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

Maternity leave has gained greater salience in the past few decades as mothers have increasingly entered the workforce. More specifically we would like to ask the following question: what is the causal effect of maternity leave on the child's exam scores? To be specific, maternity leave is measured by the length of paid maternity leave (in weeks); the exam score is each child's average score from some exams at the end of 9th grade. Call the variable "maternity leave" as *mat* and the variable "exam score" as *score*.

(a) If you were looking for an observational data set to answer this question, what would it need to have?

Answer: An observational dataset would have to contain:

1. The treatment variable, the length of paid maternity leave of the mother.
2. The outcome variable, exam score of each child.
3. A rich set of variables to use as controls. For example, mother's education, family income, etc.
4. A large number of observations.

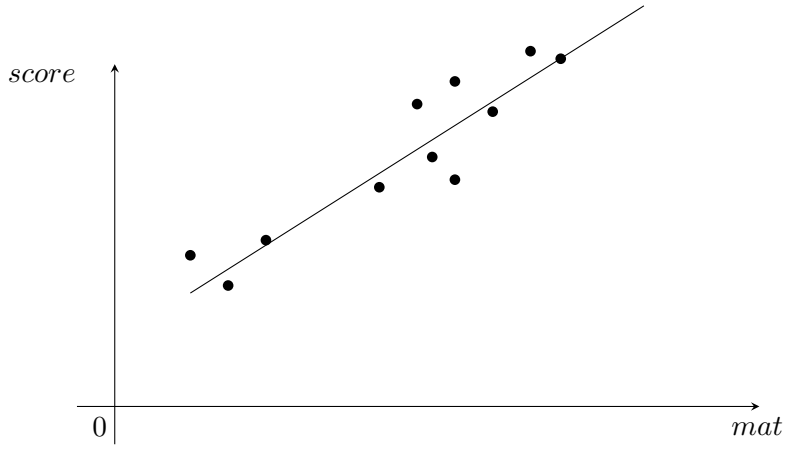
(b) Is the mother's education (call it *edu*) a confounder? Explain.

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

1. It must be associated with the treatment - Yes, it is possible that if the mother is better educated, her job is better and offers longer maternity leave.
2. It must be associated with the outcome variable - Yes. It is very likely that a better educated mother cares more about the education of her child and thus, the child tends to work harder and get higher score.
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, *edu* is not a redundant control.

Therefore *edu* is most likely a confounder.

(c) Suppose that the data set yielded the graph of averages as in the following figure. Trace the regression line of *score* on *mat*. (Don't do this in the graph below. There is one just like it in the space provided for the answer to this question.) Should a regression line be used to describe this data? Explain your answer.



Answer: Examining the graph, we can observe a strong positive correlation between interview rate and GPA. Since the relationship between the treatment and the outcome variables appears to be linear, we can use a regression line to describe the data.

(d) What is the meaning of the regression line of *score* on *mat*?

Answer: The regression line is a linear predictor of the average value of score for each value of the length of maternity leave. Caution should be used when interpreting the regression line as it might not necessarily imply causality from the length of maternity leave to the score.

(e) Suppose that my data set contains the length of paid maternity leave (mat), the child's exam scores ($score$), the mother's education (edu), and the mother's working experience (exp). The variable exp is the number of years the mother has been working before the childbirth. Consider the multivariate regression line:

$$score = a + b_1 mat + b_2 edu + b_3 exp$$

how are a, b_1, b_2, b_3 calculated?

Answer: We calculate a, b_1, b_2, b_3 by solving a system of equations:

1. The average of the regression residuals must be zero:

$$\frac{1}{n}\sum_{i=1}^{n}(score_i - \widehat{score}_i) = \frac{1}{n}\sum_{i=1}^{n}(score_i - a - b_1 mat_i - b_2 edu_i - b_3 exp_i) = 0$$

2. Next, the regression residuals must also be uncorrelated with our explanatory variables

$$\frac{1}{n}\sum_{i=1}^{n}(score_i - a - b_1 mat_i - b_2 edu_i - b_3 exp_i) mat_i = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}(score_i - a - b_1 mat_i - b_2 edu_i - b_3 exp_i)edu_i = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}(score_i - a - b_1 mat_i - b_2 edu_i - b_3 exp_i)exp_i = 0$$

(f) Suppose that the model is

$$score = \beta_0 + \beta_1 mat + \beta_2 edu + \beta_3 exp + u$$

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

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

$$\mathbb{E}\left[score|mat, edu, exp\right] = \beta_0 + \beta_1 mat + \beta_2 edu + \beta_3 exp$$

First, the model says that the conditional expectation of *score* (conditional on *mat*, *edu*, *exp*) is a linear function of these controls. For example, if the length maternity leave increases from 10 weeks to 11 weeks, this model predicts that the expected score will increase by β_1; if the length maternity leave increases from 5 weeks to 6 weeks, this model predicts that the expected interview rate will also increase by β_1. Second, $\mathbb{E}[u|mat, edu, exp] = 0$ also implies that the things we do not know are expected to be the same (zero) for all regressors. For example, $\mathbb{E}[u|mat = 12, edu = 10, exp = 0] = \mathbb{E}[u|mat = 20, edu = 18, exp = 2] = 0$.

(g) Interpret β_0 and β_1 in this model.

Answer: The coefficient β_0 is the expected score of a child whose mother has no maternity leave, no education, and no working experience.
The coefficient β_1 measures how much we expect the score to vary when we increase the length of maternity leave of one mother by 1 week and leave everything else constant. That is to say, we would expect β_1 higher (or lower, if β_1 is negative) score if we were to increase the length of maternity leave of one mother by 1 week, barring any change in *edu* and *exp* and the unobservables.

(h) Suppose that the data set also contains the family income (*inc*). 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 (f)?

Answer: The new model is

$$score = \beta_0 + \beta_1 mat + \beta_2 edu + \beta_3 exp + \beta_4 inc + u$$

where $\mathbb{E}[u|mat, edu, exp, inc] = 0$. The coefficient β_0 , the intercept, is the expected value of *score* when all controls receive a value of zero. The difference now is that apart from fixing $mat = 0$, $edu = 0$, $exp = 0$, we also fix $inc = 0$. If we would expect a child from family with higher income to have higher scores, we would expect β_0 to be lower than that in the model in item (f).

(i) Write down the formula of the R^2 of the model in item (f), and interpret it. How do you expect the R^2 of the models in items (f) and (h) to compare? Why?

Answer: The formula for R^2 is given by

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{score}_i - \overline{score})^2}{\sum_{i=1}^{n}(score_i - \overline{score})^2}$$

In this regression, the R^2 is the square of the correlation between the observed (actual) value of $score_i$ and the predicted value of $score_i$, $\widehat{score}_i$. Since R^2 never decreases when more controls are added, the R^2 in part (h) must be greater or equal to the R^2 in part (f). Since *inc* is a likely confounder in the model, including it will very likely to increase the model's explanatory power, and hence R^2 will actually increase.

2 Paper Question

This question refers to this year's paper.

(a) 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.

(b) Describe the ideal experiment to explore the customer-discrimination hypothesis. Discuss its practical applicability.

Answer: Notice first that this question explicitly refers to discrimination due to sellers' physical appearance. If we would like to learn about the specific causal effect of obesity on customers' decisions, we'd need to randomly assign a group of customers in a representative market to two different groups of sellers, identical in every characteristic, except on their weight. If we observe more sales for the group of non-obese vendors then we could argue that customer discrimination is plausible. In relation to the feasibility of this experiment, since there are confounders we aren't able to distinguish in the data (e.g. motivation) it's impossible to come up with a counterfactual group of vendors. We can conclude that the experiment is not applicable to real contexts.

(c) What is the drawback of comparing average wages of obese and non-obese women across different occupations?

Answer: Women could have chosen a job with lower wage based on characteristics or preferences that we can't observe. For example, higher levels of obesity can be associated to negative medical conditions or to lower self-esteem. Also, obesity could be correlated with lower college completion rates. All these factors affect productivity and wages. Since these characteristics are confounders that we can't distinguish in this dataset, the groups are potentially not comparable. Therefore, contrasting both groups could lead to an erroneous estimation of the causal effect.

(d) Describe the obesity measure employed by the author and how it was constructed from the data. Discuss why the author considers it is a good measure even though it is con-

structed from different surveys.

Answer: The variable BMI is constructed employing the weight reported in the 1990 round and the height reported in 1982. The height is measured in 1982 and the weight is measured in 1990, but the author argues that differences in height are minimal between 17 and 20 years. Thus, differences between measured and actual height at the moment of calculating the BMI would be negligible in this sample.

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

Answer: The variable *Self* corresponds to a dummy variable indicating if the worker is self-employed. Since they are not subject to employer discrimination, similar levels of wage differences between employees and self-employed workers in the sales/services sector would constitute evidence in favor of the customer discrimination hypothesis.

(f) Discuss the limitations of the sample of women considered in this study. Does the methodology allow a complete understanding of the wage differentials in the sales/services sector? Explain.

Answer: The author argues that making use of a subgroup of women between 26 and 33 years doesn't fully capture all the consequences of obesity on earnings. On the one hand, negative medical outcomes derived from this condition are usually observed only at a later age. On the other hand, if for some reason customer discrimination affects only young women, the inclusion of a broader sample could potentially reduce the effect of obesity on the wage differentials in the sales/services sector.

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%).