EFR summary

Introduction to Econometrics,

FEB12012X

2025-2026

Lectures 1 to 8

Weeks 1 to 3

Details

Subject: Introduction to Econometrics IBEB 2025-2026

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

Date of publication: 19.03.2026

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.

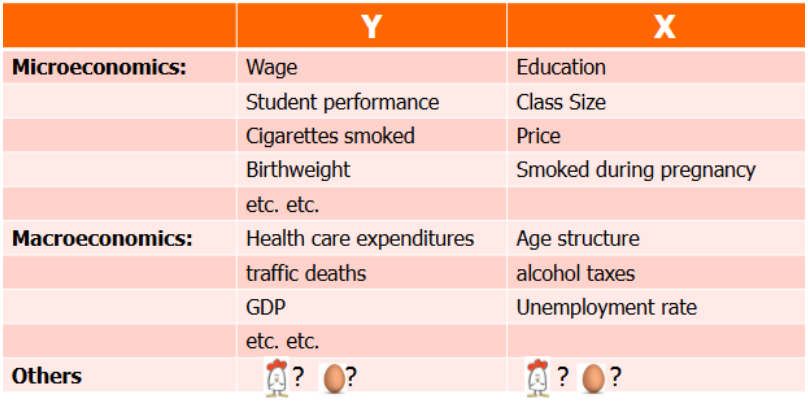
Quality of operationalization

- **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

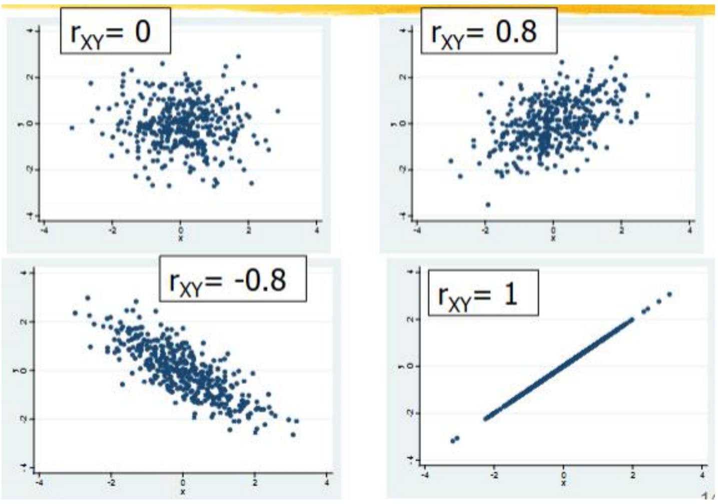
OLS: simple linear regression model

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

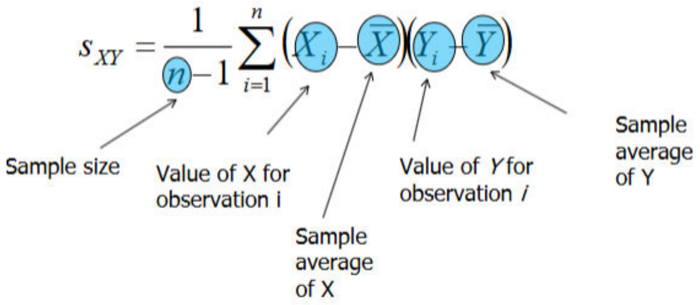


One of the ways to find out about the relationship between variables can be by constructing a **scatter plot**. The relationship can be negative or positive, or there can be no relationship.

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:



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

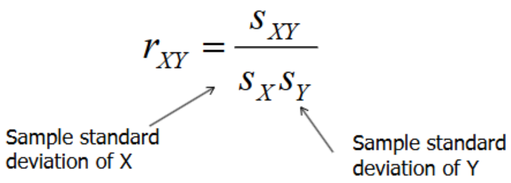
Sample size
Value of X for observation i
Sample average of X
Value of *Y* for observation *i*
Sample average of Y

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:



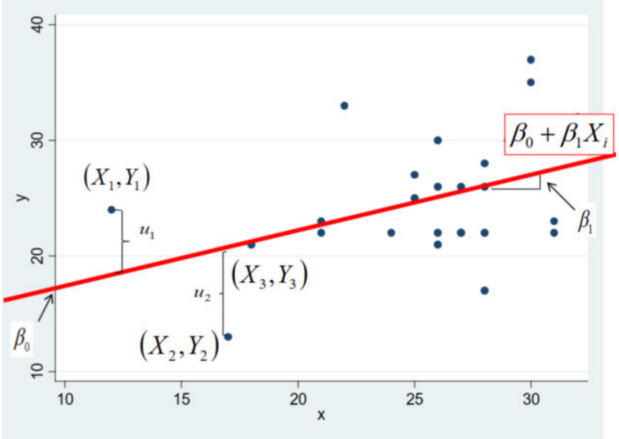
$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Sample standard deviation of X
Sample standard deviation of 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 $\widehat{\beta_0}$:

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

OLS estimator $\widehat{\beta_1}$:

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

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

- OLS predicted/fitted values: $\widehat{Y}_\iota = \widehat{\beta_0} + \widehat{\beta_1} X_i$
- Residuals: $\widehat{u}_\iota = Y_i - \widehat{Y}_\iota$

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 $\boldsymbol{R^2}$

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

- **Total sum of squares (TSS**) is the total variation in the data:

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

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

$$ESS = \sum_{i=1}^{n} \left(\widehat{Y}_\iota - \underline{Y}\right)^2$$

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

$$R^2 = \frac{ESS}{TSS} = corr\left(Y_i, \widehat{Y}_\iota\right)^2$$

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 $\widehat{Y_\iota} = Y_i$
- If R^2= 0: model (X_i) does not predict any variance in Y_i, so $\widehat{\beta_\iota} = 0$, thus $\widehat{Y_\iota} = \bar{Y}$
- 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 $\widehat{u_\iota}$

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

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.

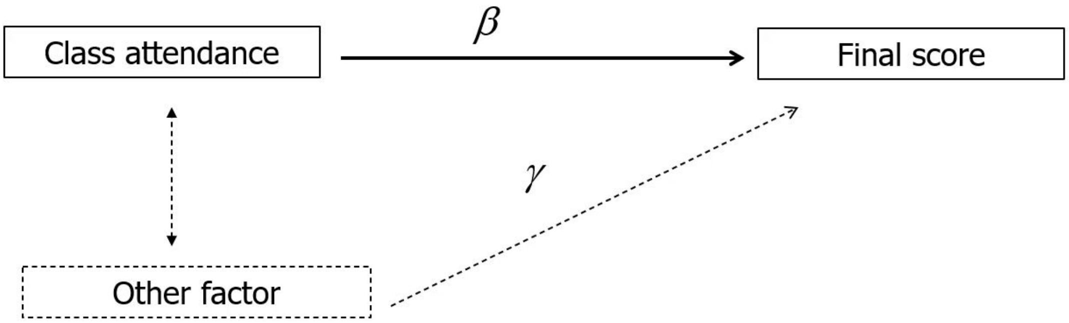
To know and assert whether this condition holds, we must be sure of the **random assignment of the variable X**.

- 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.

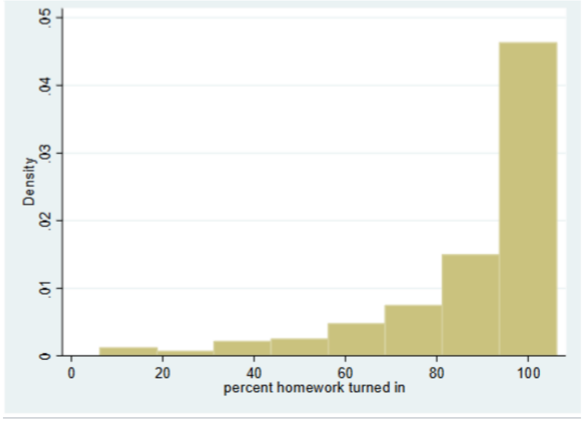
As different samples can lead to different estimates, the estimators are just some points in the sampling distribution of the estimator.

- 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.**

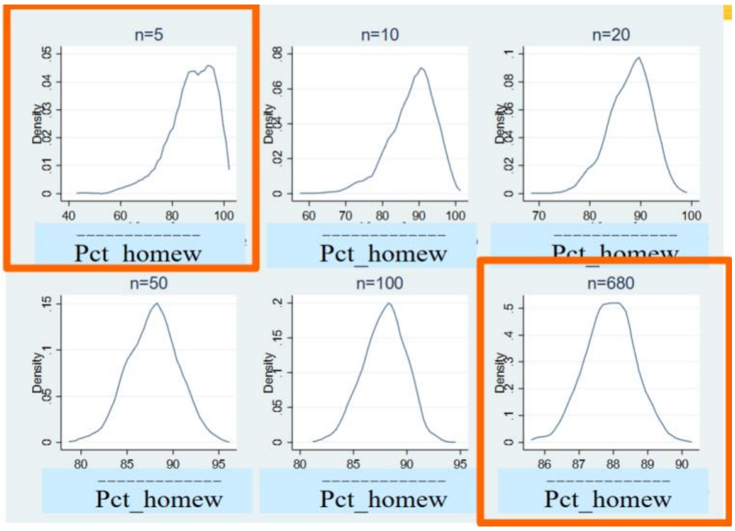
This comes directly from the **central limit theorem.**

- If Y_1, Y_2, …, 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 $\widehat{\beta_0}$

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

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

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

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

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

Under some rearrangements, we get to

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

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

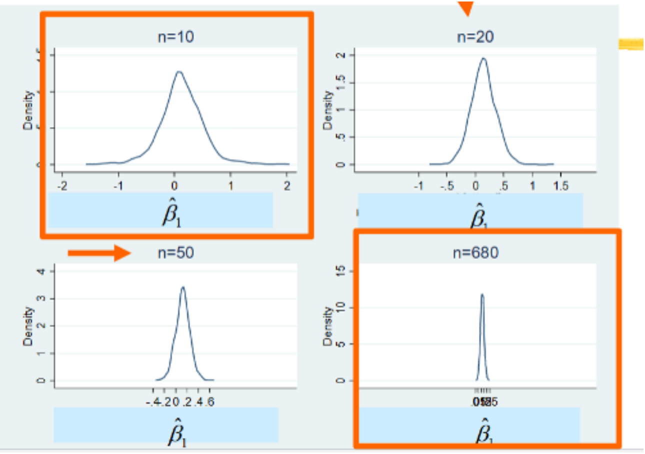
Because of the central limit theorem in large samples, $\widehat{\beta_0}$ and $\widehat{\beta_1}$ approximately follow a normal distribution $\widehat{\beta_1} \sim N\left(\beta_0; \sigma^2_{\widehat{\beta_1}}\right)$, and jointly they follow a bivariate normal distribution.

Variance of $\widehat{\beta_1}$:

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

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: $\boldsymbol{\beta_1}$ is the difference between the average when D=1 and the average when D=0. $\boldsymbol{\beta_1}$ 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 $\boldsymbol{\beta_1} = 0.839$, 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

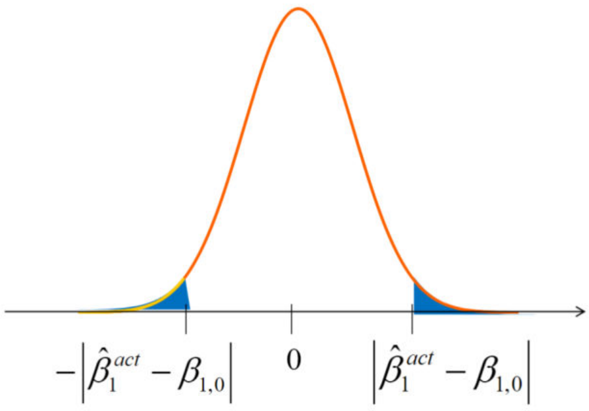
$$H_0: \beta_1 = \beta_{1,0} \quad vs \quad H_1: \beta_1 \neq \beta_{1,0}$$

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**

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

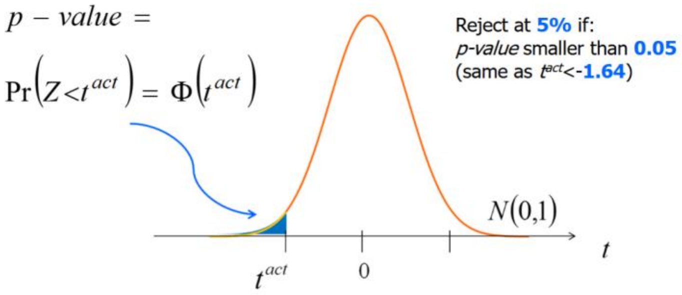
- **One sided** (either smaller or bigger than)

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

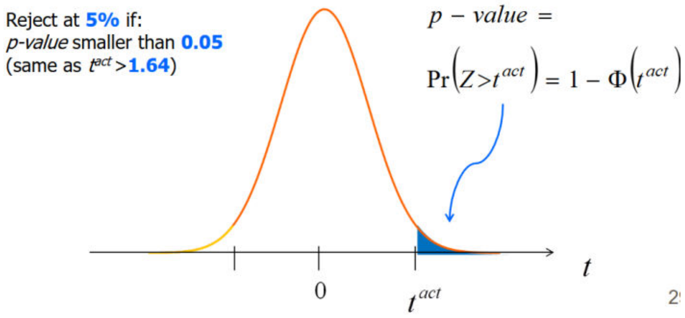
If the alternative hypothesis is that it's smaller than $(\boldsymbol{\beta_1} < \boldsymbol{\beta_{1,0}})$

- 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%)


$p - value =$
$\Pr\left(Z<t^{act}\right) = \Phi\left(t^{act}\right)$
Reject at 5% if:
p-value smaller than 0.05
(same as t^{act}<-1.64)
$N(0,1)$
t^{act}
0
t

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


Reject at 5% if:
p-value smaller than 0.05
(same as t^{act}>1.64)
$p - value =$
$\Pr\left(Z>t^{act}\right) = 1 - \Phi\left(t^{act}\right)$
0
t^{act}
t

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%

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

- 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

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

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

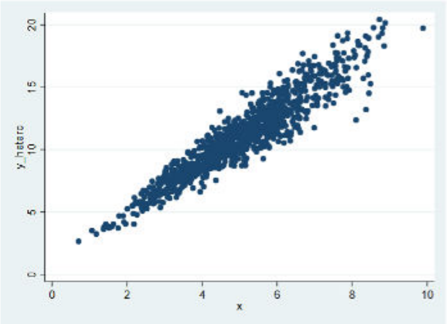
This is a special case where we assume Homoskedasticity (not an assumption)

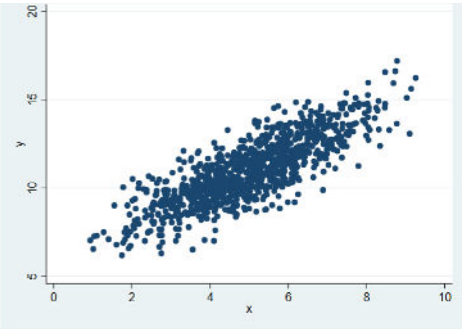
$$\mathrm{var}\left[u_i \middle| X_i = x\right] \text{ is constant for } i = 1, \ldots n$$

- 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

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

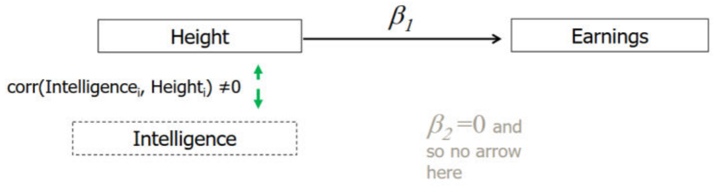
Following some steps we can also write it as

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

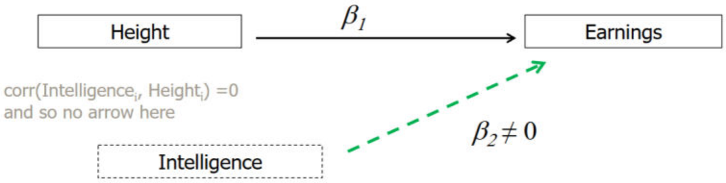
- If for now **we assume that corr(v_i, Height$_i$) = 0**, meaning there no other important variable that affects earnings

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

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


Height
β_1
Earnings
corr(Intelligence$_i$, Height$_i$) ≠0
Intelligence
$\beta_2=0$ and so no arrow here

OR



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

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

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.

if taller people are more intelligent → corr(X_1,X_2)>0

if taller people are less intelligent → corr(X_1,X_2)<0

if intelligence increases earnings → β_2>0

if intelligence decreases earnings → β_2<0

	corr(X_1,X_2)>0	corr(X_1,X_2)<0
β_2>0	**Positive** bias	**Negative** bias
β_2<0	**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)

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

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

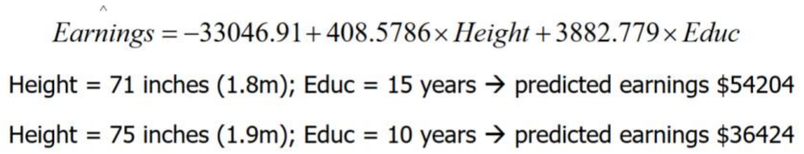
$$\begin{aligned} &E\left(Y_i \middle| X_{1i}, X_{2i}, \dots, X_{ki}\right) = \\ &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \end{aligned}$$

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

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

- 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



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

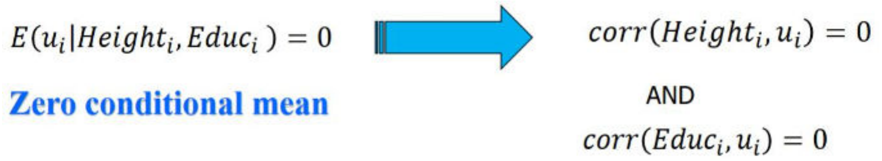
Height = 71 inches (1.8m); Educ = 15 years → predicted earnings $54204

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



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

Zero conditional mean

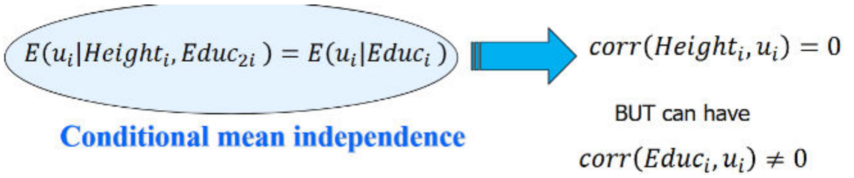
$corr(Height_i, u_i) = 0$

AND

$corr(Educ_i, u_i) = 0$

However, If we are **only** interested in obtaining an unbiased estimate (β_1) of the causal effect X1, a weaker assumption can be used

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



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

Conditional mean independence

$corr(Height_i, u_i) = 0$

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 $\hat{u}_\iota$

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

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

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

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

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.

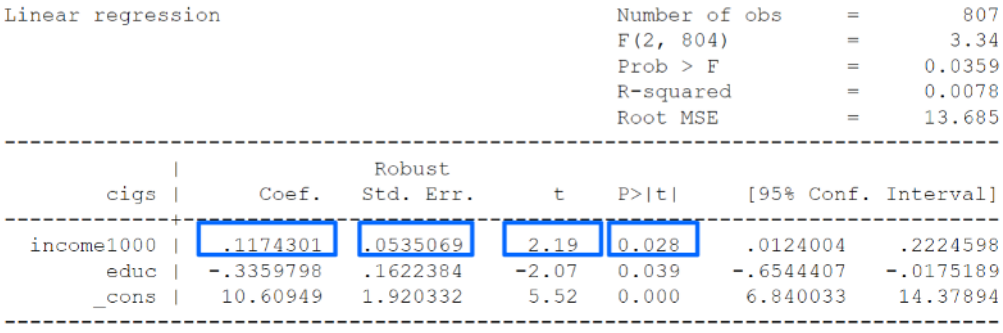
Hypothesis test for a single coefficient

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

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

- 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

------------------------------------------------------------------------------
             |               Robust
        cigs |      Coef.   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 H₀ 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

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

- 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

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

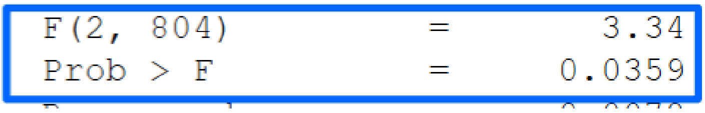
- The p-value tell us the probability that a random variable following an F distribution with q and plus infinity degrees of freedom is larger than the F-statistic (sample) that we obtain with our test, assuming H_0 is true

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 x² as a new variable and regressing on both x and x². **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 = $\alpha_1 + 2\alpha_2 \cdot x$

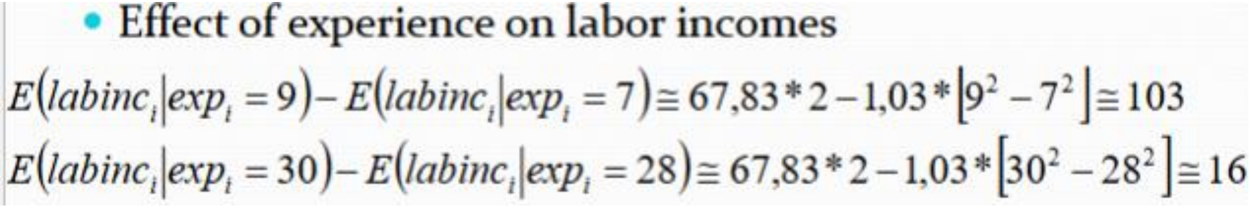
For quadratic polynomials, when the coefficient in front of squared variable is positive it represents the increasing returns to scale and when that coefficient is negative we can see decreasing returns to scale.

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

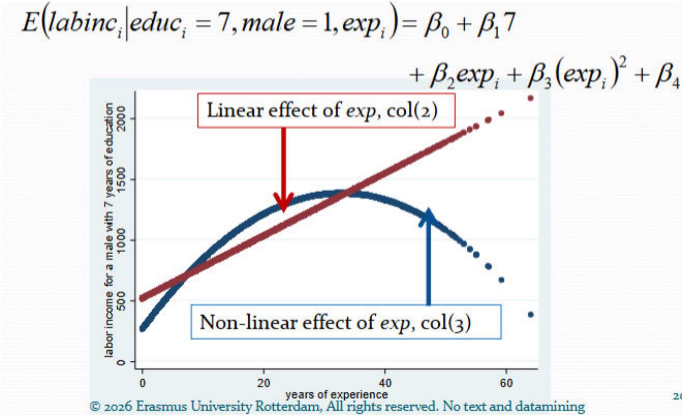


- Effect of experience on labor incomes

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

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

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:

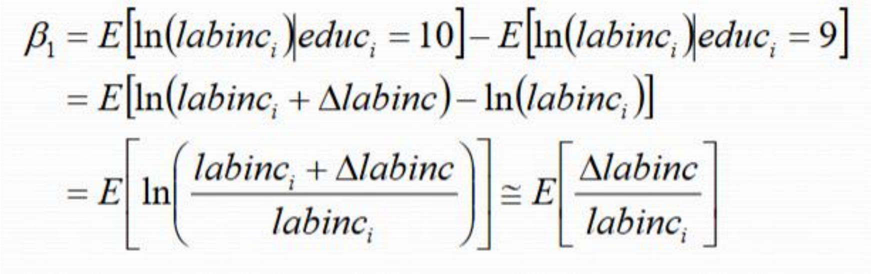
- Outliers in the right tail can be dealt with using this method. Large outliers lead to a violation of the third OLS assumption, and they are less likely to affect the model after this transformation when the large outliers are compressed.
- Used if one is interested in percentage changes.

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



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

- 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:

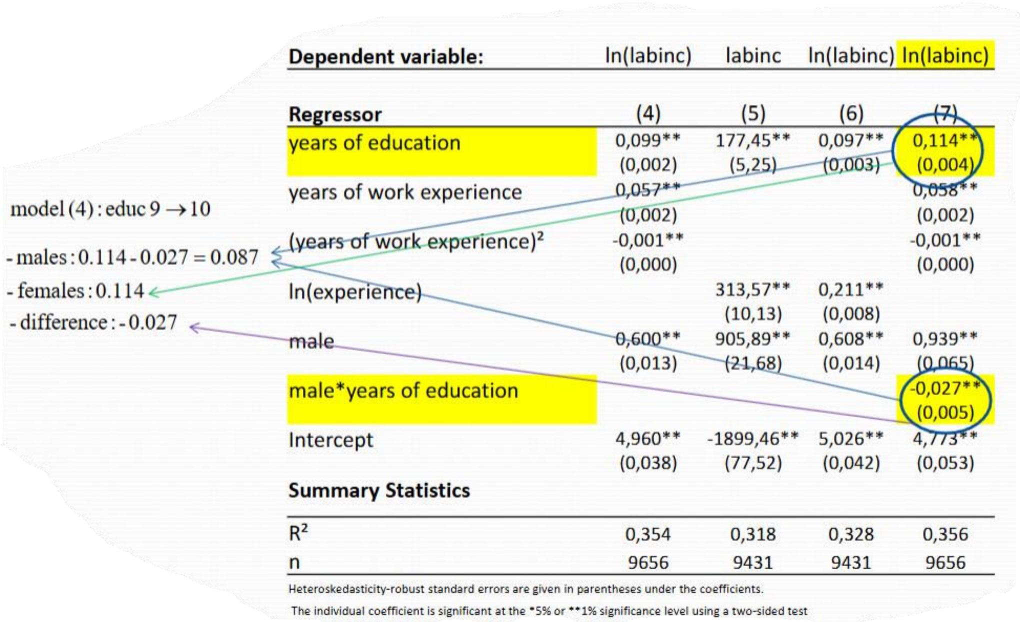
$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u_i$$

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**	177,45**	0,097**	0,114**
	(0,002)	(5,25)	(0,003)	(0,004)
years of work experience	0,057**			0,058**
	(0,002)			(0,002)
(years of work experience)²	-0,001**			-0,001**
	(0,000)			(0,000)
ln(experience)		313,57**	0,211**	
		(10,13)	(0,008)	
male	0,600**	905,89**	0,608**	0,939**
	(0,013)	(21,68)	(0,014)	(0,065)
male*years of education				-0,027**
				(0,005)
Intercept	4,960**	-1899,46**	5,026**	4,773**
	(0,038)	(77,52)	(0,042)	(0,053)
Summary Statistics				
R²	0,354	0,318	0,328	0,356
n	9656	9431	9431	9656

Heteroskedasticity-robust standard errors are given in parentheses under the coefficients.

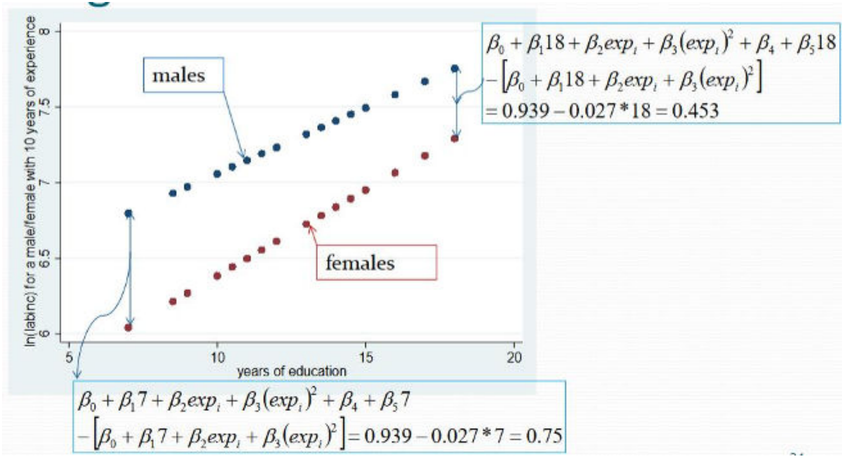
The individual coefficient is significant at the *5% or **1% significance level using a two-sided test

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



$$\beta_0 + \beta_1 18 + \beta_2 exp_i + \beta_3 (exp_i)^2 + \beta_4 + \beta_5 18 - \left[\beta_0 + \beta_1 18 + \beta_2 exp_i + \beta_3 (exp_i)^2\right] = 0.939 - 0.027 * 18 = 0.453$$

$$\beta_0 + \beta_1 7 + \beta_2 exp_i + \beta_3 (exp_i)^2 + \beta_4 + \beta_5 7 - \left[\beta_0 + \beta_1 7 + \beta_2 exp_i + \beta_3 (exp_i)^2\right] = 0.939 - 0.027 * 7 = 0.75$$

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

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

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

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

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

- 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**

$$E\left(u_i \middle| educ_i,\ldots\right) = E\left(u_i \middle| \ldots\right)$$

- 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

$$\underbrace{X_i^r}_{\text{with error}} = \underbrace{X_i}_{\text{without error}} + m_i^X \;\text{— measurement error}$$

- 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 $\alpha_1$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 $\alpha_1$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:

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

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

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

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

$$\underbrace{Y_i^r}_{\text{with error}} = \underbrace{Y_i}_{\text{without error}} + \underbrace{m_i^Y}_{\text{measurement error}}$$

We observe measured income Y_i^rinstead of true income Y_i. We estimate $\gamma_1$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.

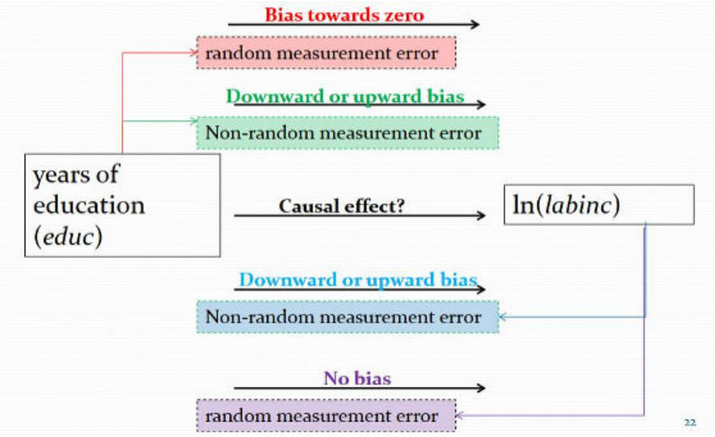
However, it does **increase variance**. The error term in the regression becomes $u_i + m_i^Y$instead of just u_i, so:

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

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.

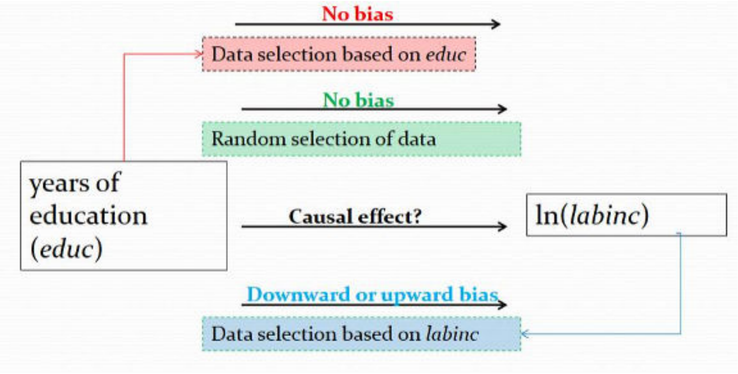
Solutions

- **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.


No bias
Data selection based on *educ*
No bias
Random selection of data
years of education (*educ*)
Causal effect?
ln(*labinc*)
Downward or upward bias
Data selection based on *labinc*

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.

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