Office hours TR 3-4
Midterm on 2/16 in class

In econometrics, in contrast with statistics, we test models.
Types of data:

- Cross section data
- Time series data
- Panel data (Longitudinal)

In this course, we will focus on cross-sectional data
We will assume iid random samples. Independence is a decent example. Identical distributions is not.
How do we do things?

1. ask a question.
2. Find the data
3. Write down the model. (What is on LHS or RHS?)
   - regressant, dependent (LHS); control, explanatory, regressors, independent (RHS)
   - Start from an economic model and work down to a statistical model.
4. Estimate the unknown parameters
5. Test the results, question the assumptions, etc.

Example: Determinants of workers' wages.
Data source: CPS95: 1289 observations of people b/t 18 and 65
Variables: Wage (hourly): mean 12.4
Education (years): mean 13.2
Potential education (age-education-6): mean 18.8
age (years): mean 37.9
Female (1 if female): mean 0.5
Nonwhite (1 if nonwhite): mean 0.2
Union (1 if union member): mean 0.2

Question: is there wage discrimination by any of these variables?

Naive solution:
average for

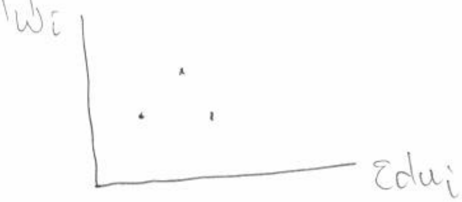
	Men	Women	White	Nonwhite	Union	Non-union
Wage	14.1	10.6				
Education	13.2	13.1				
Experience	19.1	18.5				

$\overline{Wage_{men}} - \overline{Wage_{women}}$ is statistically significant. $t = 8.5$

We need to build a model to examine the determinants of this wage difference.

$$W_i = \beta_1 1 + \beta_2 Edu_i + \beta_3 Ex_i + \beta_4 F_i + \beta_5 NonWh_i + \beta_6 Unio_i + \varepsilon_i$$

Suppose we are looking at $w_i = \beta_1 + \beta_2 Edu_i + \varepsilon_i$



We cannot expect our data to lie on a common line. This is why we introduce ε_i.

We can often take transformations of variables to make the relationship appear linear.

The general linear model is: $Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$

or $\underset{1\times 1}{Y_i} = \underset{1\times k}{X_i} \underset{k\times 1}{\beta} + \underset{1\times 1}{\varepsilon_i}$ where $X_i = [1 \ X_{2i} \ \dots \ X_{ki}]$ $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$

where $i = 1, \dots, n$.

Let $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$, $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$. Then, we have

$$\underset{n\times 1}{Y} = \underset{n\times k}{X} \underset{k\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$$

assumption 1: Linearity $Y = X\beta + \varepsilon$

assumption 2: strict exogeneity $E[\varepsilon_i | X] = 0$

In general, $E[z|w] = f(w)$ for some function.

- On average, ε_i does not depend on <u>any</u> X_i.

Implications: $E[X_j \varepsilon_i] = 0 \ \forall ij$ (X_j is orthogonal to ε_i)

- $E[X_j \varepsilon_i] = E[E[X_j \varepsilon_i | X]] = E[X_j \underbrace{E[\varepsilon_i | X]}_{=0}] = 0$

Since we usually assume $X_{1j} = 1 \ \forall j$, we have $E[\varepsilon_i] = 0 \ \forall i$.

assumption 3.1: spherical disturbances (conditional homoskedasticity)

$Var(\varepsilon_i | X) = \sigma^2$

$\Rightarrow Var(\varepsilon_i) = \sigma^2$

<u>Pf</u>: $Var(\varepsilon_i) = E[\underbrace{Var(\varepsilon_i | X)}_{=\sigma^2}] + Var[\underbrace{E[\varepsilon_i | X]}_{=0}]$

$= E[\sigma^2] = \sigma^2$

prove this equality!

assumption 3.2: $Cov(\varepsilon_i \varepsilon_j | X) = 0 \ \forall ij$. uncorrelated error terms conditional on X.

assumption 4: X is of full column rank. $Rank(X) = k$

Using matrix notation:

1] $Y = X\beta + \varepsilon$

2] $E[\varepsilon | X] = 0$

3] $Var(\varepsilon | X) = \sigma^2 I_n$

4] $Rank(X) = k$

alternatively, we can dispense with mentions of ε.

$\boxed{1'}$ $E[Y_i \mid X] = X_i \beta$ $\quad \forall i$

$\boxed{2'}$ $V(Y_i \mid X) = \sigma^2$ $\quad \forall i$

$\boxed{3'}$ $Cov(Y_i Y_j \mid X) = 0$ $\quad \forall i j$

or

$\boxed{1''}$ $E[Y \mid X] = X \beta$

$\boxed{2''}$ $V(Y \mid X) = \sigma^2 I_n$

$\boxed{4}$ Remains the same as does $\boxed{1}$.