1 Intro to statistics and Econometrics

Empirical analysis economists are forced to use **observational** data

Econometrics use **economic theory** and **economic data**

1.1 Economic data

- cross-sectional data (at one point in time)
- time series data (spans an interval)
- combining both

1.2 Causality

correlation $\neq$ causality

1.3 Probability

- Probability density function (pdf):
$$f_X(x) \tag{1}$$
- Joint distribution
$$f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x) = f_{X|Y}(x|y)f_Y(Y) \tag{2}$$
- conditional distribution
$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \tag{3}$$

1.4 How to summarize a probability distribution

- Mean
$$\mu = \mathbb{E}(X) = \int_{\mathbb{R}} xf(x)dx \tag{4}$$
- Variance
$$\sigma^2 = \mathrm{Var}(x) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}(X^2) - \mu^2 \tag{5}$$
- Standard deviation
$$\sigma = \mathrm{sd}(X) = \sqrt{\sigma^2} \tag{6}$$

1.4.1 Important properties

$$\begin{aligned}
\mathbb{E}(c) &= c && (7)\\
\mathrm{Var}(c) &= 0 && (8)\\
\mathbb{E}(aX+b) &= a\mathbb{E}(X) + b && (9)\\
\mathrm{Var}(aX+b) &= a^2\,\mathrm{Var}(X) && (10)\\
\mathbb{E}(aX+bY) &= a\mathbb{E}(X) + b\mathbb{E}(Y) && (11)\\
\mathrm{Var}(aX+bY) &= a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) - 2ab\,\mathrm{Cov}(X,Y) && (12)
\end{aligned}$$

if X and Y are independant, then $\mathrm{Cov}(X,Y) = 0$.

1.4.2 Correlation

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{sd}(X)\,\mathrm{sd}(Y)} \tag{13}$$

1.5 Standardized Normal distribution $\mathcal{N}(0,1)$

- is symmetric
- total area = 1
- 68% of area is comprised between ± 1
- 95% of area is comprised between ± 2

Each Normal distribution can be converted to the standardized form if

$$X \sim \mathcal{N}(\mu, \sigma^2) \tag{14}$$

then

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1) \tag{15}$$

1.6 Other distributions

1.6.1 χ^2 distribution

Let n independant variables $Z_i \sim \mathcal{N}(0,1)$ then

$$X = \sum_{i=1}^{n} Z_i^2 \tag{16}$$

has χ^2 distribution with n degrees of freedom

$$X \sim \chi^2_n \tag{17}$$

1.6.2 t-distribution

Let $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_n$ then

$$T = \frac{Z}{\sqrt{X/n}} \tag{18}$$

has a t distribution with n degrees of freedom.

1.6.3 F-distribution

Let $X \sim \chi^2_k$ and $Y \sim \chi^2_l$ and assume that X and Y are independant, then

$$F = \frac{X/k}{Y/l} \tag{19}$$

has an F distribution with (k,l) degrees of freedom.

1.7 Estimating the population mean

We wanna know then mean μ of the population

How good is the sample mean as an estimator for the population mean ?

- Point estimator
- Confidence interval

1.7.1 Central Limit Theorem

Let $\{Y_1, ..., Y_n\}$ be a random sample with mean μ and variance σ^2.

Then

$$Z_n = \frac{\sum Y_i/n - \mu}{\sigma/\sqrt{n}} \tag{20}$$

has an asymptotic standard normal distribution

1.7.2 Law of Large Numbers

Let $Y_1, ..., Y_n$ be independant identically distributed (iid) random variables with mean μ. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_i^n Y_i = \mu \tag{21}$$

2 Simple Regression Analysis

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \mathbf{u} \tag{22}$$

where

- y is the dependent variable
- x independent variable
- u error or disturbance term

2.1 Assumptions

- SLR1: Linear parameters
- SLR2: Zero conditional mean

  $$\mathbb{E}(u|x) = \mathbb{E}(u) = 0 \tag{23}$$

  which implies

  $$\mathbb{E}(y|x) = \beta_0 + \beta_1 x \tag{24}$$

2.2 Slope estimate

$$b_1 = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} \tag{25}$$

$$\hat{u}_i = |y_i - \hat{y}_i| \tag{26}$$

2.3 Derive OLS estimators

Minimize for b_0 and b_1

$$\sum_i (\hat{u}_i^2) = \sum_i (y_i - b_0 - b_1 x_i)^2 \tag{27}$$

2.3.1 Properties of OLS

$$\sum_i \hat{u}_i = 0 \tag{28}$$

$$\sum_i x_i \hat{u}_i = 0 \tag{29}$$

$$\overline{y} = b_0 + b_1 \overline{x} \tag{30}$$

2.4 Terminology

- Sum of squares Total

$$SST = \sum_i (y_i - \overline{y})^2 \tag{31}$$

- Sum of squares Exlained

$$SSE = \sum_i (\hat{y}_i - \overline{y})^2 \tag{32}$$

- Sum of squares Residual

$$SSR = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (\hat{u}_i)^2 \tag{33}$$

$$SST = SSE + SSR \tag{34}$$

2.5 Goodness of fit r^2

$$r^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \tag{35}$$

2.6 Functional forms

model	equation	slope	elasticity
level-level	$y = \beta_0 + \beta_1 x + u$	β_1	β_1
level-log	$y = \beta_0 + \beta_1 \ln(x) + u$	β_1/x	β_1
log-level	$\ln(y) = \beta_0 + \beta_1 x + u$	β_1/y	β_1
log-log	$\ln(y) = \beta_0 + \beta_1 \ln(x) + u$	$\beta_1 y/x$	β_1

2.7 Unbiasedness of OLS

$$\begin{aligned} b_1 &= \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} && (36) \\ &= \beta_1 + \sum_i^n (x_i - \overline{x}) \frac{u_i}{SST_x} && (37) \\ \mathbb{E}(b_1) &= \beta_1 && (38) \\ ... & && (39) \\ \mathbb{E}(b_0) &= \beta_0 && (40) \end{aligned}$$

2.8 Homoskedacity

$$\mathrm{Var}(u|x) = \mathrm{Var}(u) = \sigma^2 \tag{41}$$

and we have

$$\mathrm{Var}(b_1) = \frac{1}{SST_x}\sigma^2 \tag{42}$$

2.8.1 Unbiased estimate of the error variance (σ^2)

$$s^2 = \sum_i^n \frac{\hat{u}_i^2}{n-2} = \frac{SSR}{n-2} \tag{43}$$

3 Multiple Regression

$$y = u + \beta_0 + \sum_{i=1}^k \beta_i x_i \tag{44}$$

... allows a *ceteris paribus* interpretation

3.1 Direction of bias

	$\mathrm{Corr}(x_1, x_2) > 0$	$\mathrm{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

3.2 OLS variances

$$\mathrm{Var}(b_j) = \frac{1}{SST_j(1 - r_j^2)}\sigma^2 \tag{45}$$

4 Multiple Regression Analysis - Further Issues

4.1 Tests

we can test either

- two sided
- one sided

We test:
$H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ or build a **confidence interval**.

4.2 The F-test

Test: $H_0 : \beta_1 = 0, \beta_k = 0$, we have

$$F = \frac{SSR_r - SSR_{ur}}{SSR_{ur}} \frac{n-k-1}{q} \tag{46}$$

$F \sim F_{q,n-k-1}$

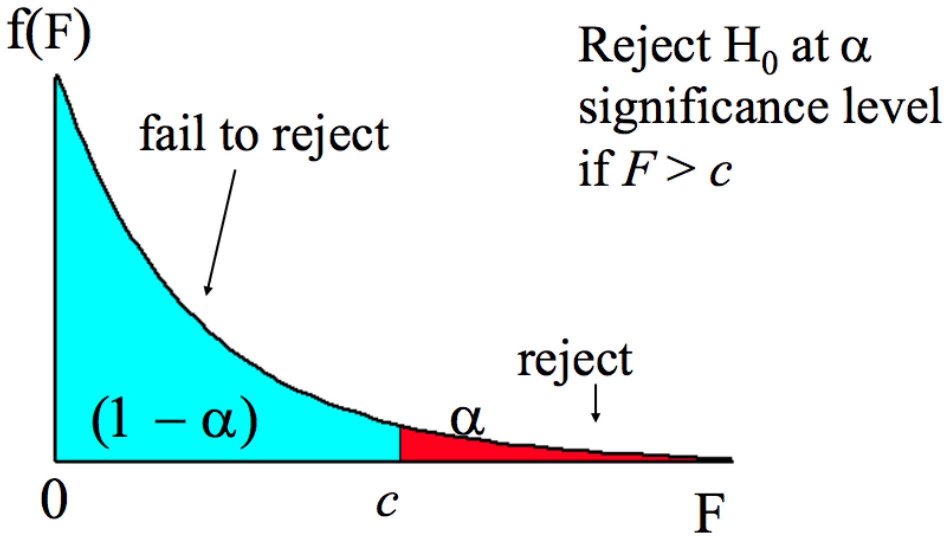

Figure 1: Test F