Yahya Alhinai

EE5373

December 10th, 2019

Avocado Prices

OBJECTIVE FOR THIS LAB:

To demonstrate that you can integrate what you have learned in this course to study an interesting set of data using appropriate data modeling techniques.

The data analyzed here are from Kagle websites:

https://www.kaggle.com/neuromusic/avocado-prices

1.Regression model:

Avocado set were conducted to get the price Avocado after training **18250 unique entries**. Avocado set has the following data frame values:

- **Average Price**: the average price of a single avocado
- **Date**: The date of the observation
- **type**: conventional or organic
- **year**: the year
- **Region**: the city or region of the observation
- **Total** Volume: Total number of avocados sold
- **4046**: Total number of avocados with PLU 4046 sold
- **4225**: Total number of avocados with PLU 4225 sold
- **4770**: Total number of avocados with PLU 4770 sold

Data Cleaning and Sanity Checking:

- I grouped all predictors of the same category into. All regions to be grouped into one predictor.
- All Date to be grouped into one predictor.
- Predictors that has no influence on the model or they were highly uncorrelated were removed.
- Elimination process will be conducted to produce a good model.

The predictors that were removed because they have not influence on the regression model are the follow:

$$Index, year$$

The predictors that were removed which has p-value that is larger than 10% are going to be removed by the backward elimination process.

We started having all the predictors selected be evaluated for the regression model. We obtained the following:

$$R^2: 0.6898 || \ Adjusted - R^2: 0.6859$$

year has a high p-value. The following R-value after removing both:

$$R^2: 0.6898 || \ Adjusted - R^2: 0.6859$$

Total.Bags, Small.Bags, Large.Bags, XLarge.Bags show low contribution to the model as well as high irrelevancy. They are to be removed:

$$R^2: 0.6893 || \ Adjusted - R^2: 0.6855$$

The following predictions coefficient is generated through RStudio after the backward elimination process were done based on the p-value be less than 10%:

```
Call:
lm(formula = AveragePrice ~ Total.Volume + X4770 + X4046 + X4225 +
    type + region, data = avocado)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09568 -0.16873 -0.01728  0.14527  1.51594

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.316e+00  1.490e-02  88.348  < 2e-16 ***
Total.Volume            2.650e-08  5.791e-09   4.575 4.79e-06 ***
X4770                  -3.905e-08  4.428e-08  -0.882 0.377866
X4046                  -3.194e-08  1.005e-08  -3.178 0.001483 **
X4225                  -4.939e-08  1.004e-08  -4.917 8.88e-07 ***
typeorganic             4.908e-01  4.243e-03 115.678  < 2e-16 ***
region                 -2.976e-01  2.093e-02 -14.219  < 2e-16 ***
Date                   -2.976e-01  2.093e-02 -14.219  < 2e-16 ***
---
Residual standard error: 0.2258 on 18022 degrees of freedom
Multiple R-squared:  0.6893,   Adjusted R-squared:  0.6855
F-statistic:   177 on 226 and 18022 DF,  p-value: < 2.2e-16
```

2. Data cleaning and sanity checking:

Only the columns with quantitative data were conducted the following operations:

Mean:

AveragePrice	Total.Volume	X4046	X4225	X4770	Total.Bags	Small.Bags	Large.Bags	XLarge.Bags
1.405978	850644	293008.4	295154.6	22839.74	239639.2	182194.7	54338.09	3106.427

Variance:

AveragePrice	Total.Volume	X4046	X4225	X4770	Total.Bags	Small.Bags	Large.Bags	XLarge.Bags
0.1621484	1.192698e+13	1.600197e+12	1.449906e+12	11548526005	972674070012	556782376191	59519391858	313038521

Maximum:

AveragePrice	Total.Volume	X4046	X4225	X4770	Total.Bags	Small.Bags	Large.Bags	XLarge.Bags
3.25	62505647	22743616	20470573	2546439	19373134	13384587	5719097	551693.7

Minimum:

AveragePrice	Total.Volume	X4046	X4225	X4770	Total.Bags	Small.Bags	Large.Bags	XLarge.Bags
0.44	84.56	0	0	0	0	0	0	0

Performing a sanity checking on the data set by doing backward elimination. It seems that we are on the right track. Nothing seems weird or not expected. Average mean with its variant seems correlated with the results obtained in the next section. Minimum value and maximum values to be considered as the outliers as they represent the extreme conditions.

3. Training and testing:

The data after backward elimination process will be trained and tested to see how well the train data represent the test data. Half of the data will be chosen randomly to represent the train set and the other half to represent the test set:

$$difference\_mean \pm (95\%\ confidence\ interval)$$
$$f = 0.5$$

In an ideal case, the mean of the difference would be zero with a tight range (approaching zero) of variation indicating a high accuracy.

RStudio computes the difference between results produced by the regression model and the actual value:

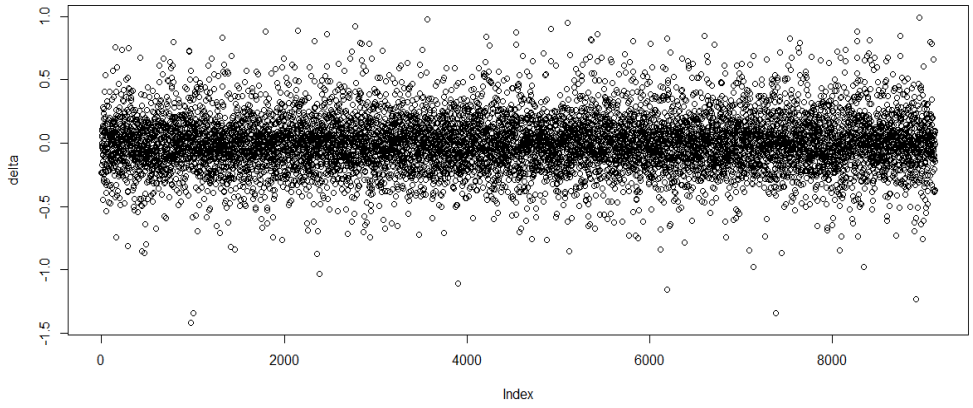
```
        One Sample t-test
data:  new
t = 0.47903, df = 9124, p-value = 0.6319
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.003439102  0.005663573
sample estimates:
  mean of x
0.001112235
```

This model predicts the house prices with 95% confidence interval:

$$\text{Difference (95\% confidence interval)} = 0.001112 \pm 0.004545$$

Prediction quality or explanation of poor predictions within the same data set:

The result generated after going through elimination prosses are great as the mean is very close to zero as well as the range is confidant interval indicates a high accuracy. This is shown when conducting a t-test when compare between the actual data and the train ones. The distribution behave like a Gaussian distribution with is good as it centered around the mean. The regression model is quite good in describing the actual value. It gives almost a perfect prediction. Below is the delta plot for the model:



Prediction quality or explanation of poor predictions across data set:

I have realized that if the trained data were significantly low as if 1%, we would still obtain almost the same results. These are the settings:

$$difference\_mean \pm (95\%\ confidence\ interval)$$
$$f = 0.01$$

The only thing increases is the variant which is to be expected with less trained data.

RStudio computes the difference between results produced by the regression model and the actual value:

```
        One Sample t-test
data:  new
t = 0.091832, df = 18066, p-value = 0.9268
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.003118601  0.003425184
sample estimates:
   mean of x
0.0001532915
```

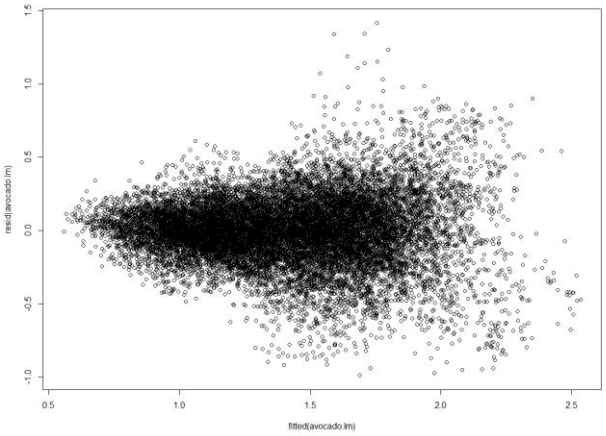
This model predicts the house prices with 95% confidence interval:

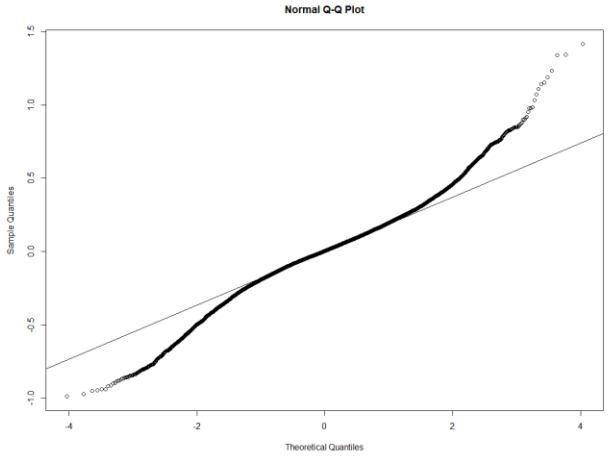
$$\text{Difference (95\% confidence interval)} = 0.000153 \pm 0.003265$$

4. Quality analysis:

- **P-values:**
  - This value indicates that the probability that the intercept is not relevant
  - The threshold was determined to be 10%. Meaning the chance that this specific intercept value is not relevant to the model is 10%

- **Residual:**
  - the residual for this regression model is decently good. The median is slightly shifted from being centered at zero. Almost ideally following Gaussian distribution and it has some of its attributes.
- **standard errors:**
  - A good Residual standard error should be about 1.5 times this standard error which is close enough in this model. This means that the residuals are distributed normally.
- **R^2 values:**
  - Our model has a value of 0.6893 which is very low. It means that this model only explains 68.93% of the data's variation.
- **residual analysis**
  - residual analysis shows if the residual values represents a quality model. Residual values represent how good our model in compared to the actual measured value: the closer to zero the better. The figure below says that the higher the price of avocado, the farther it deviates from zero which tells us the prediction the model become less accurate as the price rate increases.


resid(avocado.lm)
fitted(avocado.lm)
-1.0
-0.5
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
2.5


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-1.0
-0.5
0.0
0.5
1.0
1.5
-4
-2
0
2
4