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 are interested in whether female college students improve the academic outcomes of their male peers. Specifically, we want to estimate the effect of the proportion of female students in a freshman cohort on graduation rates for male students in that cohort.

(a) Let gra denote the graduation rate for male students in a freshman cohort, fem denote the proportion of female students in that cohort, hou denote the share of students in college housing, and $scla$ denote the proportion of classes with less than 20 students. We would like to run an OLS regression on the equation

$$gra = \beta_0 + \beta_1 fem + \beta_2 hou + \beta_3 scla + u$$

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

Answer: OLS estimators of the coefficients need to satisfy two conditions:

i. The population variables gra, fem, hou, $scla$ and u indeed satisfy the model

$$gra = \beta_0 + \beta_1 fem + \beta_2 hou + \beta_3 scla + u$$

where $E[u|fem, hou, scla] = 0$. This means that the expected graduation rate for male students in a freshman cohort is a linear function of the proportion of female students in that cohort, the share of students in college housing, and the proportion of classes with less than 20 students (linearity assumption). When we control for these variables, the unobservables in the model are expected to be zero (exogeneity assumption).

ii. The data $\{(gra_i, fem_i, hou_i, scla_i) : i = 1, ..., n\}$ was randomly collected from the population (random sampling assumption), and neither fem, hou or $scla$ is constant nor has an exact linear relationship with the other regressors (no perfect muticollinearity assumption).

(b) This model is far from realistic. At the very least, we should also include information about whether the university is specialized in engineering (say $eng = 1$ if the university is specialized in engineering, $eng = 0$ otherwise). Describe in as much depth as you can what is the bias resulting of omitting eng from the model.

Answer: Recalling Theorem 5 from the notes, if eng is a confounder which was omitted

from the regression, then the coefficient on fem becomes:

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

$\hat{\theta}_1$ is the slope coefficient of fem in a regression of eng onto $fem, hou, scla$. It is possible that there are less female students in the universities specializing in engineering. We might expect that $\hat{\theta}_1 < 0$.

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

$$\hat{\beta}_{eng} \approx \mathbb{E}(gra|eng = 1, fem, hou, scla) - \mathbb{E}(gra|eng = 0, fem, hou, scla).$$

It is possible that the requirements for graduating in the universities specializing in engineering are relatively higher. For example, the math course can be more demanding. As a result, the graduation rate is lower in such universities. We may expect $\hat{\beta}_{eng} < 0$.

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

(c) Suppose that our data was collected from four colleges: $c1$, $c2$, $c3$ and $c4$. Write the new model which incorporates college 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,

$$gra = \beta_0 + \beta_1 fem + \beta_2 hou + \beta_3 scla + \beta_4 c2 + \beta_5 c3 + \beta_6 c4 + u$$

where $c2, c3$ and $c4$ are dummy variables receiving a value if a student is from $c2$, $c3$ or $c4$, respectively. Note that $c1$ is omitted from the regression. How has the interpretation of intercept β_0 changed? In (a), β_0 means the expected graduation rate for male students in a freshman cohort in which there is no female students, no student lives in college housing, and the number of students in all the classes is greater than 20. Now we also hold $c2$, $c3$, and $c4$ to equal zero - meaning we are looking at the expected graduation rate for male students in college $c1$ if there are no female students, no student lives in college housing, and the number of students in all the classes is greater than 20. This value may be higher or lower than β_0 in (a), depending on how the quality of college $c1$ affects graduation rate.

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

Answer: The variance of $\hat{\beta}_1$ is given by

$$Var(\hat{\beta}_1) = \frac{\sigma^2/n}{\widehat{Var}(fem)(1-R_1^2)}$$

We know $\widehat{Var}(fem)$ and n are unaffected by the addition of variables. How does σ^2 change? It is unclear how σ^2 will change, but it definitely will. If the controls added were relevant, we believe that σ^2 will decrease, but it is very uncertain.

How does R_1^2 change? R_1^2 never decreases when additional controls are introduced, and will increase if the new controls are correlated with fem. We therefore have two effects, with σ^2 probably pushing the variance down and R_1^2 likely inflating it. In this model, it is likely that $c2$, $c3$ and $c4$ are not related to fem and are related to gra. It is possible that $\hat{\beta}_1$ will decrease (but it is not clear).

(e) If $\hat{\beta}_1 = 0.2$, and $SE(\hat{\beta}_1) = 0.04$, would you say that female college students improve the academic outcomes of their male peers? Why?

Answer: If female college students improve the academic outcomes of their male peers, 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 fem,

$$|t_{stat}| = |\frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}| = |\frac{0.2}{0.04}| = 5 > 1.96$$

We can reject the null hypothesis: female college students improve the academic outcomes of their male peers.

(f) The original data set had 1600 observations. How many more observations would you need in order to claim that 1% increase in the proportion of female students in a freshman cohort leads to at least 0.15% increase in graduation rates for male students in that cohort?

Answer: We wish to test the hypothesis

$$H_0 : \beta_1 = 0.15$$
$$H_1 : \beta_1 \neq 0.15$$

The rejection rule is therefore:

$$|\frac{\hat{\beta}_1 - 0.15}{SE(\hat{\beta_1})_n}| > 2$$

or

$$|\frac{0.05}{SE(\hat{\beta_1})_n}| > 2$$

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

$$SE(\hat{\beta_1})_{n=1600} = \sqrt{\frac{\hat{\sigma}^2_{\beta_1}/1600}{\hat{Var}(x_1)(1 - R_1^2)}} = 0.04$$

from which can calculate

$$\frac{\hat{\sigma}^2_{\beta_1}}{\hat{Var}(x_1)(1 - R_1^2)} = 0.04^2 \cdot 1600$$

therefore $SE(\hat{\beta}_1)_n = 1.6\sqrt{\frac{1}{n}}$, so the rejection rule must satisfy:

$$\frac{1}{32}\sqrt{n} \geq 2$$

or $n \geq 4096$ meaning we need at least 2496 additional observations to make the claim.

(g) A male student thinks that the college itself is the only factor that matters for his chance of graduation. How would you consider this statement in a scientific manner?

Answer: We wish to test the hypothesis that

$$H_0 : \beta_1 = 0, \beta_2 = 0, \beta_3 = 0.$$
$$H_1 : \text{any of them is different from zero.}$$

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

$$\frac{\left[\sum_{i=1}^n \hat{u}_{ir}^2 - \sum_{i=1}^n \hat{u}_i^2\right]/3}{\sum_{i=1}^n \hat{u}_i^2/(1600 - 6 - 1)} = 531 \cdot \frac{\sum_{i=1}^n \hat{u}_{ir}^2 - \sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n \hat{u}_i^2} > c(0.05)$$

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 $F_{3,1593}$ 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 fem onto all the other regressors in (e). We would use this formula for testing and for confidence intervals.

(i) Suppose that you have a lot of information on college characteristics, including things such as the share of students with cars on campus, the number of professors, the number of students, the proportion of graduate students, etc. Explain your decision process about including those variables or not.

Answer: Including variables about college characteristics may help reduce the omitted variable bias by controlling for confounders. However, if the variable we include are superfluous, they will increase the variance of our coefficient estimators and reduce the precision of our estimation, without improving its accuracy.
A good strategy would be to include variables in the regression that we may consider important, and work our way to less important variables while making sure the precision of our estimators remains satisfactory. For example, the share of students with cars on campus might be important since female and male student may interact more and spend more efforts in studying.

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 is 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 to reject the null if:

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

(b) 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?

Answer: The author argues that there is no statistical difference between wage penalties in both sectors, so he splits the sample into two categories: sales/services and professional/administrative/production occupations. If we find out that the variance of the residuals is different across sectors, then we're in presence of heteroskedasticity. This implies that neither the initial test nor the t-statistics in the first column of Table 4 would be valid.

(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) 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.

Answer: It's a relevant variable. Applying Theorem 5 in Class 15 we can evaluate the effect on the rest of the covariates employing the formula:

$$\hat{\beta}_i \approx \beta_i + \hat{\beta}_w \hat{\theta}_i$$

Where $\hat{\beta}_i$ is the coefficient calculated ignoring the confounder $w = Exp$; β_i is the true causal effect; $\hat{\beta}_w$ is the average effect on $y_i = ln(w_i)$ when x_i increases one unit, keeping all other explanatory variables constant; and $\hat{\theta}_i$ is the average change in the value of the omitted variable when x_i increases one unit, keeping everything else constant.

For instance let's consider the variable $Obsese$. In this case $\hat{\theta}_{Obese} = \mathbb{E}(Exp|Obese = 1, x_2, \cdots, x_k) - \mathbb{E}(Exp|Obese = 0, x_2, \cdots, x_k)$ and $\hat{\beta}_w = \mathbb{E}(ln(w_i)|Exp = 1, x_1, x_2, \cdots, x_k) -$

$\mathbb{E}(ln(w_i)|Exp = 0, x_1, x_2, \cdots, x_k)$. The latter term should be positive, and the former should be negative (obese people could be more absent of work due to medical reasons), but certainly one could also argue that $\hat{\theta}_{Obese} \approx 0$. All in all, the sign of the bias is unclear in this case.

Other relevant variables are $Tenure$ and $School$. For tenure, we have $\hat{\theta}_{Tenure} > 0$ and $\hat{\beta}_w > 0$ so we can conclude that the coefficient reported in Table 3 is upward biased. Moreover, in the case of schooling, $\hat{\theta}_{School} < 0$ (keeping everything else constant, one additional year of education means one year less in the labor market) and $\hat{\beta}_w > 0$, so the coefficient $\hat{\beta}_{School}$ would be downward biased in Table 3.

The effect on the rest of the variables follow the same reasoning.

(e) 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?

Answer: We can use the following test:

$$H_0 : \beta_{tenure} = 0$$

$$H_1 : \beta_{tenure} \neq 0$$

Table 3 gives a t-statistic of 12.12 which is high (much greater than the common critical value 1.96), so under the assumptions 1,2,3 of the linear model this estimate is precise but very close to zero. We can conclude that it has statistical significance but it's negligible in economic terms. In other words, we can be certain that there are no causal effects of tenure on the percentage variation of wages in this sample. Nevertheless, this conclusion wouldn't necessarily be correct if this model suffers from endogeneity or heteroskedasticity.

(f) 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).

Answer: If this hypothesis is true, then the unexplained part of the variation in wages would change with some observable characteristics, in particular customers' region. Thus errors would be heteroskedastic. In the presence of heteroskedasticity the standard errors calculated in Tables 3 and 4 are no longer valid (Since the author doesn't report "robust standard errors" we can infer that he assumes homoskedasticity). To address this point we would need to correct for heteroskedasticity.