Live Session - Week 3

Jeffrey Yau

Monday, December 28, 2015

Agenda

1. Conduct an exercise to review Important Concepts
2. An R exercise to discuss some practical issues / tips
3. The lm() function

Part I: Review of Important Concepts (1)

1. Mathematical Formulation
2. Important Assumptions for Classical Linear Model

- Without looking at your book, notes, and the internet(!), take a couple of minutes to write down the functional form and assumptions of the classical linear model. Which one(s) do you think are the important one and which one(s) can be satisfied easily?

Review of Important Concepts (2): The Model

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon_i$$

where $i = 1, 2, \cdots, n$ is the index of individual observations

Review of Important Concepts (3): Model Assumptions:

1. Linearity (in parameters)
    - The conditional expection function is linear in parameter
    - It is not a restrictive assumption
2. No perfect collinearity
    - Informally, the $X's$ are distinct enough
    - More formally, the matrix $X^T X$ is invertible
3. Homoskedasticity (i.e. constant variance)
    - Note that in CLM, all of the statistical assumptions can be embedded (implicitly or explicitly) in the random variable ϵ_i.
    - The variance of ϵ_i is constant: $Var(\epsilon|x_1, \cdots, x_k) = \sigma^2$
4. Zero Conditional Mean
    - $E(\epsilon|x_1, \cdots, x_k) = 0$
    - This assumption is very crucial.
5. Random Sampling
    - In the first half of the course, we focuse on cross-sectional data

Review of Important Concepts (3): Moment Conditions:

Without looking at your book, notes, and the internet(!), write down the two moment conditions for the CLM? Which one is the important one? What are the implications from these moment conditions? What assumptions are needed for these moment conditions to hold?

Review of Important Concepts (3): k+1 Moment Conditions:

$$E(y - \beta_0 - beta_1 x_1 - \cdots beta_k x_k) = 0$$

$$E(x_1(y - \beta_0 - beta_1 x_1 - \cdots beta_k x_k)) = 0 \cdots E(x_k(y - \beta_0 - beta_1 x_1 - \cdots beta_k x_k)) = 0$$

Part II: The lm function in R

- It is a good practice to read the documentation. You don't have to read everything, but at least undestand the usage, the main arguments of the function, the default functional form (as well as embedded assumptions, if any) of the function.
- https://stat.ethz.ch/R-manual/R-patched/library/stats/html/lm.html
- We will use the lm() function a lot in this course, and many of you have probably used it a lot.

Some questions on the lm() function

- Note: Please answer these questions without searching for the answers from the internet.

1. Does the default lm() function estimate a linear regression model with or without the intercept?
2. If your answer to (1) is yes, how do you run a regression without the intercept (i.e. a regression through the origin)?
3. Will the default lm() function generate an error if you include a variable that is perfectly collinear with other variables?
4. Does the default lm() function give equal(or non-equal) weights to each of the observations in the data set (you provide)?

The lm function in R (2)

lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)

- lm is used to fit linear models
- Considerable care is needed when using lm with time series.
- A formula has an implied intercept term. To remove this use either y ~ x - 1 or y ~ 0 + x.
- Non-NULL weights can be used to indicate that different observations have different variances (with the values in weights being inversely proportional to the variances)
- An object of class "lm" is a list containing at least the following components, such as -coefficients -residuals -fitted.values

An Exercise

- We will use the Prestige data set from the car library.
- When estimating regression models, always start with the question of interest?
- Please answer the following questions:

1. Run the following command library(car)
2. Use str(Prestige) and summary(Prestige) to take look at the structure of the data - number of variables, type of the variables, and number of observations
3. We are interested in variables prestige, income, and education. Let's plot some graphs: histograms of each of the variables and scatterplots of each of the pairs of the variables. For example, you can use the following functions

- hist(x)
- plot(x1,x2)

4. Run a set of 4 regressions using the lm() function and try to interpret the results. For example, lm(y ~ x1, data=YourData)

   - Interpret the results (coefficient estimates)

   a. Regress prestige on income; print a summar of the regression.
   b. Regress prestige on education; print a summar of the regression
   c. Regress prestige on both income and education; print a summar of the regression
   d. Create a variable, called it inc2, equal to income*2. Then, run the first regression again: regress prestige on income and inc2. What happens?