Data Analysis and Regression

Final Report

Name: Yi-Ching Tsai

Course: Data Analysis and Regression

Date: 7/16/2024

Introduction

The goal of my analysis is to use dataset (market_sale.csv) to explore whether the usage of marketing campaigns and promotional methods for the product have an effect in increasing the sales of the product. There are 8 variables in my dataset which are:

Sale - representing the number of product sales for each observed week.

Price - The observed week's base price for the product.

Radio - The number of radio advertisements or campaigns promoting the product for the observed week.

InStrSpending - The average expenses associated with promoting the product in stores for the observed week.

Discount - The discount rate applicable for the observed week.

TVSpending - The average expenditure on television campaigns during the observed week.

StockRate - The stock-out rate, calculated as the number of times the product was out of stock divided by the total number of product visits.

OnlineAdsSpending - The online ads spending, calculated the total amount of spend on online advertising.

The goal is to use the regression model based on these variables to make an accurate prediction on the dependent variable Y which is Sale variable.

This analysis is helpful in determining which methods are useful in increasing the sale and which are not, by doing this, the company will be able to choose the best methods and save further expenditures that are not necessary.

Analysis

When we look into the sale's breakdown, we could see that Descriptives table (A.1) shows that 25% of sale has value of 112,404.5 dollars while 50% of sale has value of 170,390.5 dollars and 75% of sale has value of 226,087.5 dollars. Maximum sale value is 393,914 dollars and minimum sale value is 1,992 dollars. The IQR (Q3-Q1) helps us to measure how spread the data is distributed, the IQR is 113,683 for sale. When compared this number to the median, we see this is not that large, so we could conclude that there is no large discrepancy among the sale. The Histogram (A.2) suggests that the data is slightly right-skewed. Comparing the mean and the median supports this suggestion because they are close together. And since it is symmetric, we could know that most values of sale are clustered near the mean which is 171327.12 US dollars. As it is symmetric, so it does not have any outliers. This histogram is unimodal since it has only one peak at 180,000 which means most of the sale values are going to be 180,000 dollars. The min value for this graph is 1,992 dollars and the max value is 393,914 dollars, meaning that the range for sale is very large. There might be several reasons contribute to this phenomenon like different amount of money spending on the ads, different base price for the product and so on.

Scatterplots (A.3) and Pearson Correlation Matrix (A.4) show that sale has a positive medium and linear correlation with both instrspending and tvspending variables (correlation values for these two variables are 0.57771 and 0.41333 respectively) but has a negative high and linear correlation with price (correlation value is -0.67337). The other variables which are discount, stockrate, radio and onlineadsspending all do not have seem to be linear association with sale, so we have to take a deeper research to look if they still need to be contained in the analysis process.

The hypothesis for this model is that all the predictors are significant in helping improve the sale amount, we will use the overall goodness of fit test to test whether my alternative hypothesis is correct.

First, we assume that: Ho (null hypothesis) is $\beta j=0$ and Ha (alternative hypothesis) is $\beta j\neq 0$. F-value (A.5) is 94772.8 and P-value (A.5) associated with F-statistic is very small (less than $\alpha=0.05$). Therefore, we can reject the null hypothesis that x-variables have no effect on sale. Overall goodness of fit test suggests that there is at least one predictor that has a significant effect on sale.

Regarding the boxplot output, since all of my variables are not categorical variables, so we will not be able to create boxplots. Because boxplots require categorical variables as an x-axis variable to be used in the code statement, so there is zero probability that we can create boxplots to do my analysis. But I have previously used descriptives table to describe the five-number summary of the data which provides the exact information as boxplot.

For the next step, we need to fit the full model in order to check for multicollinearity, outliers and influential points and if one of them seem to arise, we need to fix them. After running the code in SAS, we came up with the full model which is denoted as (A.6):
sale=125379+2875.05285*instrspending+3918.55360*discount+588.99058*tvspending-13573*stockrate-
6479.54605*price+0.06234*radio+0.06712*onlinadsspending.

From (A.6), we also know that instrspending has standardized estimate value of 0.61787 which means it is the most influential predictor to predict Y while price with standardized estimate value of -0.69379 has the least effect on Y. However, we came across the issues that there are several outliers and influential points on this model(A.7), so we need to remove them from the model and see if adj-r2 has improved. But we do not see any issues with regard to collinearity as there are no predictors with VIF value larger than 10 on the full model. After removing the outliers and influential points, we have seen that the original adj-r2 (A.8) improved from 0.9985 to 0.9987 (A.9), meaning that the model has slightly improved. Adj-r2 also tells us that 99.87% of the variation in sale can be explained by its relationship with all the predictors.

Then we have to use model assumptions on this model to check if all of these assumptions are satisfied enough to predict the data. Linearity (use histogram to check this assumption) and normality (A.10) are both satisfied, but constant variance (A.11) and independence (A.11) are unsatisfied. Since histogram shows symmetric for sale and normality probability plot shows almost 45 degree line, there is no need to do transformation.

Once we have done the above process, the next thing we need to do is to split the data into train and test set to test how well the model predicts new data. And then using model selection to come up with the final model. After model selection, we could fit the final regression model (A.12):
sale=125309+2875.25023*instrspending+3690.08702*discount+589.54930*tvs

pending-13261*stockrate-6479.90670*price.

From the final model, we know that radio and onlineadsspending are insignificant predictors, so they have been removed from the model and the remaining predictors all have p-value less than 0.05, so they have been contained in the final model. We can also have the following conclusions:

- InStrSpending is positively associated with sale. Model shows that assuming all other variables constant, for every additional amount in in-store expense, sale increases by $2,875.

- Discount is positively associated with sale. Model shows that assuming all other variables constant, for every additional rate increment in discount, sale increases by $3,690.

- TVSpending is also positively associated with sale. For every additional amount of expenditure on television campaigns, sale increases by $590.

- Stockrate is negatively associated with sale. For every additional rate increment in stock-out ratio, sale decreases by $13,261.

- Price is also negatively associated with sale. For every additional amount in base price, sale decreases by $6,480.

Although we still saw outliers and influential points in final model, adj-r2 for final model (A.13) is very high which is 0.9986, so we do not need to remove them from the final model. We can keep them as part of my observations. We also see no predictors with VIF value larger than 10, so there is no problem of collinearity with final model.

After fitting the final model, we have to compute performance statistic (RMSE, MAE, R2) for test set to test how well the model's predictive performance is. By running code in SAS, we know that RMSE is 2773.71 (A.14) and MAE is 2141.72 (A.14) for test set. R2 for test set is computed as (0.99943)*2=0.99886032 (A.15) which is very high so we know that test set performs well. Another thing we need to do is to compute whether cross-validated R2 is less than or equal to 0.3 or not because if the value is less than 0.3, we could know that the model is good for predicting data. By subtracting the training set r2 to test set r2, we know the cross-validated R2 value is (0.9986-0.9988=-0.0002) which is less than 0.3, so we can come to conclusion that this model is good for predicting the sale value.

Comparing the RMSE and R2 for train and test set can help us decide if the model is really suitable for data prediction. RMSE for test set is 2773.71 and

RMSE for train set is 3015.46277 (A.16). R2 for test set is 0.9989 and R2 for train set is 0.9986. Since RMSE is based on errorters so it needs to be small. For test set, RMSE value is lower than train set and R2 value is higher than train set, so we could know that the model is perfect for predicting the data as test set should always perform better than train set because test set helps us to predict the unknown data which is very crucial in data analysis.

Finally, we need to compute predictions on the final model. My research question is: Are expenditures in in-store and TV campaigns can actually improve the sale amount? First, we have to provide the values for two observed group. First group with instrspending value of $500 and tvspending value of $900 while another group with instrspending value of $400 and tvspending value of $700. After running the code in SAS, we got the output for first group with predicted sale amount of $ 2,095,061 with a 95% C.I of (2,086,914, 2,100,786). And for the second group, we got the predicted sale amount of $1,689,337 with a 95% C.I of (1,684,857, 1,693,818). From the calculation, we can come to conclusion that it is true that with higher amount of money in both in-store expenditure and TV campaigns, sale amount will increase significantly. Therefore, it is very effective for the company to put more money in in-store expenditure and TV campaigns if they want to increase the sales of their product.

Appendix

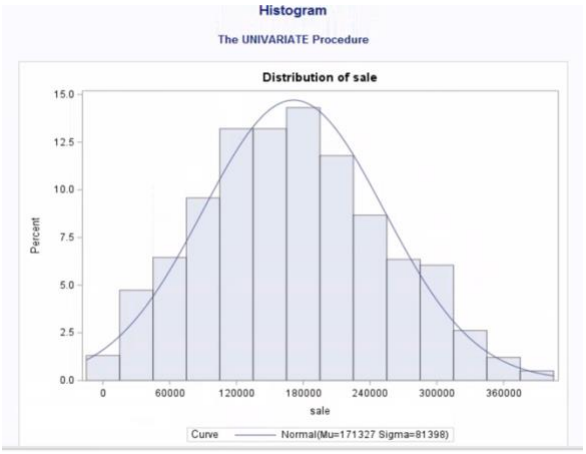
A.1

Descriptives

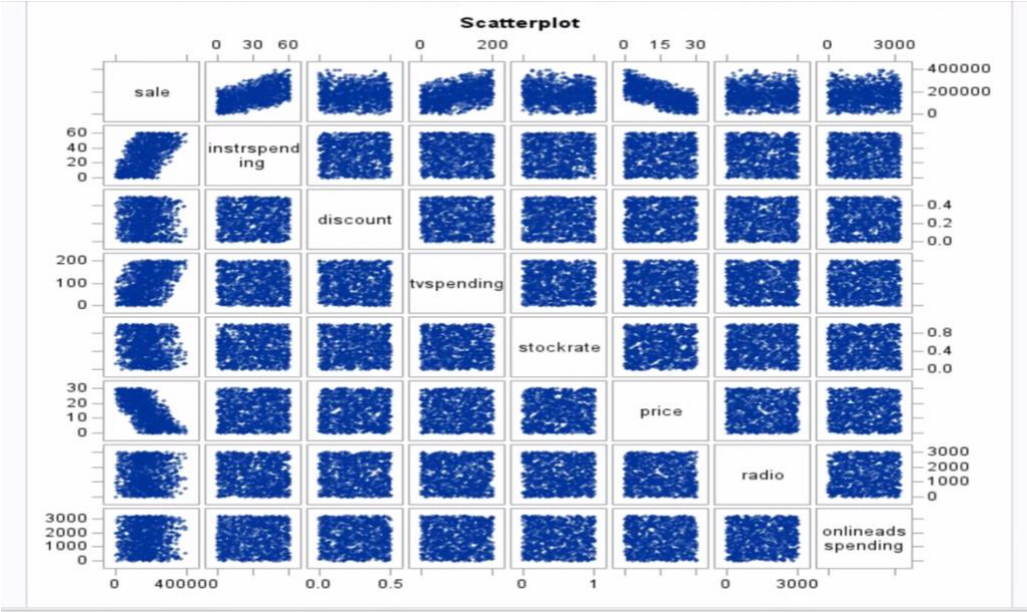
The MEANS Procedure

Analysis Variable : sale										
N	Minimum	25th Pctl	Median	75th Pctl	Maximum	Mean	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean	Std Error
992	1992.00	112404.50	170390.50	226087.50	393914.00	171327.12	81397.84	166255.63	176398.61	2584.38

A.2



A.3



A.4

Pearson Correlation Coefficients, N = 992
Prob > |r| under H0: Rho=0

	sale	instrspending	discount	tvspending	stockrate	price	radio	onlineadsspending
sale	1.00000	0.57771 <.0001	0.01100 0.7294	0.41333 <.0001	-0.07243 0.0225	-0.67337 <.0001	-0.01536 0.6289	0.04239 0.1822
instrspending	0.57771 <.0001	1.00000	0.02883 0.3644	-0.01853 0.5599	0.03175 0.3178	0.04491 0.1576	-0.08728 0.0059	0.03327 0.2952
discount	0.01100 0.7294	0.02883 0.3644	1.00000	-0.01365 0.6676	-0.01262 0.6915	0.01260 0.6918	-0.00433 0.8917	-0.03460 0.2763
tvspending	0.41333 <.0001	-0.01853 0.5599	-0.01365 0.6676	1.00000	-0.04508 0.1560	-0.01357 0.6695	-0.00092 0.9769	0.01238 0.6969
stockrate	-0.07243 0.0225	0.03175 0.3178	-0.01262 0.6915	-0.04508 0.1560	1.00000	0.03676 0.2474	-0.00141 0.9646	-0.00463 0.8841
price	-0.67337 <.0001	0.04491 0.1576	0.01260 0.6918	-0.01357 0.6695	0.03676 0.2474	1.00000	-0.05506 0.0830	-0.02297 0.4699
radio	-0.01536 0.6289	-0.08728 0.0059	-0.00433 0.8917	-0.00092 0.9769	-0.00141 0.9646	-0.05506 0.0830	1.00000	0.04542 0.1528
onlineadsspending	0.04239 0.1822	0.03327 0.2952	-0.03460 0.2763	0.01238 0.6969	-0.00463 0.8841	-0.02297 0.4699	0.04542 0.1528	1.00000

A.5

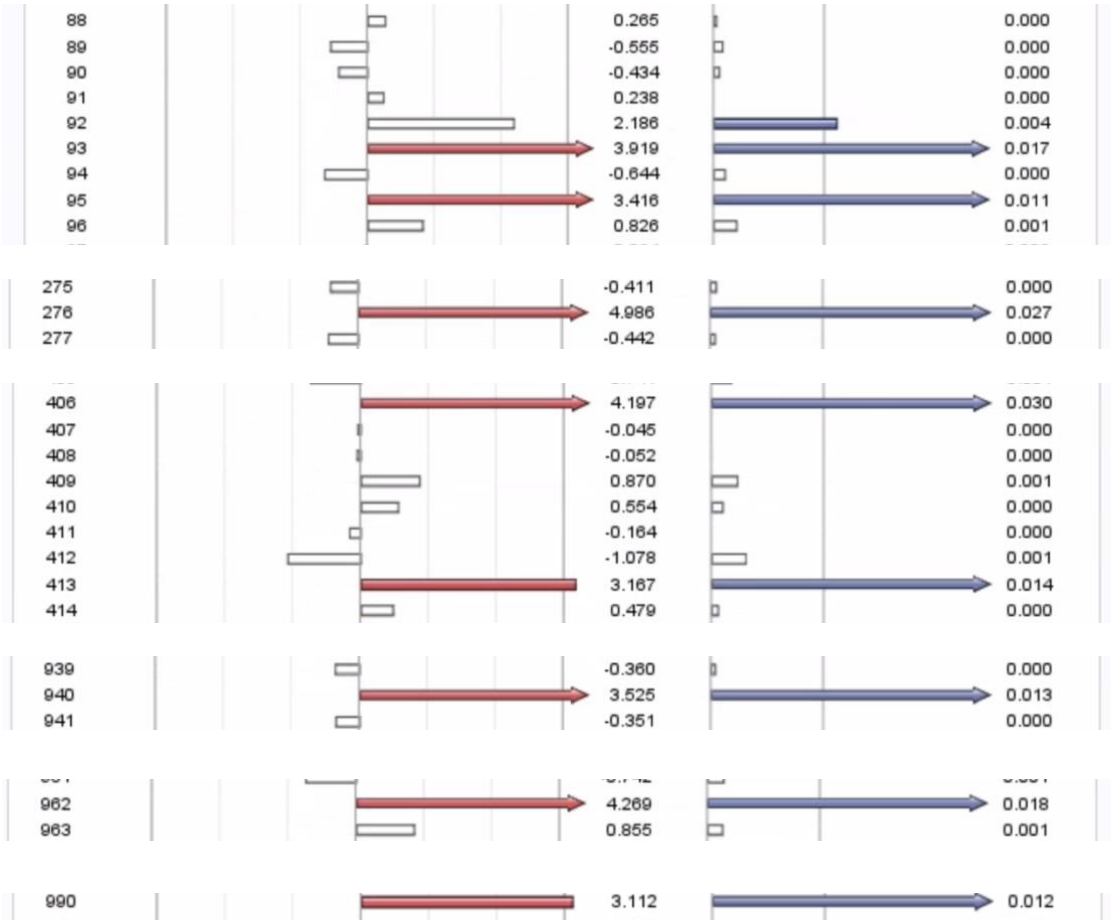
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	6.556254E12	9.366077E11	94772.8	<.0001
Error	984	9724539313	9882662		
Corrected Total	991	6.565978E12			

A.6

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	125379	468.48575	267.63	<.0001	0
instrspending	1	2875.05285	5.74544	500.41	<.0001	0.61787
discount	1	3918.55360	687.94357	5.70	<.0001	0.00700
tvspending	1	588.99058	1.75081	336.41	<.0001	0.41330
stockrate	1	-13573	348.95741	-38.90	<.0001	-0.04783
price	1	-6479.54605	11.49604	-563.63	<.0001	-0.69379
radio	1	0.06234	0.11350	0.55	0.5829	0.00067812
onlineadsspending	1	0.06712	0.10796	0.62	0.5343	0.00076476

A.7



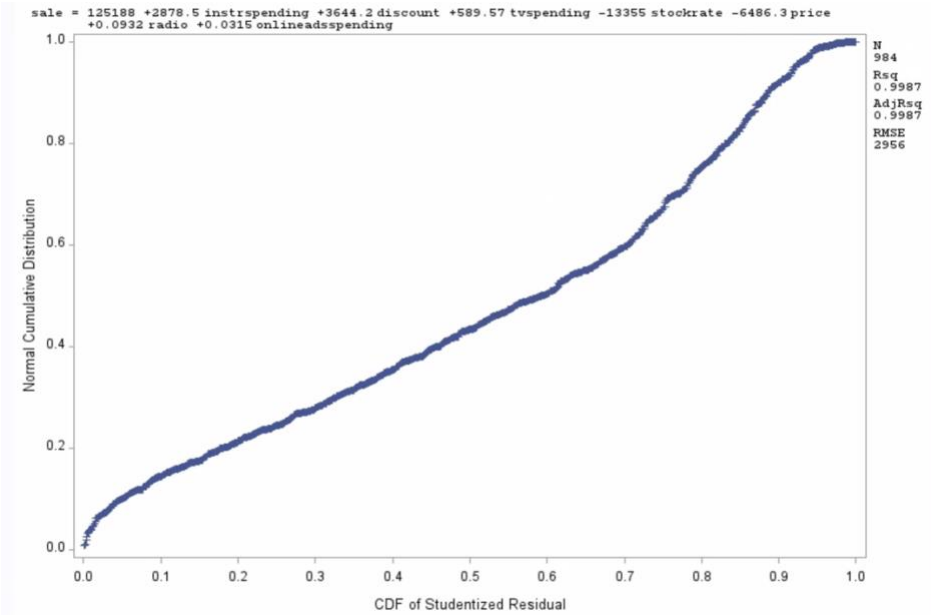
A.8

Root MSE	3143.67013	R-Square	0.9985
Dependent Mean	171327	Adj R-Sq	0.9985
Coeff Var	1.83489		

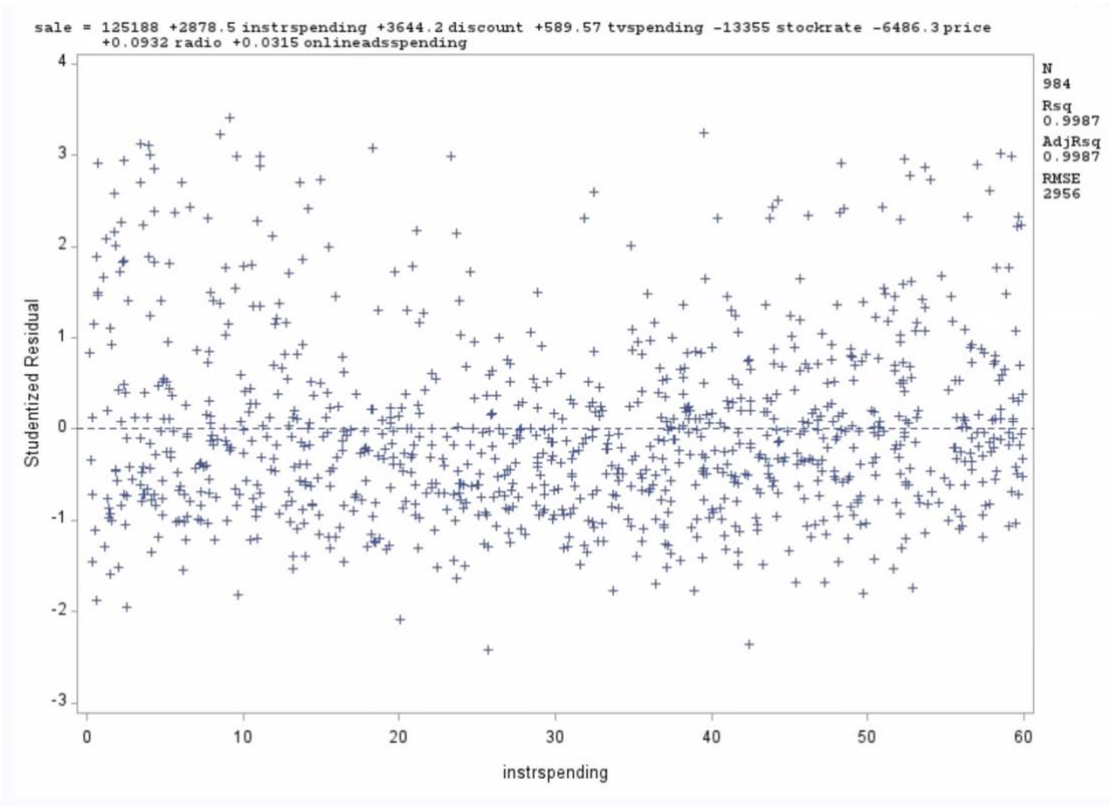
A.9

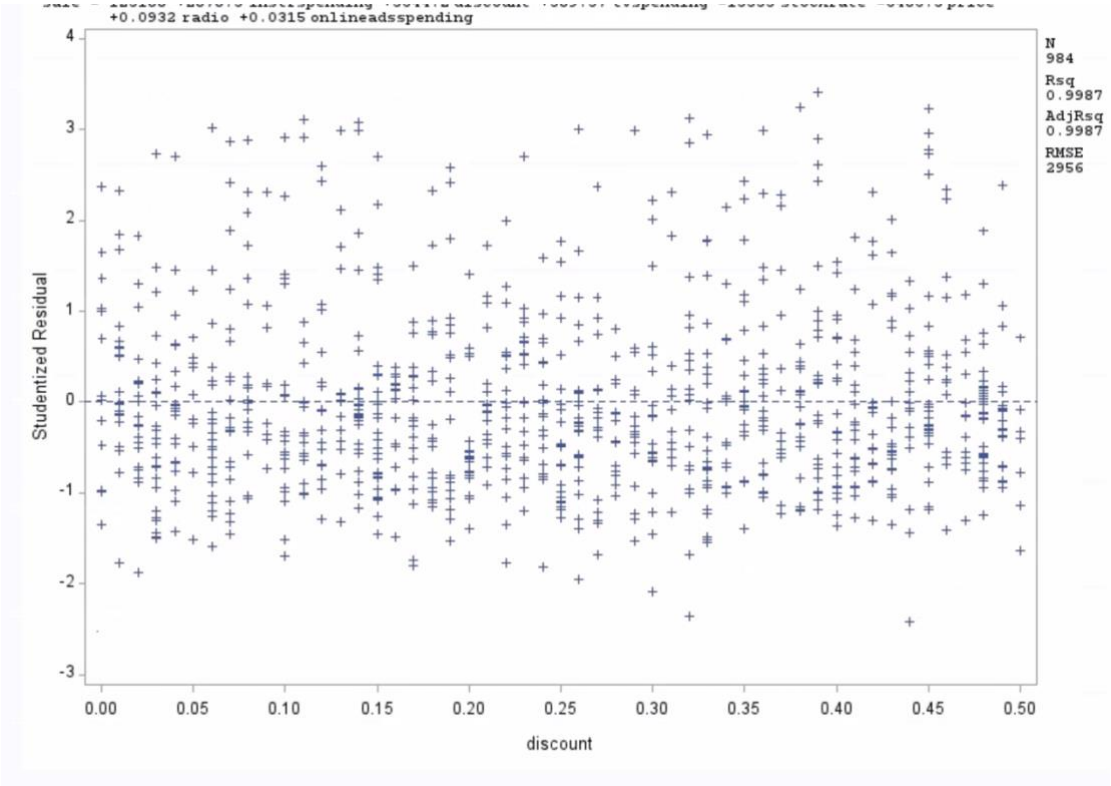
Root MSE	2955.95812	R-Square	0.9987
Dependent Mean	171896	Adj R-Sq	0.9987
Coeff Var	1.71962		

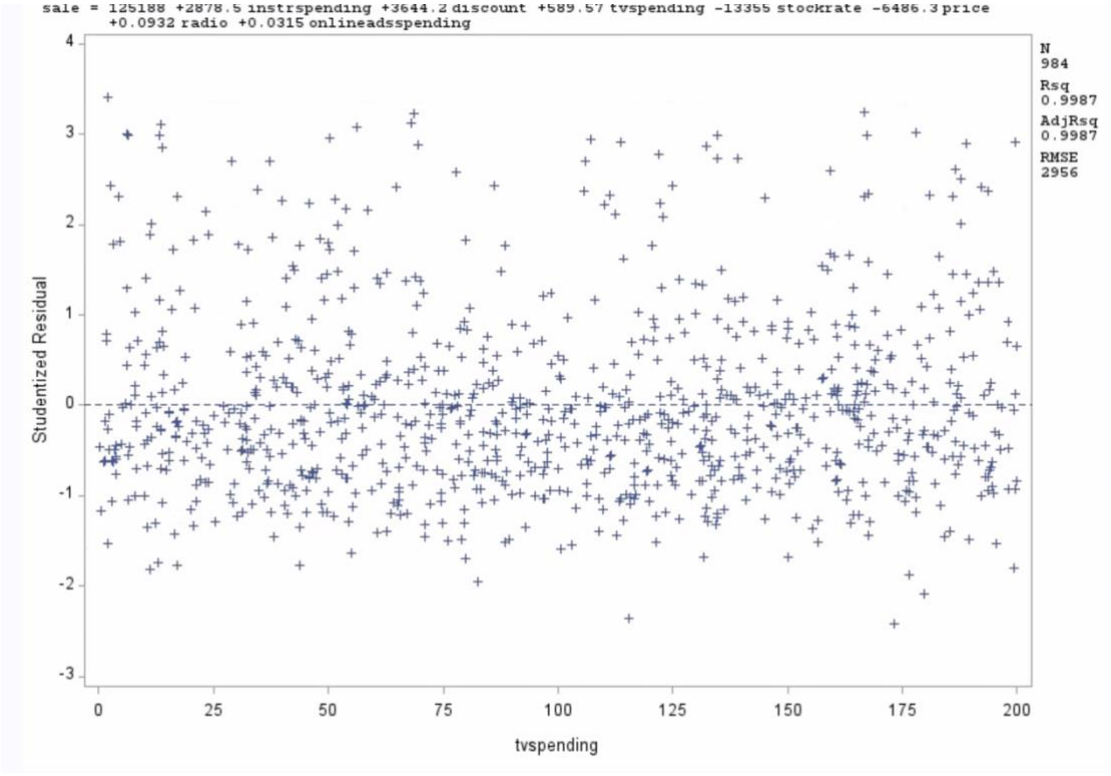
A.10

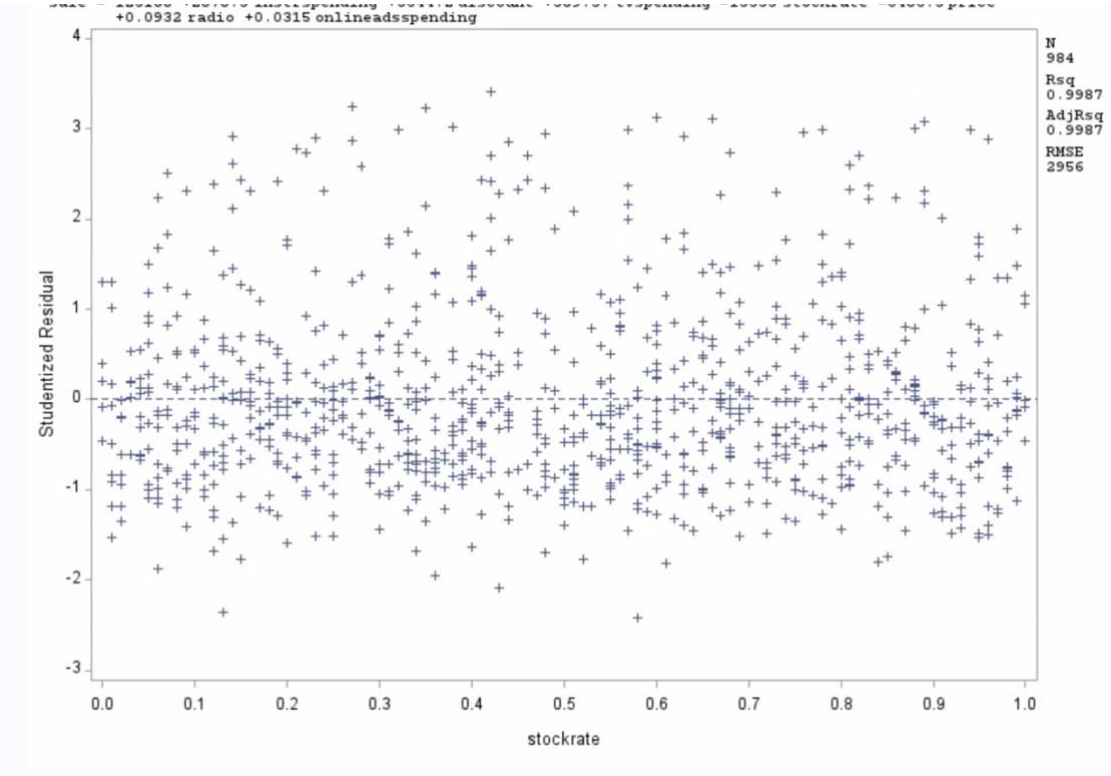


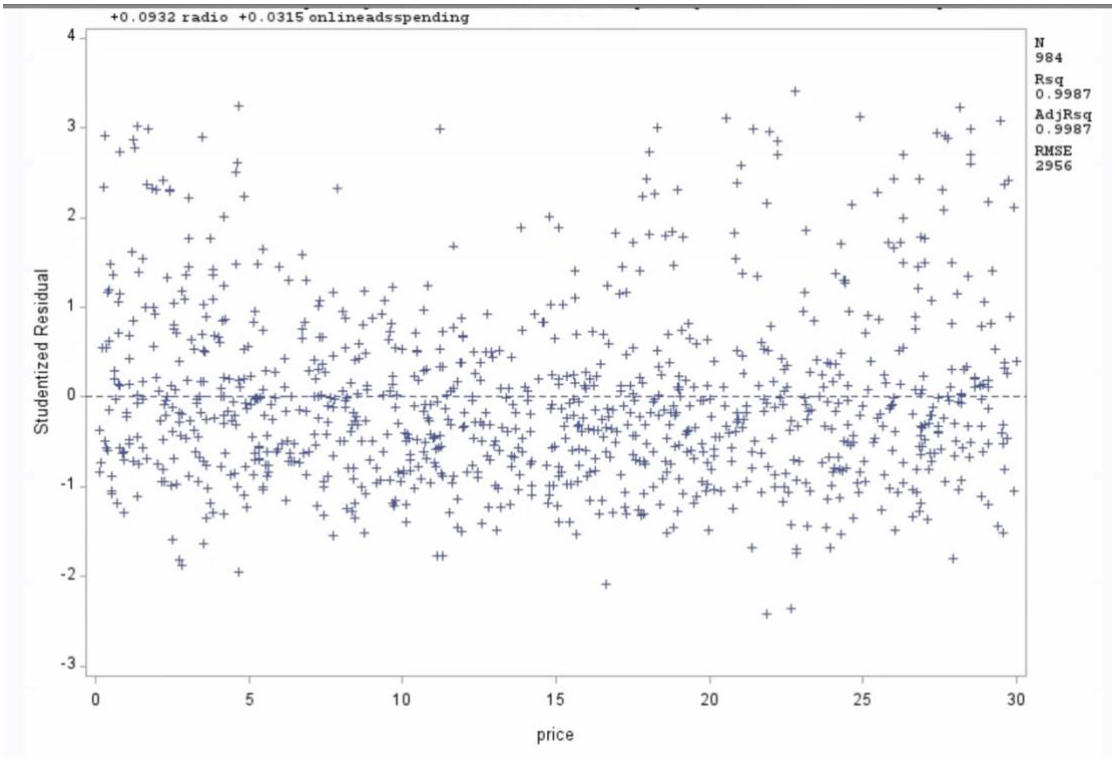
A.11

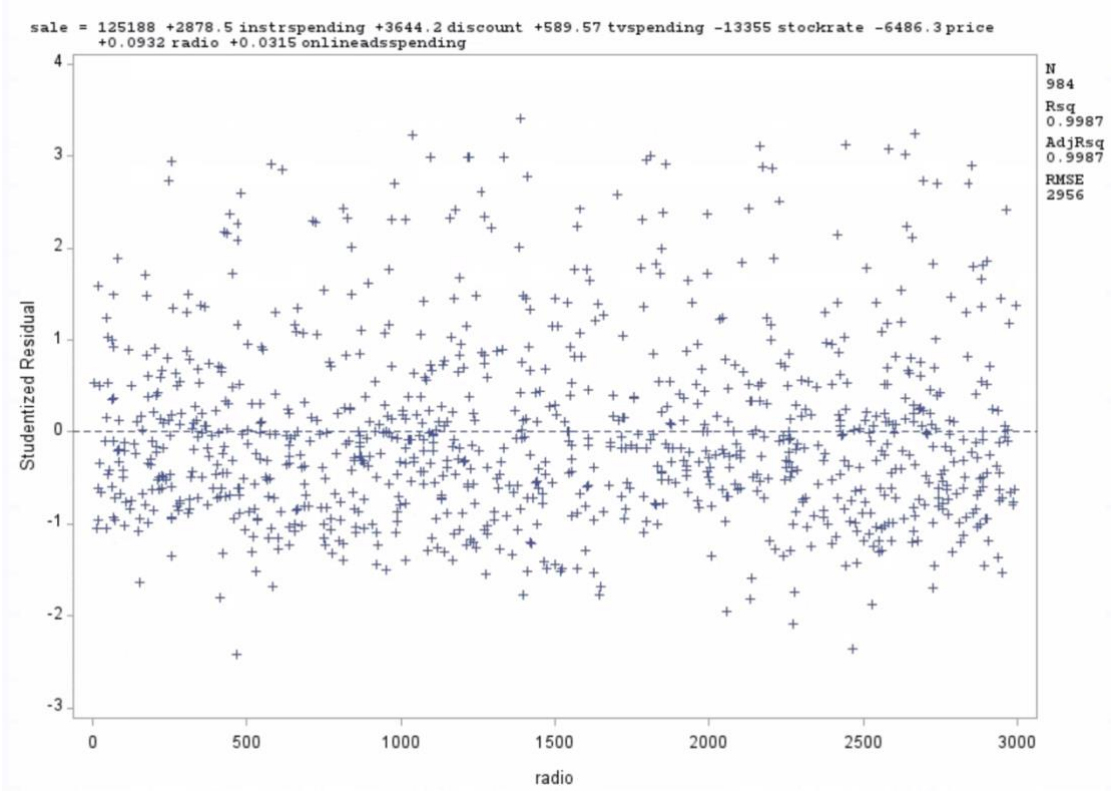

sale = 125188 +2878.5 instrspending +3644.2 discount +589.57 tvspending -13355 stockrate -6486.3 price
+0.0932 radio +0.0315 onlineadsspending
N
984
Rsq
0.9987
AdjRsq
0.9987
RMSE
2956
Studentized Residual
instrspending

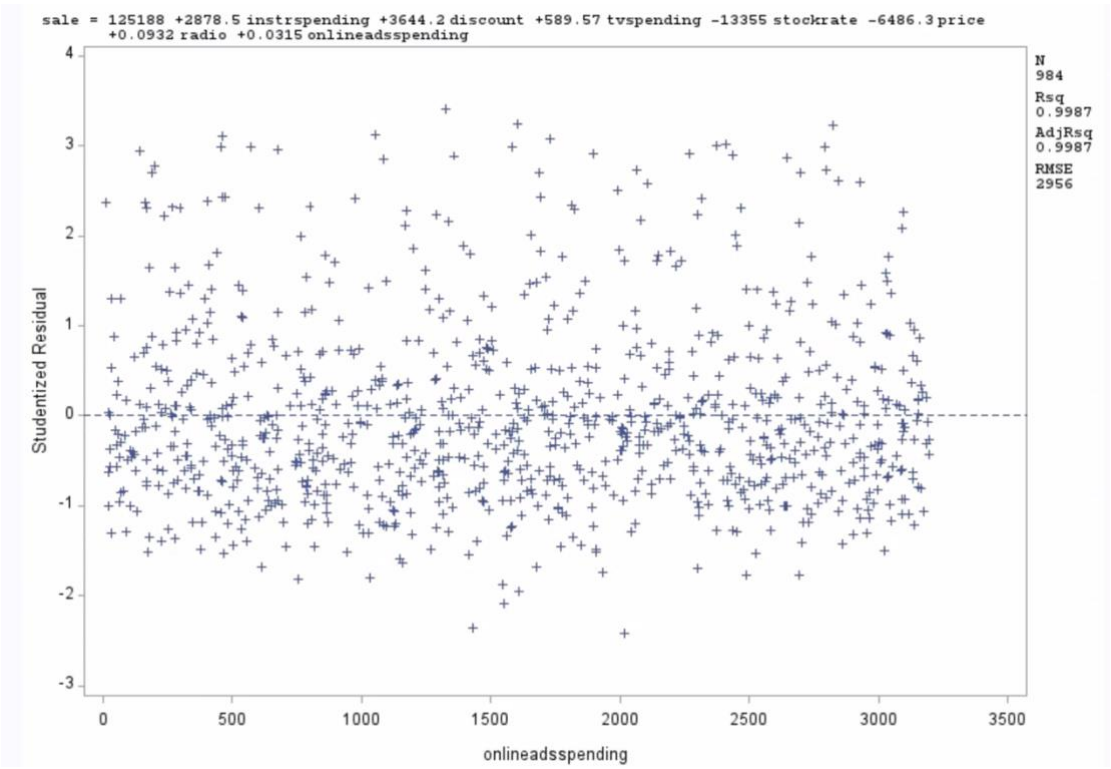

+0.0932 radio +0.0315 onlineadsspending
N
984
Rsq
0.9987
AdjRsq
0.9987
RMSE
2956
Studentized Residual
discount


sale = 125188 +2878.5 instrspending +3644.2 discount +589.57 tvspending -13355 stockrate -6486.3 price
+0.0932 radio +0.0315 onlineadsspending
N
984
Rsq
0.9987
AdjRsq
0.9987
RMSE
2956
Studentized Residual
tvspending


+0.0932 radio +0.0315 onlineadsspending
N
984
Rsq
0.9987
AdjRsq
0.9987
RMSE
2956
Studentized Residual
stockrate


+0.0932 radio +0.0315 onlineadsspending
N
984
Rsq
0.9987
AdjRsq
0.9987
RMSE
2956
Studentized Residual
price


sale = 125188 +2878.5 instrspending +3644.2 discount +589.57 tvspending -13355 stockrate -6486.3 price
+0.0932 radio +0.0315 onlineadsspending
N
984
Rsq
0.9987
AdjRsq
0.9987
RMSE
2956
Studentized Residual
radio



A.12

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	125309	434.54195	7.561476E11	83157.0	<.0001
instrspending	2875.25023	6.39846	1.83615E12	201930	<.0001
discount	3690.08702	764.09585	212073048	23.32	<.0001
tvspending	589.54930	1.94652	8.341248E11	91732.5	<.0001
stockrate	-13261	395.04888	10246252261	1126.83	<.0001
price	-6479.90670	12.95197	2.276007E12	250303	<.0001

A.13

Root MSE	3015.46277	R-Square	0.9986
Dependent Mean	171273	Adj R-Sq	0.9986
Coeff Var	1.76062		

A.14

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	246	2773.71	2141.72

A.15

Pearson Correlation Coefficients, N = 246 Prob > |r| under H0: Rho=0		
	sale	yhat
sale	1.00000	0.99943 <.0001
yhat Predicted Value of new_y	0.99943 <.0001	1.00000

A.16

Root MSE	3015.46277	R-Square	0.9986
Dependent Mean	171273	Adj R-Sq	0.9986
Coeff Var	1.76062		

A.17

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	2095061	2917	2089336	2100786	2086914	2103209	.
2	.	1689337	2283	1684857	1693818	1682010	1696664	.