Midterm 2 - Preparation

ECO 231 - Undergraduate Econometrics

Prof. Carolina Caetano

INSTRUCTIONS

The instructions below are the same that will be given to you in the exam.

1. You have received three booklets. Booklet 1 contains the exam instructions and the exam questions. Booklet 2 contains the numbered pages where you will answer question 1. Booklet 3 contains the numbered pages where you will answer question 2.

2. This exam has 2 questions, each is worth 50 points. Each item inside a question is worth the same. You have until 5 minutes before the end of the regular class time to answer it.

3. You must answer each question exactly in the space provided for it in booklets 2 and 3. You may use the back of the pages if they are empty. If you answer a question out of the order, or otherwise not on the space provided for it in the second booklet, your question will not be graded. If you need more space, you must ask for extra paper from the TA. It is your responsibility at the end of the exam to staple the extra page exactly in the right place in your exam. You may ask for draft paper if you like.

4. You are not allowed the use of notes, cheat sheets, calculators, or electronic devices of any kind. Turn your cell phone off, and put it away. If you did not bring a watch, check the board. The TAs will write down the time in the board every 15 minutes. If your answers are unclear or illegible you may lose points. You may answer in pencil.

5. If you finished your exam until 10 minutes before the end of class time, you may hand it back and leave the room. However, you may not keep booklet 1.

6. If you finished within 10 minutes of the end of class time, you must remain seated. Do not get up when the TA announces the time is up. Follow the TA's instructions about how to hand booklets 2 and 3. You may keep booklet 1 for yourself.

7. Write down your name on booklets 2 and 3. An exam without the name will not be graded.

1 Material Question

In the exam you will receive the actual situation. Here the questions are written in a generic form with variables y, x_1, etc. There are many questions below, and you must prepare for them all. However, in the exam I will only give you a subset of them, so that you can complete the exam during class time. Notice that you may still be pressed for time during the actual exam, so prepare your answers so that they are concise.

Suppose that we are interested in the effect of x_1 on y.

(a) Consider variables x_2 and x_3. We would like to run an OLS regression on the equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

Write down all the assumptions that guarantee that the OLS estimators of the coefficients of this equation are unbiased.

(b) If the model is true, is the OLS regression a good method for discovering the value of the coefficients? Why?

(c) This model is far from realistic. At the very least, we should also include information about (dummy) variable x_4. Describe in as much depth as you can what is the bias resulting of omitting x_4 from the model.

(d) Suppose that our data has information about a certain variable, which has several categories (more than 2), let's say d_1, d_2, d_3 and d_4. Write the new model which incorporates this information, and interpret β_0. Do you expect it to be higher or lower than β_0 in the model in item (a)?

(e) Suppose that the errors are homoskedasitc. What is the variance of $\hat{\beta}_1$ in the new model from item (d)? Do you expect that it will be bigger or smaller than in the original model in item (a)? Explain.

(f) Should we incorporate the interactions between x_4 and the categorical variables from item (d)? Why?

(g) Incorporate the interactions between x_4 and d_1, d_2, d_3, d_4 and write the new model. Interpret β_1. Do you expect it to be higher or lower than β_1 in the model you wrote in item (d)?

(h) If we really wanted to write a causal model of the effect of x_1 on y, what else should we include? Write down as many variables as you can up to 5. The variables should be different enough that redundancy is unlikely, and the association should be obvious

to anyone. Choose one of them, and defend why it should have been included among the controls in the model in item (g).

(i) If $\hat{\beta}_1 = a$, and $SE(\hat{\beta}_1) = b$, would you say that x_1 impacts y? Why?

(j) The original data set had n observations. How many more observations would you need in order to claim that each extra unit of x_1 accounts for an increase in y of at least c?

(k) How would you reply to the following statement: "statement about one or more coefficients" in a scientific manner? (Here I cannot give you the statement as it shows in the exam because I would just give it completely away. You must prepare to answer a question about testing).

(l) Suppose that in the model in item (a) all the assumptions for the unbiasedness of the OLS hold, but the errors are heteroskedastic. What would you do? Make sure that you mention the reasons for your decisions.

(m) The variable y is directly caused by a bunch of factors (I'll be specific in the exam). However, we are using as control variables such as x_2, x_3, etc. Explain why.

(n) The variable y depends on some factor. What control variables we should use as proxies for this factor?

(o) Suppose that you have a lot of information of a certain kind which is connected to the topic. It may be related to the topic substantially or tangentially. Explain your decision process about including those variables or not.

2 Paper Question

This question refers to this year's paper. Find it in the "Download" page in the course website. **For some of the following questions, "initial model" refers to the one implicit in Table 3.**

(a) Suppose that we want to expand the main hypothesis of the paper to the existence of customer-discrimination in the sales/services sector for women, regardless of whether they are obese or not. Assume that now your sample also includes data for men, but no new explanatory variables are added.

   (a) Write down a suitable model to answer the following scientific question: what is the causal effect on wages (percentage variation) of working in different occupation sectors for women? Use the initial model as a reference.

   (b) Using your new model, formulate the test hypothesis for the existence of a wage penalty in the services/sales sector that affects women, employing a significance level of 5%.

(b) A variable "Experience" is typically included in models studying the relation between wages and personal background. Since you don't have access to a direct measure, a researcher suggests you to consider a proxy defined as Experience=Age-18. How would you include it in the initial model? Discuss if your estimates would improve as a consequence of your choice.

(c) Why does the author group sales and service occupations in one category in Table 4? Suppose that we find out that the variance of the residuals is different across sectors. How does this finding change the results presented?

(d) Name at least two possible reasons why this model could present omitted variables bias. For each of them, discuss the availability of other covariates or feasible proxies to account for this issue.

(e) Assume we introduce the control "Experience" (correctly measured) in the initial model. Is this a relevant or a superfluous variable? If it is relevant, discuss how it could change the estimates in Table 3. If it is superfluous discuss the effects of introducing it as an additional explanatory variable.

(f) Write down the test for the following hypothesis: There's no effect on wages (in percentage points) for obese women employed either in the sales or services sector (relative to the production sector). Use a significance level of 5% and make explicit all the terms you use in the expression of your statistic.

(g) How would you test the economic relevance of the variable "Tenure" in Table 3? Given the understanding of the paper that you've developed so far, can you rely on the result of this test?

(h) It is possible that customers are not homogenous across different regions of the country. For instance, in some cities people may care more or less about physical appearance, or in other cities people may be more willing to buy online. Discuss how this statement could (couldn't) affect the implicit assumptions of the model employed (Abstain from bias considerations to answer this question).

(i) Suppose that you can observe a variable called "Self-esteem", which is a proxy to account for differences in productivity that are not related to obesity or education. Could this variable fully account for the selection bias in this study? Discuss.

(j) Suppose that you are able to obtain a set of variables that account for the characteristics of the customers. Describe which of them would be relevant for your analysis and how they would affect the estimates described in Tables 3 and 4.