POL 212 Problem Set 2: Regression and Correlation Review

Objective: To give you experience in hand-calculating the correlation coefficient, regression parameter estimates, and associated quantities of interest from a linear regression model.

Directions: Please find a data set of interest to you having one continuous response variable (dependent) and two independent variables (ideally these too should be continuous or semi-continuous). On these data, please answer the following questions. (All answers should be clearly written and where appropriate, work should be shown [if you use R as your “calculator” then you should submit your R script].)

Questions

1. What is the correlation between $Y$ and one of your $X$ variables? (Show all work) (5 points).

2. What does this correlation tell you, substantively? (10 points)

3. Using R give a scatterplot of $Y$ wrt each $X$. Clearly label your plots. Based on your inspection of the plot, what can you say about the linear association between $Y$ and the $X$? (15 points)

4. Using the normal equations, what are the coefficient estimates for the slope and intercept from a regression model where $Y$ is treated as a function of $X_k$? Present the results in a publication-quality table (I understand you will not have computed standard errors . . . that is ok for now; show all work) (20 points)

5. What does your regression model tell you, substantively? That is, provide a substantive interpretation for each variable. (25 points)

6. Now using lm, estimate the model in R. Assuming your estimates are the same as in 4 (if not, redo question 4), extract the residuals and plot them with respect to $Y$. What, if any, patterns do you see? What information can you get from this kind of plot? (15 points)

7. Show how the correlation coefficient can be used to derive $\beta$. In so-doing, provide for me a technical explanation for what a slope coefficient actually is (mathematically). (10 points)