CSC/ST 442 (Fall 2021): HW 5

Instruction

This assignment consists of 2 problems. The assignment is due on **Monday, November, 1** at 11:59pm EDT. Please submit your assignment electronically through the moodle webpage. You are encouraged (but not required) to use RMarkdown to write up your homework solution. To start using Rmarkdown read

- Section 40.2 of Introduction to Data Science
- the RStudio tutorial
- the Rmarkdown cheatsheet.

Problem 1 (20pts)

This problem uses the data set `MWwords` from the `alr4` library. The data give the frequencies of 165 common words in works from four different sources: the political writings of eighteenth century American political figures Alexander Hamilton, James Madison, and John Jay, and the book “Ulysses’ ’ by twentieth century Irish writer James Joyce.

For this problem, we will be concerned with the variables `Hamilton` (the rate per 1000 words at which the word appears) and `HamiltonRank` (the rank of the word; more frequent words having smaller ranks).

The linguist George Zipf suggests that the relationship between the frequency f_i of a word and its rank r_i is approximately of the form $f_i = \alpha r_i^{-\gamma}$ where α and γ are constants, with $\gamma \approx 1$.

A snippet of the data is given below

```
## install.packages("alr4")
library(alr4)
data(MWwords)
MWwords
```

```
## # A tibble: 165 x 8
##   Hamilton HamiltonRank Madison MadisonRank   Jay JayRank Ulysses UlyssesRank
##      <dbl>        <dbl>   <dbl>       <dbl> <dbl>   <dbl>   <dbl>       <dbl>
## 1     91.3            1    93.6           1  67.5       1    57.1           1
## 2     64.6            2    57.8           2  43.9       3    29.9           2
## 3     40.7            3    35.2           3  35.7       4    18.8           5
## 4     24.5            4    27.6           4  45.4       2    27.5           3
## 5     24.4            5    23.0           5  20.8       5    18.8           6
## 6     22.8            6    20.2           6  13.6      10    24.6           4
## # ... with 159 more rows
```

(a) Using only the 50 most frequent words in Hamilton’s work, fit a simple linear regression model with log(`Hamilton`) as the response variable.

(b) Is the empirical law $\gamma \approx 1$ likely to be “correct’ ’ ? Justify your answer.

(c) Repeat part (a) and (b), first for the 75 most frequent words in Hamilton’s work and then for the 100 most frequent words in Hamilton work’s. Is the relationship posited by Zipf still reasonable in these cases ?

(d) Now investigate Zipf law for the Simpson Frequency Dictionary dataset. You will need to do some preprocessing and possibly cleaning of the raw data (in particular the numbers in the `Frequency` column of this dataset; your `stringr` skills might come in handy here)

Problem 2 (20pts)

This problem uses the following Kelley Blue Book dataset on the selling price of used GM cars. The data is available online here

You can read in the data using the following code chunk

```
kbb <- read.csv("https://bit.ly/3GknAkw", header = T, sep = ",")
kbb
```

```
## # A tibble: 804 x 12
##    Price Mileage Make  Model   Trim     Type  Cylinder Liter Doors Cruise Sound
##    <dbl>   <int> <chr> <chr>   <chr>    <chr>    <int> <dbl> <int> <chr>  <chr>
## 1 17314.    8221 Buick Century Sedan 4D Sedan        6   3.1     4 yes    yes
## 2 17542.    9135 Buick Century Sedan 4D Sedan        6   3.1     4 yes    yes
## 3 16219.   13196 Buick Century Sedan 4D Sedan        6   3.1     4 yes    yes
## 4 16337.   16342 Buick Century Sedan 4D Sedan        6   3.1     4 yes    no
## 5 16339.   19832 Buick Century Sedan 4D Sedan        6   3.1     4 yes    no
## 6 15709.   22236 Buick Century Sedan 4D Sedan        6   3.1     4 yes    yes
## # ... with 798 more rows, and 1 more variable: Leather <chr>
```

- Using this data, what is the "best" (linear) model you can find for predicting the sell price ? Note that this is an open-ended problem.
- Writeup a short discussion summarizing the three main choices you made in choosing a suitable model for this problem as well as what you learn when trying to find a suitable model.
- The "best" model I was able to find has a residual standard error of roughly 520$ but my model could very much be overfitting to the data.