

---

## Mock Exam

*Exam Prep and Q&A on Wed 08.01 in KOL-F-101 4-6 PM.*

### Instructions to candidates

- You must write your name and student ID number on all answer pages.
  - The exam is graded out of 100 points. Points for each of the five questions are clearly labelled at the beginning of each question.
  - Please answer the long question starting on a new sheet.
  - Remember, your goal is to communicate. Full credit will be given only to the correct solution which is described clearly. Convolutd and obtuse descriptions might receive low marks, even when they are correct. Also, aim for concise solutions, as it will save you time and also help you conceptualize the key idea of the problem.
  - No calculators, mobile phones or other electronic devices are permitted. If any of these unauthorised objects are found they will be confiscated and you may face penalties.
  - Textbooks, written or recorded materials of any form are not permitted. Please have any unauthorised material securely out of view and reach. Failure to do so will result in confiscation and penalties at the discretion of the professor. If you are unsure about whether any of your materials are unauthorised please raise your hand we will check them.
  - Any communication between students during the exam, no matter whether it is about the exam or not is unacceptable and will result in harsh penalties. Attempting to view other students answers will be treated in a similar manner.
  - You are allowed to use a “Cheat Sheet” during the exam. The “Cheat Sheet” must be a single sheet of A4 paper, with *hand written* notes on *one side*. You are required to turn it in with the exam with your name on it. Failure to do so will result in penalties at the discretion of the professor.
  - You are allowed to use a dictionary.
  - $t$  and  $F$  tables are attached at the end of the exam.
  - The exam is 5 pages in total (including this page). Make sure you have all the pages.
-

## 1 Shorter Questions [35 Points]

1. **[10 Points]** Briefly describe the difference between an F-test and a t-test. Provide an example for each test different from any used in the lectures or problem sets.
2. **[10 Points]** Suppose you are hired by the superintendent of an elementary school district to decide whether to hire additional teachers. If she hires the teachers, she will reduce the number of students per teacher (the student-teacher ratio) by two. Before spending the money, however, she wants to know what impact that might have on student performance (as measured by a standardized test).

To investigate this question, you collect data on student-teacher ratios and fifth-grade test scores for 420 California school districts in 1998. In this sample, the average student-teacher ratio is 19.6 (standard deviation = 1.9) and the average test score is 654.2 (19.1). You then estimate the following model:

$$TestScores_i = \beta_1 + \beta_2 StudentTeacherRatio_i + \epsilon_i$$

Where  $i$  is a class. You find an estimated intercept,  $\hat{\beta}_1$ , of 698.9 (standard error = 10.4) and an estimated slope,  $\hat{\beta}_2$ , of -2.28 (0.52).

- (a) How would you interpret the estimated coefficient,  $\hat{\beta}_1$ ? Is it useful for your boss?
  - (b) On the basis of these results, do you conclude from your analysis that she can expect an increase of 4.56 in her district's average test scores from her plans to reduce student-teacher ratios by 2? Why or why not?
3. **[9 Points]** Consider a  $N \times K$  matrix of explanatory variables,  $X$ .
  - (a) Write down the formula for the “residual maker”,  $M_X$ . What is its dimension?
  - (b) Provide an intuitive explanation for what  $M_X$  does when “applied” to a  $N \times 1$  vector  $w$ . In other words, what does  $M_X w$  yield?
  - (c) Provide one example from the course where you used the residual maker,  $M_X$ , for a particular matrix  $X$ .

## 2 Long Question

4. **[65 Points]** As I'm sure you all know, Uber is a "ride-hailing" (i.e. taxi-like) service that connects riders to (self-employed) drivers via a smartphone app. In a recent paper, two researchers studied the effect of "surge" (or dynamic) pricing used by Uber on drivers' labor supply. The Uber platform adjusts its prices using a realtime dynamic algorithm known as "surge" pricing: this algorithm automatically raises the price of a trip when demand outstrips supply within a fixed geographic area.

The authors studied a random sample of Uber drivers in five US cities between September 2014 and July 2015, covering a total of more than 25 millions trips. They wished to test the hypothesis that higher (i.e. surge) prices increase how long drivers are willing to drive, i.e. they increase the length of their "shift" (where a shift is defined as a driver being *online* on the Uber app without a break of more than 4 hours).

Consider the following model to estimate the effects of hourly fares on drivers' hours on shift:

$$\log(\text{HoursOnShift}_{it}) = \beta_0 + \beta_1 \log(\text{HourlyFares}_{it}) + \beta_2 T_{it} + \beta_3 P_{it} + \epsilon_i \quad (1)$$

where  $\log(\text{HoursOnShift}_{it})$  is the log of the number of hours driven by driver  $i$  on shift  $t$ .  $\log(\text{HourlyFares}_{it})$  is calculated as the average hourly fare earned by the driver (i.e. it is defined as the ratio of  $i$ 's total fares earned in a session to their  $\text{HoursOnShift}_{it}$ ), while  $T_{it}$  and  $P_{it}$  are controls for temperature (measured in degrees) and precipitation (i.e. rain, measured in inches) during each shift.

The table below reports results from OLS, IV, and FE regressions of the model above, as well as an OLS ("First-stage") regression with  $\log(\text{HourlyFares}_{it})$  as a dependent variable (with the instrument used defined below). An element in the table is the estimated coefficient and its associated standard error.

- (a) Consider first the results in columns (1) and (2) and assume that the assumptions underlying the Classical Linear Regression Model (CLRM) are satisfied.
- Briefly interpret the coefficient on  $\log(\text{HourlyFares}_{it})$  in column (1). Given that the average shift length is 4 hours, considering a 50% increase in the surge price, is the estimated effect economically important?
  - In column (2) the authors control for temperature and precipitations. Why does the coefficient on  $\log(\text{HourlyFares}_{it})$  change compared to the result in column (1)?
  - Note that the coefficient of *Temperature* is not significant. Do you think it makes sense to include it with a linear specification or do you have a suggestion for a more appropriate specification for the relationship between how long a driver is willing to work and temperature?

Table 1: OLS, IV, FE and first-stage Estimates

Regressor	Dependent Variable				
		<i>log Hours On Shift</i>			<i>log Hourly Fares</i>
	OLS (1)	OLS (2)	OLS (3)	2SLS (4)	First-Stage (5)
log Hourly Fares	0.145 (0.0014)	0.197 (0.0026)	0.189 (0.0026)	0.503 (0.0057)	
log Average Hourly Fares					0.753 (0.1482)
Temperature		-0.031 (0.0142)	-0.013 (0.0251)	-0.022 (0.0321)	-0.019 (0.0287)
Precipitation		-0.048 (0.0243)	-0.021 (0.0198)	-0.013 (0.0216)	-0.012 (0.0222)
Constant	1.194 (0.0016)	1.244 (0.0025)	1.341 (0.0063)	1.671 (0.0078)	0.543 (0.0981)
<b>Fixed Effects:</b>					
Driver			X	X	X
Time			X	X	X
Observations	2'377'210	2'377'210	2'368'340	2'377'210	2'368'340
N. of Drivers	63'830	63'830	63'830	63'830	63'830
R-squared	0.007	0.013	0.038		

- iv. Referring to the formula from the lecture describing the factors that determine  $V(\hat{\beta})$ , explain which (if any) of these factors is impacted by including the extra covariates on the standard error of the  $\log(HourlyFares_{it})$  coefficient in this case. Is the change in the standard error expected or unexpected?
- (b) Why should the authors be worried? For each of the possible sources of bias below, state whether this is likely to induce a correlation between hourly fares,  $\log(HourlyFares_{it})$ , and  $\epsilon_i$ . State the direction of the bias, and briefly explain the mechanism behind it. NOTE: Normally in Supply/Demand settings there is endogeneity bias from reverse causality, but because surge pricing happens so quickly and infrequently, assume that drivers are not able to react in time such that the number of Uber drivers on the road is not affected by these prices.<sup>1</sup>

- Measurement Error

<sup>1</sup>So we are assuming NO reverse causality

- Omitted Variable Bias

- (c) In column (3) the authors control for Drivers' Fixed Effects. They do this adding a dummy for each driver.

Which of the two sources of bias discussed above do you expect this specification to address? Did introducing FE have the effect you expected on the  $\log(HourlyFares_{it})$  coefficient in the model? If so, briefly explain.

- (d) To overcome potential endogeneity problems, the authors instrument in column (4) for the variable  $\log(HourlyFares_{it})$  with  $\log(HourlyFares_{-it})$  which is the average hourly fares *of all the other drivers* in the same city and during the same hours of the driver's shift. NOTE: the authors continue to include fixed effects - they are doing FE *and* IV.

- What are the conditions of a good instrument? Do you think this variable satisfies them? Which of the two sources of endogeneity discussed above is it more likely to address? Explain.
  - Compare the results in columns (3) and (4). Did instrumenting have the effect you expected on the  $\log(HourlyFares_{it})$  coefficient in the model? If so, briefly explain. If not, what conclusions do you draw about your answer to part 4b above?
- (e) Column (5) reports first-stage results. Why might the authors have wanted to run this specification? What statistical test is most useful in these results? Implement it. What does this test tell you?