Problem sheet 4

Quantitative Methods

Preparation: In this problem sheet, you will work with real-world data from the World Bank's World Development Indicators. You will study the relationship between income-based poverty and infant mortality rates. If you have time, read through pages 139–148, sections 4.2.1 through 4.2.3, of Imai's *Quantitative Social Science*, where the main concepts of linear regression are explained. Before we construct any regression models, however, it is necessary to obtain a clean and tidy data set to work with. Have a look at chapter 12 of Wickham and Grolemund's book, especially (if you have limited time) section 12.3.1. Then have a look at this article on how to merge data sets together.

Download the data sets named `mortality.csv` and `poverty.csv` and read them into R. The former data set measures the infant (below age 5) mortality rate per 1,000 live births for a group of countries between 1960 and 2017. The latter data set measures the percentage of the population who live with less than $3.10 income (in international US dollars) per day between 1981 and 2014.

1. Explore both data sets using appropriate viewing and summary functions.

2. In a *tidy* data set, each variable has its own column, each observation has its own row, and each value has its own cell. Note that this is not the case for our two data sets. Use the `pivot_longer()` function to make both data sets tidy. Each of them should have three variables named `country`, `year`, and `mrate` (for the `mortality` data) or `prate` (for the `poverty` data).

3. Explore the `inner_join()`, `right_join()`, and `left_join()` functions to merge the two tidy data sets together into a single data frame named `data`. Which function do you think is most appropriate to use? (You can use `na.omit()` to remove missing values, which are denoted by `NA`.)

4. Let Y denote the infant mortality rate and let X denote the poverty rate. What are the expected values of Y and X?

5. Use the built-in `cor()` to calculate the correlation between Y and X.

6. Your answer to the previous questions should indicate that there a strong linear association between Y and X. We can use linear regression to model this association by positing that $Y \sim \mathcal{N}(y \mid \mu, \sigma^2)$. Here, the random variable Y, with realisation y, follows a Normal distribution with mean $\mu = \alpha + X\beta$ and variance σ^2. In other words, we are positing that Y (the "outcome" or "dependent" variable) is a linear function of X (the "regressor" or "predictor" or "independent" variable) such that μ will shift as X shifts.

The quantity of interest is β, which is a slope parameter (the slope of the line expressing the linear association between Y and X). α is an intercept term that estimates the value of Y when X is 0 (where the regression line intercepts the Y-axis). Equivalently, we can write this model as $y = \alpha + X\beta + \epsilon$, where ϵ is an error term allowing an observation to deviate from a perfect linear relationship. R's built-in function `lm()` allows you to construct such a model. Its inputs are, first, a formula of the form `y` $\sim$ `x`, where `y` is the outcome variable (mortality) and `x` is the regressor (poverty), and second, the data set (`data`). Use `lm()` to estimate β. How do you interpret the value of β?

7. Use `ggplot()` to visualise the association between Y and X. Plot each data point and also the corresponding line of best fit.

8. Now use `lm()` to regress poverty on infant mortality. What do you find?

9. Think about the limitations of this empirical analysis. What conclusions can we meaningfully draw and what new information have we gained from our models? Are the models useful?