3818 R Homework 5

*** Student Name ***

8/20/2020

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. The first step is to load the package into your current environment by typing the command library(datasets) in to the R console. 1 This loads a number of datasets including one called swiss. Type help(swiss) in the console for additional details. The basic variable definitions are as follows:

A data frame with 47 observations on 6 variables, each of which is in percent, i.e., in [0, 100].

[,1] Fertility Ig, 'common standardized fertility measure' [,2] Agriculture % of males involved in agriculture as occupation [,3] Examination % draftees receiving highest mark on army examination [,4] Education % education beyond primary school for draftees. [,5] Catholic % 'catholic' (as opposed to 'protestant'). [,6] Infant.Mortality live births who live less than 1 year.

```r
help(swiss, package="datasets")

data(swiss, package="datasets")

head(swiss)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
```

Use the summary() command to report the mean and median for the variables Fertility, Education, and Catholic.

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 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 toEducation. Keep this value stored in a scalar called beta_hat. Report this number. - 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.

```
# Code for Question 2 goes here
```

Answer:

Question 3

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

```
# Code for Question 3 goes here
```

Answer:

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!

```
# Code for Question 4 goes here
```

Answer:

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 and paste the results in the word document. Do you have the same estimates for and from Question 2? -. What is the R-squared from this regression? Interpret it in a meaningful way.

```
# Code for Question 5 goes here
```

Answer:

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.

Answer:

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`.

```
# Code for Question 7 goes here
```

Answer:

Question 8

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

```
# Code for Question 8 goes here
```

Answer:

Question 9

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

```
# Code for Question 9 goes here
```