SALES FORECAST FROM HISTORICAL DATA FOR WALMART

SUBMITTED BY,

GOPIKA JAYADEV

INTRODUCTION

- We have the historical sales data for 45 Walmart stores located in different regions.
- Task is to predict the department-wide sales for each store for the year 2011
- This would assist in planning activity, budget planning, making decisions pertaining to staffing requirements etc.
- Aim is to develop a model that effectively finds variables that influence the sales

DATA SET

Data set contains data related to the store, department, and regional activity for the given time frame

Attributes of 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

Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
1	2/5/10	42.31	2.572	NA	NA	NA	NA	NA	211.096358	8.106	FALSE
1	2/12/10	38.51	2.548	NA	NA	NA	NA	NA	211.24217	8.106	TRUE
1	2/19/10	39.93	2.514	NA	NA	NA	NA	NA	211.289143	8.106	FALSE

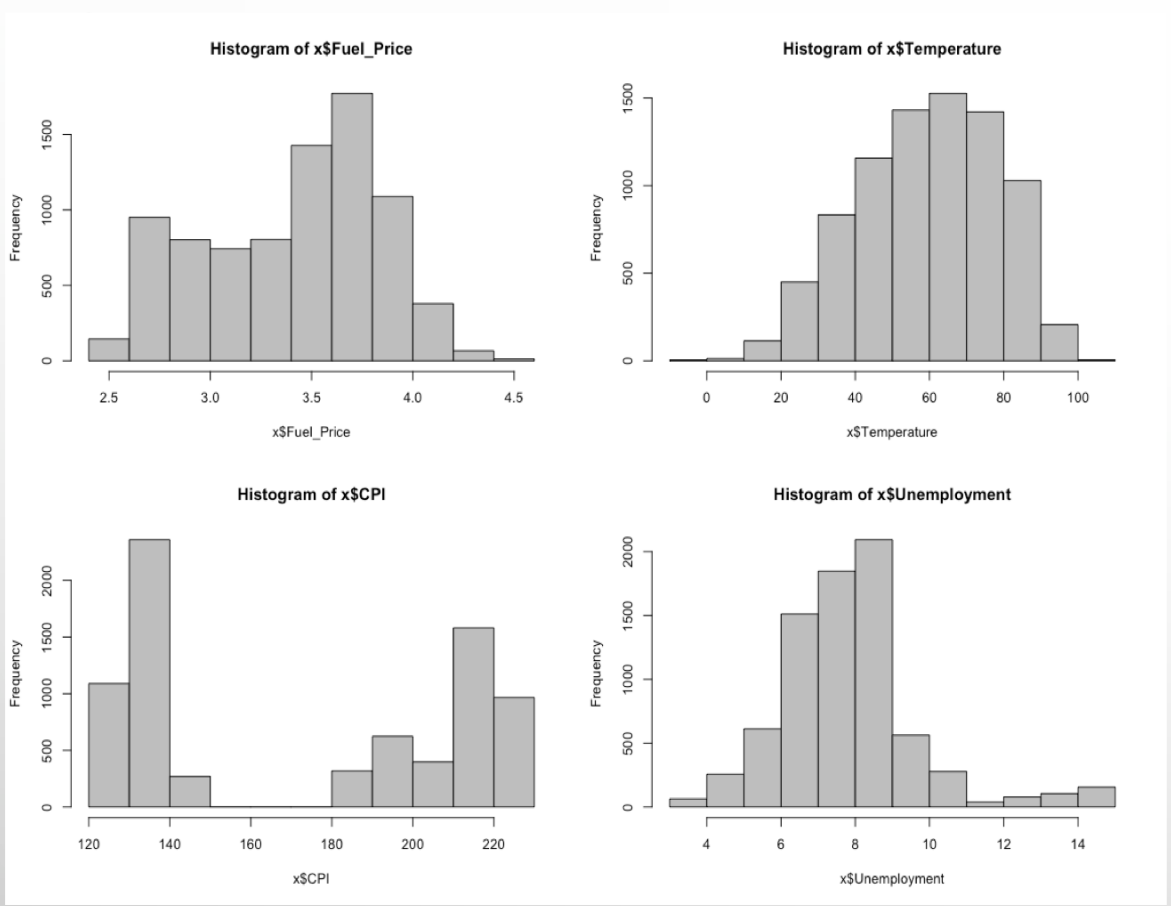
STATISTICAL SUMMARY OF THE DATA

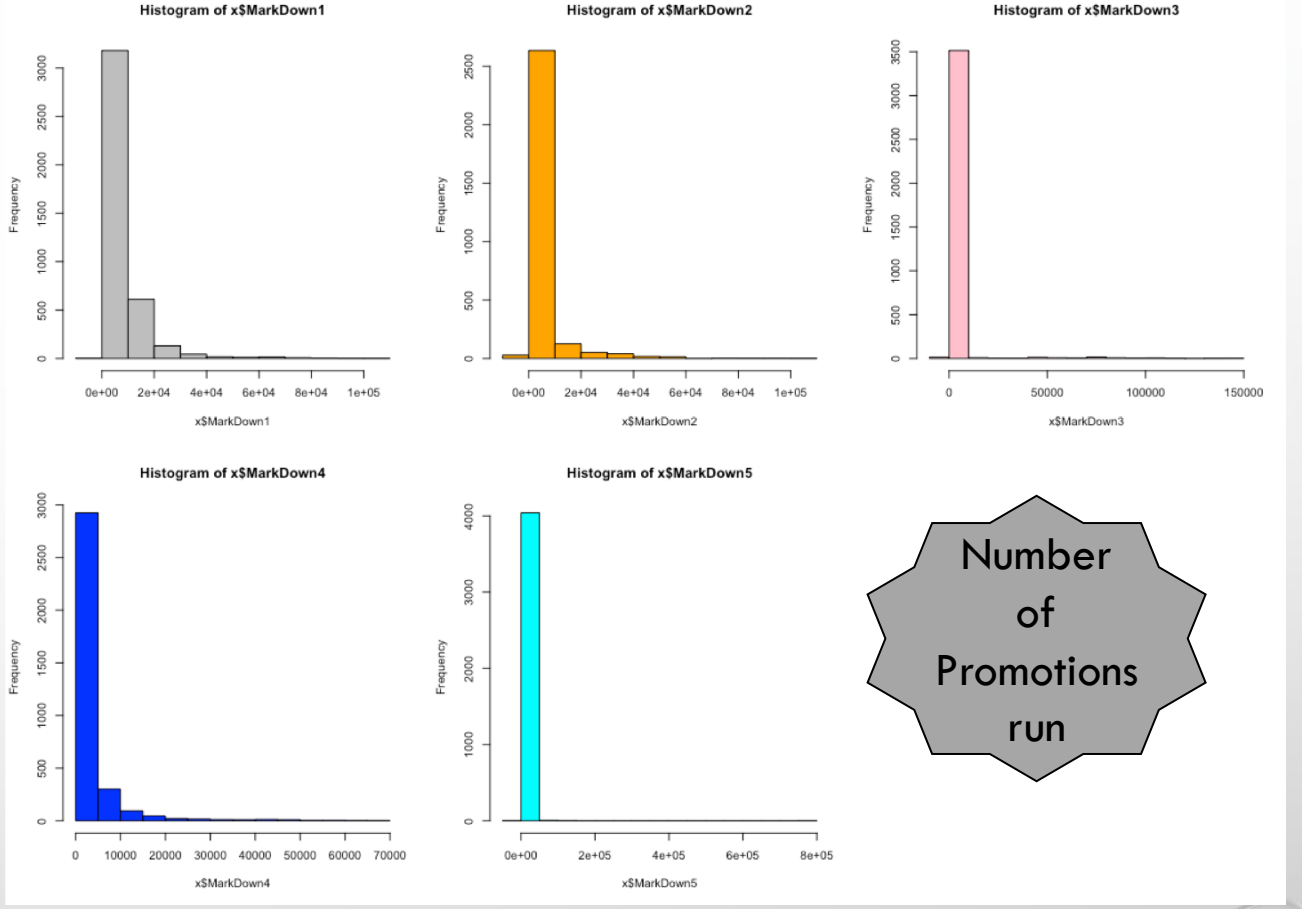
Store	Date	Temperature	Fuel_Price	CPI	Unemployment	IsHoliday
Min. : 1	2010-02-05: 45	Min. : -7.29	Min. :2.472	Min. :126.1	Min. : 3.684	Mode : logical
1st Qu.:12	2010-02-12: 45	1st Qu.: 45.90	1st Qu.:3.041	1st Qu.:132.4	1st Qu.: 6.634	FALSE:7605
Median :23	2010-02-19: 45	Median : 60.71	Median :3.513	Median :182.8	Median : 7.806	TRUE :585
Mean :23	2010-02-26: 45	Mean : 59.36	Mean :3.406	Mean :172.5	Mean : 7.827	NA's :0
3rd Qu.:34	2010-03-05: 45	3rd Qu.: 73.88	3rd Qu.:3.743	3rd Qu.:213.9	3rd Qu.: 8.567	
Max. :45	2010-03-12: 45	Max. :101.95	Max. :4.468	Max. :229.0	Max. :14.313	

Training data: Dated from Feb 2010 to November 2012.
Data fields:
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

Store	Date	Weekly_Sales	IsHoliday
Min.: 1	2010-02-05: 45	Min. : --4989	Mode : logical
1st Qu.:12	2010-02-12: 45	1st Qu.: 2080	FALSE:7605
Median: 23	2010-02-19: 45	Median : 7612	TRUE :585
Mean: 23	2010-02-26: 45	Mean: 15981	NA's :0
3rd Qu.:34	2010-03-05: 45	3rd Qu.: 20206	
Max.: 45	2010-03-12: 45	Max.: 101.95	

HISTOGRAM OF THE FEATURES IN THE DATA





Observations:

- The consumer price index varies as per region and only certain values exist, not all values in the range
- We aim to zero in on which of the featured variables have the most significance on the sales

LASSO REGRESSION MODEL

LASSO REGRESSION OBJECTIVE:

$$\min \sum_{i=1}^{T} \left(y_i - (\beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \cdots + \beta_k x_{ik})\right)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \quad \text{where } \lambda \geq 0$$

BUDGET EQUIVALENCE OF λ IN THE LASSO:

For any given value of the tuning parameter λ in the lasso, there is a parameter budget B associated with the following optimization problem:

$$\min_{\beta_j} \sum_{i=1}^{T} \left(y_i - (\beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \cdots + \beta_k x_{ik})\right)^2 + \lambda \sum_{j=1}^{k} |\beta_j|$$

$$subject\ to:$$

$$\sum_{j=1}^{k} |\beta_j| \leq B\ for\ some\ B \geq 0$$

To use this as variable selection tool, we provide a series of values for λ in the lasso, and for different values of lambda, there are different variables which would be significant.

LASSO REGRESSION IN R

Inputs to the model:

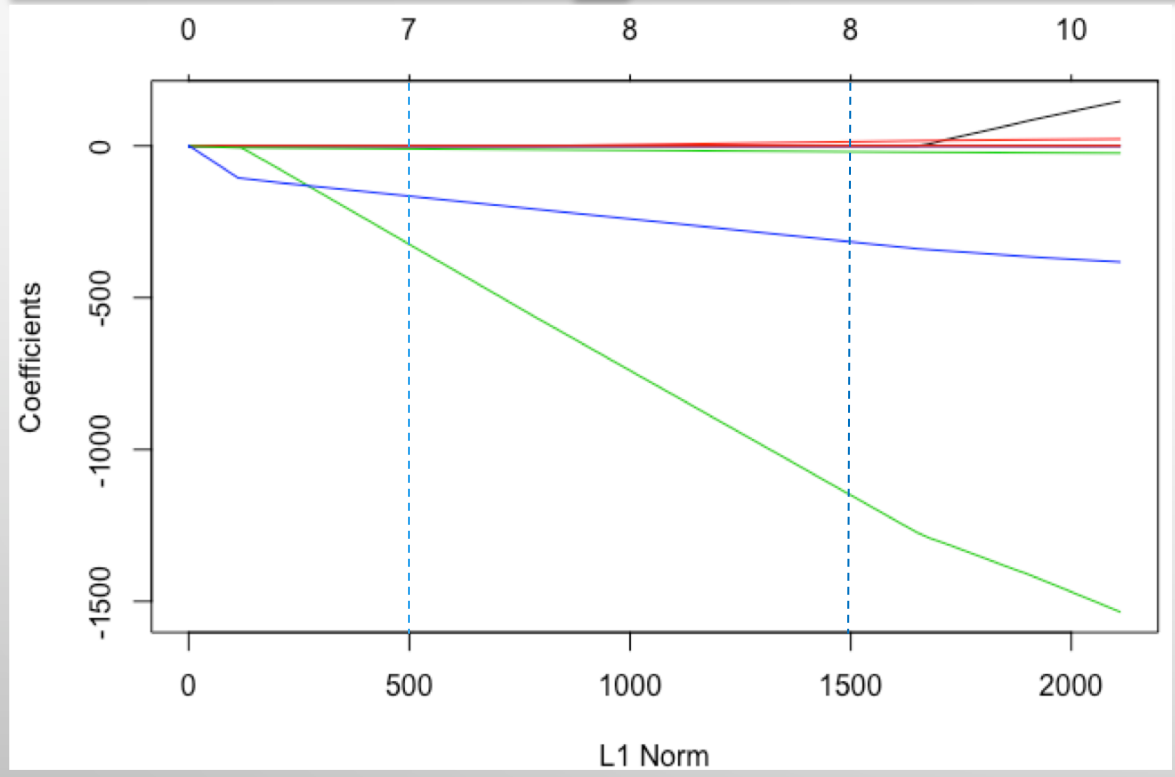
- Training data set
- Sequence of Lambda values
- Testing data set

Outputs from the model:

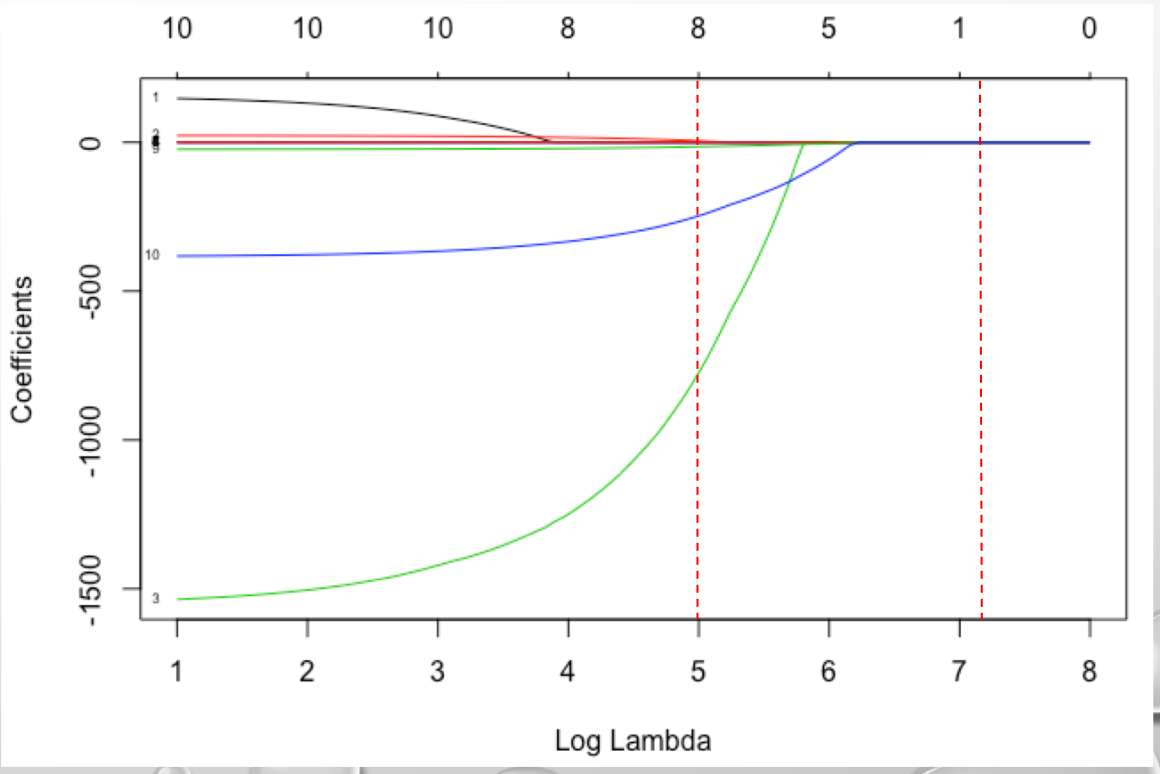
- Plots suggesting No of significant variables
- Cross Validation Plot and lambda values at min and 1SE MSE
- Intercept values for significant variables at each lambda value

Code Excerpt:

```
# Values for lambda
l.val <- exp(seq(1,10,length = 100))
# Dividing the set into train and test
set.seed(123)
train <- sample(1:nrow(x1),nrow(x1)/2)
test <- -train
# Setting up the lasso model
modelf.L <- glmnet(x1[train,],new_y[train],alpha = 1, lambda = l.val)
```



The lower x axis can be considered like a budget and the top x axis denotes the number of significant variables possible at that given budget



The lower x axis is the log of lambda value and the top x axis denotes the number of significant variables possible at each given lambda

LASSO REGRESSION RESULTS

TABLE OF SIGNIFICANCE WITH DIFFERENT LAMBDA VALUES

Log Lambda	Min = 0.9	4.00	5.50	6.00	6.50	7.00	1se = 7.04
(Intercept)	26105.80	24655.65	19672.51	16623.75	15709.78	15961.70	15981.26
IsHolidayTRUE	147.85	0.00	0.00	0.00	0.00	0.00	0.00
Temperature	22.69	16.30	0.00	0.00	0.00	0.00	0.00
Fuel_Price	-1536.39	-1248.62	-346.19	0.00	0.00	0.00	0.00
MarkDown1	0.16	0.13	0.10	0.07	0.04	0.00	0.00
MarkDown2	0.05	0.04	0.01	0.00	0.00	0.00	0.00
MarkDown3	0.16	0.15	0.11	0.08	0.04	0.00	0.00
MarkDown4	-0.04	0.00	0.00	0.00	0.00	0.00	0.00
MarkDown5	0.20	0.19	0.16	0.13	0.09	0.01	0.00
CPI	-24.04	-20.79	-10.17	-3.64	0.00	0.00	0.00
Unemployment	-382.61	-333.91	-169.44	-57.65	0.00	0.00	0.00

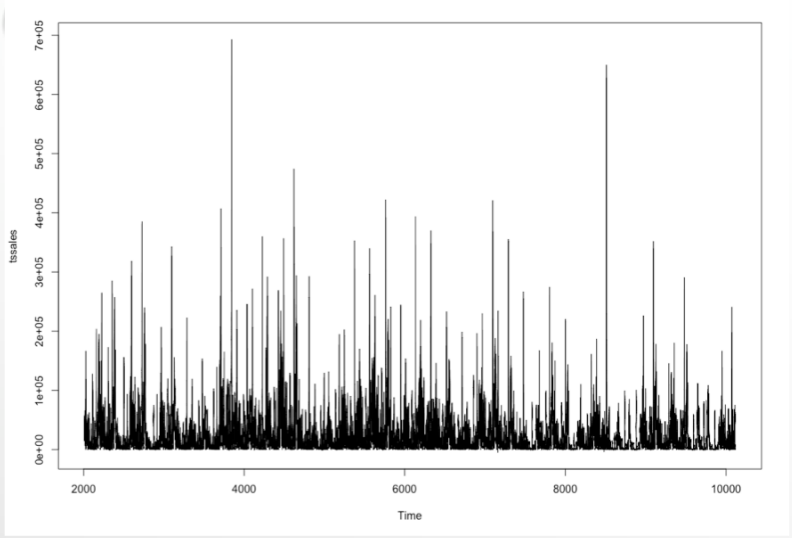
VARIABLES

Lambda is the tuning parameter.
As the parameter lambda increases, the lasso has the effect of forcing some coefficients to be exactly zero.
This property could help us in using lasso as a **Variable Selection Tool.**

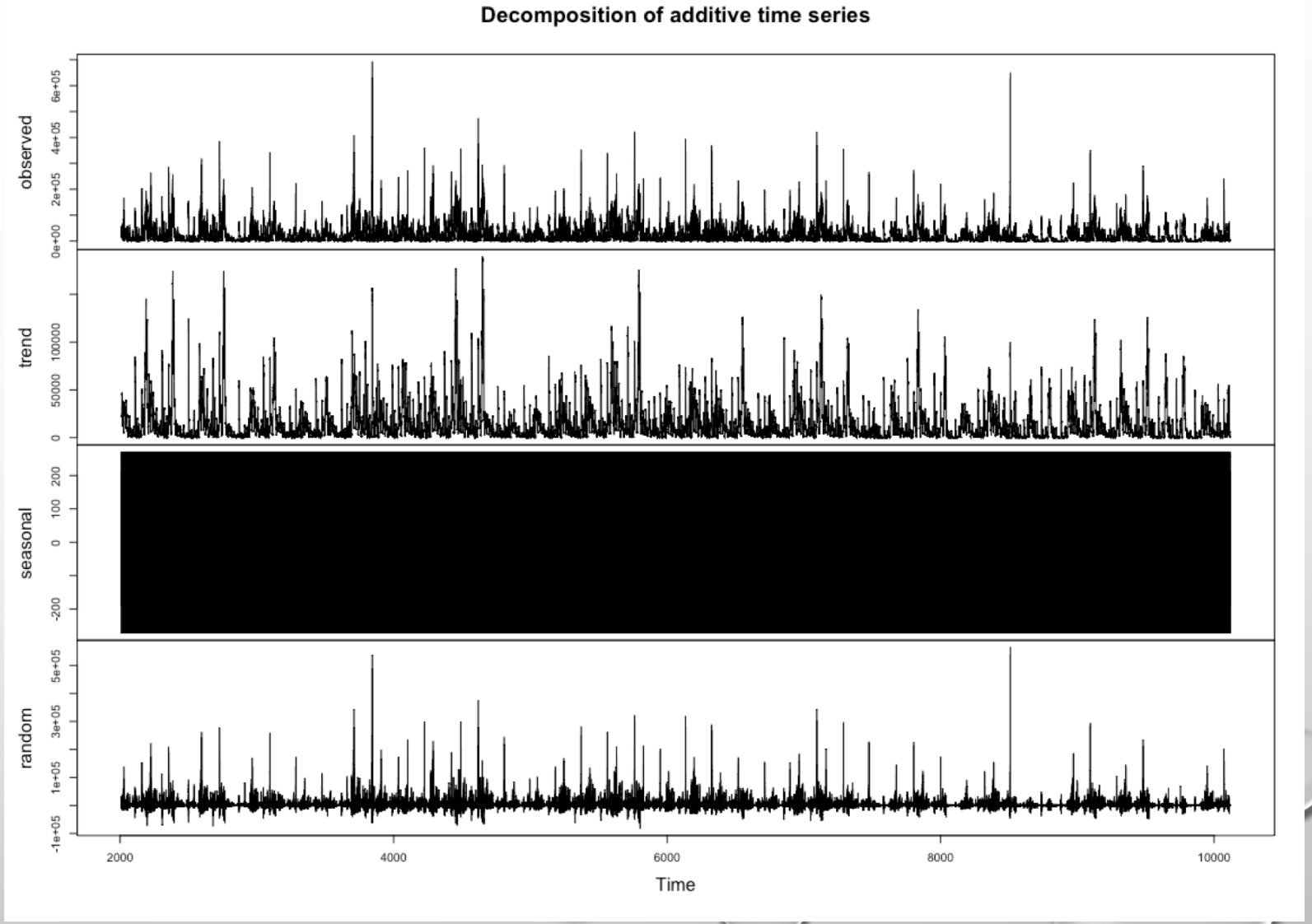
- As seen from the table, the minimum MSE error is achieved when all 10 of the variables are considered
- The MSE within 1 Standard error of the minimum MSE could be considered acceptable and there are no significant variables at this level of lambda
- As we vary the lambda, the variables which have the most significance become significant in that order
- The variables gain significance in the order given below:

  MarkDown5 > MarkDown1 = MarkDown3 > CPI = Unemployment > Fuel_Price = MarkDown2 > Temperature > IsHoliday = MarkDown4

WEEKLY SALES DATA AS A TIME SERIES MULTIPLE STORES MULTIPLE DEPARTMENTS

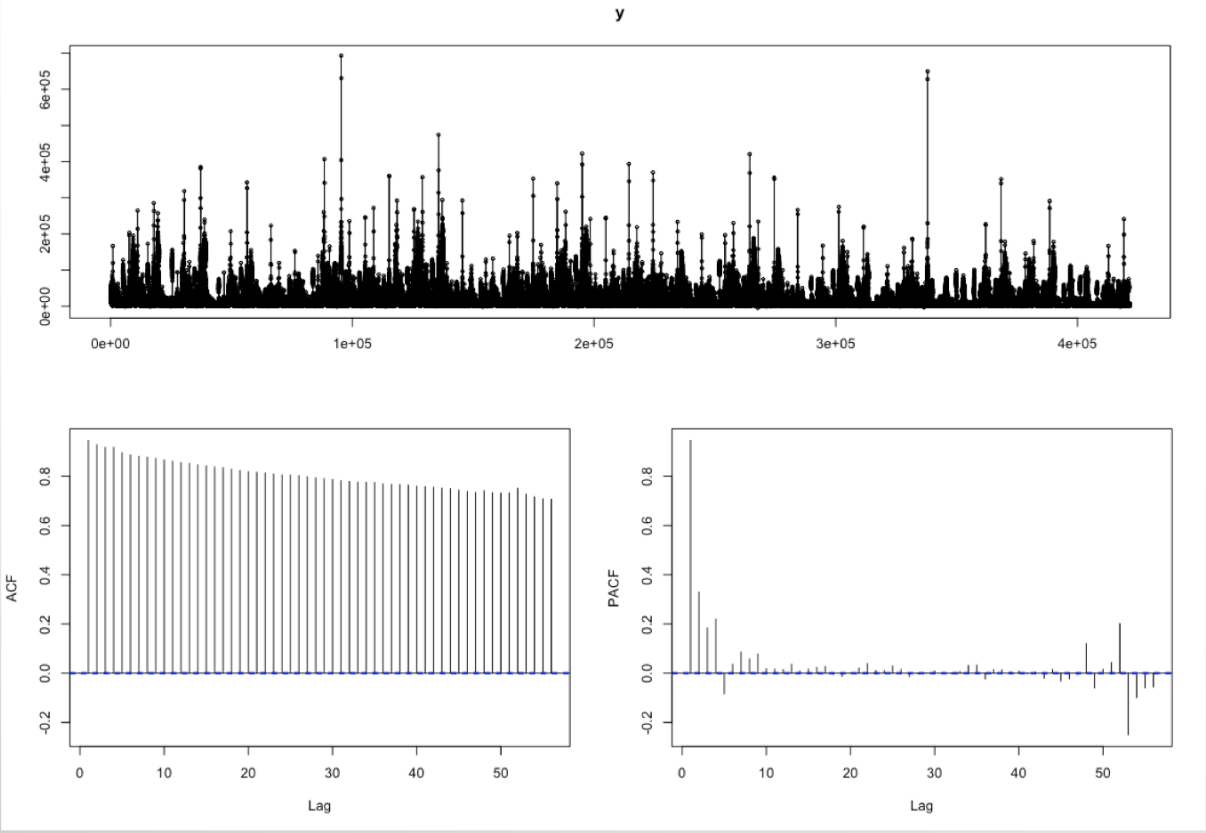


- The data is too dense to notice a strict pattern
- Decomposed plot of the time series helps to identify any underlying pattern in the data
- Multiple data points with the same time stamp makes it harder to quantify the underlying patterns

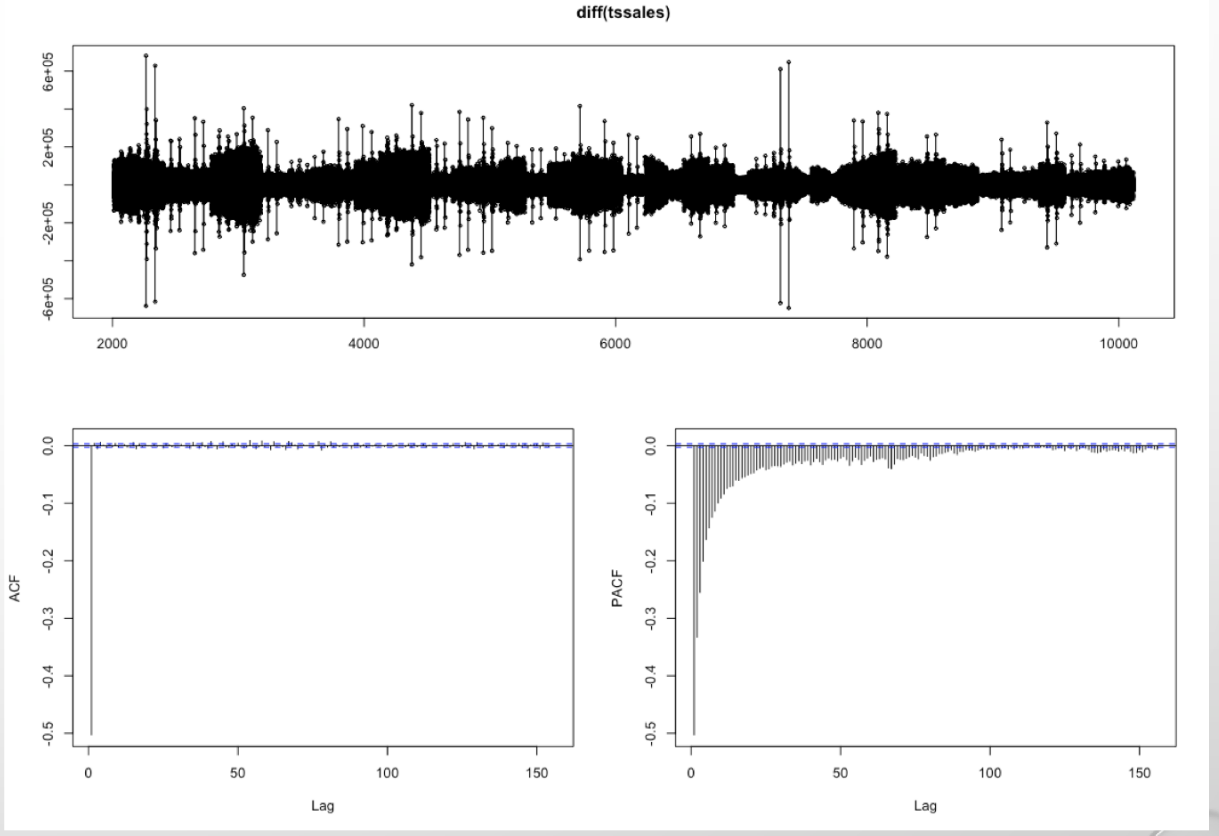


AUTO-CORRELATION AND PARTIAL AUTO-CORRELATION

WEEKLY SALES



ONCE DIFFERENCED WEEKLY SALES



The TS must be stationary in order to fit an ARIMA model, this can be tested using the Augmented Dickey-Fuller test for stationarity. When the p value <0.05, we reject the null hypothesis of non-stationarity.

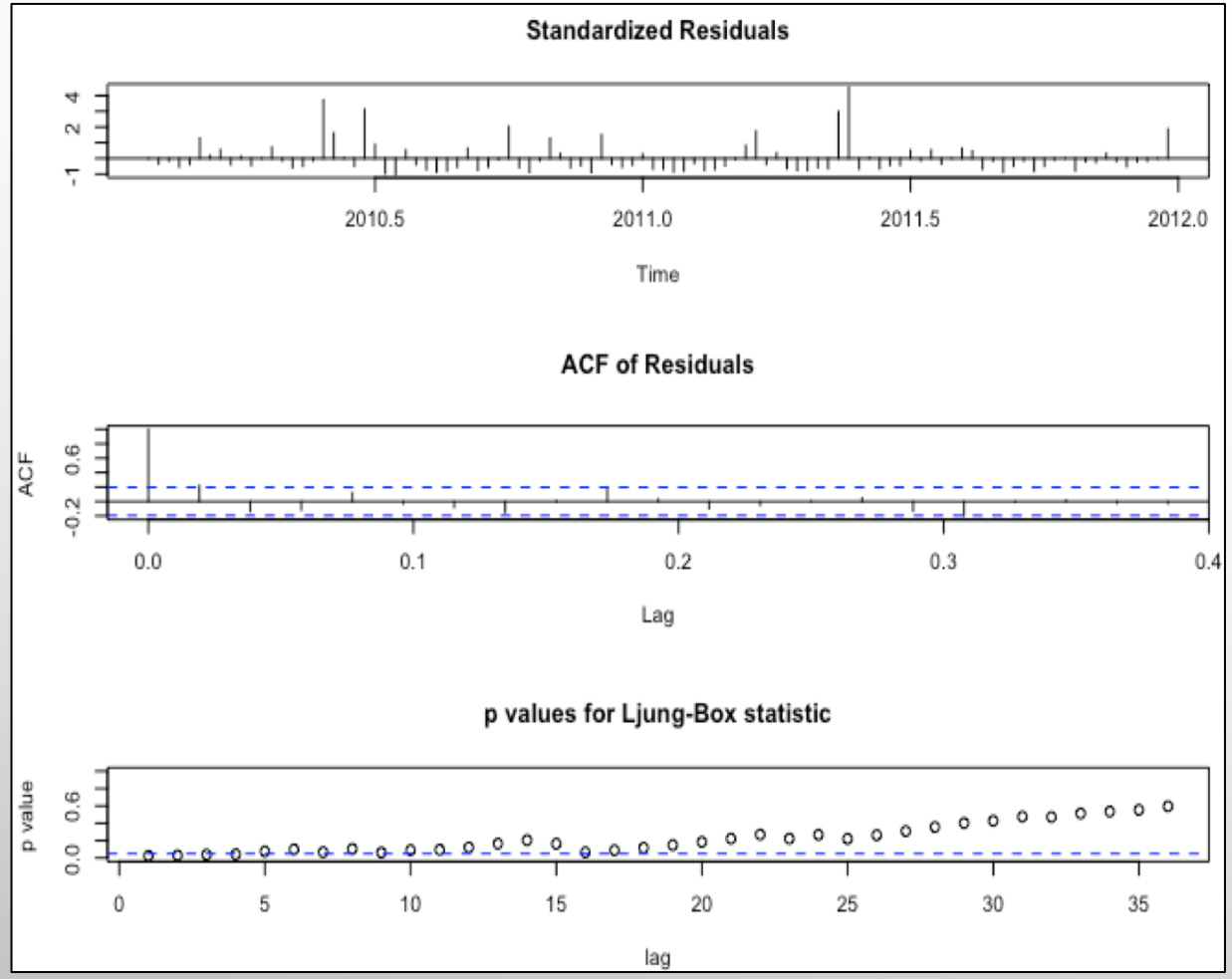
Augmented Dickey-Fuller Testdata: tssales
Dickey-Fuller = -20.575, Lag order = 74, p-value = 0.01
alternative hypothesis: stationary

ACF dies after lag 1 and PACF decays for once difference TS suggesting an ARIMA (0,1,1) model

RESIDUAL ANALYSIS

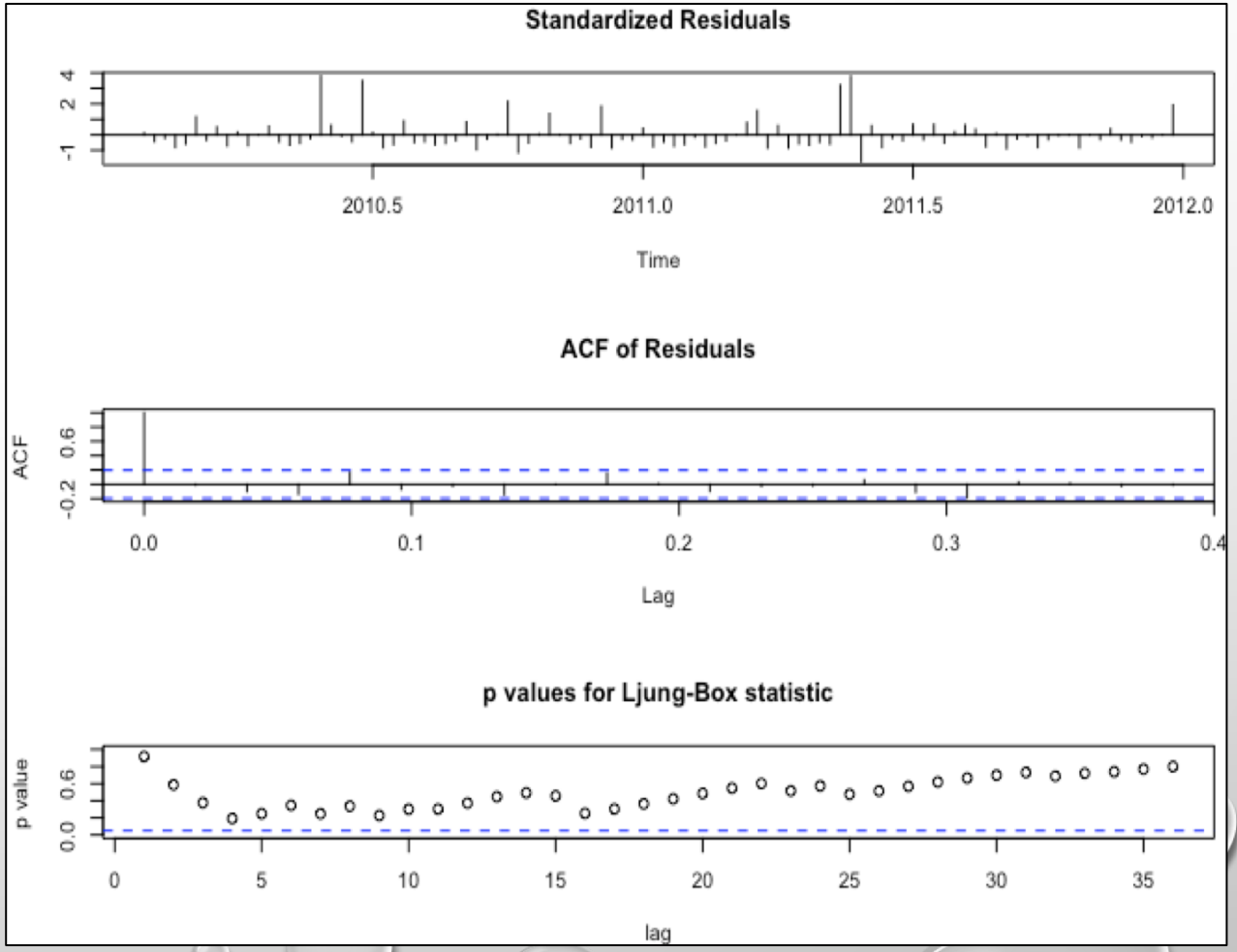
Running an auto ARIMA in R, provides the model ARIMA (0,0,1), which is expected to be the best model

ARIMA (0,1,1)



POOR RESIDUALS

ARIMA (0,0,1)



BETTER RESIDUALS

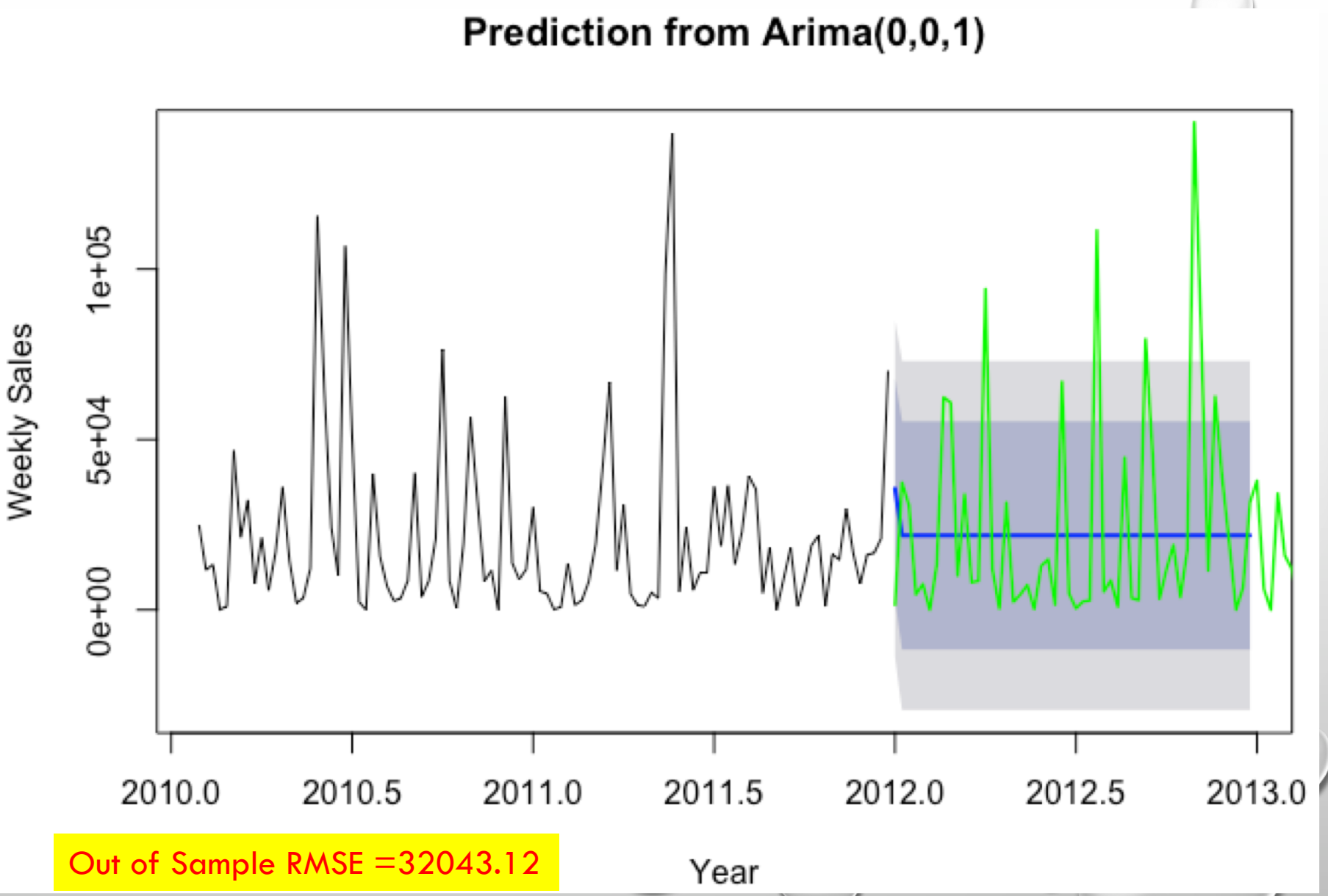
ARIMA MODEL SELECTION AND PREDICTION

COMPARING MODELS:

Fitness parameters	ARIMA (0,1,1)	ARIMA (0,0,1)
AIC	2301.46	2314.13
AICc	2301.58	2314.38
BIC	2306.65	2321.95

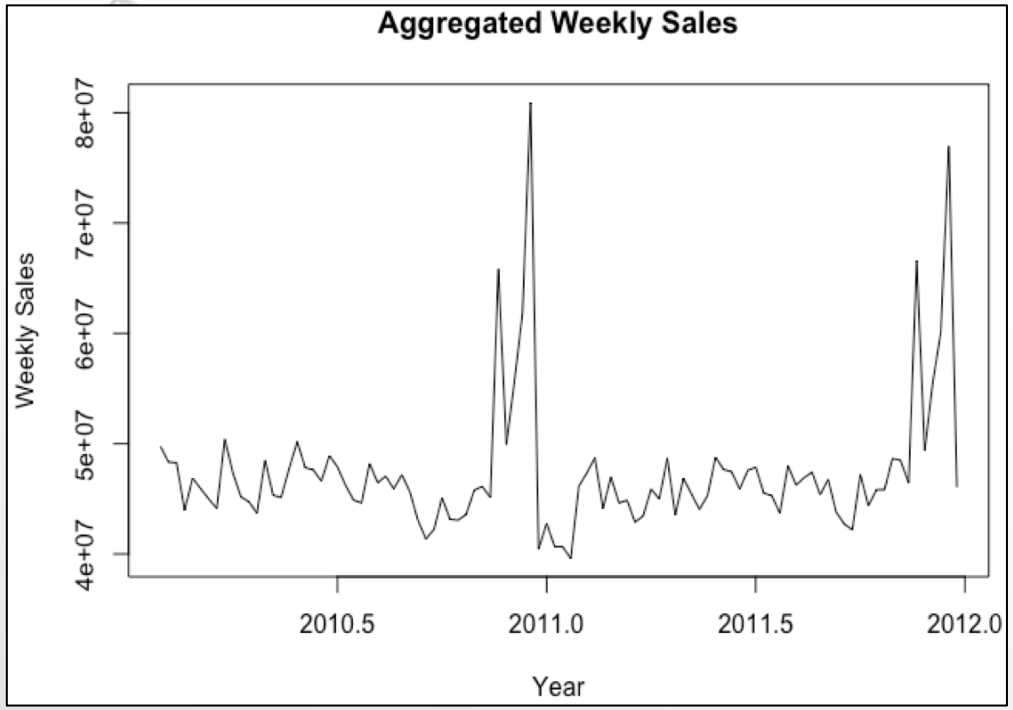
- Comparing the Fitness parameters, the ARIMA (0,1,1) looks like a better model with a lower AIC, AICc and BIC
- But the residual analysis suggest otherwise and since independence of residuals is an underlying assumption in modeling, we pick ARIMA (0,0,1)

PREDICTION IS NOT HELPFUL



Out of Sample RMSE =32043.12

AGGREGATED TOTAL WEEKLY SALES FOR ARIMA MODELING



Augmented Dickey-Fuller

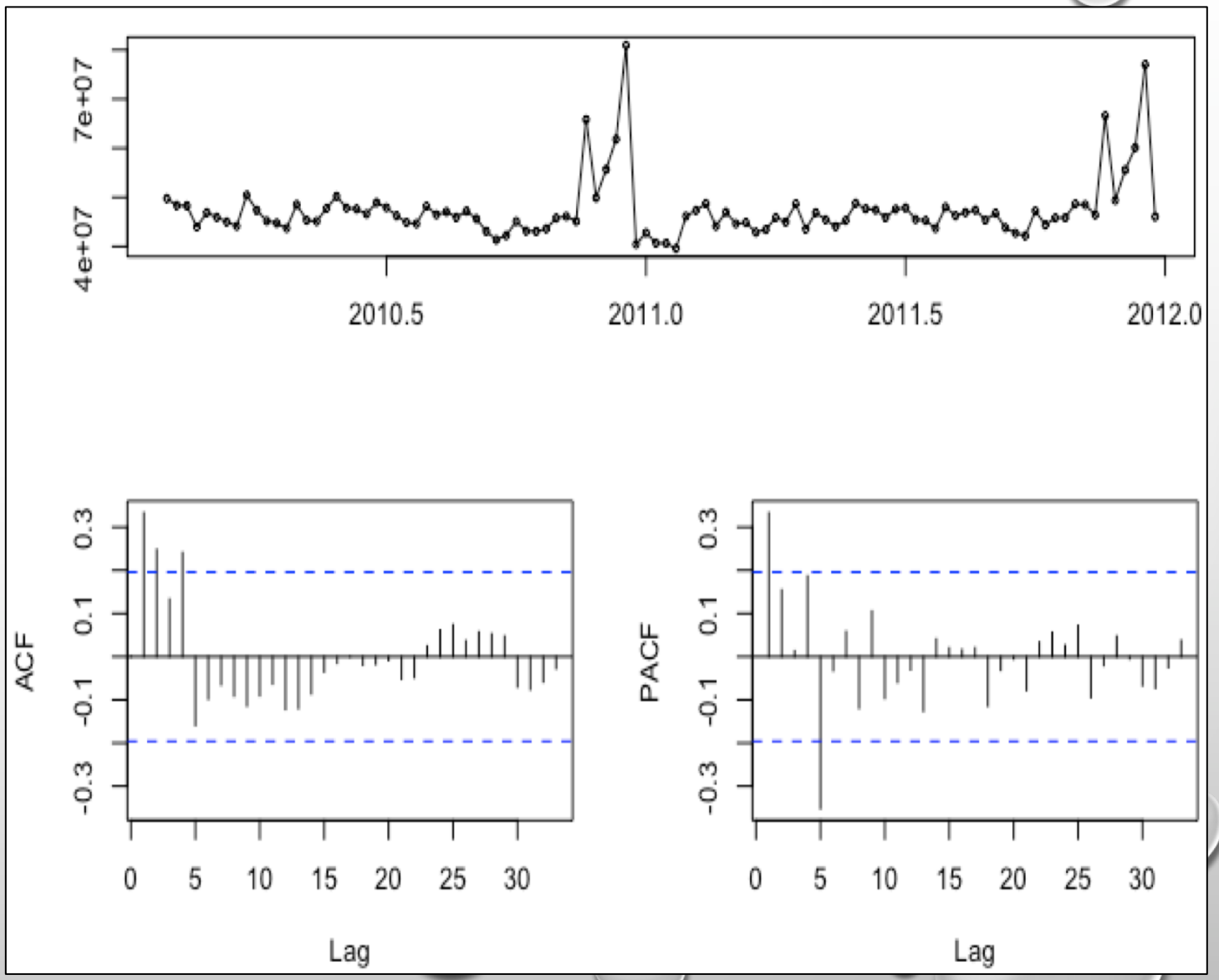
Testdata: y2.TR

Dickey-Fuller = -5.3039, Lag order = 5,

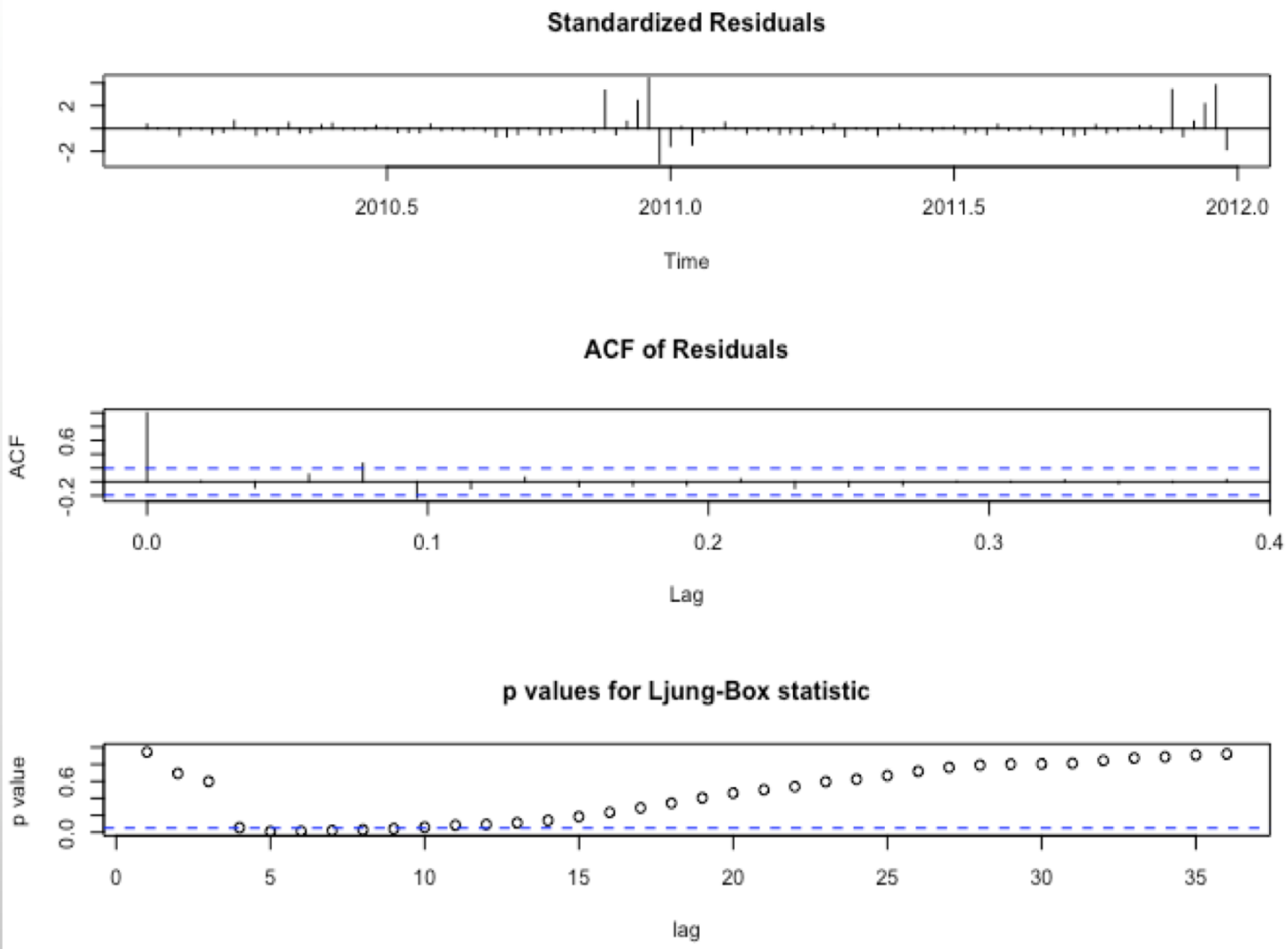
p-value = 0.01

alternative hypothesis: stationary

Stationary Time Series



RESIDUAL ANALYSIS



RESIDUALS NOT THAT GOOD BUT ACCEPTABLE

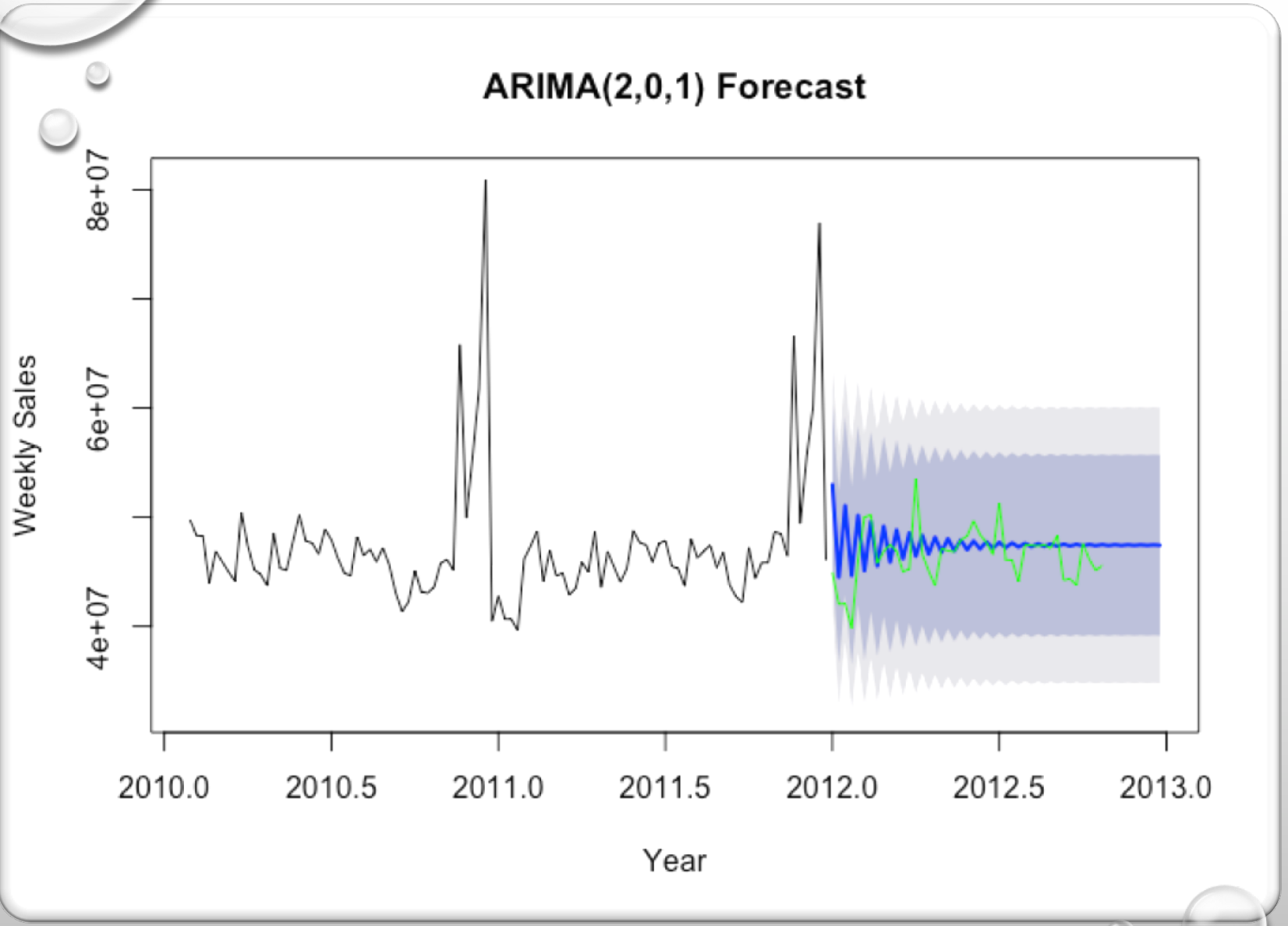
MODEL SUMMARY

Series: y2.TR
ARIMA(2,0,1) with non-zero mean
Coefficients:

	ar1	ar2	ma1	mean
	-0.4206	0.4296	0.7300	47435133.5
s.e.	0.1370	0.0963	0.1241	986343.4

sigma^2 estimated as 3.391e+13:
log likelihood=-1697.8
AIC=3405.6 AICc=3406.24 BIC=3418.63
Training set error measures:

ME	-10381.91
RMSE	3154101
MAE	5705682
MPE	-1.101533
MAPE	6.131085
MASE	2.272433
ACF1	0.006851451



ARIMA MODEL SELECTION AND PREDICTION

MODEL SELECTED:

ARIMA (2,0,1)

ARIMA EQUATION:

$$(1 - \phi_1 B + \phi_2 B^2) Y_t = (1 + \theta_1 B) e_t$$

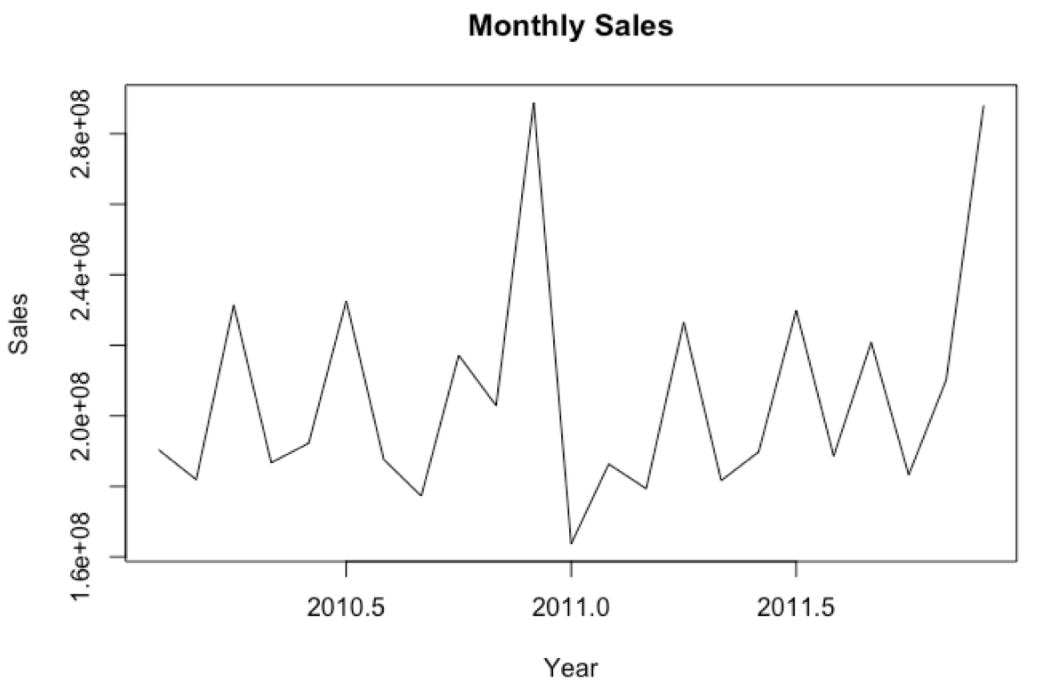
Out of sample RMSE =

$$\sqrt{\frac{\sum_{i=1}^{n} (prediction - actual)^2}{n}} = 4655633$$

PREDICTIONS BETTER THAN THE BASE MODEL

ETS MODELING

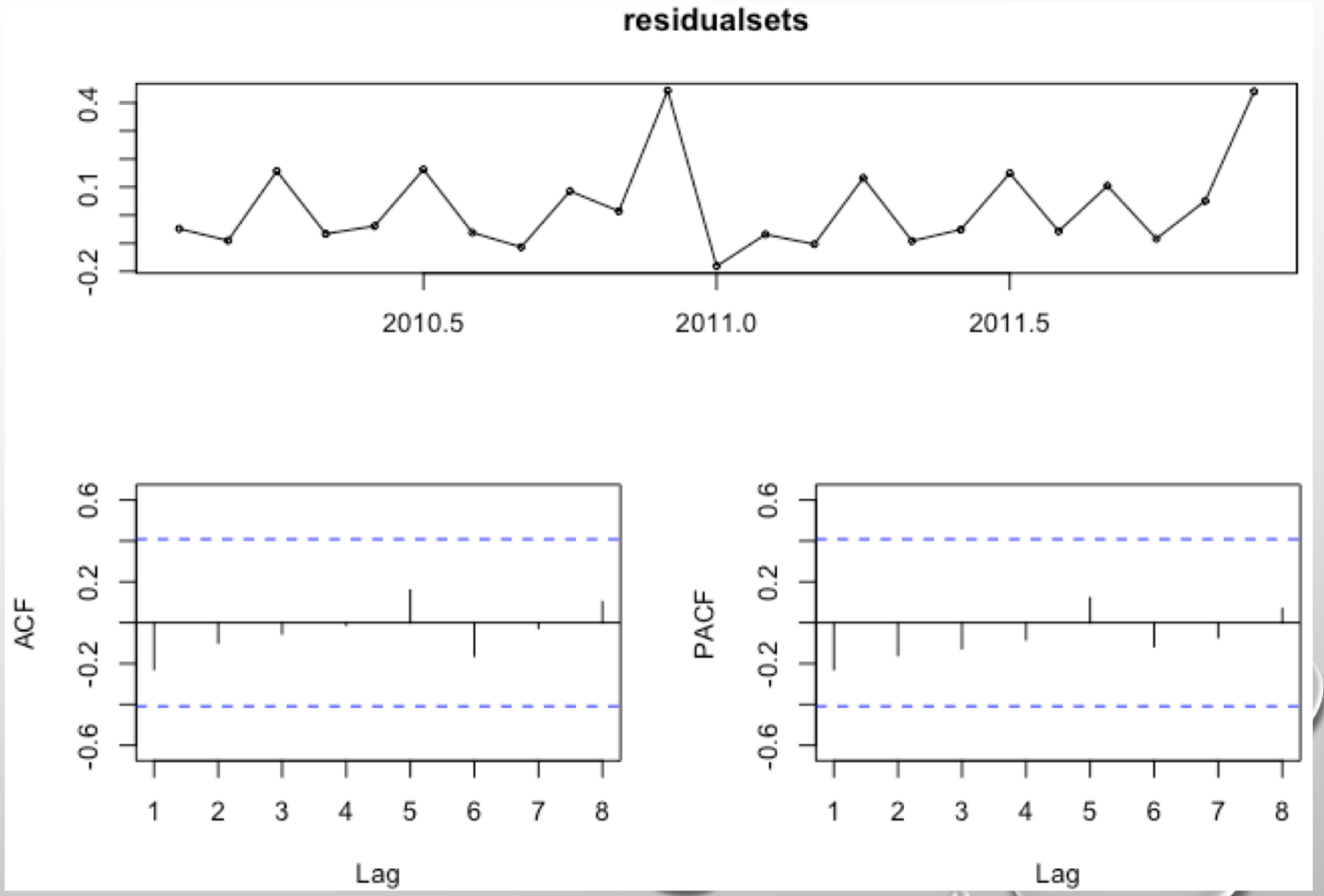
- Fitting an ETS on the weekly aggregated data is not a feasible step as ets model in R is designed to handle frequencies up to 24.
- Aggregating the data on a monthly level, we get the following:

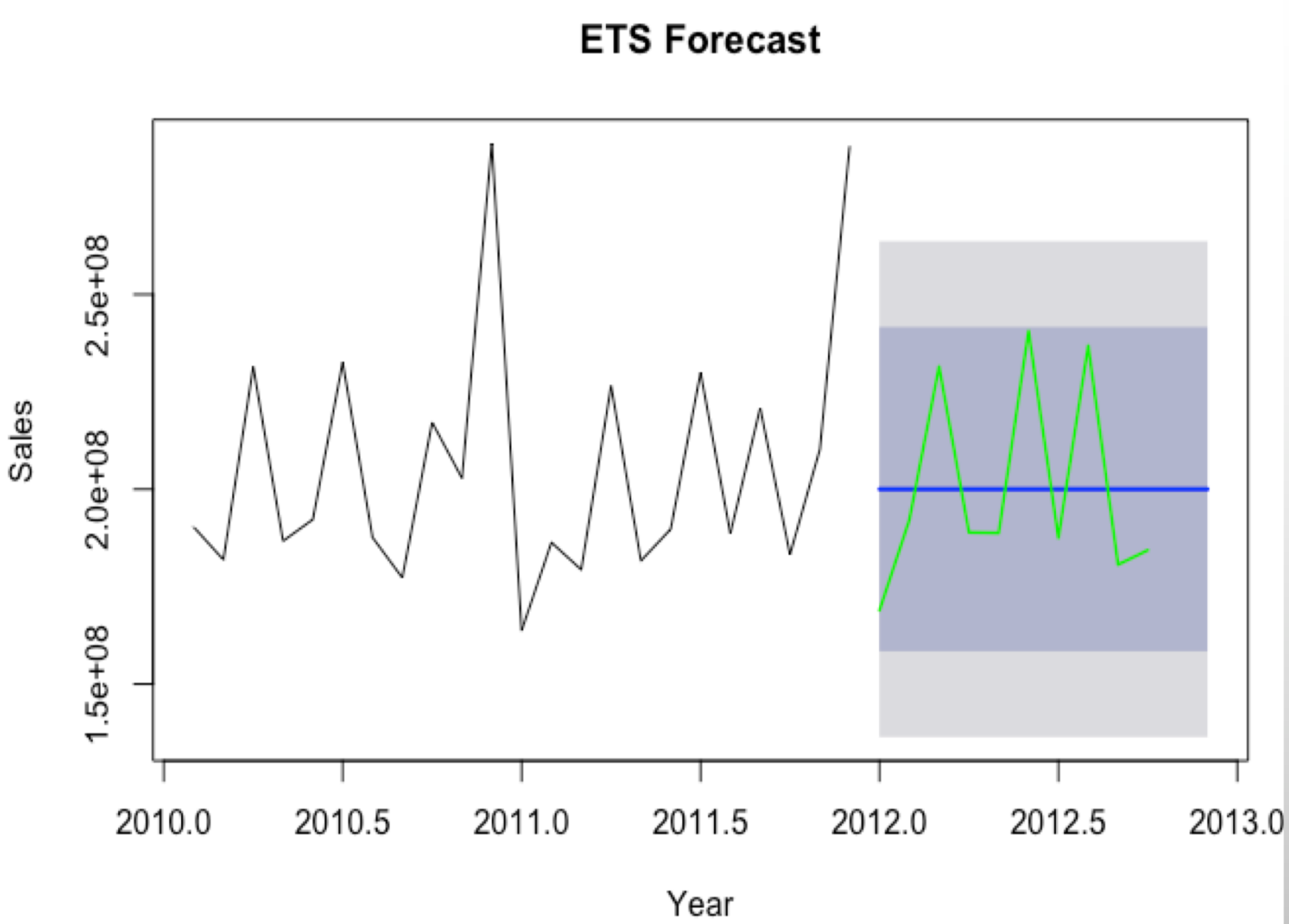


AIC	AICc	BIC
873.70	874.97	877.11

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5945779	32456610	2432899	0.859	11.06	2.476	-0.229

RESIDUAL ANALYSIS





Out of Sample RMSE = 24606260

ETS MODEL SELECTION AND PREDICTION

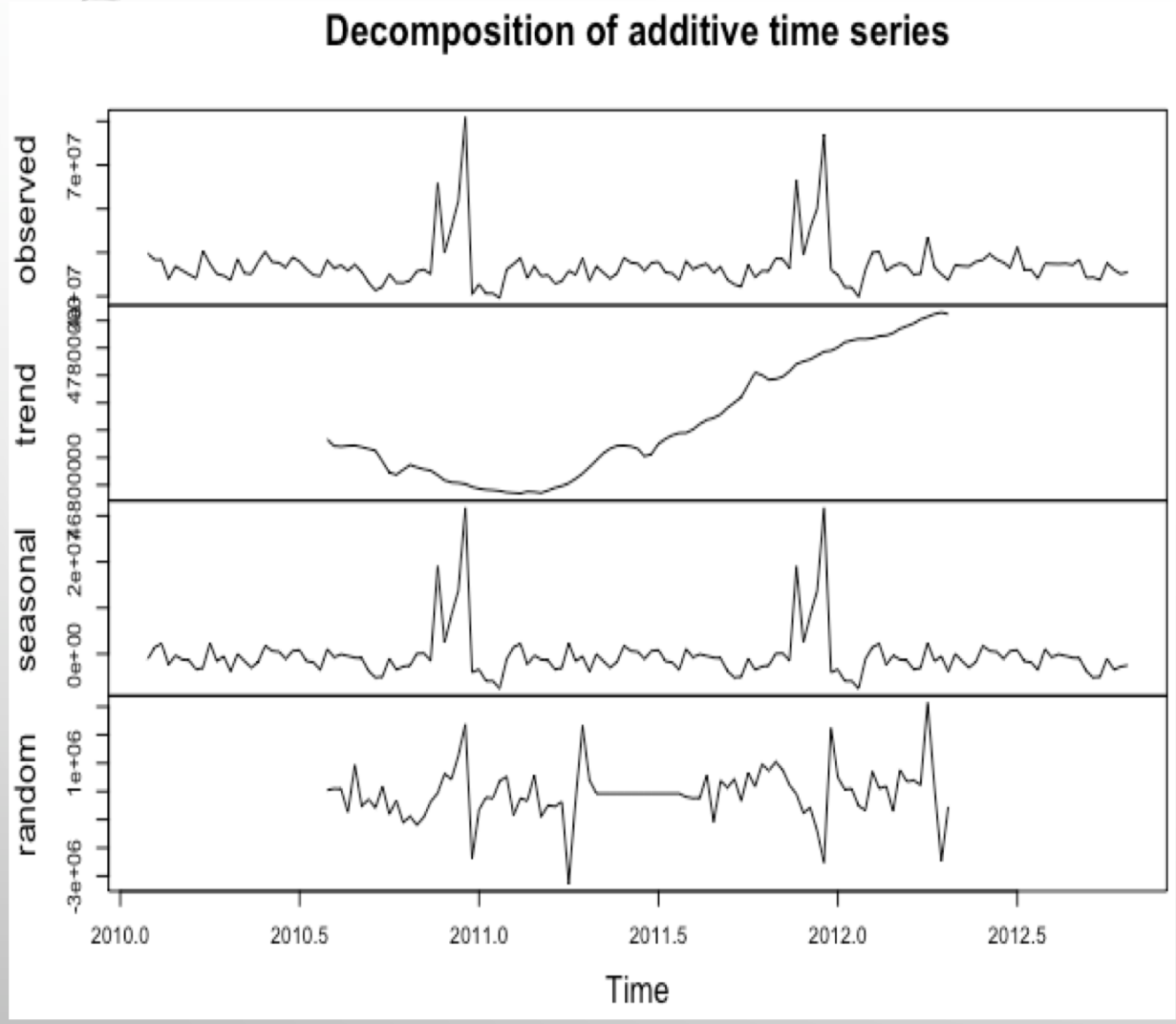
- MODEL SELECTED:

ETS(M,N,N)

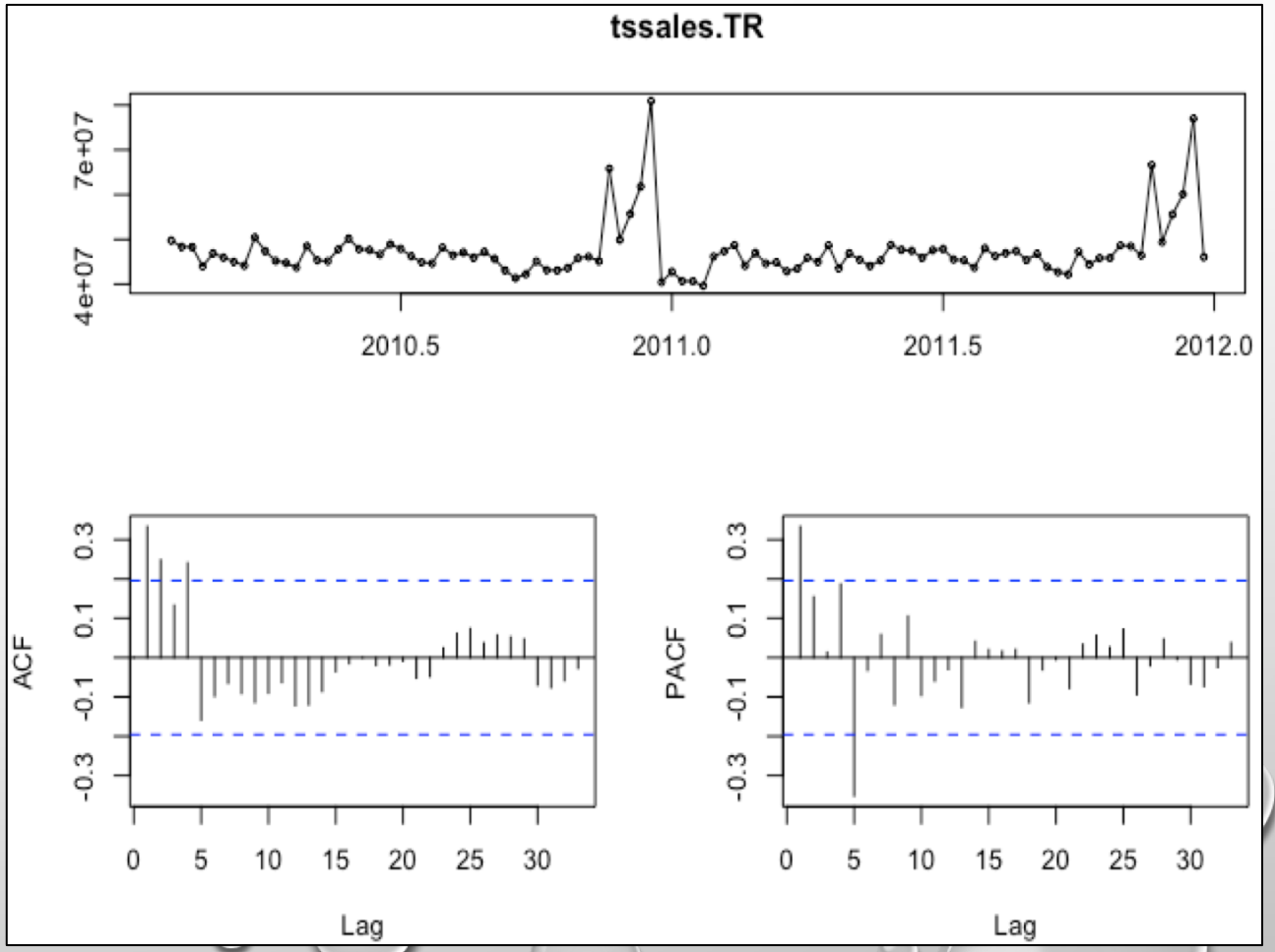
- Comparing the Fitness parameters, the ETS looks like a better model with a lower AIC, AICc and BIC
- But the out of sample prediction precision (RMSE) suggests otherwise and hence we can conclude that an **ARIMA model is a better fit**

DYNAMIC REGRESSION MODELING: DATA

- DYNAMIC REGRESSION corresponds to running a normal regression model with ARIMA errors



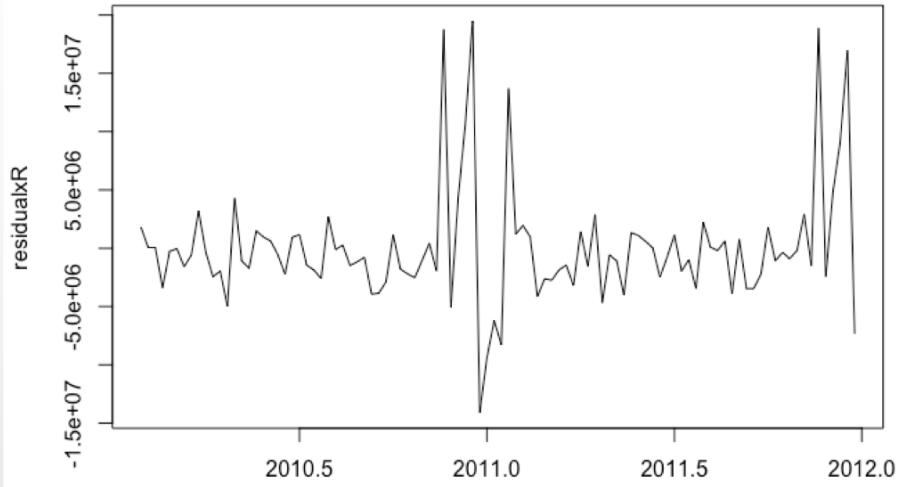
ACF and PACF of the TS

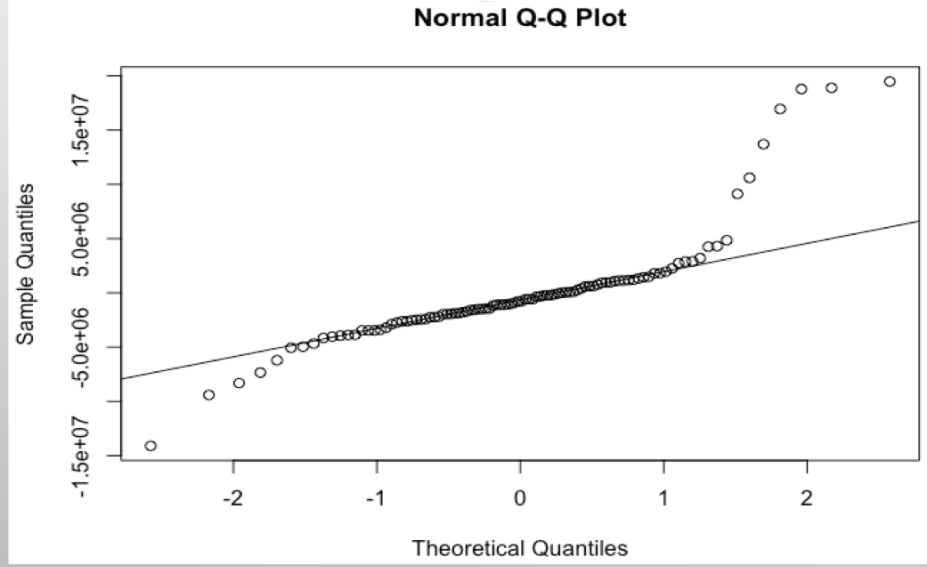


DYNAMIC REGRESSION RESIDUAL PLOTS

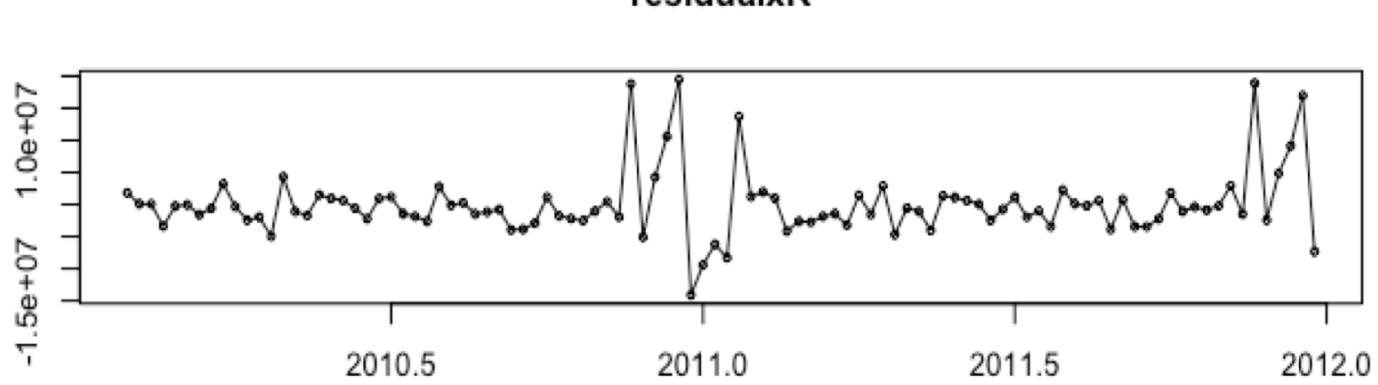
The variables suggested by LASSO are added on by one to see how the forecast improves

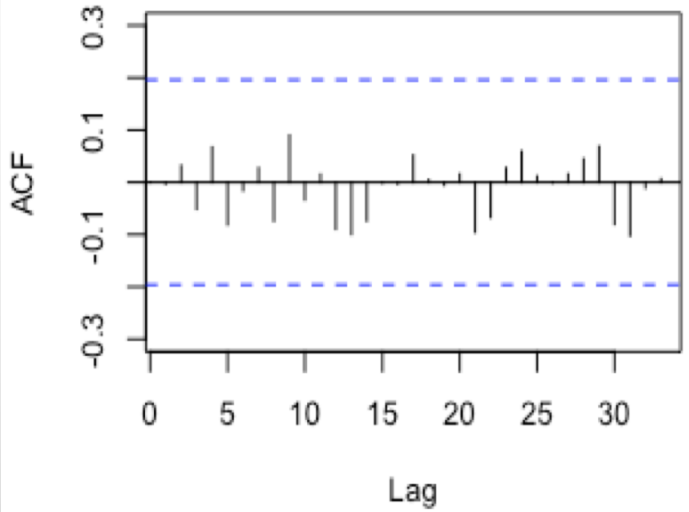
1. MarkDown5

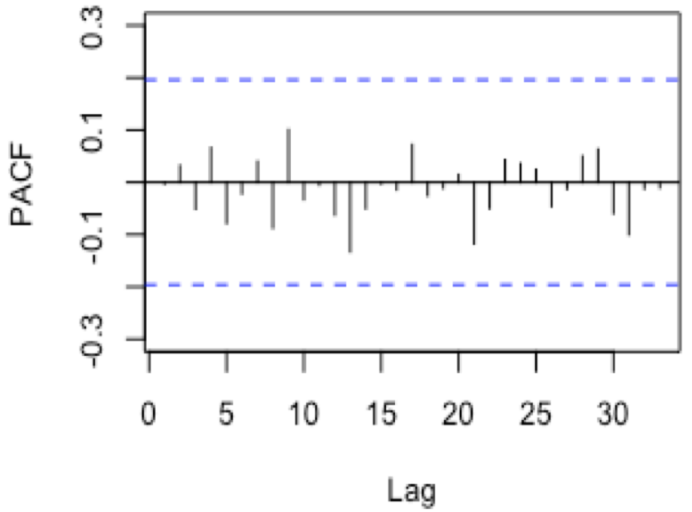


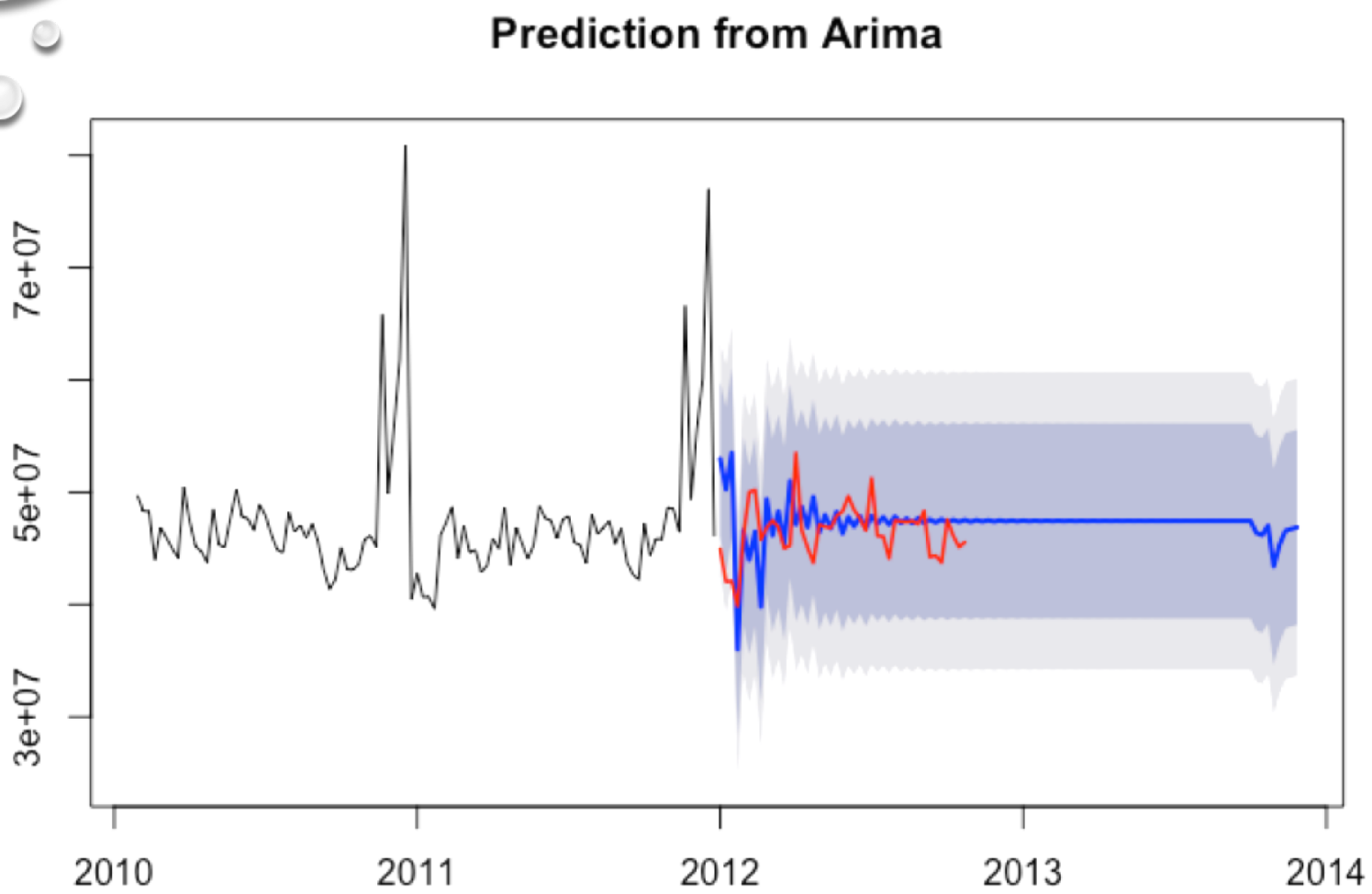


Residual Analysis









Out of sample RMSE = 3649023

DYNAMIC REGRESSION MODEL SELECTION AND PREDICTION

MODEL SELECTED:

REGRESSION WITH ARIMA(2,0,1) ERRORS

EQUATION:

$$Sales = 47556458 - 217.5031 * MarkDown5 + n_t$$

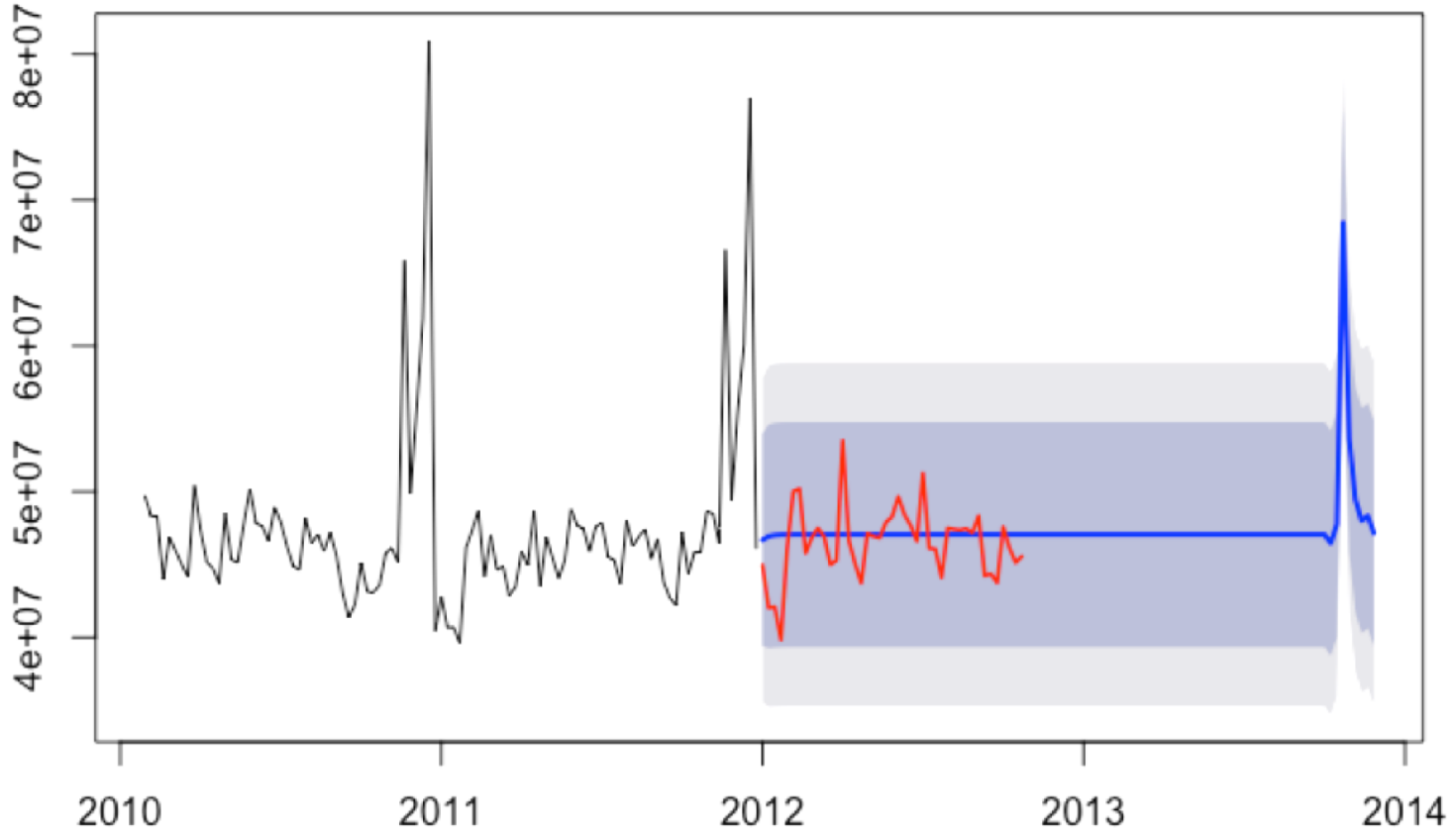
WHERE,

$$(1 + \phi_1 B + \phi_2 B^2) n_t = (1 - \theta_1 B) e_t$$

$$\phi_1 = -0.3768,\ \phi_2 = 0.4658,\ \theta_1 = 0.7226$$

BETTER PREDICTIONS THAN THE ARIMA MODEL

Variables added on by one to see how the forecast improves
1. MarkDown5 2. MarkDown1 3. MarkDown3



Out of sample RMSE = 2478771

DYNAMIC REGRESSION MODEL SELECTION AND PREDICTION

MODEL SELECTED:

REGRESSION WITH ARIMA(1 ,0,0) ERRORS

EQUATION:

$$Sales = 47051536 - 203.6625 * MarkDown1 + 367.3172 * MarkDown3 + 329.4788 * MarkDown5 + n_t$$

WHERE,

$$(1 + \phi_1 B) n_t = e_t$$

$$\phi_1 = 0.3362$$

PREDICTIONS DON'T SEEM TO IMPROVE

FUTURE WORK

- USE KNN TO PREDICT THE NA VARIABLES
- RUN A LASSO ON THE PATCHED FEATURE DATA
- RUN THE DYNAMIC REGRESSION MODEL WITH THE PATCHED VARIABLES
- ANALYZE IF THIS WILL PROVIDE A BETTER FORECAST