that as different sets of data are used, the estimates of the coefficients may vary widely and as more observations are introduced into the model, there may be shifts in the coefficients.

One final consideration of the problems that multicollinearity can cause is that with very severe but not perfect multicollinearity the true error degrees of freedom may be reduced. There may be no
indications of this happening and therefore, the researcher may believe that there are more degrees of freedom than there actually are.

**Detection of Multicollinearity**

Various techniques for detecting multicollinearity have been proposed by both econometricians and statisticians. The important methods of detection for the standardized model (1.1) will be discussed.

The most simple, with no theoretical background, is some arbitrary rule of thumb that has been established to constrain the simple correlations between any two independent variables to say less than $|r_{ij}| = .80$ or $|r_{ij}| = .90$. This will allow the most obvious type of pair wise sample interdependence to be avoided.

It has been noted in this study, however, that this high correlation alone does not necessarily mean multicollinearity will be present. In order for multicollinearity to be a problem, some other factors such as small sample size or small range of $X$ must also be present.

Another rather simple measure that will give some indication of multicollinearity is the determinant of $X'X$, denoted by $|X'X|$. Assume that $X$ has been standarized so that $X_j = 0$ and $X_j^T X_j = 1$ for all $j = 1, 2, ..., k$ and $X$ is defined as an nxk matrix of 1's. Let this standarized $X$ be denoted by $Z$. Since $Z$ is standarized, $0 \leq |Z'Z| \leq 1$. If $|Z'Z| = 0$, an exact linear relationship exists among some columns of $Z$. If $|Z'Z| = 1$, the columns of $Z$ are mutually orthogonal. Anywhere in between, that is $0 < |Z'Z| < 1$, there exists some degree of multicollinearity which becomes more severe as $|Z'Z|$ approaches zero. Although the small determinant will detect the multicollinearity, it will not tell the nature
of the linear relationship.

A more elaborate measure has been indicated by Klein (1962). It is suggested that multicollinearity is not necessarily a problem unless the inter-correlation is high relative to the over-all degree of multiple correlation. That is multicollinearity is considered harmful if

\[ r_{ij} \geq R_y \]

(2.1)

where \( r_{ij} \) is the simple correlation between two independent variables \( X_i \) and \( X_j \), and \( R_y \) is the coefficient of determination or the multiple correlation between the dependent and the independent variables. However, this measure breaks down entirely when it is extended to multiple dimensions.

Farrar and Glauber (1967) suggest that this measure might be improved by extending the concept of simple correlation between independent variables to multiple correlation within the independent variable set. A variable, \( X_i \), would then be said to be harmfully multicollinear if its multiple correlation with the other independent variables was larger than the dependent variables multiple correlation with the entire set. That is

\[ R_i > R_y \]

(2.2)

where \( R_i \) is the multiple correlation of \( X_i \) with all the other dependent variables, and \( R_y \) is the dependent variables multiple correlation with all the independent variables.

Another measure that was presented by Kmenta (1971) is known as \( R^2 (j) \) deletes. \( R^2 (j) \) is the coefficient of determination found by
regressing the dependent variable $Y$ on all the independent variables except $X_1$. Find $R^2$ and each of the $R^2_{(i)}$ as follows:

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon
\]
giving $R^2$

\[
Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \epsilon
\]
giving $R^2_{(1)}$

\[
Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon
\]
giving $R^2_{(2)}$

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon
\]
giving $R^2_{(3)}$

From this find the following measure

\[
R^2 = \max (R^2_{(1)}, R^2_{(2)}, R^2_{(3)})
\]  

(2.3)

If there is a high degree of multicollinearity in the data matrix $X$, (2.3) will be small. However if (2.3) is small, this does not necessarily mean that the data is collinear. The measure given by (2.3) could be small as a result of the variable $X_1$ being a worthless predictor. This measure does not provide a means for detecting which of the independent variables are involved in the multicollinearity.

A measure suggested by Johnston (1972) is in the form of an $F$ statistic. Here again the coefficient of determination, $R^2_{(i)}$ between each $X_1$ and the remaining $(k-1)$ variables in $X$ is used. The hypothesis being tested here is that the multiple correlation coefficients are equal to zero with the suggested statistic given by

\[
F_{i}(r_{1}, r_{1}) = \frac{R^2_{i}/(k-1)}{(1-R^2_{i})/(n-k)} \quad \text{for } i = 2, \ldots, k
\]

(2.4)

This is distributed as $F$ with $(k-1)$ and $(n-k)$ degrees of freedom.

This measure has a tendency to show most of the $F_i$'s as being statistically significant, therefore indicating most of the $X_1$'s are involved in multicollinearity. Farrar and Glauber (1967) suggest this measure
be used only as one in a series of three steps to get a hold of the overall multicollinearity picture.

They first suggest an approximate chi-square test to check for the presence and the severity of the multicollinearity in any regression model (1.1) with more than two independent variables in $\mathbf{X}$. In this case the hypothesis being tested is that the sample standardized $\mathbf{X}'s$ are orthogonal. This statistic is a transformation of $|\mathbf{Z}'\mathbf{Z}|$ which was developed by Bartlett.

$$
X^2 |\mathbf{Z}'\mathbf{Z}|(v) = -(n-1-1/6(2k+5)) \log |\mathbf{Z}'\mathbf{Z}|
$$

which is distributed approximately as chi-square with 1/2k (k-1) degrees of freedom. By transforming $|\mathbf{X}'\mathbf{X}|$ into an approximate chi-square statistic, a meaningful scale is provided for judging the singularity of the $\mathbf{X}'\mathbf{X}$ matrix. If $X^2$ is large, the matrix is close to singular. If $X^2$ is small, the matrix is close to orthogonal. If the assumption of orthogonality is rejected, that is to say some dangerous multicollinearity exists in the data, then it is necessary to locate which of the $\mathbf{X}_1$'s are collinear.

To locate the independent variables which are collinear, Farrar and Glauber suggest the use of the F statistic which Johnson has also suggested as a measure. Farrar and Glauber maintain that inspection of the $F_1$'s will indicate which of the independent variables are affecting the multicollinearity the most.

Finally they suggest to use a t test to locate the pattern of multicollinearity. To locate the variables which are responsible for the multicollinearity, the partial correlation coefficients among the
independent variables are computed and a t test is used to find the statistically significant variables. The null hypothesis being tested here is that the partial correlations are equal to zero. The t statistic is then

\[ t_{ij}(r) = \frac{r_{ij} \sqrt{n-k}}{(1-r_{ij}^2)} \]  

(2.6)

where \( r_{ij} \) denotes the partial correlation between \( X_i \) and \( X_j \). This statistic is distributed as Student's t with \( n-k \) degrees of freedom.

With these three statistics, the severity, location, and pattern of multicollinearity can be found. It should be noted at this point that the chi-square statistic is sensitive to noise other than that produced by the multicollinearity, so it is statistically significant most of the time.

Another measure of multicollinearity is obtained by finding the F statistic from the full model and t statistics from reduced models. These reduced models are obtained by deleting a single variable from the full model. If the F statistic is significant and all the t statistics are not significant, multicollinearity is indicated. The usefulness of this measure is in question since the probability of such an event happening is very small even when multicollinearity is present. Even if this event did happen, there would be no indication as to the nature of the multicollinearity from these statistics.

Other measures have been suggested recently by Webster, Gunst, and Mason (1973, 1974) and Hawkins (1973). They suggest to use the latent roots and vectors of the matrix \( A' A \) where \( A = (Y:X) \). \( A' A \) is therefore
defined as the correlation matrix of dependent and independent variables with latent roots defined by

$$|\Lambda'\Lambda - \lambda_j I| = 0 \text{ for } j = 0, 1, \ldots, k$$  \hspace{1cm} (2.7)

and latent vectors defined by

$$(\Lambda'\Lambda - \lambda_j I)Y_j = 0 \text{ for } j = 0, 1, \ldots, k$$  \hspace{1cm} (2.8)

Now if any $\lambda_j = 0$, then an exact linear dependence exists among some or all of the columns of $\Lambda$. If $\lambda_j = 0$ and $Y_j = 0$, where $Y_{0j}$ is the first element of $Y_j$, then an exact linear dependence exists among the columns of $X$, that relationship being

$$k \sum_{r=1}^{\infty} X_{ir} Y_{rj} = 0.$$  \hspace{1cm} (2.9)

In general none of the latent roots will be zero, but some may be small indicating dangerously near linear dependencies. Webster et. al. suggest that this technique will not only detect the multicollinearity, but will also determine whether the linear combination of (1.2) has any predictive value, or if the multicollinearity itself is useful in prediction.

**Solutions To Multicollinearity**

Solutions which may be adopted when multicollinearity is present in a set of independent variables depend on the severity of the multicollinearity, the availability of other sources of data, the importance of the factors which are multicollinear, and the purpose for which the model is being developed.
One solution was suggested by Kendall (1957). He suggests to reduce the model's information requirements down to the information content of the existing data. This can be accomplished if it is possible to get $X_i'X_j = 0$ for $i \neq j$. By doing this artificial orthogonalization there is a loss of information with the information that is being produced being less well defined than for the original set of variables.

Augmentation of data is frequently mentioned as the best method of handling the multicollinearity problem. Farrar and Glauber (1967) and Silvey (1969) both suggest this method. Augmentation is most effective when the multicollinearity results from the sampling technique or when it results from a over-defined model. Many times, however, more data cannot be obtained due to economic restrictions, change in the population in the study, or simply total lack of additional data. A possibility also exists that when the additional data points are collected, the population under study will be changed. This could conceivably happen if the data points that are collected are outside the region where the multicollinearity is present. These data points might be rare in the population and might have a large influence on the estimating model. If the multicollinearity is due to a physical constraint on the model, it may not be possible to augment the data without changing the population.

Another solution is to use least squares estimation restrictions. It is known that the estimated regression coefficients can be poor estimates of the individual parameters due to multicollinearity in the independent variable matrix, $X$, but this does not imply that the
estimated model is a poor predictor. If the use of the estimated model is restricted to solutions in which the multicollinearity is at least approximately true, the prediction equation works quite well. The reason for this is that, while the individual parameters may be poorly estimated, predicted values of the form

$$\sum_{j=1}^{k} X_{i,j} \beta_j$$

may be estimated with relatively small variances. The k independent variables in (2.9) are those that are involved in the multicollinearity. Therefore if the model is used to predict only in the region that is affected by the multicollinearity, least squares predictions may be used with satisfactory results. A priori information could be used to look at the values of the estimated parameters and see how bad the estimates might be. For example, if a model has been used in which no multicollinearity problems have been encountered, and a new model with additional variables has multicollinear X's among some of the original independent variables, but does not include any of the new variables, least squares estimates can be obtained with restrictions that the estimates of the coefficients on the old independent variables are the same as those obtained in the original model.

Another alternate solution is the use of least squares regression on principle components of $X'X$. Massey (1965), Lott (1973), and Greenberg (1975) suggest this method for a solution.

Ridge regression is proposed by Hoerl and Kennard (1970) and Guilkey and Murry (1975). They suggest to add a small positive con-
stant to each of the diagonal elements of $X'X$ which results in an estimator given by

$$
\hat{\beta} = (X'X + K)^{-1}X'Y
$$

(2.10)

where

$$
K = \begin{bmatrix}
  k_1 & 0 & \ldots & 0 \\
  0 & k_2 & \ldots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \ldots & k_n
\end{bmatrix}
$$

Marquardt (1970) compares a generalized inverse approach to Ridge Regression and non-linear estimation while Fabrycy (1975) suggests changing the mathematical form of the model and then using nonlinear techniques on that.

Another recent method is proposed by Webster, Gunst and Mason (1973, 1974) and Hawkins (1973). This is a modification of the principal components approach and is labeled latent root regression. The problem with this method is that after the latent root transformation, there is some question on how to interpret these variables. The new variables do not correspond with the original variables in any way.

As can be seen, there are many approaches to the solution of a problem with multicollinearity present. No best one is suggested as all of them seem to have their own shortcomings.

The rest of this paper is devoted to finding the effects of certain factors and parameters on multicollinearity and the estimates of the regression coefficient and to investigate some of the solutions to reduce the multicollinearity problem.
CHAPTER III
METHODOLOGY

As stated in the previous chapters, a high correlation among the independent variables adversely affects the estimation problem in multiple regression. It was suspected that other factors also effect the estimation problem. One of these factors is sample size. The smaller the number of observations available for estimation, the poorer the estimation process is likely to be. The amount of variability in the dependent variable should also effect the estimation problem. If there is little variability in the system, the estimates should be good. The range over which the X variables are measured should effect the estimability of the system because when the X's are only measured over a small range, very little information is available on the relationship of the line beyond the narrow range. In addition to these factors, it was felt that the adding of polynominal terms to the mathematical model might also effect the estimability of the regression coefficients.

Scope of Study

To study the effects of these factors on the estimation of a system, a $2^5$ factorial experiment was designed, that is there are two levels of each of the 5 factors. A Monte Carlo procedure was used as the research tool. A computer program was written which allowed
the user to generate a set of variables with any given correlation structure, Hurst and Knop (1972). The program uses the random number generator of the computer system and the techniques for generating normal random numbers. A mathematical model was proposed and five sets of data were generated for each of the thirty two combinations of the factors. A regression analysis was then run on this generated data with the estimates of the regression coefficients being checked against what was fed into the original model.

Two models will be used, one without a polynomial variable and one with. Since it is necessary to have at least three independent variables in the model in order for the multicollinearity statistics to be computed; the first model to be used will be defined as

$$4 + 2X_1 + 2X_2 - 2X_3 + \epsilon_1. \quad (3.1)$$

The second model has the generated variable $X_2$ present as well as the three previously defined independent variables. This model will then be defined as

$$4 + 2X_1 + 2X_2 - 2X_3 + 1X_1^2 + \epsilon_1. \quad (3.2)$$

For each models, two levels of the standard error of the dependent variable, $\xi_1$, will be defined as

$$\xi_1 \sim N(0,1.0) \quad (3.3)$$

and

$$\xi_1 \sim N(0,5.0)$$

The correlation structure of the independent variable will be defined as
\[
R_{XX} = \begin{bmatrix}
1.00 & \rho & \rho \\
\rho & 1.00 & \rho \\
\rho & \rho & 1.00
\end{bmatrix}
\] (3.4)

where \( \rho_1 = .90 \) and \( \rho_2 = .97 \) for the two levels of correlation. The mean vector for the independent variables is kept constant at

\[
\mu_X = \begin{bmatrix}
20.0 \\
20.0 \\
20.0
\end{bmatrix}
\] (3.5)

Two levels of the standard deviations of the independent variables are given by

\[
\sigma_1 = 2.0
\]

and

\[
\sigma_2 = 10.0
\] (3.6)

where \( \sigma_1 \) will denote the low level of the standard deviation and \( \sigma_2 \) will denote the high level. This gives the covariance matrices

\[
\Sigma_{11} = \begin{bmatrix}
4.00 & 3.60 & 3.60 \\
3.60 & 4.00 & 3.60 \\
3.60 & 3.60 & 4.00
\end{bmatrix}
\] (3.7)

and

\[
\Sigma_{12} = \begin{bmatrix}
100 & 90 & 90 \\
90 & 100 & 90 \\
90 & 90 & 100
\end{bmatrix}
\] (3.8)
\[
\Sigma_{21} = \begin{bmatrix}
4.00 & 3.88 & 3.88 \\
3.88 & 4.00 & 3.88 \\
3.88 & 3.88 & 4.00
\end{bmatrix}
\]

and

\[
\Sigma_{22} = \begin{bmatrix}
100 & 97 & 97 \\
97 & 100 & 97 \\
97 & 97 & 100
\end{bmatrix}
\]

where \( \Sigma_{ij} \) is defined by

1 = low level of correlation matrix

2 = high level of correlation matrix

1 = low standard deviation of X

2 = high standard deviation of X

The two sample sizes used are \( n = 25 \) and \( n = 250 \).

**Measured Variables**

It was decided that several indicators of the estimation problem would be investigated. If the multiple regression procedures are giving correct results, then the standard error of the regression coefficients should be good indicators of the precision of the estimates. Since the actual values of the population parameters are known, it is felt that a statistic could be generated indicating how far the various regression coefficients are deviating from the true population parameters. Therefore, a series of statistics was developed to measure this departure
of the sample statistic from the population parameter. The definition of the various statistics are given in equations 3.11, 3.12 and 3.13. These statistics are given the symbol T.

\[
T_T = \frac{\sum_{i=1}^{NXX} \left( \hat{\beta}_i - \beta_i \right)^2}{|\beta_i|}^{\frac{1}{2}}
\]

(3.11)

\[
T(I) = \frac{\left( \hat{\beta}_i - \beta_i \right)^2}{|\beta_i|}^{\frac{1}{2}}
\]

(3.12)

for I = 0, 1, ..., NXX

\[
T_X = \frac{\sum_{i=1}^{NX} \left( \hat{\beta}_i - \beta_i \right)^2}{|\beta_i|}^{\frac{1}{2}}
\]

(3.13)

where NXX is the total number of independent variables, \( \hat{\beta}_i \) is the estimate of the regression coefficient for the independent variable i, and \( \beta_i \) is the true parameter for the ith independent variable.

Analysis of Data

Because of the two different mathematical models, (3.1) and (3.2), gave differing number of sets of regression coefficients, it was decided that instead of running the analysis as a 2^5 factorial, two 2^4 factorial analyses within each of the mathematical models would be run.
CHAPTER IV

RESULTS

The purpose of this study is to find out how certain factors and parameters affect multicollinearity and the estimates of the regression coefficients. From this study it is found that the degree of multicollinearity and the precision of the regression estimates can be detected by examining the standard error of the regression coefficients, and that this measure is a better statistic for the aforementioned problem area than the T statistics. It has also been determined that several factors affect multicollinearity, and that there is reason to believe that Farrar and Glauber's (1967) solution to multicollinearity, that of data augmentation, may be the best solution in many circumstances.

Analysis of Variance Tests

After looking at the results of the several analysis of variances, (ANOVA) produced by the experimental procedure, several conclusions were evident. Since the mathematical models introduced the various variables with the same numerical magnitude of regression coefficient, and since the correlation structure was constant among the independent variables, it was apparent that any one of the regression coefficients was as sensitive as any of the others to the estimation problems. It was therefore decided to use only the standard error of the first regression coefficient and the T statistic for the same coefficient as indicator variables for further study.
The first regression coefficient was chosen because in the two different experiments with changing the model, variable $X_1$ was introduced as $X_1^2$ in the second model. It was therefore felt that any problems would arise due to the $X^2$ term, irregardless of the problem caused by polynomial variables.

**Interpretation of Results**

As stated previously, only the most extreme F-ratios, that is the most significant effects, will be considered. This means that all of the main effects, and four of the two-way factorial effects will be interpreted.

The treatment averages are presented in Table 3. From there, it can be seen that a small range in the independent variables will reduce the precision, that is increase the standard error of the regression coefficient, of the regression coefficients. A small range of $X$ makes the slope of $X$ harder to estimate, and therefore will increase the estimation problem.

Large variability in the dependent variable also increases the standard error of the estimate of the regression coefficients. The more variation introduced into a model, the harder it is to get good estimates.

The use of a transformed, or generated polynomial variable also adds to the estimation problems. When a variable is used to generate a new variable, the precision of the estimate of the original variable falls noticeably. Therefore, models with generated polynomial variables will have poorer estimates than those without generated variables.
TABLE 1 -- ANALYSIS OF VARIANCE FOR MODEL (3.1)

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Mean Squares</th>
<th>F-Ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$S_{\beta_1}$</td>
<td>$T(1)$</td>
</tr>
<tr>
<td>Standard Error of $Y$ ($\sigma_\varepsilon$)</td>
<td>1</td>
<td>8.9154</td>
<td>2.7836</td>
</tr>
<tr>
<td>Sample Size (n)</td>
<td>1</td>
<td>5.5666</td>
<td>2.2800</td>
</tr>
<tr>
<td>Correlation Structure (r)</td>
<td>1</td>
<td>2.0182</td>
<td>.6930</td>
</tr>
<tr>
<td>Standard Deviation of $X$ ($\sigma_X$)</td>
<td>1</td>
<td>8.3212</td>
<td>3.3009</td>
</tr>
<tr>
<td>$\sigma_\varepsilon \times n$</td>
<td>1</td>
<td>3.0260</td>
<td>1.0090</td>
</tr>
<tr>
<td>$\sigma_\varepsilon \times r$</td>
<td>1</td>
<td>1.2311</td>
<td>.3057</td>
</tr>
<tr>
<td>$\sigma_\varepsilon \times \sigma_X$</td>
<td>1</td>
<td>4.3419</td>
<td>1.4625</td>
</tr>
<tr>
<td>$n \times r$</td>
<td>1</td>
<td>.8183</td>
<td>.5547</td>
</tr>
<tr>
<td>$n \times \sigma_X$</td>
<td>1</td>
<td>2.6449</td>
<td>1.3925</td>
</tr>
<tr>
<td>$r \times \sigma_X$</td>
<td>1</td>
<td>.9110</td>
<td>.2181</td>
</tr>
<tr>
<td>$\sigma_\varepsilon \times n \times r$</td>
<td>1</td>
<td>.5901</td>
<td>.2457</td>
</tr>
<tr>
<td>$\sigma_\varepsilon \times r \times \sigma_X$</td>
<td>1</td>
<td>.6362</td>
<td>.0965</td>
</tr>
<tr>
<td>$\sigma_\varepsilon \times n \times \sigma_X$</td>
<td>1</td>
<td>1.5667</td>
<td>.6162</td>
</tr>
<tr>
<td>$n \times r \times \sigma_X$</td>
<td>1</td>
<td>.3499</td>
<td>.2029</td>
</tr>
<tr>
<td>Error</td>
<td>65</td>
<td>.02379</td>
<td>.1374</td>
</tr>
</tbody>
</table>

* Indicates statistical significance at the 5 percent level of significance.
### TABLE 2 -- ANALYSIS OF VARIANCE FOR MODEL (3.2)

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Mean Squares</th>
<th>F-Ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sβ1</td>
<td>T(1)</td>
</tr>
<tr>
<td>Standard Error of Y (σε)</td>
<td>1</td>
<td>126.2578</td>
<td>55.20*</td>
</tr>
<tr>
<td>Sample Size (n)</td>
<td>1</td>
<td>98.4311</td>
<td>43.03*</td>
</tr>
<tr>
<td>Correlation Structure (r)</td>
<td>1</td>
<td>10.9163</td>
<td>.05</td>
</tr>
<tr>
<td>Standard Deviation of X(σₓ)</td>
<td>1</td>
<td>225.6773</td>
<td>111.78*</td>
</tr>
<tr>
<td>σε x n</td>
<td>1</td>
<td>43.7570</td>
<td>19.13*</td>
</tr>
<tr>
<td>σε x r</td>
<td>1</td>
<td>4.8690</td>
<td>2.13</td>
</tr>
<tr>
<td>σε x σₓ</td>
<td>1</td>
<td>100.3196</td>
<td>43.86*</td>
</tr>
<tr>
<td>n x r</td>
<td>1</td>
<td>8.9478</td>
<td>3.91</td>
</tr>
<tr>
<td>n x σₓ</td>
<td>1</td>
<td>79.5404</td>
<td>34.77*</td>
</tr>
<tr>
<td>r x σₓ</td>
<td>1</td>
<td>8.5117</td>
<td>3.72</td>
</tr>
<tr>
<td>σε x n x r</td>
<td>1</td>
<td>3.9920</td>
<td>1.75</td>
</tr>
<tr>
<td>σε x r x σₓ</td>
<td>1</td>
<td>3.7974</td>
<td>1.66</td>
</tr>
<tr>
<td>σε x n x σₓ</td>
<td>1</td>
<td>35.3614</td>
<td>15.46*</td>
</tr>
<tr>
<td>n x r x σₓ</td>
<td>1</td>
<td>7.6566</td>
<td>3.35</td>
</tr>
<tr>
<td>Error</td>
<td>65</td>
<td>2.2873</td>
<td>--</td>
</tr>
</tbody>
</table>

* Indicates statistical significance at the 5 percent level of significance.
### TABLE 3 -- TREATMENT AVERAGES FOR ($S_{g1}$)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Correlation Structure</th>
<th>Range of $X(\sigma_X)$</th>
<th>$\sigma_\varepsilon$</th>
<th>Model 3.1</th>
<th>Model 3.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>.90</td>
<td>2.0</td>
<td>1.0</td>
<td>.2696</td>
<td>.2483</td>
</tr>
<tr>
<td>25</td>
<td>.90</td>
<td>2.0</td>
<td>5.0</td>
<td>.0487</td>
<td>.0522</td>
</tr>
<tr>
<td>25</td>
<td>.90</td>
<td>10.0</td>
<td>1.0</td>
<td>.4116</td>
<td>.4213</td>
</tr>
<tr>
<td>25</td>
<td>.90</td>
<td>10.0</td>
<td>5.0</td>
<td>.1068</td>
<td>.0993</td>
</tr>
<tr>
<td>25</td>
<td>.97</td>
<td>2.0</td>
<td>1.0</td>
<td>.0846</td>
<td>.0860</td>
</tr>
<tr>
<td>25</td>
<td>.97</td>
<td>2.0</td>
<td>5.0</td>
<td>.0175</td>
<td>.0167</td>
</tr>
<tr>
<td>25</td>
<td>.97</td>
<td>10.0</td>
<td>1.0</td>
<td>.1518</td>
<td>.1583</td>
</tr>
<tr>
<td>25</td>
<td>.97</td>
<td>10.0</td>
<td>5.0</td>
<td>.0283</td>
<td>.0307</td>
</tr>
<tr>
<td>250</td>
<td>.90</td>
<td>2.0</td>
<td>1.0</td>
<td>1.348</td>
<td>1.2401</td>
</tr>
<tr>
<td>250</td>
<td>.90</td>
<td>2.0</td>
<td>5.0</td>
<td>.2433</td>
<td>.2604</td>
</tr>
<tr>
<td>250</td>
<td>.90</td>
<td>10.0</td>
<td>1.0</td>
<td>2.938</td>
<td>2.1094</td>
</tr>
<tr>
<td>250</td>
<td>.90</td>
<td>10.0</td>
<td>5.0</td>
<td>.5337</td>
<td>.4966</td>
</tr>
<tr>
<td>250</td>
<td>.97</td>
<td>2.0</td>
<td>1.0</td>
<td>.4200</td>
<td>.4300</td>
</tr>
<tr>
<td>250</td>
<td>.97</td>
<td>2.0</td>
<td>5.0</td>
<td>.0867</td>
<td>.0837</td>
</tr>
<tr>
<td>250</td>
<td>.97</td>
<td>10.0</td>
<td>1.0</td>
<td>.7459</td>
<td>.7918</td>
</tr>
<tr>
<td>250</td>
<td>.97</td>
<td>10.0</td>
<td>5.0</td>
<td>.1443</td>
<td>.1536</td>
</tr>
</tbody>
</table>
The correlation structure among the independent variables does influence multicollinearity as suggested in the texts. It should be noted however that this main effect is the least significant of the main effects in Model (3.1) and not significant at all in Model (3.2). This finding indicates that although high correlations do exist among a set of independent variables, the estimation problem may not be too bad. A look at the Coefficient of Variation for $\hat{\beta}_1$ ($CV_{\hat{\beta}_1}$) will point this out. In the case when there exists the most favorable conditions for estimability, that is large range of X, small variability in Y, large sample size, and no generated variables, but high correlation structure, the $CV_{\hat{\beta}_1} = .81$. It is believed that this level of variation is tolerable in most situations. An explanation of why the correlation structure is not significant in Model (3.2) follows in the next section.

Sample size also influences the estimations of the regression coefficients as can be seen in Table 3. As sample size increases, the precision of the estimate also increases. A further study of this result is found later in this chapter.

The significant two way factorial effects that will be considered are the variability of the dependent variable or Y and the sample size, the variability of Y and the range of X, the sample size and the range of X, and the correlation structure and the sample size.

As can be seen in Figure 1 and Table 4 and Table 8, the interaction of a small sample size with a large variability of Y produce poor estimates of the coefficients. An increase in sample size increases the precision of the estimates significantly even if the variability is large.
FIGURE 1 — EFFECT OF STANDARD ERROR OF Y AND SAMPLE SIZE ON PRECISION OF ESTIMATION ($S_{\beta_1}$).

SOLID LINE MODEL (3.1)
DASHED LINE (3.2)
FIGURE 2 -- EFFECT OF STANDARD ERROR OF $Y$ AND STANDARD DEVIATION OF $X$ ON THE PRECISION OF ESTIMATION ($s_{b_1}$).

SOLID LINE MODEL (3.1)
DASHED LINE (3.2)
FIGURE 3 -- EFFECT OF STANDARD DEVIATION OF X AND SAMPLE SIZE ON THE PRECISION OF ESTIMATION ($S_{\beta_1}$).  
SOLID LINE MODEL (3.1)  
DASHED LINE (3.2)
FIGURE 4 -- EFFECT OF STANDARD DEVIATION OF X AND CORRELATION STRUCTURE ON PRECISION OF ESTIMATION ($S_{\beta_1}$).
TABLE 4 -- EFFECT OF STANDARD ERROR OF Y AND SAMPLE SIZE ON THE PRECISION OF ESTIMATION ($s_{\beta_1}$) FOR MODEL (3.1).

<table>
<thead>
<tr>
<th>Standard error of y ($\sigma_\varepsilon$)</th>
<th>1.0</th>
<th>5.0</th>
<th>Row mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size (n)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.209</td>
<td>1.266</td>
<td>0.738</td>
</tr>
<tr>
<td>250</td>
<td>0.071</td>
<td>0.349</td>
<td>0.210</td>
</tr>
<tr>
<td>Col. mean</td>
<td>0.140</td>
<td>0.808</td>
<td>0.474</td>
</tr>
</tbody>
</table>

TABLE 5 -- EFFECT OF STANDARD ERROR OF Y AND STANDARD DEVIATION OF X ON THE PRECISION OF ESTIMATION ($s_{\beta_1}$) FOR MODEL (3.1).

<table>
<thead>
<tr>
<th>Standard error of y ($\sigma_\varepsilon$)</th>
<th>1.0</th>
<th>5.0</th>
<th>Row mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation of $X_\varepsilon$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.229</td>
<td>1.362</td>
<td>0.796</td>
</tr>
<tr>
<td>10.0</td>
<td>0.050</td>
<td>0.252</td>
<td>0.151</td>
</tr>
<tr>
<td>Col. mean</td>
<td>0.140</td>
<td>0.807</td>
<td>0.474</td>
</tr>
</tbody>
</table>
### TABLE 6 -- EFFECT OF STANDARD DEVIATION OF X AND SAMPLE SIZE ON THE PRECISION OF ESTIMATION ($s_{\hat{b}_1}$) FOR MODEL (3.1)

<table>
<thead>
<tr>
<th>Sample size (n)</th>
<th>Standard deviation of X($\sigma_X$)</th>
<th>2.0</th>
<th>10.0</th>
<th>Row mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1.242</td>
<td>0.233</td>
<td>0.738</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0.351</td>
<td>0.069</td>
<td>0.210</td>
<td></td>
</tr>
<tr>
<td>Col. mean</td>
<td>0.797</td>
<td>0.151</td>
<td>0.474</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 7 -- EFFECT OF STANDARD DEVIATION OF X AND CORRELATION STRUCTURE ON PRECISION OF ESTIMATION ($s_{\hat{b}_1}$) FOR MODEL (3.1)

<table>
<thead>
<tr>
<th>Correlation structure (r)</th>
<th>Standard deviation of X($\sigma_X$)</th>
<th>2.0</th>
<th>10.0</th>
<th>Row mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>0.531</td>
<td>0.099</td>
<td>0.315</td>
<td></td>
</tr>
<tr>
<td>0.97</td>
<td>1.062</td>
<td>0.203</td>
<td>0.632</td>
<td></td>
</tr>
<tr>
<td>Col. mean</td>
<td>0.797</td>
<td>0.151</td>
<td>0.474</td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 8 -- EFFECT OF STANDARD ERROR OF Y AND SAMPLE SIZE ON THE PRECISION OF ESTIMATION (s_{\beta_1}) FOR MODEL (3.2)

<table>
<thead>
<tr>
<th>Sample size (n)</th>
<th>Standard error of y ((\varepsilon_y))</th>
<th>1.0</th>
<th>5.0</th>
<th>Row mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td></td>
<td>0.998</td>
<td>4.989</td>
<td>2.994</td>
</tr>
<tr>
<td>250</td>
<td></td>
<td>0.258</td>
<td>1.292</td>
<td>0.775</td>
</tr>
<tr>
<td>Col. mean</td>
<td></td>
<td>0.628</td>
<td>3.140</td>
<td>1.884</td>
</tr>
</tbody>
</table>

### TABLE 9 -- EFFECT OF STANDARD ERROR OF Y AND STANDARD DEVIATION OF X ON THE PRECISION OF ESTIMATION (s_{\beta_1}) FOR MODEL 3.2.

<table>
<thead>
<tr>
<th>Standard deviation of X ((\sigma_X))</th>
<th>Standard error of y ((\sigma_\varepsilon))</th>
<th>1.0</th>
<th>5.0</th>
<th>Row mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td></td>
<td>1.188</td>
<td>5.940</td>
<td>3.564</td>
</tr>
<tr>
<td>10.0</td>
<td></td>
<td>0.068</td>
<td>0.341</td>
<td>0.204</td>
</tr>
<tr>
<td>Col. mean</td>
<td></td>
<td>0.628</td>
<td>3.140</td>
<td>1.884</td>
</tr>
</tbody>
</table>
TABLE 10 -- EFFECT OF STANDARD DEVIATION OF X AND SAMPLE SIZE ON THE PRECISION OF ESTIMATION \( (s_{\beta_1}) \) FOR MODEL (3.2)

<table>
<thead>
<tr>
<th>Sample size (n)</th>
<th>2.0</th>
<th>10.0</th>
<th>Row mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>5.670</td>
<td>0.317</td>
<td>2.994</td>
</tr>
<tr>
<td>250</td>
<td>1.458</td>
<td>0.093</td>
<td>0.775</td>
</tr>
<tr>
<td>Col. mean</td>
<td>3.564</td>
<td>0.205</td>
<td>1.884</td>
</tr>
</tbody>
</table>
The interaction of the variability of $Y$ and the range of $X$ can be detected in Figure 2 and Table 5 and Table 9. A small range of $X$ and a large variability in $Y$ will produce poor estimates of the regression coefficients. With a large range, the precision is increased for any level of the variability of $Y$.

Figure 3 and Table 6 and Table 10 indicate the effects of sample size and the range of $X$ on the estimates. A small range of $X$ and a small sample size produce poor estimates. Here again, if the sample size is increased, the precision of the estimates increases. Also a large range of $X$ increases the precision regardless of the sample size.

The interaction of the correlation structure and the range of $X$ is not significant in Model (3.2) but is highly significant in Model (3.1) and should be considered. This factorial effect is shown in Figure 4 and Table 7. A small range of $X$ and high correlation structure will give poor estimates. Again a larger range of $X$ will increase the estimate precision.

**Effect of Correlation Structure**

As stated previously, the correlation effect was not as significant as the other main effects in Model (3.1) and not significant at all in Model (3.2).

A possible reason for this result is seen by looking at the correlation between the generated variable in Model (3.2), $X_1^2$, and all the other variables. The correlation between any two variables is given in general by

$$R(X_i, X_j) = \frac{\text{COV}(X_i, X_j)}{\sqrt{\text{VAR}(X_i)} \cdot \text{VAR}(X_j)}$$
For this study, \( X_1 = X_1^2 \) and \( X_j = X_1, X_2, X_3 \). This then gives in particular

\[
R(X_1^2, X_j) = \frac{\text{COV}(X_1^2, X_j)}{\sqrt{\text{VAR}(X_1^2)} \cdot \text{VAR}(X_j)}
\]  \hspace{1cm} (4.1)

for \( j = 1, 2, 3 \)

Where

\[
\text{COV}(X_1^2, X_j) = \mu_1^2 \mu_j + 2\sigma_1 \mu_1 + \sigma_1^2 \mu_j - (\mu_1^2 + \sigma_1^2)\mu_1
\]  \hspace{1cm} (4.2)

and

\[
\text{VAR}(X_1^2) = E(X, r) = 3\sigma_1^4 + 6\sigma_1^2 \mu_1^2 + \mu_1^4
\]  \hspace{1cm} (4.3)

Calculating the appropriate correlations, the correlation matrices for model (3.2) are then given by

\[
R_{11} = \begin{bmatrix}
1.00 & .90 & .90 & .38 \\
.90 & 1.00 & .90 & .17 \\
.90 & .90 & 1.00 & .17 \\
.38 & .17 & .17 & 1.00 \\
\end{bmatrix}
\]

\[
R_{21} = \begin{bmatrix}
1.00 & .97 & .97 & .77 \\
.97 & 1.00 & .97 & .18 \\
.97 & .97 & 1.00 & .18 \\
.77 & .18 & .18 & 1.00 \\
\end{bmatrix}
\]

\[
R_{12} = \begin{bmatrix}
1.00 & .90 & .90 & .38 \\
.90 & 1.00 & .90 & .55 \\
.90 & .90 & 1.00 & .55 \\
.38 & .58 & .55 & 1.00 \\
\end{bmatrix}
\]
\[ R_{22} = \begin{bmatrix} 1.00 & .97 & .97 & .77 \\ .97 & 1.00 & .97 & .59 \\ .97 & .97 & 1.00 & .59 \\ .77 & .57 & .59 & 1.00 \end{bmatrix} \]

where \( R_{ij} \) denotes

\[ \begin{align*} i & = \begin{cases} 1 & \text{if low correlation structure} \\ 2 & \text{if high correlation structure} \end{cases} \\ j & = \begin{cases} 1 & \text{if low standard deviation of } X \\ 2 & \text{if high standard deviation of } X \end{cases} \end{align*} \]

As can be seen, the high correlation structure is not as high any more. This result might then indicate that the transformation of a variable reduces the correlation effect on the problem.

**Effect of Sample Size**

One of the solutions to multicollinearity as proposed by Farrar and Glauber (1967) is to augment the data set. They suggest no reason why this is their preferred method. In order to get a hold on a reason why this method works, the coefficient of variation of the regression coefficient, \( \beta_1 \), is examined.

A separate experiment was conducted using the worst case of multicollinearity, that is a transformed variable, small range of independent variables, high correlation structure, and high standard error of the dependent variable. Examining the standard error of regression coefficients as the sample size increases, a good idea of what happens to the precision
of the regression coefficients can be formulated. As seen in Figure 5 or Table 11, as the sample size increases from \( n = 10 \) to \( n = 50 \), the coefficient of variation drops significantly. Therefore, to augment the data as suggested by Farrar and Glauber is a very effective solution to the multicollinearity problem on the estimation of the regression coefficients only if the original sample size is relatively small, say less than 50. Thus it appears that multicollinearity problems can be alleviated to a large degree by increasing the sample size.

Table 11: COEFFICIENT OF VARIATION ON \( \beta_1 \) WITH BAD MULTICOLLINEARITY

<table>
<thead>
<tr>
<th>( n )</th>
<th>( CV_{\beta_1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>.352</td>
</tr>
<tr>
<td>15</td>
<td>.334</td>
</tr>
<tr>
<td>20</td>
<td>.389</td>
</tr>
<tr>
<td>25</td>
<td>.198</td>
</tr>
<tr>
<td>30</td>
<td>.219</td>
</tr>
<tr>
<td>40</td>
<td>.150</td>
</tr>
<tr>
<td>50</td>
<td>.135</td>
</tr>
<tr>
<td>75</td>
<td>.130</td>
</tr>
<tr>
<td>100</td>
<td>.101</td>
</tr>
<tr>
<td>150</td>
<td>.101</td>
</tr>
<tr>
<td>200</td>
<td>.082</td>
</tr>
<tr>
<td>250</td>
<td>.082</td>
</tr>
<tr>
<td>300</td>
<td>.067</td>
</tr>
<tr>
<td>500</td>
<td>.056</td>
</tr>
</tbody>
</table>
Figure 5 -- EFFECT OF SAMPLE SIZE ON COEFFICIENT OF VARIATION OF $\beta_1$. 

**Figure 5** -- EFFECT OF SAMPLE SIZE ON COEFFICIENT OF VARIATION OF $\beta_1$. 

[Graph showing the effect of sample size on the coefficient of variation of $\beta_1$.]
CHAPTER V

CONCLUSIONS

Summary

It has been found that certain factors do influence multicollinearity and the estimates of the regression coefficients in a multiple regression analysis.

The literature leads one to believe that when some independent variables are highly correlated multicollinearity is a problem, and therefore the estimates obtained are poor. This is the only condition placed on the model with nothing else being said about other conditions needing to be present for multicollinearity to be a problem. It has been found in this study that other factors also aggravate multicollinearity.

A small range of $X$ will compound the multicollinearity problem. With a small range, the slope of $X$ is harder to estimate, thus increasing the estimation problem.

The larger the variability of the dependent variable, the more variation introduced into the model, and the greater the estimation problem.

A transformed, or generated polynomial variable adversely affects the estimation of the coefficients. When a variable is used to generate another polynomial variable, the precision of the estimate of the original variable decreases significantly.
A small sample size will also compound the estimation problem. As the sample size increases, the precision of the estimates increases to a point that would be acceptable to a researcher, even when the other factors are at undesirable levels.

Finally, it was found that the standard error of the regression coefficients is a good indicator of potential estimation problems, which is what this measure is supposed to do. The standard error is as good a measure as the calculated standardized time coefficient statistic, T, and is available to the researcher.

Future Areas of Study

This study has brought out future questions to be answered. One of the questions to be investigated is that of how well the suggested multicollinearity statistics do in localizing the variables causing the multicollinearity problem. There is some indication from this study that the collinear variables can be detected.

Another area to pursue is an analytical solution to the way the precision of the estimates increases as the sample size increases.

One final area of interest might be to investigate the effect of other generated variables rather than the polynomial, on the regression coefficients estimates.
LITERATURE CITED


