Linear Regression

Suzana Iacob

15/09/2019

This report analyzes two datasets and uses visualizations and regression models to obtain predictions.

1: Forecasting Automobile Sales

Preprocessing

The dataset contains the unit sales of Hyundai Elantra and Jeep Wrangler, two types of automobiles, alongside variables which will serve as predictors for the sales. We start by loading the data and visually inspecting a small subset..

```
WranglerElantra = read.csv("WranglerElantra2018.csv")
head(WranglerElantra)
```

	X <int>	date <fctr>	M... <int>	Y... <int>	Wrangler.Sales <int>	Elantra.Sales <int>	Unemployment.Rate <dbl>	Wrangler.Que <int>
1	1	1/1/10	1	2010	4888	7690	9.8	
2	2	2/1/10	2	2010	5967	7966	9.8	
3	3	3/1/10	3	2010	8410	8225	9.9	
4	4	4/1/10	4	2010	8327	9657	9.9	
5	5	5/1/10	5	2010	9634	9781	9.6	
6	6	6/1/10	6	2010	8923	14245	9.4	

6 rows | 1-9 of 12 columns

We also inspect the datatypes to ensure our model will yield the expected results. We note that the variables appear correct except for date which needs to be converted into a categorical variable.

```
##  Date[1:108], format: "2010-01-01" "2010-02-01" "2010-03-01" "2010-04-01" "2010-05
-01" ...
```

```
## 'data.frame':    108 obs. of  11 variables:
##  $ X                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ date             : Date, format: "2010-01-01" "2010-02-01" ...
##  $ Month            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Year             : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ Wrangler.Sales   : int  4888 5967 8410 8327 9634 8923 10043 7666 7765 7908 ...
##  $ Elantra.Sales    : int  7690 7966 8225 9657 9781 14245 18215 15181 10062 9497
...
##  $ Unemployment.Rate: num  9.8 9.8 9.9 9.9 9.6 9.4 9.4 9.5 9.5 9.4 ...
##  $ Wrangler.Queries : int  31 32 34 36 36 37 41 37 35 32 ...
##  $ Elantra.Queries  : int  7 8 9 9 9 9 10 10 9 9 ...
##  $ CPI.All          : num  217 217 217 217 217 ...
##  $ CPI.Energy       : num  213 210 209 209 207 ...
```

Building an intial regression model

We begin by selecting a training set comprising all observations in 2010–2017, and a test set comprising all observations in 2018.

```
WranglerElantraTrain = subset(WranglerElantra, Year < 2018)
WranglerElantraTest = subset(WranglerElantra, Year == 2018)
```

We then build an initial model to predict monthly Wrangler sales with five independent variables: Year, Unemployment.Rate, Wrangler.Queries, CPI.Energy, and CPI.All.

```
WranglerLM = lm(Wrangler.Sales ~ Year + Unemployment.Rate + Wrangler.Queries + CPI.En
ergy + CPI.All, data=WranglerElantraTrain)
summary(WranglerLM)
```

```
##
## Call:
## lm(formula = Wrangler.Sales ~ Year + Unemployment.Rate + Wrangler.Queries +
##     CPI.Energy + CPI.All, data = WranglerElantraTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3644.3 -1194.7   -61.8  1039.6  5108.1
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4778395.01 1223727.74   3.905 0.000182 ***
## Year                -2343.54     619.86  -3.781 0.000281 ***
## Unemployment.Rate   -3178.80     649.31  -4.896 4.28e-06 ***
## Wrangler.Queries      301.88      26.74  11.288  < 2e-16 ***
## CPI.Energy             53.53      19.49   2.746 0.007280 **
## CPI.All              -234.63     180.07  -1.303 0.195909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1634 on 90 degrees of freedom
## Multiple R-squared:  0.8273, Adjusted R-squared:  0.8177
## F-statistic:  86.2 on 5 and 90 DF,  p-value: < 2.2e-16
```

We note that Year, Unemployment.Rate, Wrangler.Queries and CPI.Energy are significant to the 95% level.

We can interpret the initial model as follows: - Year has a negative relationship with Wrangler Sales, for each additional year, the Wrangler Sales decrease by 2,343 units. - Similarly, Unemployment Rate and Wrangler Sales have a negative relationship which matches intuition (as unemployment increases, car sales decrease) - Wrangler Queries and CPI Energy have a positive relationship with sales, whereas CPI.ALL is not statistically significant.

Improving the model

Prior to assessing model performance by calculating out-of-sample R-squared values, we wish to build a model that accurately represents the business context. We start by removing the CPI.ALL variable, which is the only variable that does not match the significance threshold. Removing variables one by one is a valuable strategy for selecting the relevant predictors.

```
WranglerLM2 = lm(Wrangler.Sales ~ Year + Unemployment.Rate + Wrangler.Queries + CPI.E
nergy, data=WranglerElantraTrain)
summary(WranglerLM2)
```

```
##
## Call:
## lm(formula = Wrangler.Sales ~ Year + Unemployment.Rate + Wrangler.Queries +
##     CPI.Energy, data = WranglerElantraTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3628.4 -1230.3   -67.5  1067.6  5340.3
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.793e+06  9.479e+05   6.111 2.42e-08 ***
## Year              -2.873e+03  4.696e+02  -6.120 2.33e-08 ***
## Unemployment.Rate -2.680e+03  5.264e+02  -5.091 1.91e-06 ***
## Wrangler.Queries   3.009e+02  2.683e+01  11.212  < 2e-16 ***
## CPI.Energy         3.084e+01  8.784e+00   3.510 0.000698 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1640 on 91 degrees of freedom
## Multiple R-squared:  0.824,  Adjusted R-squared:  0.8163
## F-statistic: 106.5 on 4 and 91 DF,  p-value: < 2.2e-16
```

We note that, after removing CPI.All, **all of the other variables remain significant**. Additionally, the relationships maintain signs (Year and Unemployment maintain a negative relationship, while Queries and CPI Energy, a positive one). Nevertheless, we might have multicollinearity in the model as independent variables might be correlated to one another, affecting the model coefficients. We will later construct a correlation matrix with all variables.

The question that arises is - **which predictors should we choose?** We could take the approach of selecting all variables that meet a specific threshold (here 95%), yet this might not represent an accurate depiction of reality. We should additionally use domain knowledge and expertise to gain insight into what the variables represent and whether the relationships highlighted by the statistical model make sense in the business context.

For example - unemployment and automobile sales are negatively correlated. This is what we expect since unemployment causes economic difficulties and consumers may have less income available for purchasing automobiles. This may be especially true for the sales of Jeep Wrangler, a four-wheel-drive off-road SUV,

which might not be a necessity. However, **year and car sales are negatively correlated**. The possibility exists that the Wrangler series decreased in popularity over time, yet this is not a relationship that we anticipated. Moreover, **we expect unemployment to be highly correlated with time** since unemployment is an economic indicator that changes over time alongside economic upturns or downturns. Furthermore, even if time and sales exhibit a strong relationship, we must consider this in the context of the predictive power of the model. **Past sales may not be an indicator of future sales** (as it is often observed in the stock market), and if we wish to use the model to forecast future demand, the time element may not be a representative predictor.

Let us have a closer look at the relationship of time versus sales and unemployment versus sales.

We notice Time has a strong negative relationship with unemployment, however, Wrangler Sales actually increase over time. We thus decide to **remove the Year variable from our model**.

```
WranglerLM3 = lm(Wrangler.Sales ~ Unemployment.Rate + Wrangler.Queries + CPI.Energy,
data=WranglerElantraTrain)
print(summary(WranglerLM3)$r.squared)
```

```
## [1] 0.7515637
```

```
WranglerLM3PredictTest = predict(WranglerLM3, newdata=WranglerElantraTest)
WranglerLM3SSE = sum((WranglerLM3PredictTest - WranglerElantraTest$Wrangler.Sales)^2)
WranglerSST = sum((mean(WranglerElantraTrain$Wrangler.Sales) - WranglerElantraTest$Wr
angler.Sales)^2)
WranglerLM3R2 = 1 - WranglerLM3SSE/WranglerSST
print(WranglerLM3R2)
```

```
## [1] 0.6973278
```

The R-squared has decreased, however, we believe this to be a superior model, as we have seen the time element to be misleading. Furthermore, the R-squared has the disadvantage of increasing with the number of variables in the model. Consequently, we should not utilize solely R-squared as an indicator of performance, but combine it with domain knowledge and intuition.

Modeling Seasonality

We represent seasonality using the Month variable. We must convert this into a categorical variable; if we leave it as a number, a change from December(12) to January(1) would be treated as a “decrease” in Month, which does not accurately represent our context. Moreover, the model would assume a linear effect of Month on sales, by modeling as a factor, we no longer restrict the effect to be linear.

```
WranglerElantraTrain$Month = as.factor(WranglerElantraTrain$Month)
WranglerElantraTest$Month = as.factor(WranglerElantraTest$Month)
WranglerLM4 = lm(Wrangler.Sales ~ Month + Unemployment.Rate + Wrangler.Queries + CPI.
Energy, data=WranglerElantraTrain)
summary(WranglerLM4)
```

```
##
## Call:
## lm(formula = Wrangler.Sales ~ Month + Unemployment.Rate + Wrangler.Queries +
##     CPI.Energy, data = WranglerElantraTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3270.9  -731.3   157.3   668.9  3826.4
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       10198.444   4592.791   2.221 0.029176 *
## Month2             1230.528    682.992   1.802 0.075318 .
## Month3             4160.059    701.038   5.934 7.00e-08 ***
## Month4             4785.306    716.979   6.674 2.85e-09 ***
## Month5             6553.629    752.113   8.714 2.93e-13 ***
## Month6             5559.981    814.312   6.828 1.45e-09 ***
## Month7             5250.934    858.127   6.119 3.18e-08 ***
## Month8             4650.513    778.264   5.975 5.87e-08 ***
## Month9             3127.527    692.863   4.514 2.13e-05 ***
## Month10            2475.330    685.196   3.613 0.000524 ***
## Month11            1582.450    683.345   2.316 0.023102 *
## Month12            2899.469    696.192   4.165 7.74e-05 ***
## Unemployment.Rate -1236.966    323.776  -3.820 0.000260 ***
## Wrangler.Queries     52.739     36.582   1.442 0.153253
## CPI.Energy           21.713      7.315   2.968 0.003936 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1354 on 81 degrees of freedom
## Multiple R-squared:  0.8933, Adjusted R-squared:  0.8749
## F-statistic: 48.45 on 14 and 81 DF,  p-value: < 2.2e-16
```

We interpret the model as follows: The intercept term corresponds to the level of sales in Month1 (January). The coefficients of each of the other Month-variables, represent the increase in Sales in the respective month, as compared to the January sales. The coefficients of the rest of the variables (Unemployment, etc.) represent the change in sales per unit-change in the respective variable, as calculated in the month of January.

In modeling demand and sales, it is often useful to model seasonality, that is, sales tend to be cyclical in time. For example, sales of all products increase during the winter holidays, while ice cream sales increase in the summer. We notice that the R-squared of the model improved significantly, from 0.75 to 0.89. This leads to the conclusion that adding Month was correct. Furthermore, all other variables remain significant, suggesting the model has little variability and is stable.

We now know that Month is significant, and we can interpret this by stating that summer months (Month5, 6, and 7) are associated with larger increases, hence **consumers are in greater need of SUV vehicles during the summer**. However, we note that the coefficients are always positive, hence we see no decrease and thus no cyclical pattern. **We might, in fact, be capturing the time effect on sales** (as we have already seen there is a positive relationship with sales increasing over time). We choose to retain the Month variable, yet we should be cautious when using this model for predictions.

We finally characterize the model's out-of-sample performance.

```
WranglerLM4PredictTest = predict(WranglerLM4, newdata=WranglerElantraTest)
WranglerLM4SSE = sum((WranglerLM4PredictTest - WranglerElantraTest$Wrangler.Sales)^2)
WranglerSST = sum((mean(WranglerElantraTrain$Wrangler.Sales) - WranglerElantraTest$Wr
angler.Sales)^2)
WranglerLM4R2 = 1 - WranglerLM4SSE/WranglerSST
WranglerLM4R2
```

```
## [1] 0.7400067
```

Predicting Elantra Sales

We use the same variables (with the respective Google queries) to predict the sales of Hyundai Elantra.

```
ElantraLM = lm(Elantra.Sales ~ Month + Unemployment.Rate + Elantra.Queries + CPI.Ener
gy, data=WranglerElantraTrain)
summary(ElantraLM)
```

```
##
## Call:
## lm(formula = Elantra.Sales ~ Month + Unemployment.Rate + Elantra.Queries +
##     CPI.Energy, data = WranglerElantraTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7403.1 -1407.2   259.7  1761.7  7182.7
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         706.901   3864.432   0.183 0.855314
## Month2             2252.437   1575.561   1.430 0.156675
## Month3             9083.800   1594.116   5.698 1.89e-07 ***
## Month4             6174.343   1586.938   3.891 0.000204 ***
## Month5             7794.282   1585.257   4.917 4.53e-06 ***
## Month6             7274.792   1609.526   4.520 2.09e-05 ***
## Month7             7331.742   1616.279   4.536 1.96e-05 ***
## Month8             7147.146   1611.499   4.435 2.87e-05 ***
## Month9             4621.197   1574.233   2.936 0.004332 **
## Month10            1377.266   1570.710   0.877 0.383165
## Month11            2537.340   1573.215   1.613 0.110669
## Month12            5063.976   1573.736   3.218 0.001858 **
## Unemployment.Rate -1535.397    270.186  -5.683 2.02e-07 ***
## Elantra.Queries      -4.747    127.703  -0.037 0.970442
## CPI.Energy           97.951     17.431   5.619 2.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3127 on 81 degrees of freedom
## Multiple R-squared:  0.625,  Adjusted R-squared:  0.5602
## F-statistic: 9.644 on 14 and 81 DF,  p-value: 3.655e-12
```

Interestingly, the same variables fail to predict Elantra sales with the same strength. Elantra.Queries does not meet the significance threshold, and the R-squared is only 63%.

```
summary(ElantraLM)$r.squared
```

```
## [1] 0.6250191
```

As we have done previously, we now computed the R-squared for the test set. We use the output of the *t vi hrgx* function, which gives the y-hat values on the test set, i.e. the predictions for the Elantra Sales variable using the test data. We then subtract the true test-set Sales value. The differences represent the errors, which we square and sum.

We then compute the SST as the sum of the squared differences between Sales in the test set and the mean of Sales in the training set. The mean Sales from the training set represents our baseline - the best prediction of Sales, if we had no independent variables. Hence the SST is the total error, or total variability in the model.

Consequently, R-squared will be calculated as 1 minus the SSE divided by the SST. R-squared measures the variability explained by the regression model(SSE), as a percentage of the total variability(SST).

```
ElantraLMLMPredictTest = predict(ElantraLM, newdata=WranglerElantraTest)
ElantraLMSSE = sum((ElantraLMLMPredictTest - WranglerElantraTest$Elantra.Sales)^2)
ElantraSST = sum((mean(WranglerElantraTrain$Elantra.Sales) - WranglerElantraTest$Elan
tra.Sales)^2)
ElantraLMOSR2 = 1 - ElantraLMSSE/ElantraSST
ElantraLMOSR2
```

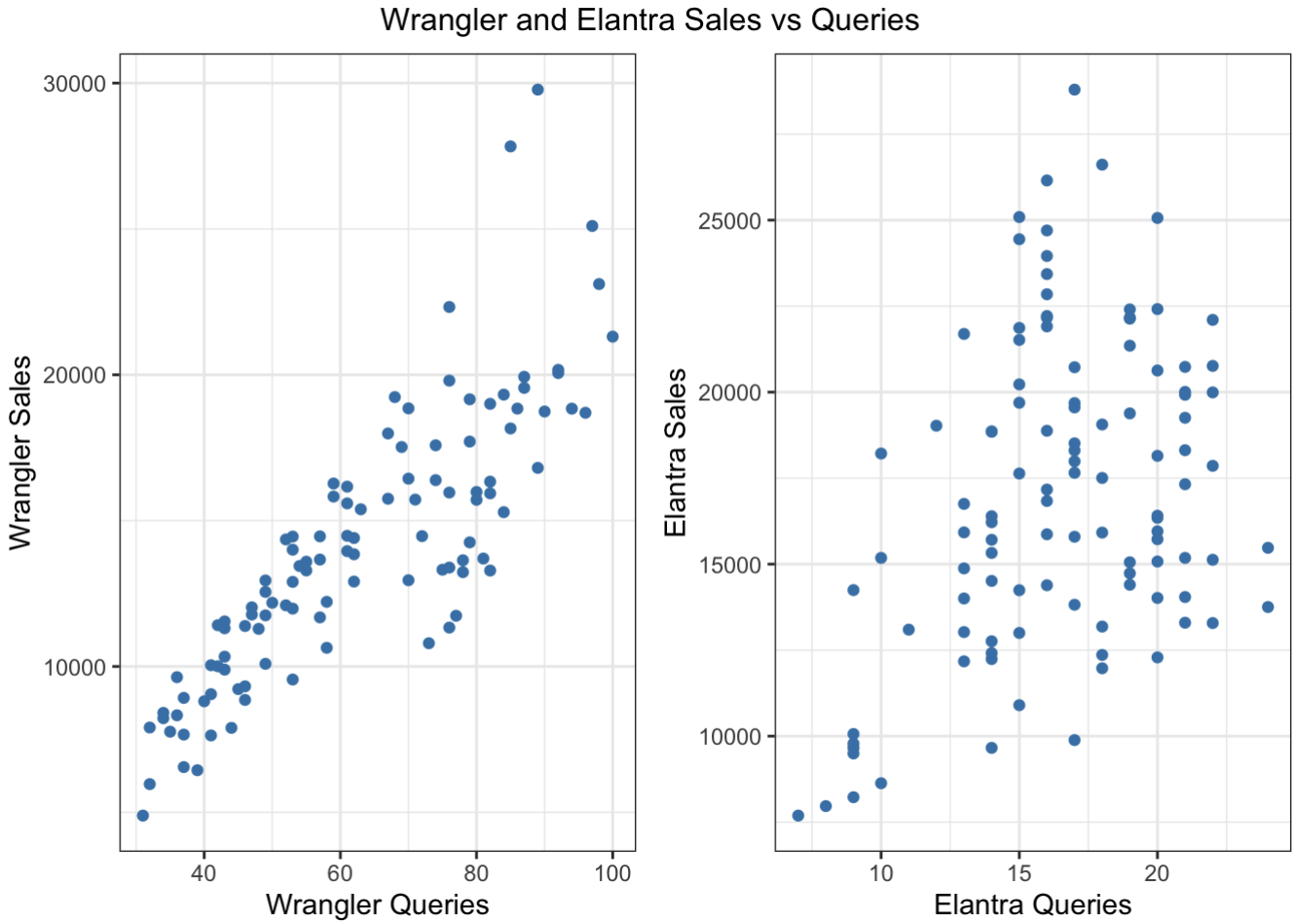
```
## [1] -4.908219
```

The R-squared of the model is negative. This means the baseline (the model predicting the mean sales for every observation in the test set) is *syxt i vjsvq mk xl i vi kvi wwsr q shi p* We should conduct further analysis to explain the lack of predictability our model has for Elantra sales.

Interpreting Results

We now plot the sales of both Elantra and Wrangler versus the relevant Google queries.

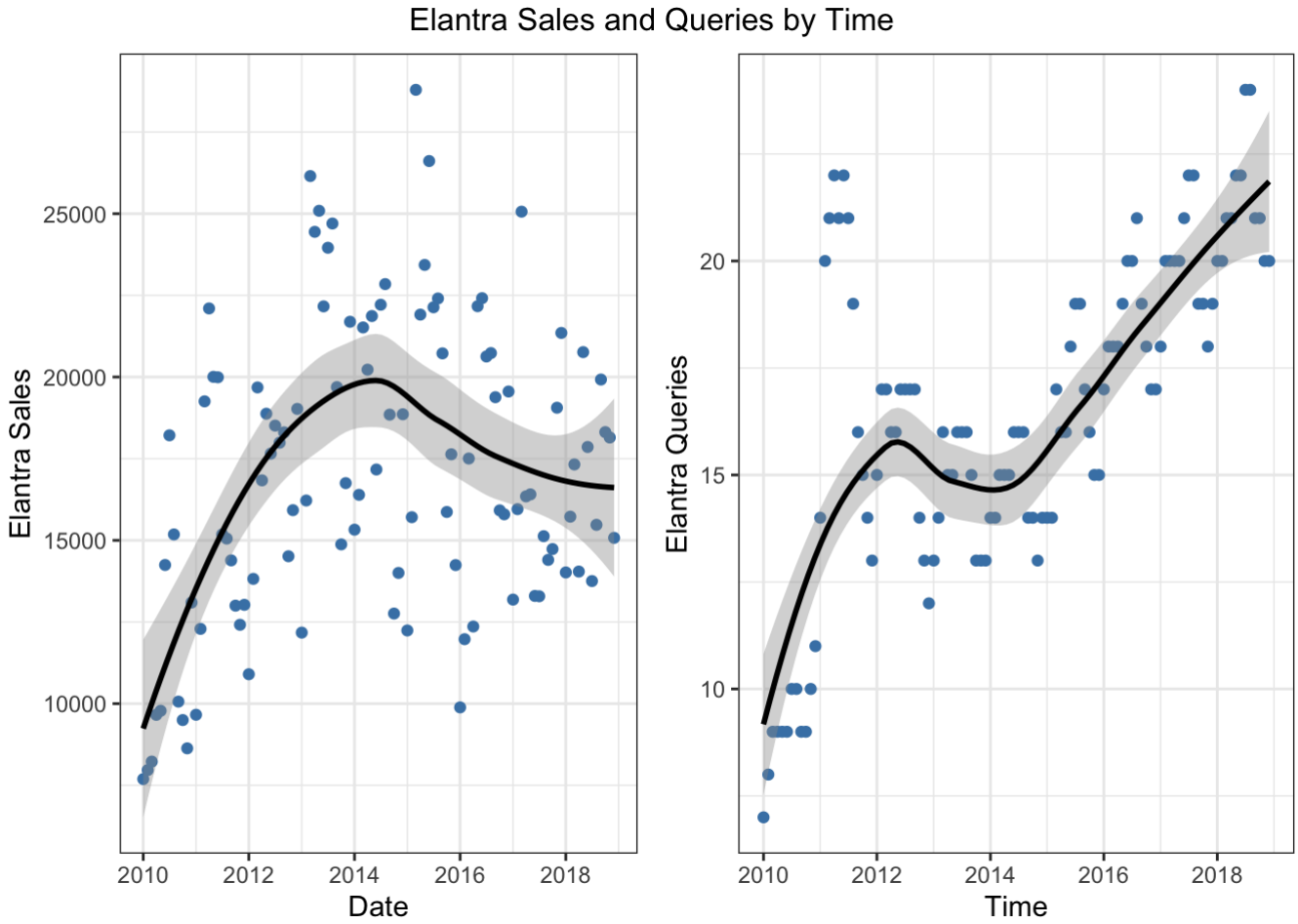
```
plot1 = ggplot(data = WranglerElantra, aes(x = Wrangler.Queries, y = Wrangler.Sales))
+
  geom_point(color = "steelblue")  +
  xlab("Wrangler Queries") + ylab("Wrangler Sales") +
  theme(legend.title=element_blank()) + theme_bw()

plot2 = ggplot(data = WranglerElantra, aes(x = Elantra.Queries, y = Elantra.Sales)) +
  geom_point(color = "steelblue")  +
  xlab("Elantra Queries") + ylab("Elantra Sales") +
  theme(legend.title=element_blank()) + theme_bw()

grid.arrange(plot1, plot2, ncol=2, top="Wrangler and Elantra Sales vs Queries")
```



Wrangler Sales clearly display a positive relationship with Google Queries, whereas Elantra Sales do not have such a strong relationship. Wrangler Sales have lower variability, while Elantra sales, despite an initial positive trend have a very unpredictable pattern.

It appears as though the sales of Elantra could not be accurately explained by the regression model. Let us inspect the sales over time, as well as the Elantra Queries over time.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Surprisingly, Elantra sales increase with time until 2014, after which they decrease, while being very scattered around the line of best fit. Elantra queries, despite the polynomial nature of the relationship, appear to have a positive trajectory.

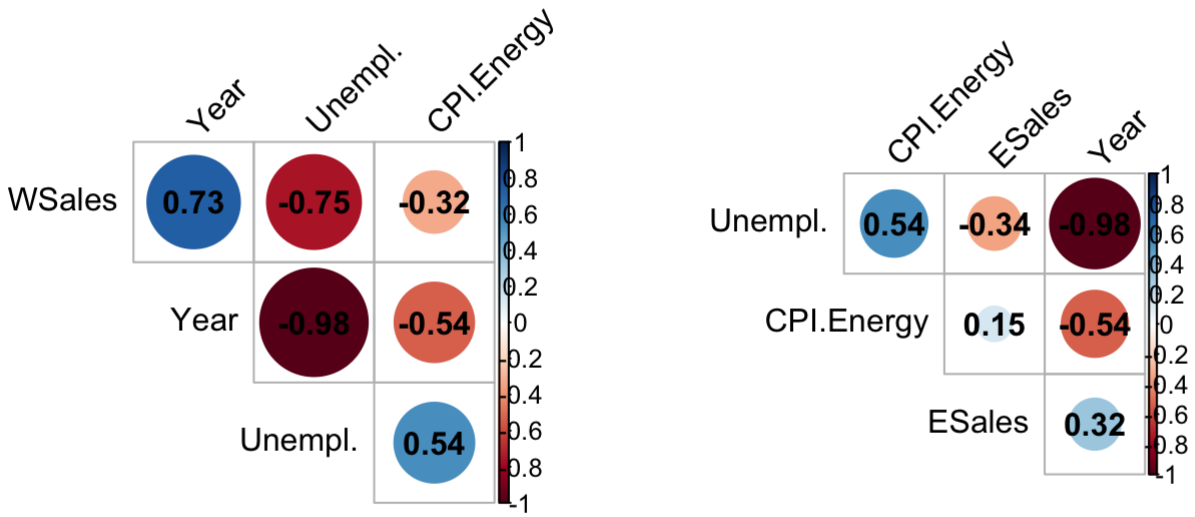
Correlations

We would like to compute the correlation coefficients between the relevant independent and dependent variables. We had previously removed Year, yet we will add it to the correlation matrix to further depict the strong correlations between Year and Unemployment.

```
WranglerCorrelation <- data.frame(WSales = WranglerElantraTrain$Wrangler.Sales,
                          Unempl. = WranglerElantraTrain$Unemployment.Rate,
                          CPI.Energy = WranglerElantraTrain$CPI.Energy,
                          Year = WranglerElantraTrain$Year)

ElantraCorrelation <- data.frame(ESales = WranglerElantraTrain$Elantra.Sales,
                          Unempl. = WranglerElantraTrain$Unemployment.Rate,
                          CPI.Energy = WranglerElantraTrain$CPI.Energy,
                          Year = WranglerElantraTrain$Year)

library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
par(mfrow=c(1,2))
corrplot(cor(WranglerCorrelation), method="circle",
         type="upper", order="hclust",
         addCoef.col = "black", # Add coefficient of correlation
         tl.col="black", tl.srt=45, #Text label color and rotation
         # hide correlation coefficient on the principal diagonal
         diag=FALSE )

corrplot(cor(ElantraCorrelation), method="circle",
         type="upper", order="hclust",
         addCoef.col = "black", # Add coefficient of correlation
         tl.col="black", tl.srt=45, #Text label color and rotation
         # hide correlation coefficient on the principal diagonal
         diag=FALSE )
```



The plot on the left displays the Wrangler Sales and other independent variables. As noted, Year-Unemployment has a high coefficient, and so does Sales-Unemployment. The remaining independent variables do not appear strongly correlated, and we should not be further concerned about multicollinearity.

In the case of Elantra sales, the strongest correlation coefficient is Unemployment (only -0.34). This corroborates the previous claims that the sales of Elantra are unpredictable.

Discussion

Wrangler Sales appear predictable to a certain extent using economic indicators such as the rate of unemployment and the energy consumer price index, alongside Google search queries. This makes sense in the business context, however, there may be many more factors contributing to sales. Below we discuss what else we may consider for predicting the sales of Hyundai Elantra, yet the same logic could be applied to other automobiles including Jeep Wrangler.

Why do we fail to predict Elantra Sales?

- **Insufficient data** - our training set only contains 96 observations which may not be sufficient data to accurately predict Elantra Sales, and gathering additional datapoints might help.
- **Missing predictors** - the most prominent possibility is that we simply do not have the independent variables to explain what impact Elantra Sales. We have seen a non-linear relationship over time, characterized by an increase followed by a decrease in sales. Hence sales do have a pattern and the Hyundai business managers might further investigate this avenue. Possibilities include:

- **Competitor Analysis** - The sales began decreasing in 2015 which could potentially be correlated with the release of a similar automobile.
- **Quality of Product** - If existing customers noticed a large number of defects, this may drive away new customers
- **Secondary Markey** - We assume the sales amount is defined as sales of new models, directly from the manufacturer/retailers. We should consider obtaining data pertaining to Elantra sales on the secondary market to understand how this behaves and if customers prefer purchasing a second-hand Elenatra model versus a new one.
- **Fuel prices** - alongside other economic matters, the price of fuel may influence consumers purchasing habits.
- **Environmental Concerns** - similarly, customers may choose environmentally-friendly vehicles such as electric cars over Elantras.

Future Model Enhancements We only considered linear regression, despite some very apparent non-linear patterns in the data, especially in the case of Elantra sales. Many other models could be attempted including polynomials, splines, logarithmic models, or other approaches such as Classification and Regression Trees(CART). CART may be particularly useful for Elantra since the linear relationship assumption appears incorrect, and CART would allow us to make splits based on decision variables, then follow the model splits to make a prediction.

From Predictions to Recommendations

A regression model could be a tool for helping decision-making in organizations, for questions such as **How many automobiles to produce in a given month?**.

Naturally, we currently lack data on the independent variables to enable us to make a prediction for the sales data in April 2018. Nonetheless, this data could be inferred through the use of forecasting. Indeed, our predictor variables are economic indicators and numerous experts and organizations concern themselves with forecasting unemployment and the CPI.

Conversely, Google queries data is more difficult to infer and our model is subject to the uncertainty. Furthermore, the causal relationship between Google queries and car sales is also questionable. Do consumers buy more because of an increased number of Google searches (e.g. through word-of-mouth, people who searched for a Wrangler online may discuss with acquaintances their intent to buy, thus driving up both queries and sales; or the Google algorithms may show more Wrangler advertisements to people who searched for this term on Google, thus driving up sales)?. Or would more sales cause more Google queries (e.g. as more Wranglers are sold, more people notice them in traffic and decide to search for them online)? Arguably we need Wrangle sales data to estimate Wrangle Google queries for a given period. Yet we can look at past queries to estimate future sales since we can safely assume that online searching precedes a purchase.

Hence a model for forecasting demand in April 2018 could include the forecasted unemployment rate and CPI for April 2018 and the Google queries for March 2018.

Inventory-related costs will include the cost of production, storage, resources, and materials costs. If supply exceeds demand storage and maintenance costs will rise. Nevertheless, cars are not fast-moving goods and we can argue that we should produce more than the forecasted demand as we will sell them in future periods.

The costs of producing less than demand include loss of opportunity of customers that would have purchased but were not able to. Also, we should consider reputational cost - consumers may not expect to not be able to purchase the car they desired.

Given that the costs of underproducing could be deemed higher than the ones of overproducing, we recommend producing more than the forecasted demand. For example, Wrangler sales in March were 27,829 units and we recommend producing at least 30,000 units, or forecasted demand + 10%. Similarly, Elantra sales were 17,323 units, and we recommend producing at least 20,000 units. Since Elantra sales are extremely unpredictable, a larger margin of error is advisable.

To deal with such uncertainty, we could look at other models that are more robust i.e. account for changes in the predictors' data (here forecasted predictors data) such as robust regression.