EFR summary

Introduction to Econometrics,

FEB12012X

2025-2026

Lectures 1 to 16

Weeks 1 to 7

Details

Subject: Introduction to Econometrics IBEB 2025-2026

Teacher: Teresa Bago d'Uva, Tom Van Ourti, Phillips Hans Franses

Date of publication: 16.04.2026

Introduction to Econometrics – IBEB – Lecture 1, week 1

Methods

Everyday vs Scientific learning

Everyday we learn via 3 different ways**, tradition, experts or personal experience**, while for learning via tradition and experts requires little effort, however the knowledge acquire might be wrong.
For learning vis personal experience, you get to understand the cause and consequences of certain action, meaning causal chain. However, it also got it's problems

- No accurate observations
- Overgeneralisation: selective observations
- Illogical reasoning
  - Science is probabilistic
  - Illogical reasoning: correct answer with wrong methods

Scientific learning requires a lot of time, where we aim to learn whether something is true or not

- We extend existing knowledge
- Learning via scientific methods
- Using theory, data and analysis

Association versus Causal Effect

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

An example to illustrate this is the paper on Mortality effect of left-handedness. A research was carried that recorded all deaths in Southern California in 1990, where they found out that left-handed people died 9 years earlier, however there was a crucial assumption violated, omitted variables.

- In the earlier days people were forced to use their right hand, so when researchers collected data they recorded right hand, even though they were born left hand, this naturally increased the average age of right hand

- While the average age at death of left handed people were younger because the older naturally left hand were recorded right hand
- As you can see in this scenario we did not take into the fact that older generations were forced to use right hand, this means the 9 years result is **only an association and not a causal effect**

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

These are the things that we have to do before we perform actual research

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

- E.g. Suppose we want to find out, do higher wages lead to a higher opportunity cost of time? if this is the case we expect people with higher wages to invest more on their health
- It's easy to define wages, however what about health, physical or mental...
- It is important to be precise

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.

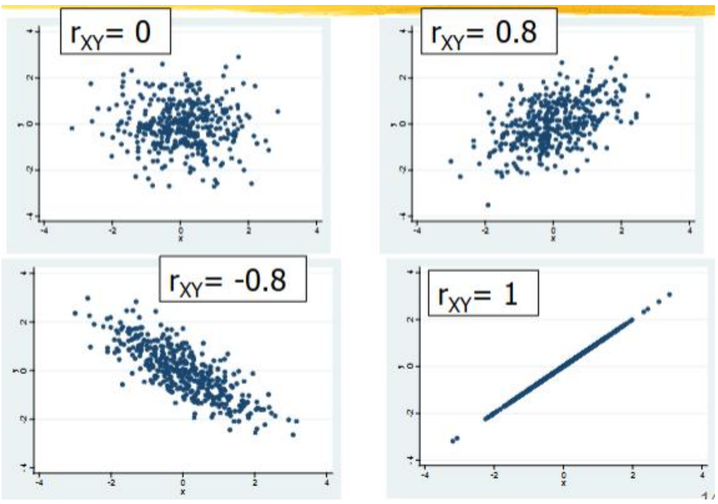
Introduction to Econometrics – IBEB – Lecture 2, week 1

OLS: simple linear regression model

Regression models are potentially used in any relationship between 2 variables that might be of interest, specifically suited for continuous dependent variable Y as functions of any kind of variable X, example include:

	Y	X
Microeconomics:	Wage	Education
	Student performance	Class Size
	Cigarettes smoked	Price
	Birthweight	Smoked during pregnancy
	etc. etc.	
Macroeconomics:	Health care expenditures	Age structure
	traffic deaths	alcohol taxes
	GDP	Unemployment rate
	etc. etc.	
Others	? ?	? ?

Relationships between variables



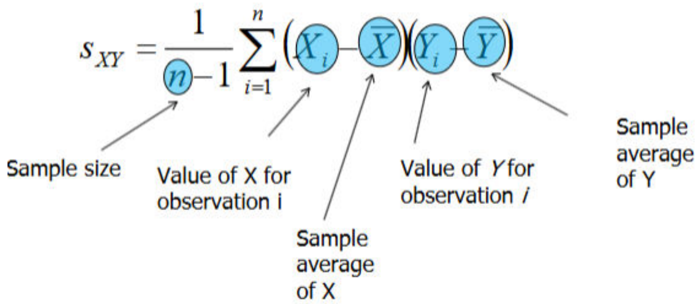
One of the ways to find out about the relationship between variables can be by constructing a **scatter plot**. The relationship can be negative or positive, or there can be no relationship.

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:

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$



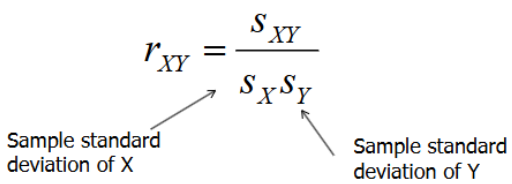
Where n is sample size, X_i is the value of X for observation i (similarly Y_i) and $\underline{X}$ is the sample average of X (similarly $\underline{Y}$). The units of sample covariance = units X * units Y.

The covariance tells us if X and Y tend to move in the same (+) or opposite directions (-), if it's 0 means they are independent

- units of measure are the units of X times the units of Y, which is not very intuitive so we have the sample correlation

Sample correlation:

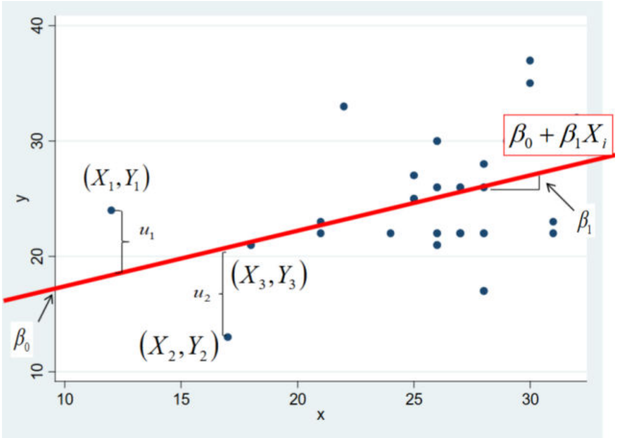
$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$



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. It shows the strength of the relationship between X&Y.

- The numerator is units of X time units of Y and denominator the same, thus correlation coefficient is unitless.

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. β_1 is the slope of the regression line and β_0 is the intercept.

The line of best fit

The line of best fit is based on minimizing the following equation:

$$\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

This formula tells us the **sum of the squared distance between data points I's and the fitted line**

- It is squared because we take into account the positive and negative differences
- And, as our goal is to minimize the equation, by squaring, we put more/less weight on points that are close/away from line

OLS estimator $\widehat{\beta_0}$:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

OLS estimator $\widehat{\beta_1}$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}$$

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 β_1 is measured in $\frac{units\ Y}{units\ X}$. Linear regression model coefficient shows causality only under OLS assumptions (lecture 3), and if those do not hold it shows association and should not be used for policy design.

Goodness of fit measures

Note: these two values tell us about how good our regression model is at explaining the data, it does not tell us anything about the relationship between X and Y

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:

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

- **Explained sum of squares (ESS)**: It shows the variation in the data explained by the model

$$ESS = \sum_{i=1}^{n} (\widehat{Y}_i - \underline{Y})^2$$

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

$$R^2 = \frac{ESS}{TSS} = corr(Y_i, \widehat{Y}_i)^2$$

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

- Range of R^2: $0 \leq R^2 \leq 1$
- If R^2= 1: model predicts Y_i perfectly, so $\widehat{Y_i} = Y_i$
- If R^2= 0: model (X_i) does not predict any variance in Y_i, so $\widehat{\beta_1} = 0$, thus $\widehat{Y_i} = \bar{Y}$
- Unitless

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.

- Measured in units of the dependent variable (often y-axis)

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

Where $S_{\hat{u}}^2$ is the sample variance of the residuals $\widehat{u_i}$

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

Examples

Suppose we have, number of classes attended out of 32 (X) and final exam score (Y)

A R^2 of 0.02 means that 2% of sample variance of Y is explained by X, meaning 2% of the difference in exam scores between students can be explained by the number of lectures they attend

- And 98% of the differences are explained by other factors, such as exam difficulty etc…

A SER of 4.667 tells us on average the model’s prediction are about 4.667 units away from the real value, in this case measured in exam scores

- We can compare it against the std. dev. = 4.710, as you can see, they are practically the same this means that almost all the variation in the data are unexplained by the model

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.

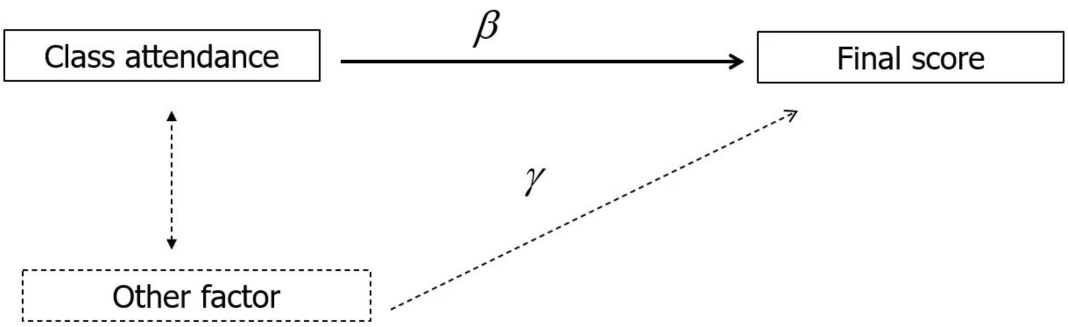
$$E(u_i|X_i) = 0$$

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 influence Y_i, 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.

As you can see in the diagram below if Class attendance is correlated with other factors that affect Final score, then the slope coefficient is not purely due to Class attendance



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.

This assumption does **not hold** when the observations are dependent, such as in time series data or panel data,

- where for example the GDP of this year might very well be influenced by the GDP of last year
- It also does not hold when sample is **not representative**

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.
- Otherwise this is a plausible assumption

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.

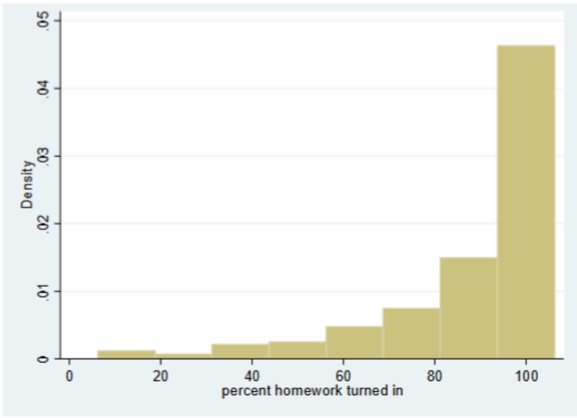
As different samples can lead to different estimates, the estimators are just some points in the sampling distribution of the estimator.

- 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 β_1 estimator $(\widehat{\beta_1})$ approximate to a **normal distribution.**

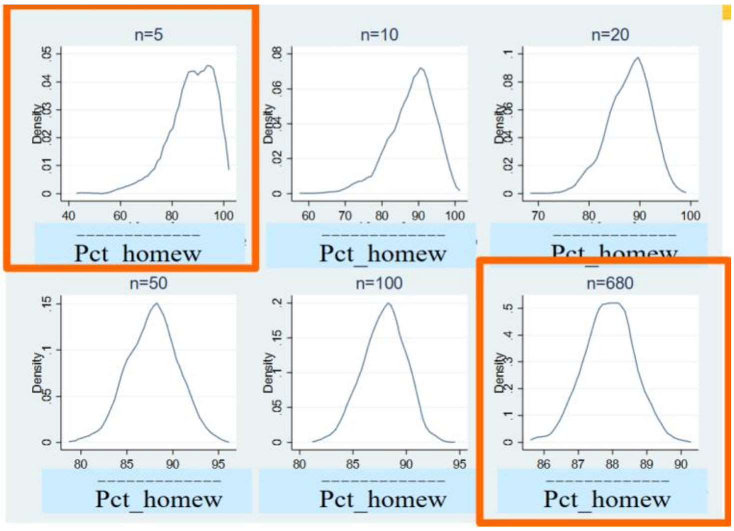
This comes directly from the **central limit theorem.**

- If $Y_1, Y_2, \ldots, Y_n$ are Independently and identically distributed with certain mean and variance, as n approaches infinity, $\bar{Y}$ follows approximately a normal distribution
- If all the requirement of the central limit theorem (above) are also met for $\widehat{\beta_1}$, then as n goes to infinity $\widehat{\beta_1}$ follows approximately a normal distribution
- **Note:** the 1 to 3 assumptions above also need to hold

Suppose we have the distribution of the percentage of homework turned, which is not normal



In the following figure we get various samples and calculate their sample mean, so that we can get the distribution of the sample mean for all the different samples, where changes to the n, sample size, are made for each distribution



As you can see the larger the sample size the distribution of the sample means approaches normal distribution, even though the initial variable distribution is not normal

Property of OLS estimators

Unbiasedness

When the estimators are **unbiased**. Therefore, the mean sampling distribution $\widehat{\beta_1}$equals β_1 and similar for $\widehat{\beta_0}$

$$E(\hat{\beta}_0) = \beta_0$$

$$E(\hat{\beta}_1) = \beta_1$$

To prove that under the OLS assumptions the OLS estimators are unbiased, we start

We saw before: $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

So $$E(\hat{\beta}_1) = E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}\right]$$

Under some rearrangements, we get to

$$E(\hat{\beta}_1) = \beta_1 + E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})E(u_i \mid X_1, \ldots, X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

And if **assumption 2** holds, so if observations are i.i.d. this means the error for observation I only depends on it's own X value, not on the X values of other observations

- Student i's unexpected factors affecting their score (like being tired, lucky guesses, stress) should only relate to **their own number of lectures attended**, not other students' attendance.

Plus, **assumption 1**, making it equal to 0, so $E(\widehat{\beta_1}) = \beta_1$

$$E(u_i \mid X_1, \ldots, X_n) = E(u_i \mid X_i) \quad \text{If Assumption 2}$$

$$E(u_i \mid X_i) = 0 \quad \text{If Assumption 1}$$

This means that, unbiasedness of $\widehat{\beta_1}$ is satisfied if assumptions 1 and 2 hold.
Note: a similar derivation can be done for $\widehat{\beta_0}$

Variance of estimators and consistency

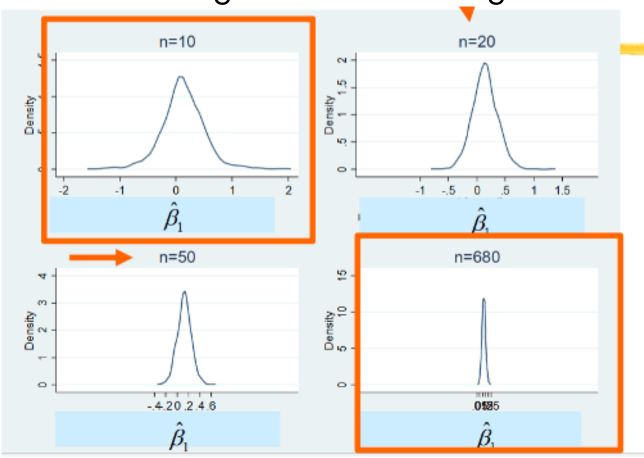
Because of the central limit theorem in large samples, $\widehat{\beta_0}$ and $\widehat{\beta_1}$ approximately follow a normal distribution $\widehat{\beta_1} \sim N\left(\beta_0; \sigma^2_{\widehat{\beta_1}}\right)$, and jointly they follow a bivariate normal distribution.

Variance of $\widehat{\beta_1}$:

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n} \frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}$$

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.

Graphically, as you can see below, as the sample size increase the variance of the OLS estimator gets smaller making the distribution tighter



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

$$E(Y_i \mid X_i) = E(\beta_0 + \beta_1 X_i + u_i \mid X_i) = \beta_0 + \beta_1 X_i + E(u_i \mid X_i)$$

- which under Assumption 1 further simplifies to $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$

Most common interpretation of β_1:

- when X goes up by 1, the $E(Y_i \mid X_i)$ goes up by β_1
- when X goes up by ΔX, then $E(Y_i \mid X_i)$, goes up by $\Delta X \beta_1$

Interpretation of the intercept

Generally β_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: β_1 is the difference between the average when D=1 and the average when D=0. β_1 is the change in average Y when D=1 compared to D=0.

Example

Dummy variable: attended more than 25 classes or not

- D = attend_25 = 1, if attend > 25
- D = attend_25 = 0, if attend < 25

If $\beta_1 = 0.839$, this means that, attending more than 25 classes increases mark with 0.839 points on average, compared to attending at most 25

- In this case the constant term tell us the average mark if student attended at most 25 classes

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

Statistical inference

We have already seem that from previous lectures that different sample gives different estimates of the slope coefficient

- We also know that, if we could draw all possible random samples on average, we would obtain the true value (OLS estimator unbiased)

Suppose from a sample we get that the estimate of the slope is 0.121, however it is also possible that true slope is actually zero

- With statistical inference we try to answer, how confident can we be that the true effect is not actually
  - 0, 1 or even 0.2?

Hypothesis tests and confidence intervals in linear regression

Notation

In previous lectures we considered **OLS estimator** $\widehat{\boldsymbol{\beta}_1}$ the same thing as the **OLS estimate** $\widehat{\boldsymbol{\beta}}_1^{act}$, however for this lecture

- OLS estimator, refers to a random variable that is different for each sample
- OLS estimate, refers to the actual estimate we obtain with our sample
- Finally, the value of β in the hypothesis we are testing denoted as $\beta_{1,0}$

Two-sided hypothesis test

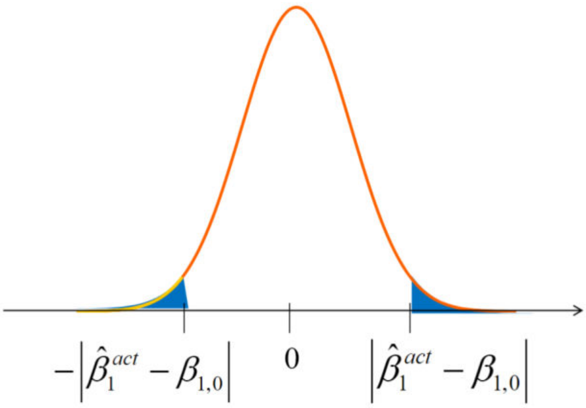
$$H_0: \beta_1 = \beta_{1,0} \quad vs \quad H_1: \beta_1 \neq \beta_{1,0}$$

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), if H_0 is true

t statistic and p-value

P-value: probability of obtaining $\widehat{\beta_1}$ 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: $t = \frac{\widehat{\beta_1} - \beta_{1,0}}{SE(\widehat{\beta_1})}$

Decision rule and Rejection region

Using the significance level of 5%:

Reject H_0 if

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

Two-sided or one-sided hypothesis test?

We can have 3 different types of alternative hypotheses:

- **Two sided**

$$H_1 : \beta_1 \neq \beta_{1,0}$$

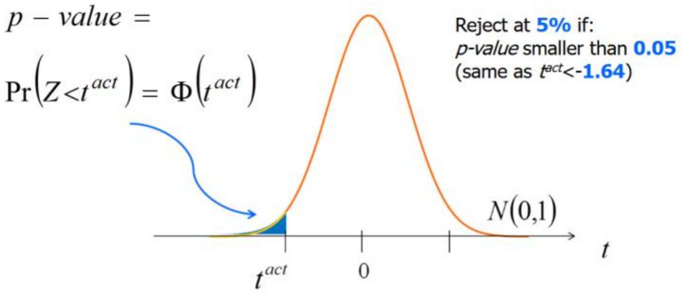
- **One sided** (either smaller or bigger than)

$$\beta_1 > \beta_{1,0} \text{ or } \beta_1 < \beta_{1,0}$$

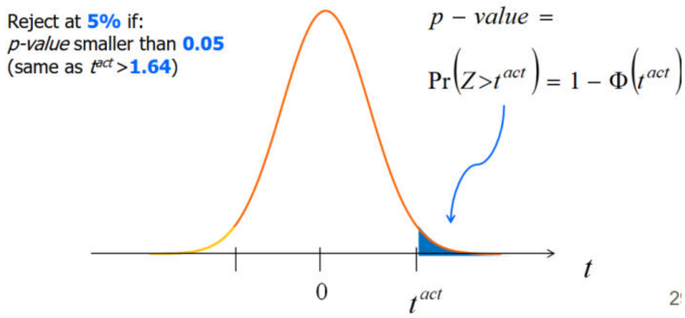
If the alternative hypothesis is that it's smaller than $(\boldsymbol{\beta_1} < \boldsymbol{\beta_{1,0}})$

- So, the p-value will tell us the probability that the t-statistic is below the t_{act}, if null is true, the blue area in the graph

- Since we know that the t-statistic in large samples follows approximately a standard normal distribution
- And we reject the null hypothesis if p-value < significance level set (in this case 5%)



A similar explanation can be given for alternative hypotheses, where $\boldsymbol{\beta_1} > \boldsymbol{\beta_{1,0}}$



Confidence intervals

A 95% confidence means that **from all samples that can be drawn, the interval contains the true value of β_1 in 95% of the cases**.

- Unlike the predicted $\hat{\beta}_1$ which is a point estimate, the confidence is **an interval estimate,** which has a upper and a lower bound, just like a two-sided test
- Since the t-statistic follows a standard normal distribution approximately and in large sample, we know that the probability that the t-statistic is between -1.96 and 1.96 equals 95%

$$\Pr\left(-1.96 < \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} < 1.96\right) = 0.95$$

- We have two bounds, we split the 5% significance level in 2, so we have 97.5%, which give us the critical values

After some steps we get

$$\Pr[\hat{\beta}_1 - 1.96 \times SE(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + 1.96 \times SE(\hat{\beta}_1)] = 0.95$$

So the 95% confidence interval is

$$\left[\hat{\beta}_1 - 1.96 \times SE\left(\hat{\beta}_1\right), \hat{\beta}_1 + 1.96 \times SE\left(\hat{\beta}_1\right)\right]$$

- Interpretation: the set of all values that cannot be rejected using a 2-sided hypothesis at 5% significance level
- Suppose [0.059, 0.183] is the interval estimate, this means that we reject null hypothesis that for values of β_1 greater than 0.183 or smaller than 0.059, otherwise we do not reject null hypotheses

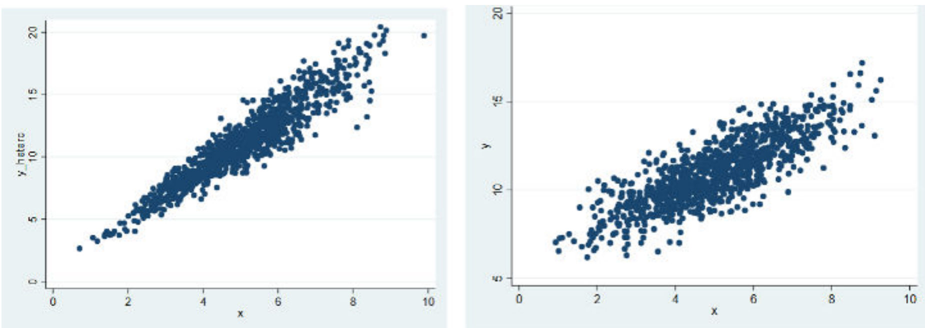
Homoskedasticity & Heteroskedasticity

This is a special case where we assume Homoskedasticity (not an assumption)

$$\text{var}\left[u_i \middle| X_i = x\right] \text{ is constant for } i = 1, \ldots, n$$

- Meaning **constant variance, so it does not depend on x**
- If this holds
    - the formula of standard errors of $\widehat{\beta_1}$ can be simplified
    - the OLS estimator has minimal variance amongst all unbiased linear estimators (efficient).

Left graph represents Heteroskedasticity, Right is Homoskedasticity
- For Homoskedasticity we see that the dots are equally distributed for all values of x
- Unlike for Heteroskedasticity we see that for greater values of x there are more dots



However, it is quite rare for Homoskedasticity to hold, so what we have been using is **heteroskedasticity-robust**, which is valid even if homoskedasticity does not hold

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.

- It is possible that a very small estimate is statistically significant, but not economically significant

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.

Suppose we want to find the causal effect of height on earnings, now we include another factor intelligence which also influences earnings, thus it is included in the error term

$$u_i = \beta_2 . Intelligence_i + v_i$$

- Where v_i is the error that include all other factors that affect earnings except for height and intelligence

Thus, we have the following

$$\text{corr}(u_i, \text{Height}_i) = \text{corr}(\beta_2 . \text{Intelligence}_i + v_i, \text{Height}_i)$$

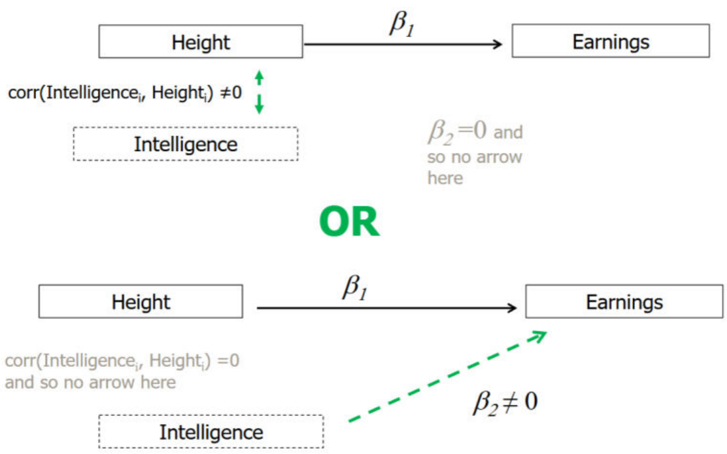
Following some steps we can also write it as

$$\beta_2 . \text{cov}(\text{Intelligence}_i, \text{Height}_i) + \text{cov}(v_i, \text{Height}_i) = 0$$

- If for now **we assume that corr(v_i, Height$_i$) = 0**, meaning there no other important variable that affects earnings

Therefore, the OLS estimator β_1 is **unbiased** if either:

- $\beta_2 = 0$
- Corr(Intelligence$_i$, Height$_i$) = 0
- Or both



Since we know that if OLS estimator is unbiased, then $E(\hat{\beta}_1) = \beta_1$, meaning the rest is the **bias**

$$E(\hat{\beta}_1) = \beta_1 + \underbrace{\beta_2 \frac{corr(X_1, X_2)}{s_{X_1}} . s_{X_2}}_{\text{bias}}$$

Direction of bias

The bias, when a simple model includes X1 and omits X2, can be positive or negative depending on the sign of β_2 and correlation between X1 and X2.

	corr$(X_1,X_2)>0$ (**if** taller people are more intelligent)	corr$(X_1,X_2)<0$ (**if** taller people are less intelligent)
$\beta_2>0$ (**if** intelligence increases earnings)	**Positive** bias	**Negative** bias
$\beta_2<0$ (**if** intelligence decreases earnings)	**Negative** bias	**Positive** bias

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

With a multiple regression model now, we will have more variables that influence Y in our model, where the population regression line represents average relationship between independent variables X_{1i}, and X_{2i} and Y_i (for 2 regressor case)

$$E(Y_i | X_{1i}, X_{2i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

When X_{1i}, goes up by ΔX_{1i}, keeping X_{2i} constant, then $E(Y_i|X_{1i}, X_{2i})$, goes up by $\Delta X_{1i} \cdot \beta_1$

$$\begin{aligned} E(Y_i \mid X_{1i} + \Delta X_{1i}, X_{2i}) &= \beta_0 + \beta_1 (X_{1i} + \Delta X_{1i}) + \beta_2 X_{2i} \\ &= \beta_0 + \beta_1 X_{1i} + \beta_1 \Delta X_{1i} + \beta_2 X_{2i} \\ &= (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}) + \Delta X_{1i} \beta_1 \\ &= E(Y_i \mid X_{1i}, X_{2i}) + \Delta X_{1i} \beta_1 \end{aligned}$$

Example interpretation

$$\widehat{Earnings} = -33046.91 + 408.5786 \times Height + 3882.779 \times Educ$$

- Keeping education fixed, one more inch is estimated to increase earnings by $408.6 on average

For k regressors the regression line will be

$$\begin{aligned} &E(Y_i | X_{1i}, X_{2i}, \dots, X_{ki}) = \\ &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} \end{aligned}$$

Where similar to the single linear regression, the OLS estimator tries to obtain $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ that minimizes the following equation

$$\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki}\right)^2$$

- Which is the sum of the squared distance between data points I's and the fitted line, but in this case for k regressors

Example predicted values

$$\widehat{Earnings} = -33046.91 + 408.5786 \times Height + 3882.779 \times Educ$$

Height = 71 inches (1.8m); Educ = 15 years → predicted earnings $54204

Height = 75 inches (1.9m); Educ = 10 years → predicted earnings $36424

Assumptions for the multiple regression model

If we are interested in obtaining an unbiased estimate (β) of the causal effect of all variables that we have in the model

1. **Zero-conditional mean**: $E(u_i|X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$, is necessary
   - As we have seem in previous examples

$E(u_i|Height_i, Educ_i) = 0$ → $corr(Height_i, u_i) = 0$

Zero conditional mean

AND

$corr(Educ_i, u_i) = 0$

However, If we are **only** interested in obtaining an unbiased estimate (β_1) of the causal effect X_1, a weaker assumption can be used

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

$E(u_i|Height_i, Educ_{2i}) = E(u_i|Educ_i)$ → $corr(Height_i, u_i) = 0$

Conditional mean independence

BUT can have

$corr(Educ_i, u_i) \neq 0$

- Suppose the ZCM does not hold, but Conditional mean independence does, this means we can interpret the effect of height on earnings as a causal effect, capturing only effect of height
- However, education is only a **partial association**, because we assume that education is not uncorrelated to the error term, meaning it captures not only the effect of education, but also other factors related to education (e.g. occupation)
- This is not a problem if our **variable of interest** is only Height, making education as a **control variable**

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).

- Meaning amount increase/decrease compared to the dropped-out dummy

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 actual values of Y.

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

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

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

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

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

$$ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \underline{Y}\right)^2$$

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:

$$\underline{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSR}{TSS}$$

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, it only measures **how well the model explains variation**, and prediction of Y, but says **nothing** about whether assumption holds

Introduction to Econometrics – IBEB – Lecture 6, week 2

Multiple regression model: Example smoking and income

For the following sections we will be using an example where we are interested in **the causal effect of income on the amount of smoking**

Suppose that we carry out the regression model and we get that

$$Cigs_i = \beta_0 + \beta_1 Income1000 + u_i$$

- There is a positive corr(Cigs, Income1000) = 0.0532
- Increase of annual income by $1000 increases number of cigarettes smoked by 0.0799 on average per day
- P-value = 0.110, so cannot reject the null hypothesis that income has no effect at 10% (2-sided)
- Model explains (R^2) 0.28% of the total variation in Cigs

A possible problem in this model is **not** very low R^2 or that effect of income is not significant, because R^2 says nothing about whether assumptions hold and also a model is not bad if X does not explain Y

- The main problem is that **if income is correlated with error term**

One possible variable could be Education, as it might both affect smoking and is correlated with income, so a possible solution is to estimate a multiple regression model including education

$$Cigs_i = \beta_0 + \beta_1 Income_i + \beta_2 Educ_i + u_i$$

In this case we take a weaker assumption, the conditional mean independence, which is necessary for unbiased estimator of β_1, as we have seem in previous lecture

$$E(u_i|Income_i,Educ_i) = E(u_i|Educ_i) \Rightarrow corr(Income_i, u_i) = 0$$

Estimating this multiple regression model gives us

$$\hat{Cigs}_i = 10.61 + 0.1174 \times Income_i - 0.3360 \times Educ$$

According to the model, an increase of annual income by 1000 dollars causes an increase in the number of cigarettes smoked by 0.1174 on average, keeping education fixed

- This means that simple model suffered from **downwards Omitted variable bias**

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: $|t^{act}| > 1.96$ or p-value < 0.05

If we carry out an hypothesis test for the multiple regression model example that we have in the section above

$$t = \frac{0.1174-0}{0.0535} = 2.19 \qquad p - value = 2\Phi(-|2.19|) = 0.28$$

- We see that not only the coefficient goes up, but also it becomes significant, evidence of OMV in simple model

Linear regression

Number of obs	=	807
F(2, 804)	=	3.34
Prob > F	=	0.0359
R-squared	=	0.0078
Root MSE	=	13.685

cigs	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
income1000	.1174301	.0535069	2.19	0.028	.0124004	.2224598
educ	-.3359798	.1622384	-2.07	0.039	-.6544407	-.0175189
_cons	10.60949	1.920332	5.52	0.000	6.840033	14.37894

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.

- If we simply use separate tests and reject H_0 if $|t_1| > 1.96$ or $|t_2| < 1.96$, then the size of this test is **not** 5%, so it's wrong to say that we reject **joint hypothesis** at 5% significance level

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.

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t1,t2}t_1t_2}{1 - \hat{\rho}_{t1,t2}^2}\right)$$

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

Special case: uncorrelated t-statistics

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2}t_1t_2}{1-\hat{\rho}_{t_1,t_2}^2}\right) = \frac{1}{2}\left(t_1^2 + t_2^2\right)$$

- So F-statistic increases with t_1 and t_2, same as above reject null if F is larger than the critical value
  - At 10% if F > 2.30, 5% if F > 3.00, 1% if F > 4.61

Special case: single restriction

$$F = t_1^2 \quad \sim F_{1,\infty}$$

- Reject H_0 if F is larger than critical value:
  - At 10% if F>2.71, 5% if F>3.84, 1% if F>6.63

P-value

$$p - value = \Pr\left[F_{q,\infty} > F^{act}\right]$$

- The p-value tell us the probability that a random variable following an F distribution with q and plus infinity degrees of freedom is larger than the F-statistic (sample) that we obtain with our test, assuming H_0 is true

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.

Note: that the q and plus infinity degrees of freedom is only an approximation that works well in large sample, Stata uses exact degrees freedom, so instead of +infinity, it is n – k – 1, where k is number of restrictions (so, 2 in this case)

```
F(2, 804)          =      3.34
Prob > F           =    0.0359
```

Testing single restrictions involving multiple coefficients

$$H_0 : \beta_1 = \beta_2 \quad \text{vs} \quad H_1 : \beta_1 \neq \beta_2$$

We can use 2 approaches to test this

- Test restriction directly using F-statistic, where number of restriction is 1 because we only have 1 "="
- Transform the regression to incorporate the restriction

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

A **linear regression model** assumes that the effect of a one-unit change in any regressor on the dependent variable is **constant**, regardless of the current value of that regressor.

The expected change in income from going from 9 to 10 years of experience is exactly β_2, and so is the change from going from 29 to 30 years. These are identical by construction.

- But realistically, early in your career, each year of experience brings big productivity gains. After 25–30 years, additional experience may add less (or even none) in terms of real productivity. This suggests a **curved** relationship between experience and wages.

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.

We are simply treating x^2 as a new variable and regressing on both x and x^2.
This creates non-linearity, because the *marginal effect* of x, how much the outcome changes when x increases by one unit, now depends on the current value of x

- Differentiate with respect to x: the marginal effect = $\alpha_1 + 2\alpha_2 \cdot x$

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

Example
The key variables are:

- **labinc** — monthly net labour income in euros
- **educ** — years of education
- **exper** — years of work experience
- **male** — a binary indicator (1 = male, 0 = female)

In the following equations you can see how coefficient difference from 7-9 years of experience is a lot higher thant he coefficient difference from 28-30 years of experience

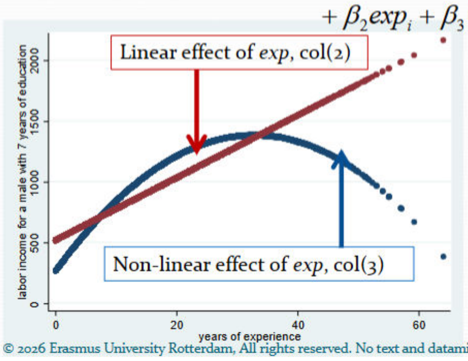
- Effect of experience on labor incomes

$$E(labinc_i | exp_i = 9) - E(labinc_i | exp_i = 7) \cong 67{,}83 * 2 - 1{,}03 * \left[9^2 - 7^2\right] \cong 103$$

$$E(labinc_i | exp_i = 30) - E(labinc_i | exp_i = 28) \cong 67{,}83 * 2 - 1{,}03 * \left[30^2 - 28^2\right] \cong 16$$

Always **plot your estimated regression function** when using non-linear models. Numbers in a table are hard to interpret intuitively; a graph immediately reveals whether the curve is U-shaped, inverted-U-shaped, accelerating

$$E(labinc_i | educ_i = 7, male = 1, exp_i) = \beta_0 + \beta_1 7 + \beta_2 exp_i + \beta_3 (exp_i)^2 + \beta_4$$



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 β_1: a 1-unit change in X corresponds to $(\beta_1 \times 100\,\%)$ change in Y (semi-elasticity).

The derivation goes as follows

$$\begin{aligned}\beta_1 &= E\left[\ln(labinc_i) | educ_i = 10\right] - E\left[\ln(labinc_i) | educ_i = 9\right] \\ &= E\left[\ln(labinc_i + \Delta labinc) - \ln(labinc_i)\right] \\ &= E\left[\ln\left(\frac{labinc_i + \Delta labinc}{labinc_i}\right)\right] \cong E\left[\frac{\Delta labinc}{labinc_i}\right]\end{aligned}$$

- In other words, β_1 captures the *proportional* change in income, not the absolute change

Linear-log model

Instead of logging the dependent variable, we can log a regressor. Consider replacing *exper* with *ln(exper)*

This captures the idea that each additional year of experience matters less as experience accumulates, the same intuition as the quadratic model, but in a different functional form.
Interpretation of β_2: A **1% increase** in experience leads to a **β_2/100 unit** change in income. If β_2 = 313.57, so a 1% increase in experience raises income by **€3.14**.

- We divide by 100, because ln(exper + 1% of exper) - ln(exper) ≈ 0.01, so the change in the outcome is $\beta_2 \times 0.01 = 313.57/100 \approx 3.1$

Log-log model

Both independent (X) and dependent (Y) variables are transformed logarithmically.

The coefficient β_2 is now a pure **elasticity**: a **1% increase in experience** leads to a **β_2 percent** change in income.

- If $\beta_2 = 0.211$ → a 1% increase in experience raises income by **0.21%**.

Note: Never compare R^2 across models with different dependent variables. The R^2 from a model with ln(labinc) as the outcome and the R^2 from a model with labinc in levels are not comparable, they measure the share of variance explained in *different quantities*.

Interaction effect

An **interaction effect** captures the idea that the effect of one variable **depends on the value of another variable**.

- We are no longer asking "what is the effect of education on income?" but rather "does the effect of education on income differ for men and women?"

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.

Suppose the following model

$$\ln(labinc_i) = \beta_0 + \beta_1 educ_i + \beta_2 exp_i + \beta_3 (exp_i)^2 + \beta_4 male_i + \beta_5 male_i * educ_i + u_i$$

And we get the following results:

Dependent variable:	ln(labinc)	labinc	ln(labinc)	ln(labinc)
Regressor	(4)	(5)	(6)	(7)
years of education	0,099** (0,002)	177,45** (5,25)	0,097** (0,003)	0,114** (0,004)
years of work experience	0,057** (0,002)			0,058** (0,002)
(years of work experience)²	-0,001** (0,000)			-0,001** (0,000)
ln(experience)		313,57** (10,13)	0,211** (0,008)	
male	0,600** (0,013)	905,89** (21,68)	0,608** (0,014)	0,939** (0,065)
male*years of education				-0,027** (0,005)
Intercept	4,960** (0,038)	-1899,46** (77,52)	5,026** (0,042)	4,773** (0,053)
Summary Statistics				
R^2	0,354	0,318	0,328	0,356
n	9656	9431	9431	9656

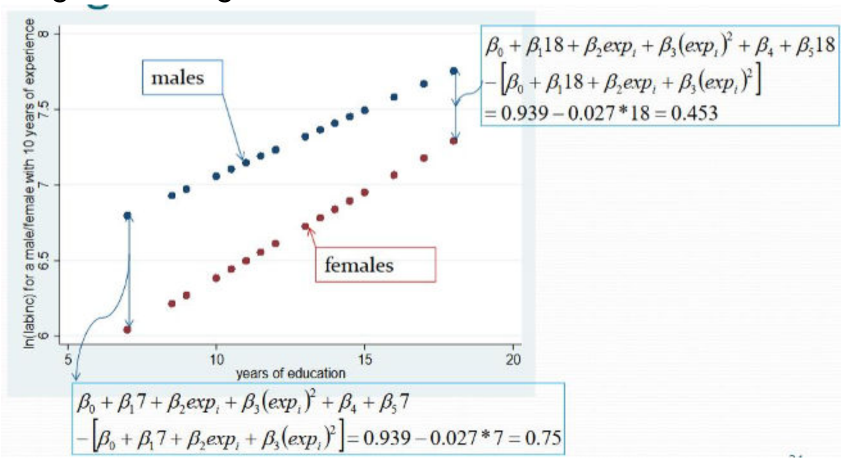
Heteroskedasticity-robust standard errors are given in parentheses under the coefficients.
The individual coefficient is significant at the *5% or **1% significance level using a two-sided test

model (4) : educ 9 → 10

- males : 0.114 - 0.027 = 0.087
- females : 0.114
- difference : -0.027

The interaction term β_5 = **-0.027** (statistically significant, t ≈ -5.4) tells us that the return to education is **2.7 percentage points lower for men than for women**

- This means women benefit more from each year of education in terms of percentage income gains than men do



The two lines have **different slopes** — the female line is steeper. The gap between the lines (male premium) starts large (at low education) and narrows as education increases, because the interaction term is negative.

When OLS fails

All the non-linear models introduced above, polynomials, log transformations, interactions, share one property: they are **linear in the parameters**.

- This means the β coefficients enter the model additively, each multiplied by some function of the data, but not multiplied by each other or raised to powers

$$E(labinc_i|\ldots) = \beta_0 + \beta_1 educ_i + exp_i^{\beta_2} + \beta_4 male_i$$

Here, β_2 appears as an exponent, the model is **non-linear in the parameter β_2**.
OLS cannot estimate this; it requires non-linear estimation methods (such as non-linear least squares or maximum likelihood).

- This distinction is crucial: **non-linearity in the variables is welcome in OLS; non-linearity in the parameters is not.**

Introduction to Econometrics – IBEB – Lecture 8, week 3

Internal validity

<u>Association is not causation</u> 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.

- That is, can we trust that our regression identifies a causal effect *within* our sample?

Using the example we had last lecture, the effect of education (*educ*) on labour income (*labinc*) using German

- The population: German workers, with their specific education levels and work experience
- The setting: the German labour market, schooling system, etc...

For internal validity, we need OLS to give us an **unbiased and consistent** estimate of the causal effect. The formal requirement is the **conditional mean independence assumption**

$$E(u_i|educ_i,\ldots)=E(u_i|\ldots)$$

- Meaning, knowing someone's education level gives us no additional information about variables in the error term, no correlation

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

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).

Example

Schooling raises cognitive ability. So IQ is not purely a background characteristic, it is partly *caused* by education itself.

- If you control for IQ, you are blocking a genuine causal pathway
- You would be asking: "what is the effect of education on income, *holding IQ fixed*?", but that is not the question we want to answer.
- We want the *total* effect of education, including the part that works through raising IQ.
- Controlling for IQ **over-controls** and gives you a misleadingly small estimate.

Your *parents'* occupation is not caused by your own education. It is a background characteristic that was determined before you made any schooling decisions.

- If parental occupation affects your income (through networks, inheritance of skills) and correlates with how much education you received (richer parents send kids to school longer), then it satisfies both OVB conditions, and it is a **legitimate control** because it is not on the causal pathway from your education to your income.

Solutions when good controls are not available, unfortunately there are no easy fixes, we mainly rely on:

- **Panel data**: observing the same individuals repeatedly over time.
- **Instrumental variables (IV)**: finding a variable that affects education but has no direct effect on income

- **Randomised controlled trials / quasi-experimental designs**: where assignment to treatment is random, breaking the correlation between the regressor and the error term entirely

Threat 2: Errors-in-variables

Errors-in-variables arises when your data does not perfectly measure the true underlying variable. In the real world, data is collected through surveys, administrative records, and self-reports,

- people misremember, lie, round numbers, or simply make mistakes when answering questionnaires
- **what does this imperfect measurement do to our OLS estimates?**

Measurement error in the independent variable

Suppose the true (unobserved) variable is X_i, someone's true years of education. But what we actually observe in our dataset is

$$X_i^r = X_i + m_i^X$$

(X_i^r: with error; X_i: without error; m_i^X: measurement error)

- Where m_i^X is the **measurement error,** the difference between what we recorded and the truth

But since X_i is unobserved, we can only estimate α_1 from

$$Y_i = \alpha_0 + \alpha_1 X_i^r + v_i$$

- When the measurement error is **non-random**, it can correlate with Y, with X, or with both. There is no general insight here, the bias in α_1 can go in any direction depending on the signs and magnitudes of those correlations.
- Since measurement error is not observable, we cannot say anything definitive about the direction of bias. It could go up or down

$$\alpha_1 = \frac{cov(Y,X^r)}{var(X^r)} = \frac{cov(Y,X+m^X)}{var(X+m^X)} = \frac{cov(Y,X)+cov(Y,m^X)}{var(X)+var(m^X)+2cov(X,m^X)}$$

- When the error is **random**, or classical measurement error, this means:

$$cov(Y,m^X) = cov(X,m^X) = 0$$

$$\alpha_1 = \frac{cov(Y,X)+\cancel{cov(X,m^X)}}{var(X)+var(m^X)+\cancel{2cov(X,m^X)}}$$

$$\alpha_1 = \frac{var(X)}{var(X)+var(m^X)}\frac{cov(Y,X)}{var(X)} = \frac{var(X)}{var(X)+var(m^X)}\beta_1$$

Look at what this tells us. The fraction $\frac{var(X)}{var(X)+var(m^X)}$ is always **between 0 and 1** because:

- The numerator $var(X)$ is always smaller than the denominator $var(X) + var(m^X)$
- Adding noise $(var(m^X))$ always makes the denominator bigger
- Meaning the estimate is always pulled toward zero

Measurement error in the Dependent variable

Now suppose the error is in the outcome variable

$$Y_i^r = Y_i + m_i^Y$$

Y_i^r: with error; Y_i: without error; m_i^Y: measurement error

We observe measured income Y_i^r instead of true income Y_i. We estimate γ_1 from

$$Y_i^r = \gamma_0 + \gamma_1 X_i + w_i$$

- For **non-random error** when the error correlates with X, bias is again unpredictable
- The bias depends entirely on the sign and magnitude of $cov(m^Y, X)$, no general conclusion is possible.

$$\gamma_1 = \frac{cov(Y^r,X)}{var(X)} = \frac{cov(Y+m^Y,X)}{var(X)} = \frac{cov(Y,X)+cov(m^Y,X)}{var(X)}$$

- For **random error**, the $cov(m^Y, X) = 0$

$$\gamma_1 = \frac{cov(Y,X)+\cancel{cov(m^Y,X)}}{var(X)} = \beta_1$$

No bias, Random measurement error in Y does not distort the coefficient estimate at all.

- Because the noise in Y is unrelated to X. It does not systematically push the regression line in any direction, it just adds scatter around the true line without tilting it.

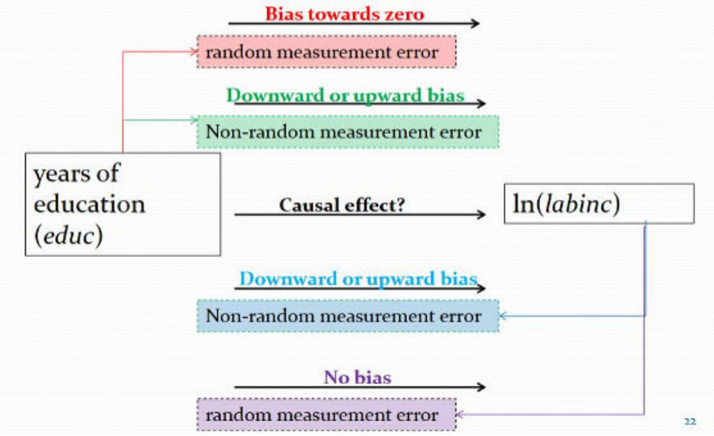
However, it does **increase variance**. The error term in the regression becomes $u_i + m_i^Y$ instead of just u_i, so:

$$var(u_i + m_i^Y) \geq var(u_i)$$

Your estimates are still unbiased but **less precise** — your standard errors are larger, confidence intervals are wider, and it becomes harder to detect statistically significant effects.

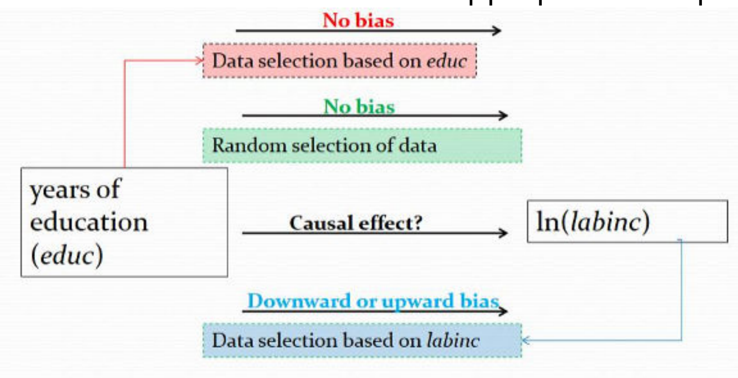
Solutions

- **Random error in Y** → no correction needed for the coefficient, though precision suffers
- **All other cases** → the standard solution is again **instrumental variables** — using an instrument that is correlated with the true X but not with the measurement error



Threat 3: Sample selection bias

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 (more detail in lecture 7). 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.

External validity

So far the entire lecture has been about **internal validity** — can we trust that our regression identifies a causal effect *within* our sample? Now we ask a completely different question

- Even if our study is perfectly internally valid — can we take those findings and apply them somewhere else?
- **External validity**: A study is externally valid if its inferences can be **generalised to other populations and settings**

It is important to check whether the population and settings are comparable for external validity, in our example

- **Population:** are the dependent and independent variables comparable, a study of German workers may not generalise to Dutch workers if:
    - The distribution of education levels is different
    - The type of jobs available differs
    - Cultural attitudes toward work and wages differ
- **Settings**: are the relationships between variables the same

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

Appropriate sampling

Appropriate sampling means obtaining a sample that is representative of the **population under study**. Why does this matter?

- It avoids **sample selection bias** — one of the key threats to internal validity.
- Random selection gives you **i.i.d. (independent and identically distributed) observations**, which is needed for OLS standard errors to be consistent and reliable

2 Essential Steps:

1. **Define the population.** You need to clearly specify *who* or *what* you're making claims about.
    - Sometimes the population literally cannot be defined, for example, homeless people have no address register to sample from, and prehistoric humans left no census records.
    - If you can't define the population, you can't draw a representative sample from it
2. **Draw a representative sample from that population.** A **representative sample** has the same distribution of characteristics as the 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**

Key takeaways

- A representative sample is essential for internal validity and for avoiding sample selection bias.
- You must first define the population, then draw a random sample from it.
- Probability sampling (random selection) allows valid statistical inference; non-probability sampling does not guarantee representativeness.
- If the population can't be defined, a representative sample is impossible

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**

$$ln(labinc_{i9}) - ln(labinc_{i8}) = \beta_0(1-1) + \beta_1(educ_i - educ_i) + \beta_2(exp_{i9} - exp_{i8}) + \beta_3(exp_{i9}^2 - exp_{i8}^2) + \beta_4(male_i - male_i) + \beta_5(educ_i male_i - educ_i male_i) + (\alpha_i - \alpha_i) + (v_{i9} - v_{i8})$$

$$ln(labinc_{i9}) - ln(labinc_{i8}) = \beta_2(exp_{i9} - exp_{i8}) + \beta_3(exp_{i9}^2 - exp_{i8}^2) + (v_{i9} - v_{i8})$$

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)

For this lecture we strip the German income equation down to its simplest form only education as a regressor

$$\ln(\text{labinc})_i = \beta_0 + \beta_1 \text{educ}_i + u_i$$

The total variation in education across people comes from **two sources**

- **Endogenous variation**, the part that's correlated with the error term.
    - For example, high-ability people choose more education *because* of their ability. This variation is "contaminated" and causes OVB.
- **Exogenous variation**, the part that's independent of the error term.
    - For example, some people got more education not because of their ability, but because a policy change forced them to stay in school longer. This variation is "clean."

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:** Regress the endogenous variable (education) on the instrument (school reform)
    - This gives you **predicted education**, the part of education that is explained purely by the school reform. Since the school reform is exogenous (independent of v, the predicted values are also clean

$$educ_i = \pi_0 + \pi_1 schoolreform_i + v_i$$

$$\hat{educ}_i = E(educ_i | schoolreform_i) = \hat{\pi}_0 + \hat{\pi}_1 schoolreform_i$$

- **2nd Stage:** Regress the dependent variable (ln wages) on the predicted education from stage 1.

$$\ln(labinc_i) = \beta_0^{TSLS} + \beta_1^{TSLS} \hat{educ}_i + e_i$$

School reform explanation:

The instrument used in the lecture is a **compulsory schooling reform** in Germany. Different German states raised the minimum school-leaving age from 14 to 15 at different times

- people who happened to live in a state that raised the school-leaving age earlier were *forced* to get more education, not because of their ability or motivation, but because of a policy change they had no control over. This is **exogenous variation** in education

Conditions for valid IV:

1. **Instrument relevance**: the instrument must actually predict the endogenous variable. In our example: the school reform must actually affect years of education
    a. This is **testable**, you just run the first-stage regression and check whether the instrument's coefficient is statistically significant

$$\operatorname{cor}(schoolreform_i, educ_i) \neq 0$$

$$educ_i = \pi_0 + \pi_1 schoolreform_i + v_i$$

2. **Instrument exogeneity**: The instrument must be unrelated to the error term in the main equation. In our example: the school reform must not affect wages through any channel *other than* education
    a. This is **NOT testable**, you can never verify it directly because the error term is unobservable. You have to argue it on logical grounds

$$\operatorname{cor}(schoolreform_i, u_i) = 0$$

$$\ln(labinc_i) = \beta_0 + \beta_1 educ_i + u_i$$

Dependent variable in 1985:	ln(labinc)	educ	ln(labinc)	ln(labinc)
Regression method	ols	ols	TSLS	ols
Regressor				
years of education	0,084** (0,004)		0,016 (0,069)	
schoolreform		0,520** (0,145)		0,008** (0,036)
Intercept	5,844** (0,047)	10,844** (0,052)	6,587** (0,749)	6,758** (0,013)
Summary Statistics				
R^2	0,110	0,004	0,038	0,000
n	2867	2867	2867	2867

Heteroskedasticity-robust standard errors are given in parentheses under the coefficients. The individual coefficient is significant at the *5% or **1% significance level using a two-sided test

$$\beta_1^{TSLS} = \frac{E[\ln(labinc_i)|schoolreform_i]}{E[educ_i|schoolreform_i]} = \frac{0.008}{0.520} = 0.016$$

30

There's also a **shortcut** to compute the TSLS coefficient. You can run a simple regression of ln(labinc) on schoolreform directly, which gives a coefficient of 0.008.

- Then divide it by the coefficient of the regression of education on shoolreform

OLS estimate > TSLS estimate

OLS Estimate:

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

IV Estimate:

- Internally valid (assuming relevance and exogeneity hold)
- Exploits only a *tiny* share of total variation in education — just the part driven by the school reform
- Externally valid only for the **specific group** whose education was changed by the reform

Supply & demand example

The lecture introduces another application of IV: the classic **supply and demand identification problem**.

Suppose you observe equilibrium prices and quantities of butter over 11 time periods. You plot them on a graph and see a scattered cloud of points. Can you draw a demand curve through those points?

- **No.** Because each point is an *equilibrium*, the intersection of supply and demand. Both curves are shifting over time.
- If supply shifts right (more butter available), price falls and quantity rises, you move along the demand curve.
- But if demand also shifts right (people want more butter), price rises and quantity rises, you move along the supply curve.
- With both curves shifting simultaneously, the data points trace out neither curve.

This is **simultaneous causality bias**: price affects quantity demanded, but quantity supplied also affects price. OLS can't disentangle which curve you're estimating.

- **IV solution:** Find an instrument that shifts *only one* curve. For estimating the demand curve, you need something that shifts supply but not demand.

The lecture's example: **rainfall in dairy-producing regions**.

- **Relevance:** Less rain → less grass → fewer cows → less butter for any given price → supply curve shifts left. Rain clearly affects butter supply.
- **Exogeneity:** Rain in dairy regions shouldn't directly affect how much butter consumers *want* at a given price. It only affects demand *through* its effect on supply (via price changes).

When only the supply curve shifts (driven by rainfall), the equilibrium points trace out the demand curve, because you're sliding along a stable demand curve as supply moves. The IV isolates supply-driven price variation to identify the demand relationship

Summary of OV and TSLS

IV/TSLS is the most general tool for addressing failures of conditional mean independence. It can handle omitted variable bias, errors-in-variables, sample selection bias, *and* simultaneous causality — all with the same framework.

The two conditions, relevance (testable) and exogeneity (not testable), must both hold.

- Finding good instruments requires economic theory, institutional knowledge, and creative thinking.
- The best instruments often come from **quasi-experiments**: natural or policy variation that randomly or quasi-randomly shifts the endogenous variable.

The major downside: you can never prove exogeneity

Introduction to Econometrics – IBEB – Lecture 10, Week 4

Experiments vs. Quasi-experiments vs. Association

Experiments

- Treatment assigned randomly
- On purpose
- Used in economics but more common in other disciplines (example: psychology)
- **Example:** The **Philadelphia Foot Patrol Experiment (2009)** randomly assigned extra police foot patrols to some city blocks but not others. Because the assignment was random and deliberate, researchers could credibly claim that differences in crime rates were *caused by* the foot patrols, not by pre-existing differences between neighbourhoods

Quasi Experiment

- Treatment assigned "as good as random"
- NOT on purpose
- More often used in economics
- **Example:** The **COPS program** allocated federal grants for police hiring across U.S. cities. Researchers have determine whether cities decide to participate in the program or not, if that happened in an "as if random" way. If yes, they used this "as-if-random" variation to study how police spending affects crime, even though no one designed the grant allocation as an experiment

Association

- Treatment assigned non-randomly
- NOT on purpose
- Common in ALL disciplines
- **Example:** A scatter plot of violent crime rate vs. police spending per capita across U.S. cities shows a *positive* correlation, cities that spend more on police also tend to have more crime. Does that mean police *cause* crime?
    - Of course not. Cities with more crime *choose* to spend more on police. The treatment assignment (police spending) is driven by the

very thing we're trying to measure (crime), so the association is contaminated by reverse causality and omitted variables.

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.

The example: does getting a mammogram reduce a woman's chance of dying from breast cancer?

- To answer this *for a single individual*, you'd need to observe two things simultaneously, her mortality if she gets the mammogram, and her mortality if she doesn't.
- But that's impossible. A woman either gets screened or she doesn't; you can never observe both realities for the same person at the same time

The **individual causal effect** would be $y_{1i} - y_{0i}$. But we can only ever observe one of these — whichever actually happened

- $\mathbf{y_{1i}}$ = the outcome (e.g., mortality) for person i if she receives the treatment (mammogram)
- $\mathbf{y_{0i}}$ = the outcome for person i if she does not receive the treatment

Since we can't measure individual causal effects, we aim for the next best thing: the **Average Treatment Effect (ATE)**, which is the expected difference in potential outcomes across the entire population

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	$E(y_{1i} \mid t_i = 1)$	$E(y_{0i} \mid t_i = 0)$
	Unobserved	$E(y_{0i} \mid t_i = 1)$	$E(y_{1i} \mid t_i = 0)$

In the diagram above, it says that we can observe both the y_{1i} and y_{0I}, as indicated by the blue and red arrow, however nothing on the red and green arrow

- We Will see that only using the red and blue arrow terms, we can get the ATE

If treatment is assigned **randomly**, then whether you end up in the treatment group or control group is completely independent of your characteristics. Mathematically, random assignment guarantees:

$$E(y_{0i}|t=1) = E(y_{0i}|t=0)$$
$$E(y_{1i}|t=1) = E(y_{1i}|t=0)$$

- Meaning on average the control and treatment group are identical
- Following this we can derive the following equation

$$\begin{aligned} E(y_{1i} - y_{0i}) &= P(t=1)[E(y_{1i}|t=1) - E(y_{i0}|t=1)] \\ &\quad + P(t=0)[E(y_{1i}|t=0) - E(y_{0i}|t=0)] \\ &= P(t=1)[E(y_{1i}|t=1) - E(y_{i0}|t=0)] \\ &\quad + P(t=0)[E(y_{1i}|t=1) - E(y_{0i}|t=0)] \\ &= E(y_{1i}|t=1) - E(y_{0i}|t=0) \end{aligned}$$

This is just the difference in observed average outcomes between the treatment and control groups

- This theoretical result translates directly into a very simple regression, the difference estimator

$$y_i = \beta_0 + \beta_1 t_i + u_i$$

- y_i is the outcome (e.g., mortality)
- t_i is a binary treatment indicator (1 = got mammogram, 0 = didn't)
- β_1 is the estimated ATE

Most crucial condition that random assignment provides is:

$$\text{Random assignment} \rightarrow E(u_i | t_i) = 0$$

- Because t is randomly assigned it must be that it is uncorrelated with the error terms
- there's no omitted variable bias, no selection bias, no simultaneous causality
- So β_1 from this simple regression is an **unbiased estimate of the causal effect** of treatment on outcome.

Threats to Internal and External Validity of experiments

Randomization through experiments gives you a beautiful, unbiased starting point. The threats below are all the ways that starting point can get corrupted between the design phase and the final estimate.

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.

For **small samples**, random assignment cannot guarantee that control and treatment group have the same characteristics on average

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)

We can use whether to buy a DVD or not buy a DVD and regress it with age

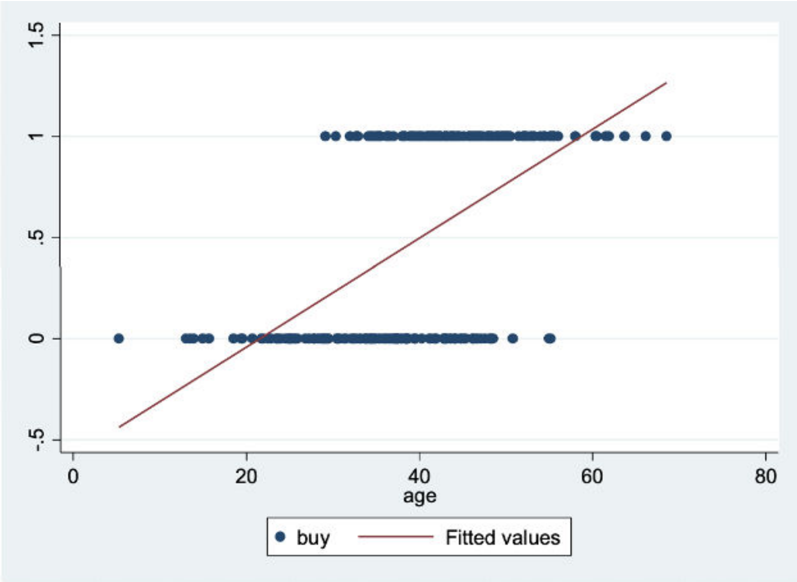
- Abnormal: all points are clustered horizontally around 1 and 0
- It is of vital importance to incorporate heteroskedastic-robust estimates.
- The predicted dependent variable of this model reflects the probability of Y=1.
- The β_1 reflects the change in probability Pr(Y=1) that occurs corresponding 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.
    - This is because the Expected value of buy (Y) is just probability of buy

$$E[buy] = \Pr(buy = 1) * 1 + \Pr(buy = 0) * 0$$
$$E[buy] = \Pr(buy = 1) + 0$$
$$E[buy] = \Pr(buy = 1)$$

- So if we had the following regression output, thus for a 50 year old

$$Buy = -0.58 + 0.027 * 50 = 0.77$$
$$\Pr(buy) = -58\% + 2.7\% * 50 = 77.0\%$$

- This tells us the probability of buying, given age = 50



Probit Model

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

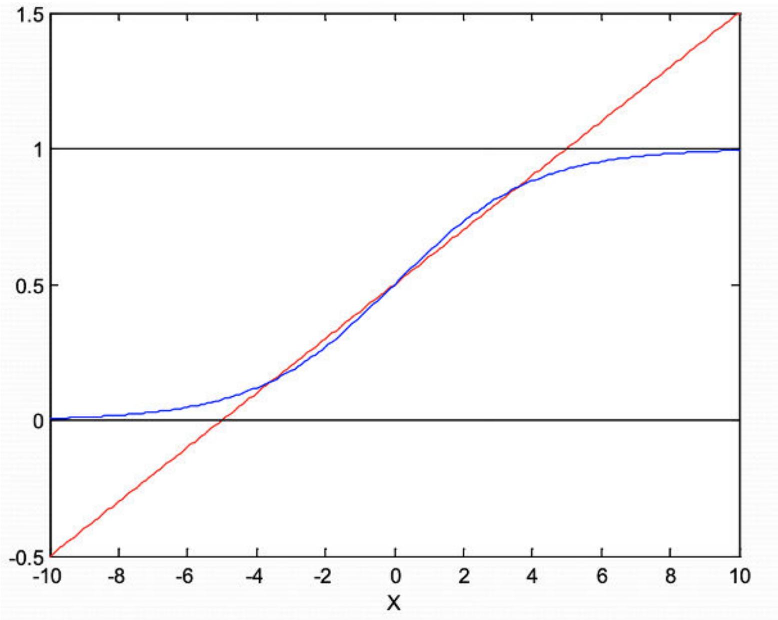
Probit models:

- 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.

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



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 β_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.

Examples

Age 50

$$\Pr(buy) = \frac{1}{1+e^{-(\beta_0+\beta_1 * age)}} = \frac{1}{1+e^{-(-6.77+0.168*50)}} = 0.836$$

Age 51

$$\Pr(buy) = \frac{1}{1+e^{-(\beta_0+\beta_1 * age)}} = \frac{1}{1+e^{-(-6.77+0.168*51)}} = 0.858$$

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

Key takeaways

- In practice, LPM, probit, and logit give **similar predicted probabilities**, especially in the middle of the data range.
- The key difference is **effect sizes**: constant for LPM, non-constant for probit/logit. This matters for interpretation
- Probit and logit are estimated via **Maximum Likelihood**, which reports Pseudo R^2, log-likelihood, and Wald chi^2 instead of the familiar OLS statistics.
- When the dependent variable has **more than two categories**, you extend to multinomial logit/probit (for unordered choices) or ordered probit (for ordered choices).

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.

There are three fundamental questions you can ask with time series data:

1. **Properties of the series itself:** Is Donald improving over time? Is there a trend? This is about understanding the characteristics of the data, does it go up, go down, cycle, or stay flat?
2. **Interaction between time series:** Did hiring a new advisor actually help? Here you're asking whether one time series (a policy change) *causes* a change in another (approval ratings). This is about relationships between variables over time.
3. **Forecasting:** What will his approval be next month? Can we use the historical pattern to predict future values? This is the forward-looking question.

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

Notation

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

- Y_t is a value of Y at time t (similarly X_t), t=1,2, … ,T
- $\{Y_1, Y_2, \ldots, Y_T\}$ is time series of Y
- 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.

Example 1
From August (month 8) onwards, Donald hired a new advisor named Jim. To test whether this made a difference, we create a **dummy variable**

$$Y_t = \beta_0 + \beta_1 \text{New\_advisor}_t + \varepsilon_t$$

The Stata regression output gives:
- **β_1 = -10.6** (p = 0.006), meaning approval ratings dropped by about 10.6 percentage points after the new advisor was hired

Example 2
We can also run a regression with time itself as the explanatory variable, we ask whether the approval rates improve over time

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

The regression output gives:
- **β_1 = -1.06** (p = 0.087), meaning approval drops by about 1 percentage point per month (not statistically significant at 5%)

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.

You can extend this to include more lags. A **second-order autoregression (AR(2))** is

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

Notation of Differences

They are important because it measures changes

- First difference: $\Delta(Y_t) = Y_t - Y_{t-1}$
- Double first difference(growth rate of the growth rate:
$$\Delta^2(Y_t) = \Delta\Delta(Y_t) = \Delta(Y_t - Y_{t-1}) = Y_t - 2Y_{t-1} + Y_{t-2}$$
- 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
$$100 \times \left(ln(Y_t) - ln(Y_{t-12})\right)$$

Worked Example: Annualized Growth in 2012Q4

- Given quarterly GDP data with 2012Q3 = 100 and 2012Q4 = 102:
- Annualized growth = 100 × 4 × ln(102/100) = 100 × 4 × 0.02 = **8%**
- Forgetting to multiply by 4 (for quarterly data) → gives 2% (wrong)
- Multiplying by 12 instead of 4 → gives 24% (wrong, that's for monthly data)
- Using the wrong base period → 100 × ((102/51) - 1) = 100% (wrong comparison entirely)

Autocorrelation

Autocorrelation is the correlation of a variable with its own past values.

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 $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon_t$, β_1 is the first-order partial autocorrelation and β_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 ε_t then we can estimate ρ from the equation $\varepsilon_t = \rho_1 \varepsilon_{t-1} + u_t$ using OLS.

If we don't have ε_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 β_0 and β_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

Formal test for first-order autocorrelation in the residuals

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

d is approximately 2 – 2p, meaning d ranges from 0 to 4
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

When ρ_1 > 0, the errors are positively correlated, which means $Cov(\varepsilon\_t, \varepsilon\_\{t-1\}) > 0$. OLS assumes this covariance is zero. The result:

- Coefficients ($\hat{\beta}$) are still unbiased, because E(ε_t) = 0 still holds
- Standard errors are wrong, OLS underestimates them, making t-statistics too large and p-values too small. You get false significance

Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors, introduced by Newey and West, correct for both heteroskedasticity *and* autocorrelation in the error term. They produce valid standard errors even when errors are correlated over time.

Introduction to Econometrics– IBEB – Lecture 13, week 6

Dynamic Models

Recall from Lecture 12 that when you estimate a simple regression:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

You may find that the residuals ε_t are **autocorrelated**, that is, ε_t is correlated with ε_{t-1}. The pattern is:

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

When this happens, your **standard errors are wrong** (even though the coefficient estimates themselves aren't biased), and Lecture 12 taught you to fix this using **HAC (Newey-West) standard errors**

- But if it happens, for forecasting purposes it is better to modify/improve the model, which is what we will be doing

Autoregressive Models (AR)

When using the past observations of the variable itself in a regression.

- The order of an AR is the **maximum lag** (p) of the equation 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 β_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 β_0 is the intercept or constant and β_1 is the coefficient of Y_{t-1})

Autocorrelation

Property of the AR(1)model is how autocorrelations decay with lag length

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

We can see the derivation here:

$Y_t = \beta Y_{t-1} + \varepsilon_t$ (Note: correlation Y_{t-1} and ε_t is 0)

$$Cov(Y_t Y_{t-1}) = Cov(\beta Y_{t-1} + \varepsilon_t, Y_{t-1}) = \beta Var Y_{t-1} = \beta Var Y_t$$
$$Corr Y_t Y_{t-1} = \frac{Cov(Y_t Y_{t-1})}{Var Y_t} = \beta$$

$$Cov(Y_t Y_{t-2}) = Cov(\beta Y_{t-1} + \varepsilon_t, Y_{t-2}) = \beta Cov(Y_{t-1} Y_{t-2}) = \beta Cov(Y_t Y_{t-1})$$

$$Corr(Y_t Y_{t-2}) = \frac{Cov(Y_t Y_{t-2})}{Var Y_t} = \frac{\beta Cov(Y_t Y_{t-1})}{Var Y_t} = \beta^2$$

Note: choose AR(p) model for the highest significant order found from the partial autocorrelation, meaning if AR(2) is significant and AR(3) is not, then choose AR(2)

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 this also increases the complexity of the model

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

So far, AR models only used **past values of Y** to explain Y. But in many real-world situations, Y is also influenced by some **external variable X**, and crucially, X may affect Y not just immediately, but also **with a delay**

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

The model is called "distributed" because the effect of X is **spread out (distributed) across multiple time periods**, and "finite" because it only goes back q periods

The Delay Multiplier (s-period delay multiplier)
- If X changes by 1 unit at time t-s, how much does Y change at time t?
- For example, if $\beta_2 = 0.2$, it means: a one-unit increase in X two periods ago raises Y today by 0.2 units.

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

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

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)

Put lagged Y and lagged X on the right-hand side simultaneously. This gives us the **ARDL model**, the most general and realistic dynamic model in this lecture.

ARDL model of order (p, q)

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

- p = number of lags of Y included
- q = number of lags of X included
- β coefficients belong to the **autoregressive part** (lagged Y)
- δ coefficients belong to the **distributed lag part** (current and lagged X)

Long-Term Value

Long term expectation for AR(p) model:

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

Use: $E[Y_t] = E[Y_{t-1}] = E[Y_{t-2}] = \ldots = E[Y_{t-p}]$

$$E[Y_t] = E[\beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \varepsilon_t]$$
$$E[Y_t] = E[\beta_0] + E[\beta_1 Y_{t-1}] + \ldots + E[\beta_p Y_{t-p}] + E[\varepsilon_t]$$
$$E[Y_t] = \beta_0 + \beta_1 E[Y_{t-1}] + \ldots + \beta_p E[Y_{t-p}] + 0$$
$$E[Y_t] = \beta_0 + \beta_1 E[Y_t] + \ldots + \beta_p E[Y_t] + 0$$
$$E[Y_t] - \beta_1 E[Y_t] - \ldots - \beta_p E[Y_t] = \beta_0$$
$$E[Y_t](1 - \beta_1 - \ldots - \beta_p) = \beta_0$$
$$E[Y_t] = \frac{\beta_0}{(1 - \beta_1 - \ldots - \beta_p)}$$

Extend to ARDL(p,0) model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \delta_0 X_t + \varepsilon_t$$

- Long term expection of Y_t

$$E(Y_t) = \frac{\beta_0 + \delta_0 X_t}{(1 - \beta_1 - \cdots - \beta_p)}$$

- The long-term expectation of Y_t if X permanently goes up with 1:

$$\frac{\beta_0 + \delta_0 (X_t + 1)}{(1 - \beta_1 - \cdots - \beta_p)} = E(Y_t) + \frac{\delta_0}{(1 - \beta_1 - \cdots - \beta_p)}$$

Extend to ARDL(p, q) model:

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

- Long terms expection of Y_t:

$$E(Y_t) = \frac{\beta_0+\delta_0 X_t+\delta_1 X_{t-1}+\cdots+\delta_q X_{t-q}}{(1-\beta_1-\cdots-\beta_p)}$$

- The long-term expectation of Y_t if X permanently goes up with 1:

$$E(Y_t) + \frac{\delta_0 + \delta_1 + \cdots + \delta_q}{(1 - \beta_1 - \cdots - \beta_p)}$$

Short-Term Effects ARDL

- Not easy to see

Error Correction Format:

Write this using Δ_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 δ_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)

Recall from Lecture 13 the **Distributed Lag (DL)** model, which said that the current value of Y depends on the current and past values of some observed variable X. Now imagine replacing X with something you can never directly observe: the **error terms** (also called shocks or innovations) themselves.

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

- where ε_t is **white noise**: it has $E[\varepsilon_t] = 0$, mean zero, $E[{\varepsilon_t}^2] = \sigma^2$ constant variance σ^2, and $E[\varepsilon_t, \varepsilon_{t-i}] = 0$, zero covariance across time (i.e., each shock is completely unpredictable and uncorrelated with any other shock)

This model represents lags of ε_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**.

We identify the MA model by looking at when the Covariance suddenly become 0, we can see this in the following example

$$Y_t = \varepsilon_t + \theta \varepsilon_{t-1}$$
$$Y_{t-1} = \varepsilon_{t-1} + \theta \varepsilon_{t-2}$$
$$Y_{t-2} = \varepsilon_{t-2} + \theta \varepsilon_{t-3}$$

$$\begin{aligned} &CovY_tY_{t-1} \\ &= Cov(\varepsilon_t \varepsilon_{t-1}) + \theta Cov(\varepsilon_t \varepsilon_{t-2}) + \theta Cov(\varepsilon_{t-1} \varepsilon_{t-1}) \\ &+ \theta^2 Cov(\varepsilon_{t-1} \varepsilon_{t-2}) = 0 + 0 + \theta \sigma_\varepsilon^2 + 0 \neq 0 \end{aligned}$$

$$\begin{aligned} &CovY_tY_{t-2} \\ &= Cov(\varepsilon_t \varepsilon_{t-2}) + \theta Cov(\varepsilon_t \varepsilon_{t-3}) + \theta Cov(\varepsilon_{t-1} \varepsilon_{t-2}) \\ &+ \theta^2 Cov(\varepsilon_{t-1} \varepsilon_{t-3}) = 0 + 0 + 0 + 0 = 0 \end{aligned}$$

ARMA (p,q) Model

ARMA(p,q) simply combines both AR and MA components

- 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.

- Instead of modelling Y_t, you model ΔY_t = Y_t - Y_{t-1} (the change in Y)
- where d is the degree of 'differencing' and I is integration of order d.

$$\Delta^d Y_t = \beta_0 + \beta_1 \Delta^d Y_{t-1} + \ldots + \beta_p \Delta^d Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

Forecasting

If your model describes how Y_t depends on past values of Y and X, then at time T (the **forecast origin**, last period for which you have real data), you can form a prediction of Y_{T+1} by plugging in what you know and replacing what you don't know with its expected value.

- The unknown future error term always drops out: $E[\varepsilon_{T+1}] = 0$

One step ahead forecast from AR(1) Model

- At forecast origin T, we can observe Y_t, then we can predict Y_{t-1}

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

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

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

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

Multiple step ahead forecast from AR(1)

- Now that we have the forecast for Y_{t-1} we can use this to get the 2 step ahead forecast, and using that result we can get the 3 step ahead forecast and so on

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

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

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

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 ε) 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.

- It is positive when you under-predicted and negative when you over-predicted. A single forecast error doesn't tell you much, you need to evaluate performance across many forecast

$$Forecast error = Y_{T+1} - \hat{Y}_{T+1}$$

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**

There are actually two sources of uncertainty in any forecast:

1. **Uncertainty from the error term**, even if you knew the true model parameters perfectly, the future shock ε_{T+1} is genuinely random and unknowable.
2. **Uncertainty from parameter estimates**, your β̂ values are estimated with sampling error, not the true β values.

This is why the **out-of-sample MSFE is always larger than the in-sample MSE** (the residual variance from your regression fit). The in-sample fit only accounts for error-term uncertainty; the out-of-sample MSFE additionally reflects parameter estimation uncertainty

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:

$$[\hat{Y}_{T+1} - 1.96 \times \hat{\sigma};\ \hat{Y}_{T+1} + 1.96 \times \hat{\sigma}\,]$$

The Variance

For AR(1) this is simply:

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

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):

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

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

Pseudo-Out-of-Samples

You want to pick the best model *before* the future arrives. You'd like to compare MSFE values across competing models, but MSFE requires actual future data.

- The solution: **pseudo-out-of-sample forecasting**.

The idea is to pretend you are standing at an earlier point in time. If your data runs from 2001 to 2016:

1. Use data from 2001–2012 to estimate the model ("training sample").
2. Use the estimated model to forecast 2013, 2014, 2015, 2016 (the "pseudo out-of-sample" period).
3. Since you actually *know* what happened in 2013–2016, compute the forecast errors and MSFE.
4. Repeat for each competing model.
5. Choose the model with the lowest pseudo-out-of-sample MSFE

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

In a standard cross-sectional regression, if you find that A and B are correlated, you genuinely cannot tell which direction the causation runs, or whether some third variable C drives both. The correlation could mean A → B, or B → A, or C → A and C → B simultaneously.

- **Time series offers something extra:** time only moves forward. If A happened before B, then at least B cannot have caused A (the future cannot cause the past)

Example

$$\text{Model is } Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \varepsilon_t$$

Granger causality Wald tests

Equation	Excluded	chi2	df	Prob > chi2
inflation	unemployment	3.0133	1	0.083
inflation	ALL	3.0133	1	0.083
unemployment	inflation	5.2276	1	0.022
unemployment	ALL	5.2276	1	0.022

- Panel 1 (top row): H_0: Unemployment does NOT Granger-cause Inflation. p = 0.083 > 0.05 → Fail to reject H_0. Unemployment does not significantly help predict future inflation (at 5% level).
- Panel 2 (bottom row): H_0: Inflation does NOT Granger-cause Unemployment. p = 0.022 < 0.05 → Reject H_0. Inflation does significantly help predict future unemployment. Inflation Granger-causes unemployment.

Granger causality is not true causality

- It can be that both variables are affected by another time-varying variable C
- **We'd ideally want:** A **randomised controlled trial**, randomly assign different "doses" of X to different units at different times

(Semi) Solution:

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

Introduction to Econometrics – IBEB – Lecture 15, week 7

Non-Stationarity

In all previous lectures, we assumed that our time series Y_t is **stationary.** The most visually obvious form of non-stationarity is a **trend**, when the long-run expected value of a series is not constant but instead rises (or falls) over time.

Spurious Regression

A **spurious regression** occurs when you regress one non-stationary series on another, and you find a "statistically significant" relationship that is entirely an artefact of shared trends, rather than any true causal or structural connection between the variables

- Does AXE deodorant sales cause world population growth? Of course not
- The real culprit is a shared **trend**: both world population and AXE sales grew over 1986–2012. When you regress one trending series on another, you almost always find a "significant" relationship, not because the variables are causally linked, but because they both went up over time

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

Recall from Lectures 12–14 that the AR(1) model is:

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

For this model to be **stationary**, we required the key condition: **$-1 < \beta_1 < 1$**.

- As β_1 moves from 0.5 → 0.9 → 1.0, the plot changes from tight oscillation around a mean, to sluggish but eventual return, to complete wandering

Setting $\beta_1 = 1$ (and no intercept for now) gives the **random walk**:

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

We can unroll this recursively to understand what Y_t really is:

$$Y_{t-1} = Y_{t-2} + \varepsilon_{t-1}$$

$$\Rightarrow Y_t = Y_{t-2} + \varepsilon_t + \varepsilon_{t-1}$$

Continuing all the way back to the starting value Y_0:

$$Y_t = Y_0 + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \cdots + \varepsilon_1 = Y_0 + \sum_{i=1}^{t} \varepsilon_i$$

- **Y_t is literally the accumulated sum of all past shocks.** Every single shock since time 0 is permanently embedded in the current value..

The Expected Value:

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

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
- This violates the constant-variance requirement for stationarity. The random walk is **non-stationary** because its variance is time-dependent and explodes over time.

$$Var(Y_t) = Var(Y_0) + Var(\sum_{i=1}^{t} \varepsilon_i) = 0 + \sum_{i=1}^{t} Var(\varepsilon_i) = t\sigma^2$$

Non-stationary!

Drift (trends)

When there is a constant term or the intercept α.

- The model is then:

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

$$Y_{t-1} = \alpha + Y_{t-2} + \varepsilon_{t-1}$$

$$Y_t = \alpha + \boldsymbol{\alpha} + \mathbf{Y_{t-2}} + \boldsymbol{\varepsilon_{t-1}} + \varepsilon_t$$

$$Y_{t-2} = \alpha + Y_{t-3} + \varepsilon_{t-2}$$

$$Y_t = \alpha + \alpha + \boldsymbol{\alpha} + \mathbf{Y_{t-3}} + \boldsymbol{\varepsilon_{t-2}} + \varepsilon_{t-1} + \varepsilon_t$$

....

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

The Expected Value:

- Now the mean is NOT constant, it changes with t. This produces an **upward trend** (if $\alpha > 0$) or a **downward trend** (if $\alpha < 0$)

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

The Variance:

- Same as the one without drift

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

Dickey-Fuller Test

It is difficult to tell from a plot alone whether a series is a random walk or just a stationary AR(1) with a high β_1 close to 1. The plots of AR(1) with $\beta_1 = 0.9$ and a random walk with $\beta_1 = 1$ look almost identical visually, both wander for long stretches. We need a formal statistical test

- **The idea**: in the AR(1) model $Y_t = \beta Y_{t-1} + \varepsilon_t$, test whether $\beta = 1$
- If $\beta = 1$ → random walk → **non-stationary**. If $\beta < 1$ → mean-reverting → **stationary**
- This is called a **unit root test**
- **Important**: standard statistical tools are not applicable as mean and variance are not constant over time in this case

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 γ:

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

where $\gamma = \beta - 1$

Hypotheses:

$$H_0: \gamma = 0 \text{ the time series is a random walk } (\beta = 1)$$
$$H_1: \gamma < 0 \text{ the time series is stationary } (\beta < 1)$$

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)

The basic DF test assumes the process generating Y_t is exactly AR(1). But what if Y_t is actually AR(2) or AR(p)?
- The **Augmented Dickey-Fuller (ADF) test** fixes this by adding lagged first-differences

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

Derivation example with AR(2)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon_t$$
$$Y_t - Y_{t-1} = \beta_0 + (\beta_1 + \beta_2 - 1)Y_{t-1} - \beta_2(Y_{t-1} - Y_{t-2}) + \varepsilon_t$$
$$\Delta Y_t = \alpha + \gamma Y_{t-1} + \delta \Delta Y_{t-1} + \varepsilon_t$$

Solutions to Non-Stationarity

Once you've confirmed non-stationarity, you have two main solutions depending on the **source** of the non-stationarity

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.

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

A **structural break** occurs when the regression function itself changes at some point in time τ. The model's coefficients (intercept, slopes) are not the same before τ and after τ

- The relationship between **inflation and unemployment** before vs. after World War II
- The relationship between **GBP/EUR exchange rate and UK exports** before vs. after Brexit

If a break occurs at time τ, the model is literally two different models:

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

Rather than estimating two separate regressions, we use a **dummy variable**:

$$D_t = \begin{cases} 1 & \text{if } t \geq \tau \\ 0 & \text{if } t < \tau \end{cases}$$

The single combined regression is:

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

If t < τ

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + 0\gamma_0 + 0\gamma_1 Y_{t-1} + 0\gamma_2 X_{t-1} + \varepsilon_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \varepsilon_t$$

If t ≥ τ

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

So,

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

Where,

$$\beta'_0 = (\beta_0 + \gamma_0), \quad \beta'_1 = (\beta_1 + \gamma_1), \quad \delta'_1 = (\delta_1 + \gamma_2)$$

Testing for a break = testing jointly whether $\gamma_0 = \gamma_1 = \gamma_2 = 0$ using an F-test

Chow Break Test

If you **know the break date τ in advance** (e.g., Brexit was a specific date, WW2 ended in 1945), use the **Chow test**:

1. Create the dummy variable D_t as defined above.
2. Run the full regression with all D-interactions.
3. Conduct an **F-test for the joint significance** of γ_0, γ_1, γ_2.

H_0: No structural break ($\gamma_0 = \gamma_1 = \gamma_2 = 0$) **$H_1$**: At least one parameter changed at τ

Quandt Likelihood test (QLR)

The **QLR test** is the appropriate test when the break date is unknown:

1. For every candidate break date τ in the test sample, run the Chow regression and compute the F-statistic.
2. Take the **maximum F-statistic** across all candidate dates — this is the QLR statistic.
3. The break is most likely located near the date with the **maximum F-value**.
4. Compare the QLR statistic to special critical values (not the standard F critical values).

Note: you cannot perform QLR to find the date of the structural break and then test again with Chow test.

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!
- can we model σ^2_t as a time series itself, rather than just correcting standard errors?

ARCH Model

Consider a simple model of financial returns where there are no predictors (you can't forecast returns, but you can forecast their riskiness)

$$Y_t = \beta_0 + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma_t^2)$$

The insight of Robert Engle is that **a large variance is probably correlated with the size of past errors**. If last period's return was a big shock (large $|\varepsilon_{t-1}|$), then this period is also likely to be volatile

- So, we move from the default to the extension

$$\begin{aligned} &\text{Default:} && \sigma_t^2 = \alpha_0 \\ &\text{Extension:} && \sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \end{aligned}$$

The AutoRegressive Conditional Heteroskedasticity (ARCH) model

Without ARCH:

$$\varepsilon_t \sim N(0, \alpha_0)$$

With ARCH(1):

$$\varepsilon_t \sim N(0,\ \alpha_0 + \alpha_1 \varepsilon_{t-1}^2)$$

The distribution of today's error is still normal, but its **spread (variance) depends on what happened last period**. After a big shock, the distribution fans out; after a quiet period, it tightens

- This can be seem by how the data not fitting the normal distribution, where we have large tails

Test if Variance is ARCH

We want to test whether $\alpha_1 = 0$ (no ARCH) vs. $\alpha_1 \neq 0$ (ARCH present).
Problem: We cannot directly observe σ^2_t or ε^2_{t-1}.
Solution: Use the **squared residuals** from your regression as proxies:

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

With that, we want to know if the α_1 is significant.

To estimate an ARCH model the maximum likelihood method is used that simultaneously estimates $\hat{\sigma}^2_t$ 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 σ^2_t 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): $\sigma^2_t = \alpha_0 + \alpha_1 \varepsilon^2_{t-1}$
GARCH(1): $\sigma^2_t = \alpha_0 + \alpha_1 \varepsilon^2_{t-1} + \varphi_1 \sigma^2_{t-1}$

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. (2026). Lecture 1: *Methods 1* [PowerPoint slides]. Retrieved from: Mon - Organization of the course and Introduction: Introduction to Econometrics

Bago d'Uva, T. (2026). Lecture 2: *Methods 2* [PowerPoint slides]. Retrieved from: Tue - Methods & Exercises - OLS: simple linear regression estimation and goodness of fit: Introduction to Econometrics

Bago d'Uva, T. (2026). Lecture 3: *Methods 3* [PowerPoint slides]. Retrieved from: Wed - Methods & exercises - OLS: simple linear regression assumptions and properties: Introduction to Econometrics

Bago d'Uva, T. (2026). Lecture 4: *Methods 4* [PowerPoint slides]. Retrieved from: Mon - Methods & Exercises videos - OLS: simple linear regression hypothesis tests and confidence intervals: Introduction to Econometrics

Bago d'Uva, T. (2026). Lecture 5: *Methods 5* [PowerPoint slides]. Retrieved from: Tue - Methods & Exercises videos - OLS: OVB, multiple linear regression, assumptions; goodness of fit: Introduction to Econometrics

Bago d'Uva, T. (2026). Lecture 6: *Methods 6* [PowerPoint slides]. Retrieved from: Wed - Methods & Exercises videos - OLS: hypothesis tests, confidence intervals and model specification: Introduction to Econometrics

Van Ourti, T. (2026). Lecture 7: *Methods 7* [PowerPoint slides]. Retrieved from: Tue - Methods & Exercises videos - Nonlinear regression functions: Introduction to Econometrics

Van Ourti, T. (2026). Lecture 8: *Methods 8* [PowerPoint slides]. Retrieved from: Wed - Methods & Exercises videos - Internal and external validity: Introduction to Econometrics

Van Ourti, T. (2026). Lecture 9: *Methods 9* [PowerPoint slides]. Retrieved from: Mon - Methods & Exercises videos - Sampling, panel data, and IV: Introduction to Econometrics

Van Ourti, T. (2026). Lecture 10: *Methods 10* [PowerPoint slides]. Retrieved from: Tue - Methods & Exercises videos - Experiments and Quasi-experiments: Introduction to Econometrics

Van Ourti, T. (2026). Lecture 11: *Methods 11* [PowerPoint slides]. Retrieved from: Tue - Methods & Exercises videos - Binary choice models: Introduction to Econometrics

Hans Franses, P. (2026). Lecture 12: *Time Series* [PowerPoint slides]. Retrieved from: Wed - Methods & Exercises videos - Time series: Introduction to Econometrics

Hans Franses, P. (2026). Lecture 13: *Dynamic models* [PowerPoint slides]. Retrieved from: 📺 Mon - Methods & Exercises videos - Dynamic models: Introduction to Econometrics

Hans Franses, P. (2026). Lecture 14: *Forecasting* [PowerPoint slides]. Retrieved from: 📺 Tue - Methods & Exercises videos - Forecasting: Introduction to Econometrics

Hans Franses, P. (2026). Lecture 15: *Non-stationarity* [PowerPoint slides]. Retrieved from: 📺 Tue - Methods & Exercises videos - Non-stationarity: Introduction to Econometrics

Hans Franses, P. (2026). Lecture 16: *Volatility clustering* [PowerPoint slides]. Retrieved from: 📺 Wed - Methods & Exercises videos - Volatility clustering: Introduction to Econometrics