Midterm 2

ECO 231 - Undergraduate Econometrics

Prof. Carolina Caetano

INSTRUCTIONS

Reading and understanding the instructions is your responsibility. Failure to comply may result in loss of points, and there will be no leniency on that respect.

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. You cannot use your phone as a clock. 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

Suppose that we want to estimate the causal effect of the Florida Student Access Grant (FSAG) on college access. FSAG provides a fixed amount of money that can be used to cover the tuition fee in college and a student can apply for FSAG if his family income is below $\$40,000$ per year.

(a) Let *col* be a binary variable that takes on the value of one if a student enters college and zero otherwise, $fsag$ be a binary variable that takes on the value of one if a student is awarded FSAG, and zero otherwise, *gpa* be a student's GPA in his high school senior year, and *inc* be a student's family income. We would like to run an OLS regression on the equation

$$col = \beta_0 + \beta_1 fsag + \beta_2 gpa + \beta_3 inc + u$$

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

Answer: If all the assumptions for the OLS estimators to be unbiased are satisfied, and homoskedasticity is also satisfied, the OLS regression line provides the best linear predictor of the conditional expectation in the sense that its mean square error is minimum among all linear predictors. If homoskedasticity is not satisfied, the OLS regression line still provides a good linear predictor of the conditional expectation, but it is not the best predictor any more.

(b) This model is far from realistic. At the very least, we should also include the information about whether students participated in a gifted program during high school. Let $gift$ be a binary variable that takes on the value of one if a student participated in a gifted program and zero otherwise. Describe in as much depth as you can what is the bias resulting of omitting $gift$ from the model.

Answer: Recalling Theorem 5 from the notes, if $gift$ is a confounder which was omitted from the regression, then the coefficient on $fsag$ becomes:

$$\hat{\beta}_1 \approx \beta_1 + \hat{\beta}_{gift}\hat{\theta}_1$$

$\hat{\theta}_1$ is the slope coefficient of $fsag$ in a regression of $gift$ onto $fsag, gpa, inc$. It is possible that students who have been in a gifted program are more likely to get the FSAG. We might expect that $\hat{\theta}_1 > 0$.

$\hat{\beta}_{gift}$ is

$$\hat{\beta}_{gift} = \mathbb{E}(col|gift = 1, fsag, gpa, inc) - \mathbb{E}(col|gift = 0, fsag, gpa, inc).$$

It is possible that students who have been in a gifted program work harder and have better grades, so they are more likely to be admitted to colleges. We might expect that $\hat{\beta}_{gift} > 0$.

Therefore, $\hat{\beta}_1$ would be biased upwards.

(c) Suppose that our data was collected from four high schools: $h1$, $h2$, $h3$ and $h4$. Write the model that incorporates high school information, and interpret β_0. Do you expect it to be higher or lower than β_0 in the model in item (a)?

Answer: Rewriting the model,

$$col = \beta_0 + \beta_1 fsag + \beta_2 gpa + \beta_3 inc + \beta_4 h2 + \beta_5 h3 + \beta_6 h4 + u$$

where $h2, h3$ and $h4$ are dummy variables receiving a value if a student is from $h2$, $h3$ or $h4$, respectively. Note that $h1$ is omitted from the regression. How has the interpretation of intercept β_0 changed? In (a), β_0 is a student's expected probability of going to college when he is not awarded FSAG, has zero GPA in high school senior year and his family income is zero. Now we also hold $h2$, $h3$, and $h4$ to equal zero - meaning we are looking at a student's expected probability of going to college when he graduates from $h1$, he is not awarded FSAG, has zero GPA in high school senior year and his family income is zero. This value may be higher or lower than β_0 in (a), depending on how the quality of high school $h1$ affects the probability of going to college.

(d) Should we incorporate the interactions between $fsag$ and $h1, h2, h3, h4$? Why?

Answer: We should include the interaction terms between $fsag$ and the high school dummies if we are concerned that FSAG eligibility may have a different effect on the probability of entering colleges in different high schools. For example, it is possible that it is more difficult to get FSAG in some high schools than others. Then a student having FSAG from those high schools may work harder and thus, he has a higher chance of going to college. Incorporating the interactions, the model becomes:

$$\begin{aligned} col = \beta_0 + \beta_1 fsag + \beta_2 gpa + \beta_3 inc + \beta_4 h2 + \beta_5 h3 + \beta_6 h4 \\ + \beta_7 h2 \times fsag + \beta_8 h3 \times fsag + \beta_9 h4 \times fsag + u \end{aligned}$$

(e) If $\hat{\beta}_1 = 0.04$, and $SE(\hat{\beta}_1) = 0.02$, would you say that FSAG eligibility has a positive effect on college enrollment? Why?

Answer: If FSAG eligibility has a positive effect on college enrollment, we would

expect $\beta_1 > 0$. Formally, we wish to test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

We will reject H_0 if our test t-statistic is larger (in absolute terms) than the test critial value. For a 0.05 significance level, the critical value is 1.96. We can calculate the t-statistic of the coefficient on $fsag$,

$$|t_{stat}| = |\frac{\hat{\beta}_1 - 0}{SE(\hat{\beta_1})}| = |\frac{0.04}{0.02}| = 2 > 1.96$$

We could say that the test is inconclusive. (If you know how to do one-sided test, you would see that we can probably reject, but not strongly.)

(f) The original data set had 1600 observations. How many more observations would you need in order to claim that FSAG eligibility increases the chance of college enrollment by 2%?

Answer: We wish to test the hypothesis

$$H_0 : \beta_1 = 0.02$$
$$H_1 : \beta_1 \neq 0.02$$

The rejection rule is therefore:

$$|\frac{\hat{\beta}_1 - 0.02}{SE(\hat{\beta_1})_n}| > 2$$

or

$$|\frac{0.02}{SE(\hat{\beta_1})_n}| > 2$$

Now we need to describe $SE(\hat{\beta}_1)_n$ as a function of sample size, n. We know that for $n = 1600$,

$$SE(\hat{\beta_1})_{n=1600} = \sqrt{\frac{\hat{\sigma}^2_{\beta_1}/1600}{\hat{Var}(x_1)(1 - R^2_1)}} = 0.02$$

from which can calculate

$$\frac{\hat{\sigma}^2_{\beta_1}}{\hat{Var}(x_1)(1 - R^2_1)} = 0.02^2 \cdot 1600$$

therefore $SE(\hat{\beta}_1)_n = 0.8\sqrt{\frac{1}{n}}$, so the rejection rule must satisfy:

$$\frac{1}{40}\sqrt{n} \geq 2$$

or $n \geq 6400$ meaning we need at least 4800 additional observations to make the claim.

(g) A student thinks that his GPA in the high school senior year is the only factor that matters for his chance of going to college. How would you consider this statement in a scientific manner?

Answer: We wish to test the hypothesis that

$$H_0 : \beta_1 = 0, \beta_3 = 0, \beta_4 = 0, \beta_5 = 0, \beta_6 = 0, \beta_7 = 0, \beta_8 = 0, \beta_9 = 0.$$
$$H_1 : \text{any of them is different from zero.}$$

We test this hypothesis using the F-test. The rejection rule is given by:

$$\frac{\left[\sum_{i=1}^n \hat{u}_{ir}^2 - \sum_{i=1}^n \hat{u}_i^2\right]/8}{\sum_{i=1}^n \hat{u}_i^2/(900-9-1)} = 111.25 \cdot \frac{\sum_{i=1}^n \hat{u}_{ir}^2 - \sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n \hat{u}_i^2} > c(0.05)$$

where $\hat{u}_{ir}^2$ are the residuals from the restricted model,

$$\hat{u}_{ir} = gra - \hat{\beta}_0 - \hat{\beta}_4 c2 - \hat{\beta}_5 c3 - \hat{\beta}_6 c4$$

and $\hat{u}_i^2$ are the residuals of the unrestricted model described in part(e). $c(0.05)$ is the 95.5 percentile of the $F_{8,890}$ distribution. If our rejection rule is satisfied - namely, if our F-statistic is greater than then the test's critical value - we would reject the hypothesis that our regressors other than the college have no effect on gra.

(h) 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.

Answer: If the errors are heteroskedastic, OLS is no longer BLUE. But we can still use it as long as we correct the variance estimator. The formula now becomes

$$\widehat{Var}(\hat{\beta}_1) = \frac{\frac{1}{n-k-1}\sum_i \hat{r}_{1i}^2 \hat{u}_i^2/n}{[\frac{1}{n}\sum_i \hat{r}_{1i}^2]^2},$$

where u_i is the residual of the OLS regression in (e), and r_{1i} is the residual of the OLS regression of $fsag$ onto all the other regressors in (e). We would use this formula for testing and for confidence intervals.

(i) College enrollment should depend on students' ability. What control variables we should use as proxies for students' ability?

Answer: A person's ability is hard to measure, and is not likely to be present in most datasets. We can use parents' education or occupation, or the student's IQ as proxies. Such information is often available and can be collected.

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.

(1) 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.

Answer: A suitable model to study the scientific question would be:

$$ln(w_i) = \beta_0 + \beta_1 Woman_i + \sum_{j=1}^{4} \gamma_j \left(Woman_i \times Sector_{ij}\right) + \beta_2 T_i + \beta_3 S_i + \beta_4 R_i + \beta_5 Age_i + \sum_{j=1}^{3} \phi_j Region_{ij} + \sum_{j=1}^{4} \delta_j Sector_{ij} + e_i$$

$$E(e_i|x_i) = 0$$

Where $Woman_i$ is a dummy indicating if the individual is a woman. We measure the effect of being a woman for a particular occupation with the coeffcients γ_j. Similarly to the model in midterm 1, this coefficient indicates the additional effect of being a woman and working in a particular occupation (Since we're based on the initial model these effects are relative to the production sector). If this value is close to zero, it means that there are no gender wage gap in that sector.

(2) 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%.

Answer: Given the model in the previous part we want to test:

$$H_0 : \gamma_{j^*} = 0$$

$$H_1 : \gamma_{j^*} \neq 0$$

Where j^* is the index referring to the sales/services sector. The t-statistic will be given by $t = \frac{\hat{\gamma}_{j^*}}{SE(\hat{\gamma}_{j^*})}$ and the rejection rule would be reject the null if:

$$\left| \frac{\hat{\gamma}_{j^*}}{SE(\hat{\gamma}_{j^*})} \right| > 1.96$$

(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.

Answer: If we just add *Experience* then there will be perfect multicollinearity with age and the constant term. But if we include $Experience^2$ instead, then there will be an improvement in the identification of the model. To see why, notice that occupations are aggregated in 5 categories according to Table 1. Adding $Experience^2$ as a control we can capture the different effects across occupations grouped in the same category.

(c) 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.

Answer: There are several possible omitted variables (observables or unobservables) in this model. A comprehensive list can't be included now since we look forward to seeing original ideas on your final project. This question will be graded according to the accuracy and relevance of your choices and discussion.

(d) 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.

Answer: Using the same model as in Table 3 (no dummy for the production sector), we write the test as:

$$H_0 : \beta_{services \times obese} = \beta_{sales \times obese} = 0$$

$$H_1 : \beta_{services \times obese} \neq 0 \quad or \quad \beta_{sales \times obese} \neq 0$$

The F-test in this case is:

$$F = \frac{\left[\sum_{i=1}^{n} \hat{u}_{ir}^2 - \sum_{i=1}^{n} \hat{u}_i^2\right] / q}{\sum_{i=1}^{n} \hat{u}_i^2 / (n-k-1)}$$

Where $q = 2$, $k = 17$, $\hat{u}_i$ are the residuals in the initial OLS regression and $\hat{u}_{ir}$ are the residuals in the restricted model (that is, an OLS regression imposing $\beta_{services \times obese} = \beta_{sales \times obese} = 0$). Given these values (as well as n) the rejection rule is to reject H_0 if:

$$F > c(\alpha = 0.05)$$

The critical value $c(\alpha = 0.05)$ corresponds to the 95th percentile of the $F_{q,n-k-1}$ distribution.

(e) 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.

Answer: Not completely, although it would help. Women could still be self-selecting into different occupations for reasons other than self-esteem or confidence. This would be the case if, for example, obese women prefer working in certain sectors due to medical reasons or higher flexibility. Also, if their condition restricts them to perform some tasks, they would choose jobs in which these tasks are less required. Thereby, even if we control for self-esteem, there could exist more determinants of occupational sorting.

(f) 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.

Answer: A set of possible controls for customers would be the following: education, income, race, age, obesity, region. You can also think about proxies for preferences or changes in perceived beauty in the last years. All of the previous variables will be relevant, so their inclusion could potentially account for some of the confounders in this model. As you've studied, the inclusion of more explanatory variables has an impact on the variance of the estimators. Nevertheless, the bias-variance tradeoff doesn't seem so relevant in this case, because we have already detected potential sources of omitted variables bias and the inclusion of new information about the customers helps us with this issue. Finally, we can estimate the bias present in the initial estimates in Tables 4 and 5 applying Theorem 5 from Class 15.