An Exploratory Analysis of Retail Site Selection Using Multiple Linear Regression

Econ 5339 Final Project on Buxton Challenge

Jie Zhang, Rahul Dwivedi, Juncheng Yang

Department of Information Systems & Operations Management
Department of Economics
College of Business
University of Texas at Arlington

May 3, 2016

Project Summary

1. We first screen the stores that were not closed. The sample size reduced to 174.[1]
2. We first run stepwise regression on categorical variables and retail store age, we find the age of the store, high visibility have significant positive impact on the sales.
3. We identify potential significant geographical variables using correlation table. We choose the variables have p-value less then 0.01.
4. We use L_1 norm lasso regression to explore the selected geographic variables that may influence sales, we identify that Count of Malls or Shopping Centers with >300K SQFT of Gross Leasable Area (GLA) has significant impact on sales

[1] The problematic issue of the data is, some of the store still generate sales in 2015 even if their close date is within 2014, this is ridiculous.

Project Summary

1. We first screen the stores that were not closed. The sample size reduced to 174.[1]
2. We first run stepwise regression on categorical variables and retail store age, we find the age of the store, high visibility have significant positive impact on the sales.
3. We identify potential significant geographical variables using correlation table. We choose the variables have p-value less then 0.01.
4. We use L_1 norm lasso regression to explore the selected geographic variables that may influence sales, we identify that Count of Malls or Shopping Centers with >300K SQFT of Gross Leasable Area (GLA) has significant impact on sales

[1] The problematic issue of the data is, some of the store still generate sales in 2015 even if their close date is within 2014, this is ridiculous.

Project Summary

1. We first screen the stores that were not closed. The sample size reduced to 174.[1]
2. We first run stepwise regression on categorical variables and retail store age, we find the age of the store, high visibility have significant positive impact on the sales.
3. We identify potential significant geographical variables using correlation table. We choose the variables have p-value less then 0.01.
4. We use L_1 norm lasso regression to explore the selected geographic variables that may influence sales, we identify that Count of Malls or Shopping Centers with >300K SQFT of Gross Leasable Area (GLA) has significant impact on sales

[1] The problematic issue of the data is, some of the store still generate sales in 2015 even if their close date is within 2014, this is ridiculous.

Project Summary

1. We first screen the stores that were not closed. The sample size reduced to 174.[1]
2. We first run stepwise regression on categorical variables and retail store age, we find the age of the store, high visibility have significant positive impact on the sales.
3. We identify potential significant geographical variables using correlation table. We choose the variables have p-value less then 0.01.
4. We use L_1 norm lasso regression to explore the selected geographic variables that may influence sales, we identify that Count of Malls or Shopping Centers with >300K SQFT of Gross Leasable Area (GLA) has significant impact on sales

[1] The problematic issue of the data is, some of the store still generate sales in 2015 even if their close date is within 2014, this is ridiculous.

Introduction

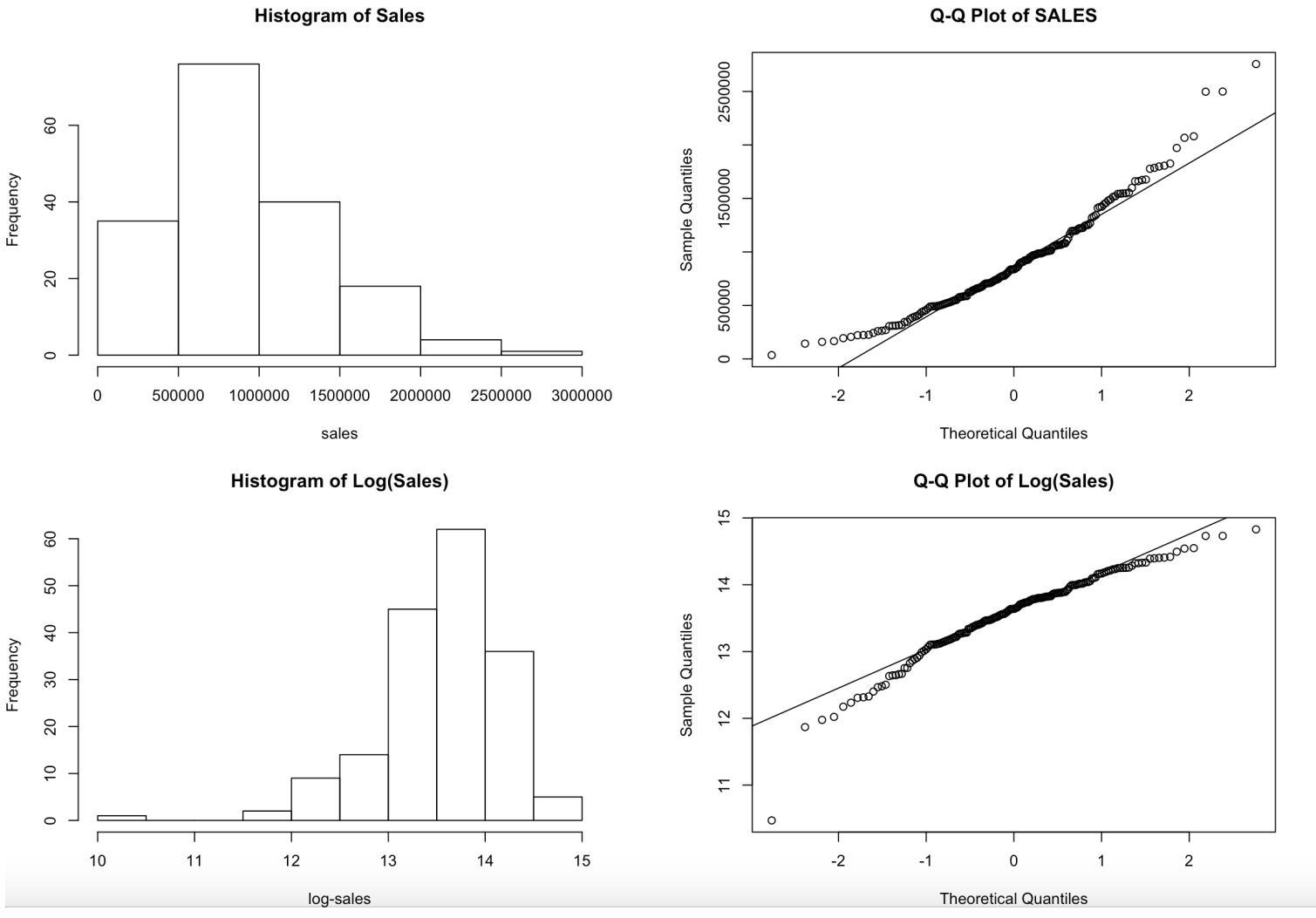
- Retail site selection has been an essential part of business strategy for retail industry.
- Using a retailer's existing locations from Buxton challenge, we build econometric models to forecast revenue of potential new locations

Table: Descriptive Statistics for 2015 Sales

n	mean	sd	median	min	max	range	skew	kurtosis	se
174	911149.2	491821.8	838116.5	35287.42	2756106	2720819	0.91	1.04	37284.91

Sales data

Histgram of Sales



Stepwise Regression

- We speculate if the age of the store can influence its sales revenue, so we calculate the duration of the store from the opening date to Dec 2015.
- Then we use sales as dependent variable; STATE, REGION, Age, DENSITY, SQFT, NBR_MACHINES, PARTY_ROOM, PATIO, BUILDING_TYPE HIGH_VISIBILITY as independent variable.
- Using stepwise regression we can find that only two variables, store age and visibility have positive and significant impact on sales

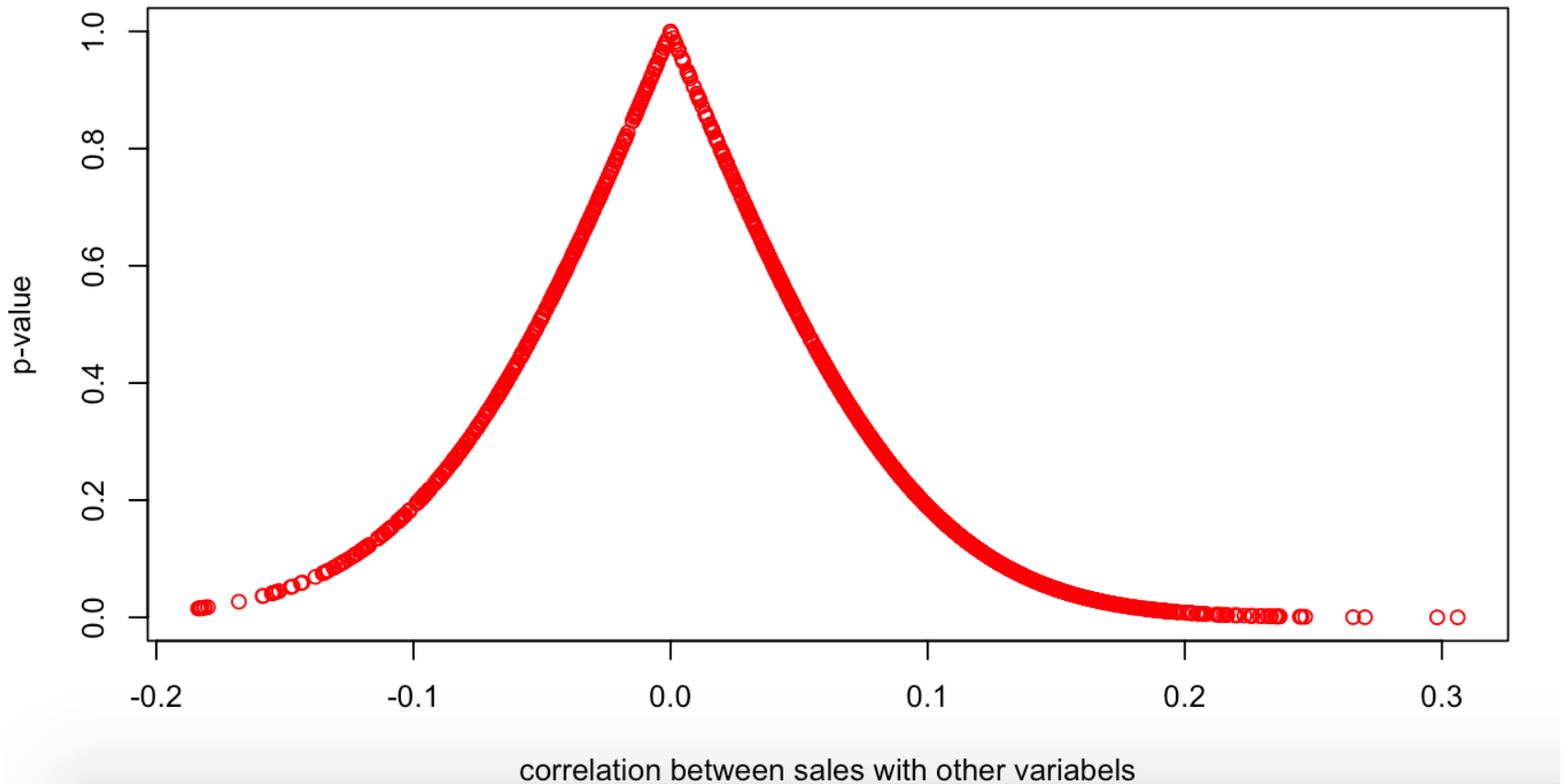
Table: Stepwise Regression

	Dependent variable:
	SALES.2015
monthsofage	16,124.010***
	(3,026.068)
factor(HIGH_VISIBILITY)Y	195,213.500***
	(69,736.890)
Constant	325,463.700***
	(102,467.700)
Observations	174
R^2	0.177
Adjusted R^2	0.167
Residual Std. Error	448,768.600 (df = 171)
F Statistic	18.393*** (df = 2; 171)
Note:	*p<0.1; **p<0.05; ***p<0.01

Correlation Table

- In order to search variables that are significant correlated sales, we calculated the correlation coefficient of geographical variable and sales.
- We choose the variables with p-value less than 0.01(74), combine them with age and high visibility.

Correlation Coefficient and p-value



Lasso Regression for Variable Selection

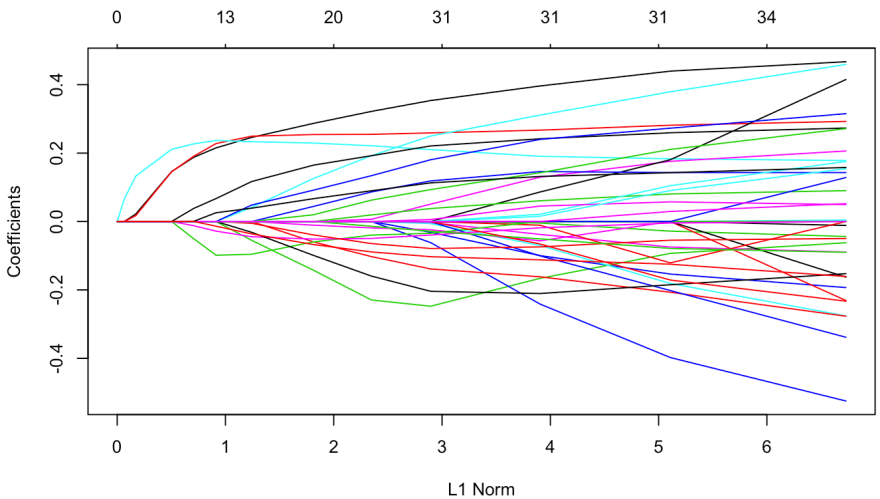
The lasso estimate is defined as

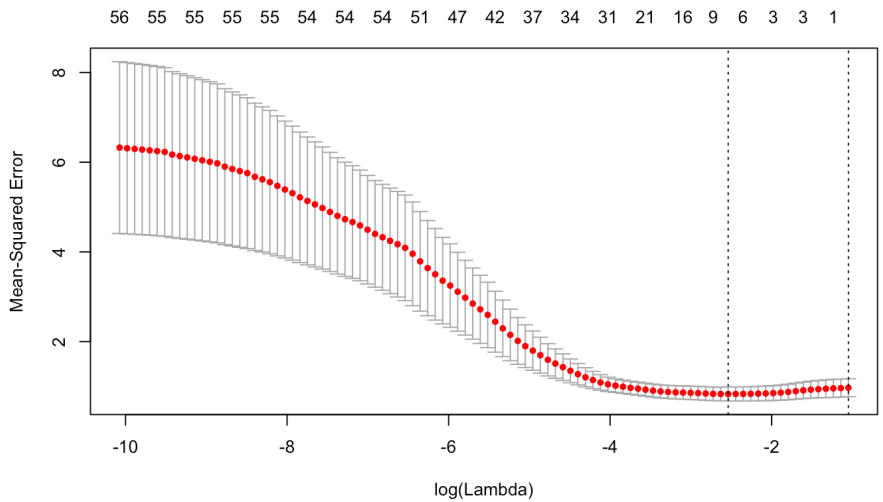
$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \sum_{p}^{j=1} |\beta_j|$$

$$= \arg\min_{\beta \in \mathbb{R}^p} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

The tuning parameter λ controls the strength of the penalty, and we get $\hat{\beta}^{\text{lasso}} =$ the linear regression estimate when $\lambda = 0$, and $\hat{\beta}^{\text{lasso}} = 0$ when $\lambda = \infty$

For λ in between these two extremes, we are balancing two ideas: fitting a linear model of y on X, and shrinking the coefficients. This is why lasso can perform variable selection.


0 13 20 31 31 31 34
Coefficients
0.4
0.2
0.0
-0.2
-0.4
0 1 2 3 4 5 6
L1 Norm


56 55 55 55 55 54 54 54 51 47 42 37 34 31 21 16 9 6 3 3 1
Mean-Squared Error
8
6
4
2
-10 -8 -6 -4 -2
log(Lambda)

We perform lasso on the new combined data set, 74 geographical variables, age of store and high visibility. we get the below variables are significant

selected significnat variable using lasso

age (months)	HIGH_VISIBILITY	CMDSC_PHARMACY_1RO	CNT_MALLS_300K_0_5RO	CYB05V003_8TO

Full Sample Regression using selected Variables

	Dependent variable:	
	SALES.2015	
	(1)	(2)
monthsofage	15,029.380***	15,822.080***
	(3,012.428)	(2,986.120)
HIGH_VISIBILITY)Y	229,171.600***	208,254.900***
	(70,452.190)	(68,965.670)
CMDSC_PHARMACY_1RO	−96,194.930	
	(77,455.550)	
CNT_MALLS_300K_0_5RO	135,140.600**	141,896.400**
	(58,826.550)	(58,367.630)
CYB01V003_8TO	−26.443	
	(22.567)	
Constant	366,674.600***	275,698.600***
	(116,021.300)	(103,080.500)
Observations	174	174
R^2	0.219	0.205
Adjusted R^2	0.196	0.191
Residual Std. Error	441,069.300 (df = 168)	442,461.100 (df = 170)

Forecasting for Optional Sites

We finally use ONLY THREE variables to forecast the sales of the optional stores and all the existing 174 stores. and the reduced form regression is:

$$sales = 275698.6 + 15822.08 \times \text{monthsofage} + 208254.9 \times \text{HIGH\_VISIBILITY} + 141896.4 \times \text{CNT\_MALLS\_300K\_0\_5RO}$$

One Year Forecasted Sales for Optional Sites

SID	NUMBER	NAME	STATE	REGION	Forecasted Sales for One Year
15839156	NBR061	CHERRY PEMBROKE	MS	ESC	465563.56
15415346	**NBR072**	**MARKET HILLS**	**TN**	**ESC**	**749356.36**
12890987	NBRB526	MARKET SCOFIELD	GA	SA	673818.46
13473985	NBRB529	ARTESIA SANTA	WA	WP	673818.46
13450510	NBRB543	CLEANERS HAMMOCK	FL	SA	673818.46

Forecasting all the existing store sales

We try to forecast the sales of the existing stores using these coefficient, and we compared the real sales and forecasted sales. We calculated absolute percent error for each store. Below, we outline the contingency table for the 174 existing stores, then we can see that our method can achieve satisfactory forecasting accuracy.

Accuracy Measures for Forecasting

Absolute Percent Error	NO. of Stores	Percent
0-0.1	34	19.54%
0.1-0.2	23	13.22%
0.2-0.3	25	14.37%
0.3-0.4	24	13.79
0.4-0.5	15	8.62%
0.5-0.6	15	8.62%
0.6-0.7	6	3.45%
0.7-0.8	5	2.87%
0.8-0.9	2	1.15%
0.9-1.0	2	1.15%
1.0-2.0	16	9.20%
2.0-3.0	4	2.30%
3.0-4.0	1	0.57%
5.0-6.0	1	0.57%
34.0-35.0	1	0.57%

Conclusion, Discussion and Limitations

1. Our method can achieve satisfactory forecasting outcome for most store sales, but we do find some outlines.
2. Since we use the correlation table and p-value to identify single significant item, we cannot identify joint significant items (but our approach can be used for fast detection of important factors)
3. Other more sophisticated methods can be applied for future study, for example we can cluster the stores first and then regress on each cluster, after than we can use majority voting to determine significant variables.
4. we can also use other lasso methods like adaptive lasso, group lasso etc.

Questions?

Thank You!