Midterm Exam Instructions
Introduction to Econometrics, Spring 2017

Prof. Adam Kapor

1. This exam consists of four pages, including this one. Please verify that your copy is complete.

2. Write your answers into the blue book. You should not need more than one. No credit is given for answers that are not in the blue book.

3. You must write legibly to receive credit.

4. No calculator or "cheat sheet" is allowed. You do not need to simplify numerical formulas (e.g. for test statistics or confidence intervals.)

5. You have 75 minutes. The number of points you will receive for correctly answering a particular question is clearly indicated in the exam. The time required to answer each question might differ.

6. No explanation requires more than two sentences. Write down the answer, not everything you know.

7. Some explanations are better than others. If we ask which regression line would you prefer, and you say "the one that fits the data better" that is not as good as "when I compare the values of ..., I see that ..."

Question 1

(short answers, 3 points each)

1. Let $\{X_1, \ldots, X_n\}$ denote independent and identically distributed draws from a population with mean μ. To estimate μ, suppose that you use the following estimator:

$$\hat{\mu} = X_n$$

Is this estimator unbiased for μ? Is this estimator consistent for μ?

answer X_n is unbiased because $E(x_n) = \mu$ but it is not consistent (does not converge in probability to μ).

2. Suppose that you are interested in testing a joint null hypothesis consisting of three restrictions, say $\beta_1 = \beta_2 = \beta_3 = 0$ in multiple regression. Assume that you have three individual t-statistics for $\beta_j = 0$, where $j = 1, 2, 3$. Consider the following testing procedure: reject the joint null hypothesis if at least one of t-statistics exceeds 1.96 in absolute value. If t-statistics are independent of each other, what is the probability of rejecting the joint null hypothesis when it is true? [Hint. $0.95^2 = 0.9025$, $0.95^3 = 0.8574$, $0.95^4 = 0.8145$, $0.95^5 = 0.7738$.]

answer When the null is true, each t-test would fail to reject the null with probability .95. The probability of correctly failing to reject the null under the suggested procedure is $(.95)^3 = 0.8574$. The null is falsely rejected with probability $1 - (.95)^3 = 1 - 0.8574$.

3. Consider the linear regression model with a single regressor

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

Recall that the OLS estimator of β_1 has the form

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

Show that

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

answer Plug in $Y_i = \beta_0 + \beta_1 X_i + u_i$ and $\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u}$ into the first equation, then simplify using simple algebra. Note: You did not need to include all steps. Minute detail is included to aid understanding.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(\beta_0 + \beta_1 X_i + u_i - (\beta_0 + \beta_1 \bar{X} + \bar{u}))}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

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

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

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

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

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

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

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

4. Using the following regression result

$$\ln(\widehat{Earnings}) = 3 + 0.01 Experience,$$

predict the percentage increase in earnings for 3 additional years of experience.

answer In an estimated log linear model of the general form $ln(Y_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i$, the expected effect of increasing X by three units is approximately a $(3)(100)\hat{\beta}_1$ percentage change in Y. Applying that formula to our case, the answer is $(3)(100)(.0) = 3\%$.

5. What is sample selection bias? Explain using an example.

answer Sample selection bias arises from choosing a non-random sample from the population with respect to the dependent variable. That is, if our study is regressing Y on a set of covariates, then sampling from a subset of the population that only has certain values of Y will cause sample selection bias.

Question 2

You are estimating a model of the form

$$y = \beta_0 + \beta_1 x_1 + u.$$

You are concerned that you have omitted a variable x_2 which is correlated with x_1 and is part of u.

1. (10 points) suppose that $E(x_2|x_1) = \alpha x_1$, for some number α. Suppose that the true data-generating process is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + w,$$

where the usual least-squares assumptions are satisfied. What is the bias of the OLS estimate $\hat{\beta}_1$ obtained from the specification $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$?

answer There were many ways to answer this question– from starting with one of the OVB equations to working directly with the equation for bias. Below, I first present the latter solution, then the former. Again, not all steps were needed in the exam to get full credit.

$$Bias = E[\hat{\beta}_1 - \beta_1] = E[\frac{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})u_i}{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})^2}]$$

By law of iterated expectations:

$$= E[E[\frac{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})u_i}{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})^2}|x_1]]$$

$$= E[\frac{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})E[u_i|x_1]}{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})^2}]$$

Noting that in this case, $u_i = \beta_2 x_{2,i} + w_i$,

$$= E[\frac{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})E[\beta_2 x_{2,i} + w_i|x_1]}{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})^2}]$$

$$= E[\frac{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})E[\beta_2 x_{2,i}|x_1]}{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})^2}] + E[\frac{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})E[w_i|x_1]}{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})^2}]$$

$$= \beta_2 E[\frac{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})E[x_{2,i}|x_1]}{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})^2}]$$

$$= \beta_2 E[\frac{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})\alpha x_{1,i}}{\sum_{i=1}^{n}(x_{1,i} - \overline{x_1})^2}]$$

$$= \beta_2 \alpha$$

The second way to solve is simply to realize that this is omitted variable bias for a simple OLS model and use one of the known equations:

$$bias = \beta_2 \frac{Cov(x_1, x_2)}{Var(x_1)}$$

And, since the relationship is linear,

$$\frac{Cov(x_1, x_2)}{Var(x_1)} = \alpha^*$$

so,

$$bias = \beta_2 \alpha$$

*You could either know this, or work it out as follows:

$$\frac{Cov(x_1, x_2)}{Var(x_1)} = \frac{E[(x_1 - E[x_1])(x_2 - E[x_2])]}{E[(x_1 - E[x_1])^2]} = \frac{E[E[(x_1 - E[x_1])(x_2 - E[x_2])|x_1]]}{E[(x_1 - E[x_1])^2]}$$

$$= \frac{E[(x_1 - E[x_1])E[(x_2 - E[x_2])|x_1]]}{E[(x_1 - E[x_1])^2]} = \frac{E[(x_1 - E[x_1])(E[x_2|x_1] - E[E[x_2]|x_1])]}{E[(x_1 - E[x_1])^2]}$$

$$= \frac{E[(x_1 - E[x_1])(\alpha x_1 - \alpha E[x_1])]}{E[(x_1 - E[x_1])^2]} = \alpha$$

2. (5 points) You decide to test for omitted variable bias as follows: first, regress y on x_1 and a constant, and compute the residual $\hat{u}_i = y_i - \hat{y}_i$. Then compute the sample covariance

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{u}_i - \overline{\hat{u}})(x_{i1} - \overline{x_1}),$$

where $\overline{z}$ denotes the sample mean of a random variable z. If the covariance is large, you will interpret this as evidence of failure of least-squares assumption 1. Will this approach work? What value of the covariance will you obtain?

answer The test will not work. The residuals here are only good estimates for errors under the assumption that LSA 1 holds. So, we cannot test LSA1 using something that relies on LSA1. More, specifically, the value of this covariance will always equal 0 by construction.

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{u}_i - \overline{\hat{u}})(x_{i1} - \overline{x_1})$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - (\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x_1}))(x_{i1} - \overline{x_1})$$

Since, by definition, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x_1}$,

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1})(x_{i1} - \overline{x_1})$$

Which is equal to zero by one of the first order conditions of OLS!

Question 3

(30 points total, 5 per item) Suppose that you've been hired by a firm that sells peanut butter and jelly, among other products, and wants to use your econometric expertise to price the goods correctly. To measure demand, the firm has conducted a series of randomized price experiments in supermarkets.

The way that experiments worked was as follows: the firm drew a set of supermarkets at random, and asked permission to conduct an experiment. Some supermarkets did not give permission, but others did, resulting in a sample $i = 1, \ldots, N$ of supermarkets that agreed to participate.

In half of these supermarkets, the prices of peanut butter and jelly were set to p^0_{pb} and p^0_{jelly}, respectively. In the other half, in addition to this marketing effort, the firm adjusted the price of jelly by a random amount $\Delta p_{jelly} < 0$, so that the price was lower than in the "control" supermarkets. In half of these "low jelly price" supermarkets (that is, in a quarter of the original sample) the firm also lowered the price of peanut butter by a random amount Δp_{pb}, so that the price was lower than it would have been absent this treatment.

You have data on prices p_{jelly} and p_{pb} and quantities sold q_{jelly} and q_{pb} at each supermarket $i = 1, \ldots, N$.

1. As you walk into the meeting, two executives are arguing. "Peanut butter and jelly are obviously complements! You need both to make a PB&J," says the first executive. "Nonsense!" replies the other executive. "When I was young we'd eat peanut butter OR jelly sandwiches! The two products are probably completely unrelated." They both turn to look at you. Can you resolve the debate about whether the two goods are complements?

   In particular: (i) what model would you estimate, (ii) what are your null and alternative hypotheses, (iii) what test statistic would you construct, and (iv) under what conditions would you reject the null? Remember that peanut butter and jelly are said to be complements if the derivative of peanut butter consumption with respect to the price of jelly is negative.

   answer (i) $q_{pb} = \beta_0 + \beta_1 p_{pb} + \beta_2 p_{jelly} + u$ (ii) $H_0 : \beta_2 = 0; H_1 : \beta_2 < 0$ (iii) $t = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta})}$ (iv) $pval < \alpha$ or $t < Z_\alpha$. We accepted two-sided tests as well. Also, note that $Z_{.05} = -1.64$.

2. Suppose that you had estimated a linear demand curve for peanut butter as follows:

$$q_{\text{pb}} = \beta_0 + \beta_1 * p_{\text{pb}} + u$$

   Suppose that the true demand curve is linear. Would the least-squares estimate $\hat{\beta}_1$ be unbiased and consistent? Does your answer depend on the answer to part (1) of this question?

answer Yes, but only if there were no omitted variables, such as p_{jelly}. If jelly and peanut butter are complements, then p_{jelly} is an omitted variable and the coefficient estimates will be biased and inconsistent.

3. You'd like to estimate the elasticity of jelly demand with respect to its own price. What model would you estimate to answer this question? What is the coefficient of interest?

answer

$$log(q_{jelly}) = \beta_0 + \beta_1 log(p_{jelly}) + \beta_2 log(p_{pb}) + u$$

β_1 is the coefficient of interest.

4. You want to evaluate the hypothesis that the elasticity of jelly with respect to its own price is twice the elasticity of jelly with respect to the price of peanut butter. How would you rearrange the model so that this hypothesis is rejected when the coefficient on the first variable is far from zero?

answer

$$H_0 : \beta_1 - 2\beta_2 = 0$$

$$log(q_{jelly}) = \beta_0 + \beta_1 log(p_{jelly}) + \beta_2 log(p_{pb}) + u$$

$$log(q_{jelly}) = \beta_0 + \beta_1 log(p_{jelly}) - 2\beta_2 log(p_{jelly}) + 2\beta_2 log(p_{jelly}) + \beta_2 log(p_{pb}) + u$$

$$log(q_{jelly}) = \beta_0 + (\beta_1 - 2\beta_2) log(p_{jelly}) + \beta_2 (log(p_{pb}) + 2log(p_{jelly})) + u$$

So, run the regression of $log(q_{jelly})$ on $log(p_{jelly})$ and $(log(p_{pb}) + 2log(p_{jelly}))$.

5. It turns out that the firm has an additional dataset with observational data on sales and prices. This dataset contains the same variables, but the prices vary as a result of the choices that stores make in order to maximize profits, rather than because of an experimental manipulation. Would it be a good idea to use this data to estimate the price elasticity? How would you expect the estimated coefficient to compare to what you obtained using the experimental data?

answer No, it is not a good idea, because price and quantity are now simultaneously determined in the market. Reverse causality bias would be present. We would expect elasticities to be biased upwards.

6. Based on your measurement of the price elasticity of demand, the firm decides that prices are too low, and chooses to raise them by a large amount in all stores. Provide two reasons why your analysis may lack external validity with respect to the effects of the proposed price increase.

answer Possible answers include: 1) Population studied is different than the population the policy is being implemented in, 2) Extrapolation– large price increase was not studied, 3) Misspecification.