Mekelle University

Statistics Department

ECONOMETRICS USING R

By

TEDROS GEBREGERGS

Course Guide Book
Course Titles/Code: Econometrics (Stat3061)
Credit: 5 EtCTS
Module title/code: Statistical Modeling II (Stat-M3061)
Course Type: Core

May 2021

Abstract

(the spacing is set to 1.5)

no more than 250 words for the abstract

- a description of the research question/knowledge gap – what we know and what we don't know
- how your research has attempted to fill this gap
- a brief description of the methods
- brief results
- key conclusions that put the research into a larger context

Contents

1 Introduction

1.1 Definition of Econometrics

Definition: Literal meaning, the word 'econometrics' means measurement in economics: **"econo-metrics"**.

- In general, econometrics is the application of statistical and mathematical methods to the analysis of economic data with a purpose of giving empirical content to economic theories **and verifying or refuting them**.

- Econometrics aims at giving **empirical content** to economic relationships.

* The **exciting** thing about econometrics is its concern for **verifying or refuting** economic laws. These economic laws or hypotheses are testable with economic data.

"Econometricsis about how we can use theory and data from economics, business, and the social sciences, along with tools from statistics, to answer **how much** questions."

In economics we express our ideas about relationships between economic variables using the mathematical concept of a function. For example, to express a relationship between income and consumption, we may write

$$CONSUMPTION = f(INCOME)$$

which says that the level of consumption is some function, $f(*)$, of income.

The supply of an agricultural commodity such as beef might be written as

$$Q^s = f(P, P^c, P^f)$$

where Q^s is the quantity supplied, P is the price of beef, P^c is the price of competitive products in production (e.g., the price of hogs), and P^f is the price of factors or inputs (e.g., the price of corn) used in the production process.

Each of the above equations is a general economic model that describes how we visualize the way in which economic variables are interrelated. Economic models of this type guide our economic analysis. For most economic decision or choice problems, it is not enough to know that certain economic variables are interrelated, or even the direction of the relationship. In addition, we must understand the magnitudes involved. That is, we must be able to say how much a change in one variable affects another.

1.2 Economic Model Building

- The first task an econometrician faces is that of **formulating an econometric model**.
- **What is a model**?
  - A model is a simplified representation of a real-world process.
  - The goal of a model is to provide a simple low-dimensional summary of a dataset.
  - The goal of a model is not to uncover truth, but to discover a simple approximation that is still usef
- Eg., 'the demand for oranges depends on the price of oranges' is a simplified representation since there are a host of other variables
- simple models are easier:
  - To understand
  - To communicate
  - To test empirically with data
- parsimonious model
- two criticisms with simple model
  - The model is oversimplified
  - The assumptions are unrealistic

An econometric model consists of the following:

a) A set of **behavioural equations** derived from the economic model. These equations involve some observed variables and some 'disturbances'.
b) A statement of whether there are **errors** of observation in the observed variables.
c) A **specification of the probability distribution** of the 'disturbances '.

1.3 Scope and Objective of Econometrics

- Developing statistical methods for the estimation of economic relationships,
- Testing economic theories and hypothesis,
- Evaluating and applying economic policies,
- Forecasting,
- Collecting and analyzing non experimental or observational data.

Components of Econometrics

- Econometric inputs:
  - Economic Theory
  - Mathematics
  - Statistical Theory
  - Data
  - Interpretation
- Econometric outputs:
  - Estimation - Measurement
  - Inference - Hypothesis testing
  - Forecasting - Prediction
  - Evaluation

We learned statistical methods so why do we need a separate discipline?

- The reason is as follows: econometrics focuses on the analysis of **non experimental** economic data.
- Non experimental data (or observational data) are not obtained through **controlled experiments** on economic agents(consumers, firms, households, sectors, countries, etc.)
- Experimental data are collected in **laboratory environments in natural sciences**.
- but for social sciences, econometrics develops special methods to handle non-experimental data

1.4 Steps in Econometric Analysis

- Formulation of theory or hypothesis,

- Specifcation of economic (mathematical) model,
- Specifcation of econometric model,
- Collecting data,
- Estimation of parameters,
- Hypothesis tests,
- Forecasting/Prediction),
- Evaluation of results for policy analysis or decision making

1. We begin with an **economic model** which is a set of **assumptions** that describes the behaviour of an **economic phenomenon**.
2. Formulation (specification) of an economic model: a **set of equations** derived from the economic model that involve some **observed variables and some 'disturbances'**.
3. Collection of **relevant data** on variables implied by the econometric model.
4. **Estimation of model parameters** using mathematical statistics and probability theory.
5. We **conduct tests** to verify whether:
   - The specification of the model is correct
   - Model assumptions are valid
6. Based on step:
   - If the **model failed to pass the specification testing and diagnostic checking step**, then **revise** .
   - If the model passes , then one has to **proceed**with testing any hypothesis of interest (e.g. which of the explanatory variables significantly affect the response variable?).
7. We use the estimated model for **predictions and policy**

ECONOMIC MODEL vs. ECONOMETRIC MODEL

- Example1: **consumption**, c, a linear function of **income**, i

$$c = f(i) = \beta_0 + \beta_1 income \quad Economic\ model (Mathematical\ Model)$$

$$c = f(i) = \beta_0 + \beta_1 income + \epsilon \quad Econometric\ mode$$

- Example2 - Job Training and Worker Productivity

$$wage = f(educ; exper; training)$$

where $wage : hourly\ \ wage(in\ \ dollars)\quad educ : level\ \ of\ \ education(in\ \ years)$
$exper : level\ \ of\ \ workforce\ \ experience(in\ years)\quad training : weeks\ spent\ in\ job\ training.$

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training : \quad Economic\ Model$$

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + \epsilon : \quad Econometric\ Model$$

2 Review of Regression and Problems in Econometrics

2.1 Simple linear regression

Introduction

Regression analysis is one of the most commonly used tools in econometric work.
Definition: Regression analysis is concerned with describing and evaluating the relationship between a given variable (often called the **dependent variable**) and one or more variables which are assumed to influence the given variable (often called **independent** or **explanatory variables**).
Regression analysis is the method to discover the relationship between one or more response variables (also called dependent variables, explained variables, predicted variables, or regressands, usually denoted by y) and the predictors (also called independent variables, explanatory variables, control variables, or regressors, usually denoted by $x_1, x_2, ..., x_k$).

2.1.1 Population Regression Function

```
demo1<-read.csv("PRF.csv",header = T,sep = ",")
demo2<-read.csv("demo2.csv",header = T,sep = ",")
```

Table 1: Weekly Family Income X

X	X80	X100	X120	X140	X160	X180	X200	X220	X240	X260
2	55	65	79	80	102	110	120	135	137	150
3	60	70	84	93	107	115	136	137	145	152
4	65	74	90	95	110	120	140	140	155	175
5	70	80	94	103	116	130	144	152	165	178
6	75	85	98	108	118	135	145	157	175	180
7	NA	88	NA	113	125	140	NA	160	189	185
8	NA	NA	NA	115	NA	NA	NA	162	NA	191
Total	25	462	445	707	678	750	685	1043	966	1211
E(Y|X)	65	77	89	101	113	125	137	149	161	173

The data in table1 refers to a total population of 60 families in a hypothetical community and their weekly income (X)(columns) and weekly consumption expenditure (Y)(first 7 rows), both in dollars. The 60 families are divided into 10 income groups (from $80 to $260) and the weekly expenditures of each family in the various groups are as shown in the table(or read `demo2.csv`). Therefore, we have 10 fixed values of X and the corresponding Y values against each of the X values; so to speak, there are 10 Y sub populations.

There is considerable variation in weekly consumption expenditure in each income group, which can be seen clearly from Figure 1. Despite the variability of weekly consumption expenditure within each income group, on the average, weekly consumption expenditure increases as income increases. corresponding to the weekly income level of $80, the mean consumption expenditure is $65, while corresponding to the income level of $200, it is $137.

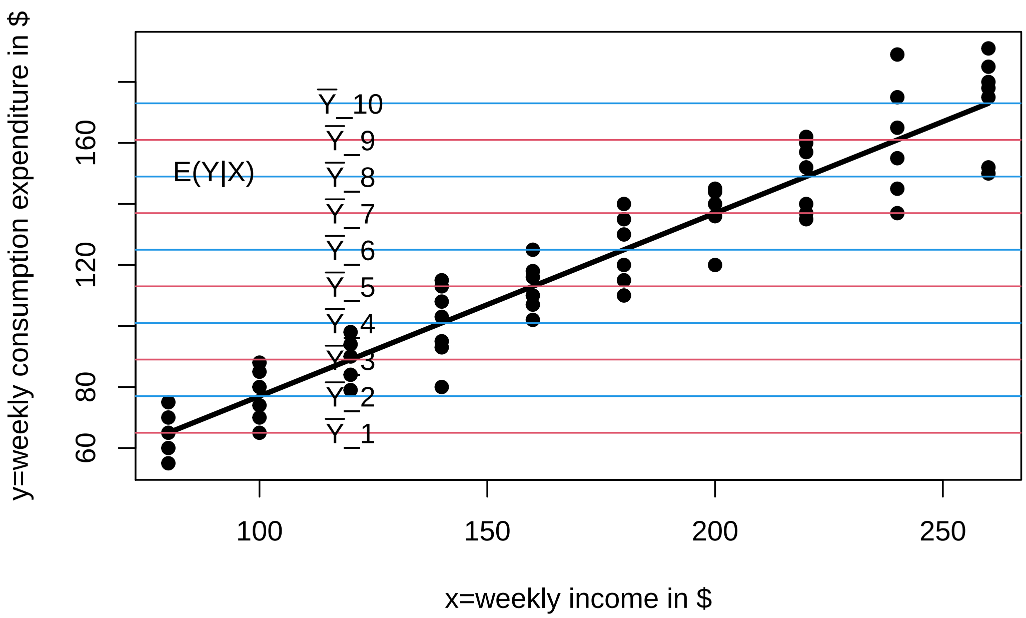

Figure 1: Conditional distribution of expenditure for various levels of income

In all we have 10 mean values for the 10 sub populations of Y. We call these mean values **conditional expected values**,as they depend on the **given values** of the (conditioning) variable X. Symbolically, we denote them as $E(Y|X)$, which is read as the expected value of Y given the value of X. **Unconditional expected value** of weekly consumption expenditure, E(Y), is obtained adding the weekly consumption expenditures for all the 60 families in the population and dividing this number by 60 (7272/60 = 121.2), we get the number $$$121.20.

When we ask the question, "What is the expected value of weekly consumption expenditure of a family," we get the answer $121.20 (the unconditional mean). But if we ask the question, "What is the expected value of weekly consumption expenditure of a family whose monthly income is,say, $140," we get the answer $101 (the conditional mean).

```
mean(demo2$y[demo2$x==140])
```

```
## [1] 101
```

A population regression curve(**population regression line (PRL)**) is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variable(s). More simply, it is the curve connecting the means of the sub populations of Y corresponding to the given values of the regressor X. It can be depicted as in Figure 2.

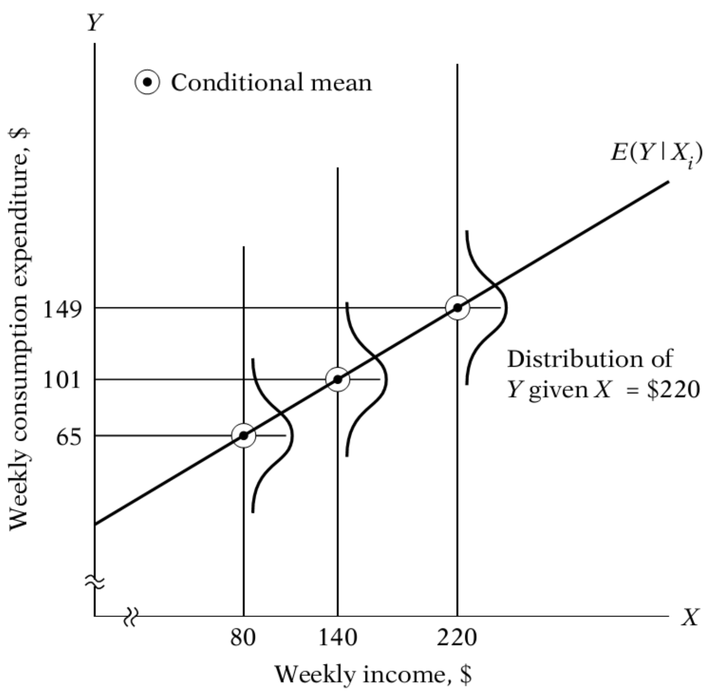

Figure 2: Population regression line

This figure shows that for each X(i.e., income level) there is a population of Y values (weekly consumption expenditures) that are spread around the (conditional) mean of those Y values. For simplicity, we are assuming that these Y values are distributed symmetrically around their respective (conditional) mean values. And the regression line (or curve) passes through these (conditional) mean values.

2.1.1.1 The concept of population regression function(PRF)

From the preceding discussions , it is clear that each conditional mean $E(Y|X_i)$ is a function of X_i, where X_i is a given value of X. Symbolically,

$$E(Y|X_i) = f(X_i).......................................(2.1.1)$$

where $f(X_i)$ denotes some function of the explanatory variable X. Equation (2.1.1) is known as the conditional expectation function (CEF)or **population regression function** (PRF) or population regression (PR)for short. It states merely that the expected value of the distribution of Y given X_i is functionally related to X_i. In simple terms, it tells how the mean or average response of Y varies with X.

What form does the function $f(X_i)$ assume? This is an important question because in real situations we do not have the entire population avail-able for examination. The functional form of the PRF is therefore an **empirical question**.

we may assume that the **PRF E(Y|Xi)** is a linear function of X_i,say, of type

$$E(Y|X_i) = \alpha + \beta X_i.......................................(2.1.2)$$

where α *and* β are **unknown but fixed** parameters known as the regression coefficients.

In regression analysis our interest is in estimating the **PRFs** like (2.1.2),that is, estimating the values of the unknowns α *and* β on the basis of observations on Y and X.

2.1.1.2 Stochastic specfication of PRF

It is clear from Figure 2.1 that, as family income increases, family consumption expenditure on the average increases, too.

Given the income level of X_i, an individual family's consumption expenditure is clustered around the average consumption of all families at that X_i, that is, around its conditional expectation. Therefore, we can express the deviation of an individual Y_i around its expected value as follows:

$$\mathcal{E}_i = Y_i - E(Y|X_i)$$
$$\Longrightarrow Y_i = E(Y|X_i) + \mathcal{E}_i \qquad (2.1.3)$$

where the deviation $\mathcal{E}_i$ is an **unobservable random variable** taking positive or negative values. Technically, $\mathcal{E}_i$ is known as the stochastic disturbance or stochastic error term.

Given the income level,We can say that the expenditure of an individual family can be expressed as the sum of two components:

- 1)$E(Y|X_i)$, which is simply the mean consumption expenditure of all the families with the same level of income. This component is known as the **systematic,or deterministic** component, and
- 2)$\mathcal{E}_i$, which is the random, or non systematic component.

If $E(Y|X_i)$ is assumed to be linear in X_i, as in (2.1.2), equation (2.1.3) may be written as

$$Y_i = E(Y|X_i) + \mathcal{E}_i$$
$$= \alpha + \beta X_i + \mathcal{E}_i \qquad (2.1.4)$$

The above equation shows that the consumption expenditure of a family is linearly related to its income plus the disturbance term.

Thus, the individual consumption expenditures, given $X = \$80$(see Table 1), can be expressed as

$$Y_1 = 55 = \alpha + \beta(80) + \mathcal{E}_1$$
$$Y_2 = 60 = \alpha + \beta(80) + \mathcal{E}_2$$
$$Y_3 = 65 = \alpha + \beta(80) + \mathcal{E}_3$$
$$Y_4 = 70 = \alpha + \beta(80) + \mathcal{E}_4$$
$$Y_5 = 75 = \alpha + \beta(80) + \mathcal{E}_5$$

Now if we take the expected value of (2.1.3) on both sides, we obtain

$$\begin{aligned} E(Y_i|X_i) &= E(E(Y|X_i)) + E(\mathcal{E}_i|X_i) \\ &= E(Y|X_i) + E(\mathcal{E}_i|X_i) \text{.............................(2.1.5)} \\ &\implies E(Y_i|X_i) - E(Y|X_i) = E(\mathcal{E}_i|X_i) \\ &\implies E(\mathcal{E}_i|X_i) = 0 \end{aligned}$$

Thus, the assumption that the regression line passes through the conditional means of Y(see Figure 2.2) implies that the conditional mean values of $\mathcal{E}_i$ (conditional upon the given X's) are zero.

Example: Suppose the relationship between consumption (Y) and income (X) of households is expressed as:

$$Y_i = 0.6X_i + 120$$

where Y_i=consumption of i^{th} household and X_i=income of i^{th} household

Here, on the basis of income, we can predict consumption. For instance, if the income of a certain household is 1500 Birr, then the estimated consumption will be:

$$\text{consumption} = 0.6(1500) + 120 = 1020 \text{ Birr}$$

Note that since consumption is estimated on the basis of income, consumption is the dependent variable and income is the independent variable.

The error term

Consider the above model: Y = 0.6X + 120. This functional relationship is deterministic or exact, that is, given income we can determine the exact consumption of a household. But in reality this rarely happens: different households with the same income are not expected to consume equal amount and this may be due to difference in real wealth or varying tastes, or unforeseen events that induce households to consume more or less.

Thus, we should express the regression model as equation (2.1.4).

Generally the reasons for including the error term include:

1. **Omitted variables:** a model is a simplification of reality. It is not always possible to include all

relevant variables in a functional form . For instance,

- the omission of relevant factors that could influence consumption, other than income, like real wealth or varying tastes, or unforseen events that induce households to consume more or less.
- we may construct a model relating demand and price of a commodity. But demand is influenced not only by own price: income of consumers, price of substitutes and several other variables also influence it.

  The omission of these variables from the model introduces an error.

2. **Measurement error:** inaccuracy in collection and measurement of sample data.

- households may not report their consumption or income accurately

3. **Sampling error:** Consider a model relating consumption (Y) with income (X) of households. The sample we randomly choose to examine the relationship may turn out to be predominantly poor households. In such cases,our estimation of α and β from this sample may not be as good as that from a balanced sample group.
4. **wrong choice of a linear relationship** between consumption and income, when the **true** relationship may be **nonlinear**.

Note that the size of the error $\mathcal{E}_i$ is not fixed: it is **non-deterministic** or **stochastic** or **probabilistic** in nature. This in turn implies that Y_i is also probabilistic in nature. Thus, the probability distribution of Y_i and its characteristics are determined by the values of X_i and by the probability distribution of $\mathcal{E}_i$.

Thus, a full specification of a regression model should include a specification of the probability distribution of the disturbance (error) term. This information is given by what we call basic assumptions or assumptions of the classical linear regression model (CLRM).

Consider the model:

$$Y_i = \alpha + \beta X_i + \mathcal{E}_i \qquad i = 1, 2, ..., n.$$

Here the subscript i refers to the i^{th} observation. In the CLRM, Y_i and X_i are observable while $\mathcal{E}_i$ is not. If i refers to some point or period of time, then we speak of time series data. On the other hand, if i refers to the i^{th} individual, object, geographical region, etc., then we speak of cross-sectional data.

2.1.1.3 Assumptions of the classical linear regression model

$$Y_i = \alpha + \beta X_i + \mathcal{E}_i \qquad i = 1, 2, ..., n.$$

$Y\ and\ X$ is observable, but not $\mathcal{E}$

- **A1:** The true model is:$Y_i = \alpha + \beta X_i + \mathcal{E}_i$
- the relationship between Y_i and X_i is **linear**, which is linear in the parameters and
- the deterministic component $(\alpha + \beta X_i)$ and the stochastic component $(\mathcal{E}_i)$ are **additive**.
- **A2:** The error terms have **zero mean**:E($\mathcal{E}_i$)=0.
- this assumption tells us that the mean of the Y_i is:

  $$E(Y_i) = \alpha + \beta X_i$$

  ,which is **non-stochastic**.
- is needed to insure that **on the average** we are on the **true line**.
- **A3: Homoscedasticity**(error terms have constant variance): $Var(\mathcal{E}_i) = E(\mathcal{E}_i^2) = \sigma^2$ for all i
- This assumption tells us that**every disturbance** has the **same variance** σ^2 whose value is **unknown**, that is, regardless of whether the X_i are **large or small**, the dispersion of the disturbances is the **same**. For example, the variation in consumption level of low income households is the same as that of high income households.
- This insures that every observation is **equally reliable**.
- **A4: No error autocorrelation**(the error terms $\mathcal{E}_i$ are statistically independent of each other): cov($\mathcal{E}_i, \mathcal{E}_j$)=E($\mathcal{E}_i\mathcal{E}_j$)=0 $for\ i \neq j$.
- It states that the disturbances are uncorrelated.
- Knowing the i^th disturbance does not tell us anything about the j^th disturbance.
- For example,

1) In the consumption case, the **unforseen** disturbance which caused the i^th household to

consume more, (like **a visit of a relative**), has **nothing** to do with the unforseen disturbances of any other household. However, this achieves in random sample of households

2) the fact that output is higher than expected today should not lead to a higher (or lower) than expected output tomorrow.

- **A5:** X_i are **deterministic**(non-stochastic):X_i and $\mathcal{E}_i$ are independent for all i, j
- It states that X_i are **not random variables**, and **hence** the probability distribution of $\mathcal{E}_i$ is in no way affected by the X_i.
- $\Longrightarrow \sum \mathcal{E}_i x_i = 0$
- Fixed in repeated sampling
- **A6: Normality:** $\mathcal{E}_i$ are normally distributed with mean zero and variance σ^2 for all i (often written as: $\mathcal{E}_i \sim$ N(0,σ^2)).

  We need this assumption for parameter estimation purposes and also to make inferences on the basis of the normal (t and F) distribution.

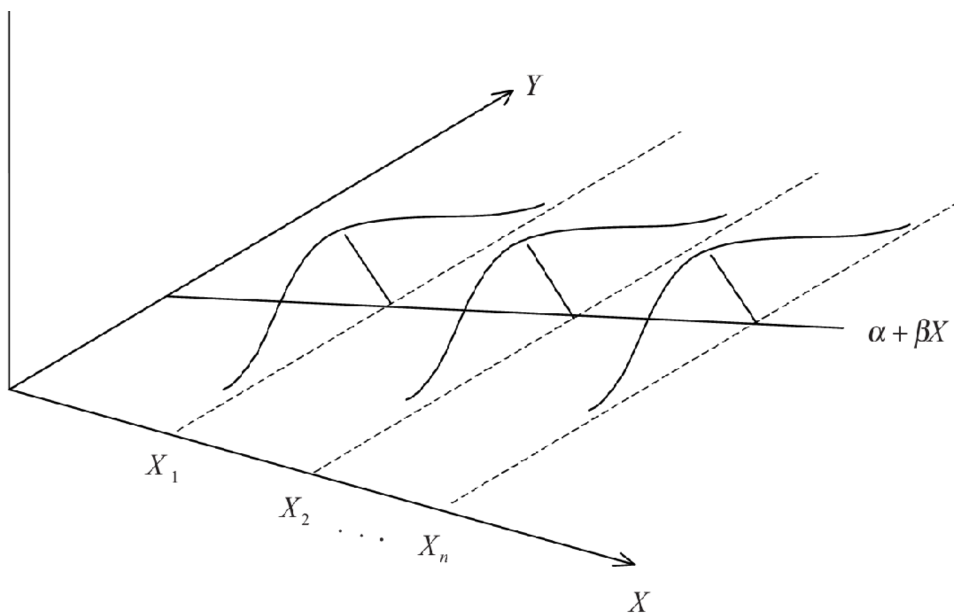

Figure 3: Random Disturbances Around the Regression

2.1.2 The ordinary least squares (OLS) method of Estimation

- Least Square method is one of the paramater estimation method which minimizes the **SS of Residuals**
  - Sample Regression Line given by:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i \qquad , i = 1, 2, ..., n.$$

where α and β are estimated by $\hat{\alpha}$ and $\hat{\beta}$, respectively, and $\hat{Y}$ is the estimated value of Y.

- Residual is the deviation b/n observed and estimated y.

$$\hat{\mathcal{E}}_i = Y_i - \hat{Y}_i \qquad , i = 1, 2, ..., n.$$

- a **good fit** minimizes the error between the **estimated points on the line** and the **actual observed points** , and hence we search which line is?

Residuals

```r
knitr::include_graphics("pop2.png")
```

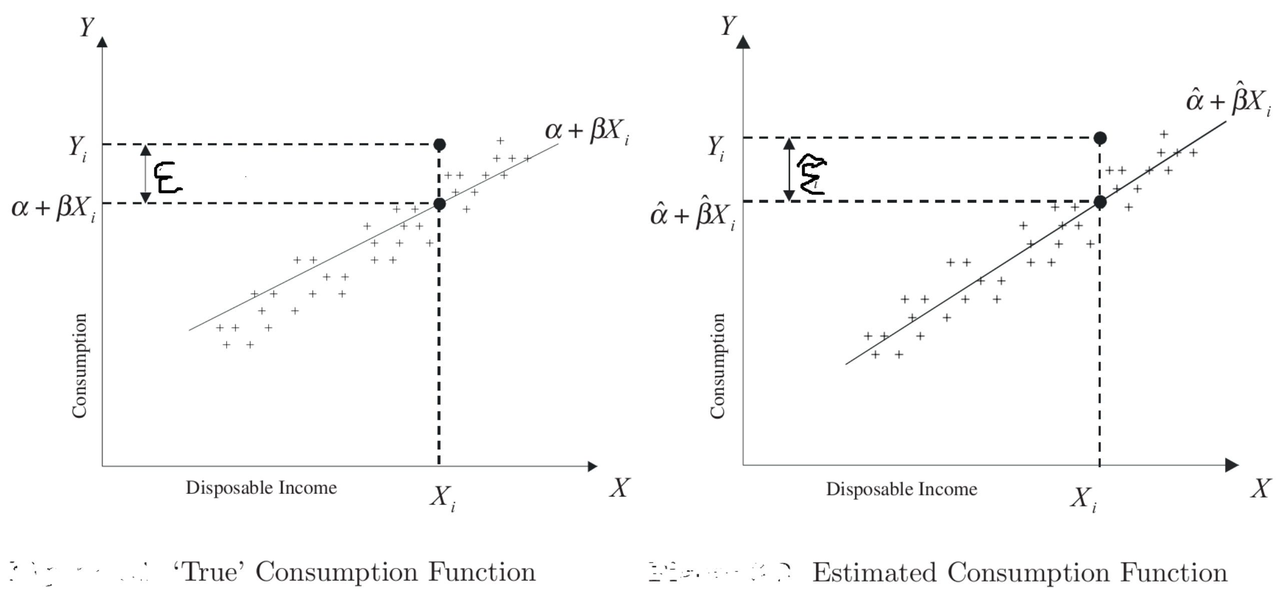

Figure 4: True Regression Line and Estimated Line

The sum of squares of the errors (SSE)

$$SSE = \sum \hat{\mathcal{E}}_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

- By partial diff of the SSE wrt $\hat{\alpha}$ and $\hat{\beta}$ and equating to zero

$$\frac{\partial SSE}{\partial \hat{\alpha}} = -2 \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0$$
$$\frac{\partial SSE}{\partial \hat{\beta}} = -2 \sum X_i (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0$$

- normal equations:

$$\begin{aligned} \sum Y_i \quad &= n\hat{\alpha} + \hat{\beta} \sum X_i \\ \sum X_i Y_i &= \hat{\alpha} \sum X_i + \hat{\beta} \sum X_i^2 \end{aligned}$$

- Thus, 2 equations with 2 unknowns $\hat{\alpha}$ and $\hat{\beta}$. Solving for $\hat{\alpha}$ and $\hat{\beta}$

$$\hat{\beta} = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n \sum X_i^2 - (\sum X_i)^2} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta} = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n \sum X_i^2 - (\sum X_i)^2} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

where $\bar{X} = \frac{1}{n} \sum X_i$ and $\bar{Y} = \frac{1}{n} \sum Y_i$

- Hence,$\hat{\alpha}$ and $\hat{\beta}$ are **ordinary least-squares (OLS) estimators** of α and β.

- The line $\hat{Y} = \hat{\alpha} + \hat{\beta} X$ is called the **least squares line** or the **estimated regression line** of Y on X.

Model in deviations form

$$Y_i = \alpha + \beta X_i + u_i \tag{1}$$

Applying summation & dividing by n to both sides:

$$\sum_{i=1}^{n} \frac{Y_i}{n} = \sum_{i=1}^{n} \frac{\alpha}{n} + \sum_{i=1}^{n} \frac{\beta X_i}{n} + \sum_{i=1}^{n} \frac{u_i}{n}$$

$$\Rightarrow \bar{Y} = \alpha + \beta \bar{X} + \bar{u} \tag{2}$$

Subtracting equation (2) from (1) we get:

$$Y_i - \bar{Y} = \beta(X_i - \bar{X}) + (u_i - \bar{u}) \tag{3}$$

Letting $x_i = X_i - \bar{X}$,$y_i = Y_i - \bar{Y}$ and $\mathcal{E}_i = (u_i - \bar{u})$, eq(3) as:

$$y_i = \beta x_i + \mathcal{E}_i \tag{4}$$

Eq (4) is the simple linear regression model in **deviations form**. - The OLS estimator of β from equation (4) is given by:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

The Gauss-Markov Theorem

- Under assumptions (1) – (5) of the **CLRM**, the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are **Best Linear Unbiased Estimators (BLUE)**.
- The theorem tells us that of all estimators of α and β **which are linear and which are unbiased**, the estimators resulting from OLS have the **minimum variance**, that is, $\hat{\alpha}$ and $\hat{\beta}$ are the best (most efficient) linear unbiased estimators (BLUE) of α and β.

<u>Note:</u> If some of the assumptions stated above do not hold, then OLS estimators are **no more**

BLUE!!!

Here we will prove that $\hat{\beta}$ is the BLUE of β. The proof for $\hat{\alpha}$ can be done similarly

a) To show that $\hat{\beta}$ is a linear estimator

The OLS estimator of β can be expressed as:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \sum a_i y_i$$

where $a_i = \frac{x_i}{\sum x_i^2}$,$x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$.

- Thus, we can see that $\hat{\beta}$ is **a linear estimator** as it can be written as a **weighted average** of the individual observations on Y.

b) To show that $\hat{\beta}$ is an unbiased estimator of β

Note: An estimator $\hat{\theta}$ of θ is said to be **unbiasedif**: $E(\hat{\theta}) = \theta$.

Consider the model in deviations form: $y_i = \beta x_i + \mathcal{E}_i$.

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i(\beta x_i + \mathcal{E}_i)}{\sum x_i^2} = \frac{\beta \sum x_i^2 + \sum x_i \mathcal{E}_i}{\sum x_i^2} = \beta + \frac{\sum x_i \mathcal{E}_i}{\sum x_i^2} \qquad (*)$$

Now we have:

- E(β)=β (since β is a constant)
- E($\sum x_i \mathcal{E}_i$)=$\sum x_i E(\mathcal{E}_i)$= $\sum x_i$(0)=0 (since x_i is non-stochastic (A5),and $E(\mathcal{E}_i) = 0$ (A2))

Thus:

$$E(\hat{\beta}) = \beta + E(\frac{\sum x_i \mathcal{E}_i}{\sum x_i^2}) = \beta + \frac{\sum x_i E(\mathcal{E}_i)}{\sum x_i^2} = \beta + 0 = \beta$$

$\Rightarrow \hat{\beta}$ is an unbiased estimator of β.

$\Rightarrow \hat{\beta}$ is an unbiased estimator of β.

c) To show that $\hat{\beta}$ has the smallest variance out of all linear unbiased estimators of β

Note:

1. The OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are calculated from a **specific sample** of observations of the dependent and independent variables. If we consider a **different sample** of observations for Y and

X, we get different values for $\hat{\alpha}$ and $\hat{\beta}$. This means that the values of $\hat{\alpha}$ and $\hat{\beta}$ **may vary from one sample to another**, and hence, are **random variables**.

2. The **variance** an estimator (a random variable) $\hat{\theta}$ of θ is given by:

$$Var(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

3. The expression: $\left(\sum_{i=1}^{n} x_i\right)^2$ can be written in expanded form as:

$$\left(\sum_{i=1}^{n} x_i\right)^2 = \sum_{i=1}^{n} {x_i}^2 + \sum_{i\neq j}^{n} x_i x_j$$

This is simply the sum of squares (x_i^2) plus the sum of cross-product terms ($x_i x_j$ for $i \neq j$).

From equation (*) we have:

$$\hat{\beta} - \beta = \frac{\sum x_i \mathcal{E}_i}{\sum x_i^2}$$

The variance of $\hat{\beta}$ is thus given by:

$$\begin{aligned}
Var(\hat{\beta}) = E(\hat{\beta} - \beta)^2 &= E\left(\frac{\sum x_i \mathcal{E}_i}{\sum x_i^2}\right)^2 \\
&= \frac{1}{(\sum x_i^2)^2} E\left(\sum_{i=1}^{n} x_i^2 \mathcal{E}_i^2 + \sum_{i\neq j}^{n} x_i \mathcal{E}_i x_j \mathcal{E}_j\right) \\
&= \frac{1}{(\sum x_i^2)^2} \left(\sum_{i=1}^{n} x_i^2 E(\mathcal{E}_i^2) + \sum_{i\neq j}^{n} x_i x_j E(\mathcal{E}_i \mathcal{E}_j)\right) \\
&= \frac{1}{(\sum x_i^2)^2} \left(\sum_{i=1}^{n} x_i^2 (\sigma^2) + \sum_{i\neq j}^{n} x_i x_j (0)\right) \quad ....(**) \\
&= \frac{1}{(\sum x_i^2)^2} \left(\sigma^2 \sum_{i=1}^{n} x_i^2\right) = \frac{\sigma^2}{\sum x_i^2}
\end{aligned}$$

Note that (**) follows from A(3)and(4),that is, $var(\mathcal{E}_i) = E(\mathcal{E}_i^2) = \sigma^2$ for all i and $cov(\mathcal{E}_i, \mathcal{E}_j) = (\mathcal{E}_i \mathcal{E}_j) = 0 \ for \ i \neq j.$

Thus,

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum X_i^2}$$

We have seen above (in proof (a)) that the OLS estimator of β can be expressed as:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \sum a_i y_i$$

where $a_i = \frac{x_i}{\sum x_i^2}$.
Now let β^* be another linear unbiased estimator of β given by:

$$\beta^* = \sum c_i y_i$$

where $c_i = \frac{x_i}{\sum x_i^2} + d_i$ and d_i are arbitrary constants (real numbers).

β^* can be written as:

$$\begin{aligned} \beta^* &= \sum c_i y_i = \sum \left(\frac{x_i}{\sum x_i^2} + d_i \right) (\beta x_i + \mathcal{E}_i) \quad (since\ y_i = \beta x_i + \mathcal{E}_i) \\ &= \beta \frac{\sum x_i^2}{\sum x_i^2} + \beta \sum d_i x_i + \frac{\sum x_i \mathcal{E}_i}{\sum x_i^2} + \sum d_i \mathcal{E}_i \end{aligned}$$

Taking expectations we have:

$$\begin{aligned} E(\beta^*) &= E\left(\beta + \beta \sum d_i x_i + \frac{\sum x_i \mathcal{E}_i}{\sum x_i^2} + \sum d_i \mathcal{E}_i \right) \\ &= \beta + \beta \sum d_i x_i \quad (since\ E(x_i \mathcal{E}_i) = x_i E(\mathcal{E}_i) = 0, E(d_i \mathcal{E}_i) = d_i E(\mathcal{E}_i) = 0) \end{aligned}$$

Thus, for β^* to be unbiased (that is, for E(β^*)=β to hold) we should have:

$$\sum d_i x_i = 0 \qquad ....(***)$$

The variance of β^* is given by:

$$\begin{aligned}
Var(\beta^*) &= E(\beta^* - \beta)^2 = E(\sum c_i\mathcal{E}_i)^2 \\
&= E\left(\sum c_i^2\mathcal{E}_i^2 + \sum_{i\neq j} c_ic_j\mathcal{E}_i\mathcal{E}_j\right) = \sum c_i^2 E(\mathcal{E}_i^2) + \sum_{i\neq j} c_ic_jE(\mathcal{E}_i\mathcal{E}_j) \\
&= \sum c_i^2\sigma^2 + \sum_{i\neq j} c_ic_j(0) = \sigma^2\sum c_i^2 \\
&= \sigma^2\sum\left(\frac{x_i}{\sum x_i^2} + d_i\right)^2 = \sigma^2\sum\left(\frac{x_i^2}{\sum(x_i^2)^2} + \frac{2x_id_i}{\sum x_i^2} + d_i^2\right) \\
&= \sigma^2\frac{\sum x_i^2}{\sum(x_i^2)^2} + \sigma^2\frac{2\sum x_id_i}{\sum x_i^2} + \sigma^2\sum d_i^2 \quad (but \ \sum d_ix_i = 0 \ from(***)) \\
&= \frac{\sigma^2}{\sum x_i^2} + \sigma^2\sum d_i^2 = Var(\hat{\beta}) + \sigma^2\sum d_i^2
\end{aligned}$$

Thus, we have shown that:

$$Var(\beta^*) = Var(\hat{\beta}) + \sigma^2\sum d_i^2$$

Since $\sum d_i^2$ (which is a sum of squares of real numbers) is always greater than or equal to zero, we have:

$$Var(\beta^*) \geq Var(\hat{\beta})$$

This implies that the **variance of $\hat{\beta}$ is the smallest** as compared to the variance of any other linear unbiased estimator of β.

Hence, we conclude that $\hat{\beta}$ is the BLUE of β

2.1.3 Maximum likelihood (ML) method of estimation

Probability distribution of error terms

The OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are both linear functions of the error term, which is random by assumption. For example:

$$\hat{\beta} = \beta + \frac{\sum x_i \mathcal{E}_i}{\sum x_i^2} = \beta + \sum a_j \mathcal{E}_i$$

where $a_i = \frac{x_i}{\sum x_i^2}$ and $x_i = X_i - \bar{X}$.

Therefore, the probability distributions of the OLS estimators will depend upon the assumptions made about the probability distribution of the error term. The nature of the probability distribution of the error term is important for hypothesis testing (or for making inferences about α and β) and also for estimation purposes.

In regression analysis, it is usually assumed that the error terms follow the normal distribution with mean 0 and variance σ^2.

Since $\mathcal{E}_i = Y_i - \alpha - \beta X_i$, the probability distribution of $\mathcal{E}_i$ would be:

$$\begin{aligned} p(\mathcal{E}_i) &= \frac{1}{\sqrt{2\Pi\sigma^2}} exp\left\{-\frac{1}{2}\left(\frac{\mathcal{E}_i - E(\mathcal{E}_i)}{sd(\mathcal{E}_i)}\right)^2\right\} \\ &= \frac{1}{\sqrt{2\Pi\sigma^2}} exp\left\{-\frac{1}{2}\left(\frac{\mathcal{E}_i - (0)}{sd(\sigma)}\right)^2\right\} \\ &= \frac{1}{\sqrt{2\Pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}\left(Y_i - \alpha - \beta X_i\right)^2\right\} \text{..................................(2.1.6)} \end{aligned}$$

Here $sd(\mathcal{E}_i)$ is the standard deviation of $\mathcal{E}_i$, that is, $sd(\mathcal{E}_i) = \sqrt{Var(\mathcal{E}_i)} = \sigma$

Consider the linear model: $Y_i = \alpha + \beta X_i + \mathcal{E}_i$. Under the assumption that the error terms $\mathcal{E}_i$ follow the normal distribution with mean 0 and variance σ^2,Y_i **is also normally distributed** with:

$$Mean = E(Y_i) = E(\alpha + \beta X_i + \mathcal{E}_i) = \alpha + \beta X_i$$

$$Variance = Var(Y_i) = Var(\alpha + \beta X_i + \mathcal{E}_i) = Var(\mathcal{E}_i) = \sigma^2$$

Thus, the probability distribution of Y_i can be written as:\

$$
\begin{aligned}
p(Y_i) &= \frac{1}{\sqrt{2\Pi\sigma^2}} exp\left\{-\frac{1}{2}\left(\frac{Y_i - E(Y_i)}{sd(Y_i)}\right)^2\right\} \\
&= \frac{1}{\sqrt{2\Pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}\left(Y_i - \alpha - \beta X_i\right)^2\right\}
\end{aligned}
$$

ML estimation focuses on the fact that different populations generate different samples, and any one sample being scrutinized is more likely to have come from some population than from others.

The ML estimator of a parameter β is the value of $\hat{\beta}$ which would most likely generate the observed sample observations $Y_1,\ Y_2,\ ...,\ Y_n$. The ML estimator maximizes the likelihood function L which is the product of the individual probabilities(since $Y_1,\ Y_2,\ ...,\ Y_n$ are randomly selected implying independence) taken over all n observations given by:

$$
\begin{aligned}
L(Y_1, Y_2, ..., Y_n, \alpha, \beta, \sigma^2) &= P(Y_1)P(Y_2)...P(Y_n) \\
&= \frac{1}{(\sqrt{2\Pi\sigma^2})^n} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \alpha - \beta X_i\right)^2\right\}
\end{aligned}
$$

Our aim is to maximize this likelihood function L with respect to the parameters α,β and σ^2. To do this, it is more convenient to work with the natural logarithm of L (called **the log-likelihood function**) given by:

$$
\log L = -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \alpha - \beta X_i\right)^2
$$

Taking partial derivatives of log L with respect to α,β & σ^2 and equating to zero, we get the ML estimators.

By partial differentiation of the log L with respect to α and β and equating the results to zero we get:

$$
\begin{aligned}
\frac{\partial \log L}{\partial \beta} &= -\frac{1}{2\sigma^2}\sum\left(Y_i - \alpha - \beta X_i\right)(-X_i) = 0 \\
\frac{\partial \log L}{\partial \beta} &= -\frac{1}{2\sigma^2}\sum\left(Y_i - \alpha - \beta X_i\right)(-1) = 0
\end{aligned}
$$

Re-arranging the two equations, and replacing β by β_{ML} and α by α_{ML}, we get:

$$\beta_{ML} = \frac{n\sum X_iY_i - (\sum X_i)(\sum Y_i)}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum X_iY_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} = \hat{\beta}$$

$$\alpha_{ML} = \bar{Y} - \beta_{ML}\bar{X} = \hat{\alpha}$$

By partial differentiation of the log L with respect to σ^2 and equating to zero we get:

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2}\left(\frac{1}{\sigma^2}\right) - \frac{1}{2}\sum (Y_i - \alpha - \beta X_i)\left(\frac{-1}{(\sigma^2)^2}\right) = 0$$

Replacing σ^2 by σ^2_{ML} and simplifying, we get:

$$\sigma^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha_{ML} - \beta_{ML}X_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{E}}_i^2$$

<u>Note</u>

1) The ML estimators $\hat{\alpha}_{ML}$ and $\hat{\beta}_{ML}$ are identical to the OLS estimators, and are thus best linear unbiased estimators (BLUE) of α and β, respectively.

2) The ML estimator $\hat{\sigma}^2_{ML}$ of σ^2 is biased.

<u>Proof</u>

$$\hat{\mathcal{E}}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\alpha} - \hat{\beta}X_i)$$
$$= (\alpha + \beta X_i + \mathcal{E}_i) - (\hat{\alpha} - \hat{\beta}X_i) = (\alpha - \hat{\alpha}) + (\beta - \hat{\beta})X_i + \mathcal{E}_i \text{.....................................(2.1.7}$$

$Y_i = \alpha + \beta X_i + \mathcal{E}_i$

$$\Rightarrow \bar{Y} = \alpha + \beta\bar{X} + \bar{\mathcal{E}}$$
$$\Rightarrow \alpha = \bar{Y} - \beta\bar{X} - \bar{\mathcal{E}} \text{.....................................(2.1.8)}$$

We know that the OLS estimator of α is given by:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \text{.....................................(2.1.9)}$$

Subtracting (2.1.9) from (2.1.8) we get:

$$(\alpha-\hat{\alpha})=(\hat{\beta}-\beta)\bar{X}-\bar{\mathcal{E}}=-(\beta-\hat{\beta})\bar{X}-\bar{\mathcal{E}} \text{..................................} (2.1.20)$$

Substituting relation (2.1.20) in (2.1.7) we get:

$$\begin{aligned}\hat{\mathcal{E}}_i=-(\beta-\hat{\beta})\bar{X}-\bar{\mathcal{E}}+(\beta-\hat{\beta})X_i-\mathcal{E}_i&=(\beta-\hat{\beta})(X_i-\bar{X})+(\mathcal{E}_i-\bar{\mathcal{E}})\\ &=(\beta-\hat{\beta})x_i+e_i\end{aligned}$$

where $x_i=X_i-\bar{X}$ and $e_i=\mathcal{E}_i-\bar{\mathcal{E}}$. Squaring both sides and taking summations we have:

$$\sum_{i=1}^{n}\hat{\mathcal{E}}_i^2=(\beta-\hat{\beta})^2\sum_{i=1}^{n}x_i^2+2(\beta-\hat{\beta})\sum_{i=1}^{n}x_ie_i+\sum_{i=1}^{n}e_i^2 \text{..................................} (2.1.21)$$

From the two-variable model in deviations form we have:

$$\begin{aligned}\hat{\beta}-\beta=\frac{\sum x_i\mathcal{E}_i}{\sum x_i^2}=\frac{\sum x_i(\mathcal{E}_i-\bar{\mathcal{E}})}{\sum x_i^2}=\frac{\sum x_ie_i}{\sum x_i^2}\qquad &\left(since\ \frac{\sum x_i\bar{\mathcal{E}}}{\sum x_i^2}=\frac{\bar{\mathcal{E}}\sum x_i}{\sum x_i^2}=\frac{\bar{\mathcal{E}}(0)}{\sum x_i^2}=0\right)\\ \Rightarrow\sum x_ie_i=(\hat{\beta}-\beta)\sum x_i^2=-(\beta-\hat{\beta})\sum x_i^2 &\text{..................................} (2.1.22)\end{aligned}$$

Substituting (2.1.22) in (2.1.21) we have:

$$\sum_{i=1}^{n}\hat{\mathcal{E}}_i^2=(\beta-\hat{\beta})^2\sum_{i=1}^{n}x_i^2-2(\beta-\hat{\beta})^2\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}e_i^2=-(\beta-\hat{\beta})^2\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}e_i^2$$

$$\bullet E\left(-(\beta-\hat{\beta})^2\sum_{i=1}^{n}x_i^2\right)=-\sum_{i=1}^{n}x_i^2E\left[(\beta-\hat{\beta})^2\right]=-\sum_{i=1}^{n}x_i^2Var(\hat{\beta})=-\sum_{i=1}^{n}x_i^2\left[\frac{\sigma^2}{\sum_{i=1}^{n}x_i^2}\right]=-\sigma^2$$

$$\begin{aligned}\bullet\qquad E\left(\sum_{i=1}^{n}e_i^2\right)&=E\left(\sum_{i=1}^{2}(\mathcal{E}_i-\bar{\mathcal{E}})^2\right)=E\left(\sum_{i=1}^{2}\mathcal{E}_i^2-n\bar{\mathcal{E}}^2\right)\\ &=\sum_{i=1}^{2}E(\mathcal{E}_i^2)-nE(\bar{\mathcal{E}}^2)=n\sigma^2-n\left(\frac{\sigma^2}{n}\right)=n\sigma^2-\sigma^2=(n-1)\sigma^2\end{aligned}$$

Then it follows that:

$$E\left(\sum_{i=1}^{n}\hat{\mathcal{E}}_i^2\right)=E\left(-(\beta-\hat{\beta})^2\sum_{i=1}^{n}x_i^2\right)+E\left(\sum_{i=1}^{n}e_i^2\right)=-\sigma^2+(n-1)\sigma^2=(n-2)\sigma^2$$

Thus,

$$E(\sigma^2_{ML}) = E\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{E}}_i^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}\hat{\mathcal{E}}_i^2\right) = \left(\frac{n-2}{n}\right)\sigma^2 \neq \sigma^2$$

From the above result it follows that an unbiased estimator of σ^2 is:

$$\hat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\mathcal{E}}_i^2$$

2.1.4 Statistical inference in simple linear regression model

Estimation of standard error

To make statistical inferences about the true (population) regression coefficient β, we make use of the estimator $\hat{\beta}$ and its variance $Var(\hat{\beta})$. We have already seen that:

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2}$$

where $x_i = X_i - \bar{X}$. Since this variance depends on the unknown parameter σ^2, we have to estimate σ^2. As shown above, an unbiased estimator of σ^2 is given by:

$$\hat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\mathcal{E}}_i^2$$

Thus, an unbiased estimator of $Var(\hat{\beta})$ is given by:

$$\hat{Var}(\hat{\beta}) = \frac{\hat{\sigma^2}}{\sum x_i^2} = \frac{\sum \hat{\mathcal{E}}_i^2}{(n-2)\sum x_i^2}$$

The square root of $\hat{Var}(\hat{\beta})$ is called the standard error of $\hat{\beta}$, that is,

$$s.e.(\hat{\beta}) = \sqrt{V\hat{a}r(\hat{\beta})} = \sqrt{\frac{\hat{\sigma^2}}{\sum x_i^2}}$$

Tests of significance of regression coefficients

Consider the simple linear regression model:

$$Y_i = \alpha + \beta X_i + \mathcal{E}_i$$

If there is no relationship between X and Y, then this is equivalent to saying β= 0 (β is not significantly different from zero). Thus, the null hypothesis of no relationship between X and Y is expressed as:

H:β=0

The alternative hypothesis is that there is a significant relationship between X and Y, that is,

H:$\beta \neq 0$

In order to reject or not reject the null hypothesis, we calculate the **test statistic** given by:

$$t = \frac{\hat{\beta} - \beta_0}{s.e.(\hat{\beta})} = \frac{\hat{\beta} - 0}{s.e.(\hat{\beta})} = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

and compare this figure with the value from the student's t distribution with (n-2) degrees of freedom for a given significance level α.

Decision rule: If $|t| > t_{\frac{\alpha}{2}}(n-2)$ then we **reject the null hypothesis**, and conclude that there is a significant relationship between X and Y.

Confidence interval for β

Confidence interval provides a range of values which are likely to contain the true regression parameter. With every confidence interval, we associate a level of statistical significance (α). The confidence intervals are constructed in such a way that the probability of the interval to contain the true parameter is (1 -α). Symbolically,

$$P[-t_{\frac{\alpha}{2}(n-2)} < t < t_{\frac{\alpha}{2}(n-2)}] = 1-\alpha$$
$$\Rightarrow P[-t_{\frac{\alpha}{2}(n-2)} < \frac{\hat{\beta}-\beta_0}{s.e(\hat{\beta})} < t_{\frac{\alpha}{2}(n-2)}]$$
$$\Rightarrow P[\hat{\beta} - t_{\frac{\alpha}{2}(n-2)} * s.e(\hat{\beta}) < \beta_0 < \hat{\beta} + t_{\frac{\alpha}{2}(n-2)} * s.e(\hat{\beta})] = 1-\alpha$$

Thus, a$(1-\alpha)100\%$ confidence interval for β is given by:

$$\hat{\beta} \pm t_{\frac{\alpha}{2}(n-2)} * s.e(\hat{\beta})$$

** Test of model adequacy**

Is the estimated equation a useful one? To answer this, an objective measure of some sort is desirable. The **total variation** in the dependent variable Y is given by:

$$Variation(Y) = \sum (Y_i - \bar{Y})^2$$

Our goal is to partition this variation into two: one that accounts for variation due to the regression equation (explained portion) and another that is associated with the unexplained portion of the model. We can write $Y_i - \bar{Y}$ as:

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

Squaring both sides and taking summations we have:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}).....(*)$$

Remark 1:

$\sum_{i=1}^{n} \hat{\mathcal{E}}_i = 0$ where $\hat{\mathcal{E}}_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$

Proof:

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i) &= \sum Y_i - \sum \hat{\alpha} - \hat{\beta}\sum Y_i \\ &= n\bar{Y} - n\hat{\alpha} - n\hat{\beta}\bar{X} \\ &= n[(\bar{Y} - \hat{\beta}) - \hat{\alpha}] = n(\hat{\alpha} - \hat{\alpha})\end{aligned}$$

Remark 2:

$\sum_{i=1}^{n} \hat{\mathcal{E}}_i X_i = 0$

Proof:

$$\begin{aligned}\sum_{i=1}^{n} \hat{\mathcal{E}}_i X_i &= \sum (Y_i - \hat{\alpha} - \hat{\beta}X_i)X_i \\ &= \sum Y_i X_i - \sum \hat{\alpha} X_i - \sum \hat{\beta} X_i^2 \\ &= \sum Y_i X_i - \sum \hat{\alpha} X_i - \sum \hat{\beta} X_i^2 \\ &= \sum Y_i X_i - n\hat{\alpha}\bar{X} - \sum \hat{\beta} X_i^2 \\ &= \sum Y_i X_i - n(\bar{Y} - \hat{\beta}\bar{X})\bar{X} - \sum \hat{\beta} X_i^2 \\ &= \sum Y_i X_i - n\bar{X}\bar{Y} - \hat{\beta}[\sum X_i^2 - n\bar{X}^2] \\ &= \sum Y_i X_i - n\bar{X}\bar{Y} - \frac{\sum Y_i X_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}[\sum X_i^2 - n\bar{X}^2] = 0\end{aligned}$$

From remarks 1 and 2 it follows that:

$$\begin{aligned}\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= \sum \hat{\mathcal{E}}_i \hat{Y}_i - \sum \hat{\mathcal{E}}_i \bar{Y} \\ &= \sum \hat{\mathcal{E}}_i(\hat{\alpha} + \hat{\beta}X_i) - \bar{Y}\underbrace{\sum \hat{\mathcal{E}}_i} = \hat{\alpha}\underbrace{\sum \hat{\mathcal{E}}_i} + \hat{\beta}\underbrace{\sum \hat{\mathcal{E}}_i X_i} = 0\end{aligned}$$

Thus, the cross product term in equation (*) vanishes, and we are left with:

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{n}(\hat{Y}-\bar{Y})^2$$

$$Variation\ in\ Y \qquad Residualvariation \qquad Explainedvariation$$

$$TSS \qquad = \qquad ESS \qquad + \qquad RSS$$

In other words, the total sum of squares (TSS) is decomposed into regression (explained) sum of squares (RSS) and error (residual or unexplained) sum of squares (ESS).

Computational formulas:

•The total sum of squares (TSS) is a measure of dispersion of the observed values of Y about their mean. This is computed as:

$$TSS=\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}y_i$$

•The regression (explained) sum of squares (RSS) measures the amount of the total variability in the observed values of Y that is accounted for by the linear relationship between the observed values of X and Y. This is computed as:

$$RSS=\sum(\hat{Y}_i-\bar{Y})^2=\hat{\beta}^2[\sum_{i=1}^{n}(X_i-\bar{X})^2]=\hat{\beta}^2\sum_{i=1}^{n}x_i^2$$

•The error (residual or unexplained) sum of squares (ESS) is a measure of the dispersion of the observed values of Y about the regression line. This is computed as:

$$ESS=\sum(Y_i-\hat{Y})^2=TSS-RSS$$

If a regression equation does a good job of describing the relationship between two variables, the explained sum of squares should constitute a large proportion of the total sum of squares. Thus, it would be of interest to determine the magnitude of this proportion by computing the ratio of the explained sum of squares to the total sum of squares. This proportion is called the sample **coefficient of determination** R^2. That is:

Coefficient of determination $=R^2=\dfrac{RSS}{TSS}=1-\dfrac{ESS}{TSS}$

The coefficient of determination can also be computed as:

$$R^2 = \frac{\hat{\beta}\sum x_i y_i}{\sum y_i^2} \quad where\ xi = Xi - \bar{X}\ and\ yi = Y_i - \bar{Y}$$

Tests for the coefficient of determination (R^2)

The largest value that R^2 can assume is 1 (in which case all observations fall on the regression line), and the smallest it can assume is zero. A low value of R^2 is an indication that:

• X is a poor explanatory variable in the sense that variation in X leaves Y unaffected, or

• while X is a relevant variable, its influence on Y is weak as compared to some other variables that are omitted from the regression equation, or

• the regression equation is misspecified (for example, an exponential relationship might be more appropriate.

Note

1) The proportion of total variation in the dependent variable (Y) that is explained by changes in the independent variable (X) or by the regression line is equal to:$R^2 * 100\%$.

2) The proportion of total variation in the dependent variable (Y) that is due to factors other than X (for example, due to excluded variables, chance, etc) is equal to:$(1 - R^2)100\%$

Thus, a small value of R^2 casts doubt about the usefulness of the regression equation. We do not, however, pass final judgment on the equation until it has been subjected to an objective statistical test. Such a test is accomplished by means of analysis of variance (ANOVA) which enables us to test the significance of R^2 (i.e., the adequacy of the linear regression model).

The ANOVA table for simple linear regression is given below:

ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio
Regression	RSS	1	RSS/1	$F_{cal} = \frac{RSS/1}{ESS/(n-2)}$
Residual	ESS	n-2	$\frac{ESS}{(n-2)}$	
Total	TSS	n-1		

To test for the significance of R^2, we compare the variance ratio with the critical value from the F distribution with 1 and (n-2) degrees of freedom in the numerator and denominator, respectively, for a given significance level α. **Decision:** If the calculated variance ratio exceeds the tabulated value, that is, if $F_{cal} > F_{\alpha(1,n-2)}$, we then conclude that R^2 is significant (or that the linear regression model is adequate).

Note: The F test is designed to test the significance of all variables or a set of variables in a regression model. In the two-variable model, however, it is used to test the explanatory power of a single variable (X), and at the same time, is equivalent to the test of significance of R^2.

Example

Consider the following data on the percentage rate of change in electricity consumption (millions KWH) (Y) and the rate of change in the price of electricity (Birr/KWH) (X) for the years 1979 – 1994. Use **chp1.csv**.

```
#electricity consumption
ele_cons=read.csv("ele_cons.csv",header = T,sep = ",")
(n<-length(ele_cons))
```

```
## [1] 3
```

```
x_i<-ele_cons$X-mean(ele_cons$X)
y_i<-ele_cons$Y-mean(ele_cons$Y)
(x_bar<-mean(ele_cons$X))
```

```
## [1] 1.280625
```

```
(y_bar<-mean(ele_cons$Y))
```

```
## [1] 23.42687
```

```
(sum_x_i_sq<-sum(x_i^2)) # sum of square of deviation of x
```

```
## [1] 92.20109
```

```
(sum_y_i_sq<-sum(y_i^2)) # sum of square of deviation of y
```

```
## [1] 13228.7
```

```
(sum_xy<-sum(x_i*y_i)) #Sum of cross product of x*y
```

```
## [1] -779.2348
```

```r
(b_hat<-sum_xy/sum_x_i_sq) # Slop using deviation form
```

```
## [1] -8.45147
```

```r
(a_hat<-y_bar-b_hat*x_bar)
```

```
## [1] 34.25004
```

Therefore, the estimated regression equation is:

$\hat{Y} = \hat{\alpha} + \hat{\beta}X \Leftrightarrow \hat{Y} = 34.25 + -8.451X$ **Test of model adequacy:**

$TSS = \sum y_i^2$

```r
tss=sum(y_i^2)
```

$$TSS = \sum y_i^2 = 1.32287 \times 10^4$$

$$RSS = \hat{\beta}^2 \sum x_i^2$$

```r
(rss<-b_hat^2*sum(x_i^2)) # Regression Sum of Square
```

```
## [1] 6585.679
```

```r
(ess<-tss-rss) # Error sum of square
```

```
## [1] 6643.016
```

```r
(r_sq<-rss/tss) # Coeffient of determination
```

```
## [1] 0.4978329
```

```r
# Mean square Error(EMS)=ESS/(n-k)
ems<-ess/(nrow(ele_cons)-2)
# Mean square of Regression(RMS)=RSS/(k-1)
rms<-rss/1
(f_cal=rms/ems)
```

```
## [1] 13.87916
```

$$RSS = \hat{\beta}^2 \sum x_i^2 = 6585.68$$

$$\Rightarrow R^2 = \frac{RSS}{TSS} = 6585.68/1.32287 \times 10^4 = 0.4978$$

Thus, we can conclude that:

About 50% of the variation in electricity consumption is due to changes in the price of electricity. The remaining 50% of the variation in electricity consumption is not due to changes in the price of electricity, but instead due to chance and other factors not included in the model.

ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio
Regression	6585.68	1	6585.68	13.8791637
Residual	6643.02	16-2=14	474.5	
Total	1.32287×10^4	16-1=15		

For $\alpha = 0.05$, the critical value from the F-distribution is:

$$F_{\alpha(1,n-2)} = F_{0.05(1,14)}$$

```
(qf(p=0.05,df1=1,df2=14,lower.tail = F)) # Tabulated F-value
```

```
## [1] 4.60011
```

```
 #Or p-value
(pf(f_cal,df1=1,df2=14,lower.tail = F)) # T
```

```
## [1] 0.002260424
```

Decision: Since the calculated variance ratio exceeds the critical value, we reject the null hypothesis of no linear relationship between price and consumption of electricity at the 5% level of significance. Thus, we then conclude that $R^2$is significant, that is, the linear regression model is adequate and is useful for prediction purposes.

Estimation of the standard error of the coeffiecents and test of its significance

An unbiased estimator of the error variance

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{16}\hat{\mathcal{E}}_i^2 = \frac{1}{n-2}\sum_{i=1}^{16}(Y_i - \hat{Y})^2 = \frac{ESS}{n-2}$$

```
y_hat<-a_hat+b_hat*ele_cons$X # Fitted values
e_hat<-ele_cons$Y-y_hat # Residuals
(sigma_hat_sq<-sum(e_hat^2)/(nrow(ele_cons)-2))
```

```
## [1] 474.5012
```

```
# Or
(sigma_hat_sq<-ess/(nrow(ele_cons)-2))
```

```
## [1] 474.5012
```

Thus, an unbiased estimator of $Var(\hat{\beta})$ is given by

$$\hat{Var(\hat{\beta})} = \frac{\hat{\sigma}^2}{\sum x_i^2}$$

```
(var_b_hat<-sigma_hat_sq/sum_x_i_sq)
```

```
## [1] 5.146372
```

$$\hat{Var(\hat{\beta})} = \frac{\hat{\sigma}^2}{\sum x_i^2} = 474.5011535/92.2010938 = 5.1463723$$

```
s.e_b_hat<-sqrt(var_b_hat)
```

The standard error of $\hat{\beta}$ is:

$$s.e(\hat{\beta}) = \sqrt{\hat{Var(\hat{\beta})}} = \sqrt{5.146} = 2.2686$$

The hypothesis of interest is: $H_0 : \beta = 0$ VS $H_1 : \beta \neq 0$ We calculate the test statistic:

```r
(t_cal<-(b_hat-0)/s.e_b_hat)
```

```
## [1] -3.725475
```

For $\alpha = 0.05$, the critical value from the student's $tdistribution$ with (n-2) degrees of freedom is:

```r
(qt(p=0.05/2,df=nrow(ele_cons)-2,lower.tail = F))
```

```
## [1] 2.144787
```

```r
#Or using p-value
 (pt(t_cal,df=nrow(ele_cons)-2,lower.tail = T))
```

```
## [1] 0.001130212
```

Decision: Since $|t| > t_{\frac{\alpha}{2}(n-2)}$, we reject the null hypothesis, and conclude that β is significantly different from zero. In other words, the price of electricity significantly and negatively affects electricity consumption.

The interpretation of the estimated regression coefficient $\hat{\beta}$=-8.451 is that for a one percent drop (increase) in the growth rate of price of electricity, there is an 8.45 percent increase (decrease) in the growth rate of electricity consumption.

2.2 Multiple Linear Regressions

2.2.1 Introduction

So far we have seen the basic statistical tools and procedures for analyzing relationships between two variables. But in practice, economic models generally contain one dependent variable and two or more independent variables. Such models are called **multiple regression models**.

Economic relationships usually include more than one regressor. For example, a demand equation for a product will usually include real price of that product($\hat{Y}_i$) in addition to real income as well as real price of a competitive product and the advertising expenditures on the product. In this case the sales of the product is modelled as

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$$

where Y_i denotes the i^{th} observation on the dependent variable Y, in this case the sales of this product. own price(X_1), the competitor's price(X_2) and advertising expenditures(X_3)

Example: a) In demand studies we study the relationship between the demand for a good (Y) and price of the good (X_2), prices of substitute goods (X_3) and the consumer's income (X_4). Here, Y is the dependent variable and X_2, X_3 and X_4 are the explanatory (independent) variables. The relationship is estimated by a multiple linear regression equation (model) of the form:

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

where $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$ are estimated regression coefficients.

b) In a study of the amount of output (product), we are interested to establish a relationship between output (Q) and labour input (L) '&' capital input (K). The equations are often estimated in log-linear form as:

$$\log(\hat{Q}) = \hat{\beta}_1 + \hat{\beta}_2 \log(L) + \hat{\beta}_3 \log(K)$$

c) In a study of the determinants of the number of children born per woman (Y), the possible explanatory variables include years of schooling of the woman (X_2), woman's (or husband's) earning at marriage (X_3), age of woman at marriage (X_4) and survival probability of children at age five (X_5). The relationship can thus be expressed as:

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5$$

Model assumptions

Dependent variable: Y of size nx1

Independent (explanatory) variables: $X_2, X_3, ..., X_k$ each of size nx1

Assumptions

1. The true model is: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \mathcal{E}_i$.
2. The error terms have **zero mean**: $E(\mathcal{E}_i)$.
3. **Homoscedasticity:** $var(\mathcal{E}_i) = E(\mathcal{E}_i^2) = \sigma^2$ for all i.
4. **No error autocorrelation:** $cov(\mathcal{E}_i, \mathcal{E}_j) = E(\mathcal{E}_i \mathcal{E}_j) = 0$ for $i \neq j$.

5. Each of the explanatory variables $X_2, X_3, ..., X_k$ is **non-stochastic**.
6. **No multicollinearity:** No exact linear relationship exists between any of the explanatory variables.
7. **Normality:** $\mathcal{E}_i$ are normally distributed with mean zero and variance σ^2 for all i ($\mathcal{E}_i \sim N(0, \sigma^2)$).

The only additional assumption here is that there is no multicollinearity, meaning that there is no linear dependence between the regress or variables $X_2, X_3, ..., X_k$.

Under the above assumptions, ordinary least squares (OLS) yields best linear unbiased estimators(BLUE) of $\beta_1, \beta_2, ..., \beta_k$.

2.2.2 Estimation of parameters and standard errors

Example: Consider the following model (K = 3):

$$
\begin{aligned}
Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \mathcal{E}_i ....(A1) \\
\frac{\sum Y_i}{n} = \beta_1 + \beta_2 \frac{\sum X_{2i}}{n} + \frac{\mathcal{E}_i}{n} + \beta_3 \frac{\sum X_{3i}}{n} \\
\bar{y} = \beta_1 + \beta_2 \bar{x}_{2i} + \beta_3 \bar{x}_{3i} + \bar{\mathcal{E}} ....(A2) \\
Equation\ (A1) - (A2) \\
Y_i - \bar{y} = \beta_1 - \beta_1 + \beta_2 (X_{2i} - \bar{X_2}) + \beta_3 (X_{3i} - \bar{X_3}) + (\mathcal{E}_i - \bar{\mathcal{E}}) \\
Hence\ the\ model\ in\ deviation\ form: \\
y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + e_i \ \ where\ e_i = \mathcal{E}_i - \bar{\mathcal{E}}
\end{aligned}
$$

OLS estimator of β_2 *and* β_3 is an estimator which minimizes the sum square of error(ESS) in the deviation form as follows:

$$ESS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_2 x_{2i} - \beta_3 x_{3i})^2$$

The error sum of squares (ESS) is:

$$ESS = \sum \mathcal{E}_i^2 = \sum (y_i - \beta_2 X_{2i} - \beta_3 X_{3i})^2$$

Partially differentiating the ESS with respect to β_2 and β_3, equating to zero and simplifying, we get:

$$\frac{\partial ESS}{\partial \beta_2} = 0 \Rightarrow 2\sum (y_i - \beta_2 x_{2i} - \beta_3 x_{3i})(-x_{2i}) = 0$$

$$\Rightarrow \sum y_i x_{2i} - \beta_2 \sum x_{2i}^2 - \beta_3 \sum x_{3i} x_{2i} = 0$$

$$\sum y_i x_{2i} = \beta_2 \sum x_{2i}^2 + \beta_3 \sum x_{3i} x_{2i} \ldots\ldots\ldots\ldots\ldots\ldots.(B1))$$

$$\frac{\partial ESS}{\partial \beta_3} = 0 \Rightarrow 2\sum (y_i - \beta_2 x_{2i} - \beta_3 x_{3i})(-x_{3i}) = 0$$

$$\Rightarrow \sum y_i x_{3i} - \beta_2 \sum x_{2i} x_{3i} - \beta_3 \sum x_{3i}^2 = 0$$

$$\sum y_i x_{3i} = \beta_2 \sum x_{2i} x_{3i} + \beta_3 \sum x_{3i}^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots.(B2))$$

$$Using\ NE\ (*): \quad \beta_2 = \frac{\sum y_i x_{2i} - \beta_3 \sum x_{3i} x_{2i}}{\sum x_{2i}^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(B3)$$

$$Substituting\ value\ of\ \beta_2\ in\ normal\ equation (B2):$$

$$\sum y_i x_{3i} = \left(\frac{\sum y_i x_{2i} - \beta_3 \sum x_{3i} x_{2i}}{\sum x_{2i}^2}\right) \sum x_{2i} x_{3i} + \beta_3 \sum x_{3i}^2$$

$$\Rightarrow \hat{\beta}_3 = \frac{[\sum x_{3i} y_i][\sum x_{2i}^2] - [\sum x_{2i} y_i][\sum x_{2i} x_{3i}]}{[\sum x_{2i}^2][\sum x_{3i}^2] - [\sum x_{2i} x_{3i}]^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots.(B4)$$

$$Substituting\ value\ of\ \hat{\beta}_3\ in\ equation\ (B3)$$

$$\hat{\beta}_2 = \frac{[\sum x_{2i} y_i][\sum x_{3i}^2] - [\sum x_{3i} y_i][\sum x_{2i} x_{3i}]}{[\sum x_{2i}^2][\sum x_{3i}^2] - [\sum x_{2i} x_{3i}]^2}$$

An estimator of β_1 is also obtained from the non-deviation form:

$$\frac{\partial ESS}{\partial \beta_1} = 0 \Longrightarrow \frac{\partial \sum \mathcal{E}_i^2}{\partial \beta_1} = 2\sum(Y_i - \beta_1 - \beta_2 X_{2i} - \beta_3 X_{3i})(-1) = 0$$

$$\sum Y_i = n\beta_1 + \beta_2 \sum X_{2i} + \beta_3 \sum X_{3i}$$

$$Dividing\ by\ n\ both\ sides:$$

$$\Rightarrow \hat{\beta} = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$

2.2.2.1 Variances of estimated regression coefficients

- First we have to determine the **unbiased estimator of the variance of the errors** σ^2 given by:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\mathcal{E}_i}^2}{n-k} = \frac{\sum_{i=1}^n \hat{\mathcal{E}_i}^2}{n-3} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-3}$$

where $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$

- $\hat{\beta}_2$,the OLS estimator of β_2, the **partial derivative** of Y_i with respect to X_{2i}
- we can interpret $\hat{\beta}_2$ as a simple linear regression coefficient.

Steps for

1. Run the regression of X2 **on all the other X's** , and obtain the residuals $\hat{v}_2$,i.e., $X_{2i} = \hat{X}_{2i} + \hat{v}_2$
2. Run the simple regression of Y on $\hat{v}_2$, the resulting estimate of the slope coefficient is $\hat{\beta}_2$

- The first regression essentially **cleans out the effect of the other X's from X2**, leaving the variation **unique to X2** in $\hat{v}_2$
- Then $var(\hat{\beta}_2) = \sigma^2 / \sum_{i=1}^n \hat{v}_{2i}^2$
- Let R_2^2 be the R^2 for the **regression of X2 on all the other X's**, then $R_2^2 = 1 - \frac{ESS}{TSS} = 1 - \frac{\sum_{i=1}^n \hat{v}_{2i}^2}{\sum_{i=1}^n x_{2i}^2} \Rightarrow \sum_{i=1}^n \hat{v}_{2i}^2 = \sum_{i=1}^n x_{2i}^2(1 - R_2^2)$
- $Var(\hat{\beta}_2) = \dfrac{\sigma^2}{\sum_{i=1}^n \hat{v}_{2i}^2} = \dfrac{\sigma^2}{\sum_{i=1}^n x_{2i}^2(1 - R_2^2)}$

Only when k=3

- The variances of estimated regression coefficients $\hat{\beta}_2$ and $\hat{\beta}_3$ are estimated, respectively, as:

$$\hat{V}(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{(1-r_{23}^2)\sum X_{2i}^2}$$
$$and \quad \hat{V}(\hat{\beta}_3) = \frac{\hat{\sigma}^2}{(1-r_{23}^2)\sum X_{3i}^2}$$

where r_{23} is the coefficient of correlation between X_2 and X_3, that is:

$$r_{23} = \frac{\sum x_{2i}x_{3i}}{\sqrt{(\sum x_{2i}^2(\sum x_{3i}^2))}}$$

Taking the square roots, we obtain the standard errors of $\hat{\beta}_2$ and $\hat{\beta}_3$: s.e.($\hat{\beta}_2$)=$\sqrt{\hat{V}(\hat{\beta}_2)}$, s.e.($\hat{\beta}_3$)=$\sqrt{\hat{V}(\hat{\beta}_3)}$.

2.2.3 The coefficient of determination and test of model adequacy

The coefficient of determination (\$ R^2 \$) can be calculated as usual as:

$$R^2 = \frac{RSS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{\mathcal{E}}_i^2}{\sum_{i=1}^{n}Y_i^2}$$

R^2 measures the proportion of variation in the dependent variable Y that is explained by the explanatory variables (or by the multiple linear regression model). It is a **goodness-of-fit statistic**. A test for the significance of R^2 or a test of model adequacy is accomplished by testing the hypotheses:

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_A : H_0 \text{ is not true}$$

The test statistic is given by:

$$F_{cal} = \frac{RSS/(k-1)}{TSS/(n-k)} = \frac{RSS/(3-1)}{TSS/(n-3)}$$

where K is the number of parameters estimated from the sample data (K = 3 in our case since we estimate β_1 ,β_2 and β_3) and n is the sample size. We say that the linear model is adequate in explaining the relationship between the dependent variable and one or more of the independent variables if:

$$F_{cal} = F_\alpha(k-1, n-k)$$

- Since **OLS minimizes** the residual sums of squares, **adding one or more variables to the regression cannot increase** this residual sums of squares
- Makes $\sum \hat{\mathcal{E}}^2$ non-increasing and R^2 non-decreasing, since $R^2 = 1 - \dfrac{\sum \hat{\mathcal{E}}^2}{\sum y_i^2}$
- $\bar{R}^2 = 1 - [\frac{\sum \hat{\mathcal{E}}^2/(n-k)}{\sum y_i^2/(n-1)}]$ adjusted by their degrees of freedom
- this variable will increase $\bar{R}^2$ **only if the reduction** in $\sum \hat{\mathcal{E}}^2$ **outweighs this loss**
-

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

Unlike R^2, $\bar{R}^2$ may increase or decrease when new variables are added into the model.

2.2.4 Tests on the regression coefficients

To test whether each of the coefficients are significant or not, the null and alternative hypotheses are given by:

$$H_0 : \beta_j = 0$$
$$H_A : \beta_j \neq 0$$

for j = 2, 3. The test statistic is:\

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}, j2, 3$$

Decision rule:

If $|t_j| > t_{\alpha/2}(n-3)$, we reject H_0 and conclude that β_j is significant, that is, the regress or variable X_j, j=2, 3 significantly affects the dependent variable Y.

Example

Consider the following data on per capita food consumption (Y), price of food (X_2) and per capita income (X_3) for the years 1927-1941 in the United States. Retail price of food and per capita disposable income are deflated by the Consumer Price Index and the data is stored as `food_cons.csv`. Fit a multiple linear regression model:

```
food_cons<-read.csv("food_cons.csv",header = T,sep=",")
```

Table 2: Food Consumption

Year	Y	X2	X3
1927	88.9	91.7	57.7
1928	88.9	92.0	59.3
1929	89.1	93.1	62.0
1930	88.7	90.9	56.3
1931	88.0	82.3	52.7
1932	85.9	76.3	44.4

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \mathcal{E}_i, i = 1, 2, ..., 15$$

To simplify the calculations, it is better to work with deviations: $y_i = Y_i + \bar{Y}$, $x_{2i} = X_{2i} + \bar{X}_2$ and $x_{3i} = X_{3i} + \bar{X}_3$.

```
#Writing in Deviation form(yi,x2i,x3i)
food_cons$x2i=food_cons$X2-mean(food_cons$X2)
food_cons$x3i=food_cons$X3-mean(food_cons$X3)
food_cons$yi=food_cons$Y-mean(food_cons$Y)
```

Table 3: In deviation form

x2i	x3i	yi
5.8	1.1733333	-0.0066667

x2i	x3i	yi
6.1	2.7733333	-0.0066667
7.2	5.4733333	0.1933333
5.0	-0.2266667	-0.2066667
-3.6	-3.8266667	-0.9066667
-9.6	-12.1266667	-3.0066667

The estimated model is:

```r
####################################
# Total
fit=lm(Y~X2+X3,data=food_cons)
#summary(fit)$coef
```

Table 4: Estiamted Coefficents

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	86.0831835	3.8730420	22.226246	0.0000000
X2	-0.2159581	0.0519701	-4.155432	0.0013336
X3	0.3781274	0.0338264	11.178463	0.0000001

$$\hat{Y} = 86.0832 \ - 0.216X_2 \ + \ 0.3781X_3$$

```r
## Var(b3_hat)
rsdl<-resid(fit)
(sigma_hat<-sum(rsdl^2)/(nrow(food_cons)-3)) #k=3
```

```
## [1] 0.7139392
```

The estimated errors (residuals) are:

$$\hat{\mathcal{E}}_i = y_i - \hat{Y} = y_i - 86.0832 - -0.216X_2 \ - \ 0.3781X_3$$

The error sum of squares (ESS) $= \sum \hat{\mathcal{E}}_i^2 = 8.5672707$

An estimator of the error variance σ^2 is:

$$\hat{\sigma^2} = \frac{\sum \hat{\mathcal{E}_i}^2}{n-3} = \frac{8.5672707}{12} = 0.7139392$$

The coefficient of determination is:

```
summary(fit)$r.squared
```

```
## [1] 0.9142667
```

$$R^2 = 0.914$$

1. R^2= 0.914 indicates that 91.4% of the variation (change) in food consumption is attributed to the effect of food price and consumer income.
2. $1 - R^2$= 0.086. This indicates that 8.6% of the variation in food consumption is due to factors (variables) not included in our specification.

Tests of model adequacy

A test of model adequacy is accomplished by testing the null hypothesis:

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_A : H_0 \text{ is not true}$$

The test statistic for this test is given by:

$$F_{cal} = \frac{RSS/(k-1)}{ESS/(n-k)} = \frac{91.362/(3-1)}{8.567271/(15-3)} = 63.98448$$

We compare this F-ratio with $F_\alpha(K-1, n-K) = F_\alpha(2, 12)$ for some significance level α.

- For α=0.01, $F_\alpha(K-1, n-K) = F_{0.01}(2, 12)$=6.93
- For α= 0.05, $F_\alpha(K-1, n-K) = F_{0.05}(2, 12)$=3.89

Since the test statistic is greater than both tabulated values, the above ratio is significant at the conventional levels of significance (1% and 5%). Thus, we reject the null hypothesis and conclude that the model is adequate, that is, variation (change) in per capita food consumption is **significantly attributed to** the effect of food price and/or per capita disposable income.

Estimation of standard errors of estimated coefficients

The standard errors of estimated regression coefficients for $\hat{\beta}_2$ is estimated as:

First regress X2 on the others X's, and obtain the correlation

```
############################
fit_x2<-lm(X2~X3,data = food_cons)
#The R_s for regressing X2 on the other X's is given by:
#R2=1-ESS/TSS
#k=2
# Residual for regressing X3 on other X's(, that is in this case X2)
v2<-resid(fit_x2)
#R2=1-(ESS/TSS)
(R_sq_2<-1-(sum(v2^2)/sum(food_cons$x2i^2)))
```

```
## [1] 0.2556877
```

```
#var_b2=sigma_hat/()
var_b2=sigma_hat/sum(v2^2)
(s.e_b2=sqrt(var_b2))
```

```
## [1] 0.05197006
```

```
#OR sum(v2^2)=ESS(1-R_sq)
var_b2=sigma_hat/(sum(food_cons$x2i^2)*(1-R_sq_2))
(s.e_b2=sqrt(var_b2))
```

```
## [1] 0.05197006
```

$$s.e.(\hat{\beta}_2) = 0.0519701$$

The standard errors of estimated regression coefficients for $\hat{\beta}_3$ is estimated as:

First regress X_3 on the others X's, and obtain the correlation

```
fit_x3<-lm(X3~X2,data = food_cons)
#The R_s for regressing X3 on the other X's is given by:
  #R2=1-ESS/TSS
  #k=2
  # Residual for regressing X3 on other X's(, that is in this case X2)
  v3<-resid(fit_x3)
#R2=1-(ESS/TSS)
R_sq_3<-1-(sum(v3^2)/sum(food_cons$x3i^2))
#var_b3=sigma_hat/()
var_b3=sigma_hat/sum(v3^2)
(s.e_b3=sqrt(var_b3))
```

```
## [1] 0.03382642
```

```
#OR sum(v3^2)=ESS(1-R_sq)
var_b3=sigma_hat/(sum(food_cons$x3i^2)*(1-R_sq_3))
(s.e_b3=sqrt(var_b3))
```

```
## [1] 0.03382642
```

$$s.e.(\hat{\beta}_3) = 0.0338264$$

Tests of significance of regression coefficients

a) Does food price significantly affect per capita food consumption?

   The hypothesis to be tested is:

$$H_0 : \beta_2 = 0$$
$$H_A = \beta_2 \neq 0$$

The test statistic is calculated as:

$$t_2 = \frac{\hat{\beta}_2}{s.e.(\hat{\beta}_2)} = \frac{-0.21596}{0.05197} = -4.155$$

For significance level $\alpha = 0.01$ and degrees of freedom $(n-3) = (15-3) = 12$, the value from the student' ' 's t-distribution is:

$$t_\alpha/2(n-3) = t_{0.005}(12) = 3.055$$

Decision: Since $|t_2| = 4.155 > 3.055$, we reject the null hypothesis and conclude that food price significantly affects per capita food consumption at the 1% level of significance.

b) Does disposable income significantly affect per capita food consumption? The hypothesis to be tested is:

$$H_0 : \beta_3 = 0$$
$$H_A = \beta_3 \neq 0$$

The test statistic is calculated as:

$$t_3 = \frac{\hat{\beta}_3}{s.e.(\hat{\beta}_3)} = \frac{0.378127}{0.033826} = 11.179$$

The 1% critical value from the student's t-distribution is again 3.055.

Decision: Since $|t_3| = 11.179 > 3.055$,we reject the null hypothesis and conclude that disposable income significantly affects per capita food consumption at the 1% level of significance.

Generally we have the following:

- Food price significantly and **negatively** affects per capita food consumption, while disposable income significantly and **positively** affects per capita food consumption.
- The estimated coefficient of food price is -0.21596. Holding disposable income constant, a one dollar increase in food price results in a 0.216 dollar decrease in per capita food consumption.
- The estimated coefficient of food price is 0.378127. Holding food price constant, a one dollar

increase in disposable income results in a 0.378 dollar increase in per capita food consumption.

As can be seen from `Table 4`, the p-values for price and income are both less than 0.01. Thus, we can conclude that both variables significantly affect consumption at the 1% level of significance. From the signs of the estimated regression coefficients we can see that the direction of influence is opposite: price affects consumption negatively while income affects consumption positively. The constant term (intercept) is also significant.

Note: In general, if the p-value > 0.05, then we doubt the importance of the variable!

2.2.5 Matrix form of the multiple linear regression model

Consider the model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \mathcal{E}_i$$

Since we have n observations, we can write the model for each observed value as:

$$\begin{aligned}
Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + ... + \beta_k X_{k1} + \mathcal{E}_1 \\
Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + ... + \beta_k X_{k2} + \mathcal{E}_2 \\
&. \\
&. \\
&. \\
Y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + ... + \beta_k X_{kn} + \mathcal{E}_n
\end{aligned}$$

The matrix form of the above model is:

$$Y = X\beta + \mathcal{E}$$

where

$$Y_{(n\times1)} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix}, \quad X_{(n\times k)} = \begin{bmatrix} 1 & X_{21} & . & . & . & X_{k1} \\ 1 & X_{22} & . & . & . & X_{k2} \\ . & . & & & & . \\ . & . & . & . & . & . \\ . & . & & & & . \\ 1 & X_{2n} & . & . & . & X_{kn} \end{bmatrix}, \quad \beta_{(k\times1)} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_n \end{bmatrix}, \quad and \quad \mathcal{E}_{(n\times1)} = \begin{bmatrix} \mathcal{E}_1 \\ \mathcal{E}_2 \\ . \\ . \\ . \\ \mathcal{E}_n \end{bmatrix}$$

Note:

a) One of the assumptions of the classical linear regression model is that there is no multicollinearity, meaning that there is no linear dependence between the regress or variables $X_2, X_3, ..., X_k$. This is the same as saying that the matrix X has **full columnrank**. Since X has K columns, we write this as:

Rank (X) = K

b) If a matrix X is of full rank K, then $X'X$ is also of full rank, i.e., Rank $(X'X)$ = K. In such cases the inverse of $X'X$ exists. Otherwise (that is, if Rank ($X'X$) < K), then $X'X$ is said to be **singular**(and its inverse does not exist).

The **mean** of the error vector $\mathcal{E}$ is:

$$E(\mathcal{E}) = \begin{bmatrix} E(\mathcal{E}_1) \\ E(\mathcal{E}_2) \\ . \\ . \\ . \\ E(\mathcal{E}_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ . \\ 0 \end{bmatrix} = 0$$

Definition :

If Y is an (n x 1) random vector, then the variance-covariance matrix of Y is given by:

$$\sum_{Y} = E\left\{[Y - E(Y)][Y - E(Y)]'\right\}$$

The variance-covariance matrix of $\hat{\beta}$ is:

$$\begin{aligned}
var(\hat{\beta}) &= E\left\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})']\right\} \\
&= E\left\{[\hat{\beta} - \beta][\hat{\beta} - \beta]'\right\}, (since E(\hat{\beta}) = \beta) \\
&= E\left\{[(X'X)^{-1}X'\mathcal{E}][(X'X)^{-1}X'\mathcal{E}]'\right\} (from(*)) \\
&= E\left\{[(X'X)^{-1}X'\mathcal{E}][\mathcal{E}'X(X'X)^{-1}]\right\} \\
&= (X'X)^{-1}X' \underbrace{E(\mathcal{E}\mathcal{E}')} X(X'X)^{-1}, ...(where, E(\mathcal{E}\mathcal{E}') = \sigma^2 I_k) \\
&= \sigma^2 (X'X)^{-1} \underbrace{X'X(X'X)^{-1}} ....(where X'X(X'X)^{-1} = I_k) \\
&= \sigma^2 (X'X)^{-1}
\end{aligned}$$

<u>Estimation of σ^2</u>

An unbiased estimator of σ^2 is given by:

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n} \hat{\mathcal{E}_\rangle}^2}{n-k} = \frac{\hat{\mathcal{E}}'\hat{\mathcal{E}}}{n-k}$$

where $\hat{\mathcal{E}} = Y - X\hat{\beta}$. Thus, an estimator of the variance-covariance matrix of $\hat{\beta}$ is

$$\hat{\sum}_{\hat{\beta}} = \hat{\sigma}^2 (X'X)^{-1} = \frac{\hat{\mathcal{E}}'\hat{\mathcal{E}}}{n-k}(X'X)^{-1}$$

Note that $\hat{\sum}_{\hat{\beta}}$ is a $(K \times K)$ matrix. The main diagonal elements of this matrix are the estimated variances of $\hat{\beta}_k, \hat{\beta}_2, ..., \hat{\beta}_k$. The square roots of the main diagonal entries of $\hat{\sum}_{\hat{\beta}}$ are the standard errors: $s.e.(\hat{\beta}_1), s.e.(\hat{\beta}_2), ..., s.e.(\hat{\beta}_k)$.

To test for the significance of the regression coefficients:

$$H_0 : \beta_J = 0$$

$$H_A : \beta_J \neq 0$$

the test statistic is:

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}, j = 1, 2, ..., k$$

<u>*Decisionrule :*</u>

If $|t_j| > t_{\alpha/2}(n - K)$, we reject H_0 and conclude that β_j is significant, that is, the regress or variable X_j,j=1, 2, . . . , K, significantly affects the dependent variable Y.

2.3 Hetroscedasticity

In this topic we will see

- the consequences of violation of constant variance assumption.
- What are the consequences for the properties of least squares estimators?
- Is there a better estimation technique?
- How do we detect the existence of Hetroscedasticity?

2.3.1 Introduction

Consider the relationship between average or mean household expenditure on food $E(y)$ and household income x described by the linear function:

$$E(y) = \alpha + \beta x$$

The unknown parameters α and β convey information about this expenditure function. The **response parameter** β describes how mean household food expenditure changes when household income increases by one unit. The intercept parameter α measures expenditure on food for a zero income level. Knowledge of these parameters aids planning by institutions such as government agencies or food retail chains.

To estimate α and β, use data **food.csv** and this data consists a sample of 40 households. Let $\mathcal{E}_i$ be the difference between expenditure on food by the i^{th} household y_i and mean expenditure on food for all households with income x_i. That is,

$$\mathcal{E}_i = y_i - E(y_i) = y_i - \alpha - \beta x_i$$

$$\implies y_i = \alpha + \beta x_i + \mathcal{E}_i$$

Model used to describe expenditure on food for the i^{th} household is written as\

This can be viwed $E(y_i) = \alpha + \beta x_i$ as that part of food expenditure explained by income x_i and $\mathcal{E}_i$ as that part of food expenditure explained by other factors.

Now, we need to check from the data: Whteher the mean function $E(y_i) = \alpha + \beta x_i$ is **equal or better** at explaining expenditure on food for low-income households than it is for high-income households or not?.

As we know that low-income households do not have the option of extravagant food tastes. Comparatively, they have few choices and are almost forced to spend a particular portion of their income on food. High-income households on the other hand could have simple food tastes or extravagant food tastes. Thus, income is relatively **less important** as an explanatory variable for food expenditure of high-income households.

It is harder to guess their food expenditure. This can also be described as that the probability of getting large positive or negative values for $\mathcal{E}$ is higher for high incomes than it is for low incomes. Factors other than income can have a larger impact on food expenditure when household income is high. So, how can we model this phenomenon? A random variable, in this case $\mathcal{E}$, has a higher probability of taking on large values if its variance is high. Thus, we can capture the effect we are describing by having $Var(\mathcal{E})$ depends directly on income x. An equivalent statement is to say $Var(y)$ increases as x increases. Food expenditure y can deviate further from its mean $E(y) = \alpha + \beta x$ when x is large. In such a case, when the variances for all observations are not the same, we say that heteroskedasticity exists. Alternatively, we say the random variable y and the random error $\mathcal{E}$ are heteroskedastic. Conversely, if all observations come from probability density functions with the same variance, we say that homoskedasticity exists, and y and $\mathcal{E}$ are homoskedastic.

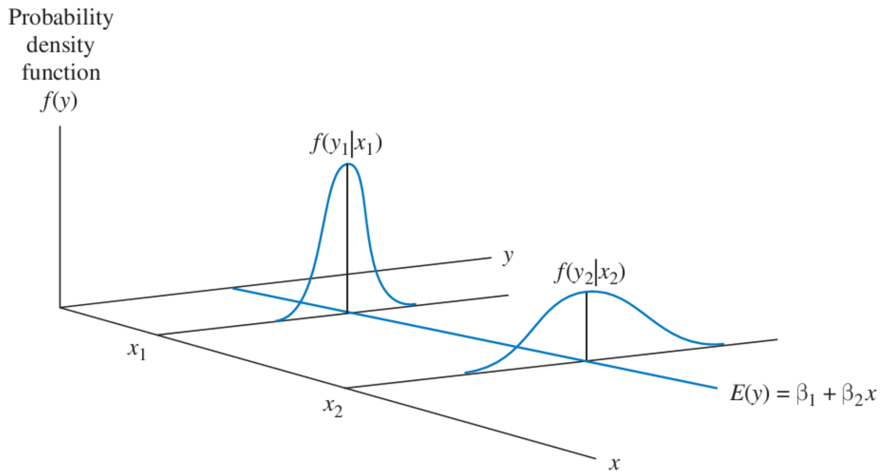

Figure 5: Heteroskedastic errors.

From the above Figure , at $x = x_1$, the probability density function $f(y_1|x1)$ is such that y_1 will be close to $E(y1)$ with high probability. When we move to x_2, the probability density function $f(y_2|x2)$ is more spread out; we are less certain about where y_2 might fall, and larger values are possible. One of the assumptions of CLRM is that random error terms are uncorrelated with mean zero and constant variances σ^2:

$$Var(\mathcal{E}_i) = E(\mathcal{E}_i^2) = \sigma^2, (i = 1, 2, 3, ..., n)$$

This assumption tells us that the variance remains constant for all observations. But there are many situations as we saw in the above in which this assumption may not hold. For example, the variance of the error term may increase or decrease with the dependent variable or one of the independent variables. Under such circumstances, we have the case of **heteroscedasticity**. Generally, under heteroscedasticity we have:

$$E(\mathcal{E}_i^2) = k_i\sigma^2, k_i \ are \ not \ all \ equal \ or$$
$$Var(y_i) = Var(\mathcal{E}_i) = g(x_i)$$

The least squares estimated equation from the observations in the file **food** is shown in the graph

```
fit<-lm(food_exp~income,data=food)


plot(food$income,food$food_exp,pch=19,xlab = "x=weekly income in $100",ylab="y=weekly fo
points(food$income,fitted(fit),type="l",lwd=2)
#text(10,500,expression(hat(y)==83.42 + 10.21x))
text(15,500,expression(paste(hat(y)==83.42 + 10.21,"x")))
```

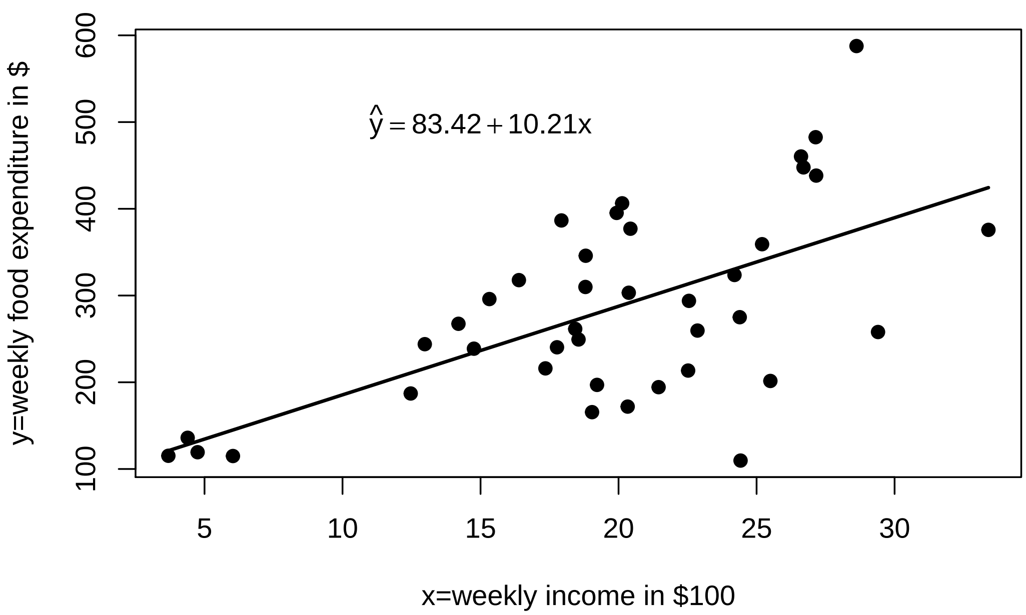

Figure 6: Least squares estimated food expenditure function and observed data points

Heteroscedasticity is more plausible at cross-sectional data than time series data. Here the assumption of homoscedasticity is not very plausible since we expect less variation in consumption for low income families than for high income families. At low levels of income, the average level of consumption is low and the variation around this level is restricted: consumption can not fall too far below the average level because this might mean starvation, and it can not rise too far above the average level because the asset does not allow it. These constraints are likely to be less binding at higher income levels.

2.3.2 Consequences of heteroscedasticity

Since the existence of heteroskedasticity means that the least squares assumption $var(\mathcal{E}_i) = \sigma^2$ is violated, we need to ask what consequences this violation has for our least squares estimator, and what we can do about it.

There are two implications:

1) The least squares estimator is still a linear and unbiased estimator, but it is no longer best. There is another estimator with a smaller variance.
2) The standard errors usually computed for the least squares estimator are incorrect(estimated variances of the OLS estimators are biased). Confidence intervals and hypothesis tests that use these standard errors may be misleading.

Consider the simple linear regression model (in deviation form):

$$y_i = \beta x_i + \mathcal{E}_i, (i = 1, 2, 3....n)$$

where $\mathcal{E}_i$ satisfies all assumptions of the CLRM except that the error terms are heteroscedastic, that is, $E(\mathcal{E}_i^2) = k_i\sigma^2$ and k_i are not all equal.

The OLS estimator of β is:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Then we have:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \beta\frac{\sum x_i^2}{\sum x_i^2} + \frac{\sum x_i \mathcal{E}_i}{\sum x_i^2} = \beta + \frac{\sum x_i \mathcal{E}_i}{\sum x_i^2}$$
$$\Rightarrow E(\hat{\beta}) = \beta + \frac{\sum x_i E(\mathcal{E}_i)}{\sum X_i^2} = \beta, where\ E(\mathcal{E}_i) = 0$$

Thus, $\hat{\beta}$ is an unbiased estimator of β even in the presence of heteroscedasticity.

Recall that the variance of the OLS estimator $\hat{\beta}$ when there is no heteroscedasticity (or under homoscedasticity) is given by:

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2} \qquad \text{......................(2.3.1)}$$

Under heteroscedasticity we have:

$$\begin{aligned}
Var(\hat{\beta})_{HET} = E(\hat{\beta} - E(\hat{\beta}))^2 &= E(\hat{\beta} - \beta)^2 \\
&= (-1)^2 E(\beta - \hat{\beta})^2 \qquad \text{......................}(Unbiasedness) \\
&= E\left(\frac{\sum x_i \mathcal{E}_i}{\sum x_i^2}\right)^2 \\
&= E\left(\frac{\sum x_i^2 \mathcal{E}_i^2}{(\sum x_i^2)^2} + \frac{\sum_{i \neq j} x_i x_j \mathcal{E}_i \mathcal{E}_j}{(\sum x_i^2)^2}\right) \\
&= \frac{\sum x_i^2 E(\mathcal{E}_i^2)}{(\sum x_i^2)^2} + \frac{\sum_{i \neq j} x_i x_j E(\mathcal{E}_i \mathcal{E}_j)}{(\sum x_i^2)^2}, \quad (where\ E(\mathcal{E}_i^2) = k_i \sigma^2 \ \ and\ E(\mathcal{E}_i \mathcal{E}_j) = 0) \\
&= \frac{\sigma^2 \sum k_i x_i^2}{(\sum x_i^2)^2} = \frac{\sigma^2}{\sum x_i^2} \frac{\sum k_i x_i^2}{\sum x_i^2} \\
&= Var(\hat{\beta}) \left[\frac{\sum k_i x_i^2}{\sum x_i^2}\right] \qquad \text{.....................(2.3.2)}
\end{aligned}$$

$if\ you\ assume\ var(\mathcal{E}_i)) = \sigma_i^2$

$$\begin{aligned}
&= \frac{\sum (X_i - \bar{x})^2 \sigma_i^2}{[\sum (X_i - \bar{x})^2]^2} \\
&= \frac{\sum (x_i)^2 \sigma_i^2}{[\sum (x_i)^2]^2} \qquad \text{.....................(2.3.3)}
\end{aligned}$$

Consequently, if we proceed to use the least squares estimator and its usual standard errors when $Var(\mathcal{E}_i) = k_i \sigma^2 = \sigma_i^2$, we are committing **two erros**.

1) we are using an estimate of equation (2.3.1) to compute the standard error of $\hat{\beta}$ when we should be using an estimate of (2.3.2).

2) it is using s^2 to estimate a common σ^2 when in fact the σ_i^2's are different.

From (2.3.1) and (2.3.2), it can be seen that the two variances, $Var(\hat{\mathcal{E}})$ and $Var(\hat{\mathcal{E}})_{HET}$ will be equal only if $k_i = 1 \ \ \forall_i$ that is, only if the errors are homoscedastic.

1) If $\frac{\sum k_i X_i^2}{\sum X_i^2} < 1$, then OLS will overestimate the variance of $\hat{\beta}$.
2) If $\frac{\sum k_i X_i^2}{\sum X_i^2} > 1$, then OLS will underestimate the variance of $\hat{\beta}$.

Thus, under heteroscedasticity, the OLS estimators of the regression coefficients are not BLUE and efficient(with minimum variance). We have to find an alternative estimator that has the minimum variance property.

2.3.3 Detecting Heteroskedasticity

How do I know if heteroskedasticity is likely to be a problem for my model and my set of data? Is there away of detecting heteroskedasticity so that I know whether to investigate other estimation techniques? We consider three ways of investigating these questions.
The tests consider a test for heteroskedasticity based on a **variance function**. As we know that the mean function $E(y_i)$ for a general multiple linear regression is given by:

$$E(y_i) = \beta_1 + \beta_2 x_{2i} + ... + \beta_k x_{ki}$$

A general form for the variance function also given by:

$$var(y_i) = \sigma_i^2 = E(\mathcal{E}_i^2) = h(\alpha_1 + \alpha_2 z_{2i} + ... + \alpha_s z_{si}$$

1. Residual Plots

One way of investigating the existence of heteroskedasticity is to estimate your model using least squares and to plot the least squares residuals. If the errors are heteroskedastic, they may tend to exhibit greater variation in some systematic way.
In a regression with more than one explanatory variable we can plot the least squares residuals against each explanatory variable, or against $\hat{y}_i$, to see if they vary in a systematic way.

Example

Consider the data on weekly food expenditure (Y) and weekly income in \$100(X) for 20 households(both in thousands of Dollars)(use "**consmption.csv**").

```
consmption<-read.csv("consmption.csv",header=T,sep=",")
head(consmption,1)
```

```
##   household income expen
## 1         1   22.3  19.9
```

```
fit<-lm(expen~income,data=consmption)
#plot(fit$fitted.values,fit$residuals,xlab = "x=fitted",ylab = "Residuals")
#abline(h=0,lwd=2,col=4)
# or use
plot(consmption$income,fit$residuals,xlab = "x=income",ylab = "Residuals")
abline(h=0,lwd=2,col="red")
```

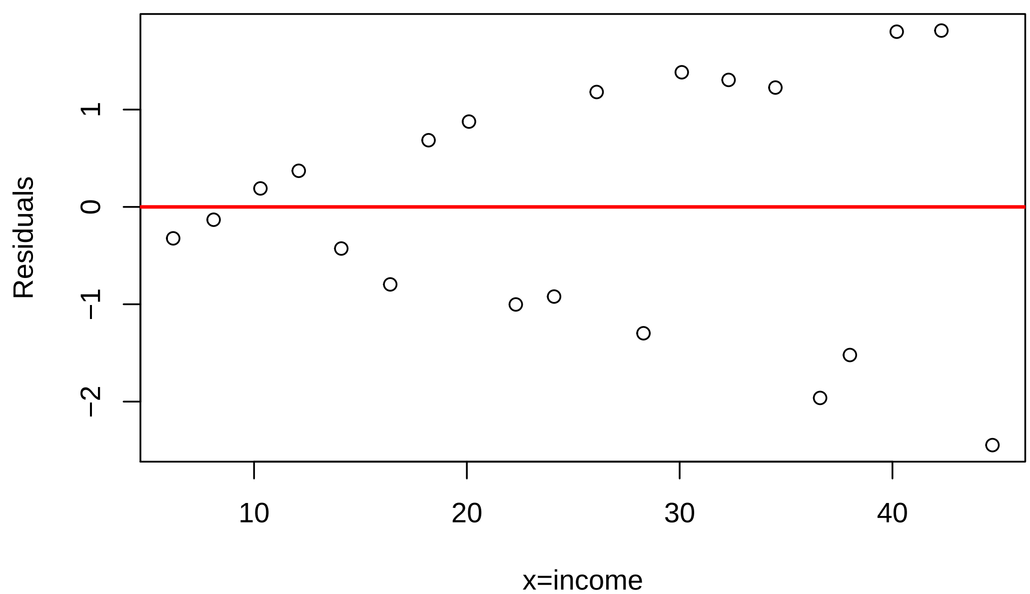

Figure 7: Residual Plot

From the graph, we suspect that the variance increases as incomes increases.

2. White's test

It is a general test for homoskedasticity where **nothing is known** about the form of this heteroskedasticity. This test is based on the difference between the variance of the OLS estimates under homoskedasticity and that under heteroskedasticity.

A test for heteroskedasticity without precise knowledge of the relevant variables and instead by defining the z's as equal to the x's, the squares of the x's, and possibly their cross-products . This test involves applying OLS to:

$$\hat{\mathcal{E}}_i^2 = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + ... + \gamma_p Z_{pi} + U_i$$

and calculating the coefficient of determination R_W^2, where $\hat{\mathcal{E}}_i$ are OLS residuals from the original model. The null hypothesis is:

$$H_0 : \gamma_1 = \gamma_2 = ... = \gamma_p = 0$$

The test statistic is:

$$\chi_{cal}^2 = nR_w^2$$

<u>*Decisionrule :*</u> Reject H_0 (the hypothesis of homoscedasticity) if the above test statistic exceeds the value from the Chi-square distribution with p degrees of freedom for a given level of significance α. If our model has only one independent variable X_i, then p = 2: $Z_{1i} = X_i$ and $Z_{2i} = X_i^2$. If we have two independent variables X_{1i} and X_{2i} , then p = 5: $Z_{1i} = X_{1i}$, $Z_{2i} = X_{2i}$, $Z_{3i} = X_{1i}X_{2i}$,$Z_{4i} = X_{1i}^2$ and $Z_{5i} = X_{2i}^2$. And so on.

Example

For the consummation data, test heteroskedasticity? (use "**consmption.csv**").

solution

To test for heteroskedasticity in the food expenditure where the variance is potentially a function of income, we test $H_0 : \delta_1 = \delta_2 = 0 \;\; Vs \quad H_1 : \delta_1 \neq 0 \; or \; \delta_2 \neq 0$ in the variance function $\sigma_i^2 = h(\alpha_1 + \alpha_2 x_i)$.

This involves applying OLS to: $\hat{\mathcal{E}}_i^2 = \delta_0 + \delta_1 X_i + \delta_2 X_i^2 + U_i$ and computing the coefficient of determination R_w^2.

```
consmption<-read.csv("consmption.csv",header=T,sep=",")
head(consmption,1)
```

```
##   household income expen
## 1         1   22.3  19.9
```

```r
fit<-lm(expen~income,data=consmption)
consmption$ei_hat<-resid(fit)
fit_w<-(lm(I(ei_hat^2)~income+I(income^2),data=consmption))
(Rw_square<-summary(fit_w)$r.squared)
```

```
## [1] 0.8781138
```

```r
x_2<-nrow(consmption)*Rw_square
```

This yields $R_w^2 = 0.8781138$. The White test statistic is:

$$\chi^2_{cal} = nR_w^2 = 20(0.8781138) = 17.5622767$$

We compare this value with $\chi^2_\alpha(p)$ for a given level of significance α. For $\alpha = 0.05$,

```r
(chi_tab<-qchisq(0.05,2,lower.tail=F))
```

```
## [1] 5.991465
```

```r
# Or using a p-value<0.05
pchisq(q=x_2,df=1,lower.tail=F)
```

```
## [1] 2.780498e-05
```

$\chi^2_{0.05}(2) = 5.9914645$.

Conclusion: we conclude that heteroskedasticity exists with the variance dependent on income.

*3. Goldfeld-Quandt test**

This test for heteroskedasticity is designed for two groups of data with possibly different variances. To introduce this case, consider a wage equation where earnings per hour (WAGE) depends on years of education (EDUC), years of experience (EXPER) and a dummy variable METRO that is equal to one for workers who live in a metropolitan area and zero for workers who live outside a metropolitan area. Using data in the file **cps2.csv**(metro= 1 if lives in metropolitan area,female= 1 if female,).

```r
cps2<-read.csv("cps2.csv",header = T,sep = ",")
head(cps2,2)
```

```
##   wage educ exper female black married union south fulltime metro
## 1 2.03   13     2      1     0       0     0     1        0     0
## 2 2.07   12     7      0     0       0     0     0        0     1
```

```r
fit<-lm(wage~educ+exper+metro,data=cps2)
```

The least squares estimated equation for this model is

$$W\hat{A}GE = -9.9139842 + 1.233964 EDUC + 0.1332437 EXPER + 1.5241042 METRO$$

The results suggest that education and experience have a positive effect on the level of wages and that given a particular level of education and experience, the average metropolitan wage is \$1.50 per hour higher than the average wage in a rural area.
The question we now ask is':' How does the variance of wages in a metropolitan area compare with the variance of wages in a rural area? Are the variances likely to be the same, or different? One might suspect that the greater range of different types of jobs in a metropolitan area might lead to city wages' having a higher variance. If the variance of metropolitan wages differs from the variance of rural wages, then we have heteroskedasticity. The variance is not constant for all observations. The Goldfeld-Quandt test is designed to test for this form of heteroskedasticity, where the **sample can be partitioned into two groups**—metropolitan and rural in this case—and we suspect the variance could be different in the two groups.
Suppose we have a model with one explanatory variable X_1 and let Y be the dependent variable. Assuming that the variance is related with X_1 and is continuous variable(). The steps involved in this test are the following:
a) Arrange the observations (both Y and X_1) in increasing order of X_1 .
b) Divide the observations into three parts: n_1 observations in the first part, p observations in the middle part, and n_2 observations in the second part ($n_1 + n_2 + p = n$). Usually p is taken to be **one-sixth** of n.
c) Run a regression on the first n_1 observations, obtain the residuals $\hat{\mathcal{E}_{1i}}$, and calculate the residual variance $s_1^2 = \sum_{i=1}^{n_1} \frac{\hat{\mathcal{E}_{1i}^2}}{(n_1-2)}$.Similarly run a regression on the second n_2 observations, obtain the

residuals $\hat{\mathcal{E}_{2i}}$, and calculate the variance $s_2^2 = \sum_{i=1}^{n_2} \frac{\hat{\mathcal{E}_{2i}^2}}{(n_2-2)}$.

Note:

The variances of the last several disturbances in the first part are likely to be similar to those of the first several disturbances in the second part. To increase the power of the test, it is recommended that the two parts be some distance apart. Thus, we drop the middle p residuals all together.

If we need to test variance for a categorical variable like(Urban/Rural), directly apply **step c** separately.

d) Calculate the test statistic: $F_{cal} = \frac{S_2^2}{S_1^2}$

e) $\underline{Decisionrule:}$ Reject the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ (and conclude that the errors are heteroscedastic) if:

$$F_{cal} > F_\alpha(n_1 - 2, n_2 - 2)$$

where $F_\alpha(n_1 - 2, n_2 - 2)$ is the critical value from the F-distribution with $n_1 - 2$ and $n_2 - 2$ degrees of freedom in the numerator and denominator, respectively, for a given significance level α.

Example

Apply the Goldfeld-Quandt test for non-constant error variance for metropolitan area and non-metropolitan area(rural)? use (cps2.csv).

solution

The test is based on a comparison of the error variances estimated from eachgroup.Using the subscript M to denote metropolitan observations and the subscript R to denote rural observations, we can write separate equations for the two groups as

$$WAGE_{Mi} = \beta_{M1} + \beta_2 EDUC_{Mi} + \beta_3 EXPER_{Mi} + \mathcal{E}_{Mi} \quad ....(2.3.4)$$

$$WAGE_{Ri} = \beta_{R1} + \beta_2 EDUC_{Ri} + \beta_3 EXPER_{Ri} + \mathcal{E}_{Ri} \quad ....(2.3.5)$$

Implicit in the above specification is the assumption that the coefficients for EDUC and EXPER(β_2 and β_3) are the same in both metropolitan and rural areas, but the intercepts differ.

Now we need to test $H_0 : \sigma_M^2 = \sigma_R^2 \quad Vs\ H_1 : \sigma_M^2 \neq \sigma_R^2$

Apply **step c:** to both equation

```
fit_M<-lm(wage~educ+exper,data=cps2[cps2$metro==1,])
fit_R<-lm(wage~educ+exper,data=cps2[cps2$metro==0,])
(s_M_sq<-sum(resid(fit_M)^2)/(nrow(cps2[cps2$metro==1,])-3))
```

```
## [1] 31.82373
```

```
(s_R_sq<-sum(resid(fit_R)^2)/(nrow(cps2[cps2$metro==0,])-3))
```

```
## [1] 15.24299
```

```
(f_cal<-s_M_sq/s_R_sq)
```

```
## [1] 2.087762
```

Calculate

$$F_{cal} = \frac{s_M^2}{s_R^2} = 31.8237318/15.2429866 = 2.0877622$$

The lower and upper critical values for a 5% significance level are
$F_{Lc} = F_{(0.025,805,189)}$ *and* $F_{Uc} = F_{(0.975,805,189)}$ are given below. We reject
$H_0 \ if \ F_{cal} < F_{Lc} \ or \ F_{cal} > F_{Uc}$

```
(F_Lc<-qf(0.025,nrow(cps2[cps2$metro==1,])-3,nrow(cps2[cps2$metro==0,])-3))
```

```
## [1] 0.8051984
```

```
(F_Uc<-qf(0.975,nrow(cps2[cps2$metro==1,])-3,nrow(cps2[cps2$metro==0,])-3))
```

```
## [1] 1.26173
```

Conclusion: Since 2.0877622 > 1.2617297, we reject H_0 and conclude that the wage variances for the rural and metropolitan regions are not equal.

*4. Breusch-Pagan test**

This involves applying OLS to:

$$\frac{\hat{\mathcal{E}}_i^2}{\hat{\sigma}^2} = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + ... + \gamma_k X_{ki} + U_i$$
$$where \quad \hat{\sigma}^2 = \sum \hat{\mathcal{E}}_i^2/n, \ the \ ML \ estimator \ of \ \sigma^2 \ \ under \ homeskedasticity.$$

and calculating the regression sum of squares (RSS). The test statistic is:

$$\chi^2_{cal} = \frac{RSS}{2}$$

<u>*Decisionrule :*</u> Reject the null hypothesis of homoscedasticity: $H_0 : \gamma_1 = \gamma_2 = ... = \gamma_k = 0$ if:

$$\chi^2_{cal} > \chi^2_{\alpha}(K)$$

where $\chi^2_{\alpha}(K)$ is the critical value from the Chi-square distribution with K degrees of freedom for a given value of α.

Example

use data of consumption.csv, and test for presence of heteroscedasticity using Breusch - Pagan test?

solution

This involves applying OLS to:

$$\frac{\hat{\mathcal{E}}_i^2}{\hat{\sigma^2}} = \gamma_0 + \gamma_1 X_i + U_i$$

```
#You have to know location of your dependent variable, when you use this function for
bp_test<-function(x=data.frame(),alpha=null)
{
  n=nrow(x)
  fit<-lm(expen~income,data=x)
  sigma_hat_2=deviance(fit)/n
  res_bp<-resid(fit)^2/sigma_hat_2
  fit_bp<-lm(res_bp~income, data=x)
  fit_bp
  ## Regresresssion SS
```

```r
  rss=sum((fitted(fit_bp)-mean(res_bp))^2)
  # x_2 test statistic
   x_cal<-rss/2
   #Probability of getting greater than this calculated value, that is, is this due to
   # p-value
   p_value<-pchisq(q=x_cal, df=1,lower.tail=F)
   #x_tab
   x_tab<-qchisq(alpha, df=1,lower.tail=F)

   if(p_value < alpha){
     cat("Decision : ","\nReject Ho: constant variance, since p-value ",p_value,"<",alph
         "\nchi_tabulated=",x_tab,"<",x_cal,"=chi_calculated")
   }
   else{
     cat(" Do not reject Ho: constant variance, with a p-value",p_value)
     cat("Decision : ","\nDo not reject Ho: constant variance, since p-value ",p_value,"
         "\nchi_tabulated=",x_tab,">",x_cal,"=chi_calculated")
   }
}
bp_test(x=consmption,alpha = 0.05)
```

```
## Decision :
## Reject Ho: constant variance, since p-value  0.006218296 < 0.05
## Or
## chi_tabulated= 3.841459 < 7.485929 =chi_calculated
```

All of the tests indicated that the disturbances are heteroscedastic. Thus, the regression coefficients obtained by OLS are **not efficient**. In such cases, we have to apply another method such as weighted least squares (WLS) estimation.

Heteroskedasticity-Consistent Standard Errors

When our model suffers from heteroskedasticity, even the least squares estimator are unbiased but no

longer best. The other is that the usual least squares standard errors are incorrect, which invalidates interval estimates and hypothesis tests. If we are prepared to accept the least squares estimator as a useful estimator, despite the fact it is not the minimum variance estimator, there is a way of correcting the standard errors so that our interval estimates and hypothesis tests are valid.
Use a standard error of the estimator from equation (2.3.2) to test your hypothesis. These standard error are known by heteroskedasticity-consistent standard errors,or heteroskedasticity robust standard errors, or simply **robust standard errors**.

2.3.4 Correction for heteroscedasticity

Consider the model:

$$Y_i = \alpha + \beta X_i + \mathcal{E}_i \text{..........................}(2.3.6)$$

where $E(\mathcal{E}_i^2) = \sigma_i^2$ is known for i = 1, 2, . . . , n. We make the following transformation:

$$\frac{\alpha}{\sigma_i} = \frac{\alpha}{\sigma_i} + \beta\frac{X_i}{\sigma_i} + \frac{\mathcal{E}_i}{\sigma_i} ...(let, \frac{\alpha}{\sigma_i} = Y_i^*,\ \frac{\alpha}{\sigma_i} = \alpha^*,\ \frac{X_i}{\sigma_i} = X_i^*\ and\ \frac{\mathcal{E}_i}{\sigma_i} = \mathcal{E}_i^*).$$

The transformed model can be written as:

$$Y_i^* = \alpha^* + \beta X_i^* + \mathcal{E}_i^* \text{..............}(2.3.7)$$

We then check if the disturbances of equation (2.3.7) satisfy OLS assumptions:

$$E(\mathcal{E}_i^*) = E(\frac{\mathcal{E}_i}{\sigma_i}) = \frac{E(\mathcal{E}_i)}{\sigma_i = 0} \text{.....}, (E(\mathcal{E}_i) = 0)$$
$$Var(\mathcal{E}_i^*) = E(\mathcal{E}_i^*)^{=} E\left(\frac{\mathcal{E}_i}{\sigma_i}\right)^2 = \frac{E(\mathcal{E}_i^2)}{\sigma_i^2} = E\left(\frac{\mathcal{E}_i}{\sigma_i}\right)^2 = \frac{\sigma_i^2}{\sigma_i^2} = 0$$

Thus, the variance of the transformed disturbances is constant. So we can apply OLS to equation (2.3.7) to get regression coefficient estimates that are BLUE (Gauss-Markov Theorem). This estimation method is known as **weighted least squares (WLS)** since each observation is weighted (multiplied) by $W_i = \frac{1}{\sigma_i}$.
Specification of the weights

The major difficulty with WLS is that σ_i^2 are rarely known. We can overcome this by making certain assumptions about σ_i^2 or by estimating σ_i^2 from the sample. The information about σ_i^2 is frequently in the form of an assumption that σ_i^2 is associated with some variable, say Z_i.

$\underline{Illustration 1:}$ In the case of micro-consumption function, the variance of the disturbances is often assumed to be positively associated with level of income. So the place of Z_i will be taken by the explanatory variable X_i income, that is,

$$\sigma_i^2 = \sigma^2 Z_i^2 = \sigma^2 X_i^2.$$

We then divide equation (1) throughout by X_i:

$$\frac{Y_i}{X_i} = \alpha\left(\frac{1}{X_i}\right) + \beta\left(\frac{X_i}{X_i}\right) + \frac{\mathcal{E}_i}{X_i}$$
$$\Rightarrow \frac{Y_i}{X_i} = \alpha\left(\frac{1}{X_i}\right) + \beta + u_i..........(2.3.8)$$

where $U_i = \frac{\mathcal{E}_i}{X_i}$ .Now we have:

$$Var(U_i) = Var\left(\frac{X_i}{X_i}\right) = E\left(\frac{X_i}{X_i}\right)^2 = E\frac{\mathcal{E}_i^2}{X_i^2} = \frac{\sigma^2 X_i^2}{X_i^2} = \sigma^2$$

Hence, the variance of the disturbance term in equation (2.3.8) is constant, and we can apply OLS by regressing $\frac{Y_i}{x_i}$ on $\frac{1}{X_i}$. Note that the estimated constant term and slope from the transformed model (2.3.8) correspond to the values of $\hat{\beta}$ and $\hat{\alpha}$, respectively.

Illustration 2: In the case of micro-consumption function, the variance of the disturbances may also be thought to be associated with changes in some 'outside' variable, say the size of the family= Z_i, that is,

$$\sigma_i^2 = \sigma^2 Z_i^2.$$

We then divide equation (2.3.5) throughout by Z_i:

$$\frac{Y_i}{Z_i} = \alpha\left(\frac{1}{Z_i}\right) + \beta\left(\frac{X_i}{Z_i}\right) + V_i$$

where $V_i = \frac{\mathcal{E}_i}{Z_i}$. It can easily be shown that $Var(V_i) = \sigma^2$. To estimate the regression coefficients, we should run a regression of $\frac{Y_i}{Z_i}$ on $\frac{1}{Z_i}$ and $\frac{X_i}{Z_i}$ without a constant term.

Example: apply WLS method on 'consumption' data?

All of the tests indicate that the disturbances are heteroscedastic. Thus, the regression coefficients obtained by OLS are not efficient. In such cases, we have to apply weighted least squares (WLS) estimation. The weights can be obtained from the

- **sample at hand** or - from some **prior knowledge. In this example let's estimate the weights σ_i from the sample (that is we are applying white's test method).

First we apply OLS estimation and obtain the residuals $\hat{\mathcal{E}}_i$. We then order the residuals based on the absolute magnitude of the explanatory variable (income). Next we divide the residuals into three parts: the first and second parts consisting of seven residuals and the third part consisting of six residuals. The variance of each part is computed as:

$$\hat{\sigma_i^2} = \frac{1}{n_i} \sum \hat{\mathcal{E}}_i^2$$

, where n_i is the number of residuals in the i^{th} part, i = 1, 2, 3. The results are:

```
rm(list = ls())
consmption<-read.csv("consmption.csv",header=T,sep=",")
  fit<-lm(expen~income,data=consmption)
  sorted_residual<-resid(fit)[order(consmption$income)]
  n1= 7; n2=7; n3=6
# variance of each part
sigma_hat1_sq<-sum(sorted_residual[1:7]^2)/n1
sigma_hat2_sq<-sum(sorted_residual[8:14]^2)/n2
sigma_hat3_sq<-sum(sorted_residual[15:20]^2)/n3
sigma_hat1_sq=round(sigma_hat1_sq,6)
sigma_hat2_sq=round(sigma_hat2_sq,6)
sigma_hat3_sq=round(sigma_hat3_sq,6)
s1=sqrt(sigma_hat1_sq); s2=sqrt(sigma_hat2_sq);s3=sqrt(sigma_hat3_sq)
```

$$\hat{\sigma_1^2} = 0.225895 \Longrightarrow \hat{\sigma_1} = 0.4752841$$
$$\hat{\sigma_2^2} = 1.330624 \Longrightarrow \hat{\sigma_2} = 1.1535268$$
$$\hat{\sigma_3^2} = 3.363024 \Longrightarrow \hat{\sigma_3} = 1.833855$$

The next step is to divide the values of the dependent variable, the independent variable and the constant term (a vector of 1's) in the i^{th} part by σ_i:

$$\frac{y_i}{\hat{\sigma_i}} = \alpha\left(\frac{1}{\hat{\sigma_i}}\right) + \beta\left(\frac{x_i}{\hat{\sigma_i}}\right) + U_i, \quad \mathit{where} \quad U_i = \frac{\mathcal{E}_i}{\hat{\sigma_i}}.$$

```
# Vector of weights
weight<-rep(c(s1,s2,s3),c(7,7,6))
xxx=consmption[,-1] # To remove Household variable
xxx$one<-1
weighted_data=xxx/weight
weighted_data<-round(weighted_data,4)
names(weighted_data)=c("y_sigma_i","x_sigma_i","one_sigma_i")
```

Table 5: transformed data by its respective sigma.

y_sigma_i	x_sigma_i	one_sigma_i
46.9193	41.8697	2.1040
67.9593	65.6449	2.1040
77.0066	66.9073	2.1040
25.4585	25.4585	2.1040
88.9994	85.6330	2.1040
13.0448	12.8344	2.1040
94.0490	81.2146	2.1040

y_sigma_i	x_sigma_i	one_sigma_i
22.6263	22.1061	0.8669
8.9291	8.9291	0.8669
34.8496	33.6360	0.8669

We then run an OLS regression of $\frac{1}{\hat{\sigma_i}}$ on $\frac{X_i}{\hat{\sigma_i}}$ and $\frac{Y_i}{\hat{\sigma_i}}$ without a constant term. The results are:

```r
# exclude the intercept (-1)
fit_wls<-lm(I(expen/weight)~I(1/weight)+I(income/weight)-1,consmption)
 print(summary(fit_wls))
```

```
##
## Call:
## lm(formula = I(expen/weight) ~ I(1/weight) + I(income/weight) -
##     1, data = consmption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2822 -0.6190  0.0360  0.9667  4.6207
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## I(1/weight)        0.9381     0.7585   1.237    0.232
## I(income/weight)   0.8881     0.0249  35.671   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.067 on 18 degrees of freedom
## Multiple R-squared:  0.9974, Adjusted R-squared:  0.9971
## F-statistic:  3465 on 2 and 18 DF,  p-value: < 2.2e-16
```

$$Y_i = \underset{(0.7585)}{0.9380998} + \underset{(0.0249)}{0.8880795X_i}$$

$$R^2 = 0.9974, \quad \hat{\sigma^2} = 2.0668$$

The plot of the residuals of the transformed model against the explanatory variable (income) is shown below. It can be seen that the spread of the residuals has no increasing or decreasing pattern, i.e., there is no heteroscedasticity.

```
plot(consmption$income,fit_wls$residuals,xlab = "income(xi)",ylab = "WLS_Residuals", pch
abline(h=0,lwd=2,col="red")
```

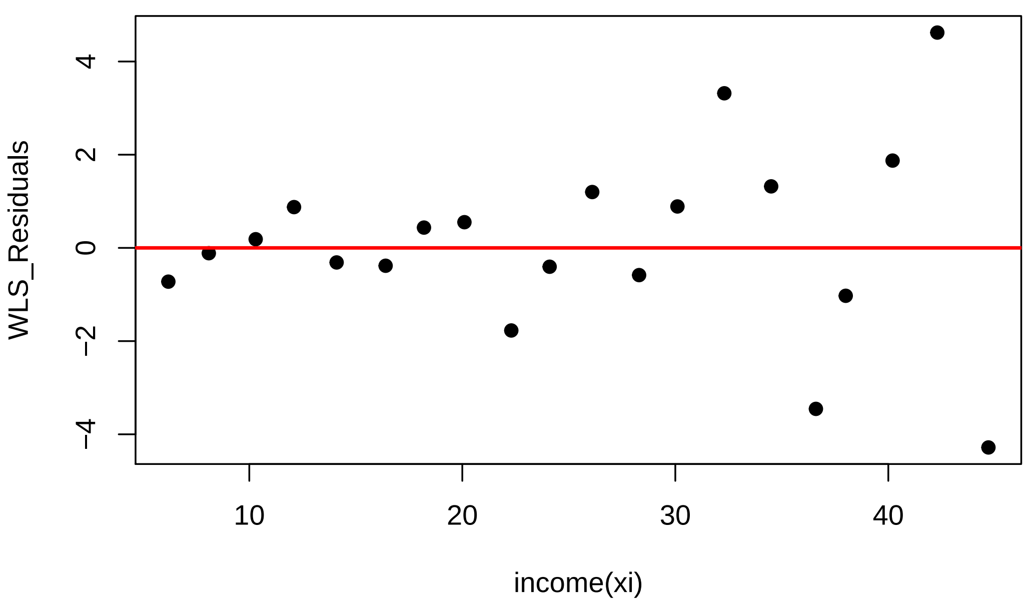

Figure 8: Residual Plot of WLS

Another method of correcting for heteroscedasticity is based on the assumption that the variance of the disturbances is positively associated with level of income X , that is,

$$\sigma_i^2 = \sigma^2 X_i^2.$$

The model we are going to estimate is then:

$$\frac{y_i}{x_i} = \alpha\left(\frac{1}{x_i}\right) + \beta\left(\frac{X_i}{x_i}\right) + \frac{\mathcal{E}_i}{x_i}$$
$$\Rightarrow \frac{y_i}{x_i} = \alpha\left(\frac{1}{x_i}\right) + \beta + U_i, \qquad ...\ where U_i = \frac{\mathcal{E}_i}{x_i}.$$

This simply means that we apply OLS by regressing $\frac{y_i}{x_i}$ on $\frac{1}{x_i}$.(**Exercise**)

2.4 Autocorrelation

objective of the lesson:

- To identify the nature of autocorrelation
- To detect the consequence of serial-autocorrelation
- To take a remedial measure for serial-autocorrelation problem
- checking spatial autocorrelation in OLS residuals
- To distinguish spatial error and lag model

Introduction

In cross-sectional studies, such as households or firms are **randomly selected** no prior reason to believe that the error term pertaining to one household or a firm is correlated with the error term of another household or firm. however, in a cross-sectional data considering a space order may exist and is called **spatial autocorrelation**. In cross-sectional analysis, the ordering of the data must have some logic, or economic interest, to make sense of any determination of whether (spatial) autocorrelation is present or not.

In time series data, where observations follow a natural ordering over time, are likely to be correlated. successive observations over short interval are also more likely to be correlated as compared to long interval difference.

2.4.1 Nature of the autocorrelation

The term autocorrelation may be defined as a correlation between order of a single variable, that is to itself. This order may across space (spatial autocorrelation) or across time(time-correlation or simply autocorrelation). One of the assumption of the CLRM is that autocorrelation does not exist in the disturbances $\mathcal{E}_i$.

$$E(\mathcal{E}_i\mathcal{E}_j) = 0$$

that is,the disturbance term relating to any observation is not influenced by the disturbance term relating to any other observation. For instance, in a study of the relationship between output and inputs of a firm or industry from monthly observations, non-autocorrelation of the disturbance implies that the effect of machine breakdown is strictly temporary in the sense that only the current month's output is affected. But in practice, the effect of a machine breakdown in one month may affect current month's output as well as the output of subsequent months. And, if there is such a dependence, we have autocorrelation. Mathematically,

$$E(\mathcal{E}_i\mathcal{E}_j) \neq 0 \ for \ all \ i \neq j$$

In regressions involving time series data, successive observations are likely to be interdependent. Reasons for occurring serial correlation includes:

1) Inertia
2) Specification Bias: Excluded Variables Case.
3) Specification Bias: Incorrect Functional Form.
4) Cobweb Phenomenon
5) Lags.
6) Manipulation of Data(such as smothing, interpolation or extrapolation)
7) Data Transformation
8) Nonstationarity

2.4.2 OLS estimation in the presence of serial autocorrelation

What happens to the OLS estimators and their variances in the presence of autocorrelation?

Consider two-variable regression model: $Y_t = \beta_1 + \beta_2 X_t + \mathcal{E}_t \ where \ t = time$

In order to see the consequences of autocorrelation, we have to specify the nature (mathematical form) of auto-correlation or the **mechanism that generates** it, i.e., $E(\mathcal{E}_t\mathcal{E}_{t+s}) \neq 0 (for\ s \neq 0)$. As first approximation or usually we assume that the errors (disturbances) follow the first-order autoregressive scheme (abbreviated as AR(1)) or generated by the mechanism:

The error process:

$$\mathcal{E}_t = \rho\mathcal{E}_{t-1} + u_t .... (2.4.1)$$

where $\rho(rho)$ is known as the **coefficient of autocovariance** and where u_t is the stochastic disturbance term (also called a white noise error term) such that it satisfied the standard OLS assumptions, namely

$$E(u_t) = 0, \quad var(u_t) = \sigma_u^2, \quad cov(u_t, u_{t+s}) = 0\ for\ s \neq 0$$

- Equation (2.4.1) says that the value of the disturbance term in period t is equal to rho times its value in the previous period plus a purely random error term.
- Equation (2.4.1) is also called an autoregressive process of order one (or **AR(1)**).
- The name autoregressive is appropriate because equation (2.4.1) can be interpreted as the regression of $\mathcal{E}_t$ on itself lagged one period.
- It is first order because $\mathcal{E}_t$ and its immediate past value are involved; that is, the maximum lag is 1. If the model were $\mathcal{E}_t = \rho_1\mathcal{E}_{t-1} + \rho_2\mathcal{E}_{t-2} + u_t$, it would be an AR(2), or second-order, autoregressive scheme, and so on.

From equation (2.4.1) we have (beginning from $\mathcal{E}_0$):

$$\mathcal{E}_1 = \rho\mathcal{E}_0 + u_1$$

$$\mathcal{E}_2 = \rho\mathcal{E}_1 + u_2$$

$$\mathcal{E}_2 = \rho(\rho\mathcal{E}_0 + u_1) + u_2$$

$$\mathcal{E}_2 = \rho^2\mathcal{E}_0 + \rho u_1 + u_2$$

$$\mathcal{E}_3 = \rho\mathcal{E}_2 + u_3$$

$$\mathcal{E}_3 = \rho(\rho^2\mathcal{E}_0 + \rho u_1 + u_2) + u_3$$

$$\mathcal{E}_3 = \rho^3\mathcal{E}_0 + \rho^2 u_1 + \rho u_2 + u_3$$

$$\mathcal{E}_t = \rho^t\mathcal{E}_0 + \rho^{t-1}u_1 + \rho^{t-2}u_2 + ... + \rho u_{t-1} + u_t$$

$$\mathcal{E}_t = \rho^t\mathcal{E}_0 + \sum_{j=0}^{t-1}\rho^j u_{t-j}$$

For this process to be stable,ρ must be less than one in absolute value, that is, $|\rho| < 1$. In such cases, as $t \longrightarrow \infty$, we have $\rho^t\mathcal{E}_0 \longrightarrow 0$. Thus, $\mathcal{E}_t$ can be expressed as:

$$\mathcal{E}_t = \sum_{j=0}^{t-1}\rho^j u_{t-j} = u_t + \rho^1 u_{t-1} + \rho^2 u_{t-2} + \rho^3 u_{t-3} + ...$$

Since $|\rho| < 1$, ρ^j keeps on decreasing as j keeps on increasing. This means that the effect of the recent past is significant and the effect keeps on diminishing the further back we go (called **fading memory**).

The mean of $\mathcal{E}$ is:

$$E(\mathcal{E}_t) = \sum_{j=0}^{\infty}\rho^j E(u_{t-j}) = 0$$

The variance of $\mathcal{E}_t$ is:

$$
\begin{aligned}
Var(\mathcal{E}_t) &= E(\mathcal{E}_t - E(\mathcal{E}_t))^2 \\
&= E(\mathcal{E}_t^2) \\
&= E(\sum_{j=0}^{\infty} \rho^j u_{t-j})^2 \\
&= E(\sum_{j=0}^{\infty} \rho^2 j u_{t-j}^2 + \sum_{j=0}^{\infty}\sum_{i\neq 0}^{\infty} \rho^j \rho^i u_{t-j} u_{t-i}) \\
&= (\sum_{j=0}^{\infty} \rho^2 j E(u_{t-j}^2) + \sum_{j=0}^{\infty}\sum_{i\neq 0}^{\infty} \rho^j \rho^i E(u_{t-j} u_{t-i})) \\
where\ (E(u_{t-j}^2) = \sigma_u^2\ and\ E(u_{t-j}u_{t-i}) = 0\ for\ i \neq j) & \\
&= \sum_{j=0}^{\infty} \rho^2 j \sigma_u^2 = \sigma_u^2 \sum_{j=0}^{\infty} \rho^2 j = \frac{\sigma_u^2}{1-\rho^2} \\
\Longrightarrow Var(\mathcal{E}_t) &= \frac{\sigma_u^2}{1-\rho^2} = \sigma^2 \\
(Assumption\ of\ constant\ error\ variance) &
\end{aligned}
$$

The covariance between $\mathcal{E}_t$ and $\mathcal{E}_{t-1}$ is derived as:

$$
\begin{aligned}
cov(\mathcal{E}_t, \mathcal{E}_{t-1}) &= E(\mathcal{E}_t \mathcal{E}_{t-1}) \\
&= E[(u_t + \rho^1 + \rho^2 u_{t-2} + ...)(u_{t-1} + \rho^1 u_{t-2} + \rho^2 u_{t-3} + ...)] \\
&= E(u_t u_{t-1}) + \rho E(u_t u_{t-2}) + \rho^2 E(u_t u_{t-3}) + ... \\
&+ \rho E(u_{t-1} u_{t-1})^* + \rho^2 E(u_{t-1} u_{t-2}) + \rho^3 E(u_{t-1} u_{t-3}) + ... \\
&+ \rho^2 E(u_{t-2} u_{t-1}) + \rho^3 E(u_{t-2} u_{t-2})^* + \rho^4\ E(u_{t-2} u_{t-3}) + ... \\
&+ \rho^3 E(u_{t-3} u_{t-1}) + \rho^4 E(u_{t-3} u_{t-2}) + \rho^5 E(u_{t-3} u_{t-3})^* + ... \\
&= \rho \sigma_u^2 + \rho^3 \sigma_u^2 + \rho^5 \sigma_u^2 + ... \\
&= \rho \sigma_u^2 (1 + \rho^2 + \rho^4 + ...)
\end{aligned}
$$

$$
cov(\mathcal{E}_t, \mathcal{E}_{t-1}) = \rho \sigma_u^2 \sum_{j=0}^{\infty} (\rho^2)^j = \frac{\rho \sigma_u^2}{1-\rho^2} = \rho \sigma^2 \text{...(2.4.2)}
$$

Similarly, it can be shown that:

$$cov(\mathcal{E}_t, \mathcal{E}_{t-2}) = E(\mathcal{E}_t, \mathcal{E}_{t-2}) = \rho^2\sigma^2.$$
$$cov(\mathcal{E}_t, \mathcal{E}_{t-3}) = E(\mathcal{E}_t, \mathcal{E}_{t-3}) = \rho^3\sigma^2.$$

Or in general:

$$cov(\mathcal{E}_t, \mathcal{E}_{t-s}) = E(\mathcal{E}_t, \mathcal{E}_{t-s}) = \rho^s\sigma^2$$

Equivalently, we have:

$$cov(\mathcal{E}_t, \mathcal{E}_s) = E(\mathcal{E}_t, \mathcal{E}_s) = \rho^{|s-t|}\sigma^2$$

Thus, the relationship between the disturbances depends on the value of the parameter ρ. From relation (2.4.2) we have:

$$\begin{aligned} cov(\mathcal{E}_t, \mathcal{E}_{t-1}) = E(\mathcal{E}_t, \mathcal{E}_{t-1}) = \rho\sigma^2 \\ \Rightarrow \rho = \frac{cov(\mathcal{E}_t, \mathcal{E}_{t-1})}{\sigma^2} \\ \Rightarrow \rho = \frac{cov(\mathcal{E}_t, \mathcal{E}_{t-1})}{\sqrt{var(\mathcal{E}_t)}\sqrt{var(\mathcal{E}_{t-1})}} \end{aligned}$$

- $\Longrightarrow \sigma$ is nothing but the **coefficient of correlation** between $\mathcal{E}_t$ and $\mathcal{E}_{t-1}$.
- ρ is estimated from the residuals of the **OLS** regression by:

$$\hat{\rho} = \frac{\sum_{t=2}^{T} \hat{\mathcal{E}}_t \hat{\mathcal{E}_{t-1}}}{\sqrt{\sum_{t=1}^{T} \hat{\mathcal{E}_t^2}}\sqrt{\sum_{t=2}^{T} \hat{\mathcal{E}_{t-1}^2}}}$$

2.4.2.1 Properties of OLS estimators under serial autocorrelation The model (in deviations form) is:

$$y_t = \beta X_t + \mathcal{E}_t$$
$$\mathcal{E}_t = \rho\mathcal{E}_{t-1} + u_t, |\rho| < 1$$

where $U_t = \mathcal{E}_t - \rho\mathcal{E}_{t-1}$ satisfies all assumptions of the **CLRM** $(E(u_t) = 0, Var(u_t) = E(u_t^2) = \sigma_u^2$ and $E(u_t u_s) = 0$ for $t \neq s)$.

The **OLS** estimator of β is:

$$\hat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2}$$

Then we have:

$$\hat{\beta} = \frac{\sum x_t \beta X_t + \mathcal{E}_t}{\sum x_t^2} = \beta + \frac{\sum x_t \mathcal{E}_t}{\sum x_t^2}$$
$$\Rightarrow E(\hat{\beta}) = \beta + \frac{\sum x_t E(\mathcal{E}_t)}{\sum x_t^2} = \beta,\ where\ E(\mathcal{E}_t) = 0.$$

Thus, $\hat{\beta}$ is still an unbiased estimator of β.

Recall that the variance of the **OLS** estimator $\hat{\beta}$ when there is no autocorrelation is given by:

$$Var(\hat{\beta}_2)_{OLS} = \frac{\sigma^2}{\sum x_t^2}..................(2.4.3)$$

Now under the **AR(1)** scheme, it can be shown that the variance of this estimator is:

$$
\begin{aligned}
Var(\hat{\beta})_{AR(1)} &= E(\hat{\beta}-\beta)^2 = E\left(\frac{\sum x_t\mathcal{E}_t}{\sum x_t^2}\right)^2 \\
&= \frac{1}{\left(\sum x_t^2\right)^2}E\left(\sum x_t^2\mathcal{E}_t^2 + 2\sum_{s>t} x_t\mathcal{E}_t x_s\mathcal{E}_s\right) \\
&= \frac{1}{\left(\sum x_t^2\right)^2}\sum x_t^2 E(\mathcal{E}_t^2) + \frac{2}{\left(\sum x_t^2\right)^2}\sum_{s>t} x_t x_s E(\mathcal{E}_t\mathcal{E}_s).....(but, E(\mathcal{E}_t\mathcal{E}_s) = \rho^|s-t|\sigma^2) \\
&= \frac{\sigma^2}{\sum x_t^2} + \frac{2\sigma^2}{\sum x_t^2}\left[\rho\frac{\sum x_t x_{t-1}}{\sum x_t^2} + \rho^2\frac{\sum x_t x_{t-2}}{\sum x_t^2} + \rho^3\frac{\sum x_t x_{t-3}}{\sum x_t^2} + ...\right] \\
&= Var(\hat{\beta})_{OLS} + \frac{2\sigma^2}{\sum x_t^2}\left[\rho\frac{\sum x_t x_{t+1}}{\sum x_t^2} + \rho^2\frac{\sum x_t x_{t+2}}{\sum x_t^2} + \rho^3\frac{\sum x_t x_{t+3}}{\sum x_t^2} + ...\right] ....(2.4.4)
\end{aligned}
$$

Therefore, when $\rho > 0$ and x_t is positively correlated with $x_{t+1}, x_{t+2}, ...$, the second term on the right hand side will usually be positive, and we have:

$$Var(\hat{\beta})_{AR(1)} = Var(\hat{\beta})_{OLS}$$

Note that the stated conditions concerning ρ and the X_t's are fairly common in economic time series, that is, usually the terms such as $\sum x_t x_{t+1}$ or $\sum x_t x_{t+2}$ are expected to be positive since it is not unreasonable to assume that successive terms in economic time series data are positively correlated (consumption high in one period usually means it will be high in the next period as well).
Thus, if the errors are autocorrelated, and yet we persist in using **OLS**, then the variances of regression coefficients will be **under-estimated** leading to narrower confidence intervals, high values of R^2 and inflated t-ratios.

2.4.2.2 The BLUE estimator in the presence of serial-autocorrelation In the case of autocorrelation can we find an estimator that is BLUE? Yes.
Continuing with the two-variable model and assuming the AR(1) process, we can show that the BLUE estimator of β_2 is given by the following expression.

$$\hat{\beta_2}^{GLS} = \frac{\sum_{t=2}^{T}(x_t-\rho x_{t-1})(y_t-\rho y_{t-1})}{\sum_{t=2}^{T}(x_t-\rho x_{t-1})^2} + C...................(2.4.5)$$

where C is a correction factor that may be disregarded in practice. And its variance is given by

$$var(\hat{\beta}_2^{GLS}) = \frac{\sigma^2}{\sum_{t=2}^{T}(x_t - \rho x_{t-1})^2} + C.....................(2.4.6)$$

where D too is a correction factor that may also be disregarded in practice.

The estimator β_2^{GLS}, as the superscript suggests, is obtained by the **method of GLS**. in GLS we incorporate any additional information we have (e.g., the nature of the heteroscedasticity or of the autocorrelation) directly into the estimating procedure by transforming the variables, whereas in OLS such side information **is not directly** taken into consideration.
Hence, under autocorrelation, it is the GLS estimator given in (2.4.5) that is BLUE, and the minimum variance is now given by (2.4.6) and not by (2.4.4) and obviously not by (2.4.3).

2.4.2.3 Consequences of Using OLS in The Presence of Serial Autocorrelation

1) The residual variance $\hat{\sigma}^2 = \dfrac{\hat{\mathcal{E}}_i}{n-2}$is likely to underestimate the true σ^2.
2) As a result, we are likely to overestimate R^2.
3) **OLS** estimators are still unbiased.
4) **OLS** estimators are consistent, i.e., their variances approach to zero, as the sample size gets larger and larger.
5) **OLS**estimators are no longer efficient.
6) The estimated variances of the **OLS** estimators are biased, and as a consequence, the conventional confidence intervals and tests of significance are not valid.

2.4.2.4 Detecting Serial-Autocorrelation 1. Graphical method

Plot the estimated residuals $\hat{\mathcal{E}}_t = y_t - \hat{y}_t$ against time. If we see a clustering of neighbouring residuals on one or the other side of the line $\mathcal{E} = 0$ then such clustering is a sign that the errors are autocorrelated.

Example

Using `auto.csv` data where the variables are about investment and value of outstanding shares for the years 1935-1953. Using graphical method check that for presence of autocorrelation?

```r
rm(list=ls())


auto=read.csv("auto.csv",header=T,sep=",")


fit<-lm(invest~vsh,data=auto)

plot(auto$year,fit$residuals,xlab = "Year",ylab = "Unstandardized Residuals",pch=19)

abline(h=0,lwd=2,col="red")

s<-seq(length(auto$year)-1)

arrows(auto$year[s], fit$residuals[s], auto$year[s+1], fit$residuals[s+1], col = 1:3,lwd
```

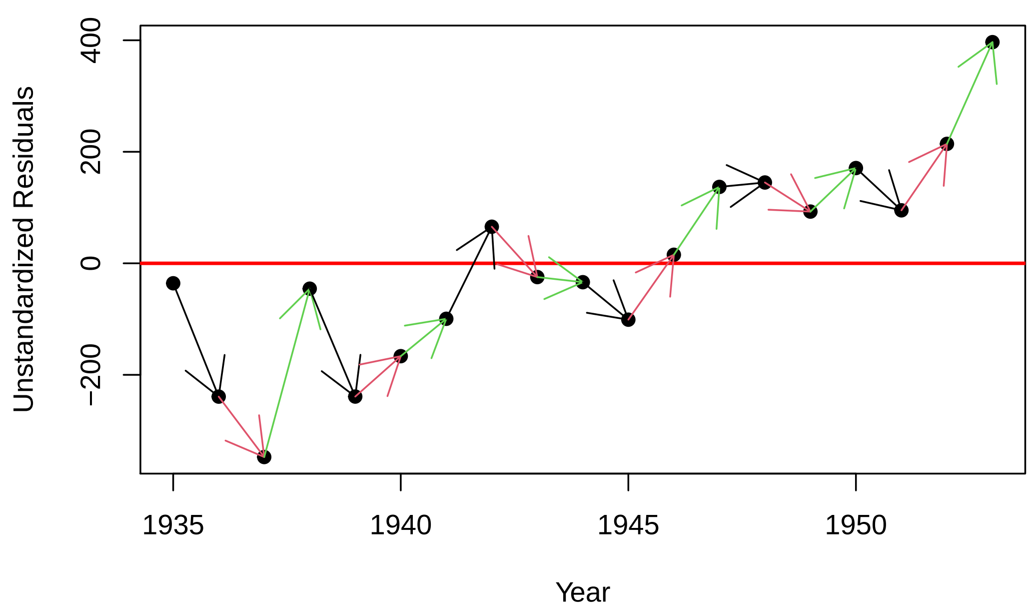

Figure 9: Plot of serial correlation

We can see a clustering of neighbouring residuals on one or the other side of the line $\hat{\mathcal{E}}_i = 0$. This might be a sign that the errors are autocorrelated. However, we do not make a final judgment until we apply formal tests of autocorrelation

2. Durbin-Watson (DW) test

The **DW** test statistic is computed as:

$$d = \frac{\sum_{t=2}^{T}(\hat{\mathcal{E}}_t - \hat{\mathcal{E}_{t-1}})^2}{\sum_{t=2}^{T}\hat{\mathcal{E}}_t^2}$$

To test of $H_0 : \rho = 0$ versus $H_A : \rho > 0$, we can use the Durbin-Watson lower (d_L) and upper (d_U) bounds (critical values).

Decision rule:

Reject H_0 if $d < d_L$

Do not reject H_0 if $d > d_U$

The test is inconclusive if $d_L < d < d_U$

Limitations of the DW test:

a) There are certain regions where the test is inconclusive.

b) The test is valid only when there is an intercept term in the model.

c) The test is invalid when lagged values of the dependent variable appear as regressors.

d) The test is valid for the **AR(1)** error scheme only.

Example

In the data `auto.csv`, check for presence of autocorrelation using DW test?

```
rm(list=ls())
auto=read.csv("auto.csv",header=T,sep=",")


dw_test<-function(x=data.frame(),dw_lower=1.180,dw_upper= 1.401){
fit<-lm(invest~vsh,data=x)
rsdl<-resid(fit)
rsdl_t=rsdl[2:19]; rsdl_t_1=rsdl[1:18] # rsdl_t=rsdl[-1];rsdl_t_1=rsdl[-19];
dw_cal<-sum((rsdl[-1]-rsdl[-19])^2)/sum(rsdl[-1]^2)
if(dw_cal< dw_lower){
   cat("Reject Ho: rho=0 "," ,since the calculated dw-test= ",dw_cal,"<",dw_lower,"=dw_l
       "that a significant presence of serial autocorrelation.")
}
if(dw_cal>dw_lower & dw_cal<dw_upper){
```

```r
    cat("The test is inconclusive")
}
if(dw_cal> dw_upper){
    cat("Do not reject Ho: rho=0")
}


}
dw_test(x=auto)
```

```
## Reject Ho: rho=0   ,since the calculated dw-test=  0.5539826 < 1.18 =dw_lower and con
##  that a significant presence of serial autocorrelation.
```

3. Breusch-Godfrey (BG) Test

Assume that the error term follows the **autoregressive scheme of order p (AR(p))** given by:

$$\mathcal{E}_t = \rho_1 \mathcal{E}_{t-1} + \rho_2 \mathcal{E}_{t-2} + ... + \rho_p \mathcal{E}_{t-p} + u_t$$

where u_t fulfills all assumption of the **CLRM**. The null hypothesis to be tested is:

$$H_0 : \rho_1 = \rho_2 = ... + \rho_p = 0$$

<u>*Steps*</u> :

1. Estimate the model:

$$Y_t = \alpha + \beta X_t + \mathcal{E}_t, ...(t = 1, 2, ..., T)$$

using **OLS** and obtain the residuals $\hat{\mathcal{E}}_t$.

2. Regress $\hat{\mathcal{E}}_t$ on X_t and $\hat{\mathcal{E}}_{t-1}, \hat{\mathcal{E}}_{t-2}, ..., \hat{\mathcal{E}}_{t-p}$ that is, run the following auxiliary regression:

$$Y_t = \alpha + \beta X_t + \rho_1 \hat{\mathcal{E}}_{t-1} + \rho_2 \hat{\mathcal{E}}_{t-2} + ... + \rho_p \hat{\mathcal{E}}_{t-p} + \xi_t$$

3. Obtain the coefficient of determination R^2 from the auxiliary regression.

4. If the sample size T is large, Breusch and Godfrey have shown that $(T-p)R^2$ follows the Chi-square (χ^2)distribution with p degrees of freedom.

Decision rule:

Reject the null hypothesis of no **AC** if $(T-p)R^2$ exceeds the tabulated value from the χ^2 distribution with p degrees of freedom for a given level of significance α.

dvantages of the BG test

a) The test is always conclusive.

b) The test is valid when lagged values of the dependent variable appear as regressors. c) The test is valid for higher order **AR** schemes (not just for **AR(1)** error scheme only).

Example

in the data `auto.csv`, check AC using BG test?

```
rm(list=ls())
auto=read.csv("auto.csv",header=T,sep=",")
T=nrow(auto)
# First apply OLS and obtain residuals
fit<-lm(invest~vsh,data=auto)
rsdl<-resid(fit)
rsdl_t=rsdl[-1];rsdl_t_1=rsdl[-19]
# to determine R2 from axiullary regression
fit1<-lm(rsdl_t~auto$vsh[-1]+rsdl_t_1)
# (T-p)R2
(bp_cal<-(T-1)*summary(fit1)$r.squared)
```

```
## [1] 9.148507
```

```
# p-value
(p_value<-pchisq(q=bp_cal, df=1,lower.tail=F))
```

```
## [1] 0.002489211
```

```r
   #x_tab
(x_tab<-qchisq(0.05, df=1,lower.tail=F))
```

```
## [1] 3.841459
```

Since the calculated test statistic(9.1485073) exceeds the critical value(3.8414588), we reject the null hypothesis $H_0 : \rho = 0$ and conclude that there is error AC(p-value=0.0024892<0.05).

4. Test based on the partial autocorrelation function (PACF) of OLS residuals

Plot the **PACF** of **OLS** residuals. If the function at lag one is outside the 95% upper or lower confidence limits, then this is an indication that the errors follow the **AR(1)** process. Higher order error processes can be detected similarly.

Example

check using pacf from auto.csv

```r
rm(list=ls())


rm(list=ls())
auto=read.csv("auto.csv",header=T,sep=",")
T=nrow(auto)
# First apply OLS and obtain residuals
fit<-lm(invest~vsh,data=auto)
summary(fit)
```

```
##
## Call:
## lm(formula = invest ~ vsh, data = auto)
##
## Residuals:
##    Min      1Q Median     3Q    Max
## -347.4 -100.3  -24.7  116.0  396.6
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -186.1598   216.2925  -0.861  0.40139
## vsh            0.1753     0.0497   3.527  0.00259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 185.1 on 17 degrees of freedom
## Multiple R-squared:  0.4225, Adjusted R-squared:  0.3885
## F-statistic: 12.44 on 1 and 17 DF,  p-value: 0.00259
```

```r
Residuals<-resid(fit)
pacf(Residuals)
```

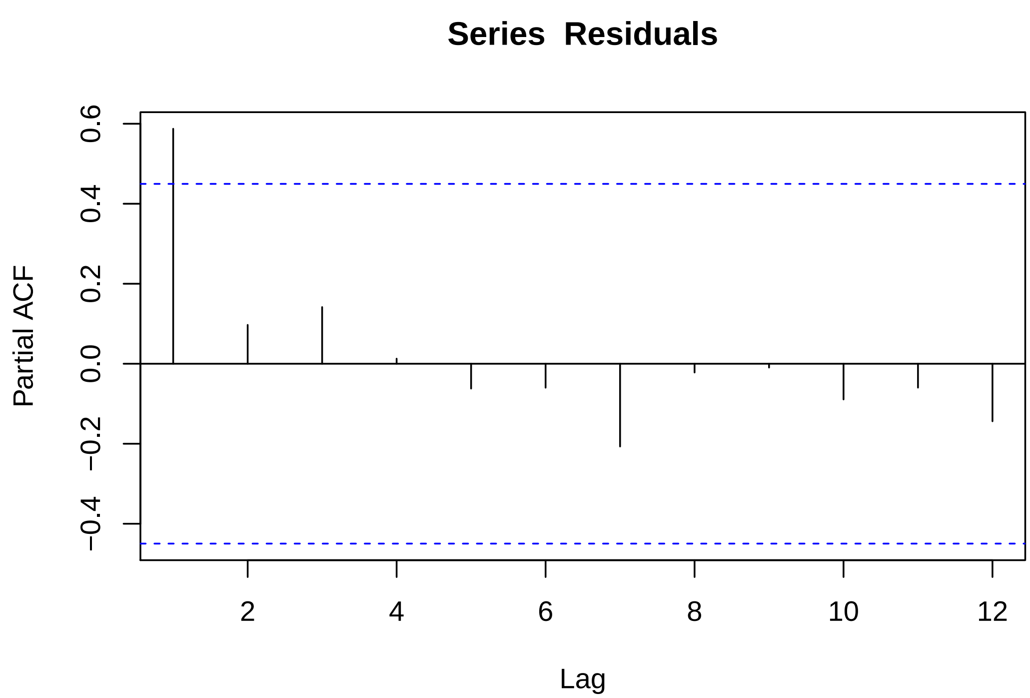

Figure 10: Partial autocorrelation function

The PACF at lag one is outside the 95% confidence limits and indicates that the errors follow the AR(1) process.

2.4.2.5 Correcting for error Autocorrelation

Consider the model:

$$Y_t = \alpha + \beta X_t + \mathcal{E}_t, ...(t = 1, 2, ..., T)...........(1)$$

where the errors are generated according to the **AR**(1) scheme:

$$\mathcal{E}_t = \rho\mathcal{E}t - 1 + u_t, |\rho| < 1$$

Here,$U_t = \mathcal{E}_t - \rho\mathcal{E}t - 1$ satisfies all assumptions of the **CLRM** (that is,$E(u_t) = 0, var(u_t) = E(u_t^2) = \sigma_u^2$ and $E(u_t u_s) = 0$ for $t \neq s$).

Suppose by applying any one of the above tests you come to the conclusion that the errors are autocorrelated. What to do next?

Lagging equation (1) by one period and multiplying throughout by ρ, we get:

$$\rho Y_{t-1} = \rho\alpha + \rho\beta X_{t-1} + \rho\mathcal{E}_{t-1}..............(2)$$

Subtracting equation (2) from equation (1), we get:

$$\begin{aligned} Y_t - \rho Y_{t-1} &= \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + (\mathcal{E}_t - \mathcal{E}t - 1) \\ let, Y_t - \rho Y_{t-1} = y_t^*, \alpha(1-\rho) = \alpha^*, \beta(X_t - \rho X_{t-1}) &= x_t^* and (\mathcal{E}_t - \mathcal{E}t - 1) = U_t \\ \Rightarrow Y_t^* &= \alpha^* + x_t^* + U_t................(3) \end{aligned}$$

The above transformation is known as the **Cochrane-Orcutt transformation**. Since $U_t = \mathcal{E}_t - \rho\mathcal{E}t - 1$ fulfils all assumption of the **CLRM**, we can apply **OLS** to equation (3) to get estimators which are **BLUE**.

Problem : The above transformation requires a knowledge of the value of ρ . Thus we need to estimate it.

Methods of estimation of ρ

a) Using the Durbin-Watson statistic.

It can be shown that as T (the sample size) gets larger, the DW statistic d approaches to $2(1-\rho)$, i.e, $d \longrightarrow 2(1-\rho)$ as $T \longrightarrow \infty$. Thus, we can use this fact to construct an estimator of ρ as:

$$\hat{\rho} = 1 - \frac{d}{2}$$

Note : This estimator is **highly inaccurate** if the sample size is **small**.

```
# from previous dw=
d=0.5539826
(rho=1-d/2)
```

```
## [1] 0.7230087
```

and hence the estimated $\hat{\rho} = 0.723$.

b) From **OLS** residuals

Regress OLS residuals $\hat{\mathcal{E}}_t$ on $\hat{\mathcal{E}}_{t-1}$ **with out a constant term**:

$$\hat{\mathcal{E}}_t = \rho\hat{\mathcal{E}}_{t-1} + u_t$$

An estimate of ρ is the estimated coefficient of $\hat{\mathcal{E}}_{t-1}$.

```
rm(list=ls())
auto=read.csv("auto.csv",header=T,sep=",")
fit<-lm(invest~vsh,data=auto)
rsdl<-resid(fit)
fit<-lm(rsdl[-1]~rsdl[-19]-1)
summary(fit)$coeff
```

```
##             Estimate Std. Error  t value    Pr(>|t|)
## rsdl[-19] 0.8049025   0.205706 3.912878 0.001119968
```

and hence the estimated $\hat{\rho} = 0.8049$.

c) Durbin's method

Run the regression of Y_t on Y_{t-1}, X_t and X_{t-1}:

$$y_t = \delta + \rho Y_{t-1} + \beta x_t + \beta\rho x_{t-1} + u_t$$

An estimator of ρ is the estimated coefficient of Y_{t-1}.

```
rm(list=ls())
auto=read.csv("auto.csv",header=T,sep=",")
y=auto$invest
x=auto$vsh
y_t1=y[-1];y_t=y[-19]
x_t1=x[-1]; x_t=x[-19]
fit<-lm(y_t~y_t1+x_t+x_t1)
summary(fit)$coeff
```

```
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 264.28691740 87.60697718  3.016734 9.239846e-03
## y_t1          0.72866112  0.06565902 11.097655 2.534061e-08
## x_t           0.04009562  0.01614676  2.483199 2.630626e-02
## x_t1         -0.07608214  0.01761327 -4.319591 7.062742e-04
```

and hence the estimated $\hat{\rho} = 0.7287$.

All tests indicated that there is serial autocorrelation. Thus, we need to apply the Cochrane-Orcutt transformation. To obtain an estimate $\hat{\rho}$ of ρ, let's use the result of regresing the OLS residuals $\hat{\mathcal{E}}_t$ on $\hat{\mathcal{E}_{t-1}}$ with out a constant term (as in b above) and is $\hat{\rho} = 0.8049$.
Then apply the following (Cochrane-Orcutt) transformation:

$$\begin{aligned} Y_t - \rho Y_{t-1} &= \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + u_t \\ &\Rightarrow Y_t^* = \alpha^* + X_t^* + u_t..........(*) \end{aligned}$$

Note that equation (*) fulfils all basic assumptions and, thus, we can estimate the parameters in this equation by an OLS procedure. Using $\hat{\rho} = 0.8049$, we obtain Y_t^* (invst_trnsf) and X_t^* (vsh_trnsf) and estimate the regression of Y_t^* on X_t^*. The results are:

```
rm(list=ls())
auto=read.csv("auto.csv",header=T,sep=",")
# lets use estiamte of rho from OLS in (b)
rho=0.8049
y=auto$invest
x=auto$vsh
y_t1=y[-1];y_t=y[-19]
invst_trnsf=y_t-rho*y_t1
x_t1=x[-1]; x_t=x[-19]
vsh_trnsf=x_t-rho*x_t1

fit_trnsf<-lm(invst_trnsf~vsh_trnsf)
summary(fit_trnsf)
```

```
##
## Call:
## lm(formula = invst_trnsf ~ vsh_trnsf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -163.37  -28.20   18.90   36.70   77.17
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.45283   18.32017   0.625 0.540688
## vsh_trnsf    0.06864    0.01598   4.294 0.000557 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 63.07 on 16 degrees of freedom
## Multiple R-squared:  0.5355, Adjusted R-squared:  0.5064
## F-statistic: 18.44 on 1 and 16 DF,  p-value: 0.0005571
```

The partial autocorrelation function of the residuals in the transformed model is shown below. It can be seen that the function lies within the upper and lower confidence limits, indicating that the autocorrelation structure has been properly dealt with.

```
Residuals<-resid(fit_trnsf)
pacf(Residuals)
```

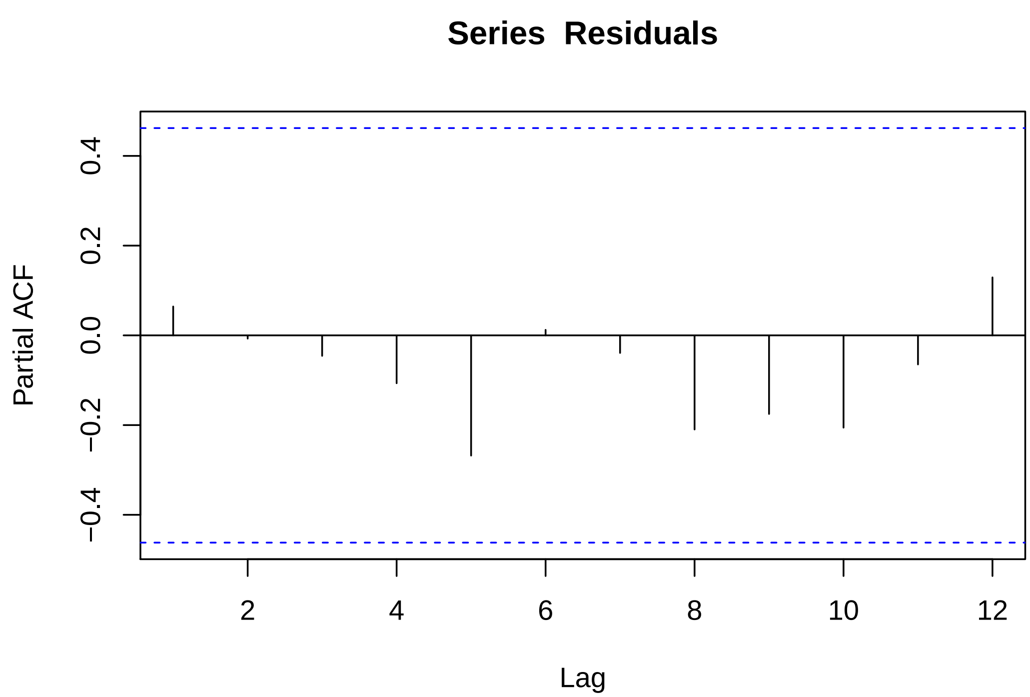

Figure 11: Partial autocorrelation function for Cochrane-Orcutt transformation

3 Non-linear Models

3.1 Introduction

Standard linear regression models are applied when the behavior of economic agents is approximated by continuous variables such as income, saving, expenditure, output, etc. But there are many situations in which the dependent variable in a regression equation simply represents a **discrete choice** assuming only a limited number of values. Models involving dependent variables of this kind are called limited (discrete) dependent variable models (also called qualitative response models).

$\underline{Examples of choices:}$

a) Whether to work or not

b) Whether to attend formal education or not

c) Choice of occupation

d) Which brand of consumer durable goods to purchase

In all of the above situations, the variables are discrete valued. In examples (a) and (b), the variables are binary (having only two possible values), whereas the variables in examples (c) and (d) are multinomial (having more than two but a finite number of distinct values). In such cases instead of standard regression models, we apply different methods of modeling and analyzing discrete data.

Qualitative choice analysis

Qualitative choice models may be used when a decision maker faces a choice among a set of alternatives meeting the following criteria:

. The number of choices if finite

. The choices are mutually exclusive (the person chooses only one of the alternatives) . The choices are exhaustive (all possible alternatives are included)

The first criterion is a binding one. We can always refine the available choices so that they can satisfy the last two criteria. Throughout our discussion we shall restrict ourselves to cases of qualitative choice where the set of alternatives is binary. For the sake of convenience the dependent variable is given a value of 0 or 1.

$\underline{Example:}$ Suppose the choice is whether to work or not. The discrete dependent variable we are working with will assume only two values 0 and 1:

$$Y_i = \begin{Bmatrix} 1 \longrightarrow if i^{th} individual is working/seeking work \\ 0 \longrightarrow if i^{th} individual is not working \end{Bmatrix}$$

where i = 1, 2, ., n. The independent variables (called factors) that are expected to affect an individual's choice may be X_1= age, X_2= marital status,X_3= gender, X_4 =education, etc. These are represented by a matrix X. If we have k factors, the vector of factors for the i^{th} individual is given by:

$$X_i(x_{1i}, x_{2i}, x_{3i}..., x_{ki})i = 1, 2, ..., n$$

The regression approach

The economic interpretation of discrete choice models is typically based on the principle of utility maximization leading to the choice of, say, A over B if the utility of A exceeds that of B. For example, let U^1 be the utility from working/seeking work and let U^0 be the utility form not working. Then an individual will choose to be part of the labour force if $U^1 - U^0 > 0$, and this decision depends on a number of factors X.

The probability that the i^{th} individual chooses alternative 1 (i.e. works/seeks work) given his/her individual characteristics X_i is:

$$P_i = Prob(Y_i = 1|X_i) = Prob[(U^1 - U^0)_i > 0] = G(X_i, \beta)$$

The vector of parameters $\beta(\beta = (\beta_1, \beta_2, ..., \beta_k)')$ measures the impact of changes in X (say, age, marital status, gender, education, etc) on the probability of labour force participation.

And the probability that the i^{th} individual chooses alternative 0 (i.e. not to work) is given by:

$$Prob(Y_i = 0|X_i) = 1 - P_i = 1 - Prob[(U^1 - U^0)_i > 0] = 1 - G(X_i, \beta)$$

Here P_i is called the **response probability** and $(1 - P_i)$ is called the **non-response probability**

The mean response of the i^{th} individual given his/her individual characteristics X_i is:

$$E(Y_i|X_i) = 1 \times Prob(Y_i = 1|X_i) + 0 \times Prob(Y_i = 0|X_i) = 1 \times G(X_i, \beta) + 0 \times G(X_i, \beta) = G(X_i, \beta)$$

The problem is thus to choose the appropriate form of $G(X_i, \beta)$.

Case 1: The linear probability model

The linear probability model defines $G(X_i, \beta)$ as: $G(X_i, \beta)$=$X_i\beta$

The regression model is thus:

$$Y_i = X_i\beta + \mathcal{E}_i............(3.1.1)$$

This is the usual linear regression model. The drawbacks of this model are:

1. The right hand side of equation (3.1.1) is a combination of discrete and continuous variables while the left hand side variable is discrete.

2. Usually we arbitrarily (or for convenience) use 0 and 1 for Y_i . If we use other values for Y_i , say 3 and 4, β will also change even if the vector of factors X_i remains unchanged.

3.$\mathcal{E}_i$ assumes only two values:

if $Y_i = 1$ then $\mathcal{E}_i = 1 - X_i\beta$ (with prob. P_i) if $Y_i = 0$ then $\mathcal{E}_i = -X_i\beta$ (with prob. $1 - P_i$)

Consequently, $\mathcal{E}_i$ is not normally distributed but rather has a discrete (binary) probability distribution defined as:

$\mathcal{E}_i$	Probability
$1 - X_i\beta$	P_i
$-x_i\beta$	$1 - P_i$

4. The expectation (mean) of $\mathcal{E}_i$ conditional on the exogenous variables X_i is (from the above table):

$$E(\mathcal{E}_i|x_i) = (1 - X_i\beta)P_i + (-x_i\beta)(1 - P_i) = P_i - X_i\beta$$

Setting this mean to zero as in the classical regression analysis means:

$$E(\mathcal{E}_i|x_i) = 0 \Rightarrow P_i = X_i\beta$$

So the original model (1) becomes:

$$Y_i = P_i + \mathcal{E}_i \Rightarrow \mathcal{E}_i = Y_i - P_i$$

That is, the binary (discrete) disturbance term $\mathcal{E}_i$ is equal to the difference between a binary variable Y_i and a continuous response probability P_i. Clearly this does not make sense.

5. We know that the probability of an event is always a number between 0 and 1 (inclusive). But here we can see that:

$$P_i = Prob(Y_i = 1|X_i) = X_i\beta$$

i .e.,P_i can take on any value (even negative numbers) leading to nonsense probabilities.

Case 2: The latent regression approach

The outcome (or observed occurrence) of a discrete choice may be considered to be an indicator of an underlying, unobservable continuous variable which may be called 'propensity to choose a given alternative'. Such a variable is characterized by the existence of a threshold, where crossing a threshold means switching from one alternative to another. For instance, a married woman's propensity to join the labour force may be directly related to the wage that she may receive in the market, which in turn may depend on her level of education, experience, etc. Whether she actually joins the work force or not is likely to depend on whether her market wage does or does not exceed her threshold or 'reservation' wage.

Example: Let an individual's propensity to enter the labour force (or to work) be an unobservable (latent) variable Y_i^* such that:

$$Y_i^* = X_i\beta + \mathcal{E}_i = X_{1i}\beta_1 + X_{2i}\beta_2 + ... + X_{ki}\beta_k + \mathcal{E}_i$$

What we actually observe here is the individual's decision to work or not (Y_i) and the set of individual factors such as age, sex, marital status, level of education, experience, etc $(X_i(X_{1i}, X_{2i}, X_{2i}, ..., X_{ki}))$ where:

$$Y_i = \left\{\begin{matrix} 1 \longrightarrow & if & Y_i^* > c \\ 0 \longrightarrow & if & Y_i^* \leq c \end{matrix}\right\}$$

where c is a floor for the propensity variable (or a threshold). The obability of labour force participation of the i^{th} individual given his/her personal background X_i is:

$$Prob(Y_i = 1|X_i) = Prob(Y_i^* > c|X_i) = Prob(X_i\beta + \mathcal{E}_i > c) = Prob(\mathcal{E}_i > c - X_i\beta|X_i)$$

Similarly,

$$\begin{aligned} Prob(Y_i = 0|X_i) &= Prob(Y_i^* \leq c|X_i) \\ &= 1 - Prob(\mathcal{E}_i > c - X_i\beta|X_i) \\ &= 1 - Prob(Y_i = 1|X_i) \end{aligned}$$

The expectation (mean) of Y_i conditional on the exogenous variables X_i is:

$$\begin{aligned} E(Y_i|X_i) &= 1 \times Prob(Y_i = 1|X_i) + 0 \times Prob(Y_i = 0|X_i) \\ &= Prob(Y_i = 1|X_i) \\ &= Prob(\mathcal{E}_i > c - X_i\beta|X_i) \end{aligned}$$

The general form of the latent regression model is:

$$E(Y_i|X_i) = Prob(Y_i = 1|X_i) = Prob(\mathcal{E}_i > c - X_i\beta|X_i) = G(X_i\beta)$$

We can predict how the probability of participating in the labour force (Y_i) will change as individuals' characteristics X_i change if we choose either the correct functional form of $G(X_i\beta)$ or the appropriate probability distribution of $\mathcal{E}_i$.

Note

The **cumulative distribution function(CDF)** of a random variable X is defined as:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

where $f_X(x)$ is the probability density function(PDF) of X. If the functional form of f_X is known, then we can evaluate $P(X \leq x)$ by integration. The PDF can be obtained from the CDF as:

$$\frac{d(f_X(x))}{dx} = f_X(x)$$

Now, if we have a latent regression model of the form:

$$Prob(Y_i = 1|X_i) = Prob(\mathcal{E}_i < c - X_i\beta|X_i)$$

and if the functional form or the probability distribution of $\mathcal{E}_i$, say,$f_{\mathcal{E}}$ is known, then the response probability can be evaluated since:

$$Prob(\mathcal{E}_i < c - X_i\beta|X_i) = F_{\mathcal{E}}(X_i\beta) = \int_{-\infty}^{X_i\beta} f_{\mathcal{E}}(t)dt$$

3.2 Intrinsically Non-linear Models

3.2.1 The probit and logit models

Setting $G(X_i\beta)$ to be the normal distribution or assuming that $\mathcal{E}_i$ follows the **normal distribution** (that is, $f_{\mathcal{E}}$ is the normal distribution) gives rise to the **probit model**. Since the normal distribution is symmetric about its mean, we have:

$$Prob(\mathcal{E}_i > -X_i\beta|X_i) = Prob(\mathcal{E}_i < X_i\beta|X_i)$$

Thus, the probability of participating in the labour force ($Y_i = 1$) given an individual's characteristics X_i is given by:

$$\begin{aligned} Prob(Y_i = 1|X_i) &= Prob(\mathcal{E}_i < X_i\beta|X_i) \\ &= \int_{-\infty}^{X_i\beta} f_{\mathcal{E}}(t)dt = \int_{-\infty}^{X_i\beta} \phi(t)dt = \Phi(X_i\beta) \end{aligned}$$

where $\phi(\cdot)$ is the standard normal PDF and $\Phi(\cdot)$ is the standard normal CDF. Thus, we can read the probabilities from the standard normal distribution table.
Setting $G(X_i\beta)$ to be the **logistic distribution** or assuming that $\mathcal{E}_i$ follows the **logistic distribution** gives rise to the **logit model**. The logistic distribution function is given by:

$$Prob(\mathcal{E}_i < X_i\beta) = \Lambda(X_i\beta) = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}}$$

Here the response probability $Prob(Y_i = 1/X_i)$ is evaluated as:

$$\begin{aligned} P_i = Prob(Y_i = 1|X_i) &= Prob(\mathcal{E}_i > -X_i\beta|X_i) \quad \textit{symmetery} \\ &= 1 - Prob(\mathcal{E}_i < -X_i\beta|X_i) \\ &= 1 - \Lambda(-X_i\beta) \\ &= 1 - \frac{e^{-X_i\beta}}{1 + e^{-X_i\beta}} \\ &= \frac{e^{X_i\beta}}{1 + e^{X_i\beta}} \end{aligned}$$

Similarly, the non-response probability is evaluated as:

$$\begin{aligned} 1 - P_i &= Prob(Y_i = 0|X_i) \\ &= 1 - \frac{e^{X_i\beta}}{1 + e^{X_i\beta}} = \frac{1}{1 + e^{X_i\beta}} \end{aligned}$$

Note that the response and non- response probabilities both lie in the interval $[0 \; , \; 1]$, and hence, are interpretable.

For the logit model, the ratio:

$$\begin{aligned} \frac{P_i}{1 - P_i} = \frac{Prob(Y_i = 1|X_i)}{Prob(Y_i = 0|X_i)} &= \frac{\frac{e^{X_i\beta}}{1+e^{X_i\beta}}}{\frac{1}{1+e^{X_i\beta}}} \\ &= e^{X_i\beta} = e^{X_{1i}\beta + X_{2i}\beta_2 + ... + X_{ki}\beta_k} \end{aligned}$$

is the **ratio of the odds** of $Y_i = 1$ against $Y_i = 0$. The natural logarithm of the odds (**log-odds**) is:

$$\ln\left[\frac{P_i}{1 - P_i}\right] = X_i\beta = X_{1i}\beta + X_{2i}\beta_2 + ... + X_{ki}\beta_k$$

Thus, the log-odds is a linear function of the explanatory variables. The above transformation has certainly helped the popularity of the logit model. Note that for the linear probability model it is P_i that is assumed to be a linear function of the explanatory variables.

Example : Suppose $P_i = Prob(Y_i = 1/X_i)$ is the probability that the i^{th} individual chooses to work given his/her individual characteristics X_i and suppose that the odds is calculated to be $e^{X_i\beta}$. The interpretation of this is that the probability of joining the labour force is **twice as likely** as staying at home given the individual characteristics X_i.

3.3 Estimation of Intrinsically non-linear Models

3.3.1 The maximum likelihood estimation

Note(**the Bernoulli distribution**)

Let Y_i be a random variable which can take on the value 1 with probability P_i (called probability of success), and the value 0 with probability $(1-P_i)$ (called probability of failure). If the observations are independent, then the probability distribution of Y_i is given by:

$$Prob(Y_i = y_i) = P_i^{yi}(1-P_i)^{1-y_i} \qquad i = 1, 2, ..., n........(3.4.1)$$

We have seen earlier that:

$$Prob(Y_i = 1|X_i){=}G(X_i\beta) \text{ and } Prob(Y_i = 0|X_i) = 1 - G(X_i\beta)$$

Thus, each observation Y_i may be treated as a single draw from a Bernoulli distribution with probability of success $(Y_i = 1)$ equal to $G(X_i\beta)$ and probability of failure $(Y_i = 0)$ equal to $1 - G(X_i\beta)$. Using equation (3.4.1), the probability distribution of Y_i is given by:

$$Prob(Y_i = 1|X_i) = [G(X_i\beta)]^{y_i}[1 - G(X_i\beta)]^{1-y_i} \qquad i = 1, 2, ..., n.$$

Since the observations are independent, the **likelihood function** is simply the product of the individual probabilities of the $Y_i's$, i=1, 2, . . ., n, that is,

$$\begin{aligned} L &= Prob(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n/X) \\ &= Prob(Y_1 = y_1) \times Prob(Y_2 = y_2) \times ... \times Prob(Y_n = y_n) \\ &= \Pi_{i=i}^{n}[G(X_i\beta)]^{y_i}[1 - G(X_i\beta)]^{1-y_i} \end{aligned}$$

Taking natural logarithms, we get the **log-likelihood function**:

$$\ln(L) = \sum_{i=1}^{n}[y_i \ln[G(X_i\beta)] + (1 - y_i)\ln[1 - G(X_i\beta)]]$$

Taking the derivative with respect to β and setting the result to zero we have:

$$\frac{\partial \ln(L)}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{y_i g_i}{G_i} + (1 - y_i) \frac{-g_i}{1 - G_i} \right] X_i = 0 ....(3.4.2)$$

where $G_i = G(X_i \beta)$ and $g_i = \frac{\partial G(X_i \beta)}{\partial \beta}$.

These equations are highly non-linear, and we need to apply numerical methods (numerical optimization methods) to obtain the solutions.

If we consider a model with only one explanatory variable X, and if $G(X_i \beta)$ is the logit model:

$$G(X_i \beta) = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}} = \frac{e^{\alpha + X_i \beta}}{1 + e^{\alpha + X_i \beta}}$$

then log likelihood function is given by:

$$\begin{aligned}
\ln(L) &= \sum_{i=1}^{n} \left[y_i \ln[\left(\frac{e^{\alpha + X_i \beta}}{1 + e^{\alpha + X_i \beta}} \right) + (1 - y_i) \ln \left(1 - \frac{e^{\alpha + X_i \beta}}{1 + e^{\alpha + X_i \beta}} \right) \right] \\
&= \sum_{i=1}^{n} \left[y_i \ln[\left(\frac{e^{\alpha + X_i \beta}}{1 + e^{\alpha + X_i \beta}} \right) + (1 - y_i) \ln \left(\frac{1}{1 + e^{\alpha + X_i \beta}} \right) \right] \\
&= \sum_{i=1}^{n} \left[y_i \ln[e^{\alpha + X_i \beta}] - y_i \ln \left(1 + e^{\alpha + X_i \beta} \right) - (1 - y_i) \ln \left(1 + e^{\alpha + X_i \beta} \right) \right] \\
&= \sum_{i=1}^{n} \left[y_i \ln[e^{\alpha + X_i \beta}] - \ln \left(1 + e^{\alpha + X_i \beta} \right) \right] \\
&= \sum_{i=1}^{n} \left[y_i \ln (\alpha + X_i \beta) - \ln \left(1 + e^{\alpha + X_i \beta} \right) \right]
\end{aligned}$$

The first order conditions are:

$$\frac{\partial \ln(L)}{\partial \beta} = 0 \Rightarrow \sum_{i=1}^{n} \left[y_i - \frac{e^{\tilde{\alpha} + \tilde{\beta} X_i}}{1 + e^{\tilde{\alpha} + \tilde{\beta} X_i}} \right] X_i = 0 ......(3.4.3a)$$

$$\frac{\partial \ln(L)}{\partial \alpha} = 0 \Rightarrow \sum_{i=1}^{n} \left[y_i - \frac{e^{\tilde{\alpha} + \tilde{\beta} X_i}}{1 + e^{\tilde{\alpha} + \tilde{\beta} X_i}} \right] = 0 ....(3.4.3b)$$

These two equations can be solved for $\tilde{\beta}$ and $\tilde{\alpha}$. Since both equations are non-linear functions of $\tilde{\beta}$ and $\tilde{\alpha}$, the solutions are obtained using numerical methods.

In the classical linear regression model we have $G(X_i \beta) = \alpha + \beta X_i$ and equations (3.4.3a) and

(3.4.3b) become:

$$\sum_{i=1}^{n}[Y_i - (\hat{\alpha} + \hat{\beta}X_i)]X_i = 0 \Longrightarrow \sum_{i=1}^{n} Y_i X_i = \hat{\alpha}\sum_{i=1}^{n} X_i + \hat{\beta}\sum_{i=1}^{n} X_i^2$$
$$\sum_{i=1}^{n}[Y_i - (\hat{\alpha} + \hat{\beta}X_i)] = 0 \Longrightarrow \sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} X_i$$

These two equations are simply the **normal equations** which can easily be solved to get the ordinary least squares (OLS) estimators of β and α.

Example

4 Simultaneous Equations Models (SEM)

Objective of the unit:

- Define the structural form of a SEM and explain its usefulness.
- Explain why the least squares estimation is not appropriate for SEM.
- Explain the difference between exogenous and endogenous variables.
- Define identification problem in a simultaneous equations models.
- Define the reduced form of a SEM and explain its usefulness.
- Explain why it is acceptable to estimate reduced-form equations by least squares.
- Describe the instrumental variable estimation procedure for estimating an equation in a simultaneous equations model, and explain how it resolves the estimation problem for least squares
- Describe the two-stage least squares estimation procedure for estimating an equation in a simultaneous equations model, and explain how it resolves the estimation problem for least squares

4.1 Introduction to SEM

In the previous lessons, we were concerned exclusively with single-equation models, i.e., models in which there was a single dependent variable Y and one or more explanatory variables, the X's. In such models the emphasis was on estimating and/or predicting the average value of Y conditional upon the fixed values of the X variables. The cause-and-effect relationship, if any, in such models therefore ran from the X's to the Y(i.e., unidirectional). However, there are situations where there is a two-way flow of influence among economic variables. This occurs if Y is determined by the X's, and some of the X's are, in turn, determined by Y. In short, there is a two-way, or simultaneous, relationship between Y and (some of) the X's, which makes the distinction between dependent and explanatory variables of dubious value.

Simultaneous equations models also differ from most of the econometric models we have considered so far, because they consist of a set of equations. For example, price and quantity are determined by the interaction of two equations, one for supply and the other for demand. Simultaneous equations models, which contain more than one dependent variable and more than one equation, require special statistical treatment. The least squares estimation procedure is not appropriate in these models, and we must develop new ways to obtain reliable estimates of economic parameters.

Example : At the macro level, aggregate consumption expenditure depends on aggregate disposable income; aggregate disposable income depends upon the national income and taxes imposed by the government; national income depends on aggregate consumption expenditure of the economy. Disregarding these sequences of relationship, if we estimate a single equation of, say, aggregate consumption on disposable income, then the estimates will be biased and inconsistent.

Example: A simple model of the market for a given commodity may involve a supply and demand function:

$$Q_t = \alpha_1 + \alpha_2 P_t + \alpha_3 P_t + U_{1t} \quad (demand) \qquad ....(4.1a)$$

$$Q_t = \beta_1 + \beta_2 P_t + U_{2t} \qquad (supply) \qquad ....(4.1b)$$

where Q is the equilibrium quantity exchanged on the market, P is equilibrium price, Y is income of consumers, and u_{1t} and u_{2t} are the disturbance terms. We also have $EU_{1t}^2 = \alpha_1^2$,$EU_{2t}^2 = \alpha_2^2$ and $EU_{1t}U_{2t} = \alpha_{12}$.

Suppose we are interested in the effect of P on Q. Can we toss out the second equation and estimate the first equation alone using OLS?

1. The equilibrium price and quantity are determined in the market by the intersection of supply and demand curves. Therefore, we can not determine equilibrium price by solving the demand equation independently.
2. A shift in the demand function produces a change in both equilibrium price and quantity if the supply curve has an upward slope.

Equations (4.1a,4.1b) are called the **structural form** of the model understudy. These equations can be solved for the 'endogenous' variables to give:

$$Q_t = \left(\frac{\alpha_2\beta_1 - \alpha_1\beta_2}{\alpha_2 - \beta_2}\right) - \left(\frac{\alpha_3\beta_2}{\alpha_2 - \beta_2}\right) Y_t + \left(\frac{-\beta_2 U_{1t} + \alpha_2 U_{2t}}{\alpha_2 - \beta_2}\right) \qquad ....(4.2a)$$

$$P_t = \left(\frac{-\alpha_1 + \beta_1}{\alpha_2 - \beta_2}\right) - \left(\frac{\alpha_3}{\alpha_2 - \beta_2}\right) Y_t + \left(\frac{-U_{1t} + U_{2t}}{\alpha_2 - \beta_2}\right) \qquad ....(4.2b)$$

The solution given by equations (4.2a) and (4.2b) is called the **reduced form** of the model. The reduced form equations show explicitly how the "endogenous" variables are **jointly dependent** on the "predetermined" variables and the disturbances of the system.

Now from equation (4.2b) we have:

$$
\begin{aligned}
E(P_tU_{1t}) &= \left(\frac{-\alpha_1+\beta_1}{\alpha_2-\beta_2}\right)E(U_{1t}) - \left(\frac{\alpha_3}{\alpha_2-\beta_2}\right)Y_tE(U_{1t}) + \left(\frac{-E(U_{1t}^2)+E(U_{1t}U_{2t})}{\alpha_2-\beta_2}\right)\\
&= \left(\frac{-\alpha_1+\beta_1}{\alpha_2-\beta_2}\right)(0) - \left(\frac{\alpha_3}{\alpha_2-\beta_2}\right)Y_t(0) + \left(\frac{-\sigma_1^2+\sigma_{12}}{\alpha_2-\beta_2}\right)\\
&= \left(\frac{-\sigma_1^2+\sigma_{12}}{\alpha_2-\beta_2}\right) \neq 0
\end{aligned}
$$

Similarly, it can be shown that:

$$E(P_tU_{2t}) = \left(\frac{-\sigma_{12}+\sigma_2^2}{\alpha_2-\beta_2}\right) \neq 0$$

Thus, in the demand equation (4.1a): $Q_t = \alpha_1 + \alpha_2P_t + \alpha_3Y_t + U_{1t}$, the variable P_t that appears as an independent or 'exogenous' variable is correlated with the disturbance term u_{1t}, and consequently, estimation of the demand equation using OLS leads to **biased** and **inconsistent estimators** of the parameters (refer to Unit 9 for details). This is referred to as simultaneity bias.

he solution is to bring the supply function into the picture and estimate the supply and demand functions simultaneously. Such models are known as **simultaneous equations models**.

Example : Wage-price model

$$
\begin{aligned}
W_t &= \alpha_0 + \alpha_1U_t + \alpha_2P_t + U_{1t} \qquad (\textit{wage equation}) \text{.............}(4.3a)\\
P_t &= \beta_0 + \beta_1W_t + \beta_2R_t + \beta_3M_t + U_{2t} \qquad (\textit{price equation}) \qquad \text{........}(4.3b)
\end{aligned}
$$

where W is rate of change in money wage, U is unemployment rate (in percentage), P is rate of change in prices, R is rate of change in cost of capital, and M is money supply. Here the price variable P enters into the wage equation (4.3a) and the wage variable W enters into the price equation (4.3b). Thus, these two variables are jointly dependent to each other, and estimation of the two equations individually by OLS yields biased and inconsistent estimators.

Note : 1. **Endogenous variables** are variables that are jointly determined by the economic model. (or are determined by the exogenous variables).

2. **Exogenous variables** are determined outside of the model and independently of the endogenous variables.

3. **Predetermined variables** are exogenous variables, lagged exogenous variables and lagged endogenous variables. Predetermined variables are non-stochastic and hence independent of the disturbance terms.

4.1.1 Structural form and reduced form of simultaneous equations model (SEM)

Consider the simple Keynesian model of income determination:

$C_t = \beta_0 + \beta_1 P_t + u_t \quad 0 < \beta_1 < 1$ $\qquad (Consumption\ function)........(4.4a)$

$Y_t = C_t + I_t$ $\qquad (Income\ identity)............(4.4b)$

where C is consumption expenditure, Y is income, I is investment (assumed to be exogenous). The above model is said to be the structural form of the SEM, and the parameters β_0 and β_1 are said to be structural parameters. Substituting (4.4a) in place of C in (4.4b) we get:

$Y_t = \beta_0 + \beta_1 Y_t + I_t + u_t$

$\Rightarrow Y_t = \frac{\beta_0}{1-\beta_1} + \frac{1}{1-\beta_1} I_t + \frac{1}{1-\beta_1} u_t$

$\Rightarrow Y_t = \pi_0 + \pi_1 I_t + w_t$ $\qquad$ ....(4.5)

Note that equation (4.5) is expressed solely as a function of the exogenous variable I_t and the disturbance term. It is referred to as the reduced form of the SEM, and the parameters π_0 and π_1 are said to be reduced form parameters. Note that the exogenous variable I_t is not correlated with the disturbance term, and hence, we can apply OLS to the reduced form equation to obtain consistent estimators of π_0 and π_1.

In general, the structural form of a simultaneous system of equations can be described as:

$$
\begin{aligned}
\beta_{11}Y_{1t} + \beta_{12}Y_{2t} + ... + \beta_{1G}Y_{Gt} + \gamma_{11}X_{1t} + \gamma_{12}X_{2t} + ... + \gamma_{1k}X_{kt} &= u_{1t} \\
\beta_{21}Y_{1t} + \beta_{22}Y_{2t} + ... + \beta_{2G}Y_{Gt} + \gamma_{21}X_{1t} + \gamma_{22}X_{2t} + ... + \gamma_{2k}X_{kt} &= u_{2t} \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \\
\beta_{G1}Y_{1t} + \beta_{G2}Y_{2t} + ... + \beta_{GG}Y_{Gt} + \gamma_{G1}X_{1t} + \gamma_{G2}X_{2t} + ... + \gamma_{Gk}X_{kt} &= u_{Gt}
\end{aligned}
$$

where the Y's are endogenous variables, X's are predetermined variables, and the u's are stochastic disturbances.

- The β's and γ's are the structural coefficients

- There are G endogenous and K predetermined variables in the system.
- Not all endogenous and predetermined variables will appear in every equation (that is, some of β's and γ's will be zero).
- In each equation, one of the β's is taken to be unity, that is, one of the endogenous variables serves as the 'dependent' variable when the equation is written out as a standard regression equation.
  - Some of the equations may be identities, that is, their coefficients are known and they contain no stochastic disturbance.

The above model in matrix form is:

$$\begin{bmatrix} \beta_{11} & \beta_{12} & . & . & . & \beta_{1G} \\ \beta_{21} & \beta_{22} & . & . & . & \beta_{2G} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \beta_{G1} & \beta_{G2} & . & . & . & \beta_{GG} \end{bmatrix} \begin{bmatrix} Y_{1t} \\ Y_{2t} \\ . \\ . \\ . \\ Y_{Gt} \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} & . & . & . & \gamma_{1k} \\ \gamma_{21} & \gamma_{22} & . & . & . & \gamma_{2k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \gamma_{G1} & \gamma_{G2} & . & . & . & \gamma_{Gk} \end{bmatrix} \begin{bmatrix} X_{1t} \\ X_{2t} \\ . \\ . \\ . \\ X_{kt} \end{bmatrix} = \begin{bmatrix} u_{1t} \\ u_{2t} \\ . \\ . \\ . \\ u_{Gt} \end{bmatrix}$$

$$\Rightarrow BY_t \quad + \quad \mathit{\Gamma} X_t = U_t$$

$$\Rightarrow -B^{-1}\mathit{\Gamma} X_t + B^{-1}U_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.6)$$

The reduced form of the system is obtained by expressing the Y's solely as a function of the predetermined variables X's and the disturbance terms:

$$Y_{1t} = \pi_{11}X_{1t} + \pi_{12}X_{2t} + \ldots.\pi_{1k}X_{kt} + v_{1t}$$
$$Y_{2t} = \pi_{21}X_{1t} + \pi_{22}X_{2t} + \ldots.\pi_{2k}X_{kt} + v_{2t}$$
$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$
$$Y_{Gt} = \pi_{G1}X_{1t} + \pi_{G2}X_{2t} + \ldots.\pi_{Gk}X_{kt} + v_{Gt}$$

The above model in matrix form is:

$$\underbrace{\begin{bmatrix} Y_{1t} \\ Y_{2t} \\ \vdots \\ Y_{Gt} \end{bmatrix}} = \underbrace{\begin{bmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1k} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{G1} & \pi_{G2} & \cdots & \pi_{Gk} \end{bmatrix}} \underbrace{\begin{bmatrix} X_{1t} \\ X_{2t} \\ \vdots \\ X_{Gt} \end{bmatrix}} + \underbrace{\begin{bmatrix} v_{1t} \\ v_{2t} \\ \vdots \\ v_{Gt} \end{bmatrix}}$$

$=Y_t \qquad = \prod \qquad X_t \qquad V_t$

$\Rightarrow Y_t = \prod X_t + V_t$.............................(4.7)

Comparing (4.6) and (4.7), the relationship between the structural and reduced form parameters is: $\prod = -B^{-1}\Gamma$ and $V_t = B^{-1}U_t$

Example: Suppose we have the following system of equations:

$q_t = a_1 + b_2 p_t + c_1 y_t + d_1 R_t + U_{1t} \qquad (demand\ function)$

$q_t = a_2 + b_2 p_t + U_{2t} \qquad (supply\ function)$

where q is equilibrium quantity exchanged on the market, p is equilibrium price, y is income of consumers, R is the amount of rainfall (Note: rainfall affects demand, i.e., if there is rain, people do not go shopping) and u_1 and u_2 are the error terms. This structural form model can be re-written as:

$q_t - b_1 - p_t - a_1 - c_1 y_t - d_1 R_t = u_{1t}$

$q_t - b_2 - p_t - a_2 - (0)y_t - (0)R_t = u_{2t}$

$$\Rightarrow \begin{bmatrix} 1 & -b_1 \\ 1 & -b_2 \end{bmatrix} \begin{bmatrix} q_t \\ p_t \end{bmatrix} + \begin{bmatrix} -a_1 & -c_1 & -d_1 \\ -a_2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ y_t \\ R_t \end{bmatrix} = \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} q_t \\ p_t \end{bmatrix} = -\begin{bmatrix} 1 & -b_1 \\ 1 & -b_2 \end{bmatrix}^{-1} \begin{bmatrix} -a_1 & -c_1 & -d_1 \\ -a_2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ y_t \\ R_t \end{bmatrix} + \begin{bmatrix} 1 & -b_1 \\ 1 & -b_2 \end{bmatrix}^{-1} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \text{.............(4.1.a)}$$

The reduced form equations can be written as:

$q_t = \pi_1 + \pi_2 y_t + \pi_3 R_t + v_{1t}$

$p_t = \pi_4 + \pi_5 y_t + \pi_6 R_t + v_{1t}$

$$\Rightarrow \begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} \pi_1 & \pi_2 & \pi_3 \\ \pi_4 & pi_5 & \pi_6 \end{bmatrix} \begin{bmatrix} 1 \\ y_t \\ R_t \end{bmatrix} + \begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix} \text{..................(4.1.b)}$$

Comparing (4.1.a) and (4.1.b), we can see that:

$$\begin{bmatrix} \pi_1 & \pi_2 & \pi_3 \\ \pi_4 & pi_5 & \pi_6 \end{bmatrix} = - \begin{bmatrix} 1 & -b_1 \\ 1 & -b_2 \end{bmatrix}^{-1} \begin{bmatrix} -a_1 & -c_1 & -d_1 \\ -a_2 & 0 & 0 \end{bmatrix}$$

=

$\frac{1}{b_2-b_1}\begin{bmatrix}-b_2 & b_1\\ -1 & 1\end{bmatrix}\begin{bmatrix}-a_1 & -c_1 & -d_1\\ -a_2 & 0 & 0\end{bmatrix}$
$=\frac{1}{b_2-b_1}\begin{bmatrix}a_1b_2-b_1a_2 & c_1b_2 & d_1b_2\\ a_1-a_2 & c_1 & d_1\end{bmatrix}$
Thus,

$$\pi_1 = \frac{a_1b_2 - b_1a_2}{b_2 - b_1}, \pi_2 = \frac{c_1b_2}{b_2 - b_1},$$
$$\pi_3 = \frac{d_1b_2}{b_2 - b_1}, \pi_4 = \frac{a_1 - a_2}{b_2 - b_1},$$
$$\pi_5 = \frac{c_1}{b_2 - b_1}, and$$
$$\pi_6 = \frac{d_1}{b_2 - b_1}$$

Note that since the reduced form of the system is obtained by expressing the endogenous variables solely as a function of the predetermined variables, OLS yields consistent estimators of the reduced form parameters. Thus, the OLS estimators $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \hat{\pi}_4, \hat{\pi}_5, \hat{\pi}_6$ from equation (4.1.b) above are unbiased and consistent.

4.2 The identification problem

Consider the supply and demand equations in the example above. We stated that the parameters of the reduced form model (that is, the π's) can be estimated using OLS consistently.

Question: Can we always recover the parameters of the structural equations (that is,a1,b1,c1,a2,b2) uniquely from the π's? In other words, can we always estimate the structural coefficients via the reduced form coefficients? This leads us to the concept of identification. Identification is a problem of model formulation rather than of model estimation or appraisal. We say a model is identified if it is in a unique statistical form, enabling unique estimators of its parameters to be subsequently made from sample data.

Example

In the supply and demand equations in the example above, we can see that:

$\pi_2 = \frac{c_1b_2}{b_2-b_1}$ and $pi_5 = \frac{c_1}{b_2-b_1} \Rightarrow b_2 = \frac{\pi_2}{\pi_5}$

$\pi_3 = \frac{d_1b_2}{b_2-b_1}$ and $\pi_6 = \frac{d_1}{b_2-b_1} \Rightarrow b_2 = \frac{\pi_3}{\pi_6}$

That is, the slope of the supply equation $q_t = a_2 + b_2 p_t + u_{2t}$ **can not be uniquely estimate.** Thus, the model is not identified. This is what we call an **identification problem.**

Note: (Status of identification)

In econometric theory, three possible situations of identifiability can arise: equation under consideration is exactly identified, over identified or under identified.

1) If there is a one to one correspondence between the reduced form and structural form parameters, then we have exact identification, that is, there is a unique solution for the structural parameters in terms of the reduced form parameters.

2) If the number of reduced form parameters exceeds the number of structural parameters, then we have over identification (no unique solution). Here there is more than sufficient information regarding the equation under consideration.

3) If the number of reduced form parameters is less than the number of structural parameters, then we have under identification (no solution). Here there is no sufficient information regarding the equation under consideration.

4.3 Conditions for identification

4.3.1 The order condition for identification

Let G be the total number of endogenous variables in the system and let k be the total number of variables (both endogenous and predetermined) missing from the equation under consideration. Then if:

a) k = G-1, the equation is exactly identified.

b) $k > G - 1$, the equation is over identified.

c) $k < G - 1$, the equation is under identified

This is known as the order condition for identification. It is a necessary but not sufficient condition for the identification status of an equation.

Example

Wage-price model

- $W_t = \alpha_0 + \alpha_1 U_t + \alpha_2 P_t + u_{1t}$ $\quad$ $(wage\ equation)$
- $P_t = \beta_0 + \beta_1 W_t + \beta_2 R_t + \beta_3 M_t + u_{2t}$ $\quad$ $(price\ equation)$

Here U, R and M are predetermined while W and P are endogenous variables. Thus G = 2.

a) Consider the wage equation. The variables R and M are missing from this equation. Thus, k = 2. The equation is over identified since $k = 2 > 1 = G\breve{\ }1$.

b) Consider the price equation. The variable U is missing from this equation. Thus, k=1. The equation is exactly identified since k = 1 = G – 1.

Example

The following represents a highly simplified model of the economy:

$C_t = \alpha_0 + \alpha_1 Y_t + \alpha_2 C_{t-1} + u_{1t}$ $\qquad (consumption)$

$I_t = \beta_0 + \beta_1 r_t + \beta_2 I_{t-1} + u_{2t}$ $\qquad (investment)$

$r_t = \gamma_0 + \gamma_1 Y_t + \gamma_2 M_t + u_{3t}$ $\qquad (money\ market)$

$Y_t = C_t + I_t + G_t$ $\qquad (income\ identity)$

where C = consumption, Y = income, I = investment, r = rate of interest, M = money supply and G = government expenditure. The variables C_t, I_t, Y_t, r_t are endogenous while the remaining are predetermined variables. Thus, G = 4.

a) Consider the consumption equation. The variables $I_{t-1}, I_t, M_t, r_t, G_t$ are missing from this equation. Thus, k = 5. The equation is over identified since $k = 5 > G\breve{\ }1 = 3$.

b) Consider the investment equation. The variables $C_{t-1}, C_t, M_t, Y_t, G_t$ are missing from this equation. Thus, k = 5. The equation is over identified since $k = 5 > G\breve{\ }1 = 3$.

4.3.2 The rank condition for identification

The rank condition states that: in a system of G equations any particular equation is (exactly or over) identified if and only if it is possible to construct at least one non-zero determinant of order (G-1) from the coefficients of the variables excluded from that particular equation but contained in the other equations of the system.

Example

Consider the simplified model of the economy above:

$C_t = \alpha_0 + \alpha_1 Y_t + \alpha_2 C_{t-1} + u_{1t}$ $\qquad (consumption)$

$I_t = \beta_0 + \beta_1 r_t + \beta_2 I_{t-1} + u_{2t}$ $\qquad (investment)$

$r_t = \gamma_0 + \gamma_1 Y_t + \gamma_2 M_t + u_{3t}$ $\qquad (money\ market)$

$Y_t = C_t + I_t + G_t$ $\qquad (income\ identity)$

This model may be re-written as:

$$-C_t + \alpha_0 + \alpha_1 Y_t + \alpha_2 C_{t-1} + 0I_t + 0r_t + 0I_{t-1} + 0M_t + 0G_t + u_{1t} = 0$$

$$0C_t + \beta_0 + 0Y_t + 0C_{t-1} - I_t + \beta_1 r_t + \beta_2 I_{t-1} + 0M_t + 0G_t + u_{2t} = 0$$

$$0C_t + \gamma_0 + \gamma_1 Y_t + 0C_{t-1} + 0I_t - r_t + 0I_{t-1} + \gamma_2 M_t + 0G_t + u_{3t} = 0$$

$$C_t + 0 - Y_t + 0C_{t-1} + I_t + 0r_t + 0I_{t-1} + 0M_t + G_t + 0 = 0$$

Note that the coefficient of a variable excluded from an equation is equal to zero. Ignoring the random disturbances and the constants, a table of the parameters of the model is as follows:

	C_t	Y_t	C_{t-1}	I_t	r_t	I_{t-1}	M_t	G_t
consumption	-1	α_1	α_2	0	0	0	0	0
investment	0	0	0	-1	β_1	β_2	0	0
money market	0	γ_1	0	0	-1	0	γ_2	0
income identity	1	-1	0	1	0	0	0	1

Now suppose we want to check the identification status of the consumption function.

a) We eliminate the row corresponding to the consumption function.

b) We eliminate the columns in which the consumption function has non-zero coefficients.

The two steps are shown below:

	C_t	Y_t	C_{t-1}	I_t	r_t	I_{t-1}	M_t	G_t
~~consumption~~	~~-1~~	~~α_1~~	~~α_2~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~
investment	0	0	0	-1	β_1	β_2	0	0
money market	0	γ_1	0	0	-1	0	γ_2	0
income identity	1	-1	0	1	0	0	0	1

Note that by doing steps (a) and (b) above, we are left with the coefficients of variables not included in the consumption function, but contained in the other equations of the system.

After eliminating the relevant row and columns, we get the following table (matrix) of parameters:

I_t	r_t	I_{t-1}	M_t	G_t
-1	β_1	β_2	0	0
0	-1	0	γ_2	0
1	0	0	0	1

..........(*)

Since the system has G = 4 equations, form the determinants of order (G-1) = 3 and examine their value.

- If at least one of these determinants is non-zero, then the consumption equation is (exactly or over) identified.
- If all determinants of order 3 are zero, then the consumption equation is under identified.

For example;

$$\Delta_1 = \begin{vmatrix} -1 & \beta_1 & \beta_2 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{vmatrix} = -\begin{vmatrix} -1 & 0 \\ 0 & 0 \end{vmatrix} - \beta_1 \begin{vmatrix} 1 & 0 \\ 0 & 0 \end{vmatrix} + \beta_2 \begin{vmatrix} 0 & -1 \\ 1 & 0 \end{vmatrix} = -1(0)+\beta_1(0)+\beta_2(1)=\beta_2 \neq 0$$

or

$$\Delta_2 = \begin{vmatrix} \beta_2 & 0 & 0 \\ 0 & \gamma_2 & 0 \\ 0 & 0 & 1 \end{vmatrix} = \beta_2(\gamma_2) \neq 0$$

Thus, we can form at least one non-zero determinant of order 3, and hence, the consumption equation is exactly or over identified.

To see whether the consumption equation is exactly or over identified, we can use the order condition. Since we have four endogenous variables $(C_t, I_t, Y_t, r_t), G = 4$. As can be seen from Table (*) above, the variables $I_{t-1}, I_t, M_t, r_t, G_t$are missing from the consumption equation, meaning k = 5. Thus, the equation is over identified since $k = 5 > G - 1 = 3$.

4.4 Estimation of simultaneous equations models: ILS, 2SLS

4.4.1 Indirect least squares (ILS) method

In this method, we first obtain the estimates of the reduced form parameters by applying OLS to the reduced form equations and then indirectly get the estimates of the parameters of the structural model. This method is applied to exactly identified equations.

Steps:

a) Obtain the reduced form equations (that is, express the endogenous variables in terms of predetermined variables).

b) Apply OLS to the reduced form equations individually. OLS will yield consistent estimates of the reduced form parameters (since each equation involves only non-stochastic (predetermined) variables that appear as 'independent' variables).

c) Obtain (or recover back) the estimates of the original structural coefficients using the estimates in step (2).

Example

Consider the following model for demand and supply of pork:

$Q_t = a_1 + a_2 P_t + a_3 Y_t + u_{1t}$ $\qquad (demand\ function)$.......................$(4.8a)$

$Q_t = a_2 + a_3 P_t + a_3 Z_t + u_{2t}$ $\qquad (supply\ function)$.......................$(4.8b)$

where Q_t is consumption of pork (pounds per capita), Pt is real price of pork (cents per pound), Y_t is disposable personal income (dollars per capita) and Z_t is 'predetermined elements in pork production'. Here P and Q are endogenous variables while Y and Z are predetermined variables. It can easily be shown that both equations are exactly identified. Thus, we can apply ILS to estimate the parameters. We first express P and Q in terms of the predetermined variables and disturbances as:

$Q_t = \left(\frac{b_2a_1 - b_1a_2}{b_2 - b_1}\right) + \left(\frac{c_1b_2}{b_2 - b_1}\right) Y_t - \left(\frac{c_2b_1}{b_2 - b_1}\right) Z_t + \left(\frac{b_2u_{1t} - b_1u_{2t}}{b_2 - b_1}\right)$..............$(4.9a)$

$P_t = \left(\frac{a_1 - a_2}{b_2 - b_1}\right) + \left(\frac{c_1}{b_2 - b_1}\right) Y_t - \left(\frac{c_2}{b_2 - b_1}\right) Z_t + \left(\frac{u_{1t} - u_{2t}}{b_2 - b_1}\right)$........................$(4.9b)$

We can re-write equations (4.9a) and (4.9b) as:

$q_t = \pi_1 + \pi_2 Y_t + \pi_3 Z_t + \epsilon_{1t}$

$p_t = \pi_4 + \pi_5 Y_t + \pi_6 Z_t + \epsilon_{1t}$

$\frac{\pi_2}{\pi_5} = \frac{c_1b_2/b_2 - b_1}{c_1b_2 - b_1} = b_2 \Rightarrow \hat{b}_2 = \frac{\hat{\pi}_2}{\hat{\pi}_5}$

$\frac{\pi_3}{\pi_6} = \frac{c_2b_1/b_2 - b_1}{c_2/b_2 - b_1} = b_1 \Rightarrow \hat{b}_1 = \frac{\hat{\pi}_3}{\hat{\pi}_6}$

$\pi_5 = \frac{c_1}{b_2 - b_1} \qquad \Rightarrow c_1 = \pi_5(b_2 - b_1) \qquad \Rightarrow \hat{c}_1 = \hat{\pi}_5(\hat{b}_2 - \hat{b}_1)$

Similarly, it can be shown that $\hat{c}_2 = \hat{\pi}_6(\hat{b}_2 - \hat{b}_1), \hat{a}_1 = \hat{\pi}_1 - \hat{b}_1\hat{\pi}_4$ and $\hat{a}_2 = \hat{\pi}_1 - \hat{b}_2\hat{\pi}_4$

Instrumental variable (IV) method

Suppose we have the model (in deviation form):

$$y_i = \beta x_i + \mathcal{E}_i$$

where x_i is correlated with $\mathcal{E}_i$. We can not estimate β by OLS as it will yield an inconsistent estimator of β(refer to errors var variables for details). What we do is search for an instrumental variable (IV) z_i that is uncorrelated with $\mathcal{E}_i$ but correlated with x_i; that is,$cov(z_i, \mathcal{E}_i) = 0$ and $cov(z_i, x_i) \neq 0$. The sample counterpart of $cov(z_i, \mathcal{E}_i) = 0$ is:

$$\begin{aligned}\frac{1}{n}\sum z_i\mathcal{E}_i = 0 &\Longrightarrow \frac{1}{n}\sum z_i(y_i - \beta x_i) = 0\\ &\Longrightarrow \frac{1}{n}\sum z_i y_i = \hat{\beta}(\frac{1}{n}\sum z_i x_i)\\ &\Longrightarrow \hat{\beta} = \frac{\frac{1}{n}\sum z_i y_i}{\frac{1}{n}\sum z_i x_i} = \frac{\sum z_i y_i}{\sum z_i x_i}\end{aligned}$$

$\hat{\beta}$ can be expressed as:

$$\hat{\beta} = \frac{\sum z_i y_i}{\sum z_i x_i} = \frac{\sum z_i(\beta x_i + \mathcal{E}_i)}{\sum z_i x_i} = \beta + \frac{\sum z_i \mathcal{E}_i}{\sum z_i x_i}$$

Now we have,

$$plim(\sum z_i\mathcal{E}_i/n) = cov(z_i, \mathcal{E}_i) = 0$$

$$plim(\sum z_i x_i/n) = cov(z_i, x_i) \neq 0$$

Thus,

$$plim(\hat{\beta}) = \beta + plim(\frac{\sum z_i\mathcal{E}_i}{\sum z_i x_i}) = \beta + \frac{plim(\sum z_i\mathcal{E}_i/n)}{plim(\sum z_i x_i/n)} = \beta + \frac{0}{\neq 0} = \beta$$

that is, the IV estimator $\hat{\beta}$ is a consistent estimator of β.

Consider the following simultaneous equations model:

$$y_1 = a_1 + b_1 y_2 + c_1 z_1 + c_2 z_2 + u_1$$

$$y_2 = a_2 + b_2 y_1 + c_3 z_3 + u_2$$

where y_1 and y_2 are endogenous while z_1, z_2 and z_3 are predetermined.

Consider the estimation of the first equation:

- Since z_1 and z_2 are predetermined, they are not correlated with u_1, that is,$cov(z_1, u_1) = 0$ and $cov(z_2, u_1) = 0$
- y_2 is not independent of u_2, that is,$cov(y_2, u_1) \neq 0$

Thus, OLS method of estimation can not be applied. To find consistent estimators, we look for avariable that is correlated with y_2 but not correlated with u_1 . Fortunately we have z_3 that satisfies these two conditions, that is, $cov(y_2, z_3) \neq 0$ and $cov(z_3, u_1) = 0$. Thus, z_3 can serve as an IV for y_2. The procedure for estimation of the first equation is as follows:

a) Regress y_2 on z_1 , z_2 and z_3; that is, using OLS estimate the model:

$$y_2 = a_{10} + a_{11} z_1 + a_{12} z_2 + a_{13} z_3 + v_1.$$

b) Obtain $\hat{y_2}$ where $\hat{y_2} = \hat{a_{10}} + \hat{a_{11}} z_1 + \hat{a_{12}} z_2 + \hat{a_{13}} z_3$.

c) Regress y_2 on $\hat{y_2}$, z_1 and z_2; that is, estimate the model:

$$y_1 = a_1 + b_1 \hat{y_2} + c_1 z_1 + c_2 z_2 + u_1$$

Note that since z_1 , z_2 and z_3 are predetermined variables, and hence, not correlated with u_1, we have:

$$cov(\hat{y_2}, u_1) = cov(\hat{a_{10}} + \hat{a_{11}} z_1 + \hat{a_{12}} z_2 + \hat{a_{13}} z_3, u_3) = 0$$

Thus, the OLS estimation using the above procedure yields consistent estimators.

Consider the second equation.

- Since z_3 is predetermined, it is not correlated with u_2, that is, $Cov(z_3, u_2) = 0$.
- y_1 is not independent of u_2, that is, $Cov(y_1, u_2) \neq 0$

Again OLS can not be applied to estimate the parameters. To find consistent estimators, we look for a variable that is correlated with y_1 but not correlated with u_2. Here we have two choices, namely,

$z_1$and z_2 that can serve as instruments.

Note : We have more than enough instrumental variables since the second equation is over identified.

In order to estimate the second equation:

a) Regress y_1 on z_1 and z_3 (if z_1 is considered as an IV for y_1) or regress y_1 on z_2 and z_3 (if z_2 is considered as an IV for y_1) using OLS and obtain $\hat{y}_1$.

b) Regress y_2 on $\hat{y}_1$ and z_3; that is, estimate the model:

Note that the solution is not unique, that is, depending on whether z_1 is considered as an IV for y_1 or z_2 is considered as an IV for y_1 , we may get different results.

Example

Using a data on some characteristics of the wine industry in Australia , which is saved as "wine.csv".

```
wine<-read.csv("wine.csv",header = T,sep = ",")
wine_var<-wine[1,] # To look the variables for short codes
names(wine)=c("year","consumption","storage_costs","price_wine","price_beer","advertisin
# Converting to log
year<-wine$year; wine=wine[,-1];wine=log(wine);wine<-cbind(year,wine)
```

It is assumed that a reasonable demand-supply model for the industry would be (where all variables are in logs):

$$Q_1 = a_0 + a_1 PW_t + a_2 PB_t + a_3 Y_t + a_4 A_t + u_t \qquad ....(demand)$$

$$Q_1 = b_0 + b_1 PW_t + b_2 S_t + v_t \qquad ....(supply)$$

where Q is real per capita consumption of wine, PW is the price of wine relative to CPI, PB is the price of beer relative to CPI, Y is real per capita disposable income, A is real per capita advertising expenditure, and S is index of storage costs. Here Q and PW are the two endogenous variables while the rest are exogenous variables.

Apply instrumental variables method of estimation?

solution

Using Instrumental variables method of Estimation

To estimate the demand function we have only one instrumental variable (IV) S. But for the estimation of the supply we have available three IVs: PB, Y and A.

I) Estimation of the supply function

The supply equation is **over-identified**. Thus, we have three possible IV's for price of wine (PW):

a) price of beer (PB),

b) advertising expense (A),

c) income (Y).

(a) Now let's take price of beer as IV to estimate the supply function:

First we regress price_wine(PW) on price_beer(PB) and storage_costs(S), and obtain the predicted values of PW (IV pb for prcwine). At last, we estimate the supply function by regressing consumption (Q) on S and (IV pb for prcwine).

The results are as follows:

```
rm(list = ls())
wine<-read.csv("wine.csv",header = T,sep = ",")
wine_var<-wine[1,] # To look the variables for short codes
names(wine)=c("year","consumption","storage_costs","price_wine","price_beer","advertisin
year<-wine$year; wine=wine[,-1];wine=log(wine);wine<-cbind(year,wine)

# regress price_wine(PW) on price_beer(PB) and storage_costs(S)
prcwine<-lm(price_wine~storage_costs+price_beer,data=wine)
#obtain the predicted values of price_wine(PW)
pred_prcwine_4_pb<-fitted(prcwine)
# estimate the supply function by regressing consumption(Q) on pred_prcwine and storag
fit_supply<-lm(consumption~pred_prcwine_4_pb+storage_costs,data=wine)
summary(fit_supply)
```

```
##
## Call:
## lm(formula = consumption ~ pred_prcwine_4_pb + storage_costs,
##     data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.71082 -0.19492 -0.03355  0.27413  0.41264
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -10.7589    44.1603  -0.244    0.810
## pred_prcwine_4_pb   0.3363    16.6039   0.020    0.984
## storage_costs       2.1310     6.8764   0.310    0.760
##
## Residual standard error: 0.3136 on 17 degrees of freedom
## Multiple R-squared:  0.7885, Adjusted R-squared:  0.7636
## F-statistic: 31.68 on 2 and 17 DF,  p-value: 1.844e-06
```

as shown above the Instrumental variable price of beer for price of wine is not significant.

(b) Now let's take advertising expense as IV to estimate the supply function:

First we regress price of wine (PW) on advertising expense (A) and storage cost (S), and obtain the predicted values of PW (IV ad for prcwine). At last, we estimate the supply function by regression consumption (Q) on S and (IV ad for prcwine).

The results are as follows:

```
rm(list = ls())
wine<-read.csv("wine.csv",header = T,sep = ",")
wine_var<-wine[1,] # To look the variables for short codes
names(wine)=c("year","consumption","storage_costs","price_wine","price_beer","advertisin
year<-wine$year; wine=wine[,-1];wine=log(wine);wine<-cbind(year,wine)

prcwine_adv<-lm(price_wine~advertising+storage_costs,data=wine)
# the predicted values of price_wine(PW) is
pred_prcwine_4_adv<-fitted(prcwine_adv)
#Then estimate the supply function by regressing consumption(Q) on storage_costs and
fit_supply_adv<-lm(consumption~pred_prcwine_4_adv+storage_costs,data=wine)
summary(fit_supply_adv)
```

```
##
```

```
## Call:
## lm(formula = consumption ~ pred_prcwine_4_adv + storage_costs,
##     data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51996 -0.14088 -0.07408  0.18592  0.51050
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -17.6494     4.6175  -3.822  0.00136 **
## pred_prcwine_4_adv   2.9282     1.6728   1.751  0.09805 .
## storage_costs        1.0584     0.7403   1.430  0.17090
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2887 on 17 degrees of freedom
## Multiple R-squared:  0.8208, Adjusted R-squared:  0.7997
## F-statistic: 38.92 on 2 and 17 DF,  p-value: 4.509e-07
```

(c) Now let's take income as IV to estimate the supply function:

First we regress price of wine (PW) on income (Y) and storage cost (S), get the predicted values of PW (IV inc for prcwine), and then estimate the supply function by regressing consumption (Q) on S and (IV inc for prcwine).

```r
rm(list = ls())
wine<-read.csv("wine.csv",header = T,sep = ",")
wine_var<-wine[1,] # To look the variables for short codes
names(wine)=c("year","consumption","storage_costs","price_wine","price_beer","advertisin
year<-wine$year; wine=wine[,-1];wine=log(wine);wine<-cbind(year,wine)

prcwine_inc<-lm(price_wine~income+storage_costs,data=wine)
# the predicted values of price_wine(PW) is
```

```r
pred_prcwine_4_inc<-fitted(prcwine_inc)
 #Then estimate the supply function by regressing consumption(Q) on storage_costs and
fit_supply_inc<-lm(consumption~storage_costs+pred_prcwine_4_inc,data=wine)
  summary(fit_supply_inc)
```

```
##
## Call:
## lm(formula = consumption ~ storage_costs + pred_prcwine_4_inc,
##     data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34213 -0.10828  0.01008  0.08265  0.37274
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -16.9793     1.3431 -12.642 4.52e-10 ***
## storage_costs        1.1627     0.2340   4.969 0.000117 ***
## pred_prcwine_4_inc   2.6762     0.4224   6.335 7.46e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1711 on 17 degrees of freedom
## Multiple R-squared:  0.9371, Adjusted R-squared:  0.9297
## F-statistic: 126.5 on 2 and 17 DF,  p-value: 6.18e-11
```

By comparing the estimated models using the three IV's, it seems that income is the best IV as the resulting estimated model has the highest coefficient of determination (R^2= 0.9370586). Since all variables are in logs, the coefficients are elasticities. Thus, quantity supplied is responsive to both price and storage costs (both p-values < 0.001). In particular, the price elasticity of supply for wine is about 2.68.

II) Estimation of the demand function

The demand equation is **exactly-identified**. Thus, we just have one available IV: storage costs. First we regress price of wine (PW) on price of beer (PB), income (Y), advertising expense (A) and storage cost (S), get the predicted values of PW (Predicted price of wine), and then estimate the demand function by regressing consumption (Q) on (Predicted price of wine), PB, Y and A.

The results are as follows:

```
rm(list = ls())
wine<-read.csv("wine.csv",header = T,sep = ",")
wine_var<-wine[1,] # To look the variables for short codes
names(wine)=c("year","consumption","storage_costs","price_wine","price_beer","advertisin
year<-wine$year; wine=wine[,-1];wine=log(wine);wine<-cbind(year,wine)

prcwine_stors<-lm(price_wine~price_beer+income+advertising+storage_costs,data=wine)
# the predicted values of price_wine(PW) is
predicted_PW<-fitted(prcwine_stors)
#Then to estimate the demand function by regressing consumption(Q)
fit_dem<-lm(consumption~predicted_PW+price_beer+income+advertising,data=wine)
summary(fit_dem)
```

```
##
## Call:
## lm(formula = consumption ~ predicted_PW + price_beer + income +
##     advertising, data = wine)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.283068 -0.092056 -0.003743  0.087104  0.292990
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -26.1950     6.5943  -3.972  0.00123 **
## predicted_PW   0.6434     0.8381   0.768  0.45459
```

```
## price_beer    -0.1397     0.8778  -0.159  0.87571
## income         4.0821     1.5943   2.560  0.02174 *
## advertising   -0.9851     0.8347  -1.180  0.25631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1545 on 15 degrees of freedom
## Multiple R-squared:  0.9547,   Adjusted R-squared:  0.9426
## F-statistic: 79.03 on 4 and 15 DF,  p-value: 6.802e-10
```

We observe that the coefficient of determination is 95.47% and the F-statistic is significant. However, most of the regression coefficients are insignificant. Furthermore, all the coefficients except that of income (Y) have the wrong signs. This is probably due to multicollinearity (MC) and check it using vif function from 'car' package.

```
#install.packages("car")
library(car)
vif(fit_dem)
```

```
## predicted_PW   price_beer       income  advertising
##     11.71816      2.55052     70.90987     32.92366
```

This shows that the variance inflation factor (VIF) for income and advertising expense are large (far greater than 10). Thus, we have to drop one of them. From practical point of view, it seems wise to drop advertising expense and re-estimate the model. The results are:

```
#library(car)
fit_dem_adv_drop<-lm(consumption~predicted_PW+price_beer+income,data=wine)
vif(fit_dem_adv_drop)
```

```
## predicted_PW   price_beer       income
##     5.213102     2.520297     7.847053
```

The problem of MC is now solved as the VIF's are greatly reduced (all less than 10).

```r
summary(fit_dem_adv_drop)
```

```
##
## Call:
## lm(formula = consumption ~ predicted_PW + price_beer + income,
##     data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33959 -0.09946  0.00757  0.06822  0.32335
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -21.1093     5.0524  -4.178 0.000710 ***
## predicted_PW   1.3803     0.5658   2.440 0.026727 *
## price_beer    -0.2524     0.8832  -0.286 0.778690
## income         2.3077     0.5368   4.299 0.000552 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1564 on 16 degrees of freedom
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.9412
## F-statistic: 102.4 on 3 and 16 DF,  p-value: 1.178e-10
```

However, the coefficients of both price of wine and price of beer have wrong signs. In particular, the coefficient of price of wine not only has the wrong sign but is also significant. This is difficult to interpret. For the other variables, the conclusion we arrive at is that the demand for wine is not responsive to the price of beer, but is responsive to income. The income elasticity of demand for wine is about 2.31.

4.4.2 Two-stage least squares (2-SLS) method

The main difference between the IV and 2-SLS methods is that in the former case the $\hat{y}_i$ are used as instruments, while in the latter case the $\hat{y}_i$ are used as regressors. Both methods yield the same result if the equation under consideration is exactly identified. The 2-SLS procedure is generally applicable for estimation of over-identified equations as it provides unique estimators.

Steps :

a) Estimate the reduced form equations by OLS and obtain the predicted $\hat{y}_i$.

b) Replace the right hand side endogenous variables in the structural equations by the corresponding $\hat{y}_i$ and estimate them by OLS.

Consider the above simultaneous equations model:

$$y_1 = a_1 + b_1 y_2 + c_1 z_1 + c_2 z_2 + u_1 ....(a)$$

$$y_2 = a_2 + b_2 y_+ c_3 z_3 + u_2 ....(b)$$

where y_1 and y_2 are endogenous while z_1 , z_2 and z_3 are predetermined.

Since $cov(y_2, u_1) \neq 0$ and $Cov(y_1, u_2) \neq 0$, we can not apply OLS. Since equation (a) is exactly identified, the 2-SLS procedure is the same as the IV method. The 2-SLS procedure of estimation of equation (b) (which is over-identified) is: . We first estimate the reduced form equations by OLS; that is, we regress y_1 on z_1 , z_2 and z_3 using OLS and obtain $\hat{y}_1$.

. We then replace y_1 by $\hat{y}_1$ and estimate equation (b) by OLS, that is, we apply OLS to:

$$y_2 = b_2 \hat{y}_1 + c_3 z_3 + u_2$$

Note

a) Unlike ILS, 2-SLS provides only one estimate per parameter for over-identified models.

b) In case of exactly identified equations, both ILS and 2-SLS produce the same parameter estimates.

Example

for the previous data ,'wine.csv', apply two stages least squares (2-SLS) method of estimation to the supply-demand function of wine. where the demand-supply model of the industry is (where all variables are in logs):

$$Q_1 = a_0 + a_1 PW_t + a_2 PB_t + a_3 Y_t + a_4 A_t + u_t \qquad ....(demand)$$

$$Q_1 = b_0 + b_1 PW_t + b_2 S_t + v_t \qquad ....(supply)$$

where Q is real per capita consumption of wine, PW is the price of wine relative to CPI, PB is the price of beer relative to CPI, Y is real per capita disposable income, A is real per capita advertising expenditure, and S is index of storage costs. Here Q and PW are the two endogenous variables while the rest are exogenous variables.

Apply 2-SLS method of estimation?

```
rm(list = ls())
wine<-read.csv("wine.csv",header = T,sep = ",")
wine_var<-wine[1,] # To look the variables for short codes
names(wine)=c("year","consumption","storage_costs","price_wine","price_beer","advertisin
# Converting to log
year<-wine$year; wine=wine[,-1];wine=log(wine);wine<-cbind(year,wine)
wine_var # Look the variable
```

```
##   year Q_real.per.capita.consumption.of.wine S_storage.costs. PW_price.of.wine PB_pri
## 1 1955                                  0.91             85.4             77.5
##   A_advertising Y_disposable.income
## 1          89.1                1056
```

solution

Estimation using two stages least squares (2-SLS)

In this method, we first find the reduced form equations by regressing each endogenous variable on all exogenous variables. Then we replace all the endogenous variables in each equation by their predicted values from the reduced forms and estimate each equation by OLS. Note that the IV estimator and the 2-SLS estimator are the same if the equation under consideration is exactly identified. In our case we have seen that the demand equation is exactly identified. Thus, the IV and 2-SLS estimators of the parameters are the same.

(I)Estimate supply function

To estimate the supply function: We first regress the price of wine (PW) on all exogenous variables PB, Y, A and S, and get the predicted values (PW 2sls). We then estimate the supply function by regressing consumption (Q) on S and PW 2sls.

The results are shown below:

```r
# regress the price_wine on price_beer, income,advertising, and storage_costs.
 pw2sls<-lm(price_wine~price_beer+income+advertising+storage_costs,data=wine)
# predicted value of price wine .
pred_prcwine<-fitted(pw2sls)
 #Estimate the supply function by regressing consumption (Q) on pred_prcwine and
fit_sup<-lm(consumption~pred_prcwine+storage_costs,data=wine)
summary(fit_sup)
```

```
##
## Call:
## lm(formula = consumption ~ pred_prcwine + storage_costs, data = wine)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.288513 -0.099838 -0.000192  0.085312  0.278820
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -16.8197     1.0890 -15.445 1.95e-11 ***
## pred_prcwine    2.6162     0.3343   7.827 4.91e-07 ***
## storage_costs   1.1876     0.1918   6.191 9.87e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1462 on 17 degrees of freedom
## Multiple R-squared:  0.954,  Adjusted R-squared:  0.9486
## F-statistic: 176.5 on 2 and 17 DF,  p-value: 4.265e-12
```

We can see from the above table that the coefficients of both price of wine and storage cost are significant. The price elasticity of supply is about 2.62.

(II) Estimate demand function

To estimate the demand equation, we have to regress the consumption on all exogenous variables of the equation, i.e.,price of wine(but use pred_prcwine)(PW), price of beer(PB), income(Y) and advertising(A).

```
fit_dem<-lm(consumption~pred_prcwine+price_beer+income+advertising,data=wine)
summary(fit_dem)
```

```
##
## Call:
## lm(formula = consumption ~ pred_prcwine + price_beer + income +
##     advertising, data = wine)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.283068 -0.092056 -0.003743  0.087104  0.292990
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -26.1950     6.5943  -3.972  0.00123 **
## pred_prcwine   0.6434     0.8381   0.768  0.45459
## price_beer    -0.1397     0.8778  -0.159  0.87571
## income         4.0821     1.5943   2.560  0.02174 *
## advertising   -0.9851     0.8347  -1.180  0.25631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1545 on 15 degrees of freedom
## Multiple R-squared:  0.9547, Adjusted R-squared:  0.9426
## F-statistic: 79.03 on 4 and 15 DF,  p-value: 6.802e-10
```

```r
## check for MC problem
vif(fit_dem)
```

```
## pred_prcwine   price_beer       income  advertising
##     11.71816      2.55052     70.90987     32.92366
```

As shown above there is multicollinearity problem. So, let's drop advertising as we did in the previous.

```r
fit_dem_drop_ad<-lm(consumption~pred_prcwine+price_beer+income,data=wine)
summary(fit_dem_drop_ad)
```

```
##
## Call:
## lm(formula = consumption ~ pred_prcwine + price_beer + income,
##     data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33959 -0.09946  0.00757  0.06822  0.32335
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -21.1093     5.0524  -4.178 0.000710 ***
## pred_prcwine   1.3803     0.5658   2.440 0.026727 *
## price_beer    -0.2524     0.8832  -0.286 0.778690
## income         2.3077     0.5368   4.299 0.000552 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1564 on 16 degrees of freedom
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.9412
## F-statistic: 102.4 on 3 and 16 DF,  p-value: 1.178e-10
```

```r
## check MC problem
vif(fit_dem_drop_ad)
```

```
## pred_prcwine   price_beer       income
##     5.213102     2.520297     7.847053
```

Since demand equation is **exactly identified**, the 2SLS estimators of the parameters are the same with IV estimators that we have done in the previous.