EFR summary

Applied Microeconometrics,

FEM11087

2023-2024

Lectures 1 to 23

Weeks 1 to 4

Details

Subject: applied microeconometrics 2022-2023

Teacher: Pilar Garcia-Gomez

Date of publication: 27.09.2024

applied microeconometrics - module 1 – linear regression analysis

Lecture 1 – Introduction to the linear regression model

EMPIRICAL ANALYSIS It is a scientific methodology where we use the data to test a theory and to estimate relationships between variables.

First, we have to define our research question. They can come from:

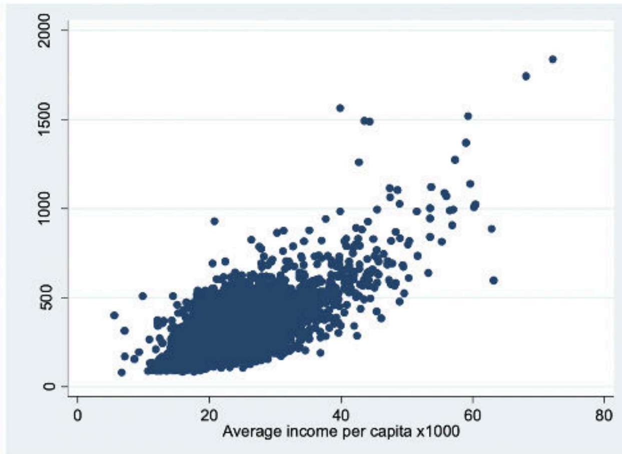
- Existing economic models.
- Via intuitive and less formal reasoning (something that inspires us to be studied from an economic point of view and with a scientific method).

SIMPLE REGRESSION MODEL It is a model that tells us somehow how a variable defines another one.
We have two variables, y and x, and we:
 want to explain y in terms of x
 want to know how y varies with changes in x

Example: how does the crime rate (y) change with changing the number of police officers (x) in a city?

Main example: HOUSE PRICES AND AVERAGE INCOME IN A NEIGHBOURHOOD[1] (Garcia-Gomez, 2022)
We could study these variables to know what policies are incrementing.

Figure n. 1[2] (Garcia-Gomez, 2022)

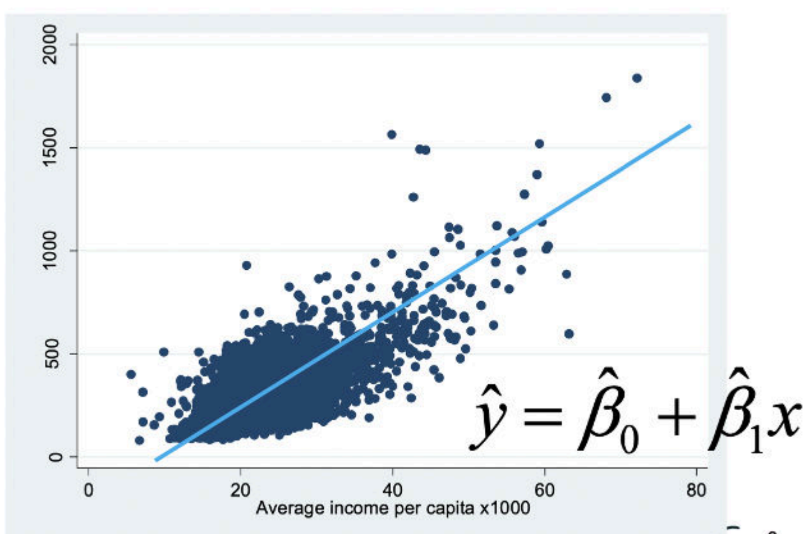


Source: CBS. Information for neighbourhoods in the Netherlands in 2012

We can observe that there is a positive association between wages and house prices (higher the income, higher the house price).
The aim of the linear regression model is to find a line that can summarize all the information given by the scatter plots, to show the predicted value of the average house price as a function of the average income per capita.

The line's formula is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Figure n. 2[3] (Garcia-Gomez, 2022)



$\hat{\beta}_0$ is the intercept (it is the house price when income = 0, which in this case is not very indicative).
$\hat{\beta}_1$ is the slope (it tells us how house prices change when income does it as well).

simple regression model

$$y = \beta_0 + \beta_1 x + u$$

- *y* is the dependent variable.
- *x* is the independent or explanatory variable.
- *u* is the error term or disturbance (also referred as the "unobserved").

u it is all that is unobserved by the researchers that has an impact on the dependent variable (in our example, everything else that affect house prices in a neighbourhood, e.g., the number of amenities present in a neighbourhood).

ceteris paribus condition

We are interested into knowing how the dependent variable changes when the *x* changes, holding all other variables fixed.
If the factors in *u* are held fixed □ $\Delta u = 0$

So □ $\Delta y = \beta_1 \Delta x$ □ this is the interpretation of a SLOPE in our linear regression model.

zero conditional mean assumption

□ The unobserved (*u*) does not change when *x* changes in term of expected values.
□ $E(x) = E(u) = 0$

This makes possible to see the line of the linear regression model in terms of **EXPECTATIONS** (what is the expected value of *y* with a given value of *x*)

$$E(x) = \beta_0 + \beta_1 x$$

Does the ceteris paribus condition works with our example?

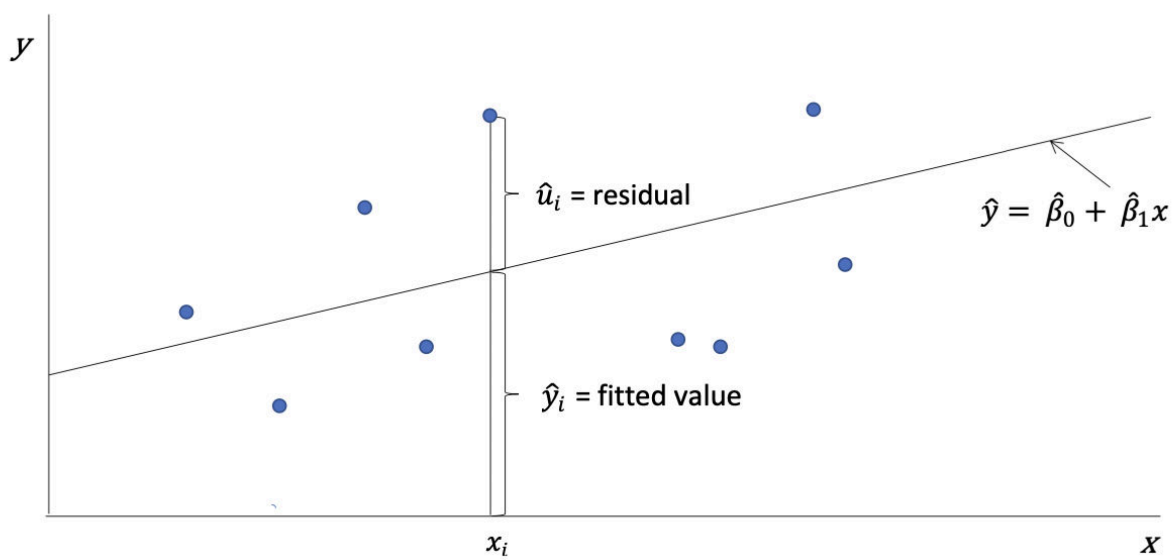
It depends if the zero conditional mean assumption holds.
For example: *u* contains the quantity and quality of amenities in a neighbourhood.

- If we assume that the number of amenities does not change given the average income in a neighbourhood, then we have $\Delta u = 0$ the zero conditional mean assumption holds.
- If the number of amenities varies depending on the wealth of the neighbourhoods, then the zero conditional mean assumption does not hold up we cannot draw ceteris paribus conclusions.

Lecture 2 - Estimation and interpretation in the linear regression model

ORDINARY LEAST SQUARE ESTIMATES

Figure n. 3[4] (Garcia-Gomez, 2022)



We want an estimate of β_0 and β_1 and we will use the ordinary least square estimates.

1. In the population we expect that *y* and *x* are related in a linear way.
    $y_i = \beta_0 + \beta_1 x_i + u_i$ this is a random sample of the population of interest.
    We will never know exactly β_0 and β_1, but we want a good estimate of those.
    First, we draw a random sample from the population, and for every individual in the random sample we can plot the value for *x* and *y* (see the figure n. 3). Given the values for *x* and *y*, we draw a fitted line, similar as before.

FITTED VALUE ($\hat{y}_i$): is the value that, for a given x_i, falls on the fitted line.

We see that there is still a difference between the observed value and the value on the fitted line: the difference is called the RESIDUAL $\hat{u}_i$.

The fitted value is just a predicted value: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

The residual is: $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ □ $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

Our aim is to have the residuals as small as possible to have the best possible fit.

2. We then obtain the estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing the square of the residuals:

$$min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

This is what the **ordinary least square** estimator does to obtain the estimates. The ordinary least square estimator is the most efficient and biased estimator. To calculate these values, we must use STATA.

With STATA we obtain an equation with which we can draw the fitted line.

MULTIPLE REGRESSION MODEL

It is more difficult to draw ceteris paribus conclusions with this model. For example, house prices are also related to the population density of an area:

$$house\ price = \beta_0 + \beta_1 income + \beta_2 density + u$$

And before we had:

$$house\ price = \beta_0 + \beta_1 income + u$$

We remember that we can draw ceteris paribus conclusions if the unobserved is not correlated to the explanatory variable. Now we have to see if the population density is correlated to income: richer householders are more likely to live in less populated areas? If yes, the zero conditional mean assumption could not be satisfied □ there would not be any ceteris paribus conclusions and would be better to estimate the simple regression model.

Everything we have seen so far also applies to the multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Multiple regression analysis allows us to control for many other factors that simultaneously affect the dependent variable.
In addition, controlling for more variables also allows us to have better predictions.

Lecture 3 – OLS assumptions – unbiasedness

UNBIASEDNESS: With the concept of "unbiasedness" we mean that the expected value of our estimated parameter is equal to the population parameter.

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

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

Please note that the hat is used for estimated values and anything without the hat is used to indicate population's values.

We need four assumptions for this property:
1. (MLR1) The model must be linear in parameters.
2. (MLR2) We must have random sampling.
3. (MLR3) There cannot be perfect collinearity.
4. (MLR4) The zero conditional mean assumption must be satisfied ($E(x) = 0$).

1. LINEARITY IN PARAMETERS ASSUMPTION

We have the population model. In this model we have the variable *x* (independent variable) and *y* (dependent variable). *y* is related to *x* and the error *u* via the following equation:

$$y = \beta_0 + \beta_1 x + u$$

All the parameters are linearly related.
Important: the assumption is about the linearity in the parameters. So, the linearity is not possible in cases like this:

$$y = \beta_0 + \beta_1 x_1 + \beta_1 \beta_2 x_2 + u$$

$\beta_1 \beta_2$ □ Interaction term between β_1 and β_2 □ not possible!

However, there can be nonlinearities in the variables:

$y = \beta_0 + \beta_1 \ln ln\left(x_1\right) + \beta_2 x_2 + \beta_3 x_2^2 + u$ (Example of quadratic variable)
$\ln ln\,(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + u$ (Example of logarithmic variable)

In these cases, we need to change our interpretation of the variables, and we are going to see a few more examples later on.

2. RANDOM SAMPLING ASSUMPTION

For this assumption to be true, we need to pick a random sample of size *n*, following the population model. If, for whatever reason, the sample will not be picked randomly, we will incur in selection bias.

3. NO PERFECT COLLINEARITY ASSUMPTION

If we pick a sample, in the following sample (and so in the population):

- Among all the independent variables, none of them should be constant, so that we have variation in all the independent variables. This is important because we use the variation to estimate the effect of our variable *x* on our variable *y*.
- There cannot be any exact linear relationship between the independent variables considered.

Example of perfect collinearity:
Given the model:

$$houseprice = \beta_0 + \beta_1 income + \beta_2 Rotterdam + \beta_3 density + \beta_4 percentageyoungs + \beta_5 percentage\ elderly +$$

Avoiding perfect collinearity means that among the independent variables there is not any exact linear relationship.

But if we assume that all and only the elderly people lived in Rotterdam (*Rotterdam = perc_elderly*), these two variables would suffer the same kind of variation, and therefore the relationship between them would be perfectly linear.
In general, there is perfect collinearity between x_1, x_2 and x_3 if x_3 can be expressed as a combination of the other two variables ($x_3 = ax_1 + bx_2$).

We can have two types of collinearity:

- **PERFECT COLLINEARITY**: in these cases, the estimation will not be successful (it could be that some software will not execute the commands given or will give inappropriate results). STATA tries to resolve the problem arbitrarily dropping the faulty variable in order to estimate a model without this problem; the problem is that it could drop one of the variables of most interest in our model. Therefore, we should first clearly and properly define our model in order to avoid any problem of this sort.
- **IMPERFECT COLLINEARITY**: when we estimate the model, the estimation works, but we get imprecise values for the estimates. We must watch out for:
    - *x*'s with high correlation between them.
    - Imperfect collinearity between variables, for example between $x_1$and x_2. This can happen when we find big F-stat, when x_1, x_2 jointly significant, but small T-statistics.

4. ZERO CONDITIONAL MEAN ASSUMPTION

Will be discussed later when talking about exogeneity.

Lecture 4 – Linear regression analysis: OLS assumptions – inference

INFERENCE – HYPOTHESIS TESTING

We want to test hypothesis about a parameter, or a group of parameters, in the population. We need not only information about the property of the estimator, so the expected value of the estimator, but also about his distribution (related to the distribution of the errors).
So, the following assumptions are related to the distribution of the errors:

- MLR5 – Homoskedasticity
- MLR6 – Normality

Under assumptions 1-6 OLS estimator is the unbiased estimator with minimum variance.

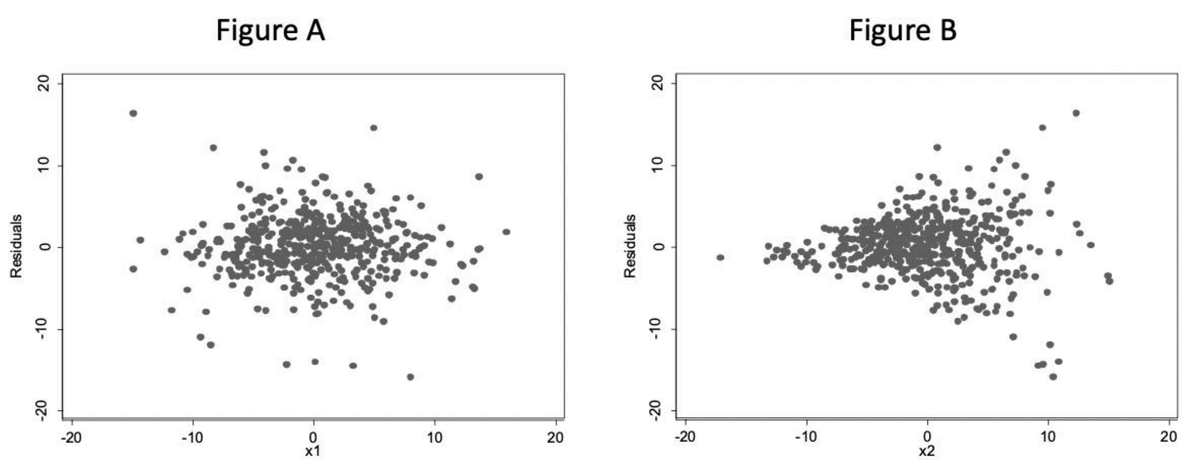
5. HOMOSKEDASTICITY ASSUMPTION

The variance of the error term does not change regardless of the values assumed by the independent variables:

$$var\left(x_1 x_2\right) = var\,(u) = \sigma^2$$

- For every individual the error term has the same importance, despite of the characteristics.
- The outcome of *y* has the same magnitude of uncertainty for every level of *x*'s.

Figure n. 4[5] (Garcia-Gomez, 2022)



In this case, the Figure A has homoskedastic residuals, while in Figure B we can see that the variance of the residuals grows with *x*.

If this assumption does not hold, we have a condition called **heteroskedasticity**:
$var\left(x_1 x_2\right) = f(x_1, x_2)$ or $f(x_1)$ or $f(x_2)$

When we have heteroskedasticity:

- OLS estimates $\hat{\beta}$ are still unbiased and inefficient.

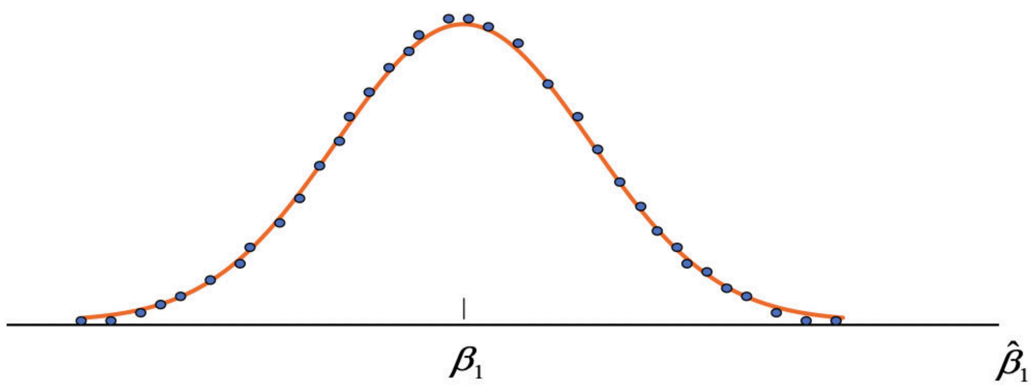
- OLS standard errors for the estimators $se(\hat{\beta})$ are incorrect.

This is a problem for inference, but it is not a big issue since with STATA we can easily adjust the *SE* and the statistics used for inference.
 ALWAYS use heteroskedasticity-robust standard errors.

6. NORMALITY ASSUMPTION

This assumption implies that the population error *u* is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and follows the distribution $u \sim Normal\,(0,\ \sigma^2)$.
This implies that if we could draw many samples of size *n* and estimate linear regression model by OLS with each of these samples plotting the estimated $\hat{\beta}_1$ that we get in every case, those follow a normal distribution centred at β_1.

Figure n. 5[6] (Garcia-Gomez, 2022)



With large samples $\hat{\beta}_1$ always approximately follows a normal distribution centered in β_1, so even if *u* does not follow a normal distribution, OLS estimator is asymptotically normally distributed (i.e., approximately normally distributed in large samples).

 When we have large sample sizes, we can always carry on using the standard tests for hypothesis testing (only IF we have them).

On the other hand, if the sample is small and we have non-normal errors, we need to worry about the normality assumption.

TO SUM UP:

- We need the first 4 assumption (MLR1-MLR4) to obtain unbiased estimates of the population parameters.

- The fifth and sixth assumptions (MLR5-MLR6) are important for inference, but we may face some problems with them. Anyway:
  - Even if assumption MLR5 is not satisfied, standard errors and test can be easily adjusted.
  - When our samples are large, the non-normality of the errors is not an issue.

Said that, now we have all the components for hypothesis testing.

Lecture 5 – Linear regression analysis: inference I. – single population parameter

We have seen in our main example using STATA that with an increase of €1.000 derives an increase of €16.000 in the average house price of the neighbourhood, ceteris paribus.
Is this effect different from 0? 16 should be far enough from 0.
But what about 15? Is it far enough from 16?

Those questions are hypothesis, which we could test comparing how different our estimates coefficient is to the number we want to test (0 or 15). We are going to see how different it is with a statistical tool/testing.

INFERENCE

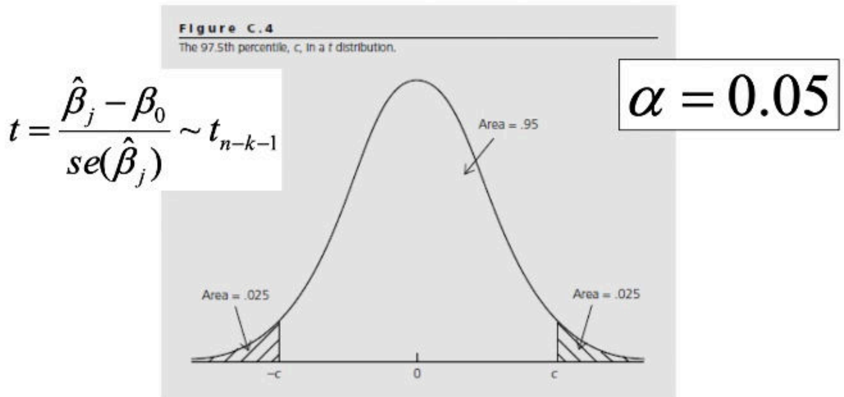
First, we are going to test a hypothesis (the null hypothesis H_0) for a single population parameter $H_0: \beta_j = \beta_0$.
To test this hypothesis (that the population parameter is equal to 0 or 15) we compute a t-statistic:

$$t = \frac{\hat{\beta}_j - \beta_0}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

The t-statistic is computed comparing the estimated population parameters minus the value that we want to test divided by the estimated standard error. The standard error gives us an idea of the level of precision of our estimate, and this is related to the fact that we have a random sample of the population.

Under the null hypothesis we know that this number is distributed following a t-distribution in which *n* is the number of observation and *k* is the number of explanatory variables.
The distribution looks like the following figure below.

Figure n. 6[7] (Garcia-Gomez, 2022)



The idea is that under the null hypothesis the t-statistic is going to be very close to 0. The more we go away from 0 the more we go towards the tails of the distribution, the less likely it is that our null hypothesis is true.

How further away are we going to accept a number, in order not to reject our null hypothesis?
We need to set a significance level (α), which is the tolerance for a type I error: it is the probability of rejecting the null hypothesis given that it is true. The most common values for the significance level α are 0.10, 0.05 and 0.01. For example, having a significance level of 5% means if the null hypothesis was true then only 5% of the random samples would provide an estimate in the area of the tail of the distribution, 0.025 on the left and 0.025 on the right. If our estimated value falls in these areas, it is very unlikely that our null hypothesis is true, therefore we reject it. We will never be certain, so we base this in this probability that it is less likely that it could happen.
We reject the null hypothesis if the absolute value of the t-statistic is larger than the critical value *c* for our significance level (1.96 for a 0.05 significance level):

$$|t| > c$$

We often use the *p-value*: which tells us what is the largest significance level at which we could carry out the test and still fail to reject the null hypothesis[8] (Garcia-Gomez, 2022). We reject the hypothesis if:

$$p - value < \alpha$$

The p-value gives us the probability that we have left on the tails given our t-statistic[9] (Garcia-Gomez, 2022).

CONFIDENCE INTERVALS

To avoid having only a point estimate, we use confidence intervals, which provide us a range of likely values for the population parameter.
We know that:

$$\frac{\hat{\beta}_j - \beta_0}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

Looking at the graph above, we can ask ourselves what the value of β that corresponds to -*c* and *c* is.
We know that testing the null hypothesis for any of those values we could not reject them because they fall within the area in the centre of the distribution (the 95% of it).
We can use this formula, where we add and subtract at the estimated parameter the value *c* (the critical value in the distribution) times the standard error we estimated for our β:

$$\hat{\beta}_j \; c \; se(\hat{\beta}_j)$$

Interpretation of the confidence intervals
<u>Important</u>: we are not 95% sure that the real value is in this interval, because, in fact, we are not, but it means that:

- From all the possible samples that we can possibly draw, in 95% of the cases the true value of the coefficient will be inside the interval.
  We just hope that our random sample is one of those containing the real value.
- Using a two-sided hypothesis, it gives us the set of all values that cannot be rejected (in this case at 5%) (the values that are in the centre of the t-distribution between -*c* and *c*).
- Again, this is NOT equivalent to the probability of 95% of having the real value inside this exact interval.

Lecture 6 – Inference II

Testing multiple restrictions: multiple hypotheses test or joint hypotheses test

- We are interested to see if a group of variables has no effect on the dependent variable. So, we are going to test this.
  For example, are the house prices affected by the demographic structure?
  $$houseprice = \beta_0 + \beta_1 income + \beta_2 perc_{young} + \beta_3 + perc\_elderly + u$$

We have multiple explanatory variables:
Our null hypothesis is that β_2 is equal to 0 as well as β_3:

- $H_0: \beta_2 = 0, \ \ \beta_3 = 0$ □ if true, the demographic structure will not have an effect in houses prices.
  We have two restrictions: if H_0 is true, then this group of variables has no effect on house prices after we controlled for the income per capita.
- $H_1: H_0$ is not true □ AT LEAST ONE of the coefficients is different than 0, or either one of them, but is sufficient that only one of the coefficients is different than 0.

To test these hypotheses, we define an F-statistic in which the null hypothesis H_0 is that the two coefficients are equal to 0; the alternative hypothesis is H_1.

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

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

Note that this formula is robust to heteroskedasticity errors.

- t_1 is the t-statistic of test $\beta_1 = 0$
- t_2 is the t-statistic of test $\beta_2 = 0$
- $\hat{p}$ is the estimator of the correlation between these two t-statistics.

Once completed we know that the F-statistic in large samples follows an F distribution.

$$F - statistic \sim F_{q, \infty}$$

(q is the number of restrictions; infinity refers to the fact that we use a very large sample)

As usual, we reject the null hypothesis if F is large in statistical terms. If F > critical value of $F_{q,\infty}$ we reject the hypothesis.

- At 10% the critical value could be 2.30.
- At 5% the critical value could be 3.00.
- At 1% the critical value could be 4.61.

We test these hypotheses with STATA.
With STATA we see that the F value is 568,95 (a lot more than 2.30), therefore we reject the null hypothesis.
In alternative we can look at the p-value, which in this case is smaller than the 5% significance level, and we get to the same conclusion.

We can use the same procedure to test a linear combination of the parameters:
H_0: $\beta_2 = \beta_3$; H_0: $\beta_2 - \beta_3 = 0$ □ test command in STATA after running the regression model.

When H_0 is not rejected, we must say that: WE FAIL TO REJECT H_0 AT THE *x*% SIGNIFICANCE LEVEL

We cannot say that H_0 is accepted at the *x*% significance level, but only that it is not rejected because it can assume different values.

Lecture 7 – Interpretation and categorical variables

CATEGORICAL VARIABLES: variables that contains qualitative information (for example, the variable "*religion*" is a categorical variable).

BINARY/DUMMY VARIABLES: categorical variables that can only take two categories.

For example, we could ask ourselves if, ceteris paribus, house prices in Rotterdam would be different from the rest of the Netherlands. In order to do so, we can create one of these two variables:

- Variable Rotterdam (it assumes the value 1 if the neighbourhood is in Rotterdam, 0 if in other parts).
- Variable Other (1 if in other parts, 0 if in Rotterdam).

And we can estimate this model:

$$House\ price\ =\ \beta_0 + \beta_1 income + \beta_3 Rotterdam + u$$

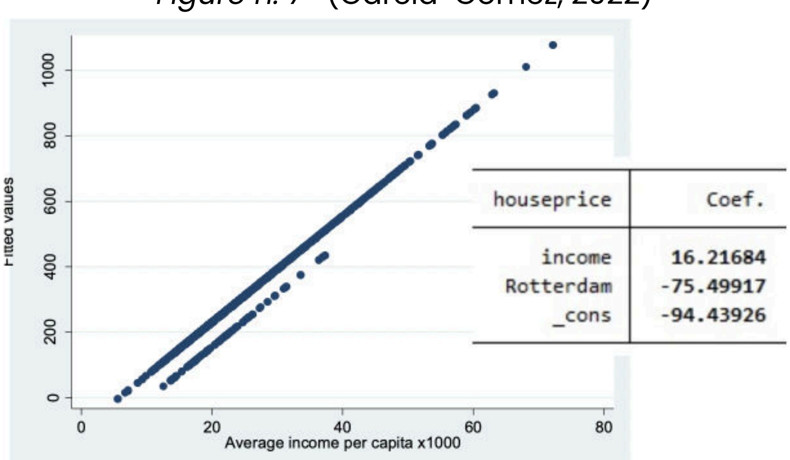
Or, alternatively, this other model:

$$House\ price\ =\ \beta_0 + \beta_1 income + \beta_4 other + u$$

The coefficient β_3 is representing the effect on the average house price of being in Rotterdam compared to other regions, ceteris paribus.

We cannot include both “Rotterdam” and “other” in the same equation because there is perfect collinearity between them, since every time “Rotterdam” is 1 “other” is 0, and vice versa □ there is no separate variation in one variable compared to the other.
If we add “Rotterdam” and “other” in the same equation, the sum will always be 1, which is the same as the constant.

Figure n. 7[10] (Garcia-Gomez, 2022)



In STATA we can choose a model or the other: the conclusions will be the same, even if we get different coefficients for each equation, and that’s because the constant is

giving us different values in each of the models (in the first model it tells us the expected house price for neighbourhoods not in Rotterdam, while in the second it tells us the expected house price for neighbourhoods in Rotterdam).
We then can look at the predictions. We have 2 lines of dots which tell us how house prices change when income does in a neighbourhood that is in Rotterdam (one line) and for a neighbourhood which is in another region (the other line).

Which one is the highest and which one is the lowest? We could look at what is the expected value as we did in the previous passage this will give us the line with the lowest intercept.

Are they parallel? It seems like they are. It is because we assume that an income variation has the same effect in Rotterdam as well as in any other region the slope is always the same.

Sometimes the categorical variable has more than two categories.

It is important that all the categories contained in the variable could be interpreted in a quantitative way, and we create a dummy variable (the variable with two variables) for each of the categories.
But if we have, for example, four categories, we cannot include all of them in the same equation because of the collinearity. Therefore, we use three categorical variables in one equation and take the one we left behind as the reference category. Once set the reference category, we must interpret the coefficients of the included dummy variables compared to the reference category. When interpreting the intercept, we also need to consider it compared to the excluded variable and remember that the intercept is the value our model assumes when the INCLUDED explanatory variables are equal to zero, but not the reference category.

We interpret the coefficients compared to the reference category because if we changed it, the estimated coefficients would also change.

Lecture 8 – Model selection in linear regression analysis

What could be the variables to include in a model?

We should begin with a (theoretical) framework:
We could, for example, examine the probability of committing a crime: Gary S. Becker elaborated in 1968 an economic model in order to describe the grade of participation in a crime by an individual.
The individual participation in a crime (our dependent variable *y*) could be described by the following variables: the hourly "wage" obtained through a criminal activity (x_1), the hourly wage obtained with a legal job (x_2), other income (x_3), the probability of getting caught (x_4), the probability of being convicted if caught (x_5), the expected sentence if convicted (x_6), the individual's age (x_7).

After choosing a model framework and having seen the data, it is possible to see which variable are available. If we cannot observe some of them, they will be part of the unobserved.

We must not start with bivariate associations, because it could be that the β, if the zero conditional mean assumption is not satisfied, is biased.
So, we could choose a variable from a bivariate association, but it is going to be biased in most of the cases.

What is the point of the analysis?
There are two possible alternatives:
1. The aim could be having the best predictive model. In this case we can use **goodness of fit** measures.
2. If the aim is to estimate the causal effect of *x* on *y*, the goodness of fit measures will not be informative: in this case it is needed to control for sufficient confounders to satisfy the zero conditional mean assumption.

First, we analyse the case where our aim is to have the best predictive models. As we said, we can use goodness of fit variables.
These variables are useful to know how well our model fits the data, which means how much of the overall of the dependent variable could be explained by our independent variables.
The standard measures of goodness of fit are the following:
- R^2
- Adjusted R^2
- Overall F-test

R^2

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Where:

TOTAL SUM OF SQUARE: $SST = \sum_{i=1}^{N} \left(y_i - \bar{y}_i\right)^2$

EXPLAINED SUM OF SQUARES: $SSE = \sum_{i=1}^{N} \left(\hat{y}_i - \bar{y}_i\right)^2$

SUM OF SQUARED RESIDUALS (UNEXPLAINED VARIATION): $SSR \sum_{i=1}^{N} \hat{u}_i^2$

R^2 captures how much of the total variation (which is the total sum of the square), is explained by our model.

We can define it in two ways, either as the ratio of the explained sum of the squares divided by the total sum of the squares, or as 1 minus the sum of the square of the residuals (that is the unexplained variation) divided by the total variation.
This number is going to be between 0 and 1 and the closest it is to 1, the highest the share of the variation that our model explains.
Our problem with this measure is that it always increases with the inclusion of variables in the model: every time we add one explanatory variable, even if it explains very little, is going to cause the R-squared to increase.

ADJUSTED R^2

Our aim is to have a model which is both simple and complete: we want to control for sufficient *x*'s, but at the same time we should not overdo the model, therefore we should add variables to the point they provide a sufficient contribution in our predictions.

$\bar{R}^2$ (adjusted R^2) corrects the increase we discussed above by penalising for the number of coefficients: if we add a new variable with a |t-stat|>1 or, in the case of a set of added variables, F-stat>1, it increases.

OVERALL F-TEST

Another useful measure is the **overall F test**: with this measure we can check whether the joint hypothesis that the coefficients of all our variables is equal to zero, compared to the alternative that at least one is different than 0. This is exactly the same F-test we have seen in the second inference lecture, even if in that case we used it for all the variables.

The unrestricted is the model that includes as many restrictions as explanatory variables in our model. This tells us whether our model explains anything at all, so at least one of the variables is statistically significant.

Overall F test of H_0: $\beta_1 = ... = \beta_k = 0$ vs H_1: at least one $\beta_j \neq 0$ with $j = 1, ..., k$

- Unrestricted model (UR): $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$
- Restricted model (R): $y = \beta_0 + u$ (with *k* restrictions)

We can do this for all the variables as well as for sub-groups of variables:
F test: H_0: $\beta_1 = ... = \beta_q = 0$ vs H_1: at least one $\beta_j \neq 0$ with $j = 1, ..., q$
- Unrestricted model (UR): $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$
- Restricted model (R): $y = \beta_0 + \beta_q + 1\, x_q + 1 + ... + \beta_k x_k + u$ (with *q* restrictions)

This is going to add information if those variables are jointly significant. We can test for this using an F-test in which we test this sub-group of variables. This is also a way of deciding whether we want to keep those variables in the model or not. Because if they are jointly statistically significant there is no need to have those variables in our model.

Lecture 9 – Exogeneity

Unbiasedness of OLS (recap):
$E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$

To be able to say this, we need the four assumptions:
- Linearity in the parameter (MLR1)
- Random sampling (MLR2)
- No perfect collinearity (MLR3)
- Zero conditional mean assumption, i.e., $E(x) = 0$ (MLR4)

The MLR4 is the assumption we need to focus on the most: does it hold in our model?

Let us consider a model with the explanatory variables x_1 and x_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

For the assumption MLR4: $E(x_1, x_2) = 0 \; Cov(x_1 = 0) \; Cov(x_2 = 0)$

The error term is independent of x_1 and x_2, so when they change, the error remains constant.

This assumption does not hold if:

- Is present an incorrect or misspecified functional form: the model is missing, for example, powers of x_1 or x_2, or, always for example, using *y* in level whether we should be using its logarithmic form.
- There is a correlation with other unobserved factors which are part of *u*.

Why $E(x_1, x_2)$ is different from 0 when we miss nonlinearities?

- True model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + u$
- Estimate: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

Where can we find x_2^2?

Everything that has an effect on *y* and is not in our model, is contained in the unobserved part of the equation.

So, is *u* independent, not correlated with our *x*'s?

x_2^2 is correlated with x_2 so in that sense the zero conditional mean assumption does not hold.

We proceed in the following order:

1. We need to estimate a model of interest (for example, a model with two explanatory variables x_1 and x_2)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

 we can test for functional form misspecification using the RESET test. In the RESET test we first estimate the model of interest (the one that we think is the true model) and then obtain fitted/predicted values.

2. After the first step, we obtain fitted/predicted values: $\hat{y}$

3. Re-estimate the model adding powers of $\hat{y}$ as independent variables.

   Example in the Wooldridge textbook: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + u$. Here the coefficients of the added powers are equal to zero, so there is no evidence of misspecification.

   Test implement in STATA: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + \delta_3 \hat{y}^4 + u$

4. F-test of joint significance of added powers of: $\hat{y}$
   - If **insignificant**: there is no evidence of misspecification
   - if **significant**: there is evidence of misspecification □ we reject the model.

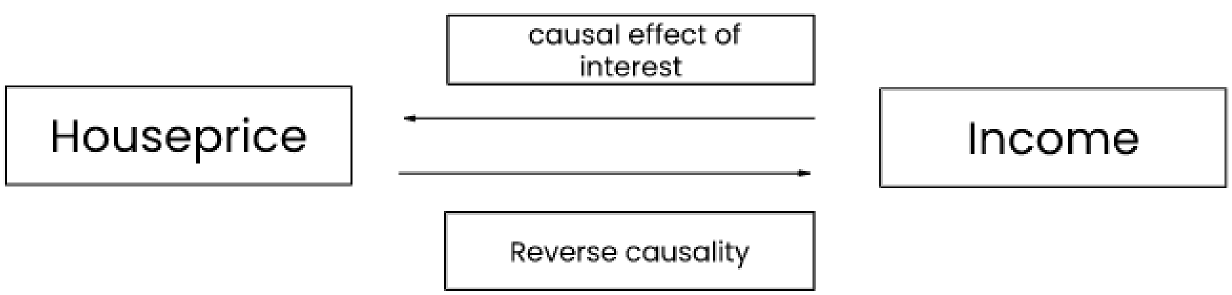
   Powers are nonlinear functions of the *x*'s. Significance means that model is missing some important nonlinearities (they were contained in the error term).

So, if those estimated deltas are statistically significant, it means that the model is missing important non linearities. We should add powers or interaction again, and re-estimate and repeat the process. So, the RESET test does not tell us what to do, but if there is a problem with misspecification in our model. It also does not tell us what to do in case of rejection, we have to try a different specification.

The other case in which the zero conditional mean assumption does not hold up, is when we have correlation with others unobserved factors.
If we take a simple model:

$$Houseprice = \beta_0 + \beta_1 income + u$$

We are interested on the fact that the average income has on the average price in a neighbourhood.



However, it could be possible (we are not sure yet) that if a house price in a neighbourhood is higher, the income is higher.

If that is the case, if *y* has an effect on *x*, we say that there is reverse causality.

If there are other factors such as: amenities in the neighbourhood, social capital, the size of the houses and other factors that have an effect on the average house price and on the income, if we estimate the model by OLS, we are estimating a mixture of all this instead of just the effect of interest. β_1 is going to catch not only the effect of interest, but also a mix of all of this.

If that is the case, the fourth assumption does not hold, i.e. income is said to be endogenous. OLS do not estimate causal effect consistently.
An explanatory variable is endogenous when it is correlated with the error term, either because there is reverse causality, or because there are other factors correlated with *x* and with *y*. and in that case OLS do not estimate causal effect consistently and what we obtain is an estimate of the association of the two variables; for example, in this case, β_1 could give us the association between income and the average house price.

There are possible solutions.
The first one, when we do not have reverse causality, is to think about other characteristics that are not included in the model but have an effect both on income and house prices and estimate a more complete model:

$$houseprice = \beta_0 + \beta_1 income + \beta_2 rotterdam + \beta_3 density + \beta_4 perc\_young + \beta_5 perc\_elderly + u$$

But is it enough? Sometimes could be, but often no.

We could try to include more relevant variables but difficult to account for all relevant variables that influence income and house prices.

There is a gold standard for estimating causal effects, which is to randomize the explanatory variable of interest.
For example, in our example, it consists in randomly allocating houses to people to see if there are different effects on house prices and income; people cannot choose where to live. Another idea could be re-allocating income in a similar way.
But in reality, we just cannot randomly tell some people to buy some houses and to others not, and we as well cannot take people's income and randomly distribute it across the population.

There are other possible solutions:

- Instrumental variables

- Panel data

We are going to look at them in future lectures.

Lecture 10 – Beyond statistical significance

This lecture is relevant not only for the linear regression, but for the entire course as well.

We start by asking the following question: is only significance important for us?

statistical significance and economical significance

Is important to distinguish the statistical significance and the economical significance, or whether the magnitude of the estimate is really relevant.

If the aim of our research is to test a hypothesis, we point out that a not statistically significant result is also a result.

For example, if we are interested in seeing if a new intervention has an effect, if we find out that we do not reject the null hypothesis that the effect is 0, this is also a result per se: we could be interested to know if we are making bad policies, for example.

Significant does not mean certain: we base our analyses on the p-value and on the confidence intervals □ if we take a random sample of the population, many time of the same size, in the 95% of the cases the true population parameter is going to be there. However, we are not sure if in this specific case our sample is there. The fact that we do not reject the null hypothesis or that we do reject it does not mean that we are accepting a given hypothesis.

Significant does not mean neither *relevant* nor *important* nor *substantial*: the size matters.
If we have a huge sample size, we are going to detect very small effects. However, if the effect is really small, we may just conclude that an intervention is not really needed, as the effect is so small that it could also be 0, even if it is statistically significant.

Our ability to find significant results also depend on other characteristics of the data like the sample size, how good are our measures, how much noise do we have.
We always have to see how much uncertainty there is around our estimate this is going to come from the confidence intervals.

Example: We want to estimate the returns to education using OLS[11] (Garcia-Gomez, 2022):

$$wages = 8 + 2.5 * years\_education$$

The wages are equal to €8 + €2.5 per year of education.
Focusing on the coefficient of years_education:

If we focus on the coefficient of years_education:

- P-value = 0.01; p-value = 0.0001

 The coefficient is statistically significant in both cases.
But if the P-value = 0.30 the coefficient is statistically insignificant.

When could this happen?
If we start from a very small sample, we get the same point estimate, but the standard errors are going to be large, but then, again, we get an estimate of 2.5.
Once we get larger and larger, the standard errors become smaller and the confidence interval as well. In this way we will obtain a more precise estimate of the effect.

If we are interested not in testing the hypothesis, but in whether there are returns to the education or not (so we do not care about testing this null hypothesis), we should be aware that there is a certain amount of uncertainty around our estimate.
For example, we make three regressions and have to make a decision based on the last one. Can we decide to get another year of education?
We get the 2.5 but we should not think only at the statistical significance. Instead, we should think at how much our wage is, and if it is going to increase either in relative or absolute terms thanks to this 2.5 increase. If €2,5 per year seems a good relative increase, it might be worth to add a few years of education at least.

economic significance (or practical significance or relevance)

We have to think if it is a large increase (for knowing this we have to see the magnitude of the coefficient).

The increase of €2.5 per year seems like a large increase, but what if, estimating a model, our β is going to be equal to 0.01 and the p-value to 0.01? It is not a statistically significant effect. Can we conclude that it is a relevant effect if our wage increases by €0.01 for every additional year of education? The effect is very small and close to 0 independently of this effect not being a statistically significant effect.

So, when we look at our results, we need to go beyond saying whether they are statistically significant or not. It is important to look at the uncertainty around our estimate and the confidence intervals. It is equally important as well to think about the magnitude. We can then look at the estimated coefficient, the absolute effect, but it also is useful to compare it with another magnitude, like average values, to inform about relative effect.

validity of results

When we think about the validity, the first question is thinking about our study per se, our samples, our analysis, and whether the study provides causal estimates for the population and setting of our study.
Does the assumption to get a causal effect hold? That refers to **internal validity**.

We talk about **external validity** when we want to use the conclusions from our study and then extrapolate them to other populations and settings.

For example, in our example we were looking at the house prices, and the income in neighbourhoods in the Netherlands. Can we extrapolate those conclusions to other countries? This will depend on how the housing market works in the different countries.

We can also think that those results were using data from 2012; can we extrapolate those conclusions to nowadays?
We can then think on how the housing market has changed. Are there reasons to think that the association between the household income and the household prices has become stronger or less strong given the dynamics over the past years?

When we have our results and it is the moment to write the conclusions, what evidence should be of our interest?
We should look for the absolute and relative effects about the importance, about what those number mean, whether they are **statistically significant or not**, the **amount of uncertainty around that estimate**, the presence of **potential biases**

(maybe our sample was not completely random, for example, and this type of discussion regards the **internal validity** of our studies, or if we extrapolate the conclusions we also have to think about the external validity.

There could be then other dimensions like, for example, the heterogeneity of effects, so whether the effects for different groups of the population are different. For example, if we think about the returns to education, we expect the returns to education to be different for men and women, for people with different ethnic backgrounds, people from different socioeconomic groups. If this is the case, we can also analyse this heterogeneity of the effects.

applied microeconometrics - module 2 – endogeneity and instrumental variables estimation

Lecture 11 (2.1) – Introduction

causality and ceteris paribus

The economist's goal is to infer a ceteris paribus relationship or to make sure that one variable has a causal effect on another variable. Another goal is knowing what is behind this causal effect.
We must be careful when we think we have found an association, because it may be tempting and we could think we had found one even if we actually don't.

Example[12] (Garcia-Gomez, 2022)
Imagine an alien observing human behaviour. He sees that some people go to the hospitals and others do not. He compares the outcome of those two groups of humans and conclude that people that go to hospitals are more likely to die. So, the alien, trying to improve human's well-being decides to close all the hospitals to save human lives.

He was inferring from an association, a causal effect, but was missing some very key information: for example, humans that go to hospitals are more likely to die beforehand because their health was worse.

We have to focus on this distinction, between associations and causal estimates □ when do OLS provide causal estimates?

□ We introduce a new instrument, the **instrumental variables regression**, that allows us to estimate a causal effect in certain conditions under some assumptions. We will discuss when those assumptions hold and if we have the right data to do so.

For the following lectures, our example will focus on what is the effect of retirement on depression[13] (Garcia-Gomez, 2022).

We see that in many countries, governments are increasing the retirement age due to the pressure on public budgets. This can alleviate how much government spend on old age pensions. But we could also wonder whether these types of policies have a spillover to other budgets.

Could these choices have an adverse effect on people's health?
If people have to work, if there is a negative effect on our health because we have to work longer, we are going to have a decreased productivity, or it could be that we end up retiring, leaving the labour force through other pathways like disability benefits.
We can also think that retiring has a bad effect on our health because we are more likely to lose our cognitive ability because we practice less, some of us gets bored, and then this could have a negative effect on mental health. If that is the case, increasing the retirement age can also have a positive effect on health □ it all comes up to it being an empirical question.

Our data come from a large European health survey, the Survey of Health, Ageing and Retirement in Europe (wave 1, 2004)[14] (A. Börsch-Supan, coordinator, 2004).
The information come from 11 countries including individuals and their spouses that are older than 50.
The dependent variable of interest is a depression scale, the EuroD (from 0 to 12). Higher values of the depression scale are associated with a worse mental health.
Our main explanatory variable of interest Is the variable "retired" (it takes value 1 if the individual is retired and 0 if not).

We will also control for other explanatory variables like age (in years), marital status and the education level, whether the individual has a low, medium, or high educational attainment.

If we estimate the regression model on STATA, we estimate an OLS with robust standard errors and then we ask ourselves what the effect of retirement on depression is. We see that the estimated effect, or association (we do not know yet what is), is positive, which means that the mental health of the individuals worsens when they retire compared to not being retired, and the magnitude is 0.7 points □ a retired individual has a mental health score of 0.7 points higher compared to a non-retired individual, ceteris paribus.

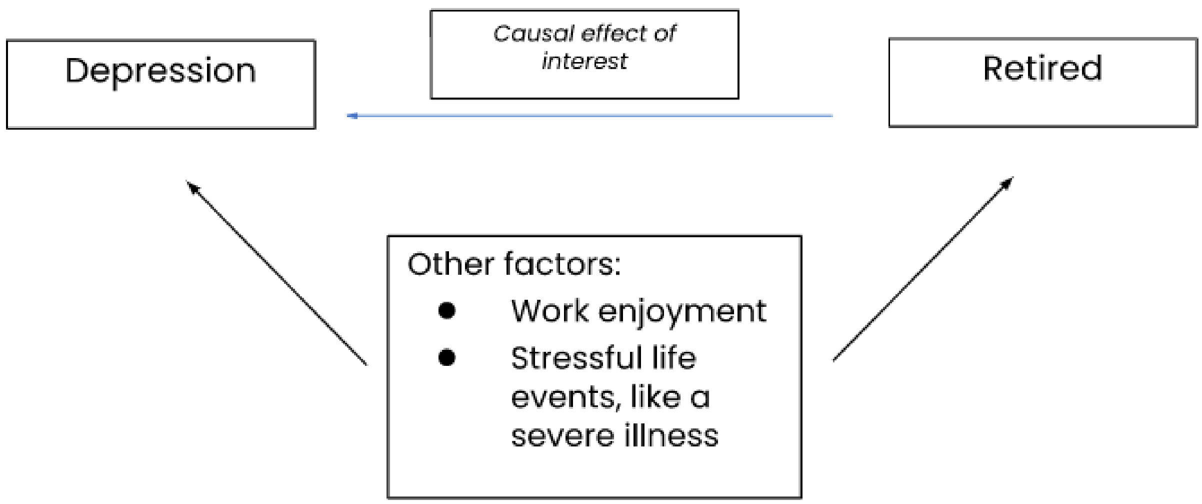
Is this a true causal effect or an association? For answering this question, we need to remember the zero conditional mean assumption: it will not be a causal effect if the zero conditional mean assumption does not hold. This happens when we have an incorrect or misspecified functional form; this is something that we can test with the RESET test, or when we have a correlation with other unobserved factors that are part of *u*. in this case our zero conditional mean assumption will not hold if there are unobserved characteristics that determine mental health and are correlated with retirement status, age, educational attainment, or marital status. If those unobserved characteristics are not correlated, the zero conditional mean assumption will hold, and if they are correlated, it will not.

So, we have to discuss whether there are correlations with other unobserved factors that are part of *u*; we will see that this will happen if we have omitted variables bias, collider bias and reverse causality.

Lecture 12 (2.2) – Omitted variable bias

The following graph shows how the following causal relationship works:

$$y = \beta_0 + \beta_1 retired + \beta_2 married + \beta_3 age + \beta_4 education + u$$



What can be present in the unobserved term?
Other factors such as the individual's level of enjoyment of its work, and meaningful stressful event of its life, like a serious illness such as cancer.
Enjoying work influences only mental health, or does it influence the probability to retire too? If the answer to this question is affirmative, then this variable is going to be correlated with retirement, and it is also going to be a determinant of depression.

Another relevant factor in this example are stressful life events: are they also likely to have an effect on whether you retire or not?
Some yes and other no.
A friend having a bad accident does not have an effect on the probability that we retire, then it is not a factor of interest for us, even if it is a stressful life event. On the other hand, there is evidence that having a severe health condition (like a cancer) increases the probability of retirement.
If this also affects our mental health (<u>but not through retirement</u>) then this is also going to have an effect on depression. Omitting any of those variables creates an **omitted variables bias**: we have an omitted variables if those variables are correlated with x and are also the determinant of y.

omitted variables bias

If we have an omitted variables bias, it could go in both directions. It could be:

- An **upward bias**
- A **downward bias**

There are going to be situations where we will not have this additional factor, otherwise, we would put this in the model. We could argue about this bias being positive or negative:

- To say that a bias is positive is the same as saying we have an upward bias: in these cases, the estimated coefficient (the value we obtain from our regression model) is larger than the population value (that we do not know: we get an estimate, and we can argue that the estimate is larger than the population value).
- To say that a bias is negative is the same as saying we have a downward bias: in these cases, the estimated coefficient is smaller than the population value.

Whether we get a positive or a negative bias, depends on the sign of the correlation of the omitted variable x_2 with x_1 (endogenous variable), the variable that in our model is correlated with the unobserved characteristics. It also depends on the sign of the correlation of the omitted variable and *y*, the dependent variable, as well.

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

[15] (J. M. Wooldridge, 2014)

When the sign of those two correlations (between x_1 and x_2) is positive, and when $\beta_2 > 0$ (which means that the omitted variable has a positive effect on our dependent variable), we have a positive bias.
We also have a positive bias when those two are negative. In any of the other two combinations we have a negative bias. For example, we get a negative bias when the correlation between the endogenous variable and the omitted variable is negative, so a correlation between x_1 and x_2 is negative, and when the correlation between the omitted variable and the dependent variable is positive. So, the coefficient of β_2 would be positive if we could have this variable in our model.

Let us think about this in our example.

What would be the bias of retired if we do not control for education? We assume for the moment that there are no other things that could be contained in *u*.

The true population regression is the following, in which we control for retired marital status, age and education:

$$y = \beta_0 + \beta_1 retired + \beta_2 married + \beta_3 age + \beta_4 education + u$$

But if in our data the variable "education" is missing, the model that we can estimate is a model in which we can control to estimate the effect of retirement on mental health.
But we can only control for marital status and for age:

$$y = \beta_0 + \beta_1 retired + \beta_2 married + \beta_3 age + u$$

Endogenous variable x_1: retired
Omitted variable x_2: education

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

(Wooldridge, 2014)

We must look at what is the correlation between x_1 and x_2 and what would be the effect of education on mental health (we remember that higher the score for mental health the worse it is).

What is the expected correlation between retired and education?
Do we expect highly educated individuals to retire earlier or later? We could expect this correlation to be negative: we expect that highly educated individuals to have less physically challenging jobs, and this could allow them to retire later.
But somebody else could have a different expectation and suppose the correlation to be positive instead of negative.
Anyway, for us the correlation is negative. Then the bias would be negative or positive? For answering this we should reflect on β_2, so we need to think about what the expected correlation between education and (bad) mental health is.

Do highly educated individuals have worse or better mental health?

For us the correlation is negative: highly educated individuals have a better mental health □ $\beta_2 < 0$

So, if β_2 is negative and the $Corr(x_1, x_2)$ is negative as well, according to our expectations, we would expect a positive bias. With different expectations, we could have a different bias.

It is recommended to reflect on our expectations and argue why we expect each of those two relationships to be positive and negative.

In the data we have, we could run an estimate on the model with education and also on the model without education.

We can then compare the coefficient of retired in the two models. The estimated coefficient in the second model (the one where we exclude the variable education), it is larger than the expected coefficient in the model that includes the education variable.

So, if the model with education is a good representation of the population model, we see that the coefficient of retired is smaller than when we exclude education.

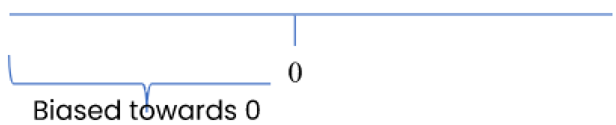
□ the model without education is a model that suffers from omitted variable bias, and then we see that we could have an upward/positive bias, because the number is higher.

It seems that probably our expectations were correct.

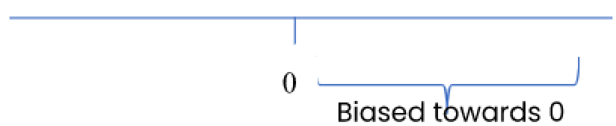
If we look at the coefficients, we cannot directly say anything about what is the correlation between education and retire. We can do that once we have access to the data. But we can see that the coefficient for edmed and edhigh are negative, so people with medium and high education have a mental health score that is lower than individuals with low education □ there is a negative correlation between education and bad mental health.

We can have upward/positive bias either if the estimated coefficient is positive or negative □ the estimated coefficient is larger than the population parameter and it does not really matter if it is on the positive or negative range of values. The same happens when we have downward/negative bias.

- Upward (or positive) bias:



- Downward (or negative) bias:



Biased towards 0: the estimated coefficient $E(\hat{\beta}_1)$ is closer to 0 than the population parameter β_1.
It is important because if we know that our estimate is biased towards 0, we can argue this is a lower bound of the estimated effect.

If we are in the range of an upward/positive bias, our estimate is going to be biased towards 0 if it is negative.

If we are in the range of a downward/negative bias, our estimate is going to be biased towards 0 if it is positive.

Should we always add more controls?
If we can control for omitted variables, if we are able to observe that they affect both *x* and *y*, we should add them to our model to be able to get this causal effect, a ceteris paribus conclusion.

We could think to add omitted variables that affect only *y*. By doing this, we will increase the goodness of fit, but adding those variables to the model is not needed for ceteris paribus conclusions.

Adding irrelevant variables (not correlated with *y*) is not a problem to get ceteris paribus conclusions, but it will have a cost in terms of efficiency: we will have a less efficient estimator □ it is better to leave them out than to include them.

Sometimes adding an additional control introduce a collider bias □ it must not be done!

Lecture 13 (2.3) – Collider bias

The collider variable is a variable that will introduce bias if we control for it[16].
It is different from what we have seen in the previous lecture, when we were worried about variables that we did not have.
The collider is a variable in our dataset that we may think it is positive to include as part of our explanatory variables, but if we include it, it will make thing worse.

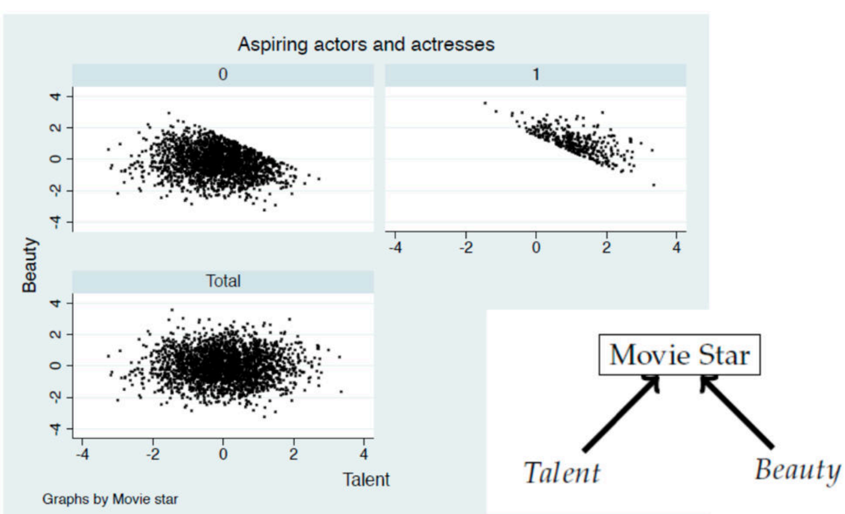
We will see some examples using **Directed Acyclical Graphs** (DAG), similar to the graphs previously used.
In these graphs the arrows indicate that there is a causal effect *x* □ *y*
"□" is a causal effect
In these graphs we also include all the common causes affecting the two variables.

The first example of a collider bias is the selection bias, which is a specific type of collider bias.

THE SELECTION BIAS

What is the importance of talent and beauty into being a movie star?
Is there a masked relation between talent and beauty in this industry?

Figure n. 1
(Cunningham, 2020 https://www.scunning.com/causalinference_norap.pdf)[17]



Let us study every aspiring actor and actress.
We can measure for each of them their beauty on a scale and what is their talent on a scale centred to zero.

We can plot all these data in a scatter plot.

Is there any relationship between these two variables? Looking at the third image in the figure n. 1 seems that there is no relationship. The ones who make it into the industry are those who are on top of the distribution of beauty and talent.

We can also observe the distribution among those who do not make it and those who make it, respectively the first and second images of figure n.1.
Focusing on the ones that do make it, it seems that beauty and talent are negatively associated: the ones who are more talented are less beauty and the other way around.
But if we do not split the distribution and select our sample, we would not find any association.
This selection creates this spurious correlation which is not really in the overall data □ this is what we call a selection bias

Now, let us return to the retirement and depression example. We assume that the DAG represents the true effect, and all the variables are dummies.

There is no arrow between retirement and depression □ There is no real effect between the two variables
The third variable, obesity, is caused by both retirement and depression.

For example, when people retire, they change eating habits, physical activity habits, etc. This will have an effect on obesity. We can make the same reasoning for depressed people.

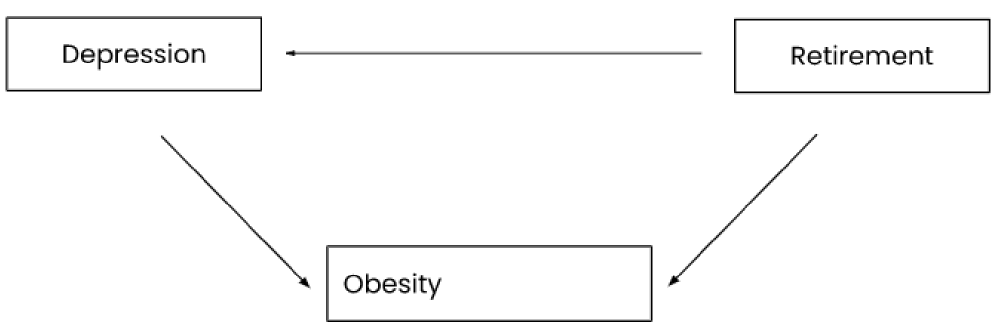
- If retirement = 1 □ it increases the chances of becoming obese (because people move less)
- If retirement = 0 □ people are less likely to be obese
- If depression = 1 □ people are more likely to be obese
- If depression = 0 □ people are less likely to be obese

What happens when we control for obesity?
What is the relationship between retirement and depression when obesity is equal to 0 and when is equal to 1?
This is what we do when controlling for the effects of other variables.

If we run a regression between retirement and depression, there would not be any association.
When obesity assumes value 1, it is likely that the person is retired because if it is, it is more likely that he is obese, and it is likely that he also is depressed, because both retirement and depression cause obesity.
When obesity assumes value 0, we have less retired people and less depressed people.

Once we control for obesity it seems that there is a relationship between retirement and depression. If we estimate now this OLS, we will likely get a positive estimated coefficient between those two variables.
□ Controlling for obesity creates a bias
□ Obesity is a collider because it creates a collider bias; this is a type of bias that exists because we introduce a variable in our model that is caused by both the explanatory and the dependent variables.

If the DAG would have been:



The reasoning would have been the same: we could get an estimate that it is not the true causal effect when we control for obesity because it is introducing the collider bias.
We could get, in this DAG, a causal effect by just including the variables "retirement" and "depression".

So, in this example, it is better not to include obesity in the model, because is a collider and introduces the collider bias in our estimates.

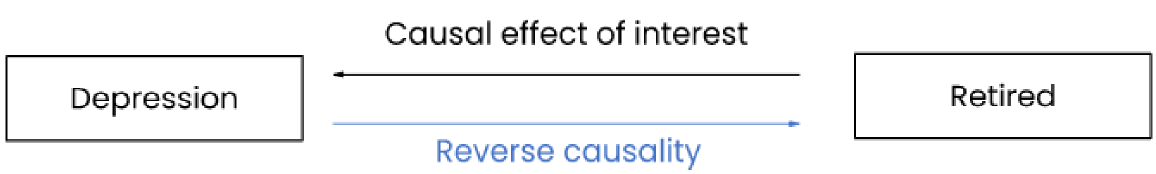
How do we select then which variables to include in our model?
There is not any existing formula that does it for us. We have to reflect on the theoretical model or the previous evidence and if there is any correlation between the two variables already known in the scientific world.

Then we have to reason about what is the expected relation between the variables. In this regard a DAG can help us because it illustrates how the different variables are related to each other.
We will need to make assumptions, as it is required every time we estimate a model and use estimators: it is all based on assumptions. We should also be able to defend those assumptions.

Lecture 14 (2.4) – Reverse causality

We currently studying the effects of retirement on mental health, and we also are controlling for marital status, age, and education.

$$y = \beta_0 + \beta_1 retired + \beta_2 married + \beta_3 age + \beta_4 education + u$$



The lower arrow indicates reverse causality.
To better understand: if our mental health worsens, we are more likely to retire.

In this case, OLS will underestimate or overestimate the effect if reverse causality is present.

□ β_1 will be biased □ the expected value will be different from the population parameter.
If depression has a negative effect on retirement, the causal effect will be underestimated.
If depression has a positive effect on retirement, the causal effect will be overestimated.

In this specific case we expect depression to have a positive effect on retirement, therefore the explanatory value of the estimated parameter is going to be larger than the true population parameter.

How can reverse causality cause a bias?
When we have an omitted variable bias, it is easier to see the reason of the correlation between the error term and the explanatory variable.
With reverse causality it is more complicated, but easy to see using a couple of equations.

$$depression = \beta_0 + \beta_1 retired + \beta_2 married + \beta_3 age + \beta_4 education + u$$

The reverse causality implies that we can estimate a model like the following:

$$retired = \gamma_0 + \gamma_1 depression + \gamma_2 married + \gamma_3 age + \gamma_4 education + v$$

Retired is the dependent variable and is described by the mental health, marital status, age, and education.

The zero conditional mean assumption means that in the first equation there is no correlation between retired and all the other variables and the error term. It means that we have to check if the covariance between *u* and retired is equal to zero or not.

So $cov(u,\ retired)$ needs to be equal to 0.

But we can also write $cov(u,\ retired)$ as following:

$$cov\left(u,\ \gamma_0 + \gamma_1 depression + \gamma_2 married + \gamma_3 age + \gamma_4 education + v\right) = 0$$

But it is impossible that $cov\left(u,\ \gamma_0 + \gamma_1 depression + \gamma_2 married + \gamma_3 age + \gamma_4 education + v\right)$ would be equal to 0 because we have the *u*, which represents all the things that explain depression other than retirement, marital status, age and education. So, this is going to be correlated with depression and this covariance cannot be equal to zero.

In this case the zero conditional mean assumption does not hold.

What should we do now?

The model where the explaining variable of interest is endogenous (without any possibilities of having this fixed) could still be very useful. Having causal effect is always preferrable, but in many cases we will only be able to get an association.
So, we must recognize that we cannot measure causal effect but simply associations. These associations often are a combination of the effects of third factors, reverse causality and the real causal effect.

This will have an effect on how we interpret our results. We have seen in our example that the zero conditional mean assumption does not hold, so we will not say that retiring will increase mental health for sure, but we need to express our interpretation in terms of an association.
In our example, we can say that on average the mental health status of retired individuals is higher than the one of non-retired individuals by 0.7 points, after controlling for the other variables.

We cannot draw ceteris paribus conclusions as we cannot be sure that all the other things do not change at same time, and we mention the variables we control for as they are included in our model.

Lecture 15 (2.5) – Estimation and interpretation of instrumental variables

instrumental variables

One of the tools we can use to estimate the causal effect is an instrumental variable, if we have the appropriate data and the required assumptions are satisfied.
With an instrumental variable we are able to isolate exogenous variation. If we think about the variation in any given variable, it consists of two parts:

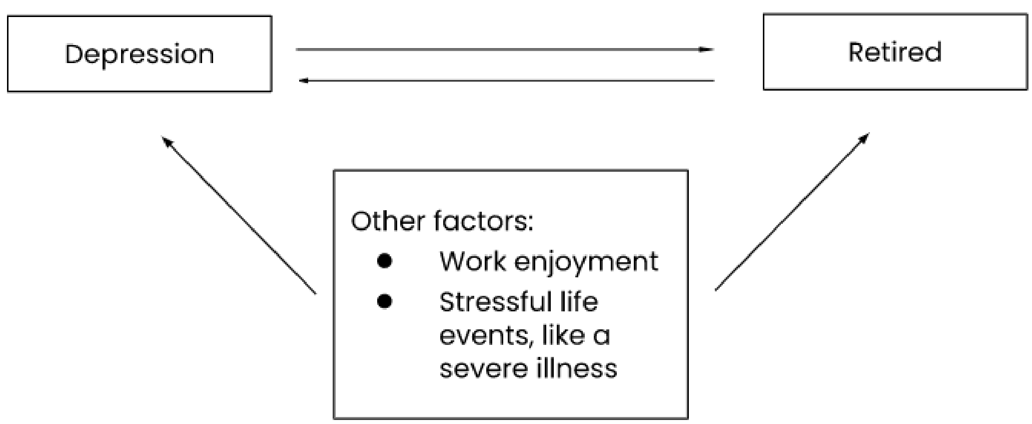
1. **Endogenous variation**: the part of the variation that is correlated with the unobserved.

2. **Exogenous variation**: the part of the variation that is independent of the unobserved.

If we think about all the reasons people retire, there are reasons that are endogenous to the individual, so they are related to what we cannot observe. Examples of this could be, if that person is enjoying his/her work, whether that person has a health impairment, etc. All these variables are related to mental health, and it also explain why people do retire. Maybe there are other reasons that motivate people to retire, related to retirement, but not related to the mental health of the individual. These reasons are what we have identified with exogenous variations.

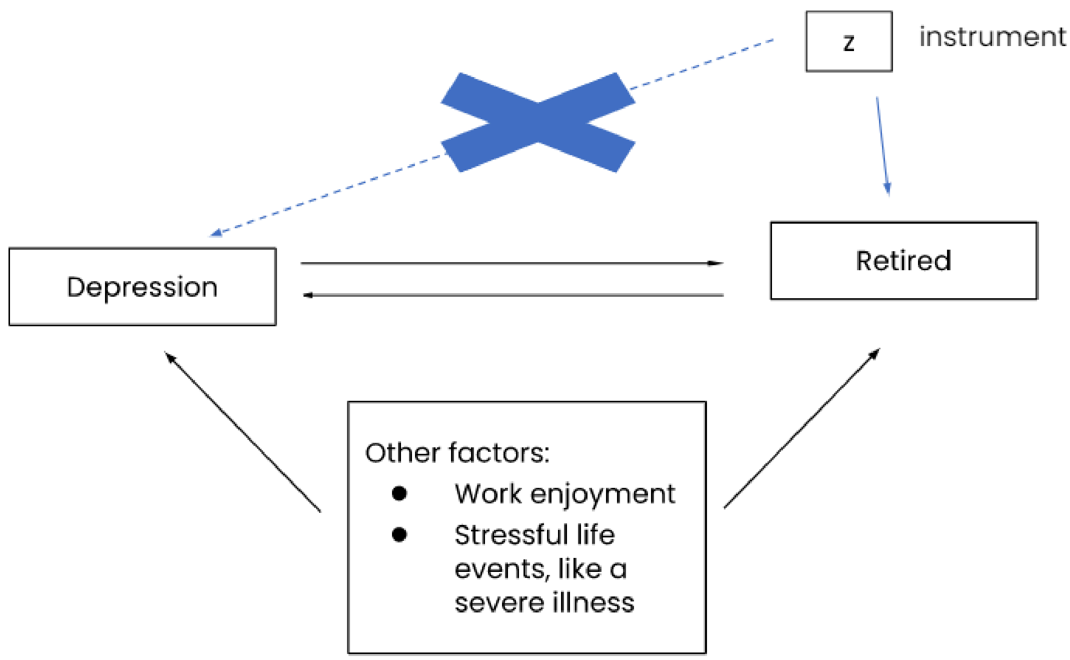
So, we have to isolate the exogenous variation to focus on the causes of why people retire and then use them to estimate a causal effect.

If we look at the following graph:

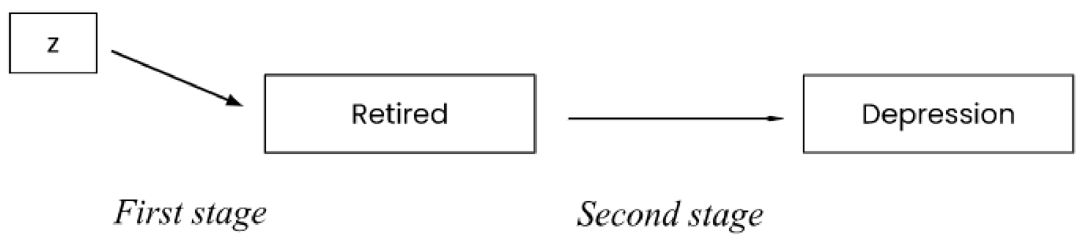


Mental health can also have an effect on the probability that someone retires. Then we have other factors that have an effect both on retired and depression.

The idea behind this instrumental variable is that this variable is able to capture some information that explains why some people retire (this variable is our instrument). This variable has not to be correlated to depression, it has only to influence retirement. This does not have an effect also on the unobserved characteristics.



The instrumental variable estimator looks first at the effect of this instrument on the probability that someone retires. This is the first stage. Once we get these predicted reasons on why people retire for exogenous reasons, then we use this exogenous variation to look at the effect of retirement on mental health. This is the second stage.



We will first look at the first arrow in the figure and then at the second. What could be a variable that we could put instead of z? What could explain that people retire but it is uncorrelated with depression? In many countries, once people reach the early retirement age or the normal retirement age they do retire at that specific age. If we plot for any country the probability that someone retires, there will be a spike at those specific ages.

We can use this variation whether someone is over their normal retirement age, for example, to predict whether someone is retiring or not, and then use this exogenous variation to look at the effect of retirement on depression.

To do this we can use a **two stage least squares estimator**.
In the <u>first stage</u>:

$$Retired = \pi_0 + \pi_1 edmed + \pi_2 edhigh + \pi_3 married + \pi_4 age + \pi_5 full + v_2$$

We estimate retired as a function of education, marital status and age, with an additional variable, the full, which is the age at which you are entitled to *full* retirement benefits in the European countries.

We have variation in this variable because it varies through different countries, and even within some countries there are different retirement age for men and women (at least at the year of the data).

From the first stage we get the predicted retirement: $\hat{retired}$

We then insert this predicted value in the <u>second stage</u>:

$$y = \beta_0 + \beta_1 \hat{retired} + \beta_2 edmed + \beta_3 edhigh + \beta_4 married + \beta_5 age + error$$

In the second stage we estimate by OLS. This model has the same explanatory variables as before, but instead of having the observed retired we have the predicted retired.

Given that in the first stage we control for educational status, educational attainment, the marital status and age, the variation that is picking from retired comes from the fact that someone has reached their normal retirement age, the *full* age.

We can estimate by OLS the first stage and then the second stage. If we were to do this, the standard errors in the second stage could not be correct because we need to consider that retired is not the observed retired, but only a prediction.
Stata does everything together to correct the standard errors in one go, so we do not have much to worry about. We only need to tell Stata which variable is our endogenous variable, and which is the instrument. We can also ask for heteroskedasticity robust standard errors as we did in OLS.

In Stata we first see the first stage regression where we see that we have this variable full. We will not interpret this output because our dependent variable "retired" is a 0-1

variable and we have not seen yet how to interpret such an outcome (we will do this when we talk about binary models).
What we cannot really say is that we see that full is statistically significant in the first stage, and that those who go over this age are more likely to be retired. This is how we can interpret the sign of this coefficient.

Moving to the second stage we have all the variables and the variable retired. Comparing its coefficient to the number we had before, the coefficient is larger and what it, still, suggest, is that those that retire have a worse mental health status compared to those that do not retire. We can interpret this coefficient in the same way as we did before: the mental health index increases by 2.2 points (on a scale 0-12) when an individual retires compared to being non retired, ceteris paribus.

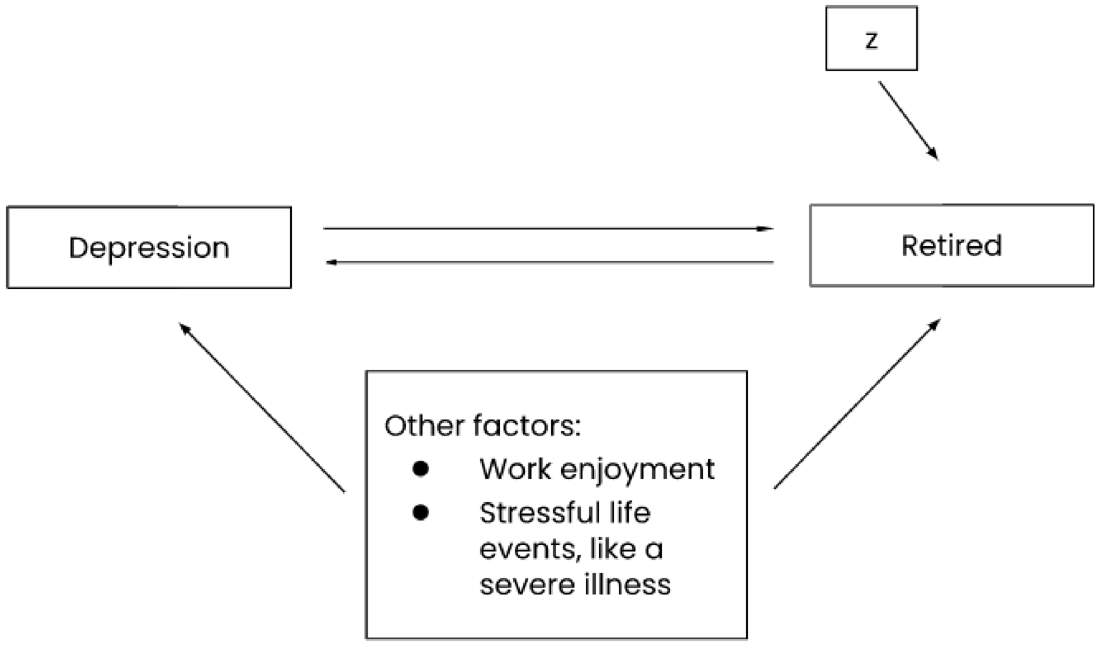
We can see that this effect is statistically significant. If we look at the confidence interval, we see that now it has broadened. One of the things that happen with instrumental variables is that because we only use a small part of all the variation, the estimate is going to be less precise, and it is important to see how much less precise this estimate becomes.

We can use the same method if we have more than one endogenous variable in the model. We can use instrumental variable regression in the same way (we only have to estimate the first stage for each of the endogenous variables and we would need more than one good instrument; we should have one for each of the endogenous variables). In fact, we always need the number of instruments to be larger or equal to the number of endogenous variables. But it is hard to usually find this exogenous source of variation, so it will be nearly impossible to be able to control for more than one endogenous variable in each model. But if, somehow, we can manage this, once we have the second stage, we interpret the coefficients in the same way we used to do when we had an OLS.

Lecture 16 (2.6) – Instrumental variables assumptions

Our model of interest is a model in which we want to estimate the effect of retirement on depression, and we want to use an exogenous variation that it is not related to depression causing retirement, or the influence of other omitted variables.

In order to do that we need to find an instrumental variable that is correlated with the endogenous variable.



The assumptions are the following:

- The instrumental variable must be correlated with the endogenous variable:

$Cov(retired,\ z) \neq 0$

  In this case we say that it is a **relevant instrument**.
  We need not only a correlation between the two variables, but a strong correlation. They need to explain a large part of the retirement decision: we need a **strong instrument**.

- The instrument must not be correlated with any other (unobserved) determinants of depression:

$Cov(u,\ z)\ = 0$

  In this case the instrument is exogenous. So, our instrument can only affect the dependent variable through the endogenous variable, after controlling for all the other variables in our morel. After this we have a **valid (exogenous) instrument**.

relevance assumption

The relevance assumption is that the covariance between the endogenous variable and the instrument is different from zero:

$$cov(endogenous\ x,\ z) \neq 0$$

This is something that we can conclude from looking at the first stage regression, where we predict our endogenous variable, using the instrument, so we can see whether the instrument is statistically significant or not. If it is not, in the second stage we will get very large instrumental variable standard errors. It is important that we have a strong correlation.

If we look at the first stage on Stata at the coefficient of full, we see that it has a t-statistic of 9.81 and that the p-value is very small too □ we can say that this instrument is highly significant.

The more variation in the variable *x* our instrument explains, the more variation will be available in the regression. Then we will have more variation to be used to look at the effect of retirement on depression. There is a little problem when we explain very little (so when our instrument is not strong), but we have a great instrument. In that case, the IV are no longer reliable.
We can only use instrumental variables if we have a strong instrument.

How do we check for weak instruments?
There is a rule of thumb that we can use when we have one endogenous variable: we must look at the F-test of the first stage, and it has to be larger than 10; if it is smaller, it suggests that the instrument is weak.

validity assumption (or exogeneity assumption)

This assumption refers to the correlation between our instrument *z* and the unobserved component. Any assumption that we make about the relationship between the unobserved components and any other variable is an assumption that we have to argue, and we will not be able to fully test it.
There are some partial tests, but they will not be covered here because none of those tests will be able to tell us if our instrument is valid or not, and they are often misused in the literature. Some people claim that passing this test means that the instrument

is valid, but then we could find other reasons for which this assumption could not be satisfied.
We have to use economic theories, our own reasoning and previous evidence to argue about the validity assumption.

In this case, for example, to justify why using the normal retirement age is exogenous or not, we can use our knowledge of the institutional setting and we know that the way in which normal retirement ages have been defined is independent of the mental health of the population.

Lecture 17 (2.7) – Instrumental variables vs OLS

To choose between IV and OLS there are the following steps to follow:

- We have to use IV only if we have a valid and strong instrument.
  Before even running any IV regression models we must think about the assumptions of our instrument. Does it satisfy them? Is it valid? Is it not correlated with the unobserved term? Is it strong?
  If we have a valid but weak instrument, we will not get a reliable IV estimate.
- We need to have large sample size.
  When we have a valid and relevant instrument, IV is consistent, but under endogeneity IV is biased. So, we must rely on this consistency property that only applies to large sample size. If it is small, we are no better off using IV instead of simple OLS.
- We choose IV if our explanatory variable is endogenous; if it is exogenous, we do not gain anything by using IV, but only lose information because IV are inefficient and do not use all the variation that we have in our explanatory variable. We throw away information if we want to gain in terms of getting a causal estimate, even though it is not a gain if we do not get a more reliable estimate.
  Our IV is going to be better the more highly correlated our instrument is with x. So, then we are going to get a smaller variance of the IV estimate. The stronger the instrument, the less we will lose in term of efficiency in terms of IV.

We have already discussed the first and the second conditions, but how do we determine if an x is endogenous or not? We do not observe the error term, and we do not observe the unobserved characteristics, so we cannot just check the correlation between the unobserved and *x* to determine whether is endogenous or not.

We can use our instrumental variable estimator. To inform about this endogeneity. We can use the same logic behind IV to test for the exogeneity of retired:

1. We first estimate the first stage equation:

$$Retired \ = \ \pi_0 + \ \pi_{1edmed} + \ \pi_2 edhigh \ + \ \pi_3 married \ + \ \pi_4 age \ + \ \pi_5 full \ + \ v_2$$

2. From the equation we get the residuals: $\hat{v}_2$. What are they picking up? The variation in retired that is not explained by the educational attainment, marital status, age, and our instrument. This is the potentially endogenous variation.
3. We then can estimate our main equation with OLS, adding residuals of step 2:

$$y = \beta_0 + \beta_1 retired + \beta_2 edmed + \beta_3 edhigh \ + \beta_{4married} + \beta_5 age + \ \delta_1 \hat{v}_2 + error$$

   We have our main equation and the residuals that are picking the potentially endogenous variation of retired. If this variation is correlated with our dependent variable, then δ_1 is going to be statistically significant.
4. We test the significant of δ_1: if it is insignificant there is no evidence of endogeneity and if it is significant there is evidence of endogeneity. It can be shown that if we do this procedure the β_1 we could get for retired, is also similar to the one that we get if we use the IV.

Stata can do all the steps above all at once. The null hypothesis is that the variables are exogenous. We get to different statistics, similar in 99.9% of the cases. In any of those cases the p-value is very small, so we reject the null hypothesis. Is retired endogenous? If we reject the null hypothesis that retired is exogenous, so this evidence that suggest that retired is endogenous.

If retired was endogenous this would be the third case in which we would prefer IV. We have been able to argue about the two assumptions in the previous lectures: we have a large sample size, and our explanatory variable is endogenous: this is a case in which we would prefer IV.

Many times, we will not have a valid and strong instrument. The golden rule for getting a causal effect is randomization, but this is not always possible neither, because it is very expensive or because there are ethical issues (imagine we want to look at the effects of going to prison on the probability of committing future crime). But there ethical about putting people in prison. We could also see the effects of providing hospital care versus not, or even retirement; we cannot tell people to retire at random.

Another solution we can apply is the use of panel data methods.

Lecture 18 (2.8) – Potential outcomes and DAGS

We will look at two useful frameworks to analyse causal effects:

- DAGs
- Potential Outcomes Framework

DAG (Directed Acyclic Graph)

DAGs are the graphical representation of a chain of causal effects[18] (Garcia-Gomez, 2022) in which the nodes are the representation of random variables. The causal effect between two variables is represented by the arrows, which also indicate the direction of causality. If an arrow goes from variable A to variable B this arrow indicates the direction in which we assume that the causal effects work out. If there are no arrows, we assume that there are no causal effects between the two variables.

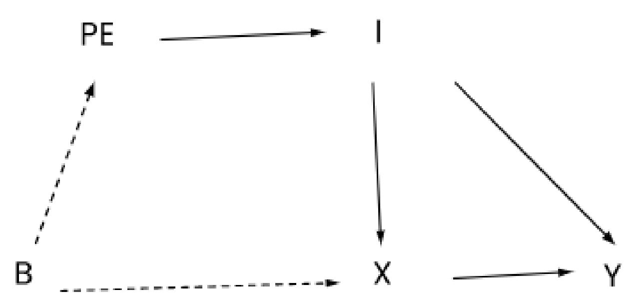
The causal effect between the independent variable x and the dependent variable y can be direct but also indirect: in this latter case we say that the causal effect is mediated by a third variable (x □ D □ y which means that the variable x influences the variable D, which has then an effect on the variable y). if we think about the causal effect of income on health, there are reasons why an additional euro per se could improve our health, but there are many more reasons on why income may indirectly have an effect on health. For example, higher income makes it possible to have a better quality, or it can grant us the access to health insurance, etc. All these effects are not direct but are mediated by third variables.

DAGs are useful to graphically represent the relationships variables and to acknowledge the assumptions and the data available in our specific setting. If *x* and *y* are connected by a solid arrow, we know that we have this information that we observe. If instead we have a dotted line, we know that it is an information we do not observe. Looking at these relationships should help us think whether this information that we do not observe is relevant or not to estimate a causal effect.

We are interested in knowing the returns on education, so how much does a schooling brings in terms of educational attainment.
We have the following variables:

- X = educational attainment
- Y = earnings
- I = family income
- PE = parental education
- B = family background

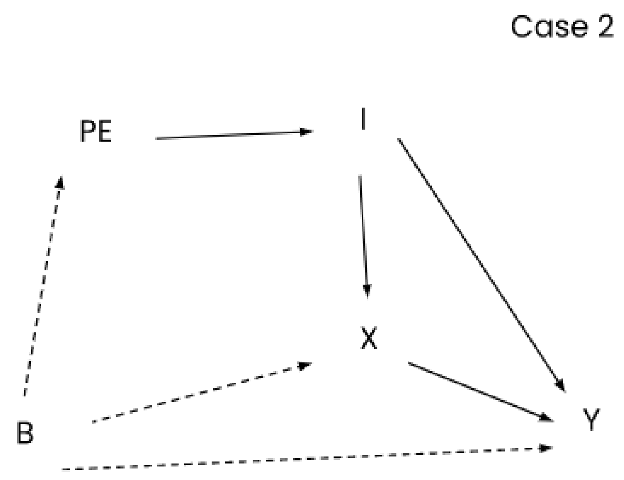
Case 1



How much an additional year of education translates into higher earnings?
In our theoretical framework, also family income plays a role: higher family income gives us easier access to certain positions, and that increases our earnings. A higher family income also affects our educational attainments because it allows us to attend better schools.
We also know that some family background characteristics influence both our educational attainment as well as parental education. Parental education has an effect as well on family income.
When we have the dotted line, we know that the family background information may be observable for individuals and their families, but not to us.
If we have this information, can we estimate the causal effect on educational attainment and earnings?

That is possible because when we observe for X and Y for family income, there are no other unobserved. This means that there is no other channel in which parental education or family background has an effect on the earnings. If this was the setting, it would be possible to estimate the causal effect.

If family background also influences earnings:



If family background also has an effect on earnings, by controlling for X and I could not be sufficient to estimate a causal effect of educational attainment on earnings, because we could have this omitted variable bias as family background also has an effect on Y and X.

The DAG is useful to clearly illustrate the assumptions that we are making, the data we are able to observe and help us see whether we estimate a causal effect or if we are missing because of the effect of these additional variables. In other examples we could also have a collider. In any case it makes it impossible to get a causal effect. We get an unbiased estimate of the causal effect, and we can talk about association or partial associations.

Potential Outcomes Framework (POF)

It is not a graphical interpretation, but a tool for when we are interested in the effect of an intervention or a variable on an outcome Y.
To keep it simple, let us think that our intervention is a binary.
For example, as we were looking throughout this topic of looking at the effects of retirement on depression, our treatment T could be retired.

If we think about the returns on education, for example we could think about our indicator as getting a college degree versus not. Treatment is a general term: it could be a medical term used in the medical field, but it can also be part of a training program, it can be retired, it can be getting college education □ it is the effect of the variable we are interested in.
If we think for every individual, we have two potential outcomes. Returning to the example of the returns to education, 1 could be the earning that that person could get if goes to college and 0 the earnings that that person does not get because he did not go to college.
Thinking about the effects of retirement on depression, for a given person in a given year, we are interested in the mental status of that person if retires and the mental status of that person if does not retire. Those are our two potential outcomes.

$$Y_i = \{Y_i(1) \;\; if \;\; T_i = 1 \; Y_i(0) \;\; if \;\; T_i = 0$$

We can then define the treatment or our causal effect as the differences in between those potential outcomes. So, the difference in our mental health, if you retire compared to not retiring. The difference in our health if we go to the hospital compared to not going. The difference of the earning if we go to college minus the earning that we could get if we do not go.

$$\Delta = Y_i(1) - Y_i(0) = (T = 1) - (Y|T = 0)$$

This is how our data could look like:

Individual	**Treated**	$Y(1)$	$Y(0)$	**Causal effect**
1	0	$Y_1(1)$	$Y_1(0)$	$Y_1(1) - Y_1(0)$
2	0	$Y_2(1)$	$Y_2(0)$	$Y_2(1) - Y_2(0)$
3	0	$Y_3(1)$	$Y_3(0)$	$Y_3(1) - Y_3(0)$
4	1	$Y_4(1)$	$Y_4(0)$	$Y_4(1) - Y_4(0)$
5	1	$Y_5(1)$	$Y_5(0)$	$Y_5(1) - Y_5(0)$
6	1	$Y_6(1)$	$Y_6(0)$	$Y_6(1) - Y_6(0)$

[19] (Garcia-Gomez, 2022)

$$ATE = E\left[Y_{j(1)} - Y_{j(0)}\right] = \frac{1}{N}\sum_{j=1}^{N}[Y_j(1) - Y_j(0)]$$

For a given individual, that individual could be treated □ we would have the outcome if that individual could be treated and the outcome if he could not be treated. The same if the individual retires or not: the causal effect is the difference.

If we want to estimate the average causal effect or the average treatment effect, we could just get the average over all these individual causal effects. So, we have the causal effect for individual one, two, and so on.

The fundamental problem to causal inference and to get a causal effect, is that we can only observe one of these two outcomes: we cannot observe the same person in two different states: at the same moment, I am either retired or not. It is not possible to look at both the states of the world, but we need to think about in which situation we are.

We want to know what the outcome of those non-observed could be, the counterfactual outcomes, which are the outcomes that could be observed if the person was in a different state (we have these gaps in our dataset). We only observe the outcomes of an individual given his or her current condition.
The goal of the analysis is to try to generate those counterfactuals.

	Outcome *without* Treatment	Outcome *with* Treatment
Control	$E[Y_i(0)\|T_i = 0]$	$E[Y_i(0)\|T_i = 1]$
Treatment	$E[Y_i(1)\|T_i = 0]$	$E[Y_i(1)\|T_i = 1]$

[20] (Garcia-Gomez, 2022)

If we think about going to the hospital/doctor example, in the data we see the output of those that did not go to the hospital (the control group), and the output of those who did go to the hospital (the treatment group). We need to be able to estimate what could be the expected output of those that did not go to the hospital if they had gone, and the other way around.

We need to ask ourselves whether our empirical method give us a valid counterfactual. If we think about the effect of going to the hospital, we can just compare the mortality outcomes of people who did go to the hospital and those who did not go to the hospital.

$$E[T = 1] - E[Y_i|T = 0]$$

The effect of going to the hospital is obtained comparing those two. If those that go to the hospital, or those that retire, or those who go to college, are different compared to those that do not receive the treatment, do not retire, do not go to college, then we cannot just use the information from this counterfactual group to estimate our causal effect.
When there are other observed or unobserved characteristics that influence both the treatment and y, in other words, our explanatory variable of interest and y, our method will not give us a valid counterfactual.

If those reasons why it is only due to other observed characteristics that confound this relation, then we can just add these explanatory variables to our model, and then OLS with this additional control explanatory variables will give us the causal effect, the average treatment effect, this right counterfactual.
However, if we also have unobserved characteristics that explain why people go to the hospital and at the same time why they die, or why people retire and at the same time their mental health, OLS provides bias estimates we need to use a different estimator.

This is another way of thinking about what it means to get a causal effect. We should always check how the assumptions of the different methods are fitting in our research. At the same time DAG will help us visualize the assumptions we are making in our specific contribution.

Applied microeconometrics - Module 3 - Introduction to empirical methods: panel data

Lecture 19 (3.1) – Panel data: introduction

Which estimation technique we should use when working with panel data?
To answer this question, we need to know:

- What panel data is
- The notations of panel data
- The distinction in the notation from cross-sectional analysis

- Unbiasedness and efficiency
- Different estimation techniques, including:
    - OLS (and the conditions under OLS will be the best estimation technique with panel data)
- The within effects estimators, which refer to fixed effects, least squared dummy variables and first differences; we also want to know which assumption they require and how these estimators solve potential OLS issues
- Random effects and correlated random effects
- How to choose across the estimators
- Definition and implication of attrition for panel data

At the end of this module, we will also introduce difference in differences, which is an approach that can solve some of the issues that within estimators might not be able to.

challenges of panel data

 We will introduce a set of new estimators, each with their own working and assumptions
 The choice of which estimator to use depends carefully on the topics of bias and statistical efficiency
 With panel data we now have the distinction between time-variant and time-invariant covariates

After this module we should be able to:
- Describe a panel data set and distinguish it from a cross-sectional
- Understand the assumptions required for each estimation technique and under which of those they are unbiased and efficient
- Knowing how to select the best a most appropriate estimation technique
- Evaluate and understand the importance of attrition in different panel data sets

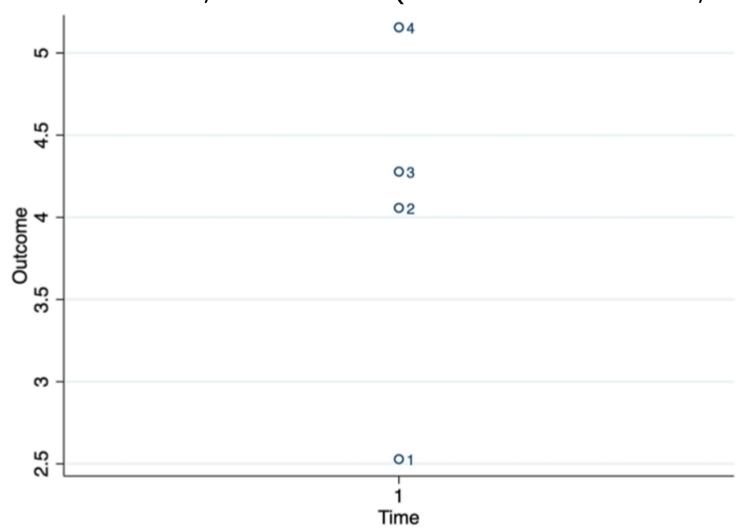
Lecture 20 (3.2) – What is panel data?

In research, data could have different structures:
- **Cross-sectional data**: the cross-sectional data are those data sets that have collected data on one observed sample of units at only one point of time.

- **Repeated cross-sectional data**: in this case the data collected are the of same type of the cross-sectional data, but on different observed samples and at different points of time. We are not following the same person, but we are collecting the same data.
- **Longitudinal/panel data**: the panel data is a data set that collects the same data on the same observed sample at different points of time □ we will collect the same data, the same variables on the same people and units over time.

In the following figure we can visualize cross-sectional data:

Figure 1 module 3, lecture 3.2[21] (Carlos Riumalló Herl, 2022)

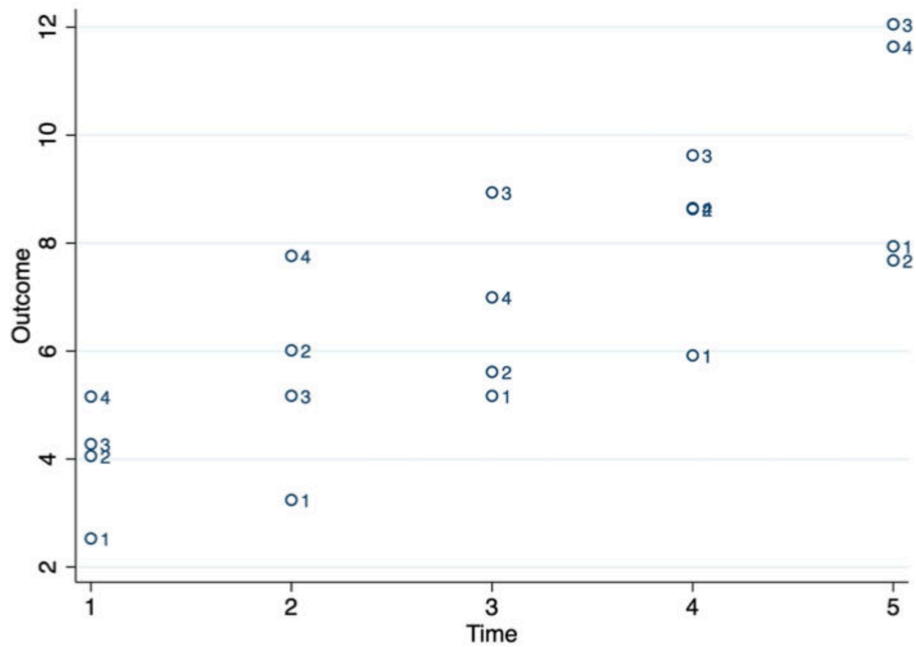


We can think of a cross sectional data as a snapshot on time.

Figure 2 module 3, lecture 3.2[22] (Carlos Riumalló Herl, 2022)

In the case of repeated cross section data (figure above), we are collecting the same data, so the same variables and the same information, but on different samples and in different time periods □ this could be imagined as two snapshots at two different people in two different points of time.

Figure 3 module 3, lecture 3.2[23] (Carlos Riumalló Herl, 2022)



With panel data we collect the same information on the same people over time.

panel data characteristics

A panel data contains repeated observations of the same unit across time.
The data is now characterized by having two dimensions:

- **Individual unit dimension** □ indicates which units we are observing (the unit could be people, countries, stocks, schools, etc.) □ $i = 1, 2, ..., N$
- **Time dimension** □ it gives us information on the period of time during which the data are collected. It could be years, month, seconds, etc. □ $t = 1, ..., T$

Now we can define panel data in two ways:

- **Balanced panel dataset**: all units are observed in all time periods □ we observe all people all the time.

- **Unbalanced panel dataset**: in this (more frequent) case the number of observations vary across individuals □ whatever the reason, we are not able to follow each individual for the entire period of time.

Some of the examples of panel data includes:
- Longitudinal surveys of Ageing (HRS, ELSA, SHARE, etc.) □ in this case the data have been collected on people every two years.
- Administrative tax records □ some countries (including Denmark and the Netherlands) collect annual tax data and make their information available for research □ each person is followed every year.
- Hospital characteristics □ in this case units are not people, and in general units do not necessarily need to be people □ in this case hospitals are observed over a period of a year; we could observe their spendings, the number of operations they do and so on.
- Financial stocks □ we could follow what is happening at a stock in every point of time □ this is a panel data set.

Figure n. 4[24] (Carlos Riumalló Herl, 2022)

	Person	Time	Age	Male	Outcome
1	1	1	47	1	2.089202
2	1	2	48	1	4.522359
3	1	3	49	1	6.552687
4	1	4	50	1	5.527746
5	1	5	51	1	7.380176
6	2	1	47	0	3.993593
7	2	2	48	0	5.705793
8	2	3	49	0	7.461751
9	2	4	50	0	7.695805
10	2	5	51	0	8.791395
11	3	1	48	1	4.401231
12	3	2	49	1	7.169417
13	3	3	50	1	6.500992
14	3	4	51	1	10.39803
15	3	5	52	1	8.463883
16	4	1	47	0	5.891078
17	4	2	48	0	6.662832
18	4	3	49	0	7.206686
19	4	4	50	0	9.569715
20	4	5	51	0	9.225389

When looking into the data □ each observation of our dataset represents one unit time □ it represents a particular outcome of one unit at a particular time set

Looking at the figure n. 4, we can see that the first observation represents the observation for person 1 at time 1, while the second observation corresponds to person 1 at time 2, and so on.

So now we have multiple observations per unit, each one in different points of time.

When we are working with panel data, we always have two fundamental variables:

- The **identifying unit variable**, in our example the variable “person” □ indicates which observation belongs to the same unit □ in our example the variable person contains the number of the ID for the unit we observe (1) □ the first five observations belong to person one because they are identified as such.
- The **time variable**, in our example the variable “time” □ with this variable we can see which observation belongs to which time frame: in this example we have observations that go from time one to time five.

The fundamental characteristic of a panel data set is that we now have a series of observations that belong to the same unit. For each of the person in the panel data of figure 4, we have five observations in five different time points. In this case the variable person allows us to identify which observation belongs to each person.

With panel data sets we also have another two types of variables:

- **Time-variant variables**
- **Time-invariant variables**

In our example, the variable *age* is a time-variant variable, as the value of age changes for each person at each point of time, while the variable *male* is a time-invariant variable, as it contains the same value over the whole timeframe in which we are following people up.

□ The concept of time-variant and time-invariant might not be absolute, but might depend on the data set we are using.
For example[25] (Carlos Riumalló Herl, 2022), if we are using a survey that collects data on older people and we have information on things like education, those might not change for older people, but it could for young people. In the end, the concept of time-variant and time-invariant might depend on the dataset that we are using.

panel data structure in practice

Whenever we are looking to a variable, there will be variations in its values.
While in the case of cross-section data we already had variations, in the case of panel data sets we need to distinguish the variation in two parts:

- How much of the total variation is due to variation within units
- How much of the total variation is due to variation between units

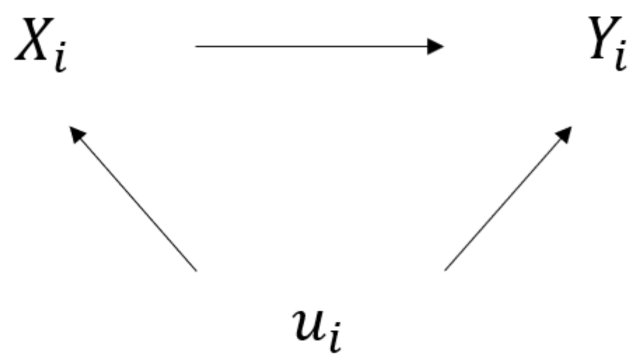
As for the within variation, we look at a single individual and we are going to look at how much the value of a certain variable changes for that individual with regards to the unit's mean value for that variable.
 The concept of within variation is relative to the unit.

In the case of between variation, we want to know how much variation of the variable exists between units. In this case we will calculate this by comparing the mean value across different units.

This analysis of the variation will tell us how much variation is occurring within individuals (how much that variable changes over time within units), or whether most of the variation exists between units.
We can ask ourselves: do people change a lot over time in reference to that certain variable, or is most of the difference due to differences between people for that variable?
Do the units per se change over time for a given variable or the difference is given by differences between the individuals?
In the applied part it will be seen how to describe and interpret these information on STATA.

Lecture 21 (3.3) – Notations

With panel data we also change the way we visualize some aspects of the models and the definitions.
This is a DAG for a cross-sectional data:



X_i has an effect on Y; we have then an unobserved error term that might affect both our covariate of interest, X and Y.

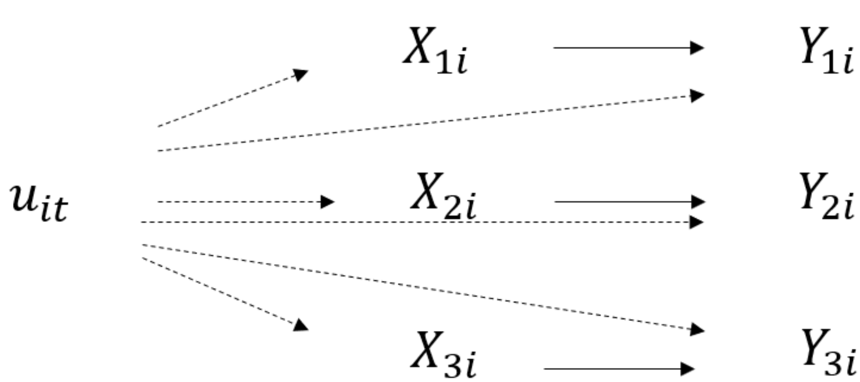
We can use the classical OLS notation:

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

We can write *y* as a linear function of β_0 and $\beta_1 X_i$ and the error term u_i.
We can add more covariates, there will be only one outcome, defined for a unit i and in this case there is no time dimension.

In this case using OLS gives us an unbiased estimate on the parameter β_1 if the expected value of the error term conditional on x is equal to 0:
□ $\hat{\beta}_1$ is unbiased if $E(X_i) = 0$

With panel data we have another dimension: time.
How can this influence the DAGs and the type of notations we use?



For each individual i now we have multiple observations. In this case we have the effect of X on Y at three different points of time, all for the same individual, i.

In this scenario we also have the error term, which has the dimensions unit and time as well. The error term not only influence X and Y in one particular time period, but it can influence the variables for the entire period of observation.

We can think of a DAG for panel data as the cross-sectional diagram multiplied for all the time periods we have.
This is a simplified version of the DAG, because we can also consider past xs: for example, X_{1i} having an impact on Y_{2i} and vice versa.

□ Now we will have multiple X's that will depend on the time period we observe, and now we have an error term that does not only influence X and Y of one time period, but of all time periods.

panel data notation: decomposing the error

The outcome now is unit and time specific. What is its notation?

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

This is a simplified model: we can always add more covariates and eventually past covariates (this is called a "**lag**"). Now the error term, that varies both across unit and time, can be decomposed in two parts:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + \varepsilon_{it}$$

Now the error term is distinguished into:

- α_i □ in the literature, this component is referred as the **unit heterogeneity**, or **individual fixed effect** or **unit fixed effect**. It is the part of the error term that is the same for all observations of one unit □ is a persistent component and the same for all observations of a unit. It is also unique to each unit and can distinguish the units between each other. It does not necessarily have to be different than zero, but if it exists it is time-invariant in the person □ <u>this is the part of the error term that is time-invariant within units</u>.
- ε_{it} □ it is referred as the **idiosyncratic error**, and it is the time-variant component of the error term. It can be different for each unit and time period. We are referring to a part of the error term that is unique for each unit and time observation and for it we will create this idiosyncratic variation. The

idiosyncratic variation could be equal to zero, but does not have to, and in this case it will distinguish which types of estimation we can use and which types we cannot use.

Example[26] (Carlos Riumalló Herl, 2022): Let us study again what is the effect of education on income: with panel data, income (the outcome for and individual i at time t) would be a linear function of the parameters β_0 and β_1 and the education of the individual i at time t.

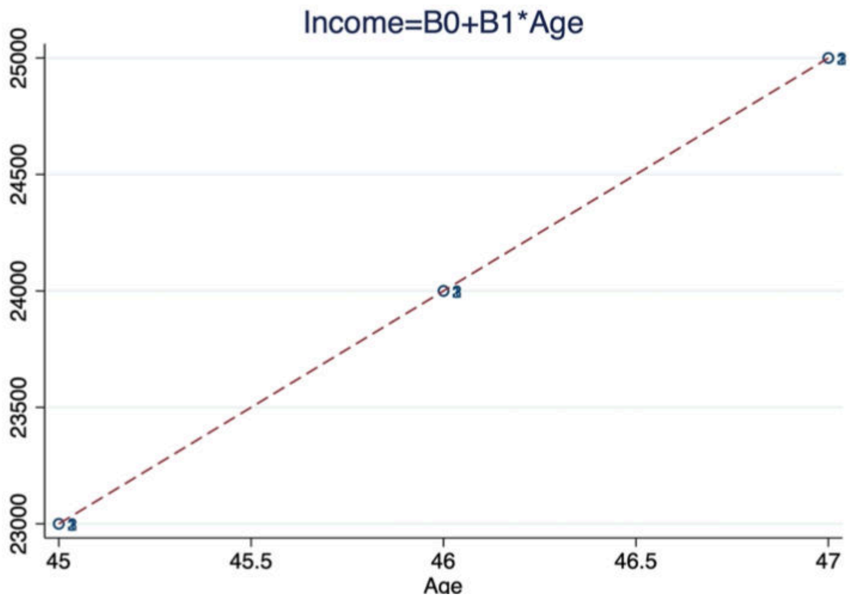
The data set for this example is the annual tax register, that allows us to follow individuals on an annual basis. From the register we have information on the education level as well as what type of income they receive.

We have the error term u_{it} (the error term for individual i at time t) that can be decomposed in the time-invariant component (the individual heterogeneity α_i) and in the time-variant component (the idiosyncratic shock ε_{it})

When we think about the time-invariant component of the error we think about all the possible unobserved factors that might influence education and income that are time-invariant. For example, in this case we could think about genetic material (as our DNA is time-invariant and it could also determine our capacity to become more educated or having an higher income). The genetic material makes each individual different and is time-invariant: it is an individual heterogeneity □ it makes each individual different in a time-invariant way.

We can then think about time-variant components that can influence our model and that could be shocks on a daily basis. Skills development could be one of these components: it could change for each individual at each time point, and it could be based on things like mental health or capacity to work □ these could be idiosyncratic shocks.

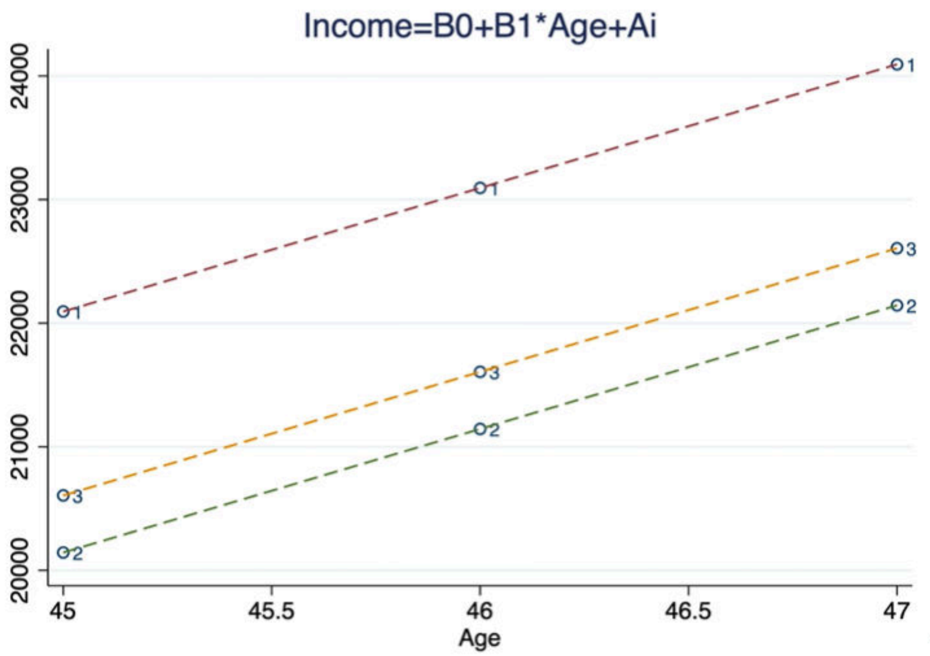
Figure n. 5[27] (Carlos Riumalló Herl, 2022)



In this case we want to see if there is any correlation between income and age.
For this example have been collected data for three people. If we look at the example without thinking about the errors, we see that income is a linear function of age. At age 45 individuals have an income of around 23000, at age 46 of 24000 and at 47 of 25000.

If we add the individual heterogeneity, so if we add a time-invariant component of the error term that make each of these individuals different, we now have a scenario where the intercept for each individual will be different and will be given by each individual heterogeneity. In this case, we have that age and income have the same linear relationship the distinction now is that the individual heterogeneity will make the intercept for each individual different this is what is called a "between difference" (how different people are between each other).

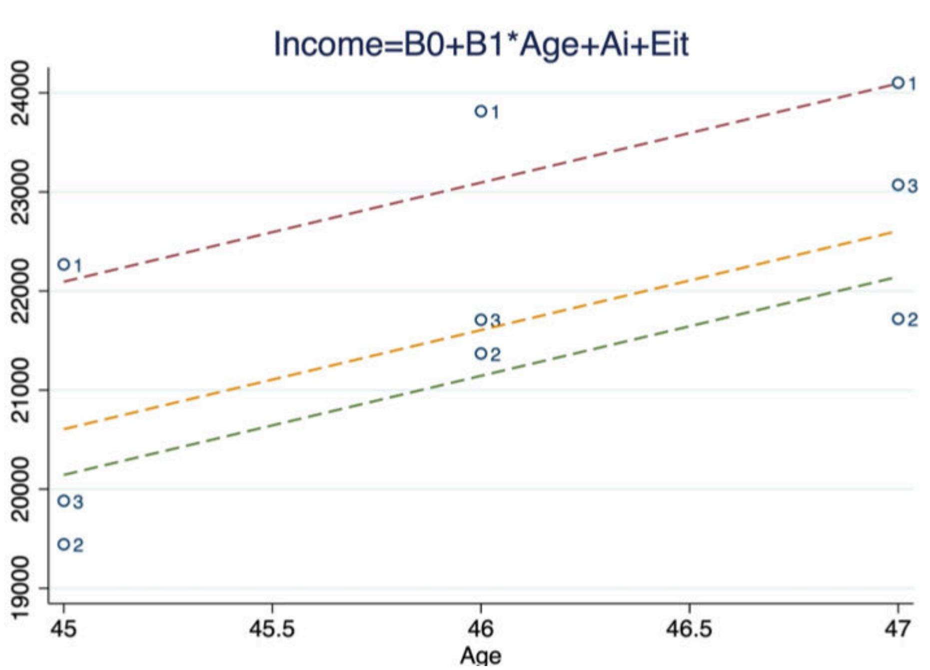
Figure n. 6[28] (Carlos Riumalló Herl, 2022)



In this setting we have that the top line is for individual 1, the second for individual 3 and the third for individual 2.

Now we add the idiosyncratic shock, an error term that varies for each individual i at time t □ it is a different value for each unique observation
In this setting we add ε_{it} in the formula:

Figure n. 7[29] (Carlos Riumalló Herl, 2022)



We can now see how the actual values get different or get spread out from the lines. Now our dots are both made up of the individual heterogeneity as well as the idiosyncratic shock.

 Whenever we are talking about the individual heterogeneity, we are referring to the time-invariant part of the error term that distinguishes units between each other and the idiosyncratic shock, that distinguishes and is unique to each point i and t and will help distinguish the within variation.

We can go further with the notation already seen:

$$Y_{it} = \beta_0 + \sum_{c=1}^{C} \beta_C X^T_{cit} + \sum_{c=C}^{C+N} \beta_C X_{cit} + \alpha_i + \varepsilon_{it}$$

When writing the model we can distinguish time-invariant and time-variant variables. In the case of the notation above, we have added two terms:
The sum from $c = 1$ to C of the X^T (which is the time-variant variable) and the sum from $c = C$ to $C + N$ of the X (the time-invariant variable).

 This will be important as some of the estimation techniques we will analyze will exploit some of these variations.

In many cases we will just write down the variables we will be using in the model without making it clear or distinguishing them as time-invariant or time-variant.
But in the case of this video this equation is meant to show us that we can distinguish in the model between the two types and decompose the error term.
We also have to remember that some characteristics can be time-invariant in our data, depending on the population and sample that we are using.

Lecture 22 (3.4) – Unbiasedness and efficiency

The decision to use one estimation technique over another is often based on whether an estimation technique is unbiased and efficient.

For this reason we are going to analyze the unbiasedness and efficiency in the contest of panel data.

An **estimator** is a method or a technique that researchers use to estimate certain parameters with a given data.
When we have to choose estimators, we have to consider two properties of interest:

- We have to see if they are unbiased □ they are if they will give us a parameter equal to the true population parameter on expectation.
- We have to see if they are efficient □ the estimators are efficient if they give us precise estimates.

Unbiasedness often relates to the value of the parameter estimated, while efficiency relates to the standard errors □ it influences how confident or not we are on rejecting the null hypothesis that the parameter is different than zero.

In the case of panel data, unbiasedness builds upon what we have already seen for cross-sectional data and refers to the lack of correlation between the error term and the variable of interest. But as we saw in the previous lecture, in the case of panel data we can decompose that error term into two parts (into the individual heterogeneity and into the idiosyncratic shock). Therefore, in the case of panel data, an estimator will be unbiased if both components of the error term are uncorrelated with the variable of interest:

- α_i must not be correlated with x_{it} □ $E[X,\ \alpha] = 0$

- ε_{it} must not be correlated with x_{it} □ $E[X, \varepsilon] = 0$

In some estimation techniques in the field of panel data some assumptions can be relaxed, but in the end a panel data estimate will be unbiased if and only if the individual heterogeneity component and the idiosyncratic shock are uncorrelated with a variable of interest (X or Y).
Failure to do so, unless we can relax those assumption in specific estimation techniques will lead to a biased estimate.

In the case of efficiency, we have to consider the structure of the error term and how the error terms are correlated over time and between each other (and over people, with other covariates).
One example that harms efficiency in OLS from cross-sectional data is heteroskedasticity □ if the error term is correlated somehow with our variable of interest, we will have heteroskedastic error terms and we will have to adjust for it, correct it, and finally have precision.
This is not different from the issues we could find in panel data. In the case of panel data, an estimation technique will be inefficient if there is a structure in the error terms that we are unable to account for correctly.
If we have inefficient estimators, this will produce incorrect standard errors and we will need to solve this problem if we do not want to reach incorrect conclusions.
In the case of panel data, when thinking about the structure of the error terms, we also need to account for the fact that we have two dimensions □ not only we have to consider whether an error term is uncorrelated with the error term of another individual, but we also have to understand and account if that error term is correlated within units.

The question that should come to mind when thinking about data structure, is whether the error terms are correlated or not across time. We will see some estimation techniques that assume that that is the case, but often is unlikely that the error terms in panel data are uncorrelated with each other (so that $Corr\left(\mu_t, \mu_s\right) = 0$).
The reason for this is simple: if we go to the construction of that error term, if we look at how the error term is constructed at time t and at time $t - 1$, as shown in the following equation, we can see that both of the error terms at time t, and at time $t - 1$, are constructed and based on the individual heterogeneity:

$$Corr\left(\mu_{it}, \mu_{it-1}\right) \neq 0 \{ \mu_{it} = \alpha_i + \varepsilon_{it} \; \mu_{it-1} = \alpha_i + \varepsilon_{it-1}$$

□ both error terms have a component within them that is the same over time. That makes it such that then the error terms are likely going to be correlated over time. This is a part of the data structure, of the error structure in panel data, that we need to account for future models that we will see. This will determine whether a model might be efficient or not.

In general, in panel data analysis, efficiency will depend on how the error terms will be structure. But still, it is likely that the error terms will be correlated with each other across time: we need to take care of this.

Lecture 23 (3.5) – Pooled OLS

We've already seen that OLS is one of the best estimation techniques.
Under which circumstances can we use (pooled) OLS, and what are the challenges when using it?

Pooled OLS is the estimation of parameters using OLS on data that combines multiple observations of units in a sample. It is like using OLS but using pool data (multiple observation for the same unit).
In this scenario, we estimate the following equation:

$$Y_{it} = \beta_0 + \sum_{c=1}^{C} \beta_C X^{T}_{cit} + \sum_{c=C}^{C+N} \beta_C X_{cit} + u_{it}$$

The outcome Y for an individual i in a time t is a linear function of our parameters, the time-invariant variables and time variant-variables and an error term.
The difference between this situation and the one with cross-sectional data, is that now our data points have two dimensions □ we have data for an individual i at a time t.
In general, pooled OLS is the same estimation done for cross-sectional data □ this estimation will be unbiased under the same assumptions and will have the same properties as a normal OLS estimation for cross-sectional data.

A difference between pooled OLS and OLS estimation is that a pooled OLS exploits all data variations including the between and within.
Also the normal OLS exploits all data variations, but with cross-sectional data there is only a between variation.

The pooled OLS is the Best Linear Unbiased Estimator (BLUE) if the following two assumptions hold:

1. The zero conditional mean assumption holds (the error term must be uncorrelated with the variable of interest).
2. There has not to be serial correlation ($Corr(X) = 0\ \forall t \neq s$).

zero conditional mean assumption

If the zero conditional mean assumption holds it is possible to estimate a parameter equal on expectation to the true population parameter.
This can happen only if both components of the error term are not correlated with a variable.
 A pooled OLS will give us an unbiased estimate with panel data if we can assume that the individual heterogeneity and the idiosyncratic shock are uncorrelated with the variable of interest.
If they are not correlated, if we use an OLS, we will get an unbiased parameter for the variable of our interest.

serial correlation

Having no serial correlation (or autocorrelation) is the second condition required to make the pooled OLS the best estimation technique.
If error terms are uncorrelated with each other, then whatever we estimate as a standard error with an OLS technique will be the appropriate standard error that we can get.
This means that conditional on X, the error term in two different time periods should be uncorrelated with itself.

If these the two conditions hold, as well as other assumptions such as having full ranked data, not having multiple correlation between variables, etc., then, by using OLS, we will get the best estimates possible. Pooled OLS under these circumstances will always be the best estimation technique, even in the field of panel data.

However, in the case of panel data, we have two issues to discuss to know if pooled OLS is the best estimating choice.

endogeneity

Are we in a circumstance where the zero conditional mean assumption holds?
If we are able to assume that the error term is uncorrelated with the variable of interest, pooled OLS might be the best option and we should test to see if it is also the most efficient option
But as we saw for cross-sectional data, being able to assume the zero conditional mean assumption is very unlikely in observational studies: the treatment allocation (why people have a certain variable), is often correlated with time-variant and time-invariant unobserved characteristics that also influence the outcome.

The term "endogeneity" means that there is no zero conditional mean assumption □ if this is the case, pooled OLS estimates will be biased. So even if there is no serial correlation, whether there would be endogeneity, pooled OLS are not an estimator to use, as the estimates will be biased.
If there is endogeneity, the other estimation techniques we choose have to account for the sources of endogeneity.
As possible solutions, we will see fixed effects and first differences as well as difference in differences. Also instrumental variables can be considered a solution, since they can also be coupled with panel data.

serial correlation (continuation)

The second issue we have to face in choosing the pooled OLS as our estimation technique with panel data, is the possible presence of serial correlation.

In the case of panel data, it is very unlikely that the error terms are uncorrelated with each other over time, so that $Corr(X) = 0$
The reason for this is that each error term for an individual i at time t, is based on a common individual heterogeneity that is similar in all observations of an individual:

$$Corr\left(\mu_{it}, \mu_{it-1}\right) \neq 0 \{ \mu_{it} = \alpha_i + \varepsilon_{it} \; \mu_{it-1} = \alpha_i + \varepsilon_{it-1}$$

The consequence is that it is likely that error terms within an individual will be serially correlated. As this leads to a different structure in the error terms, if we use an OLS, the standard error that we will get will be invalid □ whatever standard error we get it will be incorrect. This has an impact on the significance and the confidence intervals obtained. It is possible that estimates may still be unbiased, if the zero conditional mean assumption holds, but the standard errors will be incorrect.

So, when the estimates are unbiased (as the zero conditional mean assumption holds) but we have serial correlation, we must opt for solutions that address the structure of the error term; the one we will see first is random effects (which is preferred when working with panel data as it addresses the structure of the error terms).
Another solution is the clustered standard errors (which we will not analyze).

Reference list

Garcia-Gomez P. (2022) Lecture 1 – Introduction, PowerPoint slides, retrieved from: https://canvas.eur.nl/courses/40137/pages/introduction-to-empirical-methods-linear-regression-models

Garcia-Gomez P. (2022) Lecture 2 – Estimation and Interpretation, PowerPoint slides, retrieved from: https://canvas.eur.nl/courses/40137/pages/introduction-to-empirical-methods-linear-regression-models

Garcia-Gomez P. (2022) Lecture 4 – Assumptions for inference, PowerPoint slides, retrieved from: https://canvas.eur.nl/courses/40137/pages/introduction-to-empirical-methods-linear-regression-models

Garcia-Gomez P. (2022) Lecture 5 – Inference – One parameter, PowerPoint slides, retrieved from: https://canvas.eur.nl/courses/40137/pages/introduction-to-empirical-methods-linear-regression-models

Garcia-Gomez P. (2022) Lecture 7 – Interpretation categorical variables, PowerPoint slides, retrieved from: https://canvas.eur.nl/courses/40137/pages/introduction-to-empirical-methods-linear-regression-models

Garcia-Gomez P. (2022) Module 2 Lecture 10 – Beyond statistical significance, PowerPoint slides, retrieved from: https://canvas.eur.nl/courses/40137/pages/introduction-to-empirical-methods-linear-regression-models

Garcia-Gomez P. (2022) Module 2 Lecture 1 – Introduction, PowerPoint slides, retrieved from: https://canvas.eur.nl/courses/40137/pages/introduction-to-empirical-methods-endogeneity-and-instrumental-variables-estimation

Börsch-Supan A. (coordinator) (2004) The Survey of Health, Ageing and Retirement in Europe

Wooldridge J.M. (2014) Introductory Econometrics – A Modern Approach

Garcia-Gomez P. (2022) Module 2 Lecture 3 - Collider Bias, Powerpoint Slides, retrieved from: https://canvas.eur.nl/courses/40137/pages/introduction-to-empirical-methods-endogeneity-and-instrumental-variables-estimation

Cunningham S. (2020) https://www.scunning.com/causalinference_norap.pdf

Garcia-Gomez P. (2022) Module 2 Lecture 8 – Potential outcomes and DAGs, PowerPoint slides, retrieved from: https://canvas.eur.nl/courses/40137/pages/introduction-to-empirical-methods-endogeneity-and-instrumental-variables-estimation

Riumalló Herl C. (2022), Module 3 Lecture 2 - What is panel data?, Powepoint slides, retrieved from https://canvas.eur.nl/courses/40137/pages/introduction-to-empirical-methods-panel-data