MAS 637 - Applied Regression Analysis 1 Final Project

Qian Dong, Xiaoxi Yuan, Eduardo Santiago, & Philip Bachas-Daunert

Due Date: 2023-10-06

Contents

Abstract

Walmart as global retail leader has at its disposal a vast array of data reflecting its day-to-day operations which provide key insights into how different aspects of its business affect each other. Through this project, our group aims to make use of Walmart's extensive operational data to help determine what are the factors at play which affect Walmart's sales, we will also substantiate the holiday sales claims made within the problem statement by hypothesis testing, and search for significant differences in sales across the different store types. The primary objectives of our analysis will be met by conducting different analysis techniques learned in class ranging from: testing for significance amongst different variables, interpreting the impact of categorical vs numerical variables of a data set, making use of different plots to interpret our data in a visual point of reference, conduct hypothesis testing on claims made in the initial problem statement, and much more.

Introduction

With a market capitalization of over $430 billion, and over 4,600 stores in just the United States alone, Walmart by the very nature of its business is a treasure trove of operational data which lends itself to a wide array of statistical analysis to draw key insights into its business and its strategic decision making. The motivation behind selecting the Walmart case to conduct our final project stems from the valuable insight our group will be able to derive from analyzing the underlying factors affecting the sales performance of a company like Walmart, who is a leader in its industry, and in turn we will be well equipped when entering the labor market to make similar analysis for other companies' key performance metrics, revenue drivers, growth opportunities and much more. Through this project, our group will tackle different questions posed by the problem statement, first we will develop a parsimonious regression model to be able to interpret the variables at play which drive Walmart's sales. Second, we will address the holiday sales claim made that during the holiday season, sales are expected to be 10% more than when compared to non-holiday sales. More specifically, we will implement hypothesis testing techniques learned in class to validate this assertion and generate a confidence interval for the holiday sales coefficient ("IsHoliday"). Lastly, we will ascertain if in fact sales are significantly different across different store types and answer this claim with a statistical test. As we initiate our statistical analysis on Walmart's sales, we hope to successfully employ different techniques learned in class to draw actionable insights on this problem set and in turn allow us to sharpen our analytical toolbox which has become so coveted in today's vastly changing business landscape.

Analysis

We obtained the sales data of 45 Walmart stores from 2010 to 2012. From this data set, we can see the factors that may affect the sales revenue of Walmart stores: weather temperature, oil price, CPI, unemployment rate, whether it is a holiday, sales quantity, and store type. The average revenue of the store is $1,046,965, with a standard deviation of $564,366. After the initial descriptive analysis, we can calculate the correlation matrix to assess the linear relationships between the variables The correlation matrix helps us understand which independent variables are significantly correlated with the dependent variables and with each other Scatter plots are a useful tool for visualizing the relationships between the dependent and independent variables We also create scatter plots for pairs of variables to examine how they relate to each other Then We processed with linear regression modeling using the lm() function to build and interpret regression models.

Modeling

The model is defined as follows: `lm(Sales ~ Temperature + Fuel_Price + CPI + Unemployment + IsHoliday + Size + StoreType, data = d)`.

Bullet 1: Interpret the coefficients in the final model.

```
Call:
lm(formula = Sales ~ Temperature + Fuel_Price + CPI + Unemployment +
    IsHoliday + Size + StoreType, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-596469 -251071  -17468  152137 2695093

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.279e+05  5.010e+04   8.541  < 2e-16 ***
Temperature   1.293e+03  2.357e+02   5.483 4.34e-08 ***
Fuel_Price   -2.996e+04  9.160e+03  -3.271  0.00108 **
CPI          -1.390e+03  1.134e+02 -12.251  < 2e-16 ***
Unemployment -2.510e+04  2.346e+03 -10.697  < 2e-16 ***
IsHoliday     4.402e+04  1.604e+04   2.745  0.00607 **
Size          7.929e+00  1.047e-01  75.757  < 2e-16 ***
StoreTypeB    4.857e+04  1.194e+04   4.067 4.82e-05 ***
StoreTypeC    1.950e+05  1.918e+04  10.163  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 323300 on 6426 degrees of freedom
Multiple R-squared:  0.6723,    Adjusted R-squared:  0.6719
F-statistic:  1648 on 8 and 6426 DF,  p-value: < 2.2e-16
```

The following variables are significant:

- **CPI**
  - Other things remain unchanged, each additional change in CPI is associated with a decrease of approximately $1,390 in sales.
- **Fuel_Price**
  - Other things remain unchanged, each additional change in fuel price is associated with a decrease of approximately $29,960 in sales.
- **Size**
  - Other things remain unchanged, each additional change in store size is associated with an increase of approximately $7,929 sales.
- **StoreTypeB**
  - Other things remain unchanged, sales will increase $48,570 by choosing TypeB store.
- **StoreTypeC**
  - Other things remain unchanged, sales will increase $195,000 by choosing TypeC store.
- **Temperature**
  - Other things remain unchanged, each additional change in temperature is associated with an increase of approximately $1,293 in sales.
- **Unemployment**
  - Other things remain unchanged, each additional change in unemployment is associated with an increase of approximately $25,100 in sales.

Bullet 2: Our boss claims holiday sales will be 10% more than on a regular day. Is this claim correct? Conduct a hypothesis test to test the claim. Hint: you may want to generate a confidence interval for the holiday sales coefficient (isholiday).

$H_0 : S_{\text{holiday}} = 1.10 \times S_{\text{regular}}$

```
Call:
lm(formula = Sales ~ IsHoliday, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-861539 -493609  -86047  373185 2773972

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1044714       7295 143.212   <2e-16 ***
IsHoliday      32184      27586   1.167    0.243
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 564400 on 6433 degrees of freedom
Multiple R-squared:  0.0002115, Adjusted R-squared:  5.613e-05
F-statistic: 1.361 on 1 and 6433 DF,  p-value: 0.2434
```

$y = 1044714 + 32184 \times x_{\text{IsHoliday}}$

```
The min percent is:  -3.464%
```

```
The max percent is:  9.626%
```

Since the highest percentage of increase in sales in terms of whether it was a holiday or not was 9.626%, we failed to accept Hypothesis H_0 which claims that the sales in holiday will be 10% less than a regular day.

Bullet 3: It is believed that sales are significantly different across the different store types. Can you refute this claim using a statistical test?

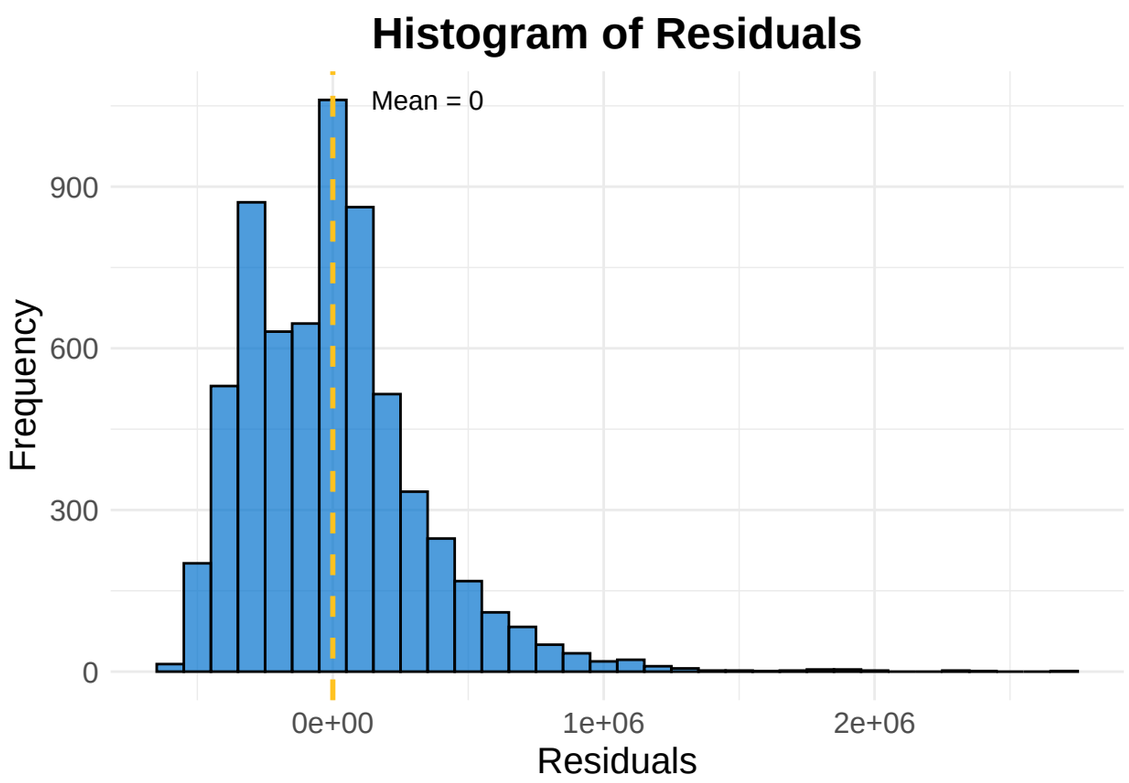
```
Call:
lm(formula = Sales ~ dummy_stores, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-1166687  -264314   -31638   196653  2926063

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1376674       8022  171.60   <2e-16 ***
dummy_storesB  -553678      12151  -45.57   <2e-16 ***
dummy_storesC  -904059      17330  -52.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 450000 on 6432 degrees of freedom
Multiple R-squared:  0.3645,    Adjusted R-squared:  0.3643
F-statistic:  1845 on 2 and 6432 DF,  p-value: < 2.2e-16
```

We assume different store types will influence significantly on sales. We made a regression model to test it. The results showed that all types of store A, B, C have a p-value < 0.05 , which means we couldn't refute the claim.


Histogram of Residuals
Mean = 0
900
600
300
0
Frequency
0e+00
1e+06
2e+06
Residuals

Conclusion

After concluding our analysis of the Walmart Sales data, our team was able to answer the following questions posed by the problem set:

1. Through analyzing the coefficients of the model, we were able to determine that the following variables were statistically significant due to them having a p-value below 0.05 (CPI, Fuel_Price, Size, the dummy variables introduced: StoreTypeB & StoreTypeC, Temperature, and Unemployment). The model also had an Adjusted R-squared of 0.6719, which indicated a relatively good fit of our regression model to the data provided, as 67.23% of the variability in the response variable "Sales" is explained by the predictor variables in the model. The overall model had a p-value of less than 2.2e-16 indicating that there was a very strong level of statistical significance amongst the variables analyzed.

2. The 2nd bullet in the Walmart Sales problem asked us to determine whether the claim that holiday sales will be 10% more than non-holiday sales. This prompted us to set the null hypothesis: Sales(holiday) = 1.10 x Sales(non-holiday) and the alternate hypothesis would be refuting this claim. Our team was able to predict Sales as a function of the variable "IsHoliday" to arrive at the estimates for the intercept and the "IsHoliday" variable. After obtaining these data points, we were able to construct a simple regression equation which helped us determine that the highest percentage increase in sales in terms of whether it was a holiday or not was 9.626%, thus we failed to accept the null hypothesis which claims that the sales in holiday will be 10% more than a regular day.

3. The last bullet in this problem prompted us to analyze if sales are significantly different across the different store types per the data in the Walmart.csv file. Again, our team was able make use of the techniques learned in class and construct 2 dummy variables (dummy_storesB & dummyStoresC) and run a regression model to determine the influence of store type on sales. The results showed that all types of stores A, B, C have a p-value below 0.05, which meant that we couldn't refute the claim that sales are significantly influenced by the different store types.