DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

PROBLEM OF OMITTED VARIABLE IN REGRESSION MODEL SPECIFICATION

by

Suay EREEŞ

June, 2009
İZMİR

PROBLEM OF OMITTED VARIABLE IN REGRESSION MODEL SPECIFICATION

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science
in Statistics

by
Suay EREEŞ

June, 2009
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**PROBLEM OF OMITTED VARIABLE IN REGRESSION MODEL SPECIFICATION”** completed by **SUAY EREEŞ** under supervision of **PROF. DR. SERDAR KURT** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

....

Prof. Dr. Serdar KURT

Supervisor

....

Assist. Prof. Dr. A. Kemal ŞEHİRLİOĞLU

(Jury Member)

....

Assist.Prof. Dr. A. Fırat ÖZDEMİR

(Jury Member)

Prof. Dr. Cahit HELVACI

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Prof. Dr. Serdar KURT for his guidance and valuable advises throughout this study.

I am also very grateful to Dr. Neslihan DEMİREL for all her guidance and valuable helps.

Finally, I would like to express my deepest gratitude to my husband Erşans EREEŞ and my all family for their encouragements and patience.

Suay EREEŞ

PROBLEM OF OMITTED VARIABLE IN REGRESSION MODEL SPECIFICATION

ABSTRACT

In many non-experimental studies, the analyst may not have access to all relevant variables, and does not include these variables into the model and omits them. To omit some variables that affect the dependent variable from the model may cause omitted variables bias. In this thesis, it is aimed to investigate the omitted variable bias, its importance, reasons, and consequences and to research the methods for dealing with omitted variable bias and RESET test which is a method for detecting omitted variable(s).

In this study, a simulation was performed by using the programs written in Minitab which is a statistical software package. Three types of populations with 1000 observations which varied depending on the correlations between the variables were generated and random samples were drawn from these populations. Though the true model had three independent variables, the models were estimated by omitting one and then two independent variables for each sample. 10,000 repetitions were generated for each of sample sizes. Therefore when correlations were changed and the number of omitted variables was increased, the effects of omitted variable bias were investigated. The amount of bias, the estimated coefficients, coefficients of determination and the adjusted coefficients of determination, standard deviations of the estimated coefficients were computed for every model and *F* statistics were also computed for applying RESET test and they were all compared for each population. Moreover, by increasing the sample size, it was investigated whether the effects of omitted variable bias were changed depending on sample size.

Keywords: Regression analysis, model specification error, omitted variable bias, RESET test

REGRESYON MODELİ BELİRLEMEDE DIŞLANAN DEĞİŞKEN SORUNU

ÖZ

Deneysel olmayan pek çok çalışmada, araştırmacı model için gerekli olan tüm değişkenlere ulaşamamakta ve bu değişkenleri modele dahil edememekte, dolayısıyla modelden dışlamaktadır. Bağımlı değişkeni önemli derecede etkileyen bazı değişkenlerin modele alınmaması dışlanan değişken yanlılığına sebep olmaktadır. Bu tezde, dışlanan değişken yanlılığı, bu yanlılığın önemi, nedeni ve sonuçları araştırılırken dışlanan değişken sorununu ortadan kaldırmak için kullanılan yöntemler incelenmiş ve ayrıca modelden dışlanan değişkenlerin varlığını saptamak üzere RESET testi kullanılmıştır.

Bu çalışmada, Minitab istatistiksel paket programı kullanılarak bir benzetim çalışması yapılmıştır. Değişkenler arasındaki korelasyon değerlerine bağlı olarak değişen 1000 verilik üç değişik tipte kitle türetilmiş ve bu kitlelerden rassal örneklemler çekilmiştir. Gerçek model üç bağımsız değişken ile kurulmuş, sırasıyla bir ve iki değişken dışlanarak her örneklem için yeni modeller elde edilmiştir. Böylece korelasyon değerleri değiştiğinde ve dışlanan değişken sayısı arttığında dışlanan değişken yanlılığının ne gibi etkileri olduğu incelenmiştir. Yanlılık miktarları, katsayı kestirimleri, belirtme katsayıları, tahmini katsayılara ilişkin standart sapmalar hesaplanmıştır. Ayrıca, *F* istatistikleri de RESET testi uygulayabilmek için elde edilmiştir. Bu işlemler 10,000 defa tekrarlanmıştır ve sonuçların birbirleriyle karşılaştırmaları yapılmıştır. Son olarak, örneklem ölçüsü arttırılarak dışlanan değişken yanlılığının örneklem ölçüsüne bağlı olarak değişip değişmediği de araştırılmıştır.

Anahtar sözcükler: Regresyon analizi, model spesifikasyon hatası, dışlanan değişken yanlılığı, RESET testi

CONTENTS

Page

CHAPTER ONE
INTRODUCTION

Regression analysis is a statistical tool for investigation of relationships between variables. In general, the investigator seeks to ascertain the casual effect of one variable upon another or others. In many non-experimental studies, however, the analyst may not have access to all relevant variables, and does not include these variables into the model. It is sometimes impossible to measure some variables such as socio economic status. Furthermore, sometimes some variables may be measurable but require too much time and abandoned. Therefore they are omitted from the model. The omission from a regression of some variables that affect the dependent variable may cause an omitted variables bias. This bias depends on the correlation between the independent variables which are omitted and included. Hence, this omission may lead to biased estimates of model parameters. The problem arises because any omitted variable becomes part of the error term, and the result may be a violation of an important assumption for being an unbiased estimator. This assumption logically implies the absence of correlation between the explanatory variables included in the regression and the expected value of the error term, because whatever the value of any independent variable, the expected value of the error term is always zero. Thus, unless the omitted variable is uncorrelated with the included ones, the coefficients of the included ones will be biased because the assumption is violated, it means that, they now reflect not only an estimate of the effect of the variable which they are associated, with but also partly the effects of the omitted variable.

The purpose of this study is to investigate omitted variable bias, its importance, reasons, and consequences.

This thesis contains five chapters. In Chapter 1, a short description of the entire study is summarized. In Chapter 2, introduction to regression analysis and methods of selection of independent variables are mentioned, because of constituting a basic for the third chapter. Problem of omitted variable, RESET test for detecting omitted variables and the methods for dealing with omitted variable bias such as proxy variable are discussed in Chapter 3. In Chapter 4, omitted variable bias and its effects on the parameters and RESET test are presented using simulation. Chapter 4 also include the simulation study to examine the effects of the larger sample size on omitted variable bias. Finally, in Chapter 5, the conclusions related to the simulation study are presented.

CHAPTER TWO
MULTIPLE REGRESSION

2.1 Introduction

Simple regression is a procedure which is used for obtaining a linear equation that predicts a dependent variable as a function of a single independent variable. However, in many situations several independent variables jointly influence a dependent variable. Multiple regression enables to determine the simultaneous effect of several independent variables on a dependent variable using the least square principle.

2.2 Multiple Regression Models

Multiple regression is a statistical method for studying the relationship between a single dependent variable and one or more independent variables. It is admittedly one of the most widely used of all statistical methods and generally used in social, biological and physical sciences. The basic uses of multiple regression are prediction and casual analysis (Mendenhall & Sincich, 2003).

Many mathematical formulas can serve to express relationships between more than two variables, but most commonly used in statistics are linear equations of the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad (2.1)$$

$\beta_0, \beta_1, \ldots, \beta_{p-1}$ are the parameters

$X_{i1}, \ldots, X_{i,p-1}$ are known constants

ε_i are independent random variables with mean zero and variance σ^2

i = 1, … , *n;* number of observations

It can also be written as:

$$Y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \varepsilon_i \tag{2.2}$$

Assuming that $E(\varepsilon_i) = 0$, the response function for regression model is:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} \tag{2.3}$$

The parameter β_k indicates the change in the mean response *E(Y)* with a unit increase in the independent variable X_k, when all other independent variables in the model are held constant (Neter, Kutner, Nachtsheim & Wasserman, 1996)

2.2.1 Least Squares Estimators

The population regression model is a useful theoretical construct, but for applications finding the real values of parameters can not be possible, therefore an estimate of the model is needed to be determined. To determine the estimated model, estimators for the unknown parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ should be found. These estimators are simply procedures for making guesses about the unknown parameters on the basis of known sample values of $Y, X_1, X_2, \ldots, X_{p-1}$. For any estimates of the parameters, denoted by $b_0, b_1, \ldots, b_{p-1}$, the value for *Y* can be estimated by

$$\hat{Y} = b_0 + b_1 X_1 + \cdots + b_{p-1} X_{p-1} \tag{2.4}$$

The coefficient estimators are obtained using equations derived by using the method of least squares (Neter, Kutner, Nachtsheim & Wasserman, 1996).

2.2.2 Method of Least Squares

The difference between the actual (observed) and predicted values for each observation is

$$e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_1 - \cdots - b_{p-1} X_{p-1} \tag{2.5}$$

e_i is called the residual for i^{th} observation and is the vertical distance between the estimated plane and the actual observation Y_i. This means, when the absolute values of e_i become larger, the estimated plane does the worse at representing the data. Since e_i indicate how closely an estimated plane comes to describing the data points, it is a reasonable approach to compare the values of e_i for choosing among alternative estimators. A mathematical function that represents the effect of squaring all of the residuals and computing the sum of squared residuals is computed. This function which is defined as sum of squared errors includes the coefficients. According to the method of least squares, the coefficient estimators are obtained as the estimators minimizing the sum of squared errors (Draper & Smith, 1966).

$$SSE = \sum e_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 \tag{2.6}$$

If the regression model has n independent variables, then the least square estimators can be solved using matrix forms.

$$\underset{nx1}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \tag{2.7}$$

$$\underset{nxp}{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \tag{2.8}$$

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} \tag{2.9}$$

The least squares normal equations for the general linear regression model:

$$X'Xb = X'Y \tag{2.10}$$

And the least squares estimators:

$$b = (X'X)^{-1}(X'Y). \tag{2.11}$$

2.2.3 Assumptions of Least Square Regression

All statistical procedures including multiple regression require the assumptions be made for their mathematical development. If these assumptions hold, then in large samples the Ordinary Least Squares (OLS) estimators have sampling distributions that are normal. In turn, this large-sample normal distribution allows for developing methods for hypothesis testing and constructing confidence intervals using the OLS estimators (Stock & Watson, 2003).

Nevertheless, violation of an assumption may potentially lead to some problems. First and more serious, the estimate of the regression coefficients may be biased in such cases, the estimates of the regression coefficients, R^2, significance tests, and confidence intervals may all be incorrect. Second, only the estimate of the standard error of the regression coefficients may be biased. In such cases, the estimated value of the regression coefficients is correct, but hypothesis tests and confidence intervals may be incorrect. Third, the estimated model would have large variances, and the estimated model would not be as efficient as it should be. These problems are all very important but fortunately, remedial measures are available for handling the problems resulting from violations of assumptions.

Many of the assumptions focus on the residuals; consequently, careful examination of the residuals can often help identify problems with regression models. All these assumptions are not only required for the OLS estimation of model parameters but are necessary for reliable confidence intervals and hypothesis tests based on *t* distributions or *F* distributions (Field, 2005).

2.2.3.1 Zero Mean Value of Error Term

The first least squares assumption is that the conditional distribution of ε_i given X_i has a mean of zero. This assumption is a formal mathematical statement about the other variables contained in ε_i and asserts that these other variables are unrelated to X_i in the sense that, given a value of X_i, the mean of the distribution of these other variables is zero.

$$E(\varepsilon_i | X_i) = 0$$

The assumption that $E(\varepsilon_i | X_i) = 0$ is equivalent to assuming that the population regression line is the conditional mean of Y_i given X_i.

The conditional mean assumption $E(\varepsilon_i | X_i) = 0$ implies that X_i and ε_i are uncorrelated, or $\operatorname{cov}(X_i, \varepsilon_i) = 0$. Because correlation is a measure of linear association, this implication does not go the other way; even if X_i and ε_i are uncorrelated, the conditional mean of ε_i given X_i might be nonzero. If X_i and ε_i

are correlated, then the conditional mean assumption is violated (Stock & Watson, 2003).

2.2.3.2 Independence of Residuals

The residuals of the observations must be independent of one another. Otherwise stated, there must be no relationship among the residuals for any subset of cases in the analysis. This assumption will be met in any random sample from a population. However, if data are clustered or temporally linked, then the residuals may not be independent. Clustering occurs when data are collected from groups. The most common situation in which this assumption might not be met is when the observations represent repeated measurements on sampling or experimental units. Such data are often termed longitudinal and arise from longitudinal studies (Cohen, 2003).

2.2.3.3 Constant Variance of Residuals (Homoscedasticity)

The conditional variance of the residuals around the regression line in the population, for any value of the independent variable X, is assumed to be constant. Conditional variances represent the variability of the residuals around the predicted value for a specified value of X. Consequently, each probability distribution for Y has the same standard deviation regardless of the X-value (Cohen, 2003).

2.2.3.4 Normality of Residuals

The residuals around the regression line, for any value of the independent variable X, are assumed to have a normal distribution (Cohen, 2003). The validity of the normality assumption can be assessed by examination of appropriate graphs of residuals (Chatterjee & Hadi, 2006).

2.2.3.5 No Multicollinearity

There are no perfect linear relationships among the independent variables. A potential problem when running a multiple regression is that two or more independent variables are very highly intercorrelated with each other. This is referred to as multicollinearity. The problem with multicollinearity is that it is likely to prevent any of the individual variables from being significant (Dewberry, 2004).

2.2.4 Properties of Least Squares Estimators

With these assumptions the least square estimator can be shown to have minimum variance among all estimators that are linear functions of the observed *Y*'s and *X*'s and that are unbiased. Unbiased estimators with minimum variance are said to be the best or most efficient estimators. Thus, the least square estimator is called BLUE (best linear unbiased estimator). The formulas and expressions of these properties are presented below by depending on simple linear regression model (Hanushek & Jackson, 1977).

2.2.4.1 Linearity

The least squares estimator is linear in *Y*. Since *Y* is a random variable, and *X* is assumed fixed, *X* is simply the weight of *Y*.

$$\begin{aligned} b_1 &= \frac{S_{XY}}{S_{XX}} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum (X_i - \overline{X})Y_i - \sum (X_i - \overline{X})\overline{Y}}{\sum (X_i - \overline{X})^2} \\ &= \frac{\sum (X_i - \overline{X})Y_i - \overline{Y}\sum (X_i - \overline{X})}{\sum (X_i - \overline{X})^2} = \frac{\sum Y_i (X_i - \overline{X})}{\sum (X_i - \overline{X})^2} \\ &= \sum \left(\frac{X_i - \overline{X}}{\sum (X_i - \overline{X})^2} \right) Y_i \end{aligned}$$

$$b_1 = \sum w_i Y_i \tag{2.12}$$

Since it is a linear function of Y_i, b_1 is a linear estimator and actually a weighted average of Y_i with w_i serving as weights (Kurt, 2000).

2.2.4.2 Unbiasedness

One intuitively desirable property of an estimator is unbiasedness or, that the expected value of the estimator equals the true population value ($E(b) = \beta$). If we could draw many samples and estimate the parameters for each sample, then the means of the estimator would equal the true population value in the unbiased case. That is, there is no systematic overestimation or underestimation of the true coefficients.

Because the properties of weights w_i:

$$\sum w_i = 0 \qquad \sum w_i X_i = 1$$

$$\begin{aligned} b_1 &= \sum w_i(\beta_0 + \beta_1 X_1 + \varepsilon_i) = \beta_0 \sum w_i + \beta_1 \sum w_i X_i + \sum w_i \varepsilon_i \\ &= \beta_1 + \sum w_i \varepsilon_i \end{aligned} \tag{2.13}$$

$$E(b_1) = \beta_1 + \sum w_i E(\varepsilon_i) \tag{2.14}$$

Since w_i is non-stochastic, they can be treated as constant. Since $E(\varepsilon_i) = 0$ by assumption obtain

$$E(b_1) = \beta_1 \tag{2.15}$$

Therefore it is said that b_1 is an unbiased estimator of β_1 (Kurt, 2000).

2.2.4.3 Best

The meaning of best estimator is that the least square estimator has minimum variance. There are many linear unbiased estimators for *b*, but the least square estimator is the most efficient by reason of having minimum variance (Hanushek & Jackson, 1977). It was given in equation 2.12 that

$$b_1 = \sum w_i Y_i$$

where $w_i = \dfrac{X_i - \overline{X}}{\sum \left(X_i - \overline{X}\right)^2}$. Let us now define an alternative linear estimator of β_1 as follows:

$$b_1^* = \sum k_i Y_i \qquad (2.16)$$

where k_i are also weights, not necessarily equal to w_i.

$$E\left(b_1^*\right) = \sum k_i E\left(Y_i\right) = \sum k_i \left(\beta_0 + \beta_1^* X_i\right) = \beta_0 \sum k_i + \beta_1^* \sum k_i X_i \qquad (2.17)$$

Now, for β_1^* to be unbiased, these conditions must be satisfied: $\sum k_i = 0$ $\sum k_i X_i = 1$

Also, we may write

$$\begin{aligned}
\mathrm{var}(b_1^*) &= \mathrm{var}\sum k_i Y_i = \sum k_i^2 \, \mathrm{var} Y_i \qquad (\mathrm{var} Y_i = \mathrm{var}\,\varepsilon_i = \sigma^2) \\
&= \sigma^2 \sum k_i^2 = \sigma^2 \sum \left[k_i - \frac{X_i}{\sum X_i^2} + \frac{X_i}{\sum X_i^2} \right]^2 \\
&= \sigma^2 \sum \left[k_i - \frac{X_i}{\sum X_i^2} \right]^2 + \sigma^2 \frac{\sum X_i^2}{\left(\sum X_i^2\right)^2} \qquad (2.18) \\
&\quad + 2\sigma^2 \sum \left[k_i - \frac{X_i}{\sum X_i^2} \right] \left[\frac{X_i}{\sum X_i^2} \right] \\
&= \sigma^2 \sum \left[k_i - \frac{X_i}{\sum X_i^2} \right]^2 + \sigma^2 \left[\frac{1}{\sum X_i^2} \right]
\end{aligned}$$

Since the last term is constant, the variance of b_1^* can be minimized only by manipulating the first term. So, if we let,

$$k_i = \frac{X_i}{\sum X_i^2} \qquad (2.19)$$

then

$$\mathrm{var}(b_1^*) = \frac{\sigma^2}{\sum X_i^2} = \mathrm{var}(b_1) \qquad (2.20)$$

In words, with weights $k_i = w_i$, which are the least squares weights, the variance of the linear estimator b_1^* is equal to the variance of the least squares estimator b_1; otherwise $\mathrm{var}(b_1^*) > \mathrm{var}(b_1)$. It means, b_1 has a minimum variance (Kurt, 2000).

2.3 Explanatory Power of a Multiple Regression Model

Independent variables explain the behavior of the dependent variable. By linear function of the independent variables, it is possible to find the variability in the dependent variable. A measure of the proportion of the variability in the dependent variable has been developed and named multiple coefficient of determination and denoted by the symbol R^2.

Error sum of squares was given in equation 2.6. Regression sum of squares

$$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 \qquad (2.21)$$

Total sum of Squares

$$SST = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (2.22)$$

$$SST = SSR + SSE \qquad (2.23)$$

Total sample variability = Explained variability + Unexplained variability

Since the coefficient of determination is the proportion of the total sample variability which is explained by the regression model,

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad 0 \le R^2 \le 1 \qquad (2.24)$$

By the way, when additional independent variables are added to a multiple regression model, the explained sum of squares (*SSR*) will increase even if the additional independent variable is not an important variable. In such a case, the increased value of R^2 would be misleading and it is acceptable to use adjusted coefficient of determination which is defined as

$$\begin{aligned} R^2_{adj} &= 1-\left(\frac{(n-1)}{n-(k+1)}\right)\left(\frac{SSE}{SST}\right) \\ &= 1-\left(\frac{(n-1)}{n-(k+1)}\right)\left(1-R^2\right) \end{aligned} \qquad (2.25)$$

where n is sample size and k is the number of regressors. In a multiple linear regression model, adjusted R square measures the proportion of the variation in the dependent variable accounted for by the independent variables. Unlike R square, adjusted R square allows for the degress of freedom associated with the sums of the squares. Therefore, even though the residual sum of squares decreases or remains the same as new explanatory variables are added, the residual variance does not. For this reason, adjusted R square is generally considered to be a more accurate goodness-of-fit measure than R square. The adjusted R^2 provides a better comparison between multiple regression models with different numbers of independent variables (Mendenhall & Sincich, 2003).

2.4 Model Building

Model building is an important issue, since writing a model will provide a good fit to a set of data and will give good estimates of the mean value of Y and good predictions of future values of Y for given values of the independent variables.

Researchers often collect a data set with a large number of independent variables, each of which is a potential predictor of some dependent variable, Y. When it is wanted that to build a multiple regression model, the problem of deciding which X's in a large set of independent variables to include in the model is common. Therefore, using variable selection methods is necessary in order to provide good fit of data and good estimates of parameters (Jobson, 1991).

2.4.1 Variable Selection Methods

In exploratory studies, an algorithmic method for searching among models can be informative, if the results are used warily. To make the model useful for predictive purposes it may be wanted the model to include as many *X*'s as possible so that reliable fitted values can be determined. However, on the other hand, because of the costs involved in obtaining information on a large number of *X*'s and subsequently monitoring them, it may be wanted the equation to include as few *X*'s as possible. Further more, the selection process becomes more challenging as the number of independent variables increases, because of the rapid increase in possible effects and interactions. There are two competing goals: The model should be complex enough to fit the data well, but simpler models are easier to interpret.

On the other hand, on reducing the model the error term may change to reflect the omission of important independent variables. If important independent variables are deleted mistakenly from the model, their effects are included in the model error terms. In this instance coefficient estimates may change impressively and reflect biases incurred by eliminating these variables (Mason, Gunst, & Hess, 2003).

However, there is no unique statistical procedure to reduce the number of independent variables to be used in the final model, and personal judgment will be a necessary part of any of the statistical methods discussed (Chatterjee & Hadi, 2006).

2.4.1.1 All Possible Regressions Procedure

Variable selection techniques have been developed in the literature for the purpose of identifying important independent variables. The most popular of these procedures are those that consider all possible regression models given the set of potentially important predictors. Such a procedure is commonly known as an all possible regressions selection procedure. The techniques differ with respect to the criteria for selecting the best subset of variables.

The purpose of the all possible regression approach is to identify a small group of regression models that are "good" according to a specified criterion so that a detailed examination can be made of these models, leading to the selection of the final regression model to be employed (Mendenhall & Sincich, 2003).

Different criteria for comparing the regression models may be used with the all possible regressions selection procedure. Four criteria are widely used in practice: R^2, *MSE*, C_p, PRESS.

R^2 or SSE Criterion: R^2 criterion calls for the use of the coefficient of multiple determination R^2 in order to identify several "good" subsets of *X* variables, in other words, subsets for which R^2 is high. The R^2 criterion, as shown in the equation (2.24), is equivalent to using the error sum of squares *SSE* as the criterion. With the *SSE* criterion, subsets for which *SSE* is small are considered "good". Since the denominator *SST* is constant for all possible regression models, R^2 varies inversely with *SSE*.

It is known that *SSE* can never increase as additional *X* variables are included in the model. Hence, R^2 will be a maximum when all potential *X* variables are included in the regression model. The aim at using the R^2 criterion is to find the point where adding more *X* variables is not worthwhile because it leads to a very small increase in R^2. Often, this point is reached when only a limited number of *X* variables is included in the regression model. Clearly, the determination of where diminishing returns set in is a judgmental one. In practice, the best model found by the R^2 criterion will rarely be the model with the largest R^2 (Mendenhall & Sincich, 2003).

Adjusted R^2 or MSE (Mean Square Error) Criterion: It was mentioned that since R^2 does not take account of the number of parameters in the regression model and since R^2 can never decrease as the number of potential *X* variables increases, the adjusted coefficient of multiple determination R^2_{adj} has been suggested as an alternative criterion. The equation of R^2_{adj} in (2.25) can be written as

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-(k+1)} \right) \frac{SSE}{SST} = 1 - \frac{MSE}{SST \big/ n-1} \tag{2.26}$$

where *n* is sample size and *k* is the number of regressors. This coefficient takes the number of parameters in the regression model into account through the degrees of freedom. It can be seen from the equation that R^2_{adj} increases if and only if *MSE* decreases since *SST* / (*n* – 1) is fixed for the given *Y* observations. Hence, R^2_{adj} and *MSE* provide equivalent information. We shall consider here the criterion *MSE*, again showing the number of the parameters in the regression model as a subscript of the criterion. The smallest *MSE* for a given number of parameters in the model can, indeed, increase as k increases. This occurs when the reduction in *SSE* becomes so small that it is not sufficient to offset the loss of an additional degree of freedom. Users of the *MSE* criterion seek to find a few subsets for which *MSE* is at the minimum or so close to the minimum that adding more variables is not worthwhile.

C_p ***Criterion:*** This criterion is a function of the mean squared error concerned with the total mean squared error (*TMSE*) of the *n* fitted values for each subset regression model. The mean squared error concept involves the total error in each fitted value:

$$TMSE = E\left\{ \sum_{i=1}^{n} \left[\hat{Y}_i - E(Y_i) \right]^2 \right\} = \sum_{i=1}^{n} \left[E(\hat{Y}_i) - E(Y_i) \right]^2 + \sum_{i=1}^{n} Var(\hat{Y}_i) \tag{2.27}$$

The objective is to compare the *TMSE* for the subset regression model with σ^2, the variance of the random error for the true model, using the ratio

$$\Gamma = \frac{TMSE}{\sigma^2}$$

Small values of Γ imply that the subset regression model has a small total mean square error relative to σ^2. Unfortunately, both *TMSE* and σ^2 are unknown, but a sample estimates of these quantities can be used. It can be shown that a good estimator of Γ is given by

$$C_p = \frac{SSE_p}{MSE(X_1, \cdots, X_{p-1})} - (n - 2p) \tag{2.28}$$

where *n* is the number of observations and *p* is the number of estimated parameters, SSE_p is the *SSE* for the estimated model, $MSE(X_{1,} \cdots, X_{p-1})$ is an unbiased estimator of σ^2 (Neter, Kutner, Nachtsheim & Wasserman, 1996).

In using the C_p criterion, it is sought to identify the subsets of *X* variables for which the C_p value is small and the C_p value is near *p*. Subsets with small C_p values have a small total mean squared error, and when the C_p value is also near *p*, the bias of the regression model is small (Mendenhall & Sincich, 2003).

PRESS Criterion: The PRESS (prediction sum of squares) criterion is a measure of how well the use of the fitted values for a subset model can predict the observed responses Y_i (Neter, Kutner, Nachtsheim & Wasserman, 1996).

$$PRESS = \sum_{i=1}^{n} \left[Y_i - \hat{Y}_{(i)} \right]^2 \tag{2.29}$$

where $\hat{Y}_{(i)}$ denotes the predicted value for the i^{th} observation obtained when the regression model is fit with the data point for the i^{th} observation omitted (or deleted) from the sample. Thus, the candidate model is fit to the sample data *n* times, each time omitting one of the data points and obtaining the predicted value of Y for that data point. Since small differences $Y_i - \hat{Y}_{(i)}$ indicate that the model is predicting well, a model with a small PRESS is chosen (Mendenhall & Sincich, 2003).

2.4.1.2 Stepwise Regression Procedure

The stepwise regression procedure is probably the most widely used of the automatic search methods. This search method develops a sequence of regression models, at each step adding or deleting an X variable. The criterion for adding or deleting an X variable can be stated equivalently in terms of error sum of squares reduction, coefficient of partial correlation, t statistic, or F statistic (Neter, Kutner, Nachtsheim, & Wasserman, 1996).

The forward selection method starts with an equation containing no independent variables, just constant term, and adds terms consecutively until further additions do not improve the fit (Agresti, 2002). At any stage in the selection process, forward selection method adds the variable which has the highest partial correlation, increases R^2 the most, and gives the largest absolute t or F statistic (Christensen, 2002). The minimum P-value for testing the term in the model is also a sensible criterion for adding variable (Agresti, 2002).

The backward elimination procedure starts with the full equation and drops one variable at every stage. The variables are dropped based on their support to the reduction of error sum of squares. This has the same meaning with deleting the variable which has the smallest t-test in the equation. Assuming that there are some variables which have insignificant t-tests, the procedure drops the variable with the smallest insignificant t-test. The procedure is terminated when all the t-tests are significant or all variables which have insignificant t-tests have been deleted (Chatterjee & Hadi, 2006).

The stepwise method is essentially a composite of the forward and backward methods. In this method, a variable which has entered in the earlier stages of selection may be eliminated at later stages.

An essential difference between automatic search procedures and the all possible regressions procedure is that the automatic search procedures end with the identification of a single regression model as “best”. With the all possible regressions procedure, on the other hand, several regression models can be identified as good for final consideration. The identification of a single regression model may hide the fact that several other regression models may also be “good”. Finally, the goodness of a regression model can only be established by a thorough examination using a variety of diagnostics (Neter, Kutner, Nachtsheim, & Wasserman, 1996).

CHAPTER THREE
OMITTED VARIABLES

3.1 Introduction

In ordinary regression models, the consistency of standard least squares estimators depends on the assumption that the explanatory variables are uncorrelated with the error term. This assumption is prone to be violated, especially when important explanatory variables are excluded from the model. Often, such omissions are unavoidable due to the inability to collect necessary variables for the model. The consequence is not only possible for estimating the effects of important variables, but also the estimates for other effects in the model may be biased and thus misleading. This problem is often called an omitted variable bias (Kim & Frees, 2006).

Most regressions conducted by economists can be critiqued for omitting some important independent variables which may cause the estimated relationships to change. Why some variables are omitted? Variables are often omitted when they cannot be measured, when it is impossible to sufficiently specify the list of potential additional variables, when it is impossible to model how the omitted variables interact with the included variables, and when the influence of the omitted variables are not known (Leightner & Inoue, 2007).

When significant independent variables are omitted from the model, the least squares estimates will usually be biased and the usual inferential statements from hypothesis tests or confidence intervals can be seriously misleading. Thus, omitted variable is a serious problem however, an omitted variable is only a problem under a specific set of circumstances. If the regressor is correlated with a variable that has been omitted from the analysis but that determines the dependent variable in part, then the OLS estimator will have omitted variable bias (Stock & Watson, 2003).

3.2 Omitted Variable Bias

The omission from a regression of some variables that affect the dependent variable may cause an omitted variable bias. Every omission doesn’t always result biassedness. Omitted variable bias occurs when two conditions come true: first, the omitted variable is a determinant of the dependent variable and second, the omitted variable is correlated with the included variables (Stock & Watson, 2003).

If a variable that is related to the dependent variable but uncorrelated with any measured independent variable is omitted, the result is a poorer fitting model with a larger error term. The regression coefficients for the measured independent variables, however, are not biased just by the omission of such a variable. In contrast, if the omitted variable is related to the dependent variable and correlated with a measured independent variable, then it can be said that the regression coefficient for the measured independent variable can be biased (Sackett, Laczo, & Lippe, 2003). Since it is impossible to include all relevant variables in a regression equation, omitted variable bias is unavoidable; however it is possible to mitigate this bias (Clarke, 2005).

The problem arises because any omitted variable becomes part of the error term, and the result may be a violation of the assumption necessary for the minimum *SSE* criterion to be an unbiased estimator. This assumption is the first least squares assumption which is $E(\varepsilon_i | X_i) = 0$ incorrect. It was described in chapter two that the error term ε_i in the linear regression model with a single regressor represents all variables, other than X_i, that are determinants of Y_i. If one of these other variables is correlated with X_i, this means that the error term (which contains this variable) is correlated with X_i. In other words, if an omitted variable is a determinant of Y_i, then it is the error term, and if it is correlated with X_i, then the error term is correlated with X_i. Since ε_i and X_i are correlated, the conditional mean of ε_i given X_i is nonzero. This correlation therefore violates the first least squares assumption which is given in Section 2.2.3.1, and this causes a serious problem which is the OLS estimator has omitted variable bias. This bias does not vanish even in very large samples, and the OLS estimator is inconsistent (Stock & Watson, 2003).

The omitted variable bias formula is a very useful tool for judging the impact on regression analysis of omitting important influences on behavior which are not observed in the data set. In small sample form, the bias formula was developed and popularized by Theil (1957, 1971), and has been used extensively in empirical research (Stoker, 1983).

To visualize the omitted variable bias, suppose that the model with two independent variables is the true model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{3.1}$$

However, suppose again instead that *Y* is regressed on X_1 alone, with X_2 omitted because of being unobservable. Then, the term $\beta_2 X_2$ is moved into the error term and the estimated model is

$$\hat{Y} = b_0 + b_1 X_1 \tag{3.2}$$

and therefore

$$Y = b_0 + b_1 X_1 + e^* \tag{3.3}$$

where e^* is the error term and equals to $(\beta_2 X_2 + \varepsilon)$ (Ramsey, 1969). As before ε is uncorrelated with X_1, but if X_2 is correlated with X_1, the error term $(\beta_2 X_2 + \varepsilon)$ will be correlated with the included variable X_1. Therefore, the least square assumption will be violated and as a consequence of this violation, the OLS estimator will be biased and inconsistent, if X_2 is correlated with X_1. Unless X_2 is correlated with X_1, however, there will be no correlation between the error term and the independent variable X_1, therefore the bias will not arise from omitting the variable X_2.

The property of being unbiasedness, mentioned in the previous chapter, means that the expected value of the estimator equals the true population value. Therefore, it is investigated whether $E(b_1) = \beta_1$ when the model has omitted variable. If the true model is as equation (3.1) and we estimate as equation (3.2), then the least square estimator is (Williams, 2008)

$$\begin{aligned} b_1 &= \frac{\hat{C}ov(X_1, Y)}{\hat{V}(X_1)} \\ &= \frac{\hat{C}ov(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon)}{\hat{V}(X_1)} \\ &= \frac{\hat{C}ov(X_1, \beta_0) + \beta_1 \hat{C}ov(X_1, X_1) + \beta_2 \hat{C}ov(X_1, X_2) + \hat{C}ov(X_1, \varepsilon)}{\hat{V}(X_1)} \\ &= \frac{0 + \beta_1 \hat{V}(X_1) + \beta_2 \hat{C}ov(X_1, X_2) + 0}{\hat{V}(X_1)} \\ &= \beta_1 + \beta_2 \frac{\hat{C}ov(X_1, X_2)}{\hat{V}(X_1)} \end{aligned} \tag{3.4}$$

$$E(b_1) = \beta_1 + \beta_2 \frac{\sigma_{12}}{\sigma_1^2} \tag{3.5}$$

If the omitted X_2 is correlated with X_1, then the estimate of β_1 will be biased. Because it now reflect not only the effect of itself but also partly the effects of the omitted variable. But, if the X_1 and X_2 are uncorrelated, then omitting one does not result in biased estimates of the effect of the other. Furthermore, if β_2 = 0, this means that the model is not mis-specified and X_2 does not belong in the model because it has no effect on Y (Williams, 2008).

The amount of bias in the estimation with omitted X_2 is $\beta_2 \frac{\sigma_{12}}{\sigma_1^2}$. As it can be seen, β_1 may increase or decrease according as the sign of β_2 and sign of the value of covariance. The direction of the bias, in other words whether b_1 tends to over or under estimate β_1 is solely a function of the signs of β_2 and σ_{12}. If both are positive or both are negative, b_1 will be biased upward; if one is negative and one is positive, b_1 will be biased downward.

It is straightforward to deduce the directions of bias when there is a single included variable and one omitted variable. It is important to note, furthermore, that if more than one variable is included, then the terms in omitted variable formula involve multiple regression coefficients, which themselves have the signs of partial, not simple, correlations (Greene, 2003). The omitted variable bias formula for the models that have three independent variables is given by Hanushek and Jackson (1977). The proof implies that if the true model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \tag{3.6}$$

and we estimate

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 \tag{3.7}$$

and therefore

$$Y = b_0 + b_1 X_1 + b_2 X_2 + e^* \qquad \text{where } e^* = \beta_3 X_3 + \varepsilon \tag{3.8}$$

The least square estimators

$$b_1 = \frac{V_2 C_{1Y} - C_{12} C_{2Y}}{V_1 V_2 - C_{12}^2}$$
$$= \frac{(1/N)\sum_{i=1}^{N}\left\{\left[V_2\left(X_{i1}-\overline{X}_1\right)-C_{12}\left(X_{i2}-\overline{X}_2\right)\right]\left(Y_i-\overline{Y}\right)\right\}}{V_1 V_2 - C_{12}^2} \tag{3.9}$$

where V_1 and V_2: the variances of X_1 and X_2; C_{ij}: the covariances of the variables i^{th} and j^{th}. From the true model for Y and from averaging the Y_i over the sample, it is known that

$$Y_i - \overline{Y} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i - \left(\beta_0 + \beta_1 \overline{X}_1 + \beta_2 \overline{X}_2 + \overline{\varepsilon}\right)$$
$$= \beta_1\left(X_{i1}-\overline{X}_1\right) + \beta_2\left(X_{i2}-\overline{X}_2\right) + \left(\varepsilon_i - \overline{\varepsilon}\right) \tag{3.10}$$

where $\overline{\varepsilon}$ is the mean of all error terms implicit in the sample. By substitution,

$$b_1 = \frac{1}{ND}\left\{\begin{array}{l} \sum_{i=1}^{N}\left[V_2\left(X_{i1}-\overline{X}_1\right)-C_{12}\left(X_{i2}-\overline{X}_2\right)\right] \\ \left[\beta_1\left(X_{i1}-\overline{X}_1\right)+\beta_2\left(X_{i2}-\overline{X}_2\right)+\left(\varepsilon_i-\overline{\varepsilon}\right)\right] \end{array}\right\}$$

$$= \frac{1}{D}\left\{\begin{array}{l} \frac{\beta_1}{N}V_2\sum\left(X_{i1}-\overline{X}_1\right)^2 - \frac{\beta_1}{N}C_{12}\sum\left(X_{i2}-\overline{X}_2\right)\left(X_{i1}-\overline{X}_1\right) + \\ \frac{\beta_2}{N}V_2\sum\left(X_{i1}-\overline{X}_1\right)\left(X_{i2}-\overline{X}_2\right) - \frac{\beta_2}{N}C_{12}\sum\left(X_{i2}-\overline{X}_2\right)^2 + \\ \frac{1}{N}\sum\left[V_2\left(X_{i1}-\overline{X}_1\right)-C_{12}\left(X_{i2}-\overline{X}_2\right)\right]\left(\varepsilon_i-\overline{\varepsilon}\right) \end{array}\right\} \tag{3.11}$$

where $D = V_1 V_2 - C_{12}^2$. The first summation can be written as $\beta_1 V_2 (1/N)\sum\left(X_{i1}-\overline{X}_1\right)^2 = \beta_1 V_2 V_1$.

Similar treatment the succeeding terms gives

$$b_1 = \frac{\beta_1 V_2 V_1 - \beta_1 C_{12}^2 + \beta_2 V_2 C_{12} - \beta_2 C_{12} V_2}{D} + \frac{V_2 C_{1\varepsilon} - C_{12} C_{2\varepsilon}}{D}$$

$$= \frac{\beta_1 \left(V_1 V_2 - C_{12}^2\right)}{D} + \frac{V_2 C_{1\varepsilon} - C_{12} C_{2\varepsilon}}{D} \qquad (3.12)$$

$$b_1 = \beta_1 + \frac{V_2 C_{1\varepsilon} - C_{12} C_{2\varepsilon}}{D}$$

Similarly,

$$b_2 = \beta_2 + \frac{V_1 C_{2\varepsilon} - C_{12} C_{1\varepsilon}}{D} \qquad (3.13)$$

Since in this case the error term equals e^*, the equations (3.12) and (3.13) change as below

$$b_1 = \beta_1 + \frac{V_2 C_{1e^*} - C_{12} C_{2e^*}}{V_1 V_2 - C_{12}^2} \qquad (3.14)$$

$$b_2 = \beta_2 + \frac{V_1 C_{2e^*} - C_{12} C_{1e^*}}{V_1 V_2 - C_{12}^2} \qquad (3.15)$$

Substituting $e^* = \beta_3 X_3 + \varepsilon$ into the covariance expressions involving e^* gives

$$C_{1e^*} = \frac{1}{T}\sum \left(X_1 - \overline{X}_1\right)\left(e^* - \overline{e}^*\right) = \frac{1}{T}\sum \left(X_1 - \overline{X}_1\right)\left(\beta_3 X_3 + \varepsilon - \beta_3 \overline{X}_3 - \overline{\varepsilon}\right)$$

$$= \frac{1}{T}\beta_3 \sum \left(X_1 - \overline{X}_1\right)\left(X_3 - \overline{X}_3\right) + \frac{1}{T}\sum \left(X_1 - \overline{X}_1\right)\left(\varepsilon - \overline{\varepsilon}\right)$$

$$= \beta_3 C_{13} + C_{1\varepsilon}$$

$$C_{1e^*} = \beta_3 C_{13} + C_{1\varepsilon} \tag{3.16}$$

$$C_{2e^*} = \beta_3 C_{23} + C_{2\varepsilon} \tag{3.17}$$

Taking the expected value of b_1 and b_2, assuming fixed *X* and $E(\varepsilon) = 0$

$$\begin{aligned} E(b_1) &= \beta_1 + \beta_3 \left(\frac{V_2 C_{13} - C_{12} C_{23}}{V_1 V_2 - C_{12}^2} \right) + E\left[\frac{V_2 C_{1\varepsilon} - C_{12} C_{2\varepsilon}}{V_1 V_2 - C_{12}^2} \right] \\ &= \beta_1 + \beta_3 b_{31} \end{aligned} \tag{3.18}$$

$$\begin{aligned} E(b_2) &= \beta_2 + \beta_3 \left(\frac{V_1 C_{23} - C_{12} C_{13}}{V_1 V_2 - C_{12}^2} \right) + E\left[\frac{V_1 C_{2\varepsilon} - C_{12} C_{1\varepsilon}}{V_1 V_2 - C_{12}^2} \right] \\ &= \beta_2 + \beta_3 b_{32} \end{aligned} \tag{3.19}$$

where

$$b_{31} = \frac{(r_{31} - r_{21} r_{32})}{1 - r_{21}^2} \sqrt{\frac{V_3}{V_1}} \qquad \text{and} \qquad b_{32} = \frac{(r_{32} - r_{21} r_{31})}{1 - r_{21}^2} \sqrt{\frac{V_3}{V_2}}$$

where r_{ij} mean the correlations between sample values. As a result of this proof, it can be seen that the models that have three independent variables may have the omitted variable bias.

The biases in the estimation with omitted X_3 are $\beta_3 b_{31}$ and $\beta_3 b_{32}$. As it is seen from the formula, to obtain the direction of bias can be difficult. This is because X_1, X_2 and X_3 can all be pair wise correlated. The direction of the bias, in other words whether b_1 and b_2 tend to over or under estimate of β_1 and β_2 is solely a function of the signs of β_3 and of b_{31} and b_{32}. If both are positive or both negative, b_1 (or b_2) will be over estimated; if one is negative and one is positive, b_1 (or b_2)

will be under estimated. Hence, the direction of bias in b_1 and b_2 does not have to be the same.

3.3 Detection of Omitted Variables with RESET Test

Detection of omitted variables plays an important role in specification analyses. Several techniques are developed for this purpose. One of the oldest specification tests for linear regression models, that is still widely used, is Regression Equation Specification Error Test (RESET), which was originally proposed by Ramsey (1969) and is known as the Ramsey RESET test (Clements and Hendry, 2002). This test is primarily a test designed to detect omitted variables and is a model misspecification test.

Ramsey's RESET Test tests the hypothesis that no relevant independent variables have been omitted from the regression model (Watson, 2002). Even if the Ramsey test signals that some variable(s) are omitted, it obviously doesn't tell which ones are omitted. Besides this, nonetheless gave satisfactory values for all of the more traditional test criteria such as goodness of fit, high t-ratios and correct coefficient signs and test for first order autocorrelation (Evans, 2002).

Furthermore, the RESET test is not only used to detect omitted variables, but also is used to check for the following types of errors, except for omitted variables:

- Nonlinear functional forms
- Simultaneous-equation bias
- Incorrect use of lagged dependent variables (Evans, 2002)

The idea is that the various powers of the fitted values will reveal whether misspecification exists in the original equation by determining whether the powers of the fitted values are significantly different from zero. More specifically, in developing a misspecification test, Ramsey recommends adding a number of additional terms to the regression model and then testing the significance of these. It means that it is necessary to include in the regression model some functions of the regressors, on the basis that, if the model is misspecified, the error term would capture these variables either directly or indirectly through other variables omitted from the regression. Then, a test for the significance of these additional variables is used. It follows from the Milliken-Graybill Theorem (1970) that the usual test statistic will be exactly *F*-distributed with *k* and *(n-k-r-*1*)* degrees of freedom under the null hypothesis, if the errors are independent, homoskedastic, and normally distributed . If these additional variables are found to be significant, then it is said that the model is misspecified and some variables are omitted.

The test is developed as follows. Suppose that the standard linear model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \tag{3.20}$$

Ramsey now proposes the creation of a vector, defined as

$$\left(\hat{Y}_i^2, \hat{Y}_i^3, \hat{Y}_i^4, \ldots, \hat{Y}_i^k\right)$$

where the value of *k* is chosen by the researcher, and suggests that the powers of $\hat{Y}$ be included in the equation in addition to all the other X_i terms that are already in the regression (Evans, 2002).

If the true model is as equation (3.6), and the estimated model is as equation (3.7), then by adding powers of the fitted values of *Y* to the original model, a new model is estimated

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \delta_1 \hat{Y}^2 + \delta_2 \hat{Y}^3 + u \tag{3.21}$$

Then, in order to test the significance of these additional variables, the following hypotheses are constructed

$H_0 : \delta_1 = 0,\ \delta_2 = 0$
$H_1 : \delta_1 \neq 0,\ \delta_2 \neq 0$

The meanings of these hypotheses are:

H_0 : the model has no omitted variable

H_1 : the model has omitted variable(s)

Test statistic:

$$F = \frac{(SSE_{old} - SSE_{new})/k}{SSE_{new}/(n-k-r-1)} \approx F(\alpha, k, n-k-r-1) \qquad (3.22)$$

where k is the number of new regressor and r is the number of old regressor and SSE_{old} is the sum of squared error for the estimated model, and SSE_{new} is the sum of squared error for the model added powers of the fitted values of Y (Newbold, Carlson & Thorne, 2003).

F-test provides an exact test for the null hypothesis (Verbeek, 2004). Decision rule implies that if the calculated *F* is greater than the *F* given by the critical value of *F* for some desired rejection probability (e.g. 0.05), the null hypothesis is rejected. Rejection of the null hypothesis implies the original model is inadequate and can be improved.

Consequently, if the model can be significantly improved by artificially including powers of the predictions of the model, then the original model must have been inadequate and some important variables must have been added to the model (Newbold, Carlson & Thorne, 2003).

RESET test is available in some software packages as STATA and R. STATA applies RESET test via the "ovtest" or "ovtest, rhs" commands after a reg command. The ovtest which is standing for "ommited variables test" uses the second through fourth powers of the fitted values. The rhs option uses the second through fourth powers of independent variables. Both the RESET test with powers of the fitted values of approval and the test with the powers of the independent variables produce significant *F* tests for specification error. Furthermore, R applies RESET test via the “reset” or “resettest” commands and uses the second and third powers of the independent variables or fitted values or first principal component.

3.4 Methods for Dealing with Omitted Variable Bias

There are two types of methods to deal with the omitted variable bias which are theoretical methods and practical methods.

3.4.1 Theoretical Methods

How the analyst should proceed can be found out by looking at the errors of models with omitted variable. The terms b_{31} and b_{32} in equations (3.18) and (3.19) are the functions of the characteristics of the particular sample. Although X_3 is not observed and included in the data set, each observation has some implicit values for this variable associated with it. The variance of these implicit values for X_3 affects the values of b_{31} and b_{32} for a given set of values for X_1 and X_2. Since the terms b_{31} and b_{32} refer to the sample used for the estimation, it is possible to reduce b_{31} and b_{32} through appropriate choice of sample. If it can be found a sample where X_3 does not vary which means $V_3 = 0$, then b_{31} and b_{32} will be zero, and therefore the bias will be removed, completely. Thus, selection of the sample is an important issue.

By the way, it can be understood that the problems of specification are related to the size of β_3. The biases in the estimation with omitted X_3 are $\beta_3 b_{31}$ and $\beta_3 b_{32}$. Thus, the biases become more severe as the excluded variable becomes more important in explaining *Y*, for example the biases become larger in absolute magnitude of β_3. Choosing independent variables to include to the model is a very critical point for proper specification. A priori knowledge based upon theory, past empirical results form the basis for making decisions on the size of different coefficients for variables omitted from models (Barreto & Howland, 2006).

The correlations between the unmeasured sample values of this omitted variable and the included variables, denoted by r_{31} and r_{32}, affect the values of b_{31} and b_{32} for a given set of values for X_1 and X_2. Therefore, one method of reducing bias is to reduce the relationships in the sample between the omitted and the included variables. It means this method involves collecting observations in which the excluded variable is uncorrelated with the included variables. In such a sample r_{31} and r_{32} are equal to zero and this makes b_{31} and b_{32} zero and in this manner it was provided unbiased estimates of β_1 and β_2. The only difficulty with this procedure is that if the included independent variables are at all correlated, the excluded variable must be randomized with respect to all the exogenous variables or all the coefficients will be biased, regardless of the correlation between the excluded variable and any particular *X*. In real data sets, it is hard enough to find situations where an omitted variable is uncorrelated with any included variable. This is the focal point for physical science research since laboratory experiments can be designed to reduce or eliminate the correlations with excluded variables from the experiment. Social scientists, however, do not often have the luxury of experimental design. Hence, they can not usually use this method.

The remedy for these misspecification problems is obvious, but not necessarily easy. The excluded variable can either be included or a sample can be collected in which the covariance between included and omitted variables is zero, either because they are uncorrelated or because the excluded variable has no variance. However, each of these solutions requires that the misspecification be recognized prior to the collection of the data. In most real world applications, the misspecification arises because researchers failed to recognize the importance of a variable, not because they were unable to obtain a measure for the excluded variable or a sample where it was uncorrelated with included variables. This will be particularly true in social science areas that do not have a well-developed priory theory. Consequently, in some areas as social science the likelihood of misspecification is increased because there is little formal theory to guide the researcher in selecting variables and ascertaining what needs to be held constant. The researcher then must be particularly careful in selecting the original variables.

One of the most important implications of the theoretical development is that the inclusion of the important variables is essential, even if one is not interested in the estimated effects of all of the variables. In order to arrive at good estimates of the parameters of interest, it may be necessary to include other variables of lesser usefulness in the given problem. Recognition of the significance of a variable in a behavioral relationship does not necessarily imply that the analyst can or wishes to interpret its coefficient, only which one wishes to avoid biasing the coefficients of real interest (Hanushek & Jackson, 1977)

3.4.2 Practical Methods

The danger of omitted variables has been a recurrent issue in the social sciences. Boardman and Murnane (1979) underscored the potential bias and inconsistency of the ordinary least squares (OLS) estimators, and promoted a panel data approach. Ehrenberg and friends incorporated instrumental variable approaches for the analysis of the High School and Beyond (Ehrenberg & Brewer, 1994) and the National Education Longitudinal Study of 1988 (Ehrenberg, Goldhaber, & Brewer, 1995). Several other studies have considered a variety of procedures to address problems related to omitted variables.

Some methods in order to prevent the problem of omitted variables are presented in the following sections.

- Proxy Variable
- Instrumental Variable
- Panel Data
- Reiterative Truncated Projected Least Squares

3.4.2.1 Proxy Variable

Some variables, such as socioeconomic status, and quality education, and ability are so vaguish that it may be impossible even in principle to measure them. Others might be measurable, but require so much time and energy that in practice they have to be abandoned. Sometimes you are frustrated because you are using survey data collected by someone else, and an important variable has been omitted. Sometimes another variable is used in place of the omitted variable. Such a measurement variable is called a proxy variable.

Because of these circumstances, if the researcher cannot obtain the variable of interest, then he must search whether proxy variables are available. When another variable, which's observations are obtainable and highly correlated with the omitted variable and this variable is thus available as a proxy (McCallum, 1972).

When only proxy variables are available for a subset of the independent variables, one must choose between the strategies of including the set of proxy variables in the regression or omitting them. A number of reasons show that it is usually a good idea to use a proxy variable to stand in for the missing variable, rather than omitting it entirely. It is shown that the bias of the estimates of the coefficients of the observable variables obtained by omitting the unobservable variable is always greater than the bias resulting from using proxy. In fact, it is better to use even a poor proxy than to use none at all and omit the variable (Wickens, 1972).

There are two good reasons for tring to find a proxy. First, the variable simply can be left out, then the regression is likely to suffer from omitted variable bias but the statistical tests will be invalidated. Second the results from your proxy regression may indirectly reveal the influence of the omitted variable.

As it was described, omitted variable bias can be eliminated or at least mitigated, if a proxy variable is avaliable for the excluded variable. Suppose that the true model is as (3.6), where X_3 is unobservable. Suppose that X_3^* is available as a proxy for X_3. The proxy variable X_3^* must have some relationship with X_3. Now, suppose that the relationship is written as:

$$X_3 = \delta_0 + \delta_1 X_3^* + v_3 \tag{3.23}$$

There are some conditions that the proxy variable X_3^* should satisfy (Byun, 2005). Because, it will solely give unbiased estimator. They are:

- ε is not correlated with X_1, X_2 and X_3. This is the standard assumption for the true model.

- ε is not correlated with X_3^*. Condition 1 and 2 imply that

  $$E(\varepsilon | X_1, X_2, X_3, X_3^*) = 0$$

- v_3 is not correlated with X_1, X_2 and X_3^*. This condition is necessary for “good” proxy variable.

  $$E(X_3 | X_1, X_2, X_3^*) = E(X_3 | X_3^*) = \delta_0 + \delta_3 X_3$$

- One another condition the proxy variable should satisfy is that the proxy variable should be redundant (sometimes called ignorable) in the structural equation (Wooldridge, 2002). Proxy variable X_3^* is irrelevant in the true model, once X_1, X_2 and X_3 have been included. It is X_3 that directly affects *Y*, not X_3^*. The most natural statement of redundancy of X_3^* is:

$$E\left(Y \middle| X_1, X_2, X_3, X_3^*\right) = E\left(Y \middle| X_1, X_2, X_3\right)$$

3.4.2.2 Instrumental Variable

The instrumental variable method is a way to consistently estimate the true coefficients of the regression model in spite of the endogenous variables which are the independent variables correlated with the error term, likely due to one or more omitted variables. Omitting a relevant variable causes endogeneity, because if an omitted variable is a determinant of Y_i, then it becomes a part of the error term, and if it is correlated with at least one of X_i, then the error term is correlated with *X*. Therefore this variable is called endogenous variable.

An instrumental variable, often defined by the letter *Z* in equations, is used as an “instrument” or “tool” to isolate the part of *X* that is correlated with the error term. Because if the information in *X* that is not correlated with the error term can be isolated, then this information can be used to obtain an unbiased estimate of true regression parameters. This method is particularly used when there are no satisfactory proxy variables for the omitted variables (Schreck, 2009).

A good instrumental variable, *Z*, has the following properties which are necessary for getting unbiased coefficient estimates (Wooldridge, 2002)

- Unlike a proxy variable, *Z* should be uncorrelated with the omitted variable. Therefore *Z* is independent of the error term, so that the instrument can isolate the variation in *X*. This property is known as “instrument exogeneity” (Schreck, 2009).

$$Cov(Z, \varepsilon) = 0$$

- *Z* is correlated with the endogenous variable *X*, hence the instrument can capture some of its variation. This property is also known as "instrument relevance" (Stock & Watson, 2003).

$$Cov(Z, X) \neq 0$$

- *Z* is strongly correlated, rather than weakly correlated, with the endogenous variable X .

If an instrument fails the first condition, it means that *Z* is correlated with the error term, the instrument is said to be an invalid instrument. If an instrument fails the second condition, the instrument is said to be an irrelevant instrument. The third condition fails when very low correlation exists between the instrument and the endogenous variable being instrumented; likewise the instrument is called a weak instrument (Cameron & Trivedi, 2005).

Instrumental variable estimator provides a way to obtain consistent parameter estimates. This method, widely used in econometrics and rarely used elsewhere, is conceptually difficult and easily misused. However this method can not always be applied, because necessary instruments may not always be available (Stock & Watson, 2003).

3.4.2.3 Panel Data

Panel data can be used to obtain consistent estimators in the presence of omitted variables (Gossy, 2008). A panel data set contains repeated observations over the same units (individuals, firms) collected over a number of periods. The availability of repeated observations on the same units allows economists to specify and estimate more complicated and more realistic models than a single cross-section or a single time series would do (Verbeek, 2004).

Time series is a data set containing observations on a single phenomenon observed over multiple time periods. In time series data, both the values and the ordering of the data points have meaning. Cross-sectional data is a data set containing observations on multiple phenomena observed at a single point in time. In cross-sectional data sets, the values of the data points have meaning, but the ordering of the data points does not. Panel data is, however, a data set containing observations on multiple phenomena observed over multiple time periods. Alternatively, the second dimension of data may be some entity other than time (Hsiao, 2003). Therefore, panel data are not only suitable to model or explain why individual units behave differently but also to model why a given unit behaves differently at different time periods (Verbeek, 2004).

The main idea of panel data models is to regard any unobserved factor affecting the dependent variable as consisting of two effects: those that are constant and those that vary over time (Gossy, 2008).

Panel data provide means to eliminate or reduce the omitted-variable bias through the various data transformations when the correlations between included explanatory variables and the random error terms follow certain specific patterns (Arminger, Clogg, & Sobel, 1995). In certain cases the availability of panel data can actually simplify the computation and inference (Cameron & Trivedi, 2005).

Panel data can reduce the effects of omitted variable bias, or in other words, estimators from a panel data set may be more robust to an incomplete model specification (Hsiao, 2003).

3.4.2.4 Reiterative Truncated Projected Least Squares

Traditional techniques for dealing with omitted variables use proxy variables or instrumental variables. However the correct use of proxy variables and instrumental variables involves knowing (1) how the omitted variable’s affect on the dependent variable should be modeled and (2) how the correlation between the instruments or proxies and the omitted variable should be modeled. This necessary knowledge is often impossible to obtain. By building on Branson and Lovell, Leightner created a new analytical technique named Reiterative Truncated Projected Least Squares (RTPLS) that produces reduced form estimations while greatly reducing the influence of omitted, unknown, and immeasurable variables. Unlike the use of proxies or instrumental variables, RTPLS does not require knowing how the omitted variable is correlated to the dependent variable and how the omitted variable is correlated with proxies or instruments (Leightner & Inoue, 2007).

3.5 The Relationship between Omitted Variable and Multicollinearity

Multicollinearity which means that two or more independent variables are highly correlated with each other, can have a powerful effect upon model specification and particularly, on statistical tests of model specification (Hanushek & Jackson, 1977).

One of the methods which are used to avoid multicollinearity is to drop the collinear variable, but it is the risk of mis-specifying model and having omitted variable bias (Crown, 1998). Where there is multicollinearity, it is especially dangerous to omit one of the interrelated variables from the model (Upton, 1987). Dropping variables seems the most obvious solution and may work in some cases where not interested in individual parameter values. But, the coefficient on the remaining collinear variable will absorb most of the effect of the omitted variable. Therefore this solution results the problem of omitted variable bias (Crown, 1998). The consequences of omitting collinear variable are potentially more serious than those of multicollinearity because specification error may introduce bias into the model (Berry & Feldman, 1990).

It was mentioned that the biases in the estimation with omitted X_3 are $\beta_3 b_{31}$ and $\beta_3 b_{32}$. It was also given in the equations of (3.18) and (3.19) that the correlations between the omitted variable and the included variables, denoted by r_{31} and r_{32}, affect the values of b_{31} and b_{32} for a given set of values for X_1 and X_2. Thus, it is obvious that, when omitted variable is correlated with the included variables, the values of r_{31} and r_{32} will be high, then as a result of this the omitted variable bias will arise. Consequently, when there is multicollinearity, exclusion of the collinear variable from the model causes omitted variable bias, because of the increasing of the values of r_{31} and r_{32} (Berry & Feldman, 1990).

On the other hand, when data on the omitted variable exist but were ignored, the standard *t*-test of the null hypothesis $\beta_j = 0$ $(j = 1, \ldots, p-1)$ is a test of the specification that includes *X*, and performance of that statistical test provides information about appropriate model specification. But multicollinearity confounds this test and weakens the ability to judge among model specifications. Since multicollinearity reduces the precision of the estimates because of it increases their variance, it becomes difficult to develop tests that are good at distinguishing between alternative values of a parameter and alternative specifications of the model (Hanushek & Jackson, 1977).

Therefore, it is suggested strongly not to omit a variable simply when it appears to be highly correlated to another variable because omitting a variable is often far worse (Burt, Barber, Rigby, & Cooper, 2009). The researcher must be more cautious in evaluating and interpreting the results and must provide much more information about the behavior being modeled. This information can come only from theoretical considerations and previous empirical work (Hanushek & Jackson, 1977).

CHAPTER FOUR
SIMULATION STUDY

4.1 Introduction

In the previous chapters the definitions of multiple regression and omitted variable were given. In this chapter, it will be given that how omitted variable bias can affect the model.

First, in this chapter, three kinds of populations with 1000 data were generated from the multivariate normal distribution. In each population, three independent variables X_1, X_2 and X_3, dependent variable *Y* and the error term were generated. The differences between the populations are the correlations between the variables. One of these populations has no correlated variables and is named “L-pop”; the other population has two variables that are correlated with each other and is named “M-pop”; and the other population has all the variables highly correlated with each other and is named “H-pop”. The purpose of generating populations with different correlated variables is to investigate the omitted variable bias in three different situations.

Second, random samples were drawn from these populations with sample size of *n* = 30. Then, regression procedure was applied to these samples. All the independent variables were included to the model firstly, then one variable (X_3) was omitted and then two variables (X_2 and X_3) were omitted from the model. The model was built in every omission in order to investigate the omitted variable bias. Furthermore, when two variables were omitted from the model, RESET test was applied in order to show how RESET test work. The computations were executed using a Minitab macro program. This macro program was run 10,000 times and the results were recorded.

The study with sample size of *n* = 30 were also applied with sample size of *n* = 50 in order to check whether larger sample size affects the omitted variable bias.

4.2 Correlations between Variables

The correlation matrixes of *Y*, X_1, X_2, and X_3 for each population are given in Table 4.1, Table 4.2 and Table 4.3.

Table 4.1 Correlation coefficients for L-pop

	Y	X_1	X_2	X_3
Y	1			
X_1	0.587	1		
X_2	0.639	0.247	1	
X_3	0.578	0.118	0.167	1

The population named L-pop has no high correlation between independent variables as Table 4.1 shows. By the way, the simple correlation coefficients between X_i and *Y* are not high.

Table 4.2 Correlation coefficients for M-pop

	Y	X_1	X_2	X_3
Y	1			
X_1	0.502	1		
X_2	0.818	0.176	1	
X_3	0.800	0.129	0.889	1

Table 4.2 shows that there is a high correlation between X_2 and X_3; $r_{32} = 0.889$. Furthermore, the correlations between X_i and *Y* are absolutely high, especially the correlation between X_2 and *Y*.

Table 4.3 Correlation coefficients for H-pop

	Y	X_1	X_2	X_3
Y	1			
X_1	0.739	1		
X_2	0.815	0.404	1	
X_3	0.926	0.747	0.903	1

Finally, as seen from Table 4.3, that a high correlation exists between all the independent variables and similar to M-pop, the simple correlation coefficients between X_i and Y are high.

4.3 Omitted Variable Bias when Sample Size 30

Random samples were drawn from each of populations with sample size of n = 30. Then, regression procedure was applied to these samples. All of the true coefficients of independent variables are adjusted to be equal to one.

4.3.1 When One Variable is Omitted

After 10,000 samples with n = 30 are drawn from each of these populations and regression procedure is applied, X_3 is omitted from the model. The results in regard to the regression analysis which is applied to the different populations are shown in Table 4.4.

Table 4.4 Mean values of the amount of bias, the coefficients and the standard deviations

	L-pop	**M-pop**	**H-pop**
b_{31}	0.076	- 0.028	0.460
b_{32}	0.145	0.904	0.715
b_1	1.036	0.959	1.363
b_2	1.258	1.929	1.790
$s(b_1)$	0.290	0.228	0.224
$s(b_2)$	0.285	0.222	0.221

In Table 4.4, b_{31} means that the amount of bias on b_1 when X_3 is omitted and similarly b_{32} means that the amount of bias on b_2 when X_3 is omitted.

For L-pop, when X_3 is excluded from the model, there is approximately 4% change in the coefficient of X_1. Since the correlation between X_3 and X_2 is much more than the correlation between X_3 and X_1, the ratio of bias on the coefficient of X_2 is approximately 0.26.

For M-pop, when X_3 is omitted, it becomes the part of the error term and since the correlation between X_2 and X_3 is high, then the error term is correlated with X_2. Since the error term and X_2 are correlated, the assumption which implies that the conditional mean of ε_i given X_i is nonzero (given in Section 2.2.3.1) is violated, and this causes omitted variable bias on b_2. That is, the estimate of β_2 is biased upward, because X_3 is omitted. On the other hand, since there is low correlation between X_1 and X_3, almost 4% bias is emerged on b_1 and likewise b_1 is biased downward.

For H-pop, since the omitted X_3 is correlated with the other two independent variables X_1 and X_2, the estimates of the β_1 and β_2 are substantially different from the real values which are equal to one. The amount of bias in the estimates are

$b_{31} = 0.460$ and b_{32} = 0.715. Therefore, the omission of X_3 which is an important variable for the model causes bias, as expected. b_1 and b_2 consist of the effects of β_3 and are biased.

The results of explanatory power of the models for each population are given in Table 4.5.

Table 4.5 The explanatory powers of the models for each population

	L-pop		**M-pop**		**H-pop**	
Omission	R^2	R^2_{adj}	R^2	R^2_{adj}	R^2	R^2_{adj}
Before	0.8112	0.7894	0.8422	0.8240	0.8715	0.8567
After (X_3)	0.6139	0.5853	0.8048	0.7904	0.8655	0.8555

In L-pop, before omitting any independent variable, $R^2 = 0.8112$ and $R^2_{adj} = 0.7894$; but after X_3 is omitted, $R^2 = 0.6139$ and $R^2_{adj} = 0.5853$. The reduction in the values of R^2 and R^2_{adj} is obvious. Therefore the variation in the dependent variable is not fully measured without it and significance of the model decreases.

In M-pop, the value of R^2 is equal to 0.8048. This value was equal to 0.8422 before X_3 was omitted. This means, although X_1, X_2 and X_3 explain 84% of the model, X_1 and X_2 without X_3 explain 81% of the total sample variation of *Y*. Similarly, although the value of R^2_{adj} is 0.824, this value decreases to 0.7904 after omitting.

In both of the populations, M-pop and H-pop, since explained variability (SSR) decreases, when X_3 is omitted, the values of R^2 and R^2_{adj} are less than before. However, as seen from Table 4.5, there are no noticeable differences among the values before and after omitting.

Although the estimates of β_i parameters have omitted variable bias, the values of R^2 and R^2_{adj} are high and does not change significantly. Basically, it is expected that these values should decrease and tell a lack of fit of the model to the data. Although X_3 has an important role in explaining Y ($r_{Y3} = 0.926$), the values of R^2 and R^2_{adj} does not give any information about omitted variable. The reason of these values does not change significantly depending on omitting an important variable may be that the included variables have high correlations with dependent variable *Y*. Thus, the results of R^2 and R^2_{adj} assert that these included variables can explain the model sufficiently, although there is an omitted variable.

On the other hand, the reason of decreasing in these values distinctly in L-pop is that the included independent variables have almost low correlations between dependent variable, and hence the explanatory power of the model, without X_3, is not enough.

To better understand the relationship between the bias and R^2, the following graphs are drawn for each population. These graphs show the relationship between R^2 and the bias on b_1 when X_3 is omitted.

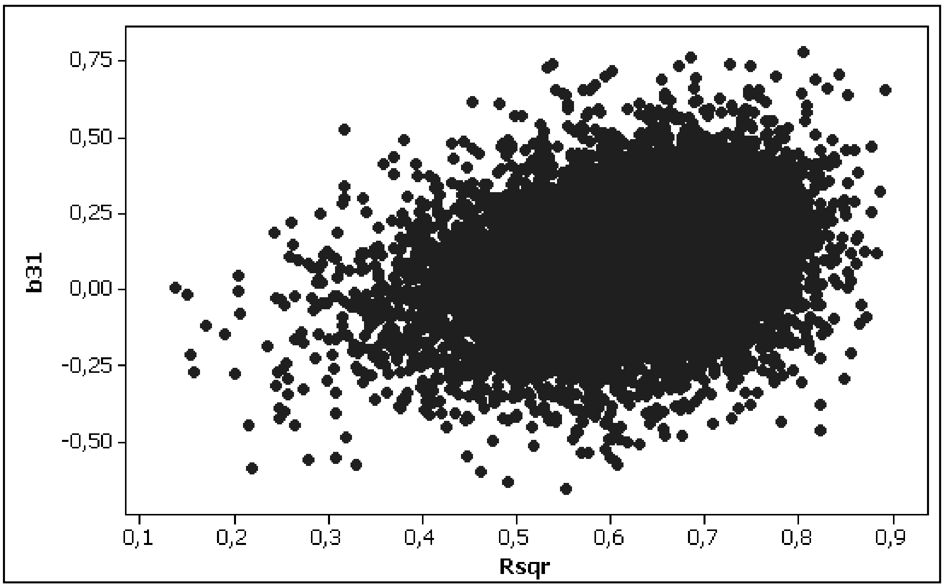

Figure 4.1 A scatterplot of the bias on b_1 versus R^2 for L-pop

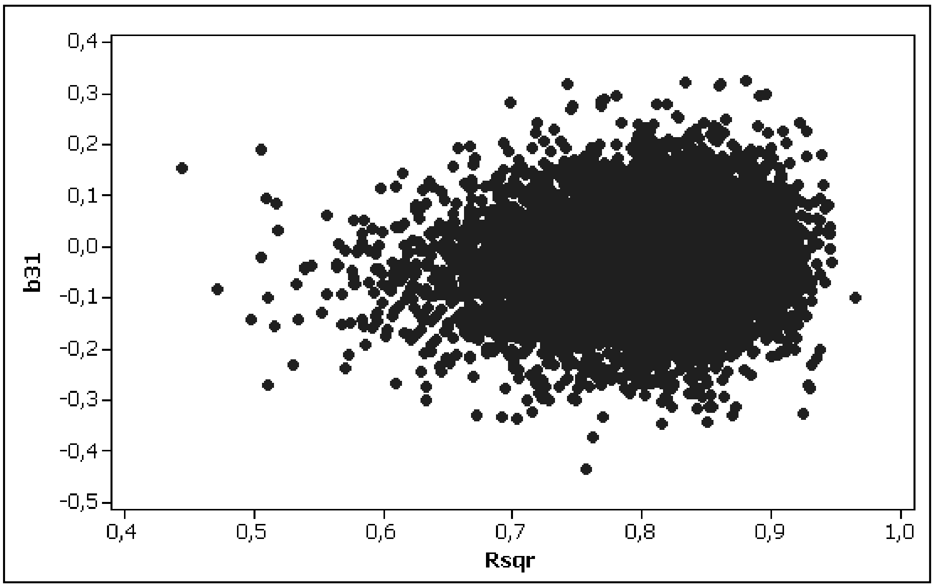

Figure 4.2 A scatterplot of the bias on b_1 versus R^2 for M-pop

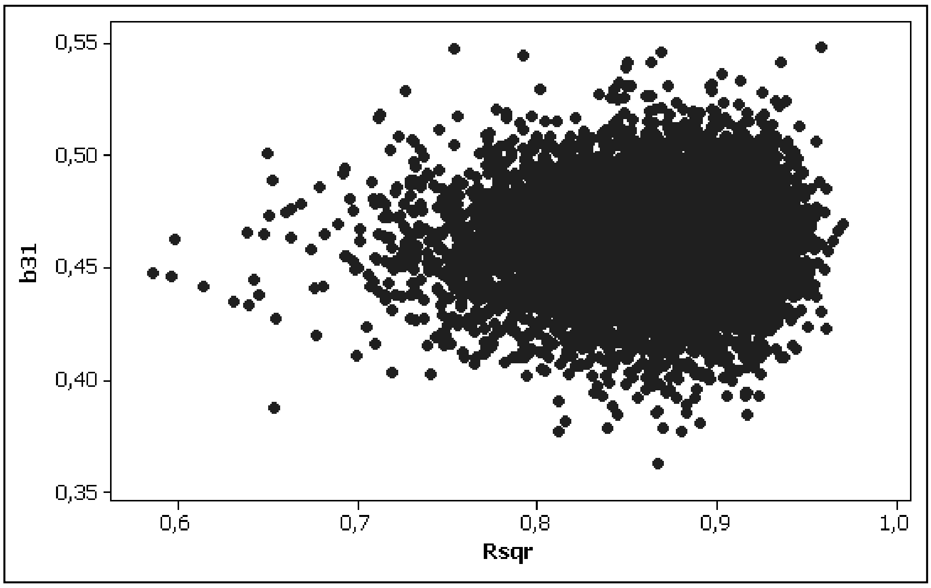

Figure 4.3 A scatterplot of the bias on b_1 versus R^2 for H-pop

As Figure 4.1, Figure 4.2 and Figure 4.3 show, while the values of R^2 increase, the bias may increase or decrease, as expected. Moreover, the graphs for the relationship between R^2 and the bias on b_2 are similar to these graphs.

Stock and Watson (2003) confirm this case. They say that a high R^2 or R^2_{adj} does not imply that there is no omitted variable and similarly a low R^2 or R^2_{adj} does not mean there is omitted variable.

Consequently, it can be said that it is dangerous to judge the usefulness of the model based solely on these values, R^2 and R^2_{adj}.

4.3.2 When Two Variables are Omitted

10,000 samples with *n* = 30 are drawn from these populations and regression procedure is applied. X_2 and X_3 are omitted from the model. The results in regard to the regression analysis which is applied to the different populations are shown in Table 4.6.

Table 4.6 Mean values of the amount of bias, the coefficient and the standard deviation

	L-pop	**M-pop**	**H-pop**
b_{31}	0.099	0.115	0.705
b_{21}	0.217	0.162	0.391
b_1	1.309	1.274	2.057
$s(b_1)$	0.356	0.429	0.368

In Table 4.6, b_{31} means that the amount of bias on b_1 when X_3 is omitted and similarly b_{21} means that the amount of bias on b_1 when X_2 is omitted.

When the results given in Table 4.6 have been checked, for L-pop, it can be seen that the amount of bias on b_1 caused by omitting X_3 is 0.099 and the amount of bias on b_1 caused by omitting X_2 is 0.217.

For M-pop, as supposed, since $r_{31} = 0.129$ and $r_{21} = 0.176$, the amounts of bias, particularly, are not high. However, unlike the situation of omitting one variable, both of the amounts of bias are added to the estimate, so that, the estimate of true coefficient is biased.

For H-pop, since the omitted X_3 is highly correlated with the included X_1, the bias is high and equal to 0.705 and furthermore, since the other omitted variable X_2 is not highly correlated with X_1, the bias is not as much as for X_3's and equal to 0.391. Besides, as seen from the table, b_1 is quite different from the true coefficient, because b_1 contains both of the omitted variables effects. The rate of bias on b_1 is approximately 106%.

Table 4.7 The explanatory powers of the models for each population

	L-pop		**M-pop**		**H-pop**	
Omission	R^2	R^2_{adj}	R^2	R^2_{adj}	R^2	R^2_{adj}
Before	0.8112	0.7894	0.8422	0.8240	0.8715	0.8567
After (X_3)	0.6139	0.5853	0.8048	0.7904	0.8655	0.8555
After (X_2, X_3)	0.3496	0.3264	0.2659	0.2397	0.5443	0.5281

As seen from the table, in each population, when two variables are excluded from the model, unlike the case that one variable is excluded, R^2 and R^2_{adj} are reduced excessively. This means, the model which is built with only X_1 does not fit the data very well. X_2 and X_3 have important roles in explaining *Y*, but X_1 does not, as it is seen from Table 4.1, Table 4.2 and Table 4.3. Hence, because of the low correlation between X_1 and *Y*, the values of R^2 and R^2_{adj} are decreased. Consequently, it can be said that if the correlation between the included variable and the dependent variable is low, then R^2 and R^2_{adj} are decreased and signal about omitted variables. However, if the correlation between these included and dependent variables is high, then R^2 and R^2_{adj} do not tell anything about omission.

4.3.3 RESET Test for Sample Size 30

In this study, RESET test is applied when two variables, X_2 and X_3, are excluded from the model to find out how it works and whether it confirms the omissions from the model. As described in the literature, RESET test is principally improved to detect omitted variables.

First, by adding second and third powers of the fitted values of Y to the original model, a new model is built. 10,000 samples with n = 30 are drawn from the populations and these procedures are applied 10,000 times. The hypothesis that no relevant independent variables have been omitted from the regression model is tested by testing the significance of additional variables, $\hat{Y}^2, \hat{Y}^3$. F test for the significance of these additional variables is used as Ramsey who is the developer of the RESET test suggests.

Ramsey RESET test results using powers of the fitted values of Y are given in Table 4.8.

Table 4.8 The statistics for F – values in regard to Ramsey RESET test

	Mean	**Min – Max**
L-pop	597.15	6.86 – 14376.1
M-pop	408.02	9.90 – 26938.6
H-pop	163.79	0.95 – 9957.3

Regarding all of these statistics, from Table 4.8, it is seen that, for every population, computed values of F are substantially great.

The critical value for F is $F_{\alpha,k,n-k-r-1}$ = 5.53 where $\alpha = 0.01$, k = 2, n = 30, r = 2.

Since the computed values of F exceed the critical value, the null hypothesis is rejected for each population. The combined effects of these additional variables do improve the model. This means, one or more variables should be included to the model. Hence, RESET test detects that some variable(s) omitted from the model. As described in the literature, RESET test is not able to discover which variables omitted. However, it gives a caution about omission.

Incidentally, as seen from the table, for H-pop, the minimum value of *F* is equal to 0.951 and less than the critical value. But, when looking at the data, the percentage of being less than the critical value for H-pop is 1%. Therefore, it can be said that, this case does not change the result.

Comparisons of the explanatory powers of the new model which is built by powers of the fitted values of *Y* and old model which is built by only X_1 are given in Table 4.9.

Table 4.9 The explanatory powers of the models for each population

	L-pop		**M-pop**		**H-pop**	
Model	R^2	R^2_{adj}	R^2	R^2_{adj}	R^2	R^2_{adj}
Old	0.3496	0.3264	0.2659	0.2397	0.5443	0.5281
New	0.9602	0.9556	0.9352	0.9278	0.9210	0.9193

Considering these statistics, to add second and third powers of the fitted values of *Y* to the original model increases the values of R^2 and R^2_{adj}, and it can be said that to add new variables to the model increases the explanatory power of the model.

4.4 Omitted Variable Bias when Sample Size 50

The samples that contain substantially more data are drawn to check whether larger sample size affects the omitted variable bias. Random samples were drawn from each of populations with sample size of $n = 50$. Then, regression procedure was applied to these samples. All of the true coefficients of independent variables are adjusted to be equal to one.

4.4.1 When One Variable is Omitted

After 10,000 samples with $n = 50$ are drawn from these populations and regression procedure is applied, X_3 is omitted from the model. The results in regard to the regression analysis which is applied to the different populations are shown in Table 4.10.

Table 4.10 Mean values of the amount of bias, the coefficients and the standard deviations

	L-pop	**M-pop**	**H-pop**
b_{31}	0.079	- 0.029	0.459
b_{32}	0.147	0.904	0.715
b_1	1.037	0.965	1.356
b_2	1.264	1.925	1.791
$s(b_1)$	0.220	0.173	0.169
$s(b_2)$	0.216	0.168	0.167

In Table 4.10, b_{31} means that the amount of bias on b_1 when X_3 is omitted and similarly b_{32} means that the amount of bias on b_2 when X_3 is omitted.

For L-pop, when X_3 is omitted from the model, approximately 4% bias on the coefficient of X_1 is emerged. Since the correlation between X_3 and X_2 is much more than the correlation between X_3 and X_1, the ratio of bias on the coefficient of X_2 is approximately 0.26.

For M-pop, when X_3 is omitted, it becomes the part of the error term and since the correlation between X_2 and X_3 is high, then the error term is correlated with X_2. Since the error term and X_2 are correlated, the assumption of the least square is violated, and this causes omitted variable bias on b_2. The percentage of bias is approximately 93%. On the other hand, when the amounts of bias are compared, it is seen that the bias on b_1 is less than the bias on b_2, since the correlation between X_1 and X_3 is less than the correlation between X_2 and X_3.

For H-pop, since the omitted X_3 is correlated with the other two independent variables X_1 and X_2, the estimates of the β_1 and β_2 are substantially different from the real values which are equal to one. The amount of bias in the estimate with omitted X_3 are $b_{31} = 0.459$ and b_{32} = 0.715. Therefore, omission of X_3 which is an important variable for the model causes bias, as supposed. b_1 and b_2 consist of the effects of β_3 and are biased.

The results of explanatory power of the models for each population are given in Table 4.11.

Table 4.11 The explanatory powers of the models for each population

	L-pop		**M-pop**		**H-pop**	
Omission	R^2	R^2_{adj}	R^2	R^2_{adj}	R^2	R^2_{adj}
Before	0.8089	0.7965	0.8389	0.8284	0.8689	0.8603
After (X_3 **)**	0.6102	0.5937	0.8034	0.7949	0.8650	0.8593

In L-pop, before omitting any independent variable, $R^2 = 0.8089$; but after X_3 is omitted, $R^2 = 0.6102$. After X_3 was omitted, as shown in the table, both of the values R^2_{adj} and R^2 are significantly less than before.

In both of the populations, M-pop and H-pop, it can be said that, since explained variability (SSR) decreases, when X_3 is omitted, the values of R^2 and R^2_{adj} are less than before. However, as it is seen at Table 4.11, there are no noticeable differences between the values before and after omitting.

Table 4.10 shows the estimates of β_i parameters have omitted variable bias. In spite of the fact that, the values of R^2 and R^2_{adj} are high and does not change significantly. Basically, it is expected that these values should decrease and tells a lack of fit of the model to the data. Although X_3 has an important role in explaining

Y ($r_{Y3} = 0.926$), the values of R^2 and R^2_{adj} does not give any information about omitted variable. The reason of these values does not change significantly depending on omitting an important variable may be that the included variables have high correlations with dependent variable Y. Thus, the results of R^2 and R^2_{adj} assert that these included variables can explain the model sufficiently, although there is an omitted variable.

On the other hand, the reason of decreasing in L-pop is that the included independent variables have low correlations between dependent variable, and the explanatory power of the model, without X_3, is not enough.

Consequently, it can be said that it is dangerous to judge the usefulness of the model based solely on these values, R^2 and R^2_{adj}.

4.4.2 When Two Variables are Omitted

10,000 samples with n = 50 are drawn from these populations and regression procedure is applied. This time, X_2 and X_3 are omitted from the model together. The results in regard to the regression analysis which is applied to the different populations are shown in Table 4.12.

Table 4.12 Mean values of the amount of bias, the coefficient and the standard deviation

	L-pop	**M-pop**	**H-pop**
b_{31}	0.116	0.129	0.743
b_{21}	0.248	0.182	0.410
b_1	1.359	1.310	2.111
$s(b_1)$	0.273	0.327	0.281

In Table 4.12, b_{31} means that the amount of bias on b_1 when X_3 is omitted and similarly b_{21} means that the amount of bias on b_1 when X_2 is omitted.

For L-pop, it can be seen from the table, the amount of bias on b_1 caused by omitting X_3 is 0.116 and the amount of bias on b_1 caused by omitting X_2 is 0.248. Therefore, the total bias on b_1 is 0.359, since b_1 contain the effects of both of the omitted variables.

For M-pop, as expected, since $r_{31} = 0.129$ and $r_{21} = 0.176$, the amounts of bias, particularly, are not too high. However, unlike the situation of omitting one variable, both of the amount of bias are added to the estimation, so that, the estimate of true coefficient is biased.

For H-pop, since the omitted X_3 is highly correlated with the included X_1, the bias is high and since the other omitted variable X_2 is not highly correlated with X_1, the bias is not as much as X_3's. Moreover, as it is seen from the table, b_1 is quite different from the true coefficient, because b_1 includes both of the omitted variables effects. The percentage of bias is approximately 111%.

Table 4.13 The explanatory powers of the models for each population

	L-pop		**M-pop**		**H-pop**	
Omission	R^2	R^2_{adj}	R^2	R^2_{adj}	R^2	R^2_{adj}
Before	0.8089	0.7965	0.8389	0.8284	0.8689	0.8603
After (X_3**)**	0.6102	0.5937	0.8034	0.7949	0.8650	0.8593
After(X_2**,** X_3**)**	0.3468	0.3331	0.2598	0.2444	0.5432	0.5336

As seen from the table, in each population, when two variables are excluded from the model, unlike the case one variable is excluded, R^2 and R^2_{adj} are reduced excessively. This means, the model which is built with only X_1 does not fit the data very well. X_2 and X_3 have important roles in explaining *Y*, but X_1 does not, as seen from the correlation tables. Hence, the low correlation between X_1 and *Y* is the reason of reduced R^2 and R^2_{adj}. Therefore it can be said that if the correlation between the included variable and the dependent variable is low, then R^2 and R^2_{adj} are decreased and signal about omitted variables. However, if the correlation between these included and dependent variables is high, then R^2 and R^2_{adj} do not tell anything about omission.

4.4.3 RESET Test for Sample Size 50

RESET test is applied when n = 50 and when two variables, X_2 and X_3, are excluded from the model to find out how it works and whether it confirms the omissions from the model.

The process which is used when n = 30 is followed. As Ramsey who is the developer of the RESET test suggests, first, by adding second and third powers of the fitted values of *Y* to the original model, a new model is built. 10,000 samples with n = 50 are drawn from the populations and these procedures are applied 10,000 times. The hypothesis that no relevant independent variables have been omitted from the regression model is tested by testing the significance of additional variables. *F* test for the significance of these additional variables is used.

Ramsey RESET test results using powers of the fitted values of *Y* are given in Table 4.14 .

Table 4.14 The statistics for 10,000 *F* - values in regard to Ramsey RESET test

	Mean	**Min – Max**
L-pop	615.86	29.63 – 9740.5
M-pop	389.73	25.60 – 7679.8
H-pop	165.45	4.85 – 4046.8

Regarding all of these statistics, from this table, it is seen that, for every population, computed values of *F* are substantially great.

The critical value for F is $F_{\alpha,k,n-k-r-1}$ where $\alpha = 0.01$, n = 50, k = 2, r = 2 is approximately 5.00. Since the computed values of F exceed the critical value, the null hypothesis is rejected for each population.

The combined effects of these additional variables do improve the model. This means, one or more variables should be included to the model. Hence, RESET test detects that some variable(s) omitted from the model. As described in the literature, RESET test is not able to discover which variables omitted. However, it gives a caution about omission.

Incidentally, as it can be seen from the table, for H-pop, the minimum value of F is equal to 4.85 and less than the critical value. But, when looking at the data, the percentage of being less than the critical value for H-pop is 0.1%. Therefore, it can be said that, this case does not change the result.

Comparisons of the explanatory powers of the new model which is built by powers of the fitted values of Y and old model which is built by only X_1 are given in Table 4.15.

Table 4.15 The explanatory of the models for each population

	L-pop		**M-pop**		**H-pop**	
Model	R^2	R^2_{adj}	R^2	R^2_{adj}	R^2	R^2_{adj}
Old	0.3468	0.3331	0.2598	0.2444	0.5432	0.5336
New	0.9539	0.9509	0.9285	0.9238	0.9077	0.9016

According to the results, to add second and third powers of the fitted values of Y to the original model increases the values of R^2 and R^2_{adj}, and it can be said that to add new variables to the model increases the explanatory of the model.

CHAPTER FIVE
CONCLUSIONS

In this study, the omitted variable bias is examined as theoretically and investigated in which conditions the omitted variable bias occurs and how affects the model and estimation by simulation.

In the simulation study, three types of populations with 1000 data which varied depending on the correlation values between the variables were generated to show the effects of the different correlations on the bias. Random samples were drawn from these populations with sample size of n = 30 and n = 50. Though the true model had three independent variables, the models were estimated by omitting one and then two independent variables for each sample. 10,000 repetitions were generated for each of sample sizes of 30 and 50. Therefore the effects of omitted variable bias were investigated in each situation. The amount of bias, the estimated coefficients, coefficients of determination and the adjusted coefficients of determination, standard deviations of the estimated coefficients are computed for every model and F statistics are also computed for applying RESET test.

It was described in the literature that, when a relevant variable is omitted from the model, the effects of this omitted variable can not be estimated and the estimators for other variables in the model may be biased and thus misleading. Because, if a relevant variable is omitted, it becomes the part of the error term and if the correlation between the omitted and the included variables is high, then the error term is correlated with the included variable. Thus, the assumption which implies that the conditional mean of ε_i given X_i is nonzero is violated, and this causes omitted variable bias in the coefficient of included variable. In this study, it is seen that when a high correlated variable with the other variables in the model is omitted from the model, it causes bias in the included variable, and this bias changes depending on the values of correlation. A high correlation increases the amount of bias and similarly a low correlation decreases the amount of bias. In brief, the correlation between the omitted and the included variables and the bias in the estimated coefficients are directly proportional.

At the same time, when the values of R^2 and R^2_{adj} are calculated and considered, it is seen that although the estimators of β_i parameters have omitted variable bias, the values of R^2 and R^2_{adj} are high and does not change significantly. Basically, it is expected that these values should decrease and tell a lack of fit of the model to the data. Even though the omitted variable has an important role in explaining *Y*, the values of R^2 and R^2_{adj} does not signal about omitted variable. The reason of these values does not change significantly depending on omitting an important variable may be that the included variables have high correlations with dependent variable *Y*. Thus, the results of R^2 and R^2_{adj} assert that these included variables can explain the model sufficiently, although there is an omitted variable. On the other hand, these values may decrease distinctly when a relevant variable is omitted. The reason of this decreasing may be that the remaining independent variables have low correlations between dependent variable, when the relevant variable is omitted. Therefore, it can be said that a high or a low R^2 or R^2_{adj} does not give any information about whether there is an omitted variable. Consequently, it can be seen clearly from the results that it is dangerous to judge the usefulness of the model based solely on these values, R^2 and R^2_{adj}.

Problem of omitting relevant variables is a remarkable issue. It brings a lot of trouble and causes misleading results. Therefore, the investigator should check whether there are omitted variables. For this purpose, Ramsey (1969) developed RESET test, as mentioned in Chapter 3. Simulation results show that, RESET test, which is applied when two variables are omitted from the model, detects that some variables are omitted from the model. As defined in Chapter 3, this test does not tell how many or which variables are omitted. However, considering computed *F* values and comparing them with the critical values, the null hypothesis which implies that the model has no omitted variable is rejected and RESET test signals the omission, truthfully.

In general, it is said that the researchers achieve greater power with increases in sample sizes. Larger sample sizes result in increasingly more precise estimates of parameters (Meyers, Gamst & Guarino, 2006). Finally, the omitted variable bias is investigated with different sample size and it is seen that when sample size is increased, the results are not changed. This means that even though the sample size is increased, the existing omitted variable bias does not disappear. Hence, as Stock and Watson (2003) defined, it can be said that to change the sample size is not the solution for the omitted variable bias.

REFERENCES

Agresti, A. (2002*). Categorical data analysis* (2nd ed.). John Wiley and Sons.

Arminger, G., Clogg, C. C., & Sobel, M. E. (1995). *Handbook of statistical modeling for the social and behavioral sciences.* Springer.

Barreto, H., & Howland, F. M. (2006). *Introductory econometrics using Monte Carlo Simulation with Microsoft Excel.* Cambridge University Press.

Berry, W. D., & Feldman, S. (1990). *Multiple regression in practice* (7th ed.). Sage.

Boardman, A. E., & Murnane, R. J. (1979). Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education*, *52,* 113–121.

Burt, J. E., Barber, G. M., Rigby, D. L., & Cooper, M. (2009). *Elementary statistics for geographers* (3rd ed.). Guilford Press.

Byun, Y. (2005). *Introduction to econometrics*. Lecture notes.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press.

Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example* (4th ed.). John Wiley and Sons.

Christensen, R. (2002). *Plane answers to complex questions: The theory of linear models* (3rd ed.). Springer.

Clarke, K. A. (2005). The Phantom Menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, 22, 341–352.

Clements, M. P., & Hendry, D. F. (2002). *A companion to economic forecasting.* Blackwell Publishing.

Cohen, J. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates.

Crown, W. H. (1998). *Statistical models for the social and behavioral sciences: Multiple regression and limited-dependent variable models.* Greenwood Publishing Group.

Dewberry, C. (2004). *Statistical methods for organizational research: Theory and practice.* Routledge.

Draper, N. R., & Smith, H. (1966). *Applied regression analysis.* John Wiley and Sons.

Ehrenberg, R.G., & Brewer, D.J. (1994). Do school and teacher characteristics matter? Evidence from high school and beyond. *Economics of Education Review, 13,* 1–17.

Ehrenberg, R.G., Goldhaber, D.D., & Brewer, D.J. (1995). Do teachers’ race, gender, and ethnicity matter? Evidence from NELS:88. *Industrial and Labor Relations Review, 48,* 547–561.

Evans, M. K. (2002). *Practical business forecasting.* Wiley-Blackwell.

Field, A. P. (2005). *Discovering statistics using SPSS* (2nd ed.). Sage.

Gossy, G. (2008). *A stakeholder rationale for risk management: Implications for corporate finance decisions.* Gabler Verlag.

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Pearson Education, New Jersey.

Hanushek, E. A., & Jackson, J. E. (1977). *Statistical methods for social scientists.* Academic Press,Inc.

Hsiao, C. (2003). *Analysis of panel data* (2nd ed.). Cambridge University Press.

Jobson, J. D. (1991). *Applied multivariate data analysis: Regression and experimental design.* Springer.

Kim, J., & Frees, E. W. (2006). Omitted variables in multilevel models. *Psychometrika, 71* (4), 659–690.

Kurt, S. (2000). *Lecture notes on regression analysis*. Dokuz Eylül University.

Leightner, J. E., & Inoue, T. (2007). Tackling the omitted variables problem without the strong assumptions of proxies. *European Journal of Operational Research, 178,* 819–840.

Mason, R. L., Gunst, R. F. & Hess, L. J. (2003). *Statistical design and analysis of experiments: With applications to engineering and science* (2nd ed.). John Wiley and Sons.

McCallum, B. T. (1972). Relative asymptotic bias from errors of omission and measurement, *Econometrica, 40* (4).

Mendenhall, W., & Sincich, T. (2003). *A second course in statistics: Rregression analysis* (6th ed.). Pearson Education, Inc.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research: Design and interpretation* (2nd ed.). Sage.

Milliken, G. A. & Graybill, F. A. (1970). Extensions of the general linear hypothesis model. *Journal of the American Statistical Association, 65* (330), 797-807.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Irwin.

Newbold, P., Carlson, W. L., & Thorne, B. (2003). *Statistics for business and economics* (5th ed.). Pearson Custom Publishing.

Ramsey, J. B. (1969). Tests for the specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)* 31 (2), 350-371.

Ryan, B., & Joiner, B. L. (2001). *Minitab Handbook* (4th ed.). Duxbury.

Sackett, P. R., Laczo, R. M., & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: The omitted variables problem. *Journal of Applied Psychology, 88* (6), 1046-1056.

Schreck, P. (2009). *Corporate social performance: Understanding and measuring economic impacts of corporate social responsibility*. Springer.

Stock, J. H., & Watson, M. W. (2003). *Introduction to econometrics.* Pearson Education.

Stoker, T. M. (1983). *Omitted variable bias and cross section regression*. Massachusetts Institute of Technology (MIT) Press.

Theil, H. (1957). Specification errors and the estimation of economic relationships. *Review of the International Statistical Institute, 25,* 41 - 51.

Theil, H. (1971). *Principles of econometrics.* John Wiley and Sons, Amsterdam.

Upton, M. (1987). *African farm management* (2nd ed.). CUP Archive.

Verbeek, M. (2004). *A guide to modern econometrics* (2nd ed.).John Wiley and Sons.

Watson, P. K. (2002). *A practical introduction to econometric methods: Classical and modern.* University of the West Indies Press.

Wickens, M. R. (1972). A note on the use of proxy variables. *Econometrica, 40* (4), 759-761.

Williams, R. (2008). *Specification error*. Lecture Notes.

Wooldridge, J. M., (2002). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.