Lab 1: Plotting in R; Regression

First Name: Bill **Last Name:** Zhou **NetID:** WZ325

In this lab, you will learn how to create simple plots in R. We will also review the basics of regression we learned in lectures using the ToyotaCorolla data set. Specifically, we will predict the price of a car model based on some of its features such as mileage, fuel type, weight, age and horse power, etc.

Lab 1 will be graded. Upload your work on Gradescope by Feb. 11th, 11:59pm. Regrade requests should be sent via Gradescope.

1 Basic Linear Algebra

R offers linear algebra operations, although its syntax might take a little to get used to. To create a vector, recall that we are using the c() function:

```
> a <- c(1, 3, 2)
```

To create a matrix, we can start by creating a "flat" view of it (i.e. all its elements in a vector), and then rearrange them by specifying the number of columns/rows using the ncol, nrow parameters, e.g:

```
> A1 <- matrix(c(1, 2, 3, 4, 5, 6), ncol = 3)
> A2 <- matrix(c(1, 2, 3, 4, 5, 6), nrow = 2)
```

You may notice that R will start "filling" the matrix column-wise. You may instruct it to fill by row by specifying byrow=TRUE:

```
> A2 <- matrix(c(1, 2, 3, 4, 5, 6), nrow = 2, byrow=TRUE)
```

To obtain the transpose of a vector or a matrix, use the t() function.

```
> t(a)
> t(A1)
```

Vector and matrix addition:

```
> b <- c(2, 8, 9)
> a + b
> B <- matrix(c(1, 2, 3, 4, 5, 6), ncol = 3)
> A + B
```

Multiplying a vector/matrix by a number:

```
> 3 * a
> 2 * A
```

Matrix-vector and matrix-matrix multiplication are possible via the operator %*%, as shown below[1]:

[1] Whenever in doubt, refer to the table of operators on R's website: https://cran.r-project.org/doc/manuals/r-release/R-lang.html#Operators

```
> A %*% a
```

$$\begin{bmatrix}1&2&8\\3&2&9\end{bmatrix}\begin{bmatrix}1\\3\\2\end{bmatrix}=\begin{bmatrix}23\\27\end{bmatrix}.$$

Multiplication of matrices:

```
> A <- matrix(c(1, 3, 2, 2, 8, 9), ncol = 2)
> B <- matrix(c(1, 3, 2, 4), ncol = 2)
> A %*% B
```

$$\begin{bmatrix}1&2\\3&8\\2&9\end{bmatrix}\begin{bmatrix}1&2\\3&4\end{bmatrix}=\begin{bmatrix}7&10\\27&38\\29&40\end{bmatrix}.$$

Matrix-matrix element-wise multiplication:

```
> C <- matrix(c(5, 3, 1, 4), ncol = 2)
> B * C
```

$$\begin{bmatrix}1&2\\3&4\end{bmatrix}\circ\begin{bmatrix}5&1\\3&4\end{bmatrix}=\begin{bmatrix}5&2\\9&16\end{bmatrix}.$$

Inner product of vectors:

```
> a <- c(1, 3, 2)
> b <- c(2, 8, 9)
> sum(a * b)
```

$$\langle a,b\rangle=\sum_{i=1}^{n}a_ib_i=1\times 2+3\times 8+2\times 9=44.$$

Euclidean norm:

```
> sqrt(sum(a^2))
```

$$\|a\|_2=\sqrt{\sum_{i=1}^{n}a_i^2}=\sqrt{1^2+3^2+2^2}=3.741657$$

Below is shown how to concatenate matrices (of compatible sizes). The initials stand for "row" and "column", respectively.

```
> rbind(B, C)
> cbind(B, C)
```

2 Plotting in R

Consider the following function:

$$y=0.1x^3-x^2-6x$$

How do we plot this function in R?
First create a vector **x** using the **seq** funciton and create **y** accordingly:

```
> seq(from = -10, to = 20, by = 0.01)
```

Use the plot function to plot x-y. Call help(plot) and help(plot.default) to see the parameter settings. Specifically, pay attention to what the type, xlim, ylim, main, xlab, ylab arguments mean.
For now we would like to create a line rather than scatter plot of y vs. x. We will give a title "x-y" to the plot, set the range of the x-label to be exactly [−10, 20], and name the x-label "x axis" and the y-label "y axis". What should the above arguments be in this case?

plot (x, y, main = 'x-y', xlim = c(-10,20), xlab = 'x-axis', ylab = 'y-axis'))

Suppose you want to plot another function and put the two plots you created in a single page, use this command:

```
> par(mfrow = c(2, 1))
```

3 Linear Regression

Download the `ToyotaCorolla.csv` data set from blackboard. This data set is in .csv (comma-separated values) text format. it contains the information of more than 1400 trade-in vehicles. Our goal is to predict the price of a vehicle for resale in the market using the variables known like age and mileage of it.

You may now read the .csv file into R as a data.frame object named `corollas` using `read.table` or `read.csv`. If you use `read.table`, what are the arguments you use for `header` and `sep`?

I used read.cvs (filepath , head= True)

Reading the dataset correctly should give you an R data frame. As in Lab 0, you may use the `tibble` library for pretty printing and summaries of your data:

```
> library("tibble"); corollas <- as_tibble(corollas)
```

How many rows does the dataset have? How many variables? What function will you use to see the names of these variables?

1436 rows , 38 variables , names(data)

How do you check for missing values? How many missing values are there in the data set?

is.na(data) 0

We will not include all variables as predictors in our regression model. For now, we're interested only in the following variables of interest: `Price`(the response variable), `Age_08_04`, `KM`, `Fuel_Type`, `HP`, `Met_Color`, `Doors`, `Quarterly_Tax` and `Weight`. (Check the **ToyotaCorollaMeta.xls** file for what these variables are.)

Choose the variables we are interested in to create a new object:

```
> corollas2 <- corollas[, c("Price","Age_08_04", "KM", "Fuel_Type", "HP",
"Met_Color", "Doors", "Quarterly_Tax", "Weight") ]
```

Equivalently, if using `dplyr` and having converted `corollas` to a `tibble`:

```
> corollas2 <- select(corollas, Price, Age_08_04, KM, Fuel_Type, HP, Met_Color,
Doors, Quarterly_Tax, Weight)
```

Check whether the categorical predictors have been read in correctly as "factor" type:

```
> is.factor(corollas2$Fuel_Type)
> is.factor(corollas2$Met_Color)
```

What function will you use to change them to categorial data?

corollas$ Met_Color = as.factor (corollas$Met_Color)

Make sure the Fuel_Type column takes 3 categorical values. How can this be accomplished without manually inspecting all the rows?

summary (corollas2 $ Fuel _Type)

We are now ready to use R's linear regression utilities. We will be using lm() from the stats package, which should be included by default. Typing help(lm) will give you an overview of the package; to read about the syntax that we will be using momentarily, navigate to the "Details" section of the help page.
Let us start by fitting the linear regression model:

```
> corollasLM = lm(formula=Price~., data=corollas2)
> summary(corollasLM)
```

This syntax means that the Price variable is the outcome and all the other variables in corollas2 should be used as predictors.

Look at the output of the summary function. For now you need only pay attention to the first column of the table. These are the β coefficients in the model.
Note we have 9 predictors instead of 8. The Fuel_Type variable becomes two predictors: Fuel_TypeDiesel and Fuel_TypePetrol. Why?

We transform different catugorical data into
one-hot (dummy) data
Since we have three classes in total for Fuel-Type
Thus we need two predictor Fuel_TypeDiesel &
Fuel_TypePetrol to indicate whether the Fuel Type
is Diesel/Petrol,

Why don't we have another predictor Fuel_TypeCNG?

When both Petrol and Diesel are 0, it means the Fuel Type has to be CNG, and the factor gonna actually included in the total intercept

Report your $\hat{\beta}$ coefficients. ($\hat{\beta}_{\text{intercept}}$ =? $\hat{\beta}_{\text{Age\_08\_04}}$ =? etc.)

$\beta_{\text{intercept}} = -7.699 \times 10^3$
$\beta_{\text{Age-08-04}} = -1.226 \times 10^2$
$\beta_{\text{KM}} = 1.703 \times 10^{-2}$
$\beta_{\text{Met Color}} = 3.343 \times 10^1$
$\beta_{\text{Doors}} = -7.81 \times 10^1$
$\beta_{\text{Quarterly-Tax}} = 1.229 \times 10^1$

$\beta_{\text{Fuel-Diesel}} = 5.26 \times 10^2$
$\beta_{\text{Fuel-Petrol}} = 2.422 \times 10^3$
$\beta_{\text{HP}} = 2.258 \times 10^1$
$\beta_{\text{Weight}} = 1.976 \times 10^1$

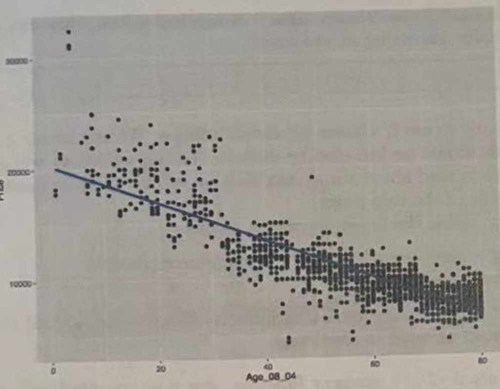


Figure 1: Plotting a linear model with ggplot2.

What is the Standard Error (SE) of the coefficient estimate for the predictor Weight?

1.188

What are the values of RSE and R^2 statistics for your regression model? What is the average of the response values $\bar{y}$ and how do you compute it? What can you say from these statistics?

0.868

$\bar{y}$=10730, I compute by $\overline{\text{price}} - \overline{\text{Residual}}$ Since average residual ~~is super~~ small, I will say generally this fit makes sense.

The ggplot2 library offers a nice way to use the lm() function (among others) for exploring data. For example, if you suspect that Age_08_04 is the predictor that mainly influences Price, you can use the following command to plot both variables, fit a linear model of the form $y = \beta_0 + \beta_{\text{Age\_08\_04}} x$ and show the line fit:

```
> library("ggplot2")
> ggplot(corollas2, aes(x=Age_08_04, y=Price)) +
+   geom_point() +
+   stat_smooth(method="lm", color="red")
```

This will produce an image like Figure 1.

3.1 Outliers vs. high leverage points

"Unusual" points can easily mess up a linear model, even if the variables obey the linearity assumption for the majority of samples. There are at least two possible sources of error, in terms of samples:

- **outliers**, whose response variable is unusual given the predictor(s),

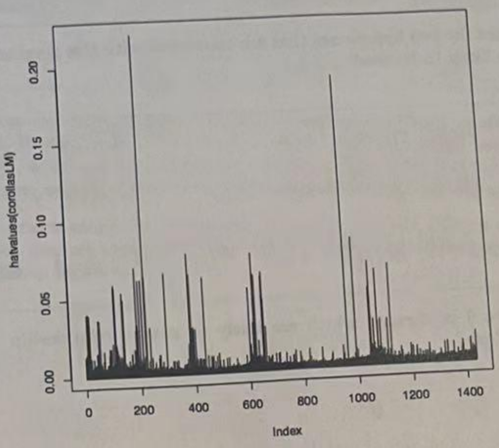


Figure 2: Sample leverage

- **high-leverage points,** *whose predictors are unusual.*

The concept of leverage is in fact useful for both diagnoses. Data samples with high leverage exercise more "control" over the regression line than samples with lower leverage, and removing them may be beneficial to the analysis. In R, they are available via the `hatvalues()` function, which outputs a vector of leverage scores. We can generate a histogram from them as in Figure 2.

```
> lev <- hatvalues(corollasLM)
> plot(lev, type="h")
```

On the other hand, to decide whether or not a data point is a candidate outlier point, we can use the *studentized residual test*, which produces the studentized residuals:

$$t_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - \texttt{lev}_i}},$$

where e_i are the residuals. In R, this is available via the `rstudent()` function:

```
> studRes <- rstudent(corollasLM)
```

Via the familiar 3σ rule of the normal distribution, we may label a data point as "outlier" if its studentized residual satisfies $|t_i| > 3$. In your take-home questions, you will explore how the removal of outliers and high-leverage points affects linear regression.

4 Take Home Questions

1. What is the p-value for the whole linear model?

p-value $< 2.2 \times 10^{-6}$

What are the two hypotheses that are associated with this p-value? Which hypothesis is more likely to be true?

① All these predictars are effective to the result
② All these preditors aren't effective to the result (null hypothesis)
I would reject null-hypothesis and take ①

2. Out of the 9 predictors, which are likely to have a relationship with the response variable Price?

If we set $p = 0.05$ as hypothesis threshold
Age_08_04, KM, Fuel_typePetrol, HP, Doors, Quataly_Tax
~~and~~ weight have relationship with Price

3. In our regression model, what are the units for the different components of β?

Price / Unit of that predictor itself (if it has a unit)

4. How much would the price be affected by a unit change in the weight of a car? Is this rate (dollar/kg) affected by other factors such as color of the car, age of the car, etc.?

+1 kg, +19.76 bucks value.
No, since variables are independent to each other in linear model

5. In our regression model, we are using multiple predictors like the number of doors, age and weight to predict the resale value of a Toyota Corolla. Assume there are two 4-door cars that weigh 1200 kilograms and 1000 kilograms respectively, and two 2-door cars that weigh 1200 kilograms and 1000 kilograms respectively. Can the model we are using capture the situation where the price difference between the two 4-door cars is larger than that between the two 2-door cars? Explain your answer.

No, variables (preditors) are independent to each other, in our linear model, we don't have terms like $\beta_{ij} \cdot x_i \cdot x_j$ terms in our model to represent the relation between different preditors.

6. How many parameters are there in a (standard) linear regression model with 4 continuous predictors? Include all unknown quantities that characterize the model and have to be estimated.

4+1 = 5 , ~~tone~~ one beta for each predictor and one for intercept

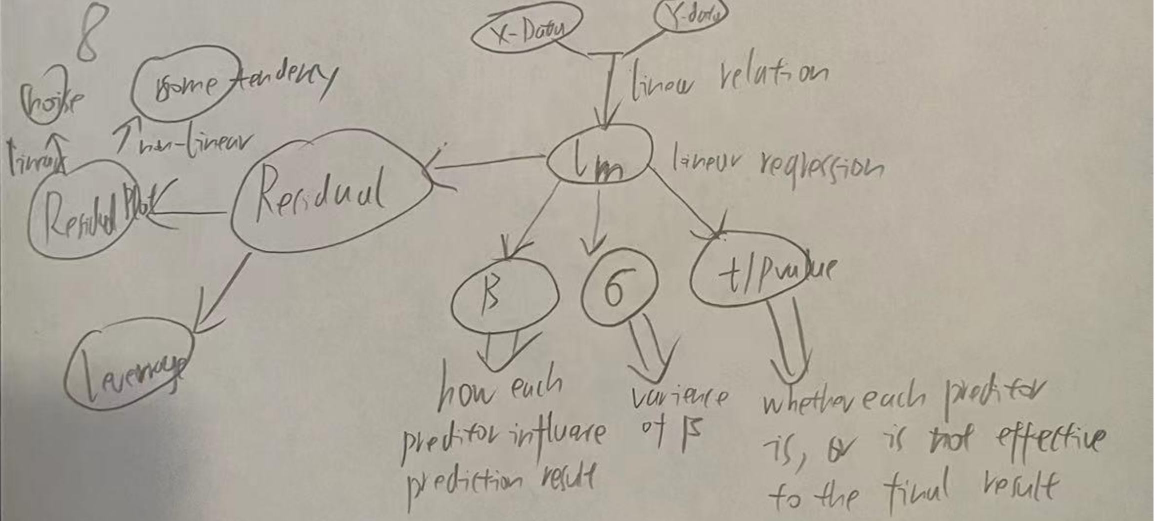
7. Find outliers and high leverage points after fitting the linear model to the data in `corollas2`. How many outliers are there? How many data points have leverage higher than $10 \cdot \frac{p+1}{n}$, where p is the number of predictors? *You might find it helpful to augment the data frame with leverage scores and studentized residuals using* `dplyr`.

8. Make a **concept map**[2] for the topics covered by this lab. While there is no single right answer, you are expected to demonstrate that you understand the relationships between various topics in the lab.

7(1) there are four outliers if we set |5| as residual outlier threshold

8(2) five leverage higher the $\frac{10(p+1)}{n}$ according to my plot

8



[2] https://en.wikipedia.org/wiki/Concept_map