Nicolas Kardous

IND 242 Homework 1

Problem 3

**Part a)**

We start with a linear regression of all four variables and see how it relates to sales.

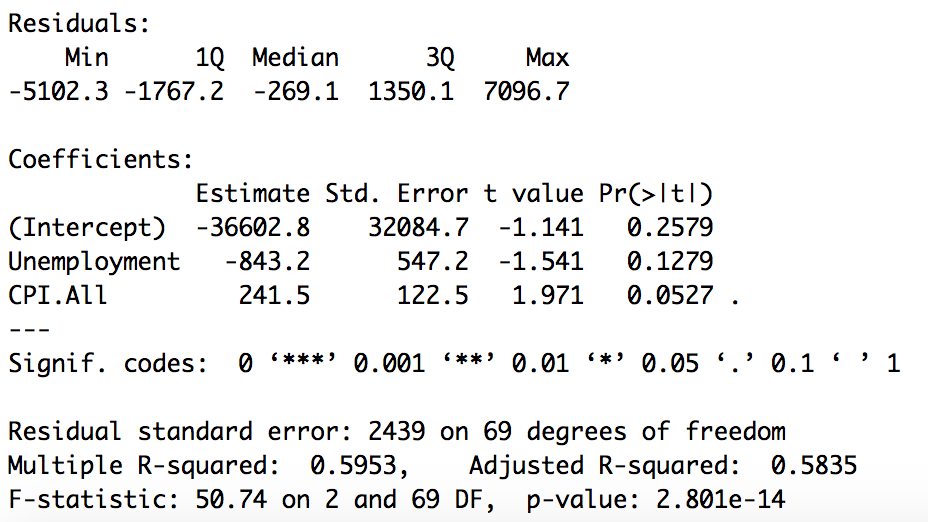
So we start with **Unemployment, WranglerQueries, CPI.All and CPI.Energy:**

The correlation between WranglerQueries and CPI.All is 0.8546 and the correlation between WranglerQueries and Unemployment is -0.873. Both of these values are highly positively or negatively correlated and so we will remove WranglerQueries from our regression model. Additionally, due to the multicollinearity with WranglerQueries and Unemployment, and the large Variance inflation factors (VIF), this would be a good decision. Also, the p value for WranglerQueries was found to be 7.15e-06, which is very small.

After getting rid of the WranglerQueries variable, we are left with **Unemployment, CPI.All and CPI.Energy:**

We first try removing CPI.All because it was not close to the 95% confidence interval threshold. For Unemployment, CPI.All and CPI.Energy we have an R2 value of 60.1%. If we get rid of CPI.All and find the regression for Unemployment and CPI.Energy, we have an R2 value of 58.6%. Finally, if we choose to have the variables Unemployment and CPI.All, the R2 value we have is 59.5%. Because all these values of R2 are relatively similar, we need to find another indicator. When running regression with (Unemployment, CPI.All), (Unemployment, CPI.Energy) and (CPI.All, CPI.Energy) we find that 2 out of 3 of them will have a really low p value for one of the independent variables. The only group of variables that don’t have a low p value are Unemployment and CPI.All. **Thus, our final model is Unemployment and CPI.All**

The values we have in our linear regression for Unemployment and CPI.All are as follows:

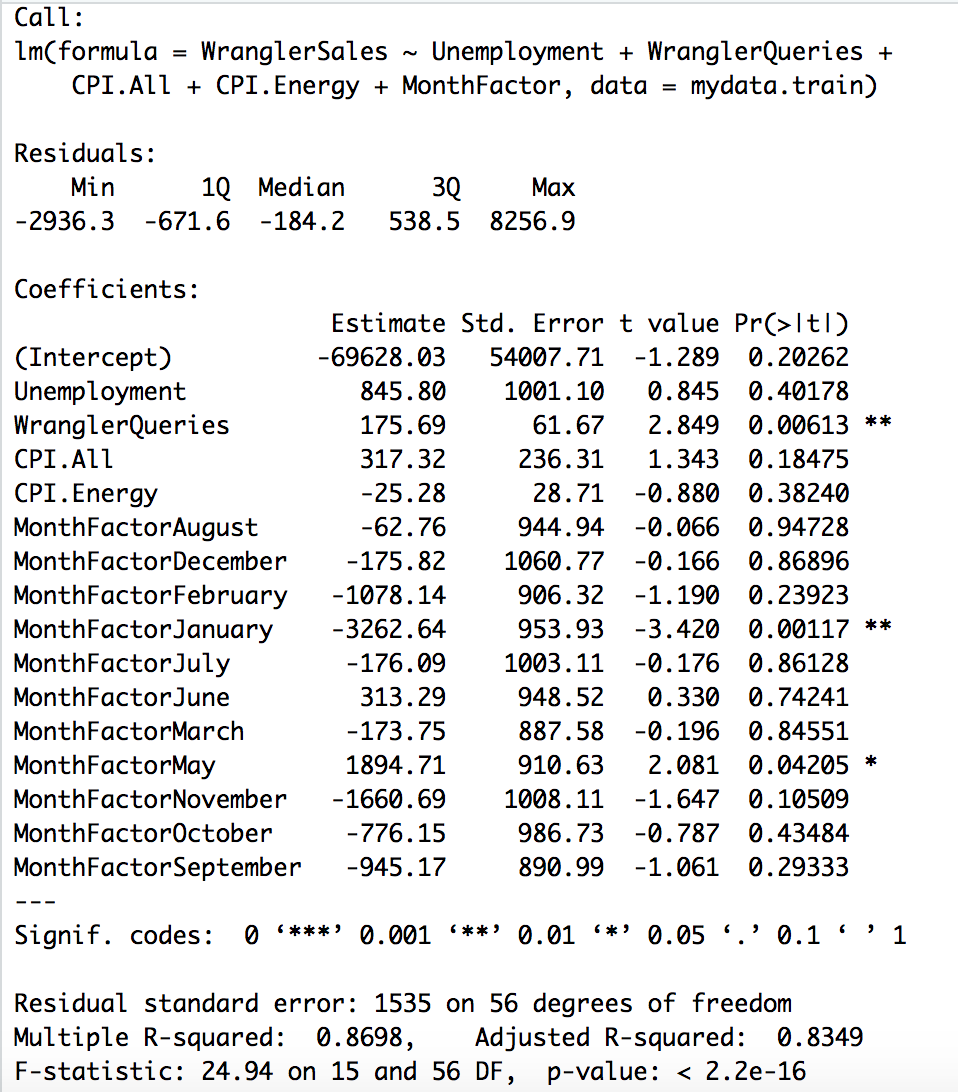


The regression equation we have is:

Even though R2 for Unemployment, CPI.All and CPI.Energy is 60.6% and higher than the other R2 values, what drove me away from using a model with 3 variables was that the coefficient for Unemployment was positive. However, when looking at models for (Unemployment, CPI.All) and (Unemployment, CPI.Energy), we found that Unemployment had a negative coefficient. Thus, it makes sense in our final model that we have a negative coefficient for Unemployment and a positive coefficient for CPI.All. This should also make sense because we could assume that when Unemployment rates are higher, people are less likely to be purchasing Wranglers. Similarly, when the Consumer Price Index is higher or (positive), there is a positive correlation to Wrangler Sales.

Given this, I believe the model predicts the training set observations very well. We could justify the model on a quantifiable basis by finding the SSE, SST and the OSR2 for it comparing it to the testing data. In this case, we found an OSR2 value of 0.5619.

**Part b)**

1. For our new model, we start with the 5 independent variables Unemployment, CPI.All, CPI.Energy, WranglerQueries and MonthFactor. Similar to part a) of the question, the correlation between WranglerQueries and CPI.All is 0.8546 and the correlation between WranglerQueries and Unemployment is -0.873. Both of these values are highly positively or negatively correlated, which would bring issues of multicollinearity and a high VIF. However, when we get rid of WranglerQueries from our model, the R2 value we have only decreased by 0.001. Additionally, the p values we have also decreases compared to the p values from the model with four variables and the MonthFactor. Similarly, if we choose to keep WranglerQueries and get rid of Unemployment, our R2 value and p values also decreases.

From this, we have the optimal model of Unemployment, WranglerSales, CPI.All, CPI.Energy and MonthFactor.

From these coefficients, we have the following regression equation:

Looking at the coefficient for the MonthFactor dummy variables, we see that they are all negative except for the month of June. This shows that there are more sales in a given month compared to the baseline month, if we keep other variables fixed. What is also interesting to note is that our coefficient for Unemployment is now positive.

1. The training set R2 for the new model is 0.8698, which is much more than the 0.595 R2 value from part a). To find which variables are significant, we need to look at which variables have a high p value. Looking at the chart above, the significant variables are: Unemployment, CPI.All, CPI.Energy, MFAugust, MFDecember, MFFebruary, MFJuly, MFJune, MFMarch, MFNovember, MFOctober and MFSeptember.
2. Adding the independent variable MonthFactor improves the quality of the model. This is because of the overall higher R2 value of 0.8698. We are able to get a significantly better fit without multicollinearity. Moreover, the other independent variables have a higher p value compared to the p values of the model from part a). For example, the p value of Unemployment is 0.4. It is also a better model because we are able to keep our other independent variables while maintaining that they are significant. We looked at the possibility of deleting WranglerSales, but that would reduce our R2 by a substantial amount.
3. A reasonable alternative the model seasonality is having a model according to fall, winter, spring and summer. This is because a number of the dummy variables (months) are not statistically significant, and so this could help us get around that.

**Part c)**

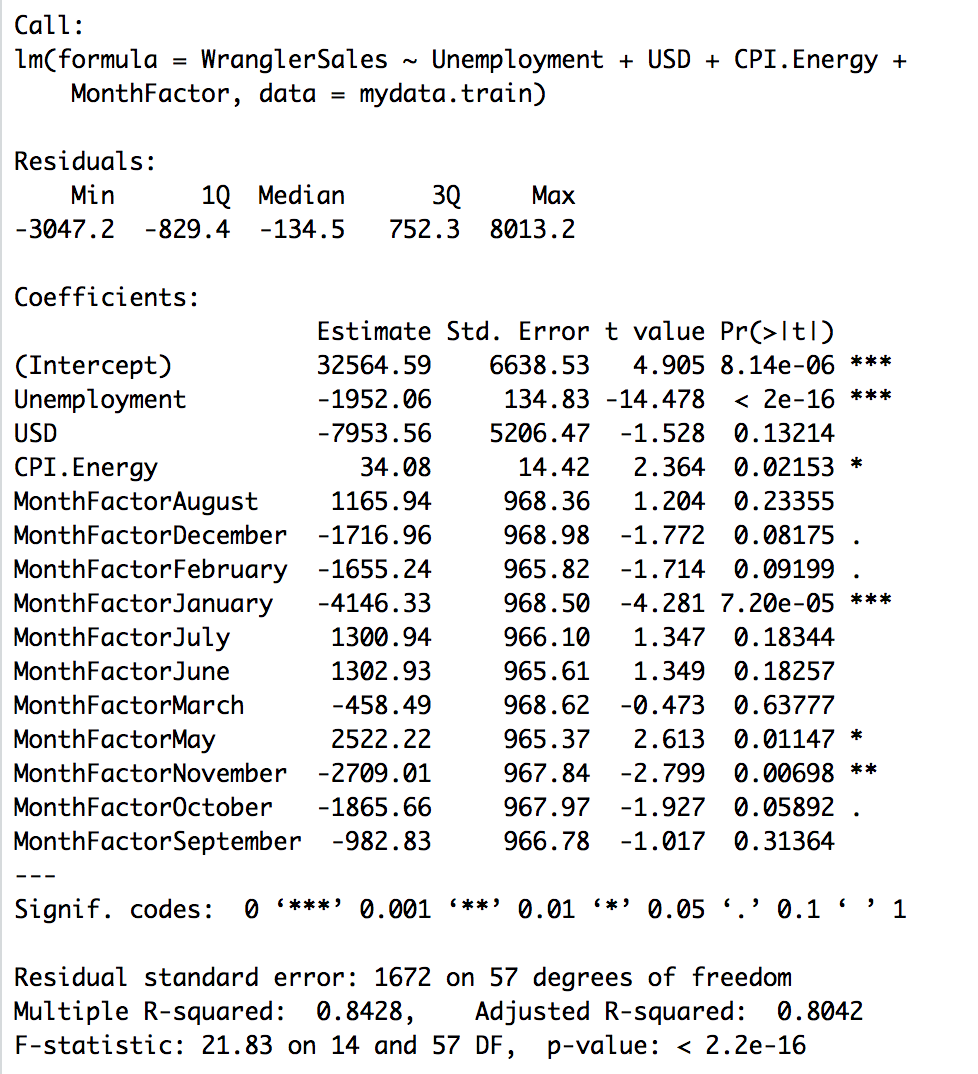
The final model I am using includes the independent variables Unemployment, CPI.Energy and Month Factor. I started off with the original model found in part c with the four variables and MonthFactor. I first got rid of WranglerQueries because it had a high correlation to Unemployment and CPI.All, and it contributed to multicollinearity. Secondly, I was stuck with a model where the coefficient for Unemployment was positive. When I got rid of CPI.All, the model I was left with gave a negative coefficient for Unemployment. Thus, I’m happy with the model of using Unemployment, CPI.Energy and MonthFactor, as it has a high R2 value of 0.8364 and also has high p values. From this, we also found an OSR2 value of 0.792.. This model will be useful to Jeep/FCA because it has an OSR2 value > 0, which means that we improved compared to the baseline model, and thus the model is useful.

**Part d)**

The monthly variables I am using is the value of the USD compared to the British pound. I am using this because I am trying to see if there is a correlation between the strength of the USD and purchasing habits for Wrangler. Additionally, with Brexit coming up, I am trying to see if a depreciation in the British pound would affect Wrangler sales.

The values of our new model are shown below.

For this new model, we are using the variables Unemployment, USD (USD value compared to 1 GBP), CPI.Energy and MonthFactor. I believe the new variable added some predictive value because it shows that the Unemployment variable is no longer statistically significant with a low p value. However, our R2 value reduced from 0.869 to 0.8428. Additionally, when we find the new value for OSR2, we get 0.633, which is lower than the OSR2 value of 0.792 found in part c).

Code:

# IND242HW1 Problem 3

# Nicolas Kardous

# Part a)

mydata = read.csv("Wrangler242-Fall2019.csv")

mydata.train <- filter(mydata, Year <= 2015)

mydata.test <- filter(mydata, Year > 2015)

LR1a <- lm(WranglerSales ~ Unemployment + WranglerQueries + CPI.All + CPI.Energy, data=mydata.test)

cor(mydata.train$WranglerQueries,mydata.train$Unemployment)

cor(mydata.train$WranglerQueries,mydata.train$CPI.All)

cor(mydata.train$WranglerQueries,mydata.train$CPI.Energy)

summary(LR1a)

LR1b <- lm(WranglerSales ~ Unemployment + CPI.All + CPI.Energy, data=mydata.train)

summary(LR1b)

LR1c <- lm(WranglerSales ~ Unemployment + CPI.All, data=mydata.train)

summary(LR1c)

# compute OSR^2

Predictions <- predict(LR1c, newdata=mydata.test)

# this builds a vector of predicted values on the test set

SSE = sum((mydata.test$WranglerSales - Predictions)^2)

SST = sum((mydata.test$WranglerSales - mean(mydata.train$WranglerSales))^2)

OSR2 = 1 - SSE/SST

# Part b)

LR2a <- lm(WranglerSales ~ Unemployment + WranglerQueries + CPI.All + CPI.Energy + MonthFactor, data=mydata.train)

summary(LR2a)

cor(mydata.train$WranglerQueries,mydata.train$MonthFactor)

cor(mydata.train$WranglerQueries,mydata.train$CPI.All)

cor(mydata.train$WranglerQueries,mydata.train$CPI.Energy)

LR2b <- lm(WranglerSales ~ Unemployment + CPI.All + CPI.Energy + MonthFactor, data=mydata.train)

summary(LR2b)

LR2c <- lm(WranglerSales ~ Unemployment + CPI.All + MonthFactor, data=mydata.train)

summary(LR2c)

# Part c)

LR3a <- lm(WranglerSales ~ Unemployment + WranglerQueries + CPI.All + CPI.Energy + MonthFactor, data=mydata.train)

summary(LR2a)

LR3b <- lm(WranglerSales ~ Unemployment + CPI.All + CPI.Energy + MonthFactor, data=mydata.train)

summary(LR3b)

LR3c <- lm(WranglerSales ~ Unemployment + CPI.Energy + MonthFactor, data=mydata.train)

summary(LR3c)

# Compute OSR^2

Predictions <- predict(LR3c, newdata=mydata.test)

# this builds a vector of predicted values on the test set

SSE = sum((mydata.test$WranglerSales - Predictions)^2)

SST = sum((mydata.test$WranglerSales - mean(mydata.train$WranglerSales))^2)

OSR2 = 1 - SSE/SST

# Part d)

LR4 <- lm(WranglerSales ~ Unemployment + USD + CPI.Energy + MonthFactor, data=mydata.train)

summary(LR4)

# Compute OSR^2

Predictions <- predict(LR4, newdata=mydata.test)

# this builds a vector of predicted values on the test set

SSE = sum((mydata.test$WranglerSales - Predictions)^2)

SST = sum((mydata.test$WranglerSales - mean(mydata.train$WranglerSales))^2)

OSR2 = 1 - SSE/SST