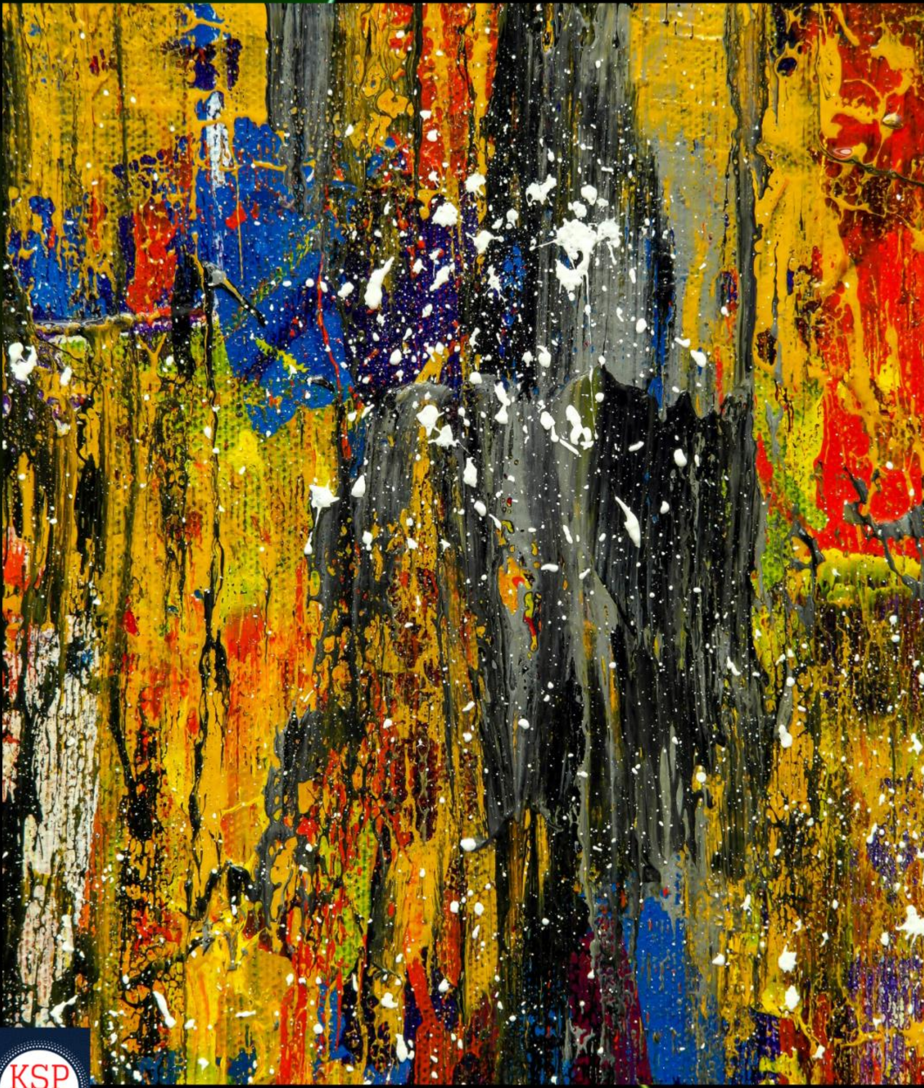
NOMAN ARSHED

Applied Cross-Sectional Econometrics



KSP BOOKS

Applied Cross-Sectional Econometrics

Noman Arshed
University of Management and Technology
Lahore, Pakistan

KSP Books
econsciences.com

Applied Cross-Sectional Econometrics

Noman Arshed

KSP Books
econsciences.com

ISBN: 978-605-7736-94-9 (e-Book)
KSP Books 2020
Applied Cross-Sectional Econometrics

Author: ***Noman Arshed***
Department of Economics, School of Business and Economics, University of Management and Technology Lahore Pakistan.

econsciences.com

Disclamier

Since econometrics is a study which builds and tests the quantified models of human behavior. Hence the author has used the help of Al Qur'an and Sunnah of Prophet Muhamad ﷺ as these narrations are addressing humans. Also it is the best source of case-based learning. Use the examples mentioned in Al Qur'an and Sunnah helps to build the background perspective, without it, the scenario building might be tedious.

Other than the above-mentioned purpose, there is no intention of the author to prove that statistics is originated from Al Qur'an or Al Qur'an needs the help of statistics to prove that it is the book of truth.

Acknowledgements

Needless to say, the structure and explanation of this book are improved by the students of my Applied Econometrics (MS Economics) class, who forced me to go to optimal depth for each model in order to make it simplified enough for non-econometric specialized research can grasp the concept and use the model. Secondly, I especially thank Dr. Rukhsana Kaleem (Dean, School of Business and Economics), Dr. Muhammad Naveed Tahir (Assistant Professor, Department of Economics), Dr . Muhammad Shahid Hassan (Assistant Professor, Department of Economics) and Afia Mushtaq (Assistant Professor, Department of Economics) who encouraged and supported me in the development of the course of Applied Econometrics for MPhil which later lead to the compilation of this book.

Preface

The project to write a book on applied Econometrics when I was granted with the Econometrics II course to teach to MPhil Economics, the contents of this book were developed from the lecture material which is competitive to course contents of The University of Edinburgh UK, a university where I did MSc in Economics, exploring other books, personal experience and the critical discussion by the students. The variety of contents which this book covers meet no competition with other universities in the city.

This book is an attempt to provide straight forward application based illustration of popular econometric models which are popular and available in the literature. I started this work with the idea that a research practitioner who is not versed with the basics of mathematics and statistics. He might not be able to understand the complex econometric model. This book provides firstly with some background to the model regarding what are the conditions which lead to this model selection. Secondly, basic mathematical derivations which are necessary for the concept. Lastly, STATA software-based example and its interpretation. The approach this book uses it that it delivers the concepts of the econometric models as well as it provides guidelines to use the STATA software using coding.

This book is especially designed for the MPhil / PhD students of all social science disciplines. And researchers who want to avail the skills of latest econometric models to be used in subjects like Sociology, Psychology, Finance and Banking.

This book uses a unique way to categorize the econometric models, which makes it different from other Econometrics books available in the market. In the first chapter, it provides an example of the simple regression model, and explains what information it provides and what information it lacks, the information which is lacking is called post regression issues in Econometrics. Unlike other econometric text books, it advocates the regression issues as missing information which model needs to incorporate rather than presenting them as a disease in a model. Then this book practically explains what each issue means and then categorizes the advanced model based on its incorporation (solution) to the regression issue.

This book constitutes of chapter 1, which provides brief and necessary background knowledge of Econometrics and regression analysis. The second part includes chapters 2–5, which are provided illustrations for the cross-sectional based models.

N. Arshed

Lahor, Pakistan

Reviews

There is a dearth of a book by local Pakistani authors on Econometrics in Pakistan. Advanced Cross-Section Models provide a fresh addition to the subject and will be an important reference for students who would like to learn about non-linear methodologies for data mining. The book covers both time series and panel data models that are the most used and popular empirical methodologies in Economics. I highly recommend this book to both practitioners and students of Economics.

Dr. Dawood Mamoon. Professor & Chairperson, Department of Economics, School of Business and Economics, University of Management and Technology Lahore, Pakistan

I have reviewed the book on Advanced Cross-Section Models written by Noman Arshed. The knowledge and application of econometric techniques are essential for research students. The book covers all the important topics in the search area at an advanced level. The book is written in a simple way and interesting style to make students comprehend the difficult concepts most easily. The practical examples are useful to grasp knowledge. The author has also made a commendable effort to add command of the software he used for analysis. The availability of data and command will assist students to learn quickly. The author discusses the sequences of models step by step and do not leave any aspect unexplained. There is well-written justification for the selection of models. The possible econometric problems have been well explained in the book. In my view, this is a very good addition to the books available in the market on the subject. I congratulate the author for producing a simple version of the advanced Econometrics in the simplest way.

Dr. Rukhsana Kaleem. Professor & Dean, School of Business and Economics, University of Management and Technology, Lahore, Pakistan.

The technically challenging aspects of applied econometric naturally abate its appeal and produce perpetual anxiety among economics' students. And despite ample quality work on the subject, the transmission of econometric knowledge with ease and excitement remains a depressing concern. This is where Mr. Noman's book strikes the most: it brings theory and techniques in such a brilliant way that completely levels econometric phobia. His work can just start you believing the possibility of being independent in the world of econometric analysis. Mr. Noman's innovative brilliance equally offers an ambitious opportunity for practitioners whose lacking empirical skills do not let them investigate several policy issues in depth.

Aqeel Ahmad. MS Economics, Department of Economics, School of Business and Economics, University of Management and Technology Lahore, Pakistan

Contents

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Modelling in the presence of multicollinearity 122

List of Tables

List of Figures

1 Introduction to regression analysis

Learning Outcomes

- History of regression analysis
- What regression actually does?
- What information regression provides and what it assumes for simplification?
- Interpreting regression coefficients

Introduction

When Adam and Eve were alone in this world, they enjoyed the blessing of Allah. Since they were only people who had the blessings, so there was no benchmark, because of that they were thankful unconditionally. Under such situations, where there is no reference, or simple benchmark statistics will suffice in explaining the behavior of 'being thankful to Allah'.

Following that, two sons of Adam argued with each other.When there a comparison made in terms of who is righteous (Al Qur'an 5:26-27). Similarly, there is an example of Iblis who compared humans and angels (Al Qur'an 2:34, 38:76). Under such conditions, where there is comparisons are made, relative statistics and correlations are used. The examples variable are 'righteousness' and 'superiority.'

Humans further evolved into worldly benefits. They went beyond comparison to dependency. It is precisely described in Al Qur'an (10:12) that man only prays when they are inflicted. Under such conditions, the actions are dependent on some conditions which are sometimes controllable and sometimes not. In the case of the complexity of human nature,where all of his actions are dependent on the prevailing situation, calls for the use of regression analysis.

So this example sums up the three levels of human behavior, because of evolution now every action has a prerequisite. And humans want to maximum his objectives. This has forced us to use the regression analysis approach when we are concerned with humans[1].

Why regression analysis

It is in human nature that the more the good is the more utility he receives. By comparing the empirical data of utility and the quantity, their co-movement highlight the fact that these events are proportional to each other in which utility is dependent upon the quantity of good. Such proportional associations can be equated using constant of proportionality.

$$Utility : Quantity$$
$$Utility = k * Quantity$$

Further mathematics helps in tracing the proportional relation into a line graph. This graphical illustration increase the degree of information in terms of intercept and slope. Figure 1 highlights the simple proportionality relationship whereby human can maximize utility by infinite quantity.

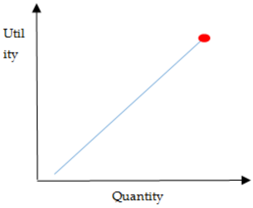
Further evolution in economics highlighted two aspects, first, is the concept of law of diminishing returns (shown in figure 2), forming a quadratic function. Which has forced us

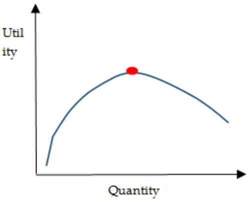
[1] If you look at the research work on physicians, or chemists you will note that they are on the first level as materials do not have thinking ability they do not compare or make their actions dependent on other materials.

to optimize the consumption pattern. This optimization of consumption is done by equating the utility quantity derivative to zero (i.e. Quantity = -j /2k).

$$Utility = j * Quantity + k * \ Quantity^2$$
$$\frac{d\ Utility}{d\ Quantity} = j + 2k * Quantity = 0$$
$$Quantity = \frac{-j}{2k}$$

The second aspect is the resource/income scarcity based constraints (shown in figure 3), forces us to optimize our decision under constraints. The Langrangian constraint optimization approach enables us to achieve utility optimization.





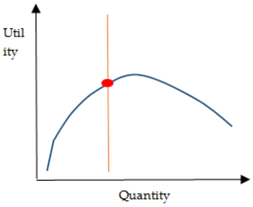


Figure 1. *Simple Proportionality*

Figure 2. *Optimal Consumption*

Figure 3. *Constraint Optimization*

Regression analysis

Regression analysis was first developed for application by Legendre (1805) and Gauss (1809). This method then formalized into theory by Gauss (1821). The word regress comes from a medical term which means that the ailment has retraced its steps and came back. Here the term regression analysis for social sciences can be explained as

> *Regression analysis is a statistical approach to retrace the patterns in the dependent variable (phenomena) linearly using the proposed independent variables*

> *(phenomenon) such that we can identify which of the independent variables has played its part in the dependent variable and with how much intensity.*

What regression analysis does is that, it tries to fit a line that passes through the scatter plot of dependent and independent variable by ensuring that the majority encompasses the majority of the incidences (dots). This can be done by ensuring the net distance of the dots above the line is equal to the net distance of dots below the line[2].

In simple terms, regression analysis formulates a cause and effect relationship. It then quantifies how much effect will occur if, cause variable is changed by 1%. So regression can provide the possible determinants (independent variables) which can change dependent variable. Several examples of cause and effect are provided in Al Qur'an (69:9, 25-34)

While plotting the line, it gives estimates of two statistics. First one is called intercept, which shows the starting point of the line, and the second one is the slope, which tells the rate of rise or fall of the line.

[2] For simplification we are showing only one independent variable, as we cannot show more than two axes.

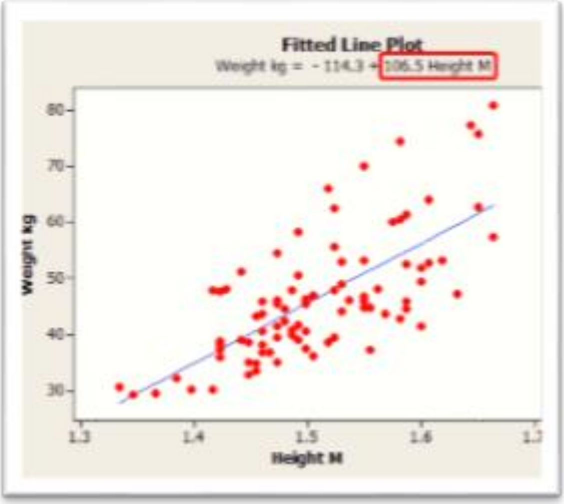


Figure 4. *Fitted line on scatter plot*

Role of regression analysis

The role of regression analysis and Econometrics can be seen from the following verse of Holy Qur'an

> *Indeed Allah will not change the condition of a people until they change what is in themselves.* Al Qur'an (13:11).

Here situation means what is happening around in the present. And probably most of it has been carried on from the past and practised by our forefathers. This verse indicates us to change our situation, so what is required to change the situation? Following are the steps involved.

1. We have to devise a theory about the situation using conceptualization reality and study of literature. Nature has been built with a pattern so that humans can formulate the theory. It can be seen from the verses of Holy Qur'an.

> *[He is] the cleaver of day break and has made the night for rest and the sun and moon for calculation. That is the determination of the Exalted in Might, the Knowing*. Al Qur'an (6:96).

The sun and the moon [move] by precise calculation. Al Qur'an (55:5)

2. After making an economic theory, we can use descriptive statistics to measure the situation with the help of mathematical economics.

3. As soon as we can measure, we can devise a benchmark to see how good or how bad the situation.

4. With the help of the benchmark, we find the need to improve the situation, as we can estimate the consequences using regression analysis if they are not improving. It can be seen from the Holy Qur'an that, if people take the situation non-seriously, they will face the consequences.Also in a Sahih al Bukhari, book 78 hadith 160, Prophet Muhammad (S.W.T) stated that a believer does not make the same mistakes twice.

> *And the evil consequences of what they did will appear to the, and they will be enveloped by what they used to ridicule*. Al Qur'an (45:55).

Narrated by Abu Huraira:

The Prophet (S.W.T) said, "*A believer is not stung twice (by something) out of one and the same hole.*"

5. Since we know the consequences, we try to find the determinants which can control or change the situation using regression analysis.

6. This process will bring clarity regarding the situation, and we will be better able to devise a policy to improve it. As Holy Qur'an [10:36] states that decisions made using guesses will not be as good as precise calculation.

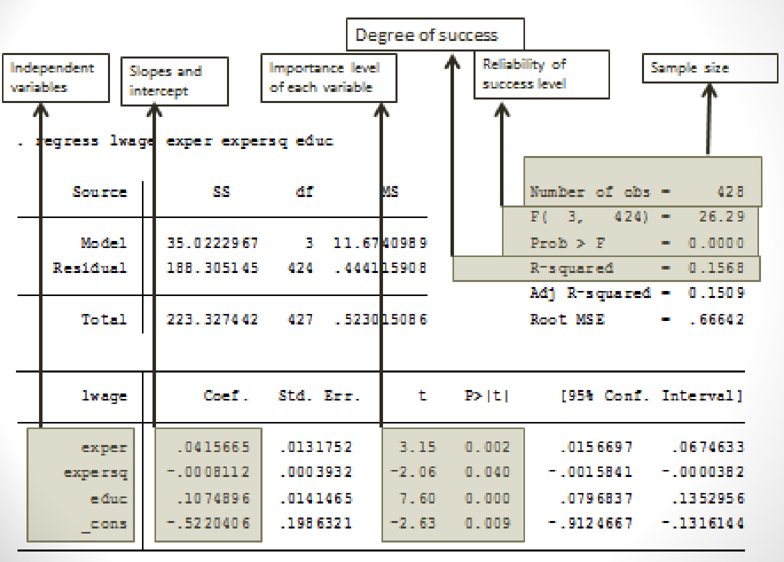
> *And most of themfollow not except assumption. Indeed, assumption avails not against the truth at all. Indeed, Allah is Knowing of what they do*. Al Qur'an (10:36).

Ordinary least square methods

In a regression analysis approach, ordinary least squares are the most popular method which can be used. Another popular method is a method of moments. This OLS is defined as

> *Ordinary Least Squares is the method, to find the optimal values of slope and intercept coefficient, by minimizing the sum of the square form of error terms (difference of each point with from the estimated line).Which will help in identifying the role of independent variables on the dependent variable.*

Following are the different components of the OLS estimate output, which highlights the sections which are important and meaningful.



```
. regress lwage exper expersq educ

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  3,   424) =   26.29
       Model |  35.0222967     3  11.6740989           Prob > F      =  0.0000
    Residual |  188.305145   424  .444115908           R-squared     =  0.1568
-------------+------------------------------           Adj R-squared =  0.1509
       Total |  223.327442   427  .523015086           Root MSE      =  .66642

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0415665   .0131752     3.15   0.002     .0156697    .0674633
     expersq |  -.0008112   .0003932    -2.06   0.040    -.0015841   -.0000382
        educ |   .1074896   .0141465     7.60   0.000     .0796837    .1352956
       _cons |  -.5220406   .1986321    -2.63   0.009    -.9124667   -.1316144
------------------------------------------------------------------------------
```

Figure 5. *Sections of Regression Output*

Estimation and role of uncertainty

> *And they ask you [O Muhammad], about the soul. Say, "The soul is of the affair of my Lord. And mankind have not been given of knowledge except a little".*

Al Qur'an (17:85)

> *And within the land are neighboring plots and gardens of grapevines and crops and palm trees, [growing] several from a root or otherwise, watered with one water; but We make some of them exceed others in [quality of] fruit. Indeed in that are signs for a people whom reason*. Al Qur'an (13:4).

Since we are humans, we can make errors in calculation or in some cases, we can be wrong in the approach of calculation. For this, we introduce a component of uncertainty in the regression, we call it error term, or residuals (ε).

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

There are three terms in this regression which can be differentiated.

1. Actual Dependent variable (Y), it is the actual measured situation which is fed for analysis
2. Estimated Dependent variable ($\alpha + \beta_1 X_1 + \beta_2 X_2 = \hat{Y}$), this is the portion of the situation which can retrace using our independent variables ($X_1 \& X_2$). This is actually the portion of the Y which we can control using the means and methods which we know.
3. Error term (ε = Y -$\hat{Y}$), this is the portion of the situation which we cannot comprehend; we do not know how to control it as per the understudy theory. It is the influence of nature.

So what regression analysis does is it splits the actual situation into two components first one is the portion of situation which we know about and second is the portion which is the random, unpredictable changes in the situation.

Properties of ordinary least squares

There are some inherent properties of Ordinary Least Squares which lead it to be used as a regression analysis approach. As a whole, it is called Best Linear Unbiased Estimator (BLUE).

Unbiased

…witness in justice, and do not let the hatred of a people prevent you from being just… Al Qur'an (5:8).

OLS uses the arithmetic mean (AM) approach to calculate of this OLS estimate from the sample are expected to be same as the estimates from reality. In statistics, it is called as the expected value of the sample mean will be equal to the population mean.

$$E(\bar{x}) = \mu_x$$

Efficient

… Be just; that is nearer to righteousness. … Al Qur'an (5:8)

OLS uses the standard deviation approach to calculate the standard errors of the estimates, as standard deviation provides the smallest possible dispersion measure because of this OLS estimates are efficient.

Consistent

O you who have believed, be persistently standing firm for Allah, … Al Qur'an (5:8).

Since increasing the sample size reduces the standard deviation, hence bigger the sample more precise and consistent, the OLS estimates will be.

$$\delta_x = \sqrt{{(x - \bar{x})^2}/{n - 1}}$$

Where if n →∞ then δ_x → 0

Sufficient

Sufficiency is the property of the estimate which uses all the available information since OLS uses arithmetic mean and standard deviation; it makes OLS a sufficient estimator.

Linear

OLS provides linear estimates (marginal impact) of independent variables on the dependent variables.

Example of regression analysis

Consider the example of determinants of student performance and use the number of study hours as independent variables. For this, we get information for one student and inquire for his GPA and the study hours. He replies with a 3 GPA with 3 hours a day study. According to this information we can write a relation and 3 hours study leads to 3GPA

Performance : Study hours
3 GPA : 3 Hours
3 GPA = β x 3 Hours
β = 1 GPA / Hour

The β is the coefficient of proportion, which converts the proportion to equality in mathematics.It is called slope which tells how much performance will be yielded by one hour of study in statistics.And it is called the marginal impact of one increasing one hour of study on performance in mathematical economics. So in terms of interpretation, β will tell if we have a student of 3.5 GPA, how many hours he might have studied (*3.5GPA x β = 3.5 hours*). It also tells if we have a student who studies 2.5 hours per day how much GPA we expect of him (*2.5 Hours x β = 2.5 GPA*). But there is one issue that we are predicting every one using one observation only, this might lead to wrong predictions. For this we might increase the sample size. For simplicity, we will calculate it using three observations.

GPA = β x Study hours
Student 1: 3.5 = β_1 x 4
Student 2: 3 = β_2 x 3.5
Student 3: 2.8 = β_3 x 3

Using this we can calculate the values of the slopes. And write it in matrix form. [3]

$$\begin{bmatrix} 3.5 \\ 3 \\ 2.8 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} x \begin{bmatrix} 4 & 3.5 & 3 \end{bmatrix}$$

$$\begin{bmatrix} 3.5 \\ 3 \\ 2.8 \end{bmatrix} = \begin{bmatrix} 0.875 \\ 0.857 \\ 0.933 \end{bmatrix} x \begin{bmatrix} 4 & 3.5 & 3 \end{bmatrix}$$

So we have three slopes/coefficient of proportions, for the sake of robustness, we will take the arithmetic mean of the slopes. Here the average of three slopes is 0.888.

$$\begin{bmatrix} 3.5 \\ 3 \\ 2.8 \end{bmatrix} = 0.888\, x \begin{bmatrix} 4 \\ 3.5 \\ 3 \end{bmatrix}$$

So we have generalized the marginal impact of this theory which says that higher study hours lead to a higher GPA. In reality, we do it using a big enough sample so that it seems reliable. Now we write the theory.

GPA = 0.888 Study Hours

Since we have developed or quantified the theory using our sample, now we will see its forecasting performance, to see how successfully the actual GPA, is predicted.

[3] Matrices cannot be multiplied directly; one of them has to be transposed.

Predicted GPA = 0.888 Study Hours

$$\begin{bmatrix} 3.552 \\ 3.108 \\ 2.664 \end{bmatrix} = 0.888\ x \begin{bmatrix} 4 \\ 3.5 \\ 3 \end{bmatrix}$$

Since the actual GPA is different from the predicted GPA, we will calculate the error of prediction.

Error of prediction = Actual GPA – Predicted GPA

$$\begin{bmatrix} -0.052 \\ -0.108 \\ 0.136 \end{bmatrix} = \begin{bmatrix} 3.5 \\ 3 \\ 2.8 \end{bmatrix} - \begin{bmatrix} 3.552 \\ 3.108 \\ 2.664 \end{bmatrix}$$

Now, as a researcher, we will try to investigate why there is this error. Indeed, we will see what it the arithmetic mean of it, if it is zero we will say that all goods and bads are cancelled out, our model predicts reality well. Here the average of the error of prediction is -0.024, it looks to be small, but still, it is some information, which can be studied. In Econometrics, we call it an aggregate effect of all those factors which affect the GPA other than the study hours.It is so because the study hours failed to explain this much about the actual GPA, since GPA does change hence there must be some other aspect which can explain it. We call it the intercept. Ee add this information in the model aftersubtracting it from the error of prediction, the remaining error which is left behind is called error term or residuals.[4]

GPA = -0.024 + 0.888 Study Hours + residuals

$$\begin{bmatrix} 3.5 \\ 3 \\ 2.8 \end{bmatrix} = -0.024 + 0.888\ x \begin{bmatrix} 4 \\ 3.5 \\ 3 \end{bmatrix} + \begin{bmatrix} -0.028 \\ -0.084 \\ 0.112 \end{bmatrix}$$

[4] So error term or residuals are actually demeaned prediction errors. And intercept is actually the arithmetic mean of prediction error.

Here you can see the mean value of the residuals is zero, which means that this is the portion of the actual GPA which is totally random we cannot explain zero.[5]

Similarly, we can calculate the standard errors at the place where we calculated the intercept and average slope.

$$SE = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \Big/ \sqrt{n}$$

$$SE_{slope} = \sqrt{\frac{(0.875-0.888)^2+(0.857-0.888)^2+(0.933-0.888)^2}{2}} \Big/ \sqrt{3}$$

$$SE_{slope} = \sqrt{\frac{(-0.013)^2+(-0.031)^2+(0.045)^2}{2}} \Big/ 1.732$$

$$SE_{slope} = 0.023$$

$$SE_{intercept} = \sqrt{\frac{(-0.052+0.024)^2+(-0.108+0.024)^2+(0.136+0.024)^2}{2}} \Big/ \sqrt{3}$$

$$SE_{intercept} = \sqrt{\frac{(-0.028)^2+(-0.084)^2+(0.112)^2}{2}} \Big/ 1.732$$

$$SE_{intercept} = 0.058$$

GPA = -0.024 + 0.888 Study Hours + residuals
SE 0.058 0.023

We have intercept and slope, which tell the value of the respective characteristic. Higher the value show higher the impact while the standard errors are from the dispersion of actual values from their mean, higher the value show less reliable the impact.

See in the above relation, the value of the intercept is smaller as compared to the slope also the value of standard error is smaller too. Here it is not humanly possible to decide out of both slope and intercept which one is more reliable as

[5] This is assumption 3 of OLS.

the smaller value of intercepts contradicts in the decision as compared to the outcome of the standard error.

Here statisticians have developed a new indicator which uses both the arithmetic mean (intercept or slope) and their respective standard errors and calculates it against a benchmark. It is called t value, usually, in the regression output, the benchmark is zero.

$$t = {}^{slope-(benchmark=0)}\Big/_{standard\,error}$$

So after calculation we have.

GPA = -0.024 + 0.888 Study Hours + residuals
SE 0.058 0.023
t -0.41 38.61

Now we the process which is used to check the slope or intercept is reliable is called hypothesis testing. In the coming section, we will use this approach to check reliability.

Hypothesis testing

> *Hypothesis testing is a scientific procedure, which is based on sample evidence and uses a probability distribution to determine whether the proposed claim (hypothesis) is feasible (true on the base of data) or not with some level of assurance (fixed level of confidence or fixed level of error).*
>
> *O you who have believed, if there come to you a disobedient one with information, investigate, lest you harm a people out of ignorance and become, over what you have done, regretful*. Al Qur'an (49:6).

Hypothesis test comprises of following steps which are usually done, as most of the statisticians and econometricians are well versed regarding this procedure, so they leapfrog the process and give the results. The list of steps is as follows

1. Null and Alternative Hypothesis
2. Level of imprecision allowed
3. Selection of Test statistics
4. Benchmark value / Decision Criterion
5. Calculation and hypothesis selection

Example

Suppose that university management wanted to evaluate the impact of a number of time spend while studying on the Grade Point Average of the student as an awareness program to promote students to spend more time studying. Since because of time and cost issues,the university cannot interview all the students of the university (which will be considered as a population of study). So they opt for a sample of 3 students. From these student's number of hours studied per week and GPA was asked and regression analysis was used to calculate the marginal impact. The result of the regression analysis as mentioned in section 1.6 is below.

GPA = -0.024 + 0.888 Study Hours + residuals
SE 0.058 0.023
t -0.41 38.61

In the summary of the report provided by the university, management stated that

"The time spent by the student in studies is never fruitless."

Here is this case we have some data which is

n = 3, b = 0.888 and se = 0.023 [6]

[6] b is sample notation for the slope coefficient and β is the population notation for slope coefficient.

Here the claim by the university management reveals that they are advocating the fact that the more time you spend, the more GPA you would expect. It means the slope coefficient will definitely more than zero. [7]

So we write the possible hypothesis for the claim as

H_0; b = 0 (Baseline, which is usually opposite to claim)
H_1; b > 0 (Claim made on the basis of data or the requirement)

We can see here that we have made a set of hypothesis based on the claim which was made by the management. Here H_0 is called null hypothesis, which usually consists of equality (representing status quo) in it or the opposite of the claim. And H_1 of H_a is the alternative statement/hypothesis, which isconstructed based on a claim or requirement provided, and remember equality sign can never come in this alternative hypothesis. This construction of hypotheses is the first step in solving hypothesis testing. Once we have two statements which are opposite to each other, we know that only one of them is true.

Now in the second step, we need to specify how much error should we allow? As the error is inevitable.We cannot make it 0 so what we can do is adjust our calculation according to the level of error, we can allow; so that wrong results do not mislead us. Usually, in the questions related to hypothesis testing the Level of significance (chances of type one error) (α) is provided. We have to use it. Otherwise, we can use anyone from 1%, 5% or 10%. Now using this α, we are creating a boundary which creates a distinction between

[7] Since according to the claim slope cannot be less than 0 it means we are only talking about the positive side of zero, so it is one tailed test. If the claim was that it is not equal to zero and it can have any value less or high than 0 then it would have been two tailed.

the sample values which could have the population mean mentioned in the null hypothesis, from the sample values which are deemed too far from this means such that while inspecting them. We could reject the mean equal to null hypothesis even though it is true, so it is the region of type I error.

In the third step, we need to tell which test should be used, which depends upon the data. T test is just a modified version of Z test it only adjusts the data which is small in sample and we are not sure that it will be normal or not. Z test is for the Normal distribution and T test for the Student T distribution. So we will use a relevant table. Usually, in regression analysis, we use the T test as it is very rare to have known population variance.

Table 1. *Test selection criterion*

	Sample n≥ 30	Sample n <30
Population variance known	$z = \frac{\bar{x} - \mu}{\delta/\sqrt{n}}$	$z = \frac{\bar{x} - \mu}{\delta/\sqrt{n}}$
Population variance unknown	$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

In the fourth step, we need to calculate/formulate a decision criterion which can define regions in the normal distribution where H_0 is accepted and where H_1 is accepted. It is also called defining critical values. These critical values are made on the basis of level of significance (allowing error) and hence it can help to findtruth. In the case of 5% error allowed and the sample size is 3, the critical value of one tailed test is 3.182.

So if

$$t_{calculated} < 3.182;\ H_0\ is\ accepted$$

In this above case the regression estimate of slope is not reliable; it is value is just by chance.

$$t_{calculated} > 3.182;\ H_1 is\ accepted$$

In this above case, the regression estimate of slope is statistically realistic; this sample value can be implied on population.

The last step is calculated using the test, which is decided in the third step.

$$t = {}^{slope-(benchmark=0)}\big/_{standard\ error}$$

$$t = {}^{(0.888-0)}\big/_{0.023}$$

$$t = 38.61$$

Here we can see that the calculated t value of 38.61 is higher than the critical value of 3.182. Which means that the estimates generated from the sample are reliable; they are not just by chance. We can use these results to make a policy or claim regarding the population.

This is the procedure in the background when we need to decide on any sample statistic. Nonetheless, the experienced statisticians bypass this procedure using the p value criterion. The advantage of the p value is that there is no need to open the critical value tables every time to decide the reliability. We will just use the value of α (allowed error) as threshold p value.

$$p_{calculated} > \alpha;\ H_0\ is\ accepted$$
$$p_{calculated} < \alpha;\ H_1\ is\ accepted$$

Types of variables

Variables are proxy data for an economic phenomenon which we discuss in economic theories. Some variables are very close to the phenomenon others are approximate of them. Econometricians try to develop better variables, but during their development process, they use what is available.

Table 2. *Variables and their terminology*

Type	Forms	Purpose
Use based		
Dependent variable	Can be any type discussed below	A variable which is expected to change because of other variables.
Independent variable	Can be any type discussed below	A variable which is expected to change the dependent variable. And this effect is point of focus of the study. They are usually those which are controllable by the policy makers
Control variable	Can be any type discussed below	A variable which is expected to change the dependent variable. They are controlling the environment of the model, as they is some cases not controllable by the policy makers
Instrument variables	Can be any type discussed below	They are discussed in chapter 4. They are used in background of regression.
Time based		
Auto Regressive (AR)	Lags of Dependent variable	It checks persistence (the effect of past of dependent variable on present).
Distributed Lag (DL)	Independent variables & their lags	It checks effects of other variables and their past
Moving average (MA)	Error term & its lags	It checks sensitivity to present and past shocks
Trend	Time variable	To incorporate continuous quality change, like experience, or technology gain
Growth based		
Difference	First difference of any variable (growth rate)	To check the effect of growth
Deviation	Difference of a variable with certain constant (threshold) or mean (mean deviation)	To check effect of change in situation as compare to some reference
Qualitative / limited quantitative		

Dummy (binomial)	Qualitative variable	To incorporate two qualities, one regime change, or one break in pattern.
Dummy (multinomial)	Qualitative variable	To incorporate more than two qualities, one regime change, or one break in pattern.
Discrete	Limited Quantitative variable	To use variables like population, number of events etc.
Volatility based		
Volatility (GARCH)	Variance of a variable (Error Term)	To incorporate effects of changes in dispersion Discussed in chapter 10
Non-linear pattern or Interaction		
Non Linear IV	Square or cube form of variable	To incorporate the effects of non-linear variable (increasing or decreasing returns)
Interactive Dummy	Dummy multiplied with independent variable or with another dummy	It is used to see the complimentary or substitutability effect of variable and qualitative variable
Cross product	Two different independent variables multiplied with each other	It is used to see the complimentary or substitutability effect of two variables
Empirical pattern		
Stationary variables	Can be any type discussed above	They are discussed in chapter 7, this is the variable which is moving naturally no influence of other variables
Non Stationary variables	Can be any type discussed above	They are discussed in chapter 7, this is the variable which is influenced by other variables or policy.

These are common terminologies used with the variables. Table 2 has tried to enlist all possible cases, so that it dissipates any confusion that exists in the names of the variables.

Types of regression data sets

Below it the constructed example of study performance. Now, this model can be estimated in reality by three ways.

$$\text{Study Performance} = \alpha + \beta_1 \text{ (no of hours studied)} + \beta_2 \text{ (age of the student)} + \mu$$

Table 3. *Converting model into sub-types*

Cross sectional model	Time series models	Panel data models
Studying how hours and age of different students effect their study performance (GPA)	Studying how hours and age of same student affects his performance throughout time.	Studying how hours and age of students effect performance in terms of their differences and aggregated throughout time

Cross sectional approach

This is a data set where several subjects are studied (i.e. individuals, firms or countries) at some known point in time. Here the number of subjects becomes a number of observations.

This data set approach is used, when the purpose of research is to estimate the generic differences between the students. In such cases the policymakers are interested in evaluating the effects of students coming from different types of the background study, in order to ensure a system where the study performance in the university is not influenced by this differences. This data set is common for survey based studies and with or without effect evaluations.

Time series approach

This is the data set where a single subject is studied across several known points in time. Here the number of time periods becomes number of observations.

This data set approach is used, when the purpose of the research is to estimate the dependent variable is evolved over time and how the present and historical patterns of other variables influence it. Usually the subject is a country or a firm. In such cases, policymakers are interested in influencing the rate of change of dependent variable. This data set is common for secondary data sets and before and after evaluations.

Panel data approach

This is the data set where several interconnected subjects are studies across several known points in time. Here the product of number of subjects and number of points in time becomes the number of observations.

This data set used the properties of both cross sectional and time series approach. This approach can be used to confirm the global application of any particular theory. If the number of subjects are in majority, then this approach exhibits characteristics of cross sectional approach (i.e. behavior model) and if the time periods are in majority, then this approach exhibits characteristics of time series approach (i.e. equilibrium model). This data set can do before & after and with & without evaluation at same time.

Summary

This chapter focuses on the evolution of ordinary least squares approach, its application example with calculations and interpretations. It was made clear that what information does OLS provide and what information it failed to provide.

The presence of information, which OLS do not cover leads to issues in the estimations. It is represented in terms of failure in any one of its property. The chapter ended with the three types of data sets and their implications.

Application questions

i. Define regression analysis and how does OLS does this function?

ii. Find the relevancy of the following concepts with regression analysis?

i. Sequence and Series

ii. Ratio and Proportion

iii. Constraint Optimization

iv. Line Plot

v. Differentiation

iii. What if we run regression without intercept, what are the reservations?

iv. What is the difference between error term and intercept?

v. What is the purpose of Z or T Test?

vi. What is the purpose of different data sets like time series, cross sectional and panel data?

vii. See Figure 4, the intercept shown is negative but the dependent variable show is the weight of the car, what does it mean?

viii. T tests usually test against 0 but if we want to test against 1, what necessary calculations needed (hint: see the formula of T values)?

ix. Consider the following model where advertisement impressions are a function of advertisement expenditure and its data shown in table 3. Calculate intercept and slope coefficient in STATA.

a. What is the expected/theoretical relationship between these two variables?

b. Use the hypothesis testing procedure to determine if advertisement expenditures are significant

c. There is a suspicion that the effect of advertisement expenditures do not have a linear effect use relevant means to test this suspicion.

Table 4. *Advertisement Impressions and Expenditure*

Firm	Impressions to Million People	Expenditure in Millions $
Pepsi	32.1	50.1
Fed'l Express	21.9	22.9
Burger King	60.8	82.4
Coca Cola	78.6	40.1
McDonald's	92.4	185.9
Ford	40.1	166.2
Calvin Klein	12.0	5.0

Source: Gujarati (2009). [Retrieved from].

x. The law of one price theory states that the price ratio between the two countries effect their exchange rate. Based on the data of Pakistan and USA following are the estimation results

Y = 5.0 – 1.13 X
se (2.4) (0.50)
R squared = 0.86

Here Y is Exchange rate of Pakistan and X is domestic prices / USA prices

a. Interpret R squared and what does it represent?
b. Interpret coefficient of variable X

xi. For the estimated equation below, interpret all the mentioned variables.

$$\begin{aligned} Consumption = {} & 500 + 0.7 Income - 0.02\, Income^2 \\ & + 0.2\, Dummy_1 + 10 Trend + 0.2 Trend^2 \\ & + 0.1(Income * Dummy_1) \\ & + 0.05(Trend * Dummy_1) \\ & + 0.5 Consumption_{-1} \\ & + 0.05 Dummy_1 * Dummy_2 \\ & + 0.002 \Delta Income + \ error\ term \\ & - 0.05\ error_{t-1} \end{aligned}$$

Where
Income is in 100US$

Dummy$_1$ [Here Dummy$_1$ = 1 for Male, Dummy$_1$ = 0 for Female]
Dummy$_2$ [Here Dummy$_2$ = 1 for Urban, Dummy$_2$ = 0 for Rural]

Further study

Armstrong, J.S. (2011). Illusions in regression analysis. *International Journal of Forecasting*, 3, 689-694.

Frost, J. (2017). *Five Regression Analysis Tips to Avoid Common Problems*. Statistics by Jim. [Retrieved from].

Frost, J. (2017). *How to interpret regression models that have significant variables but a low R-squared*. Statistics by Jim. [Retrieved from].

Frost, J. (2017). *Five reasons why your R-squared can be too high*. Statistics by Jim. [Retrieved from].

Soyer, E., & Hogarth, R.M. (2012). The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting*, 28(3), 695-711. doi. 10.1016/j.ijforecast.2012.02.002

2 Need of advanced models

Learning Outcomes

- Why simple regression is no longer feasible?
- What are possible issues in simple regression and how to detect them?
- What are applied meanings of each regression issues and what information we can extract from them?

Introduction

Advanced models are designed on the basis of indicators of abnormality to incorporate the abnormality. *"Then do they no look at the camels - how they are created? And at the sky – how it is raised? And at the mountains – how they are erected? And at the earth – how it is spread out?"* Al Qur'an (88:17-20)

With the development of Ordinary Least Square (OLS) approach in the early 1800's, people started to modify their behavior based on the information extracted from the estimation results. It was later work done by Gauss (1821) where the work started regarding the limitations of OLS. Later, several economists, mathematicians and statisticians contributed in developing the assumptions of Ordinary Least Squares.

Assumptions of ordinary least squares

...In that Empire, the craft of Cartography attained such Perfection that the Map of a Single province covered the space of an entire City, and the Map of the Empire itself an entire Province. In the course of Time, these Extensive maps were found somehow wanting, and so the College of Cartographers evolved a Map of the Empire that was of the same Scale as the Empire and that coincided with it point for point. Less attentive to the Study of Cartography, succeeding Generations came to judge a map of such Magnitude cumbersome, and, not without Irreverence, they abandoned it to the Rigors of sun and Rain. In the western Deserts, tattered Fragments of the Map are still to be found, sheltering an occasional Beast or beggar; in the whole Nation, no other relic is left of the Discipline of Geography.

–Of Exactitude in Science, Jorge Luis Borges (1972)

Following are some assumptions which needed to be fulfilled so that the above splitting of the dependent variable is reliable and correct.

1. The model must construct a straight line / linear to retrace the dependent variable

$$Y_i = \alpha_1 + \beta_1 X_i + u_i$$

This assumption clarified that the independent variables could be non-linear, but the parameters α and β must be linear. This restriction is required as OLS cannot estimate models with non-linear parameters.

Figure 1 shows the regression line with linear coefficients and variables, while figure 3 shows the example of non linear function.

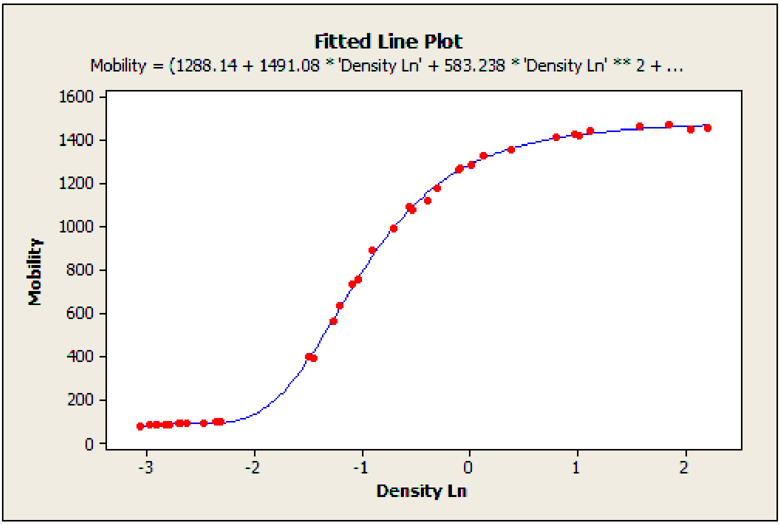
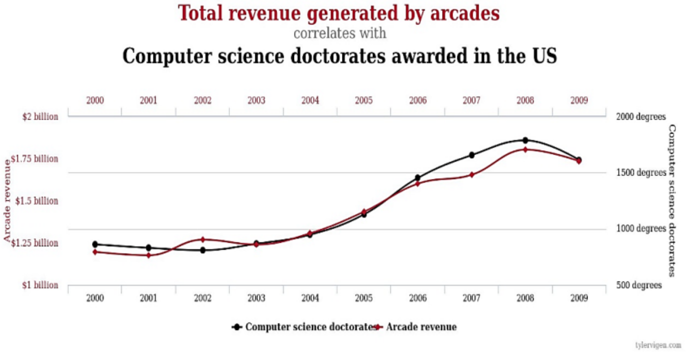


Figure 6. *Non-linear in coefficients regression*

2. The independent variables must be controllable by the policymakers

Regression must include those variables which are either controllable by the policymakers, or they are signifying the environment. This is required as relation does not mean anything if it cannot be used for certain gain.

The sensible policy requires controllable instruments



Tyler Vigen[1] provides interesting bivariate relationships. Following graphs shows a relationship with 98% correlation. While

[1] [Retrieved from].

estimating regression on determinants of a number of doctorates in computer sciences following relationship will signify that the government should promote students to visit arcades. Such type of relationships is formed by coincidence and called spurious regressions.

3. The mean of error term which is left over must be zero to make it totally uncomprehend-able

$$E(u_i|X_i) = 0$$

Zero mean value means that any information that is excluded from the model is assumed to have no net information. This means that researcher has tried his best to comprehend the dependent variable. If mean value of residuals are non-zero this will signify that author has ignored certain set of information making the results biased.

4. The variance of the error term must be constant so that it becomes comparable

$$var(u_i|X_i) = E[u_i - E(u_i|X_i)]^2 = \delta^2$$

OLS requires that the variance of the residuals (consequently variance of coefficients) must be constant in order to ensure comparability between any two set of observations. Figure 4 shows the cases of non – constant and constant variance, where assumption 3 still holds.

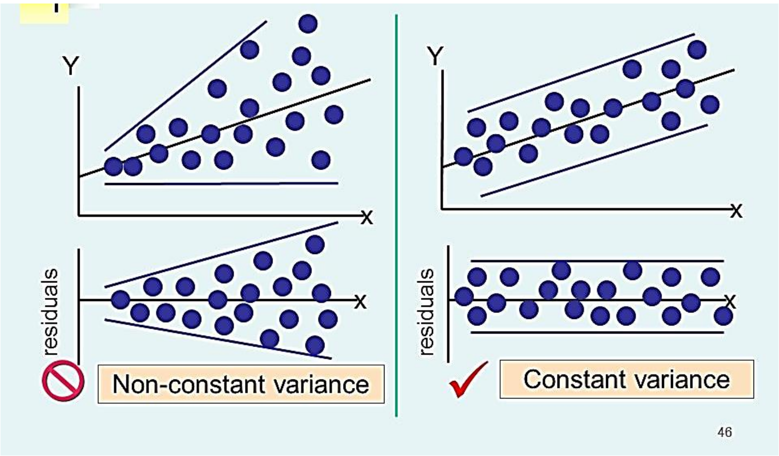


Figure 7. *Nature of variance of the model and residuals*

5. The error term must be random and uncorrelated with itself

$$cov(u_i u_j | X_i X_j) = E\{[u_i - E(u_i)] | X_i\}\{[u_j - E(u_j)] | X_j\} = 0$$

Regression errors must be random in nature. If they are a function of the time it denotes that there must be some excluded factors which are causing residuals to behave in a certain pattern. A good OLS model will not ignore any important factor (independent variable).

6. The varianceof the independent variable should not influence the variance of the error term (zero co-variance)

$$cov(u_i X_i) = E[u_i - E(u_i)][X_i - E(X_i)] = 0$$

There should be a clear difference between information (X) and random error (u). A good OLS model will ensure that there is no relation between what is included in the model and what is not included in the model.

7. The sample size should be more than the slopes and intercept. It is a mathematical requirement where data should be more than unknowns.

8. There must be some change in independent variables; if there is no change, they cannot trace the dependent variable.If the dependent variable is changing without a change in an independent variable, then it is a clear indicator that the independent variable is not useful.

9. Theory must be correctly specified in terms of functional form. Such that there must not be any error because of wrong specification.

$$Y_i = \alpha_1 + \alpha_2 X_i + u_i$$
$$Y_i = \beta_1 + \beta_2 \frac{1}{X_i} + u_i$$

Where Y_i = the rate of money wages and X_i = the unemployment rate. Both of these functions can represent the Philips curve, but the second one is a better representative. OLS requires that best possible specification must be used as only God knows the perfect specification.

10. The independent variables must not strongly relate to each other so that we can differentiate easily which independent variable has a significant impact.

Correlation and Imprecision

Class teacher while estimating the determinants of overall class performance, and assessing the role of individual students in the class. The teacher asked a question, but two students answered simultaneously (two independent variables changed together). Because of this high correlation,the teacher cannot assess the class performance perfectly, as there can be many possible answers.

Here the question arises why there is a need to put assumptions on the real-life policy estimation tool? Doesn't it make the outcome more questionable? As the outcome is based on the assumption that the research is hiding behind.

Scope of ordinary least squares

Here we will try to answer the objection that people hold against econometricians that they '*manipulate the model until they achieve the desired outcomes*'.

Let us build on the model constructed in the previous chapter using OLS. Below models indicate that an increase in study hours improves the student performance. [2]

$$student\,performance_{it} = a_{1i} + a_2 hours\ studied_{it} + e_t$$

This model is constructed by the collection of data from the students. If the data is too small, then we will not have enough variation in hours studied to compare with student performance, this will lead to *micronumerosity*. And if, the sample is not properly constructed, there will be too many intelligent students or too many dull students,then the sample will become imbalanced where there is a majority of middle intelligence students.This is indicated via *non-normal residuals*.

With this model, people optimized their study hour to achieve student performance. However, a time came they experienced diminishing returns, such that increasing hours studied either decreased student performance or no effect overall.

Now, this looks odd that the most important variable becomes insignificant. Researchers proposed that there must be other factors which can be used to offset the diminishing returns. Statistically, this can be indicated via cross sectional correlation a.k.a. *cross sectional autocorrelation*. This is checked via Durbin Watson Test is cross sectional models or Cross Sectional Dependence Test.

[2] For the start we are not discussing the construction of Panel Data model, it will be discussed later in chapter 11.

Researchers solved this problem by exploring more variables and expanding the model shown below.

$$\begin{aligned} student\ performance_{it} &= a_{1i} + a_2 hours\ studied_{it} + a_3 study\ group_{it} \\ &+ a_4 study\ pattern\ of\ friend_{it} + e_t \end{aligned}$$

Now people examined that the students who are young tend to be less consistent (their GPA are more dispersed) as compared to the students who are elder. This shows that students are following error learning behavior such that with an increase in the time they are better expert in achieving their objective (which is higher GPA). This will make an overall dispersion of the model to fall (as depicted by a fall in the standard deviation of e_t), causing *time series heteroskedasticity* in the model which can be sorted by addition of age of a student as a proxy of error learning behavior.

$$\begin{aligned} student\ performance_{it} &= a_{1i} + a_2 hours\ studied_{it} + a_3 study\ group_{it} \\ &+ a_4 study\ pattern\ of\ friend_{it} + a_5 age_{it} + e_t \end{aligned}$$

Now, model looks sound but, with the time the student will try to copy the study pattern of his intelligent friend.So that he can increase the GPA, what will happen then, both hours studied and study pattern of a friend will become correlated, causing *multicollinearity*. This problem is complex; its solutions are addressed in chapter 6. After this, motivation can come into play, such that if a student gets a good GPA in first semester, he/she spends more time in study.It will cause dependent on effecting independent variable. This dynamic is named as *endogeneity*.

Here I have tried to explain that OLS model is redundant, now as its contribution is fully utilized. The modification in people behavior has led toincompatibility of OLS. For this new advanced versions are constructed which incorporate this modification.

Following are the definition and explanations of all possible reasons which can make OLS model incompatible. [3]

Micronumerosity

Using of non-representative data or too small sample size to create a policy for all the population will represent the issue which is statistically called micronumerosity.

It is a situation in real life. Doctors face this situation, where they have a single patient (sample size = 1), from which he checksthe value of several variables (like BP, Heat rate, weight etc.) to diagnose the probability of the disease (estimated dependent variable). In statistical data, the symptoms of micronumerosity arethat all the slopes and overall model will be insignificant. In such cases, where data cannot be increased, the experience of the researcher precedes the statistical measurement.Just like the Doctor has prescribed the patient only on the base of one sample observation.

Multicollinearity

Consider the multiple regression model with more than one independent variable. This model states that both study hours per week and the age of the student affects the GPA.

$$GPA = \alpha + \beta_1 Study\ Hours + \beta_2 Age\ of\ Student + e$$

There can be cases where the older student might be engaged in the family activities and jobs that is why allocating less time in study. Such that

$$Study\ Hours = f(age\ of\ the\ student)$$

[3] Possible issues of regression estimates [Retrieved from].

This will make the independent variables related to each other in reality, such that the coefficient β_1 will be biased. This situation is called multicollinearity in statistics. The symptoms of this situation are that the slopes might be insignificant, or they have the opposite sign as expected.

It can be of two types too, time series multicollinearity and cross sectional multicollinearity but since there is no way to calculate them separately, so we do not distinguish them too.

Non-normality

Normality suggests certain conditions of the dataset where there is a given amount of extreme values in the data, and a certain amount of homogeneity (called kurtosis) and data should not have any grouping other than the center (mean) of the data (called skewness).

So if the cross sectional data are based on too many heterogeneous students such that extreme values are beyond the limit, then there will be normality issue. In time series, it can be explained in terms that there are too abrupt changes like a student getting a full-time job before he is an adult that will make data non normal. This situation can be diagnosed using Jarque & Bera (1980) Test.

Mis-specification

There can be a nonlinear effect of age, like higher the age more experienced he becomes so more chances that he gets higher marks.

$$GPA = \alpha + \beta_1 Study\ Hours + \beta_2 Age\ of\ Student + \beta_3 (Age\ of\ Student)^2 + e$$

Ignoring the square form will make model miss specified. The imperfect specification will result abnormal behavior or the residuals.

Heteroscedasticity

It means that the variance of the model is not constant; it becomes a function of some factor most probably independent variables. It is also violence in OLS assumption.

Cross sectionalheteroscedasticity

> *O mankind, indeed We have created you from male and female and made you peoples and tribes that you may know one another. Indeed the most noble of you in the sight of Allah is the most righteous of you. Indeed Allah is Knowing and Acquainted.* Al Qur'an (49:15).

It exists in cross section or panel data models only. It occurs because of differences in the cross sections i.e. students in this model. It shows that when we have too much heterogeneous sample, and we have not incorporated their differences, then model ends up having this issue.

Time seriesheteroscedasticity

This exists in the time series or panel data models only. It comes only of the individual is behaving differently in time. We call it error learning model like a tailor will make more errors at the start of his career, but after a few years, he will make very few errors. On country wise data, this problem can be an indication of a change in technology level in the country.

Autocorrelation

It means that the residuals are functions of its past. They are not random as depicted in OLS assumptions. It has two types (Frees, 1995).

Cross sectional Autocorrelation

It only occurs in cross sectional and panel data models. This means the in the cross sectional model where each cross sectional observation is a different person and in such case if the error is related to the other error. It can only occur if the two students are studying together.One is intelligent, and the other is not, so their marks are dependent on each other. So this shows a missing variable of coordination between students. Gujarati (2009) states that if there is cross sectional autocorrelation it means that there aremissing important variables.

This spatial spillover effect is common in panel data sets because of the phenomenon of trade integration, financial integration and other networking between countries (Holly *et al.*, 2010).

Time series autocorrelation

> *And when it is said to them, 'Come to what Allah has revealed and to the Messenger," they say, "Sufficient for us is that upon which we found our fathers.' Even though they father knew nothing, nor where they guided?* Al Qur'an (5:104).
>
> *And when it is said to them, "Follow what Allah has revealed," they say, "Rather, we will follow that which we found our fathers doing." Even though their fathers understood nothing, nor were they guided?* Al Qur'an (2:170).
>
> *for we only do what our forfathers used to do.* Al Qur'an (26:137).

This problem only comes in time series and panel data models. It means that the residuals are correlated to its past residuals of the same person. This shows the interdependency of the grading. Maybe because of the fact that his grading depends on the grading is past semester which is a prerequisite course. If he performs well in a prerequisite course he can perform well in the advanced course. In statistics, this issue is called non-stationarity of the dependent variable and may be independent too.

Structural Restrictions to Data

In most cases, there does not exist any solution to the post regression problems, or at least its remedy is too complex or it remedy model has not been invented yet.

Consider the example of determinants of damage caused by river floods in districts of Punjab Pakistan. In this case, the damage caused by the flood in the furthest possible district is dependent on how the flood has passed through previous districts.

Fulfilment of standard OLS assumption requires that the entering of the floodwater in each district should be random.For this, all districts would be aligned and facing the flood simultaneously, but in reality, it is not possible.

Under such cases, we use robust regression approaches as the source of the issue is structural.

Un-stability

Un-stability is a sudden change in environment. Consider the case if the students get a scholarship so that he does not have to spend much time on the job for his studies which can allow him to spend more time in the studies. So the results are different in before and after of this even if we do no incorporate this structural break, the model will become unstable.

Endogeneity

Endogeneity comes when the dependent variable itself is causing one of the independent variables. For example, the higher grade is motivating the student to spend more time in the studies.

Time Spent in Studies = f(Study Performance)

Contemporaneous correlation

This is a rare issue; it occurs if error terms of two different models are related to each other like if we can write two equations

Study Performance = α + β_1 (no of hours studied) + β_2 (age of the student) + μ_1

Wages = λ + δ_1 (no of hours spend) + δ_2 (experience) + μ_2

Here since the total time is limited and better wages can help him spend less time in job and more time in studies. Hence, both residuals are expected to be correlated with each other for the same person in the cross sectional model and can also be same in time series or panel data model.

Introduction to STATA

This book has used STATA as primary software to analyze the econometric examples and provided the coding with the example and at the end of each chapter which can be used to recreate the example for learning purposes. The link to the data file is also provided in each chapter. Following are some merits of STATA because of which it is used as a medium for analysis.

1. It is the most versatile software for the estimation of theories using regression analysis
2. The first advantage of this package is that the estimation results can easily be replicated as compared to any other software where the steps needed to be written or remembered. Once you teach your students how to replicate the work, they do not need to remember the process and can actually focus on the Econometrics.
3. The second advantage is that students do not need to have the latest version of STATA, all new features can easily be installed in the older versions.
4. Third advantage is that in the coding method, it generates three files, first data file, second, the code file and

the last is the output file, foreign universities require all three of them to check the originality of the work.

5. Lastly, the help manual and online discussion rooms of STATA are very detailed and helpful in learning new skills.

Post regression diagnostics for cross sectional data

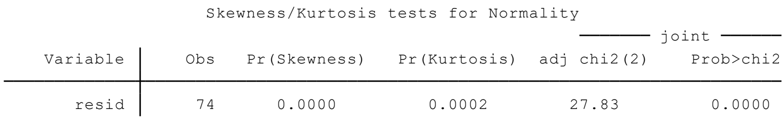
Consider the model,the activity is based onthe determinants of miles per gallon of cars using the auto.xls dataset. The independent variables include the weight of the car and the dummy variable representing the domestic and foreign car. Please remember that this activity is for cross sectional data only, we will refer to the time series or the panel data case when that particular data is studied.

$$MPG_i = \alpha + \beta_1 weight_i + \beta_2 foreign_i + \varepsilon_i$$

Normality test

After the regression, the residuals can be stored into a new variable using the following command and the new variable will then be tested for normality.

predict residuals, resid
sktest residuals



```
                    Skewness/Kurtosis tests for Normality
                                                          ------ joint ------
    Variable |       Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------------
       resid |        74     0.0000         0.0002         27.83         0.0000
```

Here the joint prob. > chi² value is used to determine the normality. Null hypothesis here is that residuals are normal (skewness = 0 & kurtosis = 3) and alternative hypothesis is that the residuals are not normal. Since the p value is less than 0.05 so at 5% we accept the alternative hypothesis.

Similarly, normality can be confirmed graphically using the graph shown below. Here we can see that the dotted line is the normal distribution as compared to that the solid line is the actual distribution of residuals, we can see that actually has higher peak, which is left of the peak of normal distribution (skewness ≠ 0) and fatter tails (kurtosis ≠ 3).

```
kdensity residuals, norm scheme(sj)
```

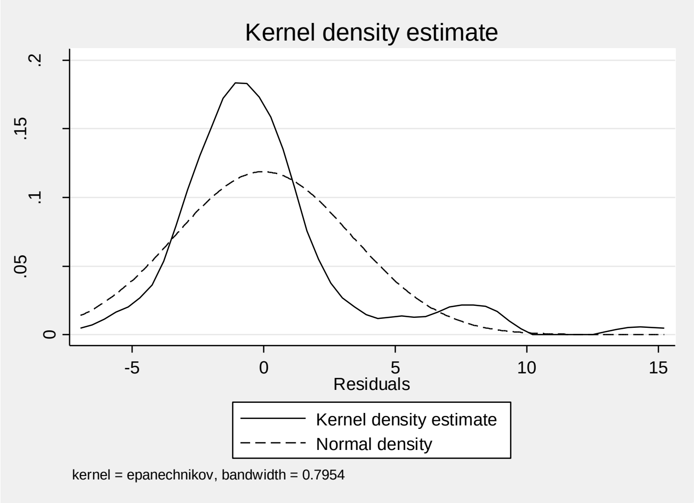


Figure 8. *K Density Graph for normality*

Other tests include pnorm, qnorm and swilk, which can identify if the variable is normal or not.

Heteroscedasticity test

Since it is a cross sectional data so the heteroscedasticity test is actually testing cross sectional heteroscedasticity. Following command once mentioned after the regression results, it will check the presence of heteroscedasticity.

estat hettest

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of mpg

         chi2(1)      =     7.89
         Prob > chi2  =   0.0050
```

Here the prob. > chi^2 is used to decide. Null hypothesis is already mentioned in the test the alternative hypothesis is that the model has non-constant variance (i.e. there is heteroscedasticity). Since p value is less than 0.05 so at 5% we can conclude that there is heteroscedasticity in the model.

Here residual versus fitted (rvfplot) values plot can be used to have a qualitative assessment for the presence of heteroscedasticity.

Mis-specification test

Ramsey proposed the misspecification test. This shows the presence of non-linear version of independent variable missing. Following command, when mentioned after the regression results, will check the model being misspecified or not.

estat ovtest

```
Ramsey RESET test using powers of the fitted values of mpg
       Ho:  model has no omitted variables
                  F(3, 68) =      2.10
                  Prob > F =      0.1085
```

Here p value is higher than 0.05 so using the 5% criterion, we can conclude that the null hypothesis is accepted, which means model is not misspecified.

Mis-specification can also be detected using the linktest command. It uses linear and squared form of the estimated dependent variable.

Multicollinearity test

Multicollinearity in the model shows the level of association between the independent variables. It is checked using the VIF (variance inflating factor) value. Following command is used after the regression results will be used to check multicollinearity.

estat vif

```
    Variable |       VIF       1/VIF  
-------------+----------------------
     foreign |      1.54    0.648553
      weight |      1.54    0.648553
-------------+----------------------
    Mean VIF |      1.54
```

Gujarati (2009) explains that if the VIF value is above 10,it shows the presence of multicollinearity.Since both VIF values are smaller than 10 so there is no multicollinearity in the model.

Autocorrelation test

Since the data is in cross sectional form, so technically, this autocorrelation test is checking the presence of cross sectional autocorrelation in the data. STATA does not provide Durbin Watson value for cross sectional data, so it should be calculated indirectly to replicate the following formula. Or the following module can be installed to generate the Durbin Watson value.

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

lmadw mpg weight foreign, lags(1)

```
=================================================
* Durbin-Watson Autocorrelation Test            *
=================================================
 Ho: No Autocorrelation - Ha: Autocorrelation

-----------------------------------------------------------------
* Rho Value for                AR(1) =    -0.2134
* Durbin-Watson Test           AR(1) =     2.4211   df: (3 , 74)
-----------------------------------------------------------------
```

Ideally, the Durban Watson value must be equal to 2 so that there is no autocorrelation; it ranges between 0 and 4. Here it is, slightly bigger, so there is a hint of autocorrelation. In cross sectional models, autocorrelation only indicates that there are some important variables missing which is making residuals nonrandom.

Non-linearity test

The augmented component plus residual plot (acprplot) can be used to investigate individual independent variables if they have been correctly assumed to have a linear effect or not. Figure 9 shows the scatter plot of the single independent variable against the dependent variable.If the scatterings are closely placed with the line it means that the model has correctly assumed linear relation. A similar conclusion can be drawn from component plus the residual plot (cprplot).

acprplot weight, title(Non-Linearity Plot) subtitle(Independent variable : Weight) scheme(sj)

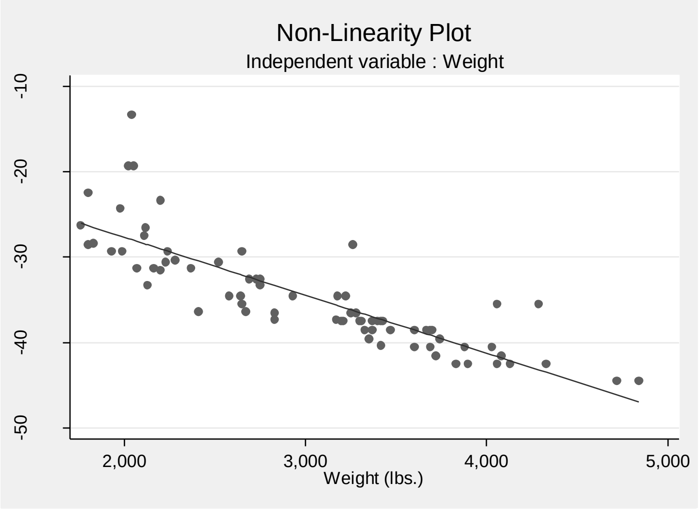


Figure 9. *ACR Plot for Cross sectional regressions*

Detection of outliers

Data can have some outlier observations which is located far from remaining observations, and it has undesirable effects on the slope of coefficients. There are two ways to identify it.

lvr2plot, title(Leverage-Residual square Plot) subtitle(outlier detection) scheme(sj) [4]

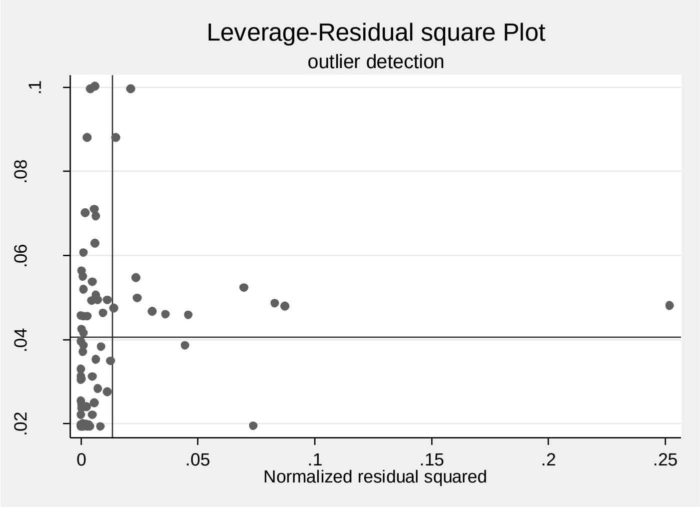


Figure 10. *Leverage versus Squared Residual Plot*

In figure 10, the most problematic observations are which are above the horizontal reference line and right of the vertical reference line. There are about 10 observations which are shown has outliers and they might wrongly influence the coefficients.The added variable plot shown below, detects the outlier observations for each slope coefficient. STATA provides other plots like residual versus predictor (rvpplot) plot or residual versus fitted (rvfplot) plot to identify outlier observations in the data.

[4] You can use lvr2plot, mlabel(make) mlabp(0) m(none) mlabsize(small) command to display the name of each dot so that you can locate the problematic observation.

avplots, title (Added variable Plot) subtitle (outlier detection) scheme(sj)[5]

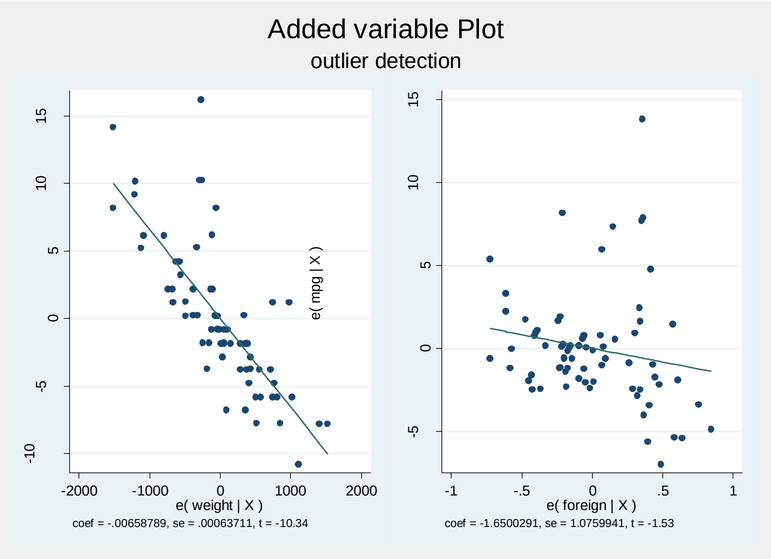


Figure 11. *Added variable plot*

Endogeneity test

The problem of endogeneity exists when the reverse version of the model also has significant F test values.

$$Weight_i = \alpha + \beta_1 MPG + \beta_2 foreign_i + \varepsilon_i$$

F test = 101.67 P value = 0.00

$$Foreign_i = \alpha + \beta_1 weight_i + \beta_2 MPG_i + \varepsilon_i$$

F test = 21.05 P value = 0.00

[5] Similarly avplots, mlabel(make) mlabp(0) m(none) mlabsize(small) command can display the names on each dot to identify problematic observation for each slope.

Since both models have significant F values, it means reverse models are valid too, so there is endogeneity.

Auxiliary information

After regression analysis, we can extract some useful information like partial R squares. It will help in identifying the share of each variable's contribution in explaining the dependent variable.

reg mpg weight foreign
esta tesize
reg mpg weight foreign

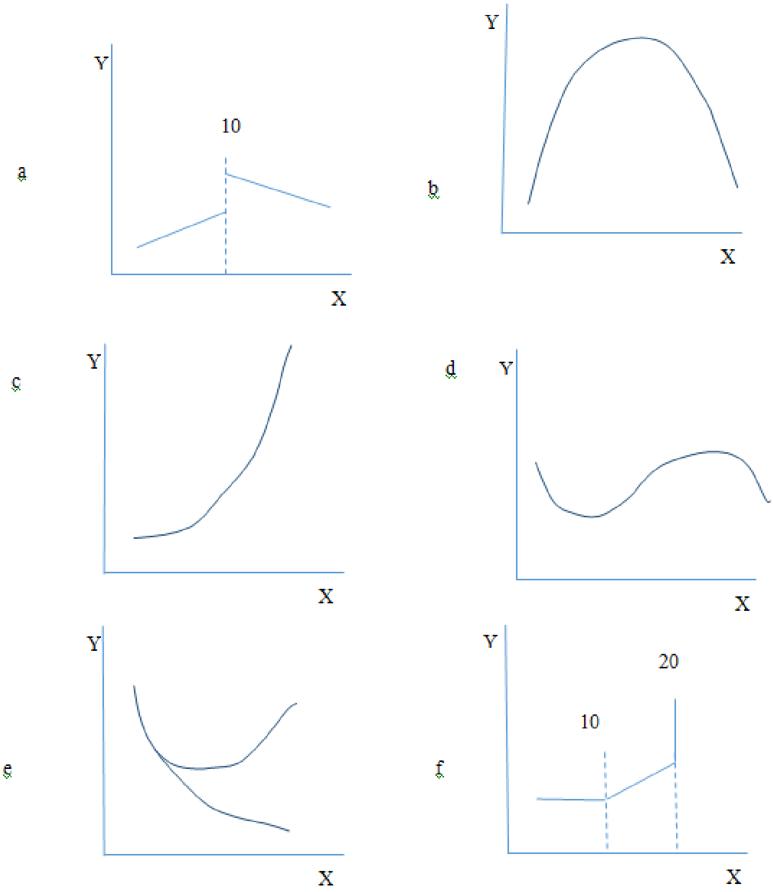
```
Effect sizes for linear models

─────────────────────────────────────────────────────────────────────
             Source │  Eta-Squared      df     [95% Conf. Interval]
────────────────────┼────────────────────────────────────────────────
              Model │   .6627029         2     .5214871    .7399246
                    │
             weight │   .6009445         1     .4498421    .6954668
            foreign │   .0320593         1            0     .143994
─────────────────────────────────────────────────────────────────────
```

Case Study I

Since we have already defined what dummy variables are how to interpret them? Here are some graphical patterns, some are changing smoothly and some are changing abruptly. First of all we will discuss how they are made, they are actually scatter plots of two variable on which we have traced a line which is representing the orientation of the scatter plot, just like figure 4.

Based on the diagram, use the dummy variable and other means to write the equation that can represent that pattern.

Patterns of Relationship between two variables



Case Study II

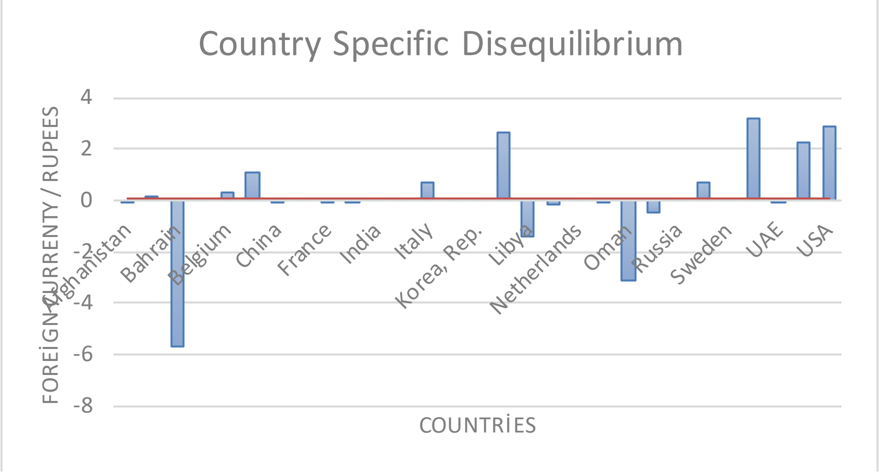
How we adapt from estimation models

Consider a panel data model of the exchange rate. Here exchange rate (ER) depends on domestic prices (CPID), foreign prices (CPIF), domestic interest rate (IRD), foreign interest rate (IRF), domestic money (M2D), foreign money (M2F), domestic GDP (GDPD) and foreign GDP (GDPF). The estimation approach will be discussed in the last chapter, while we focus on interpretation and application here.

$$ER_{it} = \alpha_i + \beta_1\, CPIF_{it} + \beta_2\, CPID_{it} + \beta_3\, IRF_{it} + \beta_4\, IRD_{it} + \beta_5\, GDPF_{it} + \beta_6\, GDPD_{it} + \beta_7\, M2F_{it} + \beta_8\, M2D_{it} + \mu_t$$

Fully Modified Least Squares (FMOLS) Dependent Variable: LER		
Independent Variable	Coefficient	P-value
LCPID	0.84	0.00
LCPIF	-0.87	0.00
IRD	0.0002	0.68
IRF	-0.005	0.00
LGDPD	-0.16	0.00
LGDPF	0.26	0.00
LM2D	-0.11	0.00
LM2F	0.01	0.00
Regression Diagnostics		
Jarque – Bera	1039	0.00
R-squared	0.99	
Total panel (unbalanced) observations	n = 24, t = 29, total = 546	

The actual data of exchange rate is called actual (Y) exchange rate, but the estimated exchange rate $(\hat{Y})$ is called equilibrium (theoretical) exchange rate. So if Y >$\hat{Y}$ then it is expected that Y will fall in future and vice versa. Following graphs shows Y - $\hat{Y}$ value, it can be observed that model has predicted that exchange rate of few countries is expected to change.



But you can see from the table below, there are several currencies which did not performed as prediction. What might be the reason? Does it show that model is faulty?

Country	ER in 2015	ER in 2016	Actual Change	Y - Ŷ
Afghanistan	1.68	1.54	-0.14	-0.01
Austria	8.33	8.47	0.14	0.19
Bahrain	273.40	278.72	5.32	-5.66*
Bangladesh	1.32	1.34	0.02	0.00
Belgium	3.01	2.87	-0.14	0.29*
Canada	80.37	79.07	-1.30	1.10*
China	16.51	15.77	-0.73	-0.02
Finland	113.97	115.93	1.96	0.09
France	18.54	17.63	-0.91	-0.04
Germany	62.21	59.12	-3.09	-0.05
India	1.60	1.56	-0.04	0.01*
Indonesia	0.01	0.01	0.00	0.00*
Italy	0.06	0.06	0.00	0.67
Japan	0.85	0.96	0.11	0.02
Kuwait	341.70	346.82	5.13	2.63
Korea, Rep.	0.09	0.09	0.00	0.00*
Libya	74.43	75.38	0.95	-1.43*
Malaysia	26.32	25.26	-1.06	-0.17
Netherlands	113.97	115.93	1.96	0.09
Norway	12.75	12.48	-0.27	-0.08
Oman	267.36	272.56	5.20	-3.14*

Qatar	28.24	28.79	0.55	-0.43*
Russian	1.69	1.56	-0.12	0.11*
Spain	113.97	115.93	1.96	0.71
Sweden	12.19	12.24	0.05	0.11
Switzerland	106.82	106.37	-0.45	3.18*
UAE	27.99	28.54	0.54	-0.12*
UK	157.06	141.50	-15.56	2.24*
US	102.80	104.80	2.00	2.87

Summary

As soon as humans started to make policies from the estimates of ordinary least squares, this altered their behavior patterns which later made OLS redundant. From now on, estimating behaviors requires a detailed investigation of OLS, if any of the post regression diagnostics indicate new information, it advocates modification of estimation approach.

There is a vast nomenclature of variable names, few of them are overviews in this chapter, whose details will come in relevant chapters. Also there are three types of data structures, like cross sectional data, time series data and panel data. They are used on two bases, first because of data limitation and second because of specific objectives.

In the remaining chapters, we will discuss each of OLS limitations one by one, and come up with possible solutions.

Application questions

i. For what theoretical instances intercept can be removed from the regression. What about normalized variable or demeaned variable regression, do we need intercept in those equations?

ii. What are possible purposes of using variables in natural logarithmic form in the regression? What will you do if you have negative, zero or missing values in the data?

iii. What are the ways two regressions can be compared with each other for the case when dependent variable is same?

iv. What are the limitations of Durban Watson autocorrelation test?

v. Suppose same model has been estimated for two countries separately shown below, what are ways we can confirm that β1& β2 are equal?

$$Y_1 = \alpha_1 + \beta_1 X_1 + \varepsilon_1$$
$$Y_2 = \alpha_2 + \beta_2 X_2 + \varepsilon_2$$

xii. For primary data provided below, construct the dummy variable for the following statements.

a. Make a dummy which is = 1 for full time employed[6] and = 0 for other

b. Make a dummy which is = 1 for part time worker[7] and = 0 for full time worker

c. Make a dummy which is =1 for unemployed and = 0 other

d. Make a dummy which is =1 for underemployed and = 0 for other employed

e. Make a dummy which is = 1 for overtime workers and = 0 for other workers

Table 5. *Construction of Dummy Variables*

Obs.	Week hours	Dummy a	Dummy b	Dummy c	Dummy d	Dummy e
1	46					
2						
3	46					
4	35					
5	56					
6	23					
7	23					
8						
9	56					
10	35					

[6] Full time employed means 8 hours work for 5 days and 6 hour work on Friday. And one day off

[7] Half number of hours than the full time worker

Further study

Arshed, N. (2016, May 14). Model selection criterion for cross sectional data. [Retrieved from].

Arshed, N. (2015, Nov 15). Possible issues of regression estimates – Applied examples. [Retrieved from].

Arshed, N. (2015 Sept. 29). Possibility of negative R squared. [Retrieved from].

Arshed, N. (2015, Sept. 13). Insignificant coefficient: Useless or perfect variable? [Retrieved from].

Arshed, N. (2015, Apr 26). Multicollinearity: A statistical vs. conceptual concept debate. [Retrieved from].

Arshed, N. (2014, Jul 17). Possible problems with regression estimates. [Retrieved from].

Arshed, N. (2013, Nov 30). What does a statistical distribution tells us? [Retrieved from].

Arshed, N, (2013, Nov 5). Problem of mixing real and nominal variables – A query. [Retrieved from].

Arshed, N. (2013, Oct 10). Type I and Type II Errors: Monte Carlo simulations. [Retrieved from].

Arshed, N. (2013, Sept 26). Monte Carlo experiments: Checking biasness of an estimator. [Retrieved from].

Arshed, N. (2013, Oct 2). Biasness of estimator – Revisited – Residual Analysis. [Retrieved from].

Coxe, S., West, S.G., & Aiken, L.S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment*, 91(2), 121-136.

Giles, D. (2015, Feb 4). Four different types of regression residuals. [Retrieved from].

Giles, D. (2016, Apr 14). Regression coefficients & units of measurements. [Retrieved from].

Giles, D. (2014, Feb 15). Some things you should know about the Jarque-Bera test. [Retrieved from].

Giles, D. (2013, May 3). When will the adjusted R-squared increase? [Retrieved from].

Hemenway, D. (2009). How to find nothing. *Journal of Public Health Policy*, *30*(3), 260-268. doi. 10.1057/jphp.2009.26

Kennedy, P.E. (2005). Oh no! I got the wrong sign! What should I do? *The Journal of Economic Education*, 36(1), 77-92. doi. 10.3200/JECE.36.1.77-92

Replication Network (2017). MENCLOVA: Is it time for a Journal of Insignificant Results? [Retrieved from].

Silva, J.S., & Tenreyro, S. (2006). The log of gravity. *The Review of Economics and statistics*, 88(4), 641-658. doi. 10.1162/rest.88.4.641

Varian, H.R. (2016). How to build an economic model in your spare time. *The American Economist*, 61(1), 81-90. doi. 10.1177/0569434515627089

3 Models for heteroskedasticity

Learning Outcomes

- Exploring the problem of heteroscedasticity and its sources
- Possible estimation approaches which can counter this issue

"Never theorize before you have data. Invariably, you end up twisting facts to suit theories instead of theories to suit facts" – Sherlock Holmes

Introduction

Heteroscedasticity is described as the situation prevailing in the estimation results, which makes the variance of the model (residuals) to become non-constant. This can be because of the construction of the data or because of the nature of the variables. Some cases popular sources of heteroscedasticity are following

a. The units [1] or scale [2] of the dependent and independent variables might mismatch such that their change inherently causes variance to switch gears. Such

[1] One variable can have units in currency units and other variable have units in percentage, since the range of the currency unit variable is expected to be longer to it will vary more than the percentage unit variable.

[2] One of the variable in regression can be an index with very limited range such that this variable stays invariant for longer durations.

cases, we use transformation in the data set. The common transformation includes natural logarithm (Benoit, 2011), differencing, and growth rates (Asteriou & Hall, 2007; Ch. 2).

b. Another source can be that the intercept or certain slopes of the variables might change within the sample because of change in the environment. It could be an indicator for use for dummy variables for the discrete environment or use of moderator approach when there can be continuous environment (Hayes, 2013).

c. For discrete dummy variables, if the qualities are inter-related, it could lead to heteroscedasticity. Such type of heteroscedasticity is managed using the panel data models.

d. In some cases, the qualitative, discrete or categorical variable becomes a dependent variable which causes inaccurate estimates to lead to heteroscedasticity.

e. Lastly, there can be frequency mismatch, or there is a regime change in the sample.

This chapter is mainly focusing the second last source of heteroscedasticity as it has few variants all are addressed using an alternative to OLS.

Sometimes, mostly in the survey-based study, the aspect which is studied as a dependent variable is discrete data. Discrete data means that either it has a lower / upper limit or it is binomial/multinomial. Initially, people used the OLS approach by assuming the dependent variable to be continuous. Discussion about what is the limitation of the OLS model what makes it not suitable for using when the dependent variable becomes discrete. (Gujarati & Sangeetha, 2007; Ch 15, figure 15.1) Some models are summarized below.

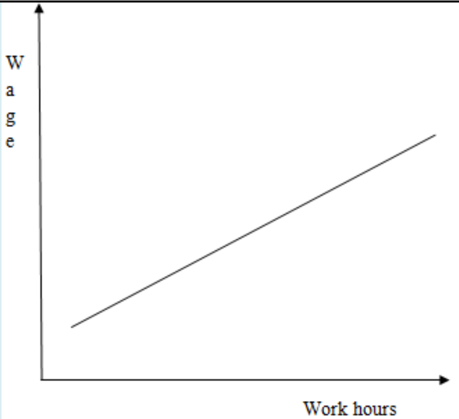
Case Study III

In this case,will illustrate the cases when heteroscedasticity is because of the composition of the independent variables, as mentioned in point b above.

We regularly study the term control variables in the estimations and research papers. This activity is about understanding the role of the control variable.

Control variables can be of two types, and they can be used in two ways

	Intercept Changing	Slope Changing	Both
Discrete Qualitative Control variable	Simple dummy	Interactive dummy	Interactive dummy
Continuous Qualitative control variable	Trend variable	Trend cross product	Trend and trend cross product
Quantitative Control Variable	Simple independent variable	Cross product	Moderator regression (both)
Quantitative Control Variable (Self) Special case	-	Square Form Or Cube Form	-

Consider the model of labor supply. Where positive slope indicates the increase in wage will promote the worker to work more. Here work hours is dependent but because of common convention for this model we keep it at x axis.	

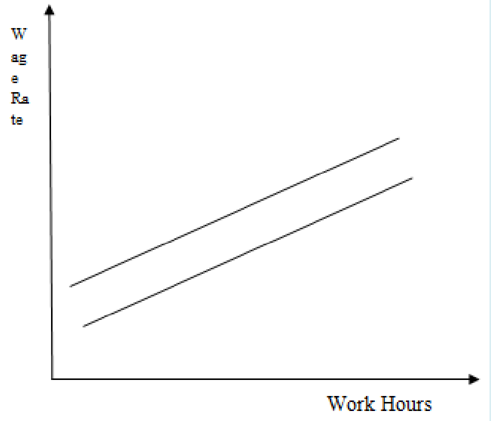
Qualitative Discrete Control

In this case we use the help of dummy variables

a) Intercept only

This case is about the difference in gender

WH = α + β WR + δ D + e
D = 1 for males D = 0 for females



a) slope only

In this case we can use the quality of promotion

WH = α + β WR + δ D * WR + e

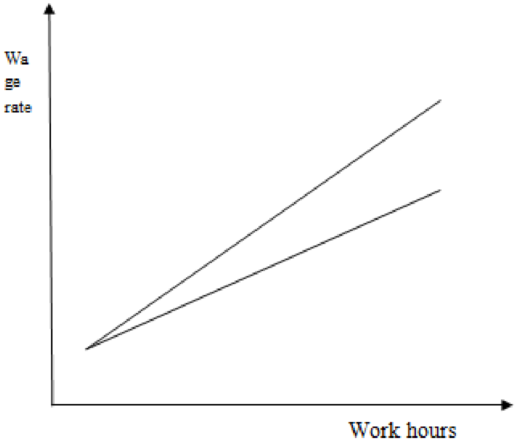
D = 1 for promoted worker
D = 0 for non-promoted worker

Al Quran (3:159)

"So by mercy of Allah, [O Muhammad], you were lenient with them. And if you had been rude [in speech] and harsh in heart, they would have disbanded from about you…."

Al Quran (20:44)

"So speak to him in soft words. May be, he accepts the advice or fears (Allah)"



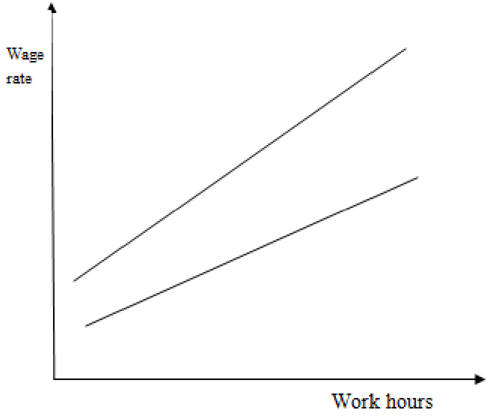
a) Both

Here we can talk about a motivated worker

$WH = \alpha + \beta\, WR + \delta\ D + \delta 2\ D * WR + e$

D = 1 for motivated worker

D = 0 for non-motivated worker



Qualitative continuous Control

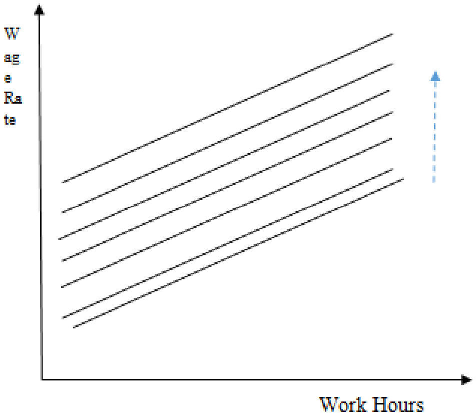
in this case we use the help of trend variables

d) Intercept only

This case is about level of experience at the hiring stage

$WH = \alpha + \beta\, WR + \delta\, T + e$

There will be infinite lines parallel for each T from T = 0 to the last value of T = n



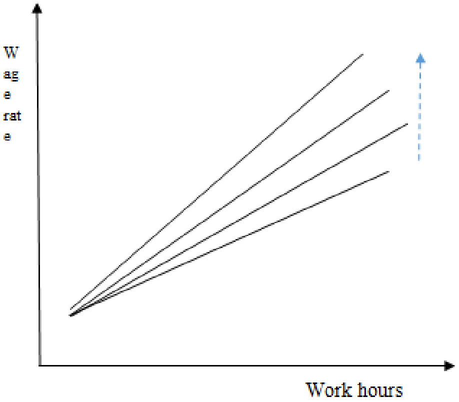
e) slope only

In this case it will show the effect of gain in experience in work

$WH = \alpha + \beta\, WR + \delta\, T * WR + e$

Here there will be infinite number of new lines rotated

Upwards for each T from T= 0 to T = n



e) Both

If we incorporate both effects of experience at start

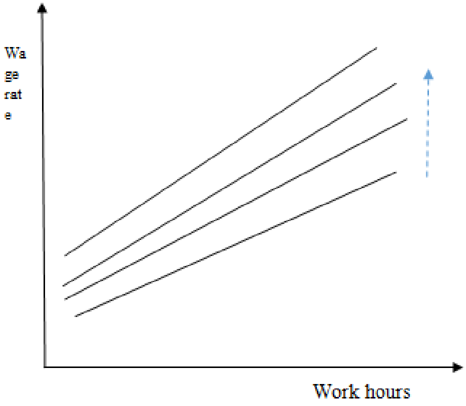
And experience during the job

$WH = \alpha + \beta WR + \delta T + \delta 2 T * WR + e$

Here there will be infinite number of new lines rotated

Upwards and shifted upwards for each T from

T= 0 to T = n



Quantitative Control

In this case we use the help of actual variables, the difference here is that this control variable is measureable while the above one was not directly measurable.

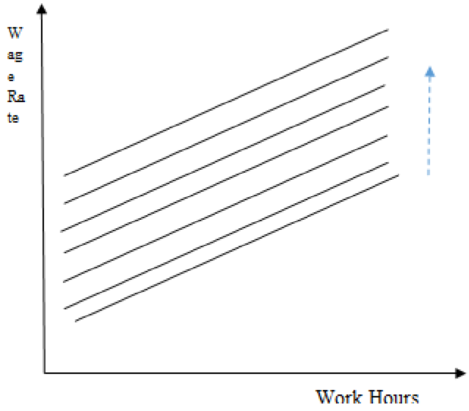
e) Intercept only

This case can be for the education level of the worker

$WH = \alpha + \beta WR + \delta EDU + e$

There will be infinite lines parallel for each Education

from Education = 0 to the last value of education = n



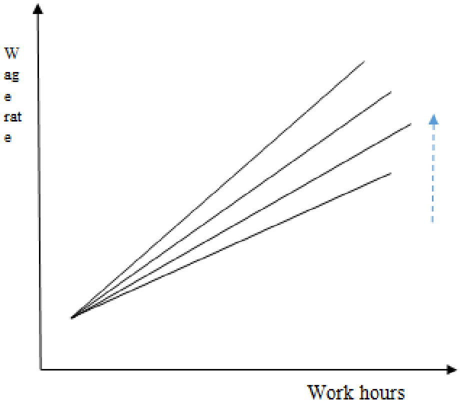
e) slope only

In this case it will show the effect of Education attainment during work

$WH = \alpha + \beta\, WR + \delta\, EDU * WR + e$

Here there will be infinite number of new lines rotated

Upwards for each EDU from EDU= 0 to EDU = n



e) Both

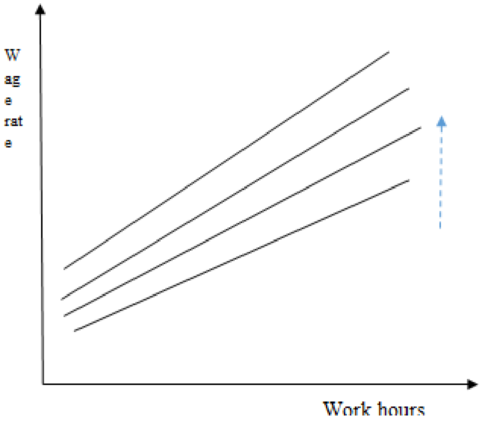
If we incorporate both effects of education attainment at start and education attainment during the job

$WH = \alpha + \beta\, WR + \delta\ EDU + \delta 2\ EDU * WR + e$

Here there will be infinite number of new lines rotated

Upwards and shifted upwards for each EDU from EDU= 0 to EDU = n

The cases b, c, e, f, h, and i are also known as Moderator model, mostly used in social sciences and cross sectional studies



Linear probability model

Even if the dependent is discrete, OLS can still fit a linear probability line and ensures a minimum sum square residuals. It has some issues

- OLS fits a linear approximation which is not appropriate
- OLS can predict more than 1 or less than 0 in the model were dependent cannot be such.

- OLS does not regard the outcome even if the dependent is discrete it provides the estimated value between 0 and 1

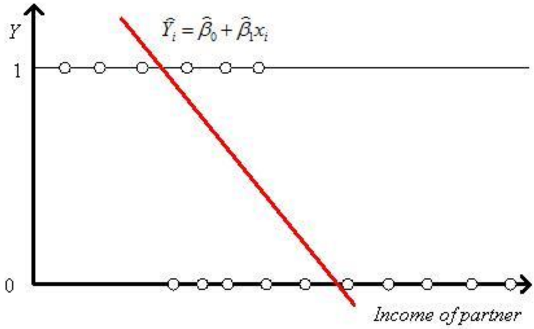


Figure 12. *Fitted line and scatter plot for dummy dependent variable*

But LPM model still can be useful for the purpose ofthe independent variable selection and coefficient sign confirmation.

In this chapter will go through 4 proposed models which are used when the dependent is limited and fits certain conditions.

Logit & Probit model

In this book, all the examples will be provided in the online links available at the start of every example. This data set can be imported in STATA using coping from excel and pasting in STATA. [3]

Berkson (1944) first introduced the logit model, in the case of estimations under the situation where the dependent variable is binomial. This model uses a non-linear approach to fit a model for only two possibilities of the dependent variable. For the example of logit and probit model, the data set is named as lbw.xls. This data is about making a model of

[3] Following link [Retrieved from]. shows a tutorial of how data can be imported in STATA

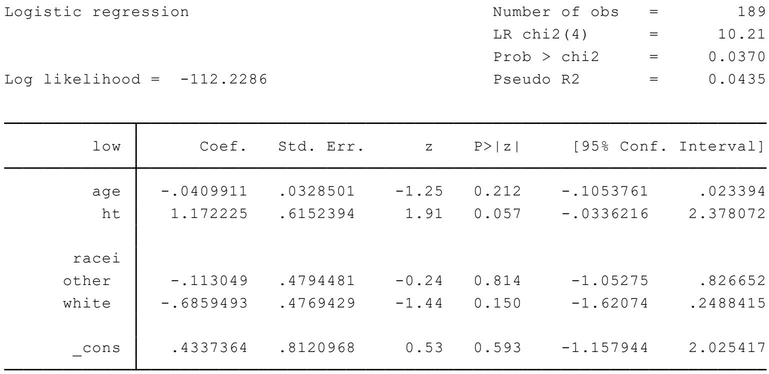
the determinants of an underweight child at birth.[4] The dependent variable in this example is low, which is 1 = child having weight less than 2500g and 0 = if it is high, which is a dummy variable.

Independent variables are

- age of mother (age)
- has a history of hypertension (ht)
- race of the family (race)

Since the race variable is categorical, but it is not coded.So following command will code the categorical variable so that it can be used in the regression. Here there are three categories in the rate variable, 0 = white rate, 1 = race and 2 = other race. You will see in the regression it will show only two categories and one of them will be used as benchmark present in the intercept.

encode race, generate (racei)
logit low age ht i.racei



```
Logistic regression                             Number of obs     =        189
                                                LR chi2(4)        =      10.21
                                                Prob > chi2       =     0.0370
Log likelihood =  -112.2286                     Pseudo R2         =     0.0435

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0409911   .0328501    -1.25   0.212    -.1053761     .023394
          ht |   1.172225   .6152394     1.91   0.057    -.0336216    2.378072
             |
       racei |
      other  |   -.113049   .4794481    -0.24   0.814     -1.05275     .826652
      white  |  -.6859493   .4769429    -1.44   0.150     -1.62074    .2488415
             |
       _cons |   .4337364   .8120968     0.53   0.593    -1.157944    2.025417
------------------------------------------------------------------------------
```

The result shows that only history of hypertension of mother (ht) has a significant impact on the low birth weight. For the case of logit model if the incidence of hypertension increases by 1% it increases the chances of children born

[4] Available at [Retrieved from].

with underweight by 1.17%. The LR test shows significant at 5% level means that model is fit.

Probit low age ht i.racei

```
Probit regression                               Number of obs     =        189
                                                LR chi2(4)        =      10.27
                                                Prob > chi2       =     0.0361
Log likelihood =   -112.201                     Pseudo R2         =     0.0438

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0253732   .0197879    -1.28   0.200    -.0641568    .0134104
          ht |   .7235584   .3812223     1.90   0.058    -.0236235     1.47074
             |
       racei |
      other  |  -.0726842   .2958839    -0.25   0.806    -.6526061    .5072376
      white  |    -.41295   .2906983    -1.42   0.155    -.9827082    .1568083
             |
       _cons |   .2682454   .4941967     0.54   0.587    -.7003623    1.236853
------------------------------------------------------------------------------
```

Post regression diagnostics

estat classification

```
Logistic model for low

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         7             5  |         12
     -     |        52           125  |        177
-----------+--------------------------+-----------
   Total   |        59           130  |        189

Classified + if predicted Pr(D) >= .5
True D defined as low != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   11.86%
Specificity                     Pr( -|~D)   96.15%
Positive predictive value       Pr( D| +)   58.33%
Negative predictive value       Pr(~D| -)   70.62%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    3.85%
False - rate for true D         Pr( -| D)   88.14%
False + rate for classified +   Pr(~D| +)   41.67%
False - rate for classified -   Pr( D| -)   29.38%
--------------------------------------------------
Correctly classified                        69.84%
--------------------------------------------------
```

The above classification test compares the actual dependent variable with the estimated dependent variable. This shows how much time the regression as correctly predicted from the two possible values of the dependent variable, here, this model has correctly specified 69.84% of the time. This table also shows that this table correctly predicts the normal-weight children up to 96.15% and low weight children by 11.86%. This test usually favors the larger group.

estat gof [5]

```
Logistic model for low, goodness-of-fit test

       number of observations =       189
 number of covariate patterns =        66
             Pearson chi2(61) =        59.88
                  Prob > chi2 =         0.5167
```

This test checked the goodness of fit of the model using Pearson Chi^2 test. This test is insignificant showing that model is fit.

lroc

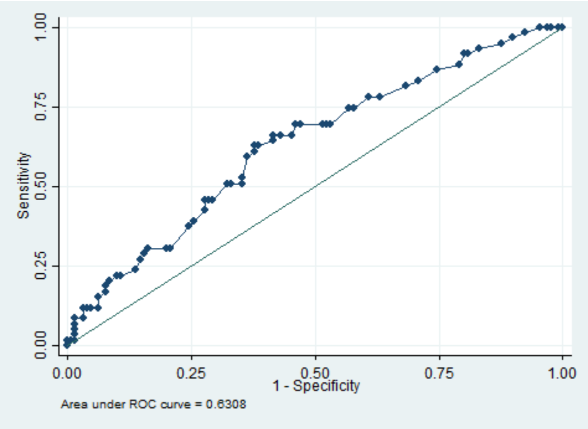


Figure 13. *ROC Curve*

[5] Here the null hypothesis of the Pearson test is that model is fit

Here the 45-degree line shows zero predictive power of the model, the more the prediction more the line will be bending upward. And the area under the curve will show the performance, value 0.5 of the area under ROC shows zero power and 1 shows perfect power.

lsens[6]

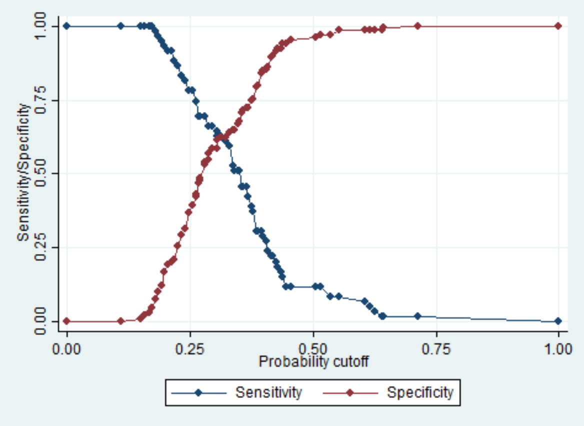


Figure 14. *Sensitivity Graph*

The above graph will plot the sensitivity and the specificity graph, and provides a cutoff value c on the x axis. This cut of value is reverse of the classification value since it is 69.84% so the cut off value is (1- 0.6984) = 0.3016.

Difference between Logit and Probit

The major difference between Logit and Probit models lies in the assumption on the distribution of the error terms in the model. In the Logit model, the errors are assumed to follow the standard logistic distribution while for the Probit, the errors are assumed to follow a normal distribution. In principle for general practice, the model formalism, both work fine and often leads to the same conclusions, especially

[6] It plots sensitivity and specificity

when the sample is large regardless of the problem complexity. When the sample is larger than 30 any distribution approximates to normal distribution[7], hence it does not matter in this economics which model is used or not.

Is Logit better than Probit, or vice versa?

Both methods will yield similar (though not identical) inferences, we can see from the above results too that both have all most same outcome.

Logit also known as logistic regression is more popular in health sciences like epidemiology partly because coefficients can be interpreted in terms of odds ratios. Whereas,Probit models can be generalized to account for non-constant error variances in more advanced econometric settings (*known as heteroskedastic probit models*) and hence are used in some contexts by economists and political scientists. If these more advanced applications are not of relevance, then it does not matter which method you choose to go with, they can be used to see the robustness of the estimates. And you can use the gof, roc and classification criteria to decide which is better.

Tobit model

This tobit model is used when the dependent variable has naturally defined upper or lower limit it can be 1 and 0 respectively and anything between it. The perfect example is GPA its lower limit is 0 and upper is 4. So if we run OLS, it will not make sure that the estimates should not exceed the 4 or stay below 0. Or in this model, you can just sensor any data for its own artificially lower or upper limit and like you are working on determinants of consumption where you

[7] Central limit theorem

need to filter the data for expenditures from 15K to 40K only and exclude remaining.

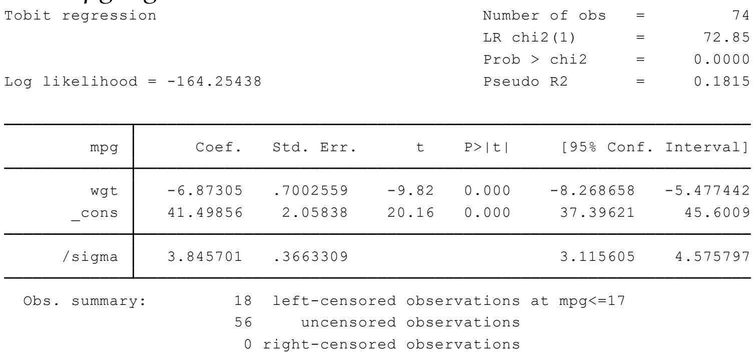
Here's the thing to note that, you can do it by yourself too by removing data and applying OLS but OLS will still not try to stay within the boundaries.

For the tobit model example, auto.xls file is used[8] This is the data for determinants of mpg of a car. And we wish to estimate the determinants of the car whose mpg is higher than 17 miles so we want to exclude heavier cars.

generate wgt = weight/1000

This code is just to generate a new variable converting weight, which was in lbs to a bigger unit of 1000 lbs.

tobit mpg wgt, ll(17)



```
Tobit regression                                Number of obs     =         74
                                                LR chi2(1)        =      72.85
                                                Prob > chi2       =     0.0000
Log likelihood = -164.25438                     Pseudo R2         =     0.1815

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wgt |   -6.87305   .7002559    -9.82   0.000    -8.268658   -5.477442
       _cons |   41.49856    2.05838    20.16   0.000     37.39621     45.6009
-------------+----------------------------------------------------------------
      /sigma |   3.845701   .3663309                      3.115605    4.575797
------------------------------------------------------------------------------
  Obs. summary:         18  left-censored observations at mpg<=17
                        56     uncensored observations
                         0 right-censored observations
```

Above is the code for tobit model, here first is dependent, then, all independents then after the comma sign you can either write nothing which will mean that tobit will assume the lowest value of the data to be the lower limit and highest value of the data to be the higher limit, here as per requirement we introduced the "lower limit" by writing ll() we can also add "upper limit" like ul(). Its application is that

[8] Available at [Retrieved from].

we are looking for determinants of middle income countries out of all the countries in the sample data. So instead of removing the irrelevant counties manually, this method will exclude it by itself.

These results show that if the weight of the automobile increases by 1000 lbs the mpg is decreased by 6.8 miles per gallon. In the obs. summary we can see that 18 observations are removed from the data as they were smaller than 17 mpg.

predict yhat, xb [9]

We can use the above predict command to generate the estimated dependent variable after these types of regressions.

Multinomial Logit model

The multinomial logit model is used when the dependent variable is from Likert scale or it has more than 2 categories in the form of a dummy variable. This method is available in SPSS.

Multinomial logit model in STATA

The data is of the insurance market where the dependent of insure variable has three categories “indemnity”, “prepaid” & “un-insure”.Where the probability of these will be estimated using the independent variables like age, gender (0 = female, 1 = male), nonwhite (0 = white, 1 = other) and site (a three category dummy). The data file used for this example is sysdsn1.xls[10]

[9] Code to generate the estimated dependent variable

[10] Available at [Retrieved from].

mlogit insure age male nonwhite i.site

```
Multinomial logistic regression                 Number of obs     =        615
                                                LR chi2(10)       =      42.99
                                                Prob > chi2       =     0.0000
Log likelihood = -534.36165                     Pseudo R2         =     0.0387

------------------------------------------------------------------------------
      insure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Indemnity    |  (base outcome)
-------------+----------------------------------------------------------------
Prepaid      |
         age |   -.011745   .0061946    -1.90   0.058    -.0238862    .0003962
        male |   .5616934   .2027465     2.77   0.006     .1643175    .9590693
    nonwhite |   .9747768   .2363213     4.12   0.000     .5115955    1.437958
             |
        site |
          2  |   .1130359   .2101903     0.54   0.591    -.2989296    .5250013
          3  |  -.5879879   .2279351    -2.58   0.010    -1.034733   -.1412433
             |
       _cons |   .2697127   .3284422     0.82   0.412    -.3740222    .9134476
-------------+----------------------------------------------------------------
Uninsure     |
         age |  -.0077961   .0114418    -0.68   0.496    -.0302217    .0146294
        male |   .4518496   .3674867     1.23   0.219     -.268411     1.17211
    nonwhite |   .2170589   .4256361     0.51   0.610    -.6171725     1.05129
             |
        site |
          2  |  -1.211563   .4705127    -2.57   0.010    -2.133751   -.2893747
          3  |  -.2078123   .3662926    -0.57   0.570    -.9257327     .510108
             |
       _cons |  -1.286943   .5923219    -2.17   0.030    -2.447872   -.1260134
------------------------------------------------------------------------------
```

In the above results, we can see that this MNL model has used the first category as base, and the coefficients of other categories are now being interpreted in terms of comparison with the base. The LR chi^2 significance shows the model is fit. While interpreting age variable for both categories, it can be said that with the increase in the age of the respondent. The odds of the respondent to choose prepaid insurance against indemnity insurance decreases by 0.01%, but we can see no chance in probability to have no insurance to indemnity insurance as age variable is insignificant here. The base category can be shifted to some other category, just like in the following command.

mlogit insure age male nonwhite i.site, base(2)[11]

```
Multinomial logistic regression                   Number of obs     =        615
                                                  LR chi2(10)       =      42.99
                                                  Prob > chi2       =     0.0000
Log likelihood = -534.36165                       Pseudo R2         =     0.0387

------------------------------------------------------------------------------
      insure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Indemnity    |
         age |    .011745   .0061946     1.90   0.058    -.0003962    .0238862
        male |  -.5616934   .2027465    -2.77   0.006    -.9590693   -.1643175
    nonwhite |  -.9747768   .2363213    -4.12   0.000    -1.437958   -.5115955
             |
        site |
          2  |  -.1130359   .2101903    -0.54   0.591    -.5250013    .2989296
          3  |   .5879879   .2279351     2.58   0.010     .1412433    1.034733
             |
       _cons |  -.2697127   .3284422    -0.82   0.412    -.9134476    .3740222
-------------+----------------------------------------------------------------
Prepaid      |  (base outcome)
-------------+----------------------------------------------------------------
Uninsure     |
         age |   .0039489   .0115994     0.34   0.734    -.0187855    .0266832
        male |  -.1098438   .3651883    -0.30   0.764    -.8255998    .6059122
    nonwhite |  -.7577178   .4195759    -1.81   0.071    -1.580071    .0646357
             |
        site |
          2  |  -1.324599   .4697954    -2.82   0.005    -2.245381   -.4038165
          3  |   .3801756   .3728188     1.02   0.308    -.3505358    1.110887
             |
       _cons |  -1.556656   .5963286    -2.61   0.009    -2.725438    -.387873
------------------------------------------------------------------------------
```

Coefficient restriction test can be applied to the variables in MNL model. The first case is to test a variable in all categories, second one is the case to test all variables in one category and the last case is a particular variable in a single category. In these tests the null hypothesis is that the variable coefficient = 0.

[11] You can fix what category you want to fix

test 2.site 3.site

```
( 1)   [Indemnity]2.site = 0
( 2)   [Prepaid]2o.site = 0
( 3)   [Uninsure]2.site = 0
( 4)   [Indemnity]3.site = 0
( 5)   [Prepaid]3o.site = 0
( 6)   [Uninsure]3.site = 0
       Constraint 2 dropped
       Constraint 5 dropped

           chi2(  4) =   19.74
         Prob > chi2 =    0.0006
```

test [Prepaid]

test [Uninsure]: 2.site 3.site

Since the MNL model provides odd ratios, not the marginal impact on the dependent variable, in such models like logit, probit and MNL we calculate the slope (dy/dx) or elasticity (ey/ex) using the margins command for each category separately.

margins, dydx(*) predict(outcome(1))

```
Average marginal effects                          Number of obs     =        615
Model VCE    : OIM

Expression   : Pr(insure==Indemnity), predict(outcome(1))
dy/dx w.r.t. : age male nonwhite 2.site 3.site

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0026655    .001399     1.91   0.057    -.0000765    .0054074
        male |  -.1302604    .045513    -2.86   0.004    -.2194643   -.0410565
    nonwhite |  -.2060514   .0519716    -3.96   0.000    -.3079139    -.104189
             |
        site |
          2  |   .0070995   .0479993     0.15   0.882    -.0869775    .1011765
          3  |   .1216165   .0505833     2.40   0.016      .022475     .220758
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Here if the age of the respondent increases by 1 year it increases the chances of indemnity by 0.002% which is significant at 10%.

margins, dydx(*) predict(outcome(2))

```
Average marginal effects                        Number of obs    =        615
Model VCE     : OIM

Expression    : Pr(insure==Prepaid), predict(outcome(2))
dy/dx w.r.t. : age male nonwhite 2.site 3.site

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0025142   .0013962    -1.80   0.072    -.0052507    .0002222
        male |   .1177644   .0446537     2.64   0.008     .0302448     .205284
    nonwhite |   .2221527   .0500585     4.44   0.000     .1240398    .3202656
             |
        site |
          2  |   .0608466   .0482687     1.26   0.207    -.0337583    .1554516
          3  |  -.1264342   .0491456    -2.57   0.010    -.2227579   -.0301105
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

margins, dydx(*) predict(outcome(3)) [12]

```
Average marginal effects                        Number of obs    =        615
Model VCE     : OIM

Expression    : Pr(insure==Uninsure), predict(outcome(3))
dy/dx w.r.t. : age male nonwhite 2.site 3.site

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0001512   .0007352    -0.21   0.837    -.0015923    .0012898
        male |    .012496   .0231963     0.54   0.590    -.0329678    .0579599
    nonwhite |  -.0161013    .026312    -0.61   0.541    -.0676718    .0354693
             |
        site |
          2  |  -.0679461   .0241128    -2.82   0.005    -.1152064   -.0206859
          3  |   .0048177   .0314084     0.15   0.878    -.0567416     .066377
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

[12] This will estimate the marginal impacts of each equation

Conclusion & discussions

Hence, these models are used in cross sectional models where the dependent is not a continuous variable. We have discussed four of such models. There are many other models which can be used in this case. Like if the dependent is discrete without any lower or upper limit then Poisson or Negative Binomial regression approach can be used.

Logit & Probit models and future implications

Since these models have only two categories of dependent variable, then, the residuals will have two categories too, which will be reverse to the dependent variable. Means if the probability of dependent = 1 is *p* then the probability of the residuals will be (*1-p*). This means that the model will be *prone to heteroskedasticity*; which can be addressed by using the robust estimation models.

Second the two categories of the dummy dependent can be skewed or ordered for this skewed logit model and ordered logit model respectively.

We cannot always make 4 categories of a quantitative variable to make a multinomial logit model. For that there is quartile regression, which can be used for the case we have to find determinants of four income groups.

Logit and probit model can be applied to panel data for this panel logit model is used where we have to compare between fixed effect logit and random effect logit model to select the appropriate.

STATA code

```
clear
**** Example for OLS logit and probit
import excel "D:\UMT notes\Applied Econometrics\lectures\lecture 2\lbw.xlsx",sheet("Sheet1") firstrow
encode race, generate (racei)
reg low age ht i.racei
estimates store ols1
estat hettest
logit low age ht i.racei
estimates store logit1
margins, eyex(age)
lroc, title(Logit) name(logit1, replace)
probit low age ht i.racei
estimates store probit1
margins, eyex(age)
lroc, title(Probit) name(probit1, replace)
estimates table ols1 logit1 probit1, stats(chi2 df N aic bic)
graph combine logit1 probit1, title(Model Selection)
**** Example for tobit model
clear
 import excel "D:\UMT notes\Applied Econometrics\lectures\lecture 2\auto.xlsx",///
 sheet("Sheet1") firstrow clear
 generate wgt = weight/1000
 tobit mpg wgt, ll(17)
 predict yhat, xb
 *** Example for  multinomial logit model
 clear
 import excel "D:\UMT notes\Applied Econometrics\lectures\lecture 2\sysdsn1.xlsx",///
 sheet("Sheet1") firstrow
 webuse sysdsn1
 mlogit insure age male nonwhite i.site
 mlogit insure age male nonwhite i.site, base(2)
test 2.site 3.site
test [Prepaid]
 margins, dydx(*) predict(outcome(1))
 margins, dydx(*) predict(outcome(2))
 margins, dydx(*) predict(outcome(3))
```

Summary

Heteroskedasticity is originated because of differences in the cross sections of the data, it can be people, firm or country. OLS originally assumes them same, but with the evolution of knowledge every cross section made their own path, now they are different. This chapter addressed the presence of heteroskedasticity which can be present in independent variables as well as in dependent variables.

For the case of heteroskedasticity in the independent variable, the solutions discussed are the transformation of variables or transformation in the specification. For the case of the dependent variable, the discussed models are Logit, Probit, Tobit and MNL. While other proposed models are Poisson and Negative binomial regression which are used when the dependent variable is limited, and lastly 'BETAREG' model is used when dependent is a ratio and is between 0 and 1.

Application questions

1. Researcher experimented on the logit/probit regression, in his dependent variable *marriage*; 1 = married person and 0 = unmarried and he has used some variables as below

$$marrige_i = \alpha_1 + \beta_1 income_i + \beta_2 gender_i + \varepsilon_i$$

He has used a gender variable where 1 = male and 0 = female. Unfortunately, he is unable to find the literature on what should be the coefficient sign of this variable should be. Please provide possible justification of what would be the possible effect of this variable.

Further study

Giles, D. (2012, Jul 13). More comments on the use of LPM. [Retrieved from].

Giles, D. (2016, Jun 25). Choosing between the Logit and Probit models. [Retrieved from].

4 Instrumental variable regression

Learning Outcomes

- Exploring the problem of endogeneity and what information does it provide
- Possible models to incorporate this information

Incorporating endogeneity

One of the assumptions of OLS model that, all the independent variables in the regression should be strictly exogenous. But what will happen if it is not true, it will create the issue that the model will become inconsistent as it will create an association of the endogenous variable with the residuals.

Today's lecture is incorporating information of endogeneity; this issue comes when one of the independent variables is actually a function of the dependent variable.

$$y = f\ (x1,\ x2,\ x3)\ \&\ x2 = f(y)$$

If the above relations are two, then it will create a non-zero covariance for first equation.

$$\operatorname{cov}\ (x2, \varepsilon) \neq 0$$

But this method has many flaws as we will use estimated ε to check the covariance which is not realistic. The other way is to compare the estimation results of OLS and 2SLS (2 stage least squares); if they are same then we can say that there is no endogeneity. In the background in order to check if the x2 is endogenous we have to find few exogenous instruments which affect x2 but are un-related to the dependent variable y. See the video provided to understand the illustration.

The model we are discussing today is the model of the determinants of wage. Here we have proposed that experience, experience squared, and education are expected to be the determinants. The data set used is the mroz.xls file [1]

First of all we will generate a log of the wage and square form of experience

```
gen lwage = ln(wage)
gen expersq = exper^2
sum lwage exper expersq educ
```
[2]

So model states that the wage is the function of education and experience of the individual.

Here we can see that any random shock that affects the education of the individual also influences the wage of that individual (theoretically), which makes education variable not exogenous. The problem here is that if we introduce an independent variable which is actually endogenous, then it gets correlated with the residuals making regression inconsistent. So we have to introduce some instruments which must be highly correlated with education but not correlated with the residuals of wage model. These instruments should not be directly effecting wage otherwise

[1] Available at [Retrieved from].

[2] This command can be used to see the summary of the variables.

they should be in the regression in the first place. Here, father education, mother education and husband's education become instruments of this regression

$$lwage_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper^2_i + u_i$$

[original estimation suspected of endogeneity]

So the background equation becomes

$$educ_i = \alpha_0 + \alpha_1 feduc_i + \alpha_2 meduc_i + \alpha_3 heduc_i + v_i$$

[Instrument regression]

Here, if the estimates of educ (education variable) are not important in instrument equation.Here we use the instrument technique to make it exogenous by using estimated educ from the instrument equation to be used as independent in original estimation, so that it becomes exogenous to the shocks which are causing lwage. [3,4]

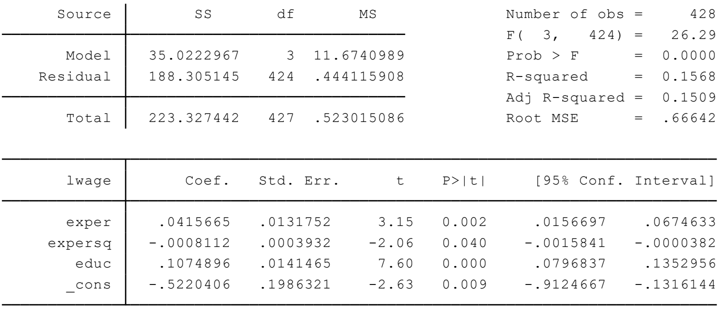
[3] The idea behind is that when we do the regression 1, it will split the dependent variable into two components, one which is estimated dependent variable which is exogenous component and the estimated error term is the endogenous component of the original dependent variable with respect to the dependent variable in the regression 2. This way if we use the estimated dependent variable from regression 1 and independent variable it will solve the problem of this endogeneity.

[4] For further details see this blog [Retrieved from].

Detecting endogeneity using IV regression

1. First run the OLS model and store the estimates

regress lwage exper expersq educ



```
      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(  3,    424)   =     26.29
       Model |  35.0222967         3  11.6740989   Prob > F        =    0.0000
    Residual |  188.305145       424  .444115908   R-squared       =    0.1568
-------------+----------------------------------   Adj R-squared   =    0.1509
       Total |  223.327442       427  .523015086   Root MSE        =    .66642

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0415665   .0131752     3.15   0.002     .0156697    .0674633
     expersq |  -.0008112   .0003932    -2.06   0.040    -.0015841   -.0000382
        educ |   .1074896   .0141465     7.60   0.000     .0796837    .1352956
       _cons |  -.5220406   .1986321    -2.63   0.009    -.9124667   -.1316144
------------------------------------------------------------------------------
```

Here you can see that model is fit as per F test, and the education variable also has significant positive impact on log of wage. But this variable is suspected of having endogeneity. In order to compare both estimates, we will store the coefficient value and its standard error using below command.

estimates store OLS

2. Then run the IVREG with proposed instruments and store the estimates

ivregress lnwage exper expersq (educ = fatheduc motheduc huseduc), first
estimates store IVREG

```
First-stage regressions
-----------------------

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  5,   422) =   63.30
       Model |  955.830608     5  191.166122           Prob > F      =  0.0000
    Residual |  1274.36565   422  3.01982382           R-squared     =  0.4286
-------------+------------------------------           Adj R-squared =  0.4218
       Total |  2230.19626   427  5.22294206           Root MSE      =  1.7378

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0374977   .0343102     1.09   0.275    -.0299424    .1049379
     expersq |  -.0006002   .0010261    -0.58   0.559    -.0026171    .0014167
    fatheduc |   .1060801   .0295153     3.59   0.000     .0480648    .1640955
    motheduc |   .1141532   .0307835     3.71   0.000     .0536452    .1746613
     huseduc |   .3752548   .0296347    12.66   0.000     .3170049    .4335048
       _cons |   5.538311   .4597824    12.05   0.000     4.634562     6.44206
------------------------------------------------------------------------------


Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  3,   424) =   11.52
       Model |   33.392737     3  11.1309123           Prob > F      =  0.0000
    Residual |  189.934705   424  .447959209           R-squared     =  0.1495
-------------+------------------------------           Adj R-squared =  0.1435
       Total |  223.327442   427  .523015086           Root MSE      =   .6693

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0803918    .021774     3.69   0.000     .0375934    .1231901
       exper |   .0430973   .0132649     3.25   0.001     .0170242    .0691704
     expersq |  -.0008628   .0003962    -2.18   0.030    -.0016415   -.0000841
       _cons |  -.1868572   .2853959    -0.65   0.513    -.7478242    .3741097
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper expersq fatheduc motheduc huseduc
------------------------------------------------------------------------------
```

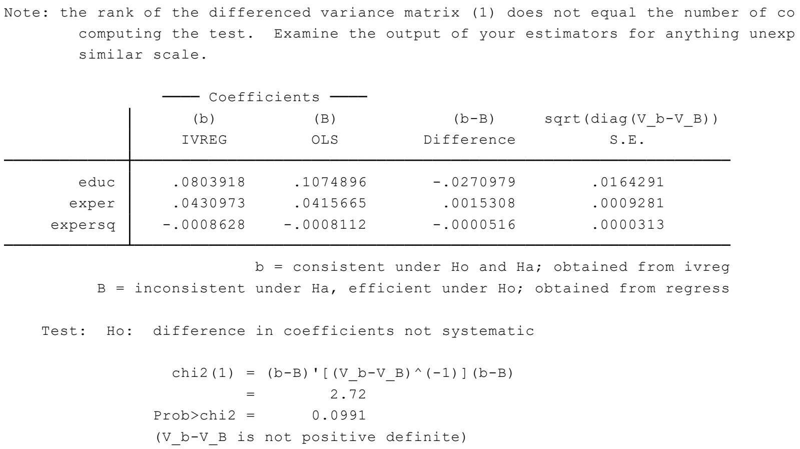
Here we can see that the first one is the instrument regression from where the estimates education variable is generated and used as independent variables in the original equation. If we compare the coefficients in both equations, it has been reduced from 0.17 to 0.080. This indicates that there was some issue in the OLS estimates which, when addressed using IV estimates, the results came out to be different. In order to compare these using Hausman test, we need to identify which one is the efficient model and which one is consistent as there is a tradeoff in our case. Since we know that OLS is a simple model so it is considered as an efficient model, while IV regression by design are consistent as even if there is a problem of endogeneity or not it will not provide wrong results.

3. Then compare the estimates using the Hausman test, which have the following hypothesis.

H_0: The difference between OLS and IVreg is not systematic, hence they are same, so we should use the simpler model as both are performing same. So OLS is an appropriate model.

H_1: The difference between OLS and IVreg is systematic, hence they are different, so we should use the IVreg as it is consistent in the presumption that model us effected by endogeneity

hausman IVREG OLS, sigmamore



```
Note: the rank of the differenced variance matrix (1) does not equal the number of co
        computing the test.  Examine the output of your estimators for anything unexp
        similar scale.

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     IVREG         OLS         Difference          S.E.
-------------+----------------------------------------------------------------
        educ |    .0803918     .1074896       -.0270979         .0164291
       exper |    .0430973     .0415665        .0015308         .0009281
    expersq |   -.0008628    -.0008112       -.0000516         .0000313
------------------------------------------------------------------------------
                          b = consistent under Ho and Ha; obtained from ivreg
           B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        2.72
                Prob>chi2 =      0.0991
                (V_b-V_B is not positive definite)
```

Here we can see that at the 10 % level, the education variable is endogenous variable. And OLS is not a consistent model so we should use IVREG model.

Here you can confirm the presence of endogeneity directly without using the OLS regression by using the following command after the IV estimates.

If this significant then it will show that, using IVreg was necessary and the proposed variable was really endogenous. IV regression has actually identified the presence of endogeneity by noting the difference between the results when we introduced the instruments.

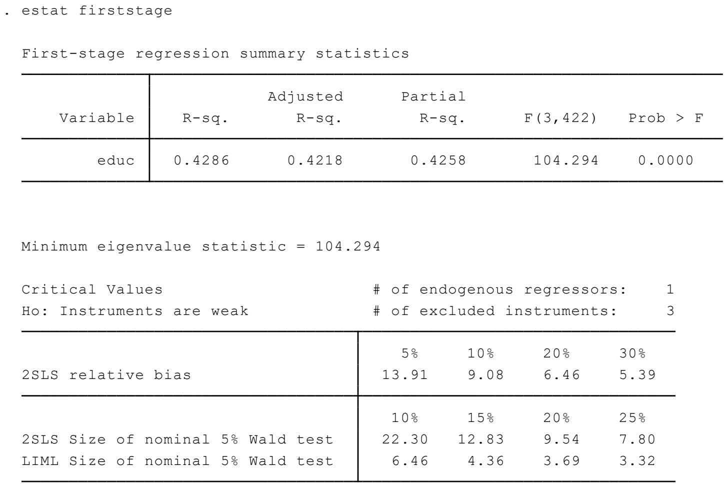
estat endog

```
Tests of endogeneity
Ho: variables are exogenous

Durbin (score) chi2(1)          =  2.74613  (p = 0.0975)
Wu-Hausman F(1,423)             =  2.73157  (p = 0.0991)
```

Both above tests shows that there is endogeneity in the model at 10% level, but if we use the 5% criteria then there is not endogeneity.

Note here that this model will work only in those cases where theoretically reverse relations are visible so that we were able to identify the endogenous variable. Note here that we have confirmed that there was endogeneity which is sorted using IV regression. Now coming tests are for the quality of the instruments utilized in the regression.

estat firststage



```
. estat firststage

  First-stage regression summary statistics
  ─────────────────────────────────────────────────────────────────────────────
               │            Adjusted      Partial
      Variable │   R-sq.       R-sq.        R-sq.     F(3,422)   Prob > F
  ─────────────┼───────────────────────────────────────────────────────────────
          educ │  0.4286      0.4218       0.4258      104.294    0.0000
  ─────────────┴───────────────────────────────────────────────────────────────


  Minimum eigenvalue statistic = 104.294

  Critical Values                      # of endogenous regressors:    1
  Ho: Instruments are weak             # of excluded instruments:     3
  ─────────────────────────────────────┬─────────────────────────────────
                                       │    5%      10%      20%      30%
  2SLS relative bias                   │  13.91     9.08     6.46     5.39
  ─────────────────────────────────────┼─────────────────────────────────
                                       │   10%      15%      20%      25%
  2SLS Size of nominal 5% Wald test    │  22.30    12.83     9.54     7.80
  LIML Size of nominal 5% Wald test    │   6.46     4.36     3.69     3.32
  ─────────────────────────────────────┴─────────────────────────────────
```

The first table shows the contribution of the education variable in the original estimation, the partial R square shows that this variable only explains 42% in the dependent variable of wages. We introduced 3 instruments as an alternative of the education of the individual, here the F test is actually the coefficient restriction test on these instruments, it is significant showing that these instruments sufficiently explain endogenous variable.

In order to check if the instruments are weak or not, the minimum eigenvalue statistic of 104.294 is compared against the critical values.Since the statistic is larger than all critical values, we can say that the alternative hypothesis is accepted that the proposed instruments not weak.

Since we have identified and rectified the problem now we need to confirm that the proposed instruments are not correlated with the error term of the original estimation. If the proposed instruments are more than the number of endogenous variables, then it might raise the issue of over identification which means that instruments are now correlated with the error term of the original model. If they both are equal, then we do not need to do this test. Following is the sargan over identification test.

estat overid

```
Tests of overidentifying restrictions:

Sargan (score) chi2(2) =  1.11504  (p = 0.5726)
Basmann chi2(2)        =  1.10228  (p = 0.5763)
```

Here both tests are insignificant, showing that the proposed instruments are not over identified they are valid.

Once the model is finalized in terms of quality of instruments, we can then check the other diagnostics. First one of them is normality test, which can be done using installed command below.

lmnad2 lwage exper expersq (educ = fatheduc motheduc huseduc), model(2sls)

```
==============================================================================
*** 2SLS-IV Non Normality Anderson-Darling Test - Model= (2sls)
==============================================================================
 Ho: Normality - Ha: Non Normality
------------------------------------------------------------------------------
- Anderson-Darling Z Test              =    5.7401     P > Z( 7.259)     1.0000
------------------------------------------------------------------------------
```

Above test indicates that the model is normally distributed. For the case of autocorrelation test, following syntax can be installed and used.

lmabpg2 lwage exper expersq (educ = fatheduc motheduc huseduc), model(2sls)

```
==============================================================================
*** 2SLS-IV Autocorrelation Breusch-Pagan-Godfrey Test - Model= (2sls)
==============================================================================
 Ho: No Autocorrelation - Ha: Autocorrelation
------------------------------------------------------------------------------
- Rho Value for Order(1)             AR(1)=  0.0236
- Breusch-Pagan-Godfrey LM Test      AR(1)=  1.9870  P-Value >Chi2(1) 0.1587
------------------------------------------------------------------------------
```

The above table shows that there is no autocorrelation in the model. Below command can be used to detect the heteroscedasticity.

lmharch2 lwage exper expersq (educ = fatheduc motheduc huseduc), model(2sls) lags(1)

```
==============================================================================
*** 2SLS-IV Heteroscedasticity Engle (ARCH) Test - Model= (2sls)
==============================================================================
 Ho: Homoscedasticity - Ha: Heteroscedasticity
------------------------------------------------------------------------------
- Engle LM ARCH Test AR(1) E2=E2_1-E2_1=  15.7986    P-Value > Chi2(1)  0.0001
------------------------------------------------------------------------------
```

The above results indicate that there is a hint of heteroscedasticity in the model. Similarly model specification can be detected using the below command. It will not be needed in present case as there is already a squared form variable added in the regression.

reset2 lwage exper expersq (educ = fatheduc motheduc huseduc), model(2sls)

This was the illustration of 2sls single equation model.For the case where there is more than one equation in the model having endogeneity issues, such cases can be estimated using the 2sls version of the model discussed in the 3sls section.

3SLS estimates

3SLS approach is used when there is a system of equation to be estimated, where some of the equations contain endogenous variables used as explanatory variables. The variables of such kind which come as dependent in one equation and independent in another equation, become related to the disturbance terms and violating the assumptions of OLS.

$$\text{cov}\ (x_i,\varepsilon) \neq 0$$

Davidson & MacKinnon (1993) and Greene (2012) illustrated the 3 stage estimation process. The example of 3SLS is a small version of the Klein (1950) model, estimating two equations simultaneously stated below.

$\text{Consump}_t = \alpha_0 + \alpha_1\ \text{wagepriv}_t + \alpha_2\ \text{wagegovt}_t + \mu_t$
$\text{Wageprivt}_t = \beta_0 + \beta_1\ \text{consump}_t + \beta_2\ \text{govt}_t + \beta_3\ \text{capital1}_t + \varepsilon_t$
$\text{Corr}\ (\text{Consump}_t, \varepsilon_t) \neq 0, \text{Corr}\ (\text{wageprivt}_t, \mu_t) \neq 0,$

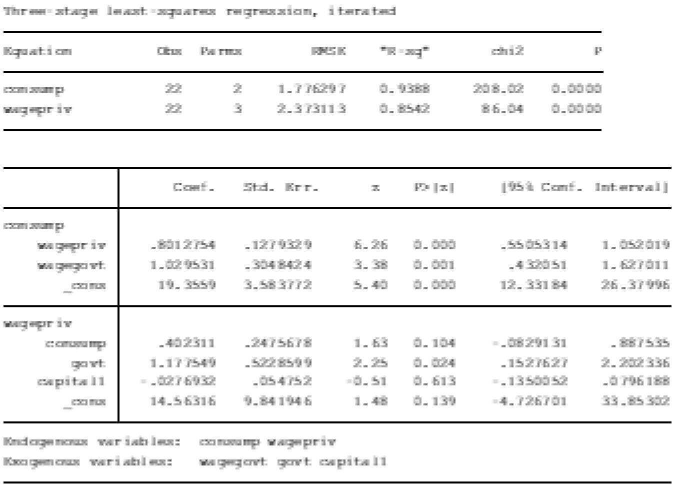
Here, comsump is total aggregate consumption, wagepriv is private wage bill, wagegovt is government wage bill, govt is total government spending, capital1 is one time period lag of capital stock. Here we can see that both dependent variables are part of independent variables in an alternative equation. This means in first equation consumption is related with μ_t with becomes an independent variable in the second equation, thus making private wage bill related to μ_t too. Similarly vice versa will make consumption to relate with ε_t. Because of this characteristic, there will be a violation of OLS assumptions. Such cases are estimated using 2SLS approach. In this approach the estimated value of the endogenous variables is estimated using all the exogenous variables and use that estimated value in the model.

Background estimates:

$\text{Consump}_t = \theta_0 + \theta_1\ \text{govt}_t + \theta_2\ \text{wagegovt}_t + \theta_3\ \text{capital1}_t + v_t$
$\text{Wageprivt}_t = \gamma_0 + \gamma_1\ \text{govt}_t + \gamma_2\ \text{wagegovt}_t + \gamma_3\ \text{capital1}_t + e_t$
3SLS estimates:
$\text{Est}\ (\text{Consump}_t) = \alpha_0 + \alpha_1\ \text{wagepriv}_t + \alpha_2\ \text{wagegovt}_t + \mu_t$
$\text{Est}\ (\text{Wageprivt}_t) = \beta_0 + \beta_1\ \text{consump}_t + \beta_2\ \text{govt}_t + \beta_3\ \text{capital1}_t + \varepsilon_t$

reg3 (consump wagepriv wagegovt) (wagepriv consump govt capital1), first ireg3

```
First-stage regressions
-----------------------

      Source |       SS           df       MS      Number of obs   =        22
-------------+----------------------------------   F(3, 18)        =      8.40
       Model |  661.220563         3  220.406854   Prob > F        =    0.0011
    Residual |  472.554316        18  26.2530175   R-squared       =    0.5832
-------------+----------------------------------   Adj R-squared   =    0.5137
       Total |  1133.77488        21  53.9892799   Root MSE        =    5.1238

------------------------------------------------------------------------------
     consump |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    wagegovt |   1.397195   .7497098     1.86   0.079    -.1778867    2.972277
        govt |   1.214701   .6203894     1.96   0.066    -.0886885    2.518091
    capital1 |    .084001   .1173772     0.72   0.483    -.1625994    .3306014
       _cons |   23.92661   22.25599     1.08   0.297    -22.83148    70.68471
------------------------------------------------------------------------------
      Source |       SS           df       MS      Number of obs   =        22
-------------+----------------------------------   F(3, 18)        =      7.83
       Model |  480.949809         3  160.316603   Prob > F        =    0.0015
    Residual |  368.542899        18  20.4746055   R-squared       =    0.5662
-------------+----------------------------------   Adj R-squared   =    0.4939
       Total |  849.492709        21  40.4520338   Root MSE        =    4.5249

------------------------------------------------------------------------------
    wagepriv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    wagegovt |   .5243061   .6620813     0.79   0.439    -.8666751    1.915287
        govt |   1.611224   .5478763     2.94   0.009     .4601782    2.762269
    capital1 |   .0422453   .1036578     0.41   0.688    -.1755317    .2600222
       _cons |   17.42221   19.65464     0.89   0.387    -23.87065    58.71507
------------------------------------------------------------------------------
```



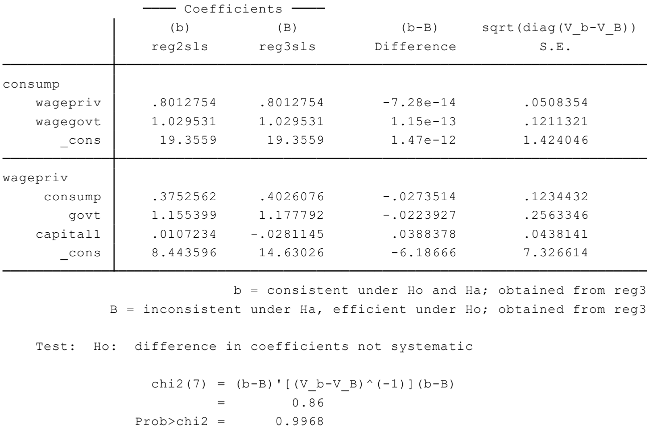
```
Three-stage least-squares regression, iterated
----------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"       chi2        P
----------------------------------------------------------------------
consump            22      2    1.776297    0.9388     208.02   0.0000
wagepriv           22      3    2.373113    0.8542      86.04   0.0000
----------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
consump      |
    wagepriv |   .8012754   .1279329     6.26   0.000     .5505314    1.052019
    wagegovt |   1.029531   .3048424     3.38   0.001      .432051    1.627011
       _cons |    19.3559   3.583772     5.40   0.000     12.33184    26.37996
-------------+----------------------------------------------------------------
wagepriv     |
     consump |    .402311   .2475678     1.63   0.104    -.0829131     .887535
        govt |   1.177549   .5228599     2.25   0.024     .1527627    2.202336
    capital1 |  -.0276932    .054752    -0.51   0.613    -.1350052    .0796188
       _cons |   14.56316   9.841946     1.48   0.139    -4.726701    33.85302
------------------------------------------------------------------------------
Endogenous variables:  consump wagepriv
Exogenous variables:   wagegovt govt capital1
------------------------------------------------------------------------------
```

The above first table shows diagnostics of two equations including, RMSE, R squared and the Chi^2 test. The second table shows the simultaneously estimated equations. In the first equation, both types of wages affect the consumption positively, whereas, private wages are only affected by government expenditures. Below command is used to store the estimates.

estimates store reg3sls

This 3sls method is different from 2sls is a way that 2sls uses the single equation method to solve the Klein model system, whereas 3sls used simultaneous equation method to solve the system. 2sls estimates are considered consistent approach to solve the endogeneity problem, whereas 3sls estimates are considered anefficient approach. So in order to decide which one is better, we use the hausman test.

reg3 (consump wagepriv wagegovt) (wagepriv consump govt capital1), first 2sls
estimates store reg2sls
hausman reg2sls reg3sls, constant alleqs



```
                    ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     reg2sls      reg3sls       Difference          S.E.
-------------+----------------------------------------------------------------
consump      |
    wagepriv |    .8012754     .8012754       -7.28e-14        .0508354
    wagegovt |    1.029531     1.029531        1.15e-13        .1211321
       _cons |     19.3559      19.3559        1.47e-12        1.424046
-------------+----------------------------------------------------------------
wagepriv     |
     consump |    .3752562     .4026076       -.0273514        .1234432
        govt |    1.155399     1.177792       -.0223927        .2563346
    capital1 |    .0107234    -.0281145        .0388378        .0438141
       _cons |    8.443596     14.63026       -6.18666         7.326614
------------------------------------------------------------------------------
                          b = consistent under Ho and Ha; obtained from reg3
           B = inconsistent under Ha, efficient under Ho; obtained from reg3

    Test:  Ho:  difference in coefficients not systematic

                  chi2(7) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        0.86
                Prob>chi2 =      0.9968
```

The Hausman test above shows the insignificant Chi^2 value, which means that the consistent estimates of 2sls and efficient estimates of 3sls are equal. Because of this equality, the 3sls estimates are now consistent and efficient, so we consider this superior to 2sls under this situation.

quietly reg3 (consump wagepriv wagegovt) (wagepriv consump govt capital1)
lmareg3

```
 *** Single Equation Autocorrelation Tests:
  Ho: No Autocorrelation in eq. #: Pij=0

  Eq. consump: Harvey LM Test =  5.4198  Rho = 0.2464  P-Value > Chi2(1) 0.0199
  Eq. wagepriv: Harvey LM Test =  5.8239 Rho = 0.2647  P-Value > Chi2(1) 0.0158
 ------------------------------------------------------------------------------
  Eq. consump: Durbin-Watson DW Test =  0.6616
  Eq. wagepriv: Durbin-Watson DW Test =  0.9698
 ------------------------------------------------------------------------------
 *** Overall System Autocorrelation Tests:
  Ho: No Overall System Autocorrelation: P11 = P22 = PMM = 0

  - Harvey  LM Test =              11.2437        P-Value > Chi2(2)   0.0036
  - Guilkey LM Test =              14.9570        P-Value > Chi2(4)   0.0048
 ------------------------------------------------------------------------------
```

Above command after installation can check for the presence of autocorrelation in an individual equation and overall model. The Harvey LM (Harvey, 1990) and Guilkey LM (Guilkey & Schmidt, 1973) test provided at the bottom of the results have significant P values confirming that there is autocorrelation in the model.

lmhreg3

```
 *** Single Equation Heteroscedasticity Tests:
   Ho: Homoscedasticity - Ha: Heteroscedasticity

  Eq. consump: Engle LM ARCH Test: E2 = E2_1 =  0.1953 P-Value > Chi2(1) 0.6585
  Eq. consump: Hall-Pagan LM Test: E2 = Yh   =  1.9987 P-Value > Chi2(1) 0.1574
  Eq. consump: Hall-Pagan LM Test: E2 = Yh2  =  1.8118 P-Value > Chi2(1) 0.1783
  Eq. consump: Hall-Pagan LM Test: E2 = LYh2 =  2.1654 P-Value > Chi2(1) 0.1411
 ----------------------------------------------------------------------------
  Eq. wagepriv: Engle LM ARCH Test: E2 = E2_1=  0.1454 P-Value > Chi2(1) 0.7029
  Eq. wagepriv: Hall-Pagan LM Test: E2 = Yh  =  2.0856 P-Value > Chi2(1) 0.1487
  Eq. wagepriv: Hall-Pagan LM Test: E2 = Yh2 =  1.9653 P-Value > Chi2(1) 0.1610
  Eq. wagepriv: Hall-Pagan LM Test: E2 = LYh2=  2.1436 P-Value > Chi2(1) 0.1432
 ----------------------------------------------------------------------------
 *** Overall System Heteroscedasticity Tests:
  Ho: No Overall System Heteroscedasticity

 - Breusch-Pagan LM Test         =   6.0869       P-Value > Chi2(1)   0.0136
 - Likelihood Ratio LR Test      =   7.1259       P-Value > Chi2(1)   0.0076
 - Wald Test                     =  11.2427       P-Value > Chi2(1)   0.0008
 ----------------------------------------------------------------------------
```

The above command after installation can check for the presence of heteroscedasticity in the individual equation and overall model. Though, there is no hint of heteroscedasticity in individual equations.But the Breusch Pegan LM, LR and Wald method show hint of heteroscedasticity in the overall model.

lmnreg3

```
*** Single Equation Non Normality Tests:
  Ho: Normality - Ha: Non Normality

 Eq. consump: Jarque-Bera LM Test  =   1.6644        P-Value > Chi2(2)  0.4351
 Eq. wagepriv: Jarque-Bera LM Test =   0.7350        P-Value > Chi2(2)  0.6924
------------------------------------------------------------------------------

*** Overall System Non Normality Tests:
 Ho: No Overall System Non Normality

*** Non Normality Tests:
- Jarque-Bera LM Test              =   1.8394        P-Value > Chi2(2)  0.3986
- Doornik-Hansen LM Test           =        .        P-Value > Chi2(2)       .
- Geary LM Test                    =  -2.7132        P-Value > Chi2(2)  0.2575
- Anderson-Darling Z Test          =  -0.9663        P-Value>Z( 0.229)  0.5906
- D'Agostino-Pearson LM Test       =   2.5894        P-Value > Chi2(2)  0.2740
------------------------------------------------------------------------------
*** Skewness Tests:
- Srivastava LM Skewness Test      =   1.7892        P-Value > Chi2(1)  0.1810
- Small LM Skewness Test           =   2.1157        P-Value > Chi2(1)  0.1458
- Skewness Z Test                  =  -1.4545        P-Value > Chi2(1)  0.1458
------------------------------------------------------------------------------
*** Kurtosis Tests:
- Srivastava Z Kurtosis Test       =   0.2240        P-Value > Z(0,1)   0.8227
- Small LM Kurtosis Test           =   0.4737        P-Value > Chi2(1)  0.4913
- Kurtosis Z Test                  =   0.6883        P-Value > Chi2(1)  0.2456
------------------------------------------------------------------------------
    Skewness Coefficient =   -0.4940     - Standard Deviation =  0.3575
    Kurtosis Coefficient =    3.1655     - Standard Deviation =  0.7017
------------------------------------------------------------------------------
    Runs Test: (14) Runs -  (24) Positives - (20) Negatives
    Standard Deviation Runs Sig(k) =  3.2501 , Mean Runs E(k) = 22.8182
    95% Conf. Interval [E(k)+/- 1.96* Sig(k)] = (16.4480 , 29.1884 )
------------------------------------------------------------------------------
```

The above command after installation can check the presence of non-normality in the overall model only. From the (Jarque & Bera, 1987) LM, Doornik-Hansen LM, (Geary, 1947) LM, (Anderson & Darling, 1954) Z test, and (D'Adnostino & Rosman, 1974) Pearson LM tests, it can be confirmed that the residuals are normal in the overall model.

GMM Approach to IV regression

Since it is already confirmed that there is endogeneity, here we can use better version of IV regression which is a GMM model as it can provide more tests to make our results efficient as GMM requires less assumptions as compared to its OLS counterparts. The advantage of GMM is that it can be applied to time series and panel data models too. GMM model was proposed by Hansen (1982).

ivreg2 lwage exper expersq (educ = fatheduc motheduc huseduc), first endog(educ)

```
First-stage regressions
-----------------------

First-stage regression of educ:

OLS estimation
--------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =      428
                                                      F(  5,    422) =    63.30
                                                      Prob > F      =   0.0000
Total (centered) SS     =  2230.196262                Centered R2   =   0.4286
Total (uncentered) SS   =        70816                Uncentered R2 =   0.9820
Residual SS             =  1274.365654                Root MSE      =    1.738

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0374977   .0343102     1.09   0.275    -.0299424    .1049379
     expersq |  -.0006002   .0010261    -0.58   0.559    -.0026171    .0014167
    fatheduc |   .1060801   .0295153     3.59   0.000     .0480648    .1640955
    motheduc |   .1141532   .0307835     3.71   0.000     .0536452    .1746613
     huseduc |   .3752548   .0296347    12.66   0.000     .3170049    .4335048
       _cons |   5.538311   .4597824    12.05   0.000     4.634562     6.44206
------------------------------------------------------------------------------
Included instruments: exper expersq fatheduc motheduc huseduc
------------------------------------------------------------------------------
F test of excluded instruments:
  F(  3,   422) =   104.29
  Prob > F      =   0.0000
Angrist-Pischke multivariate F test of excluded instruments:
  F(  3,   422) =   104.29
  Prob > F      =   0.0000
```

The above estimates are on the instrument regression. We can see that all the instruments (father education, mother education, and spouse education) are significant and another advantage that other two variables which are actually independent variables in the original regression are insignificant showing that it will not create multicollinearity later. Below is the F test for the instruments which is significant showing that the proposed instruments are good enough to split the education variable into exogenous and endogenous parts. From there we will use the estimated education variable which is exogenous part of the original regression.

```
Summary results for first-stage regressions
___________________________________________

                                          (Underid)            (Weak id)
Variable     | F(  3,   422)  P-val | AP Chi-sq(  3) P-val | AP F(  3,   422)
educ         |     104.29    0.0000 |      317.33   0.0000 |      104.29

Stock-Yogo weak ID test critical values for single endogenous regressor:
                                    5% maximal IV relative bias    13.91
                                   10% maximal IV relative bias     9.08
                                   20% maximal IV relative bias     6.46
                                   30% maximal IV relative bias     5.39
                                   10% maximal IV size             22.30
                                   15% maximal IV size             12.83
                                   20% maximal IV size              9.54
                                   25% maximal IV size              7.80
Source: Stock-Yogo (2005).  Reproduced by permission.

Underidentification test
Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)
Ha: matrix has rank=K1 (identified)
Anderson canon. corr. LM statistic       Chi-sq(3)=182.22   P-val=0.0000

Weak identification test
Ho: equation is weakly identified
Cragg-Donald Wald F statistic                                      104.29

Stock-Yogo weak ID test critical values for K1=1 and L1=3:
                                    5% maximal IV relative bias    13.91
                                   10% maximal IV relative bias     9.08
                                   20% maximal IV relative bias     6.46
                                   30% maximal IV relative bias     5.39
                                   10% maximal IV size             22.30
                                   15% maximal IV size             12.83
                                   20% maximal IV size              9.54
                                   25% maximal IV size              7.80
Source: Stock-Yogo (2005).  Reproduced by permission.

Weak-instrument-robust inference
Tests of joint significance of endogenous regressors B1 in main equation
Ho: B1=0 and orthogonality conditions are valid
Anderson-Rubin Wald test           F(3,422)=       4.48     P-val=0.0041
Anderson-Rubin Wald test           Chi-sq(3)=     13.63     P-val=0.0035
Stock-Wright LM S statistic        Chi-sq(3)=     13.21     P-val=0.0042

Number of observations               N  =        428
Number of regressors                 K  =          4
Number of endogenous regressors      K1 =          1
Number of instruments                L  =          6
Number of excluded instruments       L1 =          3
```

The above table is diagnostic tests which are there to determine the quality of the instruments used. We start with the under identification test, which is Anderson canon. Corr. LM Statistic the value is 182.22. The null and alternative hypothesis are mentioned in the test. The significance of LM test shows that the instruments are exactly identified. Which means that these instruments highly related to the endogenous variable, which is ensuring that the extracted exogenous part is good enough.

The second test is the weak identification test, it is calculated via Cragg-Donald Wald F test and its value is 104.29. The null and alternative hypothesis are provided. We can see that the calculated value is higher that all the critical values which means alternative hypothesis of instruments are not weak. This is good indication that the proposed instruments are highly correlated with the endogenous variable.

There are other weak instrument identification tests; all of these are significant showing that instruments in the reduced form equation are significant. This means these instruments are not weak; this test is robust approach to check weak identification.

These tests are also summarized in the final table, there are two more tests provided in the final table. First of them is the Sargan over identification test its value is 1.115, it is insignificant shows that the null hypothesis of instruments are not over identified is accepted. If the instruments were over identified it would mean that the proposed instruments are correlated with the error term of the main equation. The second test is the endogeneity test which is optional, GMM code must specify if this test is required. The value of endogeneity test is 2.746, which is significant at 10% showing that the alternative hypothesis of the presence of endogeneity is accepted.

The GMM regression results are interpreted the same as OLS.An increase in the education by 1%, leads to increase in the wage of the person by 0.08%, which was overestimated in OLS to 0.10%. The F test of 11.52 shows that the model is fit, all the variables are explaining the change in dependent variable.

```
IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =      428
                                                      F(  3,   424) =    11.52
                                                      Prob > F      =   0.0000
Total (centered) SS     =  223.3274418                Centered R2   =   0.1495
Total (uncentered) SS   =  829.5947848                Uncentered R2 =   0.7711
Residual SS             =  189.9347048                Root MSE      =    .6662

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0803918    .021672     3.71   0.000     .0379155    .1228681
       exper |   .0430973   .0132027     3.26   0.001     .0172204    .0689742
     expersq |  -.0008628   .0003943    -2.19   0.029    -.0016357   -.0000899
       _cons |  -.1868572   .2840591    -0.66   0.511    -.7436029    .3698884
------------------------------------------------------------------------------
Underidentification test (Anderson canon. corr. LM statistic):         182.225
                                                   Chi-sq(3) P-val =    0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):              104.294
Stock-Yogo weak ID test critical values:  5% maximal IV relative bias    13.91
                                         10% maximal IV relative bias     9.08
                                         20% maximal IV relative bias     6.46
                                         30% maximal IV relative bias     5.39
                                         10% maximal IV size             22.30
                                         15% maximal IV size             12.83
                                         20% maximal IV size              9.54
                                         25% maximal IV size              7.80
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):           1.115
                                                   Chi-sq(2) P-val =    0.5726
-endog- option:
Endogeneity test of endogenous regressors:                               2.746
                                                   Chi-sq(1) P-val =    0.0975
Regressors tested:    educ
------------------------------------------------------------------------------
Instrumented:         educ
Included instruments: exper expersq
Excluded instruments: fatheduc motheduc huseduc
------------------------------------------------------------------------------
```

ivhettest

```
IV heteroskedasticity test(s) using levels of IVs only
Ho: Disturbance is homoskedastic
    Pagan-Hall general test statistic   :  12.974  Chi-sq(5) P-value = 0.0236
```

Even though the GMM model is efficient as compared to OLS but GMM is more sensitive to heteroscedasticity. The pagan hall test statistic of 12.97 which is significant, as per the p value indicating that alternative hypothesis of heteroscedasticity is accepted. Below two are the normality and autocorrelation test on GMM estimates, both indicate that the model is normal and there is no autocorrelation.

lmnad2 lwage exper expersq (educ = fatheduc motheduc huseduc), model(gmm) hetcov(white)

```
==============================================================================
*** 2SLS-IV Non Normality Anderson-Darling Test - Model= (gmm)
==============================================================================
 Ho: Normality - Ha: Non Normality
------------------------------------------------------------------------------
- Anderson-Darling Z Test              =   5.7407     P > Z( 7.259)     1.0000
------------------------------------------------------------------------------
```

lmabpg2 lwage exper expersq (educ = fatheduc motheduc huseduc), model(gmm) hetcov(white)

```
==============================================================================
*** 2SLS-IV Autocorrelation Breusch-Pagan-Godfrey Test - Model= (gmm)
==============================================================================
 Ho: No Autocorrelation - Ha: Autocorrelation
------------------------------------------------------------------------------
- Rho Value for Order(1)               AR(1)=  0.0239
- Breusch-Pagan-Godfrey LM Test        AR(1)=  1.9874  P-Value >Chi2(1) 0.1586
------------------------------------------------------------------------------
```

Robust GMM regression

If heteroscedasticity is present, we can either solve if by changing variables or use the robust method, the robust method used alternative approach to calculate the variance covariance matrix which leads to calculation of standard errors, and this alternative approach is not sensitive to heteroscedasticity and provide consistent results.

ivreg2 lwage exper expersq (educ = fatheduc motheduc huseduc), robust gmm2s

```
2-Step GMM estimation
---------------------

Estimates efficient for arbitrary heteroskedasticity
Statistics robust to heteroskedasticity

                                                      Number of obs =      428
                                                      F(  3,   424) =     9.20
                                                      Prob > F      =   0.0000
Total (centered) SS     =  223.3274418                Centered R2   =   0.1495
Total (uncentered) SS   =  829.5947848                Uncentered R2 =   0.7710
Residual SS             =  189.9377724                Root MSE      =    .6662

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0804238   .0212634     3.78   0.000     .0387483    .1220993
       exper |   .0436998   .0151209     2.89   0.004     .0140634    .0733363
     expersq |  -.0008881   .0004154    -2.14   0.033    -.0017024   -.0000739
       _cons |  -.1861631   .2976511    -0.63   0.532    -.7695486    .3972224
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):            106.698
                                                   Chi-sq(3) P-val =    0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):              104.294
                         (Kleibergen-Paap rk Wald F statistic):        106.623
Stock-Yogo weak ID test critical values:  5% maximal IV relative bias    13.91
                                         10% maximal IV relative bias     9.08
                                         20% maximal IV relative bias     6.46
                                         30% maximal IV relative bias     5.39
                                         10% maximal IV size             22.30
                                         15% maximal IV size             12.83
                                         20% maximal IV size              9.54
                                         25% maximal IV size              7.80
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         1.042
                                                   Chi-sq(2) P-val =    0.5939
------------------------------------------------------------------------------
Instrumented:         educ
Included instruments: exper expersq
Excluded instruments: fatheduc motheduc huseduc
------------------------------------------------------------------------------
```

The above exercise using the GMM model shows that it is useful in sorting the problem of endogeneity. The last results are the robust version of GMM model, the robust version is usually used in the case where the presence of autocorrelation and heteroscedasticity is generic which cannot be sorted by changing the independent variables.

Note that in the robust GMM the over identification test is different as compared to non-robust GMM. XTIVREG2 command is also available to estimate the above instrument variable regression for panel data.

Discussion

In this chapter we have discussed a specific case of models which are used when certain independent variables are effected by dependent variables too. This two way or reverse effect caused standard regression to become inconsistent.

So what happens if there is endogeneity in the model and the dependent variable is dummy variable, in this case manual 2SLS regression estimated within logit/probit command. Similarly, these model can be used in time series where the choice of instruments increases as there are dynamic instruments available.

STATA code

```
clear
use "D:\UMT notes\Applied Econometrics\lectures\lecture 3\instrument variable regression example\mroz.dta", clear
// your above command might not work as the director might be different so it
// is better to click data editor first and then press file and open
// then use the data file with the attachment named as mroz.dta
// this data is the GMM version of the instrumental variable regression which is
// used to solve endogenous variables.
* Create the log wage variable.
gen lwage = ln(wage)
* Creating square form of experience variable
gen expersq = exper^2
* Summarize the variables used in the estimation
sum lwage exper expersq educ fatheduc motheduc huseduc
******Model*****
// this model discuses about the determinants of wage using experience and
// education. but it is observed that higher wages or shock to higher wages could
// cause education too
******* Task 1: Checking Endogeneity using IVREG *****
regress lwage exper expersq educ
estimates store olsest
** first we estimate ordinary model using Ols and store the coefficients
ivregress 2sls lwage exper expersq (educ = fatheduc motheduc huseduc)
estimates store ivest
** Then estimate the iv-reg with instruments too expected endogenous variable
hausman ivest olsest, sigmamore
******* Task 2: Confirming presence of endogeniety *****
ivregress 2sls lwage exper expersq (educ = fatheduc motheduc huseduc)
*confirms presence of endogeneity
estat endog
*confirms suitability of instruments
estat firststage
*confirms validity of extra instruments
estat overid
```

```
****** TASK 3: advanced version ************
* ssc install ivreg2
ivreg2 lwage exper expersq (educ = fatheduc motheduc huseduc), first endog(educ)
* ssc install ivhettest
ivhettest
** Interpretations to extra tests ***
** test 1: sargan overidentification test**
* its null hypothesis is : instruments are not overidentified
* here null should be accepted so that we can say that instruments are sufficient
** test 2: weak identification test **
** here null hypothesis is : instruments are not weekly identified
* here null should be accepted but it not strict requirement. It depends on high
* Correlation of instruments with the endogenous variable
** test 3: under identification test **
* here null hypothesis is : instrument is under identified
* here alternative should be accepted.
** test 4: endogeneity test **
* here null hypothesis is: there is no endogeneity
* here null should be accepted.
*** task4: robust estimation
* Now obtain the 2-step efficient GMM estimates. Use ivreg2
* with the gmm2s option:
ivreg2 lwage exper expersq (educ = fatheduc motheduc huseduc), robust gmm2s
```

Summary

Since every variable is effected by every variable, we can find situations where dependent variable also effects independent variable, which should no happen as per OLS. This chapter indicated few approaches which can be used if there is endogeneity in the model.

Application questions

1. Find Differences between OLS, ILS (indirect least squares), 2SLS, 3SLS, SURE (seemingly unrelated regressions) in terms of when we use these.
2. Construct Money Demand and Money Supply model using 3SLS.
3. Calculate the inflationary gap for any country using appropriate data.

Further Study

Acemoglu, D., Johnson, S., & Robinson, J. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *The American Economic Review, 91*(5), 1369-1401. doi. 10.1257/aer.91.5.1369

Giles, D. (2015, Feb 17). Non-Existent Instruments. [Retrieved from].

5 SURE/SEM regression for simultaneity

Learning Outcomes

- Source of simultaneity
- Why some equations are interconnected
- Foundation of panel data models

Introdution

The idea behind the SURE regression is in its name "Seemingly Unrelated Regressions" which means these individuals OLS equations seem that they are unrelated, but in reality there is contemporaneously related to each other, which in other words also mean that both models are dependent on each other. It can be of following ways;

> *Several time series equations related across the border (if the cross sections are countries). It will be more prominent if the time series equations are for the firms in the same country.*

Here an example, can be a time series model of "determinants of military expenditure of Pakistan & India". At first, they look like two different equations and two different research papers. But if we look deeply, we can see that both military expenditures are actually growing in competition with each other.

$$ME_{it} = f(determinants) + \varepsilon_{it}$$
$$ME_{pt} = f(determinants) + \varepsilon_{pt}$$

Here both residuals seem to be other, but to each other but in reality both are related, it can be confirmed by looking at their correlation. So we should use simultaneous equation model.

Several time series or panel data models relate to other models because of the same source

Here example can be like for same company determinants of ROA and ROE are interrelated, which might be because both performance indicators are related to the same cross section.

$$ROA_{it} = f(determinants) + \varepsilon_{it}$$
$$ROE_{it} = f(determinants) + \mu_{it}$$

Here both errors are seemingly unrelated, but in reality their origin is same so they must be correlated to each other and should be estimated in simultaneous equation model.

Several dependent variables of same country or firm are interrelated in terms of random shock or equilibrium

Here example can be like equations of demand of loanable funds and supply of loanable funds. Both are related to each other in equilibrium one is not complete without the other.

$$Investment_{it} = f(determinants\ of\ investment) + \upsilon_{it}$$
$$Saving_{it} = f(determinants\ of\ saving) + \omega_{it}$$

Here, though both equations are complete individually but in equilibrium both error terms should be zero as at that point saving must be equal to investment. So they must be estimated using simultaneous equation model.

SURE regression is also known as simultaneous equation model SEM or Structural equation model SEM. It is first of its kind model which connects separate equations which later builds towards advanced models. While connecting both equations it adjusts the coefficients based on the extra information, which makes these series of models more efficient as compared to individual models.

This model is called SURE and SEM in STATA, System of Equation is Eviews, and SEM is SPSS/Amos. Theoretically, In this SURE/SEM model of STATA we cannot modify the covariance structure like in Amos. The 2SLS model we studied in last lecture is also one specialized version of SURE regression, where two equations are simultaneously estimated but in that case our focus was only on the primary regression.

Detection of SURE regression

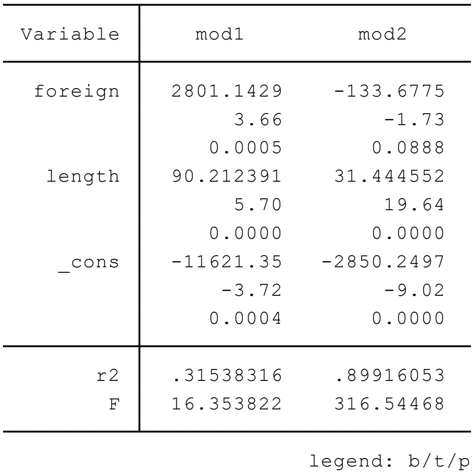
Today's exercise is about using the SURE regression and its two types, constraint and unconstrained regression models. Since we have discussed earlier with examples that sometimes two different models can be dependent or associated to each other. Apparently they might not show any issue, but it will make model inefficient as the association might affect the coefficients.

Example data set is a same auto.xls file where there the data of miles per gallon of a car is collected for 70 different cars. Illustration of the example is below. [1]

reg price foreign length

[1] Available at [Retrieved from].

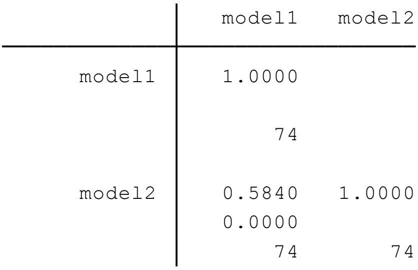
// this command estimates this price model individually
predict model1, residuals
// this command saves the residuals of above model in model1
estimates store mod1
// this command stores the regression results in memory with mod1 name
reg weight foreign length
// this command estimates the weight model individually
predict model2, residuals
// this command saves the residuals of above model in model2
estimates store mod2
// this command stores the regression results in memory with mod2 name
estimates table mod1 mod2, stats(r2 F) t p
// this command displays the results of the stored model in memory in a form which can create comparison easily



```
    Variable |     mod1          mod2
-------------+----------------------------
     foreign |  2801.1429    -133.6775
             |       3.66        -1.73
             |     0.0005       0.0888
      length |  90.212391    31.444552
             |       5.70        19.64
             |     0.0000       0.0000
       _cons |  -11621.35   -2850.2497
             |      -3.72        -9.02
             |     0.0004       0.0000
-------------+----------------------------
          r2 |  .31538316    .89916053
           F |  16.353822    316.54468
------------------------------------------
                             legend: b/t/p
```

Above are the coefficients should definitely be different as the dependent variable are different. But since the source is same as both are variables of same car, they must be related to each other, for that correlation of both residuals must be checked. Now we check if their residuals which were predicted are correlated or not

pwcorr model1 model2, obs sig
// this command is used to calculate the correlation between the specified variables provided



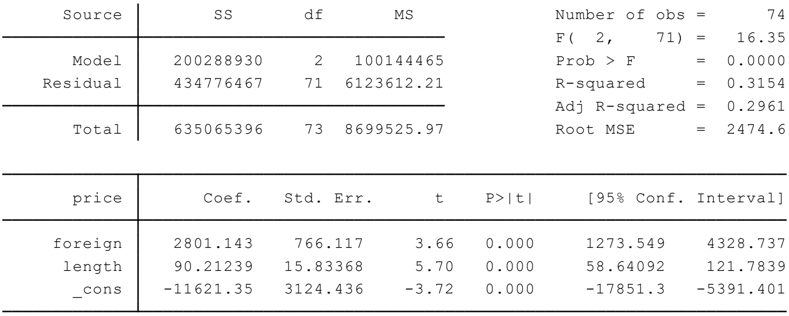
```
             |   model1   model2
-------------+------------------
      model1 |   1.0000 
             |
             |       74
             |
      model2 |   0.5840   1.0000 
             |   0.0000
             |       74       74
```

Here you can see that the correlation among the residuals is 0.5840 means that deviations of both residuals are 58% similar to each other. And see the p value below it shows that it is also significant as it is smaller than 0.05 so we can say that both models are dependent to each other.

Other post-OLS diagnostic tests

In STATA, there are some post regression diagnostic tests which can be used to identify issues in the model.

reg price foreign length
// this command estimates this price model individually



```
      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(2, 71)        =     16.35
       Model |   200288930         2   100144465   Prob > F        =    0.0000
    Residual |   434776467        71  6123612.21   R-squared       =    0.3154
-------------+----------------------------------   Adj R-squared   =    0.2961
       Total |   635065396        73  8699525.97   Root MSE        =    2474.6

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   2801.143    766.117     3.66   0.000     1273.549    4328.737
      length |   90.21239   15.83368     5.70   0.000     58.64092    121.7839
       _cons |  -11621.35   3124.436    -3.72   0.000     -17851.3   -5391.401
------------------------------------------------------------------------------
```

estat hettest

// this checks heteroskedasticity with no heteroskedasticity as null hypothesis

```
 Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
          Ho: Constant variance
          Variables: fitted values of price

          chi2(1)      =     7.54
          Prob > chi2  =   0.0060
```

The above test shows that there is significant evidence of heteroskedasticity in the model. Since it is a cross sectional data, so it is checking cross sectional heteroskedasticity.

estat ovtest

// this test checks the Ramsey RESET test of mis specification with no misspecificationspecification as null hypothesis

```
Ramsey RESET test using powers of the fitted values of price
       Ho:  model has no omitted variables
                  F(3, 68) =      2.71
                  Prob > F =      0.0515
```

The above test shows that at 5% level, there is not misspecification in the model, the proposed linear variables are sufficient.

estat vif

// this test calculates the VIF value and tells if there is multicollinearity

```
    Variable |       VIF       1/VIF
-------------+----------------------
     foreign |      1.48    0.674875
      length |      1.48    0.674875
-------------+----------------------
    Mean VIF |      1.48
```

The above table provides the VIF statistics for multicollinearity, since both of them are smaller than 10 so there is no hint of multicollinearity.

test foreign = 1500
// this command applies coefficient restriction test on the foreign variable coefficient, here null hypothesis is that the foreign variable can be equal to 1500 and alternative is that foreign variable coefficient cannot be equal to 1500

```
 ( 1)  foreign = 1500

       F(  1,    71) =    2.88
            Prob > F =    0.0938
```

Above coefficient restriction test is insignificant at 5% means that the coefficient of the foreign variable can be equal to 1500.

SURE regression

Since there is significant evidence that both regressions estimated earlier are interrelated through correlation of residuals. If this interrelation is ignored, it will lead to the exclusion of information from the model, making it inefficient. If it is the case where both interrelated equations are connected like one is demand equation and the other is supply, then ignoring the interrelation might lead to contemporaneous correlation. Below is the STATA code for SURE regression.

sureg (price foreign length) (weight foreign length), small dfk
// this command runs two regressions in two bracket sets simultaneously

```
Seemingly unrelated regression
--------------------------------------------------------------------------
Equation             Obs  Parms        RMSE    "R-sq"       F-Stat        P
--------------------------------------------------------------------------
price                 74      2    2474.593    0.3154        16.35   0.0000
weight                74      2    250.2515    0.8992       316.54   0.0000
--------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
price        |
     foreign |   2801.143    766.117     3.66   0.000     1286.674    4315.611
      length |   90.21239   15.83368     5.70   0.000     58.91219    121.5126
       _cons |  -11621.35   3124.436    -3.72   0.000    -17797.77    -5444.93
-------------+----------------------------------------------------------------
weight       |
     foreign |  -133.6775   77.47615    -1.73   0.087    -286.8332     19.4782
      length |   31.44455   1.601234    19.64   0.000     28.27921    34.60989
       _cons |   -2850.25   315.9691    -9.02   0.000    -3474.861   -2225.639
------------------------------------------------------------------------------
```

This above is a simultaneous equation model, you can see that the estimates of both models are provided together. Note that in this case the coefficients are not different may be because both equations are not very dependent on each other in terms of dependent variable, but note that the lower confidence interval is bigger (*now 1286.674 rather than 1273.549*) and the upper confidence interval is smaller (*now 4315.611 rather than 4328.737*) this means that now the difference between the lower and upper confidence interval is smaller so model as smaller standard deviation which makes it better for forecasting and models becomes efficient.

Here we can check the equation dependency directly using the command below *sureg, notable noheader corr* [2]

```
Correlation matrix of residuals:

          price  weight
  price  1.0000
 weight  0.5840  1.0000

Breusch-Pagan test of independence: chi2(1) =    25.237, Pr = 0.0000
```

Here the Breusch & Pegan (1980) test shows that p value is smaller than 0.05 so we can say that both equations are related to each other. So we should use this model instead of separate OLS.

The normality, heteroskedasticity and autocorrelation in the SURE regression method, can be checked using the commands below. The results below show that, there is autocorrelation, heteroskedasticity in the overall system.

lmareg3

```
==================================================
* System Autocorrelation Tests (sure)
==================================================
*** Single Equation Autocorrelation Tests:
 Ho: No Autocorrelation in eq. #: Pij=0

 Eq. price : Harvey LM Test = 11.6194   Rho = 0.1570  P-Value > Chi2(1) 0.0007
 Eq. weight: Harvey LM Test =  3.9363   Rho = 0.0532  P-Value > Chi2(1) 0.0473
------------------------------------------------------------------------------
 Eq. price : Durbin-Watson DW Test =  1.1994
 Eq. weight: Durbin-Watson DW Test =  1.5368
------------------------------------------------------------------------------
*** Overall System Autocorrelation Tests:
 Ho: No Overall System Autocorrelation: P11 = P22 = PMM = 0

 - Harvey  LM Test =              15.5557        P-Value > Chi2(2)    0.0004
 - Guilkey LM Test =              11.8426        P-Value > Chi2(4)    0.0186
------------------------------------------------------------------------------
```

[2] Here notable command hides the regression results and corr shows the dependency test

lmhreg3

```
==================================================
* System Heteroscedasticity Tests (sure)
==================================================
*** Single Equation Heteroscedasticity Tests:
  Ho: Homoscedasticity - Ha: Heteroscedasticity

 Eq. price : Engle LM ARCH Test: E2 = E2_1  = 10.0632 P-Value > Chi2(1) 0.0015
 Eq. price : Hall-Pagan LM Test: E2 = Yh    =  3.1259 P-Value > Chi2(1) 0.0771
 Eq. price : Hall-Pagan LM Test: E2 = Yh2   =  3.0376 P-Value > Chi2(1) 0.0814
 Eq. price : Hall-Pagan LM Test: E2 = LYh2  =  2.6720 P-Value > Chi2(1) 0.1021
------------------------------------------------------------------------------
 Eq. weight: Engle LM ARCH Test: E2 = E2_1  =  0.4143 P-Value > Chi2(1) 0.5198
 Eq. weight: Hall-Pagan LM Test: E2 = Yh    =  0.1335 P-Value > Chi2(1) 0.7148
 Eq. weight: Hall-Pagan LM Test: E2 = Yh2   =  0.1053 P-Value > Chi2(1) 0.7455
 Eq. weight: Hall-Pagan LM Test: E2 = LYh2  =  0.1716 P-Value > Chi2(1) 0.6787
------------------------------------------------------------------------------
*** Overall System Heteroscedasticity Tests:
 Ho: No Overall System Heteroscedasticity

- Breusch-Pagan LM Test         =  25.2370       P-Value > Chi2(1)   0.0000
- Likelihood Ratio LR Test      =  30.8649       P-Value > Chi2(1)   0.0000
- Wald Test                     =  73.9994       P-Value > Chi2(1)   0.0000
------------------------------------------------------------------------------
```

While comparing the normality test, 2 out of 5 tests indicate that residuals are not normal while remaining indicate that residuals are normal.

lmnreg3

```
==================================================
* System Non Normality Tests (sure)
==================================================
*** Single Equation Non Normality Tests:
  Ho: Normality - Ha: Non Normality

 Eq. price : Jarque-Bera LM Test   =  56.1031        P-Value > Chi2(2)  0.0000
 Eq. weight: Jarque-Bera LM Test   =  22.7391        P-Value > Chi2(2)  0.0000
------------------------------------------------------------------------------

*** Overall System Non Normality Tests:
 Ho: No Overall System Non Normality

*** Non Normality Tests:
- Jarque-Bera LM Test              = 559.4237        P-Value > Chi2(2)  0.0000
- Doornik-Hansen LM Test           = 1.35e+07        P-Value > Chi2(2)  0.0000
- Geary LM Test                    =  -2.2140        P-Value > Chi2(2)  0.3306
- Anderson-Darling Z Test          =   2.2182        P-Value>Z( 9.298)  1.0000
- D'Agostino-Pearson LM Test       =  91.8915        P-Value > Chi2(2)  0.0000
------------------------------------------------------------------------------
*** Skewness Tests:
- Srivastava LM Skewness Test      = 122.2459        P-Value > Chi2(1)  0.0000
- Small LM Skewness Test           =  57.2808        P-Value > Chi2(1)  0.0000
- Skewness Z Test                  =   7.5684        P-Value > Chi2(1)  0.0000
------------------------------------------------------------------------------
*** Kurtosis Tests:
- Srivastava Z Kurtosis Test       =  20.9088        P-Value > Z(0,1)   0.0000
- Small LM Kurtosis Test           =  34.6107        P-Value > Chi2(1)  0.0000
- Kurtosis Z Test                  =   5.8831        P-Value > Chi2(1)  0.0000
------------------------------------------------------------------------------
    Skewness Coefficient =    2.2262     - Standard Deviation =  0.1993
    Kurtosis Coefficient =   11.4198     - Standard Deviation =  0.3961
------------------------------------------------------------------------------
    Runs Test: (58) Runs -  (56) Positives - (92) Negatives
    Standard Deviation Runs Sig(k) =  5.7009 , Mean Runs E(k) = 70.6216
    95% Conf. Interval [E(k)+/- 1.96* Sig(k)] = (59.4478 , 81.7954 )
------------------------------------------------------------------------------
```

This above sure regression we estimated is the unconstraint model where coefficients can be any value may be same or not. In the SURE regression results, the coefficient restrictions can be applied using following command.

test foreign

```
 ( 1)  [price]foreign = 0
 ( 2)  [weight]foreign = 0

       F(  2,   142) =   17.99
            Prob > F =    0.0000
```

This F test is restriction test on the foreign variable on all simultaneous equations estimated. It is significant showing that this variable cannot be zero in both models at same time. Similarly coefficients can be test in a single equation too rather than all, it can be done in same way as done in MNL model in chapter 2.

Constraint SURE regression

In order to estimate constraint SURE regression, for that there is need of specification of the restriction. The constraint mentioned below will make sure that the coefficient of foreign variable remains same in all equations.

constraint 1 [price]foreign = [weight]foreign
sureg (price foreign length) (weight foreign length), small dfk const (1)

```
. constraint 1 [price]foreign = [weight]foreign

. sureg (price foreign length) (weight foreign length), small dfk const (1)

Seemingly unrelated regression
------------------------------------------------------------------------------
Equation             Obs   Parms        RMSE    "R-sq"      F-Stat        P
------------------------------------------------------------------------------
price                 74       2    2695.704    0.1876        9.20   0.0002
weight                74       2    256.4002    0.8941       52.06   0.0000
------------------------------------------------------------------------------

 ( 1)  [price]foreign - [weight]foreign = 0
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
price        |
     foreign |   11.92344   186.3955     0.06   0.949    -356.5451     380.392
      length |   57.34275   13.51897     4.24   0.000     30.61831    84.06719
       _cons |  -4614.851   2567.248    -1.80   0.074    -9689.815    460.1139
-------------+----------------------------------------------------------------
weight       |
     foreign |   11.92344   186.3955     0.06   0.949    -356.5451     380.392
      length |   33.16039   3.911258     8.48   0.000     25.42857    40.89221
       _cons |  -3215.998   770.8738    -4.17   0.000     -4739.87   -1692.126
------------------------------------------------------------------------------
```

In the results above the coefficient of foreign variable is same, but it is insignificant, this is actually because of the fact that in this example this restriction is not realistic, it is just illustration. Following commands can be used to generate the estimated dependent variables from which the residuals can be generated.

```
predict phat, equation(price)
predict what, equation(weight)
```

Discussions

This SURE regression is a first in the range of simultaneous regression which are ideal for the case where the dynamics of the variables are interrelated. For the financial studies people are using SEM in SPSS/AMOS. The advanced versions of the SEM model which is GSEM (generalized SEM) also support for nonlinear models for the case when the dependent variable is qualitative.

There will be some cases when equations are connected but they are following a path, for this 'pathreg' command can be used. Its tutorial is available at [Retrieved from].

The evolution of the panel data model came from the interrelation of regressions which will be discussed in chapter 10. While vector models like VECM and VAR, is the time series version of SURE regressions it will be discussed in chapter 8.

STATA code

```
    use "D:\UMT notes\Applied Econometrics\lectures\lecture 4\auto.dta", clear
    // you can use the already stored data using above method
    // or paste the data in the STATA just like we do in eviews
    **** Task 1 - checking presence of contemporaneous correlation ****
    reg price foreign length
        estat hettest
        // this is heteroskedasticity test
        // null hypothesis is no heteroskedasticity
        estat ovtest
        // this is Ramsey reset test of miss specification test
        // null hypothesis is no mis specification
        estat vif
        // this multicollinearity test
        // if vif less than 10 then no multicollinearity test
        predict model1, residuals
        // above command is used to generate a new variable which is actually
        // the residuals of the above regression and stored with name of model1
        estimates store mod1
    reg weight foreign length
        estat hettest
        estat ovtest
        estat vif
        predict model2, residuals
        estimates store mod2
    estimates table mod1 mod2, stats(r2 F) t p
    // order to test the presence of correlation we need to check for correlation
    // between two residuals
    // following command is using two variables and checking their correlation
    // here if p value is smaller than 0.05 shows that correlation is significant
    pwcorr model1 model2, obs sig
    *** task 2 - applying coefficient restriction test on the variables
    // generally the T value shows the coefficient significance test while
    // comparing it against 0
    // what if we have to check it against any other value
    reg weight foreign length
```

```
        test foreign = -100
    // the above test command is checking the command against reference of -100
    // here null hypothesis is that slope of foreign is equal -100
    // here alternative hypothesis is that slope of foreign is not equal to -100
    // if p value is less than 0.05 then alternative is selected
    ***  task3 - simultaneous equation model
    sureg (price foreign length) (weight foreign length), small dfk
    estimates store surere
    // we can directly check presence of equation dependency or contemporaneous correlation
    // below command shows the Breusch pagan test where
    // null hypothesis is that both equations are independent so OLS can be used separately
    // alternative hypothesis is that both equations are dependent so sure can be used
    // if p is smaller than 0.05 so we accept alternative hypothesis
    sureg, notable noheader corr
    estimates table mod1 mod2 surere, stats(r2 F) t p
    *** task- 4 restricted and un restricted models
    sureg (price foreign length) (weight foreign length), small dfk
    // when we run this regression we are allowing coefficients to have any value
    // making it unrestricted
    // but sometimes we need to see that a particular variable have same slope in
    // both equations
    // we apply restriction here
    sureg (price foreign length) (weight foreign length), small dfk
    test foreign
    // p value shows that both cannot be same
    // but if p value was larger than 0.05 when we could have said they can be same
    // so we make the new model in which it is same
    // it is made like this
    constraint 1 [price]foreign = [weight]foreign
    sureg (price foreign length) (weight foreign length), small dfk const (1)
    predict phat, equation(price)
    predict what, equation(weight)
```

Summary

Most of the models work in tandem like model of demand and model of supply of same product. This chapter highlighted the concept of simultaneity where in some cases models must be estimated simultaneously so that their influence on each other is incorporated.

Applied questions

1. What does Unconstrained SURE assume that Unconstrained OLS does not?
2. What does Constrained SURE assume that Unconstrained SURE does not?
3. What does Pooled OLS assume that Constrained SURE does not?
4. What does Fixed Effect assume that Constrained SURE does not?

Further study

Giles, D. (2012, Apr 14). Simultaneous equations models. [Retrieved from].

Giles, D. (2012, May 19). Estimating & simulating an SEM. [Retrieved from].

6 Modelling in the presence of multicollinearity

Learning Outcomes

- Why multicollinearity is inevitable
- Different degrees of multicollinearity

Introduction

O you who have believed, do not approach prayer while you are intoxicated until you know what you are saying… Al Qur'an (4:43).

Multicollinearity is a severe issue because it is hitting the most useful quality of any estimator, including mean and standard deviation, that it affects the property of unbiasness. We usually assume in inferential statistics that most of the estimators are truly representative of the population. There are three options which can be used under such conditions.

Detecting Multicollinearity

Before we learn the appropriate model we have to study all the tests which can be used to check the presence of multicollinearity in the model. This example from the auto.xls file can be useful in understanding and interpretation of the available tests. [1]

```
generate weight2 = weight^2
regress mpg weight weight2 displ gear turn headroom foreign price
ssc install lmcol [2]
lmcol mpg weight weight2 displ gear turn headroom foreign price
```

This command first estimates the OLS regression for the variables provided in the LMCOL test, we can note that many important variables are insignificant, but the R-squared and F test says otherwise. This command shows R square for 6 different methods all of them are high showing model must be good. This indicates that there is multicollinearity in the data.

The first table, in the multicollinearity diagnostic test shows the correlations among the independent variables.If the correlation is very high this is a hint for presence of multicollinearity but it does not confirm it. The second table of multicollinearity diagnostic criteria shows Eigen values, for those variables whose Eigen values are near to zero shows that there is multicollinearity in the model because of that variable. The other criteria available here is the Condition index (*C_Index*) which will show high multicollinearity for the variable whose C index value is larger than 15.

[1] Available at [Retrieved from].

[2] It need to be installed for the first time in the computer later it can be used without installing

```
==============================================================================
* Ordinary Least Squares (OLS)
==============================================================================
  mpg = weight + weight2 + displacement + gear_ratio + turn + headroom + foreign + price
------------------------------------------------------------------------------
  Sample Size       =          74
  Wald Test         =    164.2308   |   P-Value > Chi2(8)       =     0.0000
  F-Test            =     20.5288   |   P-Value > F(8 , 65)     =     0.0000
 (Buse 1973) R2     =      0.7164   |   Raw Moments R2          =     0.9808
 (Buse 1973) R2 Adj =      0.6815   |   Raw Moments R2 Adj      =     0.9784
  Root MSE (Sigma)  =      3.2649   |   Log Likelihood Function =  -187.7617
------------------------------------------------------------------------------
- R2h= 0.7164   R2h Adj= 0.6815  F-Test =   20.53 P-Value > F(8 , 65)  0.0000
- R2v= 0.7164   R2v Adj= 0.6815  F-Test =   20.53 P-Value > F(8 , 65)  0.0000
------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0167516   .0041614    -4.03   0.000    -.0250624   -.0084407
     weight2 |   2.15e-06   7.39e-07     2.91   0.005     6.78e-07    3.63e-06
displacement |  -.0053586   .0122473    -0.44   0.663    -.0298181    .0191009
  gear_ratio |   1.507051   1.701996     0.89   0.379     -1.89207    4.906172
        turn |  -.3659504   .1863557    -1.96   0.054    -.7381285    .0062276
    headroom |  -.1998089   .5381564    -0.37   0.712    -1.274581    .8749637
     foreign |  -3.010428   1.430563    -2.10   0.039     -5.86746   -.1533954
       price |  -.0002154    .000209    -1.03   0.307    -.0006328    .0002021
       _cons |   64.81275   11.25313     5.76   0.000     42.33869    87.28681
------------------------------------------------------------------------------
```

```
==============================================================================
*** Multicollinearity Diagnostic Tests
==============================================================================

* Correlation Matrix
(obs=74)

             |   weight  weight2 displa~t gear_r~o     turn headroom  foreign    price
-------------+------------------------------------------------------------------------
      weight |   1.0000
     weight2 |   0.9915   1.0000
displacement |   0.8949   0.9077   1.0000
  gear_ratio |  -0.7593  -0.7501  -0.8289   1.0000
        turn |   0.8574   0.8527   0.7768  -0.6763   1.0000
    headroom |   0.4835   0.4730   0.4745  -0.3779   0.4245   1.0000
     foreign |  -0.5928  -0.5663  -0.6138   0.7067  -0.6311  -0.2939   1.0000
       price |   0.5386   0.5760   0.4949  -0.3137   0.3096   0.1145   0.0487   1.0000

* Multicollinearity Diagnostic Criteria
+-------------------------------------------------------------------------------+
|   Var |  Eigenval |  C_Number |   C_Index |       VIF |     1/VIF |   R2_xi,X |
|-------+-----------+-----------+-----------+-----------+-----------+-----------|
| wei~t |    5.2981 |    1.0000 |    1.0000 |   71.6350 |    0.0140 |    0.9860 |
| wei~2 |    1.1329 |    4.6764 |    2.1625 |   86.9197 |    0.0115 |    0.9885 |
| dis~t |    0.7668 |    6.9091 |    2.6285 |    8.6638 |    0.1154 |    0.8846 |
| gea~o |    0.3679 |   14.4008 |    3.7948 |    4.1303 |    0.2421 |    0.7579 |
|  turn |    0.2044 |   25.9219 |    5.0914 |    4.6031 |    0.2172 |    0.7828 |
| hea~m |    0.1317 |   40.2412 |    6.3436 |    1.4195 |    0.7045 |    0.2955 |
| for~n |    0.0917 |   57.7627 |    7.6002 |    2.9681 |    0.3369 |    0.6631 |
| price |    0.0065 |  818.3345 |   28.6065 |    2.6029 |    0.3842 |    0.6158 |
+-------------------------------------------------------------------------------+
```

Here in this table below the tests are developed by Farrar Glauber Multicollinearity test. The first test is F-G Chi^2 Test, which is significant which means that alternative hypothesis is selected and which is mentioned that there is multicollinearity in the model. For the second table the interpretations are taken from Mitsaki (2011)[3] here if the F test is significant then this shows that this variable is contributing in multicollinearity. The third t-test table shows the pairwise t test values. For those tests whose t values are more than 2 or less than 2 then these pairs are causing multicollinearity in the model.

```
* Farrar-Glauber Multicollinearity Tests
  Ho: No Multicollinearity - Ha: Multicollinearity
--------------------------------------------------

* (1) Farrar-Glauber Multicollinearity Chi2-Test:
    Chi2 Test =  730.9648    P-Value > Chi2(28) 0.0000

* (2) Farrar-Glauber Multicollinearity F-Test:
+------------+----------+----------+----------+----------+
|   Variable |   F_Test |      DF1 |      DF2 |  P_Value |
+------------+----------+----------+----------+----------+
|     weight |  665.988 |   66.000 |    8.000 |    0.000 |
|    weight2 |  810.099 |   66.000 |    8.000 |    0.000 |
|  displace~t|   72.259 |   66.000 |    8.000 |    0.000 |
|  gear_ratio|   29.514 |   66.000 |    8.000 |    0.000 |
|       turn |   33.972 |   66.000 |    8.000 |    0.000 |
|   headroom |    3.956 |   66.000 |    8.000 |    0.022 |
|    foreign |   18.556 |   66.000 |    8.000 |    0.000 |
|      price |   15.113 |   66.000 |    8.000 |    0.000 |
+------------+----------+----------+----------+----------+

* (3) Farrar-Glauber Multicollinearity t-Test:
+----------+--------+--------+--------+--------+--------+--------+--------+--------+
| Variable | weight | weig~2 | disp~t | gear~o |   turn | head~m | fore~n |  price |
+----------+--------+--------+--------+--------+--------+--------+--------+--------+
|   weight |      . |        |        |        |        |        |        |        |
|  weight2 | 61.845 |      . |        |        |        |        |        |        |
| displa~t | 16.291 | 17.578 |      . |        |        |        |        |        |
| gear_r~o | -9.478 | -9.215 |-12.037 |      . |        |        |        |        |
|     turn | 13.537 | 13.260 | 10.020 | -7.459 |      . |        |        |        |
| headroom |  4.487 |  4.362 |  4.379 | -3.315 |  3.808 |      . |        |        |
|  foreign | -5.980 | -5.581 | -6.316 |  8.115 | -6.610 | -2.498 |      . |        |
|    price |  5.193 |  5.725 |  4.627 | -2.684 |  2.645 |  0.936 |  0.396 |      . |
+----------+--------+--------+--------+--------+--------+--------+--------+--------+
```

[3] [Retrieved from].

```
* |X'X| Determinant:
  |X'X| = 0 Multicollinearity - |X'X| = 1 No Multicollinearity
  |X'X| Determinant:       (0 < 0.0000 < 1)
---------------------------------------------------------------

* Theil R2 Multicollinearity Effect:
  R2 = 0 No Multicollinearity - R2 = 1 Multicollinearity
     - Theil R2:           (0 < 0.5631 < 1)
---------------------------------------------------------------

* Multicollinearity Range:
  Q = 0 No Multicollinearity - Q = 1 Multicollinearity
     - Gleason-Staelin Q0: (0 < 0.6288 < 1)
    1- Heo Range Q1:       (0 < 0.9878 < 1)
    2- Heo Range Q2:       (0 < 0.9563 < 1)
    3- Heo Range Q3:       (0 < 0.9948 < 1)
    4- Heo Range Q4:       (0 < 0.5389 < 1)
    5- Heo Range Q5:       (0 < 0.9960 < 1)
    6- Heo Range Q6:       (0 < 0.7468 < 1)
------------------------------------------------------------------
```

The above tests are further test which checks the multicollinearity in the model, first of this is the determinant of X'X matrix[4][5]. Here the answer is zero, which means there is multicollinearity. The second test is Theil R square test. Which should be zero for no multicollinearity. Here it is in between 0 and 1 so there are more chances of presence of multicollinearity. The third test shows a range of tests which show now multicollinearity if the answer is 0 but here all of them are near to 1 so there is multicollinearity.

graph matrix weight weight2 displ gear turn headroom price, title(Matrix Plot) subtitle(Independent Variables) scheme(sj)

This graph will make a scatter plot matrix among all the provided independent variables. It sees if there is a pattern

[4] This is a matrix of all the independent variables multiplied by its transpose and then determinant is calculated. If there is multicollinearity in the model then this determinant is greatly affected, ideally it should be 1 for no multicollinearity.

[5] See endnote for the details about the X'X matrix

among them, those pairs which show a pattern are likely to create multicollinearity.

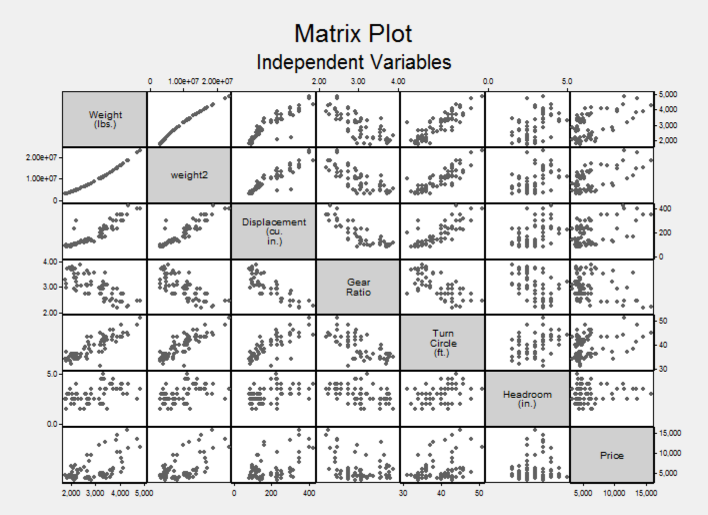


Figure 15. *Scatter matrix plot*

Stepwise regression

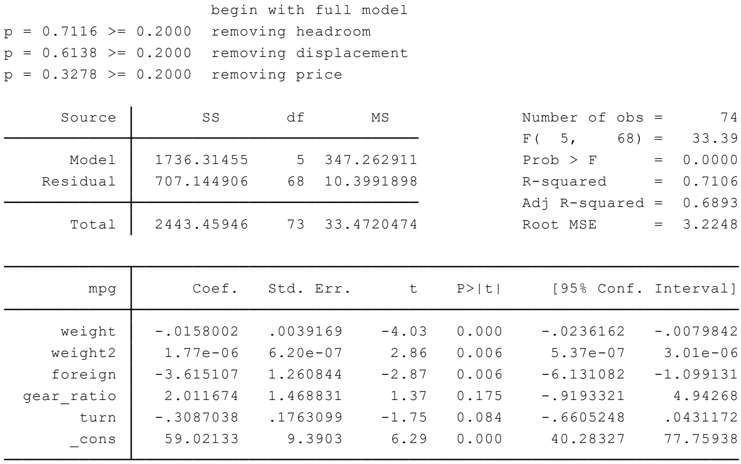
Stepwise regression is used in the case when we have too many independent variables, using all of them might cause multicollinearity. Stepwise regression will use the statistical approach to find the appropriate set of independent variable which are significant and not collinear.

Generally for this method we use step wise regression is SPSS. The advantage of this method is that, it can iteratively select variables which upon adding to the model does not show a rise in the indicators of Multicollinearity which are available in SPSS. The disadvantage of this method that it might produce a model which is totally non beneficial to the literature, it will not care which variable or set of variable must be present, hence it is up to the discretion of the researcher if he can able to explain the model generated from this approach.

This problem is solved in the STATA, where the important variables are fixed while pairs can be made. This approach can be estimated in STATA and the example is below using the auto.dta file provided [6]

First type of stepwise regression is backward selection methods which will remove the variables one by one which are problematic and use the (0.2) criteria which mean that all those variables whose P values are larger than 0.2 means they are insignificant.

stepwise, pr(.2): regress mpg weight weight2 displ gear turn headroom foreign price



```
                      begin with full model
p = 0.7116 >= 0.2000  removing headroom
p = 0.6138 >= 0.2000  removing displacement
p = 0.3278 >= 0.2000  removing price

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  5,    68) =   33.39
       Model |  1736.31455     5  347.262911           Prob > F      =  0.0000
    Residual |  707.144906    68  10.3991898           R-squared     =  0.7106
-------------+------------------------------           Adj R-squared =  0.6893
       Total |  2443.45946    73  33.4720474           Root MSE      =  3.2248

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0158002   .0039169    -4.03   0.000    -.0236162   -.0079842
     weight2 |   1.77e-06   6.20e-07     2.86   0.006     5.37e-07    3.01e-06
     foreign |  -3.615107   1.260844    -2.87   0.006    -6.131082   -1.099131
  gear_ratio |   2.011674   1.468831     1.37   0.175    -.9193321     4.94268
        turn |  -.3087038   .1763099    -1.75   0.084    -.6605248    .0431172
       _cons |   59.02133     9.3903     6.29   0.000     40.28327    77.75938
------------------------------------------------------------------------------
```

This shows that three variables are removed, which are insignificant. Now we check for the presence of the multicollinearity again by using the LMCOL command below. We have checked it in STATA, but shows there is still multicollinearity this means that the multicollinearity was complex which was rectified using this approach.

lmcol mpg weight weight2 gear turn foreign

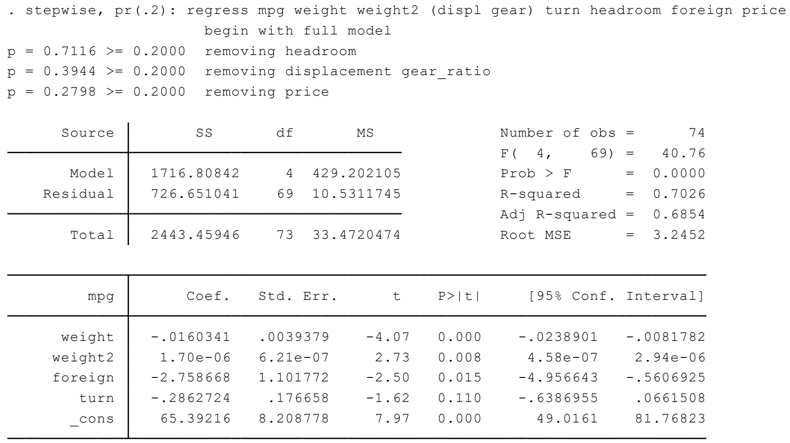
[6] Available at [Retrieved from].

This approach will only be able to remove the weak form of multicollinearity in which the problematic variables are insignificant, so by removing them leads to unbiased (free of multicollinearity) results.

Other versions of stepwise regression

This command can be used by adding brackets to pair variables which both must be present or both must be removed if necessary. This might help in cases when two variables are part of same variable like two dummies which are made from one trichotomus dummy. It is mostly defined by the theory

stepwise, pr(.2): regress mpg weight weight2 (displ gear) turn headroom foreign price



```
. stepwise, pr(.2): regress mpg weight weight2 (displ gear) turn headroom foreign price
                      begin with full model
p = 0.7116 >= 0.2000  removing headroom
p = 0.3944 >= 0.2000  removing displacement gear_ratio
p = 0.2798 >= 0.2000  removing price

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(4, 69)        =     40.76
       Model |  1716.80842         4  429.202105   Prob > F        =    0.0000
    Residual |  726.651041        69  10.5311745   R-squared       =    0.7026
-------------+----------------------------------   Adj R-squared   =    0.6854
       Total |  2443.45946        73  33.4720474   Root MSE        =    3.2452

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0160341   .0039379    -4.07   0.000    -.0238901   -.0081782
     weight2 |   1.70e-06   6.21e-07     2.73   0.008     4.58e-07    2.94e-06
     foreign |  -2.758668   1.101772    -2.50   0.015    -4.956643   -.5606925
        turn |  -.2862724    .176658    -1.62   0.110    -.6386955    .0661508
       _cons |   65.39216   8.208778     7.97   0.000      49.0161    81.76823
------------------------------------------------------------------------------
```

Another approach is that we can lock any specific variable which is important, according to a theory which must not be removed at any case. Certainly you cannot remove the variable for which you thesis is based on.

stepwise, pr(.2) lockterm1: regress mpg weight weight2 displ gear turn headroom foreignprice

```
                      begin with full model
p = 0.7116 >= 0.2000  removing headroom
p = 0.6138 >= 0.2000  removing displacement
p = 0.3278 >= 0.2000  removing price

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  5,    68) =   33.39
       Model |  1736.31455     5  347.262911           Prob > F      =  0.0000
    Residual |  707.144906    68  10.3991898           R-squared     =  0.7106
-------------+------------------------------           Adj R-squared =  0.6893
       Total |  2443.45946    73  33.4720474           Root MSE      =  3.2248

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0158002   .0039169    -4.03   0.000    -.0236162   -.0079842
     weight2 |   1.77e-06   6.20e-07     2.86   0.006     5.37e-07    3.01e-06
     foreign |  -3.615107   1.260844    -2.87   0.006    -6.131082   -1.099131
  gear_ratio |   2.011674   1.468831     1.37   0.175    -.9193321     4.94268
        turn |  -.3087038   .1763099    -1.75   0.084    -.6605248    .0431172
       _cons |   59.02133     9.3903     6.29   0.000     40.28327    77.75938
------------------------------------------------------------------------------
```

The advantage of this approach is that we can apply this to other regression models like logit, probit, and tobit. Like,

stepwise, pr(.2): logit foreign mpg weight weight2 displ gear turn headroom price

Issues with stepwise regression

Even though this method is convenient in reducing the model to relevant (significant) variables. Since it does not take into account the theoretical importance of the variables in the model, this approach may lead to following issues.

- They do not necessarily produce best model (Judd *et al.*, 2011)
- R squared might be biased and overestimated
- Confidence intervals might be falsely narrow for over efficient model (Altman & Andersen, 1989)
- It diverts the focus from the problem at hand

- Data analyst knows more than the computer, ignoring it will produce inadequate data analysis (Henderson & Vellerman, 1981).

Exploratory analysis

The method of the principle factor analysis uses the correlation among the variables generate extra information. There are two approaches in factor analysis, which are discussed below.

Factor analysis

Since we already have too many variables which are properly selected using literature review, but we have the issue that they are collinear to each other. So here we need to reduce the variables, but we do not have any criterion. Here exploratory analysis arranges the independent variables in terms of degree of explaining ability for their common regressand (unknown dependent variable). And also provided a criterion to select the best ones if we want to use the automatic method or we can make our own criteria.

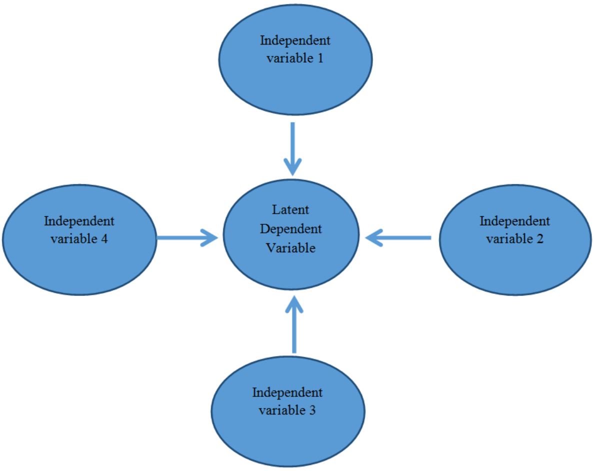


Figure 16. *Model of exploratory analysis*

This method is mostly used in cross sectional questionnaire based data, where there are too many variables at disposal and using all of them might be causing Multicollinearity.

In simple words, this principle factor analysis approach will arrange the provided variables in decreasing order of their importance which can help you to decide which variable you wish to keep and which you wish to remove.

Following illustration uses data of cross sectional data of car sales. Here we will shortlist the independent variables on the bases of the explaining power of variables. The data set provided is named as car_sales.xls.[7] First open this data set in SPSS, and select, analyze option, then dimension reduction and then factor. Here select all the proposed variables in the variable window. Here select extraction button and check the scree plot. In this window there are two approaches provided to short list the variables; the first one is Eigen value >1 method, where it selects the variables which have Eigen value more than 1, while in the second criterion we can fix how many factors we need out of the listed ones. Then press continue after that. Now press rotation button and then check varimax approach in the method window, the purpose of this approach it that it corrects the results bases on the short listed variables, press continue. Now press scores button, check the save as variables, the purpose is to store the short listed factors generated from regression approach and store it as variables to be used later and then press continue. After this, press ok to estimate the results.

Following table first assumes that all the independent variables are 100% correlated with the common unknown dependent variable, then estimates the exact correlation; here

[7] Available at [Retrieved from].

we can see that vehicle type is 93% correlated with the sales of cars similarly others.

Communalities

	Initial	Extraction
Vehicle type	1.000	.930
Price in thousands	1.000	.876
Engine size	1.000	.843
Horsepower	1.000	.933
Wheelbase	1.000	.881
Width	1.000	.776
Length	1.000	.919
Curb weight	1.000	.891
Fuel capacity	1.000	.861
Fuel efficiency	1.000	.860

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.994	59.938	59.938	5.994	59.938	59.938	3.220	32.199	32.199
2	1.654	16.545	76.482	1.654	16.545	76.482	3.134	31.344	63.543
3	1.123	11.227	87.709	1.123	11.227	87.709	2.417	24.166	87.709
4	.339	3.389	91.098						
5	.254	2.541	93.640						
6	.199	1.994	95.633						
7	.155	1.547	97.181						
8	.130	1.299	98.480						
9	.091	.905	99.385						
10	.061	.615	100.000						

Extraction Method: Principal Component Analysis.

Since there are 10 variables proposed in this method so exploratory approach created 10 components of an unknown dependent variable, and reported its Eigen value and % of variance explained by this component of the dependent variable. Here we can see that there are only three components which have Eigen value higher than 1 so three of them are short listed. We can see that theoretically, 10 variables are explaining 100% of the dependent variable, but after this approach we reduced into 3 variables means 66% reduction in the complexity of the model, contrary to it, we can see that the explanation has only been reduced by 12.3% (it can be seen by cumulative % of 87.709).

Following table is the rotated component matrix, which helps to determine which of the variables are correlated with the three components which are short listed. Here we can see that for the first component column price in thousand is most correlated with it, while for the case of the second component length of the car is most correlated and for the case of third component, vehicle type is most correlated. Hence, these three are the short listed variables. Here it's up to us to use these three variables or the three generated components. These three variables will not have issues of multicollinearity.

Rotated Component Matrix[a]

	Component		
	1	2	3
Vehicle type	-.101	.095	.954
Price in thousands	.935	-.003	.041
Engine size	.753	.436	.292
Horsepower	.933	.242	.056
Wheelbase	.036	.884	.314
Width	.384	.759	.231
Length	.155	.943	.069
Curb weight	.519	.533	.581
Fuel capacity	.398	.495	.676
Fuel efficiency	-.543	-.318	-.681

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

Structure analysis

The purpose of this structure analysis method is to make subgroups of all the available independent variable and make index for each group so that we can reduce the number of variables considerably. The difference here is that in this case we need all the variables instead of short listing it.

Like previous method, we can manually make indices if we know theoretically how many indices there are or we use automatic methods if we do not know how many indices are there. This method is called structure analysis, most of the cases the groups which are formed can be justified through the literature review.

Following diagram shows how the 5 variables are converted into two indices which then explain the dependent variable.

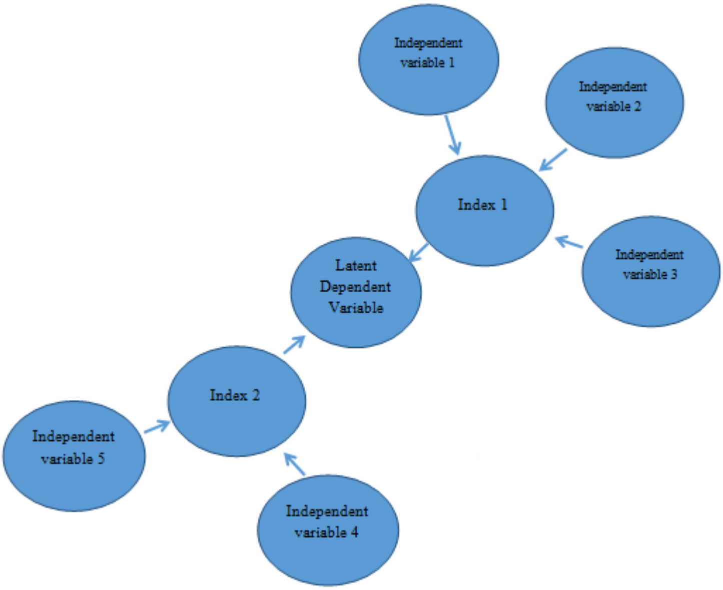


Figure 17. *Model for confirmatory analysis*

For the illustration, data about usage of differentservices of a telecommunication company are used; this data set file is named as telco.xls.[8] Here we need all the variables so we

[8] Available at [Retrieved from].

will make indices of these variables. Here open the dataset, press the analyze button on the top and go for dimension reduction and the press factor option. Here, select the proposed variables. Then select, extraction button, click the dropdown window and select principal axis factoring and check the scree plot, and press the continue button. After this, press the rotation button and check the varimax approach. Then press descriptives option and select KMO and Bartlett's test of sphericity, the purpose of this test is to check if the data is adequate enough to form indices from it, and the press continue. Now press scores and check the save as variables option and then press continue. Then, press options and check suppress small coefficients and sort by size and change the 0.10 number in the absolute value below to 0.30 and then press continue. At end, press ok to execute.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.776
Bartlett's Test of Sphericity	Approx. Chi-Square	17143.453
	df	171
	Sig.	.000

Here above it the table for KMO and Bartlett test, the KMO value must be above 0.50 and the Bartlett test must be significant so that this process of index making can be applied. If KMO is below, then this process will stop automatically.

Total Variance Explained									
Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.292	33.117	33.117	5.947	31.300	31.300	4.376	23.032	23.032
2	3.483	18.330	51.447	3.197	16.828	48.128	3.794	19.970	43.002
3	2.668	14.045	65.492	2.361	12.428	60.556	2.577	13.565	56.567
4	1.005	5.292	70.784	.896	4.718	65.274	1.654	8.707	65.274
5	.796	4.187	74.971						
6	.669	3.522	78.493						
7	.635	3.340	81.833						

8	.583	3.069	84.902
9	.475	2.498	87.400
10	.453	2.385	89.785
11	.430	2.263	92.048
12	.401	2.113	94.161
13	.380	2.002	96.163
14	.351	1.846	98.010
15	.157	.828	98.837
16	.110	.580	99.417
17	.059	.311	99.728
18	.038	.198	99.926
19	.014	.074	100.000

Extraction Method: Principal Axis Factoring.

Since 19 variables were proposed hence confirmatory approach started with 19 indices, but according to the eigenvalue> 1 criteria, there are 4 indices which are significant. We have selected 4 out of 19 indices which is 79% simplification, whereas these 4 indices are explaining 65.27% of the total variation, which is only 35% fall in efficiency.

Rotated Factor Matrix[a]

	Factor			
	1	2	3	4
Toll free last month	.802			
Call waiting	.760			
Caller ID	.758			
Call forwarding	.737			
3-way calling	.724			
Toll free over tenure	.680		.447	
Equipment last month		.888		
Equipment over tenure		.759		
Wireless last month	.551	.687		
Internet		.682		
Electronic billing		.632		
Wireless over tenure	.470	.579		
Paging service	.465	.548		
Voice mail	.443	.524		
Multiple lines		.409	.363	
Long distance over tenure			.950	
Long distance last month			.911	
Calling card last month				.938
Calling card over tenure			.502	.779

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Instead of factor matrix we use rotated factor matrix as this more robust result. Now we have to just select the variables in their respective index groups. Note that we have mentioned at the start of the test to hide the values which are less than 0.3 so that it is easier to decide.

In the case of the first index in the column one, we can see that first six variables are selected but other 4 are not selected as their correlation values are higher in other indices. For the case of second index, variables from 7 to 15 are selected in this group. For the case of third group 16th and 17th variables are selected. And the last two variables are selected in the last index.

The confirmation of the structure formed can be done using confirmatory analysis, it is done using SEM which is described in previous chapter.

Ridge regression

This method is called the Ridge Regression method available in STATA. This model is used for the case when it is compulsory to keep the collinear variables as all of them are important as per theory. In this case we have to trade off some of the efficiency of the model to correct the bias.

The idea behind the model is that, if the independent variables are collinear then the matrix X'X will not be inverted correctly as its determinant will be near to almost zero. [9,10] This means that the model is biased under the

[9] See the matrix derivation of OLS regression in that we have to take the inverse of the X'X matrix, and in order to take the inverse we have to take determinant of this matrix. Here X means set of all independent variables. So the more the independent variables are related to each other the lower the rank of the matrix will be. Rank of the matrix means the number of independent rows or number of unrelated independent variables. So lower the rank of the matrix means lower the determinant of it; and if it becomes zero then it will be impossible to estimate the regression that is why most of the software gives error of collinear independent variables when we add two same independent variables

presence of Multicollinearity, which is created because of wrong value of the inverse of X’X. [11][12]

$|X'X| \approx 0$ -- determinant almost zero under multicollinearity

So what this model does it that it introduces a scalar number whose magnitude is designed such a way that it makes the determinant non zero.

$|[X'X] + k[1]| \approx 1$ -- adjusted determinants should be almost 1

Hence theoretically we are introducing some imprecision/inefficiency in the model using a scalar (known as k value in model). Inefficiency is created because we are now using a new variable which is multiplied by k so the mean of the variable will be K times higher, this will create higher standard deviation because we know that standard deviation is lowest at mean value so if we use value some other than the mean than the standard deviation will not be lowest.

But with the insertion of this K scalar (constant number); we are ensuring that the inverse of the X’X matrix is possible and correct hence solving the biasness problem created by the correlated independent variables.

Hence we are playing with the tradeoff of biasness removal (multicollinearity removal) vs. inefficiency creation (introduction of heteroscedasticity). So first requirement to remove the multicollinearity is that the model must already have small standard errors, means that all the variables are

[10] Zero for the case of perfect Multicollinearity

[11] See the handout of derivation of OLS through matrix algebra for details

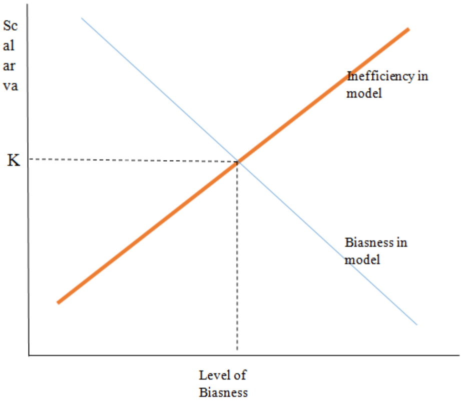
[12] See endnote for the details.

highly significant but their signs are wrong because of multicollinearity. This will only happen if there is a severe form of multicollinearity in the model.

This means that higher the k value lower the biasness (lower influence of Multicollinearity) more the inefficiency (higher Sum square residuals). Now the model requires that we need to find this value of k that is not too high, so that the model is too inefficient and not too low so that there is some biasness left.

Now here the value of the k is the balance between the degree of biasness and inefficiency in the model. Here the positive slope line is the inefficiency line which increases with the increase in the value of K explained earlier. And the thin line is the negative slope line which is biasness in the model which reduces with the increase in the value of k shown earlier.

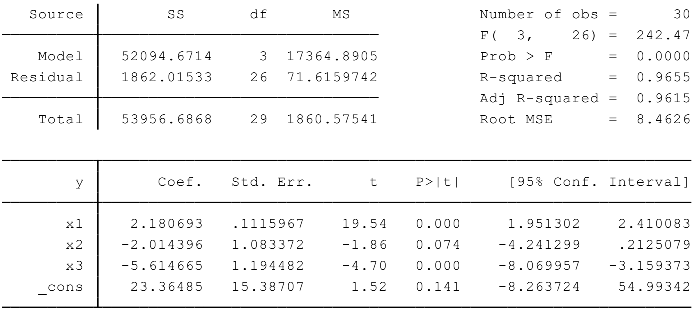
Now here we can see that if the model is already inefficient means the thicker line is on the right side, then we cannot get a high value of k, so we will not able to remove small amount of multicollinearity (if the thin line is on the left side), but we could only able to remove very high amount of multicollinearity (when the thin line is also on the right side). And when the model is efficient (the thick line is on the left side) then we can remove the small multicollinearity (when thin line is left) and high multicollinearity (when the thin line is on the right) this is because we can afford to have any value of k.



Following is the example of Ridge regression models, is described in: [Judge, *et al*., 1988; 882)], and also Theil R2 Multicollinearity Effect in: [Judge, *et al*., 1988; 872)], for Klein-Goldberger data.

This model states that domestic consumption (y) is function of wage income (x1), non-wage non-farm income (x2) and farm income (x3). Kleingoldberger.xls is used for this exercise [13]

reg y x1 x2 x3



```
      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  3,    26) =  242.47
       Model |  52094.6714     3  17364.8905           Prob > F      =  0.0000
    Residual |  1862.01533    26  71.6159742           R-squared     =  0.9655
-------------+------------------------------           Adj R-squared =  0.9615
       Total |  53956.6868    29  1860.57541           Root MSE      =  8.4626

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.180693   .1115967    19.54   0.000     1.951302    2.410083
          x2 |  -2.014396   1.083372    -1.86   0.074    -4.241299    .2125079
          x3 |  -5.614665   1.194482    -4.70   0.000    -8.069957   -3.159373
       _cons |   23.36485   15.38707     1.52   0.141    -8.263724    54.99342
------------------------------------------------------------------------------
```

We can see that R squared is high and F test is significant and surprisingly the individual variables are significant with

[13] Available at [Retrieved from].

opposite signs which shows that there is hint of severe multicollinearity in the model, as all the variables are important hence we cannot remove any one of them using stepwise regression, or we cannot make index using factor analysis since variables are few. So we use the ridge regression approach to estimate correct coefficients under the presence of the multicollinearity. There are three versions (grr1 [Judge, *et al*., 1988; 881)], grr2 [Judge, *et al*., 1988; 881)], grr3 [Strawderman, 1978]) of ridge regression which automatically estimates the value of K that reduces the multicollinearity and there is one manual method (orr [Judge, *et al*., 1988; 878)]) in which we have to manually insert the value of K. Below is the result of grr1 approach of ridge regression.

Determination of value of K

Following graph plots the value of the slopes (in y axis) for each value of K (in x axis). It can be seen that the sign of slopes gets corrected at value of K ≈ 0.25. This indicates at this point the effect of multicollinearity is removed. Here we can also see that because of multicollinearity the coefficient of x1 was over estimated and others were under estimated.

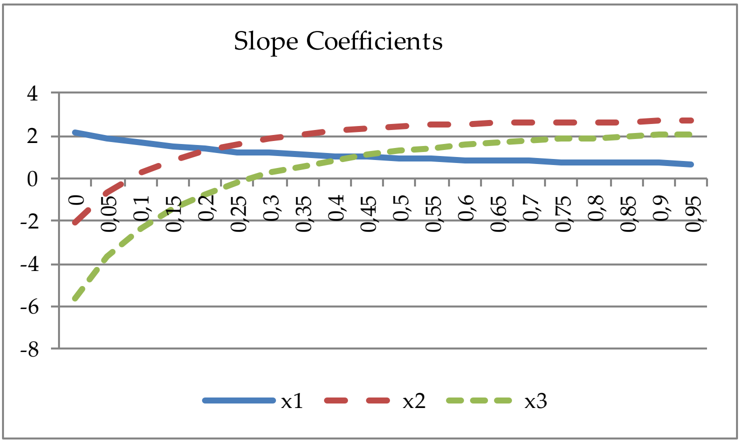
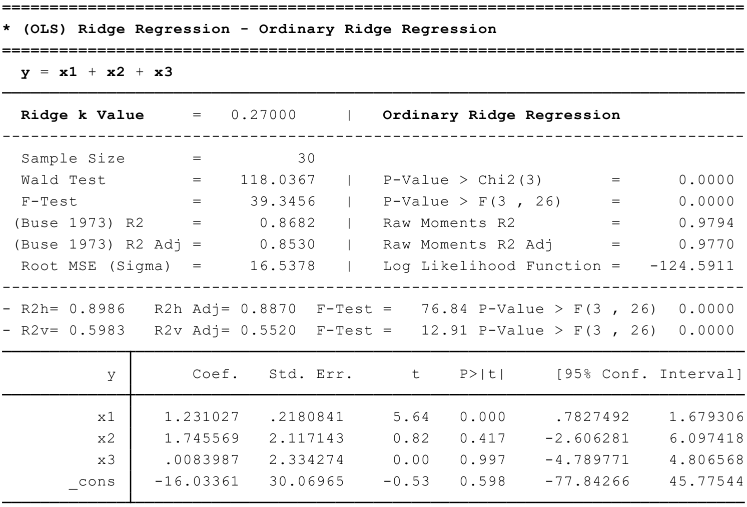


Figure 17. *Iterated valued of slope coefficients*

Estimation of ridge regression

Here it can be seen that the value k used is 0.27 in them model and we can compare the coefficients of both models this one and the OLS. See the coefficients are different which means that while correcting the multicollinearity the biasness in the coefficients gets removed and they are improved now. Note that earlier the coefficient of wage income (X1) was 2.18 means 1% increase in wage income will increase domestic consumption by 2.18% which is unbelievably high.

ridgereg y x1 x2 x3 , model(orr) kr(0.27) lmcol mfx(lin) diag



```
==============================================================================
* (OLS) Ridge Regression - Ordinary Ridge Regression
==============================================================================
  y = x1 + x2 + x3
------------------------------------------------------------------------------
  Ridge k Value     =   0.27000     |   Ordinary Ridge Regression
------------------------------------------------------------------------------
  Sample Size       =        30
  Wald Test         =  118.0367     |   P-Value > Chi2(3)       =     0.0000
  F-Test            =   39.3456     |   P-Value > F(3 , 26)     =     0.0000
 (Buse 1973) R2     =    0.8682     |   Raw Moments R2          =     0.9794
 (Buse 1973) R2 Adj =    0.8530     |   Raw Moments R2 Adj      =     0.9770
  Root MSE (Sigma)  =   16.5378     |   Log Likelihood Function =  -124.5911
------------------------------------------------------------------------------
- R2h= 0.8986   R2h Adj= 0.8870  F-Test =   76.84 P-Value > F(3 , 26)  0.0000
- R2v= 0.5983   R2v Adj= 0.5520  F-Test =   12.91 P-Value > F(3 , 26)  0.0000
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.231027   .2180841     5.64   0.000     .7827492    1.679306
          x2 |   1.745569   2.117143     0.82   0.417    -2.606281    6.097418
          x3 |   .0083987   2.334274     0.00   0.997    -4.789771    4.806568
       _cons |  -16.03361   30.06965    -0.53   0.598    -77.84266    45.77544
------------------------------------------------------------------------------
```

Here we can see that according to the C index, VIF and Eigenvalue there is no multicollinearity in the OLS model. But the Farrar and Glauber tests indicate the presence of multicollinearity.

```
==============================================================================
*** Multicollinearity Diagnostic Tests - Model= (orr)
==============================================================================

* Correlation Matrix
(obs=30)

             |       x1       x2       x3
-------------+---------------------------
          x1 |   1.0000
          x2 |   0.7539   1.0000
          x3 |   0.7483   0.6811   1.0000

* Multicollinearity Diagnostic Criteria

+-------------------------------------------------------------------------------+
|   Var |  Eigenval |  C_Number |   C_Index |       VIF |     1/VIF |   R2_xi,X |
|-------+-----------+-----------+-----------+-----------+-----------+-----------|
|    x1 |    2.4560 |    1.0000 |    1.0000 |    3.0414 |    0.3288 |    0.6712 |
|    x2 |    0.3190 |    7.6993 |    2.7748 |    2.4969 |    0.4005 |    0.5995 |
|    x3 |    0.2250 |   10.9166 |    3.3040 |    2.4485 |    0.4084 |    0.5916 |
+-------------------------------------------------------------------------------+

* Farrar-Glauber Multicollinearity Tests
  Ho: No Multicollinearity - Ha: Multicollinearity
--------------------------------------------------

* (1) Farrar-Glauber Multicollinearity Chi2-Test:
    Chi2 Test =   47.1555    P-Value > Chi2(3) 0.0000

* (2) Farrar-Glauber Multicollinearity F-Test:

+--------------------------------------------------------+
|   Variable |   F_Test |      DF1 |      DF2 |  P_Value |
|------------+----------+----------+----------+----------|
|         x1 |   27.559 |   27.000 |    3.000 |    0.009 |
|         x2 |   20.209 |   27.000 |    3.000 |    0.015 |
|         x3 |   19.555 |   27.000 |    3.000 |    0.016 |
+--------------------------------------------------------+

* (3) Farrar-Glauber Multicollinearity t-Test:

+-------------------------------------+
| Variable |       x1 |       x2 |       x3 |
|----------+----------+----------+----------|
|       x1 |        . |          |          |
|       x2 |    5.963 |        . |          |
|       x3 |    5.861 |    4.834 |        . |
+-------------------------------------+
```

The below tests also indicate for the presence of the multicollinearity in the original OLS model. This shows that there was a need of bias correction approach.

```
* |X'X| Determinant:
  |X'X| = 0 Multicollinearity - |X'X| = 1 No Multicollinearity
  |X'X| Determinant:       (0 < 0.1763 < 1)
---------------------------------------------------------------

* Theil R2 Multicollinearity Effect:
  R2 = 0 No Multicollinearity - R2 = 1 Multicollinearity
     - Theil R2:           (0 < 0.6193 < 1)
---------------------------------------------------------------

* Multicollinearity Range:
  Q = 0 No Multicollinearity - Q = 1 Multicollinearity
     - Gleason-Staelin Q0: (0 < 0.7285 < 1)
    1- Heo Range Q1:       (0 < 0.6186 < 1)
    2- Heo Range Q2:       (0 < 0.6244 < 1)
    3- Heo Range Q3:       (0 < 0.5802 < 1)
    4- Heo Range Q4:       (0 < 0.7407 < 1)
    5- Heo Range Q5:       (0 < 0.6772 < 1)
    6- Heo Range Q6:       (0 < 0.6208 < 1)
------------------------------------------------------------------
```

Below table shows the new slopes which are corrected ones and the elasticities are calculated too. Now all the coefficients are positive as expected in the theory.

```
* Marginal Effect - Elasticity (Model= orr): Linear *

+------------+--------------------+--------------------+--------------------+
|   Variable | Marginal_Effect(B) |      Elasticity(Es)|                Mean|
+------------+--------------------+--------------------+--------------------+
|         x1 |             1.2310 |             0.8321 |             66.5567|
|         x2 |             1.7456 |             0.3302 |             18.6263|
|         x3 |             0.0084 |             0.0005 |              5.7923|
+------------+--------------------+--------------------+--------------------+
 Mean of Dependent Variable =     98.4617
```

Illustration of ridge regression showed the presence of biasness and we sacrificed the efficiency of the model to correct the bias. Hence the model must have very high F values and T values so that it can absorb the decrease in efficiency for higher values of K for bias correction.

Discussions

We have discussed 3 approaches which can be used for the cross sectional or panel data sets which have small time components. We cannot use them in time series models as the multicollinearity in those models are because of trend which can be solved using cointegration models, we will study them later in chapter 7.

There are more advanced versions of ridge regression available, but make sure they are only applied in extreme forms of multicollinearity where the F test, R square is high and T values are also significant but the sign of the coefficients are wrong as compared to the theory. XTREGFEM module is available to estimate the ridge regression for panel data, with added bonus that it checks for heteroskedasticity.

Stepwise regression or factor analysis is appropriate for the case when we are using survey based data and we have too many variables available.

STATA code

```
clear
*** stepwise regression
use "D:\UMT notes\Applied Econometrics\lectures\lecture 5\stepwise regression\auto.dta", clear
generate weight2=weight^2
//Example of Theil R2 Multicollinearity Effect in:
//[Judge, et al(1988, p.872)], for Klein-Goldberger data.
regress mpg weight weight2 displ gear turn headroom foreign price
estat vif
lmcol mpg weight weight2 displ gear turn headroom foreign price
// above test check multi from different methods
graph matrix weight weight2 displ gear turn headroom price, title(Matrix Plot) subtitle(Independent Variables) scheme(sj)
//versions of stepwise regression
stepwise, pr(.2): regress mpg weight weight2 displ gear turn headroom foreign price
lmcol mpg weight weight2 gear turn foreign
// this command removes the problematic variables
stepwise, pr(.2): regress mpg weight weight2 (displ gear) turn headroom foreign price
lmcol mpg weight weight2 turn foreign
// this command makes sure if one of the bracketed one is removed the other goes too
stepwise, pr(.2) lockterm1: regress mpg weight weight2 displ gear turn headroom foreign price
// this command locks the first one because it is the most important variable
**** Principle factor analysis
// this is only the hint how it works, we will study in spss for this
pca trunk weight length headroom, components(1)
factor trunk weight length headroom, blanks(0.3)
**** Ridge regression
clear
use "D:\UMT notes\Applied Econometrics\lectures\lecture 5\ridge reg\kleingoldberger.dta", clear
**this example us using klein goldberger data
** the model is domestic consumption = f(wage income, nonwage nonfarm income, farm income)
** Source Griffiths, Hill and Judge, Learning and Practicing Econometrics, 1993, Wiley, (Table 13.1, p.433)
** Judge, Hill, Griffiths, Lutkepohl and Lee, Introduction to the Theory and Practice of Econometrics, 1988, Wiley, (Table 21.1, p. 861).
reg y x1 x2 x3
* r square is high
* but majority variables insignificant
correlate x1 x2 x3
* high correlation between x1 and x3
estat vif
* vif does not depict presence of multi
* according to partial and part correlations and
* eigen value and condition index calculated in
* spss there is multi in variable x2 and x3
ssc install ridgereg
ridgereg y x1 x2 x3 , model(orr) kr(0.5) mfx(lin) lmcol diag
** here model(orr) means
* simple ridge regression process
** here k(#) is
* the degree of biasness must be allowed to estimate under multi.
** mfx(lin) means that it will generate the marginal coefficients using linear method
** lmcol is the code to test multicollinearity
** diag means it is providing model selection criterion
ridgereg y x1 x2 x3 , model(grr1) mfx(lin) lmcol
** above one is the generalized rigde regression method
ridgereg y x1 x2 x3 , model(grr2) mfx(lin) lmcol
ridgereg y x1 x2 x3 , model(grr3) mfx(lin) lmcol
*** end of file
```

Summary

Countries use multipronged strategy to achieve the targets, since each strategy is a variable in regression, so their approach will make more than one variable move together which will cause multicollinearity in the regression analysis.

This problem is very harmful and difficult to tackle. The solution to this issue depends on the nature of multicollinearity in the model. This chapter has discussed few cases.

Application questions

1. Consider a model below, explain why multicollinearity can be expected in this model and what approach should we use to solve this problem

$$GDP_t = \alpha_1 + \beta_1 C_t + \beta_2 I_t + \beta_3 G_t + \beta_4 Nx_t + \varepsilon_t$$

Further study

Giles, D. (2013, Jun 24). Can you actually test for Multicollinearity? [Retrieved from].

End notes

Derivation of OLS matrix approach

Consider we have standard equation of the OLS. Here Y is single dependent variable so in matrix form it is a single column matrix with n rows which are equal to the sample size (n). Here X is all independent variables including the first variable which is constant 1 to represent intercept. So it has columns = number of independent variables and intercept (k) and rows are sample size and the error term has same dimension as dependent variable. Here β is also a matrix of one row and k columns.

$$Y = \beta X + e$$

So we can write it in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \hat{\beta} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Since matrices are not multiplied same way as mathematical numbers do so we have written β after the transposed X matrix. Now we can isolate the residuals as following

$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \hat{\beta}$$

Now what is regression analysis, it is the process which actually finds the optimal value of β by minimizing the e'e matrix which is actually squared residuals. So first of all we will make e'e matrix here by multiplying the above equation with its self and transposing one so that they can be multiplied.

$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}' \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \left[\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \hat{\beta} \right]' \left[\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \hat{\beta} \right]$$

$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}' \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \left[\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - 2\hat{\beta}' \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right.$$

$$\left. + \hat{\beta}' \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix} \beta \right].$$

We know that β′ = β so now we take derivative of this equation with respect to β and put it equal to zero to extract the optimal values of β

$$\frac{\vartheta e'e}{\vartheta \hat{\beta}} = \frac{\vartheta \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}' \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}}{\vartheta \hat{\beta}}$$

$$= -2 \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$+ 2 \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix} \hat{\beta} = 0$$

Form here we need to calculate the value of β so we solve the above equation

$$\begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix} \hat{\beta} = \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\hat{\beta} = \frac{\begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}{\begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}}$$

$$\hat{\beta} = \left(\begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Now the inverse of the following matrix

$$\left(\begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix} \right)^{-1}$$

$$= \frac{Adj\left(\begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix} \right)}{Det\left(\begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix}' \begin{bmatrix} 1 & x1_1 & \dots & xk_1 \\ 1 & x1_2 & \dots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \dots & xk_n \end{bmatrix} \right)}$$

For that we need to calculate the determinant in the denominator. Here we terminology needed to be understood. Rank of the matrix shows how many rows or columns of the matrix are independent (unrelated) to each other. If all the independent variables are uncorrelated to each other than the rank will be equal to the number of the variables (i.e. number of rows or columns) which is also called full rank. So here more the variables uncorrelated with each other higher will be the rank which will cause higher value of the determinant.

If there is multicollinearity in the model then the rank will be smaller which will lead to smaller determinant which will create imprecise estimate of the β value.

So the |X′X| test in the LMCOL test checks the determinant of this matrix and makes its index, if it reaches to 1 this means that it is full rank means no multicollinearity while if it reaches 0 it will show that there is perfect multicollinearity i.e. rank = 0

Now if there is extreme multicollinearity in the model then this determinant will be too small casing the making the inverse of X′X matrix approaching infinity. So what ridge regression does it that it introduces a number k like below with the identity matrix such that determinant becomes non zero, the larger the value of k the more the determinant will move away from the zero value. So in ridge regression we find the optimal value of k such that the values of β become stable, which we have checked in the slope coefficient graph. We cannot increase this k too much because even though increase in k solves biasness, but it also creates inefficiency, the standard errors will become smaller and it might lead to insignificant variables and model.

$$Det\left(\begin{bmatrix} 1 & x1_1 & \cdots & xk_1 \\ 1 & x1_2 & \cdots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \cdots & xk_n \end{bmatrix}' \begin{bmatrix} 1 & x1_1 & \cdots & xk_1 \\ 1 & x1_2 & \cdots & xk_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x1_n & \cdots & xk_n \end{bmatrix} + k \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}\right)$$

Descriptive statistics in cross sectional data

Since this data is scattered on the base of heterogeneity of the cross sections so here we have to explore the heterogeneities and inter connections of the variables with each other. Here the data used is the determinants of price of different cars.

Mean, standard deviation

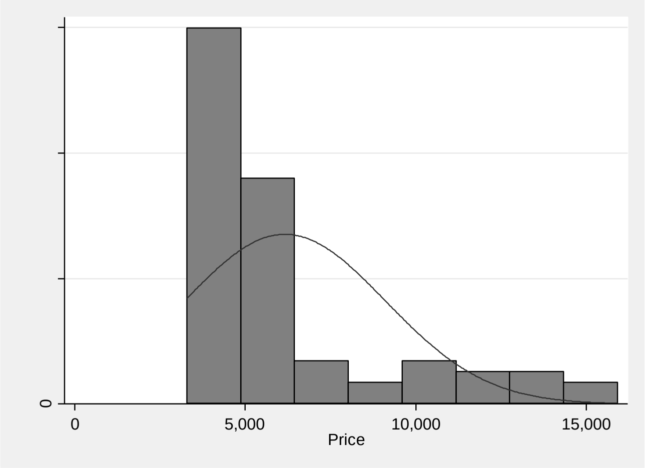
In this first set we can see the mean value of different variables and compare it with its standard deviation. If mean is larger than standard deviation then data is under dispersed (means less scattered and average value is consistent) and if mean is smaller than data is over dispersed (means average value is less consistent).

Skewness, Kurtosis and normality

Here skewness is used to see any grouping of the data, in cross sectional data it will be not zero if the cross sections are not randomly selected. If skewness is not zero then interpretation should be carefully done. For skewed variable especially dependent variable we can make histogram.

Kurtosis shows if there is any presence of outliers in the data. If kurtosis = 3 then the number of outliers present in the data are sufficient. Box plot can be made to see outliers in specific variable.

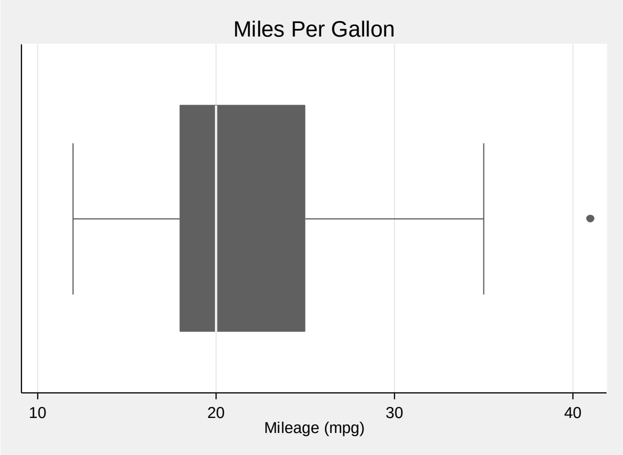
```
hist price, norm scheme(sj)
```



sum price mpg displacement, detail

```
                            Price
-------------------------------------------------------------
      Percentiles      Smallest
 1%         3291           3291
 5%         3748           3299
10%         3895           3667       Obs                  74
25%         4195           3748       Sum of Wgt.          74

50%       5006.5                      Mean           6165.257
                        Largest       Std. Dev.      2949.496
75%         6342          13466
90%        11385          13594       Variance        8699526
95%        13466          14500       Skewness       1.653434
99%        15906          15906       Kurtosis       4.819188

                        Mileage (mpg)
-------------------------------------------------------------
      Percentiles      Smallest
 1%           12             12
 5%           14             12
10%           14             14       Obs                  74
25%           18             14       Sum of Wgt.          74

50%           20                      Mean            21.2973
                        Largest       Std. Dev.      5.785503
75%           25             34
90%           29             35       Variance       33.47205
95%           34             35       Skewness       .9487176
99%           41             41       Kurtosis       3.975005

                   Displacement (cu. in.)
-------------------------------------------------------------
      Percentiles      Smallest
 1%           79             79
 5%           86             85
10%           97             86       Obs                  74
25%          119             86       Sum of Wgt.          74

50%          196                      Mean           197.2973
                        Largest       Std. Dev.      91.83722
75%          250            350
90%          350            400       Variance       8434.075
95%          350            400       Skewness       .5916565
99%          425            425       Kurtosis       2.375577
```

graph hbox mpg, ylabel(, valuelabel) title(Miles Per Gallon) scheme(sj)



Normality of variable is necessary though not an issue if sample is above 30. It provides some important information. Here significant Kurtosis can lead to heteroskedastic estimates.

sktest price mpg displacement

```
                    Skewness/Kurtosis tests for Normality
                                                        ------- joint ------
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis) adj chi2(2)    Prob>chi2
-------------+--------------------------------------------------------------
       price |     74      0.0000         0.0127        21.77        0.0000
         mpg |     74      0.0015         0.0804        10.95        0.0042
displacement |     74      0.0337         0.2048         5.81        0.0547
```

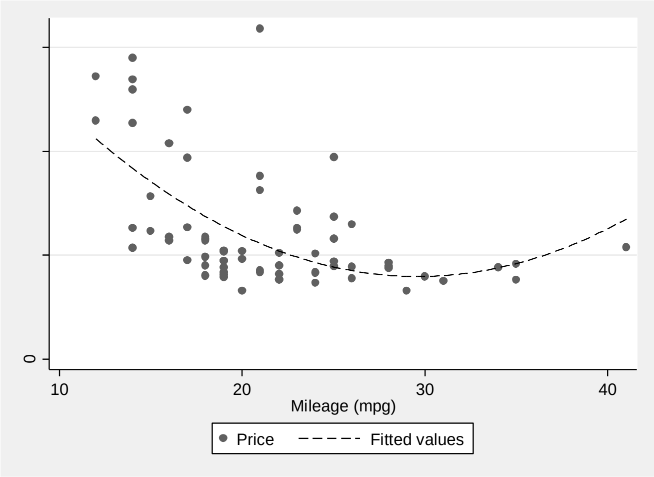
Association between quantitative variables

If dependent and independent variables are quantitative then either scatter plot, linear fit plot or quadratic flit plot can be made. Also we can calculate correlations.

pwcorr price mpg, sig

```
             |    price      mpg
-------------+------------------
       price |   1.0000
             |
             |
         mpg |  -0.4686   1.0000
             |   0.0000
             |
```

twoway (scatter price mpg) (qfit price mpg), scheme(sj)



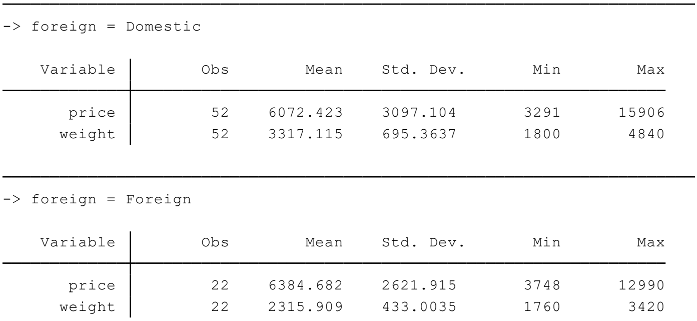
Association between quantitative and qualitative variable

If we have two qualitative variables to compare we can make cross tabulations. And for the case of quantitative and qualitative variable we can use the by sort command.

tabulate rep78 foreign, column row

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
+-------------------+

    Repair |
    Record |       Car type
      1978 |  Domestic    Foreign |     Total
-----------+----------------------+----------
         1 |         2          0 |         2 
           |    100.00       0.00 |    100.00 
           |      4.17       0.00 |      2.90 
-----------+----------------------+----------
         2 |         8          0 |         8 
           |    100.00       0.00 |    100.00 
           |     16.67       0.00 |     11.59 
-----------+----------------------+----------
         3 |        27          3 |        30 
           |     90.00      10.00 |    100.00 
           |     56.25      14.29 |     43.48 
-----------+----------------------+----------
         4 |         9          9 |        18 
           |     50.00      50.00 |    100.00 
           |     18.75      42.86 |     26.09 
-----------+----------------------+----------
         5 |         2          9 |        11 
           |     18.18      81.82 |    100.00 
           |      4.17      42.86 |     15.94 
-----------+----------------------+----------
     Total |        48         21 |        69 
           |     69.57      30.43 |    100.00 
           |    100.00     100.00 |    100.00 
```

bysort foreign: sum price weight



```
-> foreign = Domestic

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       price |         52    6072.423    3097.104       3291      15906
      weight |         52    3317.115    695.3637       1800       4840

-----------------------------------------------------------------------------------
-> foreign = Foreign

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       price |         22    6384.682    2621.915       3748      12990
      weight |         22    2315.909    433.0035       1760       3420
```

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52(3), 317-332. doi. 10.1007/BF02294359

Altman, D.G., & Andersen, P.K. (1989). *Bootstrap investigation of the stability of a Cox regression model. Statistics in Medicine*, 8(7), 771–783. doi. 10.1002/sim.4780080702

Anderson, T.W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, 22(3), 327-351. doi. 10.1214/aoms/1177729580

Anderson, T.W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34(1), 122-148. doi. 10.1214/aoms/1177704248

Anderson T.W., & Darling D.A. (1954) A test of goodness of fit. *Journal of the American Statistical Association*, 49(268), 765–769.

Anderson, T.W., & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics*, 20(1), 46-63. doi. 10.1214/aoms/1177730090

Anderson, T.W., & Rubin, H. (1950). The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics*, 21(4), 570-82. doi. 10.1214/aoms/1177729752

Angrist, J.D., & Pischke, J.S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Arshed, N., & Zahid, A. (2016). Panel monetary model and determination of multilateral exchange rate with major trading partners. *International Journal of Recent Scientific Research*, 7(4), 10551-10560.

Baltagi, B. (2008). *Econometric Analysis of Panel Data*. John Wiley & Sons.

Baum, C.F., Schaffer, M.E., & Stillman, S. (2003). Instrumental variables and GMM: Estimation and testing. *The STATA Journal*, 3(1), 1-31.

Baum, C.F., Schaffer, M.E., & Stillman, S. (2007). Enhanced routines for instrumental variables/GMM estimation and testing. *The STATA Journal*, 7(4), 465-506. doi. 10.1177/1536867X0800700402

Belsley, D. (1991) *Conditioning Diagnostics, Collinearity and Weak Data in Regression*, John Wiley & Sons, Inc., New York, USA.

Benoit, K. (2011). Linear regression models with logarithmic transformations. *London School of Economics*, 22(1), 23-36.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227), 357–365. doi. 10.1080/01621459.1944.10500699

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307-327. doi. 10.1016/0304-4076(86)90063-1

Borges, J.L. (1972). A universal history of infamy, trans. *Norman Thomas di Giovanni, London, Allen Lane*, No.141.

Breusch, T.S., & Pagan, A. R. (1980). The lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies*, 47(1), 239-253. doi. 10.2307/2297111

Brown, R.L., Durbin, J., & Evans, J.M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(2), 149-192.

Cragg, J.G., & Donald, S.G. (1993). Testing identfiability and specification in instrumental variables models. *Econometric Theory*, 9(2), 222-240. doi. 10.1017/S0266466600007519

D'Agostino, R.B., & Rosman, B. (1974) The power of Geary's test of normality. *Biometrika*, 61(1), 181-184. doi. 10.1093/biomet/61.1.181

D'Agostino, R.B., Belanger, A.J., & D'Agostino, R.B.Jr. (1990). A suggestion for using powerful and informative tests of normality. *American Statistician*, 44(4), 316–321.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Durbin, J. (1954). Errors in variables. *Review of the International Statistical Institute*, 22(1), 23-32.

Evagelia, M. (2011) *Ridge Regression Analysis of Collinear Data*. [Retrieved from].

Farrar, D., & Glauber, R. (1976). Multicollinearity in regression analysis: The problem revisited, *Review of Economics and Statistics*, 49(1), 92-107. doi. 10.2307/1937887

Gauss, C.F. (1809). Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore Carolo Friderico Gauss. sumtibus Frid. Perthes et IH Besser.

Gauss, C.F. (1821). Theory of the combination of observations which leads to the smallest errors. *Gauss Werke*, 4(1), 1-93.

Gauss, C.F. (1857). Theory of the motion of the heavenly bodies moving about the sun in conic sections: a translation of Carl Frdr. Gauss "Theoria motus": With an appendix. By Ch. H. Davis. Little, Brown and Comp.

Geary R.C. (1947) Testing for normality. *Biometrika*, 34(3-4), 209-242. doi. 10.1093/biomet/34.3-4.209

Greene, W.H. (2000). *Econometric Analysis*, (International edition).

Greene, W.H. (2008). *Econometric Analysis*. 6th ed. Upper Saddle River, NJ: Prentice–Hall.

Greene, W. H. (2012). *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.

Guilkey, David K. & Schmidt, P. (1973) Estimation of seemingly unrelated regression equations with first-order autoregressive errors. *Journal of the American Statistical Association*, 68(343), 642-647.

Gujarati, D.N. (2009). *Basic Econometrics*. Tata McGraw-Hill Education.

Hansen, L.P. (1982). Large sample properties of generalised method of moments estimators. *Econometrica, 50*(4), 1029-1054. doi. 10.2307/1912775

Harvey, A.C. (1990) *The Econometric Analysis of Time Series*, 2nd Edition, MIT Press Books, The MIT Press.

Hausman, J.A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 46(6), 1251-1271. doi. 10.2307/1913827

Hayes, A.F. (2013). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. Guilford Press.

Henderson, H.V., & Velleman, P.F. (1981). Building multiple regression models interactively. *Biometrics*, 37(2), 391-411. doi. 10.2307/2530428

Holly, S., Pesaran, M.H., & Yamagata, T. (2010). A spatio-temporal model of house prices in the USA. *Journal of Econometrics*, 158(1), 160-173. doi. 10.1016/j.jeconom.2010.03.040

Hosmer Jr, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied Logistic Regression*, Vol.398. John Wiley & Sons.

Jackson, J.E. (2003). *A User's Guide to Principal Components*. New York: Wiley.

Jarque, C.M. & Bera, A.K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, 55(2), 163-172. doi. 10.2307/1403192

Judd, C.M., McClelland, G.H., & Ryan, C.S. (2011). *Data Analysis: A Model Comparison Approach*. Routledge.

Judge, G., Hill, R.C., Griffiths, W.E., Lutkepohl, H., & Lee, T.C. (1988). *Introduction To The Theory And Practice Of Econometrics*, 2nd ed., John Wiley & Sons, Inc.

Kleibergen, F. (2007). Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics. *Journal of Econometrics*, *139*(1), 181-216. doi. 10.1016/j.jeconom.2006.06.010

Kleibergen, F., & Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, 133(1), 97-126. doi. 10.1016/j.jeconom.2005.02.011

Kleibergen, F., & Schaffer, M.E. (2007). ranktest: STATA module for testing the rank of a matrix using the Kleibergen-Paap rk statistic. [Retrieved from].

Klein, L.R. (1950). *Economic Fluctuations in the United States 1921–1941*. New York: Wiley.

Legendre, A.M. (1805). *New Methods for Determining the Orbits of Comets*, No.1. F. Didot.

Al Mamun, M., Sohag, K., & Hassan, M.K. (2017). Governance, resources and growth. *Economic Modelling*, 63, 238-261. doi. 10.1016/j.econmod.2017.02.015

Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, 59(2), 347-370. doi. 10.2307/2938260

Sargan, J.D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, 26(3), 393-415. doi. 0012-9682(195807)26:3<393:TEOERU>2.0.CO;2-R

Shapiro, S.S., & Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591–611. doi. 10.1093/biomet/52.3-4.591

Stock, J.H. & Wright, J.H. (2000). GMM with weak identification. *Econometrica*, 68(5), 1055-1096. doi. 10.1111/1468-0262.00151

Stock, J.H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, [Retrieved from].

Strawderman, W.E. (1978) Minimax adaptive generalized ridge regression estimators. *Journal American Statistical Association*, 73(363), 623-627.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, 26(1), 24-36. doi. 0012-9682(195801)26:1<24:EORFLD>2.0.CO;2-R

Wooldridge, J.M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT press.

Wooldridge, J.M. (2009). Introductory econometrics: A modern approach, South-Western Pub.

Zellner, A., & Theil, H. (1962). Three stage least squares: Simultaneous estimate of simultaneous equations. *Econometrica* 29(1), 54–78. doi. 0012-9682(196201)30:1<54:TLSSEO>2.0.CO;2-3

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298), 348-368.

Zellner, A. (1963). Estimators for seemingly unrelated regression equations: Some exact finite sample results. *Journal of the American Statistical Association*, *58*(304), 977-992.

Zellner, A., & Huang, D.S. (1962). Further properties of efficient estimators for seemingly unrelated regression equations. *International Economic Review*, 3(3), 300-313.

Applied Cross-Sectional Econometrics

Author: **Noman Arshed**
Department of Economics, School of Business and Economics,
University of Management and Technology Lahore Pakistan.

ISBN: 978-605-7736-94-9 (e-Book)
KSP Books 2020
© KSP Books 2020

Noman Arshed

Noman Arshed is currently Lecturer in Department of Economics, School of Business and Economics, University of Management and Technology Lahore Pakistan. He has completed his MS Economics from the University of Edinburgh, Edinburgh United Kingdom. He has experience in Economics, Mathematics and Statistics which lead opt the specialization of Econometrics. The Author has taught statistics and Econometrics in baccalaureate and masters level as well as taught several statistical software packages like Eviews, SPSS, STATA and Microfit. The author maintains an online blog of http://nomanarshed.wordpress.com, which is regularly used to share discussions regarding Econometrics.

The author has provided the following online link https://nomanarshed.wordpress.com/applied-econometric-models/ which can be accessed to get the data sets for the examples illustrated in the book. Readers and resource persons are welcome to visit the webpage and provide your comments and suggestion regarding this effort to simplify Econometrics.

KSP Books

e-ISBN
978-605-7736-94-9
KSP Books 2020