THE UNIVERSITY OF CHICAGO Harris Public Policy

Winter 2026 **Problem Set 2** **PPHA 311**

Due date: Wednesday, January 28, 2026 at 10:00PM. You must submit via Gradescope on Canvas. No late problem sets will be accepted.
Group work: You may work in groups, but each person must submit individual answers. These answers must reflect the individual's own work and may not be copied from others or generative AI. Please write the names of all members of your study group at the top of your submission.
Scratch work and code: Please show your work (where relevant) and include **all code and output** for this assignment with your submission. Please use brief but clear comments in the code to reference the applicable assignment section. Note: submitting the code/log of a classmate is considered a violation of our academic integrity policy and will result in a 0 on the overall assignment.

Please post any clarifying questions you have to Ed Discussion. We will do our best to answer all your questions posted on Ed Discussion 24 hours before the assignment deadline.

1 Health Spending and Life Expectancy

Many countries allocate a substantial portion of their national budgets to healthcare expenditures. Are these expenditures justified by improved health outcomes? In this question, you will work with cross-country data on health spending and life expectancy.
You should use R to answer this question and include code and output in your submission.

Refer to the datafile health.csv available on Canvas, which reports data for 2019. This dataset has the following variables:

Variable	Description
country	Country name
life_exp	Life expectancy at birth (years)
health_exp	Health expenditure per capita (USD)
gdp_pc	GDP per capita (USD)

1. Create a scatter plot between health expenditure per capita and life expectancy. Please put health expenditure per capita as the X axis and life expectancy as the Y axis. (1 point)

2. Estimate the effect of *health_exp* on *life_exp* using OLS.

$$life\_exp = \beta_0 + \beta_1 \cdot health\_exp + u$$

   Interpret $\hat{\beta}_0$ and $\hat{\beta}_1$. Your answer should also reference the statistical significance of these estimates. (2 points)

3. Calculate the predicted values of life expectancy for a country with health expenditure per capita of \$3,000 and \$6,000, respectively. You can calculate the predicted values by hand (show your work) or using RStudio (show your code and output). (2 points)

4. Do you think this regression analysis is suitable for estimating the causal effect of health spending on life expectancy? Please explain your reasoning. (2 points)

5. GDP per capita is often cited as a major determinant of both health spending and health outcomes.

Add GDP per capita (*gdp_pc*) to the model in (2) and reestimate the following regression (i.e., the following regression equation).

$$life\_exp = \beta_0 + \beta_1 \cdot health\_exp + \beta_2 \cdot gdp\_pc + \epsilon$$

Report the estimated coefficient and intercept. Interpret the coefficient on *health_exp* (including the statistical significance).

Does the estimated equation suggest that health spending improves life expectancy after controlling for GDP? Based on your answers for (2), discuss whether the data support that health spending *causes* longer life expectancy. (2 points)

6. Compare the model fit in (2) and (5) using adjusted R squared. How did the model fit change from (2)? Discuss. (2 points)

7. Construct a 95 percent confidence interval for the effect of a $1,000 increase in health expenditure per capita on life expectancy. Based on the confidence intervals, discuss whether you can reject a large positive effect. (2 points)

2 Short Answer (3 points: 1 point each)

Briefly discuss whether each of the following statements is correct or incorrect and why. If the expression is correct only under certain conditions, state those conditions.

1. You run a regression of Y on X and calculate residuals. You check that residuals are uncorrelated with X. You conclude that the regression is unbiased and identifies the causal effect of X on Y.

2. $\hat{\beta}$ will be the same if you regress X on Y, or vice versa regress Y on X, because the covariance between Y and X is the same either way.

3. You are interested in measuring the causal effects of education. You have the choice between two samples: (1) a random sample of 1,000 individuals whose education range uniformly from 0 years to 18 years, or (2) another random sample of 1,000 individuals whose education ranges uniformly from 8-16 years. The first sample will yield more precise estimates since there is more variability in the underlying data.