ASSIGNMENT 1

ENGIN 242 - Applications in Data Analytics

Aditya Peshin
Enrollment Number: 3035280249

Q1. Rescaling a linear regression problem

Ans 1.

IEOR 242 - Assignment 1

Q1.

Part a. Given the two Linear Regression models

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

and $$Y = \alpha_0 + \sum_{j=1}^{p} \alpha_j Z_j + \varepsilon$$ where $Z_j = \lambda_j X_j$

We know we can estimate the values of $\hat{\beta}$ & $\hat{\alpha}$ by:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$= X^{-1} \underbrace{X^{T^{-1}} X^T}_{I} Y \quad \text{[Matrix Identities]}$$

$$= X^{-1} Y \qquad -(1)$$

$$\hat{\alpha} = (Z^T Z)^{-1} Z^T Y$$

$$= Z^{-1} \underbrace{Z^{T^{-1}} Z^T}_{I} Y$$

$$= Z^{-1} Y \qquad -(2)$$

Now ① can be re writen as

$$\hat{\beta} = X^{-1} Y$$

$$X\hat{\beta} = Y \quad \text{[Pre multiplying with } X] \quad - ③$$

② can be rewritten as

$$\hat{\alpha} = \cancel{X} Z^{-1} Y$$

$$\text{or } Z\hat{\alpha} = Y \quad - ④$$

Equating ③ & ④ we get

$$X\hat{\beta} = Z\hat{\alpha}$$

Expanding, we get

$$\beta_0 + \sum_{j=1}^{p} \beta_j X_j = \alpha_0 + \sum_{j=1}^{p} \lambda_j \alpha_j X_j$$

Pairwise comparisson of coefficients yeilds

$$\beta_0 = \alpha_0$$

$$\beta_j = \alpha_j \lambda_j \quad \text{or} \quad \alpha_j = \beta_j / \lambda_j$$

b) No, as shown in the previous problem, rescaling the values of X_j with λ_j did not affect the equation, as by multiplying X_j with λ_j, its corresponding β_j gets divided by the same λ_j, leading to an unchanged equation

Q2. (Recentering Problem)

Part 2 (Offsetting/Recentering the model)

Old coordinate system $\rightarrow x_i, y_i$

New C.S $\rightarrow x_i - \bar{x}, y_i - \bar{y}$

Now, we know

$$\hat{\beta} = X^{-1} Y$$

$$X\hat{\beta} = Y \quad -(1)$$

And

$$\alpha = Z^{-1} W$$

$$Z\alpha = W \quad -(2)$$

Equating (1) & (2) we get

$$\hat{\beta}_0 + \sum_{j=1}^{p} \beta_j X_j = \hat{\alpha}_0 + \sum_{j=1}^{p} \alpha_j Z_j \qquad [Z_j = X_j - \bar{X}]$$

$$= \underbrace{\hat{\alpha}_0 - \sum_{j=1}^{p} \alpha_j \bar{X}}_{\text{Constant}} + \sum_{j=1}^{p} \alpha_j X_j \quad \text{(splitting eqn into 2 parts)}$$

Pairwise comparison of coefficients we get

$$\alpha_j = \beta_j \quad , j \in (1..p) \quad -(3)$$

$$\beta_0 = \alpha_0 + \sum_{j=1}^{p} \alpha_j \bar{X} \quad -(4)$$

Substituting ③ in ④, we get

$$\beta_0 = \alpha_0 - \sum_{j=1}^{p} \beta_j \bar{X}$$

$$\alpha_0 = \beta_0 + \beta_j \bar{X}$$

$$= \bar{y}$$

With respect to the new coordinate system, the new value of α_0 is:

$$\alpha_0' = \alpha_0 - \bar{y}$$

$$= \bar{y} - \bar{y} = 0$$

b) This is true because we are just shifting the graph laterally with respect to the origin and not squishing (rescaling) any of the variables.

c) $$Y_{new} = \alpha_0 + \sum_{j=1}^{p} \alpha_j (x_{new(j)} - \bar{X})$$

$$= \alpha^T (X_{new} - \bar{X})$$

1-D vector of $\bar{x}$ only

Problem 3: Forecasting Jeep Wrangler Sales

Part a)

The Linear Regression equation produced by my model is:

*WranglerSales = 257.86 * WranglerQueries - 952.18*

The coefficient 257.86 represents that around 258 wranglers are sold with every new google query about the car. The reason why I only chose this column to forecast the sales of the car is because it was the only one with a p-value less than 0.05

I started out by modelling the sales of the car with respect to all the variables, and found that the VIF for the unemployment rate and CPI.All were very high (~70).

I dropped the higher among the two, and ended up modelling with respect to CPI All, Wrangler Queries and CPI Energy, and found that all variables had acceptable VIF Values. However, the p-value for Unemployment and CPI Energy were not satisfactory (~ 0.60 and ~0.41 respectively), meaning that the model was still not perfect.

I then dropped the unemployment rate due to its higher p-value, and modelled with the remaining two variables Wrangler Queries and CPI.Energy. Both variables had acceptable VIF values, but the CPI Energy term still had an unacceptable p value of 0.3, i.e the model was not confident in its relevance to predicting Wrangler Sales.

So, I modelled the problem again, with Wrangler Queries as the only feature, and achieved acceptable values of VIF and p-value. Throughout all the testing, R had continuously marked this variable with three stars, indicating its confidence in the ability of this variable to predict Wrangler Sales.

The positive correlation between Wrangler Sales and Wrangler Queries is expected, because an increased interest in the car would lead to increased sales volumes for that month.

The value of R-squared in the final model is around 0.7895, which shows that this is a good model to predict the sales data. It isn't an excellent fit, but it is a good enough fit to understand the relationship between the feature and the output variable. Also, testing the R-Squared values for the earlier models, I realized that the value of R-Squared remained practically unchanged, reaffirming my earlier statement that this is the only feature that matters.

Part b)

i) The new regression equation is:

Wrangler Sales = Unemployment		* 845.80
+	WranglerQueries	* 175.69
+	CPI.Energy	* -25.28
+	CPI.All	* 317.32
+	MonthFactorAugust	* -62.76
+	MonthFactorDecember	* -175.82
+	MonthFactorFebruary	* -1078.14
+	MonthFactorJanuary	* -3262.64
+	MonthFactorJuly	* -176.09
+	MonthFactorJune	* 313.29
+	MonthFactorMarch	* -173.75
+	MonthFactorMay	* 1894.71
+	MonthFactorNovember	* -1660.69
+	MonthFactorOctober	* -776.15
+	MonthFactorSeptember	* -945.17

A positive coefficient in the dummy monthfactor variables indicates an inclination to buy wrangler cars in that month, whereas a negative coefficient indicates a disinclination to buy these cars in that month, with the value referring to the magnitude of the inclination/disinclination

ii) The training set R squared is 0.8698, which is characteristic of an excellent fit. R identified the Wrangler Queries, Month factor January and Month Factor May as the most significant variables.

iii) It definitely improves the quality of the model as accounting for the seasonal nature of the demand leads to a much higher value of R squared, which leads to a better fit.

iv) The other way I would model seasonality is to combine the month and year variables into a combined variable, reflecting the seasonal trends as a function of time. However, this approach might need to be a non linear approach, as if we opt for a linear regression, the seasonal “wavy” nature of the demand would get averaged out to zero.

Part c)

I decided to build a model using the monthfactor and the wrangler queries features, and I ended up with a model with training set R squared of 0.8623, with the significant features being monthfactor January and WranglerQueries.

The OSR squared of the model was 0.65, which is a considerable improvement over the earlier models.
However, the model is still not very useful, as its predictive power is not very high. The intuition is that I might be missing another critical variable that is impacting car sales.

Part d)

I decided to add monthly oil prices to my data set in order to find out if they impact the sales of the cars. I figured that it should show a negative correlation, hence my choice.

After performing the modelling and analysis, I determined that the improvement in the OSRsquared was 0.02, which is not a significant improvement. The training R squared remained the same.
The p value for the new feature was 0.39, indicating that the model was not confident in the ability of the gasoline price to predict the car sales.

This indicates that the new feature is not very helpful in improving the quality of the predictions, which means that there is still some unaccounted variable that influences the wrangler jeep sales.

R - Code:

```
# Homeowork Assignment 1 Question 3
# Part a

library(dplyr)
library(ggplot2)
library(GGally)
library(car)

#loading in the data
wrangler_orig <- read.csv("Wrangler242-Fall2019_gasPrice.csv")
View(wrangler_orig)

#Look for any corelation in the given data
ggscatmat(wrangler_orig, columns = 2:8, alpha = 0.8)

#Splitting into training and testing set
training_set <- filter(wrangler_orig, Year<= 2015)
testing_set <- filter(wrangler_orig, Year >= 2016)

model1 <- lm(WranglerSales ~ Unemployment + WranglerQueries + CPI.Energy + CPI.All, data
= training_set)
summary(model1)
vif(model1)
# model 1 has a very high VIF for CPI all and Unemployment
# removing CPI all

model2 <- lm(WranglerSales ~ Unemployment + WranglerQueries + CPI.Energy, data =
training_set)
summary(model2)
vif(model2)
# the VIF for all the variables are acceptable, however the p value for Unemployment is not

model3 <- lm(WranglerSales ~ WranglerQueries + CPI.Energy, data = training_set)
summary(model3)
vif(model3)
# the VIF for both variables are acceptabel, however the p value for CPI.Energy

model4 <- lm(WranglerSales ~ WranglerQueries, data = training_set)
summary(model4)
vif(model4)
# the p value for Wrangler Queries are acceptable,
```

```
# the value of R squared has not changed much after removing
# the other variables, so my final model only will use Wrangler Queries
# to model the linear regression

#testing model 1 for its OSR Squared
SalesPrediction1 <- predict(model1, newdata=testing_set)

SSE1 = sum((testing_set$WranglerSales - SalesPrediction1)^2)
SST1 = sum((testing_set$WranglerSales - mean(training_set$WranglerSales))^2)
OSRsquared1 = 1 - SSE1/SST1
#It is 0.45, not very helpful

#testing model 2 for its OSR Squared
SalesPrediction2 <- predict(model2, newdata=testing_set)

SSE2 = sum((testing_set$WranglerSales - SalesPrediction2)^2)
SST2 = sum((testing_set$WranglerSales - mean(training_set$WranglerSales))^2)
OSRsquared2 = 1 - SSE2/SST2
#It is 0.57, improved but not by much

#testing model 3 for its OSR Squared
SalesPrediction3 <- predict(model3, newdata=testing_set)

SSE3 = sum((testing_set$WranglerSales - SalesPrediction3)^2)
SST3 = sum((testing_set$WranglerSales - mean(training_set$WranglerSales))^2)
OSRsquared3 = 1 - SSE3/SST3
#It is 0.57, no improvement, not useful

#testing model 4 for its OSR Squared
SalesPrediction4 <- predict(model4, newdata=testing_set)

SSE4 = sum((testing_set$WranglerSales - SalesPrediction4)^2)
SST4 = sum((testing_set$WranglerSales - mean(training_set$WranglerSales))^2)
OSRsquared4 = 1 - SSE4/SST4
#It is 0.53, R squared has become worse




# Question 3 Part b (Considering Seasonality)
model5 <- lm(WranglerSales ~ Unemployment + WranglerQueries + CPI.Energy + CPI.All +
MonthFactor, data = training_set)
summary(model5)
```

```
vif(model5)


# Question 3 Part c (Building a better model)
model6 <- lm(WranglerSales ~ WranglerQueries + MonthFactor, data = training_set)
summary(model6)
vif(model6)

#testing model 6 for its OSR Squared
SalesPrediction6 <- predict(model6, newdata=testing_set)

SSE6 = sum((testing_set$WranglerSales - SalesPrediction6)^2)
SST6 = sum((testing_set$WranglerSales - mean(training_set$WranglerSales))^2)
OSRsquared6 = 1 - SSE6/SST6
#It is 0.6499, R squared has become worse



# Question 3 part d (adding oil price data)
model7 <- lm(WranglerSales ~ WranglerQueries + MonthFactor + GasolinePrice, data =
training_set)
summary(model7)
vif(model7)

#testing model 7 for its OSR Squared
SalesPrediction7 <- predict(model7, newdata=testing_set)

SSE7 = sum((testing_set$WranglerSales - SalesPrediction7)^2)
SST7 = sum((testing_set$WranglerSales - mean(training_set$WranglerSales))^2)
OSRsquared7 = 1 - SSE7/SST7
#It is 0.6499, R squared has become worse
```