Short Revision of Econometrics I

José Fajardo

Universitat Pompeu Fabra

1rst October 2018

Outline of Todays' Lecture

- Recuperation and Homework
- Conditional Expectation, Random Sampling and Mean Estimator
- Linear Regression
- Omitted Variables (If there is time)

Recuperation Exam and Homework

To be eligible to take the recuperation exam in January the student must satisfying the three conditions:

1. Not be already approved , i.e. Not satisfy the two conditions A and B.
2. Obtain at least 5 out of 20 points in the continual assessment (homework + participation).
3. Obtain at least 20 points on the December exam (i.e. 20 out of 80 points).

Recuperation Exam and Hommework

To pass in January the student should satisfy two conditions:

- A' Obtain at least 50 points in total (homework and participation during the term will be counted).
- B' Obtain at least 40 points on the January exam (i.e. 40 out of 80 points).

Homework will be due on the dates (all Mondays) indicated in the table below at 11:00. Place your homework in the box in room 20.146. No late homework will be accepted.

Conditional Distribution

- Conditional mean = mean of conditional distribution

$$E(Y|X = x) \quad \text{(important concept and notation)}$$

$$E(Y|X = x) = \sum_{i=1}^{n} y_i P(Y = y_i/X)$$

- Example: $E(MaixmumLoss|Loss < -10\%) =$ (Expected Short-Fall)
- Conditional variance = variance of conditional distribution

Random Sampling

- Population: The group or collection of all possible entities of interest.
- We will assume simple random sampling . Choose an individual (bank, entity) at random from the population
- Prior to sample selection, the value of Y is random because the individual selected is random. Once the individual is selected and the value of Y is observed, then Y is just a number, not random. The data set is $(Y_1, Y_2, \ldots, Y_n)$, where Y_i = value of Y for the i individual (bank, entity) sampled

Random Sampling

- Because individuals 1 and 2 are selected at random, the value of Y_1 has no information content for Y_2. Thus: Y_1 and Y_2 are independently distributed
- Y_1 and Y_2 come from the same distribution, that is, Y_1, Y_2 are identically distributed
- That is, under simple random sampling, Y_1 and Y_2 are independently and identically distributed (i.i.d.).
- More generally, under simple random sampling, $(Y_1, Y_2, \ldots, Y_n)$ are i.i.d.

Distribution of the mean estimator

- Because individuals 1 and 2 are selected at random, the value of Y_1 has no information content for Y_2. Thus: Y_1 and Y_2 are independently distributed
- Y_1 and Y_2 come from the same distribution, that is, Y_1, Y_2 are identically distributed
- That is, under simple random sampling, Y_1 and Y_2 are independently and identically distributed (i.i.d.).
- More generally, under simple random sampling, $(Y_1, Y_2, \ldots, Y_n)$ are i.i.d.

Distribution of the mean estimator

Under the assumption of $(Y_1, Y_2, \ldots, Y_n)$ being i.i.d., with mean μ_Y and variance σ^2_Y, we can obtain important properties for the

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

.

- $E(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \mu_Y$, (Unbiased)
- $Var(\overline{Y}) = Var(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{\sigma^2_Y}{n}$.

$$n \to \infty \Rightarrow \overline{Y} \xrightarrow{P} \mu_Y, \text{ (Consistent)}$$

- Central Limit Theorem

Uniform Random Variables

Generating “good” uniform random variables is technically complex! Use always a well validated generator from a reputable source:

- Matlab
- NAG
- Intel MKL
- AMD ACML
- not MS Excel, C rand function or Numerical Recipes.

Uniform Random Variables

Pseudo-random number generators use a deterministic (i.e. repeatable) algorithm to generate a sequence of (apparently) random numbers on $(0, 1)$ interval. What defines a good generator?

- A long period, how long it takes before the sequence repeats itself? 2^{32} is not enough, need at least 2^{40}.
- Various statistical tests to measure randomness.Well validated software will have gone through these checks.

Uniform Random Variables

Multiplicative congruential algorithms based on:

$$n_i = (a \times n_{i-1}) \mod m$$

- choice of integers a and m is crucial
- (0,1) random number given by n_i/m.
- typical period is 2^{57}, a bit smaller than m
- can skip-ahead 2^k places at low cost by repeatedly squaring a, $\mod m$

Pseudo-Random Sequence

$\gg a = 5;$
$\gg c = 3;$
$\gg m = 16;$
$\gg seed = 7;$
$\gg N = 20;$
$\gg [Useq, Zseq] = LCG(a,\ c,\ m,\ seed,\ N);$

It is not a random sequence.we call it pseudo-random sequence. It starts from Z_0 (the *seed* of the sequence). And will repeat after 16 steps. But, we can repeat in less ex.
$a = 11,\ c = 5,\ m = 16$ and $Z_0 = 3$. We repeat in half the maximal period..

It is important to make it very large.The proper choice of a and m ensures that the sequence looks random.

Pseudo-random Sequence

But, it is not the main problem. Samples should also look independent. In order to pass statistical test: “Sequence of independent samples from an uniform distribution”.
For that reason designing a good RNG is a hard task. But, good news rand is a good RNG!

However, it is important to tell Matlab rand('state',0). Since, any time you start Matlab and type rand, you obtain the same number. Then, it will reset this state vector.

Linear Regression with One Regressor

Linear Regression with One Regressor

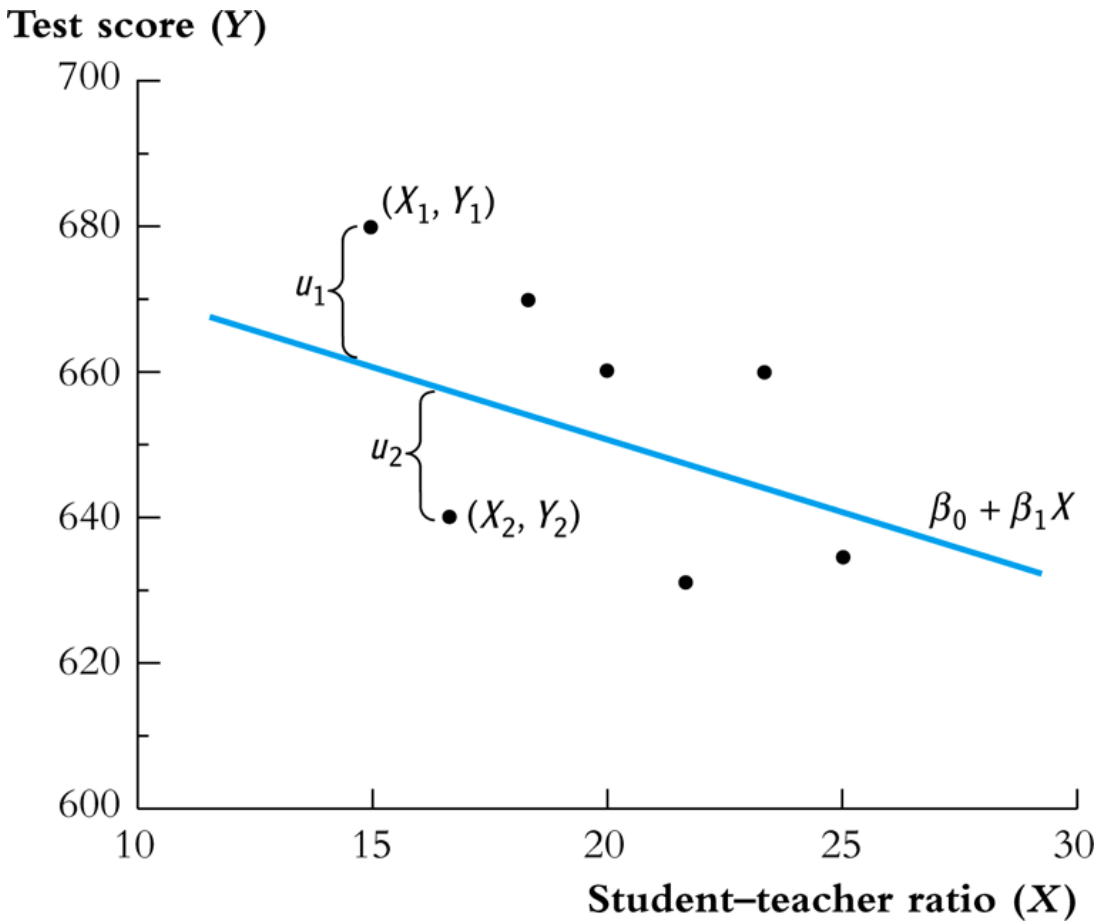
- First, we establish the empirical question (Test Score vs. Class Size, CEOs Salary vs Firm Performance, etc.)
- Then, we propose as a candidate to answer the above question, a linear relationship, i.e. $\hat{Y} = E(Y/X) = \beta_0 + \beta_1 X$.
- We do not know the population parameters β_0 and β_1.We estimate using available data!
- And, of course with suitable Assumptions.

The Population Regresion Model

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \ i = 1, \ldots, n$$

- We have n observations, $(X_i, Y_i), \ i = 1, \ldots, n$.
- X is the independent variable or regressor
- Y is the dependent variable
- β_0 = intercept
- β_1 = slope
- u_i = the regression error
- The regression error consists of omitted factors. In general, these omitted factors are other factors that influence Y, other than the variable X. The regression error also includes error in the measurement of Y.

Why Linear is a Good Candidate?



Ordinary Least Square Estimator

- How can we estimate β_0 and β_1?
- The OLS estimator solves:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

- The solution is obtained using F.O.C.

The OLS Estimator, Predicted Values, and Residuals

KEY CONCEPT 4.2

The OLS estimators of the slope β_1 and the intercept β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4.8}$$

The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \; i = 1, \ldots, n \tag{4.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \; i = 1, \ldots, n. \tag{4.10}$$

The estimated intercept $(\hat{\beta}_0)$, slope $(\hat{\beta}_1)$, and residual $(\hat{u}_i)$ are computed from a sample of n observations of X_i and Y_i, $i = 1, \ldots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

OLS Regression: STATA output

```
regress testscr str, robust
Regression with robust standard errors                Number of obs =     420
                                                      F(1, 418)     =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
------------------------------------------------------------------------------
             |               Robust
    testscr  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |   698.933   10.36436     67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

$$TestScore = 698.9 - 2.28 \times STR$$

Measures of Fit

Two regression statistics provide complementary measures of how well the regression line "fits" or explains the data:

- The regression R2 measures the fraction of the variance of Y that is explained by X; it is unitless and ranges between zero (no fit) and one (perfect fit)
- The standard error of the regression (SER) measures the magnitude of a typical regression residual in the units of Y.

Measures of Fit

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{\hat{Y}}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2}$$

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

In our empirical example $R^2 = 0.05, SER = 18.6$.

STR explains only a small fraction of the variation in test scores. Does this make sense? Does this mean the STR is unimportant in a policy sense?

Capital Asset Pricing Model (CAPM)

$$R_i = r_f + \beta_i(R_m - r_f) + \epsilon, \quad i = 1, \ldots, n.$$

Where R_i is the stock return of firm i, r_f is the fix interest rate (risk-free investment) and R_m is the market return (return of a market portfolio).

The Least Square Assumptions

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \ i = 1, \ldots, n$$

1. The conditional distribution of u given X has mean zero, that is, $E(u|X = x) = 0$. This implies that β_1 is unbiased
2. $(X_i, Y_i), \ i = 1, \ldots, n.$ are i.i.d.
   - This is true if (X, Y) are collected by simple random sampling
   - This delivers the sampling distribution of β_0 and β_1.
3. Large outliers in X and/or Y are rare.
   - Technically, X and Y have finite fourth moments
   - Outliers can result in meaningless values of β_1.

The Least Square Assumptions

- Thus, in an ideal randomized controlled experiment, $E(u|X = x) = 0$(that is, LSA 1 holds).
- In actual experiments, or with observational data, we will need to think hard about whether $E(u|X = x) = 0$ holds.
- The main place we will encounter non-i.i.d. sampling is when data are recorded over time for the same entity (panel data and time series data) .
- Assumption 4 (homoskedasticity) is added for the Gauss-Markov Theorem and for the usual OLS variance formulas to be valid. Then, $\hat{\beta}_1$ is BLUE (see appendix 5.2).
- Assumption 6 (normality), is added to round out the classical linear model assumptions.

The six assumptions are used to obtain exact statistical inference.

Distribution Properties of β_1

$$n \to \infty \Rightarrow \hat{\beta}_1 \to N(\beta_1, \frac{\sigma_v^2}{n\sigma_X^4}),$$

where $v_i = (X_i - \mu_X)\mu_i$

Omitted Variable

Omitted variable problem occurs when two conditions are true:

- When the omitted variable is correlated with the included regressor.
- When the omitted variable is determinant of the dependent variable.

Omitted Variable Bias

From the relationship

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Then,

$$\hat{\beta}_1 - \beta_1 \xrightarrow{P} \frac{\sigma_{Xu}}{\sigma_X^2}$$