HANDOUT OF

POLICY EVALUATION

2022 – 2023 EDITION

Written by Federica Di Chiara

Questa dispensa è scritta da studenti senza alcuna intenzione di sostituire i materiali universitari. Essa costituisce uno strumento utile allo studio della materia ma non garantisce una preparazione altrettanto esaustiva e completa quanto il materiale consigliato dall'Università.

INTRODUCTION

Always think in causal terms
We struggle with causal questions every day, every hour of our lives, without noticing. Consider these sentences:

- Trump/Obama administration is deleterious for the economy
- The increase in racism in Italy is Northern League's fault
- Biden's minimum wage proposal could lift millions out of poverty
- A $15 minimum wage will slow down the recovery
- A Universal Basic Income is a poor tool to fight poverty
- Had I studied anthropology, I would be happier

More formally: what is the causal effect…
… of institutions (e.g., property rights' protection) on economic development?
… of different migration policies on immigrants' crimes?
… of a tax cut (e.g. the 'bonus 80 euros') on consumption?
Etc

Quantitative evaluation of the effects of institutions, laws, policies, and other types of interventions (treatments)
Help to re-frame some habits by explicating your assumptions when 'thinking causally'

EMPIRICAL APPROACH

Understanding what is the effect of X on Y
Hypothesis: Assume X will cause Y:

- X is a potential cause – **treatment**
- Y is the outcome

Knowing the hypothesis, develop the research design, which involves 3 steps:
Situation: Identify situation useful to understand the relationship between X and Y. Other scientists can run laboratory experiments; however these are typically not available in social sciences – e.g. difficult to give a minimum wage to half of a group of workers and not to another, or groups of firms – drawbacks include the fact that it is an unequal treatment
Data collection
Measurement: need to understand what to practically do with the data to get to the causal relationship between X and Y
Third step of the empirical approach is to bring about a statistical analysis
Using statistical methods to adjust for the imperfect nature of the research design – due to the absence of perfect laboratory experiments
Try to approximate data to an experimental situation

Approach developed in mid 1990s
Before that was the Traditional econometrics: take research designs and data as given, and improve on statistical methods
Heavily based on theory and parametric specification of relationships
What followed is the Credibility revolution
Last Nobel prize winners pioneers of this type of empirical methods and designs
Contrary to traditional view that tries to model on data, this frontier research is about devising clever research designs: quasi-experimental or natural experiments
Revolution available thanks to availability of high quality data
While in the past could only use surveys, then came access to administrative data potentially on the entire population of a country: **BIG DATA**

huge volumes/detail/frequency (big data)
publicly owned: social security archives, criminal records, tax filings, etc.
private: archives of banks and insurance companies, Google/Facebook/Twitter/etc.
Proper data makes it more credible than what could have been before
Better data and research designs allow to conduct better statistical analysis using only very simple statistical methods (e.g., means comparisons)

EXAMPLE: THE MARIEL BOATLIFT

Is immigration the cause of the fall in the wage of native people?
Survey data: for each know whether they are natives or immigrants and what is their wage
Compare the wage between states with high fraction of immigrants and low fraction of immigrants might be a first proposal – there are a lot of other factors determining the wage and the share of immigrants

In 1980 Castro declared that Cubans who wanted to emigrate to the US were free to leave from the port of Mariel
Between May and September 1980 more than 125,000 people left Cuba
Miami labour force increased by 7%
David Card (1989), "The impact of the Mariel boatlift on the Miami labour market" uses this natural experiment to study the effect of immigration on natives' employment and wages
How would have wages in Miami evolved without this phenomenon?
Allows to isolate the direct effect of immigration on employment and wages that would instead be biased by other factors taking into account two different states of US
The main result: limited effects
It is a natural experiment: accident of history, something that occurs and is related to the evolution of wages and employment

e.g. want to understand the effect of fertility on female labour force participation
Compare the probability of employment of women with 0 kids and women with 1+ children
Need to take into account confounding effects: difference in participation also related to other things such as initial economic and environmental conditions, motivation, preferences, discrimination
Natural experiment that allows to assess whether women participation is influenced by having a child?
Rather than comparing women with 0 kids to women with 1 kid, compare women with 1 kid to women to 2 children
Women that both wanted to have a child, but one ended up with one and one with two

Example: measuring the evolution of economic inequality across countries and over time
Research program by Thomas Piketty and several co-authors over the last 15 years
Data source: tax records
Some results: Thomas Piketty, Capital in the XXI century: analysis of causes and consequences of economic inequality

Example: social mobility in the US
Research team from Harvard and Berkeley
Aim: measuring social mobility in the United States (gap in future income opportunities of children of rich vs. poor families)
Data source: anonymized tax records of all individuals born in 1980-81 (40 million people) linked with the tax records of their parents - connect two generations of individuals
Main results: low social mobility in the United States
Lower than in those European countries for which we have comparable data
Considerable differences across different areas within the US
American Dream – can bring people to good economic standards
But comparing data with those in Denmark – type of mobility very low in the US

STATISTICAL DATA ANALYSIS

To describe 1 phenomenon, simply need to measure it and describe it
To describe 2 phenomena, X and Y

1. Descriptive analysis of the correlation between two phenomena
2. Causal analysis: effect of X on Y

DESCRIPTIVE ANALYSIS OF 2 PHENOMENA

e.g. average earnings by educational level
Higher education associated with higher average earnings

Potential explanations
Causal effect of schooling on labour earnings (important policy implication)
Alternative explanation: other factors could influence both educational attainment and earnings (e.g. family background, intelligence)
Without additional analyses, impossible to distinguish between the two potential explanations (identification issue)
Important to understand whether there is something causal about the relationship – for policy reasons: e.g. politicians have to understand how much funding to allocate of education

Association between schooling (X) and earnings (Y)
Can be a causal relationship or it can be that there are other factors, e.g. rich parents, good school, good job – situation might be entirely explained by backward characteristic

EXAMPLE: Does hospitalization improve health?

Data from the National Health Interview Survey
Question 1: "During the past 12 months, was the respondent a patient in a hospital overnight?"
Question 2: "Would you say your health in general is excellent (5), very good (4), good (3), fair (2), poor (1)?

group	observations	health status	standard error
hospitalized	7,774	3.21	0.014
non-hospitalized	90,049	3.92	0.003
difference		-0.71	0.012

⚠ CORRELATION DOES NOT IMPLY CAUSATION!

Public debate concentrated on actual correlations:
Ethnicity and test scores: is ethnicity to blame? is this making discrimination 'fair' or 'acceptable'?
Being an immigrant and being involved in criminal activities: immigrants = criminals?

RECAP STATISTICS 1

Do people trust public institutions?
e.g. government, political parties, legal system, army, the European Union, the United Nations, etc
Trust in institution highly correlated with a lot of socio economic variables: trust in institution can explain why people vote, trust in political parties can explain a lot of variation about turnout at last elections
Questions require a numerical, quantitative answer
To get the data to answer these types of questions, surveys are particularly helpful: ESS, WVS, GSS, LITS

Trust in political parties	Freq.	Percent	Cum.
No trust at all	3,206	11.55	11.55
1	1,998	7.20	18.74
2	3,265	11.76	30.50
3	3,666	13.20	43.71
4	3,453	12.44	56.14
5	4,965	17.88	74.02
6	3,285	11.83	85.86
7	2,443	8.80	94.66
8	1,109	3.99	98.65
9	224	0.81	99.46
Complete trust	151	0.54	100.00
Total	27,765	100.00	

Trust in Political Parties: people are asked whether they trust political parties or not on a scale that goes from complete distrust to complete trust
Table shows a distribution of answers
Measure of trust is a **RANDOM VARIABLE**

Random variable (Y): variable that can take on a set of different values, each with an associated probability
It is the mathematical representation of an experiement's outcome: it is random because before drawing the units form the population, we do not know the value that is going to be observed; it is variable, because before drawing the unit from the population we only know that different values can be observed
The support set is the set of all possible values of the variable that can be associated with each and every possible outcome of an experiment, modelled by means of a random variable
Any random variable can be

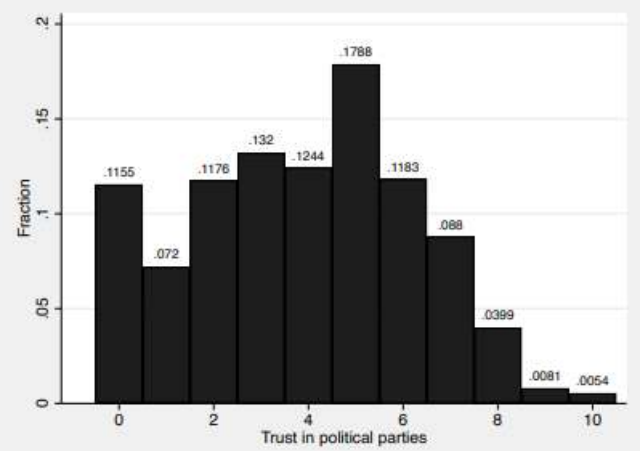
- discrete: discrete set of values 0, 1, 2… (e.g. trust) – the support set is finite or infinite but countable
- continuous: continuum of possible values (e.g. income) – the support set is infinite but not countable

Probability (p): proportion of times that a certain outcome (Y) is observed if you repeat a random process many times
Probability corresponds to the frequency in the table
Outcome (y): result of a random process, i.e. the mutually exclusive potential results of a random process
A random process is the act of asking question to surveyed people: each outcome occurs with a positive probability

DISCRETE RANDOM VARIABLES

Y is the random variable trust in political parties
It can take different outcomes from $y = 0$ (No trust) to $y = 10$ (total trust), each of which is observed p times, where p is the probability that each outcome arises (second column in STATA – tab command)
A discrete random variable takes on few values or a number of values that can be countred

It is possible to have a graphical representation of the table



Plot the probability distribution: the list of all possible values of y, y_i which are all the possible values that the random variable Y can take
The probability distribution associates to each of the possible values taken by the random variable the probability that each of them occurs (also expressed in frequency)
What is the probability that people answer total distrust in political parties?
The probability is 0.12
On the other hand, the probability that people completely trust political parties is 0.0054

$$\Pr(X = x_i) = p_i$$

All these graphs and tables are produced through STATA
Last column of the frequency distribution table delivers the value of the cumulative distribution function
for each distinct value it associates the proportion of units for which a value is smaller or equal to the distinct value itself
The cumulative distribution function allows to understand the probability that the random variable is less than or equal to a given value

$$F(Y) = \Pr(Y \leq y_k) = \sum_{i=1}^{k} p_i$$

What is the probability that the random variable is less than or equal to 3?
It is equal to 43.71% - given by the sum of all probabilities associated to values that go from 0 to 3
43% of the people that are part of the sample of the survey does not trust political parties

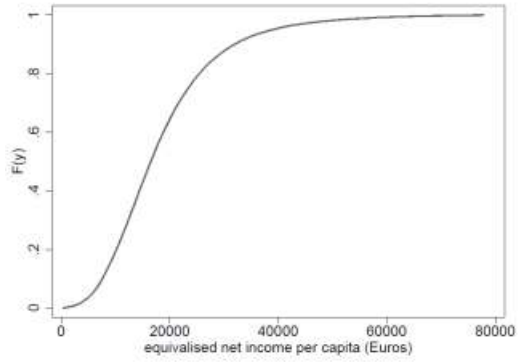
CONTINUOUS RANDOM VARIABLES

Trust in political parties is a discrete random variables
However there could also be continuous random variables, such as net income
It can take any two values withing an interval
For any two values belonging to the support set, the variable can take any value in between
The computation of the probability distribution changes, but theoretically the cumulative distribution function does not change: as before, it is the probability that the random variable is less than or equal to a particular value
What changes is the probability distribution, which is now summarised by the probability density function (p.d.f.)
Impossible to tabulate every possible value that the variable net income can take: it would be an infinitely long tabulation
Instead, can produce a density function which gives the frequency of every possible value that the variable can take: this function is the probability density function

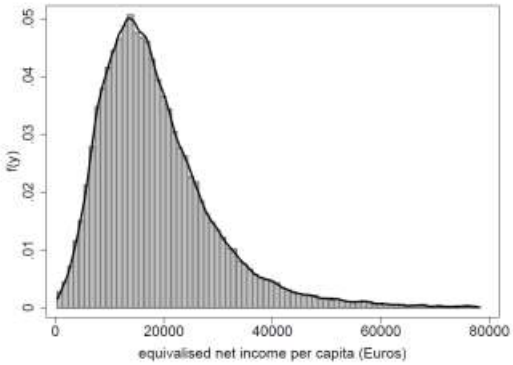
Cumulative distribution function (CDF)

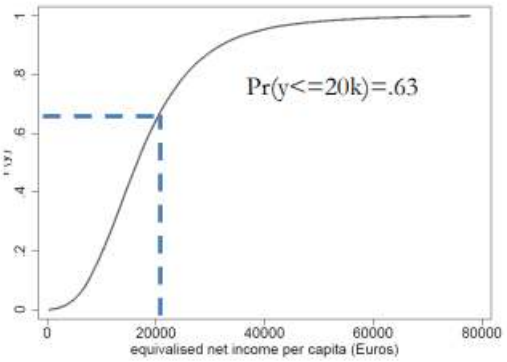
As approach the 80 000 net income per capita in Euro, the CDF tends to get value 1
Every people surveyed earned a wage that is lower or equal to 80 000
Density function plots all the possible values of the net income per capita for the people surveyed
Usually use an histogram to graphically represent a variable that is numerical and takes on several distinct values
Uses bars to portray the frequencies of the possible outcomes for a quantitative variable

Cumulative Distribution Function (CDF)



Density Function

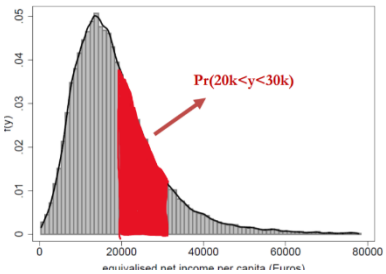




The cumulative distribution function in a continuous form is the probability that the random variable is less than or equal to a given value
In a continuum, the summatory sign are replaced by integrals

$$\Pr(y \leq b) = \int_0^b f(y)dy$$

0 here because it is the minimum value of the variable considered here: cannot have negative incomes



Probability Density Function (pdf): area under pdf between any two values is the probability that the random variable falls between those two values
Take the integral between a and b
Want to estimate the probability that the function takes any value between a and b. That is going to be equal to the area of the region delimited by the graph and the interval
Need to take the integral from a to b – gives the probability of the wage happening in the interval between the two values

$$\Pr(a \leq y \leq b) = \int_a^b f(y)dy$$

MOMENTS OF A DISTRIBUTION

Quantitative measures that are informative about the shape of a probability distribution
4 different moments or synthetic measures: a single number used for describing the feature of the variable's behaviour at the level of the sample
Measures of central tendency, measures of location, dispersion and shape of frequency distributions

First moment: expected value or mean – it is a measure of central tendency
It is the long run average value of a random variable Y: average of all possible values that a variable can take weighted by the probability that the outcome occurs
For a discrete random variable it is calculated as:

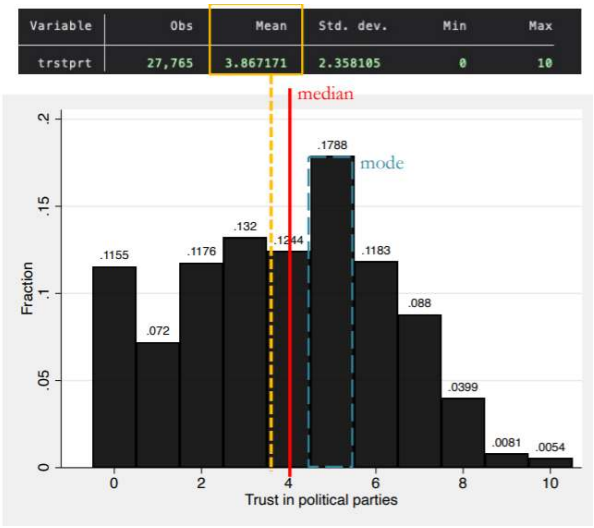
$$E(Y) = \bar{Y} = p_1 y_1 + p_2 y_2 + \cdots + p_n y_n = \sum_{i=1}^{N} p_i y_i$$

In a continuous case, the sum is replaced by the integrals

$$\bar{Y} = \int y f(y) dy$$

It gives the long run average of a random variable
In the long run, asking the question all over again, the variable will tend to a value that is equal to the mean
e.g. when flipping a coin, repeating the experiment overtime, the values will converge to 50% of cases

Mean is different from two other measures, which are more related to frequency
Median: The value y_m of the variable that splits the distribution in two equal parts
Mode: The value with the highest frequency in a distribution



Example: mode, median and mean in the variable trstprt, trut in political parties
The most frequent value is 5: mode
The median is 4: it tells that half of the respondents have a value of trust in political parties that is less than or equal to 4
Mean in this case not much indicative
The random variable can only take value 3 or 4: even with a discrete random variable, the mean can also be non-integer because it is a weighted average of discrete values

First Moment: a measure of central tendency – tells the value the variable will take on average in the long run

SECOND MOMENT: VARIANCE

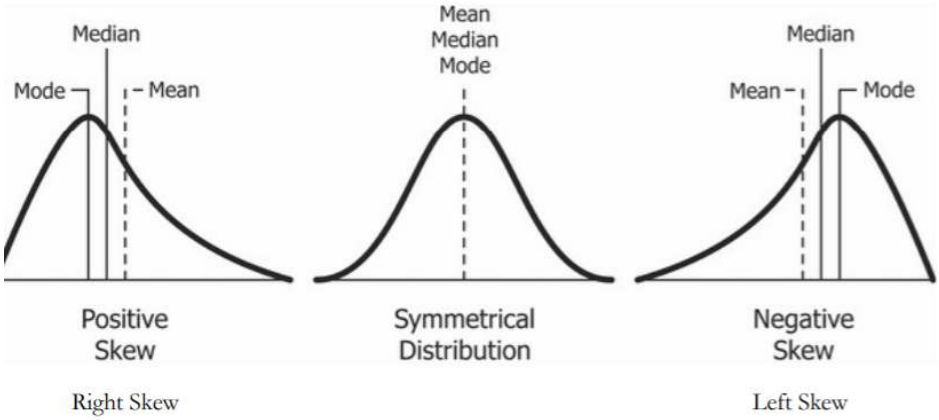
Measure of dispersion and spread of the distribution
The average of the square of the deviations from the mean – measures how spread the distribution is from the mean
Want to see how all possible outcomes are far away from the centre of the distribution
Variance is compared taking the square of the deviation from the mean, weighted by the probability that the mean itself occurs

$$Var(Y) = \sigma_Y^2 = \sum_{i=1}^{N} p_i \, (y_i - \bar{Y})^2$$
$$= p_1 \, (y_1 - \bar{Y})^2 + p_2 \, (y_2 - \bar{Y})^2 + \cdots + p_N \, (y_N - \bar{Y})^2$$

Since the variance involves the square of the deviation from the mean, numbers are difficult to interpret
Standardized version of the variance is the Standard deviation: the square root of the variance
Standard deviation and Variance provide the same information on how spread the distribution is: they are all positive numbers – cannot have a negative standard deviation

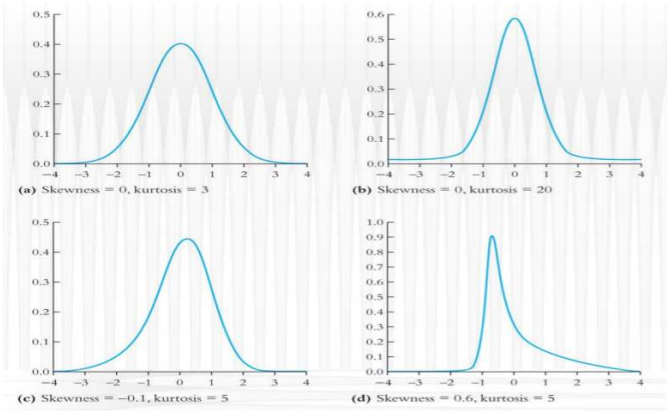
THIRD AND FOURTH MOMENTS: THE HIGHER MOMENTS

SKEWNESS: a measure of how symmetric is the distribution
skewness=0 → symmetric (normal distribution)
skewness>0 → longer right tail skewness then a perfect distribution
skewness <0 → longer left tail



In the symmetrical distribution case, mean=median=mode
Positive skewness: mode and median lower than the mean
Negative skewness: median and mode always higher than the mean – mass more skewed to the left

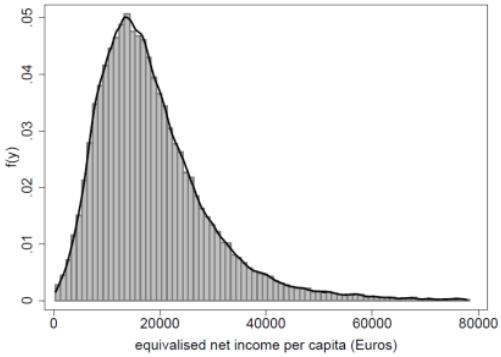
KURTOSIS: a measure that tells how important outliers are
Outliers are the values at the extreme of the distribution
Tells how many outliers there are
How much mass is in the tails of the distribution (how much variance of Y is caused by outliers)
A measure of the thickness of the tails
kurtosis=3 →Normal distribution
kurtosis>3 →More mass on tails than Normal Distribution
kurtosis <3 →Less mass on tails than Normal distribution



Case b: it is perfectly symmetrical – symmetry again equal to 0
The distribution is more concentrated in the middle – more mass in the tails with respect to the standard case
Mass more symmetric: outliers more present there than in a standard normal distribution – kurtosis larger than 3
Case c: Change in the symmetry: negative skewness – median lager than the mean
Case d: asymmetric distribution and a kurtosis that is larger than the normal case

Moments of the distribution can tell more about the shape of the distribution
Average tells where mean is
Variance tells how spread is the distribution
Skewness tells about the symmetry of the distribution
Kurtosis tells about the tails of the distribution, how important outliers are

Example: income distribution



Distribution skewed to the right: median is less than the mean – half of the population earned less than the average wage
Mode is 15 000
Kurtosis should be larger than 3
The fact that outliers are big in this situation means that we live in a very unequal world
Abstract concepts that have a very practical implementation
Concepts of the 4 moments useful when looking at the picture and just want to draw some conclusions

QUANTILES

Points of the random variable that cut the distribution in intervals with equal probabilities: equal sized, adjacent, subgroups
Median divides the population in quantiles: two subgroups that have the same probability of happening

VIQs (Very Important Quantiles):

- Median $q_{0.5}$
- Quartiles $q_{0.25}, q_{0.5}, q_{0.75}$
- Quintiles $q_{0.2}, q_{0.4}, q_{0.6}, q_{0.8}$
- Deciles $q_{0.1}, q_{0.2}, \dots, q_{0.9}$
- Percentiles $q_{0.01}, q_{0.02}, \dots, q_{0.99}$

All moments are a function of the first moment

TWO RANDOM VARIABLES: JOINT PROBABILITY DISTRIBUTION

Probability that two random variables X and Y
e.g. want to see how the random variables trust in political parties and vote in the last national elections are related
In the long term, interested in determining whether trust affected the probability of voting
The Joint Prob. Distr. of two random variables X (probability of voting in the last election) and Y (trust in political parties) is the probability that the Y and X simultaneously take on certain values
Joint probability distribution of not trusting the party and not voting in the elections (y=0, x=0)
STATA gives the cross tabulation of the two variables: frequency in number and in percent
What is the probability of observing y=0 and x-0? 12.75%

$$y_i, x_i = f(y_i, x_i) = \Pr(Y = y_i, X = x_k)$$

Conditional distribution

Distribution of one variable Y conditional on the other variable taking on a specific value
Distribution of trust in political parties among people that did not vote

. tab trust if vote==0

trust_party	Freq.	Percent	Cum.
0	3,195	58.03	58.03
1	1,826	33.16	91.19
2	485	8.81	100.00
Total	5,506	100.00	

. tab trust if vote==1

trust_party	Freq.	Percent	Cum.
0	7,998	40.92	40.92
1	8,663	44.32	85.24
2	2,884	14.76	100.00
Total	19,545	100.00	

$$f(Y|X) = \Pr(Y = y_i|X = x_k)$$

Among the respondents that did not vote, how is trust distributed? 58% of people that did not vote, did not trust political parties
Frequencies are the same: the frequency of individuals reporting that they did not vote and did not trust political parties is 3195 and that is the same when conditioning distribution on those that did not vote
The two are related: the joint probability is different from the conditional probability
Joint probability: probability of observing the two random variables
Conditional probability: probability of observing one random variable conditioning on a value of the second variable

Looking at the two conditional distribution, what can we learn?
Can we expect that the two variables move together or are they totally independent?
Significant improvement in trust in political parties among those that voted
Safe to think that these two variables might not be independent and be in some way correlated

INDEPENDENCE

Two random variables are independently distributed if knowing the value of one variable (X) provides no information about the other (Y)

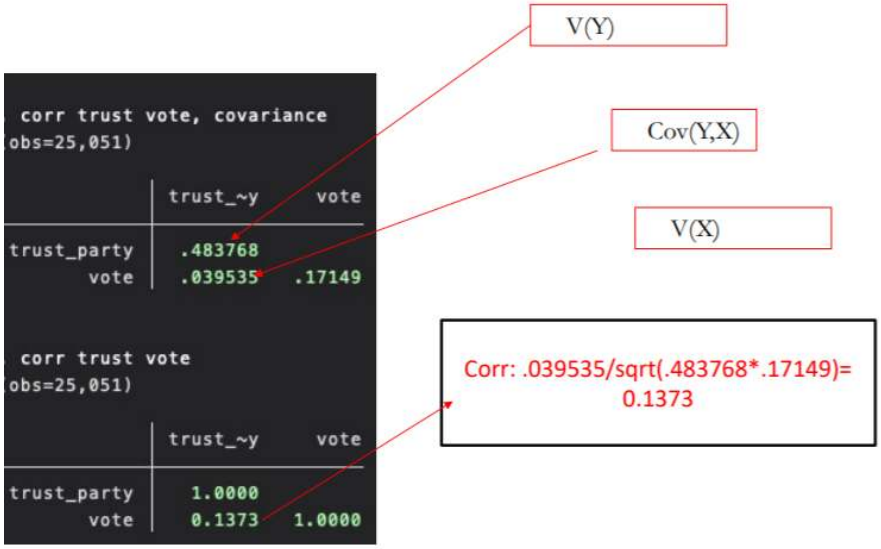
$$f(Y|X) = \Pr(Y = y_i|X = x_k) = \Pr(Y = y_i)$$

Measures of independence are the and the correlation
Covariance: measure of the extent to which the two random variables move together

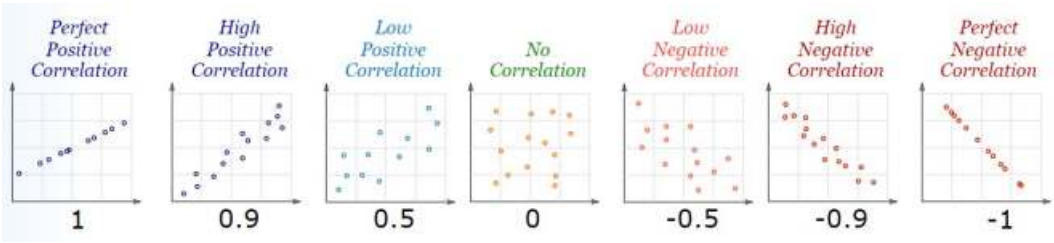
$$Cov(Y,X) = \sigma_{YX} = \sum_{i=1}^{N}\sum_{j=1}^{K}(y_i - \bar{Y})(x_j - \bar{X})\,\mathrm{f}(\mathrm{Y} = y_i, X = x_j)$$

Covariance is the sum of the deviation of the first random variable from the mean and the deviation of the second random variable from the mean
If observe positive deviations in the first random variable and positive deviations in the second random variable, it means that the covariance is positive as the two variables move together
As for the variance, the covariance measure relatively hard to interpret because it is expressed in units
Need a standardized measure
Standardized measure is the correlation: is covariance divided by the square root of the variances
Correlation is used in the data to test for independence
Correlation = 0 – the two variables are independent
Correlation = -1 – if the deviations from the mean of Y are positive, the deviations from the mean of X are negative, the two variables are negatively correlated
Correlation = 1 – deviations from the mean go in the same direction

Look at the correlation between trust and vote: based on the joint probability distribution, have an idea that these might be correlated
Produce tables on STATA
The first produces the variance and the covariance of the two variables
Variance of Y, Covariance of X
Applying the formula of the correlation, get 0.13 – STATA gives directly the correlation between trust and vote
Correlation in this case is positive: Trust and vote are not independent and they are positively correlated: if one variable goes up, the others go up as well
When trust in political parties goes up, the probability of voting also goes up



Examples of correlation between the two variables:



RECAP STATISTICS 2: Statistical inference

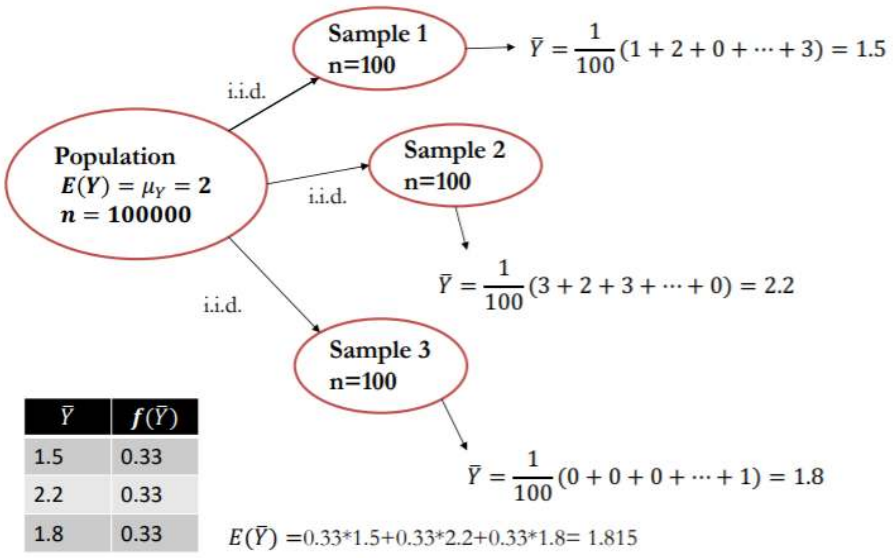
In order to answer questions, need statistics, statistical tools and data
Surveys select only a random sample of the population – surveys that includes all population is census
We cannot run a survey of a full population whenever we want to answer questions about unknown characteristics of its distribution.

Statistical Inference: we can learn about the distribution of a given quantity of interest in a population by selecting a random sample of that population
When applying statistical tools to economic problems or questions, we talk about econometrics

Econometrics is mostly about:
Estimation: computing a "Best Guess" numerical value for an unknown characteristic of a population distribution, from a sample of data. An estimator is a rule or a method for getting at an estimate of the value of a parameter/characteristic (e.g. the sample average)
Estimation is the process to get to the coefficient
Estimator is the expected value of the random variable in the population
Hypothesis testing: formulating a hypothesis about the population and use sample evidence to decide if it is "true" (with X% of probability)
Formulate a confidence set for which we reject the hypothesis made
Confidence Intervals: use the sample data to calculate a range of statistically plausible values around the best guess for the unknown population characteristic
Parameter will lie in the interval according to a predetermined probability

EXAMPLE
Want to know the mean value of Y in a population μ_Y
If we were to get the data for all the population, the true value of the mean would be μ_Y
If cannot have data for a full population, need to rely on a random sample
Draw a random sample of n independently and identically distributed (iid) observations $Y_1, Y_2, ..., Y_n$ (e.g. Flipping a coin)
Independently: events are not connected to one another: they are mutually exclusive in the population
Identically distributed: Each of them comes from the same prob. distribution: the odds are the same
Second step is to compute the sample average $\bar{Y}$
$\bar{Y}$ is the estimator of μ_Y , the expected value of the mean – it provides the best guess for μ_Y
$\bar{Y}$ is a random variable, because it is produced by the act of random sampling – can get different values for $\bar{Y}$ depending on the different process that leads to it
Repeating the random draw will get different estimates each time
The estimate is the actual number that $\bar{Y}$ spits out
Given that $\bar{Y}$ is a random variable, it is going to have its own probability distribution, known as SAMPLING DISTRIBUTION: it is the distribution of the sample averages

Properties of the estimator

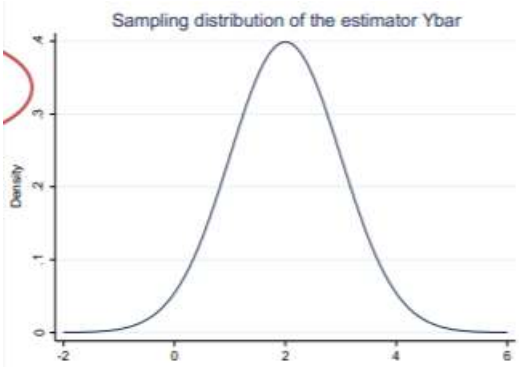


$\bar{Y}$	$f(\bar{Y})$
1.5	0.33
2.2	0.33
1.8	0.33

$E(\bar{Y})$ =0.33*1.5+0.33*2.2+0.33*1.8= 1.815

Population composed by 100 000 people – we would like to know the average hourly wage
The average from the population is equal to 2
Random sampling extracts a random sample from the 100 000 individuals
Compute the average wage for 100 people
    Get 1.5 – which is far from the true mean, equal to 2
True mean is equal to 2 – but cannot process the data about all workers, so need to do random sampling
Repeat the random process, draw a different sample – different individuals
    Mean is now 2.2
    Also pick a third sample and the average this time is 1.8
$\bar{Y}$ can take different values as it is the result of a sampling process
Associated
$E(\bar{Y})$ is the formula to calculate the discrete value of $\bar{Y}$

Can also tabulate the entire distribution of $\bar{Y}$: also have associate probabilities of observing the outcome $\bar{Y}$



Expected value of $\bar{Y}$ computed with the formula for the discrete value of the distribution
Repeating the graph n times, can get the graph of the sampling distribution of the estimator $\bar{Y}$

Need to define the characteristics of an estimator, the ones that make it the best guess
A good estimator gets as close as possible to the unknown true value
The best estimator is the one that reduces the spread of the distribution and centers it around the true value of the distribution

Unbiasedness: if you repeatedly randomly draw a sample from the same population and compute the average $\bar{Y}$, you would like that on average you get the right answer $E(\bar{Y}) = \mu_Y$ with $E(\bar{Y})$ being the average of the sampling distribution of $\bar{Y}$
The value obtained from the random sample is not different from μ_Y, the true value of the unknown mean
The difference from the value got from the random sample is not different from the value of μ
If there is bias, then the bias is given by the difference $E(\bar{Y}) - \mu_Y \neq 0$
Drawing not random samples, unbiased will not be achieved

Daily Covid cases data, along with the infection rate: million of individuals take the test and get the infection rate as a result
Average is not unbiased and will not be equal to the actual value of the population average
Self selecting into the positives: sample is not random – not independent and identically distributed observation from the population

Consistency: when sample size increases and n goes from e.g. 100s to 1000s, the uncertainty about μ_Y due to the random variation in the sample becomes very small
When n gets large enough, the value of the sample is going to converge to the true value in the population
As the sample gets bigger, the value of $E(\bar{Y})$ gets closer and closer to μ_Y

Main problem in statistics is that usually have only one sample to work with
Need to infer as much information as possible from the one sample available
Today this is changing: more data is available, researchers can have access to administrative datasets covering the whole population (e.g. Sweden, Denmark, also Italy provided access to INPS database)
Still these data is not easily found, it is expensive and it might also be difficult to run inferences on millions of observations
Solution to the problem is to use the variation in the one sample

STATISTICAL INFERENCE

Problem: we have only one sample! We have to "infer" as much information as possible from the one sample we have
Solution: we use the variation in the one sample available to approximate the sampling distribution of our estimator
Intuition: if our estimator is unbiased, the larger the sample we draw, the better we can approximate mean and variance of the sampling distribution

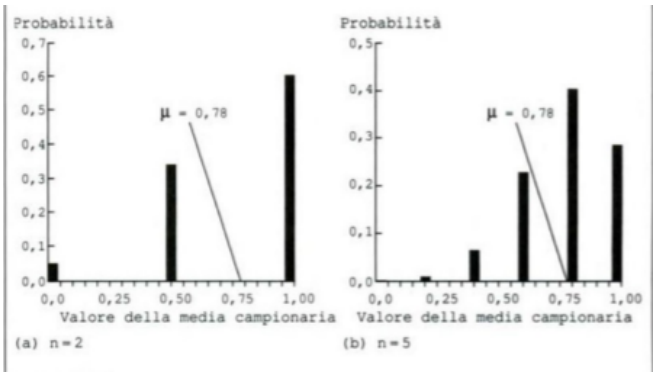
There are two tools to solve the problem and get as much information as possible from one sample
Start from a point in which we take unbiasedness for granted in the sample
Law of Large numbers: If the sample size n increases, then $\bar{Y}$ will converge to the true value of the population, and the sample variance will converge to the true value of the population

Central limit theorem: when the sample size increases, then the sampling distribution of $\bar{Y}$ can be approximated by a normal distribution that has mean equal to μ_Y and variance equal to the sample variance divided by n (the number of observation)

LAW OF LARGE NUMBERS

Take a population distributed according to a Bernoulli distribution with mean .78 (ex vote)
Particularly case of a discrete distribution in which the random variable can only take values 0 or 1
Random variable is the voting probability in the last election and the mean is 0.78
On average, at last election 78% of the population went to vote
Suppose we pick random samples with a small n, e.g. n=2

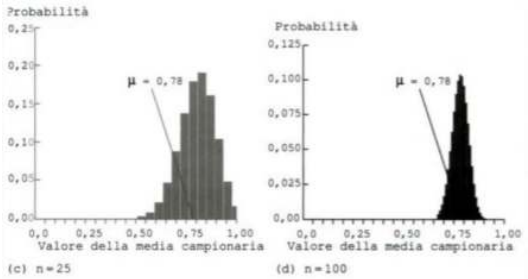
Pick just two individuals each time
Average could be 0, 1 and 0.5 (one that votes and one that doesn't)



On the graphs plot the sample averages: far away from the true value of the mean in the population

Plot the value of the estimators when drawing values from the sample

Increasing the number of people in the sample what happens is that μ is very close to the actual value of the population
When n of one sample increases, then Y will be near to μ with very high probability
Increasing to n=25, the range of values is smaller – slowly converging to the true parameter
When n=100, very tight distribution that is centred around the true parameter 0.78
Formal intuition, how the law of large numbers works



CENTRAL LIMIT THEOREM

Every random variable can be standardized

Subtract the mean and divide by the standard deviation
Creates a random variable that is distributed with mean equal to 0 and standard deviation = 1

Hard to compare random variables that have different units of measurement
Make the variable unit free, expressing the variable only in terms of how much it deviates from the mean: making it simpler to be interpreted
Central limit theorem: When n of one sample increases, then the sampling distribution of $\bar{Y}$ can be approximated by a normal distribution with mean equal to μ and standard deviation equal to the sample variance divided by n
More n needed to approximate to a normal distribution
$\bar{Y}$ can be standardized by subtracting mean and dividing by standard deviation

Standardized values of $\bar{Y}$ can be approximated by a normal distribution with those features
With mean 0 and standard deviation equal to 1
This will help in hypothesis testing

Standardized value of $\bar{Y}$ is equal to

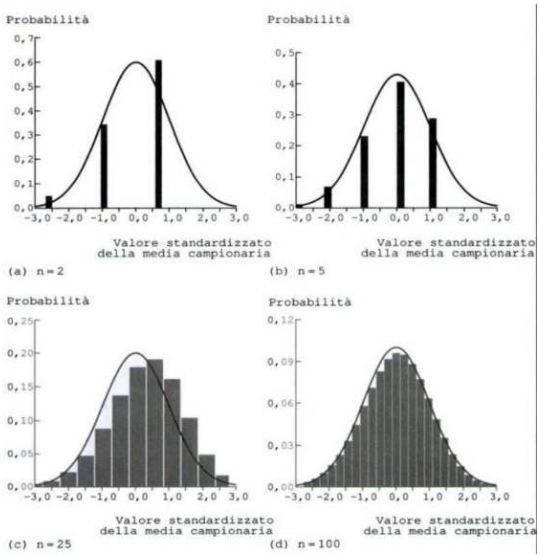
$$\frac{\bar{Y} - \mu_Y}{\frac{S_Y}{\sqrt{n}}} \sim N(0,1)$$

Standard deviation of $\bar{Y}$ is the **STANDARD ERROR**
It is an estimate of the standard deviation of the sample mean $\bar{Y}$

Bottom line: if n>30, we can always assume the "standardized value" of our estimator is distributed according to a Standardized Normal Distribution N(0,1)

Standardizing the sample mean, the same distribution as before, what changes is the x axis
Mean is equal to 0 and standard deviation is always equal to 1

One can see the convergence in the previous figure but it is more clear if we plot the standardized random variable.
It approximates very well a standard normal distribution with mean 0 and variance equal to 1
When n increases, the distribution of the standardized sample mean approaches a standard normal distribution with mean 0 and standard deviation =1
Even with n=25, get close to a standard normal distribution
With n=100, get a perfect normal distribution
Note: every rv can be standardized by subtracting the mean and dividing by its standard deviation (unit free rv!).



This theorem holds also when the original population is not distributed according to a normal (even though it takes larger sample sizes to approximate well enough)

HYPOTHESIS TESTING

Want to ask a question about the average value of trust in political parties in a particular country
Want to test what is the average wage in a population and how different it is from another country and whether this value is different from a particular value
Might want to ask a question comparing averages coming from different populations – e.g. different trust in political parties between Italy and the Netherlands?

Test the world around us with yes/no questions + statistical methods
Test an hypothesis that call the **Null hypothesis**
e.g. want to test if the average level of trust in Italy is equal to 3 (on a 0/10 scale)
Alternative hypothesis: holds when the null hypothesis is rejected
e.g. average value of trust is different from 3 – double sides hypothesis testing
Allows to check whether it is less or greater than 3 at the same time
Need to test hypothesis using a random sample of the population
Solution comes from **the t-statistic**: a (standardized) measure of the difference between observation in the sample and hypothesis; a standardized sample average used to perform hypothesis testing
Difference of the average level of trust in the population and our hypothesis that trust is equal to 3
Standardized because divided by standard error

$H_0: E(Y) = 3$
$H_1: E(Y) \neq 3$
t-statistic: take the average of Y (trust in the sample) minus the hypothesis, divided by the standard error of Y (a standardized measure of the standard deviation of Y)

$$t = \frac{\bar{Y} - 3}{SE(Y)}$$

STATA does the job for us: use the command ttest variable = H_0

Just look at Italy, less observations then the one we worked with in the previous sample (only 942 observations)
Average of trust in Italy equal to 2 – standard error is 0.07
Command produces a value of $t = -14.189$

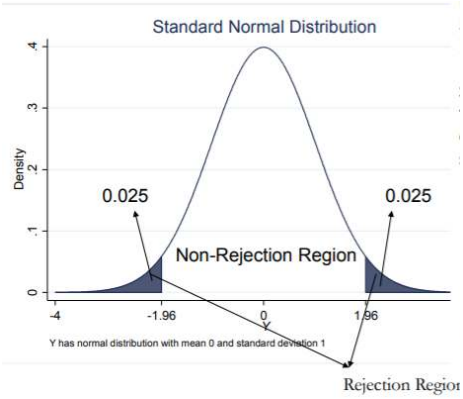
One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
trstprt	942	2.001062	.070398	2.160657	1.862906	2.139217

mean = mean(trstprt) t = -14.1899
Ho: mean = 3 degrees of freedom = 941

Need to test whether the number is informative about whether to reject or not to reject the null hypothesis
When n is large enough, the t-statistic can be distributed according to a standard normal distribution
Decide to reject the null hypothesis with a certain margin of error – assume it is 5%

May incorrectly reject H_0 5% of the time



Given this arbitrary value, known as **significance level** (usually made smaller, but not made larger), know that t is distributed according to a standard normal distribution
The 5% area of the distribution is at the tails of the standard normal distribution: so the area of rejection is the 2.5% are of the standard normal distribution on both sides, particularly for values lower than -1.96 and larger than 1.96

1.96 is the critical value: need it to test whether H_0 is true
If the t-test in absolute value is greater than 1.96, it means that the t-test statistics is in the rejection region
Therefore, we have to reject the null hypothesis
The larger the t gets, the more confident we are to reject the null hypothesis

Increasing the significance level, make it more difficult to reject the null hypothesis, because rejection area gets smaller and smaller (as significance level moves from 5 to e.g. 0.5%)
Standard distribution: density that cumulated sum up to 5% is the area at the tails - each side is 2.5%, cumulated gets to 5%
Could also be more restrictive, if needed – increase significance level of the test
5% is the significance level of the test – to make it more restrictive, increase it to 1

List of significance levels and critical values:

Significance level	Critical value
0.1	1.64
0.05	1.96
0.01	2.58
0.005	2.8

Increasing the significance level, make it harder to reject the null hypothesis
Compare the absolute value of the test statistic with the critical value
$t = -14.189$ so can reject the null hypothesis: the value of trust in political parties in Italy is not equal to 3 at a 5% significance level

A new example can be made
Is the mean level of trust equal to 2 at a 5% significance level?
$H_0: E(Y) = 2$

```
. ttest trstprt=2

One-sample t test

Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 trstprt |     942    2.001062    .070398    2.160657    1.862906    2.139217
---------+--------------------------------------------------------------------
    mean = mean(trstprt)                                     t =   0.0151
Ho: mean = 2                                 degrees of freedom =      941
```

Since the absolute value of the t-test is less than 1.96, do not reject H_0 and can conclude Political trust in Italy is equal to 2
Rule is crucial when it comes to econometrics, interpreting results of regression tables that provide information about estimates and t statistics
Whether a coefficient estimated through a regression is significant or not is determined based on the t-test

CONFIDENCE INTERVALS

Because we have a random sample, it is very unlikely that the parameter estimated is going to be exactly equal to the true parameter
Can compute the **CONFIDENCE INTERVALS**
The interval of values that with a given confidence (e.g. 95%) contains the true population mean
Can determine with a confidence interval where the true value of the parameter will fall: it will fall between the lower bound, equal to 1.86, and an upper bound that is equal to 2.13

$$C.I. = \bar{Y} \pm 1.96 * SE(\bar{Y}) = [1.863; 2.139]$$

```
. ttest trstprt=3

One-sample t test

Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 trstprt |     942    2.001062    .070398    2.160657    1.862906    2.139217
---------+--------------------------------------------------------------------
    mean = mean(trstprt)                                     t = -14.1899
Ho: mean = 3                                 degrees of freedom =      941
```

Have a random sample, estimate a number for the parameter
Since it is a random sample, there is a 95% probability that this value, the true parameter, will fall in an interval between 1.86 and 2.13
Need to make assumption about the distribution of the parameters and the likelihood that the parameter is in the confidence interval calculated
To calculate the confidence interval:
Take the mean: the lower value of the confidence level is the mean minus the critical value multiplied by the Standard error
For the upper bound it is the mean plus the critical value multiplied by the standard error

Example

Is the mean level of trust in political parties in Italy equal to the mean level of trust in the Netherlands?
Null hypothesis: the difference in the means is equal to 0
Test statistic: Average in Italy minus the average in the Netherlands minus the hypothesis (which is now 0) and divided by the standard error of the difference in the means

Alternative hypothesis: the difference in the means is not 0

Command on STATA: ttest trstprt, by (cntry)

$$H_0: \mathrm{E}(Y_{IT}) - \mathrm{E}(Y_{NL}) = 0 \text{ and } H_1: \mathrm{E}(Y_{IT}) - \mathrm{E}(Y_{NL}) \neq 0$$

$$t = \frac{\overline{Y_{IT}} - \overline{Y_{NL}}}{SE(\overline{Y_{IT}} - \overline{Y_{NL}})} = \frac{2.001 - 5.021}{0.0807} = -37.4$$

$|t|$=37.4>1.96 then we reject H_0 (Political trust in Italy is equal to political trust in NL with a 95% confidence level)

```
. ttest trstprt, by(cntry)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
IT	942	2.001062	.070398	2.160657	1.862906	2.139217
NL	1,828	5.021882	.0452062	1.932796	4.933221	5.110543
combined	2,770	3.994585	.0469273	2.469819	3.902569	4.086601
diff		-3.02082	.0807427		-3.179142	-2.862498

```
    diff = mean(IT) - mean(NL)                                    t = -37.4129
Ho: diff = 0                     degrees of freedom =     2768
```

STATA also gives the combined difference between the two values and the t-test associated with both t-statistic is 37.4, which is larger than 1.96
Can reject the null hypothesis that the difference in the level of trust in political parties is significant

RECAP STATISTICS 3: REGRESSION ANALYSIS

WHAT IS A LINEAR REGRESSION?

A linear approximation of a relationship between one dependent variable (outcome) and two or more independent variables (regression)
Regression allows to estimate parameters of linear functions
Useful method for two tasks: making inferences, particularly causal inferences that refer to a statistical method that tells how two variables move together
What happens to the dependent variable when the independent changes?
Regressions also used to make predictions

Knowing the values of the variables that enter the regression function, can predict the value of an outcome

Dependent variable is a function of something else

Dependent variable: the variable we want to explain (also called regress and or lefthand-side variable)
Independent variable: the variable used to explain the dependent variable (also called regressor or righthand-side variable)
Regression analysis is used to:

- Explain the impact of changes of an independent variable on the dependent variable
- Predict the value of a dependent variable based on the value of at least one independent variable

While a regression might look like correlation, the two are not the same
Correlation is part of the regression

Correlation

Measures, using a standardized value, the extent to which two variables are interrelated
Does not capture cause and effect – it is not able to model a function

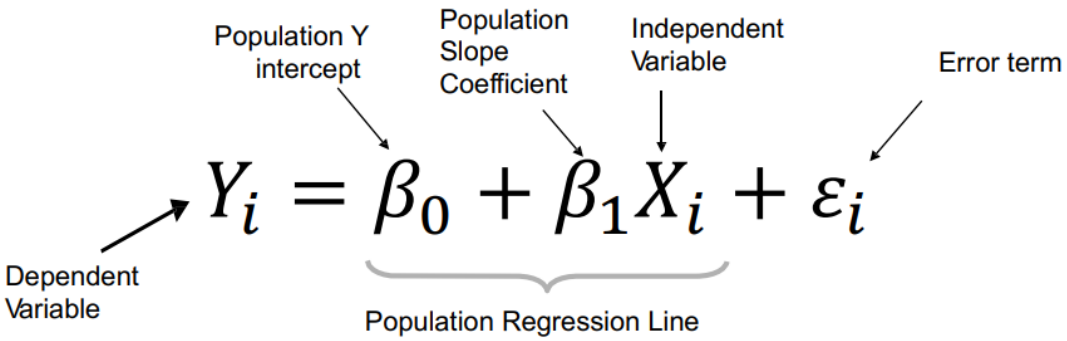
Just a measure of how two variables are linked

Same correlation coefficient if you swap x with y
Does not fit a line
Does not quantify the change in one variable associated with the change in the other: it doesn't tell what happens when there is a one-unit change in the independent variable – this is done in a regression

Regression

Relationship is evaluated through a causal model
Captures cause and effect
The regression of y on x is not the same as that of x on y
Fits a line through the data
The impact of one variable on the other is quantified.
Regression fits a line on the data
It estimates a parameter with peculiar characteristics

The relationship between X (Independent variable, e.g. education) and Y (outcome, wages) is described by a linear function
Know that somehow education and wages are correlated: but don't know what happens when change the education variable by 1 year, this the correlation doesn't tell
Changes in Y are assumed to be influenced by changes in X
Linear regression population equation model:



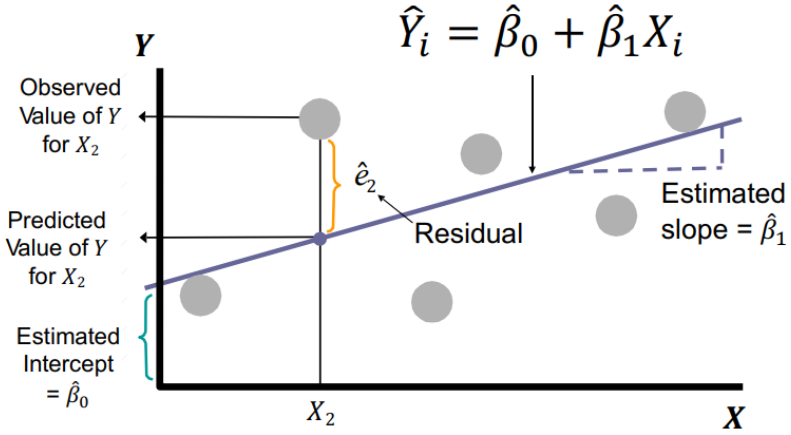
Want to model Y as a linear function of education, modelled as years of schooling
Model Y as a linear function of X means to fit a line on the data

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

Y_i= dependent variable
β_0= population Y intercept
β_1 is the slope or population coefficient of the line
X_i= independent variable
ϵ_i= error term
β_1 is the slope, so the derivative: changing X by one unit, Y will change by β_1
Error term is anything of Y that this function is not able to explain: it is the residual

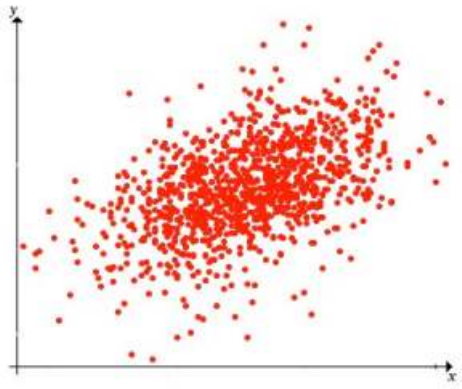
e.g. Wages are correlated and determined by the degree of education – but there are also other factors that can explain the variance of the wages observed in the: if these are not taken into account, they will end up in ϵ
Key parameters to estimate are β_0 and β_1 – knowing the two parameters, will know how the relationship moves: know the starting point and the slope of the line
Another parameter to estimate ϵ_i- how much of the relationship is not captured by the function i.e. how much relationship fails to explain all the variance of Y
Problem is that usually don't have all the population: need to rely on sample of data drawn from the population
Make assumptions: if they hold, give interpretation to the parameter of interest
When using a sample, do a sample regression: this provides an estimate of the true population regression line
Don't have all the population, just an estimate of the true values of β_0 and β_1

Imagine to have 6 observations: combination of X and Y

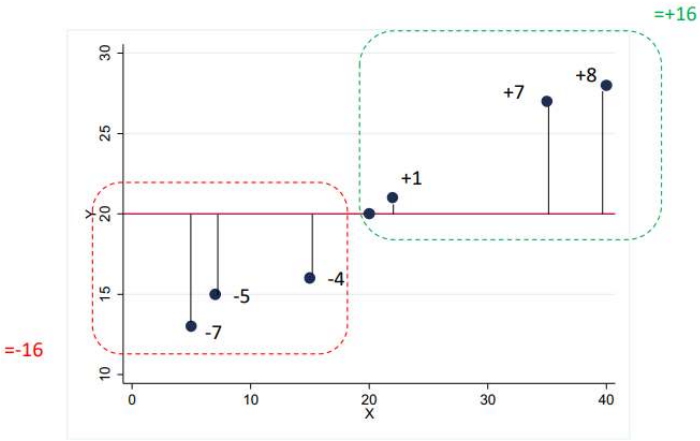


Running a regression means to find a line that best fits the data
Line has an intercept which is the distant measured by the value of Y when X is 0
This line has also a positive slope, given by β_1
The larger the β_1, the steeper the line
$\hat{e}$: estimated difference between the line and the actual value of that observation
From the graph can see that the larger the distance between the line and the points, the larger the estimated residuals

If there is distance between the observation and the line, probably the line is not the best to fit the data
Another relationship between Y and X might be:
Y: the grades in the exam
X: number of hours of study
We suspect that Bocconi students scores are related to how many hours per week they spend studying
Don't have data on all students but can extract a subsample
Data from an iid random sample (n=500)



List some of the data: for each single student correlate the hours of study and the exam score
Each observation is a combination of exam grades and hours of studying: is there a correlation?
There is a positive, not too strong correlation
No information on the hours of study: just have the distribution of exam grades
Best thing that can be done with only one variable (exam score) and no other information, the best prediction for the student's score is the mean exam score in the sample
The mean is 20
In that case, we would predict the student's score to be 20: it is a good prediction based on the information available
Mean is a constant, an intercept: a flat line
Going back to the regression equation, when X is not available, end up with a straight line
Dealing with just one variable, that might be a good prediction
However, how can we measure a good a prediction is?
There are some observations above and below the average line
One way to evaluate the "goodness of fit" of the line would be to measure the distance between the points of the observation and the line: that is given by THE SUM OF RESIDUALS
Suppose to have 8 observations°: compute for each of them the distance from the mean



Summing the deviations from the mean, however, get that their sum is equal to -16 and +16, that is 0
Make all the deviations positive and emphasize large deviations: square each of the residuals and then sum them up
Larger deviations will have more weight
7 becomes 49
1 is still 1: emphasize strong deviations

204 is the sum of squared residuals

Sum up: estimate parameter, the most simple parameter to estimate is the average
Compute deviations of each observations from the mean – square them to make them all positives
Doesn't matter whether the deviation is positive or negative
They are equally distant to the parameter
Then sum them up

Sum of squared residuals when the parameter is the mean is the total sample variance: the deviation of each single observation from the mean, squared
Formula for the variance: deviations of each single observations from the mean, squared
When the estimator is the mean, the sum of squared residuals is the total sample variation, the variance of the sample

Can do better when there is another variable explaining Y
Can apply the ordinary least squares method – OLS: minimizes the sum of squared residuals (SSR)

A regression: draws a line that fits the data
It reduces and minimizes the sum of squared residuals, i.e. it minimizes the distance of each observation from the line
The regression line resulting from the OLS estimation literally "fits" the data best by minimizing the residuals
The sample estimates of the population coefficients are those values which minimize the sum of the squared residuals

REGRESSION

Fit a line, because now also have information on X
Regression output: The sum of squared residuals is estimated and given by 4.3
204 is the total variance
Regression output also produces two important estimates of β_0 and β_1
β_0 is the variable "_cons": equal to 11

Source	SS	df	MS	Number of obs	=	7
				F(1, 5)	=	233.26
Model	199.718989	1	199.718989	Prob > F	=	0.0000
Residual	4.28101093	5	.856202186	R-squared	=	0.9790
				Adj R-squared	=	0.9748
Total	204	6	34	Root MSE	=	.92531

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
X	.4370219	.0286142	15.27	0.000	.3634667	.510577
_cons	11.00984	.6846939	16.08	0.000	9.249774	12.7699

X = 0.44 : the slope of the line
X is hours of study per week, and Y which is a number between 0 and 31
11 is the intercept of the regression line: it is the value of Y when X is 0
If study 0 hours, expect a grade of 11
X: the coefficient – look first at the sign and then the magnitude
The sign is positive: when X increases, Y should increase as well
Increasing X by 1 hour, the grade will increase by 0.43

Can give an interpretation to the intercept, but sometimes it doesn't make sense
Sometimes it doesn't make sense to have an X that is equal to 0
Suppose we want to estimate if class size has an influence on exam scores: in this case the intercept is not informative at all, it doesn't exist a case in which the class size is 0 – cannot really interpret that data
You can interpret the intercept as the predicted value of Y when X=0 – depends on the context if it makes sense to run the regression

MEASURE OF MODEL FIT

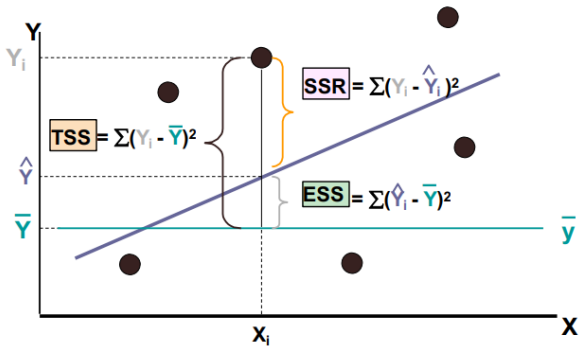
The ANOVA table or Analysis of the variance table estimates the variance observed in the data and in the outcome and it decomposes the variance in a part that is explained by the model and a part that cannot be explained (estimated by the residual)
Model sum of squares: the part of the variance that our model can explain
A part of the relation between hours of study and exam grades can be explained by the model (199), another 4.28 is due to additional variability not included in the model
Combination of the two factors gives a parameter that is informative of the goodness of fit of the model – R-squared
It is the ratio between the part of the variance that is explained by the model and the total variance

$$R^2 = \frac{ESS}{TSS}$$

Squared deviations from the fit, from the regression line
There is a way to standardize the goodness of fit – taking the ratio get rid of the unit and obtain a result that ranges from 0 to 1
If the R-squared is close to 0, the model doesn't explain anything
If the R-squared is close to 1, the model can explain the total sample variance

Graphical representation of the total sample variance: variance explained by the model and variance not explained by the model
The distance of the observation from the sample mean is the total sample variance: this is decomposed in two parts

- Sum of Squared residuals: the distance of each observation from the regression line
- Model Sum of Squares: the difference between the total sample variation and the sum of squared residuals



Can also refer to the ESS segment as the improvement that the model makes relative to the sample average
There is a distance from the observation to the average – if don't fit any regression line, that will be totally unexplained, and it will be the Sum Of Squared Residuals
Can improve the goodness of fit by fitting a regression line – reduce the distance of the observation from the new parameter, the new regression line

HYPOTHESIS TESTING

Computing the parameters is not enough – the table also produces other numbers:

- Standard error
- A measure of the t-statistic

Any standardized random variable has a distribution – this is true when the number of observation in the sample is big enough (n≥30)
β_1 as the sample average is a random variable itself when n≥30
What is done is to test whether the estimated parameter is different from 0
Hypothesis testing in a regression framework
Test the null hypothesis that $\beta_1 = 0$ against the alternative hypothesis that $\beta_1 \neq 0$
T statistic will give us the significance threshold: t is equal to the estimated β_1 minus the null hypothesis (0 here), divided by the SE

$$t = \frac{\beta_1 - 0}{SE(\beta_1)}$$

Trust in political parties: is it determined by years of education?

```
. reg  trstprt eduyr [aw=anweight]
(sum of wgt is 40,801.039823027)

      Source |       SS           df       MS      Number of obs   =    47,797
-------------+----------------------------------   F(1, 47795)     =    366.39
       Model |  2002.20128         1  2002.20128   Prob > F        =    0.0000
    Residual |  261186.635    47,795  5.46472718   R-squared       =    0.0076
-------------+----------------------------------   Adj R-squared   =    0.0076
       Total |  263188.837    47,796  5.5065034    Root MSE        =    2.3377

Trust in political parties
------------------------------------------------------------------------------
     trstprt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      eduyrs |   .0477288   .0024935    19.14   0.000     .0428415    .0526161
       _cons |    2.87179   .0342525    83.84   0.000     2.804655    2.938926
------------------------------------------------------------------------------
```

Sign. Level	Critical value
10%	1.65 (*)
5%	1.96 (**)
1%	2.58 (***)
0.5%	2.8 (***)

When increase the years of education, also increase the level of trust in the parties
Take the 47,797 observations and fit a linear regression line
Trust in political parties is regressed on years of education

Estimate two parameters: β_0, almost equal to 2.9
When people have 0 education, their level of trust in political parties is very low, about 3
Then have the second coefficient β_1 which is positive and almost equal to 0.04 in magnitude

One variable is expressed in values that go from 0 to 10 and the other is years of education
By increasing education by 1 year, increase the score that measures trust in political parties by 0.04 points

What's the sign?
What's the magnitude?
Is this coefficient statistically significant? Is it equal to 0?
Have to look at the t-test and p-value
t is equal to 19.14 – that is larger than the 1.96 critical value – so can reject the null hypothesis at 5% significance level that the value of the coefficient is equal to 0

The coefficient β_1 is different from 0

EXAMPLE

What happens to people's attitudes towards foreigners when there is a change in the visibility of immigrant communities?
Try to understand if the level of political extremism is driven by immigrants presence and in particular immigrant's visibility
Focus on Muslim communities in Germany: use a shock to Muslim visibility, given by the occurrence of the Ramadan festivity

Germany has one of the largest share of Muslim immigrants of Europe
This share has been growing for the last 50 years

At night, during the Ramadan, experience of festivals – very visible
Want to see whether the occurrence o Ramadan affects Muslim visibility
European Social Survey: has interview dates – so know whether a respondent was interviewed before or after Ramadan
Based on the interview date, could estimate the distance of the interview from the last Ramadan – measured in months
Want to see whether a change in the distance also changes the attitudes of respondents towards Muslims, immigrants and if this affects political preferences

Look at Panel 4
Fit a regression line on 2800 individuals – sample changes depending on the number of respondents to the questions
The closer to Ramadan: how does this distance affects the attitudes towards Muslims?
Dependent variable Y: have negative attitude towards Muslims
Ramadan is a measure of proximity to Ramadan measured in months
Coefficient positive: the closer to Ramadan, the more negative attitude towards migrants
Proximity to Ramadan increases the probability that a respondent has negative attitudes towards Muslims
1 month closer to Ramadan increases the probability that the respondent has negative attitudes towards Muslims by 0.04 points (4%)
Is this coefficient equal to 0? Need to do hypothesis testing
t-test is bigger than 2, so can reject the null hypothesis

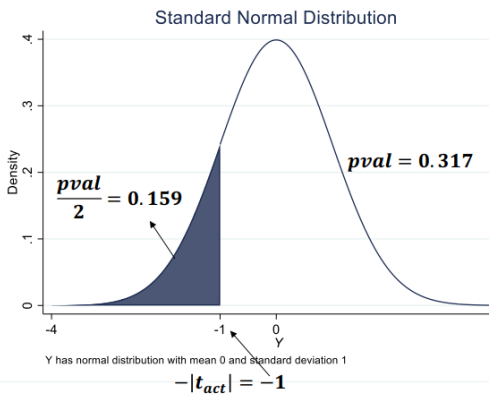
TABLE 5—RAMADAN AND INDIVIDUAL ATTITUDES

	OLS		Probit	
	(1)	(2)	(3)	Observations
Panel A. Political extremism				
Ramadan	0.0273	0.0237	0.0226	
	(0.0097)	(0.0099)	(0.0069)	2,884
Panel B. Right-wing extremism				
Ramadan	0.0118	0.0107	0.0105	
	(0.0044)	(0.0043)	(0.0033)	2,884
Panel C. Left-wing extremism				
Ramadan	0.0155	0.0130	0.0128	
	(0.0087)	(0.0086)	(0.0064)	2,884
Panel D. Anti-Muslims attitudes				
Ramadan	0.0413	0.0316	0.0427	
	(0.0175)	(0.0157)	(0.0177)	2,942
Panel E. Anti-Jewish attitudes				
Ramadan	−0.0122	−0.0133	−0.0129	
	(0.0161)	(0.0163)	(0.0164)	2,945
Panel F. Foreign-born (perceived percent)				
Ramadan	0.0856	0.0937		
	(0.0336)	(0.0345)		2,894

Coefficient
(Standard errors)

Look at Panel B:
Proximity to the last Ramadan in months: dependent variable is right-wing extremism i.e. whether the respondent says that he his right wing
Sign is positive: proximity to Ramadan seems to increase the probability that a respondent is right wing
Ratio between the coefficient and the SE is 2.68, which is greater than 1.96 and than the critical value of the test significant at 1%

Another way to do hypothesis testing is to compute the p-value
p-value is a measure derived from the t-statistics, also called as significance probability
It is the probability that by random sampling, we pick a test statistic that is as adverse to the null hypothesis as the one estimated
Probability of picking a t that in absolute value is larger than the one estimated and so that goes even more in the direction of rejecting the null hypothesis
Computed by computing the CDF of a standard normal distribution at the value actually estimated and multiply it by 2



t-statistic is estimated by a standard normal distribution
When the p-value is large we cannot reject the null hypothesis: because the probability of picking a t that is larger than the one computed

Picking a larger number by random sampling, it means we cannot be confident that the value produced is actually reliable, because picking randomly another value of t that is larger than the one estimated, the t can also be something completely due to the random sampling process
The p-value is essentially the CDF of the test-statistic at a given value (the one actually estimated)
p-value is an easy way to test the significance of the coefficient: when setting the significance level fo the test to be 5%, reject the null hypothesis only if the p-value is smaller than 0.05

Graphically:
The distribution of t has a standard normal distribution and is equal to -1
p-value tells the size of the density of the area right at -1
Multiply by 2 because we are doing a 2-sided test

p-value easy and gives the significance level of the test
STATA output: t is 19 and p-value is 0

```
. ttest trstprt, by(cntry)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      IT |   2,657    2.927362    .0450513    2.322216    2.839023    3.015701
      NL |   1,646    5.384569    .0444656    1.804009    5.297354    5.471784
---------+--------------------------------------------------------------------
combined |   4,303    3.867302    .0373422    2.449548    3.794092    3.940512
---------+--------------------------------------------------------------------
    diff |           -2.457207    .0670911                -2.58874   -2.325674
------------------------------------------------------------------------------
    diff = mean(IT) - mean(NL)                                    t = -36.6249
Ho: diff = 0                     degrees of freedom =     4301

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Significance level very close to 0 – never mistakenly reject the null hypothesis
p-value is a useful true because tells straight away what the level of significance of the test is

Regression output: several ways to do hypothesis testing
Look at the t or at the p-value: they provide the exact same information

Example of p-value applied to difference in means
Is the mean level of trust in political parties in Italy equal to the mean level of trust in the Netherlands?
Computed averages, got an estimate of the mean equal to -36
p-value is defined as the probability of picking a T in absolute values greater than the one actually computed
Reject the null hypothesis, at all significance level

p-value also allows to do one-side hypothesis testing
Can check whether we can test the null hypothesis against the probability that the difference in means is lower than 0
Given the t=-36, check the probability that t is less than -36 and compute that, taking the CDF of the t, assuming it is distributed following a standard normal distribution
p-value = 0

```
. ttest trstprt, by(cntry)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      IT |   2,657    2.927362    .0450513    2.322216    2.839023    3.015701
      NL |   1,646    5.384569    .0444656    1.804009    5.297354    5.471784
---------+--------------------------------------------------------------------
combined |   4,303    3.867302    .0373422    2.449548    3.794092    3.940512
---------+--------------------------------------------------------------------
    diff |           -2.457207    .0670911                -2.58874   -2.325674
------------------------------------------------------------------------------
    diff = mean(IT) - mean(NL)                                    t = -36.6249
Ho: diff = 0                                     degrees of freedom =     4301

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Can be different from 0, greater than 0,
Test the null hypothesis against the probability that the difference in means is lower than 0

Regression output also gives a measure of confidence interval
The range of values that with a given confidence (at 95%) contains the true coefficient

```
. reg  trstprt eduyr [aw=anweight]
(sum of wgt is 40,801.039823027)

      Source |       SS           df       MS      Number of obs   =    47,797
-------------+----------------------------------   F(1, 47795)     =    366.39
       Model |  2002.20128         1  2002.20128   Prob > F        =    0.0000
    Residual |  261186.635    47,795  5.46472718   R-squared       =    0.0076
-------------+----------------------------------   Adj R-squared   =    0.0076
       Total |  263188.837    47,796   5.5065034   Root MSE        =    2.3377

------------------------------------------------------------------------------
     trstprt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      eduyrs |   .0477288   .0024935    19.14   0.000     .0428415    .0526161
       _cons |    2.87179   .0342525    83.84   0.000     2.804655    2.938926
------------------------------------------------------------------------------
```

Estimate is equal to 0.47
STATA estimates the range of values in which the true parameter lies given the confidence of 95%
With 95% confidence we can say that the parameter lies between 0.042 and 0.052
Compute that because we can get very close to the true population parameter but we cannot be sure of having estimated it precisely – give a range of values that allows to be more precise

RECAP STATISTICS 4

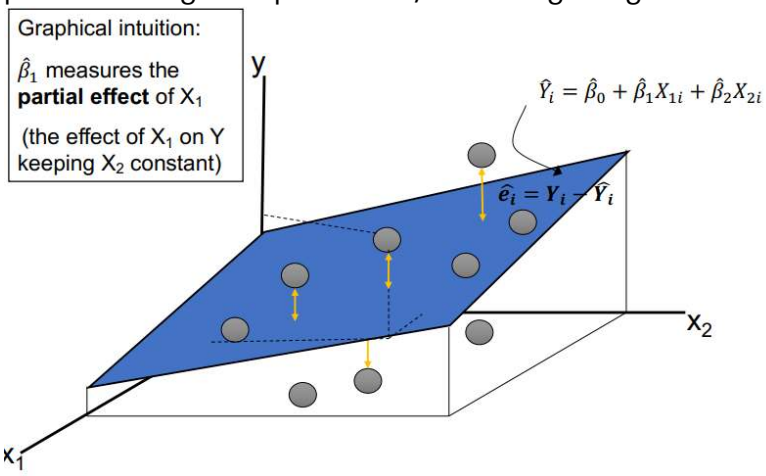
What happens when we have more information that we can exploit?
In particular, what happens if we have a variable that we think can be a good regressor of our outcome and we want to put that into the regression?
Suppose to have two variables X_1 and X_2: the latter is an additional variable that we think might be relevant to explain relationship e.g. age
Can rewrite the linear model adding the variable X_2, term that is going to be multiplied by a coefficient β_2: it gives us the effect of X on Y, what happens to Y when X changes

```
. reg  trstprt eduyr agea [aw=anweight]
(sum of wgt is 40,667.9844970075)
```

Source	SS	df	MS	Number of obs	=	47,602
				F(2, 47599)	=	284.76
Model	3097.67408	2	1548.83704	Prob > F	=	0.0000
Residual	258891.607	47,599	5.43901357	R-squared	=	0.0118
				Adj R-squared	=	0.0118
Total	261989.281	47,601	5.50386087	Root MSE	=	2.3322

trstprt	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
eduyrs	.0393914	.0025574	15.40	0.000	.0343788	.044404
agea	-.0083374	.0005801	-14.37	0.000	-.0094743	-.0072004
_cons	3.383953	.0492964	68.65	0.000	3.287332	3.480575

When looking at the regression table, still have to ask what is the sign, the magnitude and the statistical significance of each coefficient
What changes is the interpretation of β_1: it is no longer the effect of education on political parties, but it becomes the effect of education on political parties controlling for age, the new variable included in the model
Standard regression table: trust in political parties is the dependent variable
Two regressors: age and education
The first coefficient, β_1: one additional year of education increases trust in political parties by 0.04 points when age is kept constant, controlling for age



Graphically, when including an additional information, we are feeding the model with additional data
In a graph, add another dimension to the model: the line becomes a plane
To compute the goodness of fit take the distance of each observation from the new fit

The model becomes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

DUMMY VARIABLES

Binary variable that only takes value 0 or 1
Can include them in the regression as an extra regressor
Female variable: takes two values – either you are a female or you are not

Want to include this into the regression
STATA output: replacing age with the female dummy, the coefficient of education slightly changes

```
. reg  trstprt eduyr female [aw=anweight]
(sum of wgt is 40,801.039823027)
```

Source	SS	df	MS	Number of obs	=	47,797
				F(2, 47794)	=	185.58
Model	2028.12044	2	1014.06022	Prob > F	=	0.0000
Residual	261160.716	47,794	5.46429921	R-squared	=	0.0077
				Adj R-squared	=	0.0077
Total	263188.837	47,796	5.5065034	Root MSE	=	2.3376

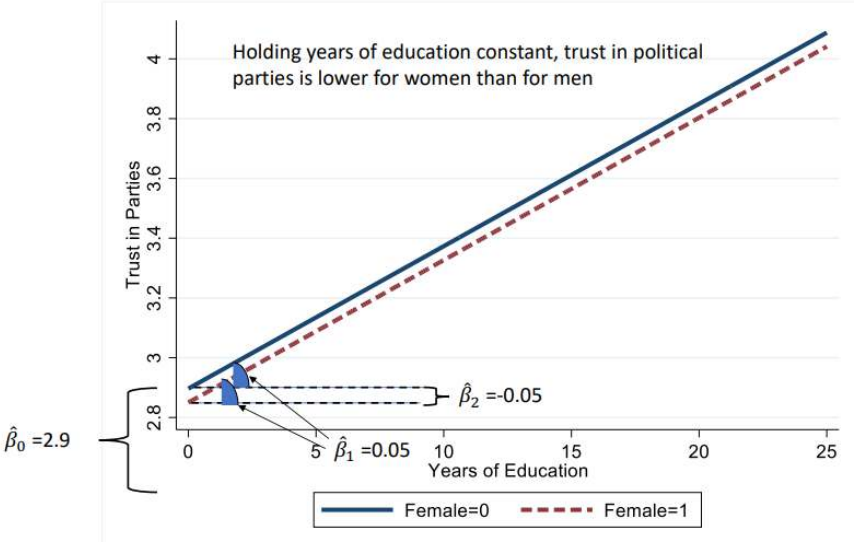
trstprt	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
eduyrs	.0476506	.0024937	19.11	0.000	.042763	.0525382
female	-.046592	.0213928	-2.18	0.029	-.0885223	-.0046618
_cons	2.896669	.0361058	80.23	0.000	2.825901	2.967437

Before β_1 was equal to 0.039, while now it is equal to 0.047
That occurs because depending on the variables included in the regression, the interpretation of β_1 changes
Coefficient for β_2 and the female variable is negative: in general, women report a lower trust in political parties
To check if the coefficient β_2 is significant, need to look at the t-test and the p-value: p-value tells us the level of significance – 0.029 which is lower than the 5% significance level, meaning that the null hypothesis that the value of the coefficient is equal to 0 can be rejected and as a consequence that the coefficient is significant

Indeed, the t-statistic has a value of 2.18 in absolute value, which is large than 1.96

Since a dummy variable can only take two values 0 and 1, we can predict the variable Y in two cases

- When $X_2 = 0$, the model reduces to $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_{1i}$ (the predicted value of Y can be estimated by looking at the predicted value of the two coefficients)
- When $X_2 = 1$, the model instead will be $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_{1i} + \widehat{\beta_2}$ (the predicted value of Y changes, because the intercept of the regression line changed)



Solid line is the first case, when $X_2 = 0$ (male group)
Intercept is equal to $\beta_0 = 2.9$, a slope that is equal to $\beta_1 = 0.48$
When looking at female respondents, the only thing that changes is the intercept – the slope doesn't change
There is a shift downwards of the intercept of the line
The coefficient estimated for the female dummy is negative
The line is going to start at $\widehat{\beta_0} + \widehat{\beta_2}$, except that $\widehat{\beta_2}$ is negative, so it shifts down
When including another regressor, what changes is the effect of education on trust of political parties, keeping constant the gender of respondents
It is possible that β_1 changes – because it is now conditional on a characteristic of the respondent group
In practice create two regression line for when the female dummy takes value 0 or 1

Direction of the shift of the regression line for the male only depends on the value of the coefficient for the female variable

Expect trust to be lower in the group of female respondents rather than the group of male respondents

This regression line (the solid one, for male group) is $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_{1i}$ – it is likely to be different from a regression line in which just include X_1 as a regressor
Difference in the two lines is the difference between male and females
If we did not control for gender however, we could have a line that is different both in terms of intercept AND slope

Not controlling for gender is likely to produce a regression output in which intercept and slope are different
But once control for gender, get the same returns of education on trust for political parties for both male and female groups
Model is just taking an average of the slope for male and female – model doesn't allow so far a difference between the two slopes

DUMMY TRAP

Suppose you want to evaluate whether the level of trust in Pol. Parties is different in these three countries, holding years of education constant
In the regression, as additional control, add the country in which the respondents live: political spectrum in Italy and the Netherlands might be different and this might lead to different results
Cannot include where people live as a continuous variable
Variable is categorical, basically a string

In order to include this variable in a regression you have to transform the variable cntry in a "set" of dummy variables (one per country)
gen france=(cntry=="FR") – variable is equal to 1 when the respondent lives in France
gen netherlands=(cntry=="NL")
gen italy=(cntry=="IT")
Then you should include the dummies in the regression, however you cannot put them all (dummy trap)
Cannot estimate all the three dummy variables simultaneously, but you need to choose one reference country and exclude its dummy from the regression.
Variables are mutually exclusive: the sum of these three variables is always equal to 1- only have people that either live in France, The Netherlands or Italy
Cannot produce coefficients for the three variables
Coefficients should be interpreted as the change in political trust between Italy and the category that is left out
e.g. choose France to be left out: this becomes the benchmark, the reference group
Include only dummies for Italy and the Netherlands

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_{IT} X_{ITi} + \beta_{NL} X_{NLi} + \epsilon_i$$

```
. reg trstprt eduyrs italy netherlands [aw=anweight]
(sum of wgt is 11,545.6068138384)

      Source |       SS           df       MS      Number of obs   =     6,142
-------------+----------------------------------   F(3, 6138)      =    291.09
       Model |  4001.20888         3  1333.73629   Prob > F        =    0.0000
    Residual |  28123.4029     6,138  4.58185124   R-squared       =    0.1246
-------------+----------------------------------   Adj R-squared   =    0.1241
       Total |  32124.6118     6,141  5.23116948   Root MSE        =    2.1405

------------------------------------------------------------------------------
     trstprt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      eduyrs |   .0563257   .0065047     8.66   0.000     .0435741    .0690773
       italy |  -.0780918     .05904    -1.32   0.186    -.1938309    .0376473
 netherlands |    2.217203   .0884183    25.08   0.000     2.043872    2.390534
       _cons |    2.340216   .0928499    25.20   0.000     2.158197    2.522234
------------------------------------------------------------------------------
```

β_{IT} is the change in trust in political parties when people live to Italy, with respect to the benchmark
So it is the difference in trust in political parties between Italy and France
β_{NL} is the difference of trust in political parties between Netherlands and France
β_1 is the effect of education on trust in political parties controlling for the country of residence

The constant now represents the average level of trust in political parties in France, holding years of education constant is 2.3 out of 10 – the country that we left out, the baseline
Coefficient for Italy and the Netherlands represent the change in trust in political parties with respect to the category left out of the regression
β_1 is positive and still significant
Italy has a level of trust, holding education constant, that is not statistically different from France
The Netherlands have a level of trust, holding education constant that is 2.2 points higher then France and is statistically significant – it measures the difference in baseline trust between the two countries

- It is a positive coefficient, it is statistically significant
- The average level of trust in Netherlands is higher than in France
- Considerable shift in the regression line

When include all the 3 dummies in the regression, STATA automatically drops one

```
. reg trstprt eduyrs italy netherlands france [aw=anweight]
(sum of wgt is 11,545.6068138384)
note: italy omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =     6,142
-------------+----------------------------------   F(3, 6138)      =    291.09
       Model |  4001.20888         3  1333.73629   Prob > F        =    0.0000
    Residual |  28123.4029     6,138  4.58185124   R-squared       =    0.1246
-------------+----------------------------------   Adj R-squared   =    0.1241
       Total |  32124.6118     6,141  5.23116948   Root MSE        =    2.1405

------------------------------------------------------------------------------
     trstprt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      eduyrs |   .0563257   .0065047     8.66   0.000     .0435741    .0690773
       italy |           0  (omitted)
 netherlands |    2.295295   .0901376    25.46   0.000     2.118593    2.471996
      france |   .0780918     .05904     1.32   0.186    -.0376473    .1938309
       _cons |    2.262124   .0852639    26.53   0.000     2.094977    2.429271
------------------------------------------------------------------------------
```

STATA automatically picks one of the variables as reference points

Italy is not excluded from the regression: it is going to be the baseline model and it is going to fall on the constant

INTERACTIONS BETWEEN INDEPENDENT VARIABLES

Allow the model to create two different slopes for the different categories that we have in our dataset (e.g. male/female; different countries)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i$$

Outcome variable is still trust in political parties, but X_{1i} is the age In years
A second variable is whether the respondent has ever been an employed for more than 3 months
β_1 is multiplying a continuous variable, β_2 is multiplying a dummy variable – the interaction between the two is captured by β_3 – tells how the slope changes when looking at different groups
The slope of the regression line is β_1: how much trust in political parties change with age
This regression might have a different intercept when we consider people that are unemployed and people that are not
Might also think that the slope of the regression line changes when we consider people that are unemployed and people that are not: do that by adding an interaction coefficient

It tells how much the slope changes when considering the unemployed category

When including the interaction term, it is also important to have the two main terms (X_{1i} and X_{2i}) separately
Regression line might change the intercept when we include a control for whether the individual is unemployed or not: when doing that just shift the line up or down
β_3 is useful to understand what happens to trust in political parties when considering the interaction between age and unemployment: it is the differential effect of trust on political parties depending on the occupational status
When one of the two interacted variables is a dummy, we can interpret estimates by distinguishing two cases
When the variable $X_{2i} = 0$, the model reduces to $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_{1i}$
When the variable $X_{2i} = 1$, the model becomes $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_{1i} + \widehat{\beta_2} + \widehat{\beta_3} X_{1i}$
$\widehat{Y_i} = (\widehat{\beta_0} + \widehat{\beta_2}) + (\widehat{\beta_1} + \widehat{\beta_3}) X_{1i}$

Tell STATA to estimate the model

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} Age_i + \widehat{\beta_2} Unemp_i + \widehat{\beta_3} Age_i * Unemp_i$$

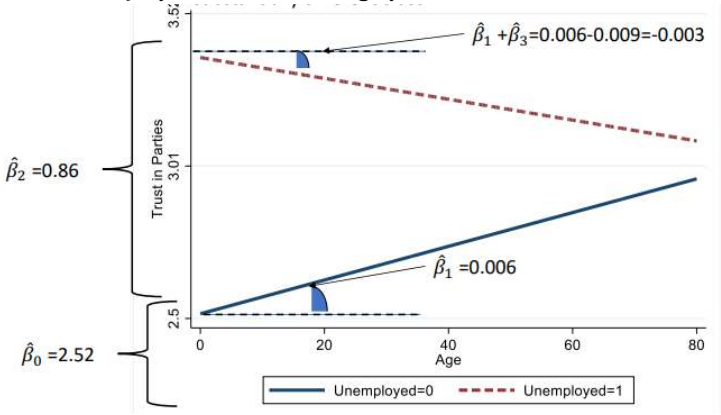
```
. reg  trstprt agea uemp3m ageaXuemp3m [aw=anweight]
(sum of wgt is 41,131.1397404819)

      Source |       SS           df       MS      Number of obs   =    47,979
-------------+----------------------------------   F(3, 47975)     =    236.60
       Model |  3853.19001         3  1284.39667   Prob > F        =    0.0000
    Residual |  260436.578    47,975  5.42858943   R-squared       =    0.0146
-------------+----------------------------------   Adj R-squared   =    0.0145
       Total |  264289.768    47,978  5.50856159   Root MSE        =    2.3299

------------------------------------------------------------------------------
     trstprt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       agea |   .0056452   .0026025     2.17   0.030     .0005444    .0107461
     uemp3m |   .8575637   .0707205    12.13   0.000     .7189507    .9961768
ageaXuemp3m |  -.0091341   .0014106    -6.48   0.000    -.0118989   -.0063693
      _cons |   2.515643   .1290521    19.49   0.000     2.262699    2.768587
------------------------------------------------------------------------------
```

When we consider $Unemp_i = 0$, the estimated line becomes $\widehat{Y_i} = 2.52 + 0.006 Age_i$
When instead $Unemp_i = 1$, there are two things that change: the intercept and the slope of the regression line
There is a shift in the intercept (constant + estimate of β_2) upwards of the regression line
If there wasn't the interaction term, the model would be just that

STATA however also produces a negative coefficient for β_3, which is statistically significant (p-value 0 and t way above 1.96)
The slope moves from 0.006 to 0.006 minus the β_3 just estimated: the slope becomes negative
The estimate of the line becomes $\widehat{Y_i} = (2.52 + 0.86) + (0.006 - 0.009)Age_i$
That is to say $\widehat{Y_i} = 3.38 - 0.03Age_i$



Graphically, the solid line represents the case in which $Unemp_i = 0$,
When $Unemp_i = 1$, there is a shift of the curve, which is the sum of $(\widehat{\beta_0} + \widehat{\beta_2})$
There is also a change in slope, which is the result of the interaction coefficient $(\widehat{\beta_1} + \widehat{\beta_3})$
Since $\widehat{\beta_3}$ is negative and larger than $\widehat{\beta_1}$, we now have a negative slope

For the employed, trust in political parties increases with age

If we just had $\widehat{\beta_2}$, there is a shift in the regression line which would mean that unemployed trust political parties more than the employed at the baseline
If unemployed at age 18, trust political parties more than if employed at 80
The change in slope means that for unemployed, when age increases, trust in political parties go down
If unemployed at 60, might have lower trust in political parties then when 18 and unemployed
The level of trust converge when age increases regardless of being employed or unemployed

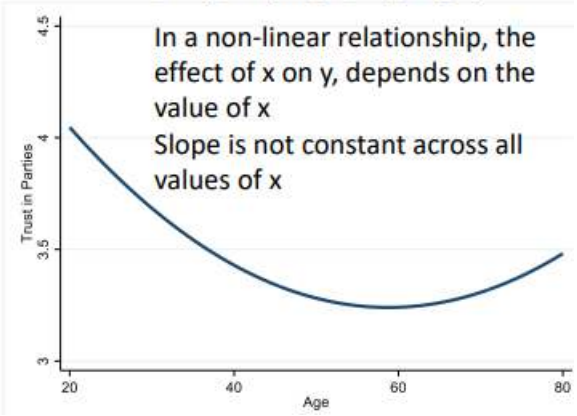
Unemployed on average trust political parties more but they have negative age gradient/slope than employed workers
Employed have instead a positive slope
Old unemployed and old employed workers have similar level of trust, but young unemployed and young employed have substantially different trust

Interaction might be important to understand the effect of a policy that might change depending on the group one belongs to

NON LINEAR FUNCTIONS OF INDEPENDENT VARIABLES

All regressors seen so far entered linearly in the model but this may not be the case
Can also tell the model to account for non-linearity, by including polynomials into the regression
e.g. we think age does not have a linear relationship with trust, but suspect it enters in a second-degree polynomial
Add to the main model Age squared
Can also include higher polynomials, like cubic

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 age_i + \hat{\beta}_2 age_i^2$$



Graphically:
Instead of modelling trust as a linear function of age, we model it as a quadratic function of age
Do that by including the square of age into the regression line
The regression line will not be linear anymore but U-shaped

STATA output when we allow age to enter quadratically into the regression line

```
. reg  trstprt agea agea2 [aw=anweight]
(sum of wgt is 41,343.2679629558)

      Source |       SS           df       MS      Number of obs   =    48,241
-------------+----------------------------------   F(2, 48238)     =    336.45
       Model |  3662.44902         2  1831.22451   Prob > F        =    0.0000
    Residual |  262550.529    48,238  5.44281538   R-squared       =    0.0138
-------------+----------------------------------   Adj R-squared   =    0.0137
       Total |  266212.978    48,240  5.51851114   Root MSE        =     2.333

------------------------------------------------------------------------------
     trstprt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       agea |   -.063003   .0029062   -21.68   0.000    -.0686991   -.0573069
      agea2 |   .0005359   .0000289    18.52   0.000     .0004792    .0005926
      _cons |   5.091959   .0666142    76.44   0.000     4.961395    5.222524
------------------------------------------------------------------------------
```

It estimates two coefficients: $\widehat{\beta_1}$ and $\widehat{\beta_2}$ which is the coefficient of Age_i^2
Relate the results to the linear case is β_1
In the case of a quadratic case, the model reduces to

$$\widehat{Y_i} = \widehat{\beta_1} + 2\widehat{\beta_2} Age_i$$

In case of a linear regression, the marginal change in Y due to a marginal change in age is given by β_1, which is nothing but the derivative of the equation

$$\frac{\partial \widehat{Y_i}}{\partial X_1} = \beta_1$$

The marginal change in Y due to a marginal change in age in the case of a quadratic model is given by the derivative of the regression equation
But now the regression equation is no longer linear, so the marginal change will be

$$\beta_1 + \frac{\partial \widehat{\beta_2} Age_i^2}{\partial Age} = \beta_1 + 2\beta_2 Age$$

β_2 coefficient just tells us the shape of the coefficient

INTERACTION BETWEEN TWO CONTINUOUS VARIABLES

Might have interaction between two continuous variables
e.g. X_1 is age and X_2 is years of education: both expressed in years and both continuous

```
. reg  trstprt agea eduyrs ageaXeduyrs [aw=anweight]
(sum of wgt is 40,667.9844970075)

      Source |       SS           df       MS      Number of obs   =    47,602
-------------+----------------------------------   F(3, 47598)     =    232.40
       Model |  3782.14046         3  1260.71349   Prob > F        =    0.0000
    Residual |  258207.141    47,598  5.42474769   R-squared       =    0.0144
-------------+----------------------------------   Adj R-squared   =    0.0144
       Total |  261989.281    47,601  5.50386087   Root MSE        =    2.3291

------------------------------------------------------------------------------
     trstprt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       agea |  -.0277039   .0018188   -15.23   0.000    -.0312688   -.0241389
     eduyrs |  -.0443527   .0078807    -5.63   0.000     -.059799   -.0289064
ageaXeduyrs |   .0015908   .0001416    11.23   0.000     .0013133    .0018684
      _cons |   4.437824   .1059536    41.88   0.000     4.230153    4.645494
------------------------------------------------------------------------------
```

The baseline is 4.43, there is a regression line with a negative slope, given by the age coefficient (-0.027)
Including education years into the regression there is a shift in the intercept: when including interaction, allow the slope to change based on the number of years of education

Regression line depicting the relationship between trust and age can change depending on the years of education
What changes now is that when we want to compute the marginal increase in Y caused by a marginal change in X, need to evaluate the effect at a specific age
Taking the derivative o f Y with respect to age, we get β_1
But now, given that there is the interaction term, we also have a $\beta_3 Education$

$$\frac{\partial \widehat{Y}_i}{\partial Age} = \beta_1 + \beta_3 Education$$

years of education	$\frac{\partial \widehat{Y}}{\partial age}$
0	-0.028
5	-0.023
10	-0.018
15	-0.013
20	-0.008

In order to know the value of $\widehat{Y}_i$, knowing the value of β_1 and β_3 provided in the regression, we also need to compute the specific value at specific years of education
Based on pre-determined years of education, it is possible to measure the change in $\widehat{Y}_i$

Interpretation: Trust in political parties deteriorates with age and becomes lower with years of education
The interpretation of the interaction: interaction answers the question "Does the level of political trust change when you become older and get more years of education?"
Is there a differential effect of age when you become more and more educated?
The coefficient of the interaction has a positive effect: it means that if you become older but also more educated, the level of trust becomes higher
Need to be able to interpret the interaction between the continuous variable: otherwise, can take education and transform it in a dummy
Interaction coefficient positive again: it now means that the relationship observed between age and trust in political parties turns out to be positive when individuals have higher education
Dummies are more intuitive and easier to use
Can always change the education or even the age variable in a dummy – any discrete variable can be transformed in a set of dummy variables
But sometimes there are continuous variables that cannot be transformed: too many values, so cannot have many sets of dummies in the regression because too hard to interpret

HOMOSCEDASTICITY VS. HETEROSCEDASTICITY

Homoscedasticity: the error term has constant variance across observations

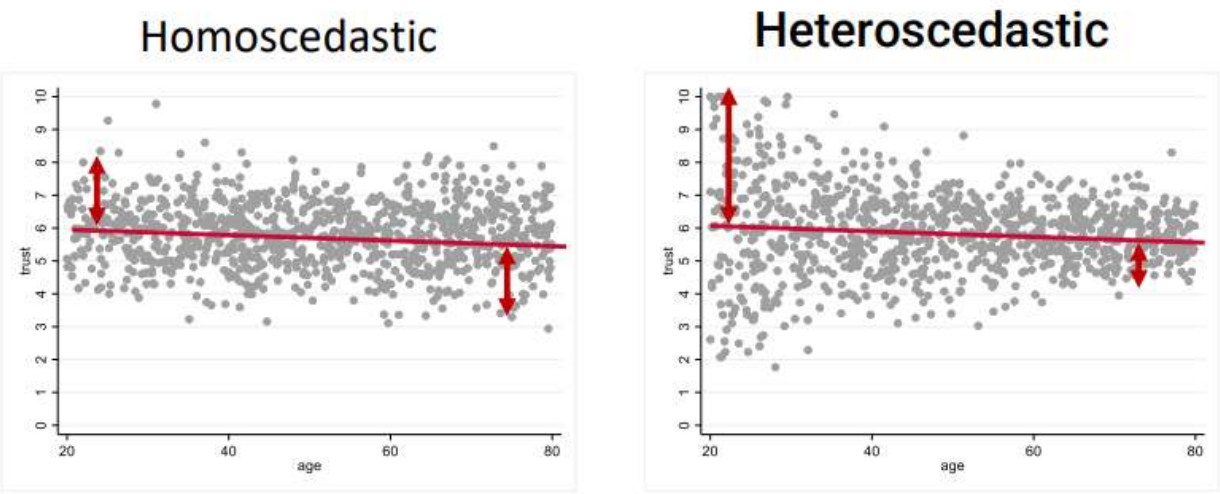
$$Var(u_i|X_i) = \sigma^2 \; \forall i$$

Variance of the standard error is the same across all possible values of a variable
Homoscedasticity implies that the model uncertainty is identical across observations and does not depend on certain values of X
Independently of the values of X, the variance of Y is of the same size

Heteroscedasticity: the variance of the error term is not constant across observations, but depends on the values of X

$$Var(u_i|X_i) = \sigma^2$$

More variance in Y depending on certain values of X e.g. more variance when people are young
Main consequence: usual standard error formula is no longer valid and inference is incorrect



There is a solution
Correct the standard errors! In STATA, add the option 'robust' to the regression command
reg y x , robust

Homoscedasticity

```
. reg trust age

      Source |       SS       df       MS              Number of obs =     1,000
-------------+------------------------------           F(1, 998)     =      4.13
       Model |  3.89424908     1  3.89424908            Prob > F      =    0.0424
    Residual |  941.492884   998  .943379643           R-squared     =    0.0041
-------------+------------------------------           Adj R-squared =    0.0031
       Total |  945.387133   999  .946333467           Root MSE      =    .97128

------------------------------------------------------------------------------
       trust |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0036896   .001816    -2.03   0.042    -.0072531   -.000126
       _cons |   6.004845   .0980322   61.25   0.000     5.812472    6.197218
------------------------------------------------------------------------------
```

```
. reg trust age, robust

Linear regression                               Number of obs     =      1,000
                                                F(1, 998)         =       3.90
                                                Prob > F          =     0.0485
                                                R-squared         =     0.0041
                                                Root MSE          =     .97128

------------------------------------------------------------------------------
             |               Robust
       trust |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0036896   .0018675    -1.98   0.048    -.0073543   -.0000248
       _cons |   6.004845   .1001739    59.94   0.000     5.808269    6.20142
------------------------------------------------------------------------------
```

Heteroscedasticity

```
. reg trust age

      Source |       SS       df       MS              Number of obs =     1,000
-------------+------------------------------           F(1, 998)     =      3.84
       Model |  5.9031014     1  5.9031014            Prob > F      =    0.0503
    Residual |  1533.46843   998  1.53654152          R-squared     =    0.0038
-------------+------------------------------           Adj R-squared =    0.0028
       Total |  1539.37153   999  1.54091245          Root MSE      =    1.2396

------------------------------------------------------------------------------
       trust |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0043688   .0022289    -1.96   0.050    -.0087427    5.10e-06
       _cons |   6.094521   .1169665    52.10   0.000     5.864992    6.324049
------------------------------------------------------------------------------
```

```
. reg trust age, robust

Linear regression                               Number of obs     =      1,000
                                                F(1, 998)         =       2.98
                                                Prob > F          =     0.0848
                                                R-squared         =     0.0038
                                                Root MSE          =     1.2396

------------------------------------------------------------------------------
             |               Robust
       trust |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0043688   .0025321    -1.73   0.085    -.0093376    .0006001
       _cons |   6.094521   .1525964    39.94   0.000     5.795074    6.393967
------------------------------------------------------------------------------
```

LINEAR PROBABILITY MODEL (LPM)

Any discrete variable can be recoded in a dummy variable
e.g. trust in political parties can be recoded in a dummy variable: values of trust of political parties lower than 3 means distrust, the rest means trust
There are several ways to treat this binary dependent variable
The most standard case is the Linear Probability Model: linear regression, applying the OLS model

- There might be a problem
- The OLS is not bounded

When we have a dummy dependent variable it means there is a probability that goes from 0 to 1

The fact that the linear probability model doesn't give any upper or lower value means that the predicted value of Y may be larger than 1 – but if the variable can only take values 0 and 1, doesn't make sense
A potential solution for that is to use two non-linear models: the PROBIT MODEL and the LOGIT MODEL
They force the predicted dependent variable to be between 0 and 1
First, they estimate the regression, getting the estimate of the key parameters: for each of these values, they compute a CDF, that must fall between 0 and 1

$$\Pr(Y = 1|age, inc, female) = CDF(\beta_0 + \beta_1 age_i + \beta_2 inc_i + \beta_3 female_i$$

Intuition: cumulative distribution function (CDF) by construction produces probabilities between 0 and 1
Probit uses CDF of standard normal distribution to model $\Pr(Y = 1|X)$
Logit uses the logistic CDF

CORRELATION IS NOT CAUSATION

Social scientists interested in understanding what is the causal effect of a given variable X (treatment) on a given outcome Y
Different ways to model the relationship between X and Y
Start with a linear regression model – assume that the relationship between Y and X can be described by means of a line

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

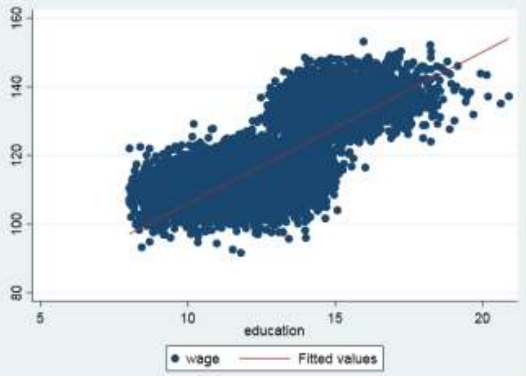
Linearity assumption: the two variables are related by means of a line
Taking the model to the data, can apply the standard OLS, minimising the sum of squared residuals and obtaining the Ordinary Least Squared estimated of coefficients α and β
β coefficient is the slope of the regression line
The slope can be rewritten as the ratio between the covariance between Y and X over the variance of X

$$\hat{\beta} = \frac{COV(X,Y)}{V(X)}$$

Formula that the minimization of squared residuals spits out
Given two variables, try to draw a line across the cloud of points describing the relationship between the two data
Try to understand what type of information the $\hat{\beta}$ coefficient is carrying
Might be tempted to give to the $\hat{\beta}$ coefficient a causal interpretation: increasing the variable X by a given quantity that will generate on average a change in Y equal to $\widehat{\Delta Y} = \hat{\beta}\Delta X$
A more careful interpretation would be: an increase in X by ΔX is associated with a change in Y equal to $\widehat{\Delta Y} = \hat{\beta}\Delta X$
β is simply describing an association between the two variables
Does the OLS-estimated coefficient $\hat{\beta}$ capture the causal effect? In most of the cases no

EXAMPLE: EDUCATION AND EARNINGS

Fictional data on daily wages in Euros and on the number of years of education these individuals accumulated
Causal effect of education on earnings is an important question for economics, social sciences and policy: schooling is one of the largest areas in which policy makers can intervene and invest
Trying to understand whether increasing access or quality of education can actually increase the earnings abilities of individuals, hence the amount they are able to contribute, is generally important
Want to understand whether an increase in education causes an increase in earnings



Start by trying to look at the correlation between the two variables
Scatterplot that shows what the association between the two variables is
The graph is also reporting a fitted line: that is describing the linear regression fit on the data when running a regression in STATA
Want to estimate the effect of additional years of education on earnings

Regression output in STATA: regress wages (Y variable) on education (X variable, plotted on the horizontal axis)

```
. reg wage education
```

Source	SS	df	MS	Number of obs	=	9,941
				F(1, 9939)	=	11768.88
Model	757831.192	1	757831.192	Prob > F	=	0.0000
Residual	639999.827	9,939	64.3927786	R-squared	=	0.5421
				Adj R-squared	=	0.5421
Total	1397831.02	9,940	140.626863	Root MSE	=	8.0245

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
education	4.425084	.04079	108.48	0.000	4.345127	4.50504
_cons	61.62703	.5377269	114.61	0.000	60.57298	62.68109

Constant value is the point at which the line crosses the vertical axis: equal to 31
β coefficient of education, i.e. the slope is equal to 4.4
The slope of the line tells what is the average increase in the daily wage associated with a one unit increase in education
Education is being measured in years: on average, 1 additional year of education is associated with an increase in the daily wage of 4.4 euros
The constant is the point at which the fitted line crosses the vertical axis – when X=0
It tells what is the average wage for an individual that has 0 years of education – not an element that has a lot of meaning in this context
In the set of data we are working with, all individuals have education that ranges from 7 to 20 years
No one in the data has 0 years of education
Make an extrapolation: take the relationship observed between daily wages and education within the range for which we have data and extrapolate that out of the sample, for individuals that have years of education that we don't actually observe in the data

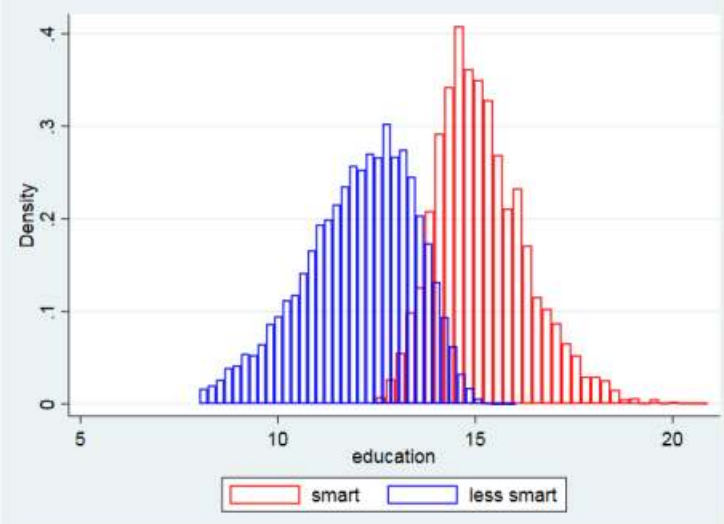
How to interpret the slope: 4.4 is the average increase in the daily wage associated with a unit increase (1 extra year) in education
Did not attach any particular causal interpretation to that coefficient
Various reasons for why need to be careful in interpreting the coefficient
When there is an association, do not imply there is no causal effect whatsoever
Just say that what we are seeing there could be or is likely going to be in part a causal effect and in part something else, just a mere association
There might be instances in which see a very strong association between two variables and absolutely no causal effect between them
The lack of knowledge ex ante for us of whether that coefficient of 4.4 is due to a causal effect or not requires to be cautious in interpreting it

Association between education and earnings
There are at least two competing ways in which it is possible to explain that result

1. Causal effect of schooling on labour earnings – imply that schooling increases earning ability of individuals and this has important policy implications
2. Omitted factors and/or selection: Other types of explanations for the positive association observed
   When look at the relationship between education and earnings do not consider that there could be other factors omitted from equation that may influence both educational attainment and increase in earnings (e.g. family background, motivation, intelligence)

OMITTED FACTOR

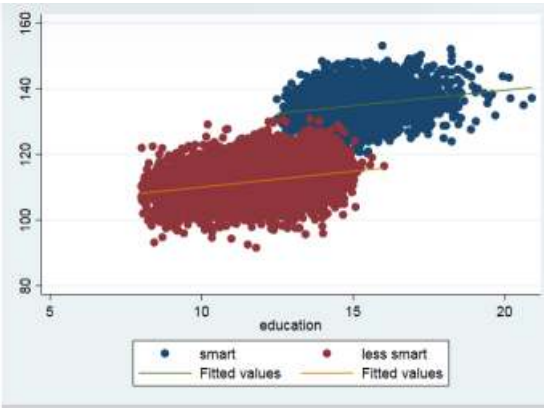
A plausible omitted factor (not considered when looking at that relation) is the ability of individuals
Consider a specific type of ability that lead people to be successful both in schooling and the labour market
When comparing people that have relatively high vs relatively low education, this on average corresponds to an earning differential
But do not consider the fact that individuals that have higher earnings might also have higher earnings ability irrespective of their education level: they are smarter and their skills are priced at a higher level in the labour market



When comparing people with high VS low education, we are also implicitly comparing people with high earnings ability and people with low earnings ability and so it could be that part of the positive association (or all of it) is explained by the fact that they are smarter rather than more educated
In the data we observe a distribution of education that is the combination of both the blue and the red distribution
This graph is the empirical distribution of years of education in the sample used
Imagine to have an indicator for whether an individual is more or less smart: this is an indicator which is difficult to obtain both because intelligence is multidimensional and because it is difficult to measure it
Splitting the sample between individuals that are smart or less smart according to the new indicator, see that there is a strong correlation between being smart and having a high education level

In the previously shown linear correlation, we are fitting a line through education in the cloud that shows the correlation between earnings and education, ignoring the fact that we actually have two groups

Taking into account the fact that there are two different groups and we modify the model, the fitted line changes
Slope of the β coefficient is much lower than before: account for the fact that part of that positive association was not at all accounted by education, rather it was explained by ability



Source	SS	df	MS	Number of obs	=	9,941
				F(2, 9938)	=	23165.68
Model	1150953.38	2	575476.691	Prob > F	=	0.0000
Residual	246877.636	9,938	24.8417827	R-squared	=	0.8234
				Adj R-squared	=	0.8233
Total	1397831.02	9,940	140.626863	Root MSE	=	4.9842

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
education	.9561081	.0374474	25.53	0.000	.8827037	1.029513
smart	19.99505	.1589462	125.80	0.000	19.68348	20.30661
_cons	100.527	.4551605	220.86	0.000	99.63483	101.4192

Assume that the relationship between education and earnings is the same irrespective of whether you are smart or not
There are TWO parallel lines plotted: they have the same slope
One extra year of education is associated with the same increase in earnings for both low and high ability individuals (slope is the same for the two groups)
Allow for a different intercept between the groups: also include in the regression the indicator for being smart
Constant term = 100: average daily wage for individuals with 0 year of education and who are not smart
No one in our sample that has 0 years of education
The coefficient on smart (dummy variable): shift the location of the slope line
The coefficient on smart = 20: being smart is associated with a wage premium of approximately 20 euros

Level effect assumed to be identical across the entire education distribution
Someone who is smart but has 0 years of education would earn 120 euros per day
Coefficient associated with education: still interpreted as the association between one extra year of education and wage – one extra year of education is now associated with approximately 1 euro increase in wage on average
Slope much flatter than before: the reason for that is that we are now accounting for the fact that we were fitting the line across two groups

Original objective is to try to understand what is the effect of education (X) on schooling (Y)
Can think of omitted factors as third factors that are not accounted for when running the SHORT REGRESSION i.e. when we think the entire story is described by the initial simple regression

Why when running the first regression we get a coefficient of 4.4 and when including the control we get a coefficient that is much lower (decreases by about ¾)?
When running the short regression we were loading onto the education variable not just the effect of education, but also on top of that the effect of being smarter
If there were not other effects apart from smart, by ignoring smart variable we were loading on education not just the true causal effect of education but also on top of that the effect of ability
Would have incorrectly concluded that having one more year of education would lead individuals to have an increase in the daily wage than what they would in reality get

Omitted variables one reason for why the β coefficient might not be accurate in describing the relationship between the dependent and the independent variable – being smart has to be correlated both with higher education and increased earnings: need a third factor that needs to be correlated with both
If there is a third factor that is only associated with X but not with Y, that is not an omitted factor
Variable can be considered an Omitted factor only if it influences both the dependent and independent variable

SELECTION
Interested in the effect of hospitalization on health
Does hospitalization improve health? (Angrist & Pischke, 2009)
Data from the National Health Interview Survey

- Question 1: "During the past 12 months, was the respondent a patient in a hospital overnight?"
- Question 2: "Would you say your health in general is excellent (5), very good (4), good (3), fair (2), poor (1)?"

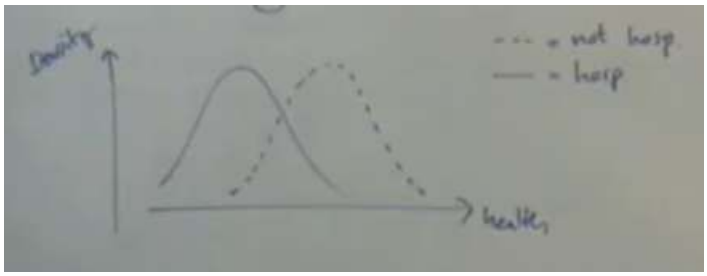
group	observations	health status	standard error
hospitalized	7,774	3.21	0.014
non-hospitalized	90,049	3.92	0.003
difference		-0.71	0.012

Might be tempted to just compare the health status of individuals that have been hospitalized and that have not been hospitalized
Hospitalization seems to reduce the quality of health of these individuals: taking the difference between the two see that those that have been hospitalized have a lower health status

Issue here is selection based on the outcome
Type of relationship here is trying to understand the effect of hospitalization on health
Regress the outcome of interest health against an indicator for whether the individual has been hospitalized or not

$$H_i = \alpha + \beta Hosp_i + \epsilon_i$$

Estimate of β would be -0.71
Do not interpret it causally because of selection



Can think about distribution of health for the people that have been hospitalized as a normal distribution (solid line), while another normal distribution describes the density of health for people not-hospitalized
Pre-treatment situation: People that end up being hospitalized have an ex-ante status of health that is lower than those that end up being hospitalized
Even if being hospitalized had a causal effect on health (represented through a rightward shift of the distribution of health for the people that have been hospitalized, i.e. when you get to the hospital, health is going to improve), we are still comparing in a regression two groups that are selected on the outcome to start with
Those that go to the hospital are those that have poorer health to begin with
Even if hospitalization was able to improve their health status, we would still not be able to capture it through a simple OLS regression
Instead, we would get a negative coefficient as in the previous example
Selection here is with respect to the outcome: people select into treatment precisely based on the outcome (the health status): this is the difference with the omitted variable problem
Omitted variable led to selection on a third element, something that was neither Y nor X

REVERSE CAUSALITY

Looking at the standard OLS regression: inclined to think that X is causing Y but it is in fact only a statistical relationship
It could as well be that Y has an effect on X
Observed relationship between Y and X reflect (in part or in total) the effect of Y on X, rather than the effect of X on Y

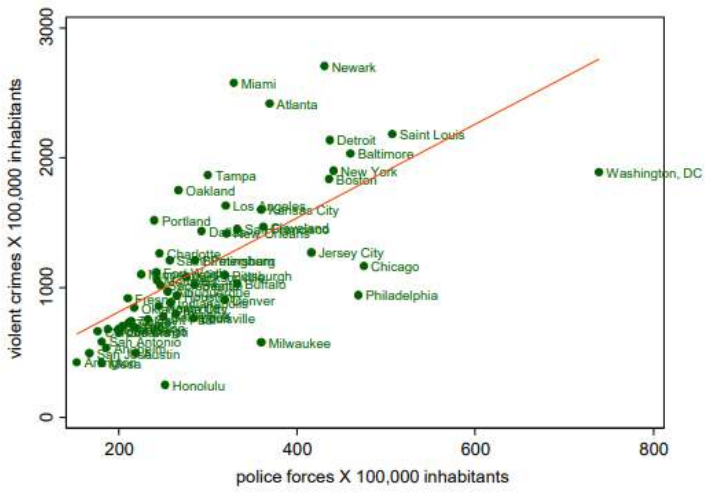
When looking at the causal relationship between two variables there are 3 problems that can arise

- Omitted variables
- Selection: third factor is ex-ante or underlying the outcome – both correlated to being hospitalized and the ex post level of health that one would have

Both are related to the idea that there is a third factor that is correlated with both treatment and outcome

- Reverse causality: Y itself might be influencing X

Example: Relationship between violent crimes and police forces
Try to understand what is the effect of increased police forces on violence and crimes
In general, would expect a negative relationship: can think of different causal channels
Positive relationship that can be mainly explained by reverse causality: where violent crimes are more prevalent, local administrations are going to increase the number of police forces that are deployed in that area



BIAS AND IDENTIFICATION PROBLEM

It is possible to formalize statistically/algebraically the issues so far analysed: helpful because it allows to think about factors that could be driving the relationship we are observing, but will also help say something about by how much we are making a mistake when running a regression
Types of derivation useful for:
Signing the bias: understand whether we are overestimating or underestimating the causal effect
Quantify the bias
Whenever we run a regression thinking that the true causal relationship of interest is captured by the short model, an OLS coefficient $\widehat{\beta_{OLS}}$ that tends to the true causal effect of interest plus an asymptotic bias

$$\widehat{\beta_{OLS}} \to \beta + asymptotic\ bias$$

The fact that the bias is asymptotic means that it doesn't vanish as we let the number of observations increase, $N \to \infty$
Situation in which we cannot solve the problem by simply resorting to bigger samples
Deeper type of problem
The 'identification problem' is concerned with understanding which part of the relationship between Y and X (i.e., coefficient $\widehat{\beta_{OLS}}$ in the OLS regression) can be attributed to the causal effect of X on Y as opposed to omitted factors, selection, and reverse causality
We are interested in identifying the causal effect of X on Y and if we get something that is different than that causal effect, we end up having an identification problem

OMITTED VARIABLE BIAS (OVB)

Interested in understanding the effect of education on earnings
Short regression:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

In fact, reality is described by a more complex model in which earnings are not just influenced by education but also by ability
Long model: the true model describing reality

$$Y_i = \alpha + \beta X_i + \gamma S_i + u_i$$

Different letter for the error term, because $\epsilon_i \neq u_i$
Inside the ϵ_i term in fact is $\gamma S_i + u_i$
This implies that we need to distinguish the coefficients in the short and the long model: in the actual results, when running the short and the long regression get different estimated values for the constant term and the slope coefficient
Therefore, rewrite the short regression as

$$Y_i = \alpha^S + \beta^S X_i + \epsilon_i$$

Estimate the $\hat{\beta}^S$ coefficient for the short model

$$\hat{\beta}^1 = \frac{Cov(Y_i, X_i)}{Var(X_i)}$$

However, now know that the Y_i is not described properly by the short model, but by the long model instead
Substitute in for Y_i using the long model

$$\hat{\beta}^S = \frac{Cov(\alpha + \beta X_i + \gamma S_i + u_i, X_i)}{Var(X_i)}$$

Covariance has a linear property: it goes through linear functions
We can split the covariance into all its linear terms: can rewrite it as the sum of the covariances between α and X_i, plus the covariance between βX_i and X_i and the covariance between all the other terms and X

$$= \frac{Cov(\alpha; X_i)}{Var(X_i)} + \frac{Cov(\beta X_i; X_i)}{Var(X_i)} + \frac{Cov(\gamma S_i; X_i)}{Var(X_i)} + \frac{Cov(u_i; X_i)}{Var(X_i)}$$

Covariance between a constant and something fixed doesn't co-vary: $\frac{Cov(\alpha;X_i)}{Var(X_i)} = 0$
Because of the linearity property, can take out the β coefficient

Therefore the second term is actually equal to $\beta \frac{Cov(X_i;X_i)}{Var(X_i)}$
Covariance of a variable with itself is the variance $\frac{Cov(X_i;X_i)}{Var(X_i)} = \frac{Var(X_i)}{Var(X_i)} = 1$
For the third term $\gamma \frac{Cov(S_i;X_i)}{Var(X_i)}$
Third assumption of the OLS: error term is independent of the Xs – covariance is assumed to be 0

$$\hat{\beta}^S = 0 + \beta \frac{Var(X_i)}{Var(X_i)} + \gamma \frac{Cov(S_i;X_i)}{Var(X_i)} + 0$$

$$\hat{\beta}^S = \beta + \gamma \frac{Cov(S_i;X_i)}{Var(X_i)}$$

β is the coefficient of education in the long regression
γ is the coefficient of ability in the long regression
$\frac{Cov(S_i;X_i)}{Var(X_i)}$ is the coefficient of a regression of ability (S) on education (X)

$$\frac{Cov(S_i;X_i)}{Var(X_i)} = \pi_1$$

Coefficient of the regression

$$S_i = \pi_0 + \pi_1 X_i + u_i$$

By running the short regression, the $\hat{\beta}^S$ estimated through the short regression is actually equal to the true β (the true causal effect) plus the product of γ (the effect of the omitted – ability, S – on the outcome) times π_1 (the correlation between the variable omitted S_i and the variable included X_i in the short regression)

$$\hat{\beta}^S = \beta + \gamma \frac{Cov(S_i;X_i)}{Var(X_i)}$$
$$\hat{\beta}^S = \beta + \gamma\pi_1$$

When ignoring the effect of the omitted factor, we are getting

- the true effect (β) +
- γ: the effect of the omitted on the outcome
- π_1: the association between the omitted (S) and the included (X)

Tells more about why we are getting something different from the actual causal effect
It also allows to understand whether we are underestimating or overestimating the true causal effect
e.g. ability not available or not well measured
When we lack the ability to measure one of the factors, by using the **omitted variable bias** formula we can understand whether by omitting that factor we are in fact overestimating or underestimating the true effect

- **Overestimate** whenever the product $\gamma\pi_1$ is positive, which happens whenever $\gamma \; and \; \pi_1$ have the same sign
- **Underestimate** whenever the product $\gamma\pi_1$ is negative, which happens when $\gamma \; and \; \pi_1$ have opposite sign

β is the true causal effect of education on earnings
When running the short regression, obtained $\hat{\beta}^S = 4.4$, while when running the long regression, $\hat{\beta}^L = 0.9$
γ is the true causal effect of ability on earnings: expect the sign of γ to be positive
π_1 is the association between education and ability: expect the sign to be positive as well
When running the short regression, we are overestimating the true causal effect of education on earnings

$$\hat{\beta}^S > \beta$$

$\hat{\beta}^S$ too large even in a world in which we cannot measure ability because we make assumptions on γ and π_1

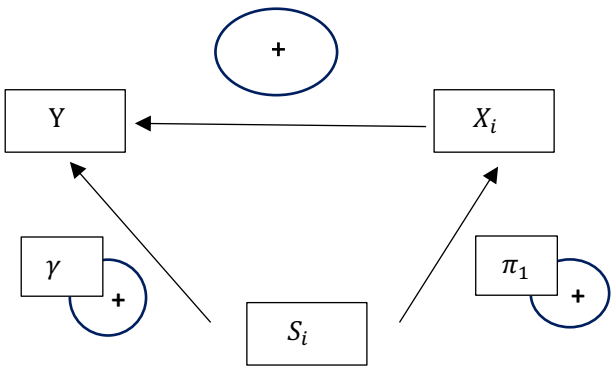
There will be bias only If both γ and π_1 are different from 0 – if one of the two is equal to 0, then there is no omitted variable bias problem
When thinking about omitted factors they qualify as such only if they are both correlated with the outcome ($\gamma \neq 0$) and they are correlated with the treatment ($\pi_1 \neq 0$)
If any of the two is 0, then there is not an omitted variable bias problem

Flip side of the coin is that whenever we want to assess what is the plausible sign of the bias we are running into when running a short regression, always need to be explicit about both γ and π_1
To understand that there is an omitted variable bias problem, also need to tell what is the sign of γ and whether it is 0 or not 0

In the example, all the coefficients are all plausibly positive – easy to make the computations
When it comes to cases in which that the main effect is negative, but it is overestimated and so it becomes negative or vice versa, need to be more careful

There is a simple visual tool to think about the omitted variable bias problem
We are interested in the effect of X on Y (education on earnings), but we are also worried about the effect of an omitted variable S_i
Expect the true relationship to be positive
For the S_i variable, it has to be associated with both Y and Xi

- The association with Y is given by γ
- The association with X_i is given by π_1



The omitted variable bias formula strictly applies only to OVB and to think about selection
Reverse causality not included: need to write a system of two equations in two variables to solve the bias
OVB is the difference between the short β and the long one
When running the short regression, we are loading into the $\hat{\beta}^S$ coefficient not just the effect of education but also the effect of ability on earnings
Need that both the two relations are there. When running the short regression load onto X also the effect that being smart has on earnings and load it into the X through the correlation between X and S
If there is no correlation, it cannot be loaded

Possible solutions of the Omitted Variable Bias problem exists

OMITTED VARIABLE BIAS (OVB)

Omitted variable bias is the difference between the $\hat{\beta}^S$ estimate of the regression that does not account for the omitted factor minus the true parameter β
The OVB tells how far we are from the true causal effect when estimating the true $\hat{\beta}$
How far we are can be inferred by the product of two elements:

- γ coefficient: the effect of the omitted variable on the outcome – what we would get if we were able to run the long regression
- $\frac{COV(S,X)}{Var(X)} = \pi_1$: the coefficient of the regression of the omitted on the included

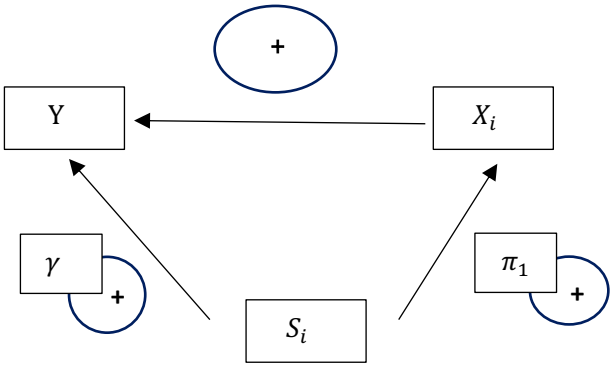
Cannot attach any causal meaning to this ratio: it is correlation between the omitted and the included (the treatment variable included in the short regression)
Even when not able to measure the omitted factor we can still learn something about the type of mistake we are committing: whether we are overestimating or underestimating the effect of interest
Need to make assumptions about the sign of γ and π_1: take the product of the sign of the two as the sign of the bias
Need both γ and π_1 to be different from 0 for any OVB to arise: an omitted variable qualifies as such only if it correlates with both the outcome and the treatment

SIGNING THE BIAS

Easiest case: β coefficient positive and the OVB positive as well

Example 1: understand the effect of education on earnings

Expect the causal effect



We expect education to have a positive effect on earnings (Y)
When regressing, we are forgetting an important factor i.e. ability (S)
Both positively correlated with earnings and education
Even absent schooling, those with higher ability are going to have higher earnings in the labour market (γ effect)
In addition, ability has a positive effect on education

$$OVB = \hat{\beta} - \beta = \gamma\pi_1 > 0$$

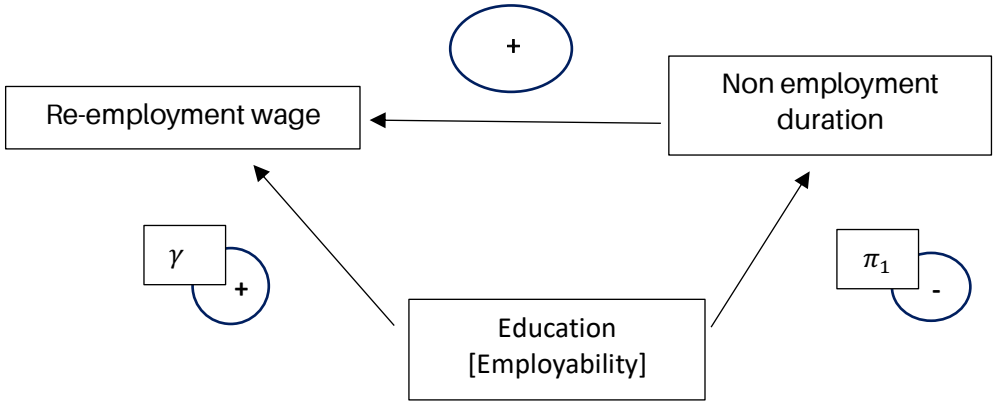
Both are positive, so the OVB is going to be positive as well
Running the short regression, we get a coefficient that is larger than the true causal effect:
$\hat{\beta}$ overestimates β
Ignoring ability in the regression, load into education variable not just the effect of education but also the fact that those to happen to have higher years of education are also the ones with higher ability
Add to education the effect of ability on earnings
Can load the effect by the extent to which ability and education are correlated (π_1)
Informally we can say that by omitting ability, education is going to compound two effect

- A direct effect of education on earnings
- An indirect effect of ability, captured by γ – how much of that γ we will be able to load onto education depends on the correlation between the two: if ability and education are not too much correlated, education will not be able to carry the effect of ability

Variation in education will be less informative about the variation in ability

Example 2: Social insurance programs

e.g. unemployment insurance : want to provide consumption smoothing to individuals that become unemployed due to the lack of a private insurance market or the lack of ability of self-ensuring themselves
When people become unemployed, typically they face a strong shock in their consumption
Crucial role of the government in providing unemployment insurance
Might also think that another reason to provide unemployment insurance is because it gives more time to find a better job to the unemployed: the better the outcome in the end
Want to understand whether it is true that the longer the unemployment period the better the employment outcome



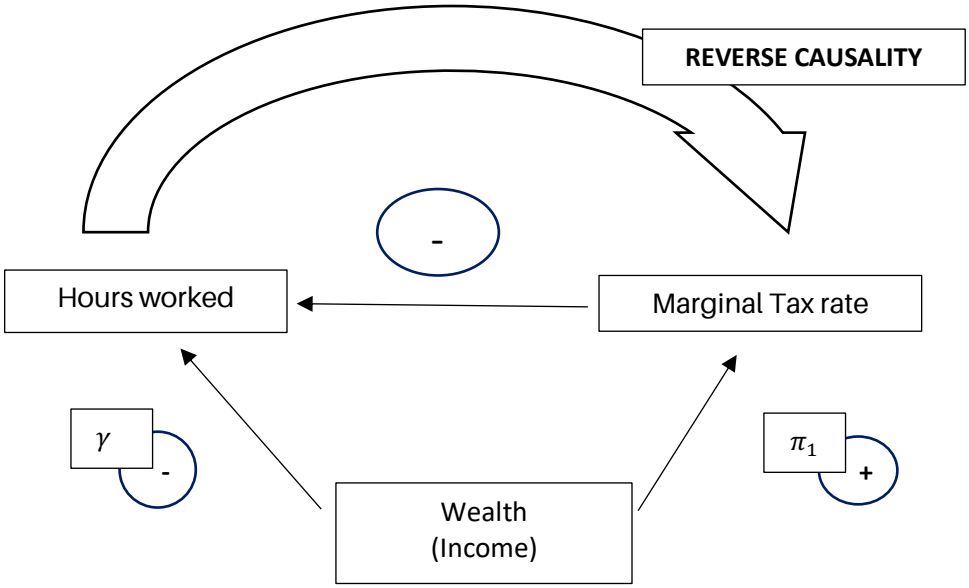
Expect the sign of the non-employment duration on the re-employment duration to be Positive
That is because the more you stay out of the market, the more the skills depreciates, the lower the wage
Already think about a confounding factor: there is selection into those that remain for a longer time unemployed
However, we are thinking only about a chain of causal effects – don't think about confounding factors
Can think that the longer we have to search, the better the future employment outcome
For simplicity, we assume that the effect between the two is positive so that the β coefficient is positive
There might be other causal channels determining a negative association between the two
When running a regression of re-employment wage on the time the individual has been out of the labour market, do not get to the true causal effect
Examples of omitted factors can be: education
A proxy for the employability of an individual
The more educated you are, the higher will be the re-employment wage regardless of the non-employment duration
Similarly, the more educated an individual is, the more we can expect him to have a lower period of non-employment

$$OVB = \hat{\beta} - \beta = \gamma\pi_1 < 0$$

When running the short regression we are underestimating the true effect of interest
The non-employment duration coefficient not only captures the effect of longer non-employment duration but also the fact that there is selection in the people that are unemployed, which are likely to be those with lower education and hence lower employability
The longer the non-employment duration so the longer will have to search and might get a better re-employment wage
In addition, will tend to be a person that has fewer employment opportunities
The true effect will be attenuated by the fact that those who stay longer unemployed are also the ones that have lower employability opportunities
Underestimation: the coefficient captured is lower than the true causal effect

Example 3: effect of higher marginal taxation on labour supply

Focus on people that are in-work: they can choose how many hours to work



Ideally there will be data for a representative number of the population that will tell what marginal tax rate the individuals face and how many hours they work over a year

Run the regression: expect a negative effect – the higher tax rate implies that the marginal return from an extra hour of work decreases, as the marginal wage decreases: less worthwhile to work an extra hour

One problem that might be present might be that we are ignoring the effect of Wealth: people that face different marginal tax rate also implicitly have different levels of income

Expect higher incomes to be associated with higher marginal tax rate: $\pi_1 > 0$

Could also think that there is an income effect: people that have higher wealth will decide to work less hours and enjoy more leisure regardless of the marginal tax rate they face: $\gamma < 0$

$$OVB = \hat{\beta} - \beta = \gamma\pi_1 < 0$$

$\hat{\beta}$ expected to be negative

But the fact that we are underestimating the effect of interest means that we are getting something that is more negative than the true causal effect

The true effect is likely to be negative: will get an underestimation

Load into the coefficient of the marginal tax rate also the fact that in progressive tax systems, the MTR will be associated with higher incomes, which through income effect will reduce the number of hours worked

Not only the substitution effect (direct effect of tax) but also the income effect

There is also a reverse causality issue here: choice of hours for a given wage is going to determine the tax rate that we face, because that determines the total income and the tax bracket in which we are going to fall

There might also be other potential omitted variables e.g. the disutility of working (how costly it is to work)

May expect that people with higher disutility of work there is going to be a lower marginal tax rate (because work less): : $\pi_1 < 0$

Also, people that have a higher disutility of work might simply want to work less hours $\gamma < 0$

There could be an other omitted variable that generates a different type of OVB, now positive

Omitting the effect of income/wealth we are underestimating the causal effect, but on top of that, omitting the disutility of work, we are overestimating the true causal effect

Possible solutions

- Including all the potential omitted factors in the OLS regression: directly control for that factors in the regression hence try to estimate the long model directly rather than the short one– solution not particularly satisfactory
- Experiments
- Quasi-experiments: try to replicate experiments in the real world, without the researcher being able to manipulate the events – use the events in the world to mimic experiments
    - Instrumental variables
    - Regression discontinuity
    - Difference-in-differences

Including additional covariates

Estimate the long model rather than the short model
We would be separately controlling for the omitted factor in the regression

$$Y_i = \alpha + \beta X_i + \gamma S_i + u_i$$

Main coefficient of interest (the one of the treatment variable) is no longer going to load the additional effect of the omitted factor
Including the S variable it would be like accounting for the direct effect of the omitted variable on the outcome: relieving X from having to carry out that extra variation
Running the long regression, Coefficient β would be only reflect the association between Y on X
Able to isolate in X only the variation that is not correlated with S
By directly including S, the coefficient β will capture just the relationship between the variation in X and Y
Variation in X is the part that is not associated with S
Variation in education net of the amount of that variation that is due to differences in ability
When running the short regression, fit the line through the cloud of data
But the variation of education is not just due to the fact that people have different levels of education but also due to the fact that people have different levels of ability
β coefficient will no longer have to capture variation in Y due to both variation in education and ability, but in the regression, the β coefficient will just capture the association between Y and education for given levels of ability

This is in principle a solution but there are two problems:

- Typically many potentially omitted factors - some of which we might not even think about
- Some omitted factors may be unobserved (e.g., ability) – will not be able to measure the S to include in the regression

Moreover when we try to include many omitted factors in the regression, we might even do worse than what we were doing by not controlling for anything
Typical case is when we include in the regression **BAD CONTROLS**
Not omitted variables – they are factors that we can call as mediating factors
Factors that pertain more to the mechanism that goes from the treatment to the outcome

Example: understand effect of education on earnings

One potential omitted factor might be the occupation of the individual
Occupation correlated with education: different years and different education choices will correspond to different roles
Occupation will also be one of the channels that determines the different earnings: different levels in the job ladder

$$Y_i = \alpha + \beta X_i + \delta W_i + u_i$$

W_i is a dummy variable that dichotomises whether an individual has a blue collar job or not
In fact, this is a mistake: get further away from the true effect of education on earnings

We are not controlling for an omitted factor: it is true that it is a variable correlated with both earnings and education, but this is also the mechanism through which education can have an effect on earnings
W_i is a mechanism → BAD CONTROL
By including the occupation in the model we are adding extra-selection in the regression
Including a covariate, keep fixed that characteristic – we were comparing people with different level of education within a certain level of ability
Here fix the level of occupation: essentially, compare individuals that have a white collar job and look within the group at individuals that have different levels of education, while trying to understand how different levels of education correlate with different levels of earnings
By doing this, we are introducing selection: look at two individuals that were both able to achieve a white collar job, but one had just a high-school degree and the other has a master's degree
Why could that be? It means that those two individuals are not similar for our purposes
People that have different occupation are likely to be very different
Load onto the β coefficient not just the effect of education, but also the fact that people that achieve different level of education conditional on having achieved the same occupation are very likely to have different levels of ability
Cannot disentangle the two effects: W is the channel that goes causally from education to earnings
By including ability

$$Y_i = \alpha + \beta X_i + \gamma S_i + u_i$$

Kept ability fixed: compare people with the same level of ability but that have different levels of education: extract from the β coefficient the effect that ability can have on earnings through its correlation with education
Keep fixed the key element that was jointly determining both wages and education

There is one case in which the use of the regression formulation can be useful to account for omitted factors: exploit the panel dimensions of the data
Can describe datasets in three categories

- Cross-sectional data: 1 observation for each unit in the sample (workers, firms, countries, etc.)
  e.g. Would have the student ID and for each individual information on the grades for policy evaluation exam, year of birth, ecc.
  Multiple unit of observations and for each of them different data
- Time-series data: 1 unit of observation for each point in time (US GDP, price of stock, etc.)
  e.g. information about one country: GDP of Italy and then we have a time dimension – have an entry of GDP for each year
  1 variable belonging to just one entity for which we have information over time
- Panel data: for each unit, repeated observations over time
  Combines both the cross-sectional and time-series data
  Have both a set of countries and follow the information for each country overtime
  e.g. ID, time dimension, earnings
  e.g. Individual 1 followed for several years, every time follow its earnings
  Repeat the process for individual 2, 3 and so on
  Can have **balanced panels** or **unbalanced panels** (different time dimension for every unit)
  Can balance the panels: reduce the dimensionality so reduce the panel data to cross sectional data
  Fixed-effects models rely on panel data
  Panel data can be useful when we want to account for specific type of omitted/confounding factors that are fixed over time – i.e. they are a fixed characteristic of the unit of observation

Panel data of i=1,2,...,N units observed over t=1,2,...,T periods
When writing down the regression of interest will have to index the observation not just by i but also by t – the time of the observation of that specific outcome
If the omitted variable is constant over time, then with panel data we can get rid of that variable through a transformation

We are interested in the effect of X on Y
Short regression is:

$$Y_{it} = \alpha + \beta X_i + \epsilon_{it}$$

But there is an omitted factor so ideally, would want to be able to run the long regression:

$$Y_{it} = \alpha + \beta X_{it} + \gamma S_i + u_{it}$$

We can safely assume that the factor doesn't vary overtime for given individuals

Can make a **within transformation**
First step: take within individual averages – rewrite the long regression so as to take individual means

$$\overline{Y_i} = \sum_t Y_{it}$$

$\overline{Y_i}$ is the time average of Y for individual i overtime
Type of available data is earnings and tax rate
$\overline{Y_i}$ is the average within an individual of their earnings over time
For individual 1: earns 300, 310, 320 in the three periods
So the average is going to be $\sum_{3t} Y_{1t} = 310$
Can rewrite the regression as

$$\overline{Y_i} = \alpha + \beta \overline{X_i} + \gamma S_i + \overline{u_i}$$

Second step: apply the within transformation
Subtract the model-in averages from the initial ones

$$Y_{it} - \overline{Y_i} = (\alpha - \alpha) + \beta(X_{it} - \overline{X_i}) + (\gamma S_i - \gamma S_i) + (u_{it} - \overline{u_i})$$
$$Y_{it} - \overline{Y_i} = 0 + \beta(X_{it} - \overline{X_i}) + 0 + (u_{it} - \overline{u_i})$$

Impossibility of controlling for S is eliminated
Solved the OVB problem exploiting the panel dimension of the data and applying the within transformation

No longer run a regression of Y_{it} on X_{it} but run a regression of $Y_{it} - \overline{Y_i}$ on $X_{it} - \overline{X_i}$
Implicitly construct two new variables equal to the value for each individual in every period minus the average for each individual and the same for X
Then regress the deviations of Y from the individual means on the deviations of X from the individual means

$$Y_{it} - \overline{Y_i} = \beta(X_{it} - \overline{X_i}) + (u_{it} - \overline{u_i})$$

There is no longer the OVB problem: but only because we are assuming that the omitted variable is constant over time
Intuition: if something is constant over time, it cannot be causally responsible for changes over time
When we consider the relationship between changes in Y and changes in X (deviations) cannot be attributed to a constant trait of the individual
Just focus on the changes within the individual: if the confounding factor is fixed overtime it cannot be responsible for changes between the individuals

Instead of doing within transformation we can do a **First difference transformation**
Rather than take deviation from the average behaviour of the individual look to how year by year changes in X affect year by year changes in Y

$$Y_{it} = \alpha + \beta X_{it} + \gamma S_i + u_{it}$$

$$Y_{i(t-1)} = \alpha + \beta X_{i(t-1)} + \gamma S_i + u_{i(t-1)}$$

And then take the difference between the two

$$Y_{it} - Y_{i(t-1)} = \alpha + \beta X_{it} + \gamma S_i + u_{it}$$

Do not compare an individual to the average behaviour that we are observing but just take the difference between the behaviour at time t and t-1 and try to correlate that with a change in the treatment over the time period
This allows again to get rid of the constant omitted factor
Less suboptimal: lose one data point

We are left with one observation for individual at the end
Third possibility is to use the **fixed effect method**
Even if do not observe it, still want to estimate γ
Don't need to have an exact quantification of S, but know that it is a fixed characteristic of all individuals
Just use a dummy variable, indicator for each individual that allows to estimate γ
Whilst γ is not going to be exactly a quantification, the relative magnitude of γ for each individual will be still useful

For doubts or suggestions on the handout:

FEDERICA DI CHIARA

+39 3279948330

@federicadichiara8

For info about our teaching division:

GIOVANNI BARBARO

+39 3277175240

@gianni_barbaro2

CARLOTTA CAROMANI

+39 3703723764

@carlottacaromani