IEE 578

Regression Analysis

Regression analysis of sale prices of a wholesale store using time series analysis.

Report by.

Individual Contribution:
Arvind Ramkumar.
Swapnil Gharat.
Sajil Odanath.

Executive Summary

This report provides an analysis and evaluation of the sale prices for the wholesale store Walmart. We are performing regression analysis to understand the trend between sale prices and the corresponding factors based on the previous research and identifying the significant regressor factors and to check for the possible interaction between the same. This report includes various regression techniques in order to achieve a satisfactory model to predict the sale prices in the future and take influential management decisions in order to increase the same. Various Regression techniques such as Multiple Regression, Model Accuracy Checking, diagnostics for outliers or influential points, indicator variables, handling multicollinearity, performing variable selection and model building and validation of the same are covered in this report. Additionally, this data being a time series data, we had to check for possible Autocorrelation in the model. Autocorrelation can be treated by addition of predictor variables or implementing various parameter estimation methods such as Cochrane-Orcutt method. We have implemented Cochrane-Orcutt method to eliminate Autocorrelation and make the data set Uncorrelated. We successfully uncorrelated the data and then performed the further statistical analysis for model validation and checked for outliers or influential points in the data set. We increased the accuracy of the model after eliminating these observations. Then we performed further analysis by fitting the full factorial model and then eliminating the insignificant interaction terms and co-efficient that led to increase in the variance. We have eliminated possible multicollinearity ill-conditioning. After all the statistical analysis we finalized the model with adequate accuracy. Based on the final model and analysis run we have mentioned certain recommendations and possible ways to increase the accuracy of the model at the end of this report. Henceforth, we have performed all the possible statistical analysis and transformations in order to obtain a satisfactory model and has been implemented stepwise throughout the report.

Appendix

Regression analysis of sale prices of a wholesale store using time series analysis.

Group members:

Sajil Odanath, Swapnil Gharat, Arvind Ramkumar.

Abstract:

Increasing Sale Prices is one of the major goals of any company or store. Increasing Sale Prices involves optimizing the available resources, and managing fuel prices, unemployment, managing employee holidays etc. The purpose of this report is to establish a relationship between the sale prices and the influential factors to take profitable managerial decisions. The paper analyses factors and their degree of correlation on sales prices using multiple regression. This experimentation consists of a historical sales data for Walmart stores located in different regions. Objective here is predicting the sales for each store.

Keywords: Sale prices, efficiency, regression analysis, variables selection, correlation, time series analysis.

1.Introduction:

Regression analysis is a statistical technique for studying linear relationships. Time series is one of the major application of regression analysis, i.e. variables are mostly time-oriented. The errors are correlated with themselves at different time periods. It also consists of similarity between the previous observations as a function of time. Presence of Auto-correlation highly affects the Least Square Estimates. It also affects the confidence intervals and the prediction intervals as the presence of auto Correlation makes them shorter. Henceforth, the standard errors or the regression coefficients obtained from such model might me quite misleading.

Auto-correlation can be approached by addition of one or more predictor variable, but if this doesn't work then Autocorrelation can be approached by methods as explained in this report.

2.Methodology:

Participants/Data Set:

This is a historical Data Set collected for the two Walmart Stores from 2010-2013. The data set consist of 286 observations. The data set consist of regressor variables like Temperature, fuel prices, CPI, Unemployment, Holiday and Store size. All of this data is continuous. Sale Price is the response variable. These are the variables based on the recommendations from the past research. This is a time series data i.e. the data is highly dependent on the previous data. So, the methodology involves creating a regression model based on the given Data set and perform statistical analysis. Check for Autocorrelation using the Durbin-Watson Test. Since there is no other predictor variable to add we perform the transformations based on the Cochrane-Orcutt Approach. Fit model by eliminating outliers or influential points. Determine significant regressors and interaction factors using full factorial analysis. Compare the models based on applied transformations. Make recommendations after performing numerical experiments on the final model.

Step1. Fitting the initial model with all the regressor variables using Ordinary Least Squares method.

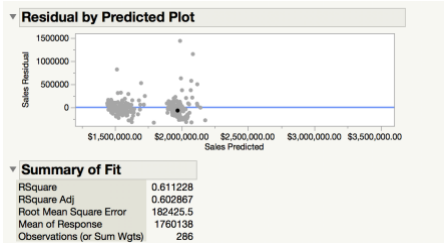


Figure 1.1

We can see from the figure 1.1 that the R-Square for the fit is 61.1%. This is quite adequate but since this is a time series data we need to check for Autocorrelation.

Residual Analysis.
Plotting the Residual vs Time Plot:

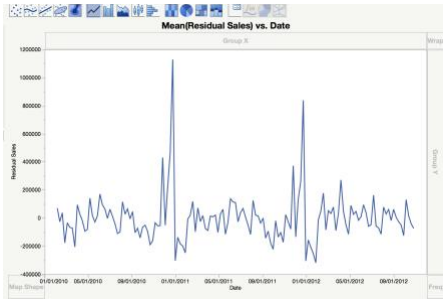
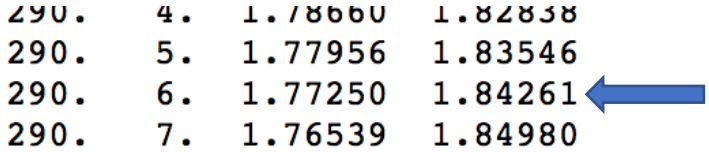


Figure 1.2

We observe potential Auto-Correlation in figure 1.2
We need to perform Durbin-Watson test to check for Auto-Correlation.

Step2.Checking for any Auto-Correlation by Durbin-Watson Test.

(a)



(b)

Figure 2.1.

We can see that d=1.60 form Figure 2.1(a). Figure 2.1(b) shows Upper limit and lower limit of d for given number of observations.
d < d_l ,Hence we Reject $H_0 : \rho = 0$, i.e. We conclude that this time series data exhibits Auto-correlation.

Step 3. The Cochrane-Orcutt Method.

To handle Auto-correlation problem we use The Cochrane-Orcutt Approach:

Since the observed correlation in the model cannot be removed by adding predictor variables we need to use appropriate parameter estimation method. We need to transform the response variable as well as the regressor variable.

$y_t' = y_t - \emptyset * y_{t-1}$

$x_t' = x_t - \emptyset * x_{t-1}$

After this transformation we need to apply ordinary least squares and again check for the Durbin-Watson Test. If the positive Autocorrelation is still indicated, we need to apply this transformation again until we the transformed model is uncorrelated.

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation
1.9808491	284	0.0094

Figure 3.1 (a)

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation
1.999617	282	0.0000

Figure 3.1 (b)

Transformations and Performing Durbin-Watson Test.

Transformation 1:

Performing Durbin-Watson test after Transformation 1.

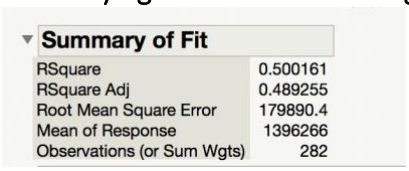
From Figure 3.1(a) we can observe that the Autocorrelation factor ∅ is still not equivalent to zero hence we need to perform one more transformation of response and regressor variables using new estimates.

Transformation 2:

Performing Durbin-Watson test after Transformation 2.

From Figure 3.1(b) we can observe that the Autocorrelation factor ∅ is still equivalent to zero hence we have successfully uncorrelated the data.

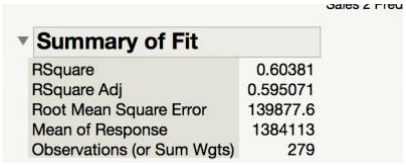
Step 4. Identifying influential or leverage points/observations.



Summary of Fit

RSquare	0.500161
RSquare Adj	0.489255
Root Mean Square Error	179890.4
Mean of Response	1396266
Observations (or Sum Wgts)	282

Figure 4.1(a)



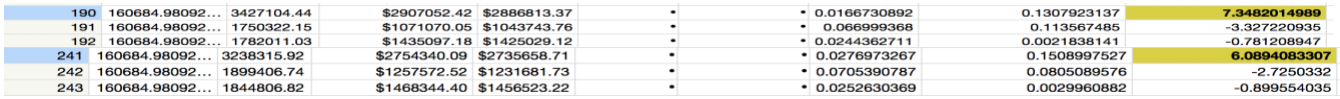
Summary of Fit

RSquare	0.60381
RSquare Adj	0.595071
Root Mean Square Error	139877.6
Mean of Response	1384113
Observations (or Sum Wgts)	279

Figure 4.1(b)

Figure 4.1: a) with all observations. b) with observation 190 and 241 eliminated.

Now that we have successfully uncorrelated the data now we can do further statistical analysis for variable selection and then deriving the final model. But first we need to check for any possible outliers or influential observations in the model.



190	160684.98092...	3427104.44	$2907052.42	$2886813.37	•	•	0.0166730892	0.1307923137	**7.3482014989**
191	160684.98092...	1750322.15	$1071070.05	$1043743.76	•	•	0.066999368	0.113567485	-3.327220935
192	160684.98092...	1782011.03	$1435097.18	$1425029.12	•	•	0.0244362711	0.0021838141	-0.781208947
241	160684.98092...	3238315.92	$2754340.09	$2735658.71	•	•	0.0276973267	0.1508997527	**6.0894083307**
242	160684.98092...	1899406.74	$1257572.52	$1231681.73	•	•	0.0705390787	0.0805089576	-2.7250332
243	160684.98092...	1844806.82	$1468344.40	$1456523.22	•	•	0.0252630369	0.0029960882	-0.899554035

As observed from studentized residuals column the value is 7.34 for observation 190 and 6.08 for 241 which is sufficiently high as compared to other observations. Eliminating these observations and again fitting the model gives us R-Square value of 60.38%(Figure 4.1(b)) and while including all the observation R-Square corresponds to 50.01%(Figure 4.1(a)). Hence, eliminating this observation makes the fit more adequate.

Step 5.

Final model including significant interaction factors.

The initial model building technique that was employed was the stepwise regression. The Forward and Backward regression with the Minimum BIC as the stopping rule, Store Size and Store Size, Temperature were significant respectively. But it failed to meet the Cp Condition.

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
9.401e+12	280	183235.96	0.4720	0.4701	12.51168	2	7639.211	7650.051

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	199174.026	1	0	0.000	1
☐	☐	Temp 2	0	1	1.91e+11	5.789	0.01678
☐	☐	Fuel 2	0	1	2.759e+8	0.008	0.92797
☐	☐	CPI 2	0	1	7.88e+10	2.359	0.12573
☐	☐	Unemployment 2	0	1	1.65e+10	0.490	0.48447
☐	☐	IsHoliday	0	1	8.51e+10	2.547	0.11162
☐	☑	Store Size 2	8.52420378	1	8.4e+12	250.270	1e-40

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
9.21e+12	279	181688.76	0.4827	0.4790	8.6062818	3	7635.478	7649.901

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	320035.342	1	0	0.000	1
☐	☑	Temp 2	-2205.9065	1	1.91e+11	5.789	0.01678
☐	☐	Fuel 2	0	1	3.545e+9	0.107	0.74378
☐	☐	CPI 2	0	1	1.02e+11	3.106	0.07909
☐	☐	Unemployment 2	0	1	3.27e+10	0.991	0.32032
☐	☐	IsHoliday	0	1	4.92e+10	1.492	0.223
☐	☑	Store Size 2	8.52094338	1	8.4e+12	254.354	3.8e-41

So, alternatively, we fit the full factorial model and using the Sorted parameter estimates, we remove the insignificant parameters based on the p value, and we found the following parameters to be significant. The VIF does also have a very small value indicating that there is no Multicollinearity either.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-678592.8	445547.7	-1.52	0.1289	.
CPI 2	5364.5726	2559.026	2.10	0.0370*	1.0019385
IsHoliday	102284.09	37477.81	2.73	0.0068*	1.0907336
(Temp 2-54.6989)*(Unemployment 2-6.03949)*(IsHoliday-0.06429)	13979.07	7042.852	1.98	0.0482*	1.0911478
Store Size 2	8.1197747	0.434968	18.67	<.0001*	1.0015616

Figure 5.1(a)

Summary of Fit

RSquare	0.56652
RSquare Adj	0.560215
Root Mean Square Error	147272.9
Mean of Response	1386159
Observations (or Sum Wgts)	280

Figure 5.1(b)

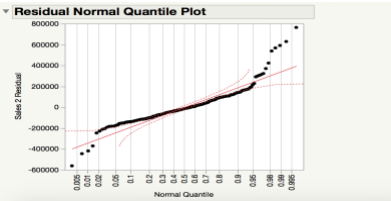


Figure 5.1(c)

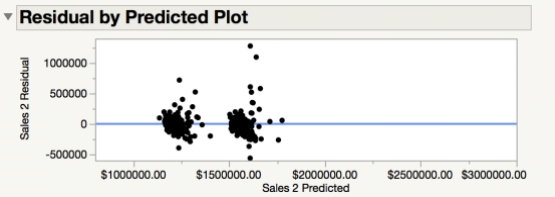


Figure 5.1(d)

3.Conclusion:

So, the final model including the significant interaction factors is as shown in figure 5.1. Model Accuracy is 56.652%. Also, checking through normal probability plot the uncorrelated data set is normally distributed. Also, the p-value as well as the VIF's are well within the range. We have achieved a satisfactory model. Adequacy check for the model using residual plot as well as normal probability plot shows that the variance is equal and well as well within range.

The Final Model:

Sales=-678592+5364.57 *CPI+ 102284*IsHoliday+ 13979.07* (Temp 2-54.69) *(Unemployment-6.03) *(IsHoliday-0.06).

4.Recommendations:

We can increase the accuracy of the model by adding the predictor variables or by increasing the number of observations by considering data for more number of stores. This recommendation is based on the observed model testing by considering Store 1 as well as data for Store 1 and 2 both together.