Regression Homework 4
US Home Sales

Janet Ye: jy1151

April 16, 2014

1 Introduction

Number of home sales can be thought of as a demand curve. Basic economic intuition tells us that the demand of home sales is inversely related to the price of homes. In this analysis, we examine the relationship between number of US home sales and several predicting variables, namely, composite housing affordability index, private construction spending, median sales price, unemployment rate, and mortgage rate.

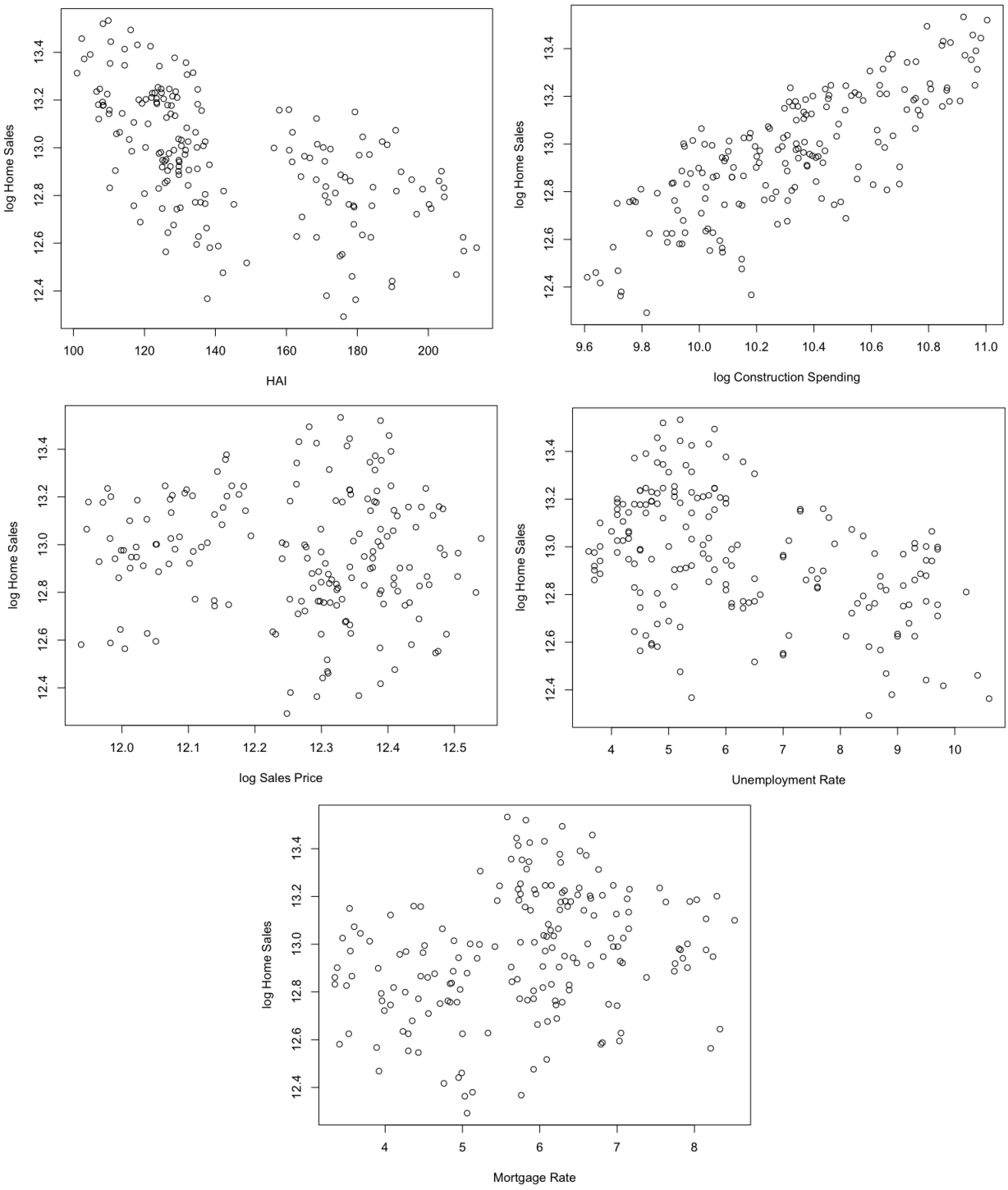
2 Data

All data used for this analysis are obtained from St. Louis FRED. The data are not previously seasonably adjusted, and measured on the first of each month ranging from January 1999 to February 2014. We begin by giving a short description about each variable.

- **Existing Home Sales**, or the target variable, are based on closing transactions of single-family, townhomes, condominiums and cooperative homes, measured in dollars.
- **Housing Affordability Index** measures the degree to which a typical family can afford the monthly mortgage payments on a typical home. Value of 100 means that a family with the median income has exactly enough income to qualify for a mortgage on a median-priced home. An index above 100 signifies that family earning the median income has more than enough income to qualify for a mortgage loan on a median-priced home, assuming a 20 percent down payment. For example, a composite housing affordability index (HAI) of 120.0 means a family earning the median family income has 120% of the income necessary to qualify for a conventional loan covering 80 percent of a median-priced existing single-family home. An increase in the HAI then shows that this family is more able to afford the median priced home.
- **Mortgage Rate** is the 30 year contract interest rates on commitments for fixed-rate first mortgages.
- **Construction Spending** is the total private construction spending on residential structures, measured in millions of dollars.
- **Median Sales Price, Unemployment Rate** are self-explanatory.

Since we are dealing with price, it makes economic sense to use to the log-log model. I logged the response variable – Existing Home Sales, since it is measured in dollars, as well as "Construction Spending" and "Median Sales Price", which are both measured in dollars. The other variables, "Housing Affordability Index", "Unemployment", and "Mortgage Rate" stay the same.

Below are scatter plots that show marginal relationship in the data.



There is a negative relationship between HAI and log Home Sales, given everything else in the data. One, however, expects this relationship to be positive, as if the more income one has (higher HAI), the more likely one would per chase a house. There also seems to be two subgroups in the data, one clustered below HAI of 140 and the other above 160.

There is a strong, positive, linear marginal relationship between log Construction Spending and log Home Sales. This makes sense: the more money spent on construction is associated with more home sales.

There is no apparent marginal relationship between log Sales Price and log Home Sales. One would expect a negative relationship – higher sales price is associated with lower home sales.

There is a weak, negative, linear marginal relationship between unemployment rate and log Home Sales. Higher Unemployment Rate is associated with lower Home Sales, which also makes sense.

Lastly, there is no apparent marginal relationship between Mortgage Rate and log Home Sales. One would expect a negative relationship, as higher mortgage rate discourages home sales.

3 Multiple Regression

Below is an output of the multiple regression model.

```
Call:
lm(formula = log(sales) ~ HAI + log(cons) + log(price) + unrate
    mort)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35868 -0.08465  0.00916  0.09850  0.28778

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.570272   1.969200   3.844 0.000169 ***
HAI          0.005719   0.001450   3.944 0.000116 ***
log(cons)    1.192932   0.085385  13.971  < 2e-16 ***
log(price)  -0.689319   0.110640  -6.230 3.33e-09 ***
unrate       0.062302   0.011472   5.431 1.84e-07 ***
mort         0.054664   0.031462   1.737 0.084056 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1303 on 176 degrees of freedom
Multiple R-squared:  0.7571,    Adjusted R-squared:  0.7502
F-statistic: 109.7 on 5 and 176 DF,  p-value: < 2.2e-16
```

```
> vif(d1.lm)
       HAI  log(cons) log(price)     unrate       mort
 19.361028   8.963405   3.121741   5.051954  16.469864
```

VIF for both HAI and mortgage are large. Below is a correlation matrix.

```
> cor(cbind(logsales,HAI,logcons,logprice,unrate,mort))
            logsales        HAI    logcons    logprice     unrate       mort
logsales  1.00000000 -0.5685299  0.7868646 -0.06151536 -0.4321349  0.2734973
HAI      -0.56852989  1.0000000 -0.8055555  0.31119256  0.8358359 -0.8139564
logcons   0.78686459 -0.8055555  1.0000000  0.10070178 -0.6828391  0.3935340
logprice -0.06151536  0.3111926  0.1007018  1.00000000  0.3805062 -0.6956739
unrate   -0.43213493  0.8358359 -0.6828391  0.38050623  1.0000000 -0.7932240
mort      0.27349728 -0.8139564  0.3935340 -0.69567394 -0.7932240  1.0000000
```

HAI has a strong negative correlation (-0.81) with log Construction Spending, a strong positive correlation (0.84) with unemployment rate, a strong negative correlation (-0.81) with mortgage rate.

Besides having a strong negative correlation with mortgage rate as mentioned above, mortgage has a strong negative (-0.79) correlation with unemployment rate.

We proceed to use best subset to determine the best model.

```
> leaps(cbind(HAI,logcons,logprice,unrate,mort),logsales,nbest=2)
$which
      1     2     3     4     5
1 FALSE  TRUE FALSE FALSE FALSE
1  TRUE FALSE FALSE FALSE FALSE
2 FALSE  TRUE FALSE  TRUE FALSE
2 FALSE  TRUE  TRUE FALSE FALSE
3 FALSE  TRUE  TRUE  TRUE FALSE
3  TRUE  TRUE  TRUE FALSE FALSE
4  TRUE  TRUE  TRUE  TRUE FALSE
4 FALSE  TRUE  TRUE  TRUE  TRUE
5  TRUE  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"
[6] "5"

$size
[1] 2 2 3 3 4 4 5 5 6

$Cp
[1]  97.912828 312.307069  84.900056  85.412663  24.637998  33.076202   7.018746
[8]  19.555118   6.000000
 > leaps(cbind(HAI,logcons,logprice,unrate,mort),logsales,nbest=2,method="r2")
 $which
       1     2     3     4     5
 1 FALSE  TRUE FALSE FALSE FALSE
 1  TRUE FALSE FALSE FALSE FALSE
 2 FALSE  TRUE FALSE  TRUE FALSE
 2 FALSE  TRUE  TRUE FALSE FALSE
 3 FALSE  TRUE  TRUE  TRUE FALSE
 3  TRUE  TRUE  TRUE FALSE FALSE
 4  TRUE  TRUE  TRUE  TRUE FALSE
 4 FALSE  TRUE  TRUE  TRUE  TRUE
 5  TRUE  TRUE  TRUE  TRUE  TRUE

 $label
 [1] "(Intercept)" "1"           "2"           "3"           "4"
 [6] "5"

 $size
 [1] 2 2 3 3 4 4 5 5 6

 $r2
 [1] 0.6191559 0.3232262 0.6398781 0.6391705 0.7258188 0.7141715 0.7528994
 [8] 0.7355953 0.7570662
> leaps(cbind(HAI,logcons,logprice,unrate,mort),logsales,nbest=2,method="adjr2")
$which
      1     2     3     4     5
1 FALSE  TRUE FALSE FALSE FALSE
1  TRUE FALSE FALSE FALSE FALSE
2 FALSE  TRUE FALSE  TRUE FALSE
2 FALSE  TRUE  TRUE FALSE FALSE
3 FALSE  TRUE  TRUE  TRUE FALSE
3  TRUE  TRUE  TRUE FALSE FALSE
4  TRUE  TRUE  TRUE  TRUE FALSE
4 FALSE  TRUE  TRUE  TRUE  TRUE
5  TRUE  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"
[6] "5"

$size
[1] 2 2 3 3 4 4 5 5 6

$adjr2
[1] 0.6170401 0.3194664 0.6358544 0.6351389 0.7211978 0.7093542 0.7473152
[8] 0.7296201 0.7501646
```

where 1 = HAI, 2 = log construction spending, 3 = log price, 4 = unemployment rate, 5 = mortgage

R-squared starts to level off at (value of 0.7528994) for the four-predictor model consisting of HAI, log Construction Spending, log Sales Price, and Unemployment Rate, but R-squared is at its highest for the model consisting of all five predictors. Adjusted R-squared is maximized at the five predictor model, a value of 0.75. Another criteria is that $C_p \leq p + 1 = 5 + 1 = 6$, where p is number of predictors. The five-predictor model satisfies this condition, with $C_p = 6$. One also wish to minimize C_p value, which is satisfied at the five-predictor model.

Below is output of AIC.

```
> extractAIC(lm(logsales~1))
[1]    1.0000 -488.4136
> extractAIC(lm(logsales~logcons))
[1]    2.0000 -662.1101
> extractAIC(lm(logsales~logcons+unrate))
[1]    3.0000 -670.2925
> extractAIC(lm(logsales~logcons+logprice+unrate))
[1]    4.0000 -717.9154
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate))
[1]    5.0000 -734.8423
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+mort))
[1]    6.0000 -735.9375
```

AIC is minimized at the five-predictor model, with AIC = -735.9375. We also wish to minimize AIC_C. Below are the outputs.

```
> n=length(logsales)
> extractAIC(lm(logsales~1))+2*2*3/(n-3)
[1]    1.067039 -488.346581
> extractAIC(lm(logsales~logcons))+2*3*4/(n-4)
[1]    2.134831 -661.975243
> extractAIC(lm(logsales~logcons+unrate))+2*4*5/(n-5)
[1]    3.225989 -670.066543
> extractAIC(lm(logsales~logcons+logprice+unrate))+2*5*6/(n-6)
[1]    4.340909 -717.574532
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate))+2*6*7/(n-7)
[1]    5.4800 -734.3623
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+mort))+2*7*8/(n-8)
[1]    6.643678 -735.293775
```

AIC_C is minimized for the five-predictor model, at -735.293775. Hence, the five-predictor model is a sensible choice. This is the same multiple regression model as before, and we list the output below again.

```
Call:
lm(formula = log(sales) ~ HAI + log(cons) + log(price) + unrate
    mort)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35868 -0.08465  0.00916  0.09850  0.28778

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.570272   1.969200   3.844 0.000169 ***
HAI          0.005719   0.001450   3.944 0.000116 ***
log(cons)    1.192932   0.085385  13.971  < 2e-16 ***
log(price)  -0.689319   0.110640  -6.230 3.33e-09 ***
unrate       0.062302   0.011472   5.431 1.84e-07 ***
mort         0.054664   0.031462   1.737 0.084056 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1303 on 176 degrees of freedom
Multiple R-squared:  0.7571,    Adjusted R-squared:  0.7502
F-statistic: 109.7 on 5 and 176 DF,  p-value: < 2.2e-16
```

The regression equation is

$$\begin{aligned} log \text{ Home Sales} = {} & 7.570272 + 0.005719 \times \text{HAI} + 1.192932 \times log \text{ Construction Spending} \\ & - 0.689319 \times log \text{ Median Sales Price} + 0.062302 \times \text{ Unemployment Rate} \\ & + 0.054664 \times \text{ Mortgage Rate} \end{aligned}$$

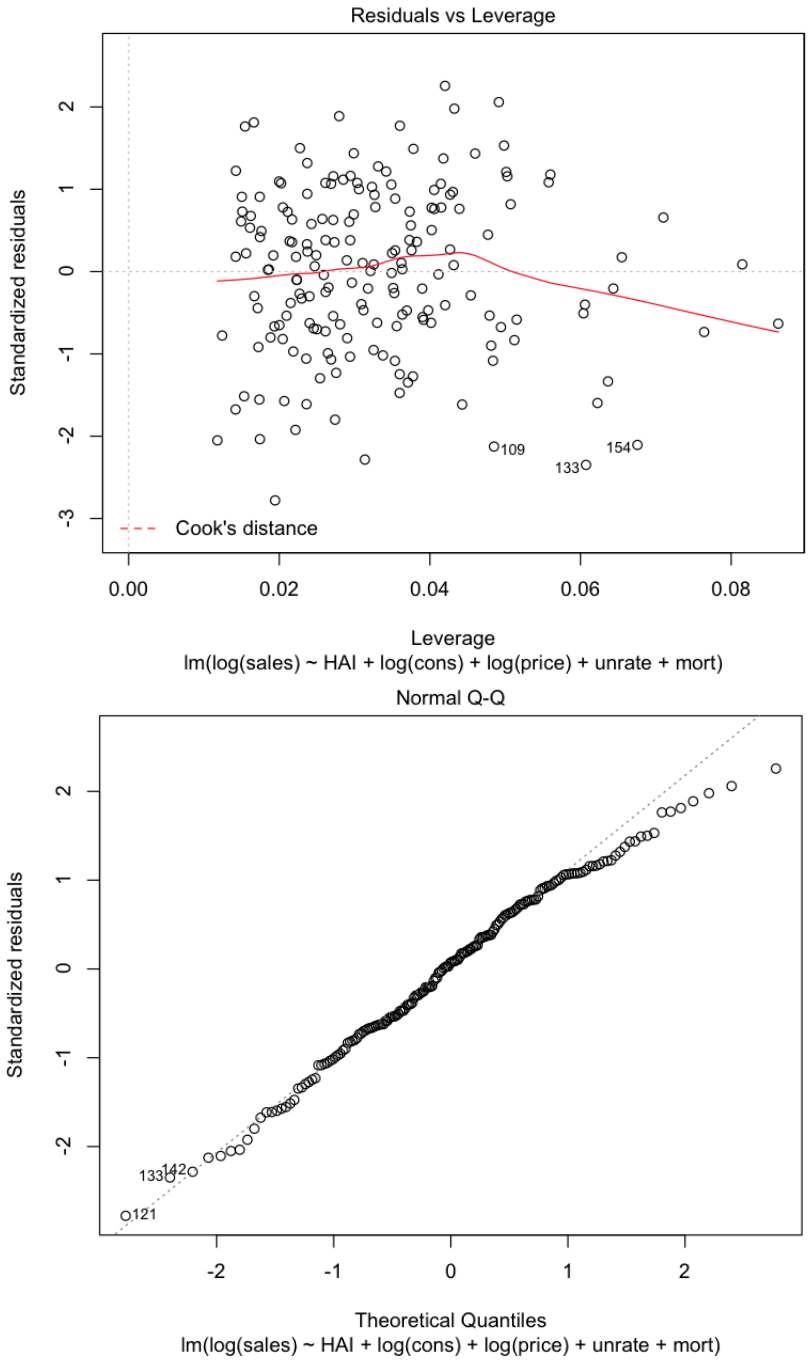
The intercept is not meaningful at all. One point increase in Housing Affordability Index is associated with 0.57% increase in home sales, holding all else constant. One percent increase in construction spending is associated with 119.3% increase in home sales, holding all else constant. One percent increase in median sales price is associated with 69% decrease in home sales, holding all else constant. One percentage point increase in unemployment rate is associated with 6.23% increase in Home Sales, holding all else constant. One percentage point increase in mortgage rate is associated with 5.47% increase in Home Sales, holding all else constant.

The residual standard error implies that our model can predict the response to within a multiplicative factor of 0.5488 to 1.822 ($= 10^{\pm 2s}$), 95% of the time.

Notice that all predictors but Mortgage are highly significant with small p value for any reasonable α. Mortgage has a p-value of 0.084, suggesting that the predictor does not add predictive power to the regression.

R-squared of 75.71% means that 75.71% of the variability in log Home Sales can be explained by the linear relationship with HAI, log Construction Spending, log Median Sales Price, Unemployment Rate, and Mortgage Rate.

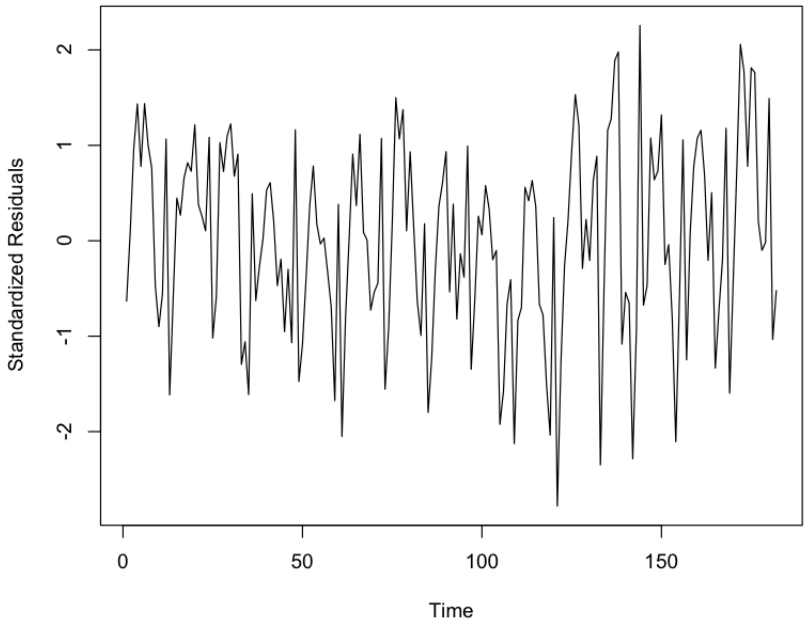
4 Residual Diagnostics



Most residuals are within ± 2.5 standardized deviation. A good guideline for leverage value is $2.5\left(\frac{5+1}{182}\right) = 0.0824$. Almost all points satisfy this criteria. R marks two bands in red dotted lines. The lines are not shown in this plot using the scale shown, meaning that the points very well satisfy Cook's D < 1.

There is a trail of points below the line on the far right. This suggests that the data is probably skewed right.

Here is a plot of standardized residual versus time plot.

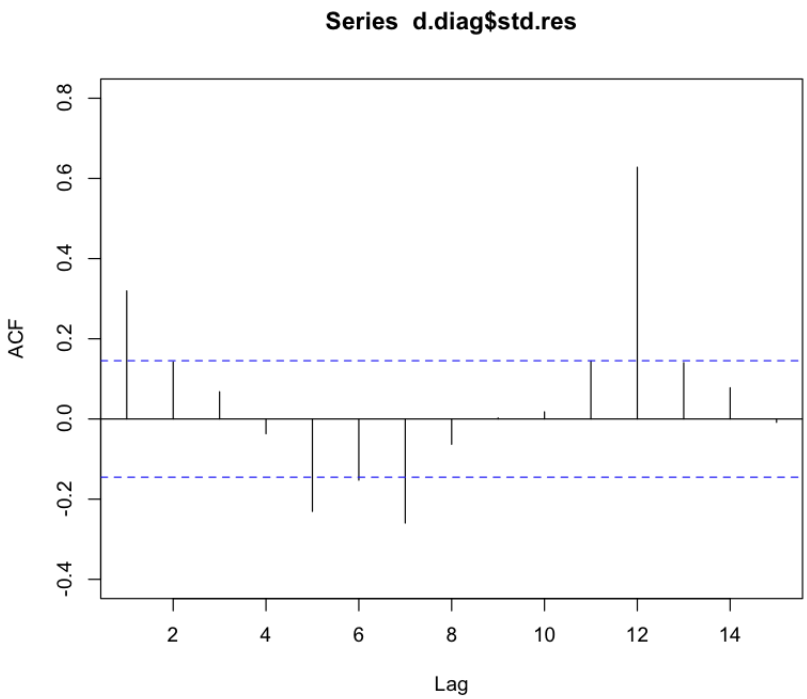


Autocorrelation is apparent. In fact, Durbin-Watson Test agrees.

```
    Durbin-Watson test

data:  logsales ~ HAI + logcons + logprice + unrate + mort
DW = 1.3542, p-value = 1.145e-06
alternative hypothesis: true autocorrelation is greater than 0
```

We have 182 data points. $DW = 1.3542$, which is highly significant. R gives a p-value of almost zero, showing that there is a sign of positive autocorrelation.

Below is an ACF plot of the standardized residuals, checking if observed autocorrelations appear to be consistent with the $AR(1)$ model.




```
> acf(d.diag$std.res, xlim=c(1,15), ylim=c(-.4,.8))
> n=length(logsales)
> acf(d.diag$std.res,plot=F)$acf[2:15]*sqrt(n)
 [1]  4.31089821  1.91716542  0.91652350 -0.49769693 -3.11146135 -2.06020387
 [7] -3.50307765 -0.85271450  0.03559569  0.23989824  1.94958150  8.47177676
[13]  1.89530413  1.05121163
```

Note that the autocorrelations decay roughly at the geometric rate consistent with an $AR(1)$ process, supporting the Durbin-Watson test. The first lag is significantly different from zero, with a $z-$statistics of 4.31. Notice that the twelfth lag spikes, and has a z-statistics of 8.47, which is highly significant. This shows there is autocorrelation, since we have monthly data, and lags spike at multiples of 12.

The Runs Test is also significant.

```
        Runs Test

data:  d.diag$std.res
Standard Normal = -3.9105, p-value = 9.209e-05
```

R gives a p-value of 9.209×10^{-5}, which is below any reasonable α. All three methods suggest autocorrelation.

5 Addressing Autocorrelation

From ACF plot, we see a seasonable effect, i.e. lag 12. We can then include seasonal indicator variables as additional predictors in the regression model.

```
Call:
lm(formula = logsales ~ HAI + logcons + logprice + unrate + mort +
    Jan + Feb + Mar + Apr + May + Jun + Jul + Aug + Sep + Oct +
    Nov)

Residuals:
      Min        1Q    Median        3Q       Max
-0.190275 -0.041393  0.001802  0.052010  0.182245

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.426439   1.275710   6.605 5.21e-10 ***
HAI          0.005237   0.001030   5.085 9.91e-07 ***
logcons      1.042320   0.071503  14.577  < 2e-16 ***
logprice    -0.603851   0.067801  -8.906 9.27e-16 ***
unrate       0.043388   0.008209   5.285 3.93e-07 ***
mort         0.041821   0.020833   2.007   0.0463 *
Jan         -0.298709   0.028264 -10.568  < 2e-16 ***
Feb         -0.195405   0.028170  -6.937 8.68e-11 ***
Mar         -0.052615   0.029311  -1.795   0.0745 .
Apr          0.008234   0.029641   0.278   0.7815
May          0.031344   0.030359   1.032   0.3034
Jun          0.056467   0.032036   1.763   0.0798 .
Jul         -0.036125   0.032802  -1.101   0.2724
Aug         -0.023833   0.033431  -0.713   0.4769
Sep         -0.154734   0.031728  -4.877 2.52e-06 ***
Oct         -0.183202   0.032742  -5.595 8.98e-08 ***
Nov         -0.179536   0.030294  -5.926 1.76e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07778 on 165 degrees of freedom
Multiple R-squared:  0.9188,    Adjusted R-squared:  0.9109
F-statistic: 116.7 on 16 and 165 DF,  p-value: < 2.2e-16

> vif(d1.lm)
     HAI   logcons  logprice    unrate      mort       Jan       Feb       Mar
27.397663 17.634699  3.288954  7.258207 20.259056  1.926810  1.913971  1.954405
     Apr       May       Jun       Jul       Aug       Sep       Oct       Nov
 1.998628  2.096558  2.334656  2.447598  2.542420  2.289961  2.438621  2.087664
```

This model is strong, with R-squared of 91.88%. However, VIF is high for HAI, log Construction Spending, and Mortgage Rate. t-statistics for Mar, Apr, May, Jun, Jul, and Aug are not statistics with p-value greater than 0.05. It makes sense to include all eleven indicator variables. Nonetheless, the best subset tests are included below.

We proceed with best subset test, where 1 = HAI, 2 = log Construction Spending, 3 = log Sales Price, 4 = Unemployment Rate, 5 = Mortgage Rate, 6 = Jan, 7 = Feb, 8 = Mar, 9 = Apr, A = May, B = Jun, C = Jul, D = Aug, E = Sep, F = Oct, G = Nov.

```
> leaps(cbind(HAI,logcons,logprice,unrate,mort,Jan,Feb,Mar,Apr,May,Jun,Jul,Aug,Sep,Oct,Nov),logsales,nbest=2)
$which
      1     2     3     4     5     6     7     8     9     A     B     C     D     E     F     G
1  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
1   TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2  FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
3  FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
3  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
4  FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
4  FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
5  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
5  FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
6  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
6  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
7  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
7   TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
8   TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
8  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
9   TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
9   TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
10  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
10  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
11  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
11  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
12  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
12  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
13  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
13  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
14  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
14  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
15  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
15  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
16  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE

$label
 [1] "(Intercept)" "1"           "2"           "3"           "4"           "5"           "6"           "7"           "8"           "9"           "A"
[13] "C"           "D"           "E"           "F"           "G"

$size
 [1]  2  2  3  3  4  4  5  5  6  6  7  7  8  8  9  9 10 10 11 11 12 12 13 13 14 14 15 15 16 16 17

$Cp
 [1] 596.08152 1197.57085  419.76338  510.01611  290.26274  360.63786  241.27597  242.95534  170.94836  200.97850  135.55910  146.38111  105.10019  109.74067   62.22166
[16]  71.90905   29.14889   45.10955   20.04942   23.13944   16.41070   16.45361   14.65075   14.86411   13.44722   14.43182   14.15417   14.18309   15.07717   15.50823
[31]  17.00000

> leaps(cbind(HAI,logcons,logprice,unrate,mort,Jan,Feb,Mar,Apr,May,Jun,Jul,Aug,Sep,Oct,Nov),logsales,nbest=2,method="r2")
$which
      1     2     3     4     5     6     7     8     9     A     B     C     D     E     F     G
1  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
1   TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2  FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
3  FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
3  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
4  FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
4  FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
5  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
5  FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
6  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
6  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
7  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
7   TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
8   TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
8  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
9   TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
9   TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
10  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
10  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
11  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
11  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
12  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
12  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
13  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
13  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
14  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
14  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
15  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
15  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
16  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE

$label
 [1] "(Intercept)" "1"           "2"           "3"           "4"           "5"           "6"           "7"           "8"           "9"           "A"
[13] "C"           "D"           "E"           "F"           "G"
```

```
$size
 [1]  2  2  3  3  4  4  5  5  6  6  7  7  8  8  9  9 10 10 11 11 12 12 13 13 14 14 15 15 16 16 17

$r2
 [1] 0.6191559 0.3232262 0.7068875 0.6624836 0.7715851 0.7369609 0.7966704 0.7958441 0.8322552 0.8174805 0.8506505 0.8453261 0.8666201 0.8643370 0.8887001 0.8839340 0.9059557
[18] 0.8981032 0.9114166 0.9098963 0.9141908 0.9141697 0.9160407 0.9159357 0.9176168 0.9171324 0.9182530 0.9182388 0.9187829 0.9185708 0.9188209
```

```
> leaps(cbind(HAI,logcons,logprice,unrate,mort,Jan,Feb,Mar,Apr,May,Jun,Jul,Aug,Sep,Oct,Nov),logsales,nbest=2,method="adjr2")
$which
      1     2     3     4     5     6     7     8     9     A     B     C     D     E     F     G
1  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
1   TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2  FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
3  FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
3  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
4  FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
4  FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
5  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
5  FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
6  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
6  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
7  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
7   TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
8   TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
8  FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
9   TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
9   TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
10  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
10  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
11  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
11  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
12  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
12  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
13  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
13  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
14  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
14  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
15  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
15  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
16  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE

$label
 [1] "(Intercept)" "1"           "2"           "3"           "4"           "5"           "6"           "7"           "8"           "9"           "A"
[13] "C"           "D"           "E"           "F"           "G"
```

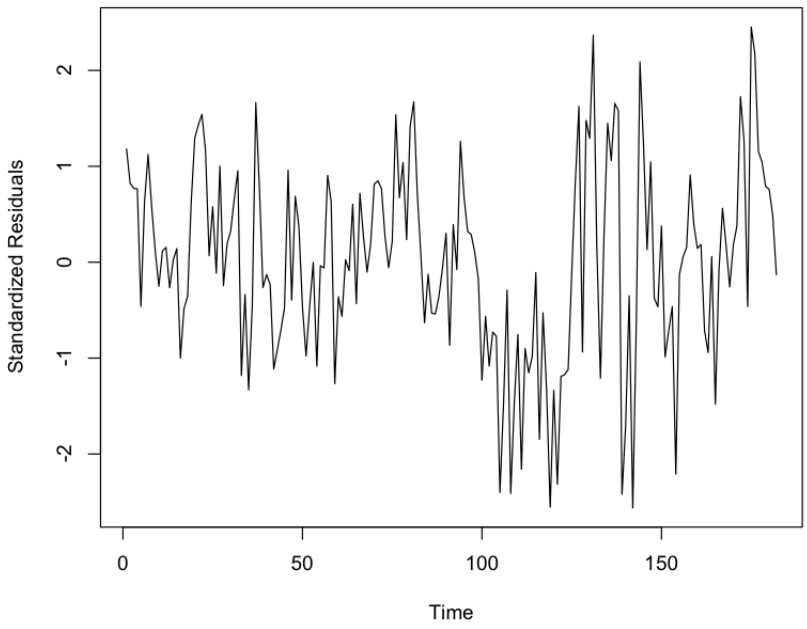
```
$size
 [1]  2  2  3  3  4  4  5  5  6  6  7  7  8  8  9  9 10 10 11 11 12 12 13 13 14 14 15 15 16 16 17

$adjr2
 [1] 0.6170401 0.3194664 0.7036125 0.6587125 0.7677354 0.7325277 0.7920753 0.7912304 0.8274897 0.8122953 0.8455299 0.8400230 0.8612542 0.8588793 0.8835533 0.8785667 0.9010348
[18] 0.8927714 0.9062363 0.9046271 0.9086385 0.9086160 0.9100791 0.9099667 0.9112419 0.9107200 0.9114000 0.9113845 0.9114440 0.9112127 0.9109489
```

R-squared begins to level off (91.14%) at 10-predictor model consisting HAI, log Construction Spending, log Sales Price, Unemployment Rate, Jan, Feb, Jun, Sep, Oct, and Nov. Adjusted R-squared is maximized (91.14%) at 14-predictor model, with all predictors except Apr and Aug. We want $C_p \leq p + 1 = 16 + 1 = 17$, which is realized at 11-predictor model, with predictors HAI, log Construction Spending, log Sales Price, Unemployment Rate, Jan, Feb, May, Jun, Sep, Oct, and Nov. C_p is minimized at the 13-predictor model, with all predictors except Apr, Jul, and Aug.

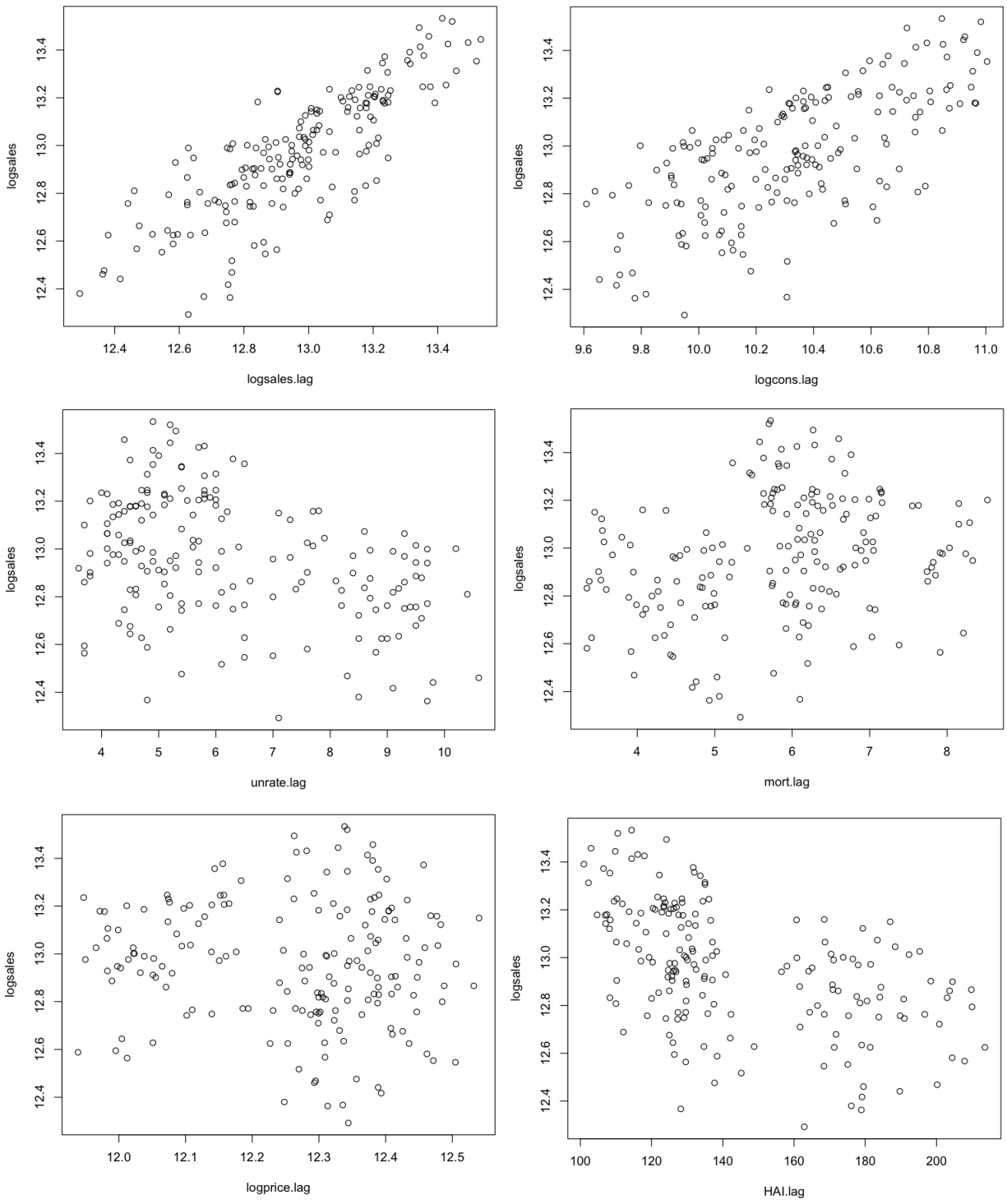
```
> extractAIC(lm(logsales~1))
[1]    1.0000 -488.4136
> extractAIC(lm(logsales~logcons))
[1]    2.0000 -662.1101
> extractAIC(lm(logsales~logcons+Jan))
[1]    3.0000 -707.7638
> extractAIC(lm(logsales~logcons+Jan+Feb))
[1]    4.0000 -751.1533
> extractAIC(lm(logsales~logcons+Jan+Feb+Nov))
[1]    5.0000 -770.3263
> extractAIC(lm(logsales~logcons+logprice+unrate+Jan+Feb))
[1]    6.0000 -803.3403
> extractAIC(lm(logsales~logcons+logprice+unrate+Jan+Feb+Nov))
[1]    7.0000 -822.4804
> extractAIC(lm(logsales~logcons+logprice+unrate+Jan+Feb+Oct+Nov))
[1]    8.0000 -841.0624
> extractAIC(lm(logsales~HAI+logcons+logprice+Jan+Feb+Sep+Oct+Nov))
[1]    9.0000 -871.9995
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+Jan+Feb+Sep+Oct+Nov))
[1]   10.0000 -900.6597
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+Jan+Feb+Jun+Sep+Oct+Nov))
[1]   11.0000 -909.5472
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+Jan+Feb+Apr+Jun+Sep+Oct+Nov))
[1]   12.0000 -908.6842
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+Jan+Feb+Mar+Apr+Jun+Sep+Oct+Nov))
[1]   13.0000 -911.6289
```

```
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+mort+Jan+Feb+Mar+Apr+Jun+Sep+Oct+Nov))
[1]   14.0000 -912.9664
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+mort+Jan+Feb+Mar+Apr+Jun+Jul+Sep+Oct+Nov))
[1]   15.0000 -913.4212
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+mort+Jan+Feb+Mar+Apr+Jun+Jul+Aug+Sep+Oct
+Nov))
[1]   16.0000 -914.2612
> extractAIC(lm(logsales~1))
[1]    1.0000 -488.4136
> extractAIC(lm(logsales~1))+2*2*3/(n-3)
[1]    1.067039 -488.346581
> extractAIC(lm(logsales~logcons))+2*3*4/(n-4)
[1]    2.134831 -661.975243
>  extractAIC(lm(logsales~logcons+Jan))+2*4*5/(n-5)
[1]    3.225989 -707.537803
> extractAIC(lm(logsales~logcons+Jan+Feb))+2*5*6/(n-6)
[1]    4.340909 -750.812399
> extractAIC(lm(logsales~logcons+Jan+Feb+Nov))+2*6*7/(n-7)
[1]    5.4800 -769.8463
> extractAIC(lm(logsales~logcons+logprice+unrate+Jan+Feb))+2*7*8/(n-8)
[1]    6.643678 -802.696598
> extractAIC(lm(logsales~logcons+logprice+unrate+Jan+Feb+Nov))+2*8*9/(n-9)
[1]    7.83237 -821.64806
> extractAIC(lm(logsales~logcons+logprice+unrate+Jan+Feb+Oct+Nov))+2*9*10/(n-10)
[1]    9.046512 -840.015915
> extractAIC(lm(logsales~HAI+logcons+logprice+Jan+Feb+Sep+Oct+Nov))+2*10*11/(n-11)
[1]   10.28655 -870.71299
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+Jan+Feb+Sep+Oct+Nov))+2*11*12/(n-12)
[1]   11.55294 -899.10680
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+Jan+Feb+Jun+Sep+Oct+Nov))+2*12*13/(n-13)
[1]   12.84615 -907.70106
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+Jan+Feb+Apr+Jun+Sep+Oct+Nov))+2*13*14/(n-14)
[1]   14.16667 -906.51757
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+Jan+Feb+Mar+Apr+Jun+Sep+Oct+Nov))+2*14*15/
(n-15)
[1]   15.51497 -909.11397
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+mort+Jan+Feb+Mar+Apr+Jun+Sep+Oct+Nov))
+2*15*16/(n-16)
[1]   16.89157 -910.07484
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+mort+Jan+Feb+Mar+Apr+Jun+Jul+Sep+Oct+Nov))
+2*16*17/(n-17)
[1]   18.29697 -910.12419
> extractAIC(lm(logsales~HAI+logcons+logprice+unrate+mort+Jan+Feb+Mar+Apr+Jun+Jul+Aug+Sep+Oct
+Nov))+2*16*17/(n-17)
[1]   19.29697 -910.96427
```

Both AIC and AIC_C are minimized using all predictors. If we choose to use all variables, autocorrelation is still noticeable in the following standardized error versus time plot, despite the fact that we included indicator variables.



Instead, we can choose to include lagged variables for log Construction Spending, Unemployment Rate, Mortgage, log Sales Price, log Home Sales. Below are some scatterplots, showing logsales versus each of the new variable created.



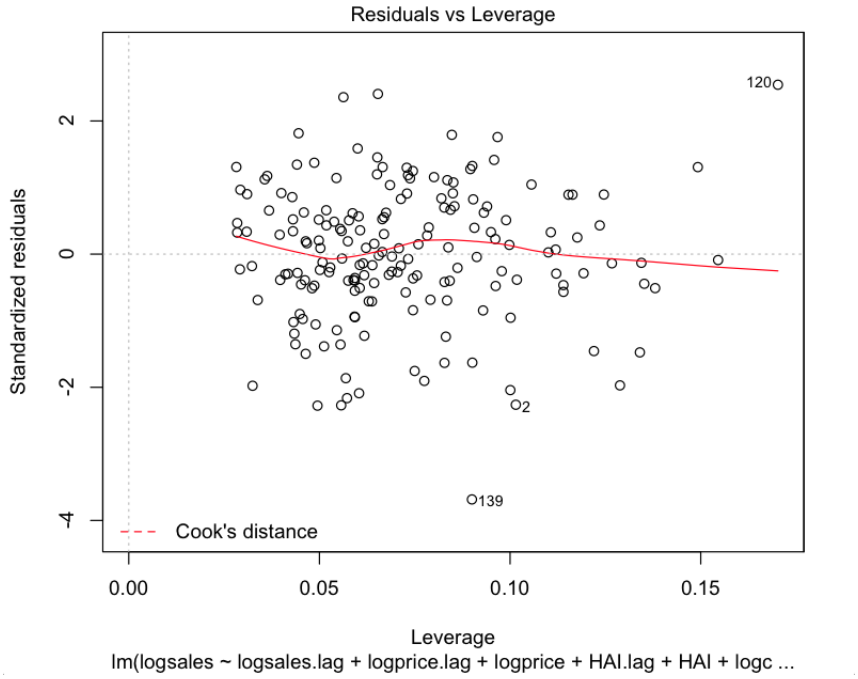
The marginal relationships between log sales versus each of lagged unemployment rate, lagged mortgage, and lagged log sales price are not clear. Again, there seems to be subgroups in logsales versus lagged HAI plot. Other relationships are fairly strong.

Since from previously, we see that every January has a peak in home sales, we also include a indicator variable for January.

Below is the regression.

```
> dadj=lm(logsales~logsales.lag+logprice.lag+logprice+HAI.lag+HAI+logcons.lag+logcons+unrate
+unrate.lag+mort+mort.lag+Jan)
> summary(dadj)

Call:
lm(formula = logsales ~ logsales.lag + logprice.lag + logprice +
    HAI.lag + HAI + logcons.lag + logcons + unrate + unrate.lag +
    mort + mort.lag + Jan)

Residuals:
      Min        1Q    Median        3Q       Max
-0.259464 -0.043580 -0.000456  0.051063  0.215725

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.775821   1.296479   4.455 1.53e-05 ***
logsales.lag  0.545105   0.052726  10.338  < 2e-16 ***
logprice.lag -0.241799   0.133785  -1.807   0.0725 .
logprice      0.005431   0.140113   0.039   0.9691
HAI.lag       0.008410   0.001727   4.871 2.55e-06 ***
HAI          -0.008237   0.001627  -5.062 1.08e-06 ***
logcons.lag  -0.570577   0.091458  -6.239 3.46e-09 ***
logcons       0.862724   0.089532   9.636  < 2e-16 ***
unrate        0.006818   0.021754   0.313   0.7544
unrate.lag    0.003772   0.021767   0.173   0.8626
mort         -0.024223   0.035205  -0.688   0.4924
mort.lag      0.012255   0.036653   0.334   0.7385
Jan          -0.212460   0.030664  -6.929 8.66e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07564 on 168 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9209,    Adjusted R-squared:  0.9153
F-statistic:   163 on 12 and 168 DF,  p-value: < 2.2e-16
 > vif(d3.lm)
 logsales.lag logprice.lag     logprice      HAI.lag          HAI  logcons.lag         cons
     5.260763    13.319105    14.401417    81.496042    68.511109    30.730246    17.286286
       unrate   unrate.lag         mort     mort.lag          Jan
    50.697572    54.638860    59.537612    65.770808     2.238703
```

There are signs of multicollinearity between variable and lagged variables, but that is expected. The model is fairly strong with R-squared of 91.25%.



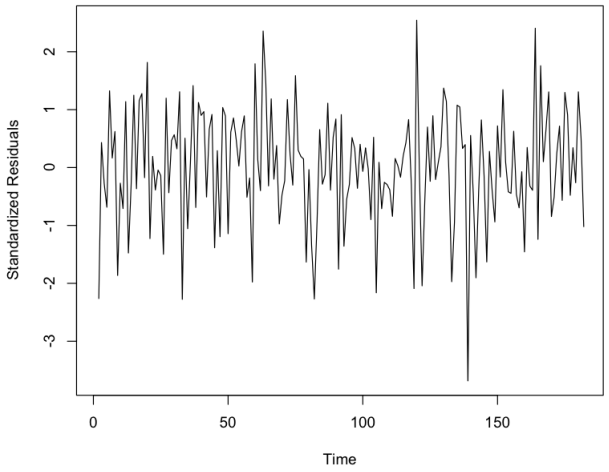
A quick regression diagnostic shows that standardized residuals are fairly randomly scatter and within ±2.5. There are two potential outliers, observation 120 and 139. Observation 120 is December 2008 and observation 139 is July 2010. Below show part of the data that con taints the two outliers.

2008-08-01	409000	129.7	33903	221000	6.1	6.48
2008-09-01	369000	142.3	30735	225200	6.0	6.04
2008-10-01	349000	145.2	29977	213200	6.1	6.20
2008-11-01	273000	148.8	25535	221600	6.5	6.09
2008-12-01	305000	162.9	20961	229600	7.1	5.33
2009-01-01	218000	176.1	18331	208600	8.5	5.06
2009-02-01	238000	171.3	16775	209700	8.9	5.13
2009-03-01	304000	168.5	18504	205100	9.0	5.00
2009-04-01	349000	177.6	20207	219200	8.6	4.81
2009-05-01	376000	171.1	20116	222300	9.1	4.86
2009-06-01	438000	160.8	20918	214700	9.7	5.42
2009-07-01	442000	156.5	22440	214200	9.7	5.22
2009-08-01	417000	161.6	24154	207100	9.6	5.19
2009-09-01	392000	164.1	22260	216600	9.5	5.06
2009-10-01	418000	170.8	23848	215100	9.5	4.95
2009-11-01	395000	175.3	20724	218800	9.4	4.88
2009-12-01	347000	178.9	17636	222600	9.7	4.93
2010-01-01	234000	179.5	16732	218200	10.6	5.03
2010-02-01	258000	178.5	15349	221900	10.4	4.99
2010-03-01	366000	173.8	17970	224800	10.2	4.97
2010-04-01	443000	170.4	20854	208300	9.5	5.10
2010-05-01	449000	168.7	21530	230500	9.3	4.89
2010-06-01	472000	161.7	22191	219500	9.6	4.74
2010-07-01	331000	164.4	22184	212100	9.7	4.56
2010-08-01	352000	171.8	22486	226600	9.5	4.43
2010-09-01	321000	179.0	20825	228000	9.2	4.35
2010-10-01	307000	181.4	22525	204200	9.0	4.23
2010-11-01	304000	183.8	19636	219600	9.3	4.30

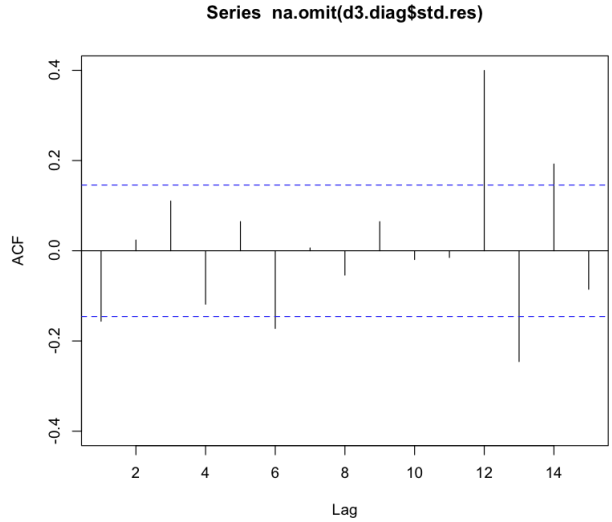
The second column represents home sales. Notice in December 2008, home sales is 305000 units, but in January 2009, it dips to 218000 units. In July 2010, home sales decreased to 331000 units from 472000 from previous month. The dip in December 2008 might be from the financial crisis. While July 2010 is indeed another unusual point.

A good guideline for the leverage value is $2.5\left(\frac{12+1}{182}\right) = 0.17$. All points satisfy this criteria. Cook's distance is also fine, for reason mentioned previously. We ignore the two outliers for now.

We then look at autocorrelation.



There is still signs of autocorrelation in the standardized residual versus time plot.



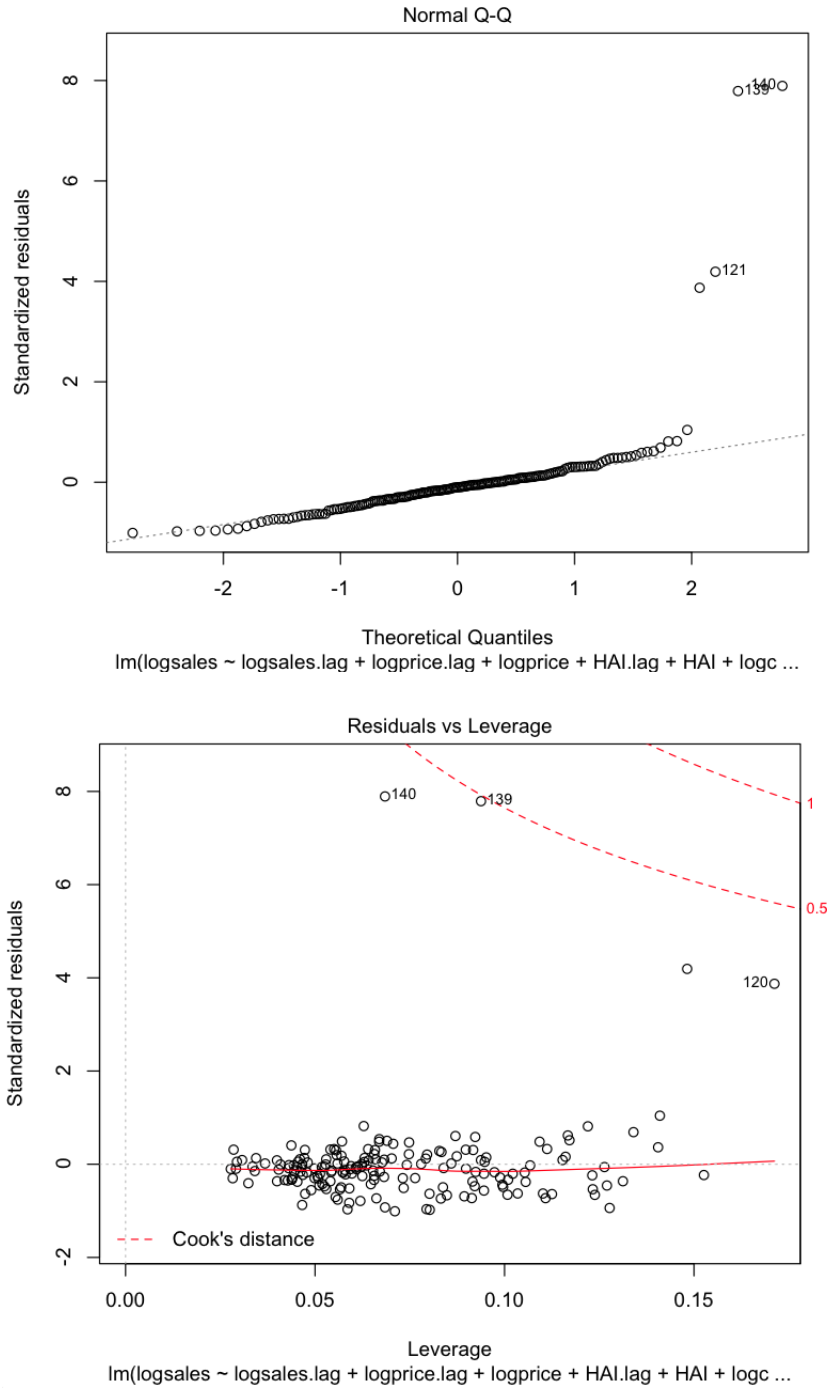

```
> acf(na.omit(d3.diag$std.res),plot=F)$acf[2:15]*sqrt(n)
 [1] -2.10772706  0.32198402  1.48757021 -1.59965093  0.87310489 -2.32436244  0.08532333
 [8] -0.72895372  0.87082475 -0.26340962 -0.20228968  5.39286864 -3.31475298  2.59507641
> runs.test(na.omit(d3.diag$std.res), cont.corr=T)

    Runs Test

data:  na.omit(d3.diag$std.res)
Standard Normal = 1.6403, p-value = 0.1009
```

The first lag in *ACF* plot is still slightly significant. The twelfth lag is still not fixed. However, runs test is not significant anymore, we have p-value of -0.1. *[How to obtain DW test for data with lagged variables? Gasoline Price does not include this part of R code]*

6 Outliers

We may address the outliers by assigning average of the periods immediately preceding and succeeding the outliers. Regression output is shown below.

```
> dadj=lm(logsales~logsales.lag+logprice.lag+logprice+HAI.lag+HAI+logcons.lag+logcons+unrate
+unrate.lag+mort+mort.lag+Jan,data=d.adj)
> summary(dadj)

Call:
lm(formula = logsales ~ logsales.lag + logprice.lag + logprice +
    HAI.lag + HAI + logcons.lag + logcons + unrate + unrate.lag +
    mort + mort.lag + Jan, data = d.adj)

Residuals:
   Min     1Q Median     3Q    Max
-46388 -16940  -4746   5513 363163

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1004010     817142   1.229   0.2209
logsales.lag    17146      33232   0.516   0.6066
logprice.lag    19904      84321   0.236   0.8137
logprice       -23455      88310  -0.266   0.7909
HAI.lag         -3110       1088  -2.858   0.0048 **
HAI              1786       1026   1.741   0.0835 .
logcons.lag    -49364      57644  -0.856   0.3930
logcons        -37389      56430  -0.663   0.5085
unrate          25258      13711   1.842   0.0672 .
unrate.lag     -22218      13719  -1.619   0.1072
mort           -32625      22189  -1.470   0.1433
mort.lag        14321      23102   0.620   0.5361
Jan            -35661      19327  -1.845   0.0668 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47670 on 168 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1523,    Adjusted R-squared:  0.0918
F-statistic: 2.516 on 12 and 168 DF,  p-value: 0.00455

> vif(dadj)
logsales.lag logprice.lag     logprice      HAI.lag          HAI  logcons.lag      logcons
    5.894525    13.403581    14.487258    80.910094    72.320746    30.391219    29.032574
      unrate   unrate.lag         mort     mort.lag          Jan
   53.728198    53.921276    60.970442    65.795590     2.260914
```



Noticing after adjusting the points, R-squared dropped to 15%. The standardized residuals are unbelievably high (roughly 8) for observation 139 and observation 140. The standardized residuals is roughly 4 for observation 120. It is obvious that we fail to address the issue of outlier. The QQ plot now sees observation 121 as an outlier. Adjusting observation 139 now makes observation 140 influential. At this point, the basic method does not seem to address the issue of outliers. Our regression seems to fail.

7 Remarks

The model is built using data that are not seasonably adjusted. Thus we run into many problems. The Fed, however, does post seasonally adjusted data, for variables such as unemployment rate and mortgage rate. I was unable to obtain seasonally adjusted mortgage rate, adjusted HAI, and adjusted construction spending,

whence unable to produce a multiple regression using adjusted data. The adjusted data are usually reported in deseasonalized form, which roughly corresponds to taking the residuals from a regression on the monthly effects and adding the overall average data back.

Despite the unsuccessful attempt to remove the outliers, our model still offers some insight.