EM Q&A Session

S. Bagagli, E. Dicarlo, M.R. Greco, A. Jenni

UZH

January 8, 2020

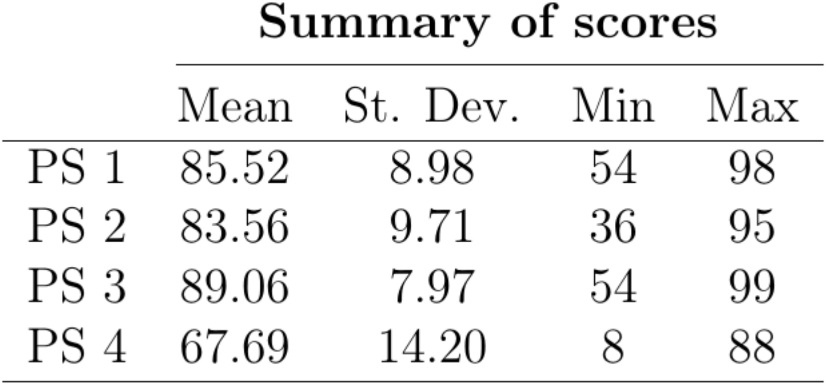
Outline

- Announcements
  - Exam
  - Problem Sets
- Mock Exam
- Your Questions

Announcements: Exam

- 2 Hours Jan 13th 10-12
- 3 Rooms
  - KOL - F - 117: surnames A to D (included)
  - KOL - F - 118: surnames E to J (included)
  - KO2 - F - 180: surnames K to Z
  - be there at 9.30
- X Short (XXpts) 1 Long (XXpts)
  - Math question (possible)
- **All** the topics except Machine Learning
- **Read carefully**

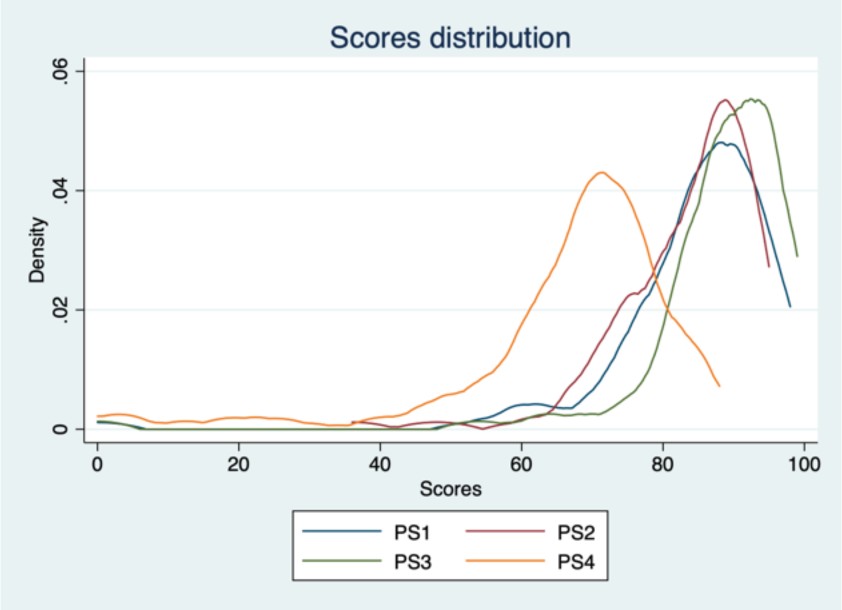
Announcements: Problem Sets



Summary of scores

	Mean	St. Dev.	Min	Max
PS 1	85.52	8.98	54	98
PS 2	83.56	9.71	36	95
PS 3	89.06	7.97	54	99
PS 4	67.69	14.20	8	88

Announcements: Problem Sets



Mock Exam
Suggested Solutions

Instructions to candidates

Read them

- You must write your name and student ID number on all answer pages.
- Please answer short and long question starting on a new page.
- 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.

Short Questions

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.
    - t-test tests single restriction in an econometric regression. (3p)
    - Example: test whether wages are higher for individuals that receive more education. Or whether political contributions from an oil company influenced a politician's votes on energy policy. (2p)

Short Questions

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.
    - An F-test tests multiple restrictions from an econometric regression. (3p)
    - Example: if you had data over time on the relationship between inputs consumed and the amount of output from a firm, you could test whether the relationship between inputs and output is the same early in the sample period versus later in the sample period.[1] Or if you had data on sports performances and multiple measures of "intelligence" (e.g. IQ score, college grade point average, etc.), test whether smarter people did better in athletics by simultaneously testing whether the coefficients on all your intelligence measures were jointly equal to zero. (2p)

[1] How: introduce a dummy variable for the later sample period and then interact this dummy with each of your inputs, giving each its own parameter. The test would be whether all these additional parameters were jointly equal to zero.

Short Questions

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 - STR) 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).

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

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).

2.a How would you interpret the estimated coefficient, $\hat{\beta}_1$? Is it useful for your boss?

- In this regression, $\widehat{\beta_1}$ has no meaningful intuitive explanation. Technically, it is the test scores when student-teacher ratio is equal to zero (i.e., there are no students in class).

2.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?

- Technically, she can expect such an increase in test scores, since the slope coefficient is statistically significant. However, it would only be correct if $E(\epsilon_i|STR_i) = 0$ (which it likely doesn't). (3p) The data can measure a correlation, but the boss wants something causal: how will her test scores change if she reduces student-teacher ratios. There could be other significant variables explaining variation in test scores across schools. If such variables do exist, they will enter the ϵ_i and $E(\epsilon_i|STR_i)$ will be non-zero, then the claim is not justified.(2p)

Short Questions

3 **[9 Points]** Consider a $N \times K$ matrix of explanatory variables, X.

3.a Write down the formula for the "residual maker", M_X. What is its dimension?

- $M_X = (I_N - X(X'X)^{-1}X')$. It is $N \times N$

3.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?

- $M_X w$ yields the residual of a regression of w on x, i.e. it gives back the e vector of estimating the following PRF $w = \mathbf{X}\beta + \epsilon$. Indeed $\mathbf{e} = w - \mathbf{X}\hat{\beta} = w - X(X'X)^{-1}X'w = M_X w$

3.c Provide one example from the course where you used the residual maker, M_X, for a particular matrix X.

- Partitioned regression

$$\hat{\beta}_k = (X_k' M_{-k} X_k)^{-1} X_k' M_{-k} y$$

- OVB

$$E(\hat{\beta}_k) \quad = \quad \beta_k + \gamma (X_k' M_{-k} X_k)^{-1} X_k' M_{-k} q$$

- Measurement Error

Long Question

4 **[65 Points]** Uber is a “ride-hailing” (i.e. taxi-like) service that connects riders to (self-employed) drivers via a smartphone app. In a paper, two researchers studied the effect of "surge" (or dynamic) pricing used by Uber on drivers’ labor supply. Uber adjusts its prices using a realtime algorithm known as "surge" pricing: it 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:

$$\log(HoursOnShift_{it}) = \beta_0 + \beta_1 \log(HourlyFares_{it}) + \beta_2 T_{it} + \beta_3 P_{it} + \epsilon_i \quad (1)$$

Long Question

$$\log(\mathit{HoursOnShift}_{it}) = \beta_0 + \beta_1 \log(\mathit{HourlyFares}_{it}) + \beta_2 T_{it} + \beta_3 P_{it} + \epsilon_i \qquad (2)$$

where $\log(\mathit{HoursOnShift}_{it})$ is the log of the number of hours driven by driver i on shift t.

$\log(\mathit{HourlyFares}_{it})$ 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 $\mathit{HoursOnShift}_{it}$), 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 $\log(\mathit{HourlyFares}_{it})$ as a dependent variable (with the instrument used defined below). An element in the table is the estimated coefficient and its associated standard error.

Figure: OLS, IV, and first-stage Estimates

		Dependent Variable			
		log Hours On Shift			*log Hourly Fares*
	OLS	OLS	OLS	2SLS	First-Stage
Regressor	(1)	(2)	(3)	(4)	(5)
log Hourly Fares	0.145	0.197	0.189	0.503	
	(0.0014)	(0.0026)	(0.0026)	(0.0057)	
log Average Hourly Fares					0.753
					(0.1482)
Temperature		-0.031	-0.013	-0.022	-0.019
		(0.0142)	(0.0251)	(0.0321)	(0.0287)
Precipitation		-0.048	-0.021	-0.013	-0.012
		(0.0243)	(0.0198)	(0.0216)	(0.0222)
Constant	1.194	1.244	1.341	1.671	0.543
	(0.0016)	(0.0025)	(0.0063)	(0.0078)	(0.0981)
Fixed Effects:					
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		

Long Question

4.a Consider first the results in columns (1) and (2) and assume that the assumptions underlying the Classical Linear Regression Model (CLRM) are satisfied.

4.a.i Briefly interpret the coefficient on $\log(HourlyFares_{it})$ 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?

Figure: OLS, IV, and first-stage Estimates

		Dependent Variable			
		log Hours On Shift			*log Hourly Fares*
	OLS	OLS	OLS	2SLS	First-Stage
Regressor	(1)	(2)	(3)	(4)	(5)
log Hourly Fares	0.145	0.197	0.189	0.503	
	(0.0014)	(0.0026)	(0.0026)	(0.0057)	
log Average Hourly Fares					0.753
					(0.1482)
Temperature		-0.031	-0.013	-0.022	-0.019
		(0.0142)	(0.0251)	(0.0321)	(0.0287)
Precipitation		-0.048	-0.021	-0.013	-0.012
		(0.0243)	(0.0198)	(0.0216)	(0.0222)
Constant	1.194	1.244	1.341	1.671	0.543
	(0.0016)	(0.0025)	(0.0063)	(0.0078)	(0.0981)
Fixed Effects:					
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		

Long Question

4.a Consider first the results in columns (1) and (2) and assume that the assumptions underlying the Classical Linear Regression Model (CLRM) are satisfied.

4.a.i Briefly interpret the coefficient on log($HourlyFares_{it}$) 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?

- Given that the average shift length is 4 hours, a 50% increase in fares implies a .145*50 = 7.25% increase in shift length = 240 minutes * .0725, or roughly 17.4 minutes. (Also 1% and .0145% is fine but then they need 50% for the economic significance)

  Economic significance: I would argue that it's not a very big effect. Remember the data is measured as the average wage *over the whole period*, so it is like if I double your wage and you say, ok I'll work 14.5% longer, that seems to me a pretty small effect.

Long Question

4.a.ii In column (2) the authors control for temperature and precipitations. Why does the coefficient on $\log(HourlyFares_{it})$ change compared to the result in column (1)?

Figure: OLS, IV, and first-stage Estimates

	Dependent Variable				
	log Hours On Shift				*log Hourly Fares*
	OLS	OLS	OLS	2SLS	First-Stage
Regressor	(1)	(2)	(3)	(4)	(5)
log Hourly Fares	0.145	0.197	0.189	0.503	
	(0.0014)	(0.0026)	(0.0026)	(0.0057)	
log Average Hourly Fares					0.753
					(0.1482)
Temperature		-0.031	-0.013	-0.022	-0.019
		(0.0142)	(0.0251)	(0.0321)	(0.0287)
Precipitation		-0.048	-0.021	-0.013	-0.012
		(0.0243)	(0.0198)	(0.0216)	(0.0222)
Constant	1.194	1.244	1.341	1.671	0.543
	(0.0016)	(0.0025)	(0.0063)	(0.0078)	(0.0981)
Fixed Effects:					
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		

Long Question

4.a.ii In column (2) the authors control for temperature and precipitations. Why does the coefficient on $\log(HourlyFares_{it})$ change compared to the result in column (1)?

- The estimated coefficient is bigger than before; it changes because we are controlling for new variables that explain hours worked and whose explanatory power was previously captured in part by $\log(HourlyFares_{it})$ whose coefficient is now "cleaned" from the negative bias generated by omitting temperature and precipitation.
  Clearly temperature and rainfall can influence whether and how long drivers want to drive and so we should include them in the econometric model. For full credit: use the bias formula to note that the direct effect of these variables on y is negative and therefore that there must be positive correlation between these vars and the hourly fare.

Long Question

4.a.iii 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?

- *Temperature* should be included as a quadratic relationship. It is likely that drivers are willing to drive less with both very low and very high temperatures.
  Other answers also possible - might try quadratics with different effects for above-average versus below-average temperatures; might try threshold effects (i.e. temperatures above 80 degrees Fahrenheit, below 30 degrees F, etc.)

Long Question

4.a.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?

- Remember the three components of the formula for $V(\hat{\beta})$

$$V(\hat{\beta}) \quad = \quad \frac{1}{N-1}\sigma^2(\frac{1}{N-1}X'X)^{-1}$$

Or $V(\hat{\beta}_2) = \sigma^2(X_2' M_{-2} X_2)^{-1}$

There might be 3 differences between (1) and (2).

First, N is the same: this does not affect standard errors. Second: adding a regressor reduces σ^2 - we take some of the noise out of the error term (assuming the regressor is relevant), lowering standard errors. Finally, adding a regressor reduces the usable variation in $\log(HourlyFares_{it})$ due to the partial regression interpretation. [2] This makes standard errors larger. In general, it's possible for them to go up or down. Here the last effect dominates. This is quite intuitive: temperature and rainfall also impact demand and could be drivers of when $\log HourlyFares$ goes up. So controlling for these appears to soak of some of its useful variation.

[2] Now there is $\log(HourlyFares_{it}) M_{-\log(HourlyFares_{it})} \log(HourlyFares_{it})$ variation to identify β

Long Question

4.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 ϵ_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. [3]

- Measurement Error
- Omitted Variable Bias

[3] So we are assuming NO reverse causality

Long Question

4.b

- Measurement Error
  - Yes.
  - Amount of hours on shift measured with error, as it measures only time the driver leaves the app open. However, we do not know how long he actually drove.
  - This is both the dependent variable and in the denominator of *Hourly Fares*.
  - In the first case, this causes no bias but only larger standard errors. Being also in our independent variable of interest, this causes attenuation bias. ME should bias our coefficient downward.
  - Careful here - since the same measurement error is in both the dependent variable *and* in the denominator of the explanatory variable, it's not a case of the standard measurement error problem. But you still get attenuation bias.

Long Question

4.b
- Omitted Variable Bias Yes. Two stories:
  - Story 1: There might be individual characteristics, such as propensity to drive of each driver, which might be positively correlated with both hours on shift and responsiveness to price.
    Why correlated with responsiveness to price? If one likes driving a lot, a smaller price increase will be enough to push him to drive more, compared to someone who does not "enjoy" driving besides the pure fact that this is his job. This should bias the estimates upward.
  - Story 2: Some drivers are just more efficient - complete jobs faster; know where to be when a new call comes in - for these they will have higher earnings over a given shift, implying a higher average price. In other words there is an individual component and a common component to the HourlyFares variable. The paper is interested in the common component (market-wide surge pricing), but the individual component could really mess things up. More efficient drivers could work either longer or shorter shifts - probably shorter? - if shorter, then this would induce a negative bias...

  NOTE: The correct answer to later questions will depend on the view they take on the direction of this bias.

Long Question

4.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.

- Solution: Yes/No. This addresses individual heterogeneity which apparently caused a positive/negative bias. But it has little effect on the results. So maybe then ME is the real problem.

Long Question

4.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.

4.d.i 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.

- Conditions are Relevance $Cov(Z, X) \neq 0$ and exogeneity $Cov(Z, \epsilon) = 0$. Instrument positively correlates with $\log(HourlyFares_{it})$ (First Stage). We've assumed away reverse causality, so that can't be causing troubles. By including fixed effects this should account for OVB meaning it's no longer in the error term. So exogeneity does hold. Using this instrument should address issues of bias from ME while FE will take care of any OVB related to individual heterogeneity.

Long Question

4.d.ii 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?

Figure: OLS, IV, and first-stage Estimates

	Dependent Variable				
	log Hours On Shift				*log Hourly Fares*
	OLS	OLS	OLS	2SLS	First-Stage
Regressor	(1)	(2)	(3)	(4)	(5)
log Hourly Fares	0.145	0.197	0.189	0.503	
	(0.0014)	(0.0026)	(0.0026)	(0.0057)	
log Average Hourly Fares					0.753
					(0.1482)
Temperature		-0.031	-0.013	-0.022	-0.019
		(0.0142)	(0.0251)	(0.0321)	(0.0287)
Precipitation		-0.048	-0.021	-0.013	-0.012
		(0.0243)	(0.0198)	(0.0216)	(0.0222)
Constant	1.194	1.244	1.341	1.671	0.543
	(0.0016)	(0.0025)	(0.0063)	(0.0078)	(0.0981)
Fixed Effects:					
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		

Long Question

4.d.ii 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?

- In case of measurement error estimates are biased downward due to attenuation bias. In this case, we would expect a higher coefficient from IV. If the effect direction supports the students answer in part (4b), then great. If it's a different effect direction, then the students should perhaps revisit the answer they gave there (tho they don't have to come up with another answer - just have to say they should think through it more.)

Long Question

4.e Columns (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?

- To check relevance of $\log(HourlyFares_{-it})$.
  The key test is the F-test on the IV. The t-stat is 5.1, thus the F-stat is its sqaure of 25+, so a very strong IV.

Your Questions

Questions

- Are 'omitted variables' and 'correlated unobservables' always synonyms? **Alex**
- Could we please again have a quick look at the proof $TSS = ESS + RSS$? **NO**
- In top01d slide 104 we say F-stat is never negative, but if the model was overfitted (K>N), would the F-stat be negative? Do we ignore this possibility in practice? **Alex**
- How relevant is Large Sample Theory for the exam? **Alex**
- When do we calculate the expected value of betas E(beta hat)? Is it the same as 'beta hat'? **Alex**
- Greg mentioned that reverse causality usually causes negative bias (when discussing the cross-border workers thesis). Can you elaborate on that give some intuition? **Ema**
- Does the Hausman endogeneity test only work for just-identified cases with one instrument for one endogenous variable or also with more than one instruments? **Sara**
- If there are gender dummies in a panel data set, an estimation using FE would not yield a β_{gender} because it doesn't vary over time. How could we still say something about m-f-effects in such cases? **Alex**

Questions

- Exercise 2, empirical application, question 1b). How do you estimate the mean of wage from table 2? **Alex**
- Exercise 2, empirical application, question 1c). Could you please explain briefly how you derived the result? **Alex**
- Do we only lose the first observation in the dataset or do we lose the first observation for each i (e.g. state)? If we subtract for example the first Texas-observation from the last New York-observation in the dataset, that wouldn't really remove the Texas-specific alpha. **Ema**
- From set 2d, slide 90: if we do the "classic" Hausman test for uncorrelated effects, a rejection of the test (i.e. a high t-statistic) means we should use FE because RE-assumption for $cov(x_it, a_i)$ doesn't hold. But what with the F-test from Wooldridge (slide 94)? Does a rejection also mean we should use FE over RE? **Sara**
- Are these slides marked with star irrelevant for the exam? (since we did't go through them in class) **They are relevant**
- Do we include "year"(time fixed effect) in the Pooled OLS? From our lecture note, we seem to include time dummy variable in the Pooled OLS. In exercise 4, we include it, too. However, in problem set 4, we did not include it. It is a little confusing. **Alex**

Questions

- Sometimes we standardized by the standard error and in other cases by the standard deviation (Exercise 2, 2c vs 2j-ii). - In which cases is one method more applicable? **Matteo**
- Is this sentence true: The value of R^2 has to increase whenever a covariate is added, since it is the sum of the individual correlations^2 (PS2,1d) **Matteo**
- Excercise 2: Comparing i) and j(i) - when do we need to fully interact and when not? In i) The R-code does not contain interactions between female and educ or age, while in j)(i) **Alex**
- Do you have more examples of unobserved heterogenity that might bias an estimator?

Good Luck!