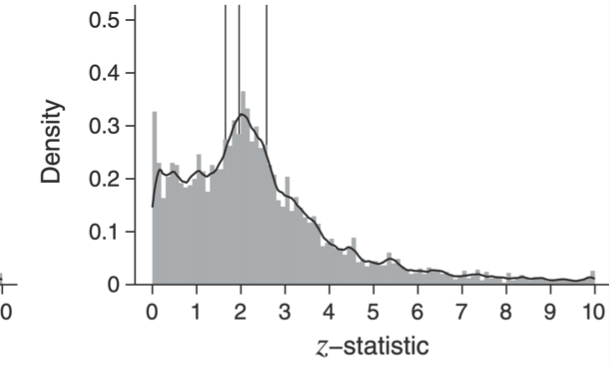
Practice Questions Midterm 2

**RCTs**

1. Explain in one sentence how an RCT help us solve the selection bias problem.
2. Explain in plain English what does expression “lack of balance in observable characteristic” mean.
3. Explain what does “lack of balance in observable characteristic” suggest about the estimation of a causals effect using an RCT?
4. RCTs are simple in logic, but difficult in logistics. In the RAND Health Insurance Experiment (HIE) discussed in class, the original design had a total of 14 treatments, but when preforming the analysis, this 14 were groups into 4 broader categories. The researchers did this to address which of the following concepts discussed in class
   1. Increase sample size so LLN and CLT applies
   2. Increase sample size to eliminate selection bias
   3. Reduce the logistical burden of tracking individuals across 6 different cities.
   4. Reduce the logistical costs of managing 14 insurance services
5. For the HIE, describe in one sentence, the key policy question that they experiment was design to answer
6. In class we discussed several logistical issues about the HIE. Which of the following, **was not** one of them?
   1. It lacked a true control groups, as everybody received some type of health insurance
   2. It had and overall cost of US $300 million.
   3. It destroys the records of treatment and control, making it impossible to measure health outcomes in the very long term.
   4. It struggled to find the administrative health records of some of the individuals in the sample.
7. What is the main difference between the treatments, as discussed in class, of the HIE and the Oregon Health Plan Experiment (OHP)?
8. Describe in 1-2 sentences the findings of the HIE
9. Describe in 1-2 sentences the findings from the OHP

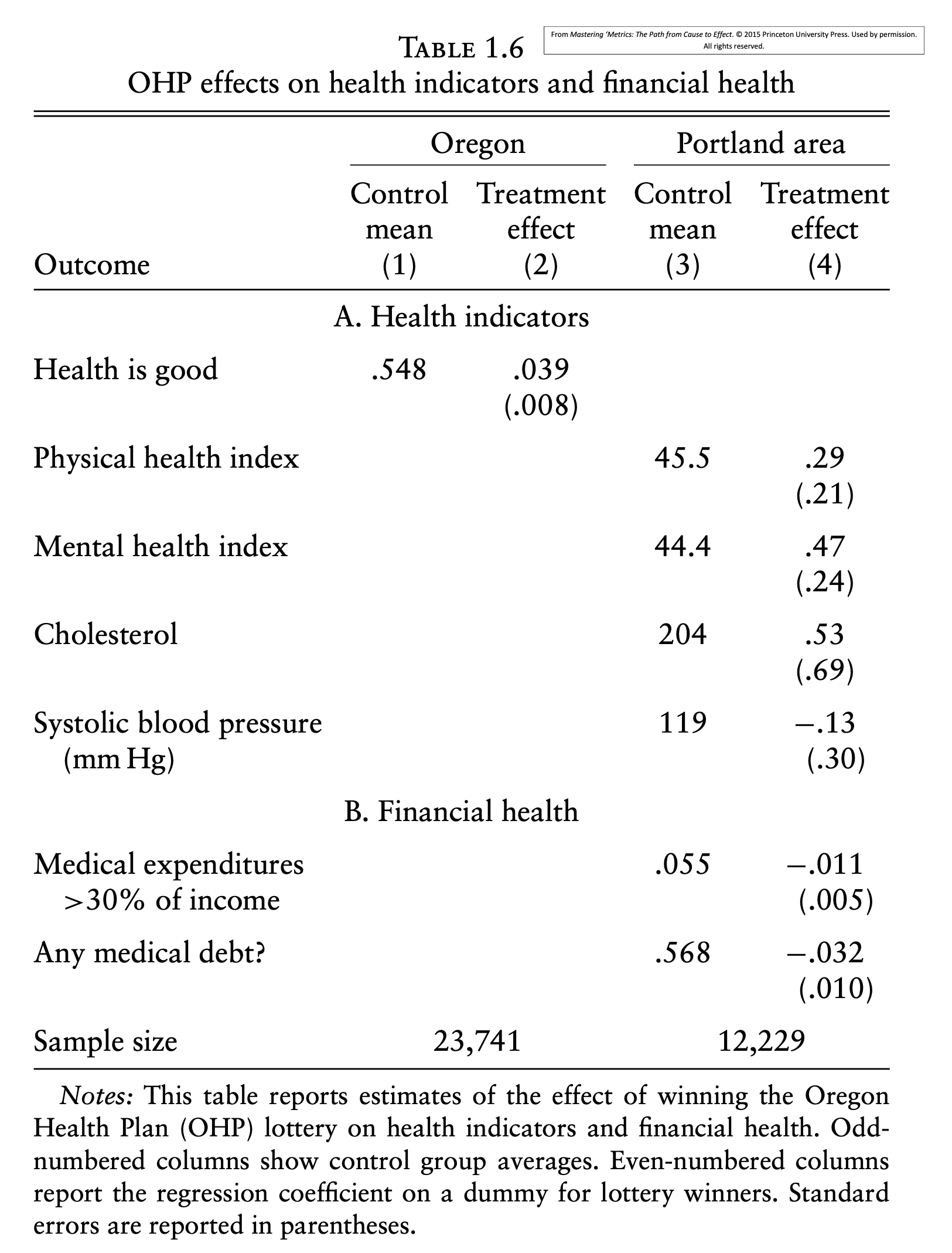
**Statistical Inference**

1. True or False: The population mean (of a variable) has a normal distribution. (justify your answer with one sentence)
2. We discussed the term “standard error” in order to distinguish between the standard deviation of which to random variables?
3. Standard errors may be complicated to derive but the idea is simple. What do they do?
4. Why do we standardize the sample mean from a Normal(mu, sigma/sqrt(n)) into a N(0,1)?
5. How do we standardize the sample mean from a Normal(mu, sigma/sqrt(n)) into a N(0,1)?
6. Which of the following values of x has a higher probability of being observed in a N(0,1)
   1. x = 0
   2. x = -1
   3. x = 2
   4. 3< x < 4
7. When conducting a hypothesis test our main goal is:
   1. Check if an observed statistic is economically meaningful
   2. Check if our hypothesis is true
   3. Check if the observed statistic is consistent with some underlying truth
   4. Check if the p-value is smaller than 0.05
8. Explain what the p-value is using the null-hypothesis and t-statistic
9. Explain what the p-value is in lay terms. Avoid using terms like null hypothesis and t-statistic.
10. (19-24) Numerical variations on any exercise from Section 6, plus asking for rough estimates of the corresponding p-values.
11. A t-statistic that test an intervention had no effect is -3.2, the standard deviation for the outcome variable is 10, and the study had 400 observations. What is estimated effect of the intervention (using a simple difference in groups)?
12. The 95% confidence interval for a sample mean [-4, 4]:
    1. Will contain the estimated sample mean 95% of the time
    2. Is more likely to contain the population value in the [-1,1] range than in the [-4, -2] range
    3. Will contain the population mean 95% of the time
    4. Is less likely to contain the population value in the [-1,1] range than in the [-4, -2] range
13. For a 90% confidence interval, explain in plain English what does the “90%” mean.
14. Define the problem of p-hacking.
15. Explain why the following figure is suggestive evidence for the existence of p-hacking in economics.
16. Briefly explain how computational reproducibility to help to reduce p-hacking.



**Regression**

1. We discuss the Table below while learning about RCTs, but now we can look with more detail at the notes: is says that this are regression coefficients. How many regressions are in these table? Write down the regression equation for the Cholesterol outcome (assume that the regression uses only the treatment variable as a regressor).



1. Extending the examples reviewed in class about the potential effect of attending a private college on earnings. Assume now that the earnings under no intervention (potential outcome, Y\_0) only differ in that individuals from rural areas tend to have lower earnings and also are more likely to attend public colleges than individuals from urban areas. How would you address this problem of selection bias in observables?
2. In the context of a regression: what does “controlling for” means in plain English?
3. What is the main reason regression alone is usually a bad research design tool to answer a causal question?
4. (35 -37, Similar to Section 7) Read a regression table, write down the corresponding equation, identify the main coefficient of interest and discuss statistical and economic significance. (repeat this exercise with at least 3 different regression tables to make sure you are comfortable with it)
5. Regression table 2
6. Regression table 3
7. Go back to lecture 1, and review the exchange between an interviewer and a commentator on the gender wage gap.
   1. Write down the underlying regression that corresponds to interviewers claim that women earn 9% less than men on average in the UK.
   2. Write down the underlying regression that corresponds to the commentator’s response
   3. Using concepts from class, particularly of regression, what is the commentator implying regarding the first estimate (in a)?
   4. Discuss how the OVB formula could help us understand the effect of including a variable like “personality” on the effect of gender on wages.

(Note: remember we are just half way through uncovering the BS in the commentator’s argument, first we need to understand his point, so then we can expose the flaws in it. If you are frustrated with the state of the debate up to here use that frustration as a motivation for learning the remaining material of the course!)

1. To decide whether or not the slope coefficient is large or small,
2. you should analyze the economic importance of a given increase in *X*.
3. the slope coefficient must be larger than one.
4. the slope coefficient must be statistically significant.
5. you should change the scale of the *X* variable if the coefficient appears to be too small.
6. To obtain the slope estimator using the least squares principle, you divide the
7. sample variance of *X* by the sample variance of *Y*.
8. sample covariance of *X* and *Y* by the sample variance of *Y*.
9. sample covariance of *X* and *Y* by the sample variance of *X*.
10. sample variance of *X* by the sample covariance of *X* and *Y*.
11. The normal approximation to the sampling distribution of  is powerful because

a. many explanatory variables in real life are normally distributed.

b. it allows econometricians to develop methods for statistical inference.

c. many other distributions are not symmetric.

d. it implies that OLS is the best estimator for  under certain conditions.

1. The OLS residuals
2. can be calculated using the errors from the regression function.
3. can be calculated by subtracting the fitted values from the actual values.
4. are unknown since we do not know the population regression function.
5. should not be used in practice since they indicate that your regression does not run through all your observations.
6. The slope estimator, *β*1, has a smaller standard error, other things equal, if
7. there is more variation in the explanatory variable, *X*.
8. there is a large variance of the error term, *u*.
9. the sample size is smaller.
10. the intercept, *β*0, is small.
11. The OLS estimator is derived by

a. connecting the *Yi* corresponding to the lowest *Xi* observation with the *Yi* corresponding to the highest observation.

b. making sure that the standard error of the regression equals the standard error of the slope estimator.

c. minimizing the sum of absolute residuals.

d. minimizing the sum of squared residuals.

1. In the simple linear regression model, the regression slope

a. indicates by how many percent *Y* increases, given a one percent increase in *X*.

b. when multiplied with the explanatory variable will give you the predicted *Y*.

c. indicates by how many units *Y* increases, given a one unit increase in *X*.

d. represents the elasticity of *Y* on *X*.

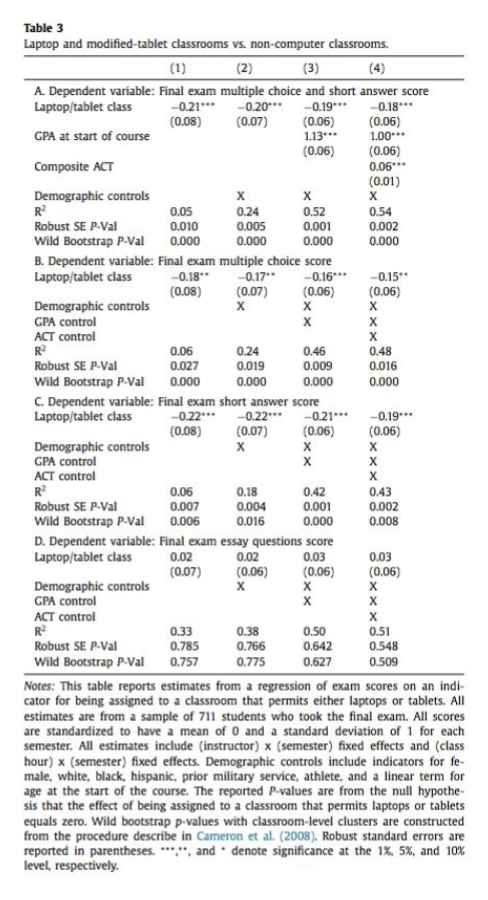
1. You have obtained a sub-sample of 1744 individuals from the Current Population Survey (CPS) and are interested in the relationship between weekly earnings and age. The regression, using standard errors, yielded the following result:

 = 239.16 + 5.20×*Age*

where *Earn* and *Age* are measured in dollars and years respectively.

* 1. Interpret the coefficient for Age.
  2. Is the effect of age on earnings large? (hint: median yearly earnings in the US where $41,000 in 2020)
  3. Why should age matter in the determination of earnings? Do the results suggest that there is a guarantee for earnings to rise for everyone as they become older? Do you think that the relationship between age and earnings is linear?
  4. The average age in this sample is 37.5 years. What is annual income in the sample?

*Table 3 below presents regression results from a study on the effect of allowing the use electronics (laptop/tablets) in a college classroom on the academic performance in the same course. Please disregard all rows with “R^2” and the “wild bootstrap P-val”; additionally, read the “Robust SE P-Val” as “p-value”).*



1. What is the difference between the different panels (A-D) in this table?
   1. The panels report estimated effects using different independent variables.
   2. The panels report estimated effects on different dependent variables.
   3. The panels report estimates using different statistical methods.
   4. a. and c. are both correct.
   5. None of the above.
2. Consider the Panel A results only. Which of the following is true of a participant’s Composite ACT?
   1. There is a significant positive association between Composite ACT and the combined final exam multiple choice and short answer score.
   2. There is a positive association between Composite ACT and final exam multiple choice and short answer score, but it is not statistically significant.
   3. Composite ACT is included as a covariate only in the model used to compute the estimates in the last column.
   4. a. and c. are both correct.
   5. None of the above
3. Write down the regression equation for column 3 in panel B, using the estimated coefficients (do not write a +b\*X1, do write 0.23 +1.23\* X1). Assume that there is only one demographic regressor: age. If you write using generic notation (Y, X1, X2, X3…) make sure to explain to describe each variable (e.g., Y: earnings, X1: when to private college, X2: own SAT /100, etc.)
4. Write down the regression equations for columns (3) and (6) in Table 2.5 of MM. Explain what is the purpose of the regressions presented in this table.

**[Additional questions on CEF, OVB and All things Regression]**

1. Explain each term in the OVB formula. Write down the equation where each is coming from.
2. We discuss in class that the CEF does not need to be a linear function (of parameters or variables). But we also **derive** one special case when the CEF is linear.
   1. Write down the CEF for this particular case.
   2. Prove that why the CEF is linear in this case.
   3. Discuss which property of CEF can you use to link this CEF with regression.
3. One reason we explore the CEF was that it has some nice properties that connects it with regression. What was the other reason?

Table 4: Regression results for a representative sample of 100,000 individuals ages between 40 and 50 years old in the US.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Log(Salaries) | | | |
|  | (1) | (2) | (3) | (4) | (5) |
| Years of education (S\_i) | 0.135  (0.051) | 0.133  (0.05) | 0.122  (0.064) | 0.092  (0.09) | 0.097  (0.065) |
| Entered labor marked in recession year (R\_i) |  | -0.05  (0.02) | -0.052  (0.022) | -0.045  (0.018) | -0.051  (0.01) |
| Age  (A\_i) |  |  | 0.02  (0.05) | 0.012  (0.002) | 0.014  (0.1) |
| Distance to wealthiest zip code in county (D\_i) |  |  |  | -0.005  (0.0001) | -0.0046  (0.00025) |
| Experience (years of work) |  |  |  |  | 0.02  (0.06) |
| Note: Distance to wealthiest zip code in county is a variable that measure the distance (in miles) of the zip code of birth to the wealthiest zip code in the county. So, if an individual was born far away from wealthy zip code, D will take a large vale (20-40 miles), on the other hand if the individual was born very close to the wealthiest zip code in their county D, would take small values (say 0-5). Enter labor market, takes the value of 1 if individual enter the labor market during a recession year, and zero otherwise. | | | | | |

(Questions 54 – 62 refer to table 4 above)

1. Describe in one sentence what this regression did in order to estimate the coefficient from above.
2. Interpret the coefficient for years of education in the first column
3. Interpret the coefficient for years of education in the third column using the idea of regression as matching.
4. Write down the regression equation for column (2)
5. Using the OVB formula, explain what is going on with the coefficient on years of education, when we move from 1 to 2, and from 3 to 4. Be explicit about the auxiliary equation that you would need in each case. (Note: by OVB formula, we mean the right-hand side, not “beta long – beta short”, given the confusion during one of the sections, we provide the solutions to one of the unobservable at the end of this document)
6. In order to capture the true causal effect, it would be helpful to have some measure of “ability” (think something like general score of “skills and talents”) and of “privilege” (think something like a general score for additional support received during upbringing and in labor market) both unobservable. Using the OVB formula (again, not the difference between beta long and beta short, but the other side of the equation), argue about the sign of the bias each of this two unobservable will generate.
7. Using regression anatomy, write down the two-regression required to generate the coefficient of education in column 5 as the coefficient of bivariate regression equation.
8. What is the t-statistic, for a null of zero association, for the coefficient on age on columns (4) and (5)? What happened with the coefficient themselves? Can you think of a possible explanation (hint: think of regression anatomy)?
9. Assume that the person that gives you the regression outputs tells you “I forgot to mentioned that the variable for years of education was actually in logs”, how should you re-interpret this coefficient?

(end of questions related to Table 4)

1. Recall the formula for standard error of regression coefficients (insert formula sigma\_e/sqrt(n) \* 1/sigma\_X tilde, where X tilde is the residual of X on a regression on all other coefficeints). When running RCTs, even with complete balance on observables, sometimes researchers include many variables in a regression equation. This does not control for OVB (given that balance has already been achieve). Why do you think they do this? When do you think that this might be a bad idea?
2. True/False or it depends (Justify in one sentence): When running a regression Y\_i = alpha +betaX + e and E(Y\_i|X) are the same thing.
3. The CEF E(Y|X) is function of:
   1. Y
   2. X
   3. Y|X
   4. X|Y
4. When discussing OVB we argue that robustness of a coefficient of interest to inclusion/exclusion of additional variables was a suggestive evidence that OVB was not a threat. Using the OVB formula, and assuming that the included/exclude variable is relevant to explain the outcome variable, what must be happening for bias to be absent?
5. True/False or it depends (Justify in one sentence): The approximation beta = log (1+beta) when beta is less 0.2 is a bad approximation because beta can take many more values.
6. State the OVB in English (no formulas) and just one sentence.
7. Looking at tables 2.3 and 2.5 of MM, can you repeat the same exercise we did for SAT, but with lnPI instead? If yes, do it. If no, explain why.
8. Derive the OVB formula with four regressors.
9. We usually cannot run population regressions. But if we could, describe one statistic that you would not need to report and justify your answer.
10. High variation in \_\_\_\_ increases the standard errors of regression coefficients, while high variation in \_\_\_ reduces it.
    1. the sample size; the residuals
    2. the residual; the sample size
    3. the regressor; the residual
    4. the residual; the regressor.

[End of practice questions]

**Solution to one part of Q58:**

Using the OVB formula, explain what is going on with the coefficient on years of education, when we move from 1 to 2, and from 3 to 4. Be explicit about the auxiliary equation that you would need in each case.

Here we will focus on moving from 3 to 4 (but the question asks you for both).

Short: Y = alpha\_s +beta1\_s (S\_i) + beta2\_s R\_i + beta\_3\_s A\_i + e\_s

Long: Y = alpha\_l +beta1\_l (S\_i) + beta2\_l R\_i + beta\_3\_l A\_i + lambda\* D + e\_l

Auxiliary:

D = pi\_0 +pi\_1 (S\_i) + pi\_2 R\_i + pi\_3 A\_i + u\_i

OVB for years of education: beta\_s – beta\_l = pi\_1 \* lambda

* Narrative on pi\_1: maybe children that grew up closer to wealthy neighborhoods (lower values of D) had access to higher quality of schooling on average which led them to have more years of education: pi\_1<0
* Narrative on lambda: in this case we can observe it, and is negative and significant (growing up further from wealthiest zip code, decreases the earning of this individuals while adults): lambda<0 [notice that when the variable is actually omitted, we also need a narrative for lambda]
* Hence OVB>0 implies beta\_s>beta\_l, omitting D would have led us to overestimate the effect of years of education on income.