Applied Econometrics for Researchers

Introduction to Econometrics; Ordinary Least Squares

H.C. Kongsted

Department of Strategy and Innovation
Copenhagen Business School
Denmark

Outline of this morning's lecture

Introduction

The value of having great data

- Data, Datasets, and Variables

Inferential Statistics

Regression

- Core: Ordinary Least Squares (OLS)

Slope and significance

- The estimate
- Its significance

Identification issues

A typical module

- We meet for intro + 8 modules, see course outline. Note: Breaks in some weeks. Exam (December 13).
- Lecture session Wednesdays 9.00-12.00. Workshop session (on-line) Fridays 9.00-12.00.
- Organized around a mixture of theory and practice
  - Theory in the lectures session;
  - Class teaching with slides, Q&A
  - Workshops provide the data used for each exercise, you solve a problem using Stata, and present/discuss your results
- Exam: in the format of a typical workshop, except that it is solved on an individual basis
- A textbook would be useful (Wooldridge, Cameron/Trivedi); links to articles provided in CANVAS

What are the (typical) data at hand?

- Surveys: of firms or individuals:
  - Ex: Valentina Tartari and I did a survey of Danish academics in 2017. Population: all academics in Denmark. Sample: $\approx$ 4,800 responses (38%)
  - Ex: Community Innovation Surveys. Sample of firms. Run in a number of developed economies. We will work with actual data from a vintage of the UK survey ($\approx$ 1,100 firms).
- Administrative registers: in firms/organizations/government
  - Data on patents and inventors from the European Patent Office: Ulrich Kaiser and I collected data for Danish firms and Danish inventors
  - Government registry data: data on individuals (and firms) in DK: Family, education, health, earnings, ...

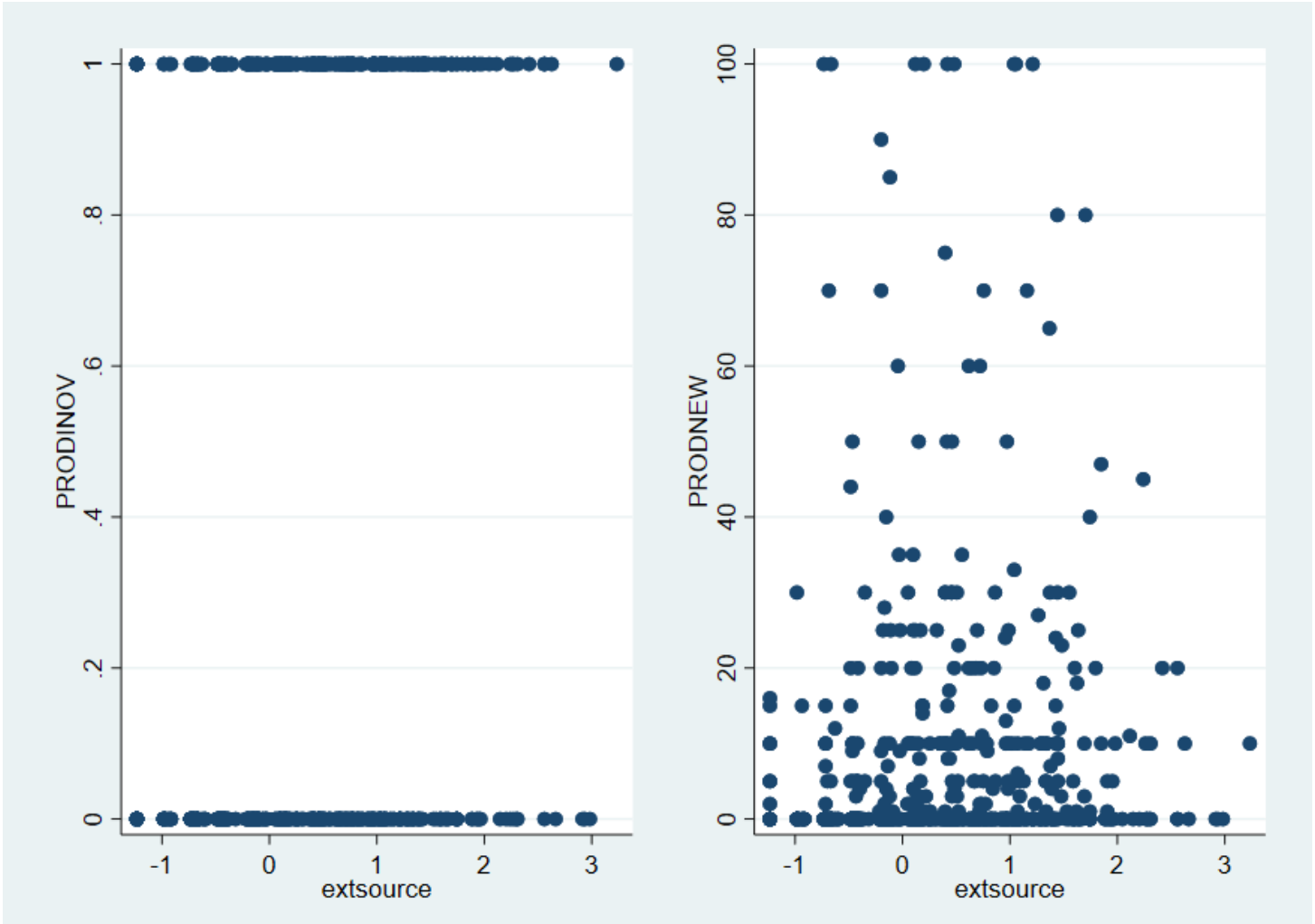
What are the (typical) data at hand?

- Scraped from the web: increasing use of data eg. from LinkedIn, facebook, etc
- Each type of data source has pros and cons, for different purposes
- Real magic (like whole new insights and/or publications in top-journals) can happen if you are able to combine data from several sources

Example: Innovation in British firms

- Modelling product innovation—and the use of external knowledge—in British firms, the potential role of internal R&D, as well as a host of "control" variables
- Example based on data from the UK Community Innovation Survey (CIS)
- Build further on this example and the CIS data more generally in lectures and workhops
- Link to UK CIS questionnaire

Example: Innovation in British firms



Different types of datasets (all relate to the same phenomenon)

- Cross-sectional: Observations drawn at same point in time: Ex: Sample of startups in DK in 2019.
- Time Series: A single observational unit, often aggregate, observed across time: Ex Data on the number of start-ups in DK, by year, from 1980 to 2018.
- Pooled Cross Sections: Cross sections observed at different times—any overlap across observational units over time is unintentional: Ex: Startups observed in their year of founding, between 2010 and 2019.
- Panel data: the *same* cross sectional units (persons, firms, ...) are followed across time: Ex: Startups in 2010 are followed for 10 years.

Types of variables

Two main types of variables:

- Continuous (Sales, Wages, Temperature,...)
- Discrete/Categorical (Industry (service, manufacturing, public sector,...), University Degree (None, B.Sc., M.Sc., PhD)

How about these examples:

- A finite number of possible categories (Ex: technology classification of patents)
- No (upper) bound (Ex: number of patents granted to a firm in a given year)

Continuous variables can be converted into discrete categorizations but not *vice versa*

Types of Scales/Variables

- Nominal: No natural ordering of values, order of listing is irrelevant (Ex. industry affiliation of a firm)
- Ordinal: There is a natural ordering of values, however "distance" is not well-defined (Ex. level of completed university degree)
- Interval: Natural ordering and numerical distance between alternatives matter (Ex.: temperature in degrees C or F)
- Ratio: Zero point and ratios meaningful (Ex.: Number of years of university education)
- Type of data to be modelled dictates the appropriate econometric method

Types of Scales/Variables

- Some tools are only strictly appropriate when studying continuous variables
- Specific tools for dealing with categorical variables and counts
- Depends on the purpose of analysis and the question asked
- Linear models will often turn out be (surprisingly) helpful

Testing hypotheses

- Econometrics is often about testing a hypothesis:
  - Do PhDs benefit from their degree in terms of wages (and why (not)?)?
  - Do R&D intensive firms produce more inventions?
  - The heavier an object is, the quicker it will hit the ground when dropped from a height of one metre in a no-air-resistance setting
- The aim is to set up a clearly defined benchmark, a "null" hypothesis (H_0). This allows you to estimate the likelihood that it can be rejected.
- However, we should also have an idea about what is the proper alternative (H_a) if we reject H_0

Types of Errors

- Type I Error: Rejecting a true hypothesis
- Type II Error: Failing to reject a false hypothesis
- First identify a region of probability in which you are "willing to" make a Type I Error
  - The probability of a Type I Error is often referred to as the level of significance, $\alpha = P(\text{Reject} H_0 | H_0)$
- Then we aim to minimize the likelihood of a Type II Error
  - 1- the probability of a Type II Error is often referred to as the *power* of the test, $1 - \pi(\theta) = P(\text{Reject} H_0 | \theta)$ – the ability of the test to detect that H_0 is not true when in fact it is not

The simplest example

- We may for instance test whether a continuous variable y is generated by a distribution in which the mean is μ_0:

$$H_0 : \mu = \mu_0, H_a : \mu \neq \mu_0$$

- Example of a two–sided test: Alternative can be greater or smaller than hypothesized value (just different).
- How can we test this simple hypothesis:
  - Collect a random sample of y's: $y_1, y_2, ..., y_n$
  - Use average value $\bar{y}$ as *estimator* of the (unknown) mean.
  - Question: Is the *observed* average sufficiently different from μ_0 that we can reject H_0?

The z-test

- z-tests are useful for testing hypotheses, strictly speaking regarding the mean of a normal distribution function
- The test relies on the calculation of a test score, z, which can be compared to a critical value signifying whether the value is high or low given our chosen level of significance

$$z = \frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{1}$$

- We can then look up the critical value in the z-table (link Z-table)
- For a 5% level of significance and a two-sided test, we consider critical values of -1.96 and +1.96.

Regression

- We need a too that enables us to quantify relationships *between* variables
- Often to quantify the *effect* of one variable, x, on another, y
- Preferably keeping all else constant.
- Some pretty causal language is being used here (why?)
- The most basic econometric tool: Ordinary Least Squares

Regression

- Regression is a statistical method used to quantify any linear association between–on the one side–a dependent variable and–on the other side–one or more independent variables
- There are many different types of regression techniques that are applicable in different scenarios depending on the characteristics of the data
- Ordinary Least Squares (OLS) is the most basic type of technique – Understanding OLS will allow you to understand the general principles of regression analysis

The Regression Equation

- We seek to understand if the variation of one variable, y, is associated with the variation of another, x
- We start with a simple linear functional form:

$$y = \beta_0 + \beta_1 x \tag{2}$$

- β_0 refers to the intercept on the y axis ($x = 0$) and β_1 is the (unknown) slope parameter (the change in y given a unit change in x)
- In practice, we will never find a perfectly fitting line between two variables. Extend by with an error term (u) – something that is left unexplained by the model

$$y = \beta_0 + \beta_1 x + u \tag{3}$$

Terminology

- Often you see the regression written as:

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{4}$$

- where the subscripts refer to each single observation i
- y may be referred to as the dependent variable, explained variable, response variable, predicted variable, or the regressand
- x may be referred to as the independent variable, explanatory variable, control variable, predictor, regressor

Notation

- In case we have multiple x variables, you may see the regression written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i \tag{5}$$

- or more compactly as

$$y_i = \beta_0 + \boldsymbol{\beta}' \mathbf{x}_i + u_i \tag{6}$$

- where bold symbols refer to vectors—a number of variables and a coefficient for each variable
- We use "explanatory" for independent variables that are of particular interest; "controls" are added to the equation to account for other variables that potentially will affect y and correlate with the explanatory variables.

Why multiple regression?

- Why don't we just stick to simple correlation and z-tests instead of doing regression analysis?
  - Correlations and simple one-variable models are likely to suffer from *spurious* correlation since relevant factors are not controlled for
    - Ex: A spurious correlation between R&D and profits, when not controlling for industry affiliation
    - Ex: Based on a z-test, we see that there is a significant difference in the wages of men and women – this could partially be explained by occupational choice
- We want to ask the question "what is the effect of a unit change in x, *keeping everything else constant?*"

Ordinary Least Squares (OLS)

- Ordinary Least Squares is the obvious first choice when your dependent variable is a continuous variable
- It is usually the "point of reference" in any application, and (surprisingly) often it provides a valid answer to the problem
- Useful to examine a bit of technique: Go back to the bivariate case to show explicit expressions and gain some intuition.

Technique of Ordinary Least Squares

The OLS estimator is obtained by minimizing the sum of squared *residuals*, $\hat{u}_i$, the difference between the estimated values, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta} x_i$, and the actually observed value of y

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \underbrace{(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))}_{\hat{u}_i}^2 \quad = \quad \min \sum_{i=1}^{n} \hat{u}_i^2 \qquad (7)$$

Technique of Ordinary Least Squares

Solving this min problem leads to the following equations from which the values of the *estimates*, $\hat{\beta}_0$ and $\hat{\beta}_1$, can be determined:

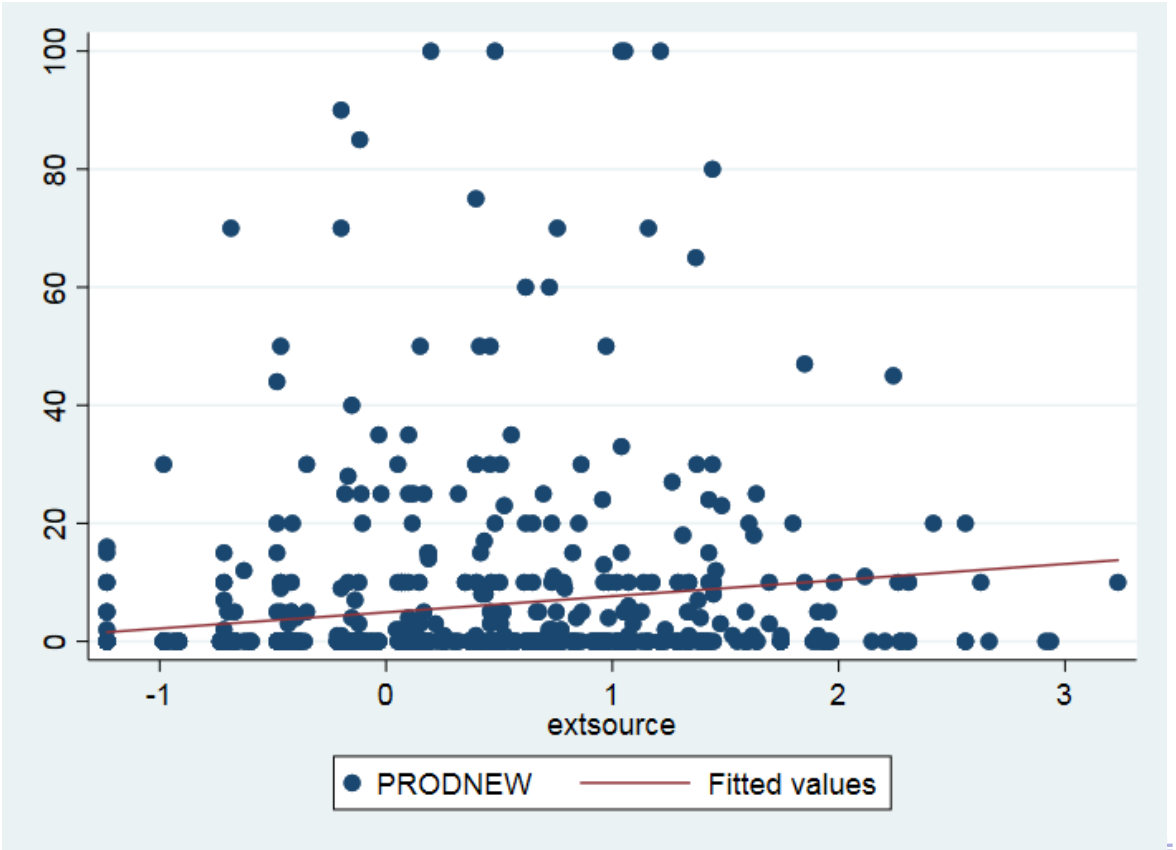
$$\sum_{i=1}^{n} y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \sum_{i=1}^{n} \hat{u}_i = 0 \quad (8)$$

$$\sum_{i=1}^{n} x_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = \sum_{i=1}^{n} x_i \hat{u}_i = 0 \quad (9)$$

$$(10)$$

By construction, the sum of the residuals is zero and the residuals are unrelated to the values of x_i. The regression "exhausts" any linear dependence between y and x.

Example: Innovation and use of external knowledge sources in British firms: Scatter plot and regression line



Example: Innovation in British firms: Regression output from Stata

```
. reg prodnew extsource
```

Source	SS	df	MS			
Model	5806.18069	1	5806.18069	Number of obs	=	773
				F(1, 771)	=	31.42
Residual	142461.411	771	184.774852	Prob > F	=	0.0000
				R-squared	=	0.0392
				Adj R-squared	=	0.0379
Total	148267.591	772	192.056465	Root MSE	=	13.593

prodnew	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
extsource	2.731872	.4873449	5.61	0.000	1.775191	3.688552
_cons	4.927718	.4889416	10.08	0.000	3.967903	5.887532

The slope parameter

- The slope parameter shows us the "best-fitting" value of the slope of the relationship between the variables
- For the simple case of just one explanatory variable, we can calculate the slope parameter by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{11}$$

- Note: The equation is simply the covariance between x and y, divided by the variance of x

The slope parameter - and so what?

- We often wish to understand if the *true but unknown value* of the slope parameter is equal to or different from zero
- In case it is zero, we can say that the two variables are not (linearly) related to each other
- We need somehow to estimate the "likelihood" that the slope is equal to zero, based on the data at hand.
- We do this by testing the following set of hypotheses:

$$H_0 \quad : \quad \beta_1 = 0 \tag{12}$$
$$H_a \quad : \quad \beta_1 \neq 0 \tag{13}$$

- How do we conclude on the validity of H_0?

Uncertainty of slope parameter

- We consider our sample *as if* it were just one out of many potential samples. This means we treat our estimate $\hat{\beta}_1$ of the slope (and the intercept) as *stochastic.*
- We need a measure of the uncertainty with which we have determined the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Such a measure can also be determined from the data.
- One determinant of the variance of the slope parameter is how "noisy" are the observations around the regression line: The variance of the error term u.
- The variance of the error term is also called the squared standard error of the model: σ^2

Uncertainty of slope parameter II

An estimate of σ^2 is obtained from the residuals as:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 \tag{14}$$

- The standard error of the estimate $\hat{\beta}_1$ is a measure of the uncertainty associated with the slope estimate

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{15}$$

- Depends positively on σ; negatively on the total variation of x.

Is it equal to zero?

- To uncover the "likelihood" that the estimated parameter is equal to zero, we consider the relative size of the estimate compared to its standard error - also known as the t-value

$$t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \tag{16}$$

- Note: This is the t-statistic for one particular hypothesis - $\beta_1 = 0$.
- We wish to understand whether t is a large value (in absolute terms)–the estimate is substantially higher than its associated standard error–to conclude on the validity of H_0.

Example: Innovation in British firms: Regression output from Stata

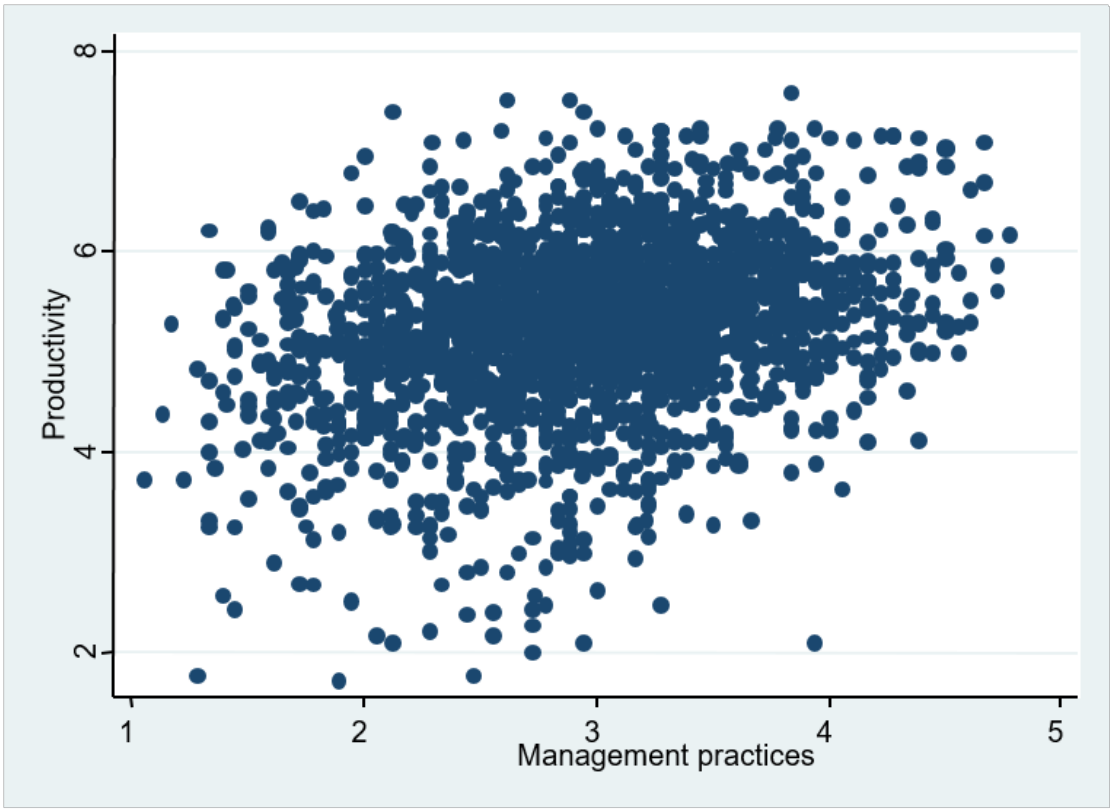
```
. reg prodnew extsource
```

Source	SS	df	MS		
Model	5806.18069	1	5806.18069	Prob > F	0.0000
Residual	142461.411	771	184.774852	R-squared	0.0392
				Adj R-squared	0.0379
Total	148267.591	772	192.056465	Root MSE	13.593

Number of obs	=	773
F(1, 771)	=	31.42

prodnew	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
extsource	2.731872	.4873449	5.61	0.000	1.775191	3.688552
_cons	4.927718	.4889416	10.08	0.000	3.967903	5.887532

Identification: What is driving this relationship?



The identification problem

- The effect of improved management practices, keeping all else constant.
- Is it actually what we would be estimating here?
- Or: is there a risk that our estimate is "confounded" by other (unobserved) factors that (also) influence productivity (the dependent variable)?
- Logically, this problem precedes estimation. And it has little to do with "having enough data".

Association versus causality

- You theorize about a (positive) relationship between firm productivity, Y, and the implementation of management practices, X, (job rotation,group/individual incentives, ...)
- You observe that Y and X are correlated over the sample.
- There can be several causal drivers of this correlation:
  - X may affect Y (as hypothesized)
  - Y may affect X (reverse causality)
  - X and Y may be jointly influenced by Z, a variable outside the model
- Theory may suggest causation; but we need to be able to "control for" (or randomize) other potential causes: main agenda for the second half of the course.

Correlated omitted variable: Simple case

- The "true" model (satisfies all ideal assumptions)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

- The actual regression equation (by OLS)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}$$

- The expected value of the OLS estimator

$$E(\hat{\beta}_1 | x_1, x_2) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2})}{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})^2} = \beta_1 + \beta_2 \frac{cov(x_1, x_2)}{var(x_1)}$$

- Biased if $\beta_2 \neq 0$ and $cov(x_1, x_2) \neq 0$!!

What's next?

- Friday: Workshop 1 - get you started on using Stata; starting the analysis of the UK CIS data set.
- Online with Zoom.
- Next week: No teaching (fall break)
- Two weeks from now: (Much more) on linear regression; show that it is a useful tool to answer a host of questions that often pop up in our research.