

Applied Health Economics (for B.Sc.)

Helmut Farbmacher, Corinna Hartung

Department of Economics
University of Munich

Winter Semester 2018

Outlook

0 Organizational issues (Timeline, Topics, etc.)

1 Linear models (OLS, Omitted variables)

0. Organizational Issues

- You have to write a term paper (around 30.000 characters; approx. 15 pages without figures and tables) based on assigned research articles.
- And you have to present (approx. 15-20 minutes) in the seminar (Monday, 17th December 2018)
- The term paper has to be written between 12 November and 9 December 2018 (4 weeks).
- Meeting with the supervisor (Corinna Hartung or me) in the second week.

0. Organizational Issues

- You have to hand in the term paper **in pdf format** on 9 December 2018, via email to farbmacher@econ.lmu.de.
- You have to hand in your presentation slides (preferably in pdf format) the day before the presentations, 6pm (at the latest), by email to farbmacher@econ.lmu.de.
- Grading: term paper (50%), presentation and active participation in the seminar (50%).

0. Organizational Issues

Topics

- Education and health
- Causal links between socioeconomic status and health
- Health policy and health behavior
- Mendelian randomization in economic research

1. Linear models

1.1 Main Concepts in Econometric Analysis

1.2 The OLS Estimator

1.1 Main Concepts in Econometric Analysis

Introduction

- In economics we are often interested in causal effects. For instance, what is the causal effect of school education x on later income y .
- Ideally, we would have an experiment, where the researcher could manipulate x and then measure y . Then the causal effect can be easily estimated as

$E(y|x = \text{good educ}) - E(y|x = \text{bad educ})$ if x is binary/discrete

$\frac{\partial E(y|x)}{\partial x}$ if x is continuous

1.1 Main Concepts in Econometric Analysis

- In economics we often do not have experimental data.
- Estimating causal effects from observational data is harder. We could, for instance,
 - control for **sufficiently** many other factors (ceteris paribus analysis)
 - use instrumental variables
 - use panel data estimators
 - ...

1.1 Main Concepts in Econometric Analysis

The Population Model

- We assume that our data come from a random sample of the relevant population.
- Suppose the observations ($i = 1, \dots, n$) in our dataset have been generated by the following population model (“data generating process” - DGP)

$$y_i = f(x_i, u_i, \beta) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + u_i = x_i' \beta + u_i$$

- Stacking the observations, we can also write $y = X\beta + u$. Here, y and u are $n \times 1$ vectors and X is a $n \times k$ matrix.

1.1 Main Concepts in Econometric Analysis

- Consider our n realizations of the population model

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_i$$

which can be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

or simply

$$y = X\beta + u$$

1.1 Main Concepts in Econometric Analysis

- Remarks:
 - β is a $k \times 1$ vector of unknown parameters, which we want to estimate.
 - Very often we are just interested in a single element of β . Say, the effect of education on wages or health.
 - Because $f(x_i, u_i, \beta) = x_i' \beta + u_i$ we know that the true functional form is linear. In practice, we don't know the true DGP and it could be far more complicated. Giving rise to semi- or non-parametric estimation techniques.
 - The population model does not necessarily have to be the model we are going to estimate (e.g., because of functional form misspecification, one of the x 's is not in our data or measured with error, . . .).

1.2 The OLS Estimator

Main assumption

- Assumption OLS.1: $E(X_i u_i) = \underset{k \times 1}{0}$. This means that u is uncorrelated with each regressor.
- A regressor x_j is called **endogenous** (in the econometric sense) if it is correlated with the error term (u), where $cov(x_j, u) \neq 0$ implies $E(x'_j u) \neq 0$. Otherwise, it is called **exogenous**.
- Potential threats of identification
 - Omitted variables
 - Measurement error in observed variables
 - Reverse causality: Not only x affects y but y simultaneously affects x

1.2 The OLS Estimator

Omitted Variables

- An example: What is the effect of education on income?
- Suppose the **population** model is

$$y = \beta x + \gamma q + u .$$

where y is wage, x is education measured in years of schooling and q is ability (unobservable for the researcher).

- We assume that our dataset has been generated by this process. However, while x is observed, q is not. We are also willing to assume that $E(x_i u_i) = 0$.

1.2 The OLS Estimator

- As data on q is missing, the best we can do is to **estimate** the following model

$$y = \alpha x + w$$

where $w = \gamma q + u$ is a composite error term.

- What is the probability limit of the OLS estimator for α ?
- Can we say something about the potential direction of the bias?

1.2 The OLS Estimator

- Wage is determined by: $y = \beta x + \gamma q + u$
- We instead estimate: $y = \alpha x + w$
- If x and q are related by: $q = \delta x + \epsilon$
- Then,

$$\begin{aligned} y &= \beta x + \gamma(\delta x + \epsilon) + u \\ &= \underbrace{(\beta + \gamma\delta)}_{\alpha} x + \underbrace{(\gamma\epsilon + u)}_w \end{aligned}$$

1.2 The OLS Estimator

$$y = \underbrace{(\beta + \gamma\delta)}_{\alpha} x + \underbrace{(\gamma\epsilon + u)}_w$$

If y is regressed on x alone, α will be the estimated slope on x

1.2 The OLS Estimator

$$y = \underbrace{(\beta + \gamma\delta)}_{\alpha} x + \underbrace{(\gamma\epsilon + u)}_w$$

- We expect that $\delta > 0$ and $\gamma > 0$ (Why?)
- The return to education will be **overestimated** if $\gamma\delta > 0$

Intuition:

- People with many years of education earn higher wages on average
- But this is partly due to the fact that people with more education are also more able on average

1.2 The OLS Estimator

- We have to include q in our regression to avoid falsely attributing the explanatory power of q to x . Example: The parameter of education is biased if we do not control for ability.
- The estimator of β will be consistent **only** when either $Cov(x, q) = 0$ or $\gamma = 0$.
- Otherwise, it will suffer from omitted-variables bias.

1.2 The OLS Estimator

- We can only estimate the real difference in earnings if we keep all other economic reasons constant (e.g., ability), we often say “we control for these other variables”.
- Of course to do so means we have to **know and observe** all relevant economic variables when we use OLS. Panel data estimators or instrumental variables estimators may help.
- The advantage of instrumental variables (IV) estimation is that we only should **know** these other variables to convince the audience that our instrument is uncorrelated with these factors. We do **not** necessarily have to **observe** these variables.
- Panel data estimators help us to get rid of **time-constant** unobserved variables.

Content of term paper and presentation

Describe

- the research question answered in the paper (and why it is interesting)
- what the difficulties are in answering this question (e.g. endogeneity)
- the strategy which the authors used to solve these difficulties (i.e. describe the methods/data and give an intuition for how they help in this specific setting)
- your conclusion

Suggestions for finding help

- Guide on how to write a term paper at www.farbmacher.de
- Text books
 - Hansen (2018) “Econometrics”, avail. online Hansen’s textbook
 - Angrist and Pischke (2009) “Mostly Harmless Econometrics”, Princeton University Press
 - Cameron and Trivedi (2010) “Microeconometrics Using Stata”, Stata Press

1.3 The 2SLS Estimator

- We are still interested in the same population model

$$y = \beta x + \gamma q + u \text{ with } w = \gamma q + u$$

- We know OLS is inconsistent because of the correlation between x and q . Without additional information, we cannot estimate β consistently.

1.3 The 2SLS Estimator

- Let's assume there is an additional variable Z which satisfies the following conditions
 - $cov(w, Z) = 0$ (exogeneity assumption)
 - $cov(x, Z) \neq 0$ (relevance assumption)
- We call such a variable Z an instrumental variable (or just instrument).

1.3 The 2SLS Estimator

- Denote the number of available instruments m and the number of endogenous variables p , then
 - a model is called “just-identified” if $m = p$
 - a model is called “over-identified” if $m > p$
- In most empirical applications $p = 1$
- More formally the assumptions to make 2SLS consistent are
 - Assumption 2SLS.1: $E(Z'w) = \underset{m \times 1}{0}$
 - Assumption 2SLS.2a: $\text{rank}(Z'x) = p$
 - Assumption 2SLS.2b: $\text{rank}(Z'Z) = m$

1.3 The 2SLS Estimator

- The two-stage least squares (2SLS) estimator can be derived in two stages.
 - First regress the endogenous variable x on Z :

$$x = Z\hat{\pi} + \hat{v}$$

$\hat{\pi}$ is such that $cov(Z, \hat{v}) = 0$.

- Now, predict $\hat{x} = Z\hat{\pi}$ and regress \hat{x} on y

$$y = [Z\hat{\pi} + \hat{v}]\beta + \gamma q + u = Z\hat{\pi}\beta + \hat{v}\beta + \gamma q + u$$

1.3 The 2SLS Estimator

- Remarks: We only used the OLS estimation technique to derive the 2SLS estimator! Although the point estimates are identical using two stages, the SEs are wrong. Why?
- To test the relevance assumption, we can test whether all elements in π are jointly zero using an F -test.
- If the model is over-identified, we can also (partly) test the exogeneity assumption.

1.3 The 2SLS Estimator

- Getting back to our application: What is the effect of schooling on earnings (private returns to education)?
- Many instruments have been discussed in the literature. Here are some examples:
 - Parents' education
 - Proximity of a collage or university (Card, 1995)
 - Quarter of birth (Angrist and Krueger, 1991)
- For each of these potential instruments, we can test (relevance) and discuss (exogeneity) the IV conditions.

1.3 The 2SLS Estimator

Final remark: Proxy variables

- Proxy variables can also provide a solution to the omitted-variable problem (e.g., IQ test results as proxy for ability)
- However, there is an important difference between instruments and proxy variables:
 - While a proxy should be highly correlated with the omitted variable to account/“control” for this omitted variable,
 - the instrument has to be uncorrelated with the omitted variable. This is because the omitted variable is part of the error term and we want the instrument to be uncorrelated with the error term.