Data 603: Walmart Model

Britain Van Bergeyk, Matthew Haddad

June 11, 2024

1 Introduction

With how crucial it is to collect and analyse data in today's era, data-based decision-making has become paramount for corporations seeking to stay competitive and meet evolving consumer demands. As Walmart is now the leading retailer worldwide[6] it is imperative that they can make informed decisions that lead to improved profitability, enhanced customer experiences, and sustained growth. This project aims to create a statistical model that could help Walmart navigate complex market dynamics and capitalize on emerging opportunities.

The goal of this research is to study the effects of a selection of econometric variables on the sales at Walmart locations in the United States. More specifically we are interested in how consumer price index, unemployment rate and fuel price as well as other factors such as temperature and holidays affect the purchasing patterns of consumers.

Our intent is to make a statistical model that can accurately predict sales trends in Walmart stores contained in our dataset. We may be able to gain insights into the importance of some of these predictors in the more general retail sphere. One of the main predictors of our model is which Walmart store, so for interpolation purposes we can only make predictions about sales at a given store in the dataset.

This research has significant importance to each and every one of us as we are all consumers affected by sales trends daily. Consequently, identifying which external factors have the most significant impact on sales is our foremost problem.

2 Methodology

2.1 Data

The data for our project was obtained from Kaggle provided for free by a user named BharatKumar0925[2]. This data was imported into pandas for the purposes of partitioning the data into a training and test set with a 70/30 split, 70% of the data was used to train the model and 30% to test it afterwards.

The dataset contains 8 fields:

i) Store: A series of 45 stores across the USA.

ii) Date: dd-mm-yyyy from 05-02-2010 to 26-10-2012.

iii) Weekly Sales: Weekly sales in USD.

iv) Holiday Flag: 1 or 0 indicating if weekly sales happened during a Holiday.

v) Temperature: Average temperature in Fahrenheit each week.

vi) Fuel Price: Fuel price in store region.

vii) CPI: Consumer Price Index

viii) Unemployment: Unemployment rate each week.

Consumer Price Index is an economic measure of the change in the price of all goods and services that households purchase for personal use [3]. This means that as the CPI increases goods tends to cost more on average for the consumer.

We are treating Weekly Sales as the dependent variable and everything else except for date are going to be treated as independent variables. Date will not be considered as a predictor, however this column will be used to check if there is independence in the error terms.

2.2 Approach

The aim of our research is to create a multiple linear regression model and perform diagnostics using the methods acquired in Data 603. With the model in hand we will make inferences about which predictors are the most important to the weekly sales trends at Walmart stores.

2.3 Workflow

The general workflow for this research is to import the data and perform wrangling tasks (e.g. partitioning into test and training sets, data type correction). With the data we will create the full first order model. We find the best first order model using step-wise t-tests. This is a test on each coefficient individually that tests the following hypothesis:

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

Where β_i refers to the coefficient of the i-th predictor. In the case of the categorical variable Store we will reject the null hypothesis for all dummy variables if any of them are determined to be significant. Note that all tests in this paper will have significance level of $\alpha = 0.05$.

With the best additive model we probe for interactions by first including every interaction then removing all insignificant ones at once according to their p-values in the t-test. With the terms all removed we compare the newest model to the full model using an F-test. This test has the hypotheses:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$

$$H_a : \text{at least one } \beta_i \neq 0,\ i = p-q+1, \cdots, p$$

Where p is the total amount of predictors in the full interaction model and q is the number of terms removed by their t-test p-values. The reason we do not perform an individual t-test for each coefficient is because the full interaction model will have 280 unique coefficients meaning that an individual t-test on each coefficient will be quite cumbersome. We will compare this reduced interaction model to the full interaction model by their adjusted R^2 to see if it has better explanation of variance.

With this interaction model we will then look for higher order terms by first observing a plot of that weekly sales against that variable and observing there is a non-linear relationship, then going through each of these variables and performing a t-test on the coefficients of the higher order variable to check for significance. This t-test is the same as the one above just on the higher order coefficients.

Once we have found all significant predictors using these methods we will begin diagnostics. This means looking into the different assumptions required for a linear regression model:

1. Linearity: We test this graphically using a residuals versus fitted plot.

2. Independence: We can check the independence of errors assumption by plotting the residuals against the date that the data point was recorded.

3. Homoskedasticity: The variances of the error terms can be tested for homoskedasticity using a scale-location plot as well as applying a Breusch-Pagan test to the data. The Breusch-Pagan test has the following hypotheses:

$$H_0 : \text{heteroskedasticity is not present}$$

$$H_a : \text{heteroskedasticity is present}$$

4. Normality Assumption: The normality assumption will be tested using a QQ-plot of the residuals alongside a Shapiro-Wilks test for normality. The Shapiro-Wilks test has hypotheses:

$$H_0 : \text{the sample data are significantly normally distributed}$$

$$H_a : \text{the sample data are not significantly normally distributed}$$

5. Multicollinearity: The multicollinearity assumption will be tested using the VIF between the different predictors.

6. Outliers: We will check for outliers using a residuals vs leverage plot as well as the Cook's distance of the residuals.

If any of these assumptions fails we will do our best to remedy it by using transformations, however if the issues persist then we will proceed with the best model we have and begin testing the model using the test set that was left untouched. Using this we will be able to estimate the error in prediction and get an idea of how effective the model is for interpolation. Even if the model turns out to not be useful for interpolation we can still glean insights from the significance of certain predictors.

2.4 Workload Distribution

Britain will perform any data-wrangling tasks required such as partitioning the data, as well as testing the final model. Matthew will perform the tests on the assumptions throughout the building process. We will both be actively involved in creating the model at all stages and will meet frequently to collaborate.

3 Results

3.1 Data Cleaning and Wrangling

The first step we took to analyzing this data was to ensure it was cleaned and wrangled appropriately. As far as cleaning goes there weren't any null values to worry about and the data types were mostly correct aside from the dates which were given as strings. So the first cleaning step was to convert them into datetime objects in pandas. The wrangling process was fairly straightforward with the only significant step being partitioning the data into a training and a test set. We did this just in case the model does not meet the assumptions to see if it could be effective nonetheless. To create the partitions we used pandas and our method was to randomly sample 70% of the data for each given store and put these records into a new dataframe called TrainingSet. The other 30% of the data was then left untouched in a dataframe called TestSet.

3.2 Model Building

3.2.1 Additive Model

The first model we built is the full additive model with all the predictors included. Store and Holiday Flag are both qualitative variables and they have 45 and 2 levels respectively. The other predictors included in the full model were Temperature, Fuel Price, Consumer Price Index and Unemployment Rate. This means that our model had 49 variables in total with 45 of them being dummy variable encodings of the two qualitative predictors.

Due to the impracticality of writing out all the dummy variable coefficients for Store we have decided to put the actual numbers in the Appendix 5.1. There are 44 dummy variables for the stores and we will label them X_{Store2}, X_{Store3}, $X_{Store4}, \cdots, X_{Store45}$. We can put all of these dummy variables into a vector

$$\mathbf{X_{STORE}} = \begin{pmatrix} X_{Store2} \\ X_{Store3} \\ \vdots \\ X_{Store45} \end{pmatrix}$$

For the coefficients we will call them generally $\beta_{Store2}, \beta_{Store3}, \cdots \beta_{Store45}$ and they will be held in a vector:

$$\boldsymbol{\beta_{STORE}} = \begin{pmatrix} \beta_{Store2} \\ \beta_{Store3} \\ \vdots \\ \beta_{Store45} \end{pmatrix}$$

The additive model can now be written with dot product notation:

$$\begin{aligned} \hat{Y_{WeeklySales}} = 1354490 + \boldsymbol{\beta_{STORE}} \cdot \boldsymbol{X_{STORE}} + 57210X_{HolidayFlag} - 810X_{Temperature} \\ -43992X_{FuelPrice} + 2678X_{CPI} - 23169X_{Unemployment} \end{aligned}$$

The actual values in $\boldsymbol{\beta_{STORE}}$ are contained in the Appendix (Figure 15).

For each quantitative variable we have that the p-values are all less than 0.05. This means that we will not remove any of the quantitative variables from our additive model as they are all significant. The Holiday Flag dummy variable also has a p-value less than 0.05 so we will keep it in the model. For the Store dummy variables there are exactly 6 of them whose p-values are greater than 0.05, however, the remaining 38 p-values are all less than 0.05 so we will also keep the Store predictor in our final additive model. All in all this means that the full additive model is also the best additive model and none of the terms should be dropped.

3.2.2 Interaction Model

The first iteration of the interaction model included all interaction terms. We then removed terms according to their p-value from the t-test

The full interaction model:

$$\begin{aligned} \hat{Y}_{\text{Weekly Sales}} = \quad & -18860000 + \boldsymbol{\beta_{STORE}} \cdot \boldsymbol{X}_{\text{STORE}} + 778600X_{\text{Holiday Flag}} + 71690X_{\text{Temperature}} \\ + \quad & 2105000X_{\text{Fuel Price}} + 97650X_{\text{CPI}} + X_{\text{Unemployment}} \\ + \quad & \boldsymbol{\beta_{STORE*Holiday_{Flag}}} \cdot \boldsymbol{X}_{\text{STORE}} \times X_{\text{Holiday Flag}} \\ + \quad & \boldsymbol{\beta_{STORE*Temperature}} \cdot \boldsymbol{X}_{\text{STORE}} \times X_{\text{Temperature}} \\ + \quad & \boldsymbol{\beta_{STORE_{Fuel_{P}rice}}} \cdot \boldsymbol{X}_{\text{STORE}} \times X_{\text{Fuel Price}} \\ + \quad & \boldsymbol{\beta_{STORE*CPI}} \cdot \boldsymbol{X}_{\text{STORE}} \times X_{\text{CPI}} \\ + \quad & \boldsymbol{\beta_{STORE*Unemployment}} \cdot \boldsymbol{X}_{\text{STORE}} \times X_{\text{Unemployment}} \\ + \quad & 1685X_{\text{Holiday Flag}} \times X_{\text{Temperature}} \\ - \quad & 47280X_{\text{Holiday Flag}} \times X_{\text{Unemployment}} \\ + \quad & 2675X_{\text{Temperature}} \times X_{\text{Fuel Price}} \\ - \quad & 3672X_{\text{Temperature}} \times X_{\text{CPI}} \\ - \quad & 10610X_{\text{Fuel Price}} \times X_{\text{CPI}} \end{aligned}$$

Where all the $\boldsymbol{\beta}$ values can be found in the appendix (Store 16, Store:CPI 19, Store:Fuel price18, Store:Temperature17, Store:Unemployment 20, Store:Holiday Flag 26)

All interactions for dummy variables STORE will be considered significant if there is at least one significant interaction.

Interaction	$\alpha = 0.05$	# That Passed
$\boldsymbol{X_{STORE}} * X_{HolidayFlag}$	$< \alpha$	1
$\boldsymbol{X_{STORE}} * X_{Temperature}$	$< \alpha$	26
$\boldsymbol{X_{STORE}} * X_{FuelPrice}$	$< \alpha$	28
$\boldsymbol{X_{STORE}} * X_{CPI}$	$< \alpha$	4
$\boldsymbol{X_{STORE}} * X_{Unemployment}$	$< \alpha$	2

The non dummy variable interactions that passed were:

Interaction	$\alpha = 0.05$
$X_{HolidayFlag} * X_{Temperature}$	p-value $= 0.004189 < \alpha$
$X_{HolidayFlag} * X_{Unemployment}$	p-value $= 0.041257 < \alpha$
$X_{Temperature} * X_{FuelPrice}$	p-value $= 0.000132 < \alpha$
$X_{Temperature} * X_{CPI}$	p-value $= 0.000103 < \alpha$
$X_{FuelPrice} * X_{CPI}$	p-value $= 0.008540 < \alpha$

The above methodology will be done again for the reduced model:

Interaction	$\alpha = 0.05$	# That Passed
$\boldsymbol{X_{STORE}} * X_{Temperature}$	$< \alpha$	27
$\boldsymbol{X_{STORE}} * X_{FuelPrice}$	$< \alpha$	25
$\boldsymbol{X_{STORE}} * X_{CPI}$	$< \alpha$	6
$\boldsymbol{X_{STORE}} * X_{Unemployment}$	$< \alpha$	4

The non dummy variable interactions that passed were:

Interaction	$\alpha = 0.05$
$X_{HolidayFlag} * X_{Temperature}$	p-value $= 0.010967 < \alpha$
$X_{Temperature} * X_{FuelPrice}$	p-value $= 3.46 \times 10^{-5} < \alpha$
$X_{Temperature} * X_{CPI}$	p-value $= 4.85 \times 10^{-4} < \alpha$
$X_{FuelPrice} * X_{CPI}$	p-value $= 7.87 \times 10^{-3} < \alpha$

The final interaction model was:

$$
\begin{aligned}
\hat{Y}_{\text{Weekly Sales}} = \quad & -9137000 + \boldsymbol{\beta_{STORE}} \cdot \boldsymbol{X}_{\text{STORE}} - 5946X_{\text{Holiday Flag}} + 53140X_{\text{Temperature}} \\
+ \quad & 1587000X_{\text{Fuel Price}} + 54310X_{\text{CPI}} - 21410X_{\text{Unemployment}} \\
+ \quad & \boldsymbol{\beta_{STORE*Temperature}} \cdot \boldsymbol{X}_{\text{STORE}} \times X_{\text{Temperature}} \\
+ \quad & \boldsymbol{\beta_{STORE_{FuelPrice}}} \cdot \boldsymbol{X}_{\text{STORE}} \times X_{\text{Fuel Price}} \\
+ \quad & \boldsymbol{\beta_{STORE*CPI}} \cdot \boldsymbol{X}_{\text{STORE}} \times X_{\text{CPI}} \\
+ \quad & \boldsymbol{\beta_{STORE*Unemployment}} \cdot \boldsymbol{X}_{\text{STORE}} \times X_{\text{Unemployment}} \\
+ \quad & 1112X_{\text{Holiday Flag}} \times X_{\text{Temperature}} \\
+ \quad & 2820_{\text{Temperature}} \times X_{\text{Fuel Price}} \\
- \quad & 301.6X_{\text{Temperature}} \times X_{\text{CPI}} \\
- \quad & 8300X_{\text{Fuel Price}} \times X_{\text{CPI}}
\end{aligned}
$$

Where all the $\boldsymbol{\beta}$ values can be found in the appendix (Store 21, Store:CPI 24, Store:Fuel price23, Store:Temperature22, Store:Unemployment 25)

We performed an F-test between this model and the full model to determine if this model is significant:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Value
Regression	50	1.2482^{12}	2.4964^{10}	1.0576
Residual	4220	9.9605^{13}	2.3603^{10}	
Total	4270	1.0085^{14}		

Table 1: Anova Table for F-test Between Reduced Interaction Model and Full Interaction

The p-value for this F-test was 0.3645 meaning we fail to reject the null hypothesis and conclude that the predictors removed were insignificant.

3.2.3 Higher Order Terms

To check for higher order terms we will first look at the pairs plot and examine if any of the relationships between the quantitative predictors and the Weekly Sales seem higher order, if they do then we will try modelling those predictors with higher order terms.

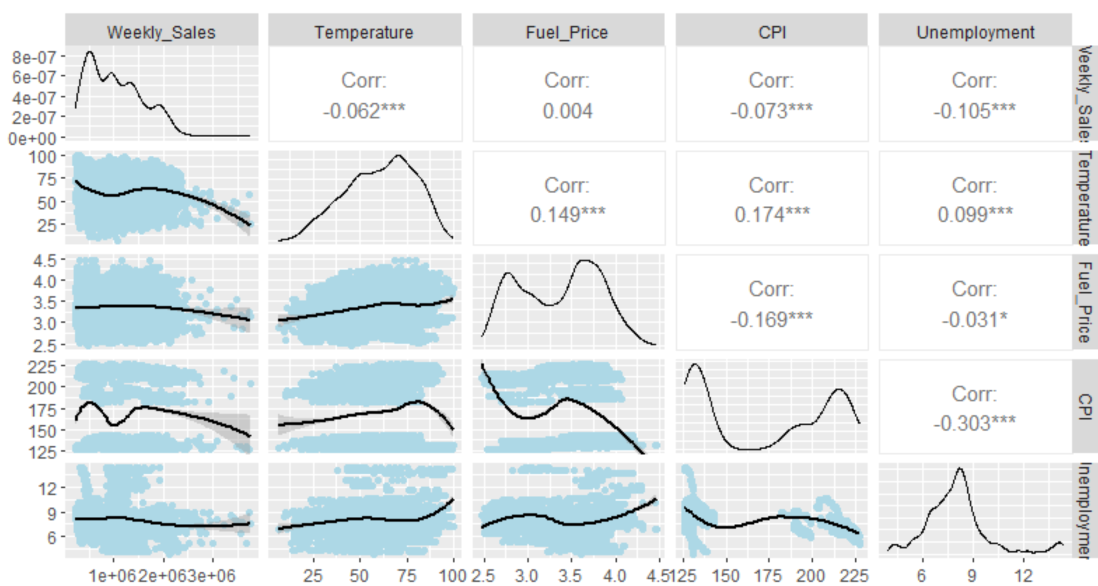

Figure 1: Scatter Plot Matrix of the Quantitative Predictors and the Dependent Variable Weekly Sales

Looking at the first column on this plot we can see the relationship between the predictors and weekly sales. It appears that Temperature and CPI may have a higher order relationship so we will try adding these into the full interaction model one at a time and do a t-test each time we add a new term, we will also compare the adjusted R^2 values before and after adding them to see if the model has improved in explaining the variance.

After adding the quadratic term for Temperature we have an R^2 of 0.9261 which is the same as the full interaction from before. The p-value of the new quadratic term is 0.062994 which is greater than our α of 0.05 so we will not add this term into the model as it is insignificant.

Now we try adding higher order terms for CPI. When adding the second order term we see that the R^2 increases to 0.9263 which is higher than the previous model. The term has a p-value of 7.76×10^{-5} which is less than our alpha of 0.05 so we will include the second order term for CPI. Now we test the cubic term, the model with the second and third order CPI terms has the same R^2 value of 0.9263, however the p-value of this term is 0.176912 which is larger than 0.05, so it is insignificant in our model and we should limit CPI to a quadratic relationship. The final model we get that includes interactions and higher order terms is given as follows:

$$
\begin{aligned}
\hat{Y_{WeeklySales}} = {} & 125224800 - 9192.481 X_{HolidayFlag} + \boldsymbol{\beta_{STORE}} \cdot \boldsymbol{X_{STORE}} + 79679 X_{Temperature} \\
& + 4594754 X_{FuelPrice} - 1247810 X_{CPI} + 3136 X_{CPI}^2 + 52310 X_{Unemployment} \\
& + \boldsymbol{\beta_{STORE*Temperature}} \cdot \boldsymbol{X_{STORE}} \times X_{Temperature} \\
& + \boldsymbol{\beta_{STORE*FuelPrice}} \cdot \boldsymbol{X_{STORE}} \times X_{FuelPrice} \\
& + \boldsymbol{\beta_{STORE*CPI}} \cdot \boldsymbol{X_{STORE}} \times X_{CPI} \\
& + \boldsymbol{\beta_{STORE*Unemployment}} \cdot \boldsymbol{X_{STORE}} \times X_{Unemployment} \\
& + 1212 X_{HolidayFlag} \times X_{Temperature} \\
& + 3401 X_{Temperature} \times X_{FuelPrice} \\
& - 433 X_{Temperature} \times X_{CPI} \\
& - 22237 X_{FuelPrice} \times X_{CPI}
\end{aligned}
$$

Where all the $\boldsymbol{\beta}$ values can be found in the appendix (Store 27, Store:Temperature 28, Store:Fuel price29, Store:CPI30, Store:Unemployment 31)

3.3 Diagnostics

3.3.1 Linearity

To check for the assumption of linearity in the residuals of our model, a residuals versus predicted values was plotted. Given the residuals are spread equally a horizontal fitted line should show indicating linearity.

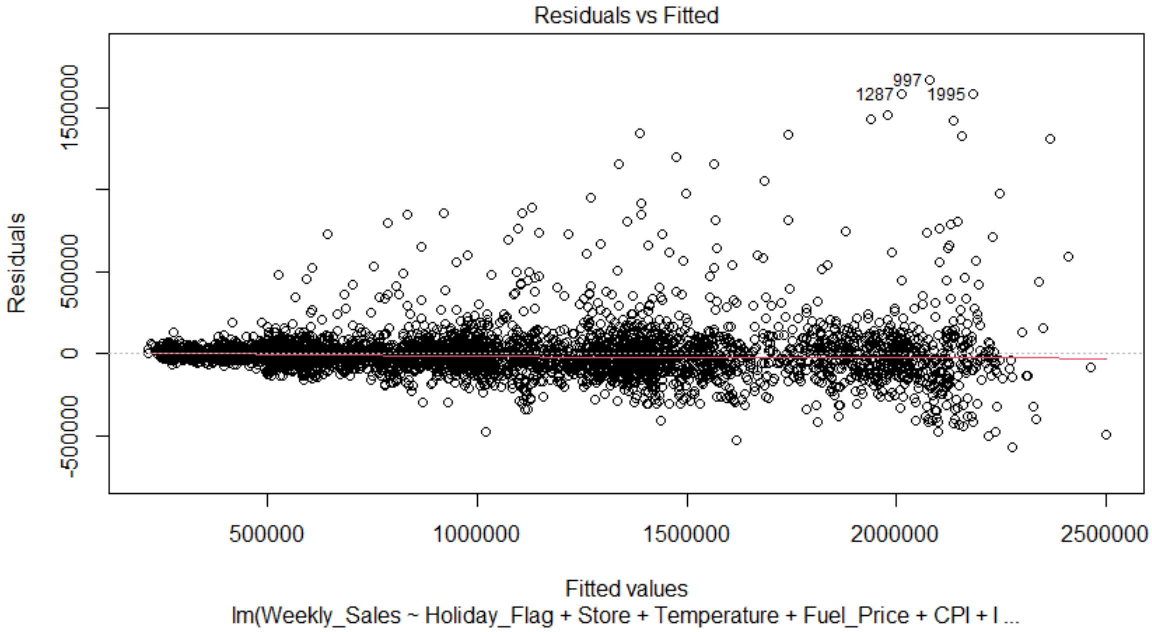

Figure 2: Scatter Plot of the residuals of the higher order model vs predicted values

From the plot above, we observed a distinct straight horizontal line across the plot. This indicates that there is no discernible pattern in the distribution of the residuals, indicating that the linearity assumption is satisfied. The model adequately captures the relationship between the independent and dependent variables.

3.3.2 Independence

The Independence assumption is checked by plotting our residuals against time. The assumption will be held if the data does not display incidents of clumping, or serial correlation.

Figure 3: Scatter Plot of the Residuals of the higher order model vs Date

From the above plot we see that each store has its residuals spread out randomly indicating that we do not have serial correlating. However, we do see two incidences of clumping during the month of December. This clumping is most likely due to the holiday season, and the increase in sales that follows suit. Because of this clumping, we cannot definitively say that the Independence assumption holds.

3.3.3 Homoskedasticity

When testing for Homoskedasticity, we are looking for equal variance in the residuals against the fitted data. For the assumption to hold we should expect to see no patterns or trends in the residuals plot.

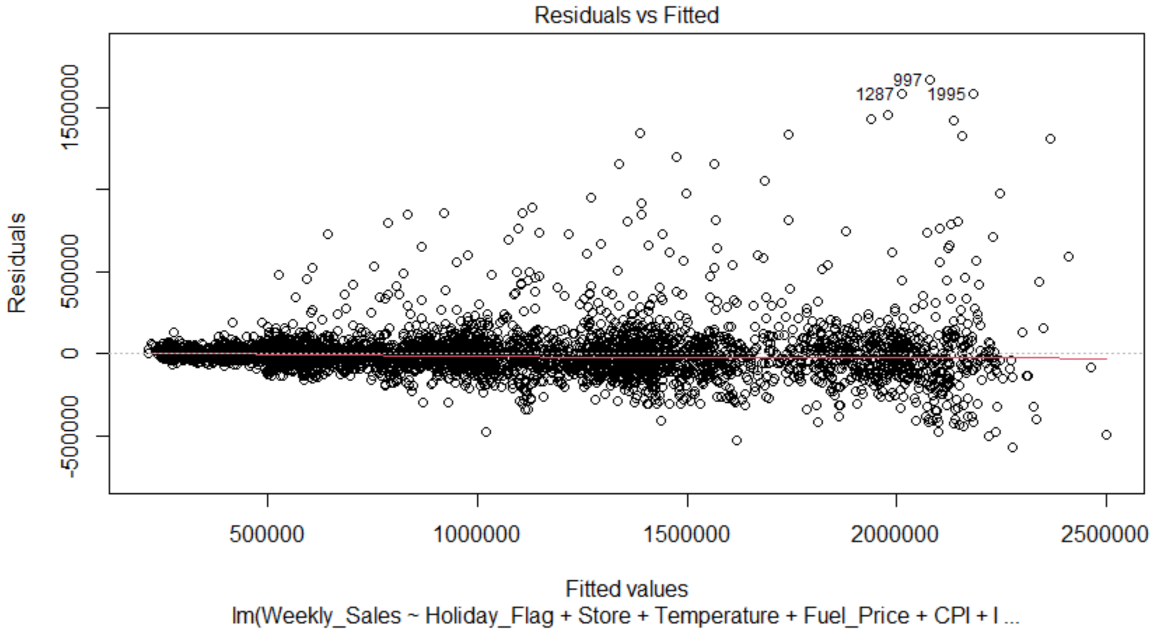

Figure 4: Scatter Plot of the Residuals vs. Fitted for Higher Order Model

We can see clearly in the plot above, that the residuals follow a conical shape, increasing in variance. This data clearly does not pass the equal variance assumption and displays heteroskedasticity. Furthermore, the Breusch-Pagan test yielded a $p-value = 2.2 \times 10^{-16} < \alpha = 0.05$, we then reject the null hypothesis that there is equal variance.

3.3.4 Normality

The normality assumption will be tested using three methods: Histogram of residuals, normal probability plot, Shapiro-Wilk test. In order for the normality assumption to hold, the residuals need to be normally distributed. Normality will be shown in the histogram if a bell curve is formed by the residuals. Normality will be shown in the normal probability plot if the residuals follow the same pattern as the normal line, and normality will be assumed if the p-value in the Shapiro-Wilk test is greater than our level of significance.

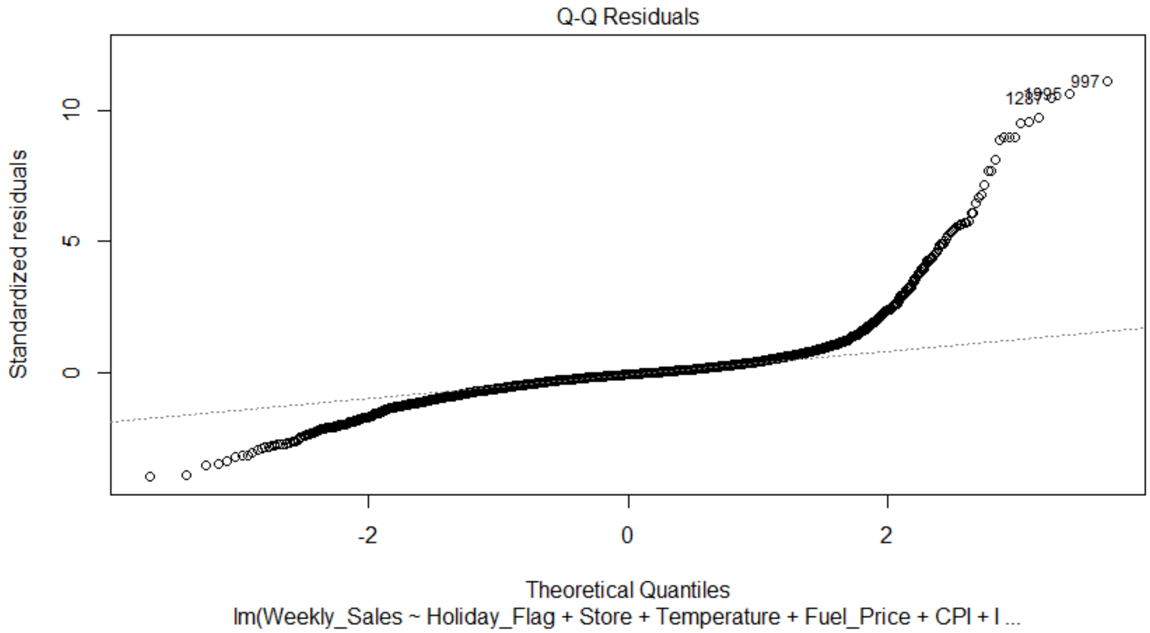

Figure 5: QQplot of the residuals versus Normal Quantile

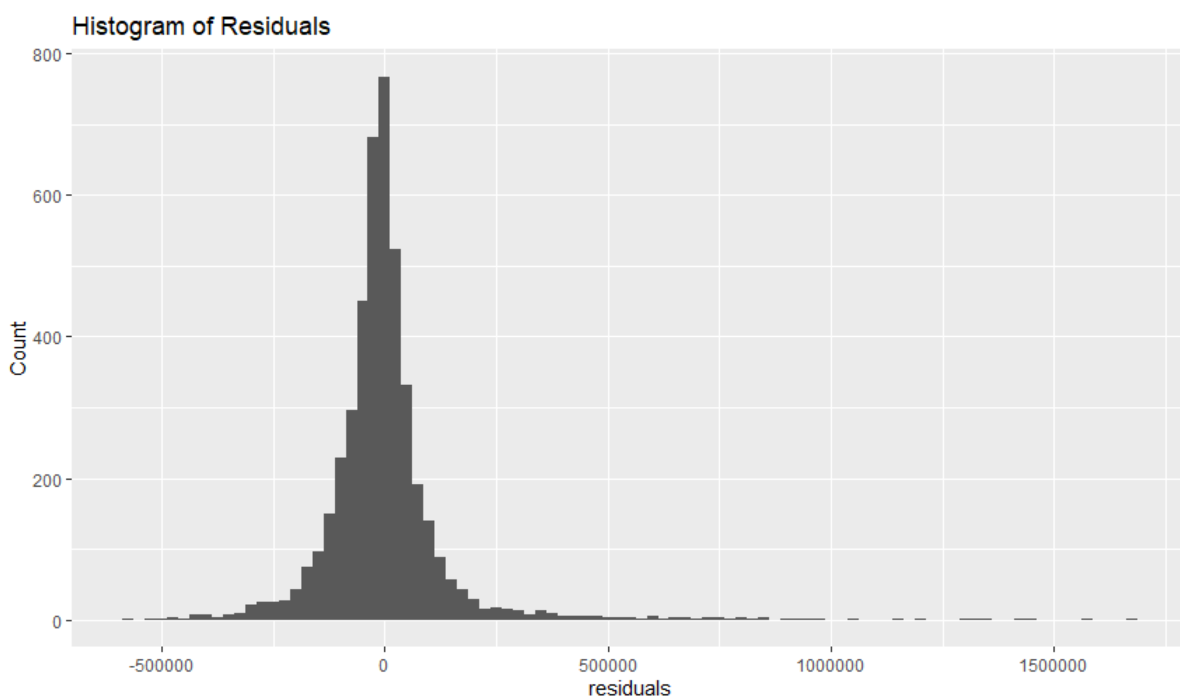

Figure 6: Histogram of Residuals in Higher-Order Model

The residuals in the histogram follow a fairly normal distribution, but the right tail of the distribution is indicative that the residuals might not be completely normal. The Q-Q plot shows snaking at the left and right tails, with a much larger deviation occurring at the right tail. The Q-Q plot does not indicate normality within the residuals. The Shapiro-Wilk test returned a $p - value = 2.2 \times 10^{-16} < \alpha = 0.05$ therefore we reject the null hypothesis that the residuals are normally distributed. The normality assumption cannot be held.

3.3.5 Multicollinearity

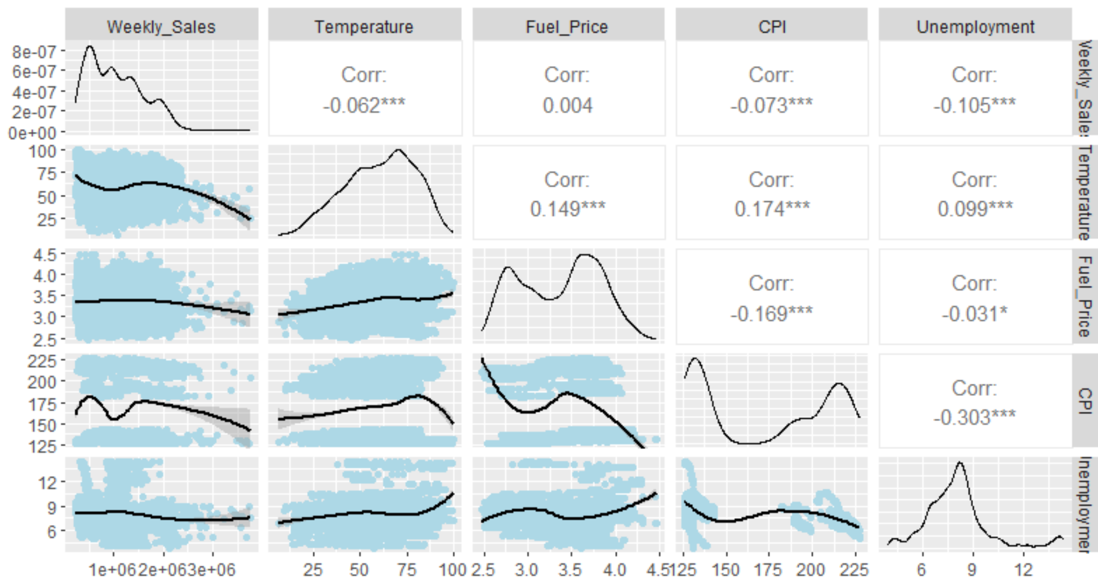

Figure 7: Scatter Plot Matrix of the Quantitative Predictors and the Dependent Variable Weekly Sales

Multicollinearity was examined using variance inflation factors (VIF). The VIF values for Temperature, Fuel Price, CPI, and Unemployment were all < 1.3 indicating no significant correlation between any two predictors. A correlation matrix was also ran to test for high $(r > 0.5)$ correlation coefficients of each predictor. All correlation coefficients were below said threshold.

3.3.6 Outliers

We will be testing outliers with cooks distance and leverage points.

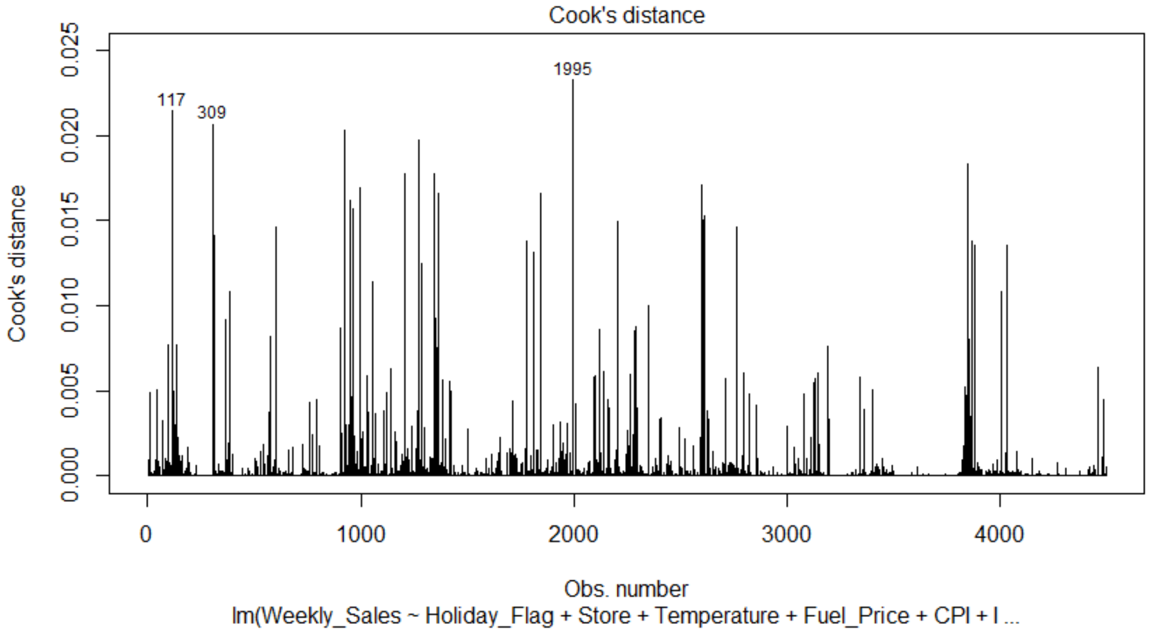

Figure 8: Bar Graph of Cook's Distance vs. Input Data

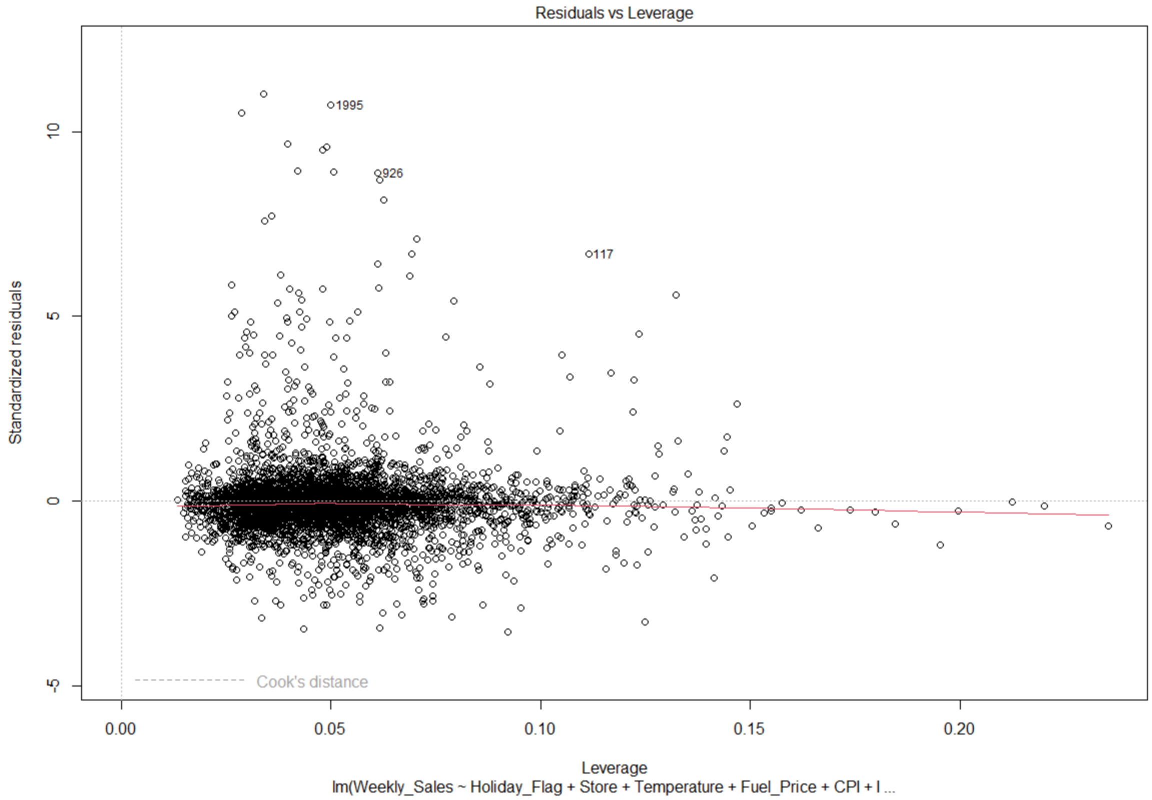

Figure 9: Scatter Plot of Residual vs. Leverage for Complete Model

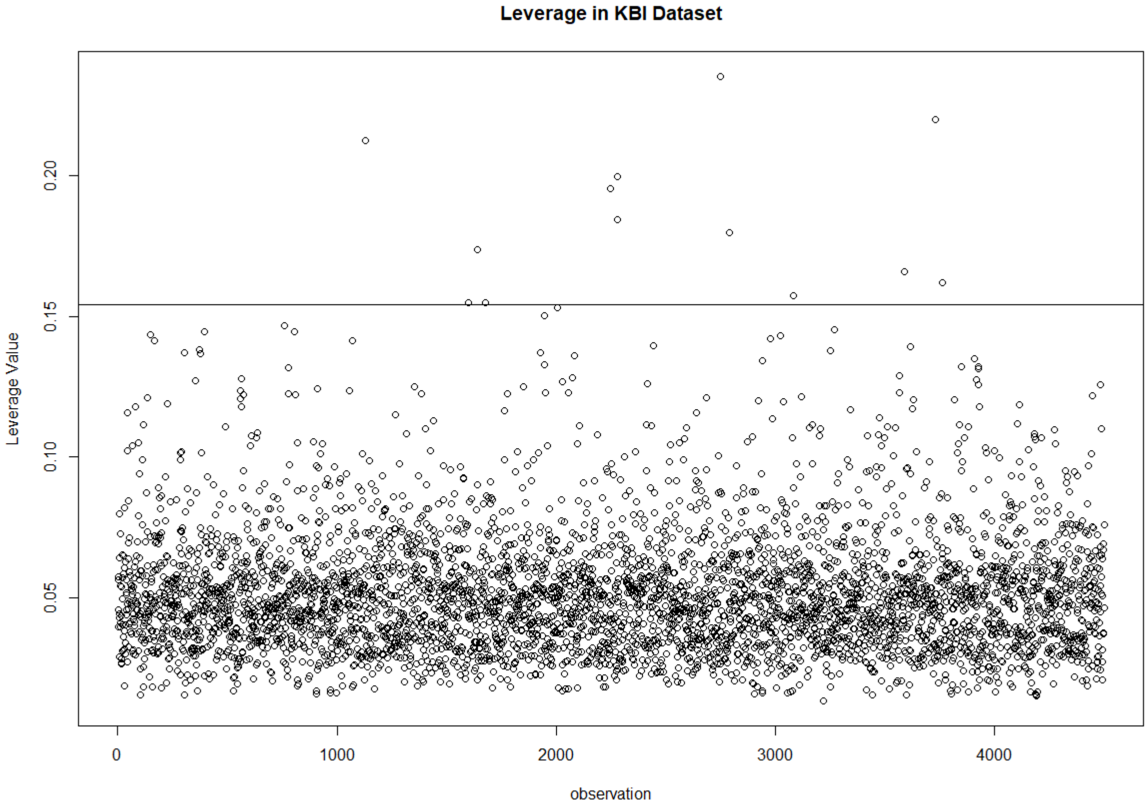

Figure 10: Scatter Plot of Leverage vs. Input Data

Figure 8 and figure 9 show no outliers that have a cooks distance greater than 0.5, this indicates no outliers. However, figure 10 shows that we have multiple points with a high leverage($> 3 * p/n$) influencing the slope of least squares. These points will be considered outliers and removed from the data.

3.4 Corrective Measures

A Box-Cox transformation was done to fix the unequal variance and normality assumptions. A transformation with $\lambda = -0.2727273$ was found. Where $\hat{Y}_{WeeklySales}$ is transformed as $(\hat{Y}^{\lambda} - 1)/\lambda = \hat{W}$

$$\begin{aligned}
\hat{W} = {} & 5.247236 + 0.0003386684X_{HolidayFlag} \\
& + \boldsymbol{\beta_{STORE}} \cdot \boldsymbol{X_{STORE}} + 0.001389164X_{Temperature} + 0.05756416X_{FuelPrice} \\
& - 0.01669500X_{CPI} + 0.00004201712X_{CPI}^2 + 0.0003534621X_{Unemployment} \\
& + \boldsymbol{\beta_{STORE*Temperature}} \cdot \boldsymbol{X_{STORE}} \times X_{Temperature} \\
& + \boldsymbol{\beta_{STORE*FuelPrice}} \cdot \boldsymbol{X_{STORE}} \times X_{FuelPrice} \\
& + \boldsymbol{\beta_{STORE*CPI}} \cdot \boldsymbol{X_{STORE}} \times X_{CPI} \\
& + \boldsymbol{\beta_{STORE*Unemployment}} \cdot \boldsymbol{X_{STORE}} \times X_{Unemployment} \\
& + 0.00001231717X_{HolidayFlag} \times X_{Temperature} \\
& + 0.00005690816X_{Temperature} \times X_{FuelPrice} \\
& - 0.000007427014X_{Temperature} \times X_{CPI} \\
& - 0.0002822293X_{FuelPrice} \times X_{CPI}
\end{aligned}$$

Where W represents the boxcox transformed data:

$$W = \frac{Y_{WeeklySales}^{-0.2727273} - 1}{-0.2727273}$$

All the $\boldsymbol{\beta}$ values can be found in the appendix (Store 32, Store:Temperature 33, Store:Fuel price 34, Store:CPI 35, Store:Unemployment 36)

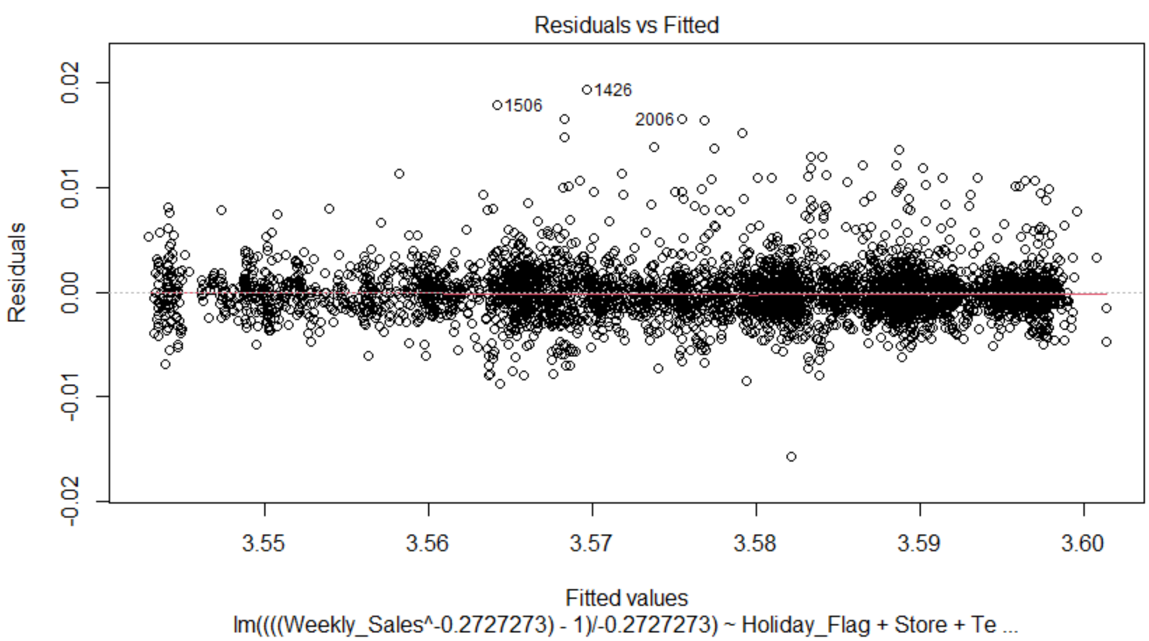

Figure 11: Scatter Plot of Residuals vs. Fitted for BoxCox Model

The breusch-Pagan test for the transformed data returns a $p-value = 2.2 \times 10^{-16} < \alpha = 0.05$ in which we reject the null hypothesis that there is homoskedasticity. Figure 11 shows a much more random variance in the residuals. Although the conical shape found in the higher order model is no longer found, there is still a small hump at fitted value 3.57 indicating that the variance is not fully random. The evidence shows that we cannot fully accept the equal variance assumption but also displays that significant improvement was made.

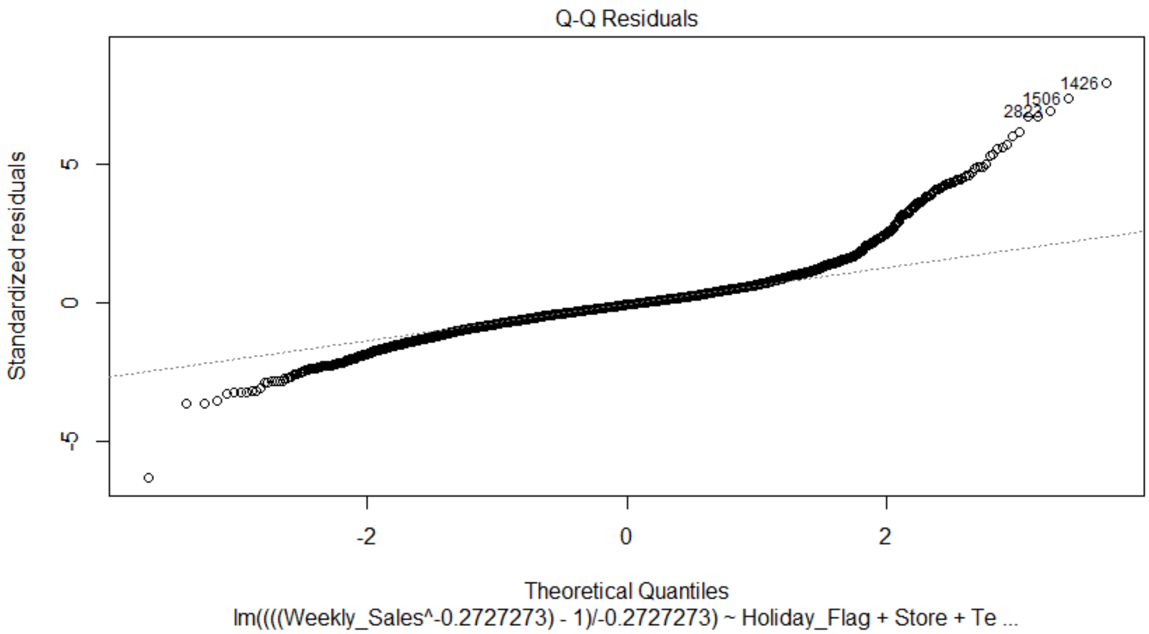

Figure 12: QQplot of residuals in BoxCox Transformed Model

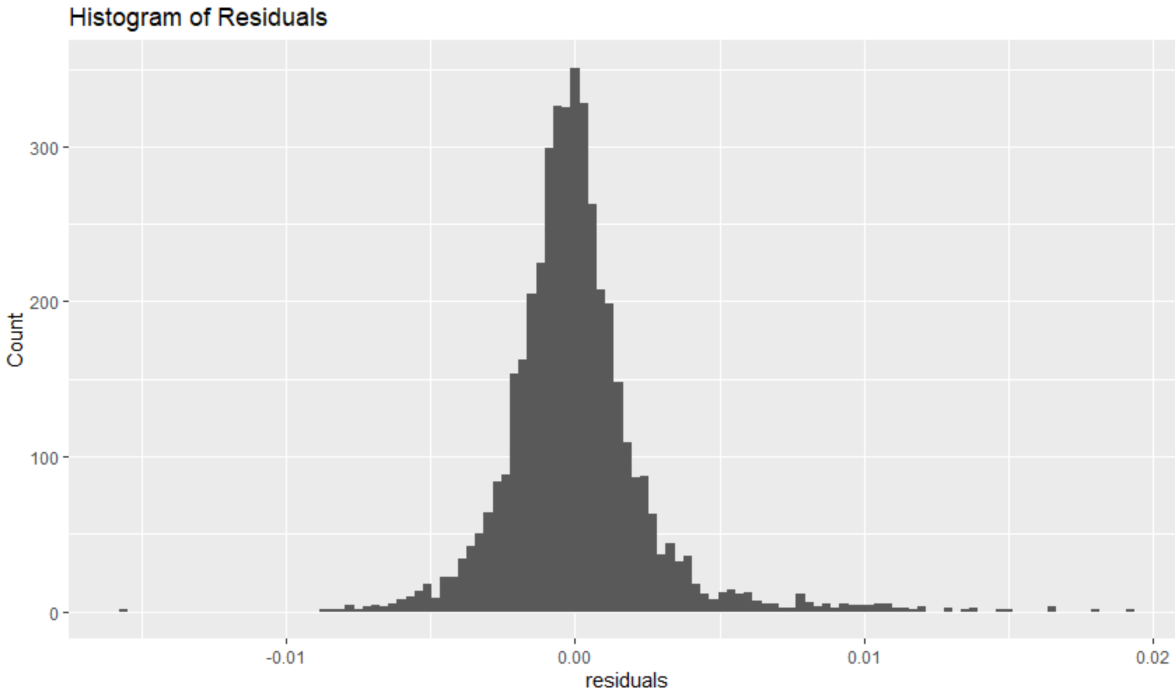

Figure 13: Histogram of the Residuals for Final Model

Figure 12 shows no improvement in the normal probability plot compared to figure 5. This indicates the normality assumption is not valid. Conversely, figure 13 shows a much tighter normal distribution when compared to figure 6. The long right side tail in figure 5 has now been reduced significantly indicating that the normality assumption may not be completely valid but much better. The Shapiro-Wilk test returns the same $p - value = 2.2 \times 10^{-16} < \alpha = 0.05$ in which we reject the null hypothesis that the residuals follow a normal distribution.

3.5 Testing

To test the efficacy of predicting with this model we employed the test set to validate the final box-cox transformed model. To do this validation we computed the standard deviation of the errors and compared this to the residual standard error within the model. If the values are similar then we will be able to conclude that the model was not overfit. Depending on the value of the error in the prediction we can deduce how effective the model is for predicting.

The first step was to perform the same box-cox transformation on the test data specifically on the Weekly Sales column. After this we can compute our predictions and take the difference between the predicted values and the actual values in the data. The distribution of these is visualized in Figure 14:

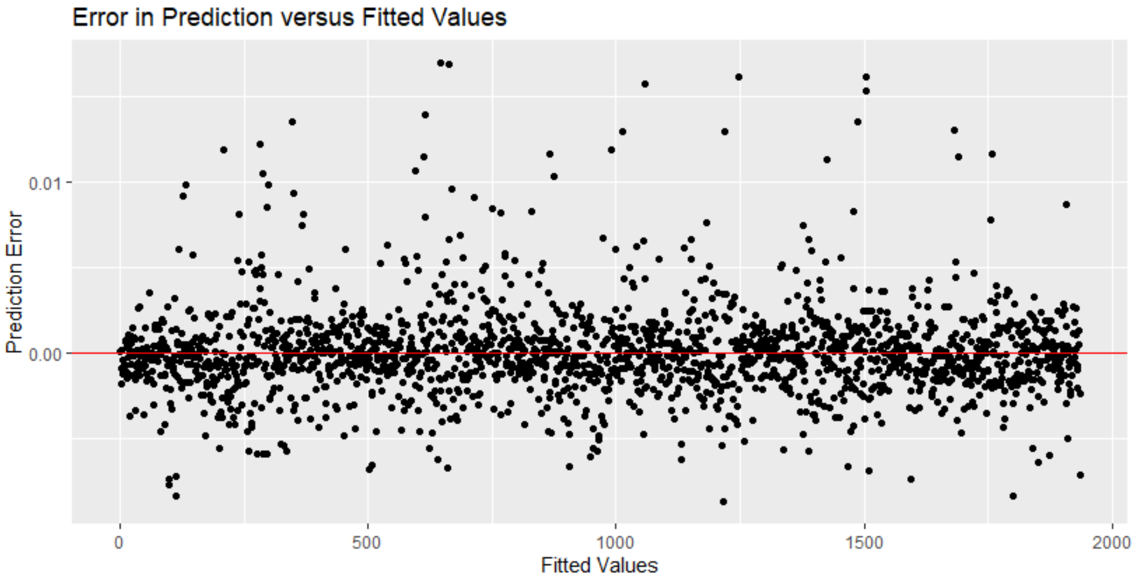

Figure 14: Scatter Plot of Prediction Errors versus Fitted Values

If we compare Figure 14 to Figure 11 we can see that the distribution of the errors is similar between the test data prediction and the model's residuals. To give a concrete comparison between the two we can compute the standard error values for both distributions and take their quotient. The RSE for the model is computed by R to be 0.002542. The standard error for the predictions was calculated using the standard deviation formula:

$$\sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

We determined it to be 0.002675304. This gives an overall ratio between standard error in the model and the predictions of 1.068412. This means that the model applied to data it was not trained on will perform almost the exact same as it performed on the data it was trained on, only around a 6% increase. This tells us that the model was not overfit and should indeed be quite effective at making predictions.

4 Conclusion and Discussion

4.1 Approach

After performing diagnostic tests and applying a power transform to the response we got a model that performed as well on validation as on the training data. This means that our model can be used for predictive purposes on data from the 45 stores in the dataset. We cannot expect the same performance on recent data nor on data for other Walmart stores or retail in general. In spite of our validation process, we suspect the model was overfit by including the store variable. Overcoming this facinorous issue given the dataset may not be feasible as the variability between stores is so great that in absence of this predictor the model has no significance.

We can see from the additive model that all first order terms were significant meaning that each of the fields in the dataset have some linear relationship with Weekly Sales at these Walmart Stores. Given that our final model is Box-Cox transformed and contains higher order and interaction terms it is very hard to provide an analysis. We can however say that Holidays and Temperature have a positive relationship with weekly Walmart sales, also observe that CPI has a negative relationship with Weekly Sales as its linear term is many orders of magnitude large than the quadratic term. This seems reasonable since as CPI increases, the cost of all goods for consumers tends to increase. This may mean that consumers would avoid shopping at retail stores such as Walmart during times of high CPI, pushing weekly sales down. There are interactions between CPI and Temperature as well as CPI and Fuel Price. It is worth noting that CPI and Fuel Price necessarily have a correlation since Fuel is a good that consumers purchase and thus included in the calculation of CPI. In spite of this we kept both terms in the model due to the low VIF calculated and a relatively small correlation. Furthermore, a statistical report on Walmart sales done by Rashmi Jeswani[1] concluded significance in interaction between temperature and CPI which is in accordance with our findings. However, Rashmi Jeswani[1] further concluded an interaction between temperature and CPI with unemployment, of which both interactions we did not find significant.

4.2 Future Work

One of the primary ways this model could be improved is by incorporating the time series data into the model itself. We saw in Figure 3 that there is a large amount of clumping in the residuals in the month of December. This is very intuitive as we would expect the holiday season to cause a much larger influx of consumers than in other months due to Christmas shopping habits. If we treated the time as another predictor we would likely have much better prediction capability since the model would contain information about when shopping blows up and slows down seasonally. Alternatively, instead of incorporating Date we may be able to add another flag for Christmas specifically. In the figure clumping occurs only around December so if we trained a new model with this flag we may no longer see the clumps in this season meaning that the independence assumption would be met [4].

Another limitation of this model is it only applies to data obtained from one of the 45 stores included in the training of the data. It may be possible to apply to other Walmart stores or even potentially other supermarkets in general, however, we would need to determine a method of associating a general store with one of the stores in the training data. We propose comparing sales data between the store of interest and all the stores in the dataset and associating this store with the nearest store in the dataset according to some metric. Then we could test the model against the data from the new store and determine if the model accurately predicts the outputs using the same methods as in section 3.

Further improvement could be found if we included different predictors from other datasets, such as types of items sold in a given week, state, size of the store. Adding more factors could improve the prediction capability of the model as well as provide further insight into what elements the weekly sales rely on. It would be increasingly important, however, to validate the model if we were to add more variables since the more predictors there are, the more likely we are to overfit the data. There is a problem in model selection called Freedman's Paradox[5] which says that as more predictors are added to the model some may spuriously pass significance tests and remain in the model.

4.3 Conclusion

In conclusion, our model has a very large R^2 and a very low RSE, however we believe that this model is only useful for interpolating data from the stores contained within this data set and in this time period. The failure of assumptions, even when transformed, indicates a lack of certainty about the significance of our predictors; we cannot guarantee which predictors are influential. In spite of this, it is possible that this data could be valuable in other areas of the retail industry. This paper provides a fascinating study into financial model building, and provides some insights into factors that affect sales at given stores in the United States.

5 Appendix

5.1 Coefficients of Store Dummy Variables

Since we are modelling with 45 levels of a categorical variable it can be quite gruesome trying to make our model look palatable while still conveying the important information. To help remedy this we are using vector notation within the body of the paper and we will not insert the actual values but they are kept here for reference.

Note that these coefficients begin with the dummy variable for store 2 since there is no dummy variable for store 1.

```
 [1] 378178 -1174469   716823 -1276265   -19332
 -947358  -698308 -1071116   624282  -223678  -167778
[12] 652021  546765  -719385 -1034705  -478216
-248122    99845   527977  -808144  -310583   -12736
[23] 20398  -855168  -345302   431600   139647
-757695 -1123152  -169019  -322088 -1024752  -323595
[34] -423915 -1179473 -1035490  -794608   -84423
-460379  -274648  -736862  -853289 -1054270  -670487
```

Figure 15: Store Dummy Variable Coefficients for Additive Model

```
 [1] -12930676.6  -2110530.6   1274805.0    178996.5  -4262764.5
   -6704830.6   -443934.0  -1473628.6  11677156.9  -4412495.5    -598952.8
    17868923.6   7383741.6   7612616.3    421076.8  11423848.3   4308579.9
[18]   6949871.9   -409352.6  -4638937.0   1663789.2  10996640.4
   4211403.7    196989.7   7592060.6  10698831.9   3612076.5   3534763.9
    3313958.8  -4137768.4  -6994332.3   7475455.3   2826272.4   7411372.9
[35]   5678560.0    846600.7   1862516.9 -18156124.5  11791698.6
   -9756325.0  10445140.2    770056.8  10735687.3    833870.5
```

Figure 16: $\boldsymbol{\beta_{STORE}}$ Coefficients For Full Interaction Model

```
 [1]  -1158.8150   2032.9360 -33577.4206   2149.9514   3385.3012
 [6]  -3784.6662    963.0289   2668.2402 -39302.0734    182.3659
[11] -30369.5500 -32634.4491  -7668.1297 -27434.7821  -4085.0106
[16] -29585.6119 -28192.4596 -28436.7746  -4458.6485   1564.8338
[21] -27748.1149 -28251.7885 -24721.1510     73.9072 -26196.7265
[26] -22458.1568 -29785.7549 -27185.1301   2315.8285    378.9533
[31]  -4994.8739 -29405.0047 -29098.5290 -22831.6278   2699.1177
[36]   1565.7623 -27461.2255   3526.2919 -27165.6543  -3221.4792
[41] -30545.6525   1620.1754 -29606.5002  -9959.1858
```

Figure 17: $\boldsymbol{\beta_{STORE*TEMPERATURE}}$ Coefficients For Full Interaction Model

```
 [1]  -284241.238   -20087.023 -1104634.010   -31334.743  -198683.834
 [6]  -205240.334   -14787.574    15324.280 -1044574.814  -116168.149
[11] -1023522.108  -849344.691  -322183.839  -837483.965  -196252.313
[16]  -752526.102 -1016840.029  -989644.339  -155790.771   -55393.654
[21]  -912166.054  -941353.559  -991187.622  -134793.203  -891277.262
[26]  -812480.704  -937670.708  -869877.507   -19143.073   -28927.939
[31]  -196414.464  -928518.168 -1031550.823  -791485.407    -1361.779
[36]   -17984.456  -939617.628  -271822.690  -851160.869  -140008.274
[41]  -860756.731   -84028.685  -861704.759  -358290.884
```

Figure 18: $\boldsymbol{\beta_{STORE*FUEL_{P}RICE}}$ Coefficients For Full Interaction Model

```
 [1]  54863.1971   2344.5541  65985.3832  -7230.1271  16483.5025
 [6]  28211.8832  -3522.9206  -1398.1738 -12583.3693  16449.6386
[11]  63458.1255 -63730.0461 -28648.2946 -11730.5872  -2987.5523
[16] -30351.0189  22995.7413  10461.6056   8867.1710  12450.4498
[21]  32044.2976 -20342.2457  30387.1177  -1128.6472   -266.9867
[26] -28760.1058  31062.4064  17662.9072 -18560.0262  14372.5534
[31]  31769.0912  -4295.9378  37219.3235 -35733.8745 -30611.3143
[36]  -8237.4096  39719.9583  76277.1538 -32063.3861  42292.1782
[41] -23807.5105  -3434.3541 -28857.0698   6595.2309
```

Figure 19: $\boldsymbol{\beta_{STORE*CPI}}$ Coefficients For Full Interaction Model

```
  [1]     6595.231  323123.461   48575.463 -399538.526   11929.820
 [6]   161134.232  142090.681   78287.876   75801.373 -410431.164
[11]   132165.845 -326675.832 -522093.287     729.515 -265874.968
[16]    41346.682 -448475.815 -304883.997 -379958.528  -12959.246
[21]   160729.019 -195019.561 -429045.721 -408137.515  -44376.493
[26]  -393800.925 -254534.553 -333708.305 -267668.139  -69028.098
[31]   123721.161  169158.980 -366406.301 -334670.600  119710.685
[36]   -58473.375  -19356.414 -367446.051  294296.540 -433756.506
[41]   307931.965 -404856.718  -82261.350 -405882.860
```

Figure 20: $\boldsymbol{\beta_{STORE*UNEMPLOYMENT}}$ Coefficients For Full Interaction Model

```
 [1] -11517177.6  -1517880.6  -5427644.4   -715945.3  -4288236.0
   -6905124.8  -1398341.6  -2111863.2   8371754.6
[10]  -3809160.3    843651.1  14162290.1   7556654.1   4594172.0
   -1538545.8   6572931.5   1866029.2   3972065.3
[19]   -452880.0  -2819284.8  -1496834.3   5120714.3    979076.5
   -220224.1   3739509.1   7828329.8   2733827.2
[28]   1362603.1   4863301.7  -2588602.1  -7279652.1   4880122.6
   246221.9   5109893.8   8328746.6   3279941.8
[37]   3078699.6 -16074328.1   5589674.8 -11243066.5   7240729.6
   2761045.9   6357585.4   1059990.6
```

Figure 21: $\boldsymbol{\beta_{STORE}}$ Coefficients For Reduced Interaction Model

```
  [1]   -1242.9979   2017.1664 -27112.9950   2791.8634
 [5]    3848.2978  -2882.4404   1518.2528   3420.0112
 [9]  -34184.4585    322.5490 -27702.1490 -26502.6567
[13]   -6142.8310 -22450.0755  -2038.6245 -23547.4463
[17]  -23516.9639 -23192.8408  -3516.8494   1772.2973
[21]  -23080.3152 -21653.1176 -20309.5931    561.9019
[25]  -21154.6861 -17774.4269 -27695.2087 -23086.2726
[29]    2584.4748    662.3179  -4208.5576 -24099.1659
[33]  -24634.0791 -18354.9806   2777.4931   1880.5147
[37]  -24611.6430   2862.7300 -20600.8771  -1457.8560
[41]  -25102.7428   1119.3959 -23379.1265  -8338.9536
```

Figure 22: $\boldsymbol{\beta_{STORE*TEMPERATURE}}$ Coefficients For Reduced Interaction Model

```
 [1] -276179.9471   -6336.3104 -925588.6862  -21795.7168
 [5] -193703.2695 -145037.4488   -8144.9499   21597.8457
 [9] -844206.9737 -103770.4611 -791652.3242 -632556.2423
[13] -239996.5315 -632361.0136 -136022.7869 -553231.3697
[17] -812827.9009 -778547.1715 -110805.7368  -34795.6890
[21] -718756.3350 -746323.4654 -802885.4425 -108116.5449
[25] -689193.4426 -628737.4620 -764367.8054 -666163.7432
[29]    -691.6592  -10414.4433 -141265.1095 -711767.5095
[33] -816346.8653 -586373.7432   33784.8838   14116.0577
[37] -707288.7355 -246924.9682 -656234.1652  -79424.7432
[41] -644063.6851  -43194.4358 -651173.0670 -276816.7283
```

Figure 23: $\boldsymbol{\beta_{STORE*FUEL_{P}RICE}}$ Coefficients For Reduced Interaction Model

```
  [1]  49389.8962    601.5304  87975.5365  -3153.9574
 [5]  17327.0043  24064.1984   1305.8097   2229.4912
 [9] -15783.0382  14712.0218  25698.9042 -65060.9547
[13] -35581.9209 -15966.3105   1548.6123 -22781.9706
[17]  14917.1058   5172.8621   6553.3226   5084.2345
[21]  29751.4708  -4734.4962  30619.9545  -1215.5301
[25]    467.1226 -31222.5945  12046.7210   7263.1557
[29] -24894.8659   8014.3057  28280.2902 -13309.5321
[33]  27810.5382 -44811.3536 -41799.1227 -18600.3769
[37]   3239.8254  67836.6789 -14312.9595  44551.1297
[41] -28746.6208 -14650.0956 -25204.5356   -484.4496
```

Figure 24: $\boldsymbol{\beta_{STORE*CPI}}$ Coefficients For Reduced Interaction Model

```
  [1]  290732.6348   12018.6166    2121.5305     768.7123
 [5]  129734.0857  245759.9908   52343.4580   38696.6966
 [9]  -28540.6061   93133.7801   38228.4117 -137382.6568
[13]  109083.3543   93654.2572  144890.8115  -54373.5258
[17]   51362.9076  -12648.9697   29222.1391  119390.6050
[21]  156486.4769  -46631.2590  -83303.1549   -1498.0657
[25]  -12257.3733   67788.3523   43934.3600   97239.9372
[29] -103675.7418   89365.0854  270475.2606   10252.6586
[33]   53068.1641  476372.8089 -104832.5917  -61992.9600
[37]     511.2793  257715.6103  -43030.5251  409496.6514
[41]  -19247.8919  -48041.9857  -14618.4436   -6990.8550
```

Figure 25: $\boldsymbol{\beta_{STORE*UNEMPLOYMENT}}$ Coefficients For Reduced Interaction Model

```
 [1]   -11986.147  -152509.655  -169313.020  -139699.758  -117449.405
 3781.908   -98531.784  -197577.261   -57589.135  -112083.105   125288.955
 -230270.938   -87155.237  -112465.873  -147720.406  -118536.055
-115844.522
[18]  -192375.336  -177704.159  -127769.314  -130353.674  -257773.808
    -30251.829
-60891.144  -117951.759  -105812.511   294003.243   -75580.848  -132692.429
-154967.323     3326.950  -212409.710   -82859.725  -161896.749
[35]  -132366.561  -149854.163    45362.237    78900.146  -249763.974
-116669.750  -179406.408   -22381.209  -264844.673  -108799.628
```

Figure 26: $\boldsymbol{\beta_{STORE*HOLIDAY_{F}LAG}}$ Coefficients For Full Interaction Model

```
 [1]  -14591760.3    1655213.0  -94409955.7   -1394798.7   -3594300.3
 -34511304.7    2400052.9     699220.8  -81314670.8    -473483.9
[11]  -88595939.3  -74688118.6  -26755886.9  -79848095.4  -28401223.4
-81796759.9  -82503816.0  -80442548.8   -7425166.5   -6322885.9
[21]  -82517468.4  -78933289.4  -83626273.0   -7112287.2  -80453726.5
-73574168.6  -86692746.2  -82959516.4    2094590.5   -5565737.0
[31]  -35123421.8  -84668880.9  -88492876.7  -75897848.4    3681777.2
-344616.5  -86359919.8  -20330877.0  -78285895.7  -39491488.2
[41]  -82262105.5   -7253195.7  -82563823.0  -33224985.2
```

Figure 27: $\boldsymbol{\beta_{STORE}}$ Coefficients For Higher Order Interaction Model

```
 [1]   -1448.62455    2509.79778  -38604.94580    2979.76887
 3810.84560   -5983.35528    1698.27609    3734.49574
 [9]  -45949.23819     756.13755  -39478.56075  -38109.78061
 -10153.69422  -33290.89759   -5175.41874  -35235.77835
[17]  -34281.18646  -34000.22266   -4621.82135    1634.84044
-33290.87037  -32425.25864  -31150.21839    -588.35977
[25]  -31928.47330  -28064.67139  -39502.24974  -33825.55811
2473.94943     586.27191   -7168.87770  -35860.61050
[33]  -36181.36611  -28549.63962    2700.27863    1702.73318
-36403.78721    2746.58142  -31352.50485   -4420.42823
[41]  -36877.43991     -98.33066  -34999.98227  -12276.58245
```

Figure 28: $\boldsymbol{\beta_{STORE*Temperature}}$ Coefficients For Higher Order Interaction Model

```
 [1]  -298857.859    -1252.978 -2167637.187   -37929.627
 -185044.274  -440844.618    24173.513    42643.385
 [9] -2087978.931   -94109.461 -2033035.973 -1856866.832
 -668271.153 -1791986.908  -420554.895 -1769749.125
[17] -1966897.903 -1936424.216  -202664.150   -65865.431
-1815818.127 -1897621.748 -1960792.919  -202732.381
[25] -1836660.956 -1736585.154 -2006640.508 -1821022.533
-24893.049   -37028.095  -445292.026 -1955815.383
[33] -2053781.077 -1690973.149   -25844.030   -31343.236
-1948038.522  -298877.492 -1807270.690  -374054.926
[41] -1883772.752  -189313.986 -1879520.910  -706063.036
```

Figure 29: $\boldsymbol{\beta_{STORE*FuelPrice}}$ Coefficients For Higher Order Interaction Model

```
 [1]  62765.2181 -14870.2601 634543.4919   -801.0675
 13756.7514 164569.7197 -17136.2931 -12025.0488 536347.4141
[10]  -1397.8200 576117.4992 479985.2912 147709.0050
494591.7180 138915.2122 518790.8521 523767.3922 515405.7836
[19]  44641.7015  20229.7209 512865.2714 502099.5120
542512.1311  36772.7080 507919.2675 455812.0130 562482.7879
[28] 515906.1956 -12622.9932  21100.0957 170198.6696
537945.1500 572544.0901 439101.4377 -20384.2286  -1386.2150
[37] 553667.7526  87601.9494 491344.4625 187930.8405
521981.5247  36023.6651 520519.2541 182739.1316
```

Figure 30: $\boldsymbol{\beta_{STORE*CPI}}$ Coefficients For Higher Order Interaction Model

```
[1]  326849.251   35438.487  -70400.131   49026.937  148592.601
235598.537   86666.521   88408.937 -100234.510
[10]  112129.725  -36844.711 -208521.677  -12282.368    8079.912
128971.458 -133799.113  -17879.710  -97187.095
[19]  -74342.012  164672.463   82410.349 -123113.421 -171060.314
-111553.333  -82185.085  -21387.180  -31832.506
[28]   25652.706  -76612.634  121347.578  258370.612  -64258.311
-21771.232  390987.043  -76715.429  -54349.400
[37]  -74777.576  278371.121 -123982.823  424842.931  -92934.748
-83284.939  -85923.623 -130180.630
```

Figure 31: $\boldsymbol{\beta_{STORE*Unemployment}}$ Coefficients For Higher Order Interaction Model

```
 [1] -0.122868887 -0.155861723 -1.210394963 -0.199202212 -0.061757393
 -0.846285554  0.001033304 -0.118383277 -1.125620140
[10] -0.011780550 -1.192499772 -1.035677467 -0.370419609 -1.061784673
-0.486161885 -1.070244729 -1.056531210 -1.075351960
[19] -0.100239825 -0.210938749 -1.107595123 -1.163578523 -1.121716362
-0.125132889 -1.081870417 -0.995469294 -1.117824801
[28] -1.163396237  0.112976784 -0.078055791 -0.487814357 -1.292620303
-1.223201216 -1.015027123  0.326939272  0.003378092
[37] -1.304461270 -0.242581636 -1.045836315 -0.533990573 -1.046912185
-0.072164283 -1.128140518 -0.459050712
```

Figure 32: $\boldsymbol{\beta_{STORE}}$ Coefficients For BoxCox Model Trained With Outliers Removed

```
 [1] -4.219725e-06 -6.891562e-05 -6.393063e-04  3.685717e-06
 5.417860e-05 -9.153412e-05
 [7]  3.284927e-07  2.774945e-05 -7.162202e-04  1.143658e-05
 -7.036596e-04 -6.367845e-04
[13] -1.890880e-04 -5.687697e-04 -4.613402e-05 -6.244883e-04
-5.855846e-04 -5.848853e-04
[19] -5.732673e-05  1.179302e-05 -5.747458e-04 -5.601394e-04
-5.470288e-04 -7.052200e-06
[25] -5.496583e-04 -5.087567e-04 -6.903188e-04 -6.068598e-04
7.069935e-06  1.582052e-06
[31] -1.319097e-04 -6.076991e-04 -6.425497e-04 -4.697497e-04
2.927307e-05 -1.563646e-05
[37] -6.607727e-04  3.772188e-05 -5.417587e-04 -8.845609e-05
-6.799954e-04 -1.243749e-05
[43] -6.035620e-04 -2.317686e-04
```

Figure 33: $\boldsymbol{\beta_{STORE*Temperature}}$ Coefficients For BoxCox Model Trained With Outliers Removed

```
 [1] -2.849659e-03 -2.304412e-03 -2.644325e-02 -2.704051e-03
 [5] -2.611774e-03 -7.250836e-03 -1.966291e-06  3.434501e-04
 [9] -2.625566e-02 -1.376057e-03 -2.584760e-02 -2.330302e-02
[13] -8.245587e-03 -2.287119e-02 -7.286213e-03 -2.053326e-02
[17] -2.603761e-02 -2.450931e-02 -2.338278e-03 -1.571486e-03
[21] -2.334969e-02 -2.391761e-02 -2.513697e-02 -3.779170e-03
[25] -2.389661e-02 -2.199168e-02 -2.474587e-02 -2.336897e-02
[29] -6.669706e-04 -4.560233e-04 -5.356959e-03 -2.574075e-02
[33] -2.690626e-02 -2.097137e-02  1.074432e-03  7.922431e-05
[37] -2.547026e-02 -3.432598e-03 -2.310148e-02 -4.331608e-03
[41] -2.208822e-02 -2.369825e-03 -2.412300e-02 -9.437667e-03
```

Figure 34: $\boldsymbol{\beta_{STORE*FuelPrice}}$ Coefficients For BoxCox Model Trained With Outliers Removed

```
 [1]  5.333031e-04  5.018449e-04  8.020110e-03  6.580173e-04
 2.353200e-04
 [6]  3.758994e-03 -1.128242e-04  3.046742e-04  7.389640e-03
 4.489104e-06
[11]  7.760171e-03  6.639155e-03  2.059331e-03  6.420278e-03
2.298115e-03
[16]  6.756223e-03  6.773303e-03  6.872150e-03  5.710591e-04
7.142174e-04
[21]  6.886655e-03  7.386696e-03  7.261221e-03  6.132169e-04
6.855346e-03
[26]  6.183272e-03  7.207885e-03  7.111585e-03 -5.412508e-04
2.932168e-04
[31]  2.337563e-03  8.120074e-03  7.953030e-03  5.482690e-03
-1.607758e-03
[36] -8.825048e-05  8.522958e-03  1.045170e-03  6.554815e-03
2.535812e-03
[41]  6.513528e-03  3.581827e-04  7.038210e-03  2.528400e-03
```

Figure 35: $\boldsymbol{\beta_{STORE*CPI}}$ Coefficients For BoxCox Model Trained With Outliers Removed

```
 [1]  0.0028246466  0.0033928231 -0.0005223857  0.0039608688
 0.0023515062
 [6]  0.0121688631  0.0018791504  0.0037487840 -0.0004753231
 0.0015493257
[11] -0.0001691588 -0.0016976674 -0.0005666298  0.0017190924
0.0036341490
[16] -0.0029620269 -0.0016088249 -0.0008628436 -0.0005239236
0.0058465363
[21]  0.0012307061  0.0018524281 -0.0017212208 -0.0013189252
-0.0012548754
[26] -0.0001508355 -0.0003672283  0.0021889279 -0.0033510299
0.0018154711
[31]  0.0040820153  0.0009211747  0.0002274952  0.0107847374
-0.0029700985
[36] -0.0012481806 -0.0013873318  0.0031882063 -0.0017786703
0.0056814130
[41] -0.0018728269 -0.0018043246 -0.0025115993 -0.0020281341
```

Figure 36: $\boldsymbol{\beta_{STORE*Unemployment}}$ Coefficients For BoxCox Model Trained With Outliers Removed

5.2 R Code

Following is the R Code used in computation of the model and the production of the figures

```r
library(olsrr)
library(mctest)
library(leaps)
library(car)
library(MASS)
library(Ecdat)
```

```
library(GGally)
library(lmtest)

dev.off()

options(max.print = 1000000)

walmar = read.csv("https://raw.githubusercontent.com/MHadd0/
    DataSets/main/TrainingSetWalmar.csv")
head(walmar)



walmar$Holiday_Flag <- as.logical(walmar$Holiday_Flag)
walmar$Store <- as.factor(walmar$Store)
walmar$Date <- as.Date(walmar$Date)

head(walmar)
str(walmar)

nocat <- walmar[-c(1,2,3,5)]
head(nocat)

# ADDITIVE MODEL
walmodel = lm(Weekly_Sales~Store+Holiday_Flag+Temperature+
    Fuel_Price+CPI+Unemployment,data=walmar)
summary(walmodel)

# Multicollinearity Testing
# pairs(~Weekly_Sales+Temperature+Fuel_Price+CPI+Unemployment,
    data=store1)

# coplot = ggpairs(nocat,
#                  lower = list(continuous = "smooth_loess",
#                                      combo =
#                                     "facethist",
#                                      discrete = "facetbar",
                                     na = "na"))

# print(coplot)
# VIF TEST, EXCLUDE CATEGORICAL
vifmodel = lm(Weekly_Sales~Temperature+Fuel_Price+CPI+Unemployment,
    data=walmar)
imcdiag(vifmodel,method="VIF") # all below 1.3


# FULL INTERACTION MODEL
walmodelintfull = lm(Weekly_Sales~(Store+Holiday_Flag+Temperature
    +Fuel_Price+CPI+Unemployment)^2,data=walmar)
summary(walmodelintfull)

# REDUCED INTERACTION MODEL
walmodelint1 = lm(Weekly_Sales~Holiday_Flag+Store+Temperature+
                        Fuel_Price+CPI+Unemployment+
```

```
                              Store*Holiday_Flag+Store*Temperature+
                              Store*Fuel_Price+
                              Store*CPI+Store*Unemployment+
                              Holiday_Flag*Temperature+
                                Holiday_Flag*Unemployment+
                                Temperature*Fuel_Price+
                                Temperature*CPI+ Fuel_Price*CPI,data=walmar)
summary(walmodelint1)

# FURTHER REDUCED INTER

walmodelint = lm(Weekly_Sales~Holiday_Flag+Store+Temperature+
                        Fuel_Price+CPI+Unemployment+
                        Store:Temperature+Store:Fuel_Price+
                        Store:CPI+Store:Unemployment+
                        Holiday_Flag:Temperature+
                        Temperature:Fuel_Price+Temperature:CPI+
                        Fuel_Price:CPI,data=walmar)
summary(walmodelint)

# Higher Order Model
highwalmar = lm(Weekly_Sales~Holiday_Flag+Store+Temperature+
                    Fuel_Price+CPI+I(CPI^2)+Unemployment
                      +Store*Temperature+Store*Fuel_Price+
                        Store*CPI+Store*Unemployment+
                        Holiday_Flag*Temperature+
                         Temperature*Fuel_Price+
                        Temperature*CPI+ Fuel_Price*CPI,data=walmar)

highwalmarclean = lm(Weekly_Sales~Holiday_Flag+Store+
    Temperature+Fuel_Price+CPI+I(CPI^2)+Unemployment
                +Store*Temperature+Store*Fuel_Price+
                  Store*CPI+Store*Unemployment+ Holiday_Flag*Temperature+
                  Temperature*Fuel_Price+
                  Temperature*CPI+ Fuel_Price*CPI,data=walmarclean)
summary(highwalmar)
plot(highwalmar,1)




#LINEARITY
ggplot(walmar, aes(x=Date, y=residuals(bcwalmar),color=Store))+
  geom_point()+
  scale_x_date(date_labels="%b-%d-%Y",date_breaks  ="6 month")+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))

# NORMALITY
shapiro.test(residuals(highwalmar))
plot(highwalmar,2)

ggplot(walmar, aes(x = residuals(highwalmar))) +
  geom_histogram(binwidth = 25000) +
  labs(title = "Histogram of Residuals",
```

```
      x = "residuals",
      y = "Count")


# VARIANCE TEST
bptest(highwalmar)

# VIF TEST
vifmodel = lm(Weekly_Sales~Temperature+Fuel_Price+CPI+
    Unemployment,data=walmar)
imcdiag(vifmodel, method="VIF")

# BOX-COX TEST
bc = boxcox(highwalmar, lambda=seq(-1,1)) # LAMBDA -0.232323
bestlambda=bc$x[which(bc$y==max(bc$y))]
bestlambda

# OUTLIERS
plot(highwalmar,4)
plot(highwalmar,5)

lev=hatvalues(highwalmar)
p = length(coef(highwalmar))
n = nrow(walmar)
outlier3p = lev[lev>(3*p/n)] # best 0.49
outlier_index <- names(outlier3p)
outlier_index <- as.numeric(outlier_index)
print(length(outlier_index))

plot(rownames(walmar),lev, main = "Leverage in KBI Dataset",
    xlab="observation",
     ylab = "Leverage Value")
abline(h = 3*p/n, lty = 1)

# CLEANED DATA
walmarclean = walmar[-outlier_index,]
nrow(walmar)
nrow(walmarclean)

# BOX-COX MODEL
bc=boxcox(highwalmar,lambda=seq(-1,1))
bestlambda=bc$x[which(bc$y==max(bc$y))]
bestlambda

boxwalmar = lm((((Weekly_Sales^-0.2727273)-1)/-0.2727273)~Holiday_Flag+
    Store+Temperature+Fuel_Price+CPI+
                    I(CPI^2)+Unemployment
                +Store*Temperature+Store*Fuel_Price+
                  Store*CPI+Store*Unemployment+ Holiday_Flag*Temperature
                  + Temperature*Fuel_Price+
                  Temperature*CPI+ Fuel_Price*CPI,data=walmar)

boxwalmarclean = lm((((Weekly_Sales^-0.2727273)-1)/-0.2727273)~Holiday_
    Flag+Store+Temperature+Fuel_Price+CPI+I(CPI^2)+
```

```
                    Unemployment+ Store*Temperature+Store*Fuel_Price+
                      Store*CPI+Store*Unemployment+ Holiday_Flag*
    Temperature+
                       Temperature*Fuel_Price+
                      Temperature*CPI+ Fuel_Price*CPI,data=walmarclean)
summary(boxwalmar)
summary(boxwalmarclean)


#LINEARITY
plot(boxwalmar,1)

plot(highwalmar,1)

# INDEPENDENCE
ggplot(walmarclean, aes(x=Date, y=residuals(boxwalmarclean),color=Store))+
  geom_point()+
  scale_x_date(date_labels="%b-%d-%Y",date_breaks  ="6 month")+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))



# NORMALITY
shapiro.test(residuals(boxwalmarclean))
plot(boxwalmarclean,2)



ggplot(walmarclean, aes(x = residuals(boxwalmarclean))) +
  geom_histogram(binwidth = 0.0003) +
  labs(title = "Histogram of Residuals",
       x = "residuals",
       y = "Count")


# VARIANCE TEST
bptest(boxwalmarclean)

#APPENDIX

walmartest = read.csv("DataScience/603proj/TestSetWalmar.csv")

walmartest$Weekly_Sales = ((walmartest$Weekly_Sales^-0.2727273)-1)/
    (-0.2727273)
walmartest$Store = as.factor(walmartest$Store)
walmartest$Holiday_Flag = as.logical(walmartest$Holiday_Flag)

str(walmartest)

head(walmartest)
walmartest
predictions = predict(highwalmarclean, walmartest)
predictionsclean = predict(boxwalmarclean, walmartest)

```

```
predictions

residualstest = walmartest$Weekly_Sales - predictions
residualstestclean = walmartest$Weekly_Sales - predictionsclean


plottingdftest <- data.frame(y = residualstest, x=1:length(predictions))
plottingdftestclean <- data.frame(y = residualstestclean,
    x=1:length(predictionsclean))

outlierstest<-ggplot(plottingdftest, aes(x = x, y = y))+
  geom_point()
nooutlierstest<-ggplot(plottingdftestclean, aes(x=x, y=y))+
  geom_point()+geom_hline(yintercept=0, color = "red")+
  labs(title = "Error in Prediction versus Fitted Values",
    x = "Fitted Values", y = "Prediction Error")
outlierstest
nooutlierstest

MSEpredclean = (1/length(predictions))*sum((walmartest$Weekly_Sales - predictionsclean)^2)
MSEpredoutlier = (1/length(predictions))*sum((walmartest$Weekly_Sales - predictions)^2)

RSEclean = sqrt(MSEpredclean)
RSEoutlier = sqrt(MSEpredoutlier)

RSEoutlier
RSEclean
summary(highwalmarclean)
plot(highwalmarclean, 1)
summary(boxwalmar)
summary(boxwalmarclean)
RSEmodel = 0.002542
unboxcox <- function(x){(-1*x*0.2727273 + 1)^(-1/0.2727273)}

unboxcox(RSEclean)-unboxcox(RSEmodel)

summary(walmodelintfull)

plot(boxwalmarclean)


predict(walmodel, walmartest)

boxwalmarcoeffs <- data.frame(coeffs = boxwalmarclean$coefficients)
boxwalmarcoeffs$coeffs[(55+(3*43)):(55+(4*43))]
boxwalmarcoeffs

boxcoxtrns <- function(x){
  (x^bestlambda - 1)/(bestlambda)
}
print(boxcoxtrns(c(min(walmarclean$Weekly_Sales), median(walmarclean$Weekly_Sales), max(walmarclean$Weekly_Sales))))
```

```

unboxcox(boxwalmarcoeffs$coeffs)
```

5.3 Python Code

This appendix contains the python code for data wrangling used

```python
import pandas as pd
import random as rd
import math

walmar = pd.read_csv("walmar.csv", parse_dates=["Date"], date_format="%d-%m-%Y")
test_set_dict = {}
training_set_dict = {}
tmp = walmar.where(walmar["Store"] == 1).dropna()
length = len(tmp)
perm = rd.sample(list(tmp.index), length)
trainingids = perm[0:math.floor(0.7*length)]
testids = perm[math.floor(0.7*length):]
test_set_dict.update(tmp.loc[testids])
training_set_dict.update(tmp.loc[trainingids])
test_set = pd.DataFrame(test_set_dict)
training_set = pd.DataFrame(training_set_dict)
for i in range(2, 46):
    test_set_dict = {}
    training_set_dict = {}
    tmp = walmar.where(walmar["Store"] == i).dropna()
    length = len(tmp)
    perm = rd.sample(list(tmp.index), length)
    trainingids = perm[0:math.floor(0.7*length)]
    testids = perm[math.floor(0.7*length):]
    test_set_dict.update(tmp.loc[testids])
    training_set_dict.update(tmp.loc[trainingids])
    test_set = pd.concat((test_set, pd.DataFrame(test_set_dict)))
    training_set = pd.concat((training_set, pd.DataFrame(training_set_dict)))

training_set.to_csv("TrainingSetWalmar.csv")
test_set.to_csv("TestSetWalmar.csv")
```

Bibliography

[1] Rashmi Jeswani. *Predicting Walmart Sales, Exploratory Data Analysis, and Walmart Sales Dashboard*. 2021. (accessed 2024-04-09).

[2] Bharat Kumar0925. *Walmart Sales Data*. URL: https://www.kaggle.com/datasets/bharatkumar0925/walmart-store-sales. Found on Kaggle (accessed 2024-04-09).

[3] Bureau of Labor Statistics. *Consumer Price Indexes Overview*. URL: https://www.bls.gov/cpi/overview.htm. (accessed 2024-04-09).

[4] Danika Lipman. “Presentation Question”. Thank you Danika.

[5] David A. Freedman Professor and David A. Freedman Professor. “A Note on Screening Regression Equations”. In: *The American Statistician* 37.2 (1983), pp. 152–155. DOI: 10.1080/00031305.1983.10482729. eprint: https://www.tandfonline.com/doi/pdf/10.1080/00031305.1983.10482729. URL: https://www.tandfonline.com/doi/abs/10.1080/00031305.1983.10482729.

[6] Tugba Sabanoglu. *World: leading retailers 2021, by retail revenue*. URL: https://www.statista.com/statistics/266595/leading-retailers-worldwide-based-on-revenue/. (accessed 2024-04-09).