Assignment 1: STAT20230

Q1. In the lecture notes (Estimation (part 3)) we computed the variance of $\widehat{\beta}_1$ and $\widehat{\beta}_0$, the estimators of β_0, β_1 for linear model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. These were given by

$$Var(\widehat{\beta}_1) = \sum_{i=1}^{n} k_i^2 \sigma^2, \quad k_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

We also wrote $\widehat{\beta}_0$ as $\sum_{i=1}^{n} y_i u_i$, which gives us that its variance is

$$Var(\widehat{\beta}_0) = \sum_{i=1}^{n} u_i^2 \sigma^2, \quad u_i = \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right).$$

However, simpler expressions of $Var(\widehat{\beta}_1)$ and $Var(\widehat{\beta}_0)$ are typically prefered. They are the following

$$Var(\widehat{\beta}_1) = \sigma^2 / SS_x, \quad Var(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x} \right)$$

where $SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Show that the expressions in terms of k_i and u_i can be rewritten as above.

Q2. Data concerning the performance of 26 National Football League teams was collected. It is suspected that the number of yards gained rushing by opponents (X) has an effect on the number of games won by a team (Y). The following information is given: $SS_x = 3608611, SS_{xy} = -25350.86, SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 326.96, n = 28, SS_e = 148.87, \bar{x} = 2110.143, \bar{y} = 6.96$.

- Compute $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and $\widehat{\sigma}^2$.
- Obtain 95% t confidence intervals for $\widehat{\beta}_0$ and $\widehat{\beta}_1$.
- Recompute the intervals with the significance level of $\alpha = 0.01$.
- Comment on the width of the intervals.
- Is a Normal confidence interval also suitable in this case? Why?
- What percentage of the total variability in Y is explained by this model?
- Find a mean response 95% CI if opponents' yards rushing is 2000 yards.

Q3. Load the `bodyfat` data set and fit a linear regression to Y: body fat (%) versus X: weight (in kg). Y is the column *brozek*, and X is in lbs, so you should first covert it into Kg. To load the data, use `library(mfp)`, `data("bodyfat")`.

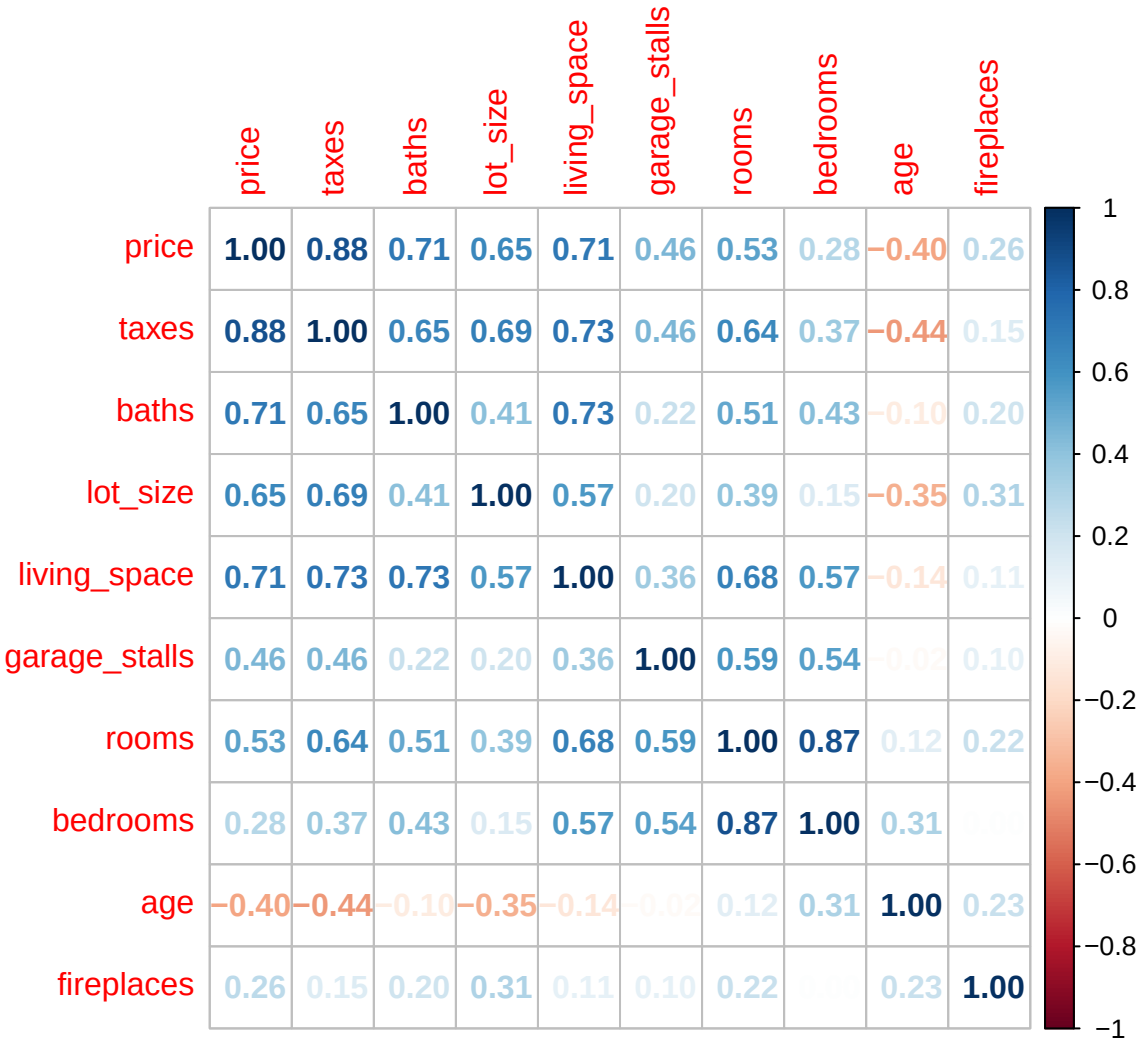
- Interpret the value of $\widehat{\beta}_1$ and explain why we should avoid interpreting $\widehat{\beta}_0$.
- Fit a second model with two covariates: X_1 : weight (kg) and X_2 : abdomen (cm). Compare the estimated effect of weight, does it change when you include X_2? Why do you think this is the case?
- Which of the two models provides a better fit to the data? Compute the coefficient of determination to answer this question.

Q4. In this question, we will simulate taking random samples from a population where we measure Y and three covariates X_1, X_2, X_3. Download the population data, file `data_simulation.csv`. Sample $n = 300$ records and estimate the regression $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$. This is done for you once in the next code chunk. Repeat this simulation 1000 times storing the parameter estimates of each sample.

```
n = 300
values = matrix(NA, ncol =4, nrow = 1000)

for(i in 1:1000){
sampled_indexes = sample(1:10000, size = n)
data_sample = data[sampled_indexes,]

model =  .....
values[i,] = ......

}
```

- Plot the estimates $(\widehat{\beta}_0, \widehat{\beta}_1\widehat{\beta}_2, \widehat{\beta}_3)^i, \quad i = 1, \ldots, 1000$ using histograms.
- Plot the correlation matrix of the results. Use `library(corrplot)` and `corrplot(cor( ... ))`.
- Are $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$ independent?

Q5. The selling price of 24 houses with different attributes was recorded in Erie, Pennsylvania. Inspect the correlation plot below and answer the following questions. The attributes are coded as `taxes` (in thousands of dollars), `baths`, `garage_stalls`, `bedrooms`, `fireplaces`, `rooms` number of baths, garage stalls, bedrooms, fireplaces and rooms in the property, `lot_size`,`living_space`: lot size and living space in thousands of square feet, and `age`: age of property in years.




```
##
## Call:
## lm(formula = price ~ baths, data = houses)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.750 -2.638 -1.194  2.013 12.250
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.875      4.438   3.127  0.00491 **
## baths         17.775      3.728   4.767 9.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.305 on 22 degrees of freedom
## Multiple R-squared:  0.5081, Adjusted R-squared:  0.4858
## F-statistic: 22.73 on 1 and 22 DF,  p-value: 9.268e-05
```

- Which attribute would you select if you can choose only one predictor? Why?
- A model was fitted between Y : `price` and X : `baths`. Use the output shown to answer "By how much

does a house price increase in average with the addition of one bath?".

- Use the output to test the hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 = 0$ using $\alpha = 0.05$. Justify and interpret your conclusion.

Q6: Data analysis

Abalone is a type of mollusk that lives on saltwater, most often found in New Zealand, Australia, South Africa, Japan and North America. Amongst one of the most expensive seafood, abalones are considered a culinary delicacy.

Figure 1: Abalone.

In this exercise, you will work with the `abalone.csv` data, which contains the following measurements:

- Gender: Male (M), Female (F), and Infant (I).
- Length (mm): Longest shell measurement.
- Diameter (mm): Perpendicular to length.
- Height (mm).
- Shucked weight (grams): Weight of meat.
- Rings (integer): +1.5 gives the age in years.

Your goal in this analysis is to create a linear model for Y : shucked weight, which is the weight of an abalone's meat content.

1. **Exploring gender:** Plot the abalone shucked weight per gender using boxplots or histograms. Does there seem to be a difference in weight of female and male abalones?
2. **Exploring rings:** The number of rings plus 1.5 gives the age in years of the abalone. Plot the distribution of age, is it symmetric?
3. **Exploring rings (2):** Create a categorization of the abalone's age into: (5 to 10 years], (10 to 15], (15 to 20], 20+. Plot the distribution of Y per age category. How does the weight distribution change with age?
4. **Correlations:** Compute the linear correlation between Y and all other numeric predictors in the data. What measurement shows the strongest correlation with shucked weight?
5. **Scatterplots:** Create scatterplots of: (1) Y versus X: length, (2): Y versus X : diameter, and (3): Y versus X : height.
6. **Outliers:** there seem to be two abalones in the data that display atypical values of height. Create a boxplot to evidence this claim and explain how they are marked.
7. **Outliers in regression:** Fit a linear model between Y and X : height and report the estimates $\widehat{\beta}_0, \widehat{\beta}_1$. Remove the outliers identified in the previous item and recompute $\widehat{\beta}_1$. What can you conclude?
8. Fit the linear model for Y using the covariates length, diameter, height and rings. Write down the equation of the fitted model and interpret its coefficient of determination.

9. Using the model of item 8, create a 95% prediction interval of a future response for the shucked weight of an abalone that has the measures: length =0.456, diameter =0.351, height = 0.102 and rings = 13.5.