Lecture 1. What is econometrics?

Econometrics is about measuring economic relations

1. What is an economic relation?
2. How do you measure it?
1. Economic relations

An economic relation is a relation between economic variables

Examples

- The FED stabilizes economic activity by increasing or decreasing short term interest rates
- The recent tax cut was passed on the hypothesis that there is a relation between the level of taxes and economic growth
- A house buyer/seller is interested in the relation between price and amenities of a house
- A company wants to know what the effect is of its ad expenditures on sales or market share
- A student wants to know what the return is to (further) education
How do we represent economic relations?

In all examples there is/are

- Dependent variable that is the outcome of interest (economic activity, economic growth, house price, market share, income)
- Independent variable(s) the effect of which on the outcome we want to measure (interest rate, tax rate, surface of house, ad expenditures, level of education)

Methods in this course assume that there is exactly one dependent variable.

The dependent variable is denoted by: \( Y \)

There may be many independent variables

Independent variables are denoted by:

\[ X_1, X_2, \ldots, X_K \]
We use mathematical expression to represent the relation between the dependent variable $Y$ and independent variables $X_1, X_2, \ldots, X_K$.

For instance, if there is only one independent variable $X$ we could assume a linear relation

$$Y = \alpha + \beta X$$

The graph of this relation is a straight line with intercept $\alpha$ and slope $\beta$.

Measuring the relation is now measuring the coefficients $\alpha, \beta$. 
Note that by redefining $X_1, X_2, \ldots, X_K$ we can also specify nonlinear relations

$$Y = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3$$

is a nonlinear relation in $X$ (a cubic in $X$).

The key is that this relation is linear in the coefficients!
2. Measuring economic relations

In this course we only consider linear relations, i.e. linear in the coefficients.

In the case of one $X$ measuring the relation is obtaining numerical values for $\alpha, \beta$.

How do we do this?

Method 1: Use a graph. A linear relation/straight line is determined by two points. With two observations on $Y, X$ we can determine $\alpha, \beta$.

Example: for two houses we obtain price and square foot. Denote these numbers by

$X_1, Y_1$ and $X_2, Y_2$
If the linear relation is correct we have

\[ Y_1 = \alpha + \beta X_1 \]

\[ Y_2 = \alpha + \beta X_2 \]

We can solve this for \( \alpha, \beta \) to find the numerical/computed values \( \hat{\alpha}, \hat{\beta} \)

\[ \hat{\beta} = \frac{Y_2 - Y_1}{X_2 - X_1} \]

\[ \hat{\alpha} = \frac{X_2 Y_1 - X_1 Y_2}{X_2 - X_1} \]

Now consider the case that we have \( n \) observations denoted by the \( n \) pairs

\[ Y_i, X_i, i = 1, \ldots, n \]

Will the method still work?
Only if all pairs $Y_i, X_i$ are exactly on the straight line!

Consider an example

Data on selling price (in 1000$) ($Y$) and living area (square foot) ($X$) of 14 houses in San Diego

The graph is called a scatterplot.
The observations are not on a straight line!

Reason

1. Relation is nonlinear. For house prices we could fit a polynomial of degree 13. Is that a good idea?
2. Other variables are important for selling price, e.g. type of house, location, no. of bedrooms, no. of bathrooms etc.

Now assume that you have an unlimited budget to study the selling price of houses and also an unlimited capacity to measure variables. Then you may be able to measure all relevant variables that affect the selling price. Let these variables be

\[ X_1, X_2, \ldots, X_K \]
Because these are all relevant variables we have the exact relation

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K \]

Hence if we have \( n \) observations then for \( i = 1, \ldots, n \)

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} \]

In the scatterplot we had only \( X_1 \) as independent variable.

Because we do not have \( X_2, \ldots, X_K \) we can rewrite the exact relation as

\[ Y_i = \beta_0 + \beta_1 X_{1i} + u_i \]

with

\[ u_i = \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} \]
Hence $u$ captures the effect of all omitted variables in the exact relation.

In the graph $u_i$ is the deviation from the straight line for observation $i$

We call $u_i$ the $i$-th residual. This is defined relative to some straight line.

$$u_i = Y_i - \alpha - \beta X_i$$

If we think that the deviations from the straight line are due to omitted variables, it makes sense to find the coefficients in the relation between $Y$ and $X$ (or $X_1$).

How do we do this?

The choice will be to compute the coefficients that best fit the data, i.e. that gives the smallest residuals.

Obvious that we can not make all residuals small at the same time. We need some criterion.
We choose

\[ S(\alpha, \beta) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2 \]

Note

1. Squares treat negative/positive residuals symmetrically.
2. Large residuals get relatively more weight.