STAT 350 Project

David Wang

301293572

Group 19

1. Introduction

The goal of the project is to build a model in order to predict the unit sales of child car seats at each location. To do so the following steps must be done: checking model adequacy, filtering unimportant variables, building models and choosing the best model. The desired outcome is a model that is most accurate in predicting the response.

The dataset contains the following variables.

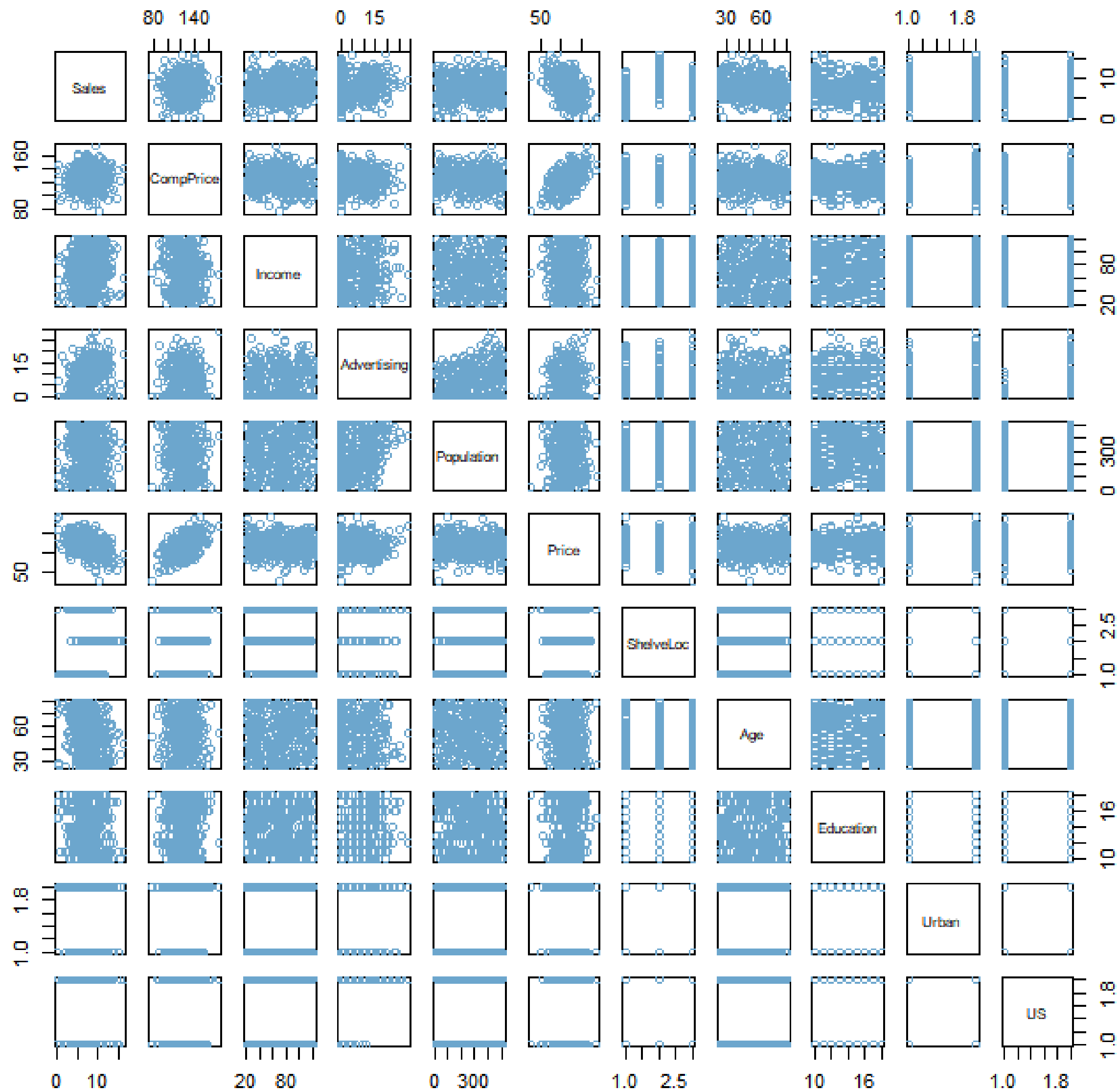
	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US
1	9.50	138	73	11	276	120	Bad	42	17	Yes	Yes
2	11.22	111	48	16	260	83	Good	65	10	Yes	Yes
3	10.06	113	35	10	269	80	Medium	59	12	Yes	Yes
4	7.40	117	100	4	466	97	Medium	55	14	Yes	Yes
5	4.15	141	64	3	340	128	Bad	38	13	Yes	No
6	10.81	124	113	13	501	72	Bad	78	16	No	Yes

Quantitative variables	Categorical variables
Sales – Unit sales (in thousands) at each location	ShelveLoc – Bad, Good, Medium quality of shelving location
CompPrice – Price charged by competitor at each location	Urban – No, yes for store in urban or rural location
Income – Community income level (thousands)	US – No, yes for store in US or not
Advertising – Budget at each location (thousands)	
Population – Size in region (thousands)	
Price – Price for seats at each site	
Age – Average age of local population	
Education – Education level at each location	

The response variable is the unit sales (in thousands) of car seats measured at 400 stores. Each observation is one store.

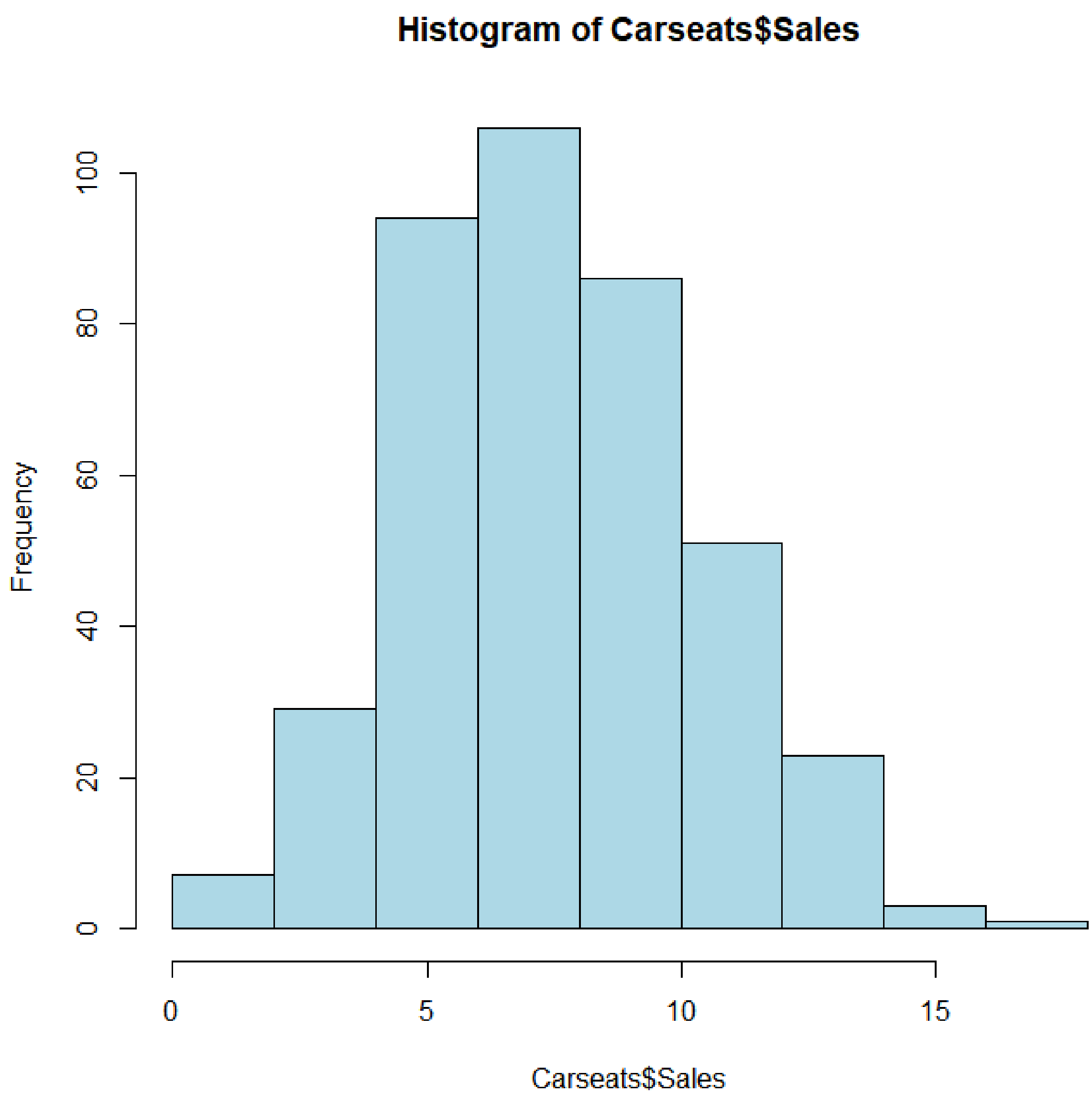
2. Preliminary testing and model adequacy

Multiple scatterplots of Variables



From the scatterplot there seem to be no linear relationship or a very weak linear relationship between most of the explanatory variables and the response variable Sales. There is a possible linear relationship between ‘Sales’ and ‘Price’ of car seats. There is no obvious outlier shown in the plots. Multicollinearity might be present due to the possible relationships between categorical variables. There are no obvious curvilinear relationships between the variables.

Histogram of Response variable Sales



The histogram of the response variable appears to be relatively normal. There are no obvious outliers in this plot.

summary(fit)

```
> summary(fit)

Call:
lm(formula = Sales ~ ., data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8692 -0.6908  0.0211  0.6636  3.4115 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
Income           0.0158028  0.0018451   8.565 2.58e-16 ***
Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
Population       0.0002079  0.0003705   0.561    0.575    
Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
Education       -0.0211018  0.0197205  -1.070    0.285    
UrbanYes         0.1228864  0.1129761   1.088    0.277    
USYes           -0.1840928  0.1498423  -1.229    0.220    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.019 on 388 degrees of freedom
Multiple R-squared:  0.8734,    Adjusted R-squared:  0.8698 
F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

The overall F test is significant at alpha = 0.05 as $F_0 = 243.4$ with a p-value of < 2.2e-16. At least one of the explanatory variables have a significant effect on the unit sales of car seats at each location.

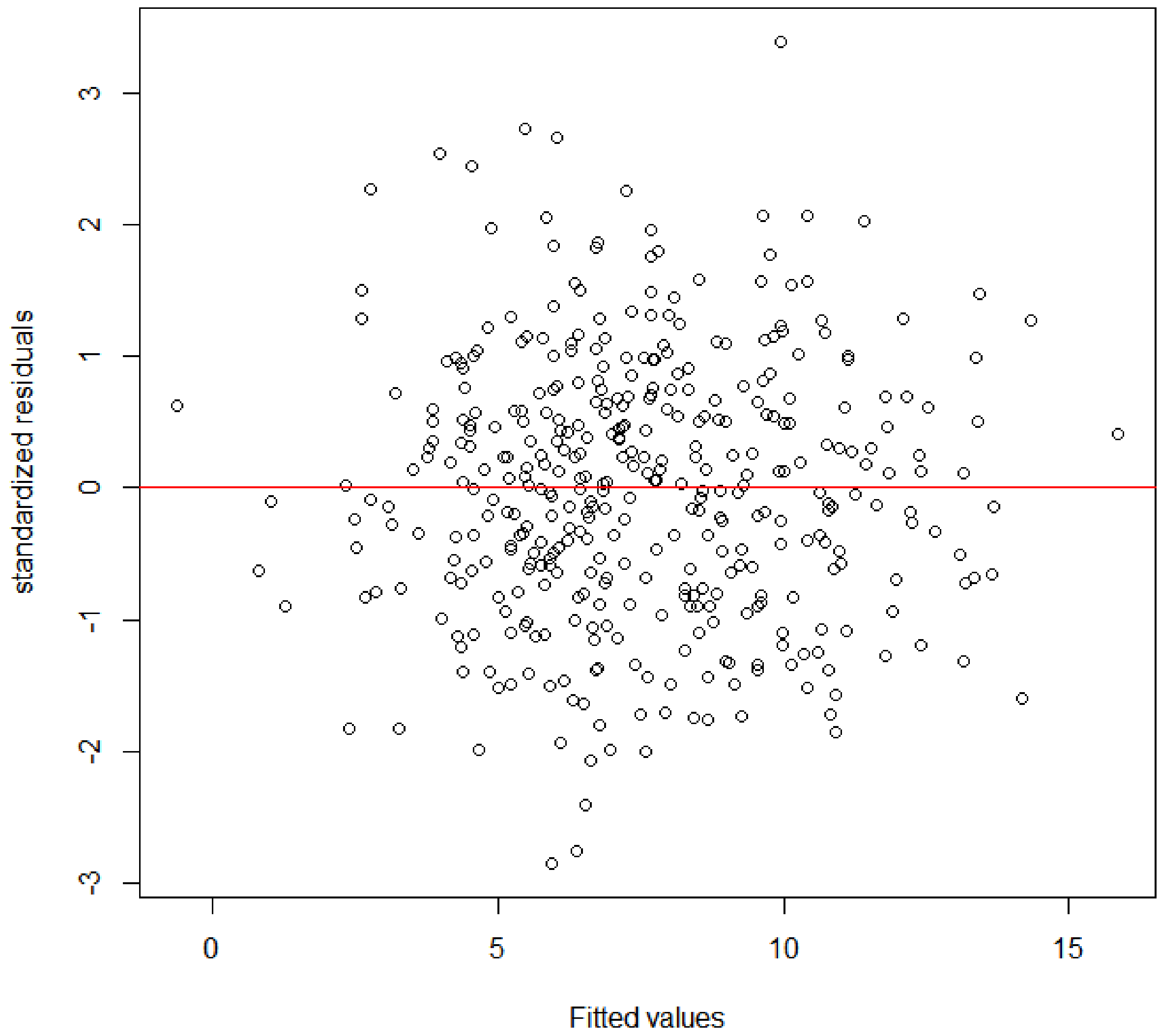
anova(fit)

```
> anova(fit)
Analysis of Variance Table

Response: Sales
             Df  Sum Sq Mean Sq   F value    Pr(>F)    
CompPrice     1   13.07   13.07   12.5855 0.0004363 ***
Income        1   79.07   79.07   76.1616 < 2.2e-16 ***
Advertising   1  219.35  219.35  211.2741 < 2.2e-16 ***
Population    1    0.38    0.38    0.3683 0.5442756    
Price         1 1198.87 1198.87 1154.7211 < 2.2e-16 ***
ShelveLoc     2 1047.47  523.74  504.4519 < 2.2e-16 ***
Age           1  217.39  217.39  209.3831 < 2.2e-16 ***
Education     1    1.05    1.05    1.0117 0.3151346    
Urban         1    1.22    1.22    1.1753 0.2789892    
US            1    1.57    1.57    1.5094 0.2199750    
Residuals   388  402.83    1.04                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

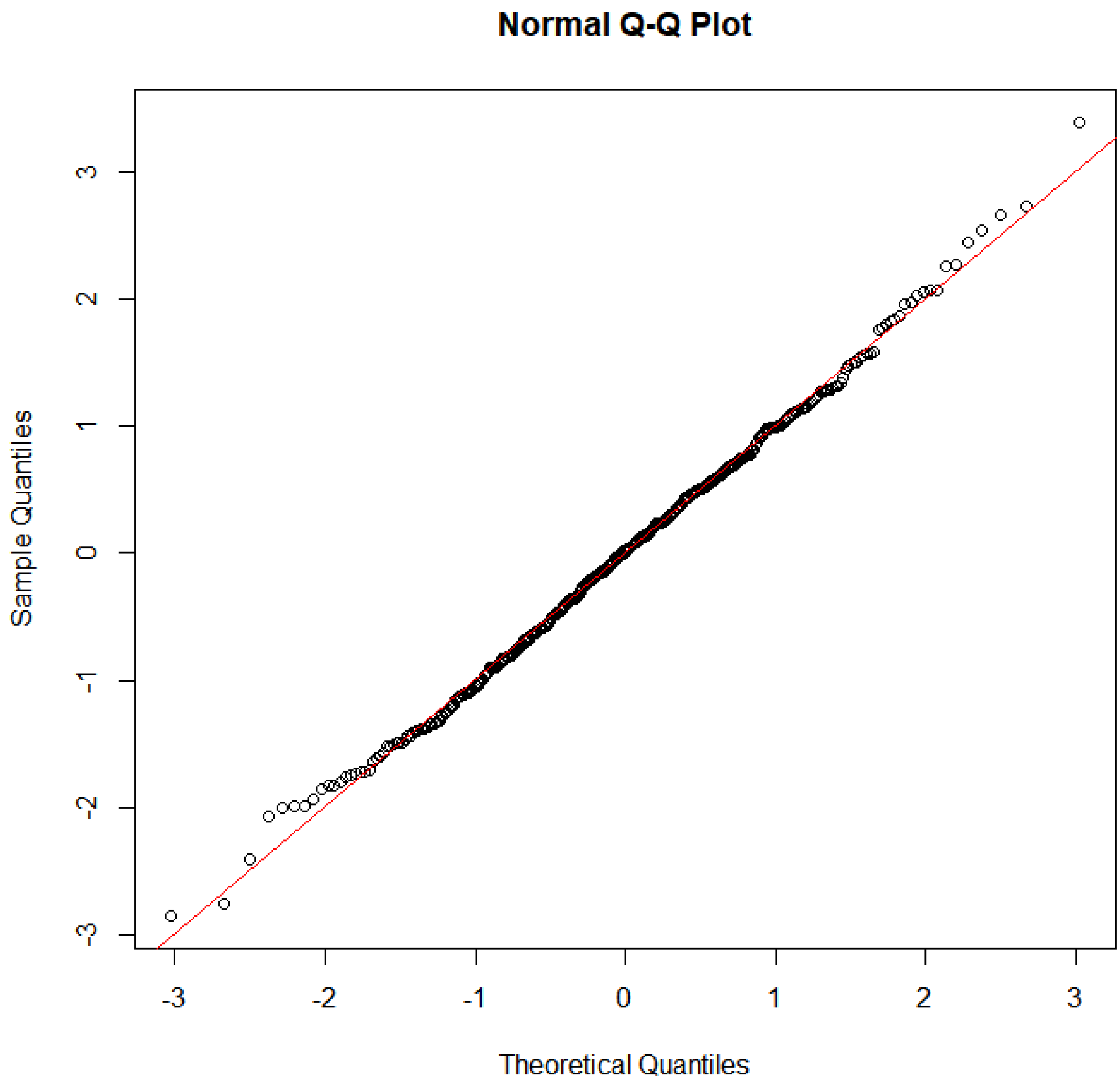
The explanatory variables deemed to have a significant effect on the unit sales of car seats at each location correspond with the results given by summary(fit).

Residuals of model



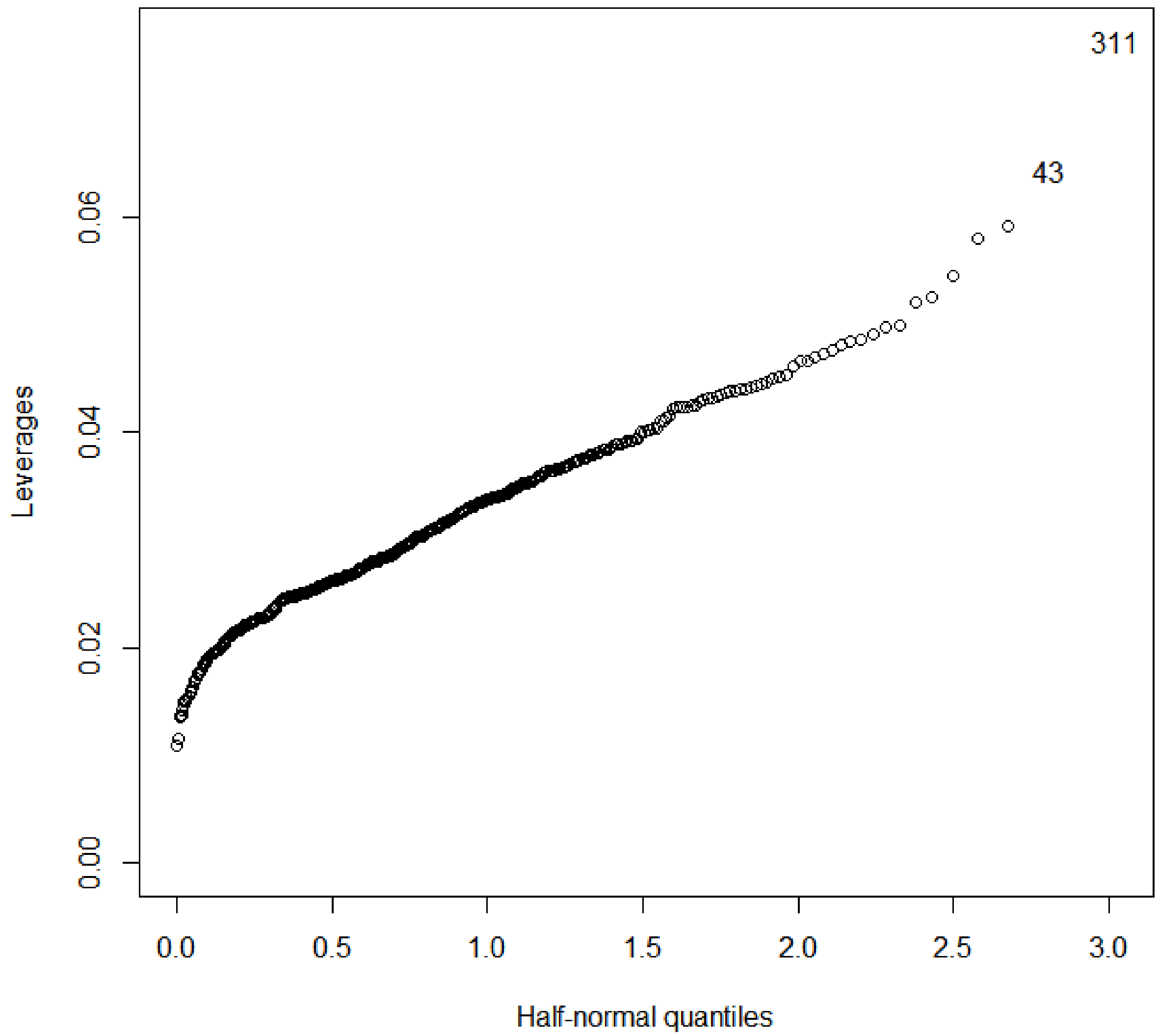
The residuals are from a model of a multiple linear regression with Sales as the response and the rest as explanatory variables. There seem to be no discernable pattern in the residuals plot. The plot shows that the variances are constant and the assumption of linearity is mostly satisfied. No transformation of the response is needed. There is one potential outlier shown in this plot.

Normal Q-Q Plot

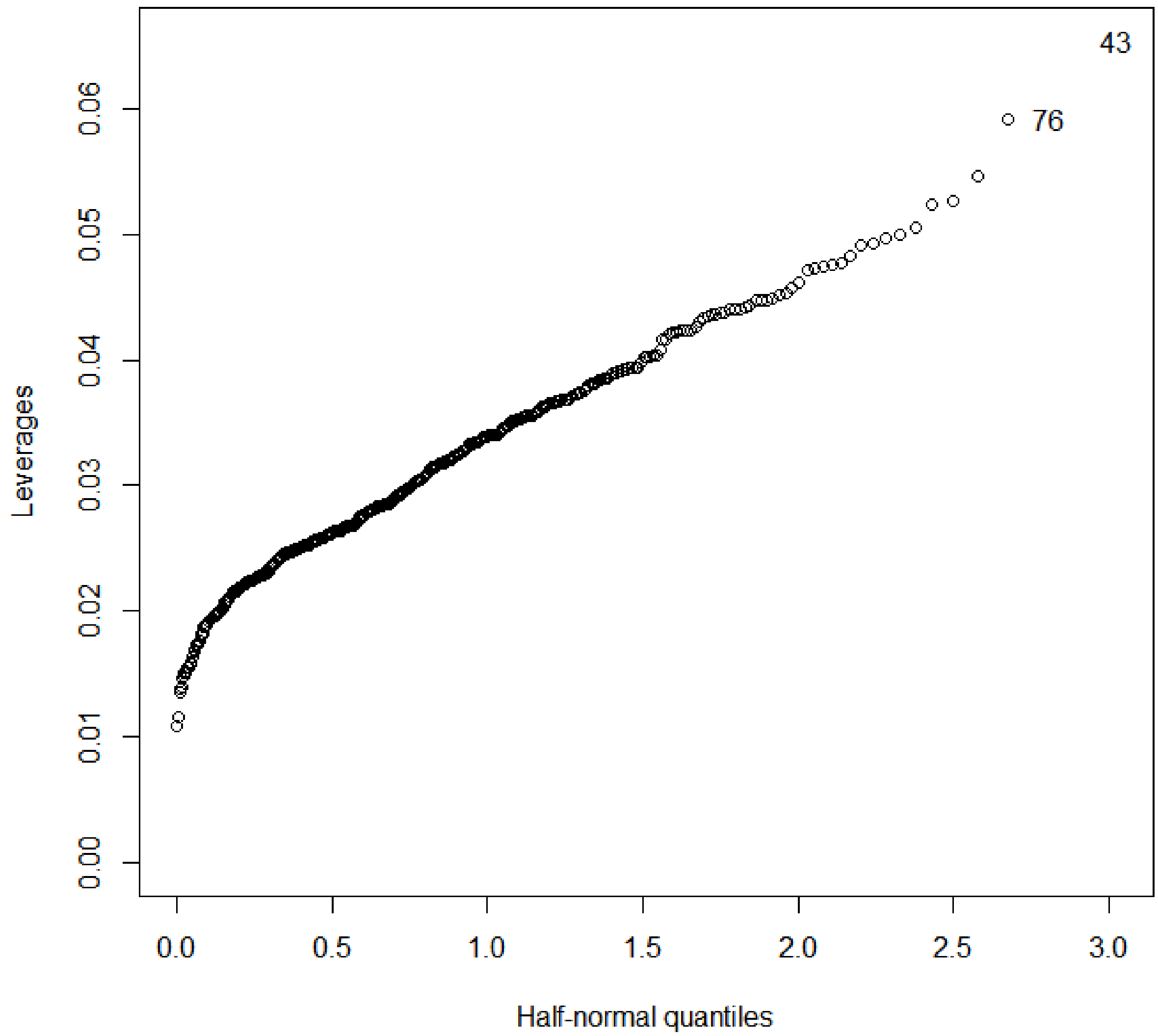


The normal Q-Q plot shows no evidence of non-normality. The residuals and normal Q-Q plot show that there is no evidence of non-constant variance or non-linearity. Again, a potential outlier can be seen. We can proceed with Multiple Linear Regression.

Checking for Leverages

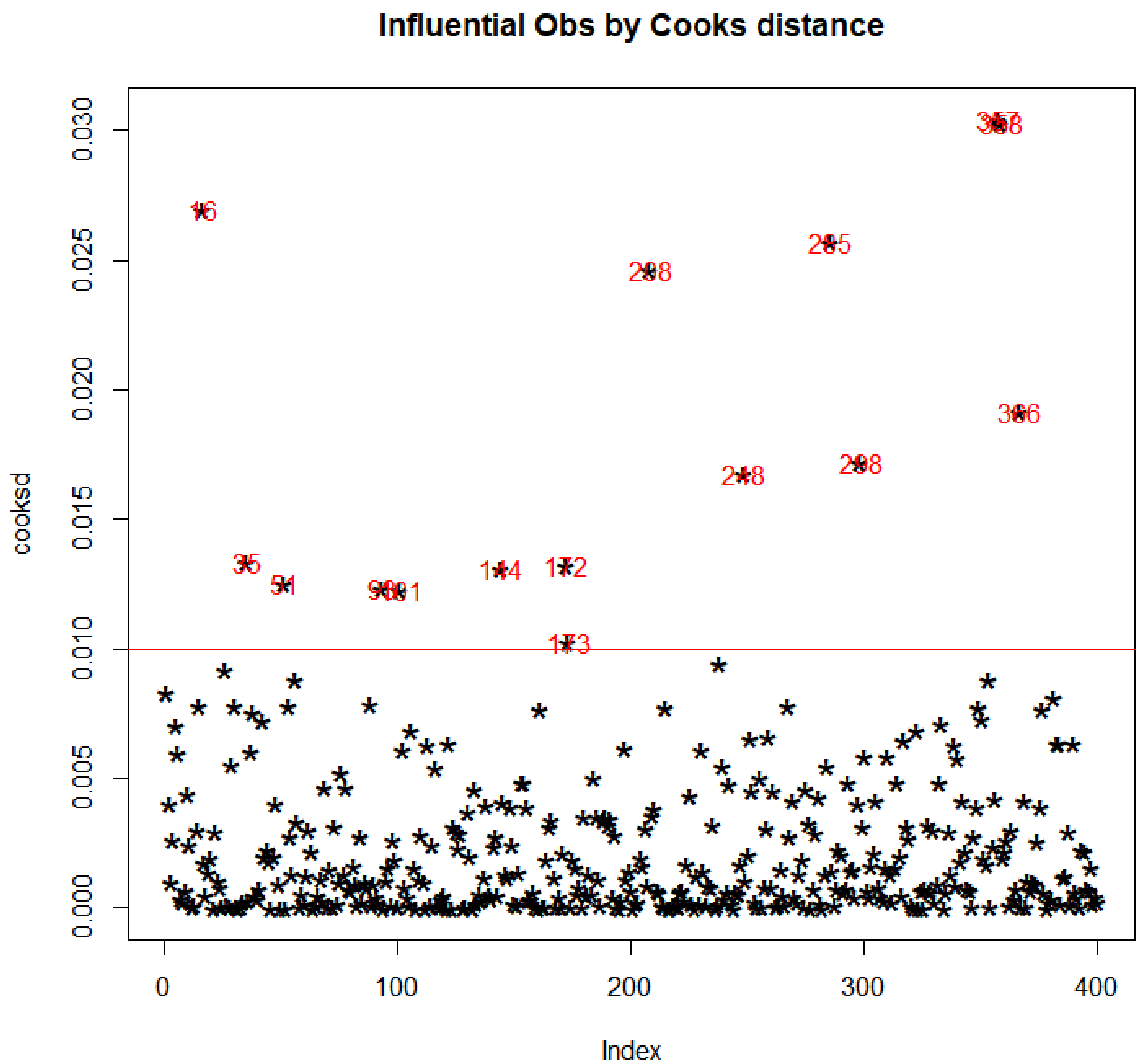


Store location #311 has an extreme leverage value. It is unclear whether location #43 is also extreme. The model needs to be refit without #311 and the above graph has to be reproduced.



After removing #311 and refitting the model the following graph is produced. The graph shows that #43 is not as extreme of a leverage value as that of #311 and therefore the point will not be removed.

Cooks' Distance check



Many more observations are identified as outliers. However, they do not match the result from half-normal plot. This shows that while #311 has a high leverage it is not necessarily an influential point.

Studentized Residuals check

```
> stud[which.max(abs(stud))]
     358
3.447684
```

Store location #358 has the largest residual. Using the general rule of thumb |d| > 3, and this shows that #358 is an outlier.

DFFITS

```
threshold1
[1] 0.3316625
```

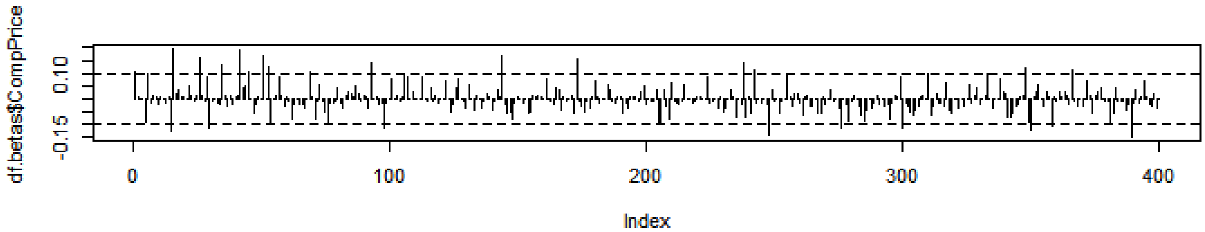
Sorting observations by dffits values

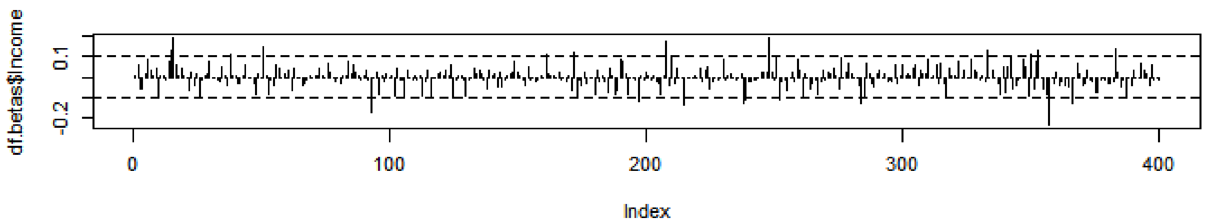
```
> df.fits[order(-df.fits['dffits(fit)']), ]
  [1]  0.6114979980  0.5737972000  0.5589772332  0.5483348198  0.4826552283  0.4511581414  0.3998285898  0.3372089259  0.3330468708
 [10]  0.3264311919  0.3263497925  0.3177732245  0.3135513266  0.3089471319  0.3069148079  0.3068499006  0.3060993436  0.3050862925
 [19]  0.2949473441  0.2802975234  0.2789391028  0.2763018667  0.2761941795  0.2761330397  0.2718285496  0.2707640677  0.2683906257
 [28]  0.2651296900  0.2642346309  0.2565295298  0.2554432611  0.2508020374  0.2414264747  0.2402789620  0.2363844737  0.2353662374
 [37]  0.2232583570  0.2228293110  0.2226535932  0.2181193500  0.2172869065  0.2074821940  0.2050241146  0.2048479087  0.2046384944
 [46]  0.2011945321  0.1995122993  0.1984575875  0.1966845378  0.1966194373  0.1945331790  0.1942239308  0.1938740031  0.1916678796
 [55]  0.1912188705  0.1908049431  0.1893711715  0.1878025140  0.1868725622  0.1855323342  0.1842613879  0.1832332091  0.1830582392
 [64]  0.1827459785  0.1821533334  0.1807019375  0.1717638389  0.1705502225  0.1702341564  0.1695362807  0.1639311614  0.1589867290
 [73]  0.1580286340  0.1577538454  0.1519478789  0.1508768342  0.1500595116  0.1425554187  0.1403221330  0.1382606878  0.1380479103
 [82]  0.1367722260  0.1357591690  0.1334377632  0.1328516567  0.1319002354  0.1289579443  0.1281292220  0.1264422942  0.1239717150
 [91]  0.1235660040  0.1221919480  0.1201800015  0.1185213644  0.1184245727  0.1149126234  0.1130593475  0.1100561254  0.1099127445
[100]  0.1098865134  0.1078288157  0.1066332037  0.1058049741  0.1040806805  0.1015706059  0.1008013819  0.0998745526  0.0986574464
[109]  0.0975038207  0.0952733616  0.0944811769  0.0938110832  0.0937164739  0.0934135538  0.0912933017  0.0902660327  0.0900057386
[118]  0.0899253211  0.0897385524  0.0897284649  0.0883802106  0.0880997833  0.0878020572  0.0868150025  0.0855918490  0.0852119927
[127]  0.0850835395  0.0837218226  0.0816434757  0.0777120843  0.0774791516  0.0773668298  0.0768439348  0.0763360088  0.0755675777
[136]  0.0752335827  0.0752269791  0.0750012285  0.0739647014  0.0733799009  0.0714552342  0.0667712385  0.0666538702  0.0659078632
[145]  0.0628322125  0.0591950757  0.0589695558  0.0577944412  0.0575223353  0.0567717504  0.0516147302  0.0507421926  0.0503068949
[154]  0.0501765959  0.0497668711  0.0496340697  0.0483381692  0.0479943411  0.0478965416  0.0471699610  0.0458813555  0.0456756767
[163]  0.0447520721  0.0409136285  0.0398524415  0.0388679344  0.0372089022  0.0365788132  0.0363711514  0.0351860797  0.0335284994
[172]  0.0326654403  0.0322679690  0.0313903593  0.0283369444  0.0257844062  0.0251121172  0.0251013224  0.0245351902  0.0237805068
[181]  0.0226244098  0.0223374297  0.0219857429  0.0202707129  0.0187806080  0.0185806278  0.0169962776  0.0149045699  0.0148313561
[190]  0.0137563817  0.0132007204  0.0112110815  0.0111361958  0.0088192908  0.0066269710  0.0059899572  0.0053842497  0.0050812527
[199]  0.0045706928  0.0040884752  0.0025958020  0.0024392280 -0.0007178068 -0.0017825864 -0.0020395101 -0.0022643309 -0.0030402943
[208] -0.0038104861 -0.0045554510 -0.0054948525 -0.0074475269 -0.0076349020 -0.0080502935 -0.0090870197 -0.0121338246 -0.0123448467
[217] -0.0138103085 -0.0173349319 -0.0198736459 -0.0234624993 -0.0249232290 -0.0252642254 -0.0255579095 -0.0260845691 -0.0267547894
[226] -0.0268503989 -0.0279414885 -0.0282292720 -0.0284499085 -0.0287074495 -0.0288687139 -0.0289254801 -0.0335630466 -0.0349453684
[235] -0.0350054352 -0.0351769994 -0.0352922790 -0.0355971493 -0.0360557470 -0.0363583384 -0.0380462280 -0.0389368228 -0.0401769847
[244] -0.0429071621 -0.0434536941 -0.0481026167 -0.0498134185 -0.0518179332 -0.0527602490 -0.0544975335 -0.0554666247 -0.0593685090
[253] -0.0610636473 -0.0635221355 -0.0646040380 -0.0655810698 -0.0657262721 -0.0677831645 -0.0689480784 -0.0693867747 -0.0699782533
[262] -0.0705730447 -0.0707693857 -0.0717727737 -0.0738597145 -0.0802207708 -0.0820970061 -0.0826781659 -0.0828099261 -0.0843169058
[271] -0.0858046601 -0.0876924046 -0.0878037898 -0.0906475322 -0.0914450389 -0.0926402269 -0.0944987220 -0.0950421837 -0.0953989816
[280] -0.0963369386 -0.0977561401 -0.0985113842 -0.0985336963 -0.0993146592 -0.1020362862 -0.1043921312 -0.1055329114 -0.1086210828
[289] -0.1108762924 -0.1133217466 -0.1140497103 -0.1145912881 -0.1150773470 -0.1169583629 -0.1184193097 -0.1213607143 -0.1240675655
[298] -0.1246217035 -0.1248513231 -0.1253568159 -0.1260186506 -0.1265597252 -0.1301390189 -0.1320526921 -0.1321909595 -0.1353719456
[307] -0.1369340850 -0.1370669996 -0.1390909699 -0.1392846129 -0.1425865933 -0.1431495977 -0.1441063146 -0.1479910399 -0.1483849028
[316] -0.1491028599 -0.1511285247 -0.1522060248 -0.1526160508 -0.1534113418 -0.1544345173 -0.1544773341 -0.1547343698 -0.1559111872
[325] -0.1606303561 -0.1611363560 -0.1618621858 -0.1628906433 -0.1631274483 -0.1651881092 -0.1654730642 -0.1664996605 -0.1697881236
[334] -0.1763140270 -0.1767372078 -0.1768215777 -0.1780534701 -0.1812738431 -0.1829044318 -0.1878657381 -0.1884363549 -0.1890689147
[343] -0.1902882138 -0.1905658055 -0.1934140632 -0.1935505425 -0.1953469787 -0.2021282059 -0.2040566517 -0.2104517276 -0.2139215343
[352] -0.2160192158 -0.2167681635 -0.2170389095 -0.2191946683 -0.2194550368 -0.2197189306 -0.2220408456 -0.2235417550 -0.2243574300
[361] -0.2262762217 -0.2277645188 -0.2301202310 -0.2329716617 -0.2340091305 -0.2347001976 -0.2358730723 -0.2407116434 -0.2409553346
[370] -0.2412010508 -0.2419118433 -0.2464213119 -0.2464281328 -0.2568287047 -0.2584539525 -0.2658925391 -0.2690014783 -0.2717776206
[379] -0.2760260109 -0.2763699507 -0.2774116856 -0.2814255838 -0.2874846674 -0.2875166683 -0.2923453016 -0.2931358062 -0.2967562066
[388] -0.3018973085 -0.3036295058 -0.3049550950 -0.3050078565 -0.3066081193 -0.3520240556 -0.3862970670 -0.3868021412 -0.3890448698
[397] -0.3976402560 -0.4016808641 -0.4584110483 -0.6097116298
```

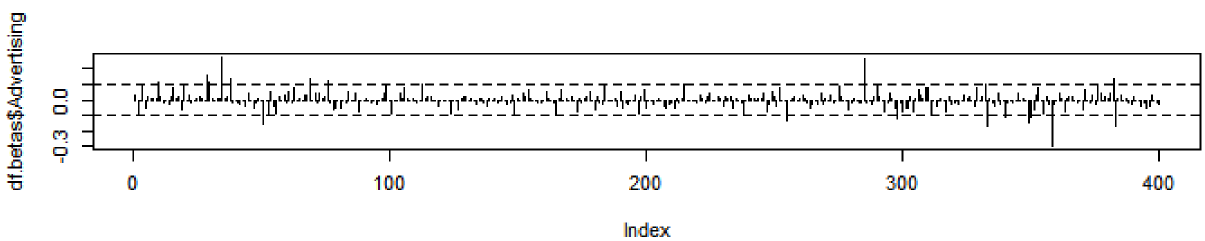
A lot of observations are identified as outliers. This could be because of the large sample size.

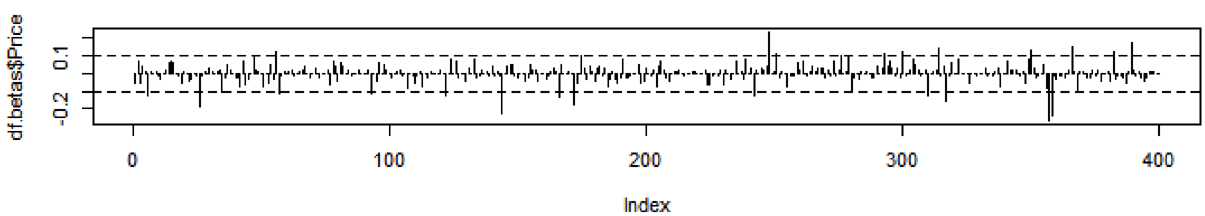
DFBETAS

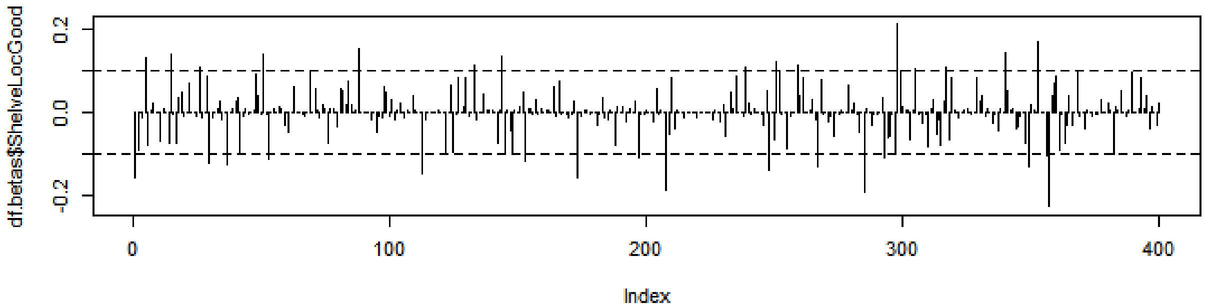
```
threshold2
[1] 0.1
```

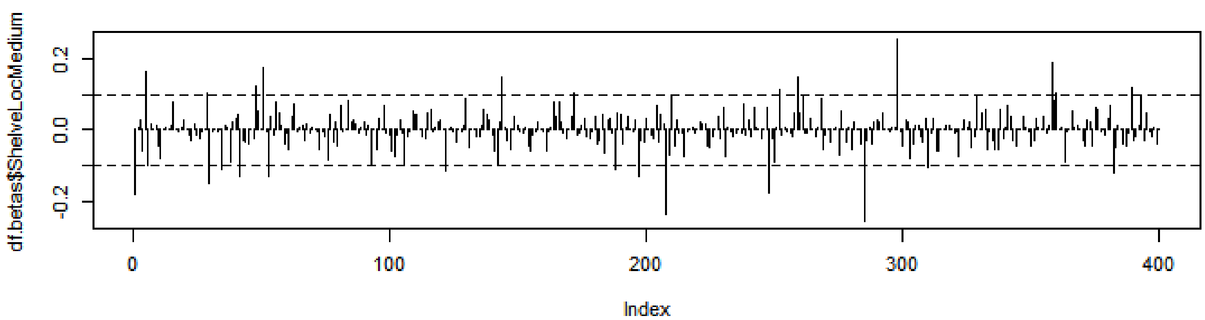

df.betas$CompPrice
-0.15
0.10
0
100
200
300
400
Index

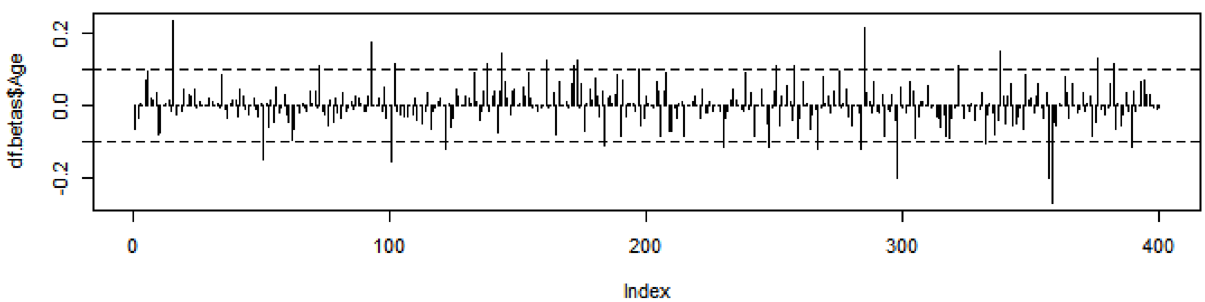

df.betas$Income
-0.2
0.1
0
100
200
300
400
Index


df.betas$Advertising
-0.3
0.0
0
100
200
300
400
Index


df.betas$Price
-0.2
0.1
0
100
200
300
400
Index







Based on the results of summary(fit) and anova(fit), the most important variables are plotted for DFBETAS. Most of the graphs show 2 outliers present at around #357 and #358 which coincide with the results by the Cooks' Distance. The threshold might be artificially deflated due the small number of predictors relative to the large sample size. The result of the graphs coincides with the results of the Cooks' Distance and therefore store locations #357 and #358 are identified as outliers and is removed from the dataset before variable selection could be done.

Variance Inflation Factor

```
vif(fit)
                GVIF Df GVIF^(1/(2*Df))
CompPrice   1.554618  1        1.246843
Income      1.024731  1        1.012290
Advertising 2.103136  1        1.450219
Population  1.145534  1        1.070296
Price       1.537068  1        1.239785
ShelveLoc   1.033891  2        1.008367
Age         1.021051  1        1.010471
Education   1.026342  1        1.013086
Urban       1.022705  1        1.011289
US          1.980720  1        1.407380
```

All of the variables have low <5 variance inflation factors. This disproves my conclusion from the scatterplot. The VIF results show that there is no evidence of multicollinearity.

3. Variable Selection

Summary(fit) and anova(fit)

From the results the variables ‘CompPrice’, ‘Income’, ‘Advertising’, ‘Price’, ‘ShelveLoc’, ‘Age’ and ‘Education’ are deemed to have a significant effect on the response variable.

Forward Selection

```
Subset selection object
Call: regsubsets.formula(Sales ~ ., data = Carseats2, nbest = 1, nvmax = 9,
    method = "forward")
11 Variables  (and intercept)
                Forced in Forced out
CompPrice           FALSE      FALSE
Income              FALSE      FALSE
Advertising         FALSE      FALSE
Population          FALSE      FALSE
Price               FALSE      FALSE
ShelveLocGood       FALSE      FALSE
ShelveLocMedium     FALSE      FALSE
Age                 FALSE      FALSE
Education           FALSE      FALSE
UrbanYes            FALSE      FALSE
USYes               FALSE      FALSE
1 subsets of each size up to 9
Selection Algorithm: forward
         CompPrice Income Advertising Population Price ShelveLocGood ShelveLocMedium Age Education UrbanYes USYes
1  ( 1 ) " "       " "    " "         " "        " "   "*"           " "             " " " "       " "      " "
2  ( 1 ) " "       " "    " "         " "        "*"   "*"           " "             " " " "       " "      " "
3  ( 1 ) "*"       " "    " "         " "        "*"   "*"           " "             " " " "       " "      " "
4  ( 1 ) "*"       " "    "*"         " "        "*"   "*"           " "             " " " "       " "      " "
5  ( 1 ) "*"       " "    "*"         " "        "*"   "*"           "*"             " " " "       " "      " "
6  ( 1 ) "*"       " "    "*"         " "        "*"   "*"           "*"             "*" " "       " "      " "
7  ( 1 ) "*"       "*"    "*"         " "        "*"   "*"           "*"             "*" " "       " "      " "
8  ( 1 ) "*"       "*"    "*"         " "        "*"   "*"           "*"             "*" " "       " "      "*"
9  ( 1 ) "*"       "*"    "*"         " "        "*"   "*"           "*"             "*" "*"       " "      "*"
```

‘Population’ and ‘Urban’ are tied as the least important variables, followed by ‘Education’, ‘US’ and ‘Income’, as they are the last to enter the model while ‘Population’ and ‘Urban’ do not enter the model at all.

Backward selection

```
> summ.mod.backward
Subset selection object
Call: regsubsets.formula(Sales ~ ., data = Carseats2, nbest = 1, nvmax = 9,
    method = "backward")
11 Variables  (and intercept)
                Forced in Forced out
CompPrice           FALSE      FALSE
Income              FALSE      FALSE
Advertising         FALSE      FALSE
Population          FALSE      FALSE
Price               FALSE      FALSE
ShelveLocGood       FALSE      FALSE
ShelveLocMedium     FALSE      FALSE
Age                 FALSE      FALSE
Education           FALSE      FALSE
UrbanYes            FALSE      FALSE
USYes               FALSE      FALSE
1 subsets of each size up to 9
Selection Algorithm: backward
         CompPrice Income Advertising Population Price ShelveLocGood ShelveLocMedium Age Education UrbanYes USYes
1  ( 1 ) " "       " "    " "         " "        " "   "*"           " "             " " " "       " "      " "  
2  ( 1 ) " "       " "    " "         " "        "*"   "*"           " "             " " " "       " "      " "  
3  ( 1 ) "*"       " "    " "         " "        "*"   "*"           " "             " " " "       " "      " "  
4  ( 1 ) "*"       " "    "*"         " "        "*"   "*"           " "             " " " "       " "      " "  
5  ( 1 ) "*"       " "    "*"         " "        "*"   "*"           "*"             " " " "       " "      " "  
6  ( 1 ) "*"       " "    "*"         " "        "*"   "*"           "*"             "*" " "       " "      " "  
7  ( 1 ) "*"       "*"    "*"         " "        "*"   "*"           "*"             "*" " "       " "      " "  
8  ( 1 ) "*"       "*"    "*"         " "        "*"   "*"           "*"             "*" " "       " "      "*"  
9  ( 1 ) "*"       "*"    "*"         " "        "*"   "*"           "*"             "*" "*"       " "      "*"  
```

The result of backwards selection is exactly the same as that of the forward selection.

All subset regression

```
Subset selection object
Call: regsubsets.formula(Sales ~ ., data = Carseats2, nbest = 1, nvmax = 9,
    method = "exhaustive")
11 Variables  (and intercept)
                Forced in Forced out
CompPrice           FALSE      FALSE
Income              FALSE      FALSE
Advertising         FALSE      FALSE
Population          FALSE      FALSE
Price               FALSE      FALSE
ShelveLocGood       FALSE      FALSE
ShelveLocMedium     FALSE      FALSE
Age                 FALSE      FALSE
Education           FALSE      FALSE
UrbanYes            FALSE      FALSE
USYes               FALSE      FALSE
1 subsets of each size up to 9
Selection Algorithm: exhaustive
         CompPrice Income Advertising Population Price ShelveLocGood ShelveLocMedium Age Education UrbanYes USYes
1  ( 1 ) " "       " "    " "         " "        " "   "*"           " "             " " " "       " "      " "  
2  ( 1 ) " "       " "    " "         " "        "*"   "*"           " "             " " " "       " "      " "  
3  ( 1 ) "*"       " "    " "         " "        "*"   "*"           " "             " " " "       " "      " "  
4  ( 1 ) "*"       " "    "*"         " "        "*"   "*"           " "             " " " "       " "      " "  
5  ( 1 ) "*"       " "    "*"         " "        "*"   "*"           "*"             " " " "       " "      " "  
6  ( 1 ) "*"       " "    "*"         " "        "*"   "*"           "*"             "*" " "       " "      " "  
7  ( 1 ) "*"       "*"    "*"         " "        "*"   "*"           "*"             "*" " "       " "      " "  
8  ( 1 ) "*"       "*"    "*"         " "        "*"   "*"           "*"             "*" " "       " "      "*"  
9  ( 1 ) "*"       "*"    "*"         " "        "*"   "*"           "*"             "*" "*"       " "      "*"  
```

The result of all subset regression is consistent with previous results. All three methods deem the quality of shelving location of car seats at each location as the most influential variable followed by the price of car seats being charged.

Fitting new linear models

New MLR models will be fitted with ‘Population’ and ‘Urban’ removed but separately exclude ‘Education’, ‘US’ and ‘Income’. The adjusted R^2, AIC and Press statistic will be calculated for each model.

```
mod2 <- lm(Sales ~. -Population -Urban, data=Carseats2)
> summary(mod2)$adj.r.squared
[1] 0.8747316
> extractAIC(mod2)
[1] 10.000000  5.242623
```

```
mod3 <- lm(Sales ~. -Population -Urban -Education,data=Carseats2)
> summary(mod3)$adj.r.squared
[1] 0.8744016
> extractAIC(mod3)
[1] 9.000000 5.314282
```

```
mod4 <- lm(Sales ~. -Population -Urban -US, data=Carseats2)
> summary(mod4)$adj.r.squared
[1] 0.8739263
> extractAIC(mod4)
[1] 9.000000 6.817574
```

```
mod5 <- lm(Sales ~. -Population -Urban -Income, data=Carseats2)
> summary(mod5)$adj.r.squared
[1] 0.8488189
> extractAIC(mod5)
[1]  9.00000 79.09912
```

```
mod6 <- lm(Sales ~. -Population -Urban -Education -US -Income, data=Ca
rseats2)
> summary(mod6)$adj.r.squared
[1] 0.8481327
> extractAIC(mod6)
[1]  7.00000 78.94239
```

All 5 models are very comparable and the differences are minute. The model that only excludes 'Population' and 'Urban' appears to be the best according to its highest adjusted R^2 value and the lowest AIC value. The AIC dramatically increased when 'Income' was removed from the model. This result contradicts with the results shown from subset and backwards/forwards selection. The previous tests showed that 'Income' was one of the lesser important variables but the AIC showed that removing 'Income' from the model is detrimental to the result.

Press Statistic

```
press1$stat
[1] 403.5527
> press2$stat
[1] 403.3295
> press3$stat
[1] 405.0966
> press4$stat
[1] 485.8136
> press5$stat
[1] 485.4826
```

The Press statistic is calculated from the models listed previously during the calculation of R^2 and AIC. The result coincides with what the AIC showed. The press statistic dramatically increased when 'Income' is removed. Both AIC and Press confirms that 'Income' is an influential variable. The press statistics for the other models are very close and therefore the model chosen by AIC will determine the best multiple linear regression model. Only the variables 'Population' and 'Urban' will be removed.

4. Model Building

The important variables as defined by earlier processes are as follows: CompPrice, Income, Advertising, Price, ShelveLoc, Age. Education and US are less important in comparison and do not make a huge effect on the model when they are removed. Population and Urban are deemed unimportant and removed.

The best MLR model will then be compared with Ridge regression and LASSO methods. Having multiple different methods allows for different perspectives on the issue of accurate prediction.

Ridge regression is similar to least squares but the method "shrinks parameter estimates" by "setting parameter values between 0 and the least squares estimate" (Loughin Lecture 6). The benefit of Ridge regression allows for a reduction of model complexity while also chasing individual errors less than MLR. Ridge regression requires tuning of λ. An optimal λ has to be found to balance between "increasing bias from shrinking parameters" and "decreasing variance by chasing errors less" (Loughin Lecture 6).

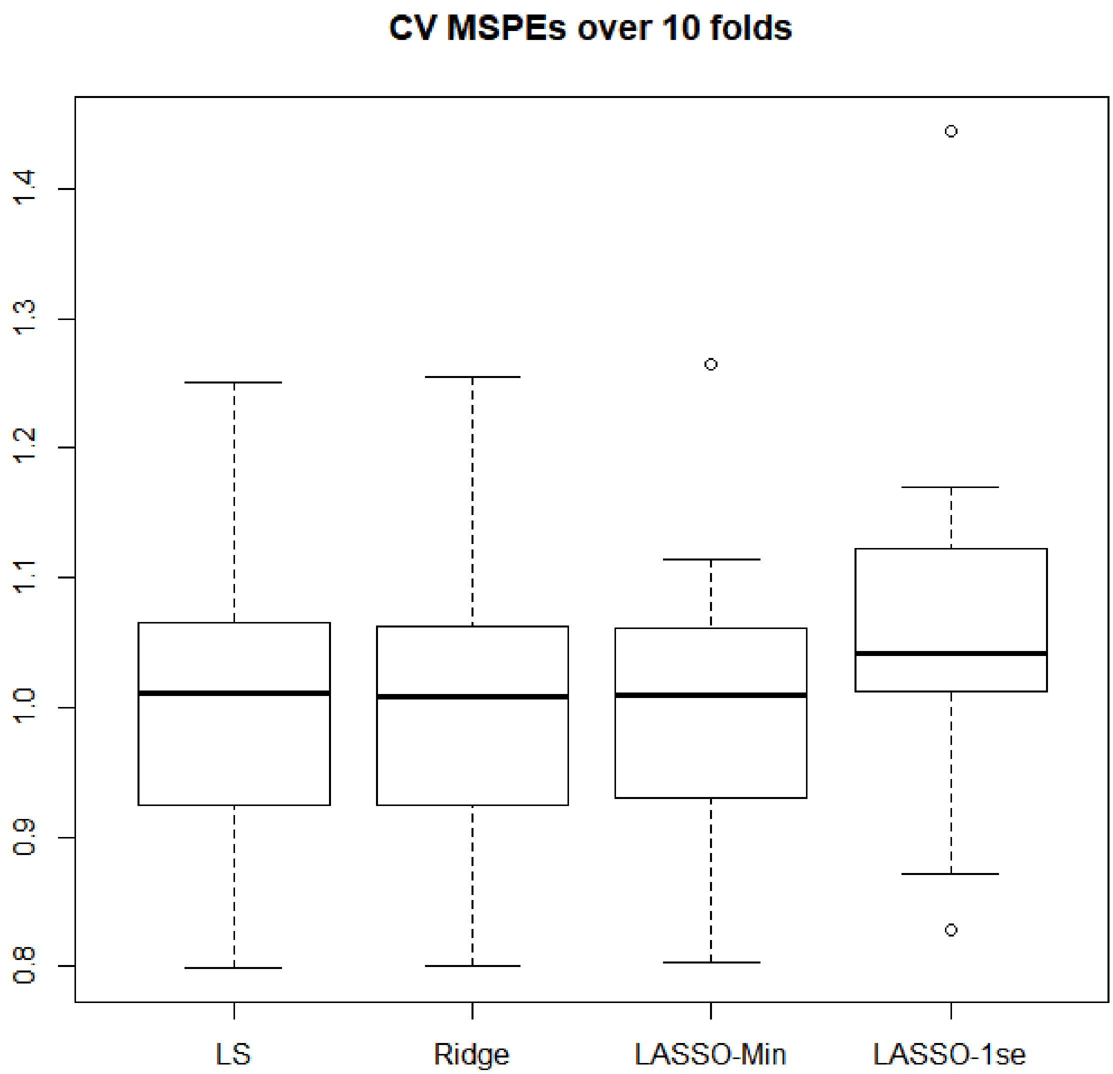
LASSO is an evolution of ridge regression as it also doubles as another method to choose variables. Similar to ridge regression the tuning of λ must be done. A large λ "shrinks parameter estimates more" and increases bias while decreasing variance (Loughin Lecture 6). A smaller λ does the opposite.

In order to tune λ, cross validation will be used. 10 folds will be done and in each fold the data will be split into training and valid sets. The model will then be built and trained through the training sets. The model will then predict the responses for the "deleted fold and compute the MSPE" (Loughin Lecture 3).

5. Model Selection

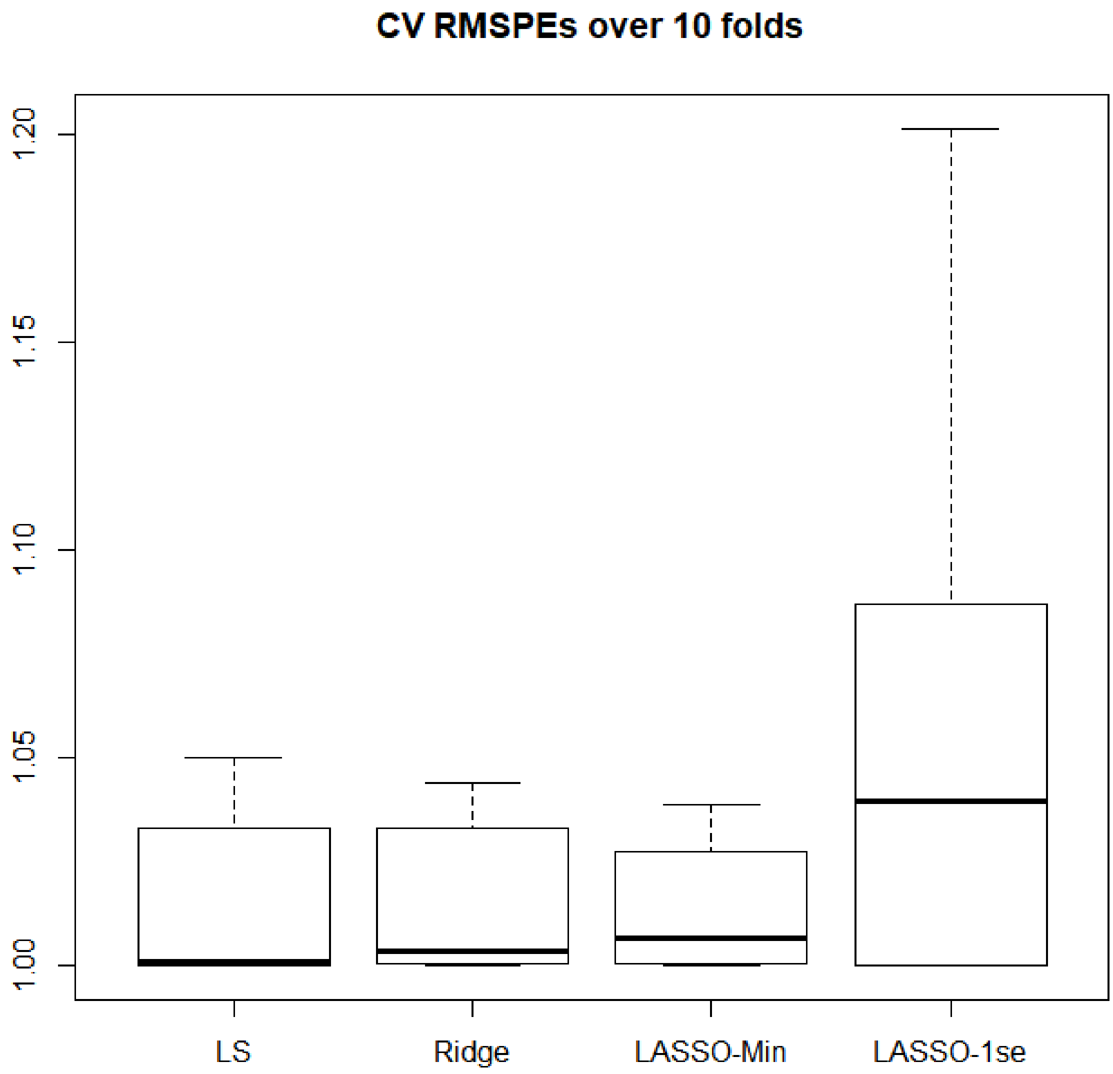
CV will produce a vector of MSPEs. The relative MSPEs will then be calculated and plotted into a boxplot. Each of the 4 methods produced RMSPEs is plotted and the lowest RSMPSE is deemed the best model. LASSO produces its list of important variables.

MSPE Boxplot

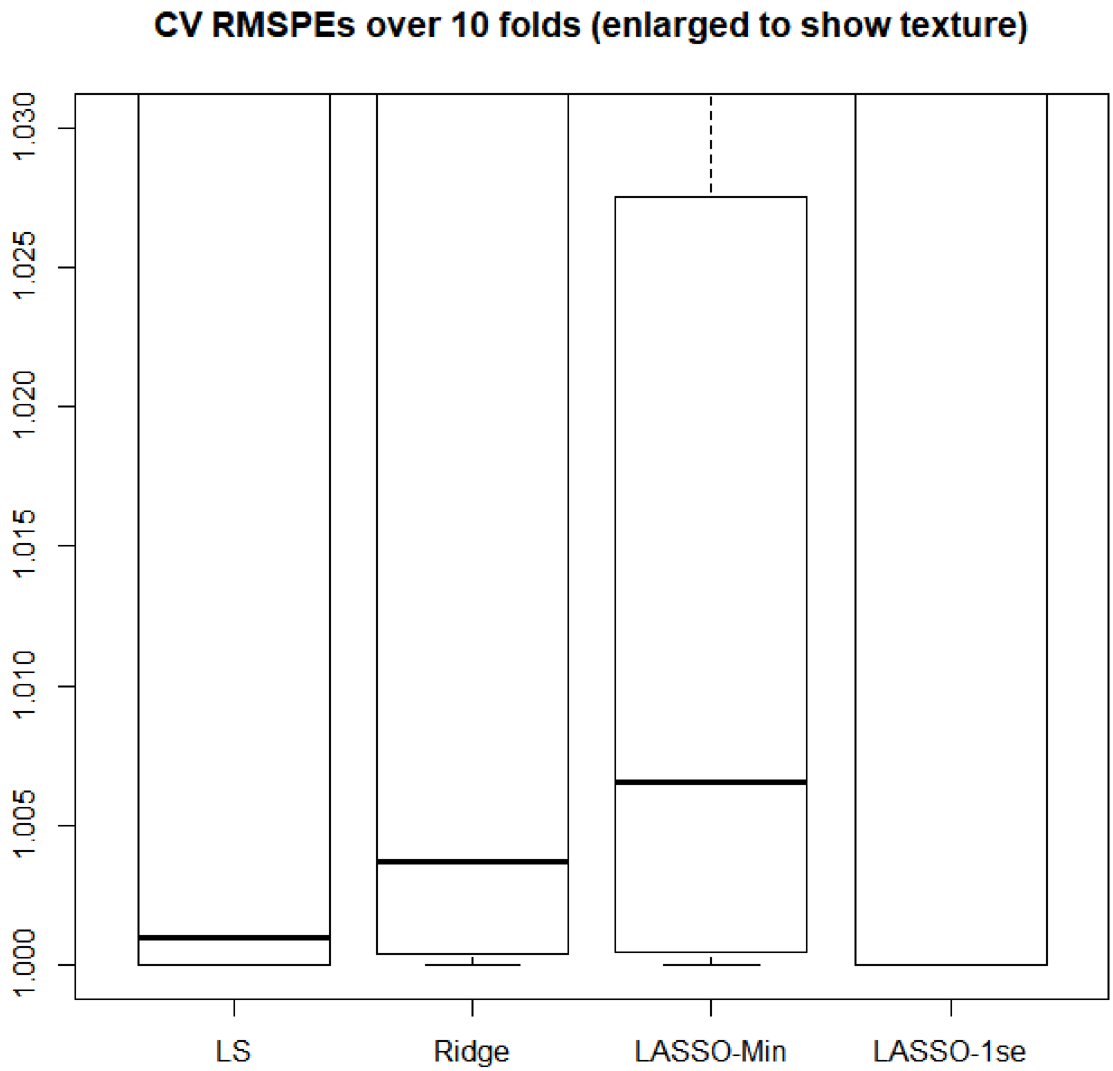


LS, Ridge and LASSO-min are fairly competitive models. LASSO-1se is trailing behind and is the worst performing model.

RMSPE Boxplot



The MLR, Ridge and LASSO-Min are similar in results. LASSO-1se is a clear loser. A closer look is required to differentiate between the former three models.



Looking at this graph the MLR model is the best performing model.

Let’s look at the variables chosen by LASSO

```
coef.LASSO.1se
10 x 1 sparse Matrix of class "dgCMatrix"
                             1
(Intercept)      5.956161996
CompPrice        0.085205886
Income           0.012675856
Advertising      0.110700639
Price           -0.088569861
ShelveLocGood    4.519441556
ShelveLocMedium  1.651528977
Age             -0.041944589
Education       -0.005764199
USYes            .
> coef.LASSO.min
10 x 1 sparse Matrix of class "dgCMatrix"
                           1
(Intercept)      5.77891584
CompPrice        0.09285372
Income           0.01492691
Advertising      0.13069006
Price           -0.09392213
ShelveLocGood    4.81184779
ShelveLocMedium  1.88155780
Age             -0.04523844
Education       -0.02652071
USYes           -0.25938401
```

The results of LASSO are consistent with my findings with variable selection from AIC and Press statistic. LASSO.1se chose to leave out ‘US’ while LASSO.min chose to include all variables. The result of the RSMPE graph showed that ‘US’ variable is still somewhat important as the LASSO.1se model performed the worst.

6. Conclusion

The work done on model adequacy and variable selection lead to the MLR model performing the best in comparison to ridge regression and LASSO. The assumptions required to execute the MLR model held up and the removal of ‘Population’ and ‘Urban’ reduced model complexity. Running summary(fit) again will allow for discussion of the variables.

```
> summary(fit.ls)

Call:
lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
    ShelveLoc + Age + Education + US, data = data.train)

Residuals:
     Min       1Q   Median       3Q      Max
-2.91084 -0.72547  0.00245  0.64349  2.67084

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.764373   0.599863   9.609  < 2e-16 ***
CompPrice        0.093502   0.004241  22.049  < 2e-16 ***
Income           0.015113   0.001885   8.016 1.66e-14 ***
Advertising      0.132756   0.010897  12.182  < 2e-16 ***
Price           -0.094361   0.002788 -33.850  < 2e-16 ***
ShelveLocGood    4.835570   0.156847  30.830  < 2e-16 ***
ShelveLocMedium  1.899541   0.128666  14.763  < 2e-16 ***
Age             -0.045493   0.003297 -13.799  < 2e-16 ***
Education       -0.028265   0.020221  -1.398   0.1631
USYes           -0.288748   0.152934  -1.888   0.0598 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.997 on 349 degrees of freedom
Multiple R-squared:  0.8796,    Adjusted R-squared:  0.8765
F-statistic: 283.3 on 9 and 349 DF,  p-value: < 2.2e-16
```

While the MLR model deems ‘Education’ and ‘US’ as variables that are not significant, the results of the Press Statistic and AIC showed that removing those variables do not improve the model in a significant margin. The result from LASSO showed that including those two variables do play a part in improving the model.

The result of ShelveLoc shows the importance of storing car seats in a good shelving location. Having them placed in good/medium quality increases the number of sales. ShelveLocBad not even being listed shows that it has a detrimental effect on unit sales. Holding all other variables constant, having good shelving location quality is associated to a 4.83-thousand-unit sales at each location while having medium shelving location quality is associated to a 1.89-thousand-unit sales at each location. The effect of the store location being in US or not plays a smaller role in comparison to that of shelving location.

For a new store owner, I would recommend him or her to focus on the quality of the shelving location of car seats to gain the most unit sales. The location of the store itself plays a small role in comparison to the price of the car seats and the price competitors are offering.

Future improvements to this project would be including more elaborate and complex models such as NNET and random forest. The tuning and testing required is beyond the scope of this project.

References

Dataset

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, https://www.statlearning.com, Springer-Verlag, New York

STAT 350 Tutorial 4,5,6,8,9,10

STAT 350 Lecture Notes

STAT 452 Lecture Notes by Professor Tom Loughin

- Lecture 3 – Evaluation models – Measuring model error
- Lecture 6 – Variable Selection: LASSO

Code for LASSO, Ridge and Cross-validation is from personal notes taken in STAT 452's tutorials.