# EFR summary

## Introduction to Econometrics, FEB12012X 2024-2025

## Lectures 1 to 16

Weeks 1 to 7

Deloitte.

DeNederlandscheBank EUROSYSTEEM

financials LAUDAME

## Details

**Subject:** Introduction to Econometrics IBEB 2024-2025

**Teacher:** T. Bago d'Uva, T. Van Ourti, P. Hans Franses

**Date of publication:** 18.04.2025

# Introduction to Econometrics – IBEB – Lecture 1, week 1

## Methods

### Everyday vs Scientific learning

**Everyday learning**
Learning through tradition or from experts requires little effort, but these sources can be wrong. On the other hand, via one's own experience, one may learn the causal relationships between things, but there's a possibility of overgeneralization due to selective observations and drawing inaccurate observations.

**Scientific learning** answers the question "Is something true or not?" This involves

1. Extending existing knowledge, which can be tested by collecting data and performing analysis (scientific methods).

### Association versus Causal Effect

When there is an association between two variables, it does not necessarily imply causation.

Association can provide useful fact descriptions, while causal effects indicate the relations between variables and, thus, can be used to understand the effectiveness of policy intervention.

### Types of data and unit of analysis

#### Types of data

**Experimental data:** used to estimate the causal effects (e.g. treatment and control group)

**Observational data:** collected for general purposes and not designed to estimate causal effects

### Time dimensional

**Time series** information on a set of indicators over time (e.g. GDP over several years)

**Cross-section** is when a sample is observed and data collected at a specific point of time

**Panel data set** combines the last two types mentioned before, when cross-sectional study is carried out over time

### Unit of analysis

For different purposes the analysis will be built on different units:

- Individuals
- Firms
- Regions
- Countries

## Operationalization and conceptualization

**Conceptualization:** means specifying what is meant by the specific terms used in research.

**Operationalization:** the process of developing specific procedures to empirically represent the concepts defined during conceptualization. In other words, it is about measuring theoretical concepts.

### Quality of operationalization

**Reliability:** Measurement methods are reliable and if the concept was to be measured repeatedly the findings of the research would be the same.

**Validity** means that a measure accurately reflects the concept it is intended to measure.

# Introduction to Econometrics – IBEB – Lecture 2, week 1

## OLS: simple linear regression model

### Relationships between variables

One of the ways to find out about the relationship between variables can be by constructing a **scatter plot**. The relationship can be negative or positive, or there can be no relationship.

### Covariance and correlation

It is not always enough to just observe the relationship, thus when we want to quantify it we can make relevant computations.

**Sample covariance**: $s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$

Where n is sample size, $X_i$ is the value of X for observation i (similarly $Y_i$) and $\overline{X}$ is the sample average of X (similarly $\overline{Y}$). The units of sample covariance = units X * units Y.

**Sample correlation**: $r_{XY} = \frac{s_{XY}}{s_X * s_Y}$

Where $s_{XY}$ is the sample covariance and $s_X$ is the sample standard deviation of X (similarly $s_Y$). A correlation of 0 reflects no correlation, a correlation of +1 reflects perfectly positive correlation and -1 reflects a perfectly negative correlation. The correlation coefficient is unitless. It shows the strength of the relationship between X&Y.

## The linear regression model

Linear regression attempts to formulate a causal effect of one variable (x) over another (y) which is unlike a mere two-sided association of correlation.

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

The error term $u_i$ represents all other factors influencing Y and measures a vertical distance between the population regression line and an observation. $\beta_1$ is the slope of the regression line and $\beta_0$ is the intercept.

## The line of best fit

The best fit line that fittingly depicts the relationship between the X and Y variables has to minimize the sum of the squared differences between observed data points and the population regression line. The differences are squared in order to prevent the positive and negative residuals from negating each other.

To minimize $\sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2$

- OLS estimator $\widehat{\beta_0}$: $\widehat{\beta_0} = \overline{Y} - \beta_1 \overline{X}$
- OLS estimator $\widehat{\beta_1}$: $\widehat{\beta_1} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{s_{XY}}{s_X^2}$

Note: unit of the coefficient of X can be stated as unit of Y by unit of X.

- OLS predicted/fitted values: $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i$
- Residuals: $\widehat{u_i} = Y_i - \widehat{Y_i}$

## Comparing correlation with linear regression model

The linear regression model is a very flexible framework that allows several directions for extensions, such as:

- Multiple X variables: having more than one independent variable simultaneously influencing Y. (multiple regression)
- Nonlinear relationships
- Discrete or binary variable

While correlation coefficient is unitless, the OLS estimator of $\beta_1$ is measured in $\frac{units\ Y}{units\ X}$. Linear regression model coefficient shows causality only under OLS assumptions, and if those do not hold it shows association and should not be used for policy design.

# Goodness of fit measures

## The $R^2$

Observed value equal: $Y_i = \widehat{Y_i} + \widehat{u_i}$, in which $\widehat{Y_i}$ is explained by the model fitted value and $\widehat{u_i}$ is unexplained residuals.

- **Total sum of squares (TSS**) is the total variation in the data: $TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$
- **Explained sum of squares (ESS)**: $ESS = \sum_{i=1}^{n} \left(\widehat{Y_i} - \overline{Y}\right)^2$

It shows the variation in the data explained by the model

- Finally, the $R^2$ is the proportion of sample variance of $Y_i$ that is explained by $X_i$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}\left(\widehat{Y_i} - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2};\ R^2 = corr\left(Y_i, \widehat{Y_i}\right)^2$$

In the case of single explanatory variable $R^2 = corr\left(Y_i, X_i\right)^2$

Range of $R^2$: $0 \leq R^2 \leq 1$

$R^2$=1: model predicts $Y_i$ perfectly

$R^2$=0: model $(X_i)$ does not predict any variance in $Y_i$

## Standard Error of the Regression (SER)

The SER shows the spread of the data points around the population regression line. Larger values indicate stronger deviation from predicted values.

$$SER = S_{\hat{u}} = \sqrt{S^2_{\hat{u}}}$$

Where $S^2_{\hat{u}}$ is the sample variance of the residuals $\hat{u}_i$

$$S^2_{\hat{u}} = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{SSR}{n-2}$$

# Introduction to Econometrics – IBEB – Lecture 3, week 1

## OLS assumptions

1. Zero conditional mean
2. The observations are independently and identically distributed
3. Large outliers are unlikely

### Assumption 1: Zero conditional mean

The **zero conditional mean assumption** implies that the expected values of the residual value given a value of X is zero.

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

The expected value of the residual is independent of X. This means that the correlation between the residual and X is zero. Thus the explanatory variable is uncorrelated with other factors that influence Y.

To know and assert whether this condition holds, we must be sure of the random assignment of the variable X. If X is not randomly assigned it is difficult to confirm the validity of the zero-conditional mean assumption.

If there is no random assignment to satisfy the OLS model, we need to suppose that X is uncorrelated with other factors, that is when X is 'as good as random'. In order to measure the pure causal effect of X on Y, the uncorrelated assumption is important. Otherwise, there would be an omitted variable bias and the pure effect would not be accounted for.

With **simultaneous causality** that is when variables influence each other, the zero-conditional mean will not hold.

## Assumption 2: Independently and identically distributed observations

Independent and identical distribution holds in the case of simple random sampling from the same population. The distribution will be identical when the observations are obtained from the same population, and the observations are uncorrelated and thus independent.

When the sample is not representative as well as for time series and panel data this assumption does not hold.

## Assumption 3: Large outliers in X and Y are unlikely

OLS is very susceptible to the influence of outliers, and thus "finite kurtosis" is an essential assumption. Mathematically, this is defined as:

$$0 < E(X_i^4) < \infty; 0 < E(Y_i^4) < \infty$$

If there are data errors it is best to eliminate large outliers by fixing or removing those data points. Fixing or dropping the data should only happen if it is suspected to be an error.

Sometimes, certain outliers can have dramatic effects on the population regression line hence it is desirable to be skeptical of extreme points.

## Sampling distribution of OLS estimators

The estimators of the constant ($\widehat{\beta_0}$) and coefficient of X ($\widehat{\beta_1}$) of the linear regression models are computed from random samples and thus are random variables themselves with a probability distribution.

As different samples can lead to different estimates, the estimators are just some points in the sampling distribution of the estimator.
If one uses all possible samples of size n from a population and applies OLS to estimate the coefficients, one will realize that large samples of the $\beta_1$ estimator ($\widehat{\beta_1}$) approximate to a normal distribution.

This comes directly from the central limit theorem. As the coefficient of X is independent and identically distributed, the expected value is the true value of the coefficient of X. This implies that the OLS estimator is unbiased:

$$E(\widehat{\beta_1}) = \beta_1$$
$$E(\widehat{\beta_0}) = \beta_0$$

Note: With large samples OLS estimators follow approximately normal distribution. Therefore, normality assumption is not needed.

## Property of OLS estimators

### Unbiasedness

When the estimators are unbiased. Therefore, the mean sampling distribution $\widehat{\beta_1}$ equals $\beta_1$ and similar for $\widehat{\beta_0}$

Unbiasedness of $\widehat{\beta_1}$ is satisfied if assumptions 1 and 2 hold. When the number of observations (sample size) increases the estimator of the coefficient of X becomes more consistent and converges towards the true value.

## Variance of estimators and consistency

Because of the central limit theorem in large samples, $\widehat{\beta_0}$ and $\widehat{\beta_1}$ approximately follow a normal distribution $\widehat{\beta_1} \sim N\left(\beta_0; \sigma^2_{\widehat{\beta_1}}\right)$, and jointly they follow a bivariate normal distribution.

Variance of $\widehat{\beta_1}$: $\sigma^2_{\widehat{\beta_1}} = \frac{1}{n} \times \frac{var\left[\left(X_i - \mu_i\right)u_i\right]}{\left[var\left(X_i\right)\right]^2}$

The variance of $\widehat{\beta_1}$ decreases when the number of observations increases, when the variance of residual factors decreases, and when the variance of the explanatory variable X increases.

=> OLS estimator unbiased and consistent
=> The sampling distribution used are hypotheses tests and confidence intervals

# Interpretation

Conditional expectation of Y, given X for the population model $Y_i = \beta_0 + \beta_1 X_i + u_i$: $E\left(Y_i \mid X_i\right) = E\left(\beta_0 + \beta_1 X_i + u_i \mid X_i\right) = \beta_0 + \beta_1 X_i + E\left(u_i \mid X_i\right)$, which under Assumption 1 further simplifies to $E\left(Y_i \mid X_i\right) = \beta_0 + \beta_1 X_i$

**Most common interpretation** of $\beta_1$: when X goes up by 1, the expected value of Y given X goes up by $\beta_1 E(Y_i|X_i + \Delta X) = E(Y_i|X_i) + \Delta X \beta_1$

## Interpretation of the intercept

Generally $\beta_0$ indicates an average Y when $X_i = 0$

The intercept may not always be interpretable and it will depend on the data whether the interpretation will be meaningful.

### Example with binary regressors

Take on only two values (Male/Female, Yes/No, Agree/Disagree)
Dummy variable: D = 0,1

Population model: $Y_i = \beta_0 + \beta_1 D_i + u$

Conditional expectation: $E(Y_i|D_i) = \beta_0 + \beta_1 D_i$

Average Y when D = 0: $E(Y_i|D_i) = \beta_0$

Average Y when D = 1: $E(Y_i|D_i) = \beta_0 + \beta_1$

**Interpretation:** $\beta_1$ is the difference between the average when D=1 and the average when D=0. $\beta_1$ is the change in average Y when D=1 compared to D=0.

# Introduction to Econometrics – Introduction to Econometrics – IBEB – Lecture 4, week 2

## Hypothesis tests and confidence intervals in linear regression

### Two-sided hypothesis test

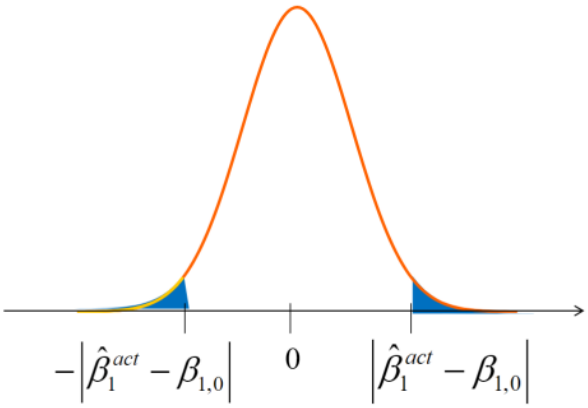
$$H_0: \beta_1 = \beta_{1,0} \quad vs \quad H_1: \beta_1 \neq \beta_{1,0}$$

Rejects null hypothesis if the estimated value $\hat{\beta}_1^{act}$ deviates substantially from the given hypothesized value $\beta_{1,0}$.

In other words, the null hypothesis is rejected if the probability of getting at least a value as extreme as the estimate $\hat{\beta}_1^{act}$ is very small (p-value)

## t statistic and p-value

**P-value**: probability of obtaining $\hat{\beta}_1$ which is even further away from hypothesized value $\beta_{1,0}$ than he obtained $\hat{\beta}_1^{act}$ (shown by the blue area).




Source: Lecture 4

t-statistic: $t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$

## Decision rule and Rejection region

Using the significance level of 5%:

Reject $H_0$ if

1. $P - value < 0.05$
2. $\left|t^{act}\right| > 1.96$ (critical value for a two-sided test)

# Confidence intervals

E.g: 95% confidence means that from all samples that can be drawn, the interval contains the true value of $\beta_1$ in 95% of the cases.

Confidence interval: $\left[\widehat{\beta_1} - 1.96 \times SE\left(\widehat{\beta_1}\right); \widehat{\beta_1} + 1.96 \times SE\left(\widehat{\beta_1}\right)\right]$

## Variance of error terms

When the variance of the error term is constant for all given values of X, then there is **homoskedasticity**. Otherwise, it is **heteroskedasticity**. When homoskedasticity holds, the formula of standard errors of $\widehat{\beta_1}$ can be simplified and the OLS estimator has minimal variance amongst all unbiased linear estimators (efficient).

However, homoskedasticity does not always hold. Therefore, it is important to use heteroskedasticity-robust standard errors.

## Significance

Statistical significance is decisive in whether to reject or not to reject the null hypothesis. Economic significance involves not only statistical significance, but also the economic effect implied by the data analysis and testing's result. Some statistical results may be significant but not economically meaningful. In this discussion of hypothesis testing for the linear regression coefficient, the key warning is that the size (the magnitude of the effect, i.e., $\widehat{\beta_1}$) matters.

# Introduction to Econometrics – IBEB – Lecture 5, week 2

## OLS: OVB, multiple linear regression, assumptions

To measure the causal effect of variable X on Y one would want the OLS estimator to be unbiased. If, however, another variable Z is correlated with X and it has a causal effect on Y too, then the coefficient estimated for X would not purely reflect the

causal effect of X on Y and would be a combination of effects, thus leading to a biased estimator.

If the correlation between the error term and variable X is not equal to zero (the zero-conditional mean assumption is violated), then the variable Z as mentioned previously would affect Y. In this case the error term is function Z and other factors: $u_i = \beta_2 \times Z + v_i$, where $v_i$ is the remaining of the error term, with everything else influencing Y except for X and Z.

When zero-conditional mean assumption holds: $corr(u_i, X_i) = 0 \Leftrightarrow \beta_2 \times corr(Z_i, X_i) = 0$ (assuming $corr(v_i, X_i) = 0$).

When the estimator of $\beta_1$ is biased we have that the expected value of $\widehat{\beta_1}$ equals $\beta_1$ plus the bias. This is summarized in the formula: $E(\widehat{\beta_1}) = \beta_1 + \beta_2 \frac{corr(X_1, X_2)}{s_{X1}} \times s_{X2}$

## Direction of bias

The bias, when a simple model includes X1 and omits X2, can be positive or negative depending on the sign of $\beta_2$ and correlation between X1 and X2.

| | $corr(X_1,X_2)>0$ | $corr(X_1,X_2)<0$ |
|---|---|---|
| $\beta_2>0$ | **Positive** bias | **Negative** bias |
| $\beta_2<0$ | **Negative** bias | **Positive** bias |

Source: Lecture 5

## Multiple regression model

By including the omitted variable into the model we try to satisfy the zero-conditional mean assumption. The omitted variable will no longer cause a correlation of error term with X1. The regression model thus can be finally expanded as follows:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + v_i$$

where the **main variable** of interest is X1, and X2 can be considered as the **control variable**. Since there is more than one coefficient that explains Y, it is a **multiple regression model**.

## Interpretation of multiple regression model

Population multiple regression model is similar to the model with a single regressor and represents the average relationship between the independent variables and Y. The interpretation is the change in Y due to the change in X1 when X2 held constant. For example, if X2 is held constant, and X1 goes up by 1 then Y on average goes up by $\beta_1$. The OLS estimators, predicted values and residuals are obtained similarly to a model with a single regressor.

## Assumptions for the multiple regression model

Similar assumptions to the one discussed for linear regression.

1. **Zero-conditional mean**: $E\left(u_i|X_{1i}, X_{2i},..., X_{ki}\right) = 0$

If we are interested in causal effect of all X1, X2...Xk, we use a weaker assumption

1. **Conditional mean independence**: $E\left(u_i|X_{1i}, X_{2i},..., X_{ki}\right) = E\left(u_i|X_{2i},..., X_{ki}\right)$

If interested in the causal effect of X1, we can make use of a weaker assumption that implies that the error term is independent of X1 and not all other variables. When conditional mean independence holds, one can interpret the effect of X1 on Y as a *causal effect*, and one can only interpret the effect of X2 on Y as a *partial association*. If the variable of interest is only X this is not a problem, we call X2 a control variable and X1 the variable of interest.

2. **Observations being independent and identically distributed**

3. **Large outliers of the variables are unlikely**

4. **No perfect multicollinearity**

Perfect collinearity between X1, X2 and X3 if there is a perfect linear relationship between 3 variables, such that $X_1 = a + bX_2 + cX_3$ with $b \neq 0 \; and \; c \neq 0$. Perfect

collinearity between explanatory variables happens in such cases as having the same variable in different units, or a dummy variable trap.

In situations when there is linear conversion of variables like change of units, it makes no reasonable sense to include both the variables (as you would essentially include the same variable twice in different measurement units).

In the case of dummy variables, it is always advisable to drop out a dummy for one category. When interpreting models with one dummy dropped out, the coefficients are always interpreted relative to the dropped-out dummy (the base/reference category).

## Sampling distribution

We need the sampling distribution for both confidence intervals and hypothesis tests. Under the 3 assumptions of OLS and no perfect multicollinearity, the estimator coefficients of the independent variables individually follow a normal distribution and collectively follow a multivariate normal distribution.

Variance of $\hat{\beta}_j$ decreases with sample size n; decreases with variance of Xj; increases with variance of error term $u_i$; increases with correlation between X's (imperfect multicollinearity), however if assumption 1 holds then the model is still unbiased.

## Measures of fit

### Standard Error of the Regression (SER)

The SER, similarly to the simple regression model, shows the spread of the data points around the population regression line. Larger values indicate stronger deviation from predicted values.

$$SER = S_{\hat{u}} = \sqrt{S^2_{\hat{u}}}$$

Where $S^2_{\hat{u}}$ is the sample variance of the residuals $\hat{u}_i$

$$S^2_{\hat{u}} = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-k-1}$$

However the SSR (sum squared residuals) is now divided by n-k-1 (where k stands for the number of independent variables that influence Y) to derive variance.

### The $R^2$

Similarly to the simple regression model

$$TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

$$ESS = \sum_{i=1}^{n} \left(\widehat{Y_i} - \overline{Y}\right)^2$$

It shows the variation in the data explained by the model

Finally, the $R^2$ is the proportion of sample variance of $Y_i$ that is explained by $X_i$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

A special characteristic of $R^2$ is that it always increases when a regressor is added to the model. In order to deflate this sensitivity the adjusted R-squared formula is as follows: $\overline{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSR}{TSS}$

Finally, it is noteworthy that the measure of fit cannot be compared and used if the dependent variables differ in the ways they are defined. Additionally, the measure of fit only represents the explained variation, but does not account for biases and whether the assumptions even hold.

# Introduction to Econometrics – IBEB – Lecture 6, week 2

## OLS: hypothesis tests, confidence intervals and model specification

When analyzing regression models, one of the worst problems encountered is when one of the three assumptions is violated, as that makes the estimator biased. If one variable is stipulated as being correlated with the variable X1, then this variable

should be included in the model as a control variable since otherwise, the model could be subject to omitted variable bias.

After introduction of this variable X2 (control variable), the zero-conditional mean must hold such that the conditional mean of other factors given variable X1 and variable X2 is 0.

## Hypothesis test for a single coefficient

Hypothesis: $H_0: \beta_j = \beta_{j,0} \; vs \; H_1: \beta_j \neq \beta_{j,0}$

T-statistic: $t = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}$

P-value: $p - value = 2\phi(-|t^{act}|)$

Reject H0 at 5% sig.level: $|t^{act}| > 1.96$

## Test of joint hypotheses

Test of joint hypotheses can be used to specify the null hypothesis that the coefficients of various variables equal to a hypothesized value, which are q restriction and the alternative hypothesis that one or more of these q restrictions does not hold. In general form the hypotheses are formulated as follows:

$H_0: \beta_j = \beta_{j,0}, \beta_m = \beta_{m,0} \ldots (total\ of\ q\ restrictions)$

$H_1: one\ or\ more\ of\ the\ q\ restrictions\ does\ not\ hold$

One must use joint hypotheses testing instead of individual one because under the assumption, the coefficients have an approximate **bivariate normal distribution** in sufficiently large samples.

In case if one only knows the individual t test and not the F test then the **Bonferroni method** can be incorporated which uses special critical values to account appropriately for the significance level.

## F-statistic

$H_0: \beta_1 = 0 \; and \; \beta_2 = 0 \; vs \; H_1: \beta_1 \neq 0 \; and/or \; \beta_2 \neq 0$

The formula for F statistics can look different from previous courses as we do not assume homoskedasticity here.

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t1,t2} t_1 t_2}{1 - \hat{\rho}^2_{t1,t2}}\right)$$

where $t_1$ and $t_2$ are the t-statistic of separate tests and $\hat{\rho}_{t1,t2}$ is the estimator of the correlation of the two t-statistics (it will happen to be 0 when there is no correlation between X1 and X2).

The distribution of the F statistics in large samples follows F distribution with degrees of freedom q (number of restrictions) in the numerator and ∞ in the denominator: $F - statistic \sim F_{q,\infty}$

Reject H0: if $F > critical\ value\ of\ F_{q,\infty}$

Common critical values for $F_{2,\infty}$: 10% sig. level =2.30, 5% =3.00, 1%=4.61

When testing whether the coefficients have no effect on Y, that is when all coefficients except the constant are zero, the hypotheses can be stated as follows:

$$H_0: \beta_1 = 0, \beta_2 = 0 \ldots \beta_k = 0 \ vs \ H_1: \beta_j \neq 0, \ at\ least\ one\ j$$

When such a null hypothesis is rejected at a given significance level, it means that coefficients are jointly significant or jointly significantly different from zero.

## Omitted variable bias

Despite incorporating another variable X2 to prevent any bias, it is still very plausible that variables X1 and X2 (explanatory and control variables) do not satisfy the zero conditional mean assumption. In this case, one can adopt the weaker assumption of conditional independence.

This implies the correlation between variable X1 and other factors is 0. This will result in the effect of variable X1 to be purely causal, but the effect of variable X2 will display partial association, thus a mixture of effects of variable X2 and other factors.

If, however, even the conditional independence does not hold then the model has an omitted variable bias and more control variables can be introduced to the model. It is called a **robustness check** when one introduces changes into the model (like including new control variables) to see if the results would differ.

# Introduction to Econometrics – IBEB – Lecture 7, week 3

## Nonlinear regression functions

If the effect measured by the slope of the regression function depends on the value of the independent variable(s), we should have a nonlinear relationship.

It is always advisable to check whether a non-linear model improves the linear model by testing whether an additional regressors are significantly different from 0, furthermore, the graph can be used to observe the evenness of the spread of points and whether there is an improvement in the fit too.

There are a number of forms of non-linear models we can employ. Here we will cover:

Form 1: Polynomials
Form 2: Natural Logarithmic Transformation of the dependent and/or independent variable(s)
Form 3: Interaction Effects

### Polynomial regression models

**Polynomials** use a linear function of a variable, where the linear function contains the variable taken to the power.

For quadratic polynomials, when the coefficient in front of squared variable is positive it represents the increasing returns to scale and when that coefficient is negative we can see decreasing returns to scale.

### Testing

If the population regression function is considered linear, then the quadratic and higher-degree coefficients would not be useful in the regression functions. To test this, we can perform an F-test where the null hypothesis is the regression being linear and the additional regressors are equal to 0; the alternative hypothesis is that at least one of the additional regressors is not equal to 0.

## Natural logarithmic transformation of the variable

This method employs the same regression model but with a logarithmic transformation of variable Y.

**2 reasons**:

- Outliers in the right tail can be dealt with using this method. Large outliers lead to a violation of the third OLS assumption, and they are less likely to affect the model after this transformation when the large outliers are compressed.
- Used if one is interested in percentage changes.

### Log-linear model

Logarithmic transformation of dependent variable (Y) only

- Interpretation of $\beta_1$: a 1-unit change in X corresponds to $\left(\beta_1 \times 100\,\%\right)$ change in Y (semi-elasticity).

### Linear-log model

Logarithmic transformation of variable X only.

- Interpretation: $1\,\%$ change in X corresponds to $0.01 \times \beta_1$ units of Y .

### Log-log model

Both independent (X) and dependent (Y) variables are transformed logarithmically.
- Interpretation: a $1\,\%$ change in X corresponds to a $\beta_1\,\%$ change in Y. In this case, $\beta_1$ is called elasticity.

## Interaction effect

The example of a model that includes interaction effect:
$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u_i$

The inclusion of $\beta_3 X_1 X_2$ term accounts for the interaction effect. It is useful to add when we believe that the effect of a variable depends on another variable.

## When OLS fails

OLS fails when there's nonlinearities in the parameters. The previous models are nonlinear in X but are linear functions of the coefficients (parameters)

# Introduction to Econometrics – IBEB – Lecture 8, week 3

## Internal and external validity

Association is not causation and when there are any policy recommendations only causal effects should hold an important value.

A study is **internally valid** if the statistical inferences about the causal relationship are valid for the population and setting studied.

There are two sets of population and setting: one that is studied and one to which inferences can be generalized upon. The **population studied** is the one from which the sample was derived. The **population of interest** is one to which the inferences are generalized on. The **setting** is the institutional, legal, social and economic background of the study.

Threats to internal validity if these do not hold:

1. The estimator of the causal effect should be unbiased and consistent.

2. The hypothesis test should have the required significance level

A study is **externally valid** if the inferences can be used to make generic inferences to other populations and settings too.

## Threat 1: Omitted variable bias (OVB)

If there is a variable that is omitted and is correlated with the variable of interest, as well as being a determinant of the dependent variable, then there is an omitted variable bias.

If the correlation between the variable of interest and omitted variable has the same sign as the effect of omitted variable on dependent variable, then there is an **upward bias**. If however, the signs are opposite then there is a **downward bias.**

### Good and bad control variables

Control variables' values should always be generated before the variable of interest's are. Stated simply, if the variable of interest has a causal effect on the new (control) variable then the new variable is not a good control. This is not applicable the other way around (that is, if the control variable affects the variable of interest).

## Threat 2: Errors-in-variables

Independent variable

- Random measurement error (classical measurement error): Bias towards 0
- Non-random: downward or upward sloping bias

Dependent variable

- Random: No bias but reduced precision
- Non-random: downward or upward sloping bias

Solutions: instrumental variables regression, and developing a mathematical model of the measurement error and using the resulting formula for correction.

## Threat 3: Sample selection

Missing data at random leads to no bias. Missing data for the regressor also leads to no bias, however, the interpretation of the coefficient would then only hold for a subset of the population for which the observations are not missing.
The exception is when there is missing data on the dependent variable, then there is a bias. The solution to this issue is the use of appropriate sampling.

## Threat 4: Simultaneous causality

There is no problem in the case of having causality that runs from the regressor to the dependent variable. However, if the reverse also holds true, then there is a bias as OLS will include both directions of causality. The potential solutions are instrumental variables regression and the design of research (randomized control trial).

## Threat 5: Functional form misspecification

If there is a non-linear relationship but we adopt a linear model in some sense, there is an omitted variable bias. Therefore, it is best to test whether a significantly different from zero non-linear coefficient exists.

## Threat 6: Inconsistency in the standard error

To avoid inconsistency in the standard error, always adopt a heteroskedasticity robust standard error and ensure independent and identically distributed observations.

## Forecasting vs causal relationship

The goal of developing a causal model is deriving the best description of behavior. The first concern is internal validity, and the second concern is external validity of the model. In contrast, for forecasting models, the models' external validity is of greater importance than internal validity. To have the best forecast for the future, the requirements are good explanatory power, stability of results, and precision.

# Introduction to Econometrics – IBEB – Lecture 9, Week 4

## Restoring Internal Validity

### Sampling

**Definition**: A method to prevent threats to internal validity of the model which consists of a representative sample of a population under study.

Why?

- Avoids sample selection bias
- Random selection => i.i.d observations => consistent standard errors

**2 Essential Steps:**

1. Define population
2. Decide how to derive that sample so that it is representative of the defined population

**2 main Sampling Techniques:**

1. **Probability Sampling:**
    - Random selection
        - all members of the population have an equal chance of being selected in the sample
        - Observations are independent
    - Representative
        - Same distribution of characteristic as the population
    - Allows use of **Probability Theory**

2. **Non-Probability Sampling**: Opposite of the previous technique
    - Non-random selection
        - Members of the population do NOT have an equal chance of being selected
    - Representativeness is NOT guaranteed
    - Does NOT allow use of **Probability Theory**

# Panel Data

**Definition:** Observing the same individuals repeatedly at different points in time. The data set is **balanced** when the duration observed is the same for all individuals, and **unbalanced** when the duration varies across them.

Dealt with using a model that emphasizes the changes in two time periods.

- Building a regression model in each time period and then finding the difference between them.
- **Only the variables whose values change over time remain in the model** => enables us to study the effect of the independent variable's change over time on the dependent variable's change over time.
- Required assumption is much weaker than OLS conditional mean independence assumption: **time difference in time-varying regressors should be unrelated to time difference in errors**

**Advantages:**

- Changes in dependent and independent variables allows removing time-invariant & unobservable omitted variable bias

**Disadvantages:**

- Time-invariant variables drop out
- Cannot remove time-varying omitted variables bias
    - Time difference in errors should still be unrelated to time difference in time-varying regressors
- Coefficients are constant over time

# Instrumental Variables (IV)

**Intuition:**
Variations in the independent variable consist of 2 parts:

1. **Endogenous variations:** correlated with error term => OVB
2. **Exogenous variations:** independent of error term => NO OVB

The method of instrumental variables gets rid of OVB by isolating the exogenous variations from the endogenous ones.

**Two Stage Least Squares (TSLS)**:

- **1st Stage:** Predicts the independent variable with the help of an instrumental variable of choice.
- **2nd Stage:** Regresses the dependent variable on the predicted independent variable derived from the first stage.

Example: IV = schoolreform
1st Stage:
$educ = \Pi_0 + \Pi_1 schoolreform + v$
$\widehat{educ} = E(educ|schoolreform) = \widehat{\Pi_0} + \widehat{\Pi_1} schoolreform$

2nd Stage:
$ln(labinc) = \beta_0^{TSLS} + \beta_1^{TSLS} \widehat{educ} + e$

**Conditions for valid IV:**

1. **RELEVANCE**: IV should have some explanatory power for the endogenous variable.
   => The correlation between the instrument and the independent variable must not be equal to zero
   => Testable

2. **EXOGENEITY**: IV must be unrelated with error terms.
   => Correlation of the instrument with the error term must equal zero.
   => Not testable

**OLS estimate > TSLS estimate** because

<u>OLS Estimate</u>:

- NOT internally valid => upward bias
- Exploits ALL variation in dependent variable
- Externally valid for ALL variation in dependent variable

<u>IV Estimate</u>:

- Internally valid provided IV relevance & exogeneity
- Exploits minor share of total variation in dependent variable
- Externally valid for the variation in dependent variable explained by the IV

# Introduction to Econometrics – IBEB – Lecture 10, Week 4

## Quasi-experiments

1. **Experiment**
- Treatment assigned randomly
- On purpose
- Used in economics but more common in other disciplines (example: psychology)

2. **Quasi Experiment**
- Treatment assigned "as good as random"
- NOT on purpose
- More often used in economics

3. **Association (forecasting)**
- Treatment assigned non-randomly
- NOT on purpose
- Common in ALL disciplines

**Example**: Marshmallow Experiment
Description: Children (age 4 to 6) are led into a room with a marshmallow on a table. The child can eat the marshmallow or wait for 15 minutes after which a second marshmallow is rewarded.
Findings: A minority eats the marshmallow immediately. One third of the remaining group manages to wait for 15 minutes. Delaying gratification predicts academic success and literacy.
This is an **association** because there are NO treatment groups clearly defined and the action each child takes is not predetermined, thus NOT on purpose.

### Average treatment effect

At an individual level, you can NEVER estimate a causal effect but with (Quasi-) experiments, you can estimate the average causal effect.

**Average Treatment Effect (ATE) = observed difference + unobserved difference**

- You need a random sample that is large enough
- Random treatment assignment avoids selection bias

Example: Mortality Experience vs Mammography Screening

- Average treatment effect [y: mortality; $t$: treatment]:

$$E(y_{1i} - y_{0i})$$

$$= P(t = 1)E(y_{1i} - y_{i0}|t = 1) + P(t = 0)E(y_{1i} - y_{0i}|t = 0)$$

- We observe the mortality experience of those women that got a mammogram, but not their mortality experience if they would not have received a mammogram
- Similarly, we observe the mortality experience of women that did not get a mammogram, but not their mortality experience when they would have received mammogram

| | | Experimental variation. Did women actually get mammogram? | |
|---|---|---|---|
| | | Treated ($t = 1$) | Control ($t = 0$) |
| Potential outcome ($y_{ti}$) | Observed | $E(y_{1i}\|t_i = 1)$ | $E(y_{0i}\|t_i = 0)$ |
| | Unobserved | $E(y_{0i}\|t_i = 1)$ | $E(y_{1i}\|t_i = 0)$ |

**Difference Estimator:**

- The control and treatment group are the same before the treatment and there is quasi-random treatment assignment.
- Typically, it is not the case in the quasi-experiments that the pre-treatment groups are identical.

## Experimental Variation & Threats to Internal and External Validity

**Experimental variation** is usually at random, however **observational variation** is non-random.

- The basic concern regarding causal inference is that you cannot observe two occurrences on the same individual at the same time

    - i.e. an individual cannot simultaneously be in the control and treatment groups
    - This is why an individual causal effect cannot be measured but it is possible to measure the average treatment effect.

## Threats to Internal Validity

**Threat 1: Failure to Randomize**

- Non-systematic ad-hoc rules entailing characteristics of name, nationality etc. should not be used to randomize the subjects
- Should be done randomly so that control and treatment group are similar
- The F-test can be done to ensure that the experiment is randomized

**Threat 2: Failure to Follow Treatment Protocol**

- Partial Compliance: The failure to follow treatment protocol leading to lack of compliance by the subjects leading to violation of the conditional mean independence assumption.

Solution:

- Use random assignment as an instrumental variable
- If there is data on the random assignment, but the data on actual treatment is missing it is also possible to estimate the Intention To Treat (ITT).
- While both can be useful as both consider random assignment, IV shows the effect of receiving the treatment, but ITT shows the effect of being selected into the treatment group.

**Other Threats to internal validity**

**Attrition:** Exclusion of some subjects from the sample due to non-random reasons.
Example:
Harmless: move out of NL is unrelated to treatment
Harmful: exclude late-stage breast cancer

**Experimental Effects:** Hawthorne and placebo effects.

Solution: double blind

- Neither the researcher nor the subjects know who is in the treatment and who is in the control group.

### Threats to External Validity

**Non-representative Sample**

- Example: Experimenting in region with high breast cancer rate

**General Equilibrium Effects**

- The experiment affects the behavior of a larger subset than initially anticipated
- Example: increasing awareness about the issue studied

# Introduction to Econometrics – IBEB – Lecture 11, week 5

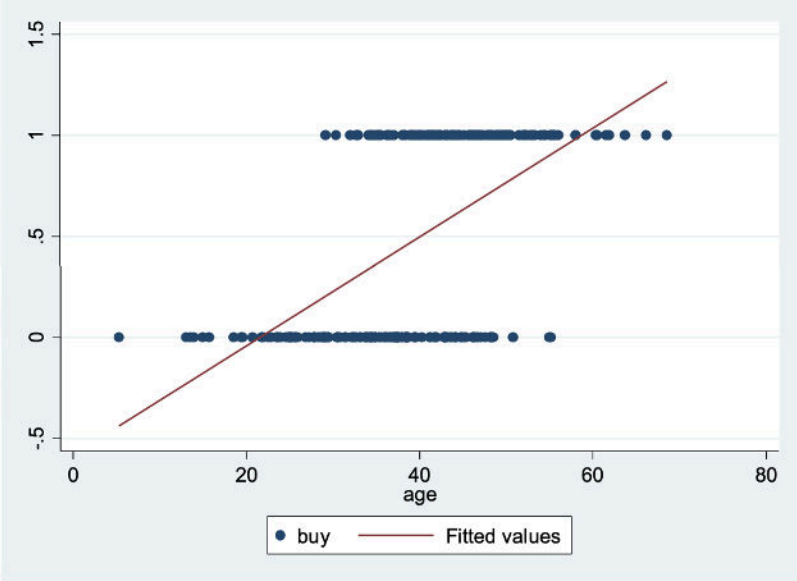
## Binary OLS

When a binary variable is used as the dependent variable, we use the numbers 0 and 1 to model the choices (usually 1 is affirmative/positive)

### Linear Probability Model (LPM)

#### Scatter plot LPM

- Abnormal: all points are clustered horizontally around 1 and 0
- It is of vital importance to incorporate heteroskedastic-robust estimates.
- The predicted dependent variable of this model reflects the probability of Y=1.
- The $\beta_1$ reflects the change in probability Pr(Y=1) that occurs corresponding to a unit change in variable X, keeping the other factors constant.
- The predicted value of Y for a certain value of X can also reflect the conditional probability of the occurrence (Y=1) in large samples.
    - The expectation is equivalent to the probability.
    - E.g: $E[buy] = Pr(buy = 1)$
    - The expected value of the dependent variable is the conditional probability that Y=1




## Probit Model

**Problem:** The LPM can sometimes observe theoretically unfeasible probabilities that do NOT fall in the range of 0%-100%.

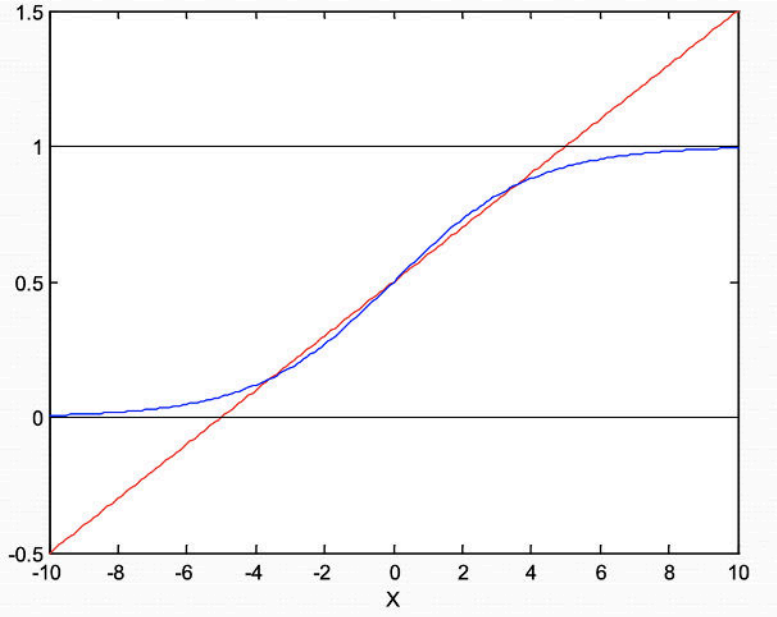
**Probit models:**

- Use the standard normal distribution function to 'bend the OLS' so that it falls in the plausible range.
- Considered good for binary regressor as it is limited in the cumulative probability range from 0% to 100%.

Approach:

- Firstly, model z-scores as a linear function of the regressors, and, assuming z-scores follow the standard normal distribution, the corresponding probability is realized.

$$z = \beta_0 + \beta_1 X$$
$$\Phi(z) = \Phi\left(\beta_0 + \beta_1 X\right)$$
$$Pr = \Phi\left(\beta_0 + \beta_1 X\right)$$




Note: The coefficient of the probit model cannot be interpreted directly. From the coefficient, we can only interpret its **sign and significance, NOT size**

- i.e. whether it increases or decreases the likelihood (probability).
- If X increases by 1 unit, then z increases by $\beta_1$ , and from that, a non-linear reference can be made about the probit =>indirect model.

The **effect size** can be easily computed by finding the conditional expectation of Y for a given value of X and then comparing the conditional expectation of Y from the initial value of X and finally taking the difference.

Example:
$Pr(buy) = \Phi(z) = \Phi(-4.1 + 0.1\,age)$
Age 50: z=0.9, Pr(buy)=82%
Age 51: z=1.0, Pr(buy)=84%
When age increases by 1 from 50 to 51, Pr(buy) increases by 2%

## Logit Model

Logit and probit models are very similar and produce almost the same results, only in the case of extreme values of X do their values deviate substantially.

The **logit model** is an alternative to the probit model as both models indirectly estimate the probability. In the case of the logit model, the **logistic function** is used:

$$Pr(Y) = \frac{1}{1+e^{-L}}$$
$$L = \beta_0 + \beta_1 X$$

- Resides on the foundation of odds: odd is defined as probability of occurrence divided by probability of non-occurrence.
- The logistic function is the natural logarithm of odds.

$$ln(odds) = ln\left[\frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \times \left(1 - \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}\right)^{-1}\right] = \beta_0 + \beta_1 X$$

It is important to realize like probit models, logit models **do NOT have constant effect sizes**, here the effect size is more substantial in the middle of the distribution rather than extremes.

## Comparison, Maximum Likelihood, and Extensions

The OLS model attempts to minimize the square of the residuals, whilst for the logit and probit models there is an attempt to maximize the likelihood efficiently.

### Multinomial Variables

- Can take on more than just two values compared to binary variables.

**Models without natural ordering:**

- Use multinomial logit and probit
- E.g. commuting preference: car, bike, public transport, …

**Models with specific ordering**.

- It is also possible to have ordered choices wherein the options themselves have some inherent ranking.
- Adopt ordered probit.
- Cannot replace 'ranking' with numbers
- E.g. How is your health in general? Very bad / Bad / Fair / Good / Very good

# Introduction to Econometrics – IBEB – Lecture 12, week 5

## Time Series

### Introduction

Definition: A sequence observed and recorded at successive points in time with equal intervals in between.

**Time series analysis:** Allows us to study the properties of the time series, the interaction between time series, or make forecasting of future occurrences.

- Point Forecast: When a single forecast is made by time series
- Interval Forecast: When there is some range of the forecasted variable

While a **cross section** is observed only once.

### Notation

$Y_t$ is a value of Y at time t (similarly $X_t$), t=1,2, … ,T
$\{Y_1, Y_2, \dots, Y_T\}$ is time series of Y
$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$ is a time series regression
The last component, also known as shock/news, is unforecastable by the model and usually marks sudden or unpredictable events.

**Note:** When performing a time series analysis, always sort the data from oldest to newest, otherwise you might make an erroneous mistake such that you might end up predicting the past.

In a graph, the correct way implies that the prior (old) periods would come to the left and the more recent ones on the right. Furthermore, rather than simply noting that the trend is upwards or downwards one must think whether it is significantly and relevantly upwards or not.

## Notation of lags

$Y_t$ is a value of Y at time t

First lag: $Y_{t-1}$ is a value of Y at time t-1

$j^{th}$ lag: $Y_{t-j}$ is a value of Y at time t-j

A first-order autoregression: $Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$

Autoregression means that this is a regression of Y on itself and first-order means that we use first lag.

## Notation of Differences

First difference: $\Delta(Y_t) = Y_t - Y_{t-1}$

Double first difference: $\Delta^2(Y_t) = \Delta\Delta(Y_t) = \Delta(Y_t - Y_{t-1}) = Y_t - 2Y_{t-1} + Y_{t-2}$

Yearly difference: $\Delta_{12}(Y_t) = Y_t - Y_{t-12}$

First (natural) log-difference: $\Delta(lnY_t) = lnY_t - lnY_{t-1}$

In this course, we will use the log differences to determine the growth rates.

**Note**: The subscript and superscript differ in their interpretation.

- Superscript (to the power) denotes the first difference that replicated j times
- Subscript denotes the difference in time t and t-j
- The double first difference above can be stated otherwise as a growth in growth (usually for exponential variables)

# Annualized vs Annual Growth

**Annualized Growth:**

- The growth per given period (other than year) is scaled to year

$$100 \times \left(ln(Y_t) - ln(Y_{t-1})\right) \times 12$$

**Annual Growth:**

- The difference from one year to the other

$$100 \times \left(ln(Y_t) - ln(Y_{t-12})\right)$$

Example:

- A change of some variable over a quarter (comparing last quarter with this quarter) could be multiplied by 400 (4 times 10: 4 quarters, 100 stands for percentage), and the result is annualized growth.
- Meanwhile, the change between this year's quarter compared to last year's quarter directly computed is annual growth.

## Autocorrelation

Correlation: $\rho = \frac{cov(X,Y)}{\sqrt{var(X) \times var(Y)}}$

First-order autocorrelation: $\rho_1 = \frac{cov(Y_t, Y_{t-1})}{\sqrt{var(Y_t) \times var(Y_{t-1})}}$

$j^{th}$ order autocorrelation: $\rho_j = \frac{cov(Y_t, Y_{t-j})}{\sqrt{var(Y_t) \times var(Y_{t-j})}}$

**Note**: As we increase the lag, the sample gets smaller as some observations have to be dropped out.

### Partial Autocorrelation

- An important concept that helps us identify an autoregression model.
- The outcome of a regression model with a time series as the dependent variable, and its jth lag as the regressor.

For $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon_t$ , $\beta_1$ is the first-order partial autocorrelation and $\beta_2$ is the second-order partial autocorrelation.

**Autocorrelation and partial autocorrelation features**

- Values are between -1 and +1
- They usually decrease in magnitude as the lag length increases. However, this is NOT necessarily true especially if one considers the case of seasonality.
- The 5% critical value for testing whether the coefficient is 0 is ±1.96 divided by the square root of the number of observations (T).

## Autocorrelated Errors

Autoregressive Form: $\varepsilon_t = \rho_1 \varepsilon_{t-1} + u_t$

If we have $\varepsilon_t$ then we can estimate $\rho$ from the equation $\varepsilon_t = \rho_1 \varepsilon_{t-1} + u_t$ using OLS.

If we don't have $\varepsilon_t$ then we can estimate $\hat{\varepsilon}_t$ first. This can be done by regressing $Y_t = \beta_0 + \beta_1 X_1 + \varepsilon_t$ to get estimates of $\beta_0$ and $\beta_1$, and then obtaining $\hat{\varepsilon}_t$ using the equation $\hat{\varepsilon}_t = Y_t - \widehat{\beta_0} - \widehat{\beta_1} X_t$

## The Durbin-Watson Statistic

Tests autocorrelated errors

$$d = \frac{\Sigma\left(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}\right)^2}{\Sigma \hat{\varepsilon}_t^2}$$

**Rule of Thumb:** d<1 is a warning for a positive autocorrelation.
$d < d_{Lower\ bound}$ : Positive autocorrelation
$d_{Lower\ bound} < d < d_{Upper\ bound}$ : Inconclusive
$d > d_{Lower\ bound}$ : Negative autocorrelation

**Note**: If the errors are autocorrelated, then the standard errors of the model turn out wrong from the standard approach. The Heteroskedasticity and Autocorrelation Consistent (HAC/Newey-West) variance can be used to get the correct standard errors.

# Introduction to Econometrics– IBEB – Lecture 13, week 6

## Dynamic Models

### Autoregressive Models (AR)

- When using the past observations of the variable itself in a regression.
- The order of an AR is the **maximum lag** of the equation (p), and may differ from the number of parameters.

P-th order autoregressive model (AR(p)):

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \varepsilon_t$$

For the AR(p) the parameter $\beta_p$ is the p-th order partial autocorrelation.

#### Long-Term Value

**Long-term expected value of an AR model:**

- Assumption 1: $Y_t$ and $Y_{t-1}$ are the observations from the same distribution, and hence $E[Y_t] = E[Y_{t-1}]$.
- Assumption 2: expected value of error term is zero.
- Derive the long-term expected value: $E[Y_t] = \beta_0/(1 - \beta_1)$
  (where $\beta_0$ is the intercept or constant and $\beta_1$ is the coefficient of $Y_{t-1}$)

#### Autocorrelation

Correlation between $Y_t$ and $Y_{t-1}$ given AR(1) model:

$$corr(Y_t, Y_{t-1}) = \beta_1$$
$$corr(Y_t, Y_{t-2}) = \beta_1^2$$
$$corr(Y_t, Y_{t-j}) = \beta_1^j$$

Partial autocorrelation is useful for higher order autocorrelation

**Note**: choose AR model for the highest significant order found from the partial autocorrelation!
**Note for STATA:** if you want to regress $Y_t$ on $Y_{t-1}$, you lose the first observation. The larger the order of the model, the more observations are lost from the sample.

## Information Criteria

**Note**: When the number of parameters in a model are increased the $R^2$ tends to increase, this, however, does not reflect an increase in the goodness of fit.

These following measures greatly assist in balancing the fit and the number of parameters:

$$Akaike\ IC\ (AIC) = \frac{-2ln(L)+2k}{T}$$
$$Schwarz\ IC\ (BIC) = \frac{-2ln(L)+k\times ln(T)}{T}$$

L: the likelihood (a function of the estimated variance of the errors)
K: the number of parameters including the constant
T: the number of observations (periods)

When fit gets better, $-\ 2ln(L)$ goes down because k goes up

You can pick which method to use but keep in mind that the BIC measure gets increasingly stricter when the number of observations is bigger than 8.

- The **lowest value of AIC and BIC is the most desired** because we are interested in the **minimal number of parameters while maximizing the fit**.
- AIC and BIC can give different results for different models => possible to have several models.

# Finite Distributed Lag Model

**Finite Distributed Lag model of Order q:**

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \ldots + \beta_q Y_{t-q} + \varepsilon_t$$

where p denotes the maximum lag of the dependent variable itself and q denotes the maximum lag of independent variable X.

$\beta_s$ = the distributed-lag weight or **s-period delay multiplier**
=> It indicates the effect of the change in $X_{t-s}$ on $Y_t$.

**j-Period Interim Multiplier:**

- The effect of a permanent change in X on Y after j periods.
- It carries on through all the parameters up until and including the j-th one

**The Total Multiplier:**

- The effect until the maximum lag q of the model
- The sum of all β

**j-Period Delay Multiplier:**

- The β that represents the j period lag

**Example**:
*What are the 2-period delay multiplier, the 2-period interim multiplier and the total multiplier?* $Y_t = 0.4 + 0.5X_t + 0.3X_{t-1} + 0.1X_{t-2} - 0.1X_{t-3} + \varepsilon_t$

A. 2-period delay: 0.1, 2-period interim: 0.8, total: 1.0
B. 2-period delay: 0.3, 2-period interim: 0.8, total: 0.8
C. 2-period delay: 0.1, 2-period interim: 0.9, total: 0.8
D. 2-period delay: 0.1, 2-period interim: 1.3, total: 1.2

# Autoregressive Distributed Lag Models (ARDL)

**ARDL model of order (p, q)**

$$Y_t = \beta_0 + \beta_1 Y_{t-i} + \ldots + \beta_p Y_{t-p} + \delta_0 X_t + \delta_1 X_{t-1} + \ldots + \delta_q X_{t-q} + \varepsilon_t$$

## Long-Term Value

2 steps:

1. Long term expectation for AR(p) model: $E[Y_t] = \frac{\beta_0}{(1-\beta_1-\beta_2-\ldots-\beta_p)}$
2. Extend to ARDL(p,q) model: $E[Y_t] = \frac{\beta_0+\delta_0 X_1+\delta_1 X_{t-1}+\ldots+\delta_q X_{t-q}}{(1-\beta_1-\beta_2-\ldots-\beta_p)}$

(where δ denotes the parameter of X.)

The long term effect of a permanent change in X by 1 on Y is reduced to:
$E[Y_{t,AR}] = \frac{\delta_0+\delta_1+\ldots+\delta_q}{(1-\beta_1-\beta_2-\ldots-\beta_p)}$

### Short-Term Effects ARDL

- Not easy to see

**Error Correction Format:**

Write this using $\Delta_1$:
Step 1: Subtract lag 1 on both sides

$$Y_t - Y_{t-1} = (\beta_1 - 1)Y_{t-1} + \delta_0(X_t - X_{t-1}) + (\delta_0 + \delta_1)X_{t-1} + \varepsilon_t$$

Step 2: Separate long-run $\frac{\delta_0+\delta_1}{1-\beta_1}$ from short-run $\delta_0$:

$$\Delta_1 Y_t = \delta_0 \Delta_1 X_t + (\beta_1 - 1)\left(Y_{t-1} - \frac{\delta_0 + \delta_1}{1-\beta_1} X_{t-1}\right) + \varepsilon_t$$

# Introduction to Econometrics – IBEB – Lecture 14, week 6

## Forecasting

### Moving Average Models (MA)

- Regresses the dependent variable $Y_t$ on the error term and the lags of the error.

$$Y_t = \beta_0 + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots + \theta_q\varepsilon_{t-q}$$

- **White Noise:** When we assume that the expected value of ε equals to 0 and the error terms (news) are uncorrelated. $E[\varepsilon_t] = 0,\ E[\varepsilon_t^2] = \sigma^2,\ E[\varepsilon_t,\ \varepsilon_{t-i}] = 0.$

This model represents lags of $\varepsilon_t$ instead of $X_t$:

- This means that errors affect Y with some lag in between
- The autocorrelation **abruptly stops** after q term, while for the AR model autocorrelations **gradually die out**.

## ARMA (p,q) Model

- Regresses the dependent variable $Y_t$ on Y up to p lags and on the error term up to q lags.
  $Y_t = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q}$

## Further Extension: ARIMA (p,d,q)

- This can be further adopted to include the difference approach.
  $\Delta^d Y_t = \beta_0 + \beta_1\Delta^d Y_{t-1} + \ldots + \beta_p\Delta^d Y_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q}$
  where d is the degree of 'differencing' and I is integration of order d.

# Forecasting

From an AR(1) Model:

We know the observation $Y_T$ at the forecast origin. Then we can predict $Y_{T+1}$.

$$Y_T = \beta_0 + \beta_1 Y_{T-1} + \varepsilon_T$$

$$Y_{T+1} = \beta_0 + \beta_1 Y_T + \varepsilon_{T+1}$$

$$E[Y_{T+1}] = E[\beta_0 + \beta_1 Y_T + \varepsilon_{T+1}] = \beta_0 + \beta_1 Y_T + 0 = \beta_0 + \beta_1 Y_T$$

$$\hat{Y}_{T+1} = \hat{\beta}_0 + \hat{\beta}_1 Y_T$$

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$$

We can calculate the expected value of $Y_{T+1}$.
With this value, we can continue and calculate the expected value of $Y_{T+2}$!

$$E[Y_{T+1}] = \beta_0 + \beta_1 Y_T$$

$$E[Y_{T+2}] = E[\beta_0 + \beta_1 Y_{T+1} + \varepsilon_{T+2}]$$
$$= \beta_0 + \beta_1 E(Y_{T+1}) + E(\varepsilon_{T+2}) =$$
$$\beta_0 + \beta_1(\beta_0 + \beta_1 Y_T) + 0 = \beta_0(1 + \beta_1) + \beta_1^2 Y_T$$

$$\hat{Y}_{T+2} = \hat{\beta}_0(1 + \hat{\beta}_1) + \hat{\beta}_1^2 Y_T$$

We can continue and calculate $\hat{Y}_{T+3}$, $\hat{Y}_{T+4}$, and so on

**Important Notes:**

- It is important to ask: **what do I know now at time T to predict $Y_{t+1}$?**
- For a fact you do NOT know $Y_{t+1}$ thus when calculating $Y_{t+2}$, for example, you can replace $Y_{t+1}$ by its prediction $\beta_0 + \beta_1 Y_t$.
- Example: if today is Tuesday and you want to predict something for Saturday, you need to predict for ALL the day before (Wednesday, Thursday, and Friday).
- When forecasting the expected value of the future of the dependent variable, the expected value of news (error term $\varepsilon$) is always zero, as news is unforeseen and unpredictable.

## How accurate are forecasts?

- **Forecast Error:** the difference between the true observed value of Y in the future period and the forecasted value of Y in the future period.

The **mean squared forecast error (MSFE)** and the **root of it (RMSFE)** can be represented as follows:

$$MSFE = \frac{\sum_{t=T+1}^{T+n} (Y_t - f_t)^2}{n}$$

$$RMSFE = \sqrt{MSFE}$$

Y: true value
f: forecasted values

**Lowest (R)MSFE = best forecasting model!**

**Factors of Uncertainty:**

1. Error terms
2. Parameter estimates

- Implies MSFE out-of-sample>MSE in-sample

# Forecast Intervals

Since we can never be 100% certain when making forecasts, we should consider for example 95% forecast interval. 1-period forecast interval of 95% is given by:

$$\left[\widehat{Y}_{T+1} - 1.96 \times \widehat{\sigma};\ \widehat{Y}_{T+1} + 1.96 \times \widehat{\sigma}\right]$$

## The Variance

For AR(1) this is simply:

$$Y_{T+1} = \beta_0 + \beta_1 Y_T + \varepsilon_{T+1}$$
$$Var\left(Y_{T+1} - f_{T+1}\right) = Var\left(\varepsilon_{T+1}\right) = \sigma^2$$

However, as we are making predictions for further in the future the variance increases (interval becomes larger) and uncertainty increases.
The example of forecasting multiple steps with AR(1):

$$Y_{T+2} = \beta_0 + \beta_1 Y_{T+1} + \varepsilon_{T+2}$$
$$Y_{T+2} = \beta_0 + \beta_1\left(\beta_0 + \beta_1 Y_T + \varepsilon_{T+1}\right) + \varepsilon_{T+2}$$
$$Y_{T+2} = \beta_0\left(1+\beta_1\right) + \beta_1^2 Y_T + \beta_1\varepsilon_{T+1} + \varepsilon_{T+2}$$
$$Var\left(Y_{T+2} - f_{T+2}\right) = Var\left(\beta_1\varepsilon_{T+1} + \varepsilon_{T+2}\right) = Var\left(\beta_1\varepsilon_{T+1}\right) + Var\left(\varepsilon_{T+2}\right) =$$
$$\beta_1^2 Var\left(\varepsilon_{T+1}\right) + Var\left(\varepsilon_{T+2}\right) = \beta_1^2\sigma_\varepsilon^2 + \sigma_\varepsilon^2 = \left(\beta_1^2 + 1\right)\sigma_\varepsilon^2$$

$$Var\left(Y_{T+3} - f_{T+3}\right) = \left(\beta_1^4 + \beta_1^2 + 1\right)\sigma_\varepsilon^2$$

## Pseudo-Out-of-Samples

- When you take some part of the data and reserve it for further analysis.

- The sample of the time series can be divided, this does not necessarily need to be in two halves as the data is not cross-sectional and varies substantially throughout.
- The first portion of the sample is used to make forecasts using the models in an attempt to see if the model can accurately predict the points of the second part of the data.

## Granger causality

- Correlation between A and B implies A=>B, or B=>A, or both.
- **An association of causality with probability**
- NOT a synonym for true causality as both variables may be affected by another time-varying variable.
- **Time-Series:** ONLY forward, thus is very complicated

**(Semi) Solution:**

- Apply different treatments to the same subject
- BIG ceteris paribus assumption

# Introduction to Econometrics – IBEB – Lecture 15, week 7

## Non-Stationarity

Up until now, we assumed the time series displayed stationarity. Now, we consider non-stationarity (inconsistent mean or variances overtime, for example when there's a trend).

### Spurious Regression

- When there should not be a relationship between variables, however, we obtain one that is significant and positive for one part of the sample and significant negative for the other part of the sample.

- NOT accounting for the non-stationarity could lead to spurious conclusions about the existence of the relationship between variables.

**Indicators of Non-Stationarity:**

- when applying the AR(1) model and the coefficient is equal to 1.
- Very large t-statistics of the coefficient of model, if it is the case this could be a sign that there is no t-distribution underlying this result.
- It is possible that by including a trend variable the effect of time will be isolated and the significance of the spurious relationship will disappear.

## Random walk

- When each new data point is determined by the error term.
- This happens when the coefficient of AR(1), which is the slope coefficient, is equal to 1.

$$Y_t = \beta_1 Y_{t-1} + \varepsilon_t \text{, where } \beta_1 = 1.$$

- When further simplified to the first order, the 0th observation is simply a sum of all the error terms of different periods.

$$Y_t = Y_0 + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \ldots + \varepsilon_1$$

- When plotting the date, the line does NOT return to a constant mean, it stays below/above for a long period of time.

**The Expected Value:**

$$E(Y_t) = E(Y_0) + E\left(\sum_{i=1}^{t} \varepsilon_i\right) = Y_0 + 0 = Y_0$$

This implies that there is NO better prediction for the future than the value of today, since the expected value of the error term is equal to zero.

**The Variance:**

- The more errors, the more variance
- Non-stationarity implies that there is a new variance for every observation.

$$Var(Y_t) = t\sigma^2$$

# Drift (trends)

- When there is a constant term or the intercept $\alpha$.
- The model is then:

$$Y_t = t\alpha + Y_0 + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \ldots + \varepsilon_1$$

**The Expected Value:**

- Non-stationarity implies that there is a new expected value for each observation

$$E(Y_t) = Y_0 + t\alpha$$

**The Variance:**

- Same as the one without drift

$$Var(Y_t) = t\sigma^2$$

# Dickey-Fuller Test

- **Important**: standard statistical tools are not applicable as mean and variance are not constant over time in this case
- Also known as a unit root test
- Test if AR coefficient $\beta$=1

The model can be manipulated as follows where the growth of the dependent variable is regressed against the lag of the dependent variable, resulting in a new parameter $\gamma$:

$$Y_t = \beta\, Y_{t-1} + \varepsilon_t$$
$$Y_t - Y_{t-1} = \beta\, Y_{t-1} - Y_{t-1} + \varepsilon_t$$
$$\Delta Y_t = \gamma Y_{t-1} + \varepsilon_t$$

where $\gamma = \beta - 1$

**Hypotheses:**

$H_0$: $\gamma = 0$ *the time series is a random walk* ($\beta = 1$)

$H_1$: $\gamma < 0$ *the time series is stationary* ($\beta < 1$)

**Versions of The Dickey-Fuller Test:**

1. DF test 1 (no intercept, no trend): $\Delta Y_t = \gamma Y_{t-1} + \varepsilon_t$

2. DF test 2 (intercept, no trend): $\Delta Y_t = \alpha + \gamma Y_{t-1} + \varepsilon_t$
3. DF test 3 (intercept and trend): $\Delta Y_t = \alpha + \lambda t + \gamma Y_{t-1} + \varepsilon_t$

**When to Use each Model:**

- DF test 1 is used when the model is around the mean of 0.
- DF test 2 is used when the model is around another mean which a constant number different than 0.
- DF test 3 is used when the model trends around a linear trend.

**Note:** It is important to choose the right test to perform as the critical values for each type are different. Mind that the test is one-sided.

The critical values at 5% are:

- DF1: -1.95
- DF2: -2.86
- DF3:-3.41

If the test statistic is less than the critical value, which means if the DF Test is to the left of the critical value, we reject the null hypothesis, hence there is no random walk detected and the data is stationary.

**Augmented Dickey-Fuller Test (ADF)**

$$\Delta Y_t = \alpha + \gamma Y_{t-1} + \delta_1 \Delta Y_{t-1} + \delta_2 \Delta Y_{t-2} + \ldots + \varepsilon_t$$

## Solutions to Non-Stationarity

1. **Detrending**

- If there is a deterministic trend and no unit root, then including an additional regressor that reflects the time period will account for the trend part.
- Steps:
    1. Regress Y on time (T) and a constant. This estimates the trendline. Subtract this to get the detrended Y.
    2. Repeat for X.
    3. Regress detrended Y on detrended X.

2. **Differencing**

- If the trend is stochastic (unit root)

- taking the first differences:

$$Y_t = \beta_0 + Y_{t-1} + \varepsilon_t$$
$$Y_t - Y_{t-1} = \beta_0 + Y_{t-1} + \varepsilon_t - Y_{t-1}$$
$$\Delta Y_t = \beta_0 + \varepsilon_t$$

## Breaks

- The regression model changes over the course of the sample.
- After a break (denoted $\tau$) we have the same variables, but different parameters:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \varepsilon_t, \qquad \text{when } t < \tau$$
$$Y_t = \beta'_0 + \beta'_1 Y_{t-1} + \delta'_1 X_{t-1} + \varepsilon_t, \qquad \text{when } t \geq \tau$$

- If we want to examine whether there was a break we must adopt a dummy variable that will take on a value of 0 if $t < \tau$, and a value of 1 when $t \geq \tau$ in the regression following regression:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + D\gamma_0 + D\gamma_1 Y_{t-1} + D\gamma_2 X_{t-1} + \varepsilon_t$$

At time $t < \tau : Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \varepsilon_t$
At time $t \geq \tau : Y_t = (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1) Y_{t-1} + (\delta_1 + \gamma_2) X_{t-1} + \varepsilon_t$

## Chow Break Test

- This is performed when we know the date of the break
- Using the dummy variables we aim to test if the new parameters $(\gamma_0, \gamma_1, \gamma_2)$ are significant or NOT.

## Quandt Likelihood test (QLR)

- This is used instead if we do not know the exact date of the break and are looking for it.
- **Note**: you cannot perform QLR to find the date of the structural break and then test again with Chow test.
- The break is around the maximum F-value.

# Introduction to Econometrics – IBEB – Lecture 16, week 7

## Volatility Clustering

All the models incorporated up till now have assumed homoskedasticity, meaning the variance of the error term is equivalent at any time t.

- Robust standard errors (HAC) adopted when this does not hold
- Now we must also consider the time on which the error term can vary!

## ARCH Model

- autoregressive conditionally heteroskedastic models
- The size of error terms is a high determinant for the variance

**Default**: $\sigma_t^2 = \alpha_0$

**Extension:** $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$

This extension implies that future variance can be predicted by past errors.

### Simple Model of Returns

$$Y_t = \beta_0 + \varepsilon_t, \qquad \text{where } \varepsilon_t \sim N\left(0; \sigma_t^2\right).$$

Previously, we did not have to worry about the $\sigma_t^2$ term as it was constant, however, now we drop this assumption to model $\sigma_t^2$ as a function of past errors:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2.$$

The model with constant variance usually displays normal distribution in very large samples. The ARCH model on the other hand has relatively thicker tails with correlated extreme errors and the mean is highly peaked.

## Test if Variance is ARCH

We can use the hypothesis testing to test whether the variance is indeed the ARCH variance.

Default: $\sigma_t^2 = \alpha_0$ (constant variance)
ARCH(1): $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$ (variance is changing)

It is important to note that we do not observe the $\sigma_t^2$, therefore we use the squared residuals instead:

$$\hat{\varepsilon}_t^{\,2} = \alpha_0 + \alpha_1 \hat{\varepsilon}_{t-1}^2.$$

With that, we want to know if the $\alpha_1$ is significant.

To estimate an ARCH model the maximum likelihood method is used that simultaneously estimates $\hat{\sigma}_t^2$ and chooses the parameters ($\beta_0$, $\alpha_0$, $\alpha_1$) that fit the data most (use Stata to solve that).

# GARCH Model

If we believe that the variance is not only dependent on the square of the error terms but also on the lags of $\sigma_t^2$ we can add them into the model, which then becomes the GARCH model, which stands for Generalized ARCH model.
It is a variance version of ARDL (distributed lag model):

ARCH(1): $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$
GARCH(1): $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \varphi_1 \sigma_{t-1}^2$

The long-term unconditional variance can be computed via the formula: $\frac{\alpha_0}{1-\alpha_1-\varphi_1}$

**Note**: it is possible to increase the GARCH(1,1) to GARCH(p,q) model by adding more autoregressive terms as well as lag terms, however, it is rarely done in practice.

# References


Van Ourti, T. (2025). Lecture 1: *Methods 1* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119869

Bago d'Uva, T. (2025). Lecture 2: *Methods 2* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119870

Bago d'Uva, T. (2025). Lecture 3: *Methods 3* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119871

Bago d'Uva, T. (2025). Lecture 4: *Methods 4* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119904

Bago d'Uva, T. (2025). Lecture 5: *Methods 5* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119594

Bago d'Uva, T. (2025). Lecture 6: *Methods 6* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119890

Van Ourti, T. (2025). Lecture 7: *Methods 7* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119930

Van Ourti, T. (2025). Lecture 8: *Methods 8* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119913

Van Ourti, T. (2025). Lecture 9: *Methods 9* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119866

Van Ourti, T. (2025). Lecture 10: *Methods 10* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100469408

Van Ourti, T. (2025). Lecture 11: *Methods 11* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119794

Hans Franses, P. (2023). Lecture 12: *Time Series* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119844

Hans Franses, P. (2023). Lecture 13: *Dynamic models* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119837

Hans Franses, P. (2023). Lecture 14: *Forecasting* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119846

Hans Franses, P. (2023). Lecture 15: *Non-stationarity* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119810

Hans Franses, P. (2023). Lecture 16: *Volatility clustering* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119835