Predicting Weekly Walmart Sales

Andrew Gooding

Jia Li Dong

Kathleen Callaghan

December 16th, 2018

Abstract

Walmart published historical sales data for 45 of its U.S. stores across various regions to the data science competition website, Kaggle. The project aim is to use the predictor variables Walmart provided from 2010 to 2013 to test their usefulness in a multiple linear regression model that predicts weekly sales and describes how individual predictors affect weekly sales. In addition, this project applies analysis of variance methods, variable exploration and variable selection methods, and tests for influential cases to determine how specific store types or specific departments affect average weekly sales. This paper further discusses the impact that each of the predictor variables has on the average weekly sales generated, including discussion of the impact of holidays, the impracticality of the markdown data, and the usefulness of interaction effects. Our goal is that this model proves useful to Walmart executives in future revenue forecasting, capital budgeting, and resource allocation.

1 Introduction

Walmart is an American multinational retail company worth over $500 billion dollars that operates a worldwide chain of discount department and grocery stores.[1] According to the 2018 Fortune Global 500 list, Walmart is not only the world's largest company by revenue but is also the world's largest employer with 2.3 million employees. As of 2018, Walmart was the U.S.'s largest grocery retailer with approximately 64% of its sales coming from its U.S. operations.[2]

In 2014, Walmart published historical sales data for 45 of its U.S. stores across various regions to the data science competition website, Kaggle.[3] Walmart challenged participants to create a model that could predict weekly sales based on several store features (e.g. type of store, size of store, department, markdowns, etc.) and macroeconomic variables (e.g. Consumer Price Index, the price of gas, the unemployment rate, etc.). The primary motivation for this project is to use simple and multiple linear regression methods to generate one or several models to describe weekly sales across various store types and departments. The secondary motivation for this project is to build a linear regression model that can predict future sales across various store and department combinations using the 2010-2013 data set. Additionally, selected holiday markdown events included in the dataset are known to affect sales but may be challenging to predict the departments affected and the extent of the impact. Specifically, we seek to test the interaction between holiday weeks and markdowns to gauge their impact on sales.

2 Methods & Materials

2.1 Data Summary

Our team was provided three classes of variables: a time-series of store-level features, static reference data by store, and a historical time-series of department-level sales within stores.

The store-level features that were provided were end-of-week observations for 45 Walmart stores over the period from Feb-2010 to July-2013:

[1] "Form 8K-Walmart Inc'. U.S. Securities and Exchange Commission. February 1, 2018.
[2] "Walmart 2018 Annual Report" (PDF) stock.walmart.com. p. 7. November 18, 2018.
[3] www.kaggle.com/c/walmart-recruiting-store-sales-forecasting. November 18, 2018.

Temperature: the temperature in degrees Fahrenheit, as measured at the store location

Fuel_Price: the average price in USD of one gallon of gasoline, as measured at the store location

MarkDown1-MarkDown5: an anonymized variable provided by Walmart which relates to discount promotions that the store was running

CPI: the US Consumer Price Index, an index which measures the price of a defined basket of goods used to measure inflation

Unemployment: the US Unemployment rate as measured by the Federal Reserve

IsHoliday: a boolean variable denoting whether or not the week contained a holiday, as defined by Walmart

The static reference data we were provided for each store were:

Type: a categorical variable denoting the type of Walmart each store is, taking values "A","B", or "C"

Size: the square footage of each store

Finally, we were provided with a time-series of Weekly_Sales by Store and Department, for the period Feb-2010 to July-2013.

2.2 Variable Exploration

A few issues with the data emerge immediately. The first is that different stores have a different number and combination of departments such that not every store has the same departments. Furthermore, the average weekly sales differs greatly across departments, so a model to predict department-level sales would need to account for this variation (Figures 3 & 4).

Second, the MarkDown data presents significant issues. It is severely sparse; roughly 50% of the MarkDown values are missing without explanation. This makes it difficult to build precise historical models, as our data is limited, and it also makes future predictive systems potentially less robust, as the presence of MarkDown data may be unreliable. Furthermore, Walmart anonymized the MarkDown data because it is proprietary. So, by design, we have no intuition as to what any of the MarkDown values actually mean. This raises an interesting question–what

is the value of a model that contains variables, which cannot be interpreted by the modeller? Should the modeller exclude these variables ex-ante? How would the modeller interpret the coefficients? These are legitimate questions, and in normal circumstances we do not feel that it is worthwhile to have unexplainable variables in a final model. In this case, however, we presume that Walmart executives have the ability to interpret these anonymized variables. For this reason, we have retained the MarkDown variables in our model exploration.

2.3 Enriching the Data

While the original feature set is relatively limited, there are ways in which simple data enrichments make sense.

First, the data is seasonal, so we want to have a method to include date as a predictor variable. WeeklySales are likely sensitive to holidays and potentially sensitive to specific holidays. One candidate for variable enhancement is the IsHoliday field. In the original dataset, Walmart only considers 4 weeks to be holiday weeks: the weeks that contain the Super Bowl, Labor Day, Thanksgiving, and Christmas. However, it is possible that other holidays (e.g. Easter, July 4th) or weeks preceding holidays (e.g. the weeks preceding Christmas) affect weekly sales. To this point, we derive two additional predictors:

Week: a categorical variable denoting the calendar week of the year (e.g. the week containing January 1st is labeled "Week 1")

IsHolidayFixed: a boolean variable indicating whether the week contained a holiday based on the extended list of holidays

Extending the feature set using these methods allows us to expand from 7 predictive variables to 9.

2.4 Variable Selection

Because we have at most nine predictor variables (*IsHoliday*, *Temperature*, *Fuel_Price*, *CPI*, *Unemployment*, *Type*, *Size*, *Week* and *IsHolidayFixed*), we use a best subsets procedure in initial variable selection and screening to reduce the number of models under consideration. We

prefer this method over a step-wise procedure because with only nine explanatory variables, the best subsets procedure allows us to compare every possible combination within each subset. We choose to narrow the list of potential models to the five best models within each subset using both Mallow's Cp and Adjusted R-squared as the model selection criteria. Adjusted R-squared describes the proportion of variation in the weekly sales explained by the predictors in the model, adjusted for the number of predictors (similar to choosing a model with the smallest mean squared error). Mallow's Cp is a measure of error in the best subset model relative to the error incorporating all variables. Adequate models using Mallow's Cp are those for which Cp is roughly equal to the number of predictors in the model and/or Cp is minimized. Using these statistics, we are able to compare the five best models in each subset and narrow the list of potential models to the top four for further exploration.

While the best subsets procedure is an effective tool for narrowing the number of potential regression models, it is important to assess that each model meets the following model assumptions:

1. Normal distribution of error terms
2. Homoscedasticity of error terms

If the model does not meet either or both assumptions, remediation is required. For models failing to meet both the assumption of normally distributed error terms and constant error variance (i.e. homoscedasticity), variable transformation is required based on the shape of the error distribution. If the error terms meet the assumption of normality but are found to be non-constant (i.e. heteroscedastic only), a weighted least squares model is appropriate using the following weight:

$$w_i = \frac{1}{\sigma_i^2}$$

While the best subsets method of variable selection helps to narrow down the set of potential regression models, further analysis is required to refine the models. To assess the significance of each predictor and whether it should be kept in the model, we perform a partial F test on each

of the variable coefficients to test the following hypotheses at a significance level of $\alpha = .05$:

$$H_0 : \beta_i = 0 \; ; \quad H_a : \beta_i \neq 0$$

We require $F^*{\sim}F(1, 6370 - p) =$, where p is the number of predictors in the model plus 1, for the null hypothesis to hold; otherwise, we reject the null hypothesis. For large values of F^* (p-value $< .05$) we conclude that the predictor variable associated with β_i has a significant impact on weekly sales and should be retained in the model.

We further refine our model selection by considering various possible interaction effects, specifically looking at possible interaction among the macroeconomic variables (*Unemployment*, *CPI*, and *Fuel_Price*). Because our models have up to nine predictors, we potentially have up to 36 pairwise interaction terms; therefore, we focus tests for interaction effects on the a-priori hypothesis that there may be interaction between *Unemployment*, *CPI*, and *Fuel_Price*, the macroeconomic variables. Because we presume that economic variables may be highly correlated with one another, these variables have a high potential for interaction effects. We test the following interaction terms: X_A= $Unemployment * CPI$, $X_B = Unemployment * Fuel\_Price$, and $X_C = Fuel\_Price * CPI$ using extra sums of squares to test the following hypotheses:

$$H_0 : \beta_A = \beta_B = \beta_C = 0 \; ; \quad H_a : \text{not all } \beta_i\text{'s are equal to } 0$$

For a level of significance of $\alpha = .05$, we require $F(.95; 3, 6305) = 2.1$. We conduct a partial F test on each interaction coefficient to assess the significance of each individual interaction term at significance level $\alpha = .05$ for the following hypotheses:

$$H_0 : \beta_A = 0 \; ; \quad H_a : \beta_A \neq 0$$
$$H_0 : \beta_B = 0 \; ; \quad H_a : \beta_B \neq 0$$
$$H_0 : \beta_C = 0 \; ; \quad H_a : \beta_C \neq 0$$

Again, we require $F^* \sim F(1, n - p)$ for the null hypothesis to hold; otherwise we reject the null hypothesis. For large values of F^* (p-value $< .05$), we conclude that the interaction term associ-

ated with β_i has a significant impact on weekly sales and should be retained in the model.

2.5 Influential Cases

First, we consider the most prominent cases that are easily detected as outliers in the diagnostic plots, such as the normal probability plots and the fitted versus residual plots. Then, we can test each of the cases to determine if it should be classified as outlying, and whether it has a significant or undue influence on the regression model.

Because the aggregated dataset contains 6435 observations overall, we consider this a large dataset, and thus, the outlying observation test is conducted based on this condition. We use DFBETA to determine outlying observations, for which observations having an absolute value of DFBETA greater than $2 \div \sqrt{6435}$ are classified as outliers. In addition, we can test whether the cases are outliers based on DFFITS, for which observations with an absolute value of DFFITS greater than $2\sqrt{p/n}$, where p is the number of parameters and n is the number of total observations, are classified as outliers. We also use Cook's Distance, where observations having a Cook's Distance that exceeds the F distribution of $1/n$ with the degrees of freedom p and $n - p$ are considered to be outlying observations. Additionally, we conduct a test using the Bonferroni critical value to determine outlying observations with a 10% family significance interval. Observations exceeding the value of the t distribution at 0.99999222999223 (or $1 - \frac{0.1}{2(6435)}$) with degrees of freedom $n - p - 1$ are classified as outlying observations.

Finally, we compare all tests to determine which observations can be classified as outlying. If it is found that the identified cases significantly influence the regression model, they should be removed from the model.

3 Results

3.1 Model Selection

Using the best subsets procedure from the leaps package to aid in initial variable selection, we narrow our models down to the four best models (based on adjusted R-squared and Mallow's

Cp) for further analysis. The regression equations are:

1. log(Weekly_Sales) ~ Temperature + CPI + Unemployment + Type + Size + Week + IsHolidayFixed (adjR2=.6806188; cp=11.532115)
2. log(Weekly_Sales) ~ Temperature + Fuel_Price + CPI + Unemployment + Type + Size + Week + IsHolidayFixed (adjR2=.6808181; cp=8.519027)
3. log(Weekly_Sales) ~ IsHoliday + Temperature + CPI + Unemployment + Type + Size + Week + IsHolidayFixed (adjR2=.6805914; cp=13.082784)
4. log(Weekly_Sales) ~ IsHoliday + Temperature + Fuel_Price + CPI + Unemployment + Type + Size + Week + IsHolidayFixed (adjR2=.6807942; cp=10)

We find that these four models have both the highest adjusted R-squared and the lowest Cp statistics. Models 1 through 3 have Cp statistics near their respective values for p, the number of parameters in the model. For Model 4, which represents the full model, we only use adjusted R-squared to evaluate its effectiveness because Cp will always equal p for the full model (Tables 1 & 2).

To assess whether the models meet the assumptions for further analysis, we plot the fitted values versus the residuals and the normal probability plot for each model. Figure 1 shows the residual plots for Model 2, which are nearly identical to the residual plots for the other three models. Based on the fitted vs. residual plots, all four models fail to meet the assumption of homoscedasticity. There is a clear megaphone shape to the plotted residuals, which indicates non-constant variance. Furthermore, we cannot safely assume that the residuals are normally distributed because they follow a clear curve. Because our models fail to meet both assumptions, a transformation is necessary.

The trend of the residuals curves up in all four cases, which indicates a log transformation on the response variable, $Weekly_Sales$, may be a good transformation. We find that doing so greatly improves the distribution of the residuals in each case and conclude that the log transformed models meet our required assumptions (Figure 2).

After performing partial F tests on each of the variables, testing for interaction effects of the macroeconomic variables (CPI, $Fuel\_Price$ and $Unemployment$) and determining multicollinear-

ity using variable inflation factors, we conclude that the following model most effectively describes weekly store sales:

$$\begin{aligned}\log(Weekly\_Sales) =& 12.88 + .0006261 * Temperature - .001217 * CPI \\ & -.009399 * Unemployment + .00000806 * Size \\ & +..08793 * IsHolidayFixedTRUE\end{aligned}$$

The adjusted R-squared is .7598 and the mean squared error is .08 (Table 3).

3.2 Influential Cases

The cases identified as potential outlying observations by the diagnostic plots for the five different models are observations 2080, 4420, 2093, 4433, and 1900 (Figure 7).

Our findings indicate that Model 1 identifies 5 outlying cases (observations 2080, 4420, 4433, 2093, and 1900), where observation 2080 is classified by two tests, DFFITS and Bonferroni critical value test. Observations 4420 and 1900 are classified by the DFFITS test as outlying. observations 4433 and 2093 are classified as outlying by two tests, DFFITS and DFBETA.Model 2 and Model 3 have the least number of outlying cases as identified by the diagnostic plots, and both observation 4433 and observation 2093 are identified by DFBETA as outliers. Model 4 also identifies all five points as outlying, where observation 2080 is classified as outlying by three tests - DFFITS, Bonferroni critical value test, and DFBETA. Observations 4433 and 2093 are classified as outlying by two tests, DFFITS and DFBETA. Observation 4420 is classified as outlying by two tests, DFFITS and Bonferroni critical value test. Observation 1900 is identified as having hidden extrapolations and classified by DFFITS as outlying. In all four models, no observations are classified as outlying using Cook's Distance.(Table 9)

The relative difference between the fitted values in the full model containing all 6435 cases and the fitted values in the corresponding model omitting the outlying cases showed that all 5 outlying cases identified do not exercise significant influence. The percentage value is less than 1/6435 (approximately 0.01554%) for all four models. Furthermore, the average absolute percentage difference for all five cases ranges from 0.002% to 0.004%, which is less than 0.01554%.

Thus, we have not omitted any outlying cases in the final model.

4 Discussion of Results

4.1 Model Summary

The results of our analysis and findings indicate that the majority of the variation in weekly store sales can be attributed to the size of the store (in square feet) (Table 5). We can see this by looking at the partial sums of squares. The macroeconomic variables, while significant, play a more marginal role; however, since the goal is to find a model that best predicts weekly sales, we believe the macroeconomic variables should remain in the model.

Additionally, a log transformation of *Weekly_Sales*, the response variable, corrects the failure of the model to meet the assumptions of normally distributed error terms and constant variation of error terms (Figure 2). The original distribution of *weekly_sales* is skewed with a tail of very high values for Weekly Sales (Figure 3), which can not be accounted for with our set of predictors. Thus, the log transformation changes how we interpret the coefficients of the final model. To determine the effect of a given predictor on *Weekly_Sales*, we must use $e^{\hat{\beta}_i}$. With all other predictors held constant, the mean change in *Weekly_Sales* is given by $e^{\hat{\beta}_i}$ for every one unit increase in the predictor value.

4.2 Holiday Data

Comparing the results of our regression models, we conclude that the categorical variables *Week*, *IsHoliday* and *IsHolidayFixed* are interchangeable. All three variables capture the same amount of variation in weekly sales. If *Week* is not included in the model, *IsHolidayFixed* is significant, however, if *Week* is included in the model, *IsHolidayFixed* is not significant. Ultimately, we keep *IsHolidayFixed* in the model over *Week* even though it accounted for slightly less variation in weekly sales. We believe that having a variable that speaks to whether a week is a holiday will be helpful for Walmart executives when making business decisions regarding the effects of holidays on weekly sales.

4.2 The Impracticality of MarkDown Data

When using the entire data set, we found that none of the MarkDown variables had any significance. When used in simple regressions, none of the MarkDown variables had an R^2 of greater than 0.05. In multiple regressions, they do not provide additional explanatory power. Additionally, the MarkDown data is extremely sparse, which reduces the explanatory precision even further.

While we conclude that MarkDown data is not useful for modelling the entire data set, we find that the MarkDown values are potentially significant in predicting Weekly_Sales during holidays. In fact, a customized model for only the holiday weeks is able to obtain an R_a^2 of 0.98 using MarkDown data (Table 7 & Figure 5)

Despite this fact, we caution against using this model for several reasons. As discussed, the *MarkDown* data is incomplete, which reduces the number of observations available to the model. For this model (which uses $MarkDown2$, $MarkDown4$, and $MarkDown5$) the MarkDown data reduces the observations from 6435 to 1524. Restricting the data to only holiday weeks reduces it further to 148 observations, roughly 3 per store. Despite the promising R_a^2, we conlcude that the results of this model are potentially spurious due to the lack of data. Further, this model makes use of a lagged value: last week's weekly sales. This short time-frame makes any practical implementation based on the model's results extremely difficult. Store managers may not have the time to re-allocate employee schedules on such short notice; it is unlikely that inventory supply-chains could be meaningfully reworked in time.

4.3 Interaction Effects

We hypothesized that the macroeconomic variables may have interaction effects in the model; however, the results are mixed. While the general test for significance of any interaction between macroeconomic variables is determined to not be significant (p-value=.99), we find that $CPI * Unemployment$ is significant (p-value=0).

Ultimately, however, $CPI * Unemployment$ is omitted from the final model as it leads to high VIF values, which indicate serious multicollinearity between CPI, $Unemployment$, and the in-

teraction term that affect the predictive ability of the regression model (Table 8).

4.4 Aggregating to Store Level

The intial objective was to predict department-level sales for each store. We attempted several approaches, none of which were able to satisfy the assumptions of linear regression–i.e. no model for department-level sales produced normally distributed residuals. The majority of the departures from normality were located in the tails of the distribution. While our attempts to model department-level sales produced accurate predictions, with normally-distributed residuals for the middle portion of the distribution, the models fail to predict *Weekly_Sales* values that are significant departures from mean values. We found some indications that these inaccurate predictions are concentrated around holidays, yet adding holiday-related predictor variables did not improve the normality of the residuals. In some sense, it is these cases–the cases where some combination of variables causes a radical departure from expected sales–that are the most valuable to predict. In addition, the majority of our predictor variables are at the store-level. This difference in granularity made it difficult to infer department-level effects–i.e. using store-level variables to predict store-level sales was a more natural fit (Figure 4).

To retain the practical value of our model and to satisfy the assumptions of a linear regression approach, we chose to aggregate department-level sales to store-level sales. Rather than attempt to predict the sales for each department within each store, we summarized a store's total sales (the sum of the store's department-level sales) by date. Therefore, store-level sales is our response variable. This aggregation reduces department-level noise and aligns the granularity of our predictors with our response. The result is the normalization of the residuals of our predictive models (Figure 3).

4.5 Model Applications

A model, like ours, which predicts weekly sales for a given store, has many practical applications. If labor requirements increase with sales, our model may be used to set employee work schedules or to decide staffing levels during holiday periods. The model's use of macroeconomic

variables may be useful for revenue forecasting and capital budgeting, allowing the company to optimize investment levels for different levels of predicted sales. An accurate prediction of a store's sales could be used for inventory planning and supply-chain management. It could also be used for expansion planning. The macroeconomic variables could be used to help decide where new stores should be located, and the store-level variables could be used in the design of the new stores themselves. Our model confirmed the intuition that larger stores tend to have higher sales on average, which could be used in a more precise evaluation of new store construction costs. Having a quantitative approximation of the value of store size makes it possible to assess the potential effects of increasing or decreasing square footage on sales totals.

5 Conclusion

In conclusion, the best model for predicting weekly store sales is

$$\begin{aligned}\log(Weekly\_Sales) =& 12.88 + .0006261 * Temperature - .001217 * CPI \\ & -.009399 * Unemployment + .00000806 * Size \\ & +..08793 * IsHolidayFixedTRUE\end{aligned}$$

where the base-line dollar amount in sales if all predictors equal zero is \$12.88. We conclude that the most important predictor of a store's total weekly sales is the size of the store. For each square foot increase in a store's size, the average weekly sales increases by $e^{.00000806} = 1.00$ dollar. Because the size of store has so much variation, its impact is significant. For example, if all other predictors are held constant, the difference in average weekly sales between a 40,000sf store and a 200,000sf store is equal to $200,000(\$1.00) - 40,000(\$1.00) = \$160,000$.

Additionally, we conclude that MarkDown data proved to be insignificant. We failed to see any significant effect in department-level weekly sales due to changes in MarkDown values. This coulud be due to the sparsity of MarkDown data provided and the lack of information about what the MarkDown data represents.

Furthermore, it is interesting to note that while holiday weeks are significant in predicting

weekly sales, the effect is much smaller than would be expected. If all other predictors are held constant, the average increase in weekly sales due to a holiday week is only $e^{.09} = 1.09$ dollars. This could be due to increased markdowns in effect during holiday weeks that lower prices on individual products so that even if the quantity of merchandise sold during holiday weeks increases, the overall effect on weekly sales is minimal.

Finally, we conclude that temperature, unemployment rate, and CPI are all significant in predicting weekly sales. The average increase in weekly sales increases by \$1.00 for every degree increase in temperature so that if all other predictors are held constant the difference in sales on a 100 degree day versus a 0 degree day is \$100. As expected, weekly sales decreases with an increase in either consumer price index or unemployment when all other predictors are held constant. It is worth noting that because the variation in the values of the macroeconomic variables is very small, they affect little change on the actual dollar amount in sales.

Appendix

Table 1: Best subset models based on Mallow's Cp

IsHoliday	Temp	Fuel Price	CPI	Unemp.	Type	Size	Week	IsHolidayFixed	Number	Cp
0	0	0	1	1	1	1	1	1	7	42.639
0	1	0	1	1	1	1	0	1	7	48.710
0	1	0	1	1	1	1	1	1	8	11.532
0	1	1	1	1	1	1	0	1	8	43.351
1	0	0	1	1	1	1	1	1	8	43.452
0	0	1	1	1	1	1	1	1	8	43.530
0	1	1	1	1	1	1	1	1	9	8.519
1	1	0	1	1	1	1	1	1	9	13.083
1	0	1	1	1	1	1	1	1	9	44.256
1	1	1	1	1	1	1	0	1	9	45.251
1	1	1	1	1	1	1	1	1	10	10.000

Table 2: Best subset models based on Adjusted R-squared

IsHoliday	Temp	Fuel Price	CPI	Unemp.	Type	Size	Week	IsHolidayFixed	Number	Adj.R2
0	1	0	1	1	1	1	1	1	8	0.681
0	1	1	1	1	1	1	1	1	9	0.681
1	1	0	1	1	1	1	1	1	9	0.681
1	1	1	1	1	1	1	1	1	10	0.681

Table 3: Summary of selected best model

Term	Coef	Std. Error	t-stat	P-Value
(Intercept)	12.876	0.029	441.560	0.000
Temperature	0.001	0.000	3.020	0.003
CPI	-0.001	0.000	-12.373	0.000
Unemployment	-0.009	0.002	-4.590	0.000
Size	0.000	0.000	140.397	0.000
IsHolidayFixedTRUE	0.088	0.009	10.213	0.000

Figure 1: Residual plots of model before log transformation of response variable

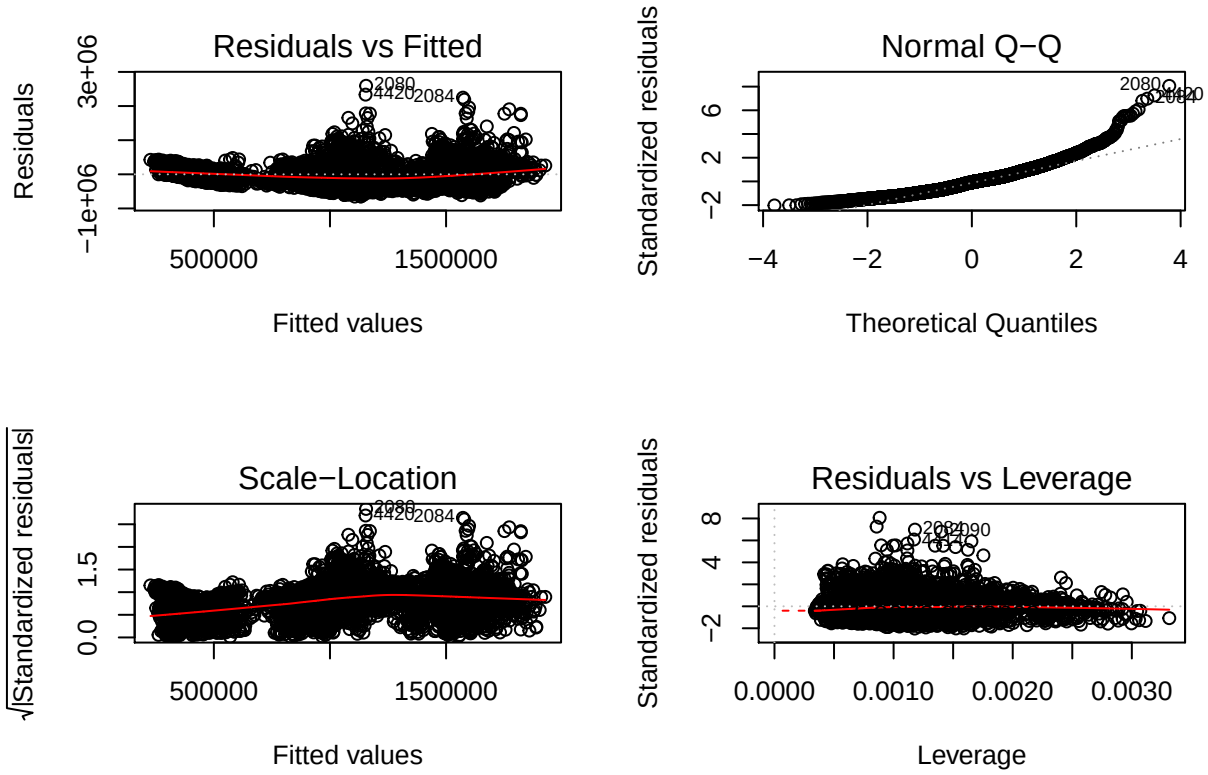


Figure 2: Residual plots of model after log transformation of response variable

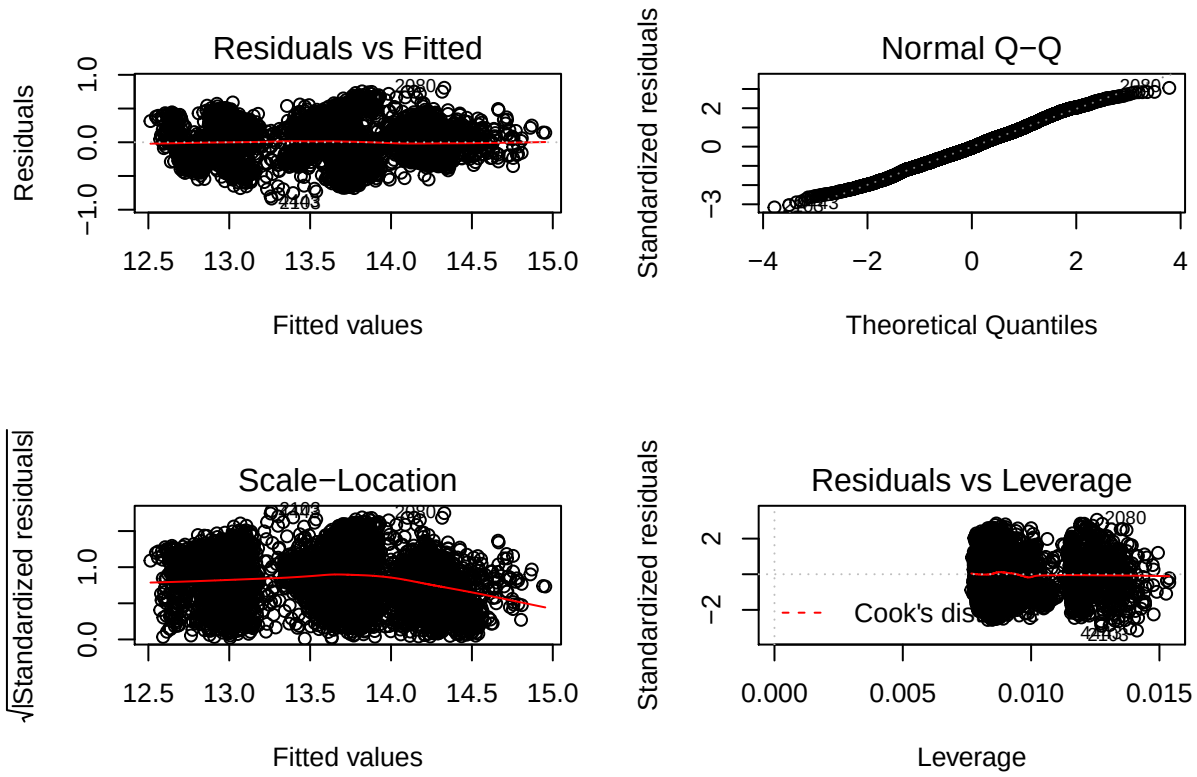


Table 4: Comparison of model statistics for top four models

Model	Adj_R2	MSE	AIC	PRESS
Model 1	0.7958	0.07	1270.575	459.0427
Model 2	0.7958	0.07	1272.382	459.1737
Model 3	0.7958	0.07	1270.575	459.0427
Model 4	0.7958	0.07	1272.382	459.1737

Table 5: Analysis of Variance table for best model

Df	Sum Sq	Mean Sq	F value	Pr(>F)
1	19.061	19.061	229.236	0
1	8.586	8.586	103.258	0
1	20.693	20.693	248.863	0
1	1635.219	1635.219	19665.974	0
1	8.673	8.673	104.309	0
6429	534.569	0.083	NA	NA

Table 6: Collinearity of MarkDown data

rowname	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5
MarkDown1	NA	0.02	-0.09	0.81	0.09
MarkDown2	0.02	NA	-0.05	-0.05	-0.01
MarkDown3	-0.09	-0.05	NA	-0.06	-0.02
MarkDown4	0.81	-0.05	-0.06	NA	0.04
MarkDown5	0.09	-0.01	-0.02	0.04	NA

Figure 3: Distribution of store-level weekly sales

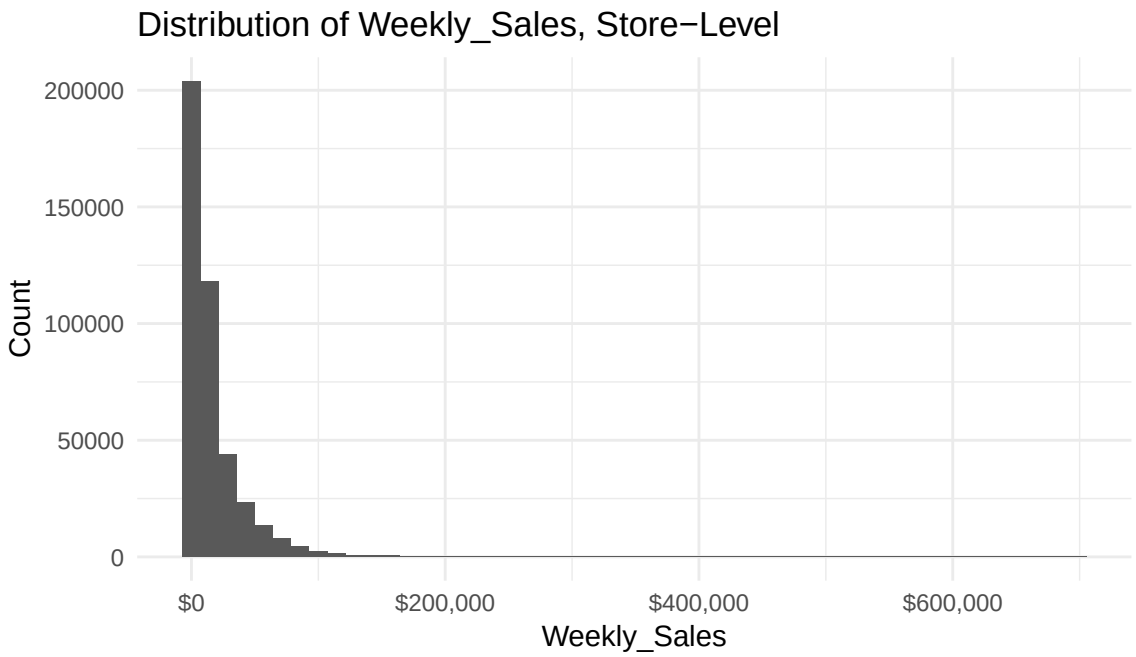


Figure 4: Distribution of department-level weekly sales

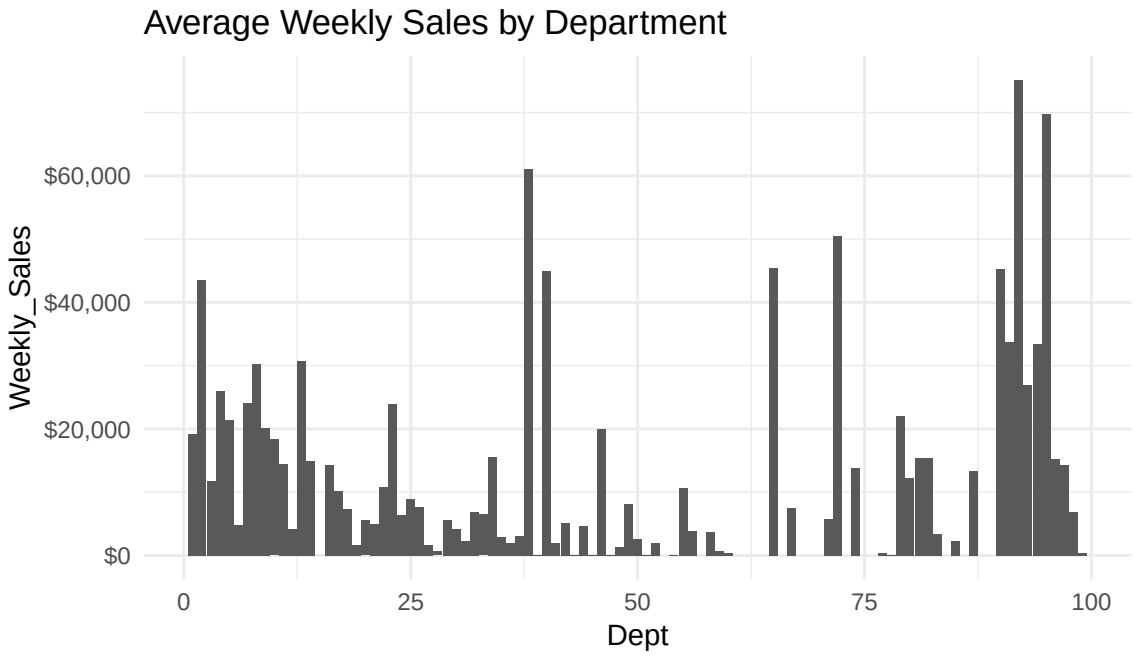


Table 7: Summary of model of holiday sales

Term	Coef	Std. Error	t-stat	P-Value
(Intercept)	0.487	0.244	1.996	0.048
MarkDown2	0.000	0.000	-4.018	0.000
MarkDown4	0.000	0.000	0.642	0.522
MarkDown5	0.000	0.000	1.583	0.116
HolidayLaborDay	0.348	0.028	12.298	0.000
HolidaySuperBowl	0.402	0.037	10.748	0.000
HolidayThanksgiving	0.762	0.028	27.105	0.000
log(Weekly_Sales_Week_Prev1)	0.938	0.019	50.661	0.000

Figure 5: Normal probability plot for model of holiday sales

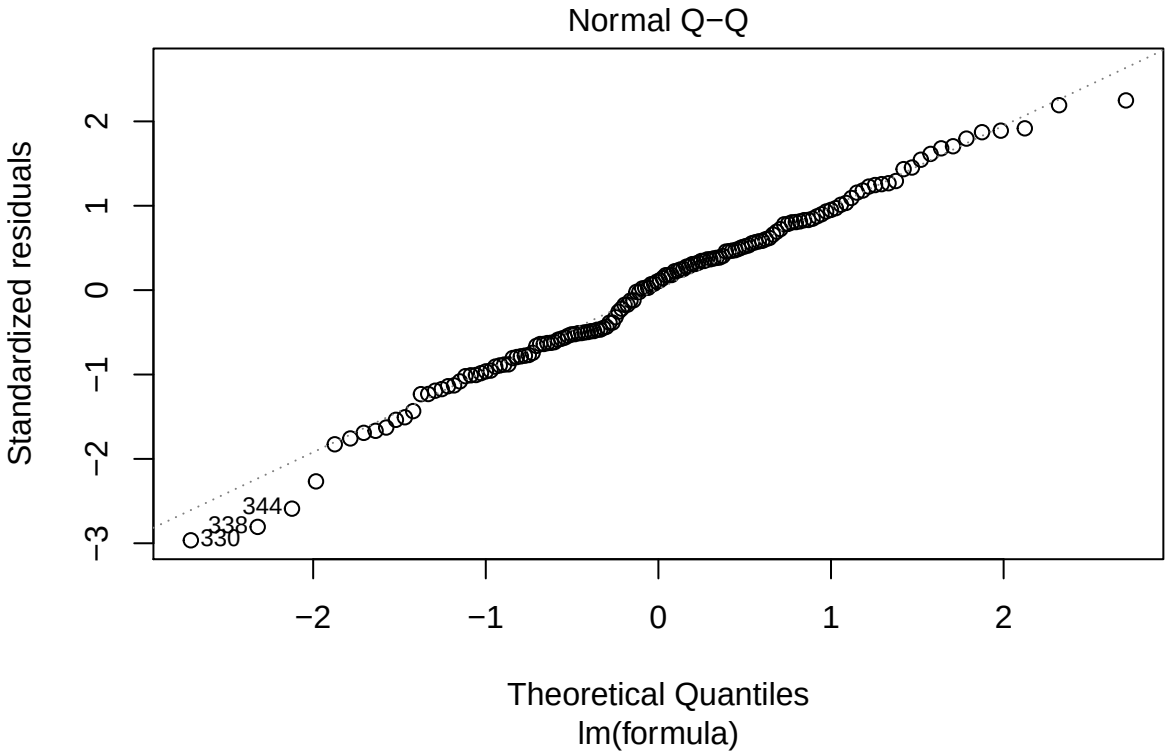


Figure 6: Number of departments per store

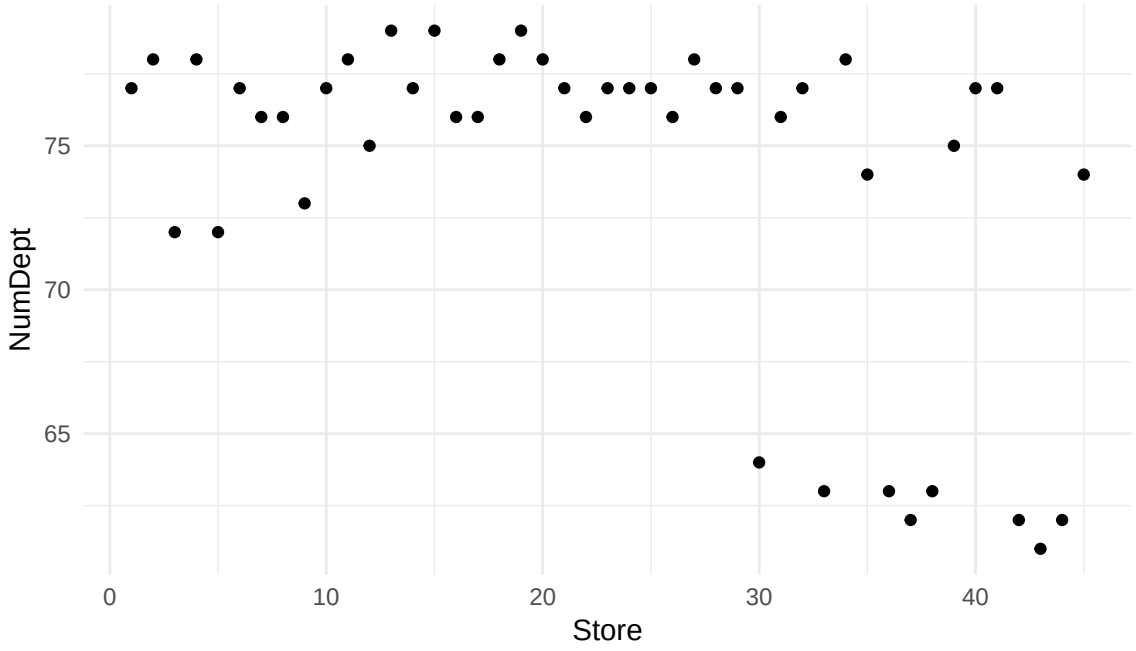
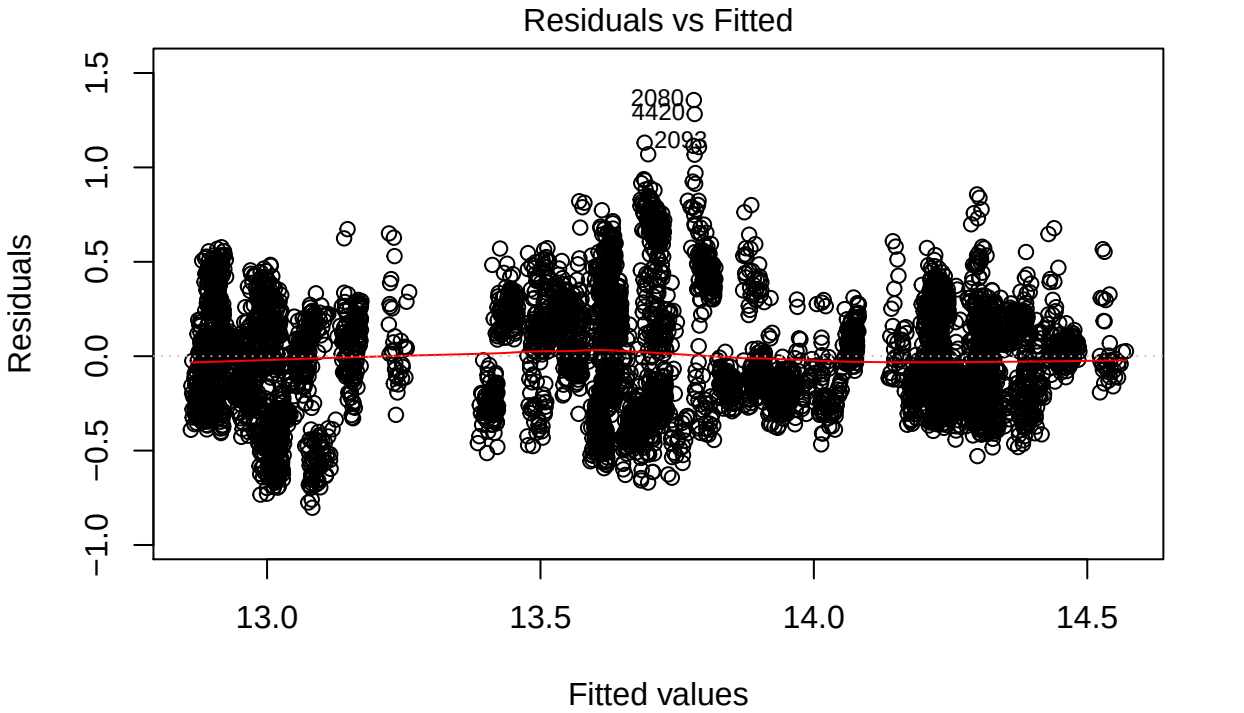
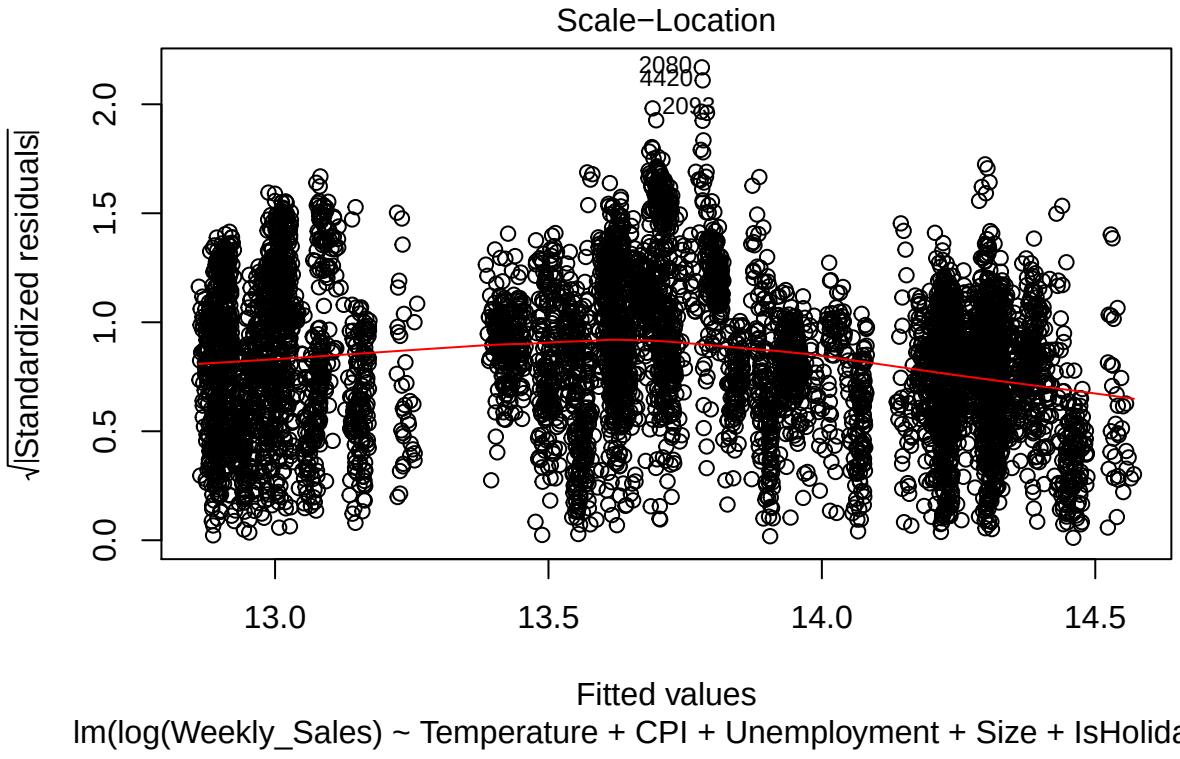


Figure 7: Outlier Plots




Scale−Location
√|Standardized residuals|
2080
44200
2093
0.0
0.5
1.0
1.5
2.0
13.0
13.5
14.0
14.5
Fitted values
lm(log(Weekly_Sales) ~ Temperature + CPI + Unemployment + Size + IsHolidayF ..

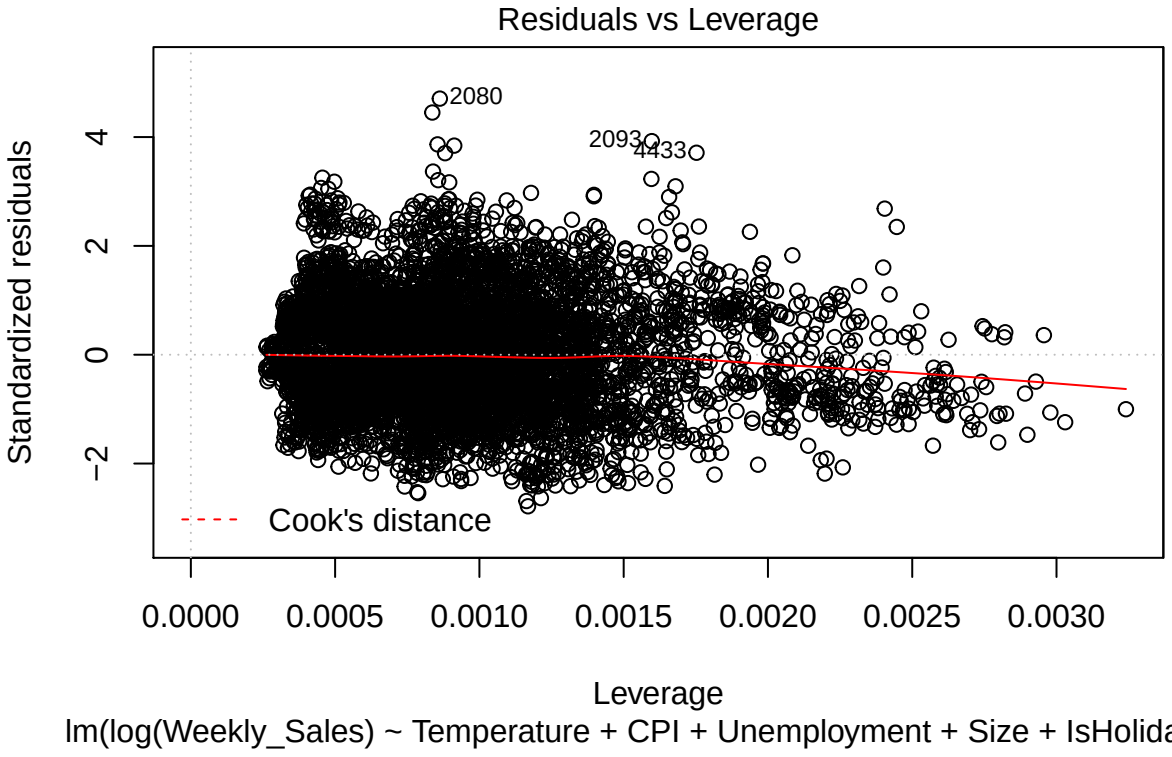

Residuals vs Leverage
Standardized residuals
2080
2093
4433
Cook's distance
−2
0
2
4
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
Leverage
lm(log(Weekly_Sales) ~ Temperature + CPI + Unemployment + Size + IsHolidayF ..

Figure 8: Variable Inflation Factors

	VIF
Temperature	1.132
CPI	1.159
Unemployment	1.142
Size	1.016
IsHolidayFixed	1.060

Table 8: Outlier Model 1

Outlying Observation Table Analysis for Model 1 (an X indicates failing the test)

Observation	DFBETA	DFFITS	Cook's Distance	Bonferroni Critical Value	Direct Leverage Calculation (for hidden extrapolation)
2080		X		X	
4420		X			
4433	X	X			
2093	X	X			
1900		X			

Table 9: Outlier Model 2

Outlying Observation Table Analysis for Model 2 (an X indicates failing the test)

Observation	DFBETA	DFFITS	Cook's Distance	Bonferroni Critical Value	Direct Leverage Calculation (for hidden extrapolation)
2080					
4420					
4433	X				
2093	X				
1900					

Table 10: Outlier Model 3

Outlying Observation Table Analysis for Model 3 (an X indicates failing the test)

Observation	DFBETA	DFFITS	Cook's Distance	Bonferroni Critical Value	Direct Leverage Calculation (for hidden extrapolation)
2080					
4420					
4433	X				
2093	X				
1900					

Table 11: Outlier Model 4

Outlying Observation Table Analysis for Model 4 (an X indicates failing the test)

Observation	DFBETA	DFFITS	Cook's Distance	Bonferroni Critical Value	Direct Leverage Calculation (for hidden extrapolation)
2080	X	X		X	
4420		X		X	
4433	X	X			
2093	X	X			
1900		X			X

Table 12: Outlier Best Model

Outlying Observation Table Analysis for Model 4 (an X indicates failing the test)

Observation	DFBETA	DFFITS	Cook's Distance	Bonferroni Critical Value	Direct Leverage Calculation (for hidden extrapolation)
2080		X		X	
4420	X	X		X	
4433	X	X		X	
2093	X	X		X	
1900		X		X	

Figure 9: VIF VALUES FOR INTERACTION MODEL

VIF- Model 5

Temperature	1.20419607875371
Fuel_Price	1.1095009413792
CPI	32.2144409812764
Unemployment	26.1778330047469
Type	2.73533156607444
Size	2.67237736748267
Week	1.07273666400897
CPI:Unemployment	40.2799086458315

Table 13: OUTLIER INFLUENCE RESULTS

Observation Number	Model Number	Percentage of Influence
2080	1	0.00314317450286986%
2080	2	0.0028313547231717%
2080	3	0.0033613122845678%
2080	4	0.0022407017833724 8%
2080	Best	0.0028348275977778 4%
4420	1	0.0029437547906130 9%
4420	2	0.0026435307591651 3%
4420	3	0.0031640230354883 2%
4420	4	0.0021269758481087 6%
4420	Best	0.0026145311808485 2%
4433	1	0.0032750289548950 3%
4433	2	0.0031318682431265 4%
4433	3	0.0034551709096619 3%
4433	4	0.0029487767271668 5%
4433	Best	0.003126264540517 5%
2093	1	0.003422691490910 2%
2093	2	0.0032181526770730 2%
2093	3	0.0036237117269879 5%
2093	4	0.0030829073238516 7%
2093	Best	0.003197795439588 3%
1900	1	0.0024298490651678 8%
1900	2	0.0023336040706455 5%
1900	3	0.0027079627858882 1%
1900	4	0.0026055367509707 2%
1900	Best	0.0023289500501651 1%

Bibliography

1. "Form 8K-Walmart Inc'. U.S. Securities and Exchange Commission. February 1, 2018.

2. "Walmart 2018 Annual Report" (PDF) stock.walmart.com. p. 7. November 18, 2018.

3. "Walmart Kaggle Competition" www.kaggle.com/c/walmart-recruiting-store-sales-forecasting. November 18, 2018.