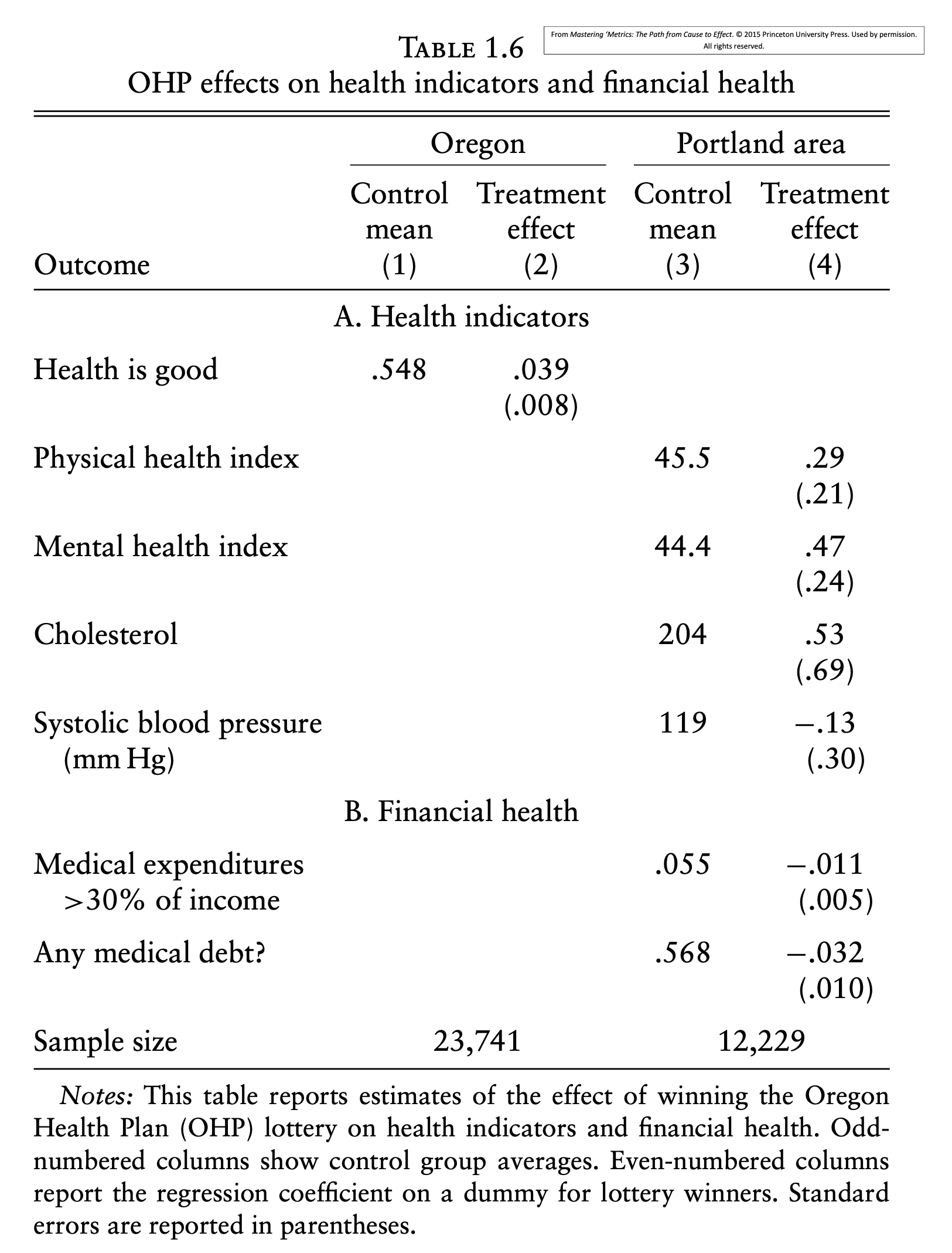
Midterm 2

1. Explain in one or two sentences how assigning an intervention on **the basis of who needs it most would** **not solve** the selection bias problem (and example of need-based assignment would be to assign health insurance to the people that have worst health, or are poorer, first).
2. For the Oregon Health Plan Experiment (OHP), describe in one sentence, one of the two key policy questions that the experiment shed light on (considered only as relevant information from Ch1).
3. What is the main difference between the **populations**, as discussed in class, of the HIE and the Oregon Health Plan Experiment (OHP)?
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 to address the fundamental problem of identification
   2. Reduce the logistical burden of tracking individuals over 14 different insurance services.
   3. Increase sample size so the estimated coefficients can be approximated by well-known distributions.
   4. Reduce variation in the error term.
5. When conducting a hypothesis test in a regression analysis our main goal is:
   1. Check if the estimated coefficient is consistent with some population parameter
   2. Check if an observed statistic is large in magnitude
   3. Check if t our t-statistic larger than 2 or smaller than -2
   4. Check if the population coefficient is consistent a range around our estimated parameter.
6. Consider the following table displaying average undergraduate GPA from a random sample of Penn State:   
   University students.  
    Mean SD N  
    GPA 3.10 .4 400
   1. Construct the 95% confidence interval for the sample mean GPA at Penn State and interpret.
   2. Calculate the t-statistic and interpret significance for the null hypothesis that population average GPA is 4.1. Repeat for a  
      population average GPA of 3.09.
   3. Provide approximate p-values for the two t-statistics in (b)
   4. Compare these two results.
7. A t-statistic that test an intervention had an effect of 1.5 is -3.2, the standard deviation for the outcome variable is 5, and the study had 100 observations. What is estimated effect of the intervention (using a simple difference in groups)?
8. The 95% confidence interval of an estimated coefficient is [1.34, 5.74]. What is the corresponding standard error of this coefficient? (hint: approximately how many times standard error can fit in this 95% confidence interval?)
9. The 95% confidence interval for a sample mean [-2, 4]:
   1. Will contain the estimated sample mean 95% of the time
   2. Is less likely to contain the population value in the [-1,1] range than in the [1, 4] range
   3. Will contain the true mean 95% of the time
   4. Is equally likely to contain the population value in the [0,1] range than in the [0, 4] range
10. P-Hacking:
    1. Define the problem of p-hacking.
    2. In the Dale and Krueger example covered in class we have seen many regression (18 if you count tables 1-3!). Could you think of additional methodological choices that could generate another set of 18 (or more) regressions (there are many ways in you can answer this, for example you can use material from nonlinearities, regression inference an even from before, if it helps you can also assume that you have other variables that you can include in your regression)?

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.
   1. How many regressions are in **panel A of this table**?
   2. Write down the regression equation for the **Physical Health index** (assume that the regression uses only the treatment variable as a regressor).
   3. Is it statistically significant? Why?



1. Under which conditions you could convincingly argue that regression is a good research resign tool to answer a causal question?
2. Question 1.3 from section 7 (“The gender gap in Canadian federal election…”), but ask the following questions:
   1. What is the corresponding equation for **Model 2?**
   2. What is the estimated equation for Model 2?
   3. Provide an interpretation of this coefficient using the idea of regression at matching.
   4. Provide an interpretation in plain English for the coefficient for the variable Women in model 2
   5. Compute the t-statistic, for standard the null of a zero coefficient, for the coefficient of the variable “Distance from contention” in model 2. Give a rough estimate of the corresponding p-value.
3. The regression residuals
4. Are aggregated in a sum to solve the regression minimization problem.
5. are unknown since we do not know the population regression function.
6. Are aggregated in a squared sum to solve the regression minimization problem.
7. should not be used in practice since they indicate that your regression does not run through all your observations.

OVB

1. Describe the OVB formula in one sentence (all English, no symbols)
2. In a video watched on lecture one, an interviewer brings up the point that the gender pay gap in the UK is 9% as evidence that modern society is still primarily dominated by men. Her point is that gender causes a wage differential.
   1. Write down the underlying regression that corresponds to interviewers' claim that women earn 9% less than men on average in the UK?

Let’s assume that another (new) commentator jumps into the conversation and says “That’s probably an underestimate once you take into account that women stay away from industries that have a bad record in the treatment of women, and that those industries tend to have lower earnings on average”

* 1. Write down the underlying regression that corresponds to the new commentator’s argument (hint: think of an Industry Machismo Index (IMI), where 0 means no machismo in the industry and 100 means maximum machismo).
  2. Discuss how the OVB formula could help us understand the effect of including a variable like “an industry machismo index” on the effect of gender on wages. Write down the long, short and auxiliary equations and interpret the OVB formula (the right hand side, not the beta\_l -  beta\_s part)

1. When discussing OVB we argue that robustness of a coefficient of interest to inclusion/exclusion of additional variables was suggestive evidence that OVB was not a threat. Using the OVB formula, and assuming that strong relationship remains between the included/excluded variable and the treatment, what must be happening for bias to be absent?

(Table 4 is used in questions 17 and 18 below)

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. | | | | | |

1. Based on OVB. Answer the following questions related to table 4:
   1. Interpret the coefficient for years of education in the second column using the idea of regression as matching.
   2. Write down the regression equation for column (3)
   3. 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)
   4. 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.

All things regression

1. Based on All Things Regression (Anatomy and others): Answer the following questions related to table 4:
   1. Write down the two-regressions required to generate the coefficient of education in column 5 as the coefficient of a bivariate regression equation.
   2. 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)?
   3. 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?
2. High variation in \_\_\_\_ increases the t-statistics regression coefficients (for 0 null hypothesis), 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.