Midterm 1

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. 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

The reduction in initial employment opportunities for recent college graduates has led many researchers to question the value of a college education. More specifically we would like to ask the following question: what is the causal effect of grade point average (GPA) in college on interview rate? To be specific, interview rate is the ratio of the number of interview requests over the number of interviews to which the applicant applied. Call the variable "GPA" as *gpa* and the variable "interview rate" as *int*.

(a) What is the point of this research question? In other words, who would be concerned with this, and why?

Answer: This research question would be of interest to parents, students, and school administrators. If GPA has an effect on interview rate, these findings are of immediate interest. For example, school administrators can implement measures to enhance teaching quality and improve students' performance; students may have more incentives to work harder to improve their GPA; policy makers may regulate the student/teacher ratio or implement some other measures to ensure school quality and, hence, increase students' GPA.

(b) Describe the ideal experiment to answer this question.

Answer: Ideal but not feasible experiment: Take two random groups of students. Force the first group to achieve a GPA of 3, and the other to achieve GPA of 4. The average difference in the interview rates across the groups is caused by the difference in GPA.

An ideal and feasible experiment: We submit 1000 randomly-generated, fictitious resumes to online job openings in a particular job categories (e.g., finance). We submitted resumes to jobs what are entry level, required a college degree, only required the submission of a resume to be considered for the job, and do not require a certificate or special training. Four resumes are submitted to each advisement and each resume is submitted to 20 advertisements. We randomly assigned the following characteristics to the fictive job seekers' resumes: a name, GPA, whether the applicant graduated with an Honor's distinction, intern experience, and gender. Four resumes are submitted to each advertisement. The researcher keeps records of the number of interview requests each resume gets.

(c) Is the variable that indicates whether the applicant graduated with an Honors distinction (call it *hon*) a confounder? The variable *hon* is a zero-one indicator equal to one when the applicant graduated with an Honors distinction and zero otherwise. Explain.

Answer: For the variable *hon* to be a confounder it must satisfy three conditions:

1. It must be associated with the treatment - Yes, it is possible that the applicants who graduated with an Honor's distinction work harder than average and thus, have higher GPA.
2. It must be associated with the outcome variable - Yes. It is very likely that graduating with an Honor's distinction signals that the applicant's ability is high, so companies are more willing to interview such applicants.
3. It must not be redundant - A variable is redundant if it is predicted by the controls. Since no other controls are mentioned in this part, *hon* is not a redundant control.

Therefore *hon* is most likely a confounder.

(d) Suppose that I want to estimate the effect of GPA (gpa) on interview rate (int). Consider the univariate regression line:

$$int = a + b_1 gpa$$

how to calculate a and b_1 using OLS?

Answer: Let n be the number of observations. Let $\overline{gpa} := \frac{1}{n}\sum_1^n gpa_i$. Let $\overline{int} := \frac{1}{n}\sum_1^n int_i$. Then

$$b_1 = \frac{\sum_1^n (gpa_i - \overline{gpa})int_i}{\sum_1^n (gpa_i - \overline{gpa})^2} \qquad a = \overline{int} - b_1\overline{gpa}.$$

(e) Suppose that the estimated regression line is $int = 0.1 + 0.12 \cdot gpa$. Comment on the following sentence: if an applicant's GPA increases by 1, then the interview rate of this applicant will increase by 0.12.

Answer: The regression line is a linear predictor of the average value of interview rate for each value of GPA. Thus if the relationship between gpa and int is indeed linear (not the causal relationship necessarily, the relationship in the data) then it says that people with a difference in GPA of one point are expected to have a difference in interview rates of 0.12.

(f) Suppose that my data set contains interview rate (int), GPA (gpa), whether the applicant graduated with an Honors distinction (hon), and gender ($female$). The variable $female$ is a zero-one indicator equal to one when an applicant is female and zero otherwise. Consider the multivariate regression line:

$$int = a + b_1 gpa + b_2 hon + b_3 female$$

Express the partialling-out formula of the regression coefficient b_1 in terms of the variables in the model. Interpret it.

Answer: In this question, we need to regress gpa on hon and $female$, and get the predicted value:

$$\widehat{gpa}_i = d_1 + d_2 hon_i + d_3 female_i.$$

For each observation i, calculate the residuals:

$$r_i = gpa_i - \widehat{gpa}_i$$

Plug the residual back into the formula

$$b_1 = \frac{\sum_{i=1}^{n} r_i int_i}{\sum_{i=1}^{n} r_i^2}$$

The coefficient of gpa measures how the expected int varies when gpa varies, but we fix the value of hon and $female$.

(g) Suppose that the model is

$$int = \beta_0 + \beta_1 gpa + \beta_2 hon + \beta_3 female + u$$

where $\mathbb{E}[u|gpa, hon, female] = 0$. What is this model saying about the world?

Answer: From the model we know that $\mathbb{E}[u|gpa, hon, female] = 0$, so we arrive at

$$\mathbb{E}\left[int|gpa, hon, female\right] = \beta_0 + \beta_1 gpa + \beta_2 hon + \beta_3 female$$

First, the model says that the conditional expectation of $interview$ (conditional on gpa, hon, $female$) is a linear function of these controls. For example, if the gpa of an applicant increases from 2.5 to 3.5, this model predicts that the expected interview rate will increase by β_1; if the GPA of an applicant increases from 1.5 to 2.5, this model predicts that the expected interview rate will also increase by β_1. Second, $\mathbb{E}[u|gpa, hon, female] = 0$ also implies that the things we do not know are expected to be the same (zero) for all values of the regressors. For example, $\mathbb{E}[u|gpa = 3, hon = 1, female = 0] = \mathbb{E}[u|gpa = 2.6, hon = 0, female = 1] = 0$.

(h) Interpret β_0 and β_1 in this model.

Answer: The coefficient β_0 is the expected interview rate of a male applicant with zero GPA and no Honor's distinction. The absurdity of its definition means we are

not particularly interested in the value the intercept receives, apart from identifying serious problems in our model.

The coefficient β_1 measures how much we expect the expected interview rate to vary when we increase the GPA of one applicant by 1 unit and leave everything else constant. That is to say, we would expect β_1 higher (or lower, if β_1 is negative) interview rate if we were to increase the GPA by 1, barring any change in hon and $female$.

(i) Suppose that the data set also contains each applicant's internship experience ($intern$). Internship experience is a zero-one indicator equal to one when an applicant has internship experience and zero otherwise. We will now include it in the model. Write the new model, and interpret β_0. Do you expect it to be higher or lower than β_0 in the model in item (g)?

Answer: The new model is

$$int = \beta_0 + \beta_1 gpa + \beta_2 hon + \beta_3 female + \beta_4 intern + u$$

where $\mathbb{E}[u|gpa, hon, female, intern] = 0$. The coefficient β_0, the intercept, is the expected value of int when all controls receive a value of zero. The difference now is that apart from fixing $gpa = 0$, $hon = 0$, $female = 0$, we also fix $intern = 0$. If we would expect an applicant with internship experience has higher interview rate, we would expect β_0 to be lower than that in the model in item (g).

2 Paper Question

This question refers to this year's paper.

(a) What is the scientific question that the author wants to address? What is the main hypothesis of this study?

Answer: The scientific question of this paper is the following: how much does changing occupation impacts the wage of obese women in percentual change. The treatment is the occupational sector and the outcome is the percentage difference in wages in comparison to another occupation. Likewise, the primary hypothesis is that customer discrimination may extend beyond race or gender differences. Specifically, customers may have a preference to be served by individuals who they find visually more appealing. If these attitudes carry over to the marketplace, these women will be compensated at a lower rate.

(b) What is the economic rationale behind the hypothesis of customer discrimination in the sales/service sector?

Answer: If customers prefer to be served or attended by individuals more appealing to them, then the employer has an incentive to under employ overweight people, or to offer them a lower wage than a non-obese person would receive in the same circumstances (ceteris paribus). This would explain the larger wage differentials for obese women in the sales/service sectors observed in the data.

(c) Write the estimated model implicit in the results of Table 3 (provided below). Interpret the results for the coefficients of the interaction between obese and occupation.

Answer: From Table 3 we can infer the following model:

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

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

Where w_i is the hourly compensation, $\{O_i, R_i\}$ are dummies taking the value 1 if i

is obese and white, respectively. $\{Age_i, S_i, T_i\}$ are variables indicating age (in years), years of education and tenure with current employer (in weeks). In addition, $\{Region_{ij}, Sector_{ij}\}$ are dummy variables indicating the region of work (south, west, north central) and the occupation sector (Professional, Sales, Administrative, Services), respectively. Finally, $x_i = \{O_i, Sector_{ij}, Region_{ij}, T_i, S_i, R_i, Age_i\}$ represents all the observable variables employed in the previous regression.

There's no dummy for the Production sector, so the negative coefficient for each interaction indicates that obese women in these 4 occupations are more penalized relatively to the production sector. The negative effect ranges from -9% to -25%, and in the sales/services sectors this difference is statistically significant at 1%.

(d) Why does the author include a variable *CB* as an explanatory variable in Table 4? What does it intent to capture? (Table 4 provided below)

Answer: The variable CB corresponds to a dummy variable indicating if the worker sets his wage by means of collective bargaining. In general, workers belonging to a union will receive more similar wages than their peers, which may partially explain the wage differentials in the data.

(e) According to the results in Table 3 (provided below), which is the wage penalty (in percentage points) for a 12 years-educated, services sector obese woman, in comparison to a non-obese woman with the same characteristics.

Answer: -0.25 + 0.07 = -0.18 (the effect of education cancels out). An average obese woman in the services sector earns approximately 18% less than a non-obese woman with the same observed characteristics.

(f) What is the conclusion of the study? Does the methodology allow a complete understanding of the wage differentials in the sales/services sector? Explain.

Answer: According to the author, occupational differences are consistent with the hypothesis of customer discrimination. He argues that other potential explanations are not supported by the data. Nevertheless, he also discusses the existence of unobservable characteristics as a possible determinant of the phenomenon and the need to account for a bigger sample of observations. Thus, the study is not conclusive about the reasons that would explain the patterns observed in the data. In addition to the author's discussion, we could also raise concerns about the presence of self-selection,

the heterogeneity of firms within a given sector, or the role of preferences in people's choices. As you can see, all these critiques are valid and they can give you a good starting point for your final project.

Table 3
OLS estimates for all women: dependent variable = ln(wage).

Variable	Unstandardized coefficient
Constant	5.70*** (44.78)
Obese	.07 (1.11)
Professional × Obese	−.11 (1.45)
Sales × Obese	−.24*** (2.72)
Administrative × Obese	−.09 (1.26)
Services × Obese	−.25*** (3.40)
Tenure	.0006*** (12.12)
School	.07*** (15.94)
Race	.03 (1.61)
Age	−.001 (0.33)
South region	−.21*** (8.60)
West region	−.08***(2.87)
North central region	−.23*** (8.74)
Professional	.23*** (6.38)
Sales	−.03 (0.82)
Administrative	.11*** (3.34)
Services	−.19*** (5.22)
Adj-R^2	.31
Observations	3079

Note: Absolute values of *t*-statistics are in parenthesis; levels of statistical significance are represented by * (10%), ** (5%), and *** (1%).

Table 4
OLS estimates by occupational grouping: dependent variable = ln(wage).

Variable	Sales and service occupations: unstandardized coefficients	Professional, administrative, and production occupations: unstandardized coefficients
Constant	5.31*** (38.70)	5.73*** (76.25)
Underweight	−.26 (1.36)	.003 (0.03)
Overweight	−.03 (0.63)	.000004 (0.001)
Obese	−.11* (1.84)	.01 (0.21)
SObese	−.25*** (4.00)	−.06 (1.51)
Tenure	.0007*** (5.94)	.0006*** (9.95)
School	.09*** (8.78)	.07*** (12.91)
CB	.22*** (3.75)	.09*** (3.51)
PBP	.07* (1.76)	.03 (1.32)
YC	−.15** (2.95)	.006 (0.16)
Self	−.26*** (4.42)	−.08 (1.25)
Nonblack	.08* (1.81)	.03 (1.07)
Sales	.11*** (2.79)	
Professional		.24*** (7.59)
Administrative		.12*** (4.03)
West	−.10* (1.66)	−.09*** (2.70)
South	−.21*** (3.84)	−.19*** (7.04)
North central	−.33*** (5.97)	−.19*** (6.21)
Adj-R^2	.24	.24
Observations	881	2077

Note: Absolute values of *t*-statistics are in parenthesis; levels of statistical significance are represented by * (10%), ** (5%), and *** (1%).