Chapter 1. The Nature of Econometrics and Economic Data

1. What is econometrics?

- Econometrics is set of tools to **analyze data** by economists or other social scientists
- We can use econometrics to **estimate economic relationships, test economic theories, evaluate policy and forecasting**
- In contrast to natural sciences, we usually have access to **non-experimental data**

2. Steps in empirical analysis

- **Step 1.** Carefully pose a question
- **Step 2.** Specify economic or conceptual model
- **Step 3.** Turn economic model into **econometric** model
- **Step 4.** Collect data on variables and use statistical methods to **estimate parameters, construct confidence intervals and test hypotheses**

ex) to study effect of **job training** on **worker productivity** (measured by hourly wage)

wage = **f(**educ, exper, ***training)***

Issues

1. How should we **measure** these economic variables?
2. What is **exact functional relationship** among variables?
3. How do we account for **unobserved factors** that make relationships among variables inexact?

Econometric Model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 training + u$$

Error term: represents all other factors that can have effects on y (wage)

Parameters **Variable of interest**

3. Structure of economic data

- **Cross-sectional data:** are collected on certain units **at given point in time**
- We assume that cross-sectional data set represents a **random sample,** that is, each unit in population has **same chance of appearing** in sample and draws are **statistically independent** of one another
- Random sampling generates observations that are **independent and identically distributed (i.i.d.)**
- Random sample is representative of **population of interest** and gives us **the best chance of learning about population**

4. Causality and notion of ceteris paribus

- Concept of **causality** is **key** in econometrics
- To establish causality, notion of **ceteris paribus** (all relevant factors equal) is crucial

Chapter 2: Simple Regression Model $y = \beta_0 + \beta_1 x + u, E\left(u|x\right) = 0$

1. **Definition of simple regression model**

how **y** varies with changes in **x**

Issues:
1. How do we allow factors **other than** x to affect y? Usually there is **no exact** relationship between variables
2. What is **functional relationship between y and x?**
3. How can we capture **ceteris paribus relationship** between y and x?

$$y = \beta_0 + \beta_1 x + u$$

Issue #1: This equation explicitly allows for other factors by ***u*** to affect **y**
Issue #2: ***y*** is assumed to be **linearly** related to ***x***
Issue #3: all other factors accepting **y** are in lumped into **u**

We want to know how y changes when x changes **holding u fixed.**

x and **u** are viewed as **random variables** having distributions in the population.

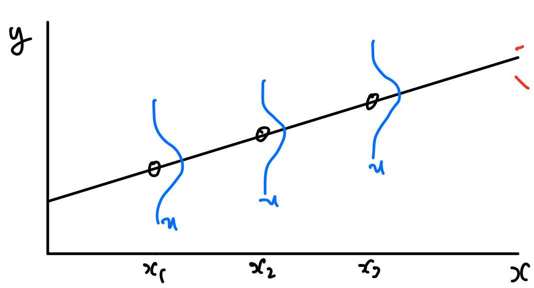
We wish to estimate parameters B0 and B1 by random sample of y and x. But we **never observe u.**
So we **restrict the way how u and x are related in the population.**

First assumption: E(u) = 0
- First, we assume the expected value of **u** is zero in population.

Second "key" assumption: E(u|x) = E(u) for each value of x
- i.e. conditional expectation of u given x does **not** depend on x
- We say u is **mean independent** of x
- This assumption is reasonable if x is assigned at random

Combining E(u) = 0 and E(u|x) = E(u) gives **E(u|x) = 0 for all values of x**

Population regression and distribution of y given x



$E\left(y|x\right) = \beta_0 + \beta_1 x$

$y = \beta_0 + \beta_1 x + u$

Population regression function

where $E(u|x) = 0$

2. Deriving ordinary least squares estimates

To estimate **B0** and **B1**, we use two conditions:

E(u) = 0 and E(u|x) = 0 gives **E(u)** and **E(xu) = 0** by Law of Iterated Expectation

y = B0 + B1x + u **(Population Model)**
u = y - B0 - B1x

E(u) = 0 becomes **E(u) = E[y - B0 - B1x] = 0**
E(xu) = 0 becomes **E(u) = E[x-(y- B0 - B1x)] = 0**

Since we do **not observe all x and y in the population,**
we **estimate expected value E( ) by sample average**

$\frac{1}{n}\sum_{i=1}^{n}$ to estimate $\widehat{\beta}_0$ $\widehat{\beta}_1$ by a random sample from population (called method of moments)

so,

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i\right)=0$$

$$\frac{1}{n}\sum_{i=1}^{n}x_i\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i\right)=0$$

two equations, two unknown

Called OLS estimates

$$\widehat{\beta}_0=\overline{y}-\widehat{\beta}_1\overline{x}$$

$$\widehat{\beta}_1=\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}=\frac{SampleCovariance\left(x_i,y_i\right)}{SampleVariance\left(x_i\right)}$$

we get **OLS regression line** as linear function of x

$$\widehat{y}=\widehat{\beta}_0+\widehat{\beta}_1x \quad\longrightarrow\quad \Delta\widehat{y}=\widehat{\beta}_1\Delta x$$

one unit change in x changes **predicted y** by **beta_hat_{1}**
estimate of "average" y

Also,

$$y_i=\widehat{\beta}_0+\widehat{\beta}_1x_i+\widehat{u}_i$$

where **residual** is defined as

$$\widehat{u}_i=y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i \quad \text{for } i=1,\ldots,n$$

B0 and B1 are called OLS estimated because **sum of squared residuals** are **minimized**

$$\sum_{i=1}^{n}\widehat{u}_i^2=\sum_{i=1}^{n}\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i\right)^2=SSR$$

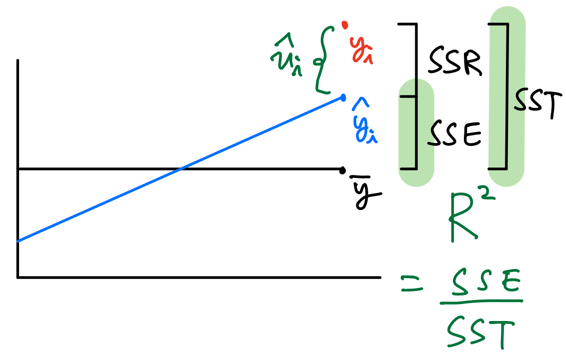
3. Properties of OLS on any sample from population

Goodness-of-fit

For each observation, write

$$y_i = \widehat{y}_i + \widehat{u}_i$$

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{u}_i$$



$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$SSE = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$$

$$SSR = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \widehat{u}_i^2$$

using the fact $\sum_{i=1}^{n} \widehat{y}_i \widehat{u}_i = 0$ we get

$$SST = SSE + SSR$$

Define **R-square** of regression: ***fraction of total variation in*** ***yi*** ***that is explained by*** ***xi***

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

ex) years of education explains about **16.5% (R^2 = 0.1648)** of variation in hourly wage

- It always holds 0 <= R^2 <= 1 by construction
- R^2 = 0 means **no linear relationship** between yi and xi
- R^1 = 1 means a **perfect linear relationship**
- As R^2 increases, yi's gets closer to OLS regression line

- R^2 is a prediction -> purely fit!
- R^2 does not explain causal relation
- Little R^2 may also have significant causal relation

4. Units of measurement and functional form

“Units of measurement”

- Change in unit of **x**: $y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{u}_i$
- Change in unit of **y**: $y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{u}_i$

R^2 does not change

5. **Expected value and variance of OLS**

Assumptions of SLR (Simple Linear Regression)

- **SLR.1 (Linear in parameters)**

  Population model is **y = B0 + B1 x + u** where B0 and B1 are (unknown) parameters, and x, u, y are random variables

- **SLR.2 (Random sampling)**

  We have **a random sample of size *n*,** *{(yixi)* : *i* = 1, …, n}, following the population model

- **SLR.3 (Sample variation in** xi**)**

  Sample outcomes on xi are **not all the same value**

- **SLR.4 (Zero conditional mean)**

  In population, **error term *u* has zero mean** given any value of regressor, **E(u|x) = 0 for all x**
  Key for showing unbiasedness of OLS estimator

Theorem (Unbiasedness of OLS) (i.e. estimates are equal to population values on average)

Under Assumptions **SLR.1–4** and conditional on **X**,

$$E\left(\widehat{\beta}_1\right) = \beta_1$$
$$E\left(\widehat{\beta}_0\right) = \beta_0$$

Remember B1 is **fixed constant** in population.
Estimator B_hat_{1} **varies across samples** and is the random outcome.

We **never know** whether we are close to the population value,
but we hope that our sample is 'typical' and produces slope estimate close to B1, but again **we never know**

Unbiasedness is a property of **estimator** (not estimate)

Key is **SLR. 4 E(u|x) = 0**
If it fails, then OLS will be biased for B1.
For example, if some omitted factor contained in **u is correlated with x**, then SLR. 4 will typically fail

- **SLR.5 (Homoskedasticity, or Constant variance)**

  Error term has same variance given any value of regressor x
  for all x, where sigma^2 is unknown

  $$Var\left(u|x\right) = \sigma^2 > 0$$

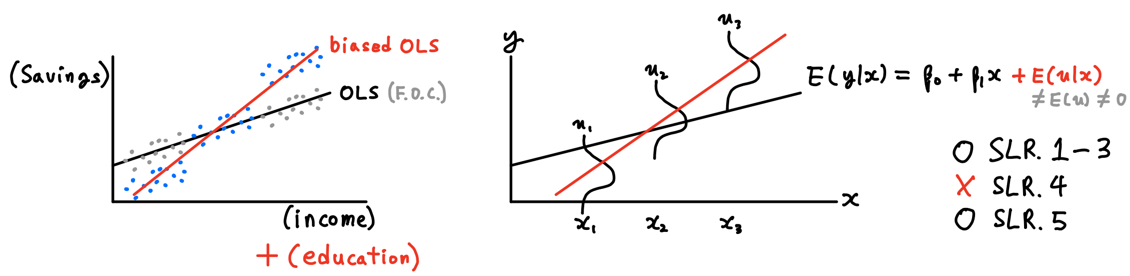
Theorem (Sampling variance of OLS) (reliability of the Standard Error of B_hat_{1})

Under **SLR.1–5**

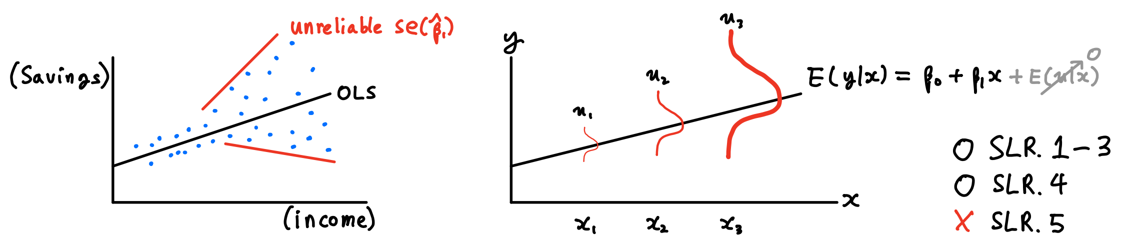
$$Var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{SST_x} \approx \frac{\sigma^2}{n\sigma_x^2}$$

1. As error variance sigma^2 increases, so does variance of B_hat_{1}
   - The more “noise” in the relationship between y and x, the harder it is to learn about B1
2. By contrast, more variation in {xi} is a **good** thing (As SSTx goes up, Variance of B_hat_{1} goes down)
3. Variance of B_hat_{1} **shrinks at 1/n rate**
   - **More data shrinks sampling variance of** B_hat_{1}

Endogeneity



Heteroskedasticity



Chapter 3. Multiple Regression: Estimation

1. Motivation for multiple regression

Multiple linear regression model is written as

$$y = \beta_0 + \beta_1 x + \ldots + \beta_k x_k + u$$

where (k + 1) unknown parameters in total.

Key assumption is

$$E\left(u | x_1, \ldots, x_k\right) = 0$$

2. Mechanics and interpretation of OLS

OLS regression line is written as

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x + \ldots + \widehat{\beta}_k x_k$$

Slope coefficients now explicitly have **ceteris paribus** interpretation, and they are called **partial effects**

Goodness-of-fit

It holds **0 <= R^2 <= 1**, but using the same dependent variable,
R^2 never falls when another regressor is added to regression.

One way to overcome this problem is to use **adjusted R^2**

$$\overline{R} = 1 - \frac{\left[SSR / \left(n - k - 1\right)\right]}{\left[SST / \left(n - 1\right)\right]}$$

Goodness-of-fit of different multiple regression models can be compared by adjusted R^2

3. Expected value of OLS estimators

- **Assumtpion MLR.1 (Linear in parameters)**

  In population, it holds

  y = B0 + B1 x1 + … + Bk xk + u   where Bj's are parameters and u is error term

  y and xj's can be nonlinear functions. Bi must be linear.

- **Assumption MLR.2 (Random Sampling) from the population**

  We have **a random sample** {(yi, xi1, …, xi) : i = 1, …, n} of size n from population
  They are representative sample from population (i.i.d.)

- **Assumption MLR.3 (No perfect collinearity) in the sample**

  None of regressor is **constant**, and there are **no exact "linear" relationships** among them

- **Assumption MLR.4 (Zero conditional mean, or Exogeneity)**

  E(u|x1, …, xk) = 0 for all (x1, …, xk)

Under **MLR.1–3,** we can compute OLS estimates.
Under **MLR.1–4**, OLS estimators are **unbiased,**

$$E\left(\widehat{\beta}_j\right) = \beta_j$$

Including irrelevant variables (regressors with zero coefficients) do **not** cause bias in any coefficients.
In other words, overspecifying the model cause no bias.

Omitted variable bias (OVB)
Leaving a variable out when it should be included in multiple regression is **serious problem.**

$$E\left(\widetilde{\beta}_1\right) = \beta_1 + \underline{\beta_2}\,\underline{\widetilde{\delta}_1}$$

$$= \beta_1 + \underbrace{(+,-)}_{\text{Correlation w/ } y}\underbrace{(+,-)}_{\text{Correlation w/ } x_1}$$

$$Bias\left(\widetilde{\beta}_1\right) = \beta_2\widetilde{\delta}_1$$

4. Variance of OLS estimators

- **Assumption MLR.5 (Homoskedasticity, or constant variance)**

  Variance of u does not change with any of x1, … xk
  Var(u|x1, …, xk) = Var(u) = sigma^2

Under **MLR.1–5,** and conditional on **X,**

$$Var\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{SST_i\left(1 - R_i^2\right)}$$

As error variance sigma^2 decreases, Variance of B_hat_{j} decreases.
One way to reduce error variance is to take more stuff out of the error, i.e. **add more regressors**

We can increase SSTj by **increasing sample size**

Rj^2 measures how linearly related xj is to other regressors.

- If xj is unrelated to all other regressors, it is easier to estimate It's ceteris paribus effect on y.
- Loosely, Rj^2 "close" to one is called the "problem" of **multicollinearity**, which causes the variance of B_hat_j to go large.
- Large Rj^2 can be offset by large SSTj, which grows roughly linearly with sample size *n*.

5. Efficiency of OLS: Gauss-Markov theorem

Gauss-Markov theorem: Under MLR. 1-5, OLS estimator is the **Best Linear Unbiased Estimator (BLUE)**
Here, "best" means "smallest variance"

If we insist on linear unbiased estimators, then we need look no further than OLS.

Chapter 4. Multiple Regression: Inference

1. Sampling distributions of OLS estimators

- **Assumption MLR.6 (Normality)**

Error term *u* is independent of (x1, …, xi) and is **normally distributed with mean zero and variance** σ^2

$u \sim Normal\left(0, \sigma^2\right)$

2. Testing hypotheses about single population

Under Assumptions **MLR. 1-6**

$$\frac{\widehat{\beta}_j - \beta_j}{se\left(\widehat{\beta}_j\right)} \sim t_{n-k-1} = t_{df}$$

t statistic (or t ratio)

To test null hypothesis that xj has no partial effect on y

$H_0 : \beta_j = 0$ we use t **static** (or t **ratio**) $t_{\widehat{\beta}_j} = \frac{\widehat{\beta}_j \ (-0)}{se\left(\widehat{\beta}_j\right)}$

We measure **how far B_hat_{j} is from zero relative to It's standard error**

If B_hat_{j} > 0, the question is, **How big does** $t_{\widehat{\beta}_j} = \frac{\widehat{\beta}_j}{se\left(\widehat{\beta}_j\right)}$ **have to be to conclude H0 is "unlikely"?**

Traditional approach to hypothesis testing
- **(1)** Choose null hypothesis H0 : Bj = 0 (or Bj <= 0)
- **(2)** Choose alternative hypothesis H1 : Bj > 0
- **(3)** Choose **significance level** for test. That is, probability of rejecting H0 when it is in fact true (Type 1 Error).
- **(4)** Choose critical value *c* so that **rejection rule** t_Bj > c lead to 5% level test

We say that **B_hat_1** is **statistically significant (at 1% level) against one-sided alternative.**
Alternatively, statistically insignificant, or **fail to reject H0 at even 10% level**.
In practice, as rule-of-thumb, we use **2** as threshold for significance (t statistics)

Coef. % Std. Errr. = t (as Bj = 0).

Computing p-values for t tests

"It is probability of observing the statistic as extreme as we did if Ho is true."
So smaller p-values provide more evidence against null.

For example, if p-value = .50, then there is 50% chance of observing t as large as we did (in absolute value).
This is not enough evidence against Ho.
If p-value = .001, then chance of seeing t statistic as extreme as we did is .1%.
We can conclude that we got very rare sample or that null. Hypothesis is highly unlikely.

It is most common to report p-values for **two-sided alternatives** (this is what Stata does)

one-sided p-value = two-sided p-value / 2

For t testing against two-sided alternative,

p-value = P(| T | > | t |) where t is the value of t statistic and T is a random variable with t_df distribution

Practical versus statistical significance

- t testing is purely about **statistical significance**
- It does not directly speak to issue of whether variable has practically or economically large effect
- **Practical or economic significance** depends on the size (and sign) of B_hat_j
- **Statistical significance** depends on t_B_hat_j

3. Confidence intervals

Loosely, CI is supposed to give "likely" range of values for corresponding population parameter

We will only consider CIs of the form,

$$\widehat{\beta}_j \pm c \cdot se\left(\widehat{\beta}_j\right)$$

where c > 0 is chosen based on **confidence level**

We will use 95% confidence level, in which case c comes from 97.5 percentile in t_df distribution.
In other words, **c is 5% critical value against two-sided alternative**.

As simple **rule-of-thumb**, for df >= 60, approximate 95% CI is **c = 2**.

Statements like "there is 95% chance that Bj is in interval [.0067, .0661] is incorrect.

What 95% CI means is that **for 95% of random samples that we draw from population, the interval we compute using the rule** $\hat{\beta}_j \pm c \cdot se(\hat{\beta}_j)$ **will include the value Bj.** But for a particular sample, we do not know whether Bj is in the interval. Bj is some fixed value, and it either **is** or **not** in the interval.