Spring, 2019
Hidehiko Ichimura

Review Questions I

In the first half of the spring term, we have studied the linear regression model, the OLS estimator for the coefficients in the linear regression model, and the basics of asymptotic analysis.

The main points I want you to understand are

1. how we can specify the linear regression model using various forms as regressors such as cross terms and dummy variables,
2. what the conditions are, under which the coefficient have a causal interpretation,
3. what the OLS estimator is, when it is well defined, and why the method is conditionally unbiased and consistent at an intuitive level,
4. concrete conditions under which the estimators are consistent and asymptotically normal,
5. how to conduct inferences using the finite sample results and asymptotic results,
6. how to conduct hypothesis tests using the finite sample results and asymptotic results, and
7. what are the concrete empirical situations under which various assumptions do not hold, and
8. what happens to the OLS estimator when the basic assumptions do not hold.

I want to emphasize that there are common set of tools underlying these issues. They are

1. matrix algebra,
2. the concept of conditional expectation operator, and
3. asymptotic analysis.

They continue to play critical roles for instrumental variable method, panel data analysis, maximum likelihood estimation, non-parametric and semiparametric analysis, etc. Tools are explained in class and discussed in the textbook and also in the lecture note. Please make sure to understand them. Once you obtain the tools, the eight issues above become applications of the tools. In turn, understanding the tools require examination of how they can be applied.

Of course concrete empirical issues motivate all these analysis. Without knowing the actual empirical issues, why the theoretical problems examined are interesting will not be understood. Some examples of empirical issues are given in the last part of the questions below.

Basics

Here is a list of questions you may be able to use to review the material covered in the first half of the course. The course covered some more material, but going through the questions will give you an idea about the outline of the course.

1. What is the interpretation of the coefficient in the standard linear regression model? What are the conditions under which this can be done?
2. What is the OLS estimator of the coefficients in the linear regression model?

3. Why is the OLS residual vector orthogonal to each of the regressors?
4. Explain intuitively why the objective function of the OLS estimator identifies the coefficients in the linear model.
5. What does the OLS estimator estimate under misspecification?
6. What are the sufficient conditions for the OLS estimator to be conditionally unbiased?
7. What are the sufficient conditions for the OLS estimator to be consistent?
8. What are the sufficient conditions for the OLS estimator to be asymptotically normal?
9. Assume random sampling and the model

$$y_i = \alpha + x_i'\beta + \varepsilon_i.$$

Under the assumptions that justify the OLS estimator to be unbiased, discuss the relationship between $\alpha + x_i'\beta$ and the conditional mean function of the dependent variable y_i conditional on the regressors x_i.

10. Is the OLS estimator consistent and asymptotically normal under heteroskedasticity? Explain your answer.
11. When there is heteroskedasticity what happens to OLS estimator?
12. When is the conditional variance of the OLS estimator of a coefficient small?
13. What should you do if you think there is heteroskedasticity in the context of a linear regression model?
14. What are the basic assumptions to justifying the OLS estimator to be BLUE?
15. What happens to the OLS estimator when the error term and some of the regressors are correlated?
16. Give five cases in which the error term and some of the regressors are correlated.
17. Explain three different ways to test the null hypothesis: $C\beta = A$ where C is an $r \times K$ matrix of full row rank.
18. Explain the circumstances in which you would use each one of the three tests.
19. Let $\hat{b}$ be an estimator of β and that $\sqrt{n}\left(\hat{b} - \beta\right)$ converge in distribution to a normal random vector with zero mean and variance covariance matrix of V. Let $\hat{V}$ be a consistent estimator of V.
    (a) How will you construct a 95% confidence interval for $c'\beta$ for a given vector c?
    (b) How will you construct a Wald type test of $c'\beta = a$ for a given vector c and a constant value a using 5% significance level?
    (c) How will you construct a 95% confidence region for $C'\beta$ for a given $(r \times k,\ r < k)$ matrix C?
    (d) How will you construct a Wald type test of $C'\beta = A$ for a given $(r \times k,\ r < k)$ matrix C and a constant value vector $(r \times 1)$ A where rows of C are linearly independent using 5% significance level?

- This problem highlights the fact that all the inference problems and hypothesis testing frameworks we use in this course have the same structure. The differences are in the construction of $\hat{b}$ and $\hat{V}$ as well as the meaning of β. This problem also makes clear that if r is two or higher, we lose the sense of direction in testing the null against a particular alternative.
- You need to be able to conduct inferences and hypothesis tests in a specific context where an estimator and c (and a) or C (and A) are given.

20. How do you estimate the variance covariance matrix of an OLS estimator under homoskedasticity?

21. How do you estimate the variance covariance matrix of an OLS estimator under heteroskedasticity?

Basics II

Here are more detailed questions.

Asymptotics

The topics we studied include (1) basics such as various concepts of convergence, $O_P(1)$, $o_p(1)$ concepts (2) LLN, CLT.

Please make sure you have clear ideas on how to use $O_P(1)$, $o_p(1)$ concepts.

1. Explain the difference between different consistency concepts of an estimator.

2. Explain the differences among consistency, unbiasedness, and asymptotic unbiasedness of an estimator.

3. Why is asymptotic normality of an estimator a useful property of an estimator?

4. Prove the following:

   (a) $O_P(1) \cdot o_p(1) = o_p(1)$ always.
   (b) $O_P(1) + O_P(1) = O_P(1)$ always.
   (c) $o_p(1) + o_p(1) = o_p(1)$ always.

5. Assume that for each i, $X_{ni} = o_p(1)$ or $X_{ni} = O_P(1)$ as $n \to \infty$. In view of the last two results above, an induction argument implies that for any finite J, $\sum_{i=1}^{J} X_{ni} = o_p(1)$ if each $X_{ni} = o_p(1)$ and $\sum_{i=1}^{J} X_{ni} = O_p(1)$ if each $X_{ni} = O_p(1)$. Show this.

6. However, when J also goes to infinity as $n \to \infty$, either of the results does not hold. Explain this by constructing an example for each case.

7. If for each i, $X_{ni} = o_p(1)$ as $n \to \infty$, then for any finite J, $\sum_{i=1}^{J} X_{ni}/J = o_p(1)$ as $n \to \infty$, but show that $\sum_{i=1}^{n} X_{ni}/n$ is not necessarily $o_p(1)$ as $n \to \infty$. (This is really the same problem as the problem above for $o_p(1)$. However, stating this way may be slightly more surprising.)

8. Consider the following random coefficient model under random sampling:

$$y_i = x_i' \beta_i$$

   where $x_i = (1, \tilde{x}_i')'$ and $\tilde{x}_i$ and β_i are independent random vectors with finite second moments. Show that when $E(x_i x_i')$ is non-singular, OLS estimator consistently estimates $E(\beta_i)$.

9. Let $\hat{b}_1$ and $\hat{b}_2$ be estimators of $\beta_0 \in R^K$ and that $\sqrt{n}(\hat{b}_j - \beta_0)$ converges in distribution to $N(0, V_j)$ for $j = 1, 2$.

   (a) Derive the asymptotic variances of $\sqrt{n}[g(\beta_j) - g(\beta_0)]$, for a continuously differentiable function g, $g : R^K \to R$ for $j = 1, 2$. (Use the Delta-method covered in Antonio's class.)

   (b) Sometimes an estimator $\hat{b}_1$ is said to be more efficient than $\hat{b}_2$ if $V_2 - V_1$ is positive definite. Use the results in (a) to explain why this is a reasonable way to rank estimators.

10. Show that if X_n converges in distribution to a normal random variable with mean 0 and variance 1 and Z_n converges in probability to a constant number $A \neq 0$, then $X_n \cdot Z_n$ converges in distribution to a normal random variable with mean 0 and variance A^2.

OLS

1. Suppose you consider a linear regression model of y_i on a constant term and a regressor x_i and z_i. Let d_i denote a dummy variable that takes value 1 if the ith person is a female and 0 if the person is a male.

   (a) Formulate a linear regression model that does not allow any difference between male and female observations.

   (b) Formulate a linear regression model that allows for a difference between male and female observations only in the constant term.

   (c) Formulate a linear regression model that allows for differences between male and female observations in the constant term and the coefficient on x_i. Coefficient on z_i is assumed to be the same between male and female.

2. Discuss how you will assess the omitted variable bias of the OLS estimator in the context of the linear regression model.

   - This problem expects an answer in general terms but you should be able to execute the omitted variable bias computation in a specific context.

3. Explain the difference in the interpretation of the β coefficient in the following linear regression model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

   where the conditional expectation of ε_i given x_i is zero.

4. Discuss the difficulties we need to face when there is multicollinearity problem in the linear regression model.

5. Why is the OLS estimator inefficient when there is heteroskedasticity?

6. What is the consequence of omitting a variable not correlated with any of the included regressors?

7. What is the consequence of including a variable whose coefficient in a linear regression model is zero? Discuss this in two cases: including a variable that is not correlated with the error term and including a variable that is correlated with the error term.

8. Suppose there is only one regressor to examine the relationship to a dependent variable. Since linear regression analysis requires specifying a functional form, we might just wish to graphically analyze the data by plotting the dependent variable against the regressor. Critically evaluate this approach in the two different contexts:

(a) the residual given x_i in the linear regression model has zero conditional mean, and

(b) the residual and x_i are correlated.

Applied

1. Suppose we wish to estimate the wheat production function specified below using a random survey sample of wheat farmers in the US:

   $$\log y_i = \alpha + \beta \cdot \log fertilizer_i/acre + \gamma \cdot \log hours\_of\_labor_i/acre + \theta \cdot \log capital_i/acre + \varepsilon_i$$

   where y_i is bushel per acre.

   (a) What does the error term represent? Give at least three different reasons why we may have an error term in this particular context.

   (b) What assumptions do we need to justify using OLS for (1) consistent estimation of the coefficients, and (2) consistent estimation of OLS standard errors?

   (c) In this particular context, discuss if the assumptions can be justified assuming that there is no misspecification in the systematic effects of inputs but that at least a part of the residual term represent the land quality and farmers choose input level knowing their land quality.

   (d) How will you form the confidence interval for the OLS estimator of β?

   (e) Suppose you learned that only about a half of the requested survey was returned. What is a crucial assumption that justifies the approach you discussed above?

2. Suppose we wish to estimate the earnings equation specified below using a random survey sample of workers in a firm:

   $$\log y_i = \alpha + \beta \cdot \log \text{age}_i + \gamma \cdot \log \text{experience}_i + \delta \cdot \log \text{education}_i + \theta \cdot \text{gender}_i + \varepsilon_i$$

   where y_i is earnings per year. Assume that gender variable takes value 1 or 0 depending on whether the ith person is a male or female, respectively. Assume that the error term conditional on regressors have mean zero.

   (a) How will you test if there is no sex discrimination in this firm under homoskedasticity?

   (b) How will you test if there is no sex discrimination in this firm under heteroskedasticity?

   (c) If you suspect that the sex discrimination may take form related with return to age or experience or education. How will you test this? Specify a model you use and explain the test statistic with degrees of freedom.

3. Suppose currently a subsidy is given to the head of a household and we want to measure the changes in the household's consumption behavior for the households with husbands being designated as the head of household when the subsidy is given to the wife. We will measure the changes in their consumption behavior by the changes in the mean of the share of their food expenditure in their total expenditure.

   (a) Discuss the difficulty of using observational data to measure the effect.

   (b) Describe the randomized experiment needed to measure this effect.

(c) Suppose the government does not want to randomly choose households whether to give the subsidy to wives or husbands but willing to randomly choose areas and give every wives or every husbands in the area the subsidy. Explain how this kind of sampling will allow us to measure the effect.

4. Suppose you wish to measure the price elasticity of gas consumption per year. You collect a random sample of households and collect gas price per kWh (kilo watt for an hour) paid (denoted by p) and amount consumed (denoted by y) along with household information and income (denoted by a vector x). Assume that a gas company is chosen by a household. In what follows, for simplicity, assume there is no nonlinear pricing. Assume that there is the following relationship between these variables

$$y = \alpha + \beta \cdot p + x'\theta + \varepsilon$$

where α, β, and θ are unknown constant parameters and ε is an unobserved variable.

(a) State conditions under which OLS estimator of this regression yields consistent estimator of α, β, and θ.

(b) Which of the assumptions you stated in (a) may be affected by the fact a gas company is chosen by a household? Explain why.

5. Suppose you wish to measure the effect of holding different health insurance on total health expenditure (denote it by y). The total health expenditure includes expenditure by a patient as well as that by the insurance company. Assume that the only difference among health insurances is the co-payment ratio (denote it by c); the percentage of expenditure you need to pay out of your own pocket. For example, if the person is not covered by a health insurance, then $c = 1$, and if the person is fully covered $c = 0$. Denote other variables by a vector x. Suppose there is the following relationship among the variables

$$\log y = \alpha + \beta \cdot c + x'\theta + \varepsilon,$$

where ε denotes an unobserved variable.

(a) Discuss how you will estimate parameters α, β, and θ in this model.

(b) Discuss how you will construct the 95% confidence interval of β.

(c) Discuss the difference of the model above and the model which includes cross terms of c and elements in vector x.

6. Two program evaluation problems are described below each with a proposed estimation method. For each evaluation problem, discuss briefly (in one or two lines for each question) what a potential problem with the proposed estimation is and why. For each issue you may find more than one problem with the described method. Choose the one you think is the more important problem. Assume that data are collected as described and that there is no problem regarding the sample size to apply large sample approximations.

(a) One wishes to measure the average effect of attending college on earnings. Using a random sample from US population one estimates the coefficients β_0,...,β_4 in the following regression using the OLS estimator:

$$\text{earnings} = \beta_0 + \beta_1 \cdot \text{college education} + \beta_2 \cdot \text{experience} + \beta_3 \cdot \text{age} + \beta_4 \cdot \text{gender} + \varepsilon,$$

where ε represents an unobserved variable.

(b) Suppose there are only two types of coffee shops, one type of shops sells for \$3 a cup and the other for \$2 a cup. Assume the shops sell just coffee and there is only one size of a cup of coffee. Assume that each shop is a local monopoly. Let p denotes the price variable which takes value 1 if the shop sells \$3 a cup coffee and 0 if the shop sells a \$2 a cup coffee. The effect of the price change is measured as the OLS estimate of the coefficient on the p dummy variable on the following regression model:

$$\text{sales} = \alpha + \beta \cdot p + \varepsilon$$

where ε represents an unobserved variable.

7. Suppose you wish to measure the effect of a de-worming program on school attendance (number of days) in a developing country. The de-worming program randomly selects 100 schools and offer free de-worming medication to all students in those schools. As a comparison group comparable data are collected from randomly selected 100 schools and not given the treatment. To make the problem simpler, assume that whether a student takes the medication or not does not affect the effect of the medication on other students.

   (a) If every student in treated schools takes the medication and every student in untreated schools does not take the medication, how can we estimate the average impact of the medication on school attendance? (If you think this is too easy, you are correct so do not be concerned.)

   (b) If every student in treated schools does not necessarily take the medication and every student in untreated schools does not take the medication, how can we estimate the average impact of the medication on school attendance for those who take the medication?

   (c) Under the same condition as in b., how can we estimate the average impact of the medication on school attendance?

8. Suppose you wish to measure the effect of education on life expectancy. Suppose in 1910 in one area the compulsory schooling increased from 6 years to 7 years. You have census data on those who were raised in this area. You want to use the change in the compulsory schooling law to estimate the effect of education on life expectancy.

   (a) How will you estimate the average effect of a year increase in education from 6 years for those students who would have received just 6 years of education without the compulsory schooling law change?

   (b) What are the assumptions you need to make to justify the method you describe in a?

   (c) What kind of additional information do you need to examine if the assumptions you make in b. are valid?