Introduction of empirical methods in modern econometrics

Session 1: OLS and Instrumental Variable

Dianzhuo ZHU

2021-09-17

Overview of the course

- **A brand-new course** on econometrics and data science, **practical-oriented with hands-on examples**
  - Learn various methods and their empirical applications
  - Get familiar with programming tools commonly used in econometrics and in data science

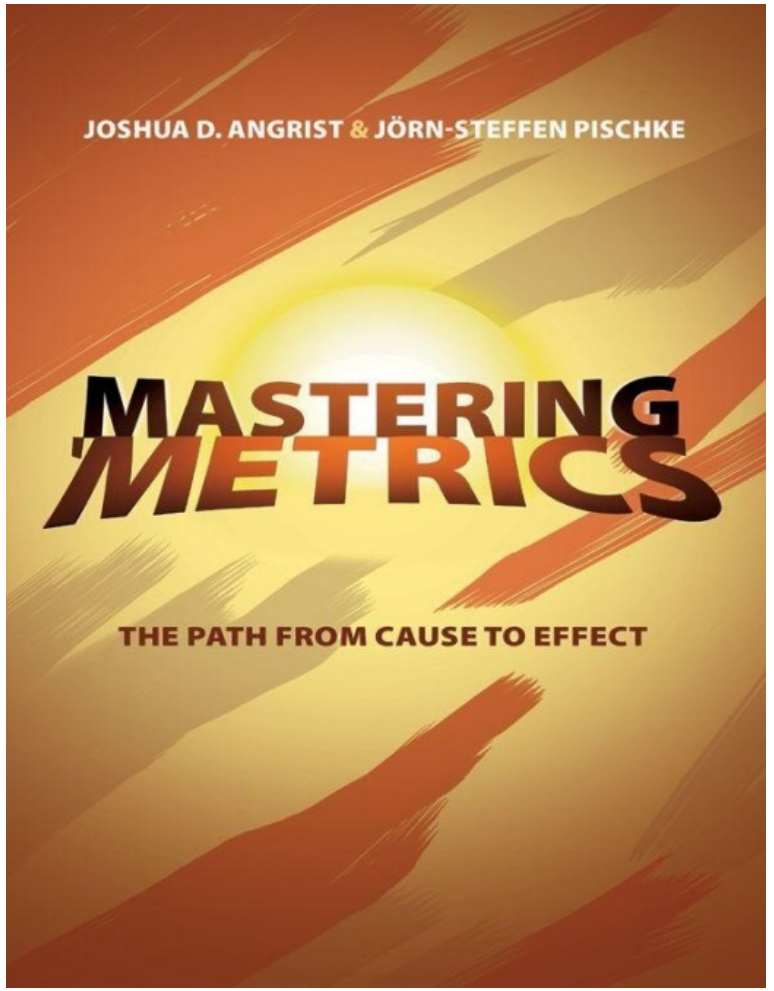
- At the end of the course, you should be able to
  - Conceive econometric models to detect causal relations and conceive machine-learning models
  - Implement econometric and machine-learning models in R and in Python
  - **Analyze, interpret and predict**

Overview of the course

- Course structure
  - 6h of catch-up sessions (optional)
  - 15h of econometrics (mandatory)
    - 6h theory, 6h TP, 3h group presentation + in-class experiment
    - Taught in English by Dianzhuo and Maria Teresa (TP)
  - 9h of data science (mandatory)
    - Theory + TP
    - Taught in French by Mustapha
- Grading policy
  - Presence: 15%
  - TP exercises: 15% * 2, in pairs
  - Group presentation: 15%, 10 people per group to read and explain an academic paper, based on a list of questions. The presenters will be randomly chosen
  - Final project : 40%, individual project, deadline to be announced

Overview of the course: econometrics

- **Crash course** on modern methods on causality
  - Instrument Variable (IV), Difference-in-Differece (DID), Regression Discontinuity (RD), Experiment (if time permits)
  - **Theory + intuition**
  - **Paper examples**
  - Could this method be used for your exercises 2 and 3 ?
- Main reference
  - *Angrist, J. D., & Pischke, J. S. (2014). Mastering'metrics: The path from cause to effect. Princeton University Press.*
  - Additional reading of each method listed at the end of each course
- Questions: dianzhuo.zhu@dauphine.psl.eu



Course outline

- Correlation vs. causation
- Ordinary Least Squares (OLS): A quick reminder
- Endogeneity
- Instrumental Variable (IV)
- Paper examples

Correlation does not imply causation!

- In regions where the police force is strengthened, crime rate is also higher
  - Police force causes crime to rise?
  - **Reverse causality!**
- More people died in France during the spring of 2020, while the air quality improved…
  - Better air quality causes more deaths?
  - **Omitted common-causal variable!** (lockdown)
- Have you read press articles that claim correlation as causation?
  - Note it down and share your example at next class

But sometimes it is not easy to distinguish

- Students who earn a better grade at school also have more books at home
  - Should we claim that parents should buy more books to boost the score of their children?
  - More books at home and better score are both related to parents' intellectual influence
  - More books at home -> reading habit -> better score

But sometimes it is not easy to distinguish

- People who subscribe to supplementary health insurance programs visit doctors more times compared to people who do not subscribe
  - Can we say that supplementary health insurance makes people less healthy?
  - Less healthy people self-select to buy supplementary insurance
  - The fact of having the insurance incentivizes people to do more precautionary visits

But sometimes it is not easy to distinguish

- Network effect
  - The more people who use Google Play Store for downloading apps, the more developers will choose to release apps on Google Play Store, and more people will use it because there are more choices…
  - How to measure the effect of user number on app availability?
  - **Simultaneity:** Supply is adjusted according to demand, but demand may also be induced or suppressed by supply capacity

Claiming causation

- Empirical economists try to quantitatively explain social events and human behavior
- Why did it happen? What factor(s) caused the result?
- Experiments: generate data
- What if we can only rely on observational data?
  - Look for an event or a policy that was not designed as an experiment, but turned out to match the criteria of one = natural experiment
  - Observational settings in which treatments are (as good as) randomly assigned among subjects
    - Instrumental variable

The OLS (Ordinary Least Squares) regression

Basic Model

Explaining a quantitative, continuous output variable with one or several explanatory variables:

$$y_i = f(x_i)$$

Ordinary least squares (OLS) one variable model:

$$y_i = a \cdot x_i + b + u_i$$

- y_i Dependent variable, outcome variable
- x_i Independent variables, explanatory variables, regressors, control variables
- u_i Error term, unobservable
- a Slope, coefficient of the explanatory variable
- b Intercept

The slides of the OLS and the endogeneity parts are retrieved and modified from the 2019-2020 econometric course of IREN, given by Nicolas Soulié and Serge Pajak

The OLS (Ordinary Least Squares) regression

Basic Model: interpretations

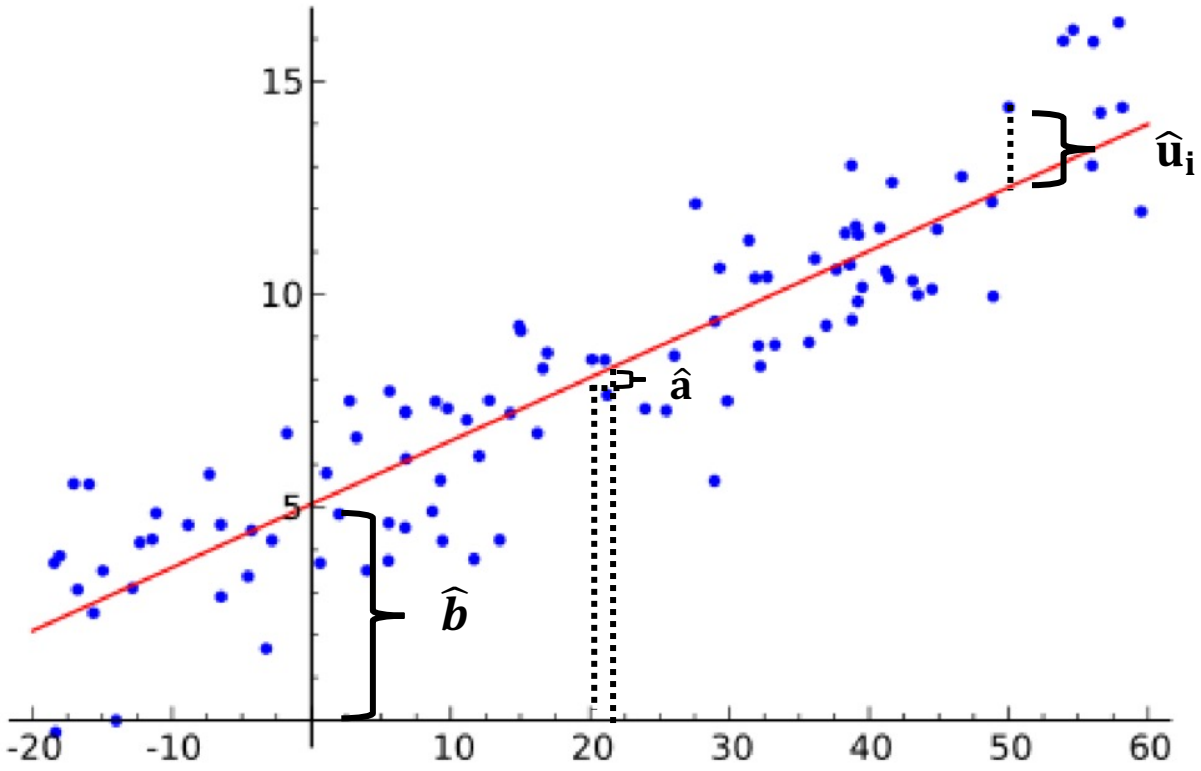
$$y_i = a \cdot x_i + b + u_i$$

- y_i and x_i are fixed, directly observed
- a and b are estimated (=computed)

  a reads "on average, in this sample, an additional unit of x adds a units to the value of y"
- u is a random variable, it captures what is specific to each observation (= not in the linear model)
- $E(y_i) = b + ax_i$: the fitted part is everything that is captured by the regressors

The OLS (Ordinary Least Squares) regression

Basic OLS Model: Visualization



Blue: observations y_i Red: fitted model $\hat{y}_i = \hat{a} \cdot x_i + \hat{b}$ Where are $\hat{a}, \hat{b}, \hat{u}_i$?

The OLS (Ordinary Least Squares) regression

Quality of the model

R^2: index of the explicative power of the model

How well a variable is explained by the explicative variables included in a model

R^2: index comprised between 0 and 100%

Large R^2 (i.e. close to 100%) mean strong explicative power (i.e. 100% of the explained variable variations are explained by the explicative variables), and inversely

The OLS (Ordinary Least Squares) regression

Quality of the model : details on goodness-of-fit / Analysis of variable

In the model with 1 variable:

Sum of the squares of errors	$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2$
Total sum of squares	$\text{SST} = \Sigma(y_i - \bar{y}_i)^2$
Sum of the squares of regression	$\text{SSR} = \Sigma(\hat{y}_i - \bar{y}_i)^2$

With the relationship:

$$SST = SSR + SSE$$

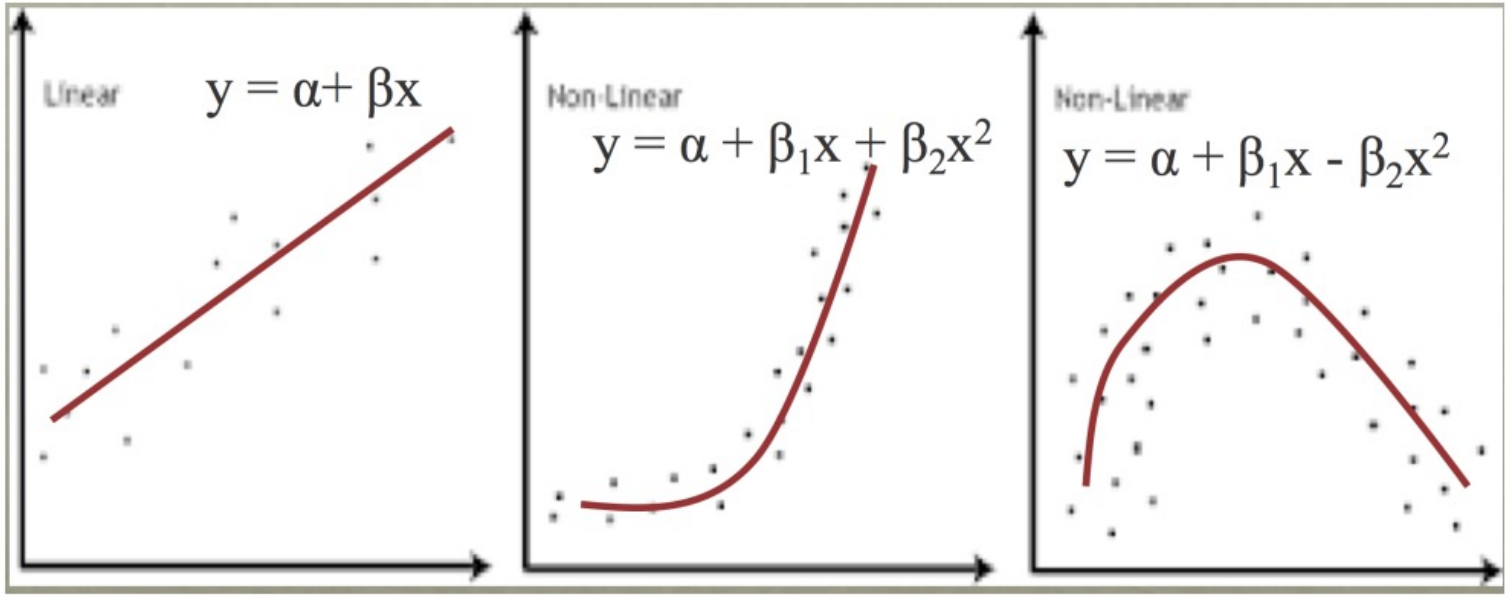
Therefore, the best fit (SSE = 0) means that SST = SSR. The R^2 measure of goodness of fit is:

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

The OLS (Ordinary Least Squares) regression

Usual OLS regression problems: linear relationship

When transforming explicative variable is not enough to improve linearity: need to create and include explicative variable at power x



The OLS (Ordinary Least Squares) regression

Gauss Markov assumptions

OLS1: True model is linear in parameters

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

OLS2: Random sampling (iid)

OLS3: x uncorrelated with u

OLS4: No linear dependence of regressors

OLS5: Homoskedasticity and no autocorrelation in the residuals

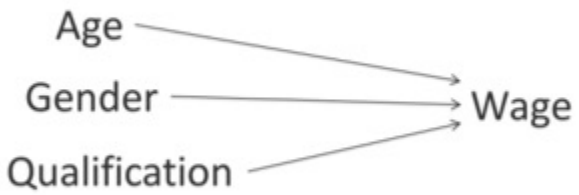
OLS is **consistent**

OLS is **efficient**

- These are the assumptions that make sure that the OLS estimators are the Best Linear Unbiased Estimatior (BLUE).
- Violating assumptions 1-4 will cause our estimation to be biased.
- Violating assumption 5 may make our estimation no longer efficient.

Endogeneity: Violation of the OLS model assumption

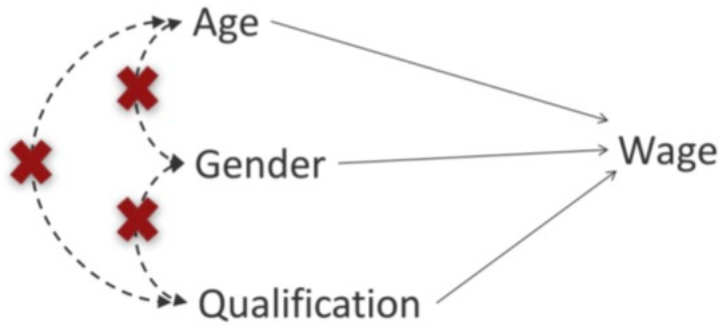
Consider an econometric model where variables v, w, x, ... explain the outcome (y)



Here, individuals' age, gender and level of qualification determine their wage level

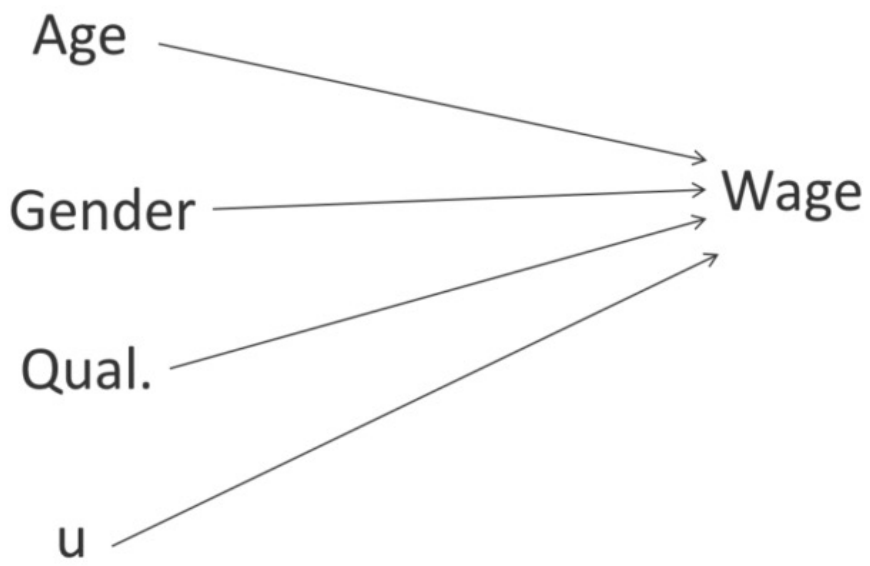
Endogeneity

- An important assumption: **independent variables (the Xs) should be exogeneous**
- Exogenous variable: Not determined by any variable IN THE MODEL
  - Not a linear transformation of other independent variables (perfect collinearity)
    - They can be correlated with each other (for example, gender and level of education/qualification are possibly correlated)
  - Not correlated with the error term u (otherwise violating Gauss-Markov assumption 3)



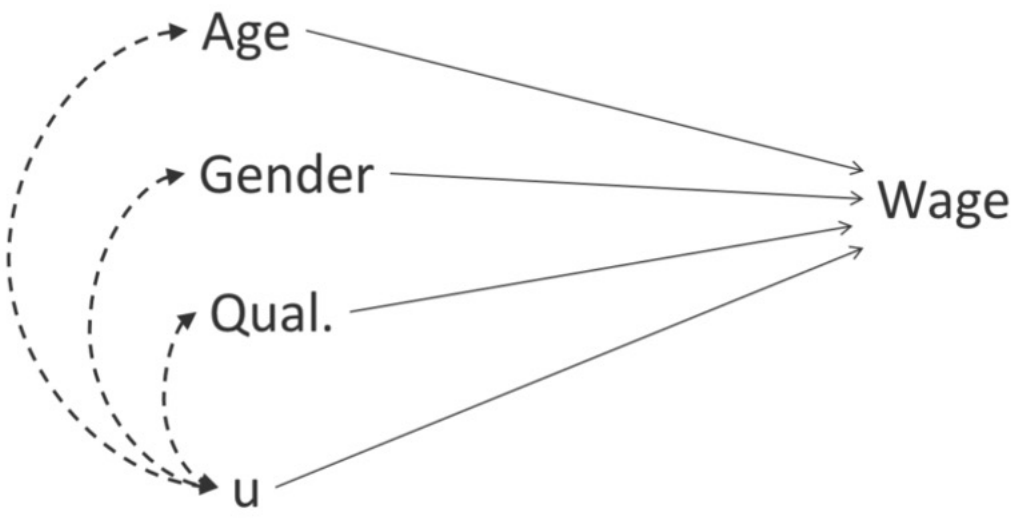
Endogeneity

In the complete model, in addition to the explicative variables there is an error term u



Endogeneity

- If the error term u is correlated with any of the three variables, then we have an endogenous issue (violation of the assumption 3)



Endogeneity

Source of endogeneity

- Omitted variable
- Reverse causality or simultaneity

What are the red flags?

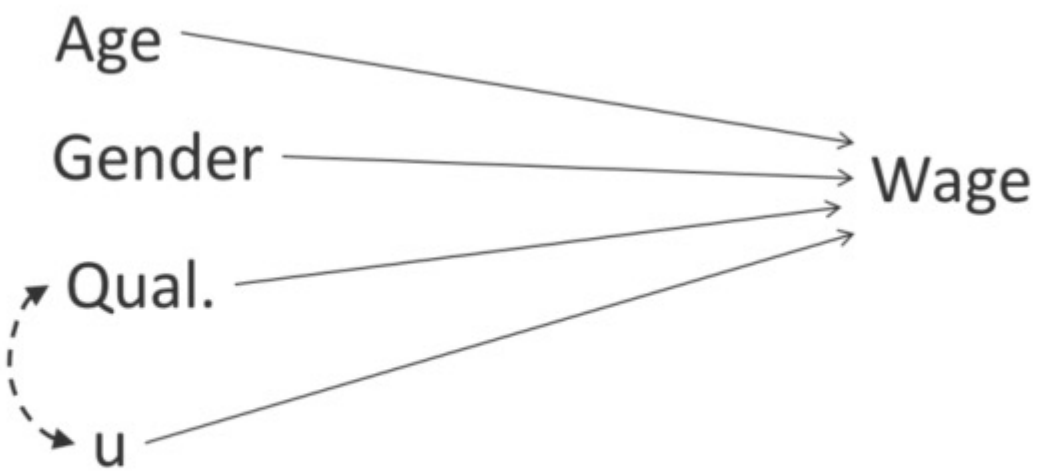
Often present with economic variables such as price and quantity, wage and number of years of education, investment and profits, elections results and economic policy, consumption and GDP, etc.

Formally

The explanatory variable is correlated with the residuals

Endogeneity

For instance: the error term (u) is correlated with the level of qualification



Then the variable 'level of qualification' is endogenous

What would be the reason for this?

Endogeneity

In this example, the coefficient for level of education could be biased due to the omission of an important variable that affects both the level of education and the wage: people specific skills (e.g. good/bad relationship management, higher/lower ability, etc.)

One can expect that high ability people:

Tend to be more educated,

Tend to earn higher wage.

So Ability affects both education and wage

Endogeneity

Consequences of omitted variable here

Without the Ability variable,

- highly capable people might have positive error term (they earn more than predicted by the average impact of school)
- and low skilled people might have negative error term (they earn even less than predicted by their level of school)

This error term might be correlated with education level as people with higher skills people tend to have higher level of education

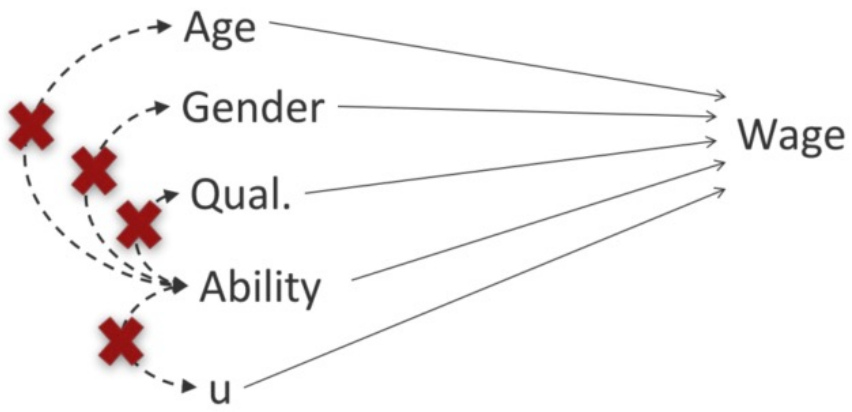
Technically, $E(u|x) \neq 0$

Then βx is not a good indicator because it includes at least two factors:

- The direct effect of schooling
- The indirect effect of ability on schooling

Endogeneity

The good model would have been:



But we may not be able to obtain this!

Several difficulties:

- How to measure ability?
- What is ability and qual (education) are highly correlated?
- How do we know that now qual and ability are exogenous now?

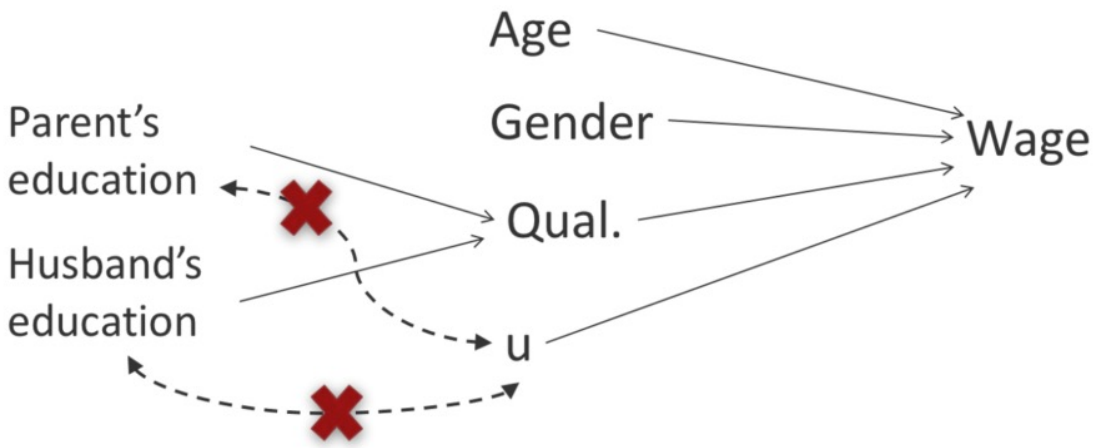
Endogeneity

- There are, of course, other possible omitted variables in the simplified model. What we have discussed in class: family richness, country, professional experience, professional position, ethnicity, etc.
- Adding more variables may improve the quality of our model
  - But adding too many variables or too many weakly explicative variables will harm the quality of the model
  - Need to find a balance
- However, not all these variables are correlated with level of education, meaning that they are not the source of endogeneity. It depends on the specific case of our data.
- Even though we have included a variable that is correlated with the endogenous one, we are not sure whether or not the error term is no longer correlated with the endogenous variable now.
- Instrumental variable helps solving the endogeneity problem and obtain the unbiased estimator of the endogenous variable

Instrumental variable

- An instrument is a variable that
  - Explains the endogenous variable
    - **No weak instrument** (reg of the endogenous variable on the instrument needs to be significant)
  - But is not correlated with the error term (*u*)
    - **Independence assumption** (the instrument is as good as randomly assigned)
- In other words: an instrument impacts the outcome only through the instrumented (endogenous) variable (**exclusion restriction**)

Instrumental variable



To find a good instrument:

Knowledge of the topic (intuition, previous literature)

+

Imagination

+

Luck

Using IV to estimate: 2SLS (Two-Stage-Least-Squares)

Given the linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \tag{1}$$

where X_1 is an endogenous variable

First stage of 2SLS: Regress the endogenous variable on the instrument and other explanatory variables and obtain the estimated value of the endogenous variable

$$\hat{X}_1 = \gamma_0 + \gamma_1 Z_1 + \gamma_2 X_2 + \gamma_3 X_3 + v \tag{2}$$

where Z_1 is the instrumental variable. We call the above equation the **"reduced form"** of X_1

- γ_1 needs to be significant (why?)
- $\hat{X}_1$ is no longer related to ϵ (why?)

Second stage of 2SLS: Plug in the fitted values of $\hat{X}_1$ derived from equation (2) into the original linear regression equation (1):

$$Y = \beta_0 + \beta_1 \hat{X}_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \tag{3}$$

now ϵ is no longer correlated with $\hat{X}_1$, and β_1 is an unbiased estimator

Until next course

- Read Angrist and Krueger (1991) again with the following questions in mind
  - Which variable is endogenous? Why?
  - Which instrumental variable is proposed?
  - How do the authors justify the validity of the instrument?
  - How is the 2SLS done?
- We will analyze this paper together in the next session (October 1st)
- Reference
  - Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. *The Quarterly Journal of Economics*, 106(4), 979-1014.

OLS and IV: To learn more

- **To better understand the theoretical part of OLS and IV**
  - Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press. (Chapters 3 and 4)
  - Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach*. Nelson Education. (For more knowledge on OLS)
  - Ben Lambert's Youtube channel
- **To learn more examples of IV**
  - Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, *15*(4), 69-85.

OLS and IV: To learn more

- **Some interesting IV examples**
  - Sibling-Sex composition: Angrist, J. D., & Evans, W. N. (1996). *Children and their parents' labor supply: Evidence from exogenous variation in family size* (No. w5778). National bureau of economic research.
  - Rainfall variation: Miguel, E., Satyanath, S., & Sergenti, E. (2004). Economic shocks and civil conflict: An instrumental variables approach. *Journal of political Economy*, *112*(4), 725-753.
  - Distance: Campante, F. R., & Do, Q. A. (2014). Isolated capital cities, accountability, and corruption: Evidence from US states. *American Economic Review*, *104*(8), 2456-81.