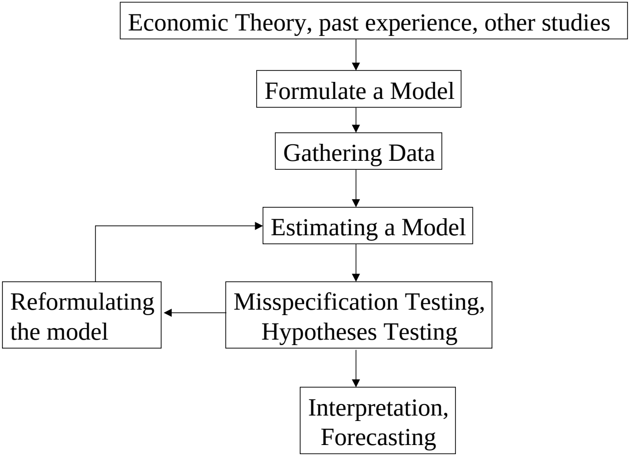
The Nature of Econometrics and Economic Data

[Wooldridge (2013) Chapter 1 and Chapter 2 (sections 2.1 and 2.2)]

- Major uses of Econometrics
- Basic Ingredients of an empirical project
- Formulate a model (example)
- The Question of Causality
- Misspecification Testing
- Types of Data
- The Simple Regression Model
    - Introduction
    - Ordinary Least Squares (OLS)
    - Deriving OLS Estimates
    - Alternative approach to derivation
    - Some definitions

1. Describing Economic Reality.
2. Testing hypotheses about Economic Theory.
3. Forecast future economic activity.

Flow chart for the Steps of an Empirical Study

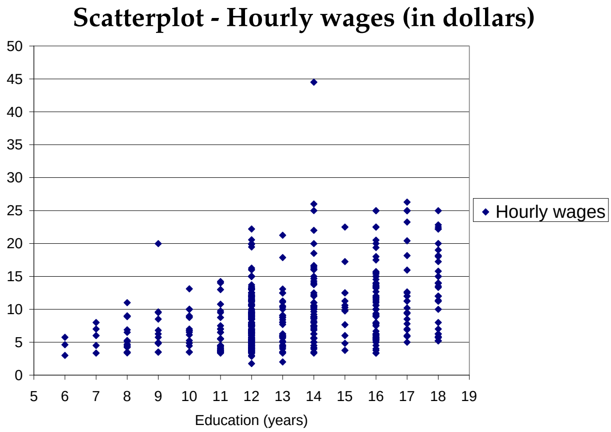


Remark: This module is not about Economic Theory and gathering data.

Economic Theory suggests interesting relations between variables.

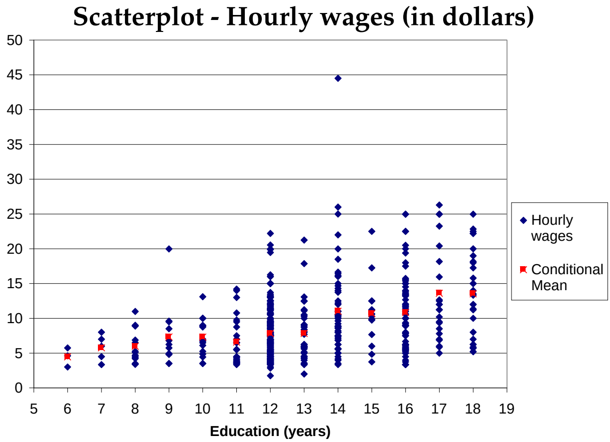
Example: Returns to education

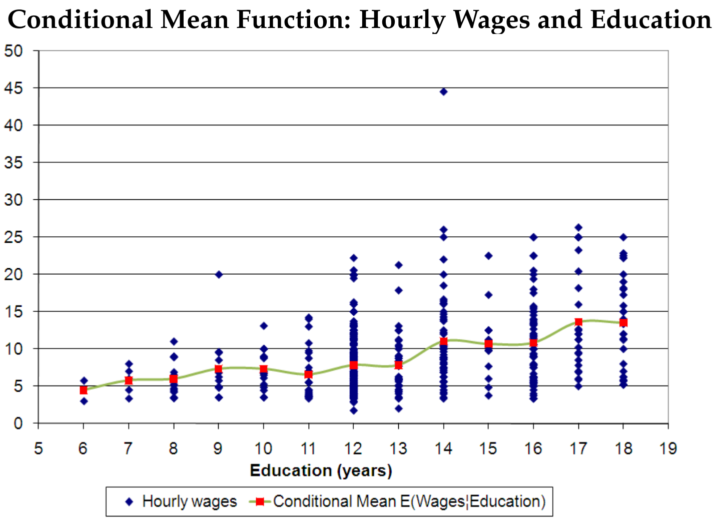
- A model of human capital investment predicts that getting more education should lead to higher wages.
- However, let us look at a data set: US national survey of people in the labour force that already completed their education, 528 people.



- People with the same years of education earn different hourly wages.
- There is a distribution for the hourly wages conditional on the years of education.

- How can we study if the evidence of the data supports Economic Theory?
- A possibility is to look at means of wages conditional on the years of Education.





We can see that the mean of wages vary with the years of Education.

- Hence, the object that we are interested in studying is the mean of wages given the years of Education: $E[Wages|Education]$.
- To simplify computations and the interpretation of results usually we assume a model for $E[Wages|Education]$.
- A possible model for $E[Wages|Education]$ is

$$E[Wages|Education] = \beta_0 + \beta_1 Education.$$

- Notice that for any value a

$$\beta_1 = E[Wages|Education = a + 1] - E[Wages|Education = a].$$

Hence, β_1 is the change of the expected value of *Wages* for one additional year of Education.

- Equivalently, the model can be written in the more familiar way

$$Wages = \beta_0 + \beta_1 Education + u,$$

  where $E\left[u|Education\right] = 0$.
- u is denoted the *error term*.
- This model is known as *The Simple Regression Model*.
- It is linear in the parameters β_0 and β_1.

The estimate of β_1, is the return to education, but can it be considered causal?

- We would like to prove that the effect is causal.
- However it is impossible to prove causality. If $\beta_1 \neq 0$ and we have a sound theoretical economic argument, this might indicate that there is a causal relation. However this is far from being a proof.

Major challenges:

- Inference procedures depend of the characteristics of the distribution of u given *Education*.
- The model
$$Wages = \beta_0 + \beta_1 Education + u$$
might be misspecified.
- Confounding Effects (omitted factors).
- Endogeneity.

David Hendry's *3 Golden rules of Econometrics*:

1. Test.
2. Test.
3. Test.

- Cross Sectional.
- Time Series.
- Panel

- Cross-sectional data is usually a random sample.
- Each observation is a new individual, household, firm, etc.. with information at a point in time.
- **Examples:** Data on expenditures, income, hours of work, household composition, assets, investments, employment, etc..
- If the data is not a random sample, we have a sample-selection problem.

A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
⋮	⋮	⋮	⋮	⋮	⋮
525	11.56	16	5	0	1
526	3.50	14	5	1	0

- Time series data has a separate observation for each time period.
- Typically Macroeconomic measures: GDP, Inflation, Prices, Exchange Rates, Interest Rates, etc..
- Financial data: Stock Prices, Bonds and other financial instruments at frequencies that range from minute to minute up to annual (useful to analyse financial markets).
- Since not a random sample, different problems to consider.
- Trends and seasonality will be important.

Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	unemp	gnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
⋮	⋮	⋮	⋮	⋮	⋮
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

- Can follow the same random individual observations over time – known as panel data or longitudinal data.
- Used to study dynamic aspects of household and firm behaviour and to measure the impact of variables that vary predominantly over time.

A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
⋮	⋮	⋮	⋮	⋮	⋮	⋮
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

$$E\left[y|x\right] = \beta_0 + \beta_1 x$$

or equivalently

$$\begin{aligned} y &= \beta_0 + \beta_1 x + u, \\ E\left[u|x\right] &= 0. \end{aligned}$$

In the model:

- β_0 is known as the *intercept parameter* or *constant term*.
- β_1 is known as the *slope parameter*.

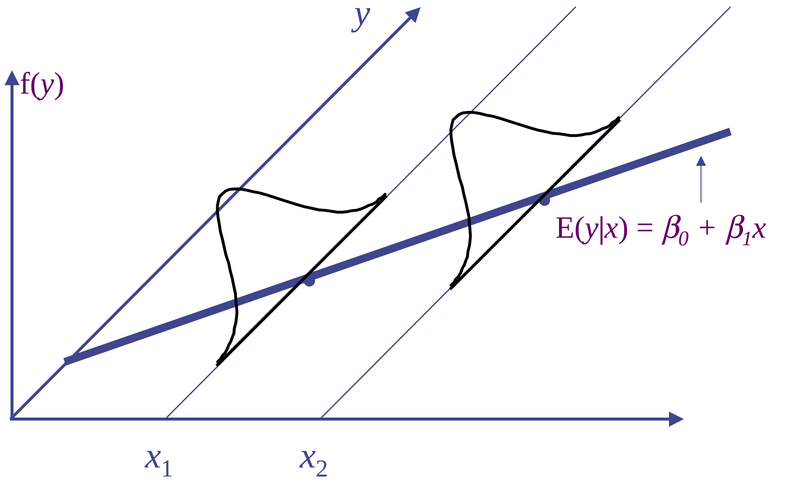
Terminology for Simple Regression

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

$$\begin{aligned} y &= \beta_0 + \beta_1 x + u, \\ E[u|x] &= 0. \end{aligned}$$

- $\beta_0 + \beta_1 x$ is the *systematic part* of y.
- u, the error term, is the *unsystematic part* of y.

$E(y|x)$ as a linear function of x, where for any x the distribution of y is centered about $E(y|x)$.



Basic idea of regression is to estimate the population parameters from a sample.

- Let $\{(x_i, y_i) : i = 1, \ldots, n\}$ denote a random sample of size n from the population.
- For each observation in this sample, it will be the case that

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

- To derive the OLS estimates we need to realize that our main assumption of $E(u|x) = 0$ also implies that

$$\begin{aligned} E(u) &= 0, \\ Cov(x,u) &= E(xu) = 0. \end{aligned}$$

- Why? Because of the *Law of Iterated Expectations*.

We can write our 2 restrictions just in terms of x, y, β_0 and β_1 , since $u = y - \beta_0 - \beta_1 x$:

$$\begin{aligned} E(y - \beta_0 - \beta_1 x) &= 0, \\ E[x(y - \beta_0 - \beta_1 x)] &= 0. \end{aligned}$$

These are called *moment restrictions*.

- We use the *Method of moments* to propose an estimator for the parameters β_0 and β_1. The moment restrictions

$$\begin{aligned} E(y-\beta_0-\beta_1 x) &= 0, \\ E[x(y-\beta_0-\beta_1 x)] &= 0, \end{aligned}$$

  correspond to population means of random variables. Hence the estimator suggested by the Method of Moments is obtained if we replace population means by sample means.
- What does this mean? Recall that for $E(X)$, the mean of a population distribution, a sample estimator of $E(X)$ is simply the arithmetic mean of the sample $\bar{X} = \sum_{i=1}^{n} X_i/n$.

- The moment restrictions in the population:

$$E(y - \beta_0 - \beta_1 x) = 0,$$
$$E[x(y - \beta_0 - \beta_1 x)] = 0.$$

- We want to choose values of the parameters that will ensure that the sample versions of our moment restrictions are true. The sample versions are as follows:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$
$$\frac{1}{n}\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

The OLS estimator is given by the pair $(\hat{\beta}_0, \hat{\beta}_1)$ that solves these equations.

The solution of this system of equations is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, and it is assumed that $\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$.

- There is an alternative justification for this estimator that justifies its name.
- This estimator is known as *Ordinary Least Squares* estimator because it is fitting a line through the sample points such that the mean of squared residuals is as small as possible.
- Consider the function

$$S = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i\right)^2 .$$

- This function takes its minimum when $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$.
- To see this notice that by using calculus to solve the minimization problem for the two parameters you obtain the following first order conditions:

$$\begin{cases} \frac{\partial S}{\partial b_0} = -\frac{2}{n} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0 \\ \frac{\partial S}{\partial b_1} = -\frac{2}{n} \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0 \end{cases} .$$

- These conditions are the same as we obtained before, multiplied by -2. Hence the solution is the same: $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$

- The the *fitted values* are defined as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

  $i = 1, ..., n.$

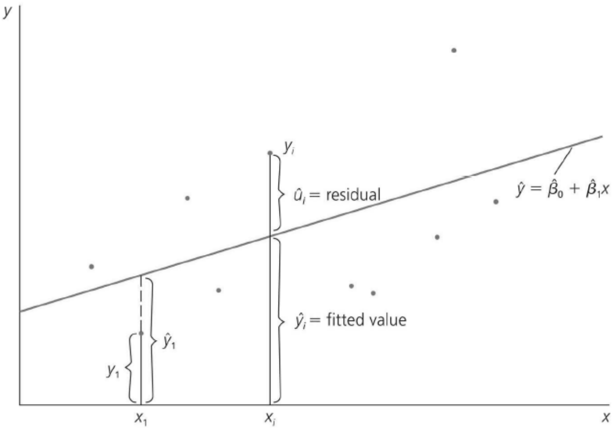
- The *residual*, $\hat{u}_i$ is the difference between the sample point and the fitted line (sample regression function)

$$\begin{aligned} \hat{u}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= y_i - \hat{y}_i, \end{aligned}$$

  $i = 1, ..., n.$

Sample regression line (fitted values), sample data points and the associated estimated error terms:

Fitted values and residuals.



Note the differences:

- Population regression line

$$E\left[y_i|x_i\right] = \beta_0 + \beta_1 x_i$$

  $i = 1, ..., n.$

- Sample regression line (fitted values)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

  $i = 1, ..., n.$

A Note on Terminology: Often we indicate that the equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

$i = 1, ..., n$, was obtained by OLS by saying that we run a regression of y on x, or that we regress y on x.

Example:

- Regression of Wages on Education

Dependent valiable: Wages

Estimation Method: Ordinary Least Squares

Sample size: 528

Regressors	Estimates
Intercept	-1.60468
Education	0.81395

Hence the fitted values are equal to

$$\widehat{Wages} = -1.60468 + 0.81395 \times Education.$$

Interpretation:

- This means that one extra year of schooling increases the average hourly wages by \$0.81395.
- The results should be interpreted with caution as the intercept of -1.60468 means that the average hourly wages of people with no education is -1.60468 which does not make sense. In the sample we do not have people with less than 6 years of education and in this case $\widehat{Wages} = -1.60468 + 0.81395 \times 6 = 3.279$.

Scatterplot and the sample regression line - Hourly wages (in dollars)