Weekly Store Sales Forecasting for Walmart

Professor: Regina G Dolgoarshinnykh

Group 27

Xiaoyang Yang (xy2231)

Fan Ye (fy2188)

Wenjia Cao (wc2477)

Chenyun Zhao (cz2321)

Xiaoran Wang (xw2313)

Abstract

In this project, we build several statistical models to predict weekly sales in each Walmart store using historical weekly sales and other features that might be influential. Before moving to data analysis, we firstly give a brief overview of this project, which includes the motivation for choosing project topic and reasoning for selecting initial prediction variables. The data analysis part in this report mainly contains three steps. The first step is preprocessing the data: this includes detection, visualization and imputation of missing values. In the second step we look deeper into the relationship between each store and its features; we tried to assign each store a score to represent different types of stores. In the third step, based on previous analysis, we use ordinary linear model, partial least square model, K nearest neighbors, regression tree to fit the data. We use cross validation to tune the model and evaluate each model based on their prediction ability and complexity.

Finally, we conclude that the optimal model is linear model using store, month, day, whether it is a holiday, CPI, unemployment, fuel price and temperature, which has overall predictive R^2 of 92.68%. After determining the optimal model, we further include Department into consideration and use elastic net to conduct linear model with penalty on coefficients. We also discuss several possible refinements for our final model in the end.

Contents

1. Overview & Motivation

The goal of our project is to predict Wal-Mart's weekly sales, given the data from Kaggle (https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting). The motivation for choosing this project is that this project can generate great economic benefits for Wal-Mart, especially when its growth is slowing down under the strike of new emerging chain stores like Target and e-commerce giant Amazon. Being able to build a model to predict weekly sales in different departments among different stores would help to optimize supply chain management and inventory management, which would further contribute to cost control and to generating more revenue.

We have a training dataset, which includes, store, department, date, weekly sales and whether this week includes holiday. In addition to this, we also include some other features that might influence weekly sales, they are:

i. Temperature. It's based on our intuition that a mild weather and an upcoming extreme weather would trigger the sales, while an unfavorable condition would make the sales drop.

ii. Whether this week includes holiday. The holidays here are defined as: Thanks Giving Day, Christmas, Super Bowl and Labor Day. It is known to us that during these holidays sales are booming.

iii. Markdown. It is based on our intuition that markdown would trigger sales.

iv. Fuel Price. Based on the fact that Wal-Mart has a large number of discount department stores and warehouse stores, which are mainly built in suburb areas. So fuel price may influence peoples' visiting frequency.

v. CPI. According to our knowledge in economics, Consumer Price Index would influence sales.

vi. Unemployment. This macroeconomic figure is a good indicator of people's desire to buy.

The last three are macroeconomic data, based on our knowledge in economics, these would influence sales in a certain degree. Such effects are further validated by Wal-Mart's clients' demographic information. The average of US Wal-Mart customers' income is below the national average, they tend to be the group of people who are more vulnerable to seasonal unemployment changes, fluctuation in fuel prices and CPI. A Wal-Mart financial report in 2006 also indicated that Wal-Mart customers are sensitive to higher utility costs and gas prices.

The challenges we meet in the earlier stage of analysis mainly come from two aspects. The first is the large number of predicted values. We have 45 different stores and each store has 99 departments and that means we are going to need 45*99 predicted values for a single date. It will be a great change for building prediction model. The second is dealing with missing values, we have large numbers of missing value in Markdown, CPI, Unemployment. We need to figure out a reliable way to deal with it in order to build a reliable model.

2. Preliminary Analysis

2.1 Data set overview

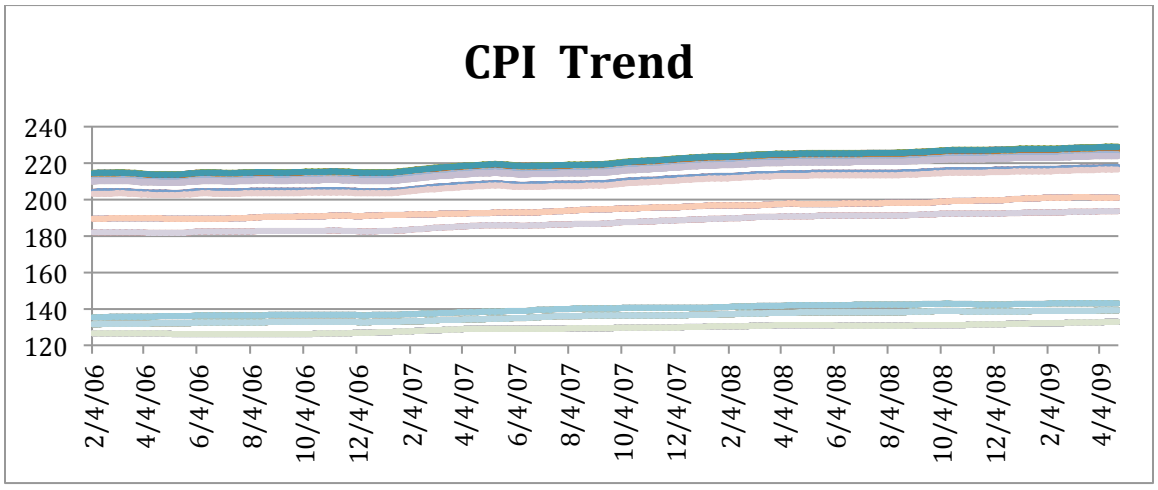
We have 4 datasets in total, these are *Features*, *Stores*, *Train*, and *Test*. In our analysis, we mainly use the first three datasets. Details of these datasets can be found in our proposal.

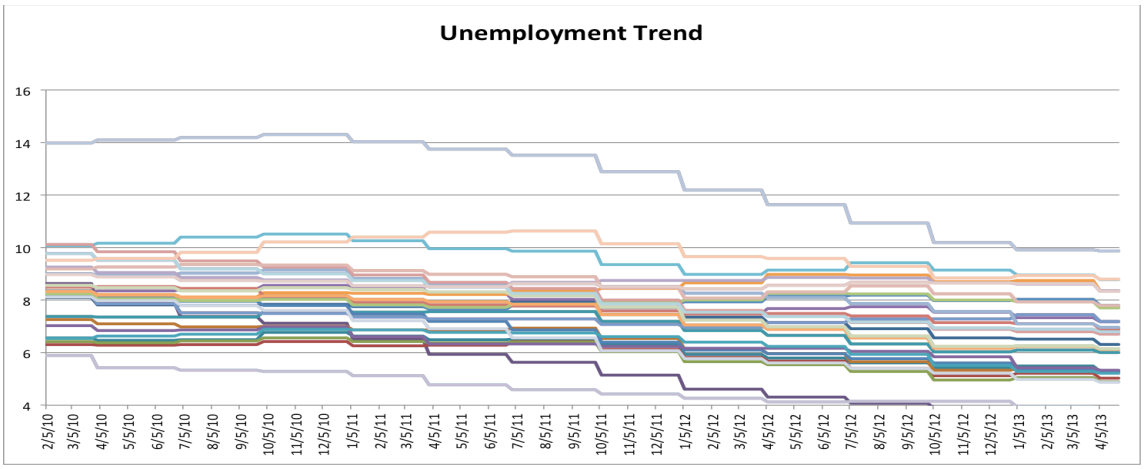
There are 45 stores, each with 99 departments. We have weekly sales data for these departments in different stores from 2010-02-05 to 2012-11-01. We also have features data of these 45 stores from 2010-01-05 to 2013-07-26. We use the data in 2010-02-05 to 2012-11-01 as our training data to build models.

2.2 Fit missing values

For these datasets, there exist some missing values in different variables through the time. From a further review, we found out that values of CPI and unemployment for all the 45 stores are all missed from 2013-04-26 to 2013-07-26.

Plot 1 shows the relationship between time and CPI. From this we can tell that CPIs of different stores have a similar trend, though they vary in their values significantly. This is consistent with empirical studies that CPI is mostly influenced by the overall economic situation of the nation as a whole, while the exact values would vary according to local development level. Together with its economical meaning, it is possible to predict its missing values using time series model.





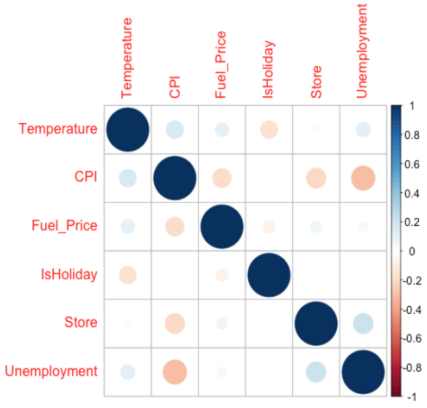
Plot 1

A same pattern is also found in unemployment rate, so we applied the same method to fill in the missing value of unemployment.

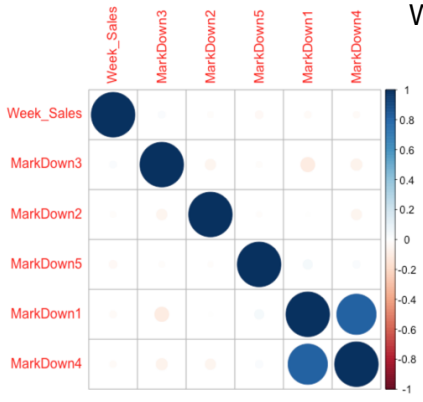
Markdowns are also variables that have missing values. Markdown data are only available after November 2010. For this part of data with Markdowns, there are missing values in different Markdown categories too. However, unlike CPI and unemployment, neither pattern nor other additional information do we have about it, so if we want to include it into the model, we need to do further analysis to see whether it really impacts weekly sales.

2.3 Exploratory analysis

Firstly we tried to see the relationship between the features to see if any dimension reduction could be made. The result is shown in *Plot 2*, these variables don't have much correlation， the highest correlation is around 0.5, appears between unemployment and CPI.

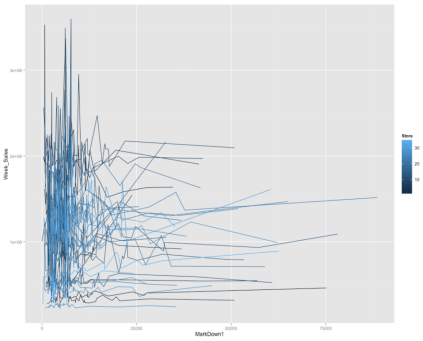


Plot 2



Plot 3

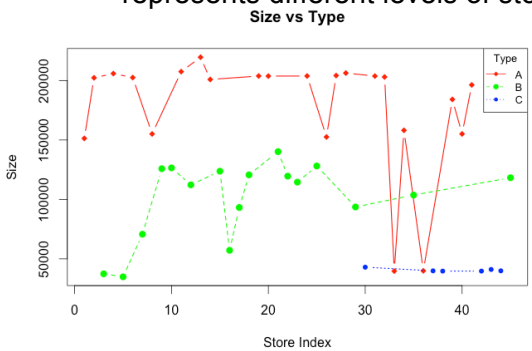
Weekly Sales against MarkDown1



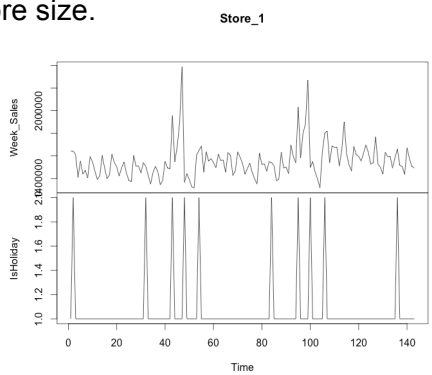
Plot 4

Since there exist a large scale of missing values in Markdowns with unknown pattern, we tried to see if they really influence weekly sales. The result is shown in *Plot 3*, we find that the correlation between weekly sales and the five Markdowns (exclude missing values) are very weak. A detailed plot of the relationship between Weekly sales and Markdown1 is shown in *Plot 4*. For most of the stores, Markdown1 has a little impact on weekly sales. Considering the complexity of our model, we no longer include markdowns in our following analysis.

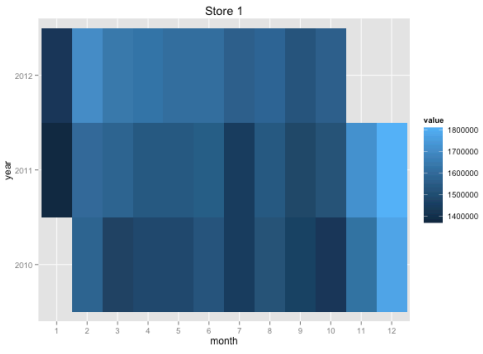
The relationship of Type and Size in each store is shown in *Plot 5*. For different types, we can see significant differences in stores' size. Although there are some 'unusual' cases in each type, like store 3 and store 5 in Type B, store 33 and store 36 in Type A, in most cases, different Type represents different levels of store size.



Plot 5



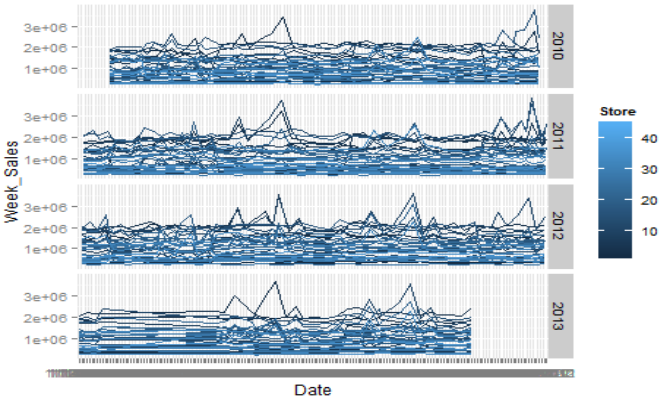
Plot 6



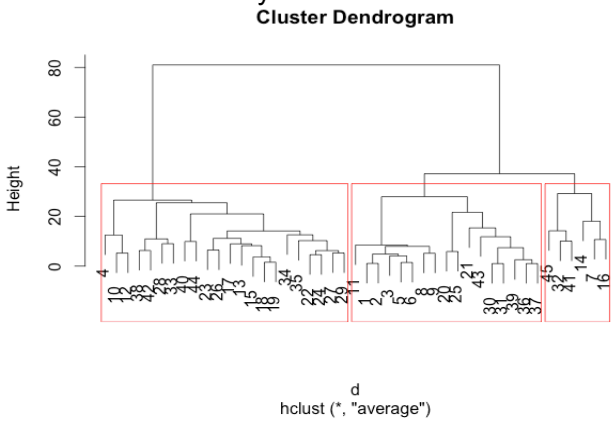
Plot 7

Using Store 1 for further exploratory analysis, we plot whether this week contains a holiday against time and weekly sales against time, as shown in *Plot 6*. *Plot 7* shows the average monthly sales from 2010-02-05 to 2012-11-01 for Store 1. We can see that weekly sales fluctuate with the existence of certain holiday, Thanks Giving Day and Christmas. There are significant seasonal effects in average monthly sales. There is a slight year effect in the increase of average monthly sales.

The situations of date against weekly sales for 45 stores in different years are shown in *Plot 8*.



Plot 8



Plot 9

From this plot, we can find that the peaks of weekly sales always come with the holiday. But the trend is not so obvious for some stores.

So for weekly sales, type, size, whether this week contains a holiday, date all have some influences. For features data, like CPI, unemployment, temperature, fuel price, based on our previous analysis, they represent different features of each store. So we will include these predictors in our next step – assigning each store a feature score.

Because given a certain store, features data are the same for different apartments. So we sum weekly sales of different department in a certain store to have the weekly sales for each store, and we set this as our independent variable.

3. Assigning feature score to each store

3.1 Original classification of 45 stores

Based on the given dataset, 45 stores are classified to 3 categories A, B and C. According to our previous analysis, this classification is mainly based on store size.

3.2 New classification of 45 stores based on features

We want to find another classification that is based on features, CPI, unemployment, fuel price and temperature. Because these features show correlation to time, in order to eliminate the influence of time, we choose hierarchical cluster instead of k-means. In order to compare hierarchical cluster classification result based on 4 features data and the original classification based on size, we set the number of clusters to be 3.The classification result is shown in Plot 9. As we can see from the plot, it is very different from the original classification. We can conclude

that these new categories represent macroeconomic feature and geographic feature of 45 stores. To be simple and accurate, we create a new categorical variable named Group to represent these new classifications.

4. Model Building and Model Selection

4.1 Model Building Method

We use weekly sales of each store as our independent variable. Through trials and errors, we have our initial set of predictors to be Store, whether this week contains holiday, month and date. In order to compare the influence of the 4 features, we have an improved set of predictors, which in addition to weekly sales, has CPI, unemployment, temperature and fuel price.

We have 6435 observations as our training data. Because we don't have any information of weekly sales in our test data, we split our training data to get a new training data and a new test data. We first randomly select 75% of our original training data to tune and fit the model using 10 fold 5 repeated cross validation, the remaining 25% of the original training dataset as a new test data to test the prediction ability of our model. The methods we applied are Linear Regression, PLS (partial least squares), KNN (K-nearest neighborhoods) and Regression Tree. For each method, we build two models based on our two different predictor datasets.

4.1.1 Linear Regression

Background:

We do not consider each department. We sum weekly sales from different department in each store, and predict the weekly sales of a specific store. This reduces our training data set to 6435 entries. Our linear model is nothing but to solve minimization problem below:

$$\hat{\beta}^{OLS} = \arg\min_{\beta}\{(Y - X\beta)^T (Y - X\beta)\} \quad (1)$$

Results:

Number of components is the tuning parameter of PLS model, tuning results of these two models and the prediction ability of tuned model are shown in *Table1* :

Table 1

	Number of components considered	Prediction RMSE	Prediction Rsquare
Without 4 features	4	149345	0.9256
With 4 features	8	147634	0.9272

Model Summary:

The result of linear model is good, especially in terms of R^2. In *Table 1* above we can see that our linear model explains almost 92% of variation. To be noted, we use dummy variable to represent different Store, while using month and date to represent specific date. The last

predictor is 'IsHoliday', which tell us if the specific day is holiday or not. We can see that it is helpful to add 4 features in our model since the RMSE, which is calculated by cross validation, decrease. However, these four features have limited contribution to our model in sense of R^2. The additional four features have moderate influence on sales, while date and store are dominant predictor.

4.1.2 PLS (Partial Least Squares)

Background:

In practical problems,, the OLS assumptions of independency between predictors couldn't be meet, especially there are several predictors. Such violation brings high variability and unstableness to the model. In case of such phenomena, for the improved predictor set, we would apply PLS to estimate parameters.

PLS method uses NIPALS algorithm, it iteratively seeks to find underlying, or latent relation. In each iteration, it assesses the relationship between the predictors (X) and response (Y) and numerically summarizes the relationship with a vector of weights (w), which is also known as a direction. The predictor data are then orthogonally projected onto the direction to generate scores (t). The scores are then used to generate loadings (p), which measure the correlation of the score vector to the original predictors. These linear combinations of predictors are called components or latent variables. PLS linear combinations of predictors are chosen to maximally summarize covariance with the response. It strikes a compromise between the objectives of predictor space dimension reduction and a predictive relationship with the response.

According to our 4 features predictors, they are in the same scales of differing magnitude. So we do not consider centering and scaling.

Results:

Number of components is the tuning parameter of PLS model, tuning results of these two models and the prediction ability of tuned model are shown in *Table 2*:

Table 2

	Number of components considered	Prediction RMSE	Prediction Rsquare
Without 4 features	11	149345	0.9256
With 4 features	13	148103	0.9268

Model Summary:

From *Table 2* we can see that, PLS model gives similar good results as ordinary linear regression model.

For the first model without the 4 features, PLS model gives exactly same result as ordinary linear regression model, which is because these predictors are not correlated; they separately represent different predictive information. However, by including the 4 features, the number of components increased from 11 to 13, there is a slight decrease in Prediction RMSE and a slight

increase in Prediction R^2, because we add more informative predictors into the model. However, PLS model with 4 features performs a little worse than ordinary linear regression model with 4 features, so the cost from reducing number of predictors is more than the benefits.

4.1.3 Regression Tree

Background:

Basic regression trees partition the data into smaller groups that are more homogenous with respect to the response. To achieve outcome homogeneity, regression trees determine:

• The predictor to split on and value of the split

• The depth or complexity of the tree

• The prediction equation in the terminal nodes

While trees are highly interpretable and easy to compute, they do have some noteworthy disadvantages. First, single regression trees are more likely to have sub-optimal predictive performance compared to other modeling approaches. An additional disadvantage is that an individual tree tends to be unstable.

In this case, we use regression tree to predict the weekly sales and also we use 10 folds cross validation to resampling.

Results:

Firstly, we choose the variables from the training data: ISHoliday, Month, Day,store and four varibles from features data: CPI, Unemployment, Temperature, Fuel_Price

RMSE was used to select the optimal model using the smallest cross-validated error, which named cp in the results. And then we also use the variables without the four variables from features dateset.

The results of Regression Tree model are as shown in *Table 3*:

Table 3

	Smallest cross validation error	Prediction RMSE	Prediction Rsquared
Without 4 features	0.000913	175166	0.8976796
With 4 features	0.0005	194386	0.8737995

Model Summary:

From *Table 3* we can see that, the results of Regression Tree mode are not as good as what we got from the above two methods. Also we can find that there is not obvious difference between the two models, which means that if we include the four features or not, it will not lead to a significant difference. But we still can figure out that if we have more variables in the model, the training error will be less, it's a bias –variance trade off problem.

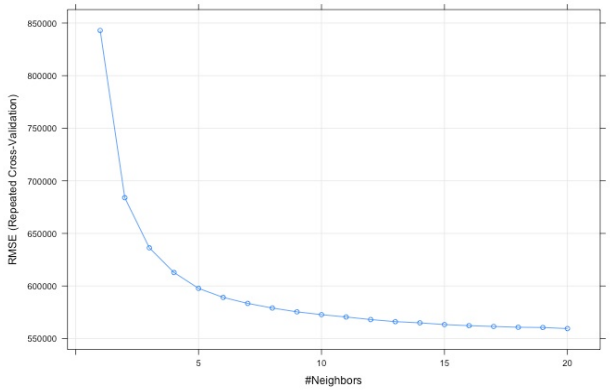
4.1.4 K-Nearest Neighbors

Background:

The KNN approach simply predicts a new sample using the K-closest samples from the training set. To predict a new sample for regression, KNN identifies that sample's KNNs in the predictor space.

Results:

Upon pre-processing the data and selecting the distance metric, the next step is to find the optimal number of neighbors. Like tuning parameters from other models, K can be determined by resampling. For our model, 20 values of K ranging between 1 and 20 were evaluated. As illustrated in *plot 10*, a K of approximately 20 yields the lowest RMSE (Using 10-fold Cross-Validation).



Plot 10

The results of KNN model are as shown in *Table 4*:

Table 4

	Number of neighbors	Prediction RMSE	Prediction Rsquared
Without 4 features	2	431300	0.4625
With 4 features	20	541296	0

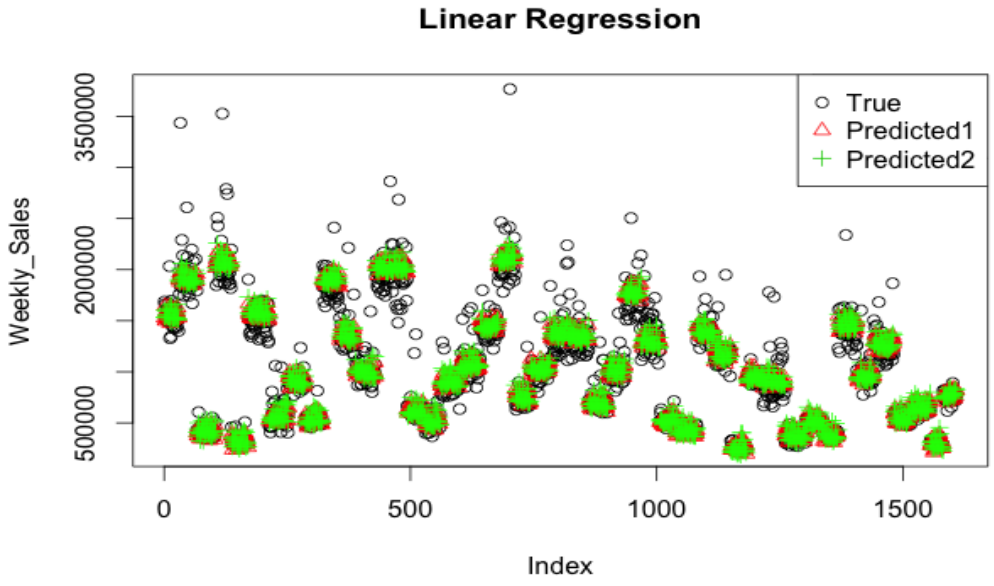
Model Summary:

The results showed that KNN model have poor behavior in our data set.

Then KNN method can have poor predictive performance when local predictor structure is not relevant to the response. Irrelevant or "noisy" predictors are one culprit. In our data set, we convert 45 stores to 44 dummy variables in the model including the dummy variables that are declared non-significant. This approach is problematic when the number of classes is large. By chance alone, as the number of classes increases, the probability of one or more dummy variables being declared non-significant increases. To put all the dummy variables in the model, as non-significant variables are known to be "noisy." Thus, this causes the poor predictive performance in KNN.

4.2 Conclusion

The linear model turns out to be optimal, in sense of statistically significant, predict accuracy and model complexity. *Plot 11* below shows two linear models we built. The red triangle indicates the model include 4 features (CPI, unemployment, temperature, and fuel price). The green plus sign represent result without these 4 features.



Plot 11

5. Model adjustment and further discussion

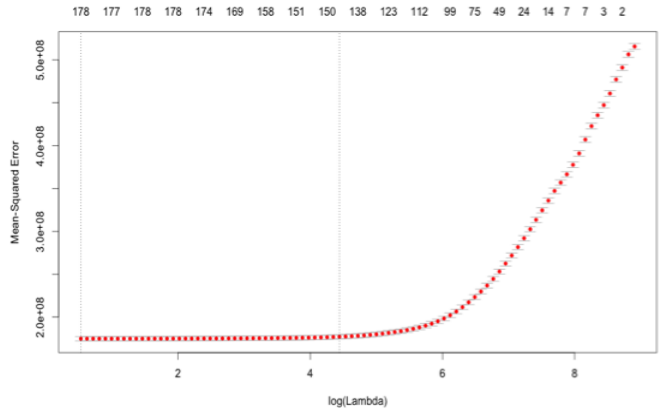
Although the ordinary linear model gives us the optimal result, we do not consider department yet. Here, we take department into consideration. We assume that for each department there is a unique linear pattern of weekly sale over time, we then convert department to dummy variables and add it to our linear model. In total, we have 181 variables. Since this is a large sparse matrix, we need to use shrinkage method to reduce the dimension while fitting our model. Ridge and lasso seems to be right choice here.

5.1 Elastic net

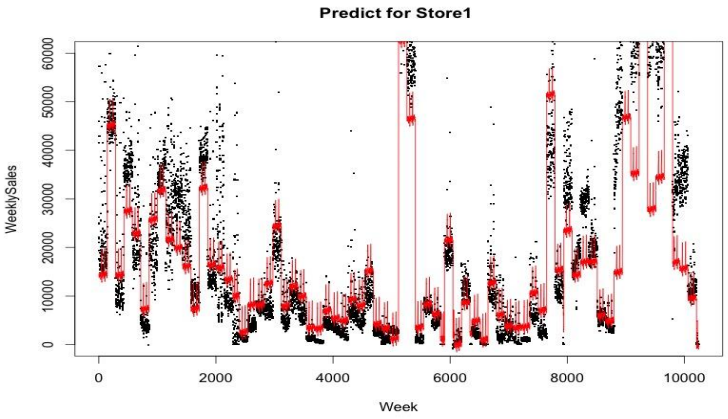
We use elastic net and package glmnet to fit our model. By choosing tuning parameter lambda using 10 folds cross validation, we can decide the number of parameters that included in our model. Below is optimization formula:

$$\hat{\beta}^{elasticNet} = \arg\min_{\beta}\{(Y - X\beta)^T (Y - X\beta) + \frac{1-\alpha}{2}||\beta||^2 + \frac{\alpha}{2}||\beta||^1\} \quad (2)$$

Where *X* is a *n* by *181* matrix, and *alpha* is parameter that control the balance of two norm constraint; *lambda* is a control for total penalty term.



Plot 12



Plot 13

Plot 12 shows relationship of lambda vs. MSE. Minimum value plus one standard estimate of lambda is 84.780, which give us almost 140 predictors out of 181 and totally 65.690% deviation explained. The MSE for this model is 177114935.

Plot 13 shows the predicted result in whole training dataset for elastic net model. The black dot is our training value, while the red line is our fitted value. We only show result of store 1 here and since sales for different department is different over 52 week (that means if we plot our result for 52 week, it would be totally a mess), we use index of week as our x label, the order of this plot is same as order of training data set.

5.2 Further discussion

Elastic net model is extremely popular these years and is very powerful when dealing with sparse matrix. The ordinary linear regression is a less flexible method. However, by adding dummy variable, we make our model more flexible and basically it tend to fit to every single data point.

One way that might improve our model is that we can break our assumption that in each store and department, weekly sale shows single linear pattern. We can fit our model first with dummy variable of Store and department, assuming that for different department in different store, the difference of sales is only a constant. Further, we add quadratic term or interaction term to make our model more sensitive to pattern inside department.

Besides, we also consider using latent variable to replace dummy variables. In our previous model, we use 52 dummy variables to represent 52 week and assume that for each week our linear model has unique constant shift. Here we aggregate our data by each week and take mean value for weekly sales from different and department. And then we use Kernel regression to smooth our predicted line. We use the predicted result as a weight for each week and use it to replace 52 weeks dummy variables. The result turns out to be very close to the method using dummy variable, which give us reasonable predicted relatively. We don't show the detail of it here.

Reference

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning – with Application in R* [M]. Springer, 2013.

[2] Max Kuhn, Kjell Johnson. *Applied Predictive Modeling* [M]. Springer, 2013.

[3] P.J. Brockwell and R.A. Davis*. Introduction to Time Series and Forecasting, 2nd edition,* Springer, 2002*.*

[4] Barber, D. *Bayesian Reasoning and Machine Learning,* Cambridge University Press, 2012

[5] http://www.geniqmodel.com/DummyVariablesTheProblemAndSolution.html

[6] https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

[7] http://www.ats.ucla.edu/stat/r/faq/R_pmm_mi.htm