**JEREMY PETERSON DUTOYA**

**PR. MENG-SHEN HSIEH**

**PMBA-8358-OLA: DATA-DRIVEN STRATEGIES FOR BUSINESS**

*Linear Regression Model Homework Assignment*

An important application of predictive models is understanding sales. In this assignment, we will predict the monthly sales in the United States of the Hyundai Elantra car. The variables are defined below:

| **Variable** | **Description** |
| --- | --- |
| Month | The observation month (1=January, 2=February, 3=March, etc.) |
| Year | The observation year |
| ElantraSales | The number of the units of the Hyundai Elantra sold in the U.S. in the given month and year |
| Unemployment | The estimated unemployment rate as a percentage in the U.S. in the given month and year |
| Queries | A (normalized) approximation of the number of Google searches for “Hyundai Elantra” in the given month and year |
| CPI.All | The consumer price index for all products for the given month and year. This is a measure of the magnitude of the prices paid by consumer households for goods and services |
| CPI.Energy | The monthly consumer price index for energy for the given month and year |
| Spring | 1 if Month=3, 4, 5; 0 otherwise |
| Summer | 1 if Month=6, 7, 8; 0 otherwise |
| Fall | 1 if Month=9, 10, 11; 0 otherwise |

*Note: the variables Spring, Summer, and Fall are called the “dummy” variables. They are created to represent each season. When the values of Spring, Summer, and Fall are all equal to 0, it refers to Month=12, 1, 2. Therefore, we don’t need an additional dummy variable to represent the Winter season.*

***Perform the following tasks:***

1. *Split the Elantra.csv data into a training and testing set. The training set has all observations for 2010, 2011, and 2012, and the testing set has all observations for 2013 and 2014.*

→ see the attached Excel File

1. *Build a regression model based on the* ***training set*** *to predict monthly Elantra sales using Unemployment, Queries, CPI.Energy, and CPI.All as the predicting variables.*
2. *Present the fitted regression equation.*

The fitted regression equation to predict monthly Elantra Sales is as follows:

y = 95385.36 - 3179.9\*Unemployment + 19.02968\*Queries + 38.50604\*CPI.Energy - 297.646\*CPI.All

1. *Provide the fitted coefficients table such as the one below:*

|  | *Coefficient* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Intercept* | *95385.36* | *170663.8* | *0.558908* | *0.58024* | *-252686* | *443456.5* | *-252686* | *443456.5* |
| *Unemployment* | *-3179.9* | *3610.262* | *-0.88079* | *0.385207* | *-10543.1* | *4183.279* | *-10543.1* | *4183.279* |
| *Queries* | *19.02968* | *11.25896* | *1.690181* | *0.101027* | *-3.93312* | *41.99249* | *-3.93312* | *41.99249* |
| *CPI.Energy* | *38.50604* | *109.6012* | *0.351329* | *0.727718* | *-185.027* | *262.0391* | *-185.027* | *262.0391* |
| *CPI.All* | *-297.646* | *704.8367* | *-0.42229* | *0.675728* | *-1735.17* | *1139.878* | *-1735.17* | *1139.878* |

1. *Is the “sign” of the fitted coefficient of each variable consistent with what we would expect for its relation with the sales? (For example, a negative sign indicates the values of the predicting variable are negatively correlated with sales amount. If the sign of the variable CPI.ALL is positive, does it make sense to you? )*

The sign of the fitted coefficient of the “Unemployment”, “Queries”, “CPI.All” predicting variables IS consistent with what I would expect for its relation with the number of sales. However, I am not completely sure about the “CPI.Energy”. Let me explain my analysis for each predicting variable in the model.

First, “Unemployment” has a negative sign indicating that the value of the predicting variable is negatively correlated with the number of sales for Elantra. This means that Unemployment negatively affects the number of sales for the car: the higher the unemployment rate, the lower the sales. This makes sense because if the unemployment rate is high this means that fewer people are working and earning money to be purchasing new cars like the Elantra for example. Put simply, Unemployed or underemployed people have reduced income, which means they have less money to spend on bigger purchases like cars, if the unemployment rate is high this means there are more people with reduced income and unable to purchase cars.

Second, Queries has a positive sign meaning that this predicting variable has a positive relationship with the number of Sales for the Elantra: the higher the number of queries, the higher the sales. Considering that the predicting variable “Queries” is a normalized approximation of the number of Google searches for “Hyundai Elantra” in the given month and the year it makes perfect sense that a high number of queries positively affects the number of sales. When people are searching for a particular car model online, it indicates that they are aware of the product and interested in it. By searching for information about the car, they are gathering information that can help them make a decision, and this may lead to a purchase. This increased awareness can definitely translate into increased sales.

Third, “CPI.Energy” has a positive sign, meaning that this predicting variable has a positive relationship with the number of Sales: the higher “CPI.Energy”, the higher the sales. “CPI.Energy” represents the monthly consumer price index for energy for the given month and year and it does not really make sense to me that the monthly consumer price index for energy has a positive relationship with the number of sales of a car because when energy prices increase, consumers may substitute away from car purchases towards more fuel-efficient vehicles or public transportation. This may result in a decrease in car sales. Also, when energy prices increase, consumers may have less disposable income to spend on other goods and services, including cars, leading to a decrease in car sales. On the other hand, an increase in sales can result in an increase in the “CPI.Energy” as more people will be paying for fuel, gas, and other energy-related expenses.

Fourth, “CPI.All” has a negative sign indicating that the values of the predicting variable are negatively correlated with the number of sales for the Elantra. This means that CPI.All negatively affect the number of sales for the car: the higher the CPI for all goods and services, the lower the sales. Considering that “CPI.All” represents the consumer price index for all products for the given month and year and is a measure of the magnitude of the prices paid by consumer households for goods and services this negative relationship makes sense to me. If the price index for all goods and services increases, this may indicate that consumers have less disposable income to spend on expensive items like cars, it may also result in higher inflation rates causing central banks to increase interest rates, making it more expensive to finance a car purchase. This may lead to a decrease in car sales as higher prices for all goods and services may cause a shift in consumer preferences towards more affordable modes of transportation, such as public transportation.

1. *What is the of the model?*

The of the model is 0.428157, meaning that roughly 43% of the variability of Y (sales for the Hyundai Elantra) can be explained by the entire set of independent predicting variables that we have considered which was “Unemployment”, “Queries”, “CPI.Energy”, “CPI.All”. This is not a great and it shows that there is 57% of the variability in sales that is not explained by this model and ultimately that we can add predicting variables in the model to better explain the variability of y.

1. *We would like to improve the above regression model by modeling seasonality. When predicting demand and sales,* ***seasonality*** *is often very important since car demands tend to be influenced by seasons. To do so, we will incorporate seasonality into the model.*
2. *Build a regression model based on the* ***training set*** *by including the 3 dummy variables,* ***Spring, Summer, Fall****, together with* ***Unemployment, Queries, CPI.Energy, and CPI.All*** *as predicting variables.*

*Provide the fitted coefficients table.*

|  | *Coefficient* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Intercept* | *195914.2* | *133532.7* | *1.467163* | *0.153475* | *-77615.1* | *469443.5* | *-77615.1* | *469443.5* |
| *Unemployment* | *-5573.59* | *2805.004* | *-1.98702* | *0.056784* | *-11319.4* | *172.1994* | *-11319.4* | *172.1994* |
| *Queries* | *-6.63741* | *9.931571* | *-0.66831* | *0.509405* | *-26.9813* | *13.70649* | *-26.9813* | *13.70649* |
| *CPI.Energy* | *236.3668* | *94.17242* | *2.509936* | *0.018132* | *43.46333* | *429.2702* | *43.46333* | *429.2702* |
| *CPI.All* | *-840.183* | *557.6772* | *-1.50658* | *0.143119* | *-1982.53* | *302.1667* | *-1982.53* | *302.1667* |
| *Spring* | *4594.154* | *1165.955* | *3.94025* | *0.000493* | *2205.804* | *6982.504* | *2205.804* | *6982.504* |
| *Summer* | *6228.202* | *1380.091* | *4.512892* | *0.000105* | *3401.213* | *9055.19* | *3401.213* | *9055.19* |
| *Fall* | *672.2826* | *1208.532* | *0.556281* | *0.582438* | *-1803.28* | *3147.847* | *-1803.28* | *3147.847* |

1. *What is the of the modified model?*

The of the modified model is 0.72711, meaning that approximately 73% of the variability/change of the sales for the Hyundai Electra can be explained by the new set of independent predicting variables that we have considered which is now “Unemployment”, “Queries”, “CPI.Energy”, “CPI.All” and “Seasonality”. This is a much better than the one our previous model provided. It shows that adding the dummy predicting variables Spring, Summer, and Fall in our model better explains the variability of y by 30%.

1. *Suppose you estimate that the unemployment rate is 8%, number of queries=250, CPI.All=235, CPI.Energy=260. Use the modified model to predict the sales for* ***Spring 2013*** *(round your number to a whole number).*

The fitted regression equation to predict monthly Elantra Sales with our new model is as follows:

y = 195914.225 - 5573.59\*Unemployment - 6.637\*Queries + 236.367\*CPI.Energy - 840.18\*CPI.All +4594.15\*Spring + 6228.2\*Summer + 672.283\*Fall

Considering the number given we expect the following sales number for Spring 2013:

y = 195914.225 - 5573.59\*8 - 6.637\*250 + 236.367\*260 - 840.18\*235 +4594.15\*1 + 6228.2\*0 + 672.283\*0

y = 18272.607

y = 18,273

1. *Based on what you see from the two models you develop, what can you conclude about predicting Hyundai Elantra sales? Justify your answer.*

Based on the two models we developed, it can be concluded that adding seasonality variables such as “Spring”, “Summer”, and “Fall” in addition to the existing variables of “Unemployment”, “Queries”, “CPI.Energy”, and “CPI.All” provides a much better explanation of the variability of sales for the Hyundai Elantra. However, 27% of the variability is still not explained by our model meaning that we could still improve it by adding other predictive variables. The first model only considers the four variables “Unemployment”, “Queries”, “CPI.Energy”, and “CPI.All” as predictors and has an R-squared value of 0.428, indicating that only 43% of the variability in sales can be explained by these variables. However, when the three dummy variables for seasonality are added to the model along with the original predictors, the R-squared value increases significantly to 0.727, indicating that approximately 73% of the variability in sales can be explained by these variables. We can then conclude that seasonality has an impact on the purchasing pattern of consumers.

But this is not the only change we can notice: the sign of a coefficient went from positive to negative, meaning that its relationship with the predicted variable changed when we added more predicting variables. To be more specific, in the first model, the sign of the Coefficient “Queries” is +19 but in the second model it becomes -6.6. When the coefficient of a predicting variable changes from positive to negative after adding other predictors to the model, it means that the relationship between the original predicting variable and the response variable is confounded by the other predictors in the model. This phenomenon is known as "confounding" and can occur when two or more predictors are correlated with each other, and each of them is independently associated with the response variable. Confounding occurs because the presence of the additional predictors changes the nature of the relationship between the original predictor and the response variable. The original predictor may have had a positive relationship with the response variable in the absence of the additional predictors, but when the other predictors are added, the original predictor's relationship with the response variable is altered. It's also worth noting that in a multiple regression model, the coefficients of the predictors represent the partial effect of that predictor on the response variable, holding all other predictors constant. Therefore, the coefficient of a predictor can change sign when other predictors are added to the model, even if the relationship between the original predictor and the response variable remains unchanged.

Let's test our fitted regression equation with the actual numbers given in the testing set for the month of July 2013 to see how close we can get to the actual number of sales obtained for this month. In other words, let's see how accurate our model is. For July 2013 we have Unemployment=7.5, Queries = 274, CPI.Energy = 246, CPI.All= 233 and Summer=1. And the actual number of Elantra Sales is 23,958. With our second model, we are obtaining: y = 195914.225 - 5573.59\*7.5 - 6.637\*274 + 236.367\*246 - 840.18\*233 +4594.15\*0 + 6228.2\*1 + 672.283\*0 = 21,791 which is a little off from the actual numbers but still pretty good to predict sales for this month. I have created 2 columns in the testing set to compare the expected Sales with the first Model and with the 2nd Model and even if the numbers obtained with the second model are not the actual numbers they are closer than the ones obtained with the first model (The Second model seems to be more accurate to predict sales in the Summer). We can then conclude that the second model is better to explain the variability of y and to estimate the sales.

Therefore, it can be concluded that seasonality is an important factor to consider when predicting sales for the Hyundai Elantra and that including seasonality variables in the model greatly improves its predictive power.