ICPSR Regression II - Problem Set 1

Ziyuan Gao

July 2024

1 Question 1

Population encompasses the complete set of individuals or items. A sample is a subset of individuals, items, or observations selected from a larger group or population to represent the characteristics of that larger group. In other words, it's a smaller, manageable portion of a population studied to make inferences about the whole population. As the data contains full information of the research therefore it is for the population.

2 Question 2

2.1 Part a

$$\bar{X} = \bar{Depth} = (32 + 10 + 41 + 23 + 69 + 14 + 116 + 132)/8 = 54.625$$

$$\bar{Y} = \bar{Magnitude} = (6.8 + 6.1 + 7.0 + 6.9 + 3.8 + 4.1 + 3.1 + 1.4)/8 = 4.9$$

$$\boldsymbol{b_2} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$$

$$= \frac{-42.9875 - 53.55 - 28.6125 - 63.25 - 15.8125 + 32.5 - 110.475 - 270.8125}{511.890625 + 1990.890625 + 185.140625 + 999.390625 + 206.140625 + 1650.390625 + 3766.890625 + 5986.890625}$$

$$= \frac{-553}{15,297.625} = -0.036$$

$$\boldsymbol{b_1} = \bar{Y} - \boldsymbol{b_2}\bar{X} = 4.9 - 54.625 * (-0.036) = 6.87$$

therefore,

$$Y = -0.036 * X + 6.87$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - 2}$$

$$= \frac{(6.8 - 5.718)^2 + (6.1 - 6.51)^2 + (7.0 - 5.394)^2 + (6.9 - 6.042)^2}{8 - 2} + \frac{(3.8 - 4.386)^2 + (4.1 - 6.366)^2 + (3.1 - 2.694)^2 + (1.4 - 2.118)^2}{8 - 2}$$

$$= 1.768$$

2.2 Part b

Interpretation of b_2: For every km increase in depth, the magnitude will decrease by -0.036 units.

3 Question 3

3.1 Part a

Means

$$\text{Mean Mortality Rate} = \frac{29.3 + 31.8 + 44.3 + 27.2 + 57.6 + 39.7 + 53.8 + 32.6}{8} = \mathbf{39.08}$$
$$\text{Mean Average Age} = \frac{39.4 + 40.1 + 44.3 + 38.2 + 48.4 + 41.9 + 45.9 + 41.2}{8} = \mathbf{42.2}$$
$$\text{Mean Average Income} = \frac{5511.8 + 4855.2 + 3825.5 + 5600.6 + 3974.4 + 3847.2 + 5081.2 + 4382.9}{8} = \mathbf{4854.4}$$

Covariances and Variances

$$\text{Cov}(X_1, Y) = \text{Cov(Average Age, Mortality Rate Per 100K)} = \mathbf{39.4161}$$
$$\text{Cov}(X_2, Y) = \text{Cov(Average Income, Mortality Rate Per 100K)} = \mathbf{-4342.229}$$
$$\text{Cov}(X_1, X_2) = \text{Cov(Average Age, Average Income)} = \mathbf{-1490.84}$$
$$\text{Var}(X_1) = \text{Var(Average Age)} = \mathbf{12.2107}$$
$$\text{Var}(X_2) = \text{Var(Average Income)} = \mathbf{532087.1}$$

Regression Coefficients

The coefficients are calculated as follows:

$$b_3 = \frac{\text{Cov}(X_2, Y) - \text{Cov}(X_1, X_2) \cdot \left(\frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)}\right)}{\text{Var}(X_2) - \left(\frac{\text{Cov}(X_1, X_2)^2}{\text{Var}(X_1)}\right)}$$
$$b_3 = \mathbf{0.0013}$$

$$b_2 = \frac{\text{Cov}(X_1, Y) - b_3 \cdot \text{Cov}(X_1, X_2)}{\text{Var}(X_1)}$$
$$b_2 = \mathbf{3.3920}$$

$$b_1 = \text{Mean Mortality Rate} - b_2 \cdot \text{Mean Average Age} - b_3 \cdot \text{Mean Average Income}$$
$$b_1 = \mathbf{-110.5924}$$

3.2 Part b

For each additional unit increase in average income, the mortality rate is expected to increase by 0.0013 units, assuming that the average age remains the same. For each additional unit increase in average age, the mortality rate is expected to increase by 3.3920 units, assuming that the average income remains the same. Age has a stronger and positive impact on mortality rate compared to income.

3.3 Part c

Yes, for b_3 it will be scaled by 1000, while b_2 and b_1 not change.

3.4 Part d

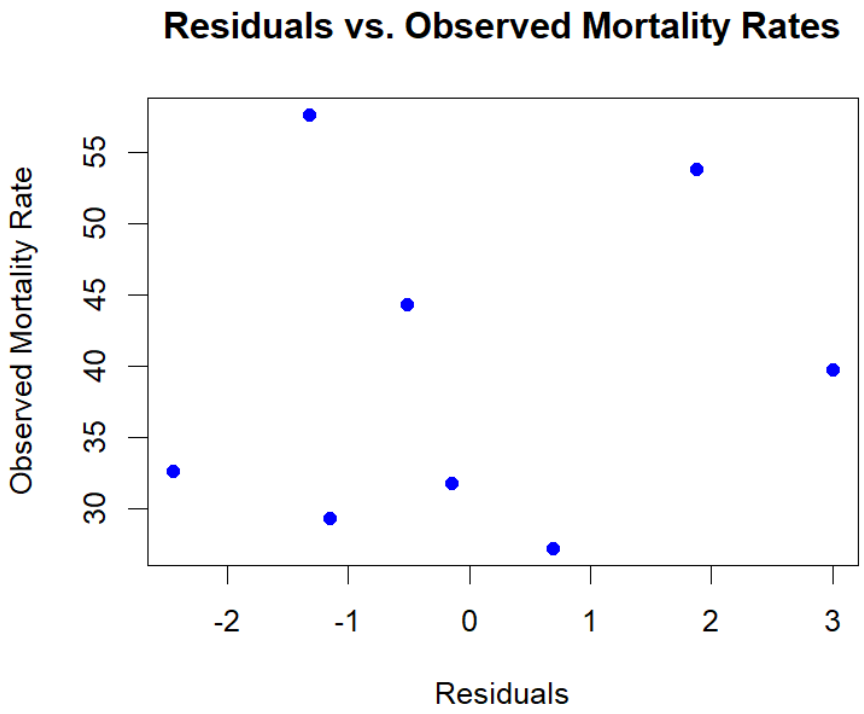

Figure 1: Residuals vs. Observed Mortality Rates

The residuals are scattered randomly and no pattern been seen.

4 Question 4

4.1 Part a

The regression coefficient b_{YX} is calculated as:

$$b_{YX} = \rho_{YX} \cdot \left(\frac{\sigma_Y}{\sigma_X}\right)$$

where:

$$\rho_{YX} = 0.167, \quad \sigma_Y = 49.06, \quad \sigma_X = 14.30$$

So,

$$b_{YX} = 0.167 \times \left(\frac{49.06}{14.30}\right) = 0.167 \times \left(\frac{49.06}{14.30}\right) = 0.5729$$

4.2

The formula for the partial regression coefficient $b_{Y,X_1 \cdot X_2}$ is:

$$b_{Y,X_1 \cdot X_2} = \frac{\rho_{Y,X_1} - \rho_{Y,X_2} \cdot \rho_{X_1,X_2}}{1 - \rho^2_{X_1,X_2}} \cdot \frac{\sigma_Y}{\sigma_{X_1}}$$

Substituting the given values:

$$b_{Y,X_1 \cdot X_2} = \frac{0.167 - 0.36 \cdot (-0.21)}{1 - (-0.21)^2} \cdot \frac{49.06}{14.30} = 0.87$$

for $b_{Y,X_2 \cdot X_1}$:

$$b_{Y,X_2 \cdot X_1} = \frac{\rho_{Y,X_2} - \rho_{Y,X_1} \cdot \rho_{X_1,X_2}}{1 - \rho^2_{X_1,X_2}} \cdot \frac{\sigma_Y}{\sigma_{X_2}}$$

Substituting the given values:

$$b_{Y,X_2 \cdot X_1} = \frac{0.36 - 0.167 \cdot (-0.21)}{1 - (-0.21)^2} \cdot \frac{49.06}{3.46} = 5.86$$

The partial slope coefficient for "Multimorbidity" is approximately 5.86. This means that, holding the average age constant, for each additional long-term disease a patient has, the number of COVID-19 related death cases increases by approximately 5.86. This indicates a significant relationship between the number of multimorbidities and the number of COVID-19 related death cases.

4.3 Part c

The change will definitely happen as there is additional independent variable "Multimorbidity" that affect the relationship to the dependent variable "of Cases".

5 Question 5

5.1

$$N = 6$$

$$\sum c_i \ = Nc = 6 * 9 = 54$$

5.2

$$\sum (X_i + Y_i) = (7 + 11) + (6 + 4) + (8 + 6) + (2 + 4) + (3 + 3) + (5 + 7) = 66$$

$$\sum (X_i) + \sum (Y_i) = (7 + 6 + 8 + 2 + 3 + 5) + (11 + 4 + 6 + 4 + 3 + 7) = 66$$

5.3

$$\sum (cX_i) = 9 * 7 + 9 * 6 + 9 * 8 + 9 * 2 + 9 * 3 + 9 * 5 = 279$$

$$c \sum (X_i) = 9 * (7 + 6 + 8 + 2 + 3 + 5) = 279$$

6 Question 6

Given the linear regression model:

$$Y = b_1 + b_2 X + \epsilon,$$

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_i)^2.$$

To find b_1 and b_2 that minimize SSE, I take partial derivatives and set them to zero.

Partial Derivative with Respect to b_1

$$\frac{\partial \text{SSE}}{\partial b_1} = -2\sum_{i=1}^{n}(Y_i - b_1 - b_2 X_i) = 0.$$

$$\sum_{i=1}^{n} Y_i - nb_1 - b_2\sum_{i=1}^{n} X_i = 0.$$

$$b_1 = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - b_2\sum_{i=1}^{n} X_i\right).$$

Partial Derivative with Respect to b_2

$$\frac{\partial \text{SSE}}{\partial b_2} = -2\sum_{i=1}^{n} X_i(Y_i - b_1 - b_2 X_i) = 0.$$

$$\sum_{i=1}^{n} X_i Y_i - b_1\sum_{i=1}^{n} X_i - b_2\sum_{i=1}^{n} X_i^2 = 0.$$

$$\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - b_2\sum_{i=1}^{n} X_i\right)\sum_{i=1}^{n} X_i - b_2\sum_{i=1}^{n} X_i^2 = 0.$$

$$\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i\sum_{i=1}^{n} X_i = b_2\left(\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right).$$

$$\text{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \frac{1}{n^2}\sum_{i=1}^{n} X_i\sum_{i=1}^{n} Y_i,$$

$$\text{Var}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} X_i\right)^2.$$

Thus:

$$\text{Cov}(X,Y) = b_2\text{Var}(X) \implies b_2 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}.$$

7 Question 7

7.1 Part a

$$\begin{aligned}\text{Mean } Y &= \frac{45.8 + 50.0 + 49.58 + 34.7 + 47.0 + 45.28 + 46.4 + 55.2 + 52.6 + 55.3 + 44.5 + 44.55}{12}\\ &= \frac{570.91}{12}\\ &= 47.58\end{aligned}$$

$$\begin{aligned}\text{Mean } X2 &= \frac{9.3 + 6.6 + 4.3 + 27.0 + 13.8 + 14.1 + 10.4 + 12.1 + 15.4 + 10.5 + 12.3 + 16.3}{12}\\ &= \frac{142.1}{12}\\ &= 11.84\end{aligned}$$

$$\begin{aligned}\text{Mean } X3 &= \frac{5.1+5.5+5.4+4.1+5.2+5.1+5.2+5.9+5.0+5.9+5.0+5.1}{12}\\ &= \frac{63.5}{12}\\ &= 5.29\end{aligned}$$

$$\begin{aligned}\text{Cov}(X2, Y) &= -20.7743\\ \text{Cov}(X3, Y) &= 2.3959\\ \text{Cov}(X2, X3) &= -1.9834\\ \text{Var}(X2) &= 32.5348\\ \text{Var}(X3) &= 0.2208\end{aligned}$$

Regression Coefficients

The coefficients are calculated as follows:

$$b3 = \frac{\text{Cov}(X3, Y) - \text{Cov}(X2, X3) \cdot \left(\frac{\text{Cov}(X2,Y)}{\text{Var}(X2)}\right)}{\text{Var}(X3) - \left(\frac{\text{Cov}(X2,X3)^2}{\text{Var}(X2)}\right)}$$

$$b3 = 11.3040$$

$$b2 = \frac{\text{Cov}(X2, Y) - b3 \cdot \text{Cov}(X2, X3)}{\text{Var}(X2)}$$

$$b2 = 0.0506$$

$$b1 = \text{Mean } Y - b2 \cdot \text{Mean } X2 - b3 \cdot \text{Mean } X3$$

$$b1 = -11.94057$$

The coefficients are:

$$\begin{aligned}b1 &= -11.94057\\ b2 &= 0.0506\\ b3 &= 11.3040\end{aligned}$$

7.2 Part b

$$N = 12$$

$$\sigma^2 = \frac{\text{Residuals\_Sum\_of\_Squares}}{N-3} = 6.7899$$

$$\text{SST}_{X_2} = \sum (X_2 - \bar{X}_2)^2 = 357.8825$$

$$\text{SST}_{X_3} = \sum (X_3 - \bar{X}_3)^2 = 2.429167$$

$$R_{X_2X_3} = \text{cor}(X_2, X_3) = -0.7399563$$

Variance of b_2:

$$\text{Var}(b_2) = \frac{\sigma^2}{\text{SST}_{X_2}(1 - R^2_{X_2X_3})} = 0.0419$$

Variance of b_3:

$$\text{Var}(b_3) = \frac{\sigma^2}{\text{SST}_{X_3}(1 - R^2_{X_2X_3})} = 6.1776$$

7.3 Part c

$$\sum(\text{Y.residuals}^2) = \sum(\hat{Y}_i - Y_i)^2 = 61.1087$$

$$R^2 = 1 - \frac{\sum(\text{Y.residuals}^2)}{\sum((Y - \bar{Y})^2)} = 0.8241$$

7.4 Part d

$$\text{SEE} = \sqrt{\frac{\sum(\text{Y.residuals}^2)}{N - 3}} = 2.6057$$

8 Question 8

8.1 Part a

We are interested in testing the following hypotheses:

$$H_0 : b_2 = 0, H_A : b_2 \neq 0$$

$$H_0 : b_3 = 0, H_A : b_3 \neq 0$$

there is no strong direction in the test(greater or smaller then). Therefore we choose a two-tailed test. I would choose 0.05 as a significane level as is industrial practice.

8.2 Part b

HRS will decrease by 1.822*20 = 36.44

8.3 Part c

The main priority of answer to partb is RATE is constant, which means is independent of values of RATE.