Problem set 2

PPHA 31102 Statistics for Data Analysis II: Regressions

Jay Ballesteros

January 28, 2026

1 Health Spending and Life Expectancy

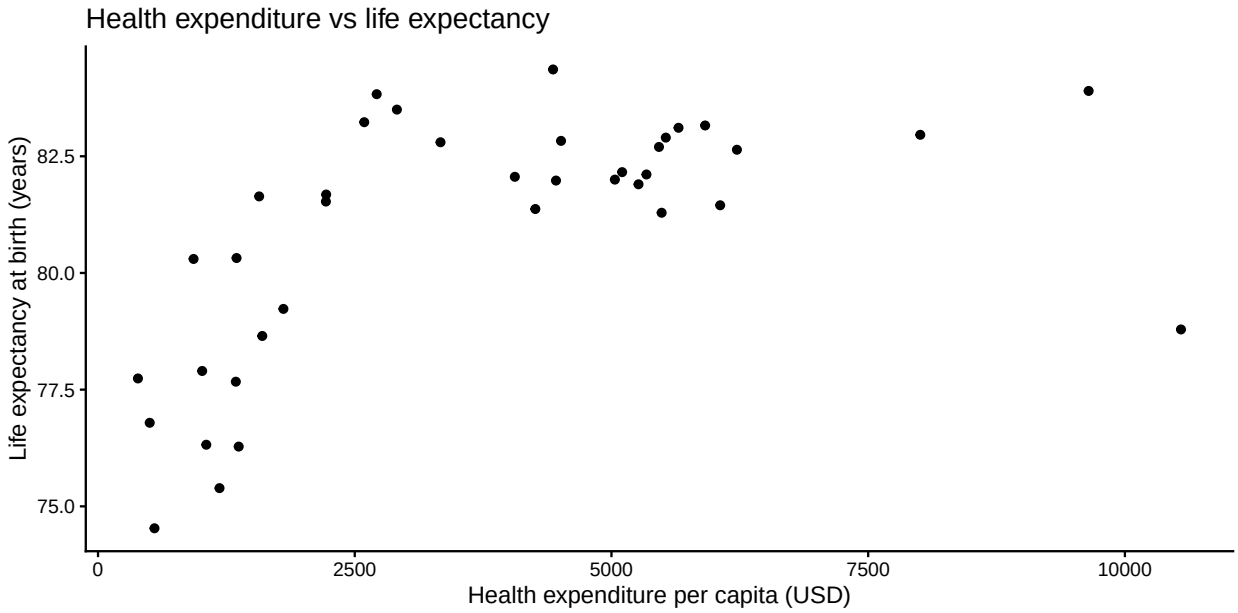
Many countries allocate a substantial portion of their national budgets to healthcare expenditures. Are these expenditures justified by improved health outcomes? In this question, you will work with cross-country data on health spending and life expectancy.

Data: 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)

Q1

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.



Q2

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.

```
q2_ols <- lm(life_exp ~ health_exp, data = health)
summary(q2_ols)
```

```
##
## Call:
## lm(formula = life_exp ~ health_exp, data = health)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1589 -0.9122  0.3145  1.1012  3.5682
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.864e+01  6.343e-01 123.982  < 2e-16 ***
## health_exp  5.984e-04  1.413e-04   4.235 0.000151 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.179 on 36 degrees of freedom
## Multiple R-squared:  0.3326, Adjusted R-squared:  0.314
## F-statistic: 17.94 on 1 and 36 DF,  p-value: 0.0001509
```

In the case of $\hat{\beta}_0$ (intercept) the result indicates that the life expectancy predicted when health expenditure per capita is \$0 USD is approximately 78.64 years. Given the p-value being close to 0 ($<$ 2e-16), the result is statistically significant.

In the case of $\hat{\beta}_1$, the result indicates that for each additional \$1 USD increase in health expenditure per capita, life expectancy increases by approximately 0.0005984 years (which is approximately 0.22 days or 5.28 hours). Given the p-value of 0.000151, the result is statistically significant as well.

Q3

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).

Considering the estimated regression equation from Q2 and the resulting coefficients:

1. For the country with health expenditure per capita of \$3,000, the predicted life expectancy is approximately

```r
coef(q2_ols)[1] + coef(q2_ols)[2] * 3000
```

```
## (Intercept)
##     80.4332
```

2. While for the country with health expenditure per capita of \$6,000, the predicted life expectancy is

```r
coef(q2_ols)[1] + coef(q2_ols)[2] * 6000
```

```
## (Intercept)
##    82.22846
```

Q4

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

No, despite high correlation, the regression is not suitable for estimating causal effect. The first reason is the assignment of the data is not randomized, and is more likely observational. Additionally, there could be other omitted variables affecting both health expenditure and life expectancy. For instance, GDP per capita, level of education, among others. These ommitted variables could lead to omitted variable bias, causing an overstimation of the true causal effect.

Q5

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:

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

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.

```
q5_ols <- lm(life_exp ~ health_exp + gdp_pc, data = health)

summary(q5_ols)
```

```
##
## Call:
## lm(formula = life_exp ~ health_exp + gdp_pc, data = health)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5219 -1.0255  0.2585  1.3054  3.6549
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.818e+01  6.729e-01 116.173   <2e-16 ***
## health_exp  1.705e-04  2.840e-04   0.601   0.5520
## gdp_pc      5.152e-05  2.991e-05   1.722   0.0938 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.122 on 35 degrees of freedom
## Multiple R-squared:  0.3847, Adjusted R-squared:  0.3496
## F-statistic: 10.94 on 2 and 35 DF,  p-value: 0.0002036
```

The estimated results are, in the case of $\hat{\beta}_0$, 78.18 years of life expectancy predicted when both health expenditure per capita and GDP per capita are $0 USD. Given the p-value being close to 0 ($<$ 2e-16), the result is statistically significant.

In the case of $\hat{\beta}_1$, the result indicates that for each additional $1 USD increase in health expenditure per capita, life expectancy increases by approximately 0.0001705 years. Given the p-value of 0.5520, the result is not statistically significant.

After including GDP per capita in the regression, the coefficient on health expenditure per capita decreased from 5.984e-04 in Q2 to 1.705e-04 and is no longer statistically significant. This result suggests that the correlation observed in Q2 was mostly driven by the bias of omitted variables. Therefore, the data do not support a causal effect of health spending on life expectancy.

Q6

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

```
summary(q2_ols)$adj.r.squared
```

```
## [1] 0.3140292
```

```
summary(q5_ols)$adj.r.squared
```

```
## [1] 0.349558
```

The adjusted R squared increase from Q2 (0.3140) to Q5 (0.3455), indicating that the model fit improved after including the GDP per capital variable. This suggests that GDP per capita explains additional variation in life expectancy, which was not captured by health expenditure per capita alone.

Q7

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.

```
confint(q5_ols, level = 0.95)
```

```
##                     2.5 %       97.5 %
## (Intercept)  7.681161e+01 7.954391e+01
## health_exp  -4.059623e-04 7.470600e-04
```

```
## gdp_pc       -9.206075e-06 1.122397e-04
```

The 95 percent confidence interval for the coefficient on health expenditure per capita is approximately (-0.0004059, 0.0007470). So multiplying the intervals by 1,000, we get (-0.4059, 0.747). Since the interval includes zero, we cannot reject the null hypothesis of no effect (at a 0.05 level).

In the case of large positive effects, given that the upper bound is 0.747 years, we can reject the hypothesis of a large positive effect. The results suggest that even if health spending does have a positive effect, it is likely to be moderate.

2 Short Answer

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.

2.1.

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.

This is incorrect. Residuals being uncorrelated with X does not guarantee that the regression suggests a causal effect (as they're always uncorrelated). To identify a causal effect, we need that the error term (u) is the one uncorrelated with X.

2.2.

$\hat{\beta}$ 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.

Incorrect. Since the denominator might change depending on which variable is the independent variable, the slope coefficients will not be the same.

2.3.

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.

This is correct. To add more variability in the independent variable (in this case education) helps to get more accurate estimates.