BOWLING GREEN STATE UNIVERSITY

REGRESSION ANALYSIS

WALMART SALES PREDICTION

Project Group 1:

Brett Conley Naresh Kulkarni

Ganesh Mandarapu Karishma Garewal

**TABLE OF CONTENTS**

**SUMMARY**

**MEET THE DATA**

**Data Dictionary**

**MODELLING**

**Data preparation**

**Variable Selection**

**Multicollinearity**

[**Assumptions of Linear Regression**](#_3znysh7)

**Interpretation of the model**

**REFERENCES**

**SUMMARY**

This data we collected from the ongoing competition in kaggle. Which is provided with historical sales data for 45 Walmart stores located in different regions. Each store contains many departments, and sales for each department in each store. In which we are considering the weekly sales as response variable.

The modelling starts with exploratory data analysis. This step is required to understand the data and to estimate how far the independent variables are useful to predict the target variable. The next step in model building is to select the significant variables which can be done by using any of the methods like all possible regression methods, forward selection, backward elimination, and stepwise regression etc. The final model is built with the significant variables which is then checked for consistency. This is achieved by checking if the formal assumptions are validated, nonexistence of multicollinearity, outliers, high leverage data points, and influential points.

The final model includes monthly sales as response variable and store number, month of sales, maximum temperature in the month, Average unemployment and minimum fuel price as predictor variables which is built based on step AIC technique. Before going for the model building we explored the data to make it ready for the model. This include the data preparation step in which we must change the variable to make sense of model. There are some categorical variables for which we created dummy variables and Date variables are changed into multiple columns for year, month and Day which will help us to analyze the monthly and weekly sales. And then we aggregated the data to bring out the monthly sales after one sample model. The model has a multiple R squared value of 94.22% which shows the model built can explain the variation significantly. The final model is then tested for multicollinearity, and formal assumptions. The final model built has no multicollinearity between the variables and all the formal assumptions have been validated. Later, we checked whether there are any outliers, high leverage points and influential points. We could find outliers in the current model. So, a new model is built with same variables without outliers and verified if there is any significant increase in the Multiple R squared value and adjusted R square value. As there is no significant increase in those values, the data values are not removed from the dataset.

**MEET THE DATA**

The dataset contains weekly sales of Walmart in the United States from 2010-02-05 to 2012-11-01. And, the data consists of 45 stores of data department wise. There are so many effective explanatory variable in the data set like temperature, fuel price, unemployment, CPI, and promotional activities which may affect the weekly sales which is our dependent variable.

**Data Dictionary**

**Stores**

This data set contains information about 45 stores, indicating type and size of the stores.

**Train Data**

Train data is having historical data of weekly sales from 2010-02-05 to 2012-11-01. Below are the fields present in the dataset.

**Store:** The store number

**Dept:** The department number

**Date:** The week

**Weekly\_Sales:** Sales for the given department in the given store

**IsHoliday:** Whether the week is a special holiday week

**Features**

Features data is containing some additional data related to department, store and the promotional activities and employment status. Below are the fields present in the features dataset.

**Store:** The store number

**Date:** The week

**Temperature:** Average temperature in the region

**Fuel\_Price:** Cost of fuel in the region

**MarkDown1-5:** Anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.

**CPI:** The consumer price index

**Unemployment:** The unemployment rate

**IsHoliday:**  Whether the week is a special holiday week

**Modeling**

**Attribute selection**

We chose to start off by executing a stepwise model selection to help ascertain which variables produce the best model for us to start with, but before executing that we had to make a few alterations to the data. First, we split our original Date value (which is the first calendar day of the week in the raw data) into the constituent parts (month, day, year). Our assumption here is that the month carries just as much – in some cases possibly more – impact on the total sales as the day of the week would. The reason for this assumption is rooted in the idea that the specific date and week of the month in which holidays occur does change from year to year in addition to the fact that for Date to be important, we would need to have Daily sales data instead of Weekly sales data to properly weight and determine. After executing the step-wise regression, the model we end up with is suggesting that all the variables must be included – including the Day Number – which yields a total of 172 variables. This feels unreasonable to us because we all know that including all variables in a model is dangerous and generally causes over-fitting issues. For these reasons, we also choose to aggregate our data as described in the next section but before moving on, we explore the results of the stepwise model selection of the raw data.

**Data aggregation**

To resolve the concerns cited in the Attribute Selection section, we drew on our practical experiences and decided to remove the weekly variable altogether and aggregating the raw data up to the Year-Month level but retaining the Store ID and Department level of granularity. Doing so caused us to re-address multiple variables that were included in the raw data previously and so the next decision was to evaluate the basic statistics of those raw values (Min, Max, Average, Median).

The variables in question after aggregation were: Temperature, Fuel Price, Unemployment Rate, and Consumer Price Index. Also, note that with the step of aggregations, we chose to remove the Holiday and Markdown variables as they are specific to dates or specific weeks. These variables will be reintroduced later in the process.

**Attribute selection – round 2**

We again choose a stepwise approach to assess the variables and with this execution we choose to include all of the stats (Min, Max, Avg., and Median) for the variables mentioned in the above Aggregation section. The result of this pass provides insight that certain variables are not significant. Most notably:

· No variants of the Consumer Pricing Index (CPI) metric are selected.

· Only the Average variant of Unemployment Rate metric is selected.

· Max, Min, and Median variants of Temperature and Fuel Price metrics are selected.

As can be seen in the notes above, the stepwise selection method on our aggregate data produces some duplicitous metrics; we quickly determine that it makes no sense to include multiple metrics tied to a single variable so we first assess the aggregate model to see if we can narrow down the variables to more meaningful set. A quick inspection of the Pure Error ANOVA results identifies that the p-values for Min and Median Temperature values are both exceed our 95% confidence (at 0.111 and 0.115 respectively) and quickly conclude that we can remove both of those to cut down to a single instance of a Temperature variable.

Additionally, we note that the Fuel Price variants are significantly different with Median, Min, and Max having p-values of 9.45E-15, 0.0014, and 0.025 respectively. That being the case, we also remove the Min and Max variants of the Fuel Price variables leaving us with a model having only a single reference for each of our variables. Unfortunately, when we review the new model, we see that Median Fuel Price now has a p-value of 0.054 which exceeds the threshold – so we determine the Min Fuel Price is the next best option and validated that it’s p-value in the new model is 0.001, which is well below our threshold.

At this point we also reassess the previously removed Temperature variables to make sure we have the best option available and find that the Max Temperature does produce the best model.

**Comparison of models**

After having executed stepwise attribute selections for both the raw data and the aggregated data, and after reducing the aggregated stepwise model, we need to decide between the models as to which one we believe will have not only the best predictive power but also which one makes the most logical sense to move forward with. To achieve this, we run the standard ANOVA process, AIC calculations and standard model summaries for each model to evaluate the strength of each attribute and of each model overall. Note that we do not use the partial F-test because that test is designed to compare models based on the same data set – because we chose to aggregate the data this test is not the best choice to drive our decision.

|  |  |  |
| --- | --- | --- |
| **Area of Concern** | **Model 1 – Raw Data** | **Model 2 – Aggregate Data** |
| **Largest p-value** | 0.011 (Fuel Price) | 0.0022 (Avg. Monthly Unemployment Rate) |
| **Adj. R-Squared** | 66% | 94% |
| **AIC** | 9,198,390 | 43,772 |
| **Number of Variables** | 172 | 58 |
| **Sum of Squared Error** | 7.380E13 | 5.076E14 |
| **Mean of Squared Error** | 1.75E8 | 3.559E11 |

Based on the evidence to this point, we conclude that the reduced aggregate model is more likely to provide a more reliable prediction and choose to move forward with Aggregate model.

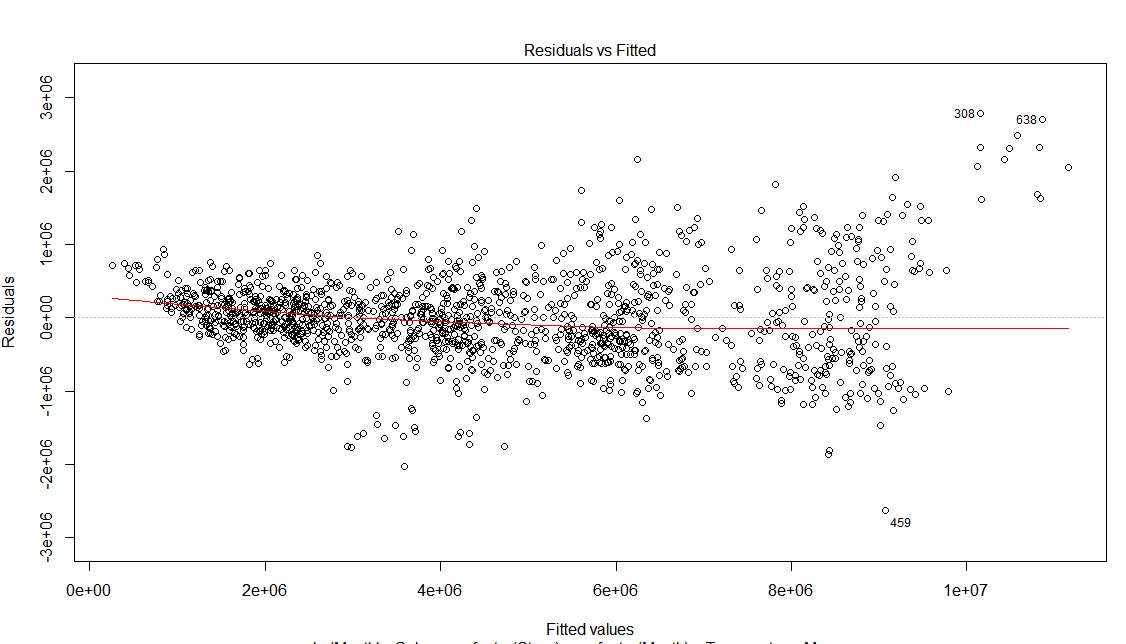
## **Assumptions of Regression model**

There are four assumptions to be verified from the model:

* The relationship is linear in explanatory variables.
* Constant variance assumption.
* Normality assumption.
* Independence assumption.

**Linearity assumption**

Linearity assumption can be verified by using plot of the model and the residuals vs fitted values graph which shown below.



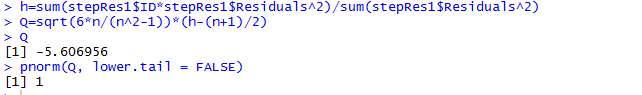
**Constant Variance Assumption**

From the above residual vs, fitter values plot itself we can check the constant variance. And it can also be verified by Szroeter's Test.

Hypothesis for the test is:

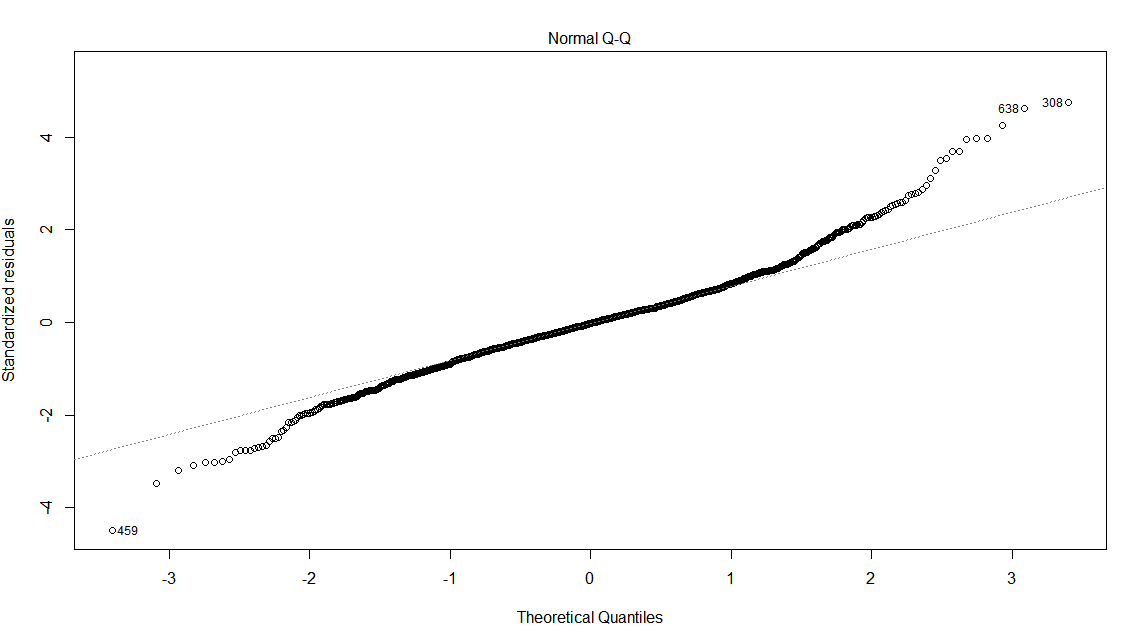
H0: Variance is constant

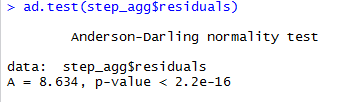
Ha: Variance is not constant



From the above result p-value is greater than the significance level which is 0.05, so we fail to reject null hypothesis. Hence the constant variance assumption is validated.

**Normality assumption:**





This is because of some outliers present in the final data. But these outliers are having some

Importance in the prediction.

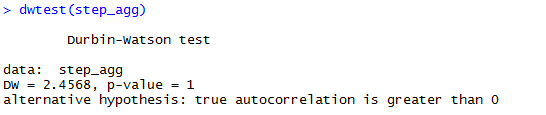
**Independence assumption:**

Using Durbin Watson test we can validate the independence assumption for the model.

Hypothesis for this test is

H0: The error terms are positively auto correlated

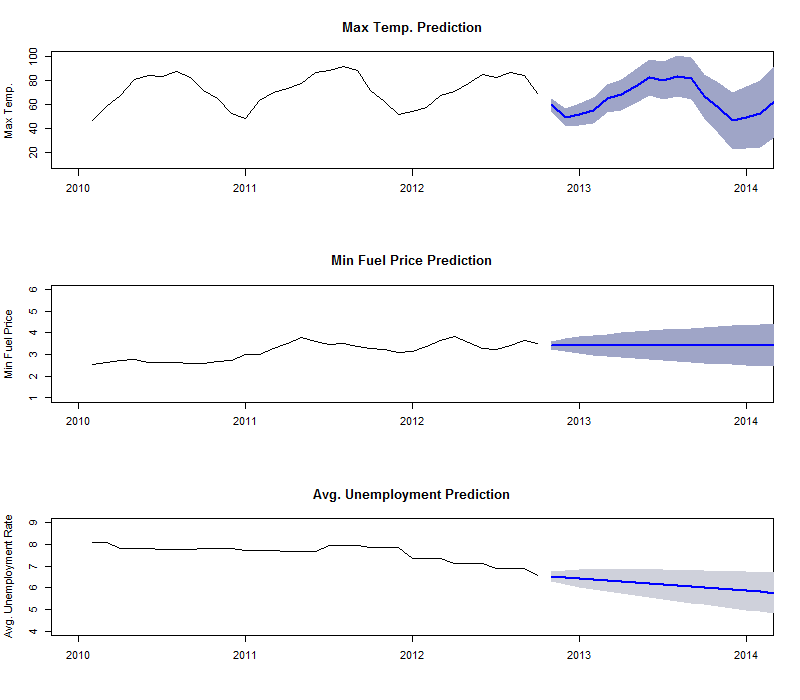
Ha: The error terms are not positively auto correlated



From the above result p-value is greater than the significance level which is 0.05, so we fail to reject null hypothesis. Hence the error terms are independent.

**Time-series predictions**

After having selected the best model to be based on the monthly aggregation data, we take a step back to decide on an approach to making predictions. Due to our own lack of experience with applied time-series analysis we are unfamiliar with all the possibilities and there does not appear to be great amount of documentation for multivariate time series executions so we ultimately decide to proceed by making multiple, single variable times series estimations of our critical values that would be necessary to fit into the model selected above. To that end, we create three time-series models to make prediction of these values as inputs to our Monthly Sales model, though these must be store-specific to be accurate. Here’s a sample using Store ID 1:

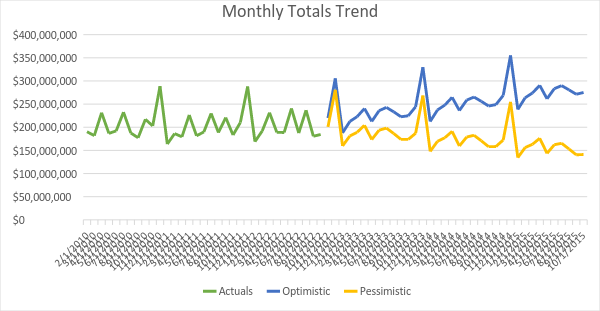


Now that the time series models have been produced, we need to make another decision about how to integrate the results from these predictive models. After inspecting the sensitivities of each of the time series models, we choose to use a 90% confidence interval for the Unemployment Rate model and an 75% confidence interval for the Max Temperature and the Minimum Fuel Price models. Furthermore, we choose to include both an “Optimistic” and a “Pessimistic” perspective buy applying the following decisions:

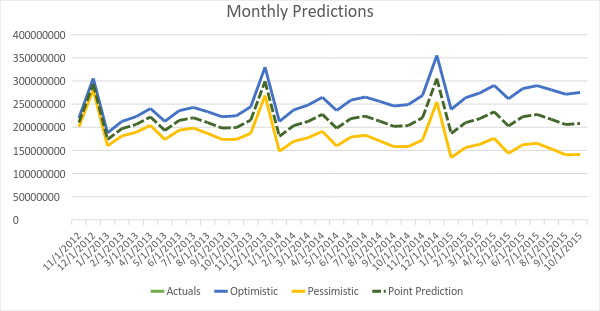
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Confidence** | **Reasoning** | **Pessimistic** | **Optimistic** |
| **Min Fuel Price** | 75% | Higher fuel prices act as a barrier to many people for going out and shopping in store. | Upper bound | Lower bound |
| **Max Temperature** | 75% | In the US, lower temperatures are generally believed to have a negative effect on overall consumer spending but since our metric is the Max temperature, we reverse this knowing that the max temperature will likely have the same effect in the hotter regions of the country. | Upper bound | Lower bound |
| **Average Unemployment Rate** | 90% | Unemployment is known to have negative correlation with overall spending in the US. | Upper bound | Lower bound |

With these decisions, we produce all the predictive values necessary to input into our linear model to produce a Monthly Sales prediction for each Store. The above decisions regarding Pessimistic and Optimistic are used as our upper and lower confidences and our point prediction for each month will be the median between the Pessimistic and Optimistic monthly sales projections.

**Results from the Model:**



Building on the previous section’s discussion of Pessimistic a and Optimistic outcomes, we achieve the Trending prediction seen in the above chart. As to be expected based on the training data, we see significant spikes in the month of December. With this interpretation, we are using the Pessimistic and Optimistic predictions as out confidence interval with our point predictions based in the center of these two values for each month as can be seen in the image below, where the dotted green line is the point prediction value.



To achieve the final output of the problem statement, these monthly totals must be divided down into individual Store-Department contributions to the total months in our historical data.

**References:**

Source of data:<https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data>