ENDOGENEITY

A Master's Thesis

by

ERDEM BAŞER

Department of
Economics
İhsan Doğramacı Bilkent University
Ankara

December 2011

To my brother, Onur Başer…

ENDOGENEITY

Graduate School of Economics and Social Sciences
of
İhsan Doğramacı Bilkent University

by

ERDEM BAŞER

In Partial Fulfilment of the Requirements for the Degree of
MASTER OF ARTS

in

THE DEPARTMENT OF ECONOMICS
İHSAN DOĞRAMACI BİLKENT UNIVERSITY

ANKARA

December 2011

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Economics.

Assistant Prof. Taner Yiğit
Supervisor

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Economics.

Associate Prof. Kıvılcım Metin Özcan
Examining Committee Member

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Economics.

Assistant Prof. Ceren Vardar Acar
Examining Committee Member

Approval of the Graduate School of Economics and Social Sciences

---
Prof. Dr. Erdal Erel
Director

ABSTRACT

ENDOGENEITY

Başer, Erdem

M.A., Department of Economics

Supervisor: Asst. Prof. Taner Yiğit

Co-Supervisor: Assoc. Prof. Kıvılcım Metin Özcan

December 2011

One of the core issues in the econometrics literature is the endogeneity problem. Endogeneity occurs when observed variables are correlated with unobserved factors. Endogeneity can arise as a result of measurement error, autoregression with autocorrelated errors, simultaneity and omitted variables. And, there are many methods such as instrumental variable, BLP, control function and maximum likelihood approaches improved in order to overcome this issue. In this study, the issue of endogeneity will be presented and the tests to identify and the methods to solve the problem will be analyzed.

Key words: Endogeneity, measurement error, omitted variable, instrumental variable analysis, simultaneity

ÖZET

ENDOJENİTE

Başer, Erdem

Master, Ekonomi Bölümü

Tez Yöneticisi: Yrd. Doç. Dr. Taner Yiğit

Ortak Tez Yöneticisi:. Doç. Dr. Kıvılcım Metin Özcan

Aralık 2011

Ekonometri literatüründeki en önemli konulardan biri endojenite (içsellik) sorunudur. Endojenite gözlenebilen değişkenler ile gözlenemeyen faktörler arasında bir ilişki olduğu zaman ortaya çıkar.. Ölçüm hatası, otomatik bağıntı hatalı otoregrasyon, eşanlılık ve dahil edilmemiş değişken endojenite sorununun doğmasına neden olur. Enstrümantal değişken, BLP, control fonksiyonu ve en büyük olabilirlik yaklaşımları bu sorunu çözmeye yönelik kullanılan metotlardandır. Bu çalışmada endojenite sorunu, uygulanan testlerin belirlenmesi ve sorunu çözmeye yönelik metotlar analiz edilecektir.

Anahtar Kelimeler: Endojenite (içsellik), ölçüm hatası, model dışı değişken, enstrümantal değişken analizi, eş anlı etkinlik

ACKNOWLEGMENT

I would like to express my gratitude to Associate Professor Kıvılcım Metin Özcan, Assistant Professor Taner Yiğit and Assistant Proffesor Ceren Vardar Acar.

I would like to thank my parents for everything they have done for me.

My beloved wife Sefika Baser needs special thanks for sharing all the good and bad things.

Finally, I would like to thank to my little lovely daughter Izgi Nilsu Baser for being my daughter.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Endogeneity problem is one of the crucial complications in econometrics field. In econometrics the problem of endogeneity occurs when the independent variable is correlated with the error term in a regression model. Endogeneity can arise as a result of measurement error, autoregression with autocorrelated errors, simultaneity and omitted variables. And, there are many methods overcoming this including instrumental variable. In this study, the issue of endogeneity will be presented and the tests to identify and the methods to solve the problem will be analyzed. In the following sections, the meaning and implications of the term will be briefly presented and the literature will be surveyed thoroughly afterwards. The tests to identify the endogeneity and the methods to eliminate the effects of endogeneity will follow in the following sections.

1.1. Classical Linear Regression Model and Endogeneity

In order to analyze the problems caused by the endogeneity, we should start with the textbook definition of the econometric assumptions underlying the regression. Endogeneity occurs when one of the independent variables of the regression equaiton is correlated with the unknown random error term (Wooldridge, 2002). The reasons that cause the exogeneous variables to become correlated with the error term is the core of this study. The classical linear regression model consists 5 key assumptions about the sample is created (Kennedy, 1998), and violations of these assumptions cause severe econometric problems which cause the estimation biased in most cases. Endogeneity bias occurs because of a few violations in these assumptions. In order to understand the nature of this bias, we will start with defining the key assumptions of an econometric model:

1. Dependent variable (Y) can be measured as a linear function of a set of independent variables (a vector X) and an unobservable random error term (ε). The coefficients of this estimation, denoted by a vector β, are assumed to be constant.

2. The expected value of the error term is equal to zero.

3. The error term must have the same variance along the sample and is supposed to be uncorrelated within the sample.

4. The values of independent variable are fixed in repeated samples.

5. The number of observations is greater than the number of independent variables and there are no exact linear relationships between independent variables.

The violation of these assumptions leads to different econometric problems:

1. One of the violations[1] that can occur in the first assumption is called “wrong regressors”. It includes either the omission of relevant independent variables or inclusion of irrelevant independent variables (Kennedy, 1998). The omission of relevant independent variables, or “**omitted variable bias**” is one of the causes of endogeneity that we will discuss more deeply in the following pages.

2. When the expected value of the error term is not equal to zero, then there is so-called “biased intercept problem”.

3. Two major econometric problems arise because of the violation of this assumption, one of which is going to be our concern in this study: heteroscedasticity and autocorrelated errors which the latter is proved to be one of the reasons for endogenetiy.

4. When the values of independent variables are not fixed in repeated samples, 3 related econometric problems arise: measurement error, autoregression and simultaneity, all three of them are considered as the sources for endogenetiy.

5. Violation of the fifth assumption is called multicollinearity, or the direct relationship among two or more independent variables.

[1] There are two more violations of the first assumption, namely nonlinearity and changing parameters as it is shown in Table 1, but they are out of the scope of this study.

Table 1: The Assumptions of the Classical Linear Regression Model[2]			
Assumption		***Mathematical expression***	***Violations***
(1)	Dependent variable is a linear function of a specific set of independent variables, plus an error term	$Y = X\beta + \varepsilon$	**Wrong regressors** Nonlinearity Changing parameters
(2)	Expected value of error term is zero	$E(\varepsilon) = 0$	Biased intercept
(3)	Errors have uniform variance and are uncorrelated	$E(\varepsilon\varepsilon)' = \sigma^2 I$	Heteroskedasticity **Autocorrelated errors**
(4)	Observations on independent variables can be considered fixed in repeated samples	X fixed in repeated samples	Errors in variables **Autoregression Simultaneous equations**
(5)	No exact linear relationships between independent variables and there are more observations than independent variables	Rank of $X = K \leq T$	Perfect multicollinearity

After this brief introduction and definition of our topic, now we can dig more into the subject by looking at the causes, tests and remedies for endogeneity. To do that, we will first discuss the bias that is causing endogeneity, then we will give the tests to identify the bias, wherever applicable, and finally we will present the solutions or methods to eliminate the bias for each of the violations in the following sections. Therefore we will start with the violation of the first assumption, omitted variable bias.

[2] Table is taken from Kennedy (1998)

1.2. Omitted Variables and Endogeneity

According to Jargowsky (2002), *"omitted variable bias is the difference between the expected value of an estimator and the true value of the underlying parameter due to failure to control for a relevant explanatory variable or variables."* In other words, when one or more variables that is supposed to be included in a model is left out, our estimation is likely to be in error.

Consider our regression model:

$$Y = X\beta + \varepsilon \qquad \text{(Eq.1)}$$

where Y is the dependent variable, X is the vector of independent (explanatory) variables, β is the vector of coefficients and ε is the unobservable random error term. As we briefly discussed above, our classical regression model in Equation 1 assumes that the regression equation represents the "population model" or sometimes called the "true model" constructed by a random sample of n observations (Wooldridge, 2002). Therefore, the assumption states that all the relevant variables are included in the regression equation to represent the "true nature" of the model. However, in reality, there might be cases that we misspecify the model by including an irrelevant variable to the model (overspecification), or more seriously, excluding a relevant variable from the model. Including an irrelevant variable has no effect on the unbiasedness of our estimator, whereas omitting a variable has serious consequences as it is shown below.

Following Clarke (2005) and Hanushek and Jackson (1977), let's suppose that our "true" regression model takes the form:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i \quad \text{(Eq. 2)}$$

While we misspecify the model by omitting variable X_4 in the following way:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon^*_i \quad \text{(Eq. 3)}$$

Therefore, our error term now becomes:

$$\varepsilon^*_i = \beta_4 X_{i4} + \varepsilon_i, \quad \text{(Eq.4)}$$

Suppose that we are interested with the estimation results for β_2. Since the expected value of the true error term is assumed to be zero, the expected value of estimated coefficient on X_2 will be:

$$E(\beta_2) = \beta_2 + \beta_4 b_{42} \quad \text{(Eq.5)}$$

where b_{42} is the regression coefficient on X_2 in the "auxiliary" regression of the excluded variable, X_4, on the included variables, X_2 and X_3 (Clarke, 2005). Therefore, the classical regression model results would be biased and the estimation will be inconsistent in a way that our estimation independent variables X_2 will be correlated with the error term as it is shown above. And this, in turn, means that one of our independent variables becomes endogeneous. Because of the bias, the variance covariance matrix of the coefficients becomes smaller (unless the omitted variable is orthogonal to the included variables (Kennedy, 1998)), and therefore inferences using these parameters will be inaccurate as the variance of the error term will be biased upward.

There are other cases that omitted variable bias may occur. One specific case, functional form misspecification, occurs when the omitted variable is a function of another explanatory variable in the model (Wooldridge, 2002).

Omitted variable bias is a serious problem and theoretically it is very easy to determine and solve the problem as it is assumed that true model is known. However in estimations with empirical data, determining the omitted variable is not a straightforward task (Clarke, 2005). In the literature, some tests and solutions for the omitted variables are offered. Ramsey (1969)'s seminal paper on the Regression Specification Error Test (RESET) is very handy in most of the cases such as omitted variables, incorrect functional form and correlation between dependent variables and the error term (Thursby and Schmidt, 1977). Ramsey RESET test is basically and F-statistic that is calculated by adding polynomials, such as square or cube, of the fitted dependent variable (that is the square of the cube of the estimated values of the dependent variable) to the original regression equation in order to detect different functional forms. If any non-linear combination of the independent variables is found to be powerful in explaining the dependent variable, then there is a misspecification in the model. When our general model takes the form in Equation 1, we can construct RESET test as follows:

Suppose that we decided to add squares and the cubic forms of the estimated dependent variable to the model. Then we need to re-estimate Eq.1 by using:

$$Y = X\beta + \beta_{k+1}\hat{y}^2 + \beta_{k+2}\hat{y}^3 + \varepsilon \qquad \text{(Eq. 6)}$$

where ŷ represents the previous estimation of the dependent variable with the original equation and it is assumed that there are k independent variables in the original equation. That is, Equation 1 is re-estimated using square and the cube of the estimated values of dependent variable. After this calculation, RESET test basically employs an F test to check whether β_{k+1} and β_{k+2} are jointly meaningful or not. That is, under the null hypothesis:

$$H_0: \beta_{k+1,}\, \beta_{k+2} = 0$$

$$H_A: \beta_{k+1,}\, \beta_{k+2} \neq 0$$

and the test statistic uses the determination coefficient, or R^2, from both equaiton and the following test statistic is computed:

$$[(R^2_1 - R^2_0)/(k-1)]/[(1 - R^2_1)/(n-k)],$$

where R^2_1 is the R^2 from Ramsey's auxiliary regression and R^2_0 is the original regression's R^2. If the null hypothesis that all regression coefficients of non-linear terms are zero is rejected, then the model suffers from misspecification.

One drawback of the RESET test is that it is a general specification test and it is not possible to clearly distinguish between omitted variables and the functional misspecification by using RESET (Wooldridge, 1995). Thursby and Schmidt (1977) improved the RESET test by combining it with other tests to aid in specification (Kennedy, 1998).

One other method to determine omitted variables bias is a variant of Hausman (1978) specification test. Hausman (1978) basically employs a chi square distribution for testing model specification. Suppose, in our regression model, for our vector of coefficients, β, we have two different sets of coefficients constructed using two different specification, namely β^0 and β^1. Hausman test asks which set has consistent and efficient estimators. Under the null hypothesis, both sets are consistent, but only β^1 is efficient, and under the alternative hypothesis β^0 is consistent and β^1 is not. Therefore the Hausman test statistic is:

$$H = (\beta^1 - \beta^0)' \, [\mathrm{Var}\,(\beta^1) - \mathrm{Var}(\beta^0)]^{-1} \, (\beta^1 - \beta^0) \qquad \text{(Eq. 7)}$$

where $'$ denotes the transpose of the difference matrix and this test is distributed as a chi-square with a degrees of freedom equal to the number of elements in the coefficient vector. This test is used for many specification purposes, not only with omitted variable case[3]. A specific version of Hausman test, which employs instrumental variables in the estimation of the second specification (β^1), is called the OV, or omitted variables, version of the Hausman test. Equivalent forms of the OV version of Hausman test use the errors from a regression of X on the instruments, or the estimated X from this regression, as the omitted variable (Kennedy, 1998). One handicap with Hausman test is that it is sensitive to several types of misspecification. Therefore, Godfrey and

[3] We will return to Hausman test when we are discussing Instrumental Variables approach in order to eliminate the endogeneity problem.

Hutton (1994) recommend testing for general misspecification before applying the Hausman test, and they develop a general misspecification test in their paper. Also, Wong (1996) finds that bootstrapping the Hausman test improves its performance.

If an omitted variable is determined by using several methods, briefly presented above, we can include the relevant variable to the equation by using several methods. This includes using economic intuition to add a new variable, prior empirical research, using proxies for the missing variables if there is no available data for the omitted variable (Wooldridge, 2002) or using more sophisticated methods such as instrumental variables[4] (IV) in order to eliminate the effects of the omitted variable bias. Wooldridge (2002) suggests four different ways to deal with omitted variables: (1) we can ignore the problem and suffer the consequences of biased and inconsistent estimators; (2) we can try to find and use a suitable proxy variable for the unobserved variable; (3) we can assume that the omitted variable does not change over time and use the fixed effects or first-differencing methods if we face the omitted variable problem in panel data; and finally (4) we can use IV approach which leaves the unobserved variable in the error term, but rather than estimating the model by OLS, it uses an estimation method that recognizes the presence of the omitted variable.

One should note however that, in reality, due to lack of data, or measurement problems of variables and such, the omitted variable bias is not

[4] Instrumental Variables method requires more detailed explanation and therefore it will be given specific attention in the following section.

easy to deal with. According to Jargowsky (2002), the definitive solution for omitted variable bias is to conduct a classical experiment in which individuals are randomly assigned to treatment and control groups. However, social sciences do not always allow for such experiments and thus, omitted variable is a possibility and probability in social sciences despite the sincere efforts of the researcher. A more sophisticated method of IV can be used in some cases, which we will turn back in the following sections.

1.3. Measurement Errors and Autocorrelation

In the empirical studies, it is not always possible to collect the data that is truly reflecting the underlying economic theory due to many reasons including data fudging. Even worse than this, in some cases even the researcher may not be aware of the true meaning of the issue at hand (Streissler, 1970). Therefore, researchers sometimes have to use an imprecise measure of an economic variable in an econometric model (Wooldridge, 2002), and this leads to a measurement error, or as known as “errors-in-variables problem”. Measurement errors, in some situations, just like in the omitted variable bias case, may cause the explanatory variables to be correlated with the random error term and therefore would inconsistent regression results.

As we stated in the beginning, the fourth assumption of the linear regression model specifies that the observations of the explanatory variables are considered fixed in repeated samples, that is, we assume fixed regressors.

However in reality, this is not the case due to either measurement errors in the regressors, or due to the necessity of using lagged dependent variables in the model, which leads to autocorrelation in the regression equation. According to Angrist and Krueger(2001), measurement error can occur for many reasons, including the limited ability of statistical agencies to collect accurate information and the deviation between the variables specified in economic theory and practice. The measurement error problem is a very well and long-known issue in econometrics that may be dated back as much as 1920s. A well-known quotation, by Josiah Stamp reflects the nature of the measurement problem very nicely:

The Government are very keen on amassing statistics - they collect them, add them, raise them to the nth power, take the cube root and prepare wonderful diagrams. But what you must never forget is that every one of those figures comes in the first instance from the village watchman, who just puts down what he damn pleases (Stamp, 1929, quoted in Kennedy (1998)).

There are books and papers that investigate the problem of mismeasurement and data accuracy. In other words, measurement errors are concerned with the effects of using incorrectly measured variables on the econometric studies. Due to the nature of the regression equation, the errors in the measurement of the dependent variable poses no threat on the consistency of the regression due to the fact that the error term itself incorporates the measurement errors in the dependent variable by definition (Kennedy, 1998; Wooldridge, 2002). However, when there are measurement errors in the

independent variables, the fourth assumption above will be violated and the regressors will no longer be fixed, they will rather be stochastic. In other words, when there is random errors in measuring an explanatory variable, then the OLS estimation of the coefficient on that variable will be biased toward zero (Angrist and Krueger, 2001). Therefore, the distribution of the variable will be correlated with the disturbance term and this will create a bias in the classical regression model as shown below.

Suppose our true regression equation is again in the form of Equation 1 with one explanatory variable:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \varepsilon_i \qquad \text{(Eq. 8)}$$

but with a difference that, our explanatory variable X is measured with an error z such that:

$$z_i = X_{i2} + \upsilon_i$$

this implies that our explanatory variable can be written as $X_{i2} = z_i - \upsilon_i$ and our original regression takes the form:

$$Y_i = \beta_1 + \beta_2(z_i - \upsilon_i) + \varepsilon_i, \text{ or } Y_i = \beta_1 + \beta_2 z_i + \varepsilon_i - \beta_2 \upsilon_i \qquad \text{(Eq.9)}$$

as can be seen from equation 8, now the original regression is biased and there is a correlation between the error term and the explanatory variable, which would cause an endogeneity problem that needs to be taken care of in order to produce consistent estimator.

Therefore the question arises here that how can one determine if there is a contemporaneous correlation between the error term and the explanatory variables due to a measurement error? A version of Hausman (1978) test is again widely used for this purpose. Since the consistency of the Instrumental Variable estimator is not affected by the measurement error[5], classical regression estimation and the IV estimation should produce similar results under the null hypothesis given in equation 7. That is, if there is no measurement error in the classical regression model, then its results will be identical to those from the IV model, otherwise, the results will be different. Hausman test is a test of statistical equality of the regression results and therefore it can be used to capture the measurement error in the classical regression model.

When there is a contemporaneous correlation between the explanatory variable(s) and the error term, the solution is usually to use an alternative estimator with desirable asymptotic properties as estimators with desirable small-sample properties are hard to find in this case (Kennedy, 1998; Wooldridge, 2002). The most common estimator used in this context is IV estimator. IV estimator provide a consistent estimate in the presence of measurement error if it is possible to find an instrument that is correlated with the explanatory variable but uncorrelated with the measurement error. As this estimator is used for many different reasons, we need to discuss this estimator in more detail in the following section.

[5] Technical details for this proposition will be given in more detail in the following section.

1.4. Instrumental Variable Estimation

As it is discussed previous sections, IV estimation is used to correct for omitted variable bias, as well as measurement error and as we will see in the following pages, it is also used to correct for the simultaneity bias. The theory of the instrumental variables as a problem of identifying and estimating one or more coefficients of a system of simultaneous equaitons was first developed by Philip G. Wright in his 1928 book The Tariff on Animal and Vegetable Oils (Stock and Trebbi, 2003). Later the method of IV is used in many different areas in econometrics including our subject matter. The instrumental variable technique is a general estimation procedure applicable to situations where the independent variable(s) are correlated with the error term. If an appropriate instrumental variable can be found for each endogenous variable that appears as a regressor in a simultaneous equation, the instrumental variable technique provides consistent estimates (Kennedy, 1998). Therefore, it becomes very useful in solving such problems as endogeneity, omitted variables and so on. To explain how the instrumental variable estimation works, let's begin with its formal mathematical derivation in the simultaneous equations context. Remembering that our classical regression model given in Eq. 1 was, in matrix notation:

$$Y = X\beta + \varepsilon$$

where Y is the dependent variable, X is the vector of independent or explanatory variables (x), β is the vector of coefficients and ε is the random error

term, assumed to be normally distributed with a zero mean and finite variance. When we try to estimate this equation using ordinary least squares (OLS, classical reg) by using number of n observations, one can write the OLS estimator for the vector of coefficients as follows:

$$\beta_{OLS} = (X'Y/X'X) = X'(X\beta + \varepsilon)/X'X = \beta + (X'\varepsilon)/(X'X) \qquad \text{(Eq. 10)}$$

where β_{OLS} denotes the estimated coefficents vector and $'$ denotes the transpose of the vector. With our initial assumptions about the OLS, the X's and ε are uncorrelated as given in the Table 1 and therefore, the OLS estimator β_{OLS} is unbiased and consistent given the 5 assumptions hold. However, so far we have seen that there are several violations of these assumptions. Therefore let's now define an "instrumental variable". An instrumental variable, z, is a variable that is correlated with the independent variable x, that is Cov (z, x) ≠ 0 but not with the error term, ε, that is, Cov (z, ε) = 0. One should note that while the covariance condition on z and x means that z must be related to the endogenous explanatory variable x; the covariance condition between z and ε is interpreted differently in different contexts. In the case of omitted variables, for example, this means that z should have no partial effect on y and z should not be correlated with other factors that impact the dependent variable, y (Wooldridge, 2002). In the case of simultaneous equations, we interpret this condition as an "exogeneity condition" by saying z is exogenous to equation 1. Using our instrument and by employing method of moments, the conditional expectation of the dependent variable Y, on z can be calculated as:

$$E[Y|z] = \beta E[X|z] + E[\varepsilon|z] \qquad \text{(Eq.11)}$$

by definition, the last term, conditional expectation of ε on z is equal to zero. If we solve Eq.11 for b and write the resulting expression in terms of sample moments, we will get:

$$\beta_{IV} = (Z'Y/Z'X) = \beta + (Z'\varepsilon)/(Z'X) \qquad \text{(Eq. 12)}$$

since z and ε are uncorrelated, the last term in eq.12 approaches to zero in the limit, which mathematically provides an asymptotically consistent estimator for the coefficient vector.

Kennedy (1998) states that the major drawback to IV estimation is that the variance-covariance matrix of the IV estimator is larger than that of the OLS estimator, by an amount that is inversely related to the correlation between the instrument and the regressor. This is the price paid for avoiding the asymptotic bias of OLS. Another problem with IV approach is to find appropriate instrumental variables which are uncorrelated with the error term. Bound et.al. (1995) argues that the use of instrumental variable in the case of a weak relationship between instruments and the endogenous explanatory variable causes large inconsistencies in the IV estimates and in the finite samples IV suffers from the same bias as OLS does. Therefore they suggest the use of partial R^2 and F statistics on the excluded instruments in the first stage of the regression.

1.5. Back to the Measurement: Using IV Method to correct for the Measurement Error and Omitted Variable Bias

After explaining the how IV is constructed and how it works, now we can go back to our discussion on how to correct the endogeneity bias coming from omitted variables and the measurement error by using IV methodology.

In the case of omitted variables, after determining the existence of an omitted variable by the tests explained above, one can use IV method to correct the bias. As it is discussed above, in social sciences there are many factors relevant to an economic behavior in question, and it is highly possible to omit one related variable from the model. If all these variables can be measured and held constant in a regression, the omitted variable bias would be eliminated. However, in practice, it is almost impossible to measure all of the relevant variables (omitted variables); and measure them accurately (measurement error) even when they are specified correctly (Angrist and Krueger, 2001).

In order to see how IV corrects the omitted variables bias, let's consider our OLS equation given in equation 1 when there are two independent variables to explain the variations in the dependent variable:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \qquad \text{(Eq. 13)}$$

Suppose that we have a third variable, the instrument, denoted by z, which is correlated with X_{i2}, but otherwise unrelated to the dependent variable, Y. That is z is uncorrelated with the omitted variables and the regression error, ε. Then an IV estimator estimate of the dependent variable is the sample analog of

$Cov(Y_i, z_i) / Cov(X_{i2}, z_i)$. IV method here allows us to estimate the coefficient of interest consistently and free from asymptotic bias from omitted variables, without actually having data on the omitted variables (Angrist and Krueger, 2001), because of the fact that we can now estimate our coefficient of interest β_2.

$$\text{Cov}(Y_i, z_i) = \beta_2 \text{Cov}(X_{i2}, z_i) + \text{Cov}(z_i, \varepsilon_t) \qquad \text{(Eq.14)}$$

since the last term in equation 14 is zero by definition, and the middle term is different from zero, we can solve for β_2 as:

$$\beta_2 = \text{Cov}(Y_i, z_i) / \text{Cov}(X_{i2}, z_i)$$

after cancelling the sample sizes in the numerator and the denominator, we get the instrumental variables estimator for β_2:

$$\beta_{2,IV} = [\Sigma(z_i - \bar{z})(y_i - \bar{y})] / [\Sigma(z_i - \bar{z})(y_i - \bar{x})] \qquad \text{(Eq.15)}$$

where the bar denotes the mean of the variable in the sample. It can be easily shown by using law of large numbers that IV estimator is consistent for our coefficient of interest provided that our assumptions about correlations are satisfied (Wooldridge, 2002). Therefore, instrumental variable approach is successfully solves the omitted variable bias.

With IV approach the most important potential problem arises if the instrument is “bad”, that is if the instrument is correlated with the omitted variables (or the error term in the original equation) (Bound et.al, 1995). Another

problem with this approach is the possibility of bias when instruments are weakly correlated with the endogenous explanatory variable. In order to eliminate weak instruments problem, several solutions are proposed in the literature, including Bound et.al (1995)'s partial R^2 and reducing the number of instruments in the equation as the bias is proportional to the degree of overidentification (that is if there are K instruments and G endogenous variables, the bias is proportional to K – G, and reducing K would decrease the bias.) (Angrist and Krueger, 2001).

In terms of the measurement error, the use of instrument variables follows almost the same steps as in the case of omitted variables with some differences. When there is a measurement error, one can include the lagged value of independent variable as it is not contemporaneously (pair-wise) correlated with the error term, with the assumption that error term is not autocorrelated (Kennedy, 1998). Another method used in the literature is the "two-group method" as suggested by , in which the observations are split into two equalsized groups on the basis of the size of the regressor and then the slope coefficient is estimated by the line joining the arithmetic means of the two groups, can be interpreted as an instrumental variables estimator with the instrumental variable taking the value -1 if the regressor value is below its median value and +1 if above its median value. The rationale behind this method is that by averaging the data in this way the measurement errors are also averaged, reducing their impact (Johnston, 1984). After investigating the problem of the omitted variables bias, and measurement errors, and the

solutions to them by using IV approach, now we can further analyze the issue of endogeneity in the case of simultaneity bias, which will be done in the next section.

1.6. Simultaneity

Simultaneity occurs in many areas in economic research due to the nature of the economics, as economic variables are highly interrelated with each other. Simultaneity can be defined as the joint determination of one or more explanatory variables with the dependent variable, typically through an equilibrium mechanism (Wooldrige, 2002). In order to deal with the simultaneity issue intrinsic in the nature of economic data, simultaneous equation models has long been developed in econometric theory. The development of the notion of simultaneity goes back as early as the work of E. Working's seminal paper "What do 'Statistical Demand Curves' Show" where he gave an early account of identification problem, while Haavelmo is the first one who realized that in the presence of jointly endogenous variables, a joint probability distribution is necessary to analyze the data (Hausman, 1984). Koopmans (1949), Koopmans and Reiersol (1950) and Koopmans, Rubin and Leipnik (1950) established conditions for identification in linear simultaneous models. Later in the literature, many other models dealing with simultaneous equations are developed and used in empirical research.

The joint endogeneity concept is the principle behind the simultaneous equation models. Therefore, endogeneity is the key concept in most simultaneous equation models, even though not all simultaneous equations systems suffer from the simultaneous equation estimation bias[6]. However in most cases, simultaneity is creating a bias needs to be solved in order to attain consistent estimators. To understand the issue more clearly, let's begin with simultaneous equaitons.

In a system of simultaneous equations, all the endogenous variables are random variables since a change in any error term changes all the endogenous variables since they are determined simultaneously. A most widely used example in textbooks of econometrics to explain the simultaneity issue is the Keynesian system with a consumption function[7]:

$$C = a + bY + \varepsilon \qquad \text{(Eq.16)}$$

and an equilibrium condition

$$Y = C + I \qquad \text{(Eq.17)}$$

[6] A good example for this is Recursive Estimation. A recursive system is one in which there is unidirectional dependency among the endogenous variables. In a recursive system, a change in the disturbance in the fourth equation, for example, affects directly the fourth endogenous variable, which in turn affects the higher-ordered endogenous variables in the system, but does not affect the lower-ordered endogenous variables. Because only lower-ordered variables appear as regressors in the fourth equation, there is no contemporaneous correlation between the disturbance and the regressors in the fourth equation. If there is no correlation between disturbances in different equations, OLS estimation is consistent, and if no lagged endogenous variables appear among the exogenous variables in the equation, it is unbiased (Kennedy, 1998). However, this is only a special case, and therefore it is not discussed in the text above.

[7] This example is taken from Kennedy (1998), p.158-159

where C, consumption and Y, income are endogenous variables and I, investment is an exogenous variable. Consider the problem of estimating the consumption function, regressing consumption on income. Suppose the error term in the consumption function increases. This directly increases consumption, this in turn will increase income through the equilibrium condition. But income is the independent variable in the consumption function. Thus, the disturbance in the consumption function and the regressor are positively correlated. An increase in the disturbance term (directly implying an increase in consumption) is accompanied by an increase in income (also implying an increase in consumption). When estimating the influence of income on consumption, however, the OLS technique attributes both of these increases in consumption (instead of just the latter) to the accompanying increase in income. This implies that the OLS estimator of the marginal propensity to consume is biased upward, even asymptotically.

As this example shows very clearly, the endogeneity bias is intrinsic in the system of simultaneous equations with a few exceptions (see the previous footnote). In other words, simultaneity includes endogeneity by its nature. Therefore it is important to present technical details of the bias to understand the underlying structure that leads to inconsistency in the classical regression model (OLS). To this end, we will use a basic two-equation structural model[8]:

$$Y_1 = \beta_1 Y_2 + \beta_2 X_1 + \varepsilon_1 \qquad \text{(Eq.18)}$$

[8] For simplicity we assume that there is no intercept in the equations.

$$Y_2 = \beta_3 Y_1 + \beta_4 X_2 + \varepsilon_2 \qquad \text{(Eq.19)}$$

where Y's are endogenous explanatory variables, and X's are exogenous variables and ε's are the error terms for each equation. Now let's turn our focus on estimating the first equation. The variables X_1 and X_2 are exogenous, so that each is uncorrelated with ε_1 and ε_2.

Now we want to show that Y_2 is generally correlated with ε_1. If we solve the two equations for Y_2 in terms of the exogenous variables and the error term and plug in plug the right-hand side of Eq.18 in for Y_1 in Eq.19, we get

$$Y_2 = \beta_3(\beta_1 Y_2 + \beta_2 X_1 + \varepsilon_1) + \beta_3 X + \varepsilon_2, \textit{ or equivalently:}$$

$$(1-\beta_3\beta_1)Y_2 = \beta_3\beta_2 X_1 + \beta_2 X_2 + \beta_3\varepsilon_1 + \varepsilon_2 \qquad \text{(Eq.20)}$$

finally in order to solve for Y_2 we need to make an assumption, which states:

$$\beta_3\beta_1 \neq 1 \qquad \text{(Eq.21)}$$

and if we divide equation 20 by $(1-\beta_3\beta_1)$ we get:

$$Y_2 = \pi_{21} X_1 + \pi_{22} X_2 + v_2, \qquad \text{(Eq.22)}$$

where

$$\pi_{21} = \beta_3\beta_2/(1-\beta_3\beta_1),\ \pi_{22} = \beta_4/(1-\beta_3\beta_1), \textit{ and } v2 = [\beta_3\varepsilon_1 + \varepsilon_2/(1-\beta_3\beta_1)]$$

(Eq.23)

Equation 23 is simply called the reduced form for Y_2. π parameters are reduced form parameters, which are the nonlinear functions of the structural parameters ($\beta_1, \beta_2, \beta_3, \beta_4$) in the structural equations 18 and 19. When Y_2 is correlated with ε_1 because of simultaneity, then it is said that OLS suffers from simultaneity bias which leads to inconsistent estimation results. Since now we have a notion of the endogenetiy bias, we can start presenting the test for endogeneity and the solution to this bias. The solution to this bias is always to use an unbiased estimator such as IV, Two Stage Least Squares, Maximum Likelihood and so on. But before reaching to the estimation with an alternative estimator, we need to discuss the tests for identification, and for endogeneity and exogeneity within the system of simultaneous equations.

To begin with, let's first discuss the tests of exogeneity used in econometric literature. One of the most widely used test for exogeneity/endogeneity of the variables is a variety of Hausman test, a test for contemporaneous correlation between the error and regressors which is described in detail in the previous section. Suppose that one of our equations has the form:

$$y = \delta Y + \beta X + e$$

where δ is the coefficient vector for endogenous explanatory variables, Y, and we wish to test all the variables in Y for exogeneity. This is done exactly as in previous section, via an omitted variable version of the Hausman test. Estimated Y values, Y*, are formed from the instruments (all the exogenous variables in the system), y is regressed on Y, X and Y*, and the coefficient

estimate of Y* is tested against zero using an F test, given in equation 7. Instead of testing for all the variables in the endogenous variables vector, one may wish to test some of the variables for exogeneity. Technical details of this derivation is skipped here but we should note that, in this case, the Hausman testing methodology becomes more complicated, Spencer and Berk (1981) constructed a variety of Hausman test for this purpose.

After the tests for exogeneity, one needs to "identify" the model in order to consistently estimate the regression. As we discussed in the beginning of this section, OLS estimator is biased when there is a "joint endogeneity" among the variables at hand which poses a problem of finding an unbiased estimator for the coefficients of interest. A natural response to this estimating problem is to suggest that the simultaneous system be solved and put into its reduced form. This means that every endogenous variable is expressed as a linear function of all the exogenous variables (and lagged endogenous variables, which are considered exogenous in this context). For the simple Keynesian example given at the beginning, the structural equations in Equation 16 and 17 can be solved to give the reduced-form equations:

$$Y = (a/(1-b)) + (1/1-b)I + 1/(1-b)\,\varepsilon$$

$$C = (a/(1-b)) + (b/1-b)I + 1/(1-b)\,\varepsilon$$

which can be rewritten in more general form as:

$$Y = \pi_1 + \pi_2 I + v_1$$

$$C = \pi_3 + \pi_4 I + v_2$$

where the π are parameters that are (nonlinear) functions of the structural form parameters and the ε are the reduced-form disturbances, functions of the structural form disturbances. Because no endogenous variables appear as independent variables in these reduced-form equations, if each reduced-form equation is estimated by OLS, these estimators of the reduced-form parameters, the p, are consistent (and if no lagged endogenous variables appear among the exogenous variables, these estimators are unbiased). Economic theory tells us that these reduced-form parameters are the long-run multipliers associated with the model. If a researcher is only interested in predicting the endogenous variables, or only wishes to estimate the size of these multipliers, he can simply use these estimators. If, however, he is interested in estimating the parameter values of the original equations (the structural parameters), estimates of the reduced-form parameters are of help only if they can be used to derive estimates of the structural parameters. However, this is not always possible; this problem is one way of viewing the identification problem.

If an equation is identified, it may be either "just-identified" or "over-identified." An equation is just-identified if the number of identifying restrictions placed on the model is the minimum needed to identify the equation; an equation is over-identified if there are some extra restrictions beyond the minimum necessary to identify the equation. After the identification of the model

is done with certain conditions, we can apply an appropriate method to solve the simultaneity bias.

In the literature, there are many different methods to estimate a simultaneous equation. Estimators that estimate the equations individually within the system of simultaneous equations are called Single equation methods, while estimators that estimate the equations in the system simultaneously are called system methods. Single equation models are usually called “limited information” methods while the system methods. Since each of these methods are the subject of a new paper, and they are widely used and known, we will just list these variables and conclude this section. There are 5 very well-known single equation methods and 2 system methods in the literature.

Single Equation Methods:

1. Ordinary Least Squares OLS
2. Indirect Least Squares ILS
3. Instrumental Variables IV
4. Two-stage Least Squares 2SLS
5. Limited information Maximum Likelihood LI/ML

System Methods

1. 3 Stage Least Squares, 3SLS

2. Full information Maximum Likelihood (FIML).

1.7. Further Studies About Correcting For Endogeneity

We can describe other methods correcting Endogeneity in choice models. Several methods have been developed recently to estimate choice models in the presence of endogenous variables. In this frame, we can discuss three approaches: BLP approach, Control Function approach and Maximum Likelihood approach. In this chapter, these approaches will be considered briefly and we will give some technical analyses as sub chapters.

We first describe the BLP approach, developed by Berry, Levinsohn and Pakes (hence the initials) through a series of publications. It was pointed out that constants can be included in the choice model to capture the average effect of the product attributes (both observed and unobserved). Then, the estimated constants can be regressed against the observed attributes in a linear regression, where endogeneity is handled in the usual way by instrumental variables estimation of the linear regression. Essentially, it has been showed that the endogeneity could be taken out of the choice model, which is inherently non-linear, and put into a linear regression model, where endogeneity can be handled through standard instrumental variables estimation. To apply this method, it is often necessary to estimate a very large number of constants in the choice model, which can be difficult using standard gradient-based methods for maximization.

The second procedure that we describe is the control function approach. The concepts motivating this approach date back to Heckman (1978) and Hausman (1978), though the first use of the term “control function” seems to have been by Heckman and Robb(1955). As we stated above endogeneity arises when observed variables are correlated with unobserved factors. This correlation implies that the unobserved factors conditional on the observed variables do not have a zero mean, as is usually required for standard estimation. A control function is the variable that captures this conditional mean, essentially “controlling” for the correlation. The procedure is implemented in two steps. First, the endogenous explanatory variable (such as price) is regressed against exogenous variables. The estimated regression is used to create a new variable (the control function) that is entered into the choice model. And then, the choice model is estimated with the original variables plus the new one, accounting appropriately for the distribution of unobserved factors conditional on both this new and the original variables.

The third procedure is maximum likelihood approach, as applied by Villas-Boas and Winer (1999) to a multinomial logit with fixed coefficients and generalized by Park and Gupta (2008) to random coefficient choice models. The procedure is closely related to the control function approach, in that it accounts for the non-zero conditional mean of the unobserved factors. However, instead of implementing the two steps sequentially (i.e., estimate the regression model to create the control function and then estimate the choice model with this control function), the two steps are combined into a joint estimation criterion.

Additional assumptions are required to allow the estimation to be performed simultaneously; however, the procedure is more efficient when those assumptions are met.

1.7.1 BLP Approach

The approach we employ here to correct for endogeneity is that proposed by Berry, Levinsohn and Pakes (1995 and 2004). Their approach is relevant when the endogeneity can be considered at a market level, applying similarly to decision makers within a given market. In our case, the markets are peer groups, where the groups are defined based on spatial proximity and social class. Each spatial and social group then, for the sake of the BLP procedure, can be considered a market.

The BLP procedure involves decomposing the error into two parts: the endogenous-causing part and the random portion. Let the utility equation be:

$$U_{in} = V(x_{in}, s_n; \beta) + \gamma F_{in} + \ddot{\varepsilon}_{in} + \dot{\varepsilon}_{in}$$

where $\ddot{\varepsilon}_{in}$ is correlated with F_{in}, and $\dot{\varepsilon}_{in}$ is uncorrelated with F_{in} (and x_{in}, s_n)

One key to the BLP procedure is to isolate the endogenous-causing components, that is, F_{in}, and $\ddot{\varepsilon}_{in}$. The terms are thus rearranged as follows:

$$U_{in} = [\gamma F_{in} + \ddot{\varepsilon}_{in}] + V(x_{in}, s_n; \beta) + \dot{\varepsilon}_{in}$$

The first term $[\gamma F_{in} + \ddot{\varepsilon}_{in}]$ represents the observable and unobservable components of utility relevant to the peer group. The second term $V(x_{in}, s_n; \beta)$

represents the remainder of the systematic utility of the individual (that is, the portion not related to the peer group). The error term $\dot{\varepsilon}_{in}$, by construction, is orthogonal to all explanatory variables in the model, including F_{in} and $\ddot{\varepsilon}_{in}$.

The second key to BLP is related to the setup of the procedure, which is the assumption that the endogeneity occurs at a market level. In our case, each peer group is a market and we add a market delineator m to denote the peer group to which a decision maker belongs. Modifying the utility equation accordingly leads to:

$$U_{in_m} = [\gamma F_{im} + \ddot{\varepsilon}_{im}] + V\left(x_{in_m}, s_{n_m}; \beta\right) + \dot{\varepsilon}_{in_m}$$

Thus, the first term $[\gamma F_{im} + \ddot{\varepsilon}_{im}]$ represents the unobservable and observable components of utility relevant to the individuals' peer group *m*. It represents the average, or common, utility of a given choice in a given group. This term varies across peer groups but does not vary across individuals in the same group. The second term $V\left(x_{in_m}, s_{n_m}; \beta\right)$ represents the systematic portion of utility that varies across decision makers. The error term $\dot{\varepsilon}_{in_m}$ is orthogonal to all explanatory variables and varies across decision makers.

The trick in the BLP procedure is to now replace the peer group effect with market specific constants α_{im} for each alternative *i* and each peer group *m* such that the new utility equation is:

$$U_{in_m} = \alpha_{im} + V\left(x_{in_m}, s_{n_m}; \beta\right) + \dot{\varepsilon}_{in_m}$$

where $\alpha_{im} = [\gamma F_{im} + \ddot{\varepsilon}_{im}]$

These constants capture the average effects of the peer group. There is no endogeneity issue in the choice model as written this way, and therefore the parameters αim and β are estimated via usual choice modeling procedures. (A very large number of markets may require that the constants be estimated via the "contraction" approach described in BLP, although our application did not require this.) Note, though, that we are interested in the social effect as represented by the parameter γ, which is not estimated via the choice model.

The final step of the BLP procedure is to estimate via linear regression the market-specific constants as explained by the field effect variable, or:

$$\alpha_{im} = \gamma F_{im} + \ddot{\varepsilon}_{im}$$

While the endogeneity issue remains (Fim is correlated with $\ddot{\varepsilon}im$), it is more straightforward to correct for endogeneity in the linear model. For this two-stage instrumental variables approach can be used. In the first stage, instrumental variables Iim (correlated with the field effect variable Fim and uncorrelated with the error $\ddot{\varepsilon}im$) are used to explain the field effect variable Fim as follows:

$$F_{im} = \theta_i + \theta_F I_{im} + v_{im},$$

where v_{im} is a random error (orthogonal to I_{im}) and θi and θ_F are estimated parameters.

In the second stage, the market-specific constants are regressed on the fitted value of the field effect from the first stage $\hat{F}_{im} = \hat{\theta}_i + \hat{\theta}_F I_{im}$ as follows:

$$\alpha_{im} = \gamma_i + \gamma_F \hat{F}_{im} + \ddot{\varepsilon}_{im},$$

As $\hat{F}_{im}$ is orthogonal to $\ddot{\varepsilon}_{im}$, this regression results in a consistent estimate of γ_F, which captures the effect of the field effect variable on the utility. This can then be inserted back into the choice model so that the choice model captures the effect of the peer group.

In summary, the BLP process removes the endogeneity from the choice model via the use of market-specific constants. The endogeneity is then dealt with in a linear regression setting (with instrumental variables) to obtain consistent estimates of the social influence effect. This consistent estimate of the field effect parameter is then reintroduced to the choice model to obtain a choice model that captures social influences. (Walker, Ehlers, Banerjee, Dugundji, 2010).

1.7.2 Control Functions

Most models that are linear in parameters are estimated using standard IV methods either two stage least squares (2SLS) or generalized method of moments (GMM). An alternative, the control function (CF) approach, relies on the same kinds of identification conditions. In the standard case where endogenous explanatory variables appear linearly, the CF approach leads to the usual 2SLS estimator. But there are differences for models nonlinear in endogenous variables even if they are linear in parameters. And, for models nonlinear in parameters, the CF approach offers some distinct advantages.

Let *y*1 denote the response variable, *y*2 the endogenous explanatory variable (a scalar for simplicity), and **z** the 1 X L_1 vector of exogenous variables (which includes unity as its first element). Consider the model:

$$y_1 = z_1\delta_1 + \alpha_1 y_2 + u_1$$

where z_1 is a 1 *X* L_1 strict subvector of z that also includes a constant. The sense in which z is exogenous is given by the *L* orthogonality (zero covariance) conditions.

$$E(\grave{z}u_1) = 0$$

this is the same exogeneity condition used for consistency of the 2SLS estimator, and we can consistently estimate δ_1 and $\alpha_1$by 2SLS.

Just as with 2SLS, the reduced form of y_2 – that is, the linear projection of y_2 onto the exogenous variables – plays a critical role. Write the reduced form with an error term as

$$y_2 = z\pi_2 + v_2$$

$$E(\grave{z}v_2) = 0$$

where π_2 is 1 *X* L_1. Endogeneity of *y*2 arises if and only if *u*1 is correlated with *v*2. Write the linear projection of *u*1 on *v*2, in error form, as

$$u_1 = \rho_1 v_2 + e_1$$

Where $\rho_1 = E(v_2u_1)/E(v_2^2)$ is the population regression coefficient. By definition, $E(v_2e_1) = 0$ and $E(\grave{z}e_2) = 0$ because u_1 and v_2.are both uncorrelated with z. Then we have,

$$y_1 = z_1\delta_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1$$

where we now view v_2 as an explanatory variable in the equation. As just noted, e_1 is uncorrelated with v_2 and z. Plus, y_2 is a linear function of z and v_2, and so e_1 is also uncorrelated with y_2.

Because e_1 is uncorrelated with z_1, y_2, and v_2: run the OLS regression of y_1 on z_1, y_2, and v_2 using a random sample. The only problem with this suggestion is that we do not observe v_2; it is the error in the reduced form equation for y_2. Nevertheless, we can write $v_2 = y_2 - z\pi_2$ and, because we collect data on y_2 and z, we can consistently estimate π_2 by OLS. Therefore, we can replace v_2 with $\hat{v}_2$, the OLS residuals from the first-stage regression of y_2 on z. Simple substitution gives

$$y_1 = z_1\delta_1 + \alpha_1 y_2 + \rho_1 \hat{v}_2 + error,$$

where, for each *i*, $error_i = \rho_i z_i(\hat{\pi}_2 - \pi_2)$, which depends on the sampling error in $\hat{\pi}_2$ unless $\rho_1 = 0$. Standard results on two-step estimation imply the OLS estimators from will be consistent for$\delta_1, \alpha_1 \; and \; \rho_1$.

The OLS estimates, $y_1 = z_1\delta_1 + \alpha_1 y_2 + \rho_1 \hat{v}_2 + error,$ are control function estimates. The inclusion of the residuals $\hat{v}_2$ "controls" for the endogeneity of y_2 in the original equation.

Basically, we can conclude that the CF approach, while likely more efficient than a direct IV approach, is less robust. The CF estimator will be inconsistent in cases where the 2SLS estimator will be consistent. On the other hand, because the CF estimator solves the endogeneity of *y2* by adding the scalar $\hat{v}_2$ to the regression, it will generally be more precise than the IV estimator. (Imbens/Wooldridge, Lecture Notes, 2007).

1.7.3 Maximum Likelihood Approach

In the maximum likelihood approach, the parameters of the model are estimated simultaneously rather than sequentially. Now, let us consider the two equations:

$$U_n = V(y_n, x_n, \beta_n) + \varepsilon_n$$

$$y_n = W(z_n, \gamma) + \mu_n$$

Rather than specifying the conditional distribution of ε_n given μ_n, the researcher specifies their joint distribution, denoted $g(\varepsilon_n, \mu_n)$. So the joint distribution of ε_n and y_n is $g(\varepsilon_n, y_n - W(z_n, \gamma))$. Denote the chosen alternative as *i.* The probability of the observed data for person *n* is the probability that the endogenous explanatory variable takes the value y_n and that alternative *i* is chosen. Conditional on β_n, this probability is:

$$P_n(\beta_n) = \int I(U_{ni} > U_{nj}\ \forall j \neq i)\ g(\varepsilon_n, y_n - W(z_n, \gamma))d\varepsilon_n$$

If β_n is random, then $P_n(\beta_n)$ is mixed over its distribution. The resulting probability P_n is inserted into the log-likelihood function: $LL = \sum_n \ln(P_n)$ .This *LL* maximized over the parameters of the model. So the basic idea, instead of estimating y_n first and using the residuals in the choice probability, the parameters of y_n and the choice model are estimated simultaneously.

In general, the maximum likelihood approach requires a specification of the joint distribution of ε_n and μ_n. Any joint distribution implies a particular conditional distribution, but any given conditional distribution does not necessarily imply a particular joint distribution. There may be numerous joint distributions that have the specified conditional distribution. Hence, if the joint

distribution can be correctly specified, the maximum likelihood approach is more efficient, simply by the fact that it is the maximum likelihood for all the parameters.

CHAPTER II

CONCLUSION

In this study, we investigated one of the most important issues in econometric theory and in empirical studies, the endogeneity bias and its implications, the methods to test for it and the solutions for this bias in order to construct a consistent estimator for the economic problem that we are interested. We saw that, endogeneity occurs due to the misspecification, omitted variables, measurement errors and simultaneity biases. We defined all these cases within the context of endogeneity and we presented tests to identify these biases and finally we looked at the literature for the solutions to the problems for every case. We have seen that endogeneity arises not only because of the “mistakes” in econometric research or data collection, but sometimes it is intrinsic to the model specification. We also noted that it is very important to eliminate the endogeneity bias in order to have consistent econometric results.

BIBLIOGRAPHY

Angrist, J.D. and A. B. Krueger, (2001), “Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments”, *Journal of Economic Perspectives,* 15, 69-85

Bound, J., D. A. Jaeger and R. M. Baker (1995), “Problems with Instrumental Variables Estimation when Correlation Between the Instruments and the Exogenous Variable is Weak”, *Journal of American Statistical Association*, 90(430), 443-450.

Clarke, K. A., (2005), “ The Phantom Menace: Omitted Variable Bias in Econometric Research*”, Conflic Management and Peace Science*, 22, 341-352

Godfrey and Hutton (1994) “Discriminating between Errors-in-Variables/Simultaneity and Misspecification in Linear Regression Models”, *Economics Letters*, 44, 35964.

Granger, C., “Investigating Causal Relationships by Econometric Models and Cross-Spectral Methods” *Econometrica,* 37, 1969, p. 424-438

Hanushek, E. A. and J. E. Jackson (1977), Statistical Methods for Social Scientists. New York: Academic Press.

Hausman, J. A., (1977) “Errors in Variables in Simultaneous Equation Models”, *Journal of Econometrics*, 5 (3), 389-401.

Hausman, J. A., (1983), “Specification and Estimation of Simultaneous Equation Models”, Chapter 7 in Handbook of Econometrics, vol. 1, eds. Z. Griliches and M.D.Intriligator, North Holland.

Johnston, J. (1984), Econometric Methods, 3rd edn. New York: McGraw-Hill.

Paul A. Jargowsky (2002), “Omitted Variable Bias”, in The Encyclopedia of Social Measurement, ed.Kimberly Kempf-Leonard, Academic Press.

Kennedy, P., A Guide to Econometrics, MIT Press, 1998.

Ramsey, J.B. (1969) "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis", Journal of Royal Statistical Society, Series B., 31(2), 350–371

Sims, C., “Exogeneity and Causal Ordering in Macroeconomic Models” University of Minnesota Center for Economic Research Discussion Paper No:76-72.

Spencer, D. and K. Berk, (1981), “A Limited Information Specification Test”, Econometrica, 49(4), 1079-1085.

Stamp, J. (1929) Some Economic Factors in Modern Life. London: King and Son.

Stock, J. H., and F. Trebbi, (2003) "Retrospectives: Who Invented Instrumental Variable Regression?" Journal of Economic Perspectives, 17(3): 177–194.

Streissler, E., (1970) Pitfalls in Econometric Forecasting. London: Institute of Economic Affairs.

Thursby, J. G., Schmidt, P. (1977). "Some Properties of Tests for Specification Error in a Linear Regression Model". Journal of the American Statistical Association, 72, 635–641

Wong (1996) “Botstrapping Hausman's Exogeneity Test”, Economics Letters 53, 13943

Wooldridge, J. M., (2002), Introductory Econometrics: A Modern Approach, Southwestern.

Zellner, A., “Statistical Analysis of Econometric Models”, Journal of American Statistical Association, 74, 1979, p.628-643.

Walker Joan, Ehlers Emily, Banerjee Ipsita, Dugundji Elenna (2010) ,”Correcting for Endogeneity in Behavioral Choice Models with Social Influence Variables”