An Exploratory Analysis of Retail Site Selection Using Multiple Linear Regression

Econ 5339 Final Project on Buxton Challenge

Jie Zhang[†*], Rahul Dwivedi[*], and Juncheng Yang[**]

[*] *Department of Information Systems & Operations Management*
The University of Texas at Arlington
[**] *Department of Economics*
The University of Texas at Arlington

May 3, 2016

Abstract

In this project, we aim to explore the factors that may impact the decision of location selection for retail stores using existing store characteristics and demographic variables. To achieve this, we first use stepwise linear regression to identify the categorical variables that can influence sales revenue, we find that age of the store (measured in months) and store visibility have statistically significant and positive impact on sales revenue for the retail stores. Then we create a correlation table to find variables that may influence sales revenue. We choose variables that have statistically significant correlation coefficent at significance level of 0.01. Then we use lasso regression for the final variable selection. Finally we use these variables to forecast sales for potential retail sites and make site selection decision.

[†]Jie Zhang (jie.zhang2@mavs.uta.edu), Rahul Dwivedi (rahul.dwivedi@mavs.uta.edu), Juncheng Yang(juncheng.yang@mavs.uta.edu). Please address correspondence on this project to Jie Zhang.

Introduction

Retail site selection has been an essential part of business strategy for the retail industry. Location of a retail store offers a unique competitive advantage as once a site location is selected, it cannot be occupied by another store. Also once a location is selected its difficult to relocate to a new location for at least a few years. Thus location selection and analysis is one of important decision made by retail stores. For the purpose of this project, based on the existing locations for one of the client's of Buxton, we build econometric models to forecast revenue of potential new locations. The data set is cross-sectional in nature containing historical performance and demographic characteristics information for the area surrounding client's locations for the year 2014 along with sales revenue for those retail locations.

Methodology

Based on the given cross-sectional data, we first want to check whether the dependent variable of sales fulfills the normality assumption or not. The histogram for these dependent variable is shown below.

Figure 1: Histogram of Sales

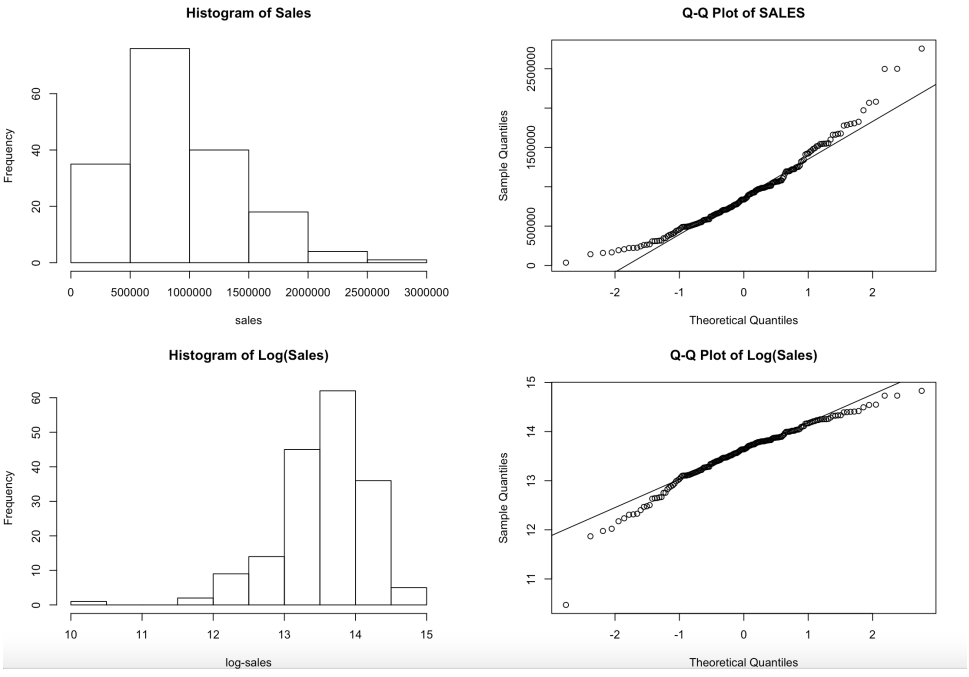


Table 1: Summary Statistics for 2015 Sales

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

As we can see from the figure above since the histogram for sales variable is not highly skewed there is no need to transform these dependent variable.

Before the final variable selection, we want to explore if some of the categorical variables are significant or not. In particular we want to speculate if the age of the store can influence its sales revenue. In order to derive the age of the store we calculate the duration of the store from the opening date to Dec 2015 as one of the independent variables. We measured the age of the store in months between the store opening date and 31 December 2015. Then we use sales as dependent variable regressed against other independent categorical variables such as STATE, REGION, Age, DENSITY, SQFT, NBR_MACHINES, PARTY_ROOM, PATIO, BUILDING_TYPE and HIGH_VISIBILITY. Using stepwise regression we found that out of the above independent categorical variables only two variables - store age and visibility have positive as well as statistically significant effect on sales. The results for step-wise regression are shown in Table 2. As we can see with inclusion of only this two variables we can achieve R^2 around 18%.

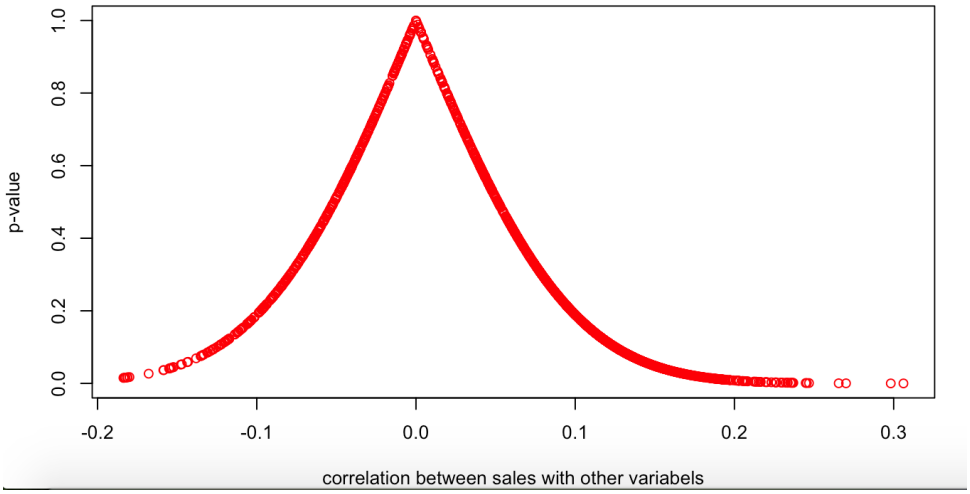
Since the ratio of number of observations to number of variables for the given data set is very low, we cannot directly run multiple regression with all those independent quantitative variables. Also, we suppose that most of the quantitative variables will have statistically insignificant relationship with the dependent variable. Hence, in order to find a list of meaningful variables from the large list of quantitative variables, we came up with the idea of using the correlation table for the purpose of identifying the variables that might be significantly correlated with the dependent variable of sales or revenue. Thus, in order to fast detect the potential variables that are significantly correlated with dependent variable of sales, we choose those variables which have p-value of less than 0.01. In other words, we choose variables which are very highly significantly correlated with the dependent variable of sales, although individually. We can also observe from Figure 3 that, as the correlation coefficient increases for a variable, it's corresponding p-value

Table 2: 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

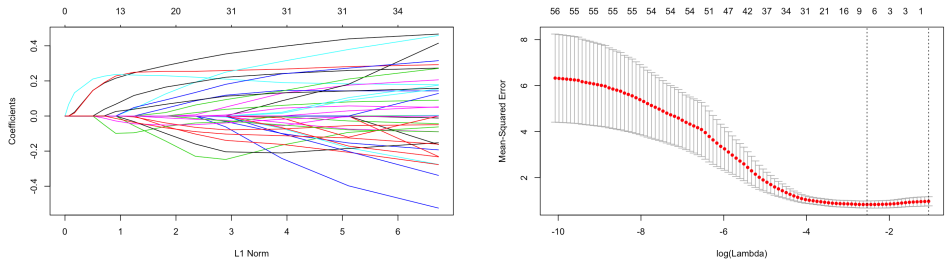
decreases. Thus the higher the correlation between a quantitative independent variable and the dependent variable the more significant these relationship.

Figure 2: Correlation coefficient and p-value



Based on the above naive technique for variable selection we were able to identify 74 quantitative variables that are significantly correlated with sales and having p-value less than 0.01. The selected variables and their description are shown in Table 3. We can also observe that some of the quantitative variables which are significant correlated with

Figure 3: Lasso Variable Selection Process



sales revenue intuitively makes sense such as distance score to pharmacy store, household maximum income group, etc.

Table 3: Potential Significant Variables

Variable Name	Description
CMDSC_PHARMACY	Distance Score (Closer Proximity and Greater Count = Higher Value) - All Pharmacies (e.g. CVS, Walgreen's)
CM_MALL	Count - All Mall Stores (e.g. Journey's, Footlocker, JCPenney)
CNT_MALLS_100K	Count of Malls or Shopping Centers with>100K SQFT of Gross Leasable Area (GLA)
CNT_MALLS_300K	Count of Malls or Shopping Centers with >300K SQFT of Gross Leasable Area (GLA)
CX01V037	Total Expenditure:Food:Food at home: Dairy products:Other dairy products Ice cream and related products
CYA04V001	Age:Total (Pop) Age 0-4
CYA04V002	Age:Total (Pop) Age 5-9
CYA04V003	Age:Total (Pop) Age 10-14
CYA04V004	Age:Total (Pop) Age 15-17
CYA08V003	Race and Ethnicity:Population American Indian & Alaska Native
CYB05V002	Race and Ethnicity:Households:Total Black/African American
CYB05V003	Race and Ethnicity:Households:Total American Indian & Alaska Native
CYB05V005	Race and Ethnicity:Households:Total Native Hawaiian / Other Pacific Islander
CYB17V022	Housing Value Housing Value $500,000-$749,999
CYD01VV05	School Enrollment (Pop 3+):Total Kindergarten
CYD02VV06	Educational Attainment:By Sex (Pop 25+):Total 10th grade
CYEC14V001	Income:Household Average (Mean) Household Income
D20200	Demand by Merchandise Line Men's wear, including accessories
D20240	Demand by Merchandise Line Children's wear, incl boys', girls', infants' & toddlers'
D20310	Demand by Merchandise Line Small electric appliances & personal care appliances
D20340	Demand by Merchandise Line Furniture, sleep equipment & outdoor/patio furniture
DIST_HWY	Distance to nearest highway (Miles)
ESTSIC54	Establishments: Major SIC Division: Retail Trade (52-59) Food Stores (54)
ESTSIC56	Establishments: Major SIC Division: Retail Trade (52-59) Apparel and Accessory Stores (56)
ESTSIC57	Establishments: Major SIC Division: Retail Trade (52-59) Home Furniture, Furnishings and Equipment Stores (57)
MAXINCGROUP	Income:Household Maximum Income Group (HHs)
MINAGEVAL	Age:Total (Pop) Minimum Age Value
MININCGROUP	Income:Household Minimum Income Group (HHs)
SMAPPR001H	Apparel:HH Children's Clothing - Bought Last 12 Months
SMAPPR075H	Apparel:HH Women's Apparel/Accessories - Bought Last 12 Months
XCYA08V004	By Percent:Race and Ethnicity:Population:Total % Asian
XCYA08V005	By Percent:Race and Ethnicity:Population:Total % Native Hawaiian / Other Pacific Islander
XCYB17V020	By Percent:Housing Value Housing Value $300,000-$399,999
XCYB17V021	By Percent:Housing Value Housing Value $400,000-$499,999
XCYD01VV17	By Percent:School Enrollment (Pop 3+):Total % College undergraduate

Finally, we use lasso regression, which is shrinkage and variable selection technique for linear regression to minimize and select the most important variables from the above 74 selected variables. From lasso regression, we found some variables to be statistically significant. These final selection of variables is shown in Table 6, and Figure 3 shows how the variables are selected.

Thus, our final selection of variables for the purpose of forecasting revenues are as follows: **monthsofage** , **HIGH_VISIBILITY** , **CMDSC_PHARMACY_1RO** ,

CNT_MALLS_300K_0_5RO , **CYB05V003_8TO**.

Results

The last step is to forecast the revenue using the selected five variables and choose the site with highest predicted revenue for the year 2015. In oder to accomplish this task, we first run regression using the above selected variable, followed by dropping the non-significant variables. The results are shown in Table 5.Eventually we identify three variables that are significant: **monthsofage**, **HIGH_VISIBILITY**, and **CNT01V003_8TO**. The reduced form regression is

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

Out of the above selected variables the significant and positive relation between age of retail store measured in months with sales of the retail store can be explained on the basis of store loyalty, as the older the store gets the more loyal will its customers becomes over time leading to higher revenues. The empirical investigation highlighting the relationship between store loyalty and its revenue had been carried out by Knox and Denison (2000) who found that there is significant association between store loyalty and customer expenditure.

The significant positive relation of the variable CNT_MALLS_300K_0_5RO which represents number of malls or shopping centers with greater than 300K square feet of gross leasable area can be explained on the basis of two important dimensions of retail store location - proximity to consumers and proximity to other stores i.e. agglomeration; as explained in Fox, Postrel and McLaughin (2007). Both are important predictors of consumer spending and hence revenues of retailers. The phenomenon of stores locating near each others is known as agglomeration and large malls as represented by the variable under question is a good example. Thus, such large malls may bring a lot of consumers

to other smaller retailers located in vicinity of the mall, such as the retail store or the yogurt shop which are located within 0.5 miles of radius. It also supports the intuition that retailers may want to be located at locations with lots of shoppers visiting shopping malls.

Finally, the significant positive relation of the categorical variable store visibility with revenue can be explained intuitively as the more visible store will have higher revenue sales. In other words, the revenue for more visible store locations will be significantly higher than stores with lower visibility as highly visible stores will get more customers leading to higher sales.

Then we make forecasted sales based on the above regression, which is shown in Figure 4.

Table 4: 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

Based on the above forecasted sales, we can see that the highest revenue for the year 2015 is predicted for the retail location named MARKET HILLS for the state TN and region ESC. The forcasted sales for these location is 749356.36.

Conclusion

In this study we apply multiple linear regression to forecast retail store sales. We find that older stores with high visibility are more likely to attract more customers to get more sales revenue. We also find that retail stores surrounding by more large malls or shopping centers have larger sales. Although there are some outliers, our method can achieve satisfactory forecasting outcome for most store sales.

Our approach do have some several limitations, 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.

For future research, we will consider more sophisticated methods, 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. we can also use other lasso methods like adaptive lasso, group lasso etc.

References

Fox, E. J., Postrel, S., & McLaughlin, A. (2007). The impact of retail location on retailer revenues: An Empirical investigation. Southern Methodist University, mimeo.

Knox, S. D., & Denison, T. J. (2000). Store loyalty: its impact on retail revenue. An empirical study of purchasing behaviour in the UK. Journal of retailing and consumer services, 7(1), 33-45.

Table 5:

	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)
F Statistic	9.421*** (df = 5; 168)	14.584*** (df = 3; 170)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 6: selected significant variables using lasso

age	HIGH_VISIBILITY	CMDSC_PHARMACY_1RO	CNT_MALLS_300K_0_5RO	CYB05V003_8TO
2.725107e-01	1.149364e-01	-3.579293e-02	8.568517e-02	-1.632991e-02

Appendix

```
par(mfrow=c(2,2),mar=c(4, 4, 4, 4))
hist(SALES.2015, xlab="sales", main="Histogram of Sales")
qqnorm(SALES.2015, main="Q-Q Plot of SALES")
qqline(SALES.2015)
hist(log(SALES.2015), xlab="log-sales", main="Histogram of Log(Sales)")
qqnorm(log(SALES.2015), main="Q-Q Plot of Log(Sales)")
qqline(log(SALES.2015))
```

```
> formula0<-SALES.2015 ~ factor(STATE)+factor(REGION)+monthsofage+factor(DENSITY_CLAS
+SQFT+NBR_MACHINES+PARTY_ROOM+factor(PATIO)+factor(BUILDING_TYPE)+factor(HIGH_VISIBIL
> reg0<-lm(formula0,data=dataset1)
> summary(reg0)

Call:
lm(formula = formula0, data = dataset1)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-731196 -207015  -17837  191124 1286734

Coefficients: (7 not defined because of singularities)
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.458e+05  5.400e+05   1.196  0.23380
factor(STATE)AZ         -6.146e+05  5.800e+05  -1.060  0.29121
factor(STATE)CA         -8.550e+05  4.212e+05  -2.030  0.04439 *
factor(STATE)CO         -9.771e+05  4.221e+05  -2.315  0.02217 *
factor(STATE)FL         -8.883e+05  4.081e+05  -2.177  0.03127 *
factor(STATE)GA         -9.953e+05  4.220e+05  -2.358  0.01982 *
factor(STATE)HI         -1.006e+06  4.971e+05  -2.023  0.04505 *
factor(STATE)IA         -7.703e+05  5.575e+05  -1.382  0.16939
factor(STATE)IL         -6.592e+05  4.712e+05  -1.399  0.16415
factor(STATE)IN         -1.200e+06  4.910e+05  -2.444  0.01585 *
factor(STATE)KY         -1.063e+06  4.666e+05  -2.277  0.02438 *
factor(STATE)LA         -1.416e+06  5.045e+05  -2.807  0.00575 **
factor(STATE)MD         -5.514e+05  5.669e+05  -0.973  0.33248
factor(STATE)MI         -2.860e+05  5.757e+05  -0.497  0.62019
factor(STATE)MN         -3.189e+05  5.823e+05  -0.548  0.58485
factor(STATE)NC         -1.107e+06  4.416e+05  -2.507  0.01338 *
factor(STATE)NJ         -1.215e+06  5.661e+05  -2.146  0.03367 *
factor(STATE)NM         -8.068e+05  4.932e+05  -1.636  0.10426
factor(STATE)NV         -1.413e+06  5.688e+05  -2.484  0.01425 *
factor(STATE)NY         -1.426e+06  5.783e+05  -2.466  0.01495 *
factor(STATE)OH         -5.193e+05  4.176e+05  -1.243  0.21591
factor(STATE)OK         -1.373e+06  4.901e+05  -2.801  0.00585 **
```

```
factor(STATE)OR                 -2.299e+05  4.940e+05  -0.465  0.64242
factor(STATE)PA                 -7.638e+05  4.655e+05  -1.641  0.10323
factor(STATE)SC                 -1.403e+06  4.989e+05  -2.812  0.00567 **
factor(STATE)TN                 -1.280e+06  4.363e+05  -2.934  0.00394 **
factor(STATE)TX                 -7.768e+05  4.207e+05  -1.847  0.06704 .
factor(STATE)UT                 -1.153e+06  4.448e+05  -2.593  0.01060 *
factor(STATE)VA                 -1.300e+06  5.054e+05  -2.572  0.01122 *
factor(STATE)WA                 -2.614e+05  4.259e+05  -0.614  0.54043
factor(STATE)WI                 -1.056e+06  5.670e+05  -1.862  0.06487 .
factor(REGION)ESC                       NA         NA      NA       NA
factor(REGION)MA                        NA         NA      NA       NA
factor(REGION)SA                        NA         NA      NA       NA
factor(REGION)WM                        NA         NA      NA       NA
factor(REGION)WNC                       NA         NA      NA       NA
factor(REGION)WP                        NA         NA      NA       NA
factor(REGION)WSC                       NA         NA      NA       NA
monthsofage                      2.660e+04  4.482e+03   5.936 2.44e-08 ***
factor(DENSITY_CLASS)3          -2.608e+05  1.026e+05  -2.541  0.01221 *
factor(DENSITY_CLASS)4          -2.204e+05  1.210e+05  -1.822  0.07077 .
SQFT                             9.095e+01  1.357e+02   0.670  0.50384
NBR_MACHINES                     5.627e+04  3.090e+04   1.821  0.07081 .
PARTY_ROOMYES                   -5.800e+04  1.002e+05  -0.579  0.56360
PARTY_ROOMYES - 2               -2.865e+05  2.851e+05  -1.005  0.31664
factor(PATIO)Y                  -1.235e+05  9.204e+04  -1.342  0.18183
factor(BUILDING_TYPE)IN-LINE     9.271e+04  7.249e+04   1.279  0.20317
factor(BUILDING_TYPE)STAND ALONE 2.004e+05  2.970e+05   0.675  0.50097
factor(HIGH_VISIBILITY)Y         1.424e+05  7.472e+04   1.905  0.05890 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 394000 on 132 degrees of freedom
Multiple R-squared:  0.5102,Adjusted R-squared:  0.3581
F-statistic: 3.354 on 41 and 132 DF,  p-value: 8.459e-08
```

```
# =====================
# Stepwise Regression
# =====================

library(MASS)
fit.step <- stepAIC(fit.cat, direction="backward", k=log(nrow(dataset2)))
fit.step$anova
summary(fit.step)

library(stargazer)
stargazer(fit.step)
> summary(fit.step)

Call:
lm(formula = SALES.2015 ~ monthsofage + factor(HIGH_VISIBILITY),
data = dataset2)

Residuals:
Min      1Q  Median      3Q     Max
-985234 -303375  -90847  173789 1832329
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)                325464     102468   3.176  0.00177 **
monthsofage                 16124       3026   5.328 3.09e-07 ***
factor(HIGH_VISIBILITY)Y   195214      69737   2.799  0.00571 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 448800 on 171 degrees of freedom
Multiple R-squared:  0.177,Adjusted R-squared:  0.1674
F-statistic: 18.39 on 2 and 171 DF,  p-value: 5.82e-08
```

```
> lasso.mod=glmnet(x[train,],y[train],alpha=1,lambda=grid)
> plot(lasso.mod)
> set.seed(1)
> cv.out=cv.glmnet(x[train,],y[train],alpha=1)
> plot(cv.out)
> bestlam=cv.out$lambda.min
> lasso.pred=predict(lasso.mod,s=bestlam,newx=x[test,])
> mean((lasso.pred-y.test)^2)
[1] 0.9383676
> out=glmnet(x,y,alpha=1,lambda=grid)
> lasso.coef=predict(out,type="coefficients",s=bestlam)[1:20,]
> lasso.coef
(Intercept)              monthsofage factor(HIGH_VISIBILITY)Y
```

```
-5.741237e-17             2.725107e-01             1.149364e-01
 CMDSC_PHARMACY_1RO              CM_MALL_1RO       CNT_MALLS_100K_1RO
-3.579293e-02             0.000000e+00             0.000000e+00
 CNT_MALLS_300K_0_5RO     CNT_MALLS_100K_0_5RO            CX01V037_8TO
 8.568517e-02             0.000000e+00             0.000000e+00
 CX01V037_0_5RO            CYA04V001_8TO            CYA04V002_8TO
 0.000000e+00             0.000000e+00             0.000000e+00
 CYA04V003_8TO             CYA04V004_8TO            CYA08V003_8TO
 0.000000e+00             0.000000e+00             0.000000e+00
 CYA08V003_0_5RO           CYB05V002_15TO             CYB05V003_1RO
 0.000000e+00             0.000000e+00             0.000000e+00
 CYB05V003_8TO          CYB05V003_0_5RO
-1.632991e-02             0.000000e+00
> lasso.coef[lasso.coef!=0]
(Intercept)              monthsofage factor(HIGH_VISIBILITY)Y
-5.741237e-17             2.725107e-01             1.149364e-01
 CMDSC_PHARMACY_1RO     CNT_MALLS_300K_0_5RO            CYB05V003_8TO
-3.579293e-02             8.568517e-02            -1.632991e-02
```