ANADOLU ÜNİVERSİTESİ BİLİM VE TEKNOLOJİ DERGİSİ –B
Teorik Bilimler

ANADOLU UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY –B
Theoretical Sciences
Cilt/Vol.: 2-Sayı/No: 1 : 1-19 (2012)

ARAŞTIRMA MAKALESİ /**RESEARCH ARTICLE**

OMITTED VARIABLE BIAS AND DETECTION WITH RESET TEST IN REGRESSION ANALYSIS

Suay EREEŞ[1*], Neslihan DEMİREL[2]

ABSTRACT

In this paper, 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). A simulation was performed and three types of populations which varied depending on the correlations between the variables were generated and random samples were drawn from these populations. When correlations were changed and the number of omitted variables was increased, the effects of omitted variable bias were investigated. Moreover, by increasing the sample size, it was investigated whether the effects of omitted variable bias were changed depending on sample size.

Keywords: Omitted variable bias, Model specification error, RESET test.

REGRESYON ANALİZİNDE DIŞLANAN DEĞİŞKEN YANLILIĞI VE YANLILIĞIN RESET TESTİ İLE TESPİTİ

ÖZ

Bu çalışmada, dışlanan değişken yanlılığı, bu yanlılığın önemi, nedenleri ve sonuçları araştırılırken dışlanan değişken sorununu ortadan kaldırmak için kullanılan yöntemler incelenmiş ve ayrıca modelden dışlanan değişkenlerin varlığını saptamak üzere RESET testi kullanılmıştır. Bir benzetim çalışması yapılmıştır ve değişkenler arasındaki korelasyon değerlerine bağlı olarak değişen üç değişik tipte kitle türetilmiş ve bu kitlelerden rassal örneklemler çekilmiştir. Korelasyon değerleri değiştiğinde ve dışlanan değişken sayısı arttığında dışlanan değişken yanlılığının ne gibi etkileri olduğu incelenmiştir. Ayrıca, örneklem ölçüsü arttırılarak dışlanan değişken yanlılığının örneklem ölçüsüne bağlı olarak değişip değişmediği de araştırılmıştır.

Anahtar Kelimeler: Dışlanan değişken yanlılığı, Model spesifikasyon hatası, RESET test.

[1,] Department of Statistics, Yaşar University, Izmir, Turkey.
[*]Yaşar University, Faculty of Science and Letter, Department of Statistics
Email: suay.erees@yasar.edu.tr
[2,] Department of Statistics, Dokuz Eylül University, Izmir, Turkey.

Recieved: 16 February 2011; **Revised:** 29 June 2011; **Accepted:** 10 August 2011

1. INTRODUCTION

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. 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 work. 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.

In this study, the problem of omitted variable and RESET test for detecting omitted variables are discussed. Furthermore, omitted variable bias and its effects on the parameters and RESET test are presented using simulation. The simulation study also include examining the effects of the larger sample size on omitted variable bias.

2. OMITTED VARIABLE

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 and Frees, 2006).

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 and Watson, 2003). Most regressions conducted by economists can be critiqued for omitting some important independent variables which may cause the estimated relationships to change. If the omitted variable bias is such a big problem, then why are some variables 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 and Inoue, 2007).

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. 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 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 and 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 true model with two independent variables is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{1}$$

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

$$\hat{Y} = b_0 + b_1 X_1 \tag{2}$$

and therefore

$$Y = b_0 + b_1 X_1 + e^* \tag{3}$$

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 and 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 $\beta_2 = 0$, 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. If the true model is as equation (1) and we estimate as equation (2), then the least square estimator is (Williams, 2008)

$$E(b_1) = \beta_1 + \beta_2 \frac{\sigma_{12}}{\sigma_1^2} \tag{4}$$

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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad (5)$$

and we estimate

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 \quad (6)$$

and therefore

$$Y = b_0 + b_1 X_1 + b_2 X_2 + e^* \quad (7)$$

where $e^* = \beta_3 X_3 + \varepsilon$. Taking the expected value of b_1 and b_2, assuming fixed X and $E(\varepsilon) = 0$

$$E(b_1) = \beta_1 + \beta_3 b_{31} \quad (8)$$

$$E(b_2) = \beta_2 + \beta_3 b_{32} \quad (9)$$

where

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

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. 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. Thursby (1989) has examined the sensitivity of some general checks for misspecification in linear regression models. He considered the Durbin-Watson, Regression Equation Specification Error Test (RESET), Chow, and differencing tests and found that the Durbin-Watson test performs well when the omitted is autocorrelated, RESET performs well against incorrect functional form, and the Chow test performs well against structural shifts. Pagan and Hall (1983) have suggested that White's test will be sensitive to omitted variables. Godfrey and Orme (1994) has showed that, Information Matrix (IM) test and White's test are insensitive and a simple RESET test, using only the squared value of the OLS predicted value from the null model, performs quite well in in recognizing the omitted variables. Since the value of the RESET test has been raised, RESET can be used at least as a reliable specific check for omitted variables and also as an appropriate general check for misspecification in linear regression models (Leung and Yu, 2000).

RESET test which is one of the oldest specification tests for linear regression models, that is still widely used, was originally proposed by Ramsey (1969) and is known as the Ramsey RESET test (Clements and Hendry, 2002). Ramsey's RESET test is primarily a test designed to detect omitted variables and is a model misspecification test. It 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 does not 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 nonlinear functional forms, simultaneous-equation bias, incorrect use of lagged dependent variables (Evans, 2002).

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 . Suppose that the standard linear model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \tag{10}$$

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 (6), and the estimated model is as equation (7), then by adding powers of the fitted values of *Y* to the original model, a new model is estimated

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \delta_1 \hat{Y}^2 + \delta_2 \hat{Y}^3 + u \tag{11}$$

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:

$$F = \frac{(SSE_{old} - SSE_{new})/k}{SSE_{new}/(n-k-r-1)} \approx F(\alpha, k, n-k-r-1) \tag{12}$$

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 et al, 2003). 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, the null hypothesis is rejected. Rejection of the null hypothesis implies the original model is inadequate and can be improved by adding some important variables to the model (Newbold et al, 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

"omitted 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.

4. SIMULATION STUDY

4.1 Introduction

In simulation study, it will be given that how omitted variable bias can affect the model. First, 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 Generating Data

A *p*-dimensional random vector $X = \left(X_1, \ldots X_p\right)'$ is defined to have the multivariate normal ditribution if and only if every nontrivial linear combination of the p-components of X has a univariate normal distribution. The distribution of X is denoted by $N_p\left(\mu, \Sigma\right)$, where μ is a $p \times 1$ mean vector with enties $\mu_i = E\left(X_i\right)$ and Σ is a $p \times p$ covariance martix whose $\left(i, j\right)$th entry is $Cov\left(X_i, X_j\right)$ (Johnson, 1987). In this simulation study, the mean vector which is the same for each population is given in Table 1 and the covariance matrices are given in Table 2, Table 3 and Table 4:

Table 1. Mean vector for three populations

	$\mu_i = E\left(X_i\right)$
X_1	1
X_2	1
X_3	1
e	0

Table 2. Covariances for L-pop

	Y	X_1	X_2	X_3	e
Y	5.148				
X_1	1.298	0.951			
X_2	1.443	0.240	0.990		
X_3	1.298	0.114	0.164	0.980	
e	1.109	-0.007	0.049	0.040	1.027

Table 3. Covariances for M-pop

	Y	X_1	X_2	X_3	e
Y	6.123				
X_1	1.174	0.893			
X_2	1.972	0.162	0.949		
X_3	1.948	0.120	0.852	0.968	
e	1.030	-0.001	0.008	0.007	1.016

Table 4. Covariances for H-pop

	Y	X_1	X_2	X_3	e
Y	8.000				
X_1	2.054	0.966			
X_2	2.294	0.396	0.991		
X_3	2.594	0.727	0.890	0.981	
e	1.058	-0.035	0.017	-0.004	1.079

4.3 Correlations Between Variables

The correlation matrixes of Y, X_1, X_2, and X_3 for each population are given in Table 5, Table 6 and Table 7.

Table 5. 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 5 shows. By the way, the simple correlation coefficients between X_i and Y are not high.

Table 6. 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 6 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 7. 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 7, 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.

4.4 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.

4.4.1 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 8.

Table 8. 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 8, 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 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 9.

Table 9. 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 9, 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.

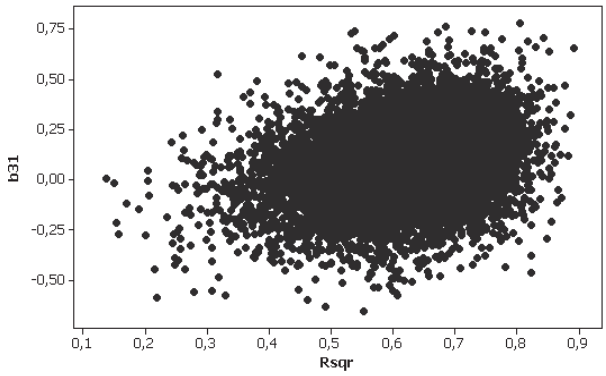

Figure 1. A scatterplot of the bias on b_1 versus R^2 for L-pop

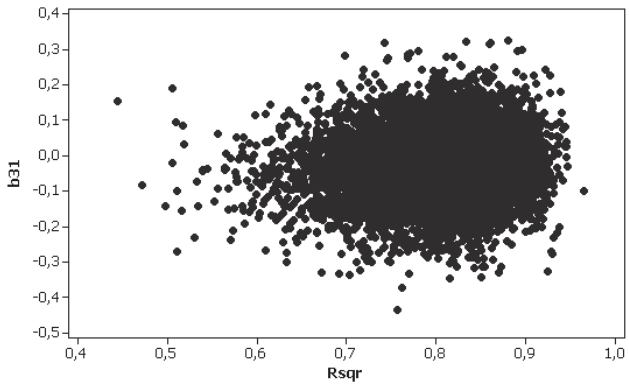

Figure 2. A scatterplot of the bias on b_1 versus R^2 for M-pop

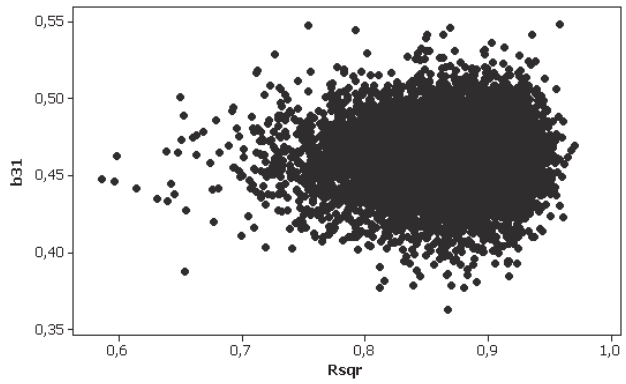

Figure 3. A scatterplot of the bias on b_1 versus R^2 for H-pop

As Figure 1, Figure 2 and Figure 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}.

4.4.2 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.

Table 10. 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 10, 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 10 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 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	0.8112	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 5, Table 6 and Table 7. 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.

4.4.3 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.

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 12.

Table 12. 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 12, 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 13.

Table 13. 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.

4.5 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.

4.5.1 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 14.

Table 14. 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 14, 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 15.

Table 15. 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 when X_3 is omitted, the values of R^2 and R^2_{adj} are less than before, but not as much as expected.

Table 14 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 this may be that the included variables have high correlations with dependent variable. 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}.

4.5.2 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.

Table 16. 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 16, 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%.

Therefore, when comparing the amount of biases for n = 30 and n = 50, it is clear that there is no difference, shown in Table 17. Because when sample size is 30, for example, for H-pop, $b_{31} = 0.460$ and b_{32} = 0.715 which are almost the same for the values in sample size is 50. The increase in the sample size also does not affect the amount of bias when two variables are omitted. Total amount of bias is 1.096 for n = 30 while it is equal to 1.153 for n = 50.

Table 17. The comparison of bias between different sample sizes

		Sample Size n = 30			**Sample Size n = 50**		
		L-pop	**M-pop**	**H-pop**	**L-pop**	**M-pop**	**H-pop**
When one variable is omitted.	b_{31}	0.076	-0.028	0.460	0.079	- 0.029	0.459
	b_{32}	0.145	0.904	0.715	0.147	0.904	0.715
When two variables are omitted.	b_{31}	0.099	0.115	0.705	0.116	0.129	0.743
	b_{21}	0.217	0.162	0.391	0.248	0.182	0.410

As seen from the Table 18, 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.

Table 18. 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

Finally, comparing the values of R^2 and R^2_{adj} between two different sample sizes shows that there is almost no difference and the sample sizes do not affect these values.

4.5.3 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 suggests, first, by adding second and third powers of the fitted values of Y to the original model, a new model is built. 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.

Table 19. 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	04.85 – 4046.8

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 19, 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 20.

Table 20. 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.

5. 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 from multivariate normal distribution with given parameters 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. Although the omitted variable has an important role in explaining *Y*, the values of R^2 and R^2_{adj} do 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 developed RESET test in 1969. 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. 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 et all, 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. Neither the amount of bias nor the values of R^2 and R^2_{adj} are altered. 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 problem of omitted variable bias.

REFERENCES

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

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

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

Godfrey, L.G. and Orme, C.D. (1994). The Sensitivity of some general checks to omitted variables in the linear model. *International Economic Review* 35(2), 489-506.

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

Johnson, M.E. (1987). *Multivariate Statistical Simulation*. John Wiley and Sons.

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

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

Leung, S.F. and Yu, S. (2000). How effective are the RESET tests for omitted variables? *Communications in Statistics-Theory and Methods* 29(4), 879-902.

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

Pagan, A.R. and Hall, A.D. (1983). Diagnostic Tests as Residual Analysis. *Econometric Reviews* 2, 159-218.

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.

Stock, J.H. and 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.

Thursby, J.G. (1989). A Comparison of Several Specification Error Tests for a General Alternative. *International Economic Review* 30(1), 217-230.

Williams, R. (2008). *Specification error*. Lecture Notes. http://www.nd.edu/~rwilliam/stats2/l41.pdf