Instructions for R Homework 5*

Question 1

For this exercise we will run a regression using Swiss demographic data from around 1888. The sample is a cross-section of French speaking counties in Switzerland. This data come with the R package `datasets`. To load this dataset type this `data(swiss, package="datasets")`.

The basic variable definitions are as follows:

VARIABLE	DESCRIPTION
Fertility	Ig, 'common standardized fertility measure'
Agriculture	% of males involved in agriculture as occupation
Examination	% draftees receiving highest mark on army examination
Education	% education beyond primary school for draftees.
Catholic	% 'catholic' (as opposed to 'protestant').
Infant.Mortality	live births who live less than 1 year.

Type `help(swiss)` in the console for additional details. Use the `summary()` command to report the mean and median for the variables Fertility, Education, and Catholic.

* For all HW assignments, I need to see all the code used

Question 2

We want to estimate the expected Fertility level in a Swiss county conditional on the county's education level. We assume the relationship is linear. So, we are interested in estimating α and β in

$$\text{Fertility}_c = \alpha + \beta \cdot \text{Education}_c + \epsilon_c.$$

If we use Ordinary Least Squares to estimate and we have the following formulas:

$$\hat{\beta} = r_{x,y} \frac{s_y}{s_x}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

where y is the left hand side variable, x is the right hand side variable, and the bar $^-$ denotes the sample mean, s is the standard deviation, and $r_{x,y}$ is the correlation between x and y.

- Find the correlation between Education and Fertility using the `cor()` function, as well as the sample standard deviation for each variable using the `sd()` function. Report these numbers.
- Use the `cor()` and `sd()` function to get an estimate for β in the equation relating Fertility to Education. Keep this value stored in a scalar called `beta_hat`. Report this number by having a line with `print(beta_hat)`.
- Now use the estimate `beta_hat`, along with the function `mean()` to get an estimate for `alpha`. Keep this value stored as a scalar called `alpha_hat`. Report this number by having a line with `print(alpha_hat)`.

Question 3

Use `alpha_hat` and `beta_hat` to predict the average fertility rate in a county where 40% of the population is educated.

Question 4

Plot the relationship between Fertility and Education using the `plot()` function with Education on the horizontal axis. Make sure to label your axis! (If you don't know how, use `?plot`).

Question 5

Now estimate the model the model relating Fertility Rate to Education using the `lm()` function in R's base code. Typically, if you want to estimate you use the syntax `lm(yvar ~ xvar, data = dataframe)`.

- Store the estimation results as follows `model_1 <- lm(...)`. This list should include a number of details include the estimated parameters, the coefficient of determination (r-squared), all of the residuals from the model, and more.
- Use the command `summary(model_1)` to report the summary of the ordinary least squares estimation. Do you have the same estimates as Question 2?
- What is the R-squared from this regression? Interpret it in a meaningful way.

Question 6

For each one of the estimated parameters reported in Question 5:

- Interpret the coefficient in a meaningful way.
- Report the results from testing the null hypothesis that the true parameter value is zero.

Question 7

Recreate the figure in Question 4, and then add the line of best fit using the `abline()` function with the coefficients from `model_1`, `model_1$coefficients`.

Question 8

Plot Education with the residuals associated with the model, `model_1$residuals`. Do the residuals show any pattern?

Question 9

Use the `mean()` command to show that the average of the residuals associated with model_1 is zero.