Lecture 1

Introduction

Objectives of econometrics course

- To develop an understanding of the use of regression analysis for quantifying economic relationships and testing economic theories.
- To equip for reading and evaluation of empirical papers in professional journals.
- To provide practical experience of using econometric software to fit economic models.

Econometrics can be defined as the branch of economics concerned with the application of methods to the measurement and analysis of economic relationships. These relationships are called functional relationships ($Y = f(\vec{X})$ - general function) which describe how one variable, usually known as the dependent variable, is determined by other variables, usually known as explanatory variables, independent variables, or regressors. The hypothesized mathematical relationship linking them is known as the regression model and a regression is a statistical way of estimating a relationship between a dependent variable and one or more explanatory variables.

Steps to construct a regression:

1. Specify the relationship to be studied and choose explanatory variables
2. Collect data
3. By running regression compute the coefficients
4. Analyze the results

There are various functional forms that can be used to describe the model. The decision about specification is based on theoretical considerations and beliefs about the sort of relationship between dependent and explanatory variables. The important thing is to justify the choice of the model and regressors theoretically. Let's look at the most common functions that will be considered in the course:

I.1) $Y = \beta_1 + \beta_2 \cdot X_2 + ... + \beta_m \cdot X_m + u$ linear relationship

Let $f(X)$ - some non-linear function, then

II. $Y = \alpha + \beta \cdot f(X) + u$ non-linear relationship

The following functions which coefficients have meaningful and commonly encountered economic interpretation (will be discussed later) belong to class II:

II.1) $\left.\begin{array}{l} Y = \beta_1 + \beta_2 \cdot \log X_2 + ... + \beta_m \cdot \log X_m + u \\ \log Y = \beta_1 + \beta_2 \cdot X_2 + ... + \beta_m \cdot X_m + u \end{array}\right\}$ semi-logarithmic relationship

II.2) $\log Y = \beta_1 + \beta_2 \cdot \log X_2 + ... + \beta_m \cdot \log X_m + u$ double-logarithmic relationship

Also (less frequently will be analyzed):

II.3) $Y = \beta_1 + \beta_2 \cdot X_2^{k_2} + ... + \beta_m \cdot X_m^{k_m} + u$, where $k_i \in R$ polynomial relationship

II.4) $Y = \beta_1 + \beta_2 \cdot (\frac{1}{X_2}) + \beta_3 \cdot X_3 + ... + \beta_m \cdot X_m + u$ inverse form relationship

Types of data:

1. Cross-sectional data consist of observations on a number of units all taken at the same (one) point in time. The units can be any set of elements similar in nature such as individuals, households, countries.

2. Time series data include repeated observations on the same unit through time (usually with fixed intervals in time). There is a certain degree of regularity in models with time series data. For example, annual data on consumption expenditure, GDP (other macroeconomic indicators).

3. Panel data consist of repeated observations on the same elements through time combining the features of cross-sectional and time series data. For example, the National Longitudinal Surveys.

Let's consider some of the issues which are important in applied analysis:

1. ***Correct specification***

Misspecification can occur when explanatory variables and the dependent variable belong to different data categories and they are measured differently. For instance, there is a misspecification problem in the regression [Growth rate of GDP] on [Population, Area of a country]. The growth rate is expressed in relative terms (also it has something to do with the qualitative side in the analysis), while regressors have absolute (volume) nature.

2. ***Endogeneity***

It occurs when one of regressors is a function of other explanatory variables and/or the dependent variable in the model. The estimation results will be invalid. For example, the decision for students to attend classes may be dependent on grade point average (GPA), so GPA acts as one of explanatory variables. However, it is reasonable to consider GPA to be dependent on whether a student attends classes => Endogeneity. Moreover, such common factor as motivation can affect both GPA and attendance.

3. ***Sample size***

Small sample size may result in large standard errors of coefficients which can lead to insignificant estimates. The problem becomes more serious when the hypothesis about stationarity for time series data is tested (the power of his test is reduced). This issue will be discussed further in the course, nevertheless, now at this stage it is useful to note that as a result it becomes harder to distinguish between the true relationship and a false regression (so called spurious) in which significant coefficients cannot indicate that the relationship exists, indeed.

Lecture 2

Simple linear regression model

As was discussed in the previous lecture econometricians wish to investigate how one variable, usually known as the dependent variable, is determined by other variables, usually known as explanatory variables, independent variables, or regressors. The hypothesized mathematical relationship linking them is known as the regression model. This lecture will analyze a model with one explanatory variable described as a simple linear regression model. First, it will look at the underlying assumptions behind this model. Then it will investigate the main technique used to estimate the relationship (known as OLS – Ordinary Least Squares). And, finally, the lecture will examine the properties of applying OLS to a simple linear regression.

Simple linear regression model:

It can be defined by the following equation:

$$Y_i = \beta_1 + \beta_2 \cdot X_i + u_i ,$$

where Y_i is the value of the dependent variable, X_i is the value of the explanatory variable, u_i is the value of the stochastic error term (disturbance term) for the i^{th} observation, and β_1 and β_2 are fixed quantities known as the parameters of the equation. It should be noted that we do not suppose to find the exact linear relationship between any economic variables, unless it exists by definition. Therefore, in order to take this fact into account the random component presented by u_i is always included into the model. There are several reasons for inclusion of the disturbance term to which discrepancies in the exact linear relationship are attributed:

1) *Omission of explanatory variables*: In real life there are many factors affecting the dependent variable. So, not only one factor is responsible to explain all the variation in the dependent variable (for example, there are several factors which influence the exam grade). Sometimes due to our model specification we fail to include all of them. Moreover, some factors are impossible to measure (how can we take account of luck in exams?);

2) *Aggregation of variables*: The relationship under study can be derived from the number of other relationships which have different parameters (for example, the Keynesian consumption function relates aggregate consumption to current disposable income where aggregate consumption is determined by the optimal choice of each individual with various parameters). Aggregation leaves this fact out of account causing discrepancies in the regression;

3) *Model misspecification*: The structure of the model can be misspecified when, for example, the dependent variable is actually determined by some not appeared factor in the regression which is highly correlated with what is included into the model;

4) *Functional misspecification*: The chosen type of mathematical function specifies the true relationship not fully but only partially for certain values of variables (linear function instead of non-linear one);

5) *Measurement error*: One or more variables can be measured incorrectly which will break the relationship.

It must be emphasized that β_1 and β_2 are population parameters which are unknown, as are the values of the disturbance term in the observations. The task of regression analysis is to obtain estimates of β_1 and β_2. Ordinary least squares method (OLS) is a good tool of estimating a linear relationship between a dependent variable and some independent or explanatory variables. We will use assumptions of the Model A for cross-sectional data with nonstochastic regressors. Nonstochastic regressors mean that their values in the observations are fixed and do not have random components. They will be used to analyze the properties of estimated coefficients.

Assumptions for model A: Model A for cross-sectional data with nonstochastic regressors.

A.1 *The model is linear in parameters and correctly specified.*

It means that the model is of the following form: $Y = \beta_1 + \beta_2 \cdot X_2 + ... + \beta_m \cdot X_m + u$: there is no built-in relationship among the βs. Note that the model can be non-linear in variables. In this case we can always transform it (for example, $X = \sqrt{Z^2 + 1}$). Correct specification means that the model is sufficiently close to the actual examined process and the difference is not important: no omitted variables, inclusion of significant variables.

A.2 *There is some variation in the regressor in the sample.*

Otherwise, if all Xs are the same then they cannot account for any of the variation in Y => no regression can be obtained.

A.3 Linearity: *The disturbance term has zero expectation*. $E(u_i) = 0$ for all i.

A.4 Homogeneity: *The disturbance term is homoscedastic*. $\sigma_{u_i}^2 = \sigma_u^2$ for all i.

Note that this condition deals with population variance versus sample variance: Population variance $\text{var}(u_i) = E[(u_i - E(u_i)^2)] = \sigma_{u_i}^2 = \sigma_u^2$; Sample variance $Var(Y) = \dfrac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n}$.

A.5 Independence: *The values of the disturbance term have independent distributions*. u_i is distributed independently of u_j for all $j \neq i$.

Gauss-Markov conditions

According to the Gauss-Markov theorem, these assumptions will guarantee that the OLS estimators of the coefficients are BLUE: best (most efficient) linear (function of the observations on Y) unbiased estimators.

A.6 The disturbance term has a normal distribution. $u_i \sim N(0, \sigma_u^2)$

Ordinary Least Squares (OLS):

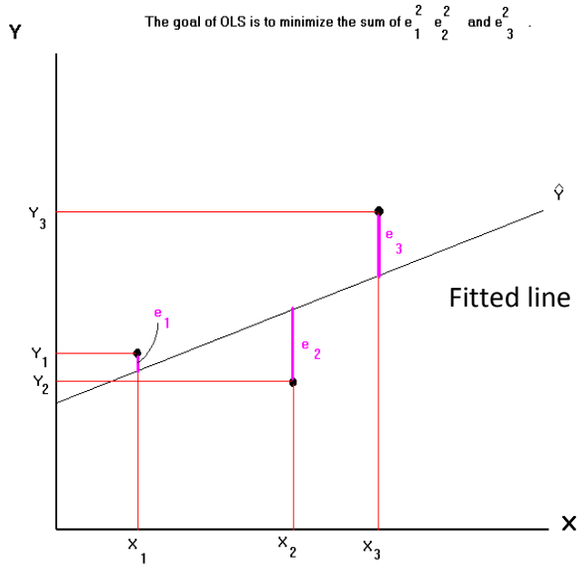
The simple linear regression model $Y_i = \beta_1 + \beta_2 \cdot X_i + u_i$ has 2 components: a non-random component $\beta_1 + \beta_2 \cdot X_i$ and the random unobserved component u_i. Given a sample of n observations the result of estimation procedure is the following fitted equation:

$$\hat{Y}_i = b_1 + b_2 X_i$$

The difference between the actual value of Y_i and its estimated value , $\hat{Y}_i$, equals $e_i = Y_i - \hat{Y}_i$ which is called a residual. It is an observed part of the disturbance term. Residuals are not the values of the disturbance term but their behavior is very much similar to the behavior of the disturbance term.

$$Y_i = b_1 + b_2 X_i + e_i$$

OLS technique minimizes the sum of squared residuals (RSS – residual sum of squares):



$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - (b_1 + b_2 X_i))^2 \to \min$$

$$\frac{\partial RSS}{\partial b_1} = 0 => 2nb_1 - 2\sum_i Y_i + 2b_2 \sum_i X_i = 0 <=>$$

$$\overline{Y} = b_1 + b_2 \overline{X} \quad (1)$$

FOC: $\frac{\partial RSS}{\partial b_2} = 0 => 2b_2 \sum_i X_i^2 - 2\sum_i X_i Y_i + 2b_1 \sum_i X_i = 0 <=>$

$$b_2 \sum_i X_i^2 - \sum_i X_i Y_i + b_1 \sum_i X_i = 0 <=> \langle from\,(1)\, b_1 = \overline{Y} - b_2 \overline{X} \rangle$$

$$b_2 \sum_i X_i^2 - \sum_i X_i Y_i + (\overline{Y} - b_2 \overline{X}) \sum_i X_i = 0 =>$$

$$b_2 (\sum_i X_i^2 - \overline{X} \sum_i X_i) = \sum_i X_i Y_i - \overline{Y} \sum_i X_i => \langle as \sum_i X_i = n\overline{X} \rangle$$

1) $$b_2 = \frac{\sum_i X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_i X_i^2 - n\overline{X}^2}$$
$$b_1 = \overline{Y} - b_2\overline{X}$$

Let's obtain more expressions for b_2:

Note that $$\sum_i (Y_i - \overline{Y})(X_i - \overline{X}) = \sum_i X_i Y_i - n\overline{X}\,\overline{Y}$$
$$\sum_i (X_i - \overline{X})^2 = \sum_i X_i^2 - n\overline{X}^2$$ =>

2) $$b_2 = \frac{\sum_i (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_i (X_i - \overline{X})^2}$$ dividing both numerator and denominator by n =>

3) $$b_2 = \frac{\sum_i (Y_i - \overline{Y})(X_i - \overline{X})/n}{\sum_i (X_i - \overline{X})^2/n} = \frac{Cov(X,Y)}{Var(X)}$$, where sample covariance and sample variance, respectively

Let $$x_i = X_i - \overline{X}$$
$$y_i = Y_i - \overline{Y}$$

4) $$b_2 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$ Therefore, it is linear with respect to y

Let $a_i = \frac{x_i}{\sum_j x_j^2}$

5) $$b_2 = \sum_i a_i y_i$$

Properties of a_i:

1. a – non-stochastic since it involves only x_i which is non-stochastic by the assumptions of the model A;

2. $$\sum_{i=1}^{n} a_i = 0$$

Proof: $\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2} \right) = \frac{1}{\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2} \sum_{i=1}^{n} \left(X_i - \overline{X}\right) = 0$

since $\sum_{i=1}^{n} \left(X_i - \overline{X}\right) = \sum_{i=1}^{n} X_i - n\overline{X} = n\overline{X} - n\overline{X} = 0$

3. $\sum_{i=1}^{n} a_i^2 = \frac{1}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} = \frac{1}{\sum_{i=1}^{n} x_i}$

Proof: $\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2} \right)^2 = \frac{1}{\left(\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2 \right)^2} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \frac{1}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}.$

4. $\sum_{i=1}^{n} a_i X_i = 1$

Proof: As we noted $\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^2 = \sum_{i=1}^{n} X_i^{\,2} - n\overline{X} \frac{\sum_{i=1}^{n} X_i}{n} = \sum_{i=1}^{n} \left(X_i - \overline{X}\right) X_i$

since $\sum \left(X_i - \overline{X}\right) = 0$ (see above). Then, using the above equation in reverse,

$$\sum_{i=1}^{n} a_i X_i = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right) X_i}{\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2} = \frac{1}{\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2} \sum_{i=1}^{n} \left(X_i - \overline{X}\right) X_i = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2} = 1.$$

Hence, $b_2 = \sum_i a_i y_i = \sum_i a_i (Y_i - \overline{Y}) = \sum_i a_i Y_i - \overline{Y} \sum_i a_i = \langle from\, 2\, property \rangle = \sum_i a_i Y_i$

6) $b_2 = \sum_i a_i Y_i$

Using derived properties:

$$b_2 = \sum_i a_i Y_i = \sum_i a_i (\beta_1 + \beta_2 \cdot X_i + u_i) = \beta_1 \sum_i a_i + \beta_2 \sum_i a_i \cdot X_i + \sum_i a_i u_i = 0 + \beta_2 \cdot 1 + \sum_i a_i u_i$$

7) $b_2 = \beta_2 + \sum_i a_i u_i$

From 7) it is easy to get that b_2 is unbiased:

$$E(b_2) = E(\beta_2 + \sum_i a_i u_i) = E(\beta_2) + E(\sum_i a_i u_i) = \langle \textit{from property 1} \rangle = \beta_2 + \sum_i a_i \cdot E(u_i) = \beta_2$$

$$E(b_1) = E(\overline{Y} - b_2 \overline{X}) = E(\overline{Y}) - \overline{X} \cdot E(b_2) = E(\frac{\sum_i \beta_1 + \beta_2 \cdot X_i + u_i}{n}) - \overline{X} \cdot \beta_2 =$$

$$= \beta_1 + \beta_2 \cdot \frac{\sum_i X_i}{n} + E(\overline{u}) - \overline{X} \cdot \beta_2 = \beta_1 + \overline{X} \cdot \beta_2 + 0 - \overline{X} \cdot \beta_2 = \beta_1$$

Lecture 3

Simple Linear Regression Model (part 2)

As was shown in the previous lecture, there are several ways to express the slope coefficient of the simple linear regression model applying OLS method. In fact, we got:

1) $$b_2 = \frac{\sum_i X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_i X_i^2 - n\overline{X}^2} \qquad (1)$$

2) $$b_2 = \frac{\sum_i (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_i (X_i - \overline{X})^2} \qquad (2)$$

3) $$b_2 = \frac{Cov(X,Y)}{Var(X)} \qquad (3)$$

In this lecture we will derive some more equivalent expressions which are useful for the analysis of statistical properties of the regression coefficients. Then we will look at the precision of estimated coefficients calculating their variances. And, finally, the lecture will examine the issue of hypothesis testing in regression analysis.

Regression coefficients: fixed and random components:

Under assumptions of the model A let's consider the following simple linear regression model: $Y_i = \beta_1 + \beta_2 \cdot X_i + u_i$. The fitted equation is $\hat{Y}_i = b_1 + b_2 X_i$. We discussed that $\beta_1 + \beta_2 \cdot X_i$ is a non-random component and the disturbance term u_i is random. In the numerator all (1), (2), (3) have Y_i which depends on the values of u_i, so we would get different values of the dependent variable if the disturbance term had been different. But let's show that it is possible to decompose b_2 into its non-random and random components.

1st way: Substituting $Y_i = \beta_1 + \beta_2 \cdot X_i + u_i$ and $\overline{Y} = \beta_1 + \beta_2 \cdot \overline{X} + \overline{u}$ for Y_i and $\overline{Y}$, respectively, in the numerator of (2) and rearranging we get:

$$b_2 = \frac{\sum_i (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_i (X_i - \overline{X})^2} = \frac{\sum_i \left(X_i - \overline{X}\right)\left([\beta_1 + \beta_2 X_i + u_i] - [\beta_1 + \beta_2 \overline{X} + \overline{u}]\right)}{\sum_i (X_i - \overline{X})^2} =$$

$$= \frac{\sum_i \left(X_i - \overline{X}\right)\left(\beta_2 [X_i - \overline{X}] + [u_i - \overline{u}]\right)}{\sum_i (X_i - \overline{X})^2} = \beta_2 \frac{\sum_i (X_i - \overline{X})^2}{\sum_i (X_i - \overline{X})^2} + \frac{\sum_i \left(X_i - \overline{X}\right)(u_i - \overline{u})}{\sum_i (X_i - \overline{X})^2} \quad \Leftrightarrow$$

$$b_2 = \beta_2 + \frac{\sum_i \left(X_i - \overline{X}\right)(u_i - \overline{u})}{\sum_i (X_i - \overline{X})^2}$$

2nd way: Let's use (3) and properties of sample variance:

$$b_2 = \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(X,\beta+\beta_2 X+u)}{Var(X)} = \frac{Cov(X,\beta)}{Var(X)} + \beta_2\frac{Cov(X,X)}{Var(X)} + \frac{Cov(X,u)}{Var(X)} =$$

$$= 0 + \beta_2\frac{Var(X)}{Var(X)} + \frac{Cov(X,u)}{Var(X)} = \beta_2 + \frac{Cov(X,u)}{Var(X)} = \underbrace{\beta_2}_{fixed} + \underbrace{\frac{\sum_i (X_i-\overline{X})(u_i-\overline{u})}{\sum_i (X_i-\overline{X})^2}}_{random}$$

For analyzing statistical properties (expected value, population variance) it is convenient to denote:

$$x_i = X_i - \overline{X} \qquad y_i = Y_i - \overline{Y} \qquad \text{and} \qquad a_i = \frac{X_i-\overline{X}}{\sum_j (X_j-\overline{X})^2} = \frac{x_i}{\sum_j x_j^2}$$

Moreover, their properties are used for the proof of Gauss-Markov theorem.

4) $$\boxed{b_2 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}} \qquad (4)$$

Therefore, it is linear with respect to y_i. Obviously, as $y_i = Y_i - \overline{Y}$, b_2 is linear with respect to the dependent variable.

5) $$\boxed{b_2 = \sum_i a_i y_i} \qquad (5)$$

Properties of a_i:

1. a – non-stochastic since it involves only x_i which is non-stochastic by the assumptions of the model A;

2. $$\sum_{i=1}^n a_i = 0$$

   Proof: $$\sum_{i=1}^n a_i = \sum_{i=1}^n \left(\frac{X_i-\overline{X}}{\sum_{j=1}^n (X_j-\overline{X})^2} \right) = \frac{1}{\sum_{j=1}^n (X_j-\overline{X})^2} \sum_{i=1}^n (X_i-\overline{X}) = 0$$

   since $$\sum_{i=1}^n (X_i-\overline{X}) = \sum_{i=1}^n X_i - n\overline{X} = n\overline{X} - n\overline{X} = 0$$ QED

3. $$\sum_{i=1}^n a_i^2 = \frac{1}{\sum_{i=1}^n (X_i-\overline{X})^2} = \frac{1}{\sum_{i=1}^n x_i}$$

Proof: $\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sum_{j=1}^{n} (X_j - \overline{X})^2} \right)^2 = \frac{1}{\left(\sum_{j=1}^{n} (X_j - \overline{X})^2 \right)^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$. QED

4. $\sum_{i=1}^{n} a_i X_i = 1$

Proof: As we noted,

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X} \frac{\sum_{i=1}^{n} X_i}{n} = \sum_{i=1}^{n} (X_i - \overline{X}) X_i$$

as $\sum (X_i - \overline{X}) = 0$ (see above). Then, using the above equation in reverse,

$$\sum_{i=1}^{n} a_i X_i = \sum_{i=1}^{n} \frac{(X_i - \overline{X}) X_i}{\sum_{j=1}^{n} (X_j - \overline{X})^2} = \frac{1}{\sum_{j=1}^{n} (X_j - \overline{X})^2} \sum_{i=1}^{n} (X_i - \overline{X}) X_i = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{\sum_{j=1}^{n} (X_j - \overline{X})^2} = 1.$$

Hence, $b_2 = \sum_i a_i y_i = \sum_i a_i (Y_i - \overline{Y}) = \sum_i a_i Y_i - \overline{Y} \sum_i a_i = \langle from\ 2\ property \rangle = \sum_i a_i Y_i$ QED

6) $\boxed{b_2 = \sum_i a_i Y_i}$ (6)

Using derived properties:

$$b_2 = \sum_i a_i Y_i = \sum_i a_i (\beta_1 + \beta_2 \cdot X_i + u_i) = \beta_1 \sum_i a_i + \beta_2 \sum_i a_i \cdot X_i + \sum_i a_i u_i = 0 + \beta_2 \cdot 1 + \sum_i a_i u_i$$

7) $\boxed{b_2 = \beta_2 + \sum_i a_i u_i}$ (7)

Unbiasedness of estimated coefficients:

From 7) it is easy to get that b_2 is unbiased:

$$E(b_2) = E(\beta_2 + \sum_i a_i u_i) = E(\beta_2) + E(\sum_i a_i u_i) = \langle from\ property\ 1 \rangle = \beta_2 + \sum_i a_i \cdot E(u_i) = \beta_2$$

$$\boldsymbol{E(b_1) = E(\overline{Y} - b_2 \overline{X}) = E(\overline{Y}) - \overline{X} \cdot E(b_2) = E(\frac{\sum_i (\beta_1 + \beta_2 \cdot \boldsymbol{X_i} + \boldsymbol{u_i})}{\boldsymbol{n}}) - \overline{X} \cdot \beta_2 =}$$

$$= \beta_1 + \beta_2 \cdot \frac{\sum_i \boldsymbol{X_i}}{\boldsymbol{n}} + \boldsymbol{E(\overline{u})} - \overline{X} \cdot \beta_2 = \beta_1 + \overline{X} \cdot \beta_2 + 0 - \overline{X} \cdot \beta_2 = \beta_1$$

Therefore, OLS method gives unbiased estimates of the regression coefficients.

Precision of the regression coefficients:

Unbiasedness is a desirable characteristic but it is not the only one to be considered. One can check that the following estimator $b_2 = \frac{Y_n - Y_1}{X_n - X_1}$ is also unbiased but it is naïve because from the whole sample of n observations the information about only 2 observations is included into the calculation. What else do we need to consider?

Another important quality of an estimator is its reliability. The estimator should have as high a probability as possible of giving a close estimate of the population characteristic, making population variance as small as possible. Population variances can be calculated as following:

$$\sigma_{b_1}^2 = \sigma_u^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right) \text{ and } \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}.$$

Proof for b_2*:*

By definition of variance:

$\sigma_{b_2}^2 = E\{(b_2 - E(b_2))^2\} = E\{(b_2 - \beta_2)^2\}$, <as $E(b_2) = \beta_2$>

From (7) $b_2 = \beta_2 + \sum_i a_i u_i$, hence $\sigma_{b_2}^2 = E\left\{\left(\beta_2 + \sum_i a_i u_i - \beta_2\right)^2\right\} = E\left\{\left(\sum_{i=1}^{n} a_i u_i\right)^2\right\}$

Expanding the quadratic:

$$\sigma_{b_2}^2 = E\left\{\sum_{i=1}^{n} a_i^2 u_i^2 + \sum_{i=1}^{n}\sum_{j \neq i} a_i a_j u_i u_j\right\} = \sum_{i=1}^{n} a_i^2 E(u_i^2) + \sum_{i=1}^{n}\sum_{j \neq i} a_i a_j E(u_i u_j).$$

Under assumption A.5 (independence of the disturbance term) $E(u_i u_j) = 0$ for $j \neq i$ and A.4 (homoscedasticity) $E(u_i^2) = \sigma_u^2$. Hence, the second term results in zero and we get:

$$\sigma_{b_2}^2 = \sum_{i=1}^{n} a_i^2 \sigma_u^2 = \sigma_u^2 \sum_{i=1}^{n} a_i^2 = \frac{\sigma_u^2}{\sum_{j=1}^{n} (X_j - \overline{X})^2} \qquad \text{QED}$$

Proof for b_1*:*

$$\sigma_{b_1}^2 = \text{var}(b_1) = \text{var}(\overline{Y} - b_2 \overline{X}) = \text{var}(\overline{Y}) + \text{var}(b_2 \overline{X}) - 2\,\text{cov}(\overline{Y}, b_2 \overline{X})$$

Let's look at each term separately:

$$\text{var}(\overline{Y}) = \text{var}(\beta_1 + \beta_2 \cdot \overline{X} + \overline{u}) = \text{<as } \beta_1 + \beta_2 \cdot \overline{X} \text{ is non-random>} = \text{var}(\overline{Y}) = \text{var}(\overline{u}) = \frac{\sigma_u^2}{n}$$

$$\text{var}(b_2 \overline{X}) = \overline{X}^2 \cdot \text{var}(b_2) = \frac{\overline{X}^2 \cdot \sigma_u^2}{\sum_{j=1}^{n} (X_j - \overline{X})^2}$$

$$2\operatorname{cov}(\overline{Y}, b_2\overline{X}) = 2\overline{X}\operatorname{cov}(\overline{Y}, b_2) = 2\overline{X}\operatorname{cov}(\beta_1 + \beta_2 \cdot \overline{X} + \overline{u}, \beta_2 + \sum_i a_i u_i) = 2\overline{X}\operatorname{cov}(\overline{u}, \sum_i a_i u_i) = <\text{as}$$

$$\overline{u} = \frac{\sum_j u_j}{n} >= 2\overline{X} \cdot \frac{1}{n}\operatorname{cov}(\sum_j u_j, \sum_i a_i u_i)$$

When we open $\operatorname{cov}(\sum_j u_j, \sum_i a_i u_i)$ using covariance rules, we can notice that cross-terms $i \neq j$ are vanished as Gauss-Markov conditions hold. For $i = j$ $\operatorname{cov}(\sum_j u_j, \sum_i a_i u_i) = \sum_i a_i \sigma_u^2$ Hence,

$$2\operatorname{cov}(\overline{Y}, b_2\overline{X}) = 2\overline{X} \cdot \frac{1}{n}\sum_i a_i \sigma_u^2 = 2\overline{X} \cdot \frac{1}{n} \cdot \sigma_u^2 \sum_i a_i = 0$$

Therefore, $\sigma_{b_1}^2 = \frac{\sigma_u^2}{n} + \frac{\overline{X}^2 \cdot \sigma_u^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2} = \sigma_u^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$ QED

Gauss–Markov theorem states that, provided that the assumptions of Model A are satisfied, the OLS estimators are BLUE: best (most efficient) linear (combinations of the Y_i) unbiased estimators of the regression parameters.

Since σ_u^2 is unknown, we cannot calculate population variances. Intuitively, the information about the behavior of the disturbance term is somehow contained in the residuals. So we can derive the estimate of σ_u^2 from the residuals. It can be shown that s_u^2 is an unbiased estimator of σ_u^2 (left as an exercise to get this result)

$$s_u^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

We can obtain estimates of the standard deviations of the distributions of b_1 and b_2 by substituting s_u^2 for s_u^2 in the variance expressions and taking the square roots. Hence,

$$\text{s.e.}(b_1) = s_u\sqrt{\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)} \quad \text{and} \quad \text{s.e.}(b_2) = \sqrt{\frac{s_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}.$$

Statistical tests for regression coefficients

On the basis of the OLS estimates obtained, it is possible to test statistical hypotheses concerning the parameters of the true relationship between variables. For this purpose, a null hypothesis (H_0) is formulated, which is then either rejected in favour the alternative hypothesis (H_1) or is not rejected, depending on the outcome of the test.

Null hypothesis: $H_0 : \beta_2 = \beta_2^0$

Two-sided alternative hypothesis $H_1 : \beta_2 \neq \beta_2^0$

If the null hypothesis is true, then the random variable $t = \frac{b_2 - \beta_2^0}{s.e.(b_2)}$ has a Student's *t*-distribution with *(n-2)* degree of freedoms. The critical value of *t*, which we will denote by t_{crit}, is the value which the absolute value of *t* will exceed with the probability equal to the significance level.

Therefore, we have the following decision rule: H_0 is rejected, if $\left|\frac{b_2 - \beta_2^0}{s.e.(b_2)}\right| > t_{\text{crit}}$, and it is not rejected, if $\left|\frac{b_2 - \beta_2^0}{s.e.(b_2)}\right| \le t_{\text{крит}}$.

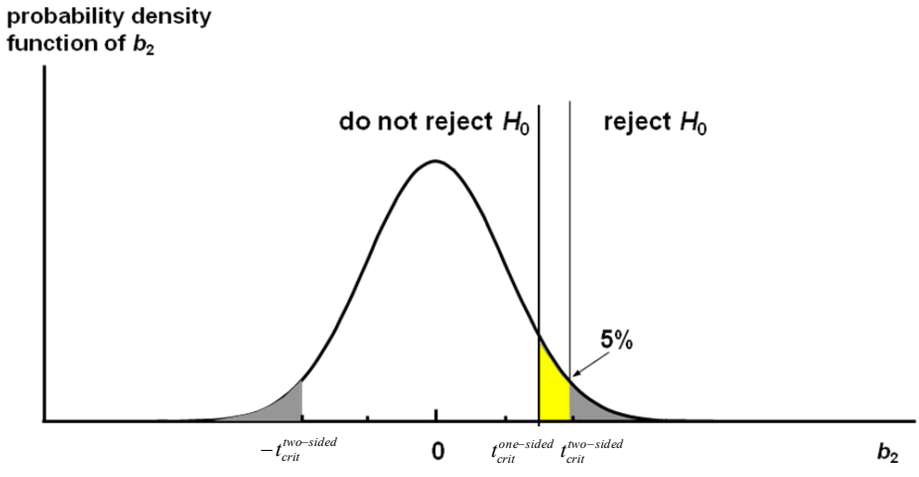
One-sided t test:

We have considered a case when the alternative hypothesis is the negation of the null hypothesis. But it is possible to make the testing procedure better if we can be more precise about the alternative hypothesis. For example, it is reasonable to suppose that the highest grade completed (HGC) affects hourly earnings (EARN) positively. Therefore, one-sided alternative should be used.

Null hypothesis: $H_0 : \beta_2 = \beta_2^0$

One-sided alternative hypothesis $H_1 : \beta_2 > \beta_2^0$ (or for the opposite direction $H_1 : \beta_2 < \beta_2^0$)

In this case test statistics $t = \frac{b_2 - \beta_2^0}{s.e.(b_2)}$ will be unchanged but the critical value t_{crit} will change. It allows to increase the power of the test compared to two-sided test with the same significance level or equivalently to decrease probability of making Type I error keeping the power of the test unaltered. Overall, the testing procedure will improve.



Confidence intervals

In addition to hypothesis testing, the Student's *t*-distribution is also used for the construction of confidence intervals. *A confidence interval* is the set of values of β which are compatible with the regression estimate obtained.

As can be implied from the above reasoning, a regression estimate b_2 and hypothetical value of β_2 are considered incompatible if the following condition is satisfied:

$$\frac{b_2 - \beta_2}{s.e.(b_2)} > t_{crit} \quad or \quad \frac{b_2 - \beta_2}{s.e.(b_2)} < -t_{crit},$$

i.e., if

$$b_2 - \beta_2 > s.e.(b_2) \cdot t_{crit} \qquad \text{or} \qquad b_2 - \beta_2 < -s.e.(b_2) \cdot t_{crit}$$

or, equivalently,

$$b_2 - s.e.(b_2) \cdot t_{crit} > \beta_2 \qquad \text{or} \qquad \beta_2 > b_2 + s.e.(b_2) \cdot t_{crit}$$

Therefore, a hypothetical value of β_2 is compatible with the regression estimate, if it satisfies the following inequality:

$$b_2 + s.e.(b_2) \cdot t_{crit} \leq \beta_2 \leq b_2 - s.e.(b_2) \cdot t_{crit}$$

By definition, this is the confidence interval.

The limits of a confidence interval depend on the choice of a significance level, as it determines the value of t_{crit}. If the significance level is 5 %, then the corresponding confidence interval is known as a 95% confidence interval; a 1 % significance level corresponds to a 99 % confidence interval, etc.

F-test for goodness of fit

This statistic is used to test the null hypothesis that all slope coefficients in a multiple regression are simultaneously equal to zero. The *F*-statistic for the overall goodness of fit is calculated as follows:

$$F = \frac{\frac{ESS}{k}}{\frac{RSS}{n - k - 1}},$$

where *k* is the number of explanatory variables.

Dividing the numerator and the denominator by *TSS* we can obtain an equivalent expression for the *F-statistic* in terms of R^2:

$$F = \frac{(ESS / TSS) / k}{(RSS / TSS) / (n - k - 1)} = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}.$$

For a simple linear regression *k* it is equal to 1, and, thus,

$$F = \frac{R^2}{(1 - R^2) / (n - 2)}.$$

After calculating *F* and setting an appropriate significance level for the test, F_{crit} (the critical value of *F*) is found in the table. If $F > F_{crit}$, the null hypothesis is rejected.

In case of a simple linear regression, the null hypothesis formulated above is reduced to $H_0 : \beta_2 = 0$, and can be equivalently tested using either an F-test or a t-test for the slope coefficient. The two tests are equivalent, because the F-statistic is equal to the square of the t-statistic and the critical value of *F*, at any given significance level, is equal to the square of the critical value of *t*. For the multiple regression there is no connection between t-tests and F-test.

Let's show that for the simple linear regression model $t^2 = F$

$$F = \frac{R^2}{(1-R^2)/(n-2)} = \frac{\dfrac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)}}{\left\{1 - \dfrac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)}\right\}/(n-2)} == \frac{\dfrac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)}}{\left\{\dfrac{\mathrm{Var}(y) - \mathrm{Var}(\hat{y})}{\mathrm{Var}(y)}\right\}/(n-2)} =$$

$$= \frac{\dfrac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)}}{\left\{\dfrac{\mathrm{Var}(e)}{\mathrm{Var}(y)}\right\}/(n-2)} = \frac{\mathrm{Var}(b_1 + b_2 x)}{\left\{\dfrac{1}{n}\sum e_i^2\right\}/(n-2)} = \frac{b_2^2 \mathrm{Var}(x)}{\dfrac{1}{n}s_u^2} = \frac{b_2^2}{\dfrac{s_u^2}{n\mathrm{Var}(x)}} = t^2$$

Alternative way to demonstrate that $t^2 = F$ you can find in lecture slides.

Lecture 4

Multiple Linear Regression Model

In the last lecture we completed the analysis of a simple linear regression model which describes how just one explanatory variable affects the dependent variable. However, in practice, more than one regressor will be available and will be responsible for the effects on the dependent variable. This results in the multiple linear regression model which will be discussed in this lecture. But it will not be only an extension of the one-variable case but also this lecture will deal with two new topics: discrimination between the effects of different explanatory variables and measurement of the joint explanatory power of the independent variables. First, we will state the underlying assumptions and consider as a special case the model with two explanatory variables. After that, we will use it to derive estimators of the regression coefficients obtained by OLS technique and interpret the model graphically. Then, we will follow the same plan as before discussing statistical properties and the precision of estimated coefficients. Finally, the lecture will examine F-test of goodness of fit for the whole equation and F-test for linear restrictions imposing on regression coefficients.

Assumptions:

Since the multiple linear regression model represents an extension of the simple linear regression model, we will still use the Model A assumptions restated for multivariable case. Let's look at them:

A.1 The model is linear in parameters and correctly specified:

$$Y = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + u$$

A.2 There does not exist an exact linear relationship among the regressors in the sample.

This assumption is very important when multicollinearity is considered.

A.3 The disturbance term has zero expectation: $E(u_i) = 0$ in every observation i

A.4 The disturbance term is homoscedastic: $pop.var(u_i) = \sigma_{u_i}^2 = \sigma_u^2$ for every i

A.5 The values of the disturbance term have independent distributions: u_i, u_j are independent for all $i \neq j$

A.6 The disturbance term has a normal distribution: $u_i \sim N(0, \sigma_u^2)$

The same assumptions as before

The model with two explanatory variables

Consider the following specification: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$

The fitted model is described as: $\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$

The regression coefficients can be derived by the same OLS procedure used in the simple linear regression analysis. The residual in i^{th} observation is expressed as follows: $e_i = Y_i - \widehat{Y}_i = Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}$. Running OLS:

$$RSS = \sum e_i^2 = \sum (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i})^2 \underset{b_1, b_2, b_3}{\longrightarrow} \min$$

The first-order conditions for an extremum $\frac{\partial S}{\partial b_1} = 0$, $\frac{\partial S}{\partial b_2} = 0$, and $\frac{\partial S}{\partial b_3} = 0$ give rise to the following equations:

$$\begin{cases} \frac{\partial S}{\partial b_1} = -2\sum (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0 \\ \frac{\partial S}{\partial b_2} = -2\sum X_{2i}(Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0 \\ \frac{\partial S}{\partial b_3} = -2\sum X_{3i}(Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0 \end{cases}$$

Therefore, we have three simple equations with three unknowns: b_1, b_2, and b_3. The first equation can be rearranged to express b_1 as a function of b_2, b_3 and the sample data on *X* and *Y:*

$$b_1 = \overline{Y} - b_2 \overline{X}_2 - b_3 \overline{X}_3$$

Using this expression, we can transform the other two equations as follows:

$$\begin{aligned} &\frac{\partial S}{\partial b_2} = -2\sum X_{2i}(Y_i - \overline{Y} + b_2 \overline{X}_2 + b_3 \overline{X}_3 - b_2 X_{2i} - b_3 X_{3i}) = 0 \\ &\frac{\partial S}{\partial b_3} = -2\sum X_{3i}(Y_i - \overline{Y} + b_2 \overline{X}_2 + b_3 \overline{X}_3 - b_2 X_{2i} - b_3 X_{3i}) = 0 \end{aligned} \qquad (*)$$

After dividing by (-2), opening the brackets and rearranging, expressions become equivalent to:

$$\begin{cases} \sum X_{2i}(b_2(\overline{X}_2 - X_{2i})) + \sum X_{2i}(b_3(\overline{X}_3 - X_{3i})) + \sum X_{2i}(Y_i - \overline{Y}) = 0 \\ \sum X_{3i}(b_2(\overline{X}_2 - X_{2i})) + \sum X_{3i}(b_3(\overline{X}_3 - X_{3i})) + \sum X_{3i}(Y_i - \overline{Y}) = 0 \end{cases} \qquad (**)$$

Let's notice that:

$$\begin{aligned} &\sum X_{2i}(\overline{X}_2 - X_{2i}) = \overline{X}_2 \sum X_{2i} - \sum X_{2i}^2 = \overline{X}_2 \cdot n\overline{X}_2 - \sum X_{2i}^2 = \\ &= n\overline{X}_2^2 - \sum X_{2i}^2 = -\sum (X_{2i} - \overline{X}_2)^2 = -nVar(X_2) \end{aligned}$$ – using the result from lecture 2

$$\begin{aligned} &\sum (X_{2i} - \overline{X}_2)(Y_i - \overline{Y}) = nCov(X_2, Y) = \sum X_{2i}(Y_i - \overline{Y}) - \overline{X}_2 \sum (Y_i - \overline{Y}) = \\ &= \sum X_{2i}(Y_i - \overline{Y}) - \overline{X}_2 \cdot 0 = \sum X_{2i}(Y_i - \overline{Y}) \end{aligned}$$

$$\begin{aligned} &\sum (X_{2i} - \overline{X}_2)(\overline{X}_3 - X_{3i}) = -nCov(X_2, X_3) = \sum X_{2i}(\overline{X}_3 - X_{3i}) - \overline{X}_2 \sum (\overline{X}_3 - X_{3i}) = \\ &= \sum X_{2i}(\overline{X}_3 - X_{3i}) - \overline{X}_2 \cdot 0 = \sum X_{2i}(\overline{X}_3 - X_{3i}) \end{aligned}$$

The same applies for the second equation in (**):

$$\sum X_{3i}(\overline{X}_2 - X_{2i}) = = -nCov(X_3, X_2)$$

$$\sum X_{3i}(\overline{X}_3 - X_{3i}) = -nVar(X_3)$$

$$\sum X_{3i}(Y_i - \overline{Y}) = nCov(X_3, Y)$$

Hence, (**) is equivalent to:

$$\begin{cases} -nb_2 Var(X_2) - nb_3 Cov(X_2, X_3) + nCov(X_2, Y) = 0 \\ -nb_2 Cov(X_2, X_3) - nb_3 Var(X_3) + nCov(X_3, Y) = 0 \end{cases}$$

Dividing by n and rearranging we get the following system of linear equations:

$$\begin{cases} b_2 Var(X_2) + b_3 Cov(X_2, X_3) = Cov(X_2, Y) \\ b_2 Cov(X_2, X_3) + b_3 Var(X_3) = Cov(X_3, Y) \end{cases}$$

Let's solve this system of simple equations using the method of determinants (Cramer's rule):

$$\Delta = \text{Var}(X_2)\text{Var}(X_3) - [\text{Cov}(X_2, X_3)]^2$$

$$\Delta_1 = \text{Cov}(X_2, Y)\text{Var}(X_3) - \text{Cov}(X_3, Y)\text{Cov}(X_2, X_3)$$

$$\Delta_2 = \text{Cov}(X_3, Y)\text{Var}(X_2) - \text{Cov}(X_2, Y)\text{Cov}(X_2, X_3)$$

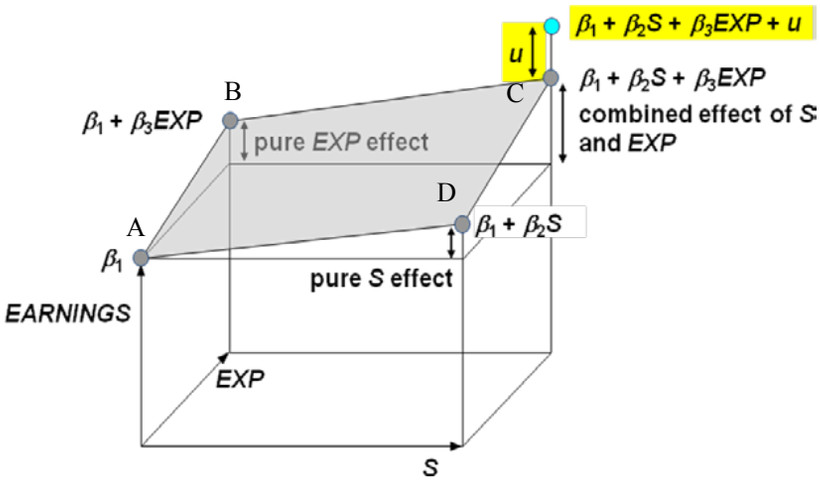
$$b_2 = \Delta_1 / \Delta = \frac{\text{Cov}(X_2, Y)\text{Var}(X_3) - \text{Cov}(X_3, Y)\text{Cov}(X_2, X_3)}{\text{Var}(X_2)\text{Var}(X_3) - [\text{Cov}(X_2, X_3)]^2}$$

$$b_3 = \Delta_2 / \Delta = \frac{\text{Cov}(X_3, Y)\text{Var}(X_2) - \text{Cov}(X_2, Y)\text{Cov}(X_2, X_3)}{\text{Var}(X_2)\text{Var}(X_3) - [\text{Cov}(X_2, X_3)]^2}$$

Graphical illustration and interpretation:

Consider the following relationship: the dependence of earnings on years of work experience (EXP) and years of schooling of a respondent (S): $EARNING = \beta_1 + \beta_2 S + \beta_3 EXP + u$.

In order to interpret this relationship graphically, let's construct the 3-dimensional space as shown below. The intercept β_1 gives earnings for respondents with zero schooling and experience. But if in the sample there are no respondents with zero schooling and experience, the intercept has not plausible interpretation. Pure S effect can be calculated holding EXP constant. It equals $\beta_1 + \beta_2 S$. Geometrically, it is a distance between point D and the upper face of the parallelepiped where AD is such that the angle between AD and the space EXPOS is equal to β_2. Similarly, pure EXP effect is $\beta_1 + \beta_3 EXP$. Under the assumption that the effects of S and EXP on $EARNING$ are linear and additive, by adding the vectors $\vec{AD}$ and $\vec{AB}$ we can get their combined effect on earnings.



Properties of regression coefficients:

As in the simple regression case, the Gauss-Markov conditions ensure unbiasedness, consistency and efficiency of OLS-estimators of the regression coefficients.

Let's show unbiasedness of coefficients for the model with 2 explanatory variables:

$$E(b_2) = E\left[\frac{\text{Cov}(X_2,Y)\text{Var}(X_3) - \text{Cov}(X_3,Y)\text{Cov}(X_2,X_3)}{\text{Var}(X_2)\text{Var}(X_3) - [\text{Cov}(X_2,X_3)]^2}\right] = <$$ Using the fact that X is nonstochastic, the denominator $= \Delta$ can be taken out from the expectation function>

$$E(b_2) = \frac{1}{\Delta}E(\text{Cov}(X_2,Y)\text{Var}(X_3) - \text{Cov}(X_3,Y)\text{Cov}(X_2,X_3)) =$$

$$= \frac{1}{\Delta}E(\text{Cov}(X_2,\beta_1 + \beta_2 X_2 + \beta_3 X_3 + u)\text{Var}(X_3) - \text{Cov}(X_3,\beta_1 + \beta_2 X_2 + \beta_3 X_3 + u)\text{Cov}(X_2,X_3)) =$$

$$= \frac{1}{\Delta}E(\Delta\beta_2 + \beta_3\text{Cov}(X_2,X_3)\text{Var}(X_3) - \beta_3\text{Cov}(X_2,X_3)\text{Var}(X_3) +$$

$$+ \text{Var}(X_3)\text{Cov}(X_2,u) - \text{Cov}(X_2,X_3)\text{Cov}(X_3,u)) =$$

$$= \beta_2 + \frac{\text{Var}(X_3)\text{E}(\text{Cov}(X_2,u)) - \text{Cov}(X_2,X_3)E(\text{Cov}(X_3,u))}{\Delta}$$

$\text{Cov}(X_2,u) = 0$ and $\text{Cov}(X_3,u) = 0$ since X is nonstochastic. Hence,

$$E(b_2) = \beta_2 + \frac{\text{Var}(X_3)\cdot 0 - \text{Cov}(X_2,X_3)\cdot 0}{\Delta} = \beta_2$$

Similar procedure for $E(b_3)$ results in $E(b_3) = \beta_3$

$$E(b_1) = E(\overline{Y} - b_2\overline{X}_2 - b_3\overline{X}_3) = E(\beta_1 + \beta_2\overline{X}_2 + \beta_3\overline{X}_3 + \overline{u}) - E(b_2)\overline{X}_2 - E(b_3)\overline{X}_3 = \beta_1$$ QED

Precision of the multiple linear regression coefficients:

In the model with 2 regressors the variance of a slope coefficient can be expressed in several ways:

1) $\sigma_{b_2}^2 = \dfrac{\sigma_u^2}{\sum\left(X_{2i} - \overline{X}_2\right)^2} \times \dfrac{1}{1 - r_{X_2,X_3}^2}$

2) $\sigma_{b_2}^2 = \dfrac{\sigma_u^2}{nVar(X_2)} \times \dfrac{1}{1 - r_{X_2,X_3}^2}$

3) Let $x_{2i} = X_{2i} - \overline{X}_2$ $\Rightarrow$ $\sigma_{b_2}^2 = \dfrac{\sigma_u^2}{\sum x_{2i}^2} \times \dfrac{1}{1 - r_{X_2,X_3}^2}$

r_{X_2,X_3} - sample correlation coefficient between X_2, X_3

Generally, for the model $Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_j X_{ji} + ... + \beta_k X_{ki} + u_i$

$\sigma_{b_j}^2 = \dfrac{\sigma_u^2}{\sum\limits_i x_{ji}^2} \times \dfrac{1}{1 - R_j^2}$ where R_j^2 is the determination coefficient of the following regression:

$X_{ji} = \alpha_1 + \alpha_2 X_{2i} + ... + \alpha_{j-1} X_{j-1i} + \alpha_{j+1} X_{j+1i} + ... + \alpha_k X_{ki} + \varepsilon_i$

It can be shown that $E\left(\dfrac{1}{n}\sum e_i^2\right) = \dfrac{n-k}{n}\sigma_u^2$ where $n-k$ is the number of degrees of freedom: k = ***# of normal equations = # of parameters to be estimated (including the intercept)***. For example, for $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ there are 3 estimated parameters: the intercept and 2 slope coefficients => d.f. = $n-3$.

Therefore, $s_u^2 = \dfrac{1}{n-k}\sum e_i^2$ – unbiased estimator of σ_u^2

Standard errors are calculated as follows:

$$\text{s.e.}(b_2) = \sqrt{\frac{s_u^2}{\sum\left(X_{2i} - \overline{X}_2\right)^2} \times \frac{1}{1 - r_{X_2,X_3}^2}}$$

$$\text{s.e.}(b_3) = \sqrt{\frac{s_u^2}{\sum\left(X_{3i} - \overline{X}_3\right)^2} \times \frac{1}{1 - r_{X_2,X_3}^2}}$$

Testing hypotheses concerning multiple regression coefficients

As usual, when the variance of the disturbance term is unknown, to test the hypothesis that a regression coefficient is equal to zero (or to an arbitrary value β_0) we use the Student's t-distribution. The critical value of t at any significance level depends on the number of degrees of freedom which is equal to *n-k*. The test procedure is the same as in the simple regression case. The null and alternative hypotheses are as follows:

$H_0 : \beta_j = \beta_0$

$H_1 : \beta_j \neq \beta_0$

Test-statistics: $t = \dfrac{b_j - \beta_0}{s.e.(b_j)}$ has a Student's t-distribution with *(n-k)* degrees of freedom.

H_0 is rejected if $\left|\dfrac{b_j - \beta_0}{s.e.(b_j)}\right| > t_{\text{crit}}$, and it is not rejected if $\left|\dfrac{b_j - \beta_0}{s.e.(b_j)}\right| \leq t_{\text{crit}}$

F-tests

To test the significance of the explanatory power of regression as a whole, or of groups of variables included in it, F-statistic is used. Under the appropriate null hypothesis, it has the two-parameter Fischer distribution, for which special tables exist. Consider the k-variable multiple linear regression model: $Y = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + u$. There are 2 types of F-test.

1) **<u>F-test of Goodness of fit</u>**

In the test for significance of the equation as a whole, the null hypothesis is that the coefficients of all variables are simultaneously equal to zero:

$H_0 : \beta_2 = ... = \beta_k = 0$

$H_1 : \exists \beta_i \neq 0$.

The appropriate F-statistic is $F(k-1, n-k) = \dfrac{ESS/(k-1)}{RSS/(n-k)} = \dfrac{R^2/(k-1)}{(1-R^2)/(n-k)}$ where k is the number of estimated parameters in the model (number of coefficients).

Under the null hypothesis, it has F-distribution with *k* and *(n-k)* degrees of freedom.

2) **<u>F-test for linear restrictions</u>** (testing the significance of joint contribution of a group of variables to the explanatory power of the model):

The general form of this test is the following:

$$F(cost, d.f.remaining) = \frac{(improvement)/(cost)}{(remaining\ unexplained)/(d.f.remaining)}$$

Let's explain this formula in words. Suppose that the explained sum of squares in regression with *k* explanatory variables is ESS_k. Having added more variables, so that their total number is now *m*, we achieve an increase in the explained sum of squares up to ESS_m. Thus, we have managed to explain additionally ESS_m-ESS_k, – our improvement (which is equal to $RSS_k - RSS_m$ because TSS does not change), having given up *(m-k)* degrees of freedom – our cost. Then, this obtained relative improvement is divided by what is remained unexplained adjusted for the degrees of freedom for the best of the models (in our example it is the model with m variables). Therefore, we get *F-statistics* and it is necessary to understand, whether this increase in the explanatory power exceeds that which can be obtained randomly. Hence, F-test is used.

<u>Consider 2 models:</u>

Unrestricted (UR) MLR with m parameters (including the constant term) and RSS_{UR}

Restricted (R) version of the previous MLR (the same data set) with k parameters and RSS_R

F-statistics $F(m-k, n-m) = \dfrac{(RSS_R - RSS_{UR})/(m-k)}{RSS_{UR}/(n-m)}$ Under the null hypothesis that additional variables do not increase the explanatory power of the model (i.e. the true values of their coefficients are all equal to zero) this statistics is distributed with *(m-k)* and *(n-m)* degrees of freedom. *(m-k)* corresponds to the number of restrictions and *(n-m)* is the difference between the number of observations and the number of coefficients for Unrestricted model.

Alternative expression for F-statistics: $F(m-k, n-m) = \dfrac{(R^2_{UR} - R^2_R)/(m-k)}{(1-R^2_{UR})/(n-m)}$

For example, let's look at the following relationship:

$$Y = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + ... + \beta_m X_m + u$$

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

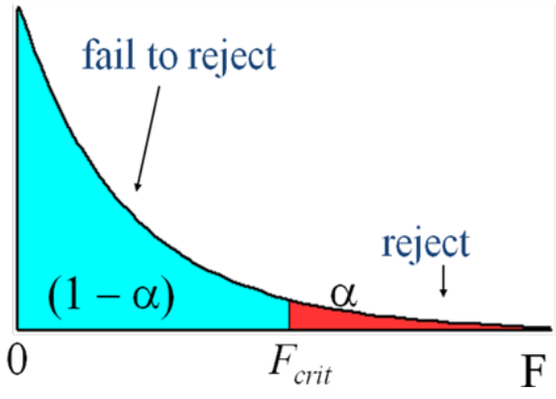
$$H_1 : \exists \beta_i \neq 0 \text{ where } i \in [k+1, m] \text{ and } i \in N$$

F-statistics: $$F(m-k, n-m) = F(\#restrictions,\ n - coefficients\ in\ UR) = \frac{(R^2_{UR} - R^2_R)/(m-k)}{(1-R^2_{UR})/(n-m)} = \boxed{= \frac{(R^2_{UR} - R^2_R)/(\#restrictions)}{(1-R^2_{UR})/(n - coefficients\ in\ UR)}}$$

The same F-statistics can be calculated using information about RSS:

$$F(m-k, n-m) = \frac{(RSS_R - RSS_{UR})/(m-k)}{RSS_{UR}/(n-m)} = \boxed{\frac{(RSS_R - RSS_{UR})/(\#restrictions)}{RSS_{UR}/(n - coefficients\ in\ UR)}}$$

In both tests the null hypothesis is rejected, if the obtained value of F exceeds the critical level (found from the tables, after the significance level is chosen).



Lecture 5. Multicollinearity. Linear Restrictions.

As we have already discussed that one of our remarks of the multiple linear regression model was the absence of an exact linear relationship among the regressors in the sample. The violation of it results in multicollinearity issue. This lecture will analyze the problem of multicollinearity according to the following plan:

1. Reasons
2. Consequences
3. Detection
4. Remedial measures

1. Reasons

The "problem" of multicollinearity is not really well-defined because it violates none of assumptions of the model A for the simple linear regression (correct specification of the model, variation in regressors in the sample, Gauss-Markov conditions. The best way to define it as a risk of obtaining erratic estimates of regression coefficients when ***the population variances of their distributions are large***, and ***it becomes hard to distinguish the influence of each factor***.

Multicollinearity

Perfect

- one explanatory variable is an exact linear function of one or more explanatory variables with no error term

Imperfect

- there is a linear relationship between the variables, but there is some error in that relationship.

Consider the following model with 2 explanatory variables:

$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i,$ where u_i satisfies all Gauss-Markov conditions;

Fitting the regression line:

$\hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i}$

It can be shown that the population variance of b_2 can be calculated as:

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{Var(X_2)} \times \frac{1}{1 - r_{X_2X_3}^2} = \frac{\sigma_u^2}{\sum_{i=1}^{n} (X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r_{X_2X_3}^2} = \frac{\sigma_u^2}{n \, \mathrm{MSD}(X_2)} \times \frac{1}{1 - r_{X_2X_3}^2}$$

From this formula it can be obtained that high correlation between explanatory variables ($r_{X_2X_3}$) is not the only one determinant of multicollinearity making $\sigma_{b_2}^2$ large. Multicollinearity is caused by high correlation between explanatory variables and (at least) one of the following factors: small sample size, small mean standard deviations of explanatory variables, large population variance of the disturbance term.

2. ***Consequences***

The presence of multicollinearity does not mean that the model is misspecified. Gauss-Markov theorem holds meaning that OLS technique yields the best linear unbiased estimates (BLUE):

a) Estimates are unbiased, consistent, and efficient, statistical tests and s.e. are valid;

b) Standard errors of estimates will increase / t-statistics will decrease:

$$s_u^2 = \frac{1}{n-k}\sum_{i=1}^{n} e_i^2$$ where k – number of parameters in regression

$$\text{s.e.}(b_2) = \sqrt{\frac{s_u^2}{\sum_{i=1}^{n}\left(X_{2i} - \bar{X}_{2i}\right)^2} \times \frac{1}{1 - r_{X_2,X_3}^2}}$$

$$t - statistics\ (b_2) = \frac{b_2}{\text{s.e.}(b_2)}$$

Hence, as r_{X_2,X_3}^2 appears in the denominator with the minus sign, $r_{X_2,X_3}^2 \uparrow$ => $\text{s.e.}(b_2) \uparrow$ $t - statistics\ (b_2) \downarrow$. It causes some problems concerning the significance from zero of estimated coefficients;

c) Estimates become very sensitive to changes in specification of the model:

Excluding insignificant variable (as t-stat are smaller as a result of multicollinearity) may lead to large changes of estimated coefficients causing omitted variable bias => estimates become biased;

d) Explanatory power of the model will be unaffected:

The estimated models can have good predictive power, even though no estimated coefficients are significantly different from zero.

3. ***Detection:***

1) Insignificant coefficients (small t-statistics) but their significance as a group (formally it can be tested using F-test or informally this can be observed by relatively high R^2 for cross-section data);

2) High correlation coefficients between explanatory variables.

Example:

Consider the following model: the dependence of earnings on abilities (measured by ASVABC), years of schooling of a respondent and family years of schooling both father and mother (HGC, HGCF, HGCM, respectively).

$$\text{Log(Earnings)} = \beta_1 + \beta_2 \cdot \text{ASVABC} + \beta_3 \cdot \text{HGC} + \beta_4 \cdot \text{HGCF} + \beta_5 \cdot HGCM$$

Estimated equation:

$$\text{LOG(EARNINGS)} = 0.98089 + 0.06211 * \text{HGC} + 0.01687 * \text{HGCF} - 0.002316 * \text{HGCM} + 0.009770 * \text{ASVABC}$$

s.e. (0.13334) (0.01014) (0.00775) (0.01027) (0.00288)

$$R^2 = 0.213505$$

Dependent Variable: LOG(EARNINGS)
Method: Least Squares
Sample: 1 570
Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.980889	0.133317	7.357585	0.0000
ASVABC	0.009770	0.002884	3.387169	0.0008
HGC	0.062113	0.010137	6.127627	0.0000
HGCF	0.016874	0.007751	2.176988	0.0299
HGCM	-0.002316	0.010271	-0.225441	0.8217

R-squared	0.213505	Mean dependent var	2.486474
Adjusted R-squared	0.207937	S.D. dependent var	0.537956
S.E. of regression	0.478769	Akaike info criterion	1.373538
Sum squared resid	129.5094	Schwarz criterion	1.411658
Log likelihood	-386.4584	Hannan-Quinn criter.	1.388411
F-statistic	38.34439	Durbin-Watson stat	1.689301
Prob(F-statistic)	0.000000		

What is wrong with this equation?

1. Insignificant HGCM: intuitively, mother's education cannot be considered as at least less important as father's education;
2. Unexpected sign of HGCM.

Consider the significance of this model using F-test:

H_0: $\beta_2=\beta_3=\beta_4=\beta_5=0$

H_1: at least of one of $\beta_2, \beta_3, \beta_4, \beta_5 \neq 0$

$$F = \frac{R^2/(5-1)}{(1-R^2)/(570-5)} = 38.34 \overset{H_0}{\approx} F(4,565)$$

$F_{critic\alpha=1\%}(4,120) = 3.48 < 38.34$ => At 1% significance level the model is significant.

The second thing that we can do is to test significance of HGC, HGCF and HGCM as a group:

H_0: $\beta_3=\beta_4=\beta_5=0$

H_1: at least of one of $\beta_3, \beta_4, \beta_5 \neq 0$

Restricted model:

$$\text{LOG(EARNINGS)} = 1.34536 + 0.02264 * \text{ASVABC} \qquad R^2_{\text{restricted}} = 0.140183$$

s.e (0.12041) (0.02264)

F-test: $F = \dfrac{(R^2 - R^2_{restricted})/3}{(1-R^2)/(570-5)} = 17.55 \overset{H_0}{\approx} F(3,565)$

$F_{critic\,\alpha=1\%}(3,120) = 3.95 < 17.55$ =>At 1% significance level HGC, HGCM, HGCF are significant as a group. So, we got the usual sign of multicollinearity (1)): particular coefficients can be insignificant but as a group they are significant.

(2)) Correlation matrix (In EViews the following command: COR HGC HGCM HGCF)

	HGC	HGCF	HGCM
HGC	1.000000	0.407582	0.404978
HGCF	0.407582	1.000000	0.632591
HGCM	0.404978	0.632591	1.000000

It can be seen that the correlation coefficient between HGCF and HGCM is relatively high. This fact is known as 'assortive mating'. Moreover, it is quite natural that if parents have more years of schooling then their children will tend to also have more years of schooling as they know that earnings are increasing function with HGC and they can afford it (payment for education).

4. Remedial measures

Sample ← → Model

- Increase sample size
- Extend the range of explanatory variables

⇩

BUT: NOT ALWAYS POSSIBLE!!!

- Make some aggregates that seem to have reasonable interpretation
- Impose a restriction !
- Cancel some variable from the model ?
- Add one more regressor

OR: Do nothing using the fact that estimated equation has good explanatory power and the only problem is artificially high standard errors. Estimates are unbiased.

Warning!!!:

- dropping one or more variables is not the best way to deal with the problem of multicollinearity because there is a risk of omitted variable bias: one variable appears to have a double effect (direct effect and a proxy one when it mimics the effect of excluding variable) => biased estimates;
- adding irrelevant variables can lead to reduction in efficiency of estimates.

Transforming the variables by aggregation and imposing restrictions – **usual solution**.

Example (continuing the previous one):

Aggregation: the effects of father's and mother's education on earnings are equally important. Due to 'assortive mating' the restriction $\beta_4=\beta_5$ seems to be reasonable. Hence, let's define $HGCP = HGCM + HGCF$.

Dependent Variable: LOG(EARNINGS)
Method: Least Squares
Sample: 1 570
Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.940096	0.129044	7.285108	0.0000
ASVABC	0.009975	0.002881	3.462760	0.0006
HGC	0.061424	0.010125	6.066715	0.0000
HGCP	0.009015	0.004239	2.126494	0.0339

R-squared	0.211464	Mean dependent var	2.486474
Adjusted R-squared	0.207285	S.D. dependent var	0.537956
S.E. of regression	0.478967	Akaike info criterion	1.372621
Sum squared resid	129.8455	Schwarz criterion	1.403117
Log likelihood	-387.1971	Hannan-Quinn criter.	1.384520
F-statistic	50.59537	Durbin-Watson stat	1.703487
Prob(F-statistic)	0.000000		

$$\text{LOG(EARNINGS)} = 0.94 + 0.0614 * \text{HGC} + 0.009 * \text{HGCP} + 0.00997 * \text{ASVABC} \qquad R^2_{1)} = 0.2115$$

s.e. (0.129) (0.0101) (0.0042) (0.00288)

At 5% significance level all coefficients are significant, but at 1% HGCP is not significant. HGCP has expected sign. The standard error of HGCP = 0.0042 becomes smaller than those of HGCF (=0.00775) and HGCM (=0.01027) from the first model. It indicates that the use of restriction has led to a gain in efficiency. Testing the restriction:

H_0: $\beta_4 = \beta_5$

H_1: $\beta_4 \neq \beta_5$

F-test: $F = \dfrac{(0.213505 - 0.211464)/1}{(1 - 0.213505)/(570 - 5)} = 1.46 \overset{H_0}{\approx} F(1,565)$

$F_{critic\,\alpha=10\%}(1,\infty) = 2.71 > 1.46$ => At 10% significance level the imposing restriction is not rejected. So the problem of multicollinearity is overcome.

One more solution to make the following restriction: $HGCSUM = HGC + HGCM + HGCF$.

$$\text{LOG(EARNINGS)} = 1.0586 + 0.01965 * \text{HGCSUM} + 0.01387 * \text{ASVABC} \qquad R^2_{2)} = 0.1846$$

s.e. (0.1283) (0.0035) (0.0028)

All coefficients are significant at all reasonable significance levels but it can be checked that the restriction is invalid (practice it!) – it is rejected at 1% level. So this is not a good solution.

Lecture 6

Variables Transformation in Regression Analysis

Previous lectures have examined only linear form models. In fact, the first assumption A1 for the model A requires that the model is linear in parameters and correctly specified. However, for many economic processes it is more common to consider nonlinear relationships among variables. This lecture will extend the analysis of estimating regressions taking into account that some types of nonlinear functions can be transformed into a linear form (so called they can be linearized). First, we will distinguish between linear and nonlinear forms looking at different types of nonlinearity. Then the lecture will deal with economic interpretations of estimated coefficients for some nonlinear models and examine the test of observing nonlinearity in the model (Ramsey's RESET test of functional misspecification**)**. After that we will discuss the issue of comparison between linear and logarithmic models. And, finally, the lecture will analyze how different types of production functions can be estimated.

Linearity and nonlinearity:

By saying that the model is linear in variables, we imply that the variables are included in the model exactly as defined as opposed to the case when they are defined as functions. Linearity in parameters means that a different parameter appears as a multiplicative factor in each explanatory variable. For example, the following model (6.1) is linear in both parameters and variables.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u \qquad (6.1)$$

At the same time, (6.2) is linear in parameters but nonlinear in variables:

$$Y = \beta_1 + \beta_2 X_2^2 + \beta_3 \sqrt{X_3} + \beta_4 \log X_4 + u \qquad (6.2)$$

For this case our assumption A1 still holds. Moreover, by making certain transformations we can get (6.2) both linear in parameters and variables. For instance, let's denote:

$$Z_2 = X_2^2, \quad Z_3 = \sqrt{X_3}, \quad Z_4 = \log X_4$$

Therefore, it is transformed into the objective function: $Y = \beta_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + u$

(6.3) form is nonlinear in parameters (because X_4 is the product of the coefficients of X_2 and X_3) but linear in variables.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_2 \beta_3 X_4 + u \qquad (6.3)$$

This model cannot be linearized (so, OLS is not applicable) but for some nonlinear in parameters specifications certain appropriate transformations can result in linear parameters models. Let's consider, for example, the following power law model (nonlinear in both parameters and variables):

$$Y = \beta_1 X^{\beta_2}$$

It is characterized by the constant elasticity which is equal to β_2.

Demonstration: By definition: The elasticity of Y with respect to X is the proportional change in Y per proportional change in X: $\frac{dY/Y}{dX/X} = \frac{dY/dX}{Y/X}$.

$$\frac{dY}{dX} = \beta_1 \beta_2 X^{\beta_2 - 1}$$

$$\frac{Y}{X} = \frac{\beta_1 X^{\beta_2}}{X} = \beta_1 X^{\beta_2 - 1}$$

Hence, elasticity $= \frac{dY/dX}{Y/X} = \frac{\beta_1 \beta_2 X^{\beta_2 - 1}}{\beta_1 X^{\beta_2 - 1}} = \beta_2$.

This model can be transformed into the linear in parameters form by taking logarithms from both sides of the equation: $\log Y = \log \beta_1 X^{\beta_2} = \log \beta_1 + \log X^{\beta_2} = \log \beta_1 + \beta_2 \log X$.

Therefore, denoting $Y' = \log Y$, $X' = \log X$, $\beta_1' = \log \beta_1$ => $Y' = \beta_1' + \beta_2 X'$ – linear in parameters

Interpretation of estimated results:

The marginal effect of minor change in X (ΔX) is equal to $\Delta Y = \frac{\partial Y}{\partial X} \cdot \Delta X$ irrespective of the regression form. It is a general procedure for interpretation of estimated results. Let's consider 4 types of estimated regressions and their corresponding interpretations of coefficients:

1) **Linear** model: $Y = b_1 + b_2 X$

$$\Delta Y = \frac{\partial Y}{\partial X} \Delta X = b_2 X$$

Interpretation: A one unit increase in X leads to b_2 units increase in Y.

2) **Semi-logarithmic** model: $Y = \beta_1 e^{\beta_2 X} \cdot \nu$

Taking logarithms from both sides it is equivalent to estimate the following equation:
$\log Y = b_1 + b_2 X$

$$\Delta Y = \frac{\partial Y}{\partial X} \Delta X = \frac{\partial Y}{\partial \log Y} \cdot \frac{\partial \log Y}{\partial X} \Delta X = \frac{\partial (e^{\log Y})}{\partial \log Y} \cdot \frac{\partial \log Y}{\partial X} \Delta X = e^{\log Y} \cdot b_2 \cdot \Delta X = Y \cdot b_2 \cdot \Delta X =>$$

$$\frac{\Delta Y}{Y} = b_2 \cdot \Delta X$$

So, if $\Delta X = 1$ then $\frac{\Delta Y}{Y} = b_2$ which implies that Y changes by $b_2 \cdot 100\%$.

Interpretation: A one unit increase in X leads to $b_2 \cdot 100\%$ increase in Y.

3) **Semi-logarithmic** model: $e^{y} = \beta_1 X^{\beta_2} \cdot \nu$

Taking logarithms from both sides it is equivalent to estimate the following equation:

$$Y = b_1 + b_2 \log X$$

$$\Delta Y = \frac{\partial Y}{\partial X}\Delta X = \frac{\partial Y}{\partial \log X} \cdot \frac{\partial \log X}{\partial X}\Delta X = \frac{\partial Y}{\partial \log X} \cdot \frac{\partial \log X}{\partial X}\Delta X = b_2 \cdot \frac{1}{X} \cdot \Delta X = b_2 \cdot \frac{\Delta X}{X} =>$$

$$\Delta Y = b_2 \cdot \frac{\Delta X}{X}$$

So, if X increases by 1% then $\Delta Y = b_2 \cdot \frac{1}{100}$ which implies that Y changes by $\frac{b_2}{100}$ units.

Interpretation: A one percentage (1%) increase in X leads to $\frac{b_2}{100}$ increase in Y.

4) **Double-logarithmic** model: $Y = \beta_1 X^{\beta_2} \cdot \nu$

Taking logarithms from both sides it is equivalent to estimate the following equation:

$$\log Y = b_1 + b_2 \log X$$

$$\Delta Y = \frac{\partial Y}{\partial X}\Delta X = \frac{\partial Y}{\partial \log Y} \cdot \frac{\partial \log Y}{\partial \log X} \cdot \frac{\partial \log X}{\partial X}\Delta X = Y \cdot b_2 \cdot \frac{1}{X}\Delta X = b_2 \cdot Y \cdot \frac{\Delta X}{X} =>$$

$$\frac{\Delta Y}{Y} = b_2 \cdot \frac{\Delta X}{X}$$

Interpretation: A one percentage (1%) increase in X leads to $b_2\%$ increase in Y

Note that for considered specifications the disturbance term ν has a lognormal distribution meaning that its logarithm is distributed normally: $\log \nu \sim Normal$ satisfying the Gauss – Markov conditions. However, if there are reasons to believe that in the true relationship the disturbance term behaves differently, then in the transformed relationship the Gauss-Markov conditions fail and, moreover, direct linearization can be impossible to perform. In this case it can be estimated by means of non-linear OLS.

5) **Quadratic** explanatory variables: $Y = b_1 + b_2 X_2 + b_3 X_2^2$

It is a linear in parameters model but the usual interpretation of b_2 cannot be applied because a unit change in X_2 affects $b_3 X_2^2$ term, so holding all other variables constant is not valid here. Instead, let's use the described general procedure:

$\frac{dY}{dX_2} = b_2 + 2b_3 X_2$ – depending on X_2 the effect is different (changing marginal effect)

Interpretation:

For b_2: It shows the effect of a unit change in X_2 on Y for the special case where $X_2 = 0$ (zero marginal effect).

Let's write the estimated model as: $Y = b_1 + (b_2 + b_3 X_2) X_2$. Hence,

For b_3: It shows the rate of change of the coefficient of X_2 per unit change in X_2.

It should be noted that the fit of quadratic specification is not significantly different from linear and semi-logarithmic specifications. But there are problems associated with extrapolation outside the data range when interpretation can become implausible because quadratic functions consist of both decreasing and increasing segments. Moreover, b_2 can be interpreted only for $X_2 = 0$ but it can be outside the sample range. Therefore, in practice, semi-logarithmic models are more commonly used.

6) **Interactive** explanatory variables: $Y = b_1 + b_2 X_2 + b_3 X_3 + b_4 X_2 X_3$.

The same as before, slope coefficients do not represent individual marginal effects because a unit change in X_2 or X_3 affects the last term $b_4 X_2 X_3$. Let's rewrite the estimated model as follows:

$Y = b_1 + (b_2 + b_4 X_3) X_2 + b_3 X_3$ and equivalently $Y = b_1 + b_2 X_2 + (b_3 + b_4 X_2) X_3$

For b_2: It shows the marginal effect of X_2 on Y, when X_3 is equal to zero.

For b_3: It shows the marginal effect of X_3 on Y, when X_2 is equal to zero.

For b_4: It shows the change in the coefficient of X_2 when X_3 changes by one unit (equivalent to the change in the coefficient of X_3 when X_2 changes by one unit.

However, $X_2 = 0$ or $X_3 = 0$ can be outside the sample range. Hence, this kind of interpretation cannot be used. Let's look at the one more way to interpret the results rescaling X_2 and X_3.

Consider the general specification: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3 + u$. Let's define:

$$X_2^* = X_2 - \bar{X}_2 \qquad X_3^* = X_3 - \bar{X}_3$$
$$X_2 = X_2^* + \bar{X}_2 \qquad X_3 = X_3^* + \bar{X}_3$$

Hence, $Y = \beta_1 + \beta_2 (X_2^* + \bar{X}_2) + \beta_3 (X_3^* + \bar{X}_3) + \beta_4 (X_2^* + \bar{X}_2)(X_3^* + \bar{X}_3) + u$

Denoting: $\beta_1^* = \beta_1 + \beta_2 \bar{X}_2 + \beta_3 \bar{X}_3 + \beta_4 \bar{X}_2 \bar{X}_3$

$$\beta_2^* = \beta_2 + \beta_4 \bar{X}_3 \qquad \beta_3^* = \beta_3 + \beta_4 \bar{X}_2$$

The model is transformed into: $Y = \beta_1^* + \beta_2^* X_2^* + \beta_3^* X_3^* + \beta_4 X_2^* X_3^* + u$

Rewriting the model as before:

$Y = \beta_1^* + (\beta_2^* + \beta_4 X_3^*) X_2^* + \beta_3^* X_3^* + u$ $\qquad$ $Y = \beta_1^* + \beta_2^* X_2^* + (\beta_3^* + \beta_4 X_2^*) X_3^* + u$

Therefore, the coefficients X_2 and X_3 show the marginal effects of the corresponding variable when the other variables are at their sample means (so we do not require for them to be zero). It solves the problem of taking one of regressors equal to zero outside the sample range.

Ramsey's RESET test of functional misspecification:

It is used to test whether some kind of nonlinearity in the model is present. The idea of the test is that if the squared fitted value of the dependent variable is included into the original specification it should capture quadratic and interactive nonlinearity (if present) without necessary giving rise to the problem of multicollinearity (because they are squared).

Procedures:

1) Run the regression: $Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u$
2) Save the fitted variables of Y. $\hat{Y} = b_1 + \sum_{j=2}^{k} b_j X_j$
3) Add $\hat{Y}^2$ to the original specification: $Y = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + \beta_{\hat{Y}^2} \hat{Y}^2 + u$
4) Usual t-test is performed for the hypothesis:

   H_0: No nonlinearity ($\beta_{\hat{Y}^2} = 0$)

   H_1: Some kind of nonlinearity is present ($\beta_{\hat{Y}^2} \neq 0$).

   If the *t* statistic for the coefficient is significant, this indicates that some kind of nonlinearity may be present.

Comparing linear and logarithmic specifications:

This section will analyze how to choose the best specification. In case of models with the same dependent variables, the model with the greatest R^2 (equivalently, lowest RSS) is selected. However, this does not hold for different dependent variables. For example, let's consider linear and logarithmic models:

$$Y = \beta_1 + \beta_2 X + u \qquad \text{and} \qquad \log Y = \beta_1 + \beta_2 X + u$$

By definition, R^2 shows the proportion of the dependent variable explained by the model but shares of Y and $\log Y$ variables are not compatible. In order to compare specifications the Box-Cox transformation is used when the following model is fitted:

$$\frac{Y^\lambda - 1}{\lambda} = \beta_1 + \beta_2 X + u$$

The model is nonlinear in parameters, so standard OLS procedure is invalid and nonlinear regression method is used (typically, it is maximum likelihood estimation).

Special cases:

1) $\lambda = 1 \Rightarrow \frac{Y^\lambda - 1}{\lambda} = Y - 1 \qquad \Rightarrow \qquad Y = \beta_1' + \beta_2 X + u$ – linear model
2) $\lambda \to 0 \Rightarrow \lim_{\lambda \to 0} \frac{Y^\lambda - 1}{\lambda} = \lim_{\lambda \to 0} \frac{Y^\lambda \cdot \log Y}{1} = \log Y$ || using L'Hospital's Rule || $\Rightarrow$ $\log Y = \beta_1 + \beta_2 X + u$ – logarithmic model.

Hence, the regression is estimated and the null hypotheses $\lambda = 1$ and $\lambda = 0$ are tested (separately). The problem can arise because depending on the chosen significance level both hypotheses might be rejected (alternatively, neither of them might be rejected). So, the answer is ambiguous.

If the objective is only to compare the fits of linear and logarithmic specifications the procedure involving Zarembka scaling is used. Let's consider steps:

1) Divide the observations on the dependent variable by their geometric mean to get the transformed variable Y^*

$$Y^* = \frac{Y}{geometric\ mean} = \frac{Y}{(Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n)^{\frac{1}{n}}}$$

2) Estimate:

$Y^* = \beta_1^{'} + \beta_2^{'} X + u$
$\log Y^* = \beta_1^{'} + \beta_2^{'} X + u$ We do not need to care about estimated coefficients (in primes)

3) Testing:

H_0: Qualities (goodness of fit measure) of models are the same;
H_1: The difference in the goodness of fit is significant.

Test-statistics: $\chi^2 = \frac{n}{2} \cdot \log \frac{larger\ RSS}{smaller\ RSS} \underset{asympt}{\overset{H_0}{\sim}} \chi^2(1)$.

Decision: If $\chi^2 > \chi^2_{crit\,\alpha\%}(1)$ then H_0 is rejected in favour of H_1 in $\alpha\%$ significance level.
If $\chi^2 < \chi^2_{crit\,\alpha\%}(1)$ then H_0 is not rejected in $\alpha\%$ significance level.

Production Functions Estimation

Consider Cobb-Douglas Production Function: $Y = A \cdot K^{\alpha} \cdot L^{\beta} \cdot e^{\gamma t}$ where
A – scaling coefficient;
$e^{\gamma t}$ – total factor productivity;
$A \cdot e^{\gamma t}$ – describes neutral technical progress independent from K and L;
α – elasticity of production function with respect to capital;
β – elasticity of production function with respect to labour.

Let's linearize by taking logarithms:

$$\ln Y = \ln A + \alpha \ln K + \beta \ln L + \gamma t$$

Taking the full differential from both sides:

$$\frac{dY_t}{Y_t} = \alpha \cdot \frac{dK_t}{K_t} + \beta \cdot \frac{dL_t}{L_t} + \gamma \cdot dt$$

In growth rates: $y_t = \alpha \cdot k_t + \beta \cdot l_t + \gamma$

Estimation:

Consider specification: $Y_t = A \cdot K_t^{\alpha} \cdot L_t^{\beta} \cdot e^{\gamma t} \cdot \nu_t$ where $\log \nu \sim Normal$ satisfying the Gauss – Markov conditions.

As we got there are 2 alternative ways of estimation:

1) $\ln Y_t = \ln A + \alpha \ln K_t + \beta \ln L_t + \gamma t + \log \nu_t$
2) $y_t = \alpha \cdot k_t + \beta \cdot l_t + \gamma + u_t$

HOWEVER, we should note that these 2 approaches are not strictly equivalent because for getting 2) we used the time derivative which is continuous function of time but in practice we just now discrete growth rates (not continuous ones). Moreover, 2) is more stable in time.

CES (Constant Elasticity of Substitution) Production Function:

$CES\ Function:\quad Y = A \cdot (u \cdot K^{-\rho} + (1-u) \cdot L^{-\rho})^{-n/\rho} e^{\gamma t}$ where

A – factor productivity: $A > 0$

u – share parameter: $0 < u < 1$

n – degree of homogeneity (is determined by returns to scale): $n > 0$

ρ – parameter which determines elasticity of substitution: $\rho \geq -1$

$Elasticity\ of\ Substitution:\quad \sigma_{LK} = \dfrac{d \ln(K/L)}{d \ln(Y'_L / Y'_K)} = \dfrac{1}{1+\rho}$ – degree of level curves' curvature.

Special cases:

For $\rho = -1$ – linear isoquants;

For $\rho = 0$ – Cobb-Douglas Production Function;

For $\rho \to \infty$ – Leontief function.

The function cannot be estimated directly by OLS. Instead, the non-linear technique is used:

$$\ln\left(\frac{Y}{L}\right) = \ln A - \left(\frac{n}{\rho}\right) \cdot \ln\left[u \cdot \left(\frac{K}{L}\right)^{-\rho} + (1-u)\right] + \gamma \cdot t$$

Lecture 8

Dummy variables

Frequently, some factors we would like to include into a regression model are of qualitative nature and, therefore, are not numerically measurable. One possible approach would be to divide observations into several groups in accordance with whether they possess a certain qualitative characteristic, and then analyze the difference between regression coefficients across respective groups. Alternatively, one could estimate a single regression for all observations, measuring the influence of the qualitative factor by introducing a so called ***dummy variable***. This variable takes the value of either zero or unity, depending on whether the given observation possesses the qualitative characteristic we want to account for. It allows to test the significance of the effect of the corresponding qualitative factor. Moreover, under certain assumptions regression estimates become more efficient. This lecture will analyze different ways to include dummy variables into the model in accordance with the initial hypothesis of how the difference in qualitative characteristics can affect the relationship. First, we will illustrate where and how dummy variables are used looking at different examples. Second, the lecture will describe different types of dummy variables, including intercept, slope, and interactive dummy variables. Then it will discuss what is the dummy variable trap and how estimation results depend on the choice of a reference category. And, finally, we will examine the Chow test that enables to compare relationships between different subsamples.

How to define and use:

Dummy variables are used to define a set of categories which are qualitative in nature. Note that dummy variables refer to only explanatory variables. In other words, the dependent variable is never dummy (in fact, if the dependent variable can take either 0 or 1 than it is called a binary choice variable).

Let's consider a qualitative variable that has N categories. The standard procedure is to choose one category as the reference category and to define dummy variables for each of the others. A reference category is used as a basis of comparison and it is a good practice to select the most normal and basic category as the reference category. The number of dummies is equal to $N-1$. However, if we do not include the intercept (constant term) into the model then N dummies are used. We will analyze this result in more detail in the section of the dummy variable trap. Moreover, for models with several qualitative variables each of which defines a corresponding group the result is the following: the number of dummy variables in a group = the number of categories according to the appropriate characteristic -1.

Examples:

- Accounting for gender: If we believe that gender can be a significant factor in explaining the dependent variable (for example, Earnings) then the following procedure and set of dummies (depending on the chosen reference category) can be used:
  1) Define a qualitative variable: gender;
  2) Categories: male and female. Hence, $N=2$;
  3) Define a reference category:

  Reference category: Female

  Dummy: $Female = \begin{cases} 1 \; for \; female \\ 0 \; for \; male \end{cases}$

  Reference category: Male

  Dummy: $Male = \begin{cases} 1 \; for \; male \\ 0 \; for \; female \end{cases}$

- Accounting for economic conditions: crisis

Dummy: $Crisis = \begin{cases} 1 \text{ for period of crisis} \\ 0 \text{ for other periods} \end{cases}$

- Accounting for seasonal effects: some variables differ significantly across the seasons of the year. For example, consider expenditures on fuel black oil for boiler houses. If we want to use the quarterly data to study the changes in expenditure across years, it is necessary to take into account the seasonal factor. This can be done with the help of dummy variables:
  There are 4 seasons (quarters) = 4 categories => 3 dummy variables are used.
  Let's define the first quarter as a reference category. Hence,

$$D_2 = \begin{cases} 1 \text{ for second quarter} \\ 0 \text{ otherwise} \end{cases}$$

Dummy variables: $D_3 = \begin{cases} 1 \text{ for third quarter} \\ 0 \text{ otherwise} \end{cases}$

$$D_4 = \begin{cases} 1 \text{ for fourth quarter} \\ 0 \text{ otherwise} \end{cases}$$

Therefore,

$D_2 = D_3 = D_4 = 0$, if the observation refers to the first quarter;

$D_2 = 1;\ D_3 = D_4 = 0$, if the observation refers to the second quarter;

$D_3 = 1;\ D_2 = D_4 = 0$, if the observation refers to the third quarter;

$D_4 = 1;\ D_2 = D_3 = 0$, if the observation refers to the fourth quarter.

Intercept and slope dummy variables:

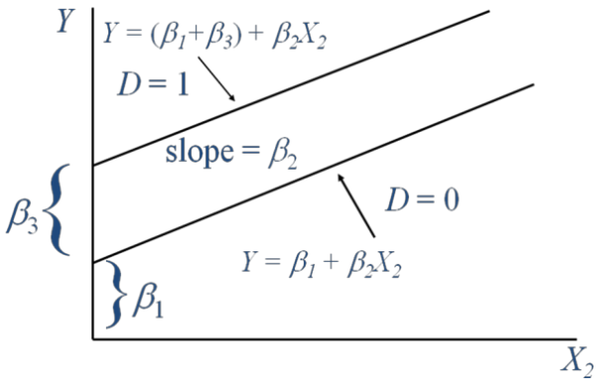
Dummy variables can be used to test for a change in intercept or a change in slope.

Intercept dummy

It assumes that the qualitative variables introduced into the regression are responsible only for shifts in the constant term. The slope of the regression line is identical for each category of the qualitative variables. In other words, marginal effects are not affected with the inclusion of a qualitative characteristic.

Consider a model with one regressor X_2 and one dummy variable D:

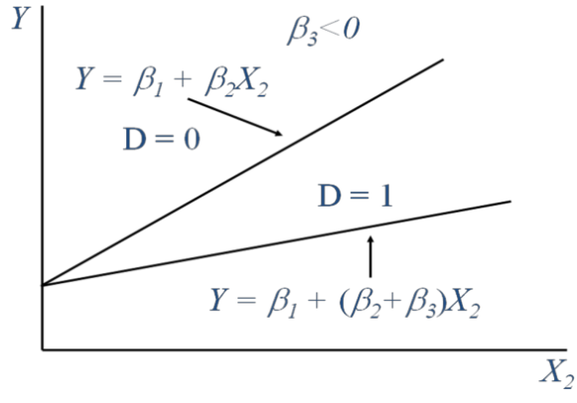
$$Y = \beta_1 + \beta_2 X_2 + \beta_3 \cdot D + u$$



Slope dummy

The assumption that each category of the qualitative variables does not influence the slope of the regression line is not always plausible. Sometimes we might want to allow the slope coefficients on other variables to vary between groups, thus accounting for different marginal effects. It can be done by creating a slope dummy, equal to a dummy variable times another variable:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 (D \cdot X_2) + u$$

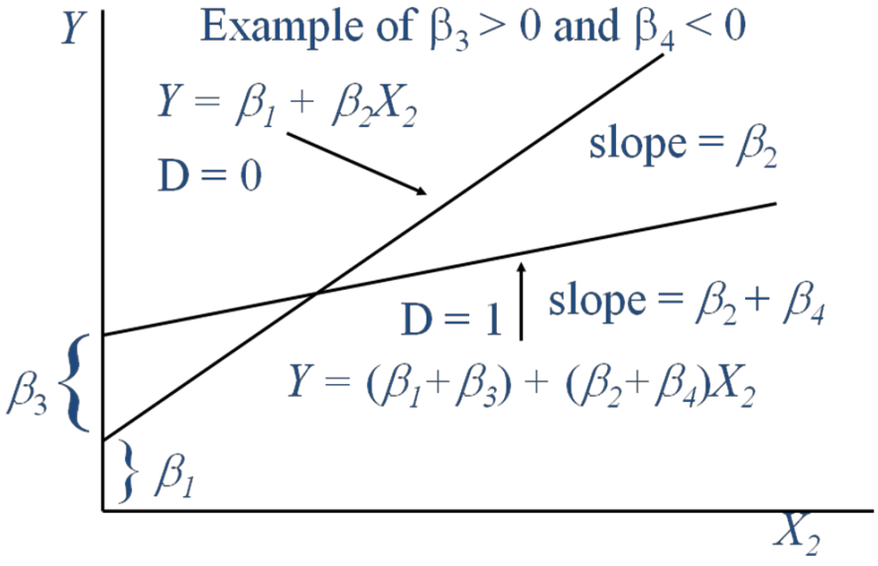


Interpretation: Reference category is $D=0$
If $D=0$, then $Y=\beta_1+\beta_2X_2+u$;
If $D=1$, then $Y=(\beta_1+\beta_3)+\beta_2X_2+u$

Interpretation: Reference category is $D=0$
If $D=0$, then $Y=\beta_1+\beta_2X_2+u$;
If $D=1$, then $Y=\beta_1+(\beta_2+\beta_3)X_2+u$

When we believe that differences in qualitative characteristics have influence on both the intercept and marginal effects of other explanatory variables, then we introduce a complete set of all dummy variables:

$$Y=\beta_1+\beta_2X_2+\beta_3\cdot D+\beta_4(D\cdot X_2)+u$$



If $D=0$, then $Y=\beta_1+\beta_2X_2+u$;
If $D=1$, then $Y=(\beta_1+\beta_3)+(\beta_2+\beta_4)X_2+u$.

Dummy variables of interaction

A dummy variable of interaction is introduced when the presence of two qualitative characteristics simultaneously brings about an additional effect on the dependent variable. Such a dummy variable is defined as the product of two initial dummy variables. For example, supposing that abilities (ASVABC), the number of schooling years (HGC), and 2 dummies: gender (MALE) and ethnicity (ETHWHITE) are factors in determining Earnings (EARN), let's introduce the interaction dummy: MALEWHITE $=MALE\cdot ETHWHITE$, where

$$Male=\begin{cases}1\ for\ male\\0\ for\ female\end{cases}\qquad ETHWHITE=\begin{cases}1\ for\ White\ ethnicity\\0\ for\ otherwise\end{cases}$$

$$LOG(EARN)=\beta_0+\beta_1\cdot ASVABC+\beta_2\cdot HGC+\beta_3\cdot MALE+\beta_4\cdot ETHWHITE+\beta_5\cdot MALEWHITE+u$$

	WHITE	NON-WHITE
MALE	$LOG(EARN)=b_0+b_1\cdot ASVABC+b_2\cdot HGC+$ $+b_3\cdot MALE+b_4\cdot ETHWHITE+b_5\cdot MALEWHITE$	$LOG(EARN)=b_0+b_1\cdot ASVABC+$ $+b_2\cdot HGC+b_3\cdot MALE$
FEMALE	$LOG(EARN)=b_0+b_1\cdot ASVABC+$ $+b_2\cdot HGC+b_4\cdot ETHWHITE$	$LOG(EARN)=b_0+b_1\cdot ASVABC+b_2\cdot HGC$

So, the inclusion of this interaction variable allows to answer the question: are there ethnic variations in the effect of the gender of a respondent on earnings? Formally,

$H_0 : \beta_5 = 0$

If β_5 is significant, then the estimate of the coefficient at the variable MALEWHITE shows, that the observed interaction between gender and ethnicity (in the sense that there is a significant ethnic variation in the effect of gender) makes possible for a white male respondent to earn $b_5\%$ more.

The dummy variable trap:

Consider the following model: $Y = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + \delta_2 D_2 + ... + \delta_s D_s + u$ (1)

So, the qualitative variable has s categories. As was discussed, the general procedure is to include $s-1$ dummies into the model. The dummy variable trap occurs when the reference category is also included together with the constant term and as a result it becomes impossible to fit the model:

$Y = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + \delta_1 D_1 + \delta_2 D_2 + ... + \delta_s D_s + u$ (2) – the dummy variable trap.

Reasons:

1) Intuitive: The intercept dummy variable shows the increase in the intercept relative to that of the reference category but, as we have already included this basic category, there is no room for comparison now. Hence, there is no logical interpretation.
2) Mathematically: exact multicollinearity issue. Let's denote the regressor at the constant term β_1 as X_1. It is identical equal to one: $X_1 \equiv 1$. Now it becomes $\sum_{i=1}^{s} D_i = 1 = X_1$, because one of the dummy variables will be equal to 1 and all the others will be equal to 0 in any observation i. Therefore, there is an exact multicollinearity => no estimates can be obtained.

Solution:

1) Estimate as (1);
2) Drop the constant term: $Y = \beta_2 X_2 + ... + \beta_k X_k + \delta_1 D_1 + \delta_2 D_2 + ... + \delta_s D_s + u$. Note that the interpretation of the coefficients will change.

Change of reference category:

Consider a model which describes how the cost of running a school depends on the number of students and the type of the school (qualitative variable). There are 4 categories: general, technical, skilled workers', and vocational schools. Accordingly, 4 dummy variables are defined: GEN, TECH, WORKER, VOC. The table shows the estimation results depending on the chosen reference category (it is general for the first case and skilled workers' for the second).

```
. reg COST N TECH WORKER VOC
------------------------------------------------------------------------------
        COST |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           N |   342.6335    40.2195     8.519   0.000     262.3978    422.8692
        TECH |   154110.9   26760.41     5.759   0.000     100725.3    207496.4
      WORKER |   143362.4    27852.8     5.147   0.000     87797.57    198927.2
         VOC |   53228.64   31061.65     1.714   0.091    -8737.646    115194.9
       _cons |  -54893.09   26673.08    -2.058   0.043    -108104.4   -1681.748
------------------------------------------------------------------------------

. reg COST N TECH VOC GEN
------------------------------------------------------------------------------
        COST |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           N |   342.6335    40.2195     8.519   0.000     262.3978    422.8692
        TECH |   10748.51   30524.87     0.352   0.726    -50146.93    71643.95
         VOC |  -90133.74   33984.22    -2.652   0.010    -157930.4   -22337.07
         GEN |  -143362.4    27852.8    -5.147   0.000    -198927.2   -87797.57
       _cons |   88469.29   28849.56     3.067   0.003     30916.01    146022.6
------------------------------------------------------------------------------
```

Main results:

The choice of reference category does not affect the substance of the regression results. The table below shows the effects on the model with intercept dummies.

DO NOT CHANGE

1) The goodness of fit measured by:
   - R^2
   - RSS
   - standard error of regression (Root MSE);
2) F-statistics for the whole equation;
3) Coefficients and t-statistics of other variables (non-qualitative ones) => also their standard errors;
4) The coefficients of dummies related to the correspondent reference category for 2 models (in our example it is WORKER for the reference category general (first regression), and it is GEN for the reference category skilled workers' (second regression)) are the same in the absolute magnitude but have different signs. Moreover, standard errors for such coefficients are the same, and, hence, t-statistics only differ in signs.

DO CHANGE

1) Interpretation of t-tests: the meaning of a null hypothesis for a dummy variable coefficient being equal to 0 is different;
2) Coefficients of dummy variables are different;
3) Standard errors of dummy variables are different, hence, no relation in t-statistics.

Note that this result does not apply for the situation described in 4): for coefficients of WORKER and GEN for models with reference categories of general and skilled workers' schools, respectively

Statistical tests:

The test for either a change in intercept or a change in slope can be performed by using standard t-tests on the dummy variable parameters.

Consider a model with both slope and intercept dummies:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 \cdot D + \beta_4 (D \cdot X_2) + u$$

$H_0 : \beta_3 = 0$

Standard t-test: $d.f. = (\#of\ observations) - (\#of\ estimated\ parmeters) = n - 4$ – for our case.

The same applies for the slope dummy:

$H_0 : \beta_4 = 0$.

The test for the ***joint explanatory power of all dummy variables*** is carried out with the help of the following *F-statistics*, calculated on the basis of the residual sums of squares in the models with dummy variables and without them:

H_0 : The coefficients of all dummy variables are simultaneously equal to zero

H_0 : The coefficient of at least at one dummy variable is non-zero.

$$F = \frac{(RSS_{no\ dumies} - RSS_{dummies}) / the\ number\ of\ dummies}{RSS_{dummies} / (the\ number\ of\ observations - the\ total\ number\ of\ parameters\ estimated)}$$

$$\overset{H_0}{\sim} F(number\ of\ dummies, (number\ of\ observations - the\ total\ number\ of\ parameters\ estimated))$$

Chow test:

Sometimes a sample of observations consists of two or more subsamples and it is difficult to decide, whether it is necessary to estimate one regression for the entire sample or separate regressions for all sub-samples. The Chow test is used to solve this problem. It tests the

following hypothesis:

H_o : *the coefficien ts are the same for all subsamples*

H_1 : *at least one coefficien t differs*

Moreover, it will be shown that the Chow test is equivalent to an F test testing the explanatory power of the dummy variables as a group (only if we include the full set of dummies).

I. Chow test for 2 subsamples each of which has k parameters to estimate ($k-1$ explanatory variables, and 1 intercept):

Subsample 1: $Y=\beta_1+\beta_2X_2+\beta_3\cdot X_3+...+\beta_k\cdot X_k+u_1$ sample size n_1 RSS_1

Subsample 2: $Y=\beta_1^{'}+\beta_2^{'}X_2+\beta_3^{'}\cdot X_3+...+\beta_k^{'}\cdot X_k+u_2$ sample size n_2 RSS_2

$$H_o:\begin{cases}\beta_1=\beta_1^{'}\\ \beta_2=\beta_2^{'}\\ .....\\ \beta_k=\beta_k^{'}\end{cases}$$

Procedures:

1) Estimate the regression for the whole sample: $n=n_1+n_2$ and RSS_0
2) F-statistics: $F(k,n-2k)=\dfrac{(RSS_0-(RSS_1+RSS_2))/k}{(RSS_1+RSS_2)/(n-2k)}$
3) Perform F-test: compare to $F^{crit}_{\alpha\%\,significance\,level}(k,n-2k)$:

   If $F(k,n-2k)=\dfrac{(RSS_0-(RSS_1+RSS_2))/k}{(RSS_1+RSS_2)/(n-2k)}>F^{crit}_{\alpha\%}(k,n-2k)$, then we can reject the null hypothesis that the relationships in both samples are the same.

Showing the equivalence:

Let's show that this test is equivalent to the following F-test for linear restrictions:

Define the dummy variable: $D=\begin{cases}1\,if\ observation\,belongs\,to\,sample\,1\\ 0\,if\ observation\,belongs\,to\,sample\,2\end{cases}$

Let's include into the regression the full set of dummies:

$$Y=\beta_1+\beta_1^{'}D+\beta_2X_2+\beta_2^{'}(D\cdot X_2)+\beta_3X_3+\beta_3^{'}(D\cdot X_3)+...+\beta_kX_k+\beta_k^{'}(D\cdot X_k)+u$$

Now the number of estimated parameters is equal to $2k$

So it becomes equivalent to test:

$H_o:\beta_1^{'}=\beta_2^{'}=...=\beta_k^{'}=0$. There are k restrictions

Unrestricted model with RSS_{UR}:

$$Y=\beta_1+\beta_1^{'}D+\beta_2X_2+\beta_2^{'}(D\cdot X_2)+\beta_3X_3+\beta_3^{'}(D\cdot X_3)+...+\beta_kX_k+\beta_k^{'}(D\cdot X_k)+u$$

OLS will choose the intercept b_1 and the *bs* coefficients of X_2 … X_k such that to optimise the fit for the $D=0$ observations. The coefficients will be exactly the same as if the regression has been run with only the subsample of $D=0$ observations. The same logic applies for $D=1$. So, $RSS_{UR}=RSS_1+RSS_2$.

Restricted model with RSS_R:

$$Y=\beta_1+\beta_2X_2+\beta_3\cdot X_3+...+\beta_k\cdot X_k+u$$

Therefore, we get the same $F(k,n-2k)=\dfrac{(RSS_R-RSS_{UR})/k}{(RSS_{UR})/(n-2k)}=\dfrac{(RSS_0-(RSS_1+RSS_2))/k}{(RSS_1+RSS_2)/(n-2k)}$

II. Generalizing for m subsamples:

Using the equivalence result, in this case there are m categories => we include $m-1$ dummies => the number of restrictions is $k(m-1)$ and the number of estimated parameters is mk. Therefore F-statistics:

$$F(k(m-1), n-mk) = \frac{(RSS_0 - (RSS_1 + ... + RSS_m))/(k(m-1))}{(RSS_1 + ... + RSS_m)/(n-mk)}$$

The Chow test is easier to perform than the test of the joint explanatory power of the group of dummy variables, but it is less informative in the sense it does not distinguish between the contributions of each dummy variable to the difference between regressions and does not test them for significance. However, the test statistics and, accordingly, the conclusions of these two tests are identical.

Lecture 9

Model Misspecification

Model specification is concerned with the choice of the functional form that is used to analyze relationships between variables. In fact, the first assumption of Model A requires the model to be correctly specified. In other words, it is assumed that we know precisely which variables should be included. However, in practice no one can be absolutely sure that the chosen functional form is a correct one; therefore specification errors are possible. Primarily, there are two reasons for their existence:

1) the formula is wrong;
2) the list of explanatory variables is wrong.

The objective of this lecture is to discuss the consequences of either including variables that should not be included in the correct specification or leaving out variables that are relevant for the analyzing relationship. So, basically, the lecture will deal with the second reason for the presence of specification errors. Looking at procedures for functional form selection is a more advanced subject that will not be considered in detail but we will grasp the approach to the choice of model specification.

Variable misspecification:

Suppose, it is necessary to choose one of the two functions (1) and (2):

$$Y = \beta_1 + \beta_2 X_2 + u \qquad (1)$$

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad (2)$$

Two kinds of misspecification are possible here:

I. Omitting an important explanatory variable, i.e. estimating the type (1) relationship when the true one is of the type (2).
II. Including unnecessary explanatory variable, i.e. estimating the type (2) relationship when the true one is of the type (1).

Two kinds of misspecification			
		True model	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Fitted model	$\hat{Y} = b_1 + b_2 X_2$	Correct specification,	Omission of a relevant variable
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$	Inclusion of an irrelevant variable	Correct specification,

I. Omission of a relevant variable:

Let's analyze statistical properties of estimated coefficients. Remember that there are several equivalent ways to express formulas for OLS coefficients. This lecture will use a sample covariance/variance approach that can be considered as an alternative way to get the same result as it is done in the lecture slides by using a different tool.

Correct specification: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ (*);

Fitted model: $\hat{Y} = b_1 + b_2 X_2$.

OLS result $$b_2 = \frac{Cov(X_2, Y)}{Var(X_2)}$$

Substituting the expression of Y in (*) for Y and using properties of variance/covariance:

$$b_2 = \frac{Cov(X_2, Y)}{Var(X_2)} = \frac{Cov(X_2, \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u)}{Var(X_2)} =$$

$$b_2 = \frac{Cov(X_2, Y)}{Var(X_2)} = \frac{Cov(X_2, \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u)}{Var(X_2)} =$$

$$= \frac{Cov(X_2, \beta_1) + \beta_2 Cov(X_2, X_2) + \beta_3 Cov(X_2, X_3) + Cov(X_2, u)}{Var(X_2)} =$$

$$= \frac{0 + \beta_2 Var(X_2) + \beta_3 Cov(X_2, X_3) + Cov(X_2, u)}{Var(X_2)} = \underbrace{\beta_2}_{\substack{true \\ value}} + \underbrace{\beta_3 \frac{Cov(X_2, X_3)}{Var(X_2)}}_{bias} + \underbrace{\frac{Cov(X_2, u)}{Var(X_2)}}_{\substack{random \\ component}}$$

As X_2 is non-stochastic and $E(Cov(X_2, u)) = E(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{n}) = \frac{\sum ((X_{2i} - \bar{X}_2) E(u_i - \bar{u}))}{n} = 0$

$$E(b_2) = E(\beta_2 + \beta_3 \frac{Cov(X_2, X_3)}{Var(X_2)} + \frac{Cov(X_2, u)}{Var(X_2)}) = \beta_2 + \beta_3 \frac{Cov(X_2, X_3)}{Var(X_2)}$$

$$bias = E(b_2) - \beta_2 = \beta_3 \frac{Cov(X_2, X_3)}{Var(X_2)} = \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2}$$

If $bias > 0$, then it is called that the estimate is biased upwards.
If $bias < 0$, then it is called that the estimate is biased downwards.

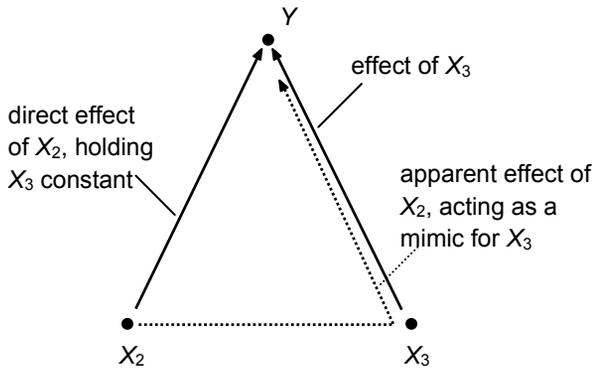
Intuitive explanation:

If X_3 is omitted, then as a result X_2 has 2 effects:

- direct effect on Y;
- apparent indirect effect acting as a proxy for missing X_3.

The strength of the apparent indirect effect depends on 2 factors:

1) the ability of X_2 to mimic the behavior of X_3 that can be derived by regressing X_3 on X_2. In fact, it is a slope coefficient in the regression $X_3 = h_1 + h_2 X_2 + u$ where $h_2 = \frac{Cov(X_2, X_3)}{Var(X_2)} = \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2}$
2) the direct effect of X_3 on Y measured by β_3 that is the significance of the omitted variable in explaining the dependent variable.



Consequences of this type misspecification: other things being equal, estimated coefficients are biased; standard errors, t tests, and F test are invalid.

Comparison of R^2 in the presence of omitted variable bias:

Consider three models:

Correct specification: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ with $R^2_{correct}$ (*i*);

Variable X_3 is omitted: $Y = \beta_1 + \beta_2 X_2 + u$ with $R^2_{om\_var3}$ (*ii*);

Variable X_2 is omitted: $Y = \beta_1 + \beta_3 X_3 + u$ with $R^2_{om\_var2}$ (*iii*).

Let's analyze how the goodness of fit measures are compared. Is it possible to determine the contribution to R^2 of each explanatory variable in (*i*) by running (*ii*) and (*iii*) and then calculating R^2 separately? If it is true, then their separate measures will determine the joint explanatory power exactly in one way: $R^2_{correct} = R^2_{om\_var3} + R^2_{om\_var2}$. BUT IT IS NOT TRUE because of the omitted variable bias. In fact, either $R^2_{correct} > R^2_{om\_var3} + R^2_{om\_var2}$ or $R^2_{correct} < R^2_{om\_var3} + R^2_{om\_var2}$. The answer depends on the direction of biases in (*ii*) and (*iii*):

$$\begin{cases} bias_{ii} > 0 \\ bias_{iii} > 0 \end{cases} <=> \begin{cases} \beta_3 \dfrac{Cov(X_2, X_3)}{Var(X_2)} > 0 \\ \beta_2 \dfrac{Cov(X_2, X_3)}{Var(X_3)} > 0 \end{cases}$$

In (*ii*) the indirect effect of X_2 acting as a proxy of X_3 is positive inflating its apparent explanatory power.

In (*iii*) the indirect effect of X_3 acting as a proxy of X_2 is also positive inflating its apparent explanatory power.

Combining these 2 results:

$$R^2_{correct} < R^2_{om\_var3} + R^2_{om\_var2}$$

$$\begin{cases} bias_{ii} < 0 \\ bias_{iii} < 0 \end{cases} <=> \begin{cases} \beta_3 \dfrac{Cov(X_2, X_3)}{Var(X_2)} < 0 \\ \beta_2 \dfrac{Cov(X_2, X_3)}{Var(X_3)} < 0 \end{cases}$$

In (*ii*) the apparent explanatory power of X_2 is reduced by the negative bias in its coefficient.

Similarly, In (*iii*) the apparent explanatory power of X_3 is reduced by the negative bias in its coefficient.

Combining these 2 effects:

$$R^2_{correct} > R^2_{om\_var3} + R^2_{om\_var2}$$

For other directions of biases the answer to the question what sum will be greater is not definite because it depends on relative strengths of indirect effects. The main point is that in general it is impossible to determine the contribution to R^2 of each explanatory variable in multiple regression analysis.

II. Inclusion of an irrelevant variable:

If X_3 has no effect on the dependent variable but $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ is estimated, then we can express the relationship with $\beta_3 = 0$. However, at the moment of estimation we do not grasp this fact (in other words, we do not use all available information). Therefore, it results in inefficient estimates. At the same time, the properties of OLS estimators, including unbiasedness, do not depend on the true values of the parameters. Formally, the model is still correctly specified, so standard errors remain valid but they tend to be larger indicating the loss in efficiency. Mathematically,

$$\sigma^2_{b_2} = \frac{\sigma^2_u}{\sum (X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r^2_{X_2, X_3}}$$

Therefore, as $0 < 1 - r^2_{X_2, X_3} < 1 => \sigma^2_{b_2} > \dfrac{\sigma^2_u}{\sum (X_{2i} - \bar{X}_2)^2}$ obtained from $Y = \beta_1 + \beta_2 X_2 + u$.

Consequences of this type misspecification: other things being equal, estimated coefficients are unbiased; standard errors, t tests, and F test are valid but efficiency is lower.

Consequences of misspecification			
		True model	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Fitted model	$\hat{Y} = b_1 + b_2 X_2$	Correct specification,	Coefficients are biased (in general). Standard errors and tests are invalid.
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$	Coefficients are unbiased (in general), but inefficient. Standard errors and tests are valid	Correct specification,

The approach to the choice of model specification

The foregoing discussion implies that it is preferable to construct an initial model with maximum number of explanatory variables, and its subsequent improvement by gradual elimination of insignificant variables and deriving more and more efficient estimators. If we go the other way round, i.e. start from the model with a minimum number of explanatory variables and add new ones, the obtained estimates will be biased from the very beginning and the standard errors will be invalid.

However, a criterion is needed to determine, which explanatory variables are insignificant and should be excluded. If we simply exclude all variables with insignificant coefficients, it is possible to make an error because of the presence of multicollinearity. Therefore, the decision to exclude variables is made on the basis of an F-test for the joint explanatory power of several variables (F-test for linear restrictions).

Proxy variables:

There are many examples when one of explanatory variables is some unobserved factor. In particular, it is either not precisely defined (as the quality of education) or it requires a lot of time to obtain data. However, skipping such factor results in omitted variable bias. The problem can be reduced or eliminated by using a proxy variable that is linearly related to the unobserved variable as much as possible.

Consider $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$.

Assume that X_2 is unobserved, so we use a proxy Z such that $X_2 = \lambda + \mu Z$.

The relationship is transformed into:
$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + ... + \beta_k X_k + u \\ &= (\beta_1 + \beta_2 \lambda) + \beta_2 \mu Z + \beta_3 X_3 + ... + \beta_k X_k + u \end{aligned}$$

All variables become observed. If the proxy relationship is an exact one, then

1. The coefficients of $X_3, \ldots, X_k$ will be the same as those we would obtain using X_2 as an explanatory variable.
2. Standard errors and *t*-statistics of these coefficients will be the same as those obtainable when X_2 is used as an explanatory variable.
3. R^2 will be the same as in the model with X_2.
4. The coefficient of Z will provide an estimate of $\beta_2 \mu$, and, consequently, it is impossible to obtain an estimate of β_2, unless μ is known.
5. The *t*-statistic for Z will be same as that we would obtain for X_2, so, we can test the significance of the explanatory variable X_2, even being unable to estimate its coefficient.

6. It is impossible to obtain an estimate of the intercept β_1, since the intercept of the estimated model is equal to $\beta_1 + \beta_2\lambda$. However, it is usually more important to estimate the regression coefficients than the intercept.

In practice, the relationship between the proxy variable and the approximated one is usually not exactly linear, but the above tendencies are still observed and should be taken into account.

Unintended proxies:

If we use a proxy variable without realizing that it is a proxy, then this situation is called unintentional use of proxy. Consequences depend on motives of estimating the regression:

Use to predict future values of the dependent variable ⇩	Use as a policy variable to influence other variables ⇩
If $correlation(X_2, Z)$ is relatively high, then it does not matter whether to include X_2 or Z for that purpose	It matters whether to use X_2 or Z. If there is no functional connection between X_2 and Z, then the proxy has no direct effect on Y

A Monte Carlo experiment: omitted variable

A Monte Carlo is a controlled experiment that allows us to check whether the results of estimated model are plausible. Finite sample distribution properties can be investigated.

We know that omitted variables result in biased estimates. This point can be illustrated by means of Monte Carlo simulation:

Correct specification: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$

Estimated model: $Y = \beta_1 + \beta_2 X_2 + u$

Stages:

1) Choose the true values of $\beta_1, \beta_2, \beta_3$;
2) Choose X_2 and X_3 in each observation;
3) Use random generating process to provide the disturbance term. Run regression: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ and calculate corresponding values of the dependent variable;
4) Use generating values of Y to run the regression with the omitted variable: $Y = \beta_1 + \beta_2 X_2 + u$. Estimate the parameters;
5) Repeat the procedure from step 3.

As a result, when we produce the experiment many times, we are able to obtain the distribution of estimated β_1 and β_2 from step 4). Then we can compare these values to the chosen true parameters of β_1 and β_2 in 1). The finite sample bias can be determined:

$$bias = \beta_3 \frac{Cov(X_2, X_3)}{Var(X_2)} = \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2}$$

Lecture 10

Heteroscedasticity

The assumption A.4 of Model A which is equivalent to the second Gauss-Markov condition requires of the disturbance term to be homoscedastic. It is the same to say that the dispersion of the disturbance term is identical in all observations. Mathematically it is written as $\sigma_{u_i}^2 = \sigma_u^2$ for all i. However, it is not always a plausible assumption: variance is generally speaking different in different observations. Heteroscedasticity is concerned with a non-constant variance of the disturbance term. Formally, it means that $\sigma_{u_i}^2 \neq \sigma_{u_j}^2$ for some $i \neq j$. This lecture will analyze heteroscedasticity according to the following plan:

1. Reasons
2. Consequences
3. Detection
4. Remedial measures

I. Reasons

The problem of heteroscedasticity can arise when the scale of different economic variables changes in the same direction. Suppose, we want to analyze how expenditure on education depends on GDP of a country. We have n observations for n countries and the following model is used:

$$EDUC\_EXP = \beta_1 + \beta_2 GDP + u$$

Obviously, both variables $EDUC\_EXP$ and GDP change their scales simultaneously: a country with a larger GDP can spend in absolute terms much more on education. Therefore, the absolute value of the expected dispersion of the dependent variable $EDUC\_EXP$ will increase with the value of the explanatory variable GDP. The **reason** is that the variances of the omitted variables and the measurement errors, which jointly determine the values of the disturbance term, rise.

Cross-sectional data often give rise to heteroscedasticity. Many economic variables tend to move in size together. For example, consider the sample that contains data on different companies. Heteroscedasticity is likely to arise because large firms will typically display much greater variation (for instance, in profits, costs, expenditures …) than smaller ones. One more example is when we analyze the relationship between consumption expenditures and family income. It is reasonable to suppose that a family with greater aggregate income will have greater variation in consumption expenditures.

II) Consequences

Let's analyze statistical properties the OLS estimators when the disturbance term is subject to heteroscedasticity. Obviously, the fact that the dispersion of the disturbance term is not constant affects standard deviations of the regression estimators because we do not use all available information about counting each observation differently. Nevertheless, the procedure of OLS estimation and checking unbiasedness do not depend on the presence of heteroscedasticity, so expected values of estimators are unchanged. The **main results** are as following:

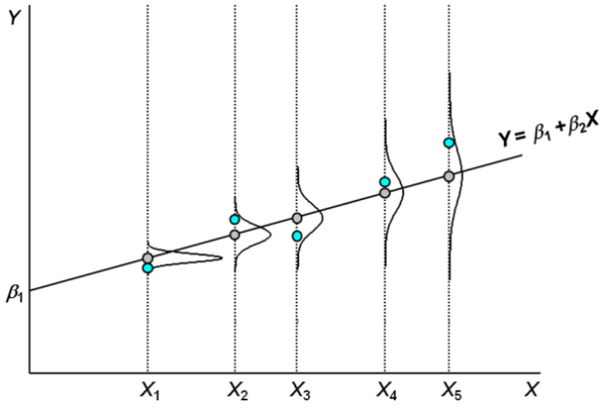
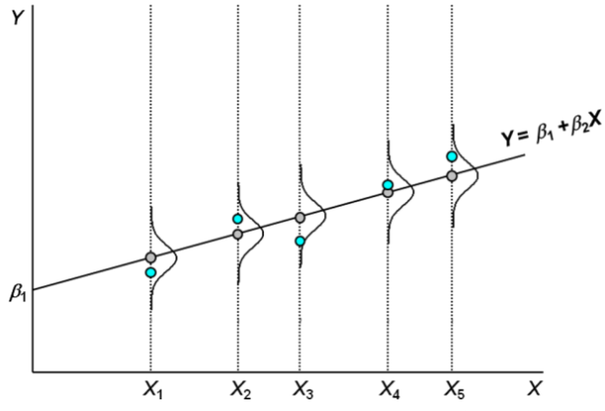
1) Standard errors of the regression coefficients are estimated wrongly (it is likely that they will be underestimated). Hence, t-test and F-test are invalid. Moreover, as the bias of standard errors will be typically negative, t-statistics will be overestimated giving rise to the misleading impression of the precision of regression coefficients;

2) OLS estimators are unbiased and consistent BUT inefficient. It becomes possible to find estimators that are still unbiased with a smaller variance.

Intuitive explanation

$$Y = \beta_1 + \beta_2 X + u$$

Let's illustrate some reasons behind the consequences of heteroscedasticity graphically:

Heteroscedasticity (different distributions of u in terms of variance)	Homoscedasticity (the same distributions of u in terms of variance)
 sample values theoretical values	 sample values theoretical values
Observations with lower variance like X_1 and X_2 tend to be better guides than ones with a higher variance like X_5 BUT OLS procedure count them equally => INEFFICIENCY	Before the sample is generated, each observation of X_i is equally reliable guide to the location of the true regression line

In other words,

The OLS criteria is $\sum_i e_i^2 \to \min$. In the presence of heteroscedasticity, obviously, it would be better to use a rule $\sum_i \gamma_i e_i^2 \to \min$, where γ_i some decreasing function with respect to $\sigma_{u_i}^2$.

Mathematically:

Let's use explicit formulae for the estimated coefficients for the simple linear regression model to analyze their statistical properties. In fact, there are several ways to do so. Let's use the following (see lecture 3):

Model: $Y_i = \beta_1 + \beta_2 X_i + u_i$, where $u_i \sim N(0, \sigma_{u_i}^2)$

OLS estimation: $b_2 = \beta_2 + \sum_i a_i u_i$, where $a_i = \dfrac{X_i - \overline{X}}{\sum_j \left(X_j - \overline{X}\right)^2} = \dfrac{x_i}{\sum_j x_j^2}$

When we take the expectation of b_2 , we have nothing to do with the fact that the variance of the disturbance term is not constant. Hence, there is the same result as before:

$E(b_2) = E(\beta_2 + \sum_i a_i u_i) = \beta_2 + E(\sum_i a_i u_i) = \beta_2 + \sum_i a_i E(u_i) = \beta_2 + \sum_i a_i \cdot 0 = \beta_2$ => unbiased.

We used that $X -$ is non-stochastic and $E(u_i) = 0$.

Precision:

$$\sigma_{b_2}^2 = E\left\{\left(b_2 - E(b_2)\right)^2\right\} = E\left\{\left(b_2 - \beta_2\right)^2\right\} = E(\sum_i a_i u_i)^2 = \sum_{i=1}^{n} a_i^2 E\left(u_i^2\right) + \sum_{i=1}^{n} \sum_{j \neq i} a_i a_j E\left(u_i u_j\right) =$$

$$= \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 0 = \sum_{i=1}^{n} a_i^{\,2} \sigma_i^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sigma_i^2}{\left(\sum_{j=1}^{n} (X_j - \bar{X})^2 \right)^2} = \frac{\sum_{i=1}^{n} x_i^{\,2} \sigma_i^2}{\left(\sum_{j=1}^{n} x_j^{\,2} \right)^2}.$$

Standard errors are biased but White (1980) shows that the following estimator is consistent:

$$\text{s.e.}(b_2) = s_{b_2} = \sqrt{\frac{\sum_{i=1}^{n} x_i^{\,2} e_i^2}{\left(\sum_{j=1}^{n} x_j^{\,2} \right)^2}} = \sqrt{\sum_{i=1}^{n} a_i^{\,2} e_i^2}$$ – heteroscedasticity-consistent standard error

Therefore, if it is impossible to identify the nature of heteroscedasticity, then in large samples heteroscedasticity-consistent standard errors make t-test and F-test asymptotically valid. However, there are some problems:

1) The obtained estimator may not perform well in finite samples;
2) OLS point estimates remain inefficient.

Heteroscedasticity	**Homoscedasticity**
$E(b_2) = \beta_2$	$E(b_2) = \beta_2$
$\sigma_{b_2}^2 = \sum_{i=1}^{n} a_i^{\,2} \sigma_i^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sigma_i^2}{\left(\sum_{j=1}^{n} (X_j - \bar{X})^2 \right)^2}$	$\sigma_{b_2}^2 = \sigma_u^2 \cdot \sum_{i=1}^{n} a_i^{\,2} = \sigma_u^2 \cdot \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\left(\sum_{j=1}^{n} (X_j - \bar{X})^2 \right)^2}$
$s_{b_2}^2 = \frac{\sum_{i=1}^{n} x_i^2 e_i^2}{\left(\sum_{j=1}^{n} x_j^{\,2} \right)^2} = \sum_{i=1}^{n} a_i^{\,2} e_i^2$	$s_{b_2}^2 = \frac{\sum_{i=1}^{n} e_i^2}{(n-2) \cdot \left(\sum_{j=1}^{n} x_j^{\,2} \right)}$
Therefore, depending on the behavior of σ_i^2, the population variance of the heteroscedasticity case can be either greater or less than the standard one of the homoscedasticity case. Generally, **the direction of the bias depends on the pattern of the heteroscedasticity**. It can be shown that when σ_i^2 and $x_i = (X_i - \bar{X})^2$ are positively correlated, the OLS estimation underestimates the true variance of the b_2.	

III) Detection

There are several tests for detection of heteroscedasticity:

1. Goldfeld-Quandt

It assumes that the standard deviation of the disturbance term is proportional to some factor, i.e. $\sigma_i = \gamma \cdot X_i$. Assume that other assumptions of Model A are satisfied.

Testing procedure:

H_0 : Homoscedasticity: $pop.var\,(u_i) = \sigma^2$

H_1 : Heteroscedasticity: $\sigma_i = \gamma \cdot X_i$

1) Arrange all observations according to the factor X;
2) Divide the sample into 3 subsamples: n_1 observations with the smallest X, n_2 observations with the largest X, and the remaining observations ($n-n_1-n_2$) that are dropped entirely. It has been empirically established that the maximum power of the Goldfeld-Quandt test is achieved when $n_1 = n_2 \approx \frac{3}{8}n$. Then 2 separate regressions for the two groups are estimated and the following F-statistics is computed:

$$F = \frac{RSS_2/(n_2-k)}{RSS_1/(n_1-k)} \overset{H_0}{\sim} F(n_2-k, n_1-k),$$ where RSS_1 *is calculated for the first subsample(smallest X)*; RSS_2 *is calculated for the second subsample(largest X)*; k *– number of estimated parameters (including constant)*

In other words, we test whether RSS_2 is significantly greater than RSS_1 using F-test. To perform the test let's compare the calculated F-statistics to $F_{\alpha\%}^{crit}$:

If $F < F_{\alpha\%}^{crit}$, then do not reject the null hypothesis of homoscedasticity at $\alpha\%$ significance level;

If $F > F_{\alpha\%}^{crit}$, then there is enough evidence of heteroscedasticity of the type $\sigma_i = \gamma \cdot X_i$ at $\alpha\%$ significance level.

Note, that this test can be used for the case when the standard deviation of the disturbance term is inversely related to some factor X i.e. $\sigma_i = \frac{\gamma}{X_i}$. In this case we get $RSS_1 > RSS_2$. Hence, $F = \frac{RSS_1/(n_1-k)}{RSS_2/(n_2-k)} \overset{H_0}{\sim} F(n_1-k, n_2-k)$. Then perform the same F-test as before.

2. White test

It is used for the detection of heteroscedasticity of the general type. This test deals with residuals because, in some sense, e_i^2 is a counterpart of σ_i^2.

Testing procedure:

H_0: Homoscedasticity

H_1: Heteroscedasticity (any type).

1) Estimate the equation and get the residuals;
2) Regress the obtained squared residuals on the explanatory variables, their squares, and their cross-products, omitting any duplicative variables (for example, if there is a dummy variable, then its square coincides with the dummy variable => duplicative). As a result, we get so called White auxiliary equation. If there is a perfect multicollinearity, then skip one of the variables. If the number of d.f. is insufficient, then skip cross-terms.
3) Evaluate the general significance of the equation using usual F-test or, alternatively, $\chi^2 - test$:

   χ^2 – statistics: $\chi^2 = n \cdot R^2 \underset{large sample}{\overset{H_0}{\sim}} \chi^2(d.f.)$ where $d.f.$ *– # of regressors in the White auxiliary equation*; n *– sample size*; R^2 *– goodnes of fit for the White auxiliary equation*

   Compare the calculated χ^2 – statistics to $\chi_{\alpha\%}^{crit}$:

If $\chi < \chi_{\alpha\%}^{crit}$, then do not reject the null hypothesis of homoscedasticity at $\alpha\%$ significance level;

If $\chi > \chi_{\alpha\%}^{crit}$, then there is enough evidence of heteroscedasticity of the general type at $\alpha\%$ significance level.

Example:

Consider the model: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$

White auxiliary equation: $e_i^2 = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{2i}^2 + \beta_5 X_{3i}^2 + \beta_6 X_{2i} X_{3i} + v_i$

χ^2 – statistics: $\chi^2 = n \cdot R^2 \underset{large sample}{\overset{H_0}{\sim}} \chi^2(5)$ as there are 5 regressors (# of estimated parameters in White auxiliary equation minus 1).

Note that 5 degrees of freedom are absorbed in the White auxiliary equation, leaving $n-5$ degrees of freedom in this case for the regression. If the regression includes many explanatory variables, then it may cause a problem of insufficiency of d.f. left.

White test	
Advantages	*Disadvantages*
• can be performed if the type of heteroscedasticity is not known; • can detect heteroscedasticity even if it is not connected with factors included in the regression.	• low power – a price for its generality; • only for large samples; • if many explanatory variables are included, then the problems of insufficient d.f. left and possible perfect multicollinearity arise => some variables should be excluded in the White auxiliary equation but what are these variables? No predictions.

These problems explain the reasons for possible different results of the White test and the Goldfeld-Quandt test.

IV) Remedial measures

Weighted Least Squares:

In order to deal with the problem of heteroscedasticity, the weighted least squares (WLS) procedure is usually used. In fact, WLS is a special case of a more general method of generalized least squares (GLS) that allows to transform variables in such a way that standard assumptions of model A are satisfied. As a result, GLS estimators are BLUE.

Let's denote standard deviation of the disturbance term in observation *i* by σ_i. If the value of σ_i in each observation was known, we would be able to eliminate heteroscedasticity by dividing both parts of the equation for each observation by the corresponding σ_i. Then the model takes the form:

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

The disturbance term $\frac{u_i}{\sigma_i}$ here is homoscedastic, since its population variance is

$$\left\{\frac{u_i}{\sigma_i}\right\} = E\left\{\left(\frac{u_i}{\sigma_i}\right)^2\right\} = \frac{1}{\sigma_i^2} E(u_i^2) = \frac{1}{\sigma_i^2} \sigma_{u_i}^2 = 1 \qquad \text{for any } i.$$

In practice, however, the variance of the disturbance term is usually unknown. Nevertheless, we can assume, for example, that σ_i is proportional to some measurable variable Z_i as in the Goldfeld-Quandt test, and then to divide each observation by the corresponding value of Z_i, so that the model becomes:

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

The new disturbance term $\frac{u_i}{Z_i}$ will have a constant variance $= \gamma^2$. It is unknown but it does not matter because crucially it is constant. Now all the assumptions of the model A are satisfied => OLS procedure will give BLUE estimates.

Alternative approach:

Heteroscedasticity can be a cause of an inappropriate mathematical specification. Suppose, in particular, that the true relationship is in fact logarithmic.

$\log Y = \beta_1 + \beta_2 \log X + u$ <=> $Y = e^{\beta_1} X^{\beta_2} e^u$

This specification means that for large values of X the absolute size of the effect of the disturbance term is large, while for small values of X it is small. In other words, u in the logarithmic model is equivalent to multiplicative one in the original specification $Y = e^{\beta_1} X^{\beta_2} e^u$.

Heteroscedasticity: Monte Carlo illustration

The idea of biased standard errors and inefficient estimators can be illustrated with the help of a Monte Carlo experiment.

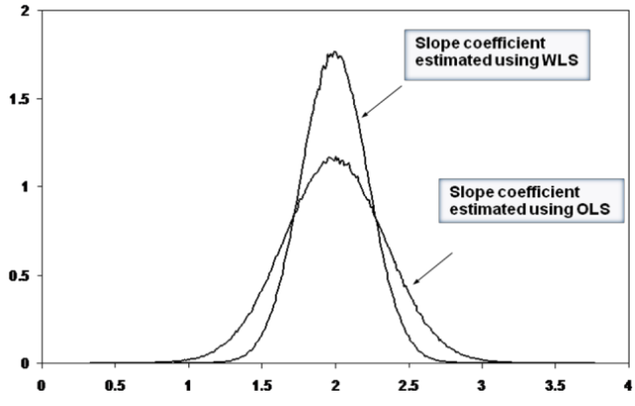
Firstly, it is necessary to generate the data so that the assumption of homoscedasticity fails (in *EViews* this can be done by generating the disturbance term according to the formula $u_i = X_i \cdot NRND$).

Secondly, having conducted a large number of experiments (1 million ones), it is possible to calculate the true standard deviations of the regression coefficients according to the formula *s.d. (b)* $= \sqrt{\frac{\sum (b_i - \beta)^2}{n}}$ and obtain their distribution. Moreover, if we specify the type of heteroscedasticity as $\sigma_i = \gamma \cdot X_i$, then it can be eliminated by means of WLS that allows us to compare efficiency of estimates.

Finally, knowing the true standard deviations, it is possible to determine the direction of bias of standard errors' estimates and analyze efficiency issues of regressors' coefficients.

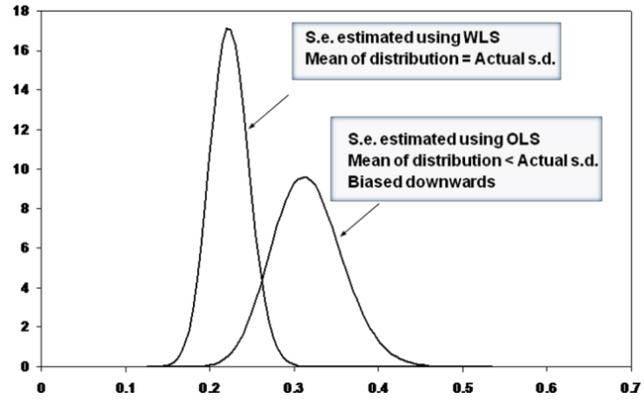
Illustration:

Inefficiency



Both OLS and WLS are unbiased but WLS is more efficient

Biased standard errors



Comparing standard errors it is evident that they are biased.

Lecture 11

Stochastic Regressors. Measurement Errors.

Until now we have analyzed regression models within the framework of Model A, where explanatory variable do not have random components (they are non-stochastic). It is equivalent to say that if we ran the same regression as before with a new sample, the values of the regressors would not change, but the dependent variable would be different due to new values of the disturbance term. This assumption is usual for natural sciences, where experiments can be repeated, but it sounds quite unrealistic for economic data because we cannot resample the given data. The reason was to simplify the analysis of regression coefficients' statistical properties (unbiasedness and efficiency).

This lecture will relax this assumption about non-stochastic regressors and proceed to Model B, where the values of the explanatory variables are assumed to be drawn randomly from defined populations. It is much more realistic way to analyze economic relationships.

Assumptions of Model B:

B.1 *The model is linear in parameters and correctly specified:* $Y = \beta_1 + \beta_2 \cdot X_2 + ... + \beta_m \cdot X_m + u$;

B.2 ***The values of the regressors are drawn randomly from fixed populations;***

B.3 *There does not exist an exact linear relationship among the regressors;*

B.4 Linearity: *The disturbance term has zero expectation*. $E(u_i) = 0$ for all i;

B.5 Homogeneity: *The disturbance term is homoscedastic*. $\sigma^2_{u_i} = \sigma^2_u$ for all i;

B.6 Independence: *The values of the disturbance term have independent distributions*. u_i is distributed independently of u_j for all $j \neq I$;

B.7 ***The disturbance term is distributed independently of the regressors:*** u_i **is distributed independently of** $X_{ji'}$ **for all** i' **and all** j**;**

4 Gauss-Markov conditions

B.8 *The disturbance term has a normal distribution:* $u_i \sim N(0, \sigma^2_u)$

Note that there are 2 new assumptions: B.2 and B.7. All other Model B assumptions are similar to those of the Model A.

Properties of the regression coefficients:

Consider the simple linear regression model with a stochastic explanatory variable: $Y_i = \beta_1 + \beta_2 X_i + u_i$, where the variable X contains stochastic components. In other words, X is drawn randomly from a population with finite mean and variance.

The properties of the estimators of the model parameters depend on the character of the relationship between the stochastic explanatory variable and the disturbance term. It is possible to distinguish between several cases:

I. The stochastic explanatory variable X is distributed independently of the disturbance term u (model B):

   a) Unbiasedness:

   1st approach to show:

$b_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \beta_2 + \sum a_i u_i$, where $a_i = \frac{X_i - \bar{X}}{\sum_j (X_j - \bar{X})^2}$.

$E(b_2) = E(\beta_2) + E(\sum a_i u_i) = \beta_2 + \sum E(a_i u_i)$.

Since X is stochastic, we cannot treat a_i as fixed => $\sum E(a_i u_i) \neq \sum (a_i E(u_i))$.

Instead, let's use the fact that if X and Y are independent random variables , then the expectation of the product can be expressed as $E\{f(X)g(Y)\} = E\{f(X)\}E\{g(Y)\}$, where g and f some functions. Let's denote:

$$f(X) = a_i, \qquad g(Y) = u_i.$$

Hence, $E(a_i u_i) = E(a_i)E(u_i)$. Using B.3 (existence of a_i) and B.4 ($E(u_i) = 0$) => $\sum E(a_i u_i) = \sum (E(a_i)E(u_i)) = \sum 0 = 0$.

$E(b_2) = E(\beta_2) + E(\sum a_i u_i) = \beta_2 + 0 = \beta_2$ – unbiased

$E(b_1) = E(\bar{Y} - b_2 \bar{X}) = E(\beta_1 + \beta_2 \bar{X} + \bar{u}) - E(b_2 \bar{X}) = \beta_1 + \beta_2 E(\bar{X}) + 0 - E(\bar{X}(\beta_2 + \sum a_i u_i)) =$

$= \beta_1 + \beta_2 E(\bar{X}) - \beta_2 E(\bar{X}) + E((\bar{X} \cdot a_i) \cdot u_i) = \beta_1$ – unbiased

We used independence of $\bar{X} \cdot a_i$ and u_i => $E((\bar{X} \cdot a_i) \cdot u_i) = 0$.

2nd approach to show:

$b_2 = \frac{Cov(X,Y)}{Var(X)} = \beta_2 + \frac{Cov(X,u)}{Var(X)}$;

Now we cannot rewrite: $E\left(\frac{Cov(X,u)}{Var(X)}\right) = \frac{1}{Var(X)} \cdot E(Cov(X,u))$.

Instead, let's express: $\frac{Cov(X,u)}{Var(X)} = \frac{\frac{1}{n} \cdot \sum (X_i - \bar{X})(u_i - \bar{u})}{Var(X)} = \frac{1}{n} \cdot \sum \left(\frac{X_i - \bar{X}}{Var(X)}\right) \cdot (u_i - \bar{u})$.

Denoting $f(X_i) = \frac{X_i - \bar{X}}{Var(X)}$ and $g(u_i) = u_i - \bar{u}$. Using independence of these 2 functions and $E(u_i - \bar{u}) = 0$ => $E\left(\frac{Cov(X,u)}{Var(X)}\right) = 0$. Hence, $E(b_2) = E(\beta_2) + 0 = \beta_2$.

b) <u>Consistency:</u>

$plim\, b_2 = plim\left(\frac{Cov(X,Y)}{Var(X)}\right) = plim\left(\beta_2 + \frac{Cov(X,u)}{Var(X)}\right) = plim\, \beta_2 + plim\left(\frac{Cov(X,u)}{Var(X)}\right)$.

Using the properties of probability limits: $plim\, \frac{A}{B} = \frac{plim\ A}{plim\ B}$ provided both $plim\ A$ and $plim\ B$ exist.

$plim(Cov(X,u)) = \text{cov}(X,u)$
$plim(Var(X)) = \text{var}(X)$ => $plim\left(\frac{Cov(X,u)}{Var(X)}\right) = \frac{\text{cov}(X,u)}{\text{var}(X)} = 0$, because X and u are independent and $\text{var}(X) \neq 0$ using B.3. Hence,

$plim\, b_2 = plim\, \beta_2 + plim\left(\frac{Cov(X,u)}{Var(X)}\right) = \beta_2 + 0 = \beta_2$ – consistent

$plim\, b_1 = plim\,(\beta_1 + \beta_2 \bar{X} + \bar{u}) - plim\, b_2\ plim\, \bar{X} = \beta_1$ – consistent

II. X and u are contemporaneously uncorrelated:

This case can be demonstrated by using the lagged dependent variable as one of the explanatory variables. For example: $Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + u_t$, where u_{t-1} has a direct influence on Y_{t-1} and Y_{t-1} is indirectly affected by all previous values of the disturbance term. Hence, B.7 is violated => OLS is not an unbiased estimator. However, consistency proof is not changed: it can be shown that $plim(Cov(Y_{t-1}, u_t) = 0$. **Biased** but **consistent**.

III. X and u are contemporaneously correlated:

In this case estimators are **biased** and **inconsistent**. In large samples the magnitude of the bias tends to $\frac{plim Cov(X,u)}{plim Var(X)} = \frac{plim Cov(X,u)}{\sigma_X^2}$. If the variance of the explanatory variable is infinitesimally large (i.e. σ_X^2 tends to infinity), the magnitude of the bias tends to zero.

Measurement errors

It frequently happens in economics that variables are measured with error. This provides an important application of the use of stochastic regressors. There are 2 different types of measurement errors:

1) **<u>Measurement error in explanatory variable:</u>**

Suppose that Y is determined by a variable Z, but Z is subject to measurement error, w. We have observations of X but the actual relationship is determined by Z:

$Y = \beta_1 + \beta_2 Z + v \quad - actual\ relationship$

$X = Z + w \quad - measured\ with\ random\ error\ w$

Assume that $w \sim Normal(0, \sigma_w^2)$, $u \sim Normal(0, \sigma_u^2)$, w and u are uncorrelated.

Substituting for Z from the second equation:

$Y = \beta_1 + \beta_2 (X - w) + v = \beta_1 + \beta_2 X + v - \beta_2 w = \beta_1 + \beta_2 X + u$, where $u = v - \beta_2 w$.

Hence, in the model $Y = \beta_1 + \beta_2 X + u$, B.7 is violated: $\text{cov}(X, u) = \text{cov}\big((Z + w), (v - \beta_2 w)\big) \neq 0$.

As a result, OLS estimators are inconsistent and the sign of the bias of the slope coefficient β_2 is determined by the sign of β_2. Let's show it:

$$b_2 = \frac{Cov(X,Y)}{Var(X)} = \beta_2 + \frac{Cov(X,u)}{Var(X)}$$

Note that since both the numerator and the denominator are functions of w and there are no expected value rules that can allow us to simplify, we cannot analyze small sample properties. Let's look at large sample ones:

$$plim\, b_2 = plim\left(\beta_2 + \frac{Cov(X,u)}{Var(X)}\right) = \beta_2 + plim\left(\frac{Cov(X,u)}{Var(X)}\right) = \beta_2 + \frac{\text{cov}(X,u)}{\text{var}(X)}$$

$$\text{cov}(X,u) = \text{cov}\big((Z + w), (v - \beta_2 w)\big) = cov(Z, v) + cov(w, v) + cov(Z, -\beta_2 w) + cov(w, -\beta_2 w) =$$

$$= 0 + 0 + 0 - \beta_2 \sigma_w^2 = -\beta_2 \sigma_w^2$$

$$\text{var}(X) = \text{var}(Z + w) = \text{var}(Z) + \text{var}(w) + 2\,\text{cov}(Z, w) = \sigma_Z^2 + \sigma_w^2 + 0 = \sigma_Z^2 + \sigma_w^2$$

Hence, $plim\, b_2 = \beta_2 - \frac{\beta_2 \sigma_w^2}{\sigma_Z^2 + \sigma_w^2} = \beta_2 - \beta_2 \cdot \frac{\sigma_w^2}{\sigma_Z^2 + \sigma_w^2}$ – inconsistent

$$bias(in\ large\ samples) = -\beta_2 \cdot \frac{\sigma_w^2}{\sigma_Z^2 + \sigma_w^2}$$

Therefore, the size of the bias depends on the relative sizes of the variances of w and Z and it is

also determined by the slope coefficient (both value and sign).

Consider the constant term $b_1 = \overline{Y} - b_2\overline{X} = \beta_1 + \beta_2\overline{X} + \overline{u} - b_2\overline{X} = \beta_1 + \beta_2\overline{X} + \overline{v} - \beta_2\overline{w} - b_2\overline{X}$.

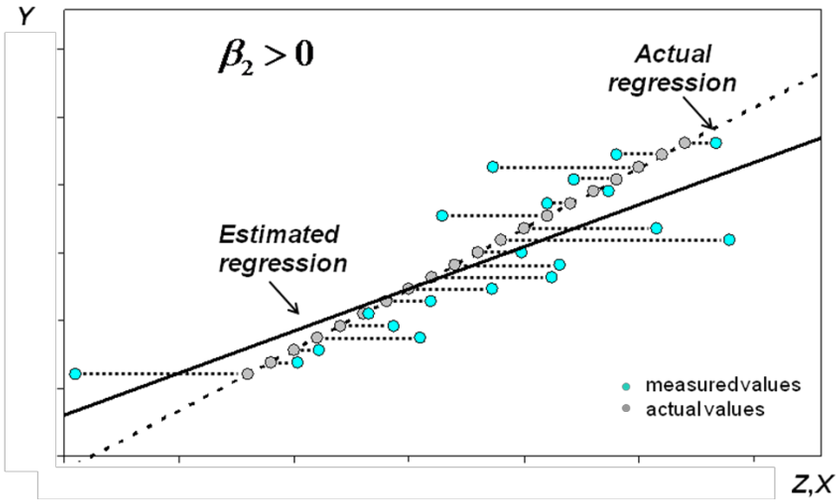
Using properties of $plim$ we get:

$plim\, b_1 = \beta_1 + (\beta_2 - plim\, b_2)\, plim\overline{X} + plim\, \overline{v} - \beta_2 plim\, \overline{w}$.

Probability limit of the sample mean of some variable Q is its expected value: $plim\overline{Q} = E(Q)$.

Hence, $\boxed{plim\, b_1} = \beta_1 + (\beta_2 - plim\, b_2)\, plim\overline{X} = \beta_1 + \beta_2 \frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2} E(X) = \boxed{\beta_1 + \beta_2 \frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2} E(Z)}$

Illustration:



In the figure the measurement error in each observation is shown by the horizontal distance between measured and actual values of the explanatory variable. The true relationship is represented by the dashed line, while the solid line shows the relationship between Y and X.

As can be seen from the graph, for positive actual relationship ($\beta_2 > 0$) the fitted regression line underestimates the true slope (it is flatter), while overestimates the constant term.

2) <u>Measurement error in dependent variable:</u>

$Q = \beta_1 + \beta_2 X + v$ – *actual dependent variable*

$Y = Q + r$ – *measured variable*

r - *mearurement error. Suppose* $E(r) = 0$

We can rewrite the model in terms of the observable variables by substituting for Q from the second equation: $Y - r = \beta_1 + \beta_2 X + v$. Hence,

$Y = \beta_1 + \beta_2 X + v + r = \beta_1 + \beta_2 X + u$, where $u = v + r$

The only difference from the actual relationship is that the new disturbance term u reflects the fact of larger variances of the coefficients: $\sigma_{b_2}^2 = \frac{\sigma_u^2}{n\sigma_X^2} = \frac{\sigma_v^2 + \sigma_r^2}{n\sigma_X^2}$. The regressor has not been affected. Hence, OLS still yields unbiased and consistent estimates provided that X is nonstochastic or that it is distributed independently of v and r.

Main results

	Measurement error in explanatory variable	Measurement error in dependent variable
Relationship	$Y = \beta_1 + \beta_2 Z + v$ – *actual relationship* $X = Z + w$ – *measured with random error* w $w \sim Normal(0, \sigma_w^2), u \sim Normal(0, \sigma_u^2),$	$Q = \beta_1 + \beta_2 X + v$ – *actual dependent variable* $Y = Q + r$ – *measured variable* r - *mearurement error and* $E(r) = 0$

	w and u are uncorrelated.	
Assumption B.7	Violated	Holds
Consistency	Inconsistent: $plim\, b_2 = \beta_2 - \frac{\beta_2 \sigma_w^2}{\sigma_Z^2 + \sigma_w^2} = \beta_2 - \beta_2 \cdot \frac{\sigma_w^2}{\sigma_Z^2 + \sigma_w^2}$ $plim\ b_1 = \beta_1 + \beta_2 \frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2} E(Z)$	Consistent
Bias	Since estimators are inconsistent, it is safe to assume that they are biased in finite samples as well. Bias in large samples $= plim(b) - \beta$ $bias(for\, b_1) = \beta_2 \frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2} E(Z)$ $bias(for\, b_2) = -\beta_2 \cdot \frac{\sigma_w^2}{\sigma_Z^2 + \sigma_w^2}$	Unbiased
s.e.	Invalid	Valid (but should account for larger variances)
t-test	Invalid	Valid
F-test	Invalid	Valid

Example: Friedman's consumption function

The most famous example of applying the measurement error analysis in economic theory is Milton Friedman's Permanent Income Hypothesis (PIH).

In the model the consumption depends not on the current income of an individual but on the permanent income (denoted by Y^P). Permanent income is the individual's estimate of the average annual revenue expected over the long term. Being a subjective concept, permanent income is unobservable. Similarly, Friedman makes a distinction between actual consumption, C_i, and permanent consumption, C_i^P that is determined by the level of permanent income. Thus, the actual relationship is described as $C^P = b_2 Y^P$. However, we possess data on the actual income and consumption, which in every given year can differ from the permanent levels by the corresponding transitory components. Let's assume that Friedman's hypothesis concerning the form of the consumption function is true, but we try to estimate a regression of consumption on actual income, instead of permanent one. The following model is considered:

True model	$Q = \beta_1 + \beta_2 Z + v$	$C^P = \beta_2\, Y^P$
Measurement errors	$X = Z + w$ $Y = Q + r$	$Y = Y^P + Y^T$ $C = C^P + C^T$
True model in measured variables	$Q = \beta_1 + \beta_2 X + v + r - \beta_2 w =$ $= \beta_1 + \beta_2 X + u$	$Q = \beta_2 Y + C^T - \beta_2 Y^T =$ $= \beta_2 Y + u$
$plim\, b_2^{OLS}$	$\beta_2 - \beta_2 \frac{\sigma_w^2}{\sigma_Z^2 + \sigma_w^2}$	$\beta_2 - \beta_2 \frac{\sigma_{Y^T}^2}{\sigma_{Y^P}^2 + \sigma_{Y^T}^2}$
Assumptions	v, w, and r distributed independently of each other, Z and Q	Y^T and C^T distributed independently of each other, Y^P and C^P

Let's study the consequences of the use of OLS for estimating the consumption function under the Permanent Income Hypothesis. We will use a **Monte Carlo** experiment for the illustration:

1) Generate a sample of 20 observations, for which permanent income takes the values 2000, increasing in steps of 100 to 3,900. Thus, $\sigma_{Y^P}^2 = 325{,}000$;

2) Generate the value of the transitory income as a random number drawn from a population following a normal distribution with zero expected value and a constant variance: $Y^T \sim 400 \cdot N(0,1)$. Accordingly, $\sigma^2_{Y^T} = 160000$;

3) Let the true value of β be 0.9.

Hence, given the choice of variables, in large samples:

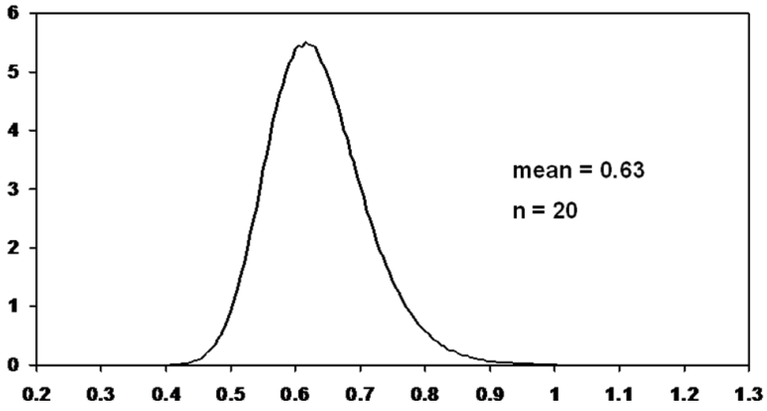
$$plim\, b_2^{OLS} = 0.9 - 0.9\frac{160{,}000}{325{,}000 + 160{,}000} = 0.9 - 0.29 = 0.61$$

Regressing the consumption on the actual income yields the following outcome

$$\hat{C} = 1{,}001 + 0.56Y$$
$$s.e.\ \ (251)\ \ (0.08)$$

In accordance with theoretical predictions, the estimated marginal propensity to consume is below the true value. Moreover, this true value is outside the 95% confidence interval which could be constructed on the basis of the regression output. Therefore, the hypothesis $H_0 : \beta = 0.9$ would be rejected at the 5% significance level, so that we would make a Type I error.

When the Monte Carlo simulation was repeated 1,000,000 times, the following distribution of the slope coefficient appears:



Clearly, there is a strong downward bias in the slope coefficient. This underestimation of the marginal propensity to consume leads to an underestimation of the expenditure multiplier that can affect policy decisions.

Note that even though the sample size is small the probability limit for the slope coefficient $plim\, b_2^{OLS} = 0.61$ is close to the mean value = 0.63.

Lecture 12

Instrumental Variables

A necessary condition for consistency of OLS estimators is that the disturbance term is distributed independently of the regressors (assumption B.7 is satisfied). It is said that there is endogeneity in the regression model if some of the explanatory variables are correlated with the disturbance term. Variables correlated with the disturbance term are defined as endogenous variables, while variables uncorrelated with the disturbance term are called exogenous variables. There are many sources of endogeneity. As we have studied earlier, endogeneity would occur if:

1) There are omitted variables that are correlated with at least one of the included explanatory variables;
2) There is a measurement error in one of the explanatory variables.

In the next lecture we will cover one more source of endogeneity when models comprise two or more simultaneous relationships. The main consequences of assumption B.7 violation consist in the following: OLS estimators become inconsistent, standard statistics are calculated wrongly and statistical tests are invalid. We can deal with these problems by applying instrumental variables (IV) estimation.

Instrumental variables

A valid instrumental variable Z is a special proxy variable for which the following conditions must be satisfied:

1) Instrument exogeneity: It is uncorrelated with the disturbance term, i.e. $cov(Z,u)=0$;
2) Instrument relevance: It is sufficiently strongly correlated with the corresponding endogenous variable, i.e. $cov(Z,X)\neq 0$

The main objective of their application is to obtain consistent estimates of the parameters once the assumption B.7 is violated for the initial specification. Let's consider a simple linear regression model:

$$\boxed{Y=\beta_1+\beta_2 X+u,\quad X \text{ is related to } u}$$

OLS will give inconsistent estimates, i.e.:

$$b_2^{OLS}=\frac{Cov(X,Y)}{Var(X)}=\frac{Cov(X,\beta_1+\beta_2X+u)}{Var(X)}=\frac{Cov(X,\beta_1)}{Var(X)}+\frac{Cov(X,\beta_2X)}{Var(X)}+\frac{Cov(X,u)}{Var(X)}=$$
$$=0+\beta_2\frac{Var(X)}{Var(X)}+\frac{Cov(X,u)}{Var(X)}=\beta_2+\frac{Cov(X,u)}{Var(X)}$$

By taking the probability limit and using its properties, we can show that the resulting estimator is inconsistent:

$$plim\left(b_2^{OLS}\right)=plim\left(\beta_2+\frac{Cov(X,u)}{Var(X)}\right)=\beta_2+\frac{plim\left(Cov(X,u)\right)}{plim\left(Var(X\right)}=\beta_2+\frac{cov(X,u)}{\text{var}(X)}\neq\beta_2,\ as\ cov(X,u)\neq 0$$

Let's introduce an instrumental variable Z, correlated with X and uncorrelated with u. Let's show that the estimator of the parameter β_2 obtained with the use of the instrumental variable, defined as

$$b_2^{IV}=\frac{Cov(Z,Y)}{Cov(Z,X)}=\frac{\sum\left(Z_i-\bar{Z}\right)\left(Y_i-\bar{Y}\right)}{\sum\left(Z_i-\bar{Z}\right)\left(X_i-\bar{X}\right)}$$ is consistent:

$$b_2^{IV} = \frac{Cov(Z,Y)}{Cov(Z,X)} = \frac{Cov(Z,\beta_1 + \beta_2 X + u)}{Cov(Z,X)} = \frac{Cov(Z,\beta_1)}{Cov(Z,X)} + \frac{Cov(Z,\beta_2 X)}{Cov(Z,X)} + \frac{Cov(Z,u)}{Cov(Z,X)} =$$
$$= 0 + \beta_2 \frac{Cov(Z,X)}{Cov(Z,X)} + \frac{Cov(Z,u)}{Cov(Z,X)} = \beta_2 + \frac{Cov(Z,u)}{Cov(Z,X)}$$

By taking the probability limit and using its properties, we can show that IV estimator is consistent:

$$plim\left(b_2^{IV}\right) = plim\left(\beta_2 + \frac{Cov(Z,u)}{Cov(Z,X)}\right) = \beta_2 + \frac{plim\left(Cov(Z,u)\right)}{plim\left(Cov(Z,X)\right)} = \beta_2 + \frac{cov(Z,u)}{cov(Z,X)} = \beta_2 + \frac{0}{cov(Z,X)} = \beta_2,$$

$as\ cov(Z,u) = 0\ and\ cov(Z,X) \neq 0\ \ by\ construction$

Note that it is not possible to demonstrate whether the estimator is unbiased because we are unable to take expectations since X is non-stochastic that is distributed not independently of u, i.e.

$$E\left(b_2^{IV}\right) = E\left(\beta_2 + \frac{Cov(Z,u)}{Cov(Z,X)}\right) = \beta_2 + E\left(\frac{Cov(Z,u)}{Cov(Z,X)}\right) \neq \beta_2 + \frac{E\left(Cov(Z,u)\right)}{E\left(Cov(Z,X)\right)}$$

The population variance of the Instrumental Variables estimator of the slope coefficient of the simple linear regression is given by the following expression (valid for the large samples):

$$\text{var}\left(b_2^{IV}\right) = \sigma^2_{b_2^{IV}} = \frac{\sigma_u^2}{\sum\left(X_i - \bar{X}\right)^2} \times \frac{1}{r^2_{X,Z}}$$

Compare this expression to that for the variance of the OLS estimator:

$$\text{var}\left(b_2^{OLS}\right) = \frac{\sigma_u^2}{\sum\left(X_i - \bar{X}\right)^2}$$

The variance of b_2^{IV} can be obtained by multiplying b_2^{OLS} by $\frac{1}{r^2_{X,Z}}$. The higher the correlation between X and Z, the smaller this multiplier and, therefore, the smaller the variance of b_2^{IV}. Therefore, facing a choice between several potential instrumental variables, it is necessary to choose the one which is most strongly correlated with X because, other things being equal, it will yield the most efficient estimators. At the same time, it would be undesirable to use an instrumental variable perfectly correlated with X, even if one could be found, because it would automatically be correlated with u as well and we would still get inconsistent estimators. We need an instrumental variable, most strongly correlated with X but uncorrelated with u.

Example: Instrumental variables in Friedman's model (Nissan Liviatan)

To illustrate the use of instrumental variables method, we will show how it can be applied to solve the problem of inconsistency arising when Friedman's consumption function is estimated (see Lecture 10). Suppose, we have data on consumption and income of a sample of households in two sequential years. We shall denote consumption and income in the first year by C_1 and Y_1 and those in the second year by C_2 and Y_2.

If Friedman's theory is correct, Y_2 can act as an instrumental variable for Y_1. Obviously, it is likely to be closely correlated with Y_1, so that one of the two requirements to a good instrumental variable is satisfied. Second, if the transitory components of the measured income in different years are uncorrelated, as Friedman assumed, Y_2 will be uncorrelated with the disturbance term in the regression of C_1 on Y_1; thus, the other condition is also satisfied.

It is also possible to use C_2 as an instrumental variable for Y_1. It is strongly correlated with Y_2

(and, therefore, with Y_1 as well) while being uncorrelated with the disturbance term in the relationship between C_1 and Y_1 (if, in accordance with Friedman's hypothesis, the transitory components of consumption are uncorrelated with one another).

Similarly, it is possible to estimate regressions for the second year consumption, using C_1 and Y_1 as instrumental variables for Y_2.

This described approach, where a lagged explanatory variable is used as an instrument, is quite often employed in econometric modelling. In fact, economic variables are frequently connected by both direct and inverse relationships. For example, in the model $Y = \beta_1 + \beta_2 X + u$ the value of X can itself depend on the corresponding value of Y (endogeneity problem). This leads to a violation of the 4-th Gauss-Markov condition, i.e. to a correlation between the explanatory variable and the disturbance term, and, consequently, to a bias in OLS estimators. If, however, a lagged value of X is used as an explanatory variable and if autocorrelation is absent, then consistent estimators can be obtained, which will also be quite reliable, since the subsequent values of X are strongly correlated (this is the case in most economic models).

Asymptotic and finite-sample distributions of the IV estimator

The distribution of the IV estimator degenerates to a spike. In fact, the expression for the variance may be rewritten as shown.

$$\operatorname{var}\left(b_2^{IV}\right) = \sigma^2_{b_2^{IV}} = \frac{\sigma_u^2}{\sum\left(X_i - \overline{X}\right)^2} \times \frac{1}{r^2_{X,Z}} = \frac{\sigma_u^2}{n\left(\frac{1}{n}\sum\left(X_i - \overline{X}\right)^2\right)} \times \frac{1}{r^2_{X,Z}} = \frac{\sigma_u^2}{n\,MSD(X)} \times \frac{1}{r^2_{X,Z}}$$

$MSD(X)$ is the mean square deviation of X. By a law of large numbers (LLN), $MSD(X)$ tends to the population variance of X that is non-zero. Therefore, as n appears in the denominator, the variance of b_2^{IV} is inversely related to n => it tends to zero for large n and its distribution collapses to the spike.

Next, let's show the asymptotic normality. Consider the distribution of $\sqrt{n}\left(b_2^{IV} - \beta_2\right)$. Now the problem of diminishing variance is eliminated. Moreover, it has a limiting distribution with zero mean and stable variance, i.e.

$$\operatorname{var}\left(\sqrt{n}\left(b_2^{IV} - \beta_2\right)\right) = \frac{\sigma_u^2}{MSD(X)} \times \frac{1}{r^2_{X,Z}}$$

$$E\left(\sqrt{n}\left(b_2^{IV} - \beta_2\right)\right) = \sqrt{n}\,E\left(b_2^{IV} - \beta_2\right) \to \sqrt{n}\left(\beta_2 - \beta_2\right) \to 0 \;\; as \;\; n \to \infty$$

It can be shown that a central limit theorem (CLT) can be applied to demonstrate that $\sqrt{n}\left(b_2^{IV} - \beta_2\right)$ has the limiting normal distribution:

$$\sqrt{n}\left(b_2^{IV} - \beta_2\right) \xrightarrow{d} N\left(0, \frac{\sigma_u^2}{\sigma_X^2} \times \frac{1}{r^2_{XZ}}\right)$$

$$\left(b_2^{IV} - \beta_2\right) \sim N\left(0, \frac{\sigma_u^2}{n\,MSD(X)} \times \frac{1}{r^2_{XZ}}\right)$$

Hence, as an approximation, for sufficiently large samples, b_2^{IV} is distributed as follows:

$$b_2^{IV} \sim N\left(\beta_2, \frac{\sigma_u^2}{n\,MSD(X)} \times \frac{1}{r^2_{X,Z}}\right)$$

This result can be used for performing usual tests. However, from the mathematical point of view such concepts as "sufficiently large samples" and "an approximation" are not well defined.

That is why the analysis is made by means of a Monte Carlo experiment. Suppose, we set up the following model where Z, V, and u are drawn independently from a normal distribution with mean zero and unit variance:

$$Y = \beta_1 + \beta_2 X + u$$
$$X = \lambda_1 Z + \lambda_2 V + u$$

We will treat and V as variables, u as the disturbance term in the model. λ_1 and λ_2 are constants. Setting $\beta_1 = 10$, $\beta_2 = 5$, $\lambda_1 = 0.5$, $\lambda_2 = 2$, the model becomes:

$$Y = 10 + 5X + u$$
$$X = 0.5Z + 2.0V + u$$

By this construction, X is not distributed independently of u, i.e.

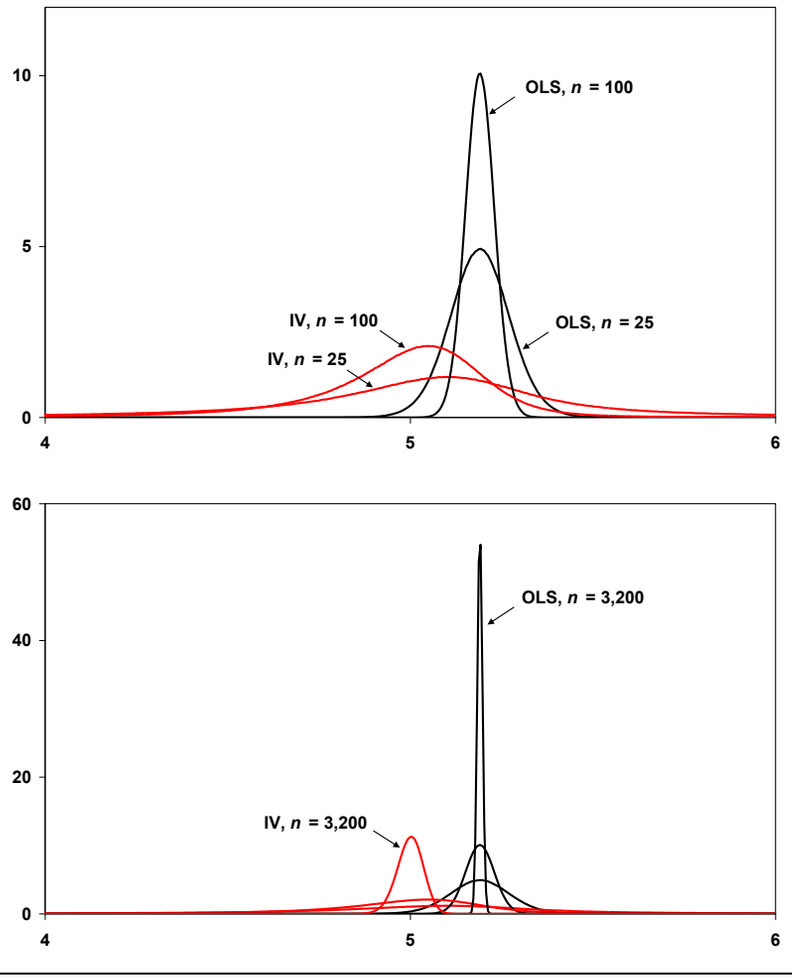
$$\text{cov}(X,u) = \text{cov}(0.5Z + 2.0V + u, u) = \text{cov}(u,u) = \text{var}(u) = 1$$

OLS will give inconsistent estimates. Let's calculate the large sample bias:

$$plim\left(b_2^{OLS}\right) = plim\left(\beta_2 + \frac{Cov(X,u)}{Var(X)}\right) = \beta_2 + \frac{cov(X,u)}{\text{var}(X)} =$$
$$= \left| \begin{matrix} \text{var}(X) = \text{var}(0.5Z + 2.0V + u) = \\ 0.25\,\text{var}(Z) + 4\,\text{var}(V) + \text{var}(u) \end{matrix} \right| = 5 + \frac{1}{0.25 + 4 + 1} \approx 5.19$$

At the same time, Z can serve as an instrument, correlated with X, but independent of u. Hence, $plim\left(b_2^{IV}\right) = 5$.

The diagrams below show the distributions of the OLS and IV estimators for different sample sizes ($n = 25$, $n = 100$, $n = 3200$) for 10 million samples:



$$plim\left(b_2^{OLS}\right) \approx 5.19$$
$$plim\left(b_2^{IV}\right) = 5$$

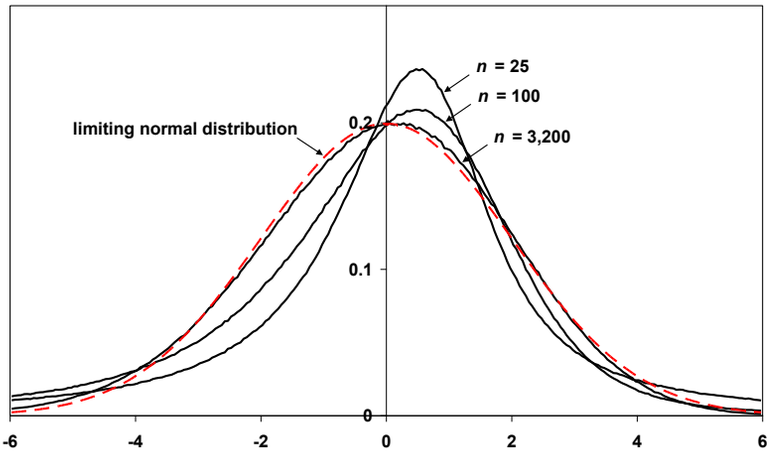
It is evident that IV estimators have greater variances, while OLS are biased. Note that for a small sample size ($n = 25$) one might prefer the OLS estimator according to some criterion such as the mean square error:

$$MSE(b_2) = \left(bias(b_2)\right)^2 + \text{var}(b_2)$$

For $n = 100$ IV estimator looks better.

For a large sample size ($n = 3200$), both estimators are tending to the predicted limits. The IV estimator is definitely better here.

Let's consider the distribution of $\sqrt{n}\left(b_2^{IV} - \beta_2\right)$ for different sample sizes. The dashed red line shows the limiting normal distribution predicted by CLT. As can be seen from the diagram below, for $n = 3200$ the distribution is very close to the limiting normal case, while for small sample sizes ($n = 25$ and $n = 100$) it is definitely non-normal. In fact, it has fat tails that increases the probability of suffering a Type I error. This distortion for small sample sizes is partly explained by low correlation between *X* and *Z* that equals 0.22 for this case. In other words, the chosen instrument is not strong. However, it is often difficult to find any credible instrument at all.



Durbin-Wu-Hausman specification test

To test the hypothesis if endogeneity the Hausman test could be used. Original test is based on the comparison of coefficients of the instrumental variable regression with the conventional OLS regression using chi-squared test with degrees of freedom equal to the number of parameters compared. Under the null hypothesis that there is no simultaneous equation bias, both OLS and IV will be consistent estimators, but OLS will be more efficient. Under the alternative hypothesis, OLS will be inconsistent, and this could be seen from the value of the test (significantly different coefficients of two regressions). The test enables to discriminate between these two possibilities.

H_0: difference in coefficients is not systematic (B.7 is valid <=> no endogeneity)

Estimator b^{IV} is consistent under H_0 and H_1
Estimator B^{OLS} is inconsistent under H_1 and efficient under H_0

However, if the test statistic is not significant, this does not necessarily mean that the null hypothesis is true. It could be that it is false, but the instruments used in IV are so weak that the differences between the IV and OLS estimates are not significant.

There is a different approach to the test of endogeneity, based on the **Davidson-McKinnon**:

1) Estimate the initial model: $Y = \beta_1 + \beta_2 X + u$;
2) Estimate the regression of instrumented variable on the instrument(s), save the residuals: $X = \alpha_1 + \alpha_2 Z + v$ where Z is a possible instrument. Let *resid* be a variable that consists of residuals in this estimated regression;
3) Add the residuals as the additional regressor in the initial model: $Y = \beta_1 + \beta_2 X + \beta_3 resid + u$
4) Check the significance of the slope coefficient β_3. If this new variable is insignificant – then the difference in coefficients is not systematic (the OLS estimates are consistent) => use the initial model. If it is significant – then the difference is systematic, use IV.

This is an asymptotic test, and the t-statistic should be compared with the critical values from the standard normal.

Lecture 13

Simultaneous Equations

Until now we have covered 2 possible sources of endogeneity: omitted variables and measurement errors. This lecture will analyze one more source of endogeneity when equations we want to estimate are closely related. In fact, many economic models include several equations which are described to hold simultaneously. Consequently, both direct and inverse relationships exist between the same variables. These models involve some circularity in equations as both dependent variables are used as explanatory variables in other equations. If each equation of such a model is estimated separately, disregarding the interdependence between the equations, the fourth Gauss-Markov condition (assumption B.7) is violated which leads to biased and inconsistent estimates (*Simultaneous Equations Bias*). Let's illustrate this with an example:

Consider a *macroeconomic equilibrium model* consisting of a consumption function and a national income identity:

$$\begin{aligned} C &= \alpha + \beta Y + u \quad (1) \\ Y &= C + I \quad (2) \end{aligned}$$

C is aggregate consumption, Y is aggregate income (endogenous variables as their values are determined interactively within the model), I is aggregate investment (exogenously determined outside the model), and u is the disturbance term. Suppose, we want to estimate the consumption function as a separate equation, and consider the properties of the OLS estimators.

The system of equations allows us to express the endogenous variables (C and Y) in terms of the exogenous ones (I) and the disturbance term:

$$\begin{aligned} Y &= \frac{\alpha}{1-\beta} + \frac{I}{1-\beta} + \frac{u}{1-\beta} \quad (3) \\ C &= \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} \cdot I + \frac{u}{1-\beta} \quad (4) \end{aligned}$$

System (1)-(2) representing the economic relationships among the variables is called the **Structural Form** of the model. The equations (3)-(4) expressing the endogenous variables in terms of the exogenous variable(s) and the disturbance terms are described as the **Reduced Form** equations.

Suppose that we estimate the first equation in the structural form of the model using OLS. We will get the following results:

$$b^{OLS} = \frac{Cov(Y,C)}{Var(Y)} = \beta + \frac{Cov(Y,u)}{Var(Y)}$$

Since both the numerator and the denominator are the functions of the disturbance term, we cannot compute the expected value to analyze unbiasedness. Instead, we can look at large sample properties

$$plim\left(b^{OLS}\right) = plim\left(\beta + \frac{Cov(Y,u)}{Var(Y)}\right) = \beta + plim\left(\frac{Cov(Y,u)}{Var(Y)}\right) = \beta + \frac{plim\left(Cov(Y,u)\right)}{plim\left(Var(Y)\right)} = \beta + \frac{\sigma_{Y,u}}{\sigma_Y^2}$$

The population covariance and variance can be obtained by substituting the result for Y from the reduced form equation, i.e.

$$\sigma_{Y,u} = \operatorname{cov}\left[\left(\frac{\alpha}{1-\beta} + \frac{1}{1-\beta}\cdot I + \frac{u}{1-\beta}\right), u\right] = \operatorname{cov}\left(\frac{u}{1-\beta}, u\right) = \frac{1}{1-\beta}\sigma_u^2$$

$$\sigma_Y^2 = \operatorname{var}\left(\frac{\alpha}{1-\beta} + \frac{1}{1-\beta}\cdot I + \frac{u}{1-\beta}\right) = \operatorname{var}\left(\frac{1}{1-\beta}\cdot I + \frac{u}{1-\beta}\right) = \frac{1}{(1-\beta)^2}(\sigma_I^2 + \sigma_u^2)$$

Hence, $plim\left(b^{OLS}\right) = \beta + \dfrac{\frac{1}{1-\beta}\sigma_u^2}{\frac{1}{(1-\beta)^2}(\sigma_I^2 + \sigma_u^2)} = \beta + (1-\beta)\cdot\dfrac{\sigma_u^2}{\sigma_I^2 + \sigma_u^2}$, i.e. *in large samples the estimator is biased*. The value and direction of large sample bias depends on the structure of the model. In our case, it is positive and it is calculated as:

$$large\,sample\,bias = (1-\beta)\cdot\frac{\sigma_u^2}{\sigma_I^2 + \sigma_u^2}$$

The example shows that applying OLS to estimating an equation which is a part of a simultaneous equations system yields, generally speaking, biased and inconsistent estimators and makes statistical tests invalid. Therefore, special methods are developed to estimate the parameters of such equations.

Identification

Identification is a characteristic of a particular equation in an entire system. Generally, an equation in Simultaneous Equations Model is exactly identified if the number of potential instruments is equal to the number of endogenous variables to be instrumented.

An equation in Simultaneous Equations Model is underidentified if no set of consistent estimators can be provided (generally, the number of potential instruments is less than the number of endogenous variables to be instrumented).

An equation in Simultaneous Equations Model is overidentified if (generally) the number of potential instruments is greater than the number of endogenous variables to be instrumented (different sets of consistent estimators are provided).

The order condition for identification:

It follows directly from this observation:

An equation is identified, if the number of exogenous variables missing from it is greater or equal to the number of endogenous variables on its right side.

Consider a model with G equations and G endogenous variables. Suppose that j endogenous variables are missing from the equation. Then $(G-j-1)$ endogenous variables are available on the right side, and at least $(G-j-1)$ instruments are needed. So the minimum number of variables missing from the equation is $j+(G-j-1) = G-1$.

The order condition implies that an equation in this model is identified, if it does not include at least $(G-1)$ variables of the model (endogenous or exogenous). If exactly $G-1$ variables are missing, the equation is likely to be exactly identified. If more than $G-1$ variables are missing, it is likely to be overidentified, and TSLS to be used.

Example:

Consider IS-LM model described by the following equations:

1) $Y_t = C_t + I_t + G_t + X_t$ income identity

2) $C_t = \alpha + \beta\cdot Y_t (+u_t)$ consumption function

3) $I_t = \delta + \nu \cdot Y_t + \varepsilon \cdot R_t (+v_t)$ investment spending

4) $X_t = \rho + \sigma \cdot Y_t + \tau \cdot ER_{t-1} (+w_t)$ net exports function

5) $M_t = \lambda + \mu \cdot Y_t + \theta \cdot R_t (+z_t)$ money market equilibrium

Endogenous variables: Y_t, C_t, I_t, X_t, R_t

Exogenous variables: G_t, ER_{t-1}, M_t

In this model all the equations (except (1) which is an identity) are overidentified, as the number of exogenous variables missing from each equations exceeds the number of endogenous variables appearing on its right side. In order to make it evident that the equation (5) is overidentified, it can be transformed so that there is an endogenous variable on the left side: $R_t = \frac{M_t}{\theta} - \frac{\lambda}{\theta} - \frac{\mu \cdot Y}{\theta} (+z_t^{'})$

Order condition: $G - 1 = 4$. In each of (2)-(5) equations the number of missing variables exceeds 4 => these equations are overidentified.

Instrumental variable (IV) estimation of simultaneous equations

The idea of this method is as follows: if there exist exogenous variables which do not appear in the equation estimated, they can be used as instruments for the endogenous variables appearing in it. For this purpose, the number of such exogenous variables should be no less than the number of the endogenous variables to be replaced.

Two-stage least squares method (TSLS)

TSLS is a technique that is widely used to deal with issues of endogeneity. It can be applied to any identified systems of equations to obtain consistent estimates of parameters (not only to exactly identified as ILS but also to overidentified equations).

Consider again the market equilibrium model:

$$\begin{cases} y_t^d = \alpha + \beta \cdot p_t + \gamma \cdot x_t + \rho \cdot t + u_t^d, \\ y_t^s = \delta + \varepsilon \cdot p_t + u_t^s \end{cases}$$

There are several potential instrumental variables for p_t in the supply equation: variables x_t and t, as well as their linear combinations $z_t = h_0 + h_1 \cdot x_t + h_2 \cdot t$, are suitable for this purpose. Thus, the supply function is overidentified.

The idea of the TSLS method is to select the linear combination which is most strongly correlated with p_t, i.e. we shall use $\hat{p}_t = a' + b' \cdot x_t + c' \cdot t$, such that $r^2_{\hat{p},p} = R^2 = 1 - \frac{RSS}{TSS_p} \to \max$.

While estimating a linear regression, we have minimized the residual sum of squares and maximized the coefficient of determination. Consequently, the squared correlation coefficient for the actual and estimated values of p is also maximized. Having calculated the fitted values of p on the basis of the regression equation, we thus get an instrumental variable, which is correlated with p as strongly as possible but, being exogenous, is not dependent on the disturbance term.

Generally, to estimate a system of equations using the two-stage least squares method, the following should be done for every identifiable equation:

Stage 1

1) A list of potential instrumental variables (exogenous variables not appearing in the equation estimated) is formed.

2) Regressions of the endogenous explanatory variables on the potential instruments are estimated.

3) The fitted values of the endogenous explanatory variables, which will later be used as instruments, are calculated on the basis of these regressions.

Stage 2

4) The instruments thus derived are substituted for the corresponding endogenous variables on the right side of the estimated equation.

5) A regression of the dependent variable on the instruments is estimated and consistent estimates of the coefficients are thus obtained.

Estimation in EViews (*for EAEF40)*

$$HGC = \beta_1 + \beta_2 ASVABC + \beta_3 HGCM + \beta_4 HGCF + u \qquad (1)$$

$$ASVABC = \alpha_1 + \alpha_2 HGC + v \qquad (2)$$

These equations are simultaneous structural equations. They represent the model of the determinants of educational attainment (HGC) and general indicator of mathematical and language abilities (ASVABC). It is supposed that HGC is related to ASVABC and educational attainments of both mother and father. At the same time, ASVABC is related to HGC. By this specification, this simultaneous equation model involves a certain amount of circularity: ASVABC determines HGC in the first equation, and in turn HGC helps to determine ASVABC in the second model. There are endogenous and exogenous variables which are needed to cut through the circularity. Endogenous variables are HGC and ASVABC whose values are determined interactively within the model. Exogenous variables are HGCM, HGCF, whose values are determined outside the model. Thus, they can be considered as instruments for HGC.

It can be seen that (1) is underidentified, while (2) is overidentified. In fact, in equation (1) there is one right-hand side endogenous variable. But there is no available instruments for ASVABC in equation (there are 2 exogenous variables available (HGCM, HGCF) but each of them is in the first equation on its own right – if we would use them as instruments as well, then it would lead to perfect multicollinearity. At the same time, in equation (2) there is one right-hand side endogenous variable and there are two available instruments for HGC from the first equation.

Let's apply TSLS to (2). In EViews it can be done directly by the following command:

tsls asvabc c hgc @ c hgcm hgcf

TSLS

Dependent Variable: ASVABC
Method: Two-Stage Least Squares
Sample: 1 570
Included observations: 570
Instrument specification: C HGCM HGCF

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.217455	4.801296	0.461845	0.6444
HGC	3.529327	0.351016	10.05460	0.0000

OLS

Dependent Variable: ASVABC
Method: Least Squares
Sample: 1 570
Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	23.03563	1.847766	12.46675	0.0000
HGC	2.003304	0.133468	15.00960	0.0000

Note that the true effect of HGC on ASVABC is underestimated under OLS (the sign of bias is negative)

Hausman test (Davidson-McKinnon modification)

It allows to test whether instrumental variables should be used:

H_0: The difference in coefficients of the instrumental variable estimation and OLS estimation is not systematic – OLS estimates are consistent;
H_1: OLS estimates are inconsistent

1) Estimate the regression of HGC on the instruments HGCM and HGCF, saving the residuals (RESID1)

Dependent Variable: HGC
Method: Least Squares
Sample: 1 570
Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.230042	0.432887	21.32206	0.0000
HGCM	0.190291	0.043133	4.411751	0.0000
HGCF	0.182065	0.031782	5.728623	0.0000

2) Add residuals as the additional regressor in (2)

Dependent Variable: ASVABC
Method: Least Squares
Sample: 1 570
Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.217455	4.223490	0.525029	0.5998
HGC	3.529327	0.308774	11.43015	0.0000
RESID1	-1.856144	0.340538	-5.450625	0.0000

The residual coefficient is significant at any reasonable significance level. It means that we reject the null hypothesis that the difference in coefficients is not systematic => there is evidence to use instrumental variables.

Lecture 14

Linear Probability Model. Binary Choice Models.

It is usually a point of interest to investigate the factors behind the decision-making of individuals or enterprises in applied econometrics. For example, the following questions can be asked:

- What characteristics affect the likelihood that an individual obtains a higher degree?
- Why do some people buy houses while others rent?
- What determines labour force participation?

In order to answer these questions, qualitative response models are proposed. The common feature of such models is that the economic outcome is a discrete choice among a set of alternatives, rather than a continuous measure of some activity the modeling procedures of which were discussed in previous lectures.

This lecture will analyze binary choice models that describe some event either happens or not. The dependent variable can take only 2 values: it is typically equal to one for all observations in the data for which the event has happened (success) and zero for the remaining observations (failure). Whenever the variable that we want to model is binary, it is natural to think in terms of probabilities. For example, the first question can be reformulated as following:

What is the probability that an individual with such and such characteristics obtains a higher degree?

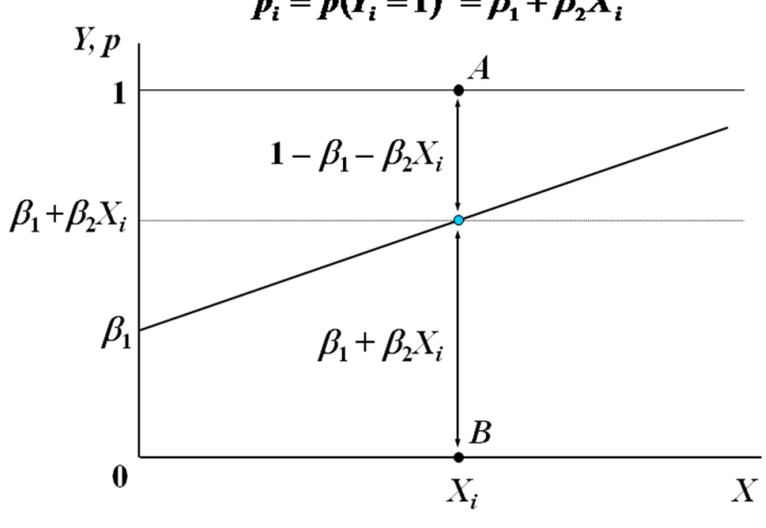
In general, conventional regression methods (OLS) cannot be used for estimation.

Linear probability model (LMP)

The simplest binary choice model is the linear probability model where the probability of the event occurring, p, is assumed to be a linear function of a set of explanatory variables. If there is only 1 regressor, the expression becomes:

$$p_i = p(Y_i = 1) = \beta_1 + \beta_2 X_i$$

According to this model, the non-stochastic component of the relationship between Y and X in observation i, $E(Y_i) = p_i \cdot 1 + (1 - p_i) \cdot 0 = p_i = \beta_1 + \beta_2 \cdot X_i$, is the same as the probability function, i.e. the expected values of Y lay on a straight line $Y = \beta_1 + \beta_2 \cdot X$. However, the true values of Y will never lay on this straight line, falling instead either on the X-axis or on the horizontal line $Y = p = 1$.



$$Y_i = E(Y_i) + u_i$$

$$Y_i = p_i + u_i = \underbrace{\beta_1 + \beta_2 X_i}_{\text{non-stochastic}} + \underbrace{u_i}_{\text{random}}$$

$$u_i = Y_i - \beta_1 - \beta_2 X_i$$

The two possible values of the disturbance term, which give rise to the observations A and B, are illustrated in the diagram. In fact,

A: $Y_i = 1 \Rightarrow u_i = 1 - \beta_1 - \beta_2 X_i$

B: $Y_i = 0 \Rightarrow u_i = -\beta_1 - \beta_2 X_i$

It is evident that the linear probability model is straightforward to estimate. However, it has a number of serious shortcomings:

1) Possible predicted values can be outside the range of possible probability $[0;1]$. In other words, the model can yield estimates of p, which are above 1 or below 0;
2) The model assumes a constant marginal effect of each of explanatory variables. In fact, in $p_i = p(Y_i = 1) = \beta_1 + \beta_2 X_i$ the slope coefficient, β_2, measures the change in the probability of "success", resulting from a change in the variable X, holding other factors fixed, i.e. $\Delta \Pr(Y = 1 | X) = \beta_2 \cdot \Delta X$. This may seem reasonable for average values of observations in a sample, $\overline{X}$, (near the center of the distribution) but for extreme cases (relatively large/small X) this partial effect on the probability of "success" is different (probably smaller). Hence, it is more realistic to take into account different marginal effects;
3) As it was shown, the distribution of the disturbance term is not continuous (only 2 values are possible) and thus is not normally distributed. Consequently, standard errors and statistical tests are invalid;
4) The disturbance term is heteroscedastic. In fact,

$$E(u_i) = P(u_i = 1 - \beta_1 - \beta_2 X_i) \cdot (1 - \beta_1 - \beta_2 X_i) + P(u_i = -\beta_1 - \beta_2 X_i) \cdot (-\beta_1 - \beta_2 X_i) =$$
$$= p(Y_i = 1) \cdot (1 - \beta_1 - \beta_2 X_i) + p(Y_i = 0) \cdot (-\beta_1 - \beta_2 X_i) = (\beta_1 + \beta_2 X_i) \cdot (1 - \beta_1 - \beta_2 X_i) +$$
$$+ (1 - \beta_1 - \beta_2 X_i) \cdot (-\beta_1 - \beta_2 X_i) = 0$$

$$\text{var}(u_i) = E(u_i^2) - [E(u_i)]^2 = E(u_i^2) - 0^2 = E(u_i^2) = p(Y_i = 1) \cdot (1 - \beta_1 - \beta_2 X_i)^2 +$$
$$+ p(Y_i = 0) \cdot (-\beta_1 - \beta_2 X_i)^2 = (\beta_1 + \beta_2 X_i) \cdot (1 - \beta_1 - \beta_2 X_i)^2 + (1 - \beta_1 - \beta_2 X_i) \cdot (-\beta_1 - \beta_2 X_i)^2 =$$
$$= (\beta_1 + \beta_2 X_i)(1 - \beta_1 - \beta_2 X_i)(\beta_1 + \beta_2 X_i + 1 - \beta_1 - \beta_2 X_i) = (\beta_1 + \beta_2 X_i)(1 - \beta_1 - \beta_2 X_i)$$

Therefore, the population variance of the disturbance term is not constant and depends on X_i, i.e.

$$\sigma_{u_i}^2 = (\beta_1 + \beta_2 \cdot X_i)(1 - \beta_1 - \beta_2 \cdot X_i)$$

Thus, the key Gauss-Markov conditions are not satisfied, which means that the linear probability model cannot be considered satisfactory for studying the probability of observing a certain quality.

At the same time, the linear probability model can be useful as a first step in the analysis of binary choices. It is much easier to fit and certain econometric problems are easier to address within its framework. For this reason it is often recommended to be used for initial, exploratory work.

Example: Let's examine how the total results of the Unified State Examination (in Russian: EGE) for 3 subjects (Mathematics, Russian and English) affect the probability of admission to the University of London (UoL) for ICEF students in 2012.

Consider a binary variable taking the value of 1, if the student has admitted to the University of London, and the value of 0, otherwise. We need to find out how likely a person with a given EGE sum for 3 subjects is to enter the University of London. We estimate the linear probability model of the following form: $UL\_PASS = \beta_1 + \beta_2 \cdot EGE\_SUM$

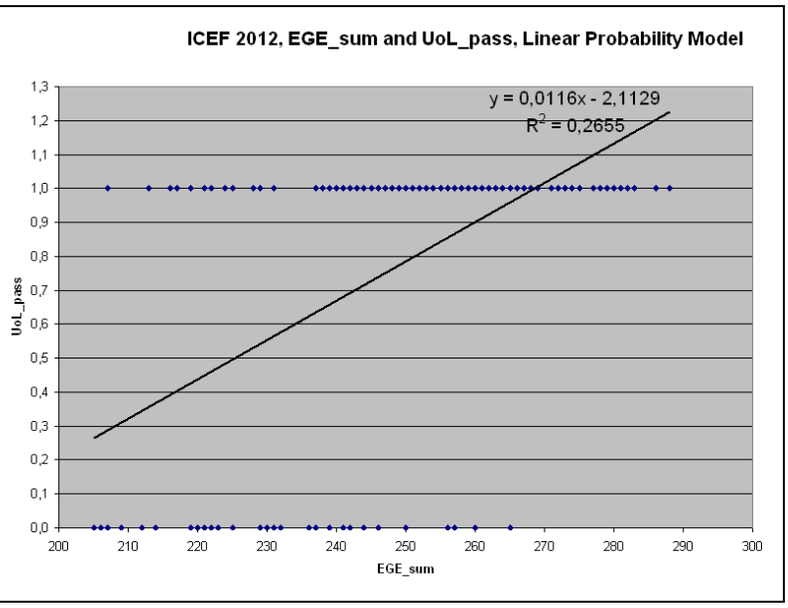
Estimated regression (LMP): $UL\_PASS = -2.1129 + 0.0116 \cdot EGE\_SUM$

$s.e.$ (0.364) (0.00147)

Dependent Variable: UL_PASS
Method: Least Squares
Sample: 1 212
Included observations: 175

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.112884	0.363689	-5.809590	0.0000
EGE_SUM	0.011595	0.001466	7.907131	0.0000

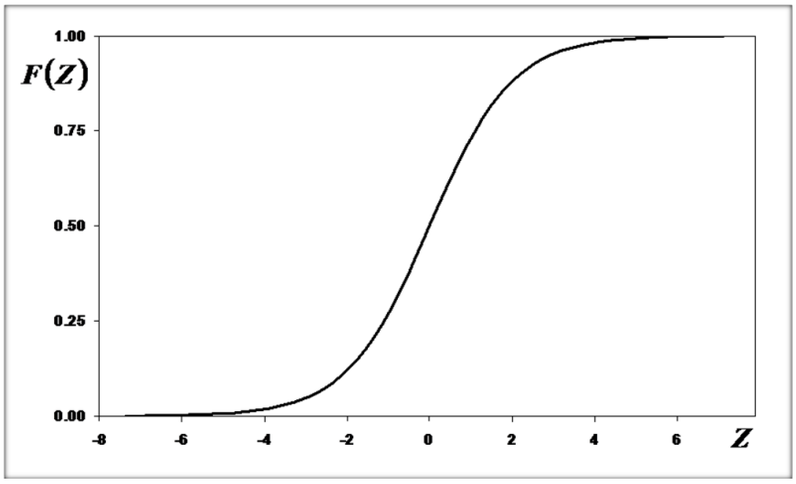
It is evident that the model gives senseless probability predictions outside the range $[0;1]$, and hence is incorrectly specified. In fact, for students with $EGE\_SUM>268$, the resulting estimating probability of admission to the University of London is greater than 1.



Models for Binary Response:

The usual way to deal with problems (1) and (2) stated in the previous section of the linear probability model is to use a sigmoid (*S*-shaped) probability function of Z, $F(Z)$, where Z is some linear function of the explanatory variables. The shape implies that for low values of Z the probability of the event occurring is low and insensitive to variations in Z. Likewise, for high values of Z this probability is high and also insensitive to variations in Z.

Sigmoid (S-shaped) probability function of Z



Formally, let's $Z_i = \beta_1 + \beta_2 \cdot X_i$. The probability model can be written as follows:

$$P(Y_i = 1 | X_i, \beta) = F(Z_i)$$
$$P(Y_i = 0 | X_i, \beta) = 1 - F(Z_i)$$

Various non-linear functions for $F(Z)$ have been proposed in the literature. By far the most common ones are the logistic distribution, yielding the Logit model, and the standard normal distribution, yielding the Probit model.

Logit analysis

The cumulative distribution function (cdf) for a logistic variable Z is given by:

$$p = F(Z) = \frac{1}{1+e^{-Z}} \quad \text{or equivalently} \quad p = F(Z) = \frac{e^{Z}}{1+e^{Z}}$$

Let's show that it has limiting bounds of 1 and 0, as Z tends to plus and minus infinity, respectively:

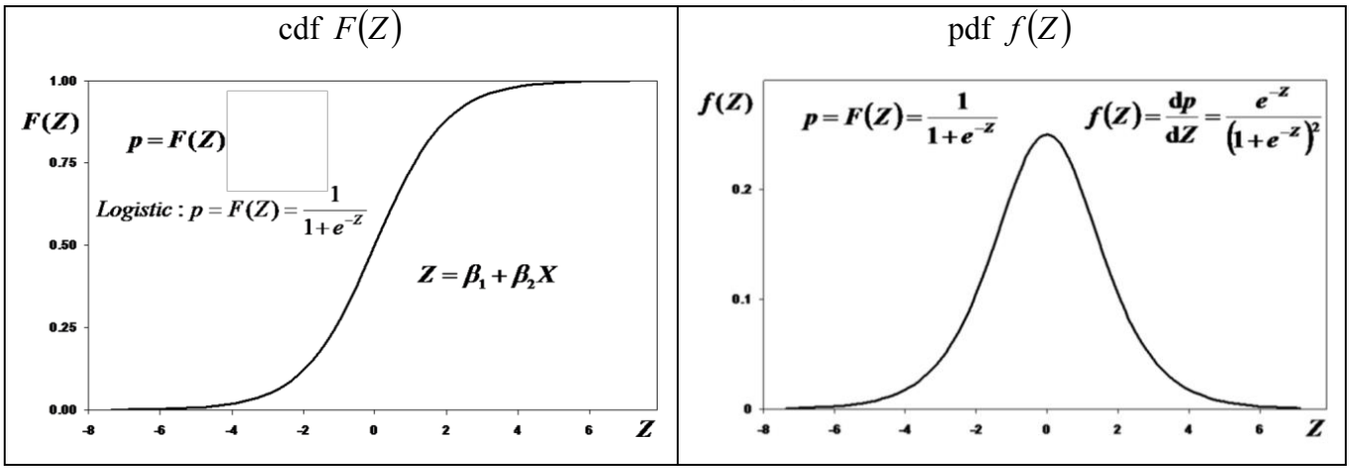
$$Z \to +\infty \Rightarrow e^{-z} \to 0 \Rightarrow F(Z) \to 1, \text{ but never exceeds 1;}$$

$$Z \to -\infty \Rightarrow e^{-z} \to \infty \Rightarrow F(Z) \to 0, \text{ but never becomes negative.}$$

Obviously, $\dfrac{1}{1+e^{-Z}} > 0 \quad for\ any\ Z$ $\int_{-\infty}^{+\infty} f(Z)dZ = F(+\infty) - F(-\infty) = 1 - 0 = 1$. Hence, all the requirements of cdf for $F(Z)$ are satisfied.

By definition, $f(Z) = \dfrac{dF(Z)}{dZ} = \left(\dfrac{1}{1+e^{-Z}}\right)' = \dfrac{(1+e^{-Z})\times 0 - 1\times(-e^{-Z})}{(1+e^{-Z})^2} = \dfrac{e^{-Z}}{(1+e^{-Z})^2}$. Hence, $\boxed{f(Z) = \dfrac{e^{-Z}}{(1+e^{-Z})^2}}$ is the probability density function (pdf) associated with $F(Z)$. It shows the sensitivity of $F(Z)$ to changes in Z. This is used to compute marginal effects.

Logit model: cdf and pdf



In the general case (with *k* explanatory variables), the Logit model takes the following form:

$$p = F(Z) = \frac{1}{1+e^{-Z}}, \text{ where } Z = \beta_1 + \beta_2 \cdot X_2 + \ldots + \beta_k \cdot X_k.$$

Interpretation: It becomes less straightforward to interpret the estimated coefficients compared to the linear probability case. In fact, marginal effects that can be defined as the effects on the response probability resulting from a change in one of the explanatory variables depend on the value of Z, which in turn depends on the values of the explanatory variables, i.e.

$$\boxed{\frac{\partial p}{\partial X_i} = \frac{dp}{dZ}\frac{\partial Z}{\partial X_i} = f(Z)\cdot\beta_i = \frac{e^{-Z}}{(1+e^{-Z})^2}\beta_i}$$

A common procedure is to evaluate the marginal effect for the sample means of the explanatory variables. Note that since $f(Z)$ is non-negative function, the marginal effect of X_i will always have the same sign as β_i. The marginal function reaches its maximum value at $Z = 0$, whereas at very small and very large values the marginal effect is small. This property of the function makes the models suitable for describing most real situations.

Continuing example ICEF 2012, EGE_SUM and UL_PASS: **Logit Model**.

Let's estimate the Logit model of the probability to the University of London admission on the total EGE sum for ICEF students in 2012 (see the next lecture for the estimation method)

.

Estimated regression (LOGIT):
$$Z = -17.322 + 0.0763 \cdot EGE\_SUM$$
$$s.e \quad (3.087) \qquad (0.013)$$

All coefficients are highly significant. At the same time, they do not have any direct intuitive interpretation. Note that for the testing procedure we should use *t*-statistics that is valid only for large samples in this case. Therefore, we usually use the normal distribution as the reference distribution for testing the significance of obtained coefficients.

The probability to the University of London admission is calculated as:

$$p(UL\_PASS = 1) = F(Z) = \frac{1}{1+e^{-Z}} = \frac{1}{1+e^{17.322-0.0763*EGE\_SUM}}$$

Marginal effect, evaluated at the sample mean (average student):

$Mean\ EGE\_SUM = \bar{X} = 247.3$. $Z = \beta_1 + \beta_2 \bar{X} = -17.32 + 0.076 \times 247.3 = 1.475$.

$p(UL\_PASS = 1) = F(Z) = \frac{1}{1+e^{-Z}} = \frac{1}{1+e^{-1.475}} = 0.814$. It means that there is 81.4% probability that a student with average EGE_SUM will be admitted to the University of London.

$$f(Z) = \frac{dp}{dZ} = \frac{e^{-Z}}{(1+e^{-Z})^2} = \frac{0.229}{(1+0.229)^2} = 0.152$$
$$\frac{\partial p}{\partial X} = \frac{dp}{dZ}\frac{\partial Z}{\partial X} = f(Z)\beta_2 = 0.152 \times 0.076 = 0.0115$$

This implies that a one point increase in EGE_SUM would increase the probability of admission to the University of London by 1.15 percent points. It is about the same as the LPM model slope $= 0.0116$.

The biggest marginal effect is evaluated at $Z = 0$ that corresponds to $EGE\_SUM \approx 228$.

$\frac{\partial p}{\partial X} = \frac{dp}{dZ}\frac{\partial Z}{\partial X} = f(Z)\beta_2 = 0.25 \times 0.076 = 0.019$. One point increase in EGE_SUM would increase the probability of admission to the University of London by 1.9 percent points.

Probit analysis:

The cdf of the sigmoid function $F(Z)$ is given by the cdf of the standardized normal distribution. Thus, the marginal function $f(Z)$ is defined as the pfd of the standardized normal distribution, i.e.:

$$\boxed{f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}}$$

The principle and logic of the Probit model are similar to ones of the Logit model.

Marginal effects are calculated as following:

$$\boxed{\frac{\partial p}{\partial X_i} = \frac{dp}{dZ}\frac{\partial Z}{\partial X_i} = f(Z)\beta_i = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}\right)\beta_i}$$

Continuing the previous example, the Probit estimation results in the following equation:

Estimated regression (PROBIT): $Z = -10.066 + 0.04434 \cdot EGE\_SUM$

$s.e. \quad (1.705) \qquad (0.0071)$

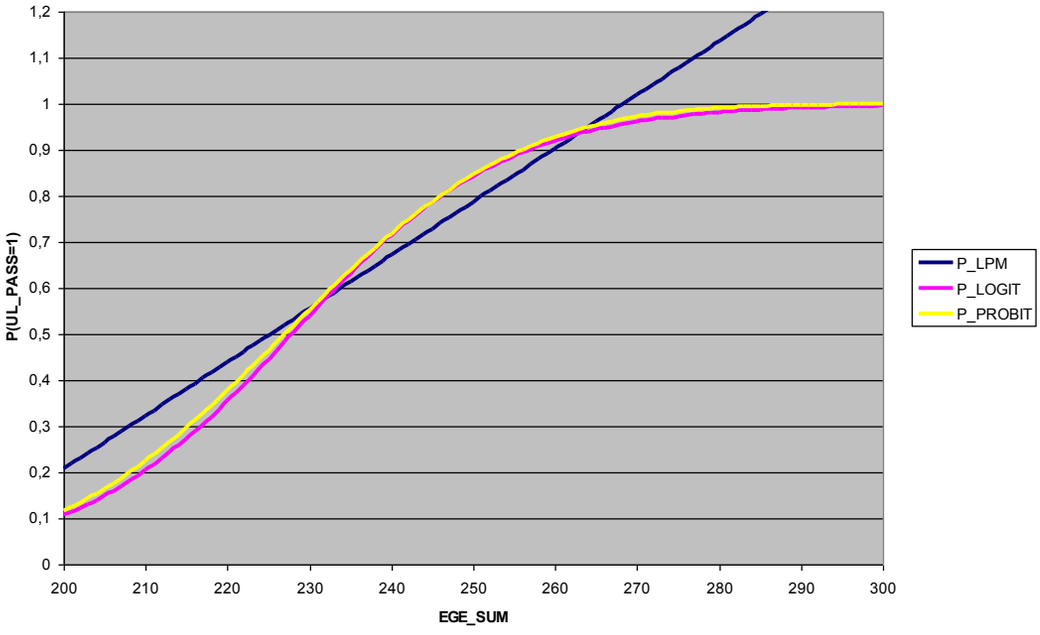
$p(UL\_PASS = 1) = F(Z)$. For example, for the sample mean value of $EGE\_SUM = 247.3$, $Z = 0.899$. Hence, $F(Z) = F(0.899) = 0.816$ (remember that for the Logit model it equals 0.814). These 2 results are quite similar.

The biggest marginal effect is evaluated at $Z = 0$, when $EGE\_SUM = \frac{10.066}{0.04434} \approx 227$. It equals $\frac{\partial p}{\partial X_i} = f(Z)\beta_i = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}0^2} \right)\beta_i = \frac{0.04434}{\sqrt{2\pi}} = 0.0177$

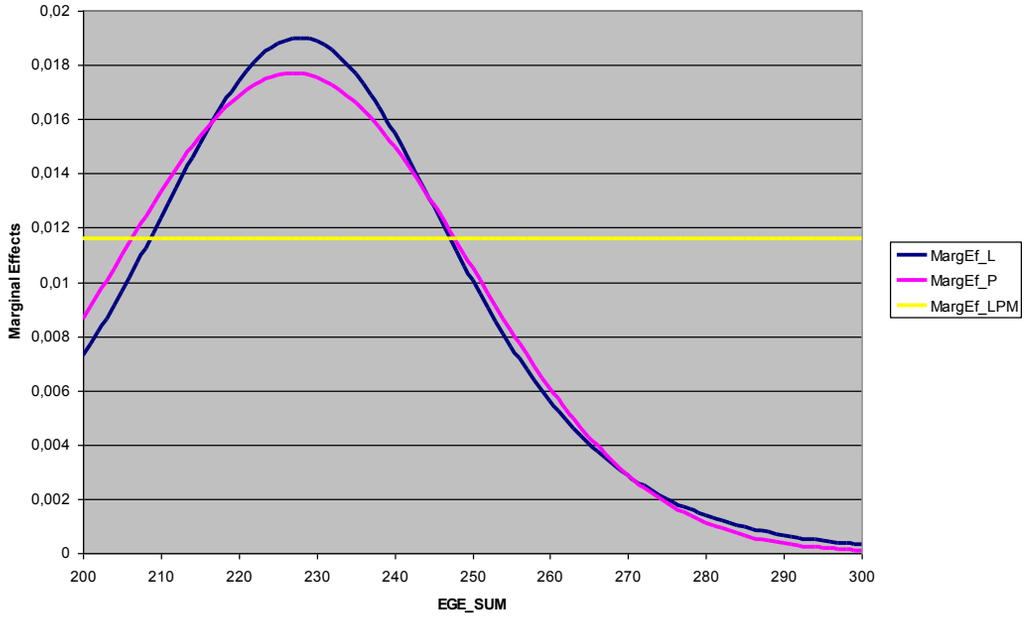
Comparison of results: LMP, Logit, Probit:

It can be seen from the diagrams below that the estimates of the probability of admission to the University of London and the corresponding marginal effects are similar for the Logit and Probit.

Probability to pass to the UoL: LPM, Logit and Probit



Marginal Effects: LPM, Logit, Probit



For most correctly specified models the results for the Logit and Probit are similar, suggesting that it does not much matter which one of the two we choose to use in our analysis. However, their tails are somewhat different, meaning that if the outcomes in the sample are divided between a large majority and a small minority (observations are concentrated in a tail of the distribution), the results can differ. Traditionally, Logit was wider used because the logistic function leads to a more easily computed model. Now, Probit is easy to compute with standard packages, and is used widely.

Lecture 15

Maximum Likelihood Estimation. Limited Dependent Variable Models

Maximum likelihood estimation is one of the most popular estimation methods in econometrics. Together with the least squares method, it is widely used in estimating various models. Moreover, for some of them (for example, binary choice models) it is the principal estimation method.

Let's begin with *an example* of a simple problem, which can be solved with the help of the maximum likelihood estimation.

Suppose, a certain random variable takes the following values: $\begin{matrix} 1 \text{ with probability } p; \\ 0 \text{ with probability } 1-p. \end{matrix}$

We need to estimate the parameter p on the basis of the following 3 observations: 1; 1; 0.

For any given p, the probability of getting such a sample is $p^2 \cdot (1-p)$. Here the maximum likelihood estimation consists in finding the value of p which maximizes this probability:

$$\max_p p^2 \cdot (1-p)$$

$$FOC: \ 2p - 3p^2 = 0 \Rightarrow p^* = \frac{2}{3}$$

$$SOC: \ 2 - 6p\big|_{p^*=\frac{2}{3}} = -2 < 0$$

Thus, $p^* = \frac{2}{3}$ is a point of the global maximum.

The leading idea: given observations, the estimate should provide the most likelihood (most probable) of the observations.

Theory:

Let X be population that is identified with some random variable. $f(x,\theta)$ is its probability density function (pdf), where $\theta \in \Theta \subset R^k$ is some parameter (generally, multidimensional) that must be estimated.

$X_1, X_2, X_3, ..., X_n \equiv sample$, where X_i is a random variable with $f(x,\theta)$ and these random variables are identically and independently distributed. Hence, the joint density is described as follows:

$$f_X(x_1, x_2, ..., x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot ... \cdot f_{X_n}(x_n) = f(x,\theta),$$

where the vector $x = (x_1, x_2, ..., x_n)$ is viewed as arguments of the density function.

Given observed values, $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$, the likelihood of θ is the function:

$$\boxed{L(\theta) = L(\theta, X) = f(X;\theta) = f(X_1, X_2, X_3, ..., X_n | \theta)}$$

It should be stressed that $L(\cdot)$ is considered as a function of θ and X is fixed. In words, $L(\theta)$ is the probability of observing the given data as a function of θ.

Let's consider the following problem: $L(\theta; X) \to \max_{\theta \in \Theta \subset R^k}$

Definition: the solution $\hat{\theta} = \hat{\theta}(X)$ of the problem $L(\theta; X) \to \max_{\theta \in \Theta \subset R^k}$ is called *the Maximum Likelihood Estimator (MLE) of the parameter* θ and is denoted as $\hat{\theta}_{ML}$. It is the value that makes the observed data the "most probable".

If the X_i are *iid*, then the likelihood simplifies to:

$$L(\theta) = L(\theta, X) = f(X_1;\theta)\cdot f(X_2;\theta)\cdot\ldots\cdot f(X_2;\theta) = \prod_{i=1}^{n} L(\theta; X_i)$$

Rather than maximizing this product which can be quite tedious, we often use the fact that the logarithm is an increasing function so it will be equivalent to maximize the log likelihood:

$$l(\theta) = \sum_{i=1}^{n} \log\left(L(\theta; X_i)\right)$$

Properties of the MLE

1) *Consistency:* $\widehat{\theta}_{ML} \to \theta$ as $n \to \infty$;
2) *Asymptotic normality:* $\sqrt{n}\left(\widehat{\theta}_{ML} - \theta\right) \to N(0, \sigma^2)$ as $n \to \infty$;
3) *Asymptotic efficiency:* For any (asymptotically normal) estimator $\widetilde{\theta}$ of the parameter θ the following inequality holds: $\lim_{n\to\infty} \frac{V(\widehat{\theta}_{ML})}{V(\widetilde{\theta})} \le 1$;
4) *Invariance:* Let $\mu = g(\theta)$ where $g(\cdot)$ is a smooth function. Then $\hat{\mu}_{ML} = g\left(\widehat{\theta}_{ML}\right)$.

Example: Estimation of parameters of ***Normal distribution***

Let X be *iid* normally distributed with mean μ and standard deviation σ, its density function is as shown:

$$f(X_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X_i-\mu}{\sigma}\right)^2}$$

The joint probability density for the n observations in the sample is just the product of their individual densities:

$$L(\mu, \sigma \mid X_1,\ldots,X_n) = \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X_1-\mu}{\sigma}\right)^2}\right) \times\ldots\times \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X_n-\mu}{\sigma}\right)^2}\right)$$

$$\log L = \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X_1-\mu}{\sigma}\right)^2}\right) + \ldots + \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X_n-\mu}{\sigma}\right)^2}\right) =$$

$$= n\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{X_1-\mu}{\sigma}\right)^2 - \ldots - \frac{1}{2}\left(\frac{X_n-\mu}{\sigma}\right)^2 = -n\log\sigma + n\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{\sigma^{-2}}{2}\sum\left(X_i-\mu\right)^2$$

FOC:
$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2}\frac{\partial}{\partial \mu}\left(-\frac{1}{2}(X_1-\mu)^2 - \ldots - \frac{1}{2}(X_n-\mu)^2\right) = \frac{1}{\sigma^2}\left[(X_1-\mu) + \ldots + (X_n-\mu)\right] =$$
$$= \frac{1}{\sigma^2}\left(\sum\left(X_i - n\mu\right)\right) = 0 \iff \sum\left(X_i - n\mu\right) = 0$$

Dividing by n, we get that ML of μ is the sample mean, i.e. $\hat{\mu} = \overline{X}$

$$\frac{\partial \log L}{\partial \sigma} = -\frac{n}{\sigma} + \sigma^{-3}\sum(X_i-\mu)^2 = 0$$

Substituting $\overline{X}$ into the equation above and simplifying, we obtain:

$$-n\hat{\sigma}^2+\sum(X_i-\bar{X})^2=0$$

$$\hat{\sigma}^2=\frac{1}{n}\sum(X_i-\bar{X})^2$$

Thus, $\boxed{\hat{\mu}_{ML}=\bar{X}}$

$\boxed{\hat{\sigma}^2_{ML}=\frac{1}{n}\sum(X_i-\bar{X})^2}$

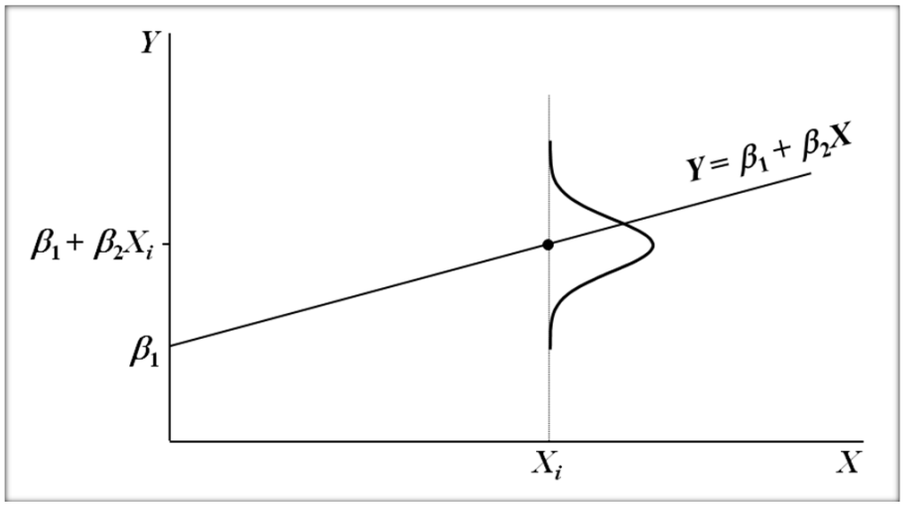
Note that the obtained ML estimator for μ is unbiased, while ML estimator for σ is biased but consistent. Formally,

$$\boxed{E(\hat{\mu}_{ML})=\mu}$$

$$\boxed{bias(\hat{\sigma}^2_{ML})=E(\hat{\sigma}^2_{ML})-\sigma^2=-\frac{\sigma^2}{n}}$$

Maximum likelihood estimation of a simple linear regression

Suppose, the true relationship between Y and X looks like $Y_i=\beta_1+\beta_2\cdot X_i+u_i$. In other words, Y_i is distributed around $\beta_1+\beta_2\cdot X_i$, according to the values of u_i. The density function for the distribution of Y_i conditional on $X=X_i$ is shown below:



Let's assume that $u_i=Y_i-\beta_1-\beta_2X_i$ is normally distributed with zero mean and variance σ^2. Note that it is the ex ante distribution, i.e. potential distribution before the observation is generated. Then, its probability density function takes the following form:

$$\boxed{f(u)=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{1}{2}(\frac{u}{\sigma})^2}}$$

Accordingly, $f(Y_i)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{Y_i-\beta_1-\beta_2X_i}{\sigma}\right)^2}$

Our choice variables are β_1,β_2,σ that we want to estimate on the basis of the n sample values of X and Y. Let's construct *the likelihood function* under the assumption that the values of the disturbance term in different observations are independent:

$$L(\beta_1,\beta_2,\sigma\,|\,Y_1,...,Y_n)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{Y_1-\beta_1-\beta_2X_1}{\sigma}\right)^2}\times...\times\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{Y_n-\beta_1-\beta_2X_n}{\sigma}\right)^2}=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{Y_i-\beta_1-\beta_2X_i}{\sigma}\right)^2}$$

Log likelihood becomes:

$$\log L = l(\beta_1, \beta_2, \sigma \mid Y_1, ..., Y_n) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y_i - \beta_1 - \beta_2 X_i}{\sigma}\right)^2} \right) = n \log\left(\frac{1}{\sigma\sqrt{2\pi}} \right) - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{Y_i - \beta_1 - \beta_2 X_i}{\sigma} \right)^2$$

$$= n \log\left(\frac{1}{\sigma\sqrt{2\pi}} \right) - \frac{\sigma^{-2}}{2} Z \quad where\ Z = RSS = \sum_{i=1}^{n} (Y_i - \beta_1 - \beta_2 X_i)^2$$

From this obtained expression, maximization of the log likelihood implies minimization of Z. Therefore, it is identical to the OLS procedure ($RSS \to \min_{\beta_1, \beta_2}$) for choosing estimators of β_1 and β_2. Hence, ML estimators of β_1 and β_2 coincide with the OLS ones.

Let's obtain the expression for $\hat{\sigma}_{ML}$ .FOC: $\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \sigma^{-3} Z = \sigma^{-3}\left(Z - n\sigma^2\right)$

Hence, $\hat{\sigma}^2_{ML} = \frac{Z}{n} = \frac{\sum_{i=1}^{n} e_i^2}{n}$

Note that this estimator is biased for finite samples (remember that we should divide by $n-k$ to get an unbiased estimator, where k is the number of estimated parameters in the regression – $k = 2$ for our case). At the same time, the ML estimator of the variance is consistent since the bias disappears as the sample size becomes large.

The goodness of fit in maximum likelihood estimation

1. ***Pseudo- R^2 (or McFadden R^2):***

For ML estimation method it happens that $R^2 = \frac{ESS}{TSS}$ is meaningless. For example, in binary choice models where the dependent variable only takes two states, 0 and 1, (these models will be discussed further in detail) TSS will take different values dependent on the coding of "success" or "failure" even though the independent variables are the same. In these regressions the possibilities for measuring goodness of fit are the *Pseudo- R^2* defined by:

$$1 - \frac{\log L}{\log L_0}$$

where $\log L_0$ is the natural logarithm of the value the likelihood function would take with only the intercept in the regression and $\log L$ is the log-likelihood. Note that $\log L < 0$ since $0 < L < 1$. The values of *Pseudo- R^2* range from 0 to 1; the closer this coefficient is to 1, the better the fit. Moreover, unlike R^2, the *Pseudo- R^2* does not have a natural interpretation

2. ***The likelihood ratio:***

The LR statistic is defined as $LR = 2\log(\frac{L}{L_0}) = 2(\log L - \log L_0)$.

The likelihood ratio is used to test the following hypothesis:

H_0: the coefficients of all explanatory variables are equal to zero;

H_1: the coefficient of at least one explanatory variable is not equal to zero.

Under the null hypothesis, the statistic LR has a χ^2-distribution with $k-1$ degrees of freedom, where k is the number of parameters estimated, and, accordingly, $k-1$ is the number of explanatory variables.

Generally: Suppose, we want to test the hypothesis $H_0 : \theta = \theta_0$ against all possible alternatives. Given a simple random sample $X_1, X_2, X_3, ..., X_n$, a natural way of judging the acceptability of the hypothesis would be to compare the likelihood functions. Let

$l_R \equiv$ Restricted log likelihood (based on the null hypothesis)

$l_{UR} \equiv$ Unrestricted log likelihood (based on the alternative hypothesis)

The LR statistic is defined as

$$LR = 2(l_{UR} - l_R) \overset{H_0}{\sim} \chi^2_q$$ where q is the number of restrictions

It is a large sample test. It has approximately χ^2-distribution with degrees of freedom equal to the number of restrictions imposed by the null hypothesis.

3. The significance of individual coefficients is tested via *Z*-statistics, whose distribution approaches the standard normal in large samples.

Maximum Likelihood Estimation of the Logit and Probit Models

Logit models are fitted by maximum likelihood estimation. Accordingly, the estimates of β_1 and β_2 are chosen so that to maximize the probability of obtaining the actual sample (we maximize the probability with respect to the parameters of the *Z*-function, since the logistic function has no parameters).

Let's write down the maximization problem of the joint probability of outcomes in the general form:

$$L = \prod_i p(Y = Y_i \,|X_i, \beta) = \prod_{i:Y_i=1} F(\beta_1 + \beta_2 \cdot X_i) \cdot \prod_{i:Y_i=0} (1 - F(\beta_1 + \beta_2 \cdot X_i)) \to \max_\beta$$

$$\text{where } F(\beta_1 + \beta_2 \cdot X_i) = \frac{1}{1 + e^{-(\beta_1 + \beta_2 \cdot X_i)}}$$

Log likelihood function is the following:

$$l(\beta) = \log L = \sum_i \left(\log p(Y = Y_i \,|X_i, \beta)\right) = \sum_{i:Y_i=1} \left(\log F(\beta_1 + \beta_2 \cdot X_i)\right) + \sum_{i:Y_i=0} \log(1 - F(\beta_1 + \beta_2 \cdot X_i)) =$$

$$= \sum_i \left(Y_i (\log F(\beta_1 + \beta_2 \cdot X_i)) + (1 - Y_i)(\log(1 - F(\beta_1 + \beta_2 \cdot X_i)))\right) \to \max_\beta$$

Solving this problem we get the Logit-estimates of the parameters of the probability model. There is no analytical solution for the coefficients estimated by MLE. MLE procedure ensures that we will get consistent estimators.

For probit model $F(\beta_1 + \beta_2 \cdot X_i) -$
cumulative function of standardized normal distribution

Censored regressions and tobit analysis.

Suppose, we investigate the dependence of a variable Y^* on a variable X. The true model is given by $Y^* = \beta_1 + \beta_2 X + u$, the variable Y^* is unobservable, but we can observe a variable Y, which behaves as follows:

$Y = Y^*$, if $Y^* > Y_L$

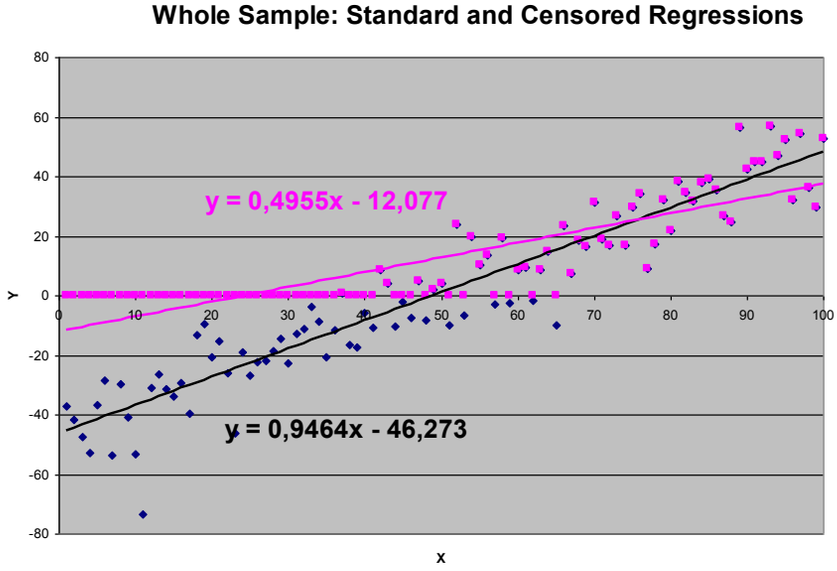
$Y = Y_L$, if $Y^* \le Y_L$

For example, we want to estimate a labour supply model in which the number of labour hours supplied, Y, is regressed on the wage rate, X. Since it is not possible to supply a negative number of hours, the data for Y is limited by below in the sample. Accordingly, in fact, we estimate the model $Y = \beta_1 + \beta_2 X + u_1$, where the disturbance term u_1 does not satisfy the Gauss-Markov conditions (specifically, the conditions 1 and 4). As a consequence, the least squares method yields biased and inconsistent estimators when applied to such a model.

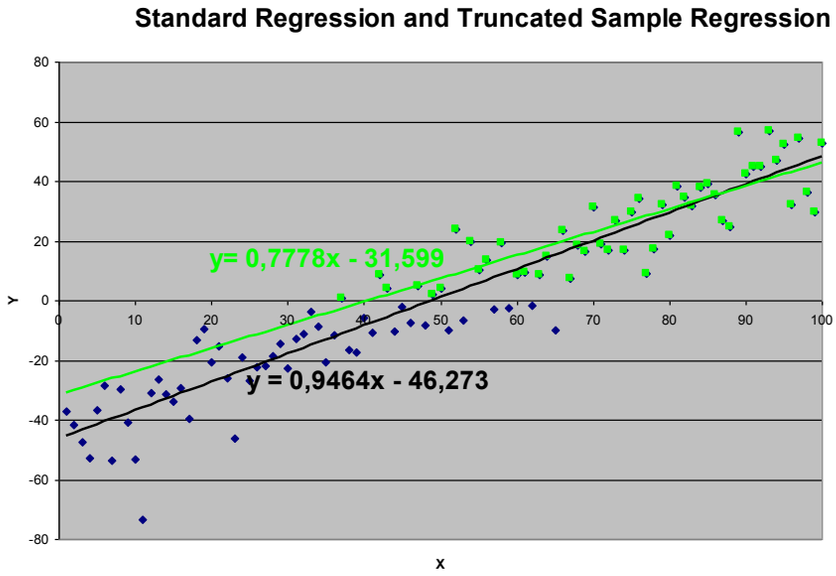
It can be illustrated by the Monte Carlo experiment. Let's generate a sample of 100 observations, for which X takes the values from 1 to 100. The true specification is the following:

$$Y^* = -50 + X + 10 * NRND$$

The following diagram describes one of the realizations of the experiment. The blue circles show the values of Y, while the pink ones plot the values of Y^* (if they differ). The blue line represents the true relationship; the pink line is the regression line, which can be drawn on the basis of the available sample data.



As can be seen from the graph, the slope coefficient of the true dependence is underestimated. In fact, for an observation to appear in the sample, Y^* must be positive, and this requires that $u \geq -\beta_1 - \beta_2 X$. Hence, in the remaining sample X and u related, and the OLS estimators are biased.
At first it can seem that the problem can be solved by means of "truncating" the sample, i.e. by taking into account only those observations in which the value of Y exceeds zero. However, the problem with this method is that the value of Y depends not only on the value of X but also on that of the disturbance term.



The truncation will result in the observations with small values of X and large values of the disturbance term being included, although the corresponding points on the line of the true dependence lay below the X-axis, because the observed values of Y are positive only due to the influence of the disturbance term. On the contrary, the observations with large X and small u will be excluded from the sample, because the disturbance term makes Y negative, even though the corresponding points on the true line lay above the X-axis. Consequently, in the resulting sample X and u will turn out to be negatively correlated, i.e. the 4th Gauss-Markov condition will be violated, and the OLS estimators will be biased and inconsistent.

To solve this problem, known as the censored regressions problem, a special statistical procedure called **tobit analysis** has been developed. It is a combination of linear regression analysis and the Probit method and is based on maximum likelihood estimation.

General description of the model: $$Y_i = \begin{cases} Y^*, & Y_L < Y^* \le Y_U \\ Y_L, & Y^* \le Y_L \\ Y_U, & Y^* > Y_U \end{cases}.$$

As $u_i = Y_i - \beta_1 - \beta_2 X \sim N(0,\sigma^2)$ => $u_i = \dfrac{Y_i - \beta_1 - \beta_2 X}{\sigma} \sim N(0,1)$ – Standard Normal distribution.

Let $f(Z)$ be pdf of Standard Normal distribution. Hence, the log-likelihood is described as the following:

$$l(\beta) = \log L = \sum_{i:Y_L<Y_i<Y_U} \log\left[f\left(\frac{Y_i - \beta_1 - \beta_2 X}{\sigma} \right) \right] + \sum_{i:Y_i=Y_L} \log\left[F\left(\frac{Y_i - \beta_1 - \beta_2 X}{\sigma} \right) \right] +$$

$$+ \sum_{i:Y_i=Y_U} \log\left[1 - F\left(\frac{Y_i - \beta_1 - \beta_2 X}{\sigma} \right) \right] \to \max$$

Tobit provides consistent estimates. It is important to note that tobit estimation is only valid if the disturbance term u is homoscedastic and normally distributed.

Lecture 16

Modeling with Time Series Data. Part 1.

Until now we have analyzed cross-sectional data within the framework of Model A and then Model B, relaxing the assumption of non-stochastic regressors. The common feature of such type of models is that in the data generating process (DGP) – when observations on a number of units are all taken at the same (one) point in time – the ordering of the observations does not influence regression results. In other words, this ordering is arbitrary.

In time series analysis, observations are usually collected at fixed intervals in time (regular spans). As a consequence, there is a so called natural ordering that is reflected by a particular time index. Moreover, there is a certain degree of regularity (persistence) in models with time series data: successive observations are correlated. This describes the behavior of many macroeconomic variables.

In this lecture we will switch to time series models, looking at the analysis within the framework of Model C.

Assumptions of Model C:

C.1 *The model is linear in parameters and correctly specified:* $Y = \beta_1 + \beta_2 \cdot X_2 + ... + \beta_k \cdot X_k + u$;

C.2 ***The time series for the regressors are (at most) weakly persistent;***

C.3 *There does not exist an exact linear relationship among the regressors;*

C.4 Linearity: *The disturbance term has zero expectation*. $E(u_t) = 0$ for all *t*;

C.5 Homogeneity: *The disturbance term is homoscedastic*. $\sigma^2_{u_t} = \sigma^2_u$ for all *t*;

C.6 Independence: *The values of the disturbance term have independent distributions*. u_t is distributed independently of $u_{t'}$ for all $t \neq t'$;

C.7 *The disturbance term is distributed independently of the regressors:* u_i *is distributed independently of* $X_{ji'}$ *for all i' and all j;*

C.8 *The disturbance term has a normal distribution:* $u_t \sim N(0, \sigma^2_u)$

Note that C.2 is a new assumption. Let's look at it in a greater detail:

C.2: Advanced technical concepts are used in the precise definition of weakly persistence. Therefore, it goes beyond the scope of the introduction to econometrics course. Intuitively, highly persistence is connected with strongly dependence. For example, C.2 is violated in the random walk model $X_t = X_{t-1} + u_t$, where $t = 1,2,...$ and u_t is an i.i.d. with zero mean and constant variance. For period $t + h > 0$ the relationship becomes: $X_{t+h} = X_{t+h-1} + u_{t+h}$, where $h > 0$. This can be rewritten as $X_{t+h} = X_t + u_{t+1} + u_{t+2} + ... + u_{t+h}$.

Note that $E(X_{t+h} | X_t) = E(X_t + u_{t+1} + u_{t+2} + ... + u_{t+h}) = X_t$. It means that the value of X_t today affects all the future values of X_{t+h} – indication of highly persistence (strong dependence). Models of the type $X_t = \rho \cdot X_{t-1} + u_t$ with $|\rho| < 1$ are weakly persistent. In fact, $E(X_{t+h} | X_t) = \rho^h \cdot X_t$ => not so strong dependence (the expectation approaches to zero for $h \to \infty$).

As a first approximation, we could use stationarity as a definition, and this criterion seems to be widely adopted in practice (stationarity will be considered in next lectures). But in theory some non-stationary models can be weakly persistent, while not all weakly persistent models are stationary.

All other Model C assumptions and the consequences of their violations are similar to those of the Model B. However, assumptions C.6 and C.7 have a greater relative importance in the context of time series models.

C.6: Due to a certain degree of persistence inherited in time series data, it frequently happens that subsequent values of disturbance terms are correlated. For cross-sectional data the violation of C.6 is a rare situation as observations are generated randomly (even if it happens, we can always rearrange a sample to get rid of this dependence).

C7: This assumption can be divided into 2 parts:

Part (1): The disturbance term in any observation is distributed independently of the values of the regressors in the same observation, and

Part (2): The disturbance term in any observation is distributed independently of the values of the regressors in the other observations.

If both parts hold, then estimates are unbiased. At the same time, (1) is necessary for consistency (but not sufficient). For cross-sectional models part (2) is usually not violated since observations are generated randomly. Hence, unbiasedness depends on part (1) that can be violated by measurement errors and estimation of simultaneous equations. However, for time series data part (2) becomes a major concern. To understand it, let's look at the simple linear regression model:

$$Y = \beta_1 + \beta_2 \cdot X_2 + u$$

Remember that the OLS slope coefficient can be decomposed into the true value and an error term:

$$b_2^{OLS} = \beta_2 + \sum a_i u_i \,, \qquad \text{where} \qquad a_i = \frac{X_i - \overline{X}}{\sum \left(X_i - \overline{X}\right)^2}$$

Taking expectations: $E\left(b_2^{OLS}\right) = E\left(\beta_2\right) + E\left(\sum a_i u_i\right) = \beta_2 + \sum E\left(a_i u_i\right)$.

Assuming that u_i is independent of a_i, where a_i is a function of all X_i, the decomposition becomes $E\left(b_2^{OLS}\right) = \beta_2 + \sum E\left(a_i u_i\right) = \beta_2 + \sum E\left(a_i\right) \cdot E\left(u_i\right) = \beta_2 + \sum E\left(a_i\right) \cdot 0 = \beta_2$. Therefore, for unbiasedness part (1) is insufficient and part (2) assumption is needed.

C.7 assumption for unbiasedness

	Cross-sectional	**Time series**
Part (1)	Required (main concern)	Required
Part (2)	Usually holds by default due to DGP	Required

Consider a model with a lagged dependent variable where the dependent variable with some lag is a part of regressors. The simplest case is the considered random walk model of the form: $Y_t = Y_{t-1} + u_t$. In the next observation: $Y_{t+1} = Y_t + u_{t+1}$

OLS gives a biased result as u_t affects Y_t that is a regressor for the next observation – part (2) condition is violated. The value of bias can be determined implementing Monte Carlo simulation. There is no analytical expression for the bias. At the same time, it can be noted that when the number of observations increases in the sample, the bias seems to disappear – evidence of consistency. Analytically, for more general case of the model:

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + u_t$$

$$b_2^{OLS} = \frac{Cov(Y_t, Y_{t-1})}{Var(Y_{t-1})}$$

$$plim\left(b_2^{OLS}\right) = plim\left(\frac{Cov(Y_t, Y_{t-1})}{Var(Y_{t-1})}\right) = \frac{plim\left(Cov(Y_t, Y_{t-1})\right)}{plim\left(Var(Y_{t-1})\right)} = \frac{\text{cov}(Y_t, Y_{t-1})}{\text{var}(Y_{t-1})} =$$

By taking probability limits: $= \frac{\text{cov}\left((\beta_1 + \beta_2 Y_{t-1} + u_t), Y_{t-1}\right)}{\text{var}(Y_{t-1})} = \beta_2 \cdot \frac{\text{cov}\left(Y_{t-1}, Y_{t-1}\right)}{\text{var}(Y_{t-1})} + \frac{\text{cov}\left(u_t, Y_{t-1}\right)}{\text{var}(Y_{t-1})} =$

$$= \beta_2 + \frac{\sigma_{Y_{t-1}, u_t}}{\sigma^2_{Y_{t-1}}} = \beta_2 + \frac{0}{\sigma^2_{Y_{t-1}}} = \beta_2$$

Hence, the estimate is consistent. We used part (1) condition $\sigma_{Y_{t-1}, u_t} = 0$ by showing consistency. It is quite reasonable to be valid as Y_{i-1} is determined before u_i is generated => they are independent.

C.7 assumption for consistency

	Cross-sectional	**Time series**
Part (1)	Required (main concern)	Required (but not sufficient – technical issue)
Part (2)	Usually holds by default due to DGP	Not Required

Static models and models with lags:

The values of economic variables can depend not only on current values of other (explanatory) variables, but also on their preceding values. When such a relationship with the preceding values of explanatory variables exists, we can speak about a model with a **distributed lag**. This characteristic is notable for time series data. Suppose, we need to estimate the parameters of the following model: $Y_t = \beta_1 + \beta_2 X_t + \beta_3 X_{t-1} + ... + u_t$. The following problems can arise:

1) ***Multicollinearity.*** Lagged values of an explanatory variable can be strongly correlated with each other.
2) ***Decrease in the sample size:*** To include an observation into the sample, the data on the values of regressors in the preceding observations are necessary. For a number of the earliest observations such data are unavailable, therefore, they are excluded from the sample, which thus decreases in size.
3) ***Reduction in the number of degrees of freedom:*** A large number of parameters are estimated and, therefore, the number of degrees of freedom is reduced. The solution can be to make an assumption concerning the distribution of β_i, which is assumed to be characterized with small number of parameters.

Illustration of multicollinearity:

Let's consider a constant elasticity function in which consumer expenditure on housing services depends on aggregate disposable personal income and its relative price index. It takes the form: $HOUS = \beta_1 DPI^{\beta_2} PRELHOUS^{\beta_3} \cdot v$. Here β_2 is the income elasticity and β_3 is the price elasticity for expenditure on housing services. By taking logarithms, the model can be viewed as a linear regression model:

$$LGHOUS = \log \beta_1 + \beta_2 LGDPI + \beta_3 LGPRHOUS + \log v.$$

Then, slope coefficients are estimated using OLS procedure. This corresponds to the cross-sectional analysis.

Time series analysis allows to introduce some simple dynamics. In fact, changes in income and price level at some period t are continuing to be reflected in expenditure on housing for subsequent time periods. There is a so called inertia in the response of housing expenditure. Therefore, new specifications are considered in which lagged values of income and price level become regressors. They are described in the table below.

Alternative dynamic specifications, housing services
(*LGHOUS* is the dependent variable, USA, 1959-2003)

Variable	(1)	(2)	(3)	(4)	(5)
LGDPI	1.03 (0.01)	–	–	0.33 (0.15)	0.29 (0.14)
LGDPI(–1)	–	1.01 (0.01)	–	0.68 (0.15)	0.22 (0.20)
LGDPI(–2)	–	–	0.98 (0.01)	–	0.49 (0.13)
LGPRHOUS	–0.48 (0.04)	–	–	–0.09 (0.17)	–0.28 (0.17)
LGPRHOUS(–1)	–	–0.43 (0.04)	–	–0.36 (0.17)	0.23 (0.30)
LGPRHOUS(–2)	–	–	–0.38 (0.04)	–	–0.38 (0.18)
R^2	0.9985	0.9989	0.9988	0.9990	0.9993

It is evident that direct inclusion of lagged explanatory variables results in severe multicollinearity. It is caused by the high correlation between current and lagged values. Moreover, long lag structure is expected for housing expenditure (2 lags is not always enough), therefore, some other approaches are needed to construct a dynamic model. To alleviate the problem, one can impose some restrictions on the regression coefficients β_i (for example, by assuming that they follow a certain distribution) or employ autoregressive distributed lag model ADL(p, q). The parameter p stands for the maximum number of lags of the dependent variable, and q is the maximum lag of explanatory variables. General expression:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 X_{t-1} + \ldots \beta_{q+2} X_{t-q} + \alpha_3 Y_{t-1} + \alpha_4 Y_{t-2} + \alpha_{p+2} Y_{t-p} + u_t - ADL(p,q)$$

Nevertheless, long-run estimates of elasticities can be obtained by looking at equilibrium relationship. For example, if the process ADL(0,2) $Y_t = \beta_1 + \beta_2 X_t + \beta_3 X_{t-1} + \beta_4 X_{t-2} + u_t$ ever reached equilibrium, the following holds:

$\overline{Y} = \beta_1 + \beta_2 \overline{X} + \beta_3 \overline{X} + \beta_4 \overline{X} = \beta_1 + (\beta_2 + \beta_3 + \beta_4)\overline{X}$, where $\overline{X}$ and $\overline{Y}$ are equilibrium values.

Hence, $\beta_2 + \beta_3 + \beta_4$ is a measure of the long-run effect. While the impact of current X_t on Y_t is measured by β_2 – short-run effect. For instance, in specification (5) the long-run income elasticity is equal to $0.29 + 0.22 + 0.49 = 1$. The long-run price elasticity equals $-0.28 + 0.23 - 0.38 = -0.43$. Corresponding short-run elasticities are 0.29 and -0.28, respectively.

Furthermore, in order to test significance of obtained estimate for the long-run effect, one needs its standard error. It is derived by the following reparametrization of the model. Adding and subtracting $\beta_3 X_t$ and $\beta_4 X_t$ from the right side equation of $Y_t = \beta_1 + \beta_2 X_t + \beta_3 X_{t-1} + \beta_4 X_{t-2} + u_t$, one can get:

$$Y_t = \beta_1 + (\beta_2 + \beta_3 + \beta_4) X_t - \beta_3 (X_t - X_{t-1}) - \beta_4 (X_t - X_{t-2}) + u_t$$

Multicollinearity is unlikely to be an issue here because X_t may not be highly correlated with $X_t - X_{t-1}$ and $X_t - X_{t-2}$.

Estimation of distributed lag models:

As was discussed, one way to deal with the problem 3) is to choose a particular time structure of the model by specifying certain behavior of coefficients over time. The most frequently used such distributions are the following:

1) Geometric distribution (Koyck distribution): coefficients of the explanatory variables have geometrically declining weights, i.e.

$\beta_1;\ \beta_2 = \beta_1 \cdot \rho;\ \beta_3 = \beta_1 \cdot \rho^2 \ldots \beta_k = \beta_1 \cdot \rho^{k-1}$, where $0 < \rho < 1$.

2) Polynomial distribution:

$\beta_s = \widetilde{\beta}_0 + \widetilde{\beta}_1 \cdot s + \widetilde{\beta}_2 \cdot s^2$, where s are integers.

It is also possible to impose restrictions on the parameters of uniform, linear, "triangular" and other distributions.

Geometrically Distributed lag (Koyck model):

General specification of the model is described as:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_2 \rho X_{t-1} + \beta_2 \rho^2 X_{t-2} + \ldots + u_t \qquad (*)$$

The problem of inclusion of infinite number of lags into the model is resolved by Koyck transformation. Multiplying both parts of one time period lagged (*) by ρ:

$$\rho Y_{t-1} = \beta_1 \rho + \beta_2 \rho X_{t-1} + \beta_2 \rho^2 X_{t-2} + \beta_2 \rho^3 X_{t-3} + \ldots + \rho u_{t-1} \qquad (**)$$

Subtracting (**) from (*), note that terms with X from the right side of both expressions are cancelled starting from the first observation's period till $t-1$:

$(*) - (**): \Rightarrow Y_t - \rho \cdot Y_{t-1} = \beta_1(1-\rho) + \beta_2 \cdot X_t + u_t - \rho \cdot u_{t-1}$.

This can be rewritten as:

$$Y_t = \beta_1(1-\rho) + \rho \cdot Y_{t-1} + \beta_2 \cdot X_t + v_t, \text{ where } v_t = u_t - \rho \cdot u_{t-1}$$

However, part (1) of the assumption C.7 that is necessary for both consistency and unbiasedness does not hold as v_t is related to Y_{t-1} because u_{t-1} is a component of v_t and u_{t-1} influences Y_{t-1}. Thus, direct implementation of OLS gives biased and inconsistent estimates. Instead, non-linear procedures of estimation are used. For instance, we can stop increasing the number of lags in (*) when the next increment in lags does not change the values of estimates (achieving stability of coefficients).

Short-run and long-run effects:

Short-run influence of X on Y: β_2

$$\overline{Y} = \beta_1 + \beta_2 \overline{X} + \beta_2 \rho \overline{X} + \beta_2 \rho^2 \overline{X} + \ldots + u_t = \beta_1 + \beta_2 \overline{X}(1 + \rho + \rho^2 + \ldots) + u_t = \beta_1 + \frac{\beta_2 \overline{X}}{1-\rho} + u_t.$$

Long-run influence of X on Y: $\dfrac{\beta_2}{1-\rho}$

Polynomial distributed lag (Almon):

The problem of Koyck procedure is that it is very restrictive where the values of coefficients decline in geometric proportions. However impact of economic variables may be better explained by a quadratic, cubic or higher order polynomial of the form:

$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + u_t,$ where m-order polynomial lag structure defined as:

$$\beta_s = \gamma_0 + \gamma_1 \cdot s + \gamma_2 \cdot s^2 + \ldots + \gamma_m \cdot s^m; \quad s = -1,0,1,2,\ldots$$

The model becomes more flexible. It can incorporate variety of lags.

Calculating for:

s=-1	$\beta_0 = \gamma_0 - \gamma_1 + \gamma_2 + \ldots$
$s = 0$	$\beta_1 = \gamma_0;$
$s = 1$	$\beta_2 = \gamma_0 + \gamma_1 + \gamma_2 + \ldots + \gamma_m;$
$s = 2$	$\beta_3 = \gamma_0 + 2\gamma_1 + 4\gamma_2 + \ldots + 2^m \gamma_m;$
……	

Setting n = 3 and m = 2, we get:

$$Y_t = \alpha + (\gamma_0 - \gamma_1 + \gamma_2) X_t + \gamma_0 X_{t-1} + (\gamma_0 + \gamma_1 + \gamma_2) X_{t-2} +$$
$$(\gamma_0 + 2\gamma_1 + 4\gamma_2) X_{t-3} + u_t = \alpha + \gamma_0 (X_t + X_{t-1} + X_{t-2} + X_{t-3}) +$$
$$\gamma_1 (-X_t + X_{t-2} + 2X_{t-3}) + \gamma_2 (X_t + X_{t-2} + 4X_{t-3}) =$$
$$\alpha + \gamma_0 Z_0 + \gamma_1 Z_1 + \gamma_2 Z_2.$$

Therefore, after estimation of α and $\gamma's$ we can derive $\beta's$, substituting $\gamma's$ into the expressions for β_s.

Lecture 17

Modeling with Time Series Data. Dynamic Processes.

Econometric models often include variables which are not directly observable. For example, an important role in economic analysis belongs to expectations (inflationary expectations, expectations of an exchange rate, etc.). However, there are no statistical data on expectations, because they are unobservable. There are many other unobservable economic variables: "permanent income" in M. Friedman's model of consumption, "the target dividend rate" in G. Lintner's model of dividend policy etc.

A common way of making a model with unobservable variables suitable for estimation and use in economic analysis is to introduce additional premises concerning the behaviour of unobservable variables and describe their interaction with observable ones. Two types of such premises will be considered in the lecture: a premise about adaptive expectations and a premise about partial adjustment. The main features of both models are given in the table below:

	Partial adjustment model	**Adaptive expectation model**
Unobserved variable:	**Dependent** variable	**Explanatory** variable
Adjustment process of:	Observed variable (actual level)	Unobserved variable
Model description:	$Y_t^* = \gamma_1 + \gamma_2 X_t + u_t$ $Y_t - Y_{t-1} = \lambda(Y_t^* - Y_{t-1})$ <=> $Y_t = \lambda Y_t^* + (1-\lambda)Y_{t-1}$ where Y_t^* – unobserved dependent variable	$Y_t = \gamma_1 + \gamma_2 X_{t+1}^e + u_t$ $X_{t+1}^e - X_t^e = \lambda\left(X_t - X_t^e\right)$ <=> $X_{t+1}^e = \lambda X_t + (1-\lambda)X_t^e$ where X_{t+1}^e – unobserved explanatory variable

They are identical from the mathematical point of view, since both include the calculation of a weighted average; the difference consists in the observability of the dependent and explanatory variables. In each models there are problems with parameter estimation, which will be discussed in the lecture.

I. Partial adjustment model:

Suppose, the unobservable ("target" or desired) value of the dependent variable is determined by the equation:

$$Y_t^* = \beta_1 + \beta_2 X_t + u_t \qquad (1)$$

However, this "target" value is reached gradually through a process of partial adjustment. There can be several ways to specify it. The common feature is that for non-observable dependent variable Y_t^*, inertia in the system makes the actual value of Y_t be a compromise (weighted average) between its value in the previous time period, Y_{t-1}, and the value of the unobserved variable justified by the value of the explanatory variable. The adjustment process can vary depending on the time period to which Y^* relates: either current value, Y_t^*, or one period lagged, Y_{t-1}^*. Let's consider 2 cases:

1) *Adjustment to the current value of unobserved variable:* the actual increase in the dependent variable from time $t-1$ to time t, $Y_t - Y_{t-1}$, is proportional to the difference between the

desired value of the dependent variable at the current period and its previous actual value, $Y_t^* - Y_{t-1}$: $Y_t - Y_{t-1} = \lambda(Y_t^* - Y_{t-1})$.

This can be transformed to $Y_t = \lambda Y_t^* + (1-\lambda)Y_{t-1}$ where λ is called the speed of adjustment. λ lies in the range from 0 to 1. $\lambda = 0$ corresponds to no change at all, while $\lambda = 1$ means the full adjustment in the current period. Substituting for Y_t^* from the specification (1) results in the specification in terms of observable variables of the ADL(1,0) form:

$$\begin{aligned} Y_t &= \lambda(\beta_1 + \beta_2 X_t + u_t) + (1-\lambda)Y_{t-1} \\ &= \beta_1\lambda + \beta_2\lambda X_t + (1-\lambda)Y_{t-1} + \lambda u_t \\ &= \alpha_1 + \alpha_2 X_t + \alpha_3 Y_{t-1} + \lambda u_t \end{aligned}$$

where $\alpha_1 = \beta_1\lambda,\ \alpha_2 = \beta_2\lambda,\ \alpha_3 = (1-\lambda).$

In the resulting model the explanatory variable and the disturbance term are not simultaneously correlated, therefore, the estimators will be consistent. However, OLS gives biased estimates as part (2) assumption of C.7 is now violated: $Y_t = \alpha_1 + \alpha_2 X_t + \alpha_3 Y_{t-1} + \lambda u_t$ $Y_{t+1} = \alpha_1 + \alpha_2 X_{t+1} + \alpha_3 Y_t + \lambda u_{t+1}$ It is evident that u_t for Y_t affects the regressor Y_t for the next period dependent variable Y_{t+1}.

Short-run effect of X ***on*** Y***:*** it is measured by α_2 coefficient. It equals $\beta_2\lambda$:

Long-run effect of X ***on*** Y***:*** it is measured by β_2 .

In fact, the relationship between the equilibrium values is written as:

$$\bar{Y} = \beta_1\lambda + \beta_2\lambda\bar{X} + (1-\lambda)\bar{Y}$$

Rearranging, getting how $\bar{Y}$ depends on $\bar{X}$:

$\lambda\bar{Y} = \beta_1\lambda + \beta_2\lambda\bar{X}$. Therefore, $\bar{Y} = \beta_1 + \beta_2\bar{X}$.

2) *Adjustment to the previous value of unobserved variable:* the actual increase in the dependent variable from time $t-1$ to time t, $Y_t - Y_{t-1}$, is proportional to the difference between the previous desired value of the dependent variable and its previous actual value, $Y_{t-1}^* - Y_{t-1}$: $Y_t - Y_{t-1} = \lambda(Y_{t-1}^* - Y_{t-1})$.

This can be transformed to $Y_t = \lambda Y_{t-1}^* + (1-\lambda)Y_{t-1}$ Substituting for Y_{t-1}^* from the lagged by one period specification (1) results in the specification in terms of observable variables of the ADL(1,1) form:

$$\begin{aligned} Y_t &= \lambda(\beta_1 + \beta_2 X_{t-1} + u_{t-1}) + (1-\lambda)Y_{t-1} \\ &= \beta_1\lambda + \beta_2\lambda X_{t-1} + (1-\lambda)Y_{t-1} + \lambda u_{t-1} \\ &= \alpha_1 + \alpha_2 X_{t-1} + \alpha_3 Y_{t-1} + \lambda u_{t-1} \end{aligned}$$

where $\alpha_1 = \beta_1\lambda,\ \alpha_2 = \beta_2\lambda,\ \alpha_3 = (1-\lambda).$

For the same reasons as in 1), results are biased but consistent.

Short-run effect of X **on** Y**:** It equals zero. The dependent variable is not explained by the current value of X .

Long-run effect of X **on** Y**:** It is measured by β_2 (the same derivation as before).

Example:

Lintner's model of dividend adjustment (1956):

In this model it is assumed that a firm has a desired level of dividends based on its expected earnings. When earnings (profit) vary, the firm will adjust its dividends slowly spreading these variations in earnings over a number of time periods. Lintner's model reflects empirically observed phenomenon that shareholders prefer smoothened dividend income. Suppose, D_t actual dividend, Π_t is profit, and D_t^* is "target" dividend. According to Lintner, the "target" dividend can be best explained by earnings to which the actual dividend is adjusting gradually through the partial adjustment process:

$$D_t^* = \alpha + \gamma \Pi_t + u_t$$

$$D_t - D_{t-1} = \lambda (D_t^* - D_{t-1}) + v_t$$

Plugging the expression for the "target" dividend into the equation of partial adjustment, we get: $D_t - D_{t-1} = \lambda\alpha + \gamma\lambda\Pi_t - \lambda D_{t-1} + \lambda u_t + v_t$

Accordingly, $\boxed{D_t = \lambda\alpha + \gamma\lambda\Pi_t + (1-\lambda)D_{t-1} + \lambda u_t + v_t}$

In this model the explanatory variable D_{t-1} and the disturbance term are related, but not simultaneously correlated, therefore, the estimators obtained will be biased, but consistent. The model can be directly estimated, since it does not include unobservable variables. G. Lintner has estimated the model on the data for the US corporate sector for 1918-1941 and has obtained the following results: $\gamma = 0.3$; $\lambda = 0.5$.

II. Adaptive expectations

There are 2 sources of dynamics for this type of models: inertia that is the drag from the past (like in the partial adjustment models) and the effect of anticipations. Economic agents form expectations about the future values of variables and then adjust their plans accordingly. For example, inflationary expectations influence the current interest rate and the demand for money, and the expected exchange rate affects the supply and demand for a currency. To estimate such type of models, it is necessary to introduce an assumption concerning the behaviour of the unobservable variable. In its simplest form, the dependent variable Y_t is explained by the anticipated value of X_{t+1}^e in the next period:

$$\boxed{Y_t = \beta_1 + \beta_2 X_{t+1}^e + u_t} \qquad (2)$$

X_{t+1}^e is subjective, so it is assumed to be described as the following process:

$$X_{t+1}^e - X_t^e = \lambda\left(X_t - X_t^e\right)$$

λ is interpreted as a speed of adjustment (adaptation). It shows the share of expectations that are based on the actual behavior of the explanatory variable X, while $1-\lambda$ of them are related to the previous period anticipated value.

The adaptive expectation process can be rewritten as: $X_{t+1}^e = \lambda X_t + (1-\lambda)X_t^e$ (*)

Substituting for X_{t+1}^e from (2), we can obtain: $Y_t = \beta_1 + \beta_2\lambda X_t + \beta_2(1-\lambda)X_t^e + u_t$.

This equation still involves the unobserved variable. The adaptive expectation relationship (*) also holds for $t-1$: $X_t^e = \lambda X_{t-1} + (1-\lambda)X_{t-1}^e$.

Substituting for X_t^e in the equation for Y_t:

$$Y_t = \beta_1 + \beta_2\lambda X_t + \beta_2\lambda(1-\lambda)X_{t-1} + \beta_2(1-\lambda)^2 X^e_{t-1} + u_t$$

Lagging and substituting s times in this way, we get:

$$\begin{aligned} Y_t = {} & \beta_1 + \beta_2\lambda X_t + \beta_2\lambda(1-\lambda)X_{t-1} + \beta_2\lambda(1-\lambda)^2 X_{t-2} + \ldots \\ & + \beta_2\lambda(1-\lambda)^{s-1} X_{t-s+1} + \beta_2(1-\lambda)^s X^e_{t-s+1} + u_t \end{aligned} \qquad (**)$$

For $s \to \infty$, $(1-\lambda)^s$ tends to zero => we can drop the unobservable final term. The specification is non-linear in parameters; therefore it can be fitted using some nonlinear estimation technique.

As $0 \le \lambda \le 1$, it follows that $0 \le 1-\lambda \le 1$ => we get a lag structure with geometrically declining weights – Koyck distribution. Using Koyck transformation the model can be expressed in terms of finite number of observable variables. Lagging (**) and then multiplying it by $1-\lambda$,

$$\begin{aligned} (1-\lambda)Y_{t-1} = {} & \beta_1(1-\lambda) + \beta_2\lambda(1-\lambda)X_{t-1} + \beta_2\lambda(1-\lambda)^2 X_{t-2} + \beta_2\lambda(1-\lambda)^3 X_{t-3} + \ldots \\ & + \beta_2\lambda(1-\lambda)^{s-1} X_{t-s+1} + \beta_2(1-\lambda)^s X^e_{t-s+1} + (1-\lambda)u_{t-1} \end{aligned}$$

Subtracting the derived expression from (**) under $s \to \infty$ (=> final terms can be dropped):

$$Y_t - (1-\lambda)Y_{t-1} = \beta_1 + \beta_2\lambda X_t - \beta_1(1-\lambda) + u_t - (1-\lambda)u_{t-1}$$

Accordingly:
$$\begin{aligned} Y_t &= \beta_1\lambda + \beta_2\lambda X_t + (1-\lambda)Y_{t-1} + u_t - (1-\lambda)u_{t-1} \\ &= \alpha_1 + \alpha_2 X_t + \alpha_3 Y_{t-1} + u_t - (1-\lambda)u_{t-1} \end{aligned}$$

where $\alpha_1 = \beta_1\lambda,\ \alpha_2 = \beta_2\lambda,\ \alpha_3 = (1-\lambda)$

Short-run effect of X ***on*** Y***:*** it is measured by α_2 coefficient. It equals $\beta_2\lambda$.

Long-run effect of X ***on*** Y***:*** it is measured by β_2.

<u>Another possible way</u> to derive this relationship is to use the original specification one period lagged: $Y_{t-1} = \beta_1 + \beta_2 X^e_t + u_{t-1}$. From this let's obtain the expression of $\beta_2 X^e_t$:

$$\beta_2 X^e_t = Y_{t-1} - \beta_1 - u_{t-1}$$

Substituting for $\beta_2 X^e_t$ in $Y_t = \beta_1 + \beta_2\lambda X_t + \beta_2(1-\lambda)X^e_t + u_t$ (for $s=1$ in (**)):

$$\begin{aligned} Y_t &= \beta_1 + \beta_2\lambda X_t + (1-\lambda)(Y_{t-1} - \beta_1 - u_{t-1}) + u_t \\ &= \beta_1\lambda + \beta_2\lambda X_t + (1-\lambda)Y_{t-1} + u_t - (1-\lambda)u_{t-1} \\ &= \alpha_1 + \alpha_2 X_t + \alpha_3 Y_{t-1} + u_t - (1-\lambda)u_{t-1} \end{aligned}$$

<u>*Estimation:*</u>

Note that, mathematically it is the same ADL(1,0) model as for the partial adjustment process apart from the compounded disturbance term. Hence, it would be difficult to differentiate between these two models in practice, despite the fact that the approaches are opposite in spirit. This is an example of observational equivalence of two theories. However, the compounded disturbance term plays important role here: OLS estimates become biased and inconsistent because both parts of C.7 are now violated. Part (1): u_{t-1} affects Y_{t-1} in the observation for Y_t. Part (2) the same reasons as for the partial adjustment process.

Therefore, the model has to be estimated step by step as the geometrically distributed lag model until the coefficients cease to differ by some specified number in 2 consecutive steps, i.e.

0) Use (**): $\begin{aligned} Y_t = {} & \beta_1 + \beta_2\lambda X_t + \beta_2\lambda(1-\lambda)X_{t-1} + \beta_2(1-\lambda)^2 X_{t-2} + \ldots \\ & + \beta_2(1-\lambda)^{s-1} X_{t-s+1} + \beta_2(1-\lambda)^s X^e_{t-s+1} + u_t \end{aligned}$;

1) Estimate the model for $s=1$;
2) Estimate the model for $s=2$;

……..

s') Estimate the model for $s=s'$ until fitted coefficients coincide with corresponding coefficients from the step $s'-1$) with the specified accuracy. Hence, choose estimation for which $s=s'-1$.

Example:

Cagan's model of hyperinflation

The basic equation of the model is the equation of the demand for real money balances as a function of the next period expected rate of inflation: $\frac{M}{P}=f(\pi_{t+1}^{e})$.

In the conditions of hyperinflation, the appropriate time unit is normally a month or even a week. The standard Cagan's model is specified as follows:

$$log\,(\frac{M}{P})=\alpha-\rho\cdot\pi_{t+1}^{e}+u_t$$

The parameter ρ in this model shows the percentage decrease in the demand for real money balances associated with a one percent increase in the expected inflation rate.

Let π_{t+1}^{e} follow a partial adjustment process:

$$\pi_{t+1}^{e}-\pi_t^{e}=\beta\left(\pi_t-\pi_t^{e}\right)$$

Then, $\log(\frac{M}{P})_t=\alpha-\rho\beta(\pi_t+(1-\beta)\pi_{t-1}+(1-\beta)^2\pi_{t-2}+...)+u_t$

Using monthly data for 7 cases of a hyperinflation, he has obtained the following estimates of the model parameters: ρ = 4,68; β = 0,20. They indicate that in a hyperinflation the demand for real money balances is not very sensitive to the expected inflation rate, and the inflationary expectations are adjusted rather slowly.

Friedman's Permanent Income Hypothesis:

Lecture 10 considered Milton Friedman's Permanent Income Hypothesis (PIH) as the application of the measurement error analysis. According to this theory, permanent consumption is proportional to permanent income: $C_t^{P}=\beta_2 Y_t^{P}$. Permanent income is a subjective concept of likely medium-run future income. Actual income and consumption is decomposed into permanent and transitory components:

$$Y_t=Y_t^{P}+Y_t^{T}$$
$$C_t=C_t^{P}+C_t^{T}$$

Since the permanent income is unobservable, suppose that it follows the adaptive expectation process: $Y_t^{P}-Y_{t-1}^{P}=\lambda(Y_t-Y_{t-1}^{P})$. This can be rewritten as follows: permanent income at t is a weighted average of actual income at t and permanent income at $t-1$, i.e. $Y_t^{P}=\lambda Y_t+(1-\lambda)Y_{t-1}^{P}$. This can be substituted into the consumption function for Y_t^{P}:

$$C_t-C_t^{T}=\beta_2\left(\lambda Y_t+(1-\lambda)Y_{t-1}^{P}\right)$$

By lagging the adaptive process $Y_{t-1}^{P}=\lambda Y_{t-1}+(1-\lambda)Y_{t-2}^{P}$ and substituting it for Y_{t-1}^{P}, we get:

$$C_t=\beta_2\lambda Y_t+\beta_2\lambda(1-\lambda)Y_{t-1}+\beta_2(1-\lambda)^2Y_{t-2}^{P}+C_t^{T}$$

Repeating this procedure s times, we obtain:

$$C_t = \beta_2 \lambda Y_t + \beta_2 \lambda (1-\lambda) Y_{t-1} + \beta_2 \lambda (1-\lambda)^2 Y_{t-2} + \ldots + \beta_2 \lambda (1-\lambda)^{s-1} Y_{t-s+1} + \beta_2 (1-\lambda)^s Y_{t-s}^P + C_t^T$$

It is non-linear in parameters specification. Friedman fitted it using non-linear iterative estimation method. In order to evaluate short-run and long-run effects, we can either use Koyck transformation or the described in the previous section another way to get rid of unobservable variables. Lagging the basic relationship one period $\beta_2 Y_{t-1}^P = C_{t-1}^P = C_{t-1} - C_{t-1}^T$. Substituting it into $C_t - C_t^T = \beta_2 \left(\lambda Y_t + (1-\lambda) Y_{t-1}^P \right)$, we obtain:

$$\begin{aligned} C_t &= \lambda \beta_2 Y_t + (1-\lambda)\left(C_{t-1} - C_{t-1}^T\right) + C_t^T \\ &= \lambda \beta_2 Y_t + (1-\lambda) C_{t-1} + C_t^T - (1-\lambda) C_{t-1}^T \end{aligned}$$

Short run marginal propensity to consume equals $mpc_{SR} = \lambda \beta_2$;

Long run marginal propensity to consume equals $mpc_{LR} = \beta_2$.

Thus, since λ is less than 1, the model is able to explain the fact that after the Second World War the long-run average propensity to consume seemed to be roughly constant despite the marginal propensity to consume being much lower.

Lectures 18-19

Autocorrelation.

The assumption C.6 of Model C which is equivalent to the third Gauss-Markov condition, requires of the values of the disturbance term to have independent distributions, i.e. u_t is distributed independently of $u_{t'}$ for all $t \neq t'$. This assumption is usually satisfied for cross-sectional data models, while it frequently turns out to be violated when one deals with time series data: error terms for time periods not too far apart may be correlated. When C.6 assumption does not hold, it is said that the disturbance term is subject to autocorrelation, or serial correlation. It can be explained by the persistence of the behaviour of factors combined in the disturbance term in time. This lecture will analyze autocorrelation according to the following plan:

1. Reasons
2. Consequences
3. Detection
4. Remedial measures

I. Reasons

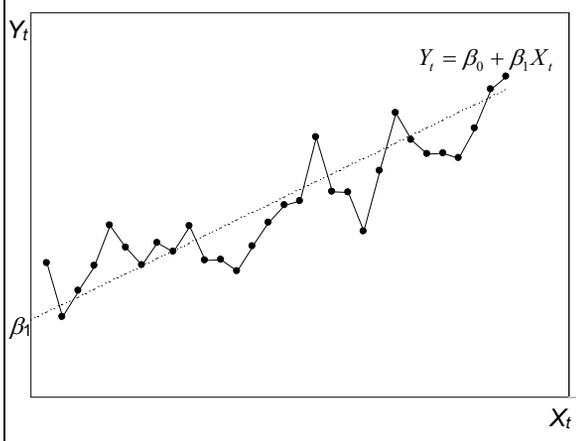
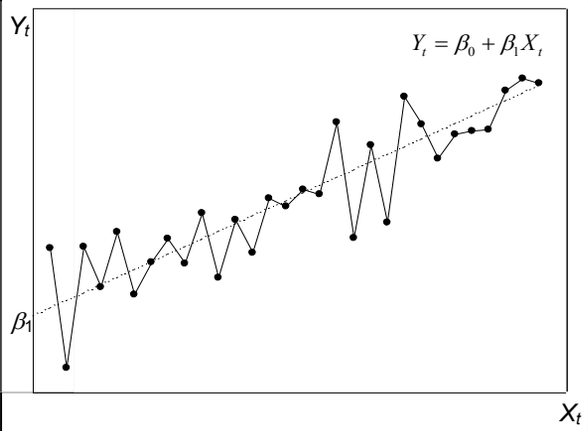
In time-series models autocorrelation of the disturbances can arise from the following reasons:

1) Omitted variables: relevant factors omitted from the regression are correlated across periods. Since the disturbance term picks these missing factors up, the dependence of omitted factors (their persistence in time) translates into apparent autocorrelation of the disturbance term. For example, let the true model be a model with 2 explanatory variables:
$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + v_t$$
However, the second factor, X_{t2}, is not included into the estimated regression:
$$Y_t = \beta_0 + \beta_1 X_{t1} + u_t$$
Thus, the effect of X_{t2} is captured by the disturbance term u_t.Suppose that the omitted variable exhibits trends over time: X_{t2} depends on $X_{t-1,2}$, $X_{t-2,2}\ldots$ Therefore, as u_t combines these effects, the assumption of the serial independence of the disturbance term is violated
2) Misspecification of the functional form: for example, suppose that we use a linear specification for the relationship between Y and X instead of a quadratic one. Consider that X is growing over time. Thus, as in the linear model the disturbance term depends on X^2 that is increasing, the error term will also exhibit such growth, giving rise to autocorrelation.
3) Measurement errors: systematic errors (in the way how variables are measured) are reflected in the disturbance term that will behave systematically, showing a pattern. This can cause serial correlation.

Illustration:

Autocorrelation can be observed in scatter diagrams.

	Positive autocorrelation	**Negative autocorrelation**
Features:	Positive values of residuals tend to be followed by positive ones, and negative	Positive values of residuals tend to be followed by negative ones, and negative

	values by negative ones. Successive values tend to have the same sign.	values by positive ones. Successive values tend to alternate in sign.
Graph:		
Frequency	Common in economic analysis. Explained by persistence of effects of omitted variables. For more observations (longer intervals) this effect is less perceived.	Uncommon in economic analysis. Sometimes this type of autocorrelation is induced by transformations of the original specification to make the model suitable for regression analysis.

Autocorrelation processes:

Consider a model: $Y_t = \beta_1 + \beta_2 X_t + u_t$. There are many forms of process to capture serial correlation:

1) First-order autoregressive autocorrelation: AR(1) – common type of autocorrelation described as:

$$u_t = \rho u_{t-1} + \varepsilon_t$$

It is autoregressive, because u_t depends on lagged values of itself. It is the first-order, because there is only one lag (dependence on its previous value). ε_t is innovation term at time t (white noise with zero mean and positive variance). For $\rho > 0$, the process is subject to positive autocorrelation. While for $\rho < 0$, there is negative autocorrelation.

For stationarity purposes we consider $|\rho| < 1$. Note that $E(u_t) = 0$.

Demonstration:

Lagging the process on time period: $u_{t-1} = \rho u_{t-2} + \varepsilon_{t-1}$. Substitute it for u_{t-1} into AR(1):

$$u_t = \rho^2 u_{t-2} + \rho\varepsilon_{t-1} + \varepsilon_t$$

Continuing to lag and substitute, we can express AR(1) in terms of innovation terms with diminishing weights: $E(u_t) = E(\varepsilon_t) + \rho E(\varepsilon_{t-1}) + \rho^2 E(\varepsilon_{t-2}) + \rho^3 E(\varepsilon_{t-3}) + \ldots = 0$

By implementing Monte Carlo simulation, we can generate plots of residuals depending on the value of ρ. It shows that for $|\rho| < 0.3$, autocorrelation is weak and practically invisible. From $|\rho| > 0.3$, autocorrelation is beginning to be apparent. From $|\rho| > 0.6$, there is an obvious pattern in the behaviour of the disturbance term that is observed more frequently than we would expect as a matter of chance. When $|\rho|$ is around 0.9, autocorrelation is strong.

2) p^{th} order autoregressive autocorrelation: AR(p)

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \ldots + \rho_p u_{t-p} + \varepsilon_t$$

Similarly, it is autoregressive but here it depends on lagged values of u_t up to the p[th] lag. It can also be demonstrated (by analogy) that if $|\rho| < 1$, then $E(u_t) = 0$.

3) q[th] order moving average autocorrelation: MA(q)

$$u_t = \lambda_0 \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2} + ... + \lambda_q \varepsilon_{t-q}$$

The disturbance term is a linear combination of innovation terms up to the q[th] lag. Immediately, it follows that $E(u_t) = 0$.

4) ARMA(p,q) – general type of process:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + ... + \rho_p u_{t-p} + \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2} + ... + \lambda_q \varepsilon_{t-q}$$

It is a combination (aggregation) of AR(p) and MA(q). Hence, $E(u_t) = 0$

II. ***Consequences of autocorrelation:***

The consequences of autocorrelation for OLS are similar to those of heteroscedasticity. In general, the regression coefficients remain unbiased, but OLS is inefficient because one can find an alternative regression technique that yields estimators with smaller variances. The other main consequence is that autocorrelation causes the standard errors to be estimated wrongly, often being biased downwards (for positive autocorrelation). Thus, *t*- and *F*-tests are no longer valid. Finally, although in general OLS estimates are unbiased, there is an important special case where they are biased (lagged dependent variable as a regressor).

Let's demonstrate some of these results for the case of the simple linear regression model:

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

Unbiasedness:

The estimated slope coefficient can be decomposed as follows:

$$b_2 = \beta_2 + \sum_{t=1}^{T} a_t u_t \qquad \text{where} \qquad a_t = \frac{X_t - \bar{X}}{\sum_{s=1}^{T} (X_s - \bar{X})^2}$$

Taking expectation: $E(b_2) = \beta_2 + E\left(\sum_{t=1}^{T} a_t u_t\right) = \beta_2 + \sum_{t=1}^{T} E(a_t u_t)$. If C.7 is satisfied, then a_t and u_t are distributed independently => expectation can be rewritten as $\sum_{t=1}^{T} E(a_t u_t) = \sum_{t=1}^{T} E(a_t) E(u_t)$.

Since $E(u_t) = 0$ independent of whether the disturbance term is subject to autocorrelation:

$$E(b_2) = \beta_2 + \sum_{t=1}^{T} E(a_t) E(u_t) = \beta_2 + 0 = \beta_2 \text{ – unbiased.}$$

Inefficiency:

We will not show this result analytically. Nevertheless, it can be mentioned that the proof of the Gauss-Markov theorem that established efficiency relies on the assumption of no autocorrelation. Since it is violated, OLS estimates are no longer BLUE and are hence inefficient.

Special case: The OLS estimators are biased and inconsistent for the model with the lagged dependent variable as a regressor with the disturbance term subject to autocorrelation:

Demonstration: $\begin{cases} Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + u_t \\ u_t = \rho u_{t-1} + \varepsilon_t \end{cases} <=> Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + \rho u_{t-1} + \varepsilon_t$

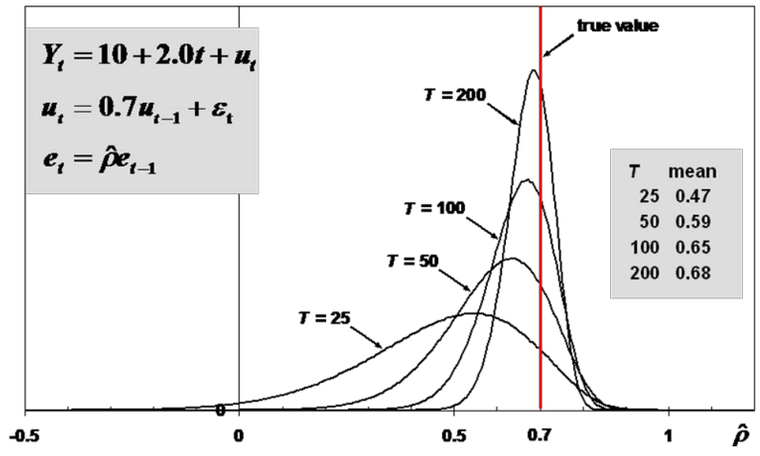
Lagging the model by one time period: $Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + \beta_3 Y_{t-2} + u_{t-1}$. It is evident that Y_{t-1} depends on u_{t-1}. Thus, assumption C.7 is violated, because the disturbance term affects Y_{t-1} in the observation for Y_t.

III. Detection: tests for autocorrelation:

1. Residual plot: a graph of the estimated residuals e_t against time.

Since the disturbance term is theoretical concept, it is non-observable. Residuals are not the values of the disturbance term but their behaviour is very much similar to the behaviour of the disturbance term. Therefore, one way to detect autocorrelation is to look at residuals: it is likely that if the disturbance term is subject to autocorrelation, then the residuals will be subject to a similar pattern of autocorrelation.

For example, if the disturbance term follows the AR(1) process, then provided that the sample size is large enough (condition for consistency is satisfied), the regression parameters will converge to their true values and residuals will be very close to the values of the disturbance term. Hence, a regression of e_t on e_{t-1} can be sufficient, at least in large samples. It can be illustrated with the Monte Carlo experiment shown on the graph below.



The true model is generated as AR(1) with $\rho = 0.7$. On the horizontal axis it shows the estimated value of this coefficient when e_t is regressed on e_{t-1}. For a small sample size the distribution of ρ is skewed to the left and heavily biased downwards. However, as the sample size increases, the bias becomes sufficiently small and it is clear that the mean value is approaching to the true value of ρ.

2. Breusch-Godfrey test:

The idea of the Breusch-Godfrey test is that in order to control the effects of any endogeneity on the residuals we should also include all explanatory variables from the original model into the residuals regression. However, the theory is complex relating to maximum-likelihood estimation. Therefore, several asymptotically-equivalent versions of the test have been proposed. We will look at the Lagrange Multiplier version of the test. It allows to identify serial correlation not only of the first order but higher orders as well. Note that it is valid only for large samples.

Consider a linear regression:

$$Y_t = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jt} + u_t$$

The disturbance term might follow an AR(p) process: $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + ... + \rho_p u_{t-p} + \varepsilon_t$.

Steps:

1) Estimate the initial regression model by OLS and calculate its residuals, e_t;

2) Estimate $e_t = \gamma_1 + \sum_{j=2}^{k} \gamma_j X_{jt} + \sum_{s=1}^{p} \rho_s e_{t-s}$ (so called auxiliary equation) using the $T - p$ observations: $p+1$ through T;

3) Calculate LM statistics: $(T-p)\cdot R_e^2$ where R_e^2 is the R-squared from the auxiliary regression. $T-p$ is the actual number of observations. LM statistics is distributed as $\chi^2(p)$ when testing for p[th] order autocorrelation.

4) H_0 : no autocorrelation

H_1 : not H_0

Perform Chi-square test: reject the null hypothesis of zero autocorrelation in favour of the alternative if $(T-p)\cdot R_e^2 > \chi^2_{\alpha\%signlevel}(p)$ and not reject otherwise. This test is valid only asymptotically.

Note that, alternatively, for $p=1$, simple *t*-test on coefficient e_{t-1} can be used with asymptotically validity. For $p>1$, *t*-test version becomes *F*-test on the lagged residuals comparing RSS_{UR} for the auxiliary regression with RSS_R for the same specification without the residual terms, i.e.

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

$H_1 : at\ least\ one\ \rho \neq 0$

Moreover, this test is asymptotically valid for MA(p) autocorrelation.

3. Durbin-Watson test:

This test is used for the detection of AR(1) autocorrelation. The null hypothesis states that in the equation $u_i = \rho \cdot u_{i-1} + \varepsilon_i$ the true value of $\rho = 0$. The following Durbin-Watson statistics is calculated from estimated residuals:

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

It is shown that $0 < d < 4$.

With a strong positive residual autocorrelation, $DW \to 0$;

With a strong negative residual autocorrelation, $DW \to 4$;

When $\rho = 0$, $DW \to 2$.

In practice it is convenient to use the approximate formula of the Durbin-Watson statistics (for large samples):

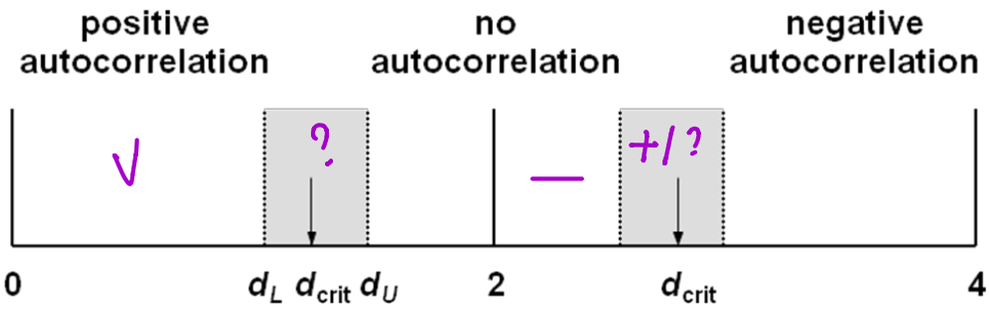
$DW = \frac{\sum_t e_t^2 - 2\sum_t e_t e_{t-1} + \sum_t e_{t-1}^2}{\sum_t e_t^2} \approx 2\cdot(1 - \frac{\sum_t e_t e_{t-1}}{\sum_t e_t^2})$, since $\sum_t e_t^2$ is approximately equal to $\sum_t e_{t-1}^2$.

Further, as $\sum_t e_t^2 \approx \sqrt{\sum_t e_t^2}\cdot\sqrt{\sum_t e_{t-1}^2}$ and the sample correlation coefficient between e_t and e_{t-1} is given by $r_{t,t-1} = \frac{\sum_t e_t e_{t-1}}{\sqrt{\sum_t e_t^2}\cdot\sqrt{\sum_t e_{t-1}^2}}$, the Durbin-Watson statistic can be approximated by the following expression: $DW \approx 2\cdot(1 - \frac{\sum_t e_t e_{t-1}}{\sum_t e_t^2}) \approx 2\cdot(1 - r_{t,t-1})$.

The distribution of the Durbin-Watson statistics depends not only on *k* (the number of explanatory variables) and *n* (the numbers of observations), but also on the sample values

explanatory variables. Therefore the exact critical value of the statistic for any given significance level will be different for different samples. Durbin and Watson showed that the distribution of d is bounded by two limiting distributions.

The decision rule is described as follows:



From the table of the Durbin-Watson statistics we take upper and lower bounds of d, given the number of observations in the sample and the number of explanatory variables for the chosen significance level.

Test for positive autocorrelation: $\begin{matrix} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{matrix}$.

Reject H_0 if $d \le d_L$. If $d \ge d_U$, we cannot reject H_0. If $d_L < d < d_U$, the test is inconclusive.

Test for negative autocorrelation: $\begin{matrix} H_0 : \rho = 0 \\ H_1 : \rho < 0 \end{matrix}$ use $4 - d$. This is done when d is greater than 2.

If $4 - d \le d_L$, there is significant negative autocorrelation. If $4 - d \ge d_U$, there is evidence of no negative autocorrelation. The test is inconclusive for $d_L < 4 - d < d_U$.

Main features and remarks of the Durbin-Watson test:

1) It detects only first-order autocorrelation;
2) It cannot be applied if the model contains lagged dependent variable (the statistics is biased towards 2 for such case);
3) It cannot be applied if the model does not contain constant term;
4) There is uncertainty zone (test is inconclusive);
5) It is appropriate for finite samples and is provided directly by standard packages.

Extension: Durbin's test for detection of AR(1) autocorrelation in the case of the lagged dependent variable used as a regressor.

Durbin proposed ***h*-test:** $h = \hat{\rho}\sqrt{\dfrac{n}{1 - n s^2_{b_{Y(-1)}}}}$, where

$\hat{\rho}$ is an estimate of ρ in the AR(1) process. Since the test is only valid for large samples, $d \to 2 - 2\hat{\rho}$. Hence, the estimator is then $\hat{\rho} = 1 - 0.5d$;

$s^2_{b_{Y(-1)}}$ is an estimate of the variance of the coefficient of the lagged dependent variable;

n is an actual number of observations.

This statistics is distributed as a normal variable with zero mean and unit variance (asymptotically). Hence, standard Z-test is performed. The problem is that $n s^2_{b_{Y(-1)}}$ can be greater than, or close to 1 (it usually happens when sample size is not very large). Then h-statistics cannot be computed.

IV. Remedial measures:

If the test for detection of autocorrelation results in the rejection of the null hypothesis, it is necessary to do the following:

1) Try to improve the model specification
2) If the final specification is considered correct but the null hypothesis is still rejected, the disturbance term is subject to autocorrelation. It can be eliminated by means of an **autoregressive transformation**.

Eliminating AR(1) autocorrelation: one explanatory variable

$$\begin{cases} Y_t = \beta_1 + \beta_2 X_t + u_t \\ u_t = \rho u_{t-1} + \varepsilon_t \end{cases}$$

The regression is also valid for $t-1$: $Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1}$. Multiply this equation by ρ to get:

$$\rho Y_{t-1} = \beta_1 \rho + \beta_2 \rho X_{t-1} + \rho u_{t-1}$$

Subtract it from the original equation:

$$Y_t - \rho Y_{t-1} = \beta_1 (1-\rho) + \beta_2 X_t - \beta_2 \rho X_{t-1} + u_t - \rho u_{t-1}$$

As $u_t - \rho u_{t-1} = \varepsilon_t$ for which all Gauss-Markov conditions are satisfied, the problem of the autocorrelated disturbance term is now eliminated. The model becomes:

$$Y_t = \beta_1 (1-\rho) + \rho Y_{t-1} + \beta_2 X_t - \beta_2 \rho X_{t-1} + \varepsilon_t$$

It is ADL(1,1) model of the form $Y_t = \lambda_1 + \lambda_2 Y_{t-1} + \lambda_3 X_t + \lambda_4 X_{t-1} + \varepsilon_t$ with a non-linear restriction: the coefficient of X_{t-1} is minus the product of the coefficients of X_t and Y_t, i.e. $\lambda_4 = -\lambda_2 \cdot \lambda_3$. Therefore, a non-linear estimation technique is used. For example, in EViews it is done by the following command: Y=C(1)*(1-C(2))+C(2)*Y(-1)+C(3)*X-C(2)*C(3)*X(-1).

Eliminating AR(1) autocorrelation: two explanatory variables

$$\begin{cases} Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \\ u_t = \rho u_{t-1} + \varepsilon_t \end{cases}$$

Using the same procedure as before: lagging the original specification, multiplying it by ρ and then subtracting the derived equation from the initial regression, we get:

$$\rho Y_{t-1} = \beta_1 \rho + \beta_2 \rho X_{2t-1} + \beta_3 \rho X_{3t-1} + \rho u_{t-1}$$

$$Y_t - \rho Y_{t-1} = \beta_1 (1-\rho) + \beta_2 X_{2t} - \beta_2 \rho X_{2t-1} + \beta_3 X_{3t} - \beta_3 \rho X_{3t-1} + u_t - \rho u_{t-1}$$

So, the model is transformed into ADL(1,1) of the form $Y_t = \lambda_1 + \lambda_2 Y_{t-1} + \lambda_3 X_{2t} + \lambda_4 X_{2t-1} + \lambda_5 X_{3t} + \lambda_6 X_{3t-1} + \varepsilon_t$ with 2 non-linear restrictions:

$$Y_t = \beta_1 (1-\rho) + \rho Y_{t-1} + \beta_2 X_{2t} - \beta_2 \rho X_{2t-1} + \beta_3 X_{3t} - \beta_3 \rho X_{3t-1} + \varepsilon_t$$

Restrictions: $\begin{cases} \lambda_4 = -\lambda_2 \cdot \lambda_3 \\ \lambda_6 = -\lambda_2 \cdot \lambda_5 \end{cases}$

There are several ways to estimate this regression in EViews:

1) Y=C(1)*(1-C(2))+C(2)*Y(-1)+C(3)*X2-C(2)*C(3)*X2(-1)+C(4)*X3-C(2)*C(4)*X3(-1);
2) Add AR(1) to the list of explanatory variables in the initial regression. Note that coefficients of lagged explanatory variables not presented in the original specification (X_{2t-1} and X_{3t-1} here) are not reported. For higher orders of autocorrelation add AR(1), AR(2), …, up to AR(p), where p is expected order of autocorrelation to deal with.

Cochrane-Orcutt iterative process:

In early days of computing, non-linear techniques was not so simple and widely applicable. Therefore, some other methods were used. The Cochrane–Orcutt iterative procedure requires the transformation of the regression model to a form in which the OLS procedure is applicable.

Consider a simple linear regression model $Y_t = \beta_1 + \beta_2 X_t + u_t$ where the disturbance term follows AR(1) process $u_t = \rho u_{t-1} + \varepsilon_t$.

This transformation is done in several steps:

1) Estimate the original equation by OLS and compute its residuals e_t;
2) Estimate the first-order serial correlation coefficient (ρ) by regressing e_t on e_{t-1}

$$\widetilde{Y}_t = Y_t - \rho Y_{t-1}$$

3) Transform the variables as follows: $\widetilde{X}_t = X_t - \rho X_{t-1}$;

$$\beta_1' = \beta_1(1-\rho)$$

4) Regress $\widetilde{Y}_t$ on $\widetilde{X}_t$ to obtain revised estimates b_1 and b_2;
5) Plug estimated b_1 and b_2 into the original regression, and then obtain a new set of estimates for residuals. Go back and repeat step 2).

This iterative procedure can be stopped when the estimates of ρ from two successive iterations differ by no more than some predetermined values, such as 0.001. The final value of ρ is then used to get estimates for transformed regression.

Making the autoregressive transformation, we lose the first observation. To avoid this, the **Price–Winsten correction** is applied: on the 4-th step of the Cochrane-Orcutt procedure the first observation multiplied by $\sqrt{1-\rho^2}$ is added to the transformed observations [2; T].

The introduction of this multiplier is explained by the need to deal with the heteroscedasticity problem, which arises due to the following effect:

$$u_t = \rho \cdot u_{t-1} + \varepsilon_t \Rightarrow \sigma_u^2 = \mathrm{var}(u_t) = \mathrm{var}(\rho \cdot u_{t-1} + \varepsilon_t) = \rho^2 \,\mathrm{var}(u_{t-1}) + 2\,\mathrm{cov}(u_{t-1}, \varepsilon_t) + \mathrm{var}(\varepsilon_t) = \rho^2 \sigma_u^2 + \sigma_\varepsilon^2$$

$$\Rightarrow \sigma_u^2 = \frac{\sigma_\varepsilon^2}{1-\rho^2}.$$

Therefore, if we added the first observation without multiplying it by $\sqrt{1-\rho^2}$, the variance of the disturbance term would be equal to σ_u^2, while in the other observations, which have been subject to the Cochrane-Orcutt transformation, this variance is equal to σ_ε^2.

Common factor test:

We have shown that for a linear regression model with the disturbance term subject to AR(1) process, the model can be transformed into ADL(1,1) with non-linear restriction imposed for estimation purposes, i.e.

Restricted model (transformed AR(1)): $Y_t = \beta_1(1-\rho) + \rho Y_{t-1} + \beta_2 X_{2t} - \beta_2 \rho X_{2t-1} + \beta_3 X_{3t} - \beta_3 \rho X_{3t-1} + \varepsilon_t$

This model can be considered as a special case of a more general model involving the same variables, i.e.

Unrestricted ADL(1,1) model: $Y_t = \lambda_1 + \lambda_2 Y_{t-1} + \lambda_3 X_{2t} + \lambda_4 X_{2t-1} + \lambda_5 X_{3t} + \lambda_6 X_{3t-1} + \varepsilon_t$

Restrictions embodied in the AR(1) process: $\begin{cases} \lambda_4 = -\lambda_2 \cdot \lambda_3 \\ \lambda_6 = -\lambda_2 \cdot \lambda_5 \end{cases}$

The common factor test is used to differentiate between these 2 cases. It involves a comparison of

RSS_R and RSS_U, the residual sums of squares in the restricted and unrestricted specifications.

H_0 : Restricted model should be used: $\begin{cases}\lambda_4 = -\lambda_2 \cdot \lambda_3 \\ \lambda_6 = -\lambda_2 \cdot \lambda_5\end{cases}$ are valid;

H_1 : Unrestricted ADL(1,1) model should be used.

Under the null hypothesis that the restrictions are valid, the test statistic has a χ^2 (chi-square) distribution with degrees of freedom equal to the number of restrictions (2 in this case).Note that it is a large-sample test.

Test statistics: $\chi^2 = n \log \frac{RSS_R}{RSS_U} \underset{asympt}{\overset{H_0}{\sim}} \chi^2(d.f. = number\ of\ restrictions)$, where

n is the actual number of observations in the regression (after adjusting endpoints);

log is the natural logarithm.

Dynamic model specification:

Initially, in order not to deal with a poorly specified model for which tests can appear to be satisfactory, even though it is misspecified (high risk of making Type II error), one should adopt a general-to-specific approach. It means that one should start with a model that is sufficiently general to avoid potential problems of under specification, and then see if it is possible to simplify it by testing restrictions on parameters. For example, in our case we should make the following steps for models with dynamic specification:

1) Estimate the model with all the lagged variables (unrestricted ADL(1,1):
$$Y_t = \lambda_0 + \lambda_1 Y_{t-1} + \lambda_2 X_{2t} + \lambda_3 X_{2t-1} + \lambda_4 X_{3t} + \lambda_5 X_{3t-1} + \varepsilon_t;$$
2) Test whether the lagged variables individually and as a group do not have significant explanatory power, i.e. $\lambda_1 = \lambda_3 = \lambda_5 = 0$. If it is not rejected, then the model can be simplified to the static case: $Y_t = \lambda_0 + \lambda_2 X_{2t} + \lambda_4 X_{3t} + \varepsilon_t$;
3) If the lagged variables do have significant explanatory power, we could perform a common factor test and see if we could simplify the model to an AR(1) specification, i.e. $\begin{cases}\lambda_3 = -\lambda_1 \cdot \lambda_2 \\ \lambda_5 = -\lambda_1 \cdot \lambda_4\end{cases}$. If the test shows that unrestricted version of ADL(1,1) should be used, then perform described tests detecting autocorrelation;
4) Test for a model with a lagged dependent variable, i.e. $\lambda_3 = \lambda_5 = 0$.

Lecture 20. Nonstationary Time Series.

The values of economic indicators representing time series data are very often related in time, and the Assumption B.2 of the Cross Section Data Model with Stochastic Explanatory Variables (stating that the consecutive values of each explanatory variable are drawn independently from the same distribution) is unrealistic. In case of Time Series implementation of B.2 Assumption would imply **strong stationarity,** supplemented by zero population covariance between any terms in time. Instead, we assume weak (or covariance) stationarity and weak persistency. Stationarity (strong stationarity) of a stochastic process implies that the X_t's are identically distributed and any correlation between adjacent terms is the same across all periods.

A series is **weakly stationary** if its expected value, variance and covariance between any two terms of the series do not depend on time (the covariance can depend on the distance between the terms though). From now and on by stationarity we will mean weak stationarity. If for a stationary process $\text{cov}(X_t, X_{t+s}) \to 0$ as $s \to \infty$, the process is called weakly dependent (or weakly persistent).

Some examples of time series processes:

- $X_t = \beta_1 X_{t-1} + \varepsilon_t$ – autoregressive process of the order 1, AR(1)
- $X_t = \beta_0 + \beta_1 X_{t-1} + \varepsilon_t$ – AR(1) with a constant
- $X_t = X_{t-1} + \varepsilon_t$ - random walk
- $X_t = \beta_0 + X_{t-1} + \varepsilon_t$ - random walk with drift (β_0 is called a drift)
- $X_t = \beta_0 + \beta_1 t + \varepsilon_t$ – deterministic trend
- $X_t = \beta_0 + \beta_1 X_{t-1} + \ldots + \beta_m X_{t-m} + \varepsilon_t$ - autoregressive process of the order m, AR(m)
- $X_t = \beta_0 + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_k \varepsilon_{t-k}$ – moving average of the order k, MA(k)
- $X_t = \beta_0 + \alpha_1 X_{t-1} + \ldots + \alpha_m X_{t-m} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_k \varepsilon_{t-k}$ - ARMA(m,k)

Checking for stationarity: some examples

AR(1) $X_t = \beta_1 X_{t-1} + \varepsilon_t$

Assume first that $|\beta_1| < 1$. Then:

$X_t = \beta_0 + \beta_1 X_{t-1} + \varepsilon_t = \beta_0 + \beta_1 (\beta_0 + \beta_1 X_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \beta_0 + \beta_1\beta_0 + \beta_1^2 X_{t-2} + \beta_1\varepsilon_{t-1} + \varepsilon_t = \ldots = \beta_0 +$

$+ \beta_1 \beta_0 + \beta_1^2 \beta_0 + \ldots + \beta_1^{t-1} \beta_0 + \beta_1^t X_0 + \beta_1^{t-1} \varepsilon_1 + \ldots + \beta_1 \varepsilon_{t-1} + \varepsilon_t$ = [sum up the constants

using geometric series formula] = $\frac{\beta_0(1-\beta_1^t)}{1-\beta_1} + + \beta_1^t X_0 + \beta_1^{t-1} \varepsilon_1 + \ldots + \beta_1 \varepsilon_{t-1} + \varepsilon_t$

Expected value:

$E(X_t) = E(\frac{\beta_0(1-\beta_1^t)}{1-\beta_1} + \beta_1^t X_0 + \beta_1^{t-1} \varepsilon_1 + \ldots + \beta_1 \varepsilon_{t-1} + \varepsilon_t) = \frac{\beta_0(1-\beta_1^t)}{1-\beta_1} + \beta_1^t E(X_0) + \beta_1^{t-1} E(\varepsilon_1)$

$+ \ldots + \beta_1 E(\varepsilon_{t-1}) + E(\varepsilon_t)$=[by Gauss-Markov condition 1, $E(\varepsilon_t)=0 \quad \forall t$]= $\frac{\beta_0(1-\beta_1^t)}{1-\beta_1} + \beta_1^t E(X_0)$

Variance:

$\text{var}(X_t) = var(\frac{\beta_0(1-\beta_1^t)}{1-\beta_1} + \beta_1^t X_0 + \beta_1^{t-1} \varepsilon_1 + \ldots + \beta_1 \varepsilon_{t-1} + \varepsilon_t) = var(\beta_1^t X_0 + \beta_1^{t-1} \varepsilon_1 + \ldots + \beta_1$

$\varepsilon_{t-1} + \varepsilon_t) = \beta_1^{2t} var(X_0) + \beta_1^{2(t-1)} var(\varepsilon_1) + \ldots + \beta_1^2 var(\varepsilon_{t-1}) + var(\varepsilon_t) + 2\beta_1^{2t-1} cov(X_0, \varepsilon_1) + \ldots$

$+2\beta_1 cov(\varepsilon_{t-1}, \varepsilon_t)$ =[Gauss-Markov 2 assumption of homoscedasticity]= $\beta_1^{2t} \text{var}(X_0) +$

$\sigma^2 (\beta_1^{2(t-1)} + \ldots + \beta_1^2 + 1) + 2\beta_1^{2t-1} \text{cov}(X_0, \varepsilon_1) + \ldots + \quad +2\beta_1 cov(\varepsilon_{t-1}, \varepsilon_t)$ = [By Gauss-

Markov conditions 3-4, ε_t is serially uncorrelated and non-correlated with X_0] = $\beta_1^{2t} \text{var}(X_0) +$

$\sigma^2 (\beta_1^{2(t-1)} + \ldots + \beta_1^2 + 1)$ =[sum up the constants using geometric progression formula] =

$\beta_1^{2t} \text{var}(X_0) + \sigma^2 \frac{\beta_0(1-\beta_1^{2t})}{1-\beta_1^2}$

Covariance:

$X_{t+s} = \beta_0 + \beta_1 X_{t+s-1} + \varepsilon_{t+s} = \beta_0 + \beta_1\beta_0 + \ldots + \beta_1^{s-1} \beta_0 + \beta_1^s X_t + \beta_1^{s-1} \varepsilon_{t+1} + \ldots + \beta_1 \varepsilon_{t+s-1} + \varepsilon_{t+s} =$

$= \frac{\beta_0(1-\beta_1^s)}{1-\beta_1} + \beta_1^s X_t + \beta_1^{s-1} \varepsilon_{t+1} + \ldots + \beta_1 \varepsilon_{t+s-1} + \varepsilon_{t+s}$

$\text{cov}(X_t, X_{t+s}) = \text{cov}(X_t, \frac{\beta_0(1-\beta_1^s)}{1-\beta_1} + \beta_1^s X_t + \beta_1^{s-1}\varepsilon_{t+1} + \ldots + \beta_1\varepsilon_{t+s-1} + \varepsilon_{t+s}) = \text{cov}(X_t, \beta_1^s X_t +$

$\beta_1^{s-1}\varepsilon_{t+1} + \ldots + \beta_1\varepsilon_{t+s-1} + \varepsilon_{t+s}) =$ [by Gauss-Markov conditions 3-4, all cross terms disappear] $=$

$\beta_1^s \text{var}(X_t) = \beta_1^{2t+s}\text{var}(X_0) + \sigma^2\frac{\beta_1^s\beta_0(1-\beta_1^{2t})}{1-\beta_1^2}$

We can see that for finite samples the expected value, population variance and covariances depend on time *t*. However, if $|\beta_1| < 1$, the AR(1) time series is asymptotically stationary. Since in that case $\lim_{t\to\infty}\beta_1^t = 0$, then for $t \to \infty$:

$E(X_t) \to \frac{\beta_0}{1-\beta_1^2}$; $\quad \text{var}(X_t) \to \frac{\sigma^2}{1-\beta_1^2}$; $\quad \text{cov}(X_t, X_{t+s}) \to \sigma^2\frac{\beta_1^s\beta_0}{1-\beta_1^2}$.

neither of three limits depend on time (though the poputation covariance of X_t and X_{t+s} depends on *s*), and in such a case we say that the series is **asymptotically stationary**. It is also weakly dependent (or weakly persistent) since $\frac{\beta_1^s\beta_0}{1-\beta_1^2} \to 0$ *as* $s \to \infty$.

Conclusion: any AR(1) with $|\beta_1| < 1$, and disturbance term satisfying Gauss-Markov conditions, is asymptotically stationary.

Random walk

In the derivations above we used the assumption that $|\beta_1| < 1$.Thus we have excluded random walk (the case with $\beta_1=1$) from our analysis. As this type of process is widely observed for time series economic variables, we will consider it separately, the general case – including drift β_0 :

$X_t = \beta_0 + X_{t-1} + \varepsilon_t = \beta_0 + \beta_0 + X_{t-2} + \varepsilon_{t-1} + \varepsilon_t = \ldots = \beta_0 t + X_0 + \varepsilon_1 + \ldots + \varepsilon_t$

Expected value:

$E(X_t) = E(\beta_0 t + X_0 + \varepsilon_1 + \ldots + \varepsilon_t) = \beta_0 t + E(X_0)$. The drift refers to the systematic change in expected value. As you can see, the expected value includes time trend $\beta_0 t$. If the drift is positive (negative), the trend is upward (downward). If there is no drift ($\beta_0 = 0$), then $E(X_t)=E(X_0)$, and it does not depend on time.

Variance

$var(X_t) = var(\beta_0 t + X_0 + \varepsilon_1 + \ldots + \varepsilon_t) = var(X_0) + t\sigma^2$ - note, the variance of the series grows with time, proportionally to t.

Covariance

$cov(X_t, X_{t+s}) = cov(X_t, \beta_0 s + X_t + \varepsilon_{t+1} + \ldots + \varepsilon_{t+s}) = var(X_t) = var(X_0) + t\sigma^2$, - the pattern of growth with time is the same as for variance.

Both the formulas for the population variance and covariance of the random walk are the same for the cases of presence or absence of the constant (drift).

Conclusion: random walk (with or without a drift) is nonstationary for finite samples, and also asymptotically nonstationary.

Deterministic trend: $X_t = \beta_0 + \beta_1 t + \varepsilon_t$

The series is named this way to differentiate it from the trend in a random walk with a drift.

Expected value: $E(X_t) = \beta_0 + \beta_1 t$. Since it depends on t, the series is non-stationary. The expected value includes the time-proportional component like the random walk with a drift, but there is one important detail. In deterministic trend the observations deviate around the trend line, and if the disturbance term ε_t is stationary (white noise type), should not go far from the trend. In contrast, the observations of random walk may or may not return to the trend line, due to rising variance.

Some more facts

Some types of time series models have been considered above. For some other types, (1-2) can be proven as an exercise. Properties (3) and (4) are taken for granted in the course.

1. Necessary condition for AR(k) stationarity: $\sum_{i=1}^{k} \beta_i < 1$
2. Sufficient condition for AR(k) stationarity: $\sum_{i=1}^{k} |\beta_i| < 1$
3. Any MA series is stationary

4. Stationarity of any ARMA depends on its AR part only

Types of nonstationary series

One can distinguish among the nonstationary time series two important types: trend stationary and difference stationary series.

Trend stationary time series are such that can be made stationary by detrending (removal of trend), using the following algorithm:.

1. Regress the series on time : $\hat{X}_t = a + bt$
2. Remove the trend: $\widetilde{X}_t = X_t - \hat{X}_t = X_t - a - bt.$

If X_t is a trend-stationary series then $\widetilde{X}_t$ is a stationary series.

If you are constructing a model in which at least one of the series is trend nonstationary, it is possible to make it (them) stationary by just adding time (and a constant, if it was initially absent) to the regressors' set.

Difference stationary series are such that can be made stationary by differencing. To illustrate, let X_t be a nonstationary series. If there exists k such that k-th difference of the series is stationary, then the series is difference stationary, or integrated, of the order *k* (denoted *I(k)*). For example, if a series X_t is nonstationary, but its first difference $\Delta X_t = X_t - X_{t-1}$ is stationary, then X_t is integrated of the order 1, I(1). If a series X_t is nonstationary, and its first difference $\Delta X_t = X_t - X_{t-1}$ is nonstationary as well, but its second difference ($\Delta^2 X_t = \Delta X_t - \Delta X_{t-1}$) is stationary, then X_t is integrated of the order 2, *I(2)*, and so on.

Most nonstationary time series of economic indicators are integrated of the order 1, or sometimes of the order 2, not more.

Useful implications are:

1. For any (non-negative) n, n-th order difference (Δ^n) of a stationary series is stationary.
2. For any (non-negative) n greater or equal to k, n-th order difference of an I(k) series is stationary.

The main problem of estimation using non-stationary time series is a danger of getting Spurious Regressions. These models demonstrate apparently significant regression coefficients in cases with no relationship between the variables involved. If to generate two independent time trends, or random walks, then regressing one on another, then the share of Type 1 errors would be much higher than just the significance level. Granger and Newbold (1974) in their well-known Monte-Carlo experiment did 100 estimations of regression of one random walk on another (generated each time independently from each other). They got significant slope coefficients 70 time under 1% significance level, and 77 time at the 5% level. Having modern software, everyone may easily reproduce the experiment getting similar results. An example of estimation of the model using EViews, for $X(0)=Y(0)=0$, where $X(t)=X(t-1)+\varepsilon_t$; $Y(t)=Y(t-1)+\mu_t$ (ε_t and μ_t are white noises with $\sigma^2_\varepsilon=\sigma^2_\mu=1$), is given below.

Dependent Variable: Y
Method: Least Squares
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.835350	0.382064	-7.421143	0.0000
X	0.437773	0.093798	4.667209	0.0000

R-squared	0.181853	Mean dependent var	-4.140088
Adjusted R-squared	0.173504	S.D. dependent var	2.864612
S.E. of regression	2.604271	Akaike info criterion	4.771980
Sum squared resid	664.6582	Schwarz criterion	4.824083
Log likelihood	-236.5990	F-statistic	21.78284
Durbin-Watson stat	0.152903	Prob(F-statistic)	0.000010

The slope coefficient seems to be significant under no actual relationship between *X* and *Y*. You should also pay attention to the low Durbin-Watson statistics which indicates here at the incorrect model specification used. If for the same data you estimate the regression of *Y(t)* on *Y(t-1), X(t)* and *X(t-1)*, then the result is:

Dependent Variable: Y
Method: Least Squares
Sample (adjusted): 2 100
Included observations: 99 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.180816	0.182112	-0.992881	0.3233
X	0.111751	0.105524	1.059013	0.2923
X(-1)	-0.084815	0.107120	-0.791776	0.4305
Y(-1)	0.935233	0.038683	24.17712	0.0000

R-squared	0.886959	Mean dependent var	-4.181907
Adjusted R-squared	0.883389	S.D. dependent var	2.848345
S.E. of regression	0.972663	Akaike info criterion	2.822007
Sum squared resid	89.87693	Schwarz criterion	2.926860
Log likelihood	-135.6893	F-statistic	248.4671
Durbin-Watson stat	1.916660	Prob(F-statistic)	0.000000

Here the result shows the correct specification estimation, and the coefficients expected to be insignificant are in fact insignificant, while the coefficient of *Y(t-1)* is significant and close to 1. Using Durbin h-statistic, you may discover also absence of the residual autocorrelation.

Though time series' stationarity does not guarantee non-appearance of spurious regressions (particularly asymptotic weak stationarity) in all cases, stationarity is usually considered as a necessary property of the time series in regression models.

Lecture 21. Testing for Nonstationarity

Graphical detection of nonstationarity.

Before considering the analytical tools for detecting nonstationarity, some (exploratory) graphical analysis can be done. Sometimes a graph can give the necessary information indicating the directions of analytical work.

Autocorrelation function ρ (k) presents the theoretical correlations between values of the series at time t (X_t) and its values at time $t+k$ (X_{t+k}), i.e. $\rho(k) = corr(X_t, X_{t+k})$

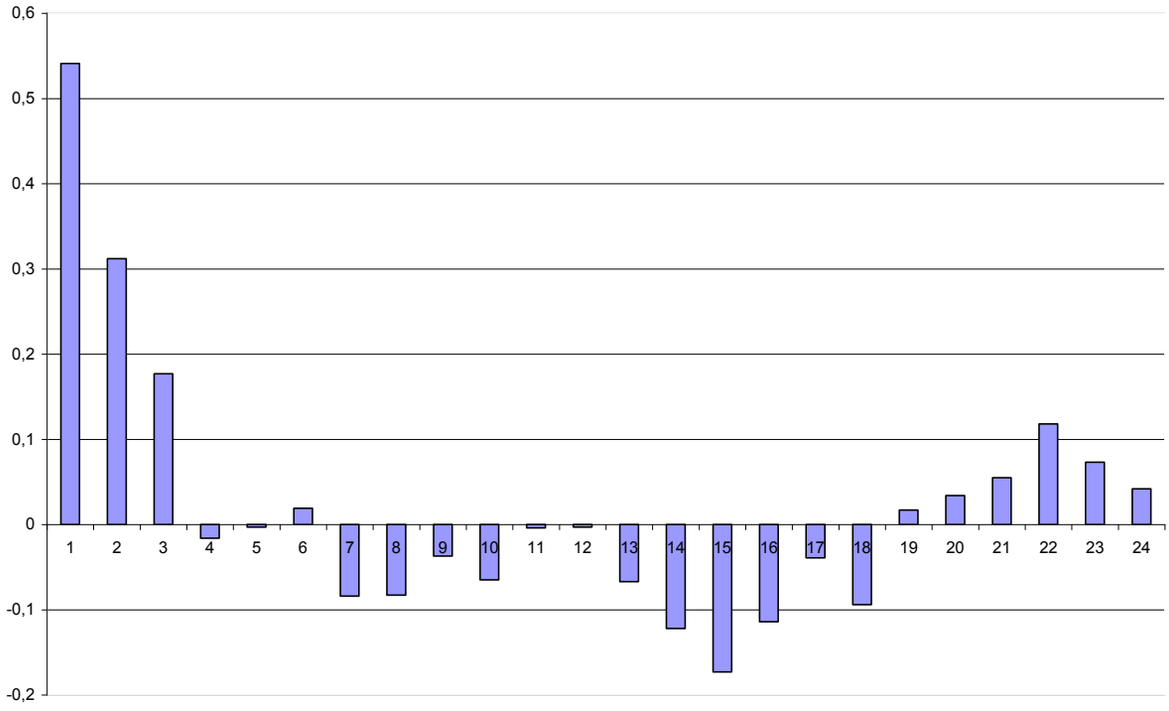
Graphical representation of Autocorrelation function is called a **correlogram**. Theoretical correlograms are not considered for nonstationary series, since their auctocorrelation functions also depend on time. Instead, one can consider expected value of the autocorrelation coefficients.

Properties of $\rho(k)$:

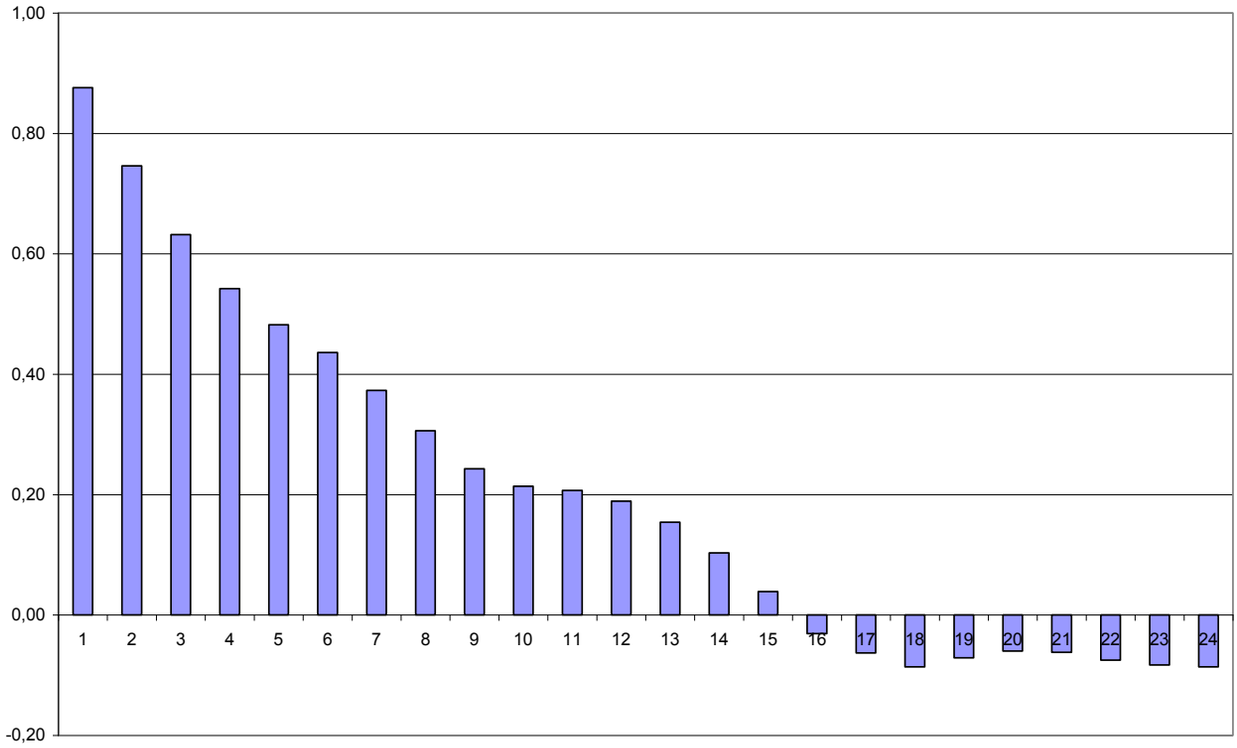
1. $\rho(0)=1$
2. Presence of a constant in the process does not influence its autocorrelation function, i.e. $X_t = \beta_0 + \beta_1 X_{t-1} + \varepsilon_t$ and $X_t = \beta_1 X_{t-1} + \varepsilon_t$ have the same autocorrelation functions.

By inspecting the **sample correlogram** of the series one can make some conclusions about its stationarity. There is no exact rule on how to determine whether the series is stationary or not with the use of a correlogram, but there are some helpful facts:

1. Correlogram of a stationary AR process declines to 0 exponentially. Let the process be $X_t = \beta_1 X_{t-1} + \varepsilon_t$. Then its autocorrelation function is $\rho_k = \beta_1^k$.

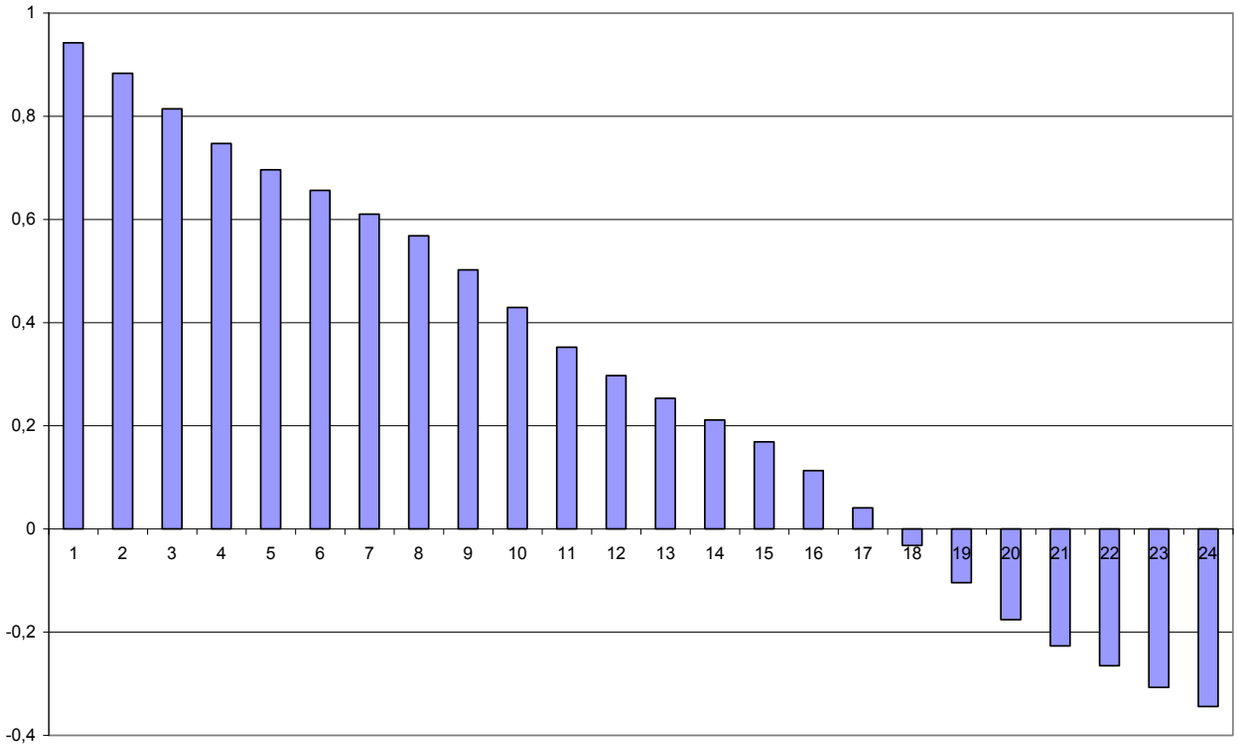


Correlogram of $X_t = 0.6\, X_{t-1} + \varepsilon_t$ ($X_0=10$; $\sigma^2_\varepsilon=1$; $T=50$).



Correlogram of $X_t = 0.9\, X_{t-1} + \varepsilon_t$ ($X_0=10$; $\sigma^2_\varepsilon=1$; $T=50$).

Compare it with the correlogram for Random Walk (the last one declines more slowly, and after getting zero does not stay around it):



Correlogram of $X_t = X_{t-1} + \varepsilon_t$ ($X_0=10$; $\sigma^2_\varepsilon=1$; $N=50$).

Correlogram of *MA(m)* process sharply falls to 0 for $k = m+1$ and is 0 for all greater values of k.

Combining these facts, we get the following:

- If the series is stationary, its autocorrelation function declines exponentially to zero
- If the series is not stationary (e.g. random walk), its autocorrelation function may decline more slowly

Unfortunately, there are some problems with the criteria mentioned above:

- Some stationary processes could have a slowly declining correlogram that looks similar to that of random walk (e.g. AR(1) with the slope coefficient close to 1)
- Correlogram of a nonstationary process can decline quickly if there are not many observations.

Formal testing for Nonstationarity

In order to test the presence of trend nonstationarity, you may regress the series on time and test the significance of time variable with usual t-test.

In order to test random walk type nonstationarity, use **Dickey-Fuller** test. Consider the model:

$$X_t = \beta_0 + \beta_1 X_{t-1} + \varepsilon_t$$

It was already shown that the series is nonstationary for finite samples, and it is asymptotically stationary if $|\beta_1|<1$. Since explosive processes are out of our consideration, we assume that $|\beta_1|$ cannot exceed 1.

The procedure may be presented as a three-step one:

Step 1. Transform, the model.

Let $\theta = \beta_1 - 1$. Subtract X_{t-1} from both sides: $X_t - X_{t-1} = \beta_0 + \beta_1 X_{t-1} - X_{t-1} + \varepsilon_t$

As $\Delta X_t = X_t - X_{t-1}$, we get $\Delta X_t = \beta_0 + \theta X_{t-1} + \varepsilon_t$

Step 2. Run the regression according to the transformed model: $\Delta \hat{X}_t = \hat{\beta}_0 + \hat{\theta} X_t$

Step 3. Test the hypothesis using t-test.

H_0: $\theta = 0$ (which corresponds to $\beta_1 = 1$). Thus, H_0 means nonstationarity.

Note that as the null hypothesis is nonstationarity, the critical values of t-statistic in the test are different from usual critical values. EViews gives all the necessary critical values in the printout.

Augmented Dickey-Fuller test.

We have considered autoregressive model with one lag, but it is possible to consider processes with more lags, e.g.: $X_t = \beta_0 + \beta_1 X_{t-1} + \ldots + \beta_k X_{t-k} + \varepsilon_t$ - $AR(k)$, or $X_t = \beta_0 + \sum_{i=1}^{k} \beta_i X_{t-i} + \varepsilon_t$

In this case, necessary condition for the series to be asymptotically stationary is that the sum of all lag coefficients is (in absolute value) less than 1, i.e. $|\sum_{i=1}^{k} \beta_i| < 1$ (we do not consider the proof

of this). To test the hypothesis of nonstationarity (H_0: $|\sum_{i=1}^{k}\beta_i|=1$) , apply **augmented Dickey-Fuller** test. The special case AR(2), when $X_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \varepsilon_t$.

The necessary condition for stationarity is $|\beta_1 + \beta_2| < 1$. The null hypothesis of nonstationarity is (H_0: $\beta_1 + \beta_2 = 1$). To do the test, we do the following:

First, transform the model: introduce a new variable θ: $\theta = \beta_1 + \beta_2 - 1$, and $\beta_1 = \theta - \beta_2 + 1$.

Substitute the above expression into the model: $X_t = \beta_0 + (\theta - \beta_2 + 1)X_{t-1} + \beta_2 X_{t-2} + \varepsilon_t$.

Regroup the terms: $X_t - X_{t-1} = \beta_0 + \theta X_{t-1} - \beta_2 X_{t-1} + \beta_2 X_{t-2} + \varepsilon_t$,

i.e. $\Delta X_t = \beta_0 + \theta X_{t-1} - \beta_2 \Delta X_{t-1} + \varepsilon_t$

Second, estimate the transformed model.

Finally, **test** the estimate of θ for significance (H_0: $\theta=0$).

The intercept and the time trend may be included, or not included in the model tested (appropriate options are given by the EViews).

The general case with arbitrary number of lags AR(k):

Transformation of the model. Let $\theta = \sum_{i=1}^{k}\beta_i - 1$, then $\beta_1 = \theta - \sum_{i=2}^{k}\beta_i + 1$.

Substitute it into the model: $X_t = \beta_0 + (\theta - \sum_{i=2}^{k}\beta_i + 1)X_{t-1} + \sum_{j=2}^{k}\beta_j X_{t-j} + \varepsilon_t$, and

$X_t - X_{t-1} = \beta_0 + \theta X_{t-1} - X_{t-1}\sum_{i=2}^{k}\beta_i + \sum_{j=2}^{k}\beta_j X_{t-j} + \varepsilon_t$. After the transformations of the model similar to the case k=2, we get:

$$\Delta X_t = \beta_0 + \theta X_{t-1} - \Delta X_{t-1}\sum_{i=2}^{k}\beta_i - \Delta X_{t-2}\sum_{i=3}^{k}\beta_i - \Delta X_{t-3}\sum_{i=4}^{k}\beta_i + \ldots + \Delta X_{t-k+1}\beta_k + \varepsilon_t$$

Number of differences in the right side equals the number of lags in the initial model minus 1. Then you run the regression for the transformed model, and test the hypothesis H_0: $\theta = 0$ (which corresponds to $\sum_{i=1}^{k}\beta_i = 1$). Again note, that as the null hypothesis is nonstationarity, the critical values of t in the test are not the same as the usual critical values.

Again, the intercept and the time trend may be included or not included in the model.

Finding out the number of lags in the process

There exist some procedures, which help to estimate the number of lags in the model. Two Information Criteria, Akaike (AIC) and Schwarz (SIC, also known as Bayes Information

Criteria, BIC), are the most popular for this. EViews also offers Hannan-Quinn Information Criterion (HQ), and some special modifications of the all three.

Akaike Information Criterion (AIC)

A statistic is calculated as: $\text{AIC}= \log\left(\frac{RSS}{T}\right)+\frac{2k}{T}$,

where k is the number of parameters estimated, and T is the number of observations. The first term decreases under the increase of the number of lags, while the second one (the penalty term) increases. The number of lags which minimizes the statistic is recommended.

Schwarz Information criterion (SIC)

The statistic is calculated as: $\text{SIC}= \log\left(\frac{RSS}{T}\right)+\frac{k}{T}\log(T)$, with the same notation. Again, the number of lags which minimizes the statistic is chosen.

Since the penalty term is greater in SIC (for T>7), it recommends smaller number of lags in general. For finite samples neither of two criteria has definite advantage, and both may be applied in practice. At the same time, for large samples SIC provides a consistent estimate of the number of lags.

Applying the ADF test in the EViews, you may set the number of lags exogenously, or apply one of the Information Criteria.

Conclusion

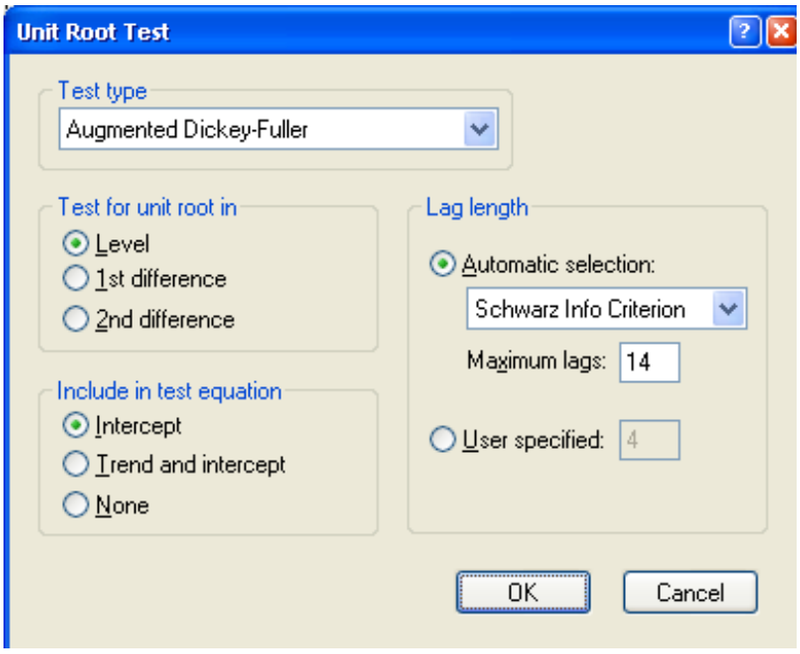
If you run a regression with nonstationary time series, you may get spurious regression. So, before running a regression always check for the presence of nonstationarity. And if you find that all or some of your series are nonstationary, you will need to make some model and/or data transformation which enable you to escape the danger. It will be shown in the next chapter.

EViews estimations

To make a detrending, add @trend to the regressor list. There is no need to generate time variable explicitly, since EViews already has a build-in trend function. E.g. the command [ls y c x @trend] will let us to estimate the equation $\hat{Y}_t = a + b\,X_t + \delta t$.

To carry out Augmented Dickey-Fuller test:

1. Open the series window.
2. View $\rightarrow$ Unit root test. A window will appear:



Lag length:

You can either specify the number of lags yourself by clicking option button «User specified» or let the software determine the number of lags by clicking «Automatic selection».

If you specify the model yourself, just type the number of lags in the textbox.

If you choose «Automatic selection», EViews will offer you one of the six info criteria mentioned earlier in this chapter. On the picture Schwarz (or Bayes) info criterion is selected.

Include in test equation:

If you press "Intercept", your model will be:

$$X_t = \beta_0 + \beta_1 X_{t-1} + \ldots + \beta_k X_{t-k} + \varepsilon_t$$

If you press "Trend and intercept", your model will be:

$$X_t = \beta_0 + \beta_1 X_{t-1} + \ldots + \beta_k X_{t-k} + \delta t + \varepsilon_t$$

EViews will test the series for both – trend and random walk stationarity.

If you press "None", your model will be:

$$X_t = \beta_1 X_{t-1} + \ldots + \beta_k X_{t-k} + \varepsilon_t$$

Test for the unit root may be done for the levels of the variable, as well as for the first or the second differences of it:

If you choose "level", EViews will test the series X_t for nonstationarity

If you choose "1st (2nd) difference", EViews will carry out the test for the 1st (2nd) difference of the series.

Lecture 22. Cointegration. Modelling with Nonstationary Time Series.

Cointegration

Let $X_1, \ldots, X_n$ be integrated series. In general, if you take a linear combination of these series, the order of integration of the linear combination will be equal to the maximum order of integration among the series. To illustrate, if series X_1 and X_2 are integrated of the order 1 and 2 respectively, then their linear combination will be integrated of the order 2. This seems to imply that a linear combination of series with the same order of integration k will be integrated of the order k. **However,** this is not always the case, and if the series have some long-run relationship, the order of integration of their linear combination can be lower.

Two series (with the same order of integration k≥1) are called **cointegrated** if there exists their stationary linear combination.

We will consider a particular case of cointegration of *I(1)* series which are rather typical for economic data. Two or more *I(1)* series are cointegrated if there exists their *I(0)*, i.e. **stationary**, linear combination. Even though each of them has a Random Walk type behaviour, they, nevertheless, stay rather close to each other in the long run, which means an existence of actual (not spurious) long run relationship.

Example

A logarithmic regression of expenditure on housing on *DPI* and the relative price of housing was estimated for the USA, 1959-2003:

$$\widehat{LGHOUS} = 0.006 + 1.03\ LGDPI - 0.48\ LGPRHOUS \qquad R^2 = 0.999.$$
$$(0.17)\ \ (0.007) \qquad\qquad (0.04)$$

The residuals for this regression are presented on the graph below.

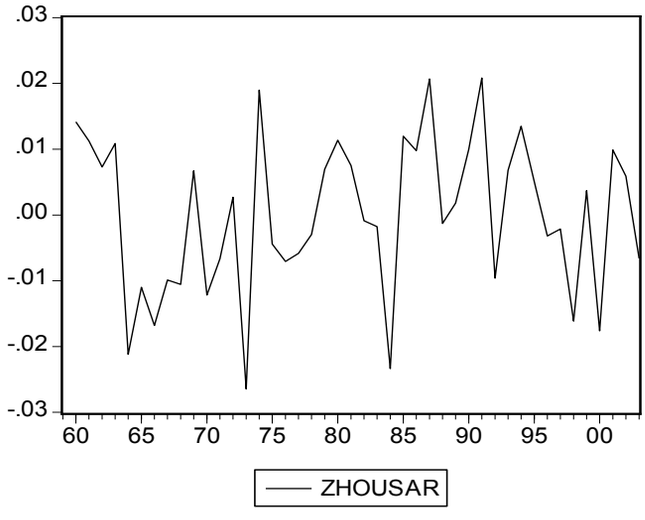
The residuals' behaviour looks more or less stationary. The ADF test statistic (intercept, no trend) is –2.91, while the asymptotic critical value for two explanatory variables is -3.34 (there are critical values for cointegrating relationships which are even higher in absolute value than those for ordinary ADF tests). So, it is not significant even at the 5 percent level, and we cannot reject a hypothesis of nonstationarity for the residuals. This would mean that the variables *LGHOUS*, *LGDPI* and *LGPRHOUS* are not cointegrated, but it is likely due to the low power of the test. The estimated coefficient of the lagged residuals is –0.33 which corresponds to the AR(1) process with ρ equal to about 0.67. It is quite possible that the residuals are stationary but autocorrelated with high ρ coefficient, and the variables are in fact cointegrated.

Durbin d-statistic equal to 0.63 indicates that it is possible that the disturbance term is subject to positive autocorrelation, with the cointegrating relationship being correct.

If the AR(1) term is added to the initial model, then we get

$$\widehat{LGHOUS} = 0.155 + 1.01\ LGDPI - 0.48\ LGPRHOUS \qquad R^2 = 0.999.$$
$$(0.35)\ \ (0.02) \qquad\qquad (0.09)$$

With the estimate of ρ equal to 0.72 (s.e. 0.16); d=1.82. All the coefficients estimated are close to those received before. New residuals look stationary:



and the ADF test statistic for them is -6.3. Though we cannot do the formal test without knowing the critical level for this particular case, which is different from the standard one, it seems that cointegrating relationship has been found.

Once more, cointegrating relationship represents a long-run link between the variables.

Fitting Models with Nonstationary Time Series

Since estimating models with nonstationary time series may rather often lead to spurious regressions, the idea is to transform the model in such a way that the series in it become stationary. We will consider three approaches: Detrending, Differencing and Error Correction Models.

Detrending

In the models with variables which include time trends, removal of the trends, or detrending, allows to avoid getting spurious regressions. On the detrending procedure, see the Lecture 19. As indicated, detrending of each variable in the model is equivalent to including the time trend as one of the explanatory variables. The coefficients would be the same for these two cases while the standard errors slightly differ (they are correct when just time variable is included).

Economic indicators rather often behave not as series including time trends, but as random walks. If you detrend a series which is in fact a random walk with a drift, then its variance still increases in time proportionally to the variable *t*, the series does not become stationary, and hence the problem of spurious regressions is not resolved.

Differencing

If having random walk time series, differencing is a procedure which can be applied:

subtracting $Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1}$ from $Y_t = \beta_1 + \beta_2 X_t + u_t$, we get

$\Delta Y_t = \beta_2 \Delta X_t + \Delta u_t = \beta_2 \Delta X_t + (\rho - 1) u_{t-1} + \varepsilon_t.$

The series ΔY_t and ΔX_t are stationary, and the coefficient β_2 can be estimated from this model. At the same time, the new disturbance term Δu_t is subject to autocorrelation, and appropriate remedial measures should be applied. Only in the case of severe autocorrelation of u_t in the initial model (ρ is close to 1) differencing also helps to reduce the autocorrelation influence. If Y_t and X_t are unrelated I(1) processes, absence of their relationship will be revealed in the differenced model, so the problem of spurious regressions will be resolved.

At the same time, there are a few more shortcomings in the differenced model: the constant disappears (though it is usually of low interest) and only short-run relationships can be investigated with it. In the long-run equilibrium $\Delta Y = \Delta X = 0$, and hence no conclusions about

long-run relationship can be made. The approach for including the long-run relationships is the Error correction model.

Error correction model

Error correction model involves transforming the original model with nonstationary time series in such a way that all the series in the transformed model are stationary, and at the same time it includes the description of both short-run and long-run relationship between the variables. But every model development has its price: here this is an assumption about the particular form of relationship between the variables, the ADL(1,1) one.

So, let the model be ADL(1,1), and X_t and Y_t are both I (1) series:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_t + \beta_3 X_{t-1} + \varepsilon_t$$

Assume that in the long-run, all variables reach their steady states. This means that

$$X_t = X_{t-1} = \overline{X}$$

$$Y_t = Y_{t-1} = \overline{Y}$$

So, in the long-run, the equation looks like this:

$$\overline{Y} = \beta_0 + \beta_1 \overline{Y} + \beta_2 \overline{X} + \beta_3 \overline{X}$$

$$\overline{Y}\,(1 - \beta_1) = \beta_0 + (\beta_2 + \beta_3)\,\overline{X}$$

$$\overline{Y} = \frac{\beta_0}{1-\beta_1} + \frac{\beta_2 + \beta_3}{1-\beta_1}\,\overline{X}$$ - long-run relationship between X and Y

Now it is assumed that the long-run relationship is the cointegrating relationship. This implies that in the expression

$$Y_t = \frac{\beta_0}{1-\beta_1} + \frac{\beta_2 + \beta_3}{1-\beta_1} X_t + \upsilon_t$$

the disturbance term υ_t is **stationary.**

The model can be transformed as follows:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_t + \beta_3 X_{t-1} + \varepsilon_t$$

Subtract Y_{t-1} from both sides:

$$Y_t - Y_{t-1} = \beta_0 + \beta_1 Y_{t-1} - Y_{t-1} + \beta_2 X_t + \beta_3 X_{t-1} + \varepsilon_t$$

$$\Delta Y_t = \beta_0 + (\beta_1 - 1) Y_{t-1} + \beta_2 X_t + \beta_3 X_{t-1} + \varepsilon_t$$

Add and subtract $\beta_2 X_{t-1}$:

$$\Delta Y_t = \beta_0 + (\beta_1 - 1) Y_{t-1} + \beta_2 X_t - \beta_2 X_{t-1} + \beta_3 X_{t-1} + \beta_2 X_{t-1} + \varepsilon_t$$

$$\Delta Y_t = \beta_0 + (\beta_1 - 1)\, Y_{t-1} + \beta_2 \Delta X_t + (\beta_3 + \beta_2) X_{t-1} + \varepsilon_t$$

Take $(\beta_1 - 1)$ out of the brackets

$$\Delta Y_t = (\beta_1 - 1)(Y_{t-1} - \frac{\beta_0}{1-\beta_1} - \frac{\beta_2 + \beta_3}{1-\beta_1} X_{t-1}) + \beta_2 \Delta X_t + \varepsilon_t$$

Thus, ΔY_t and ΔX_t are stationary, as the original series are I(1) and the expression in the brackets is also stationary, since it is the cointegrating relationship for *(t-1)* time unit.

So, all the series in the transformed model are stationary and spurious regression is no longer a threat. The only problem is that the cointegrating relationship parameters are unknown. To deal with this, **Engle-Granger two-step procedure** is used in practice:

Step1. Estimate the cointegrating relationship with OLS.

Step 2. Using the estimated relationship, fit the error correction model.

Engle and Granger showed that the results of the two-step procedure are asymptotically the same as in the case when the true cointegrating relationship is used.

Let us consider the model be ADL(1,1) with two explanatory variables, X_t and Z_t, and Y_t as the dependent variable, all are *I(1)* series:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_t + \beta_3 X_{t-1} + \beta_4 Z_t + \beta_5 Z_{t-1} + \varepsilon_t$$

After the same transformations as before, we have

$$\Delta Y_t = (\beta_1 - 1)(Y_{t-1} - \frac{\beta_0}{1-\beta_1} - \frac{\beta_2 + \beta_3}{1-\beta_1} X_{t-1} - \frac{\beta_4 + \beta_5}{1-\beta_1} Z_{t-1}) + \beta_2 \Delta X_t + \beta_4 \Delta Z_t + \varepsilon_t$$

We have already estimated the cointegrating relationship for the logarithmic function of demand for housing (USA, 1959-2003) above, and now we estimate the error correction model (EViews printout below):

Dependent Variable: DLGHOUS
Method: Least Squares
Sample (adjusted): 1960 2003

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ZHOUS(-1)	-0.311355	0.111123	-2.801888	0.0077
DLGDPI	0.938132	0.048823	19.21503	0.0000
DLGPRHOUS	-0.498342	0.122453	-4.069678	0.0002
R-squared	0.249635	Durbin-Watson stat		1.626460

Here ZHOUS(-1) is the lagged residual of the cointegrating relationship. The estimation shows that about 0.31 of the short-run deviation from the equilibrium is covered each year, and the estimates of the short-run income and price elasticities of demand for housing are 0.94 and -0.50 which are rather close to the estimates of the long-run elasticities (1.01 and -0.48).

Lecture 23. Panel Data Models.

Introduction to panel data analysis

Panel (or longitudinal) data set contains observations on the same units for several periods of time. Units can be individuals, countries, households, etc. Panel data sets are very typical in economic and social analysis, - like macroeconomic indicators for the regions of Russia in different time periods, or expenditures structure of households in different time units, or exam grades of ICEF students in different exam sessions, etc. In many countries there are regular surveys like Russian Longitudinal Monitoring Survey (RLMS), or National Longitudinal Survey of Youth (NLSY) in the USA, etc., which provide reliable panel data sets for all kinds of research and analysis.

Panel data set is **balanced** if every unit is surveyed in every time period and **unbalanced** otherwise. **The RLMS and NLSY present** unbalanced panel data, since some respondents can move, refuse, die, etc.

Benefits of having panel data

1. **Larger data set**. If in time series data T observations are available, with panel data T*n observations are available, where n is the number of units and T is the number of periods.
2. **Unobserved heterogeneity** problem (more about it later) is eliminated or mitigated.
3. **Dynamics** can be explored better (compared to cross-section). Although with cross-section one could investigate dynamics by asking retrospective questions, it is not very reliable, as people forget details over time.
4. Panel data are often of **higher quality**. For example, national surveys are usually rather well designed and well organised.

Panel Data Models

When dealing with the panel data, we may assume the following type of the DGP (data generating process):

$$Y_{i\,t} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{jit} + \sum_{p=1}^{s} \gamma_p Z_{pi} + \delta t + \varepsilon_{i\,t} \qquad (1)$$

where Xs are observed and Zs are unobserved variables, i – unit index, t – time index, j, p – summation indices for observed and unobserved variables respectively. The term δt allows

intercept to shift over time. It can be substituted for dummies, representing corresponding periods if the constant change assumption is too strong.

Note: unobserved variables are assumed to be static, i.e. do not change over time, thus Z has no time index. As there is no information on Z, we can rename the variables:

$$\sum_{p=1}^{s} \gamma_p Z_{pi} = \alpha_i$$

α_i is referred to as **unobserved heterogeneity term**. It is also assumed that the disturbance term $\varepsilon_{i\,t}$ satisfies all the Gauss-Markov conditions, in particular that $\text{cov}(\alpha_i, \varepsilon_{j\,t}) = 0\ ,\ \forall\ i, j, t$

If α_i is ignored, by considering the model

$$Y_{i\,t} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{jit} + \delta t + \varepsilon_{i\,t}\ ,$$

then OLS estimators will suffer from omitted variable bias, i.e., will be biased and inconsistent if α_i is correlated with any of the regressors. Thus, under such circumstances, α_i has to be accounted for somehow, and the Panel Data models allow to do it. Note that it is impossible for the cross section data.

Fixed effect model

The so-called Fixed Effect approach allows to transform the model (1) in such a way that the unobserved heterogeneity term is removed. Three different fixed effect type methods of transformation and estimation of the model (1) will be considered: First Differences Method, Within Groups Method, and Least Squares Dummy Variables (LSDV) Method.

First differences method is as follows:

We lag the model (1) by one period:

$$Y_{i\,t\text{-}1} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ijt-1} + \alpha_i + \delta(t\text{-}1) + \varepsilon_{i\,t\text{-}1}\ ,$$

And then subtract $Y_{i\,t\text{-}1}$ from $Y_{i\,t}$:

$$Y_{i\,t} - Y_{i\,t\text{-}1} = \sum_{j=1}^{k} \beta_j (X_{ijt} - X_{ijt-1}) + \delta + \varepsilon_{i\,t} - \varepsilon_{i\,t\text{-}1}\text{, or}$$

$$\Delta Y_{i\,t} = \sum_{j=1}^{k} \beta_j \Delta X_{ijt} + \delta + \varepsilon_{i\,t} - \varepsilon_{i\,t\text{-}1}$$

Thus, the unobserved heterogeneity term α_i disappears. However, autocorrelation of the type MA(1) arises. In general, it leads to inefficiency of the OLS estimators and invalid test

statistics, though consistency is provided. Only in the special case when the original disturbance term was subject to AR(1) autocorrelation with coefficient close to 1, then the disturbance term in the transformed model will not be autocorrelated.

Least squares dummy variables (LSDV) method

A natural way to account for α_i is to introduce a set of dummies for units. The model looks as follows:

$$Y_{it} = \sum_{j=1}^{k} \beta_j X_{jit} + \sum_{i=1}^{n} A_i D_i + \delta t + \varepsilon_{it}$$

where D_i is the dummy variable equal to 1 for i-th unit and zero otherwise. Thus, there are as many dummy variables as there are units, i.e. n. But note that such specification is possible only if intercept is omitted, otherwise one falls into the dummy trap. Alternatively, you can choose a reference category and keep the intercept.

Within groups method

This method allows to cancel the unobserved heterogeneity term by using deviations of the variables from their mean values. To apply it, for the model

$$(1)\ Y_{it} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it},$$

consider the average Y_i in time:

$$(2)\ \bar{Y}_i = \beta_0 + \sum_{j=1}^{k} \beta_j \bar{X}_{ij} + \alpha_i + \delta \bar{t} + \bar{\varepsilon}_i$$

As α_i is constant in time for each unit, it is not affected.

Them subtract (2) from (1):

(1)- (2)

$$Y_{it} - \bar{Y}_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it} - \beta_0 - \sum_{j=1}^{k} \beta_j \bar{X}_{ij} - \alpha_i - \delta \bar{t} - \bar{\varepsilon}_i$$

$$Y_{it} - \bar{Y}_i = \sum_{j=1}^{k} \beta_j \left(X_{ijt} - \bar{X}_{ij}\right) + \delta(t - \bar{t}) + \varepsilon_{it} - \bar{\varepsilon}_i$$

Thus the unobserved heterogeneity term α_i disappears, as well as the intercept β_0. This method is called «within groups» one because the variations of the dependent variable around its mean are regressed on the variations of explanatory variables around their means.

Dummy variables vs Within groups

It can be shown mathematically that within-groups method is equivalent to LSDV method. Thus, the two methods always give identical estimates. If the number of units is large, it is not convenient to introduce dummy variables. For this reason in practice within-groups method is used.

The only thing that can be unclear is degrees of freedom. In LSDV method there are n*T-k-n degrees of freedom (n*T observations, k regressors, and n dummies). At first glance, it looks as if within-groups method has n*T-k degrees of freedom. Nevertheless, transformation of the model consumes n degrees of freedom through the calculationa of averages, so the number of the degrees of freedom is, as expected, the same for both methods.

So the Fixed Effect Model allows to take into account some unobservable heterogeneity. At the same time, the approach has essential drawbacks.

The drawbacks of the Fixed Effect model are:

1. All variables that are constant in time (though different for different units) disappear and we cannot determine to what extent they influence the dependent variable.
2. Variation of the new explanatory variables is (most likely) smaller than in the original specification. Thus, the precision of the estimates of the coefficients decreases. If measurement error bias took place, it would be aggravated.
3. n (number of units) degrees of freedom are lost as a result of the model transformation or of adding extra dummies.

Random effect model

If α_i is not correlated with regressors then it can be made a part of the disturbance term, and the new disturbance term will be uncorrelated with the regressors too. If so, omitting the individual effect will **not** make OLS estimators unbiased and inconsistent, while doing so will enable to save the degrees of freedom and keep regressors which are constant in time. The method allowing to do this is called a Random Effect Method.

Assumptions on the α_i in the Random Effect Model:

- α_i is taken randomly from a fixed distribution
- α_i is independent from the regressors

Thus, it is possible to move fixed part of α_i to the constant and the rest – to the disturbance term. Hence, without loss of generality, we will assume that E(α_i) = 0. Then:

$$Y_{it} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{jit} + \delta t + u_{it} ,$$

where $u_{it} = \alpha_i + \varepsilon_{it}$.

Assume that ε_{it} satisfies all Gauss-Markov conditions.

Let us check whether the new disturbance term u_{it} satisfies Gauss-Markov conditions:

1. $E(u_{it}) = E(\alpha_i + \varepsilon_{it}) = E(\alpha_i) + E(\varepsilon_{it}) = 0 + 0 = 0$
2. $var(u_{it}) = var(\alpha_i + \varepsilon_{it}) = var(\alpha_i) + var(\varepsilon_{it}) + 2cov(\alpha_i, \varepsilon_{it}) =$[assume $cov(\alpha_i, \varepsilon_{it}) = 0$] $= var(\alpha_i) + var(\varepsilon_{it}) = \sigma_\alpha^2 + \sigma_\varepsilon^2$
3. $cov(u_{it}, u_{it-s}) = cov(\alpha_i + \varepsilon_{it}, \alpha_i + \varepsilon_{it-s}) =$[assume ε_{it} is not serially correlated] $= var(\alpha_i) = \sigma_\alpha^2$
4. $cov(X_{ijt}, u_{it}) = cov(X_{ijt}, \alpha_i + \varepsilon_{it}) = cov(X_{ijt}, \alpha_i) + cov(X_{ijt}, \varepsilon_{it}) = 0$

Thus, the Gauss-Markov condition 3 is violated. The disturbance term is autocorrelated for each unit *i,* with a particular type of autocorrelation. To overcome this problem, the so-called **Feasible Generalised Least Squares (FGLS)** method is used. This method takes autocorrelation into account and produces the best linear unbiased and consistent estimators, possible with the available information.

Random vs Fixed

Random effect model is superior to Fixed effect model for two reasons:

1. n degrees of freedom are not lost in the Random effect model
2. observations that are constant in time are not dropped

So, it is better to use Random Effect than Fixed Effect model. However, if at least one of the assumptions of Random Effect model does not hold, we have to use Fixed Effect model. Thus, the question is whether the necessary assumptions hold. As for the first assumption (randomly drawn observations), it is supposed to be guaranteed by the survey. The tests for checking the assumption on independence of the new disturbance term with the explanatory variables, are described in the section “Tests”.

Pooled regression

It can be the case that there is no unobserved heterogeneity at all, i.e.

$$\alpha_i = 0, \ \forall \ i$$

If so, the sample can be pooled (unified sample for all units and periods). This will give two benefits comparing to the Random effect model since there is no need to allow for non-existing individual effects:

- it allows to get not just unbiased and consistent but also efficient estimates, with valid tests, without special procedures like FGLS, and these properties are provided for finite samples;
- no loss of efficiency due to no testing of irrelevant constraints or effects.

Panel data summary

True situation	Fixed Effect Model	Random Effect Model	Pooled Regression
$\alpha_i = 0$ for any i	• unbiased • consistent • inefficient	• unbiased • consistent • inefficient	• unbiased • consistent • efficient
$\text{cov}(\alpha_i, X_j) = 0$ for any i,j	• unbiased • consistent • inefficient	• unbiased • consistent • efficient	• unbiased • consistent • inefficient
$\text{cov}(\alpha_i, X_j) \neq 0$ for some i,j	• unbiased • consistent • efficient	• biased • inconsistent	• biased • inconsistent

Tests

In this section tests relevant to panel data models are described. There are two main questions that one can ask on the model to use:

1. Should we use Random effect or Fixed effect model?
2. Should we use Pooled regression or Random Effect model (if Random effect is preferred to Fixed effect in the p.1)?

For the p.1, one can use the Durbin-Wu-Hausman (DWH) test, while Breush-Pagan Lagrange Multiplier test can be used for the p.2.

Durbin-Wu-Hausman test

This test is applied in many cases other than panel data, e.g. to detect the endogeneity or measurement errors problems. Its purpose is to test whether there is significant difference between the estimates of coefficients, obtained by different methods, one set always consistent while another consistent only under particular circumstances (some hypothesis H_0 implementation).

$$Y_{i\,t} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{i\,t}$$

To do the test, we have to consider first whether all α_i can be considered as random variables taken from the same distribution. Under H_0 (which means that α_i are not correlated with X_j for any *i,j*) both Fixed Effect and Random Effect models provide us with consistent estimates. However Fixed Effect model estimates will be inefficient since it involves estimating an unnecessary set of coefficients, so **Random Effect** model should be used if H_0 is not rejected.

If we reject H_0 in favour of H_a (meaning thet α_i and X_j are correlated for some *i,j*), the Random Effect estimators would be biased and inconsistent, while **Fixed Effect** estimators are consistent, and hence the **Fixed Effect model should be used** though having the drawbacks indicated above.

Under H_0 the DWH test statistic has a chi-square (χ^2) distribution. Its calculations involve matrix algebra. The number of degrees of freedom usually equals the number of coefficients compared, but can also be lower in some special cases.

Random Effect model or Pooled Regression?

In order to decide if we should use the panel data models at all, we have to test if there is unobserved heterogeneity. For this, Breush-Pagan Lagrange multiplier test could be used. The null hypothesis H_0 for it means that $\alpha_i = 0$ for any *i*. The test statistic under H_0 has χ^2 distribution with one degree of freedom.

If we do not reject the H_0 then Pooled regression should be estimated, and it provides unbiased, consistent and efficient estimates. If H_0 is rejected then we apply the Random Effect model.