Deriving the Relationship between Age and Quality on Housing Price

Pstat 126 Group Project - Winter 2018 - Instructor Dr. Xiyue Liao

Group Members:
Jaeyun Lee -- Brennan (9am Friday)
Tony Khoury -- Megan (12pm Friday)
Robert Boullon -- Ana Luo (6pm Monday)
Pranavi Gandham -- Brennan (9am Friday)

Introduction:

Real Estate Sales

The data set “Real Estate Sales” contains 521 arms-length transactions of home sales during the year 2002. Among the 12 variables from the data set, 3 variables are chosen as study of interest:

1. Quality: Index of quality of construction: 1 indicates high quality; 2 indicates medium quality; 3 indicates low quality
2. Year built: Year property was originally constructed
3. Sales price: Sales price of residence (dollars)

Use the Real Estate data set with “Sales price” as the response and the “Quality” and “Year built” as predictors. The study of interest is predicting residential home sales as a function of quality and year built of the home and surrounding property.

Question of Interest:

First, we want to know if “Quality” and “Year built” as predictors and “Sales price” as response meet the LINE condition:

1) mean of response is a Linear function,
2) Errors are Independent,
3) Errors are Normally distributed, and
4) Errors have Equal variances.

Then, we ask: Is a “best” model achievable with “Quality” and “Year built” as the predictors and “Sales price” as a response that meets the Mallows’ Cp, adjusted R^2, and AIC criterion with a 95% significance level?

Conclusion:

- Final Model: x_1, x_8, x_7, x_{10}, x_9, x_{11} and x_5

```
y <- realestate$SalePrice
x1 <- realestate$SqFeet
x2 <- realestate$Beds
x3 <- realestate$Baths
x4 <- realestate$Air
x5 <- realestate$Garage
x6 <- realestate$Pool
x7 <- realestate$Year
x8 <- factor(realestate$Quality)
x9 <- realestate$Style
x10 <- realestate$Lot
x11 <- realestate$Highway
```

- Really depends on the questions we asked, I think for us it ended up being which predictors best predict sales price (and are year and quality included in that)

- Predictor x_3 made it through AIC but was factored out after doing Adj R^2 and MSE and then another diagnostic check with C_P

“For example, is there any other possible way to improve the model such as to find predictors not in the data set? How general are your results, to what situations do they apply?”

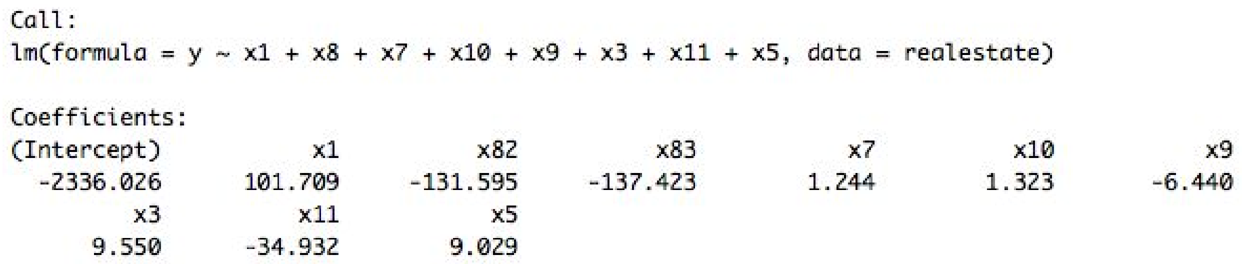
- It’s not a large data set (only 521 observations) so this is relatively tailored to this one city and really probably only applies to this one “midwestern” city

Regression Method:

- In order to determine which predictors best predict our response, “Sale Price”, (and whether “year” and “quality” are included in that) we will start with the full model containing all predictors and use the stepwise regression procedure with AIC as the criterion to narrow down the model.
- Once we obtain a new model based off of the AIC criterion from the stepwise regression procedure, we will then narrow that model down further by eliminating more predictors through the best subsets regression procedure with adjusted R-squared as the criterion. Then, we will repeat the subsets regression procedure once more with Mallow’s Cp statistic as the criterion to obtain a final model.
- Once we have obtained our final model through the process described above we will then make sure that it is an appropriate model for the data by making sure it satisfies the LINE conditions.
- If the model does not initially satisfy all LINE conditions we will appropriately transform either the response or predictors, or both depending on which one of the LINE conditions is not satisfied. After the transformation we re-check that model once more to insure that it satisfies all LINE conditions *(Linear, Independent, Normal, and Equal)*.

Regression Analysis, Results and Interpretation:

Stepwise Regression: Akaike’s Information Criterion (AIC): fit sales price to 11 predictors of real estate: x_1 = SqFeet, x_8 = Quality, x_7 = Year, x_{10} = Lot, x_9 = Style, x_3 = Baths, x_{11} = Highway, x_5 = Garage

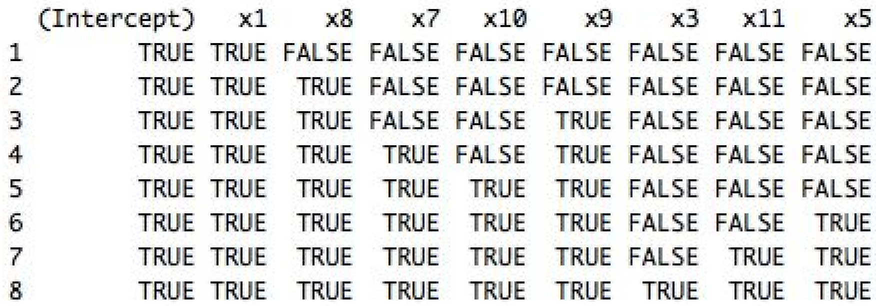


```
Call:
lm(formula = y ~ x1 + x8 + x7 + x10 + x9 + x3 + x11 + x5, data = realestate)

Coefficients:
(Intercept)           x1          x82          x83           x7          x10           x9
  -2336.026      101.709     -131.595     -137.423        1.244        1.323       -6.440
         x3          x11           x5
      9.550      -34.932        9.029
```

- Takes into account SSE, number of parameters “p” (11) and the sample size “n” (512)

- Outputs an eight predictor model as the “best” model: SqFeet, Quality, Year, Lot, Style, Bed, Highway and Garage
- Cannot use this as the final model since the conclusion is not based off any specifications

Best subsets regression:



```
  (Intercept)   x1    x8    x7   x10    x9    x3   x11    x5
1        TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2        TRUE TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
3        TRUE TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
4        TRUE TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE
5        TRUE TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
6        TRUE TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
7        TRUE TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
8        TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

R^2 *(coefficient of determination)*

```
[1] 0.6769435 0.7414778 0.7643109 0.7752927 0.7847008 0.7869074 0.7883866 0.7884390
```

- The largest increase of R^2 happens between the “best” one predictor model (x_1 = SqFeet) and the “best” two predictor model (x_1 = SqFeet, x_2 = Quality). There is a significant R^2 increase between the “best” two predictor model and the “best” three model predictor but it is not as large. However, it is larger than the increases between all other predictor models going forward. This suggests that is best to include at least *three predictors* in the model.
- We also use R^2 to see when it is unnecessary to add more predictors. The smallest difference in R^2 values happens between 0.7883866 to 0.7884390 suggesting that adding the eighth predictor would **not** be worthwhile.

Adjusted R^2

- Adjusted R^2 penalizes us for adding more predictors to the model thus making it logical to use adj. R^2 to determine the final model (how many predictors we should use).

```
[1] 0.6763211 0.7404796 0.7629433 0.7735508 0.7826106 0.7844199 0.7854991 0.7851334
```

- The largest adjusted R^2 value is 0.7854991 which corresponds to the seven model predictor.
-

MSE *(Mean Squared Error)*

- We use MSE to see if our conclusion from adjusted R^2 is correct because these two yield the same results since R^2 increases only if MSE decreases. Along with the fact that MSE quantifies the difference between our predicted response from our observed response.
- Therefore the smallest MSE value will tell us what model to use:

```
[1] 6129.895 4914.848 4489.426 4288.540 4116.964 4082.698 4062.261 4069.187
```

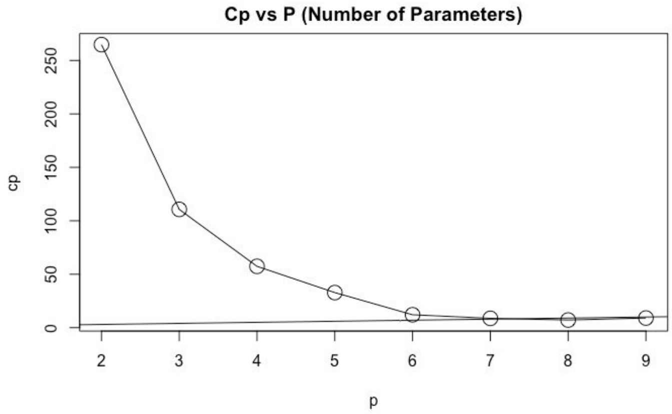
- The smallest value is 4062.261, which also correlates to the seven predictor model (R^2 and MSE match up).

C_P

- Mallow's C_P as a final check of bias.

```
[1] 264.830847 110.651144  57.392462  32.815431  12.046755   8.706721   7.126854
9.000000
```

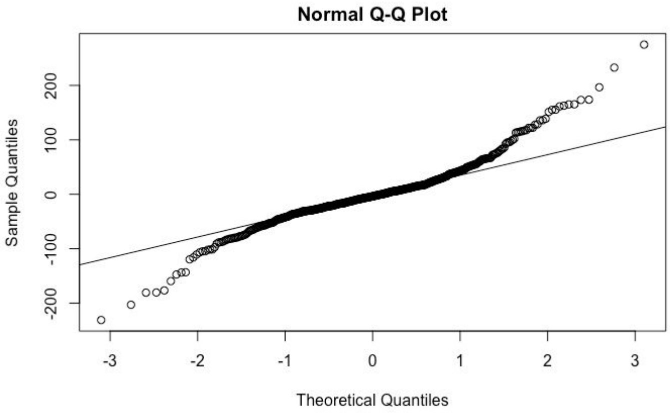
- Since there are a few models that are near P=8 (p for number of parameters) we should use the lowest C_P value which is 7.126854 again determining our previous conclusion that we should use the seven predictor model.
- C_P Plot



Thus the "best" model to determine Sales Price consists of seven predictors: SqFeet, Quality, Year, Lot, Style, Highway and Garage.

Check the "best" Model against LINE conditions:

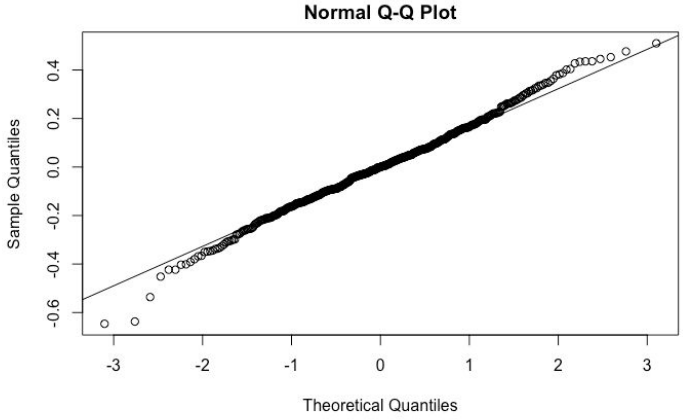
```
fit <- lm(y ~ x1 + x8 + x7 + x10 + x9 + x11 + x5)
```



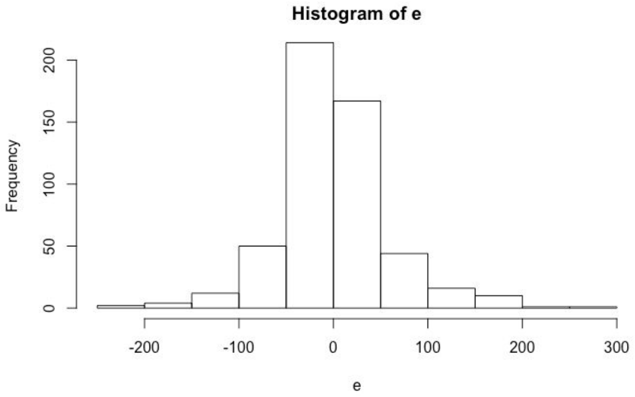
Since the Normal Q-Q Plot shows some unequal variances, in response we log Sales Price (the response).

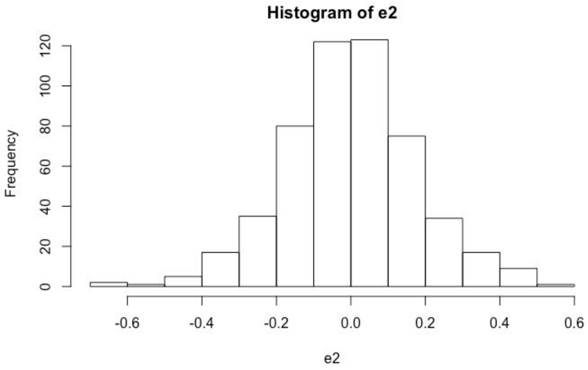
```
fitlog <- lm(log(y) ~ x1 + x8 + x7 + x10 + x9 + x11 + x5)
```

Now the data points in the Normal Q-Q Plot follow a more linear pattern.



We can also double check by checking the histograms of both fits.





There is one outlier on the left but we can ignore that due to the symmetrical pattern of the rest of the histogram.

Summary of Our Strategy

All in all we think that we were successful in following the basic formula of Regression Analysis. The formula we followed was as follows:
Strategy:

1. Before we began we had to know our goal and know our research question.
    - Our goal was to use the data gathered on the Real Estate Market in 2002 and derive valuable information pertaining to the relationships between characteristics of a house and that house's selling price.
    - Our research question was: "Does a house's quality and year constructed have a substantial effect on the selling price?"
2. Next we identified all of the possible candidate predictors.

- In our dataset there were no truly extraneous variables provided (i.e. if we were given a variable like “movie ticket sales” that had no relation whatsoever to the real estate market that would be deemed an extraneous variable), therefore from the onset all of our variables provided in the “Real Estate Sales” dataset could initially be considered candidate predictors.
- Additionally while we can assume that not all variables will be used in our final model, there were no redundant variables that we could rule out from the onset (quick recap on redundant variables: a redundant variable would be one that gives similar information as that of another variable, we already have number of bathrooms in our dataset, a redundant variable for instance would be if the dataset that also included number of sinks/toilets).
  - Furthermore we did not worry about interactions or the appropriate functional forms such as x^2 or log(x) at this stage just yet.

3. Then we performed variable selection procedures:
   - This was done in order to find a middle ground between an underspecified model and a model with extraneous or redundant variables.
   - First a **Stepwise** regression was performed using the AIC criterion (pick lowest AIC value for each “best” model).
     - Results: eight predictor model: x_1 x_8 x_7 x_{10} x_9 x_3 x_{11} x_5
       - These predictors are SqFeet, Quality, Year, Lot, Style, Bed, Highway, and Garage.
   - Next a **Best Subsets** regression was performed:
     - During this process we measured each “best” model’s coefficient of determination (by looking for largest increase in R^2), adjusted coefficient of determination (by looking for the largest value of adj. R^2), Mean Squared Error (by looking for the largest MSE value), and finally Mallow’s C_p-statistic (this time we looked for the lowest value).
     - Results: seven predictor model: x_1 x_8 x_7 x_{10} x_9 x_{11} x_5 (x_3 is removed)
       - These predictors are: SqFeet, Quality, Year, Lot, Style, Highway, and Garage.
4. Finally we Fine-Tuned the model to make sure it was correctly specified.
   - We iterated back and forth between formulating different regression models all the while checking the behavior of the residuals until we were finally satisfied with our model.
     - This is when we now transform data by taking the square root or log of a parameter in order to more accurately fit the regression model.

Concluding Remarks

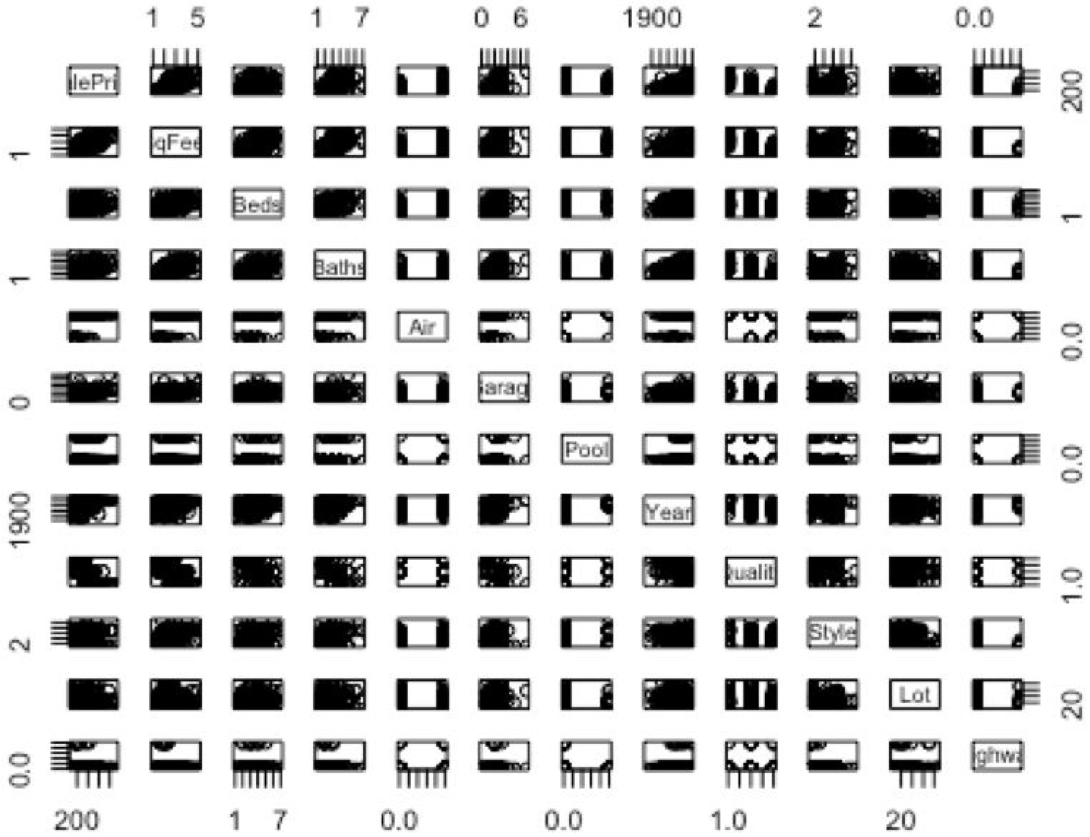
To put it in non-statistical terms we found the factors that affected the price of a house in order of strongest to weakest to be the House’s Square Footage, Quality, Age, Lot, Style, Close Proximity to a Highway, and Inclusion of a Garage. In our initial assessment we were both right and wrong in a way. Our prediction that Quality and

Year were strong factors in affecting the price of a house was true. Neither of these two variables, however, were the single best predictor: which was the square footage of a house. In retrospect this was a cautionary lesson of filtering out our own biases. We are used to assuming that a house's year of construction is very important due to newer houses being built with more fire safety and earthquake resistance in mind
Unlike California, however, the threat of wildfires and earthquakes are not present in the Midwest and it becomes plain to see why square footage was so influential on a House's price.

All in all if there were predictors we think could improve our model they would likely be things unrelated to the predictors already in our model (i.e. we would not want redundant information such as predictors that are interacting with the predictors already in our model like, say, pool-size). Some hypothetical predictors we would want are proximity to non-housing zones (if a house is across the street from an apartment complex or trailer park it's value is less), proximity to local schools (houses that are located near schools will have higher value since most home-buyers are people with families or looking to start a family; they would value not having to travel as far to take their kids to school), and lastly quality of roof since roof's need to be replaced and/or repaired periodically which is a very expensive process (also the Midwest has heavy snowfall and rainfall in comparison to California so this is likely to have a bigger influence on housing price than we're used to here).

Appendix

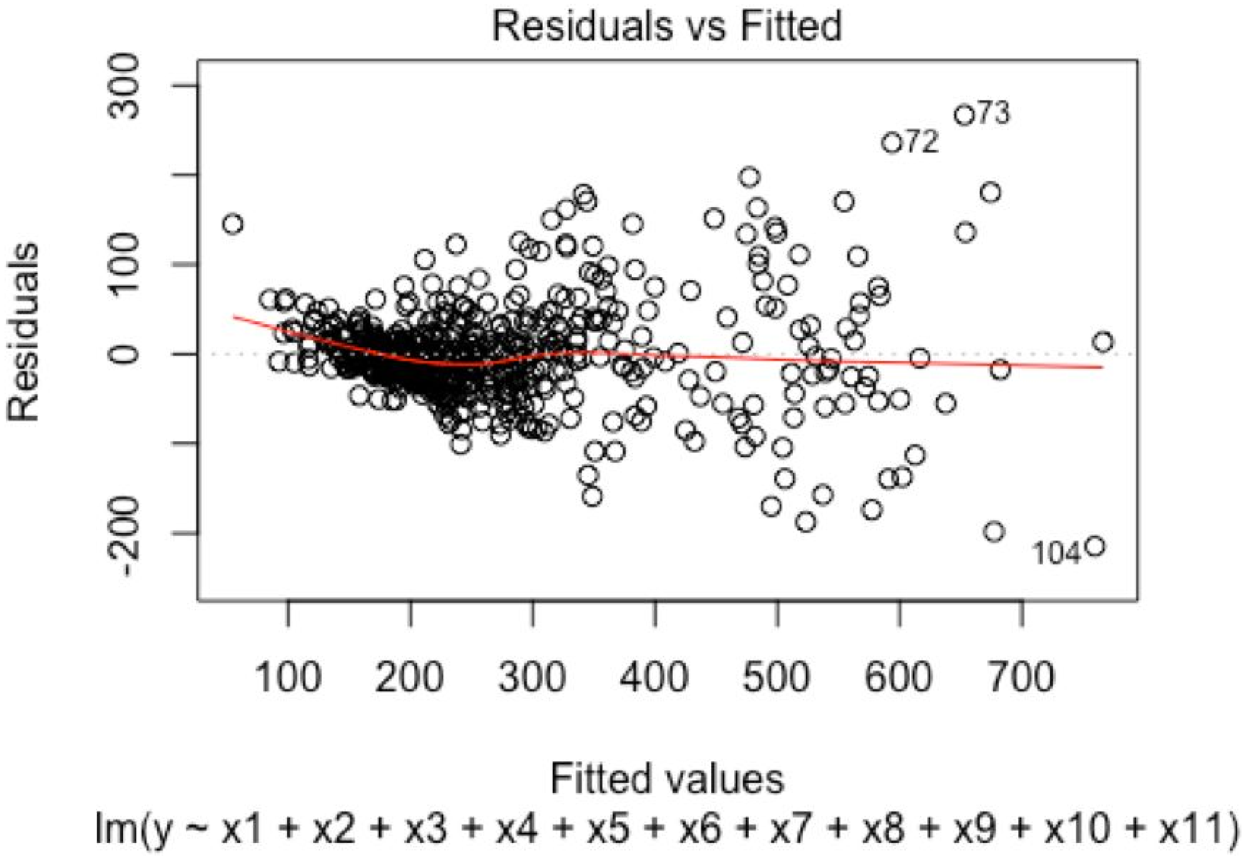
```
load(file = "realestate.RData")
library(leaps)
pairs(realestate)
```

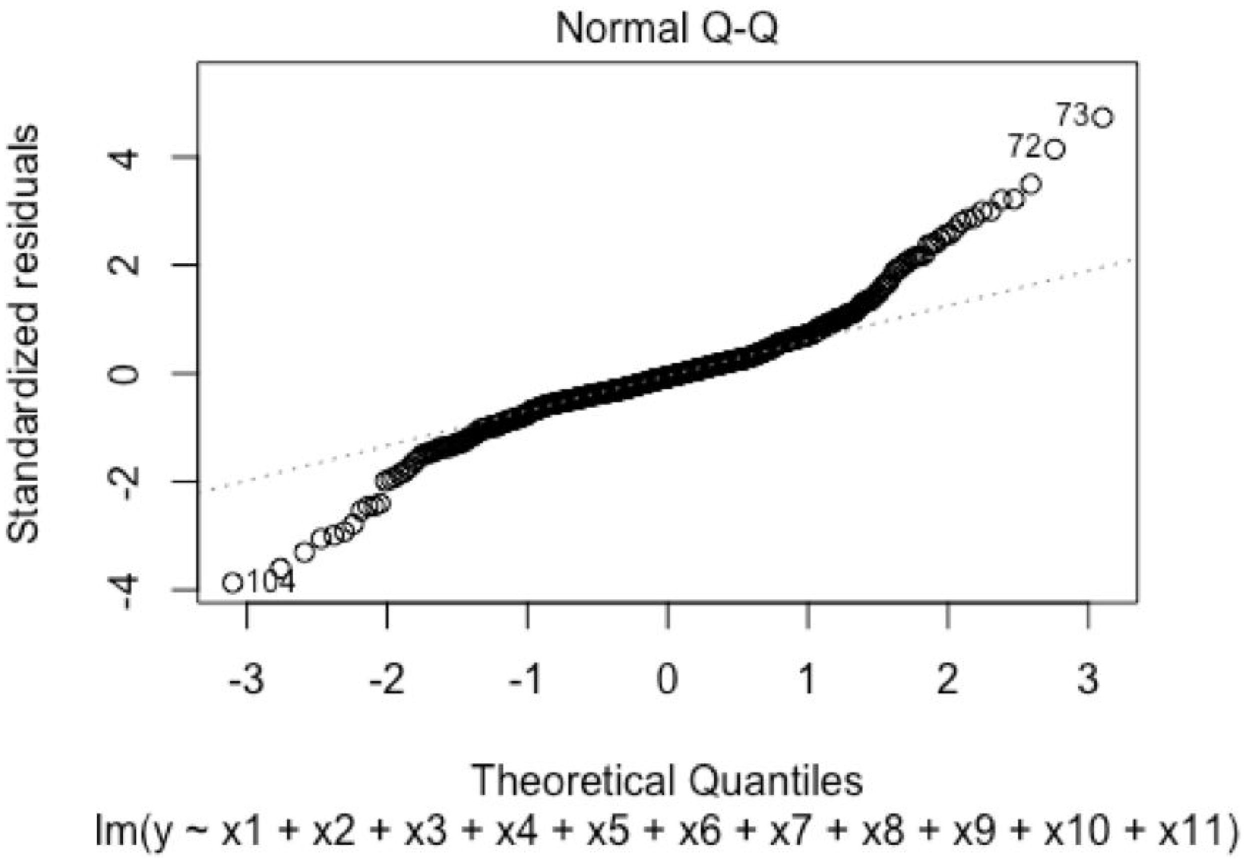


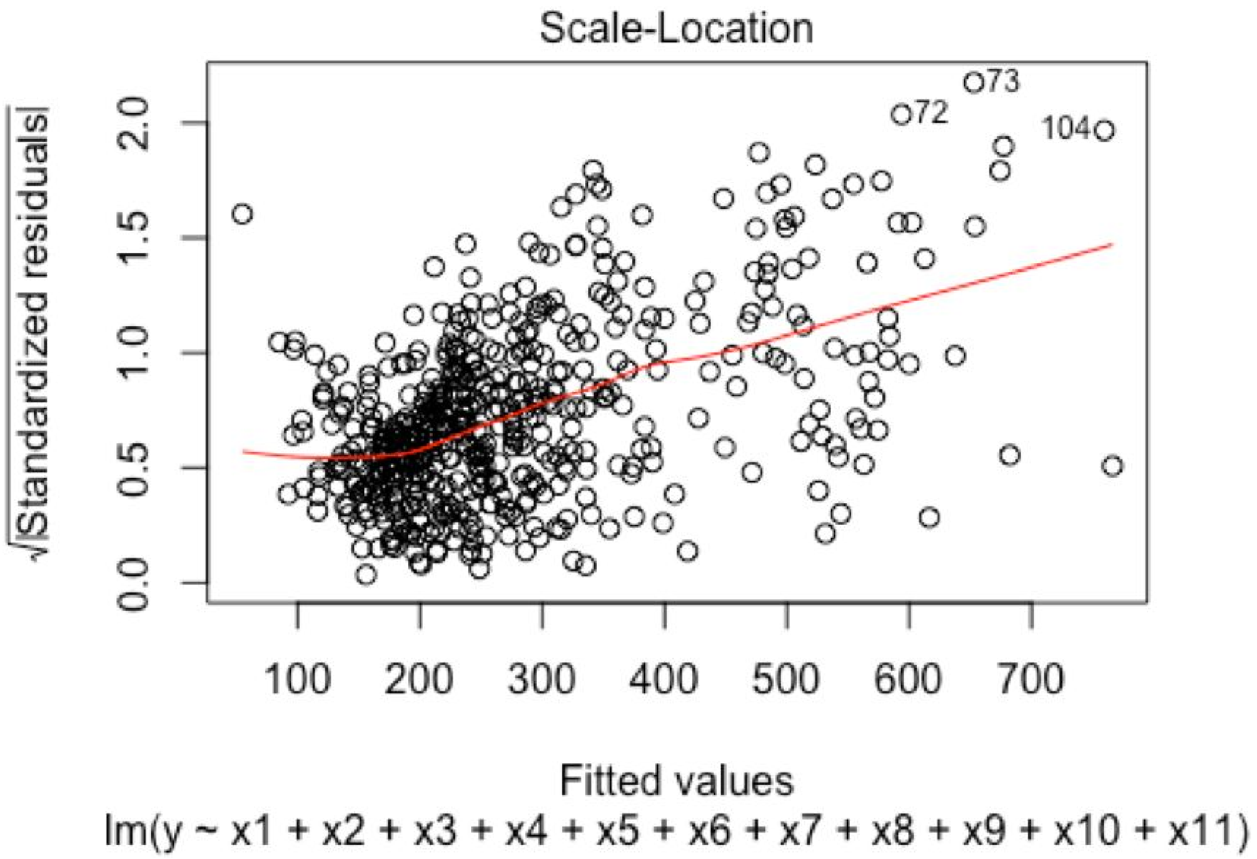
#response: sales

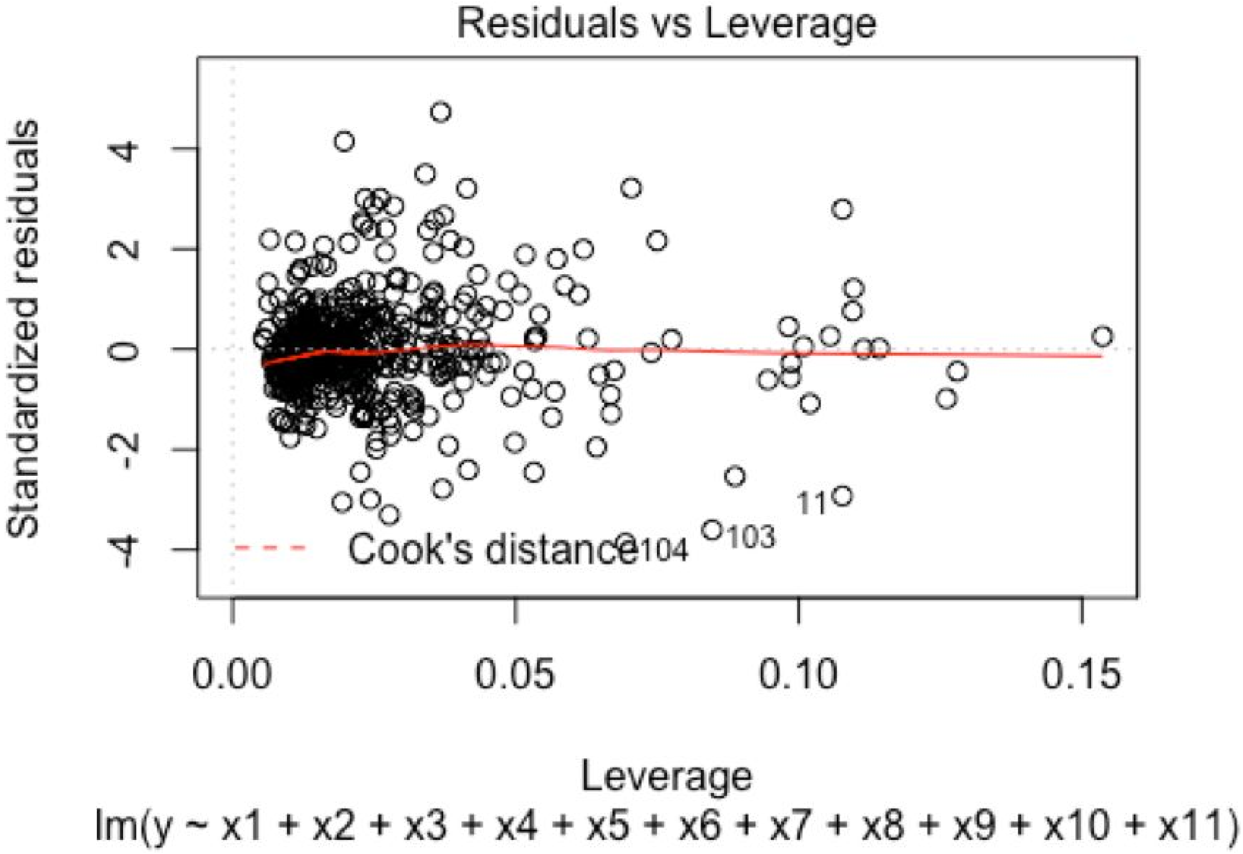
```
#set predictor variables
y <- realestate$SalePrice
x1 <- realestate$SqFeet
x2 <- realestate$Beds
```

```r
x3 <- realestate$Baths
x4 <- realestate$Air
x5 <- realestate$Garage
x6 <- realestate$Pool
x7 <- realestate$Year
x8 <- factor(realestate$Quality)
x9 <- realestate$Style
x10 <- realestate$Lot
x11 <- realestate$Highway
#fit all predictors against response
fit <- lm(y ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11, data = realestate)
plot(fit)
```


Residuals vs Fitted
Residuals
300
100
0
-200
100
200
300
400
500
600
700
72
73
104
Fitted values
lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11)


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11)
73
72
104


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
100
200
300
400
500
600
700
Fitted values
lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11)
73
72
104



```
#stepwise regression AIC
mod0 = lm(y ~ 1, data = realestate)
mod.all = lm(y ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11, data =
realestate)
step(mod0, scope = list(lower = mod0, upper = mod.all))
```

```
## Start:  AIC=5132.29

## y ~ 1
##
##        Df Sum of Sq     RSS   AIC
## + x1    1   6666447 3181416 4545.6
## + x8    2   6478958 3368904 4577.4
## + x3    1   4808012 5039850 4785.3
## + x5    1   3265611 6582252 4924.4
## + x7    1   3015679 6832183 4943.8
## + x2    1   1833933 8013929 5026.9
## + x9    1   1291487 8556376 5061.1
## + x4    1    816799 9031063 5089.2
## + x10   1    481994 9365869 5108.1
## + x6    1    215838 9632024 5122.7
## <none>              9847862 5132.3
## + x11   1     25232 9822631 5133.0
##
## Step:  AIC=4545.6
## y ~ x1
##
##         Df Sum of Sq    RSS    AIC
## + x8    2   1094724 2086691 4329.9
## + x7    1    438567 2742849 4470.3
## + x9    1    336825 2844591 4489.3
## + x5    1    253284 2928132 4504.4
## + x3    1    128267 3053149 4526.2
## + x10   1     83898 3097517 4533.7
## + x4    1     48138 3133278 4539.7
## + x2    1     14478 3166937 4545.2
## <none>              3181416 4545.6
## + x6    1      2126 3179290 4547.3
## + x11   1         5 3181411 4547.6
## - x1    1   6666447 9847862 5132.3
##
## Step:  AIC=4329.87
## y ~ x1 + x8
##
##         Df Sum of Sq    RSS    AIC
## + x7    1    123693 1962998 4300.0
## + x9    1    106084 1980608 4304.7
## + x10   1     94472 1992220 4307.7
## + x5    1     62015 2024676 4316.2
## + x3    1     48283 2038408 4319.7
## <none>              2086691 4329.9
## + x4    1      7075 2079616 4330.1
## + x6    1      1354 2085337 4331.5
## + x11   1     1148 2085544 4331.6
## + x2    1          1 2086690 4331.9
## - x8    2   1094724 3181416 4545.6
## - x1    1   1282213 3368904 4577.4
##
## Step:  AIC=4300.04
## y ~ x1 + x8 + x7
##
##         Df Sum of Sq    RSS    AIC
## + x10   1     157990 1805008 4258.3
## + x9    1     107805 1855194 4272.6
## + x5    1     34208 1928791 4292.9
## + x3    1     25520 1937478 4295.2
## <none>             1962998 4300.0
## + x11   1     2688 1960310 4301.3
## + x6    1     2080 1960918 4301.5
## + x2    1     230 1962768 4302.0
## + x4    1          3 1962995 4302.0
## - x7    1     123693 2086691 4329.9
## - x8    2     779851 2742849 4470.3
## - x1    1   1256629 3219627 4555.8
##
## Step:  AIC=4258.32
## y ~ x1 + x8 + x7 + x10
##
##         Df Sum of Sq    RSS    AIC
## + x9    1     73032 1731976 4238.8
## + x5    1     17886 1787123 4255.1
## + x3    1     15131 1789878 4255.9
## + x11   1     8811 1796198 4257.8
## <none>             1805008 4258.3
## + x6    1     5509 1799500 4258.7
## + x4    1     2857 1802151 4259.5
## + x2    1     1634 1803375 4259.8
## - x10   1     157990 1962998 4300.0
## - x7    1     187211 1992220 4307.7
## - x8    2     754677 2559685 4436.3
## - x1    1   1128557 2933565 4509.3
```

```
##
## Step:  AIC=4238.8
## y ~ x1 + x8 + x7 + x10 + x9
##
##         Df Sum of Sq    RSS    AIC
## + x3    1     20623 1711354 4234.6
## + x11   1     13885 1718092 4236.6
## + x5    1     12762 1719214 4236.9
## <none>              1731976 4238.8
## + x6    1     4157 1727820 4239.5
## + x2    1     1273 1730703 4240.4
## + x4    1     1214 1730763 4240.4
## - x9    1     73032 1805008 4258.3
## - x10   1     123217 1855194 4272.6
## - x7    1     180563 1912539 4288.5
## - x8    2     589916 2321893 4387.5
## - x1    1   1030229 2762205 4480.0
##
## Step:  AIC=4234.56
## y ~ x1 + x8 + x7 + x10 + x9 + x3
##
##        Df Sum of Sq      RSS   AIC
## + x11   1     12780 1698573 4232.7
## + x5    1     11092 1700262 4233.2
## <none>              1711354 4234.6
## + x2    1     5766 1705588 4234.8
## + x6    1     2385 1708969 4235.8
## + x4    1     900 1710453 4236.3
## - x3    1     20623 1731976 4238.8
## - x9    1     78524 1789878 4255.9
## - x10   1     111638 1822991 4265.5
## - x7    1     150826 1862180 4276.6
## - x8    2     589868 2301222 4384.9
## - x1    1     722939 2434292 4416.1
##
## Step:  AIC=4232.66
## y ~ x1 + x8 + x7 + x10 + x9 + x3 + x11
##
```

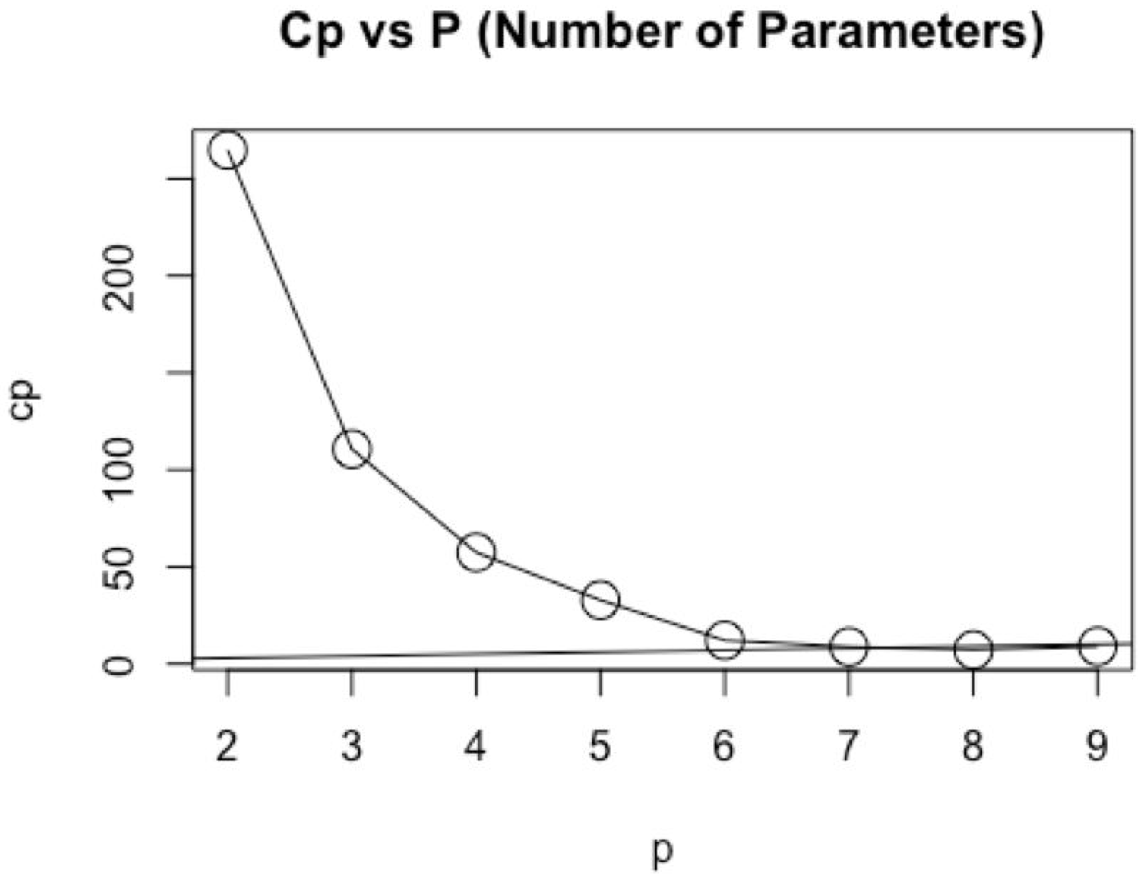
```
##        Df Sum of Sq     RSS    AIC
## + x5    1     11100 1687473 4231.2
## <none>              1698573 4232.7
## + x2    1      5617 1692956 4232.9
## + x6    1      2146 1696428 4234.0
## + x4    1       637 1697936 4234.5
## - x11   1     12780 1711354 4234.6
## - x3    1     19518 1718092 4236.6
## - x9    1     83381 1781955 4255.6
## - x10   1    117796 1816369 4265.6
## - x7    1    156722 1855295 4276.6
## - x8    2    587251 2285824 4383.4
## - x1    1    723468 2422041 4415.5
## 
## Step:  AIC=4231.24
## y ~ x1 + x8 + x7 + x10 + x9 + x3 + x11 + x5
## 
##        Df Sum of Sq     RSS    AIC
## <none>              1687473 4231.2
## + x2    1      6059 1681414 4231.4
## - x5    1     11100 1698573 4232.7
## + x6    1      1860 1685613 4232.7
## - x11   1     12789 1700262 4233.2
## + x4    1       155 1687318 4233.2
## - x3    1     17896 1705369 4234.7
## - x9    1     77942 1765415 4252.8
## - x10   1    108384 1795857 4261.7
## - x7    1    134388 1821861 4269.2
## - x8    2    558255 2245728 4376.1
## - x1    1    667068 2354541 4402.8
## 
## Call:
## lm(formula = y ~ x1 + x8 + x7 + x10 + x9 + x3 + x11 + x5, data = 
realestate)
## 
## Coefficients:
## (Intercept)           x1          x82          x83           x7  
##   -2336.026      101.709     -131.595     -137.423        1.244  
##         x10        x9         x3         x11        x5
##       1.323        -6.440     9.550      -34.932         9.029

#best subset regression
mod = regsubsets(cbind(x1, x8, x7, x10, x9, x3, x11, x5), y)
summary.mod = summary(mod)
summary.mod$which

##   (Intercept)   x1 x8    x7    x10   x9    x3    x11   x5
## 1        TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2        TRUE TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 3        TRUE TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
## 4        TRUE TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE
## 5        TRUE TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
## 6        TRUE TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
## 7        TRUE TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## 8        TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE

summary.mod$rsq

## [1] 0.6769435 0.7414778 0.7643109 0.7752927 0.7847008 0.7869074
## [7] 0.7883866 0.7884390
```

```
adjr2 <- summary.mod$adjr2
Adjr2

## [1] 0.6763211 0.7404796 0.7629433 0.7735508 0.7826106 0.7844199
## [7] 0.7854991 0.7851334

max(adjr2)

## [1] 0.7854991

n = 521
rss = summary.mod$rss
mses = c(rss[1]/(n-2), rss[2]/(n-3), rss[3]/(n-4), rss[4]/(n-5),
rss[5]/(n-6), rss[6]/(n-7), rss[7]/(n-8), rss[8]/(n-9))
Mses
```

```
## [1] 6129.895 4914.848 4489.426 4288.540 4116.964 4082.698 4062.261
4069.187

min(mses)

## [1] 4062.261
```

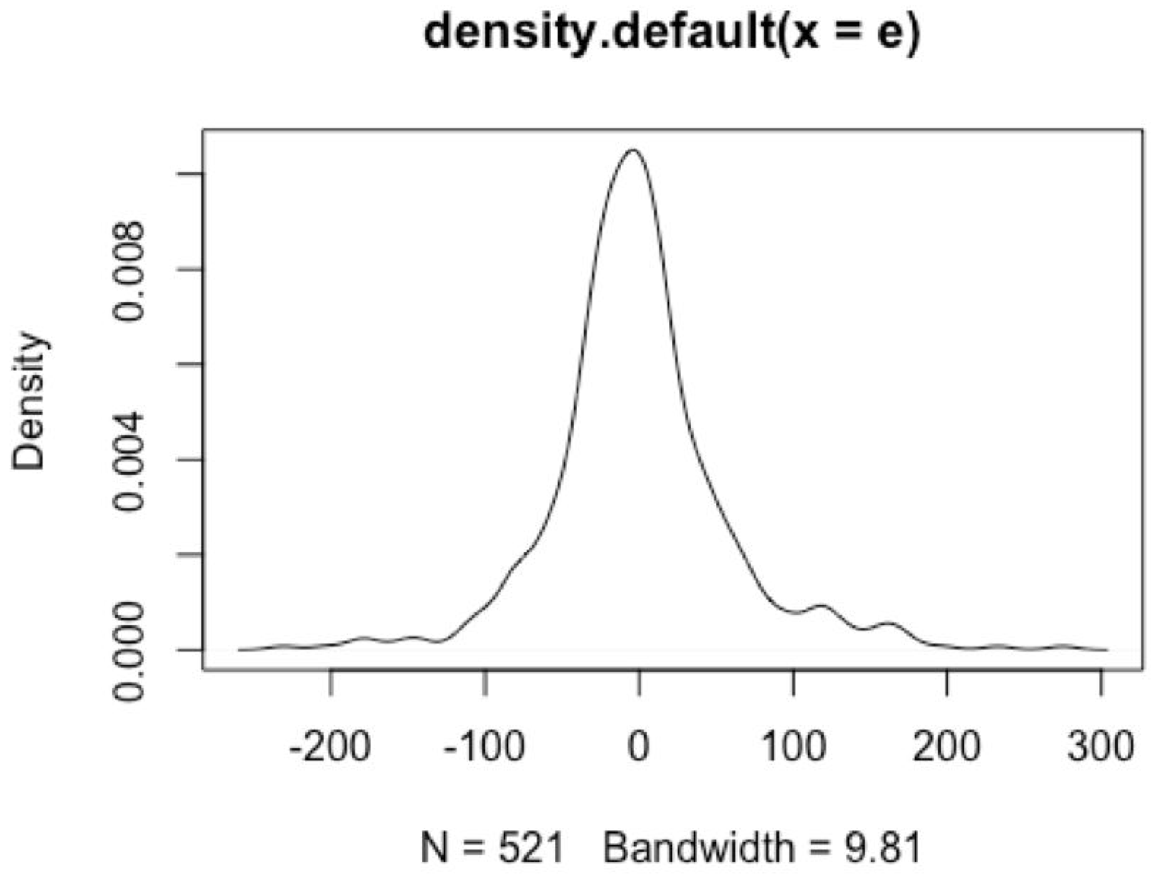
use seven predictor model: x1, x8, x7, x10, x9, x11 and x5

```
cp = summary.mod$cp
cp
## [1] 264.830847 110.651144  57.392462  32.815431  12.046755
8.706721
## [7]   7.126854   9.000000
p = 2:9
plot(p, cp, type = 'o', cex = 2, main = 'Cp vs P (Number of
Parameters)')
abline(1, 1)
```

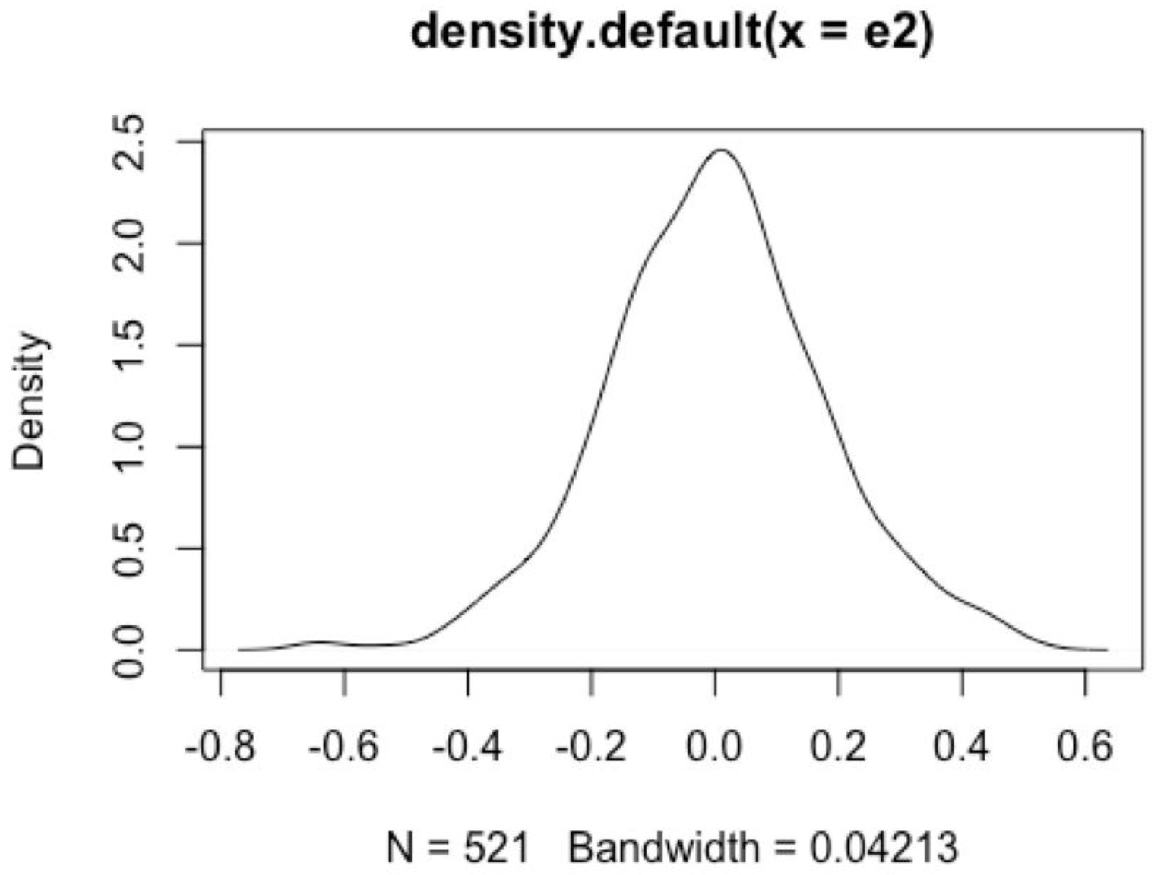

Cp vs P (Number of Parameters)

#diagnostic check: asserts that we should use the seven model predictor

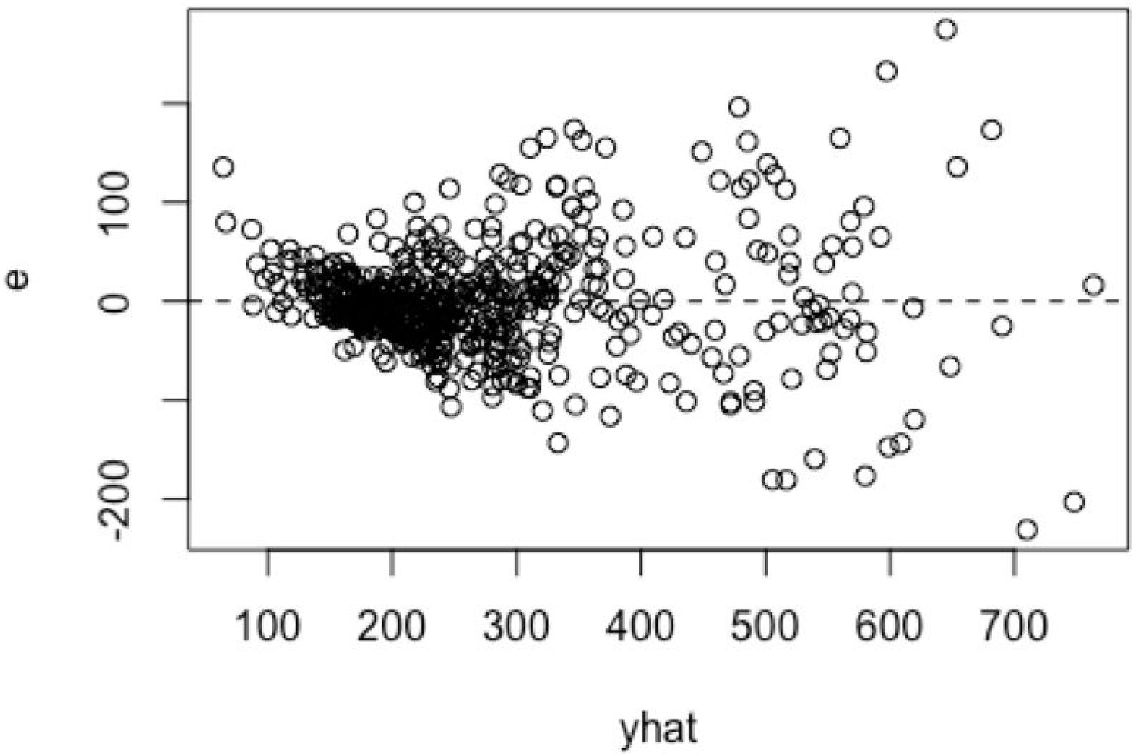
```
#fit new predictors
#compare log sales vs sales
fit <- lm(y ~ x1 + x8 + x7 + x10 + x9 + x11 + x5)
fitlog <- lm(log(y) ~ x1 + x8 + x7 + x10 + x9 + x11 + x5)

yhat <- fitted(fit)
yhatlog <- fitted(fitlog)

e = (y) - yhat
e2 = log(y) - yhatlog


plot(density(e))
```



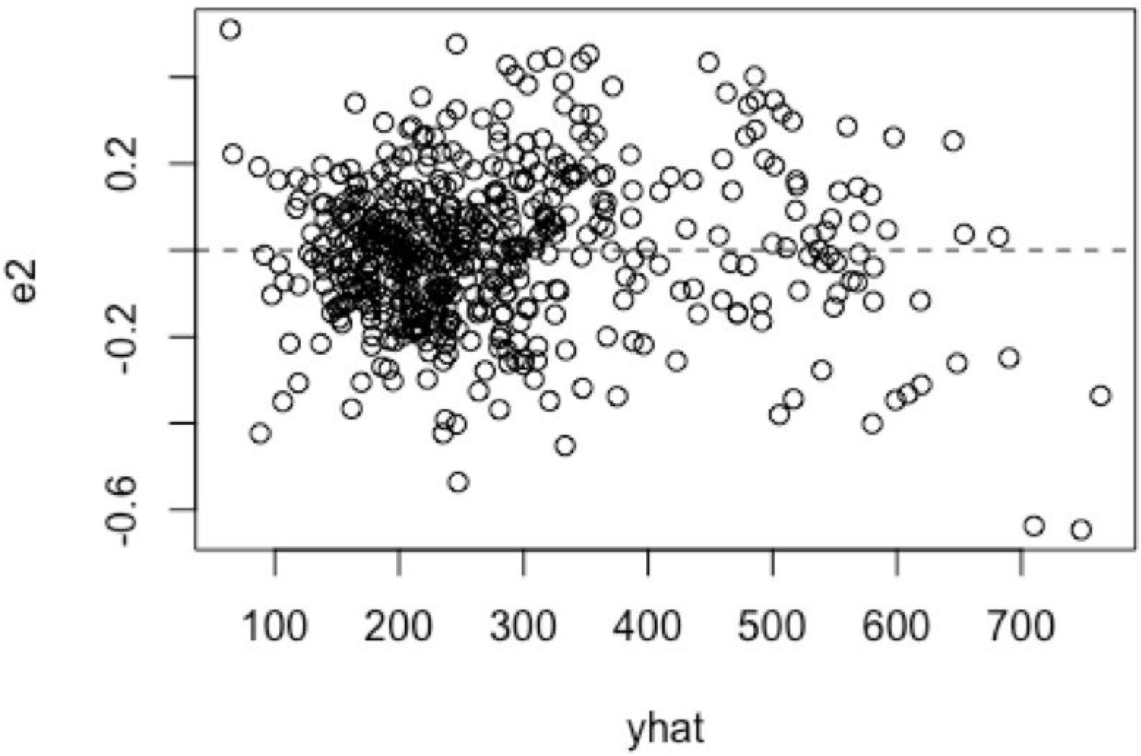
```
plot(density(e2))
```



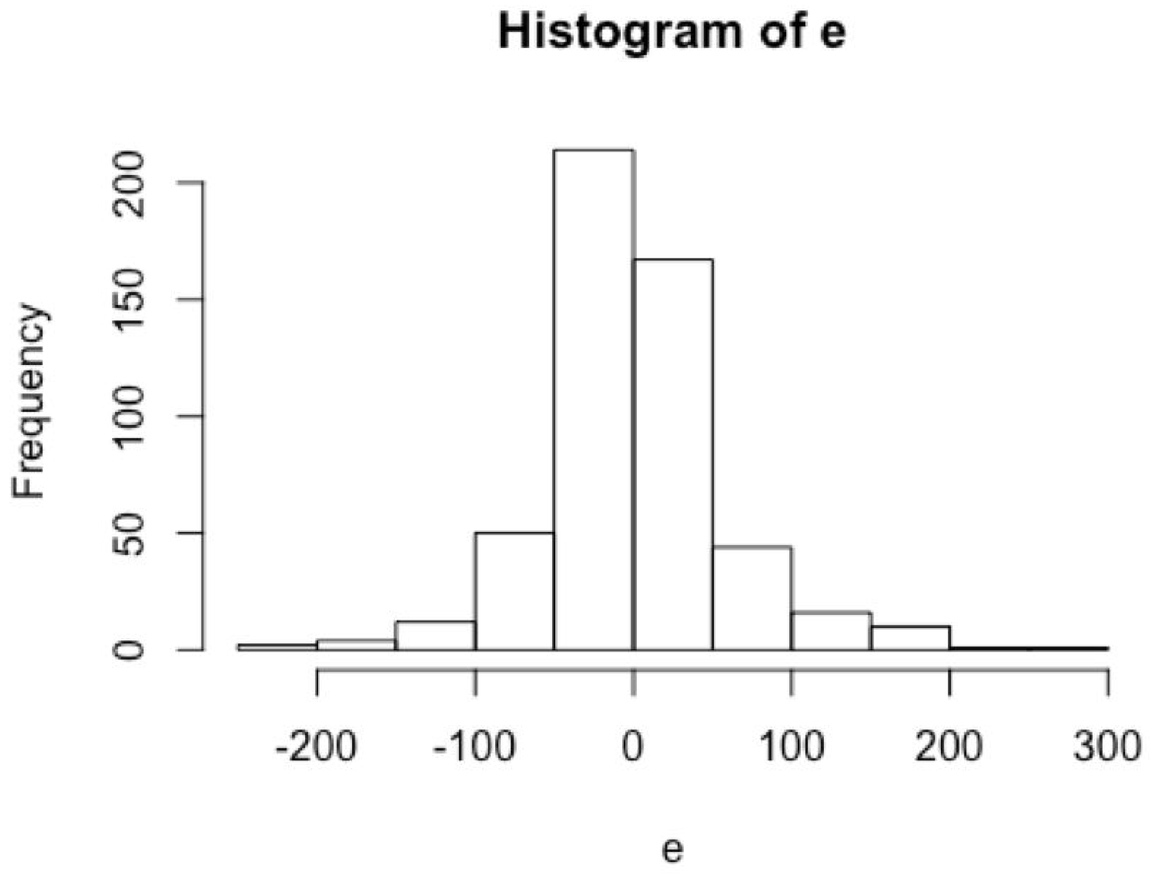
```
plot(yhat, e)
abline(h=0, lty=2)
```



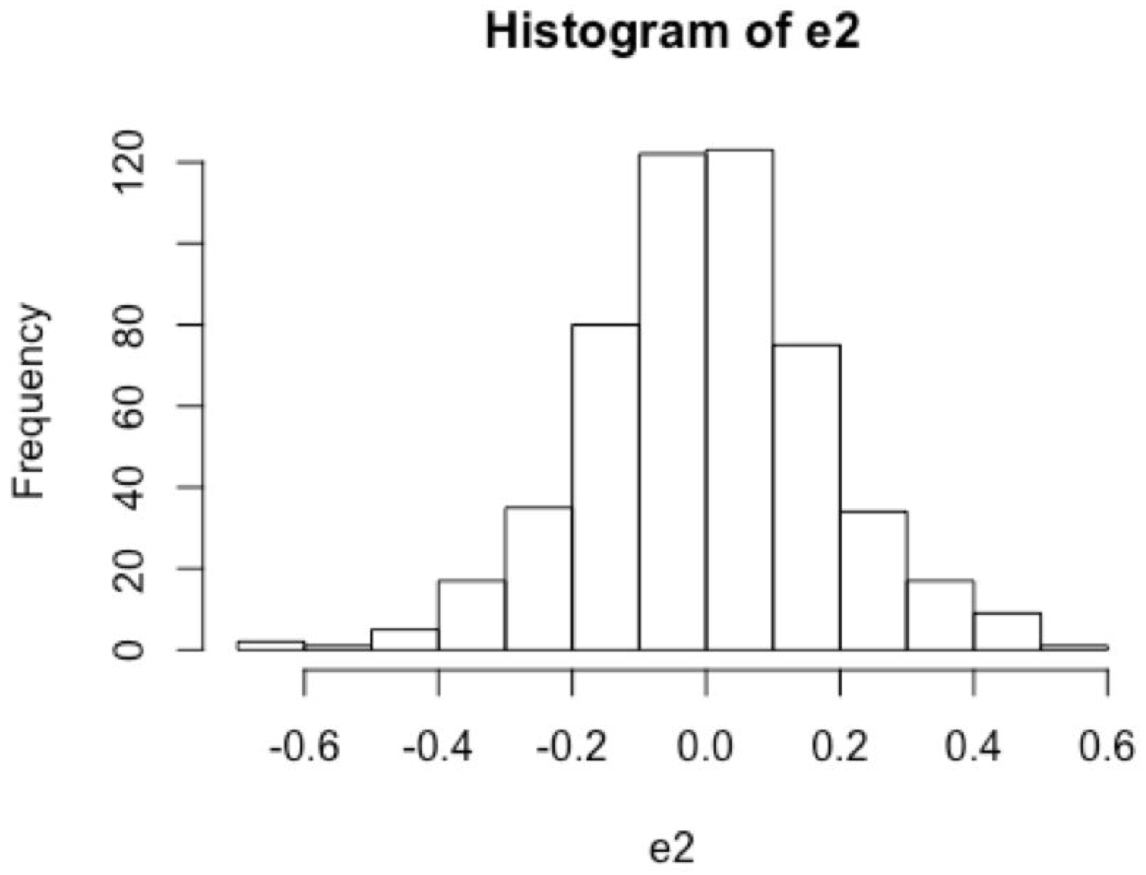
```
plot(yhat, e2)
abline(h=0, lty=2)
```



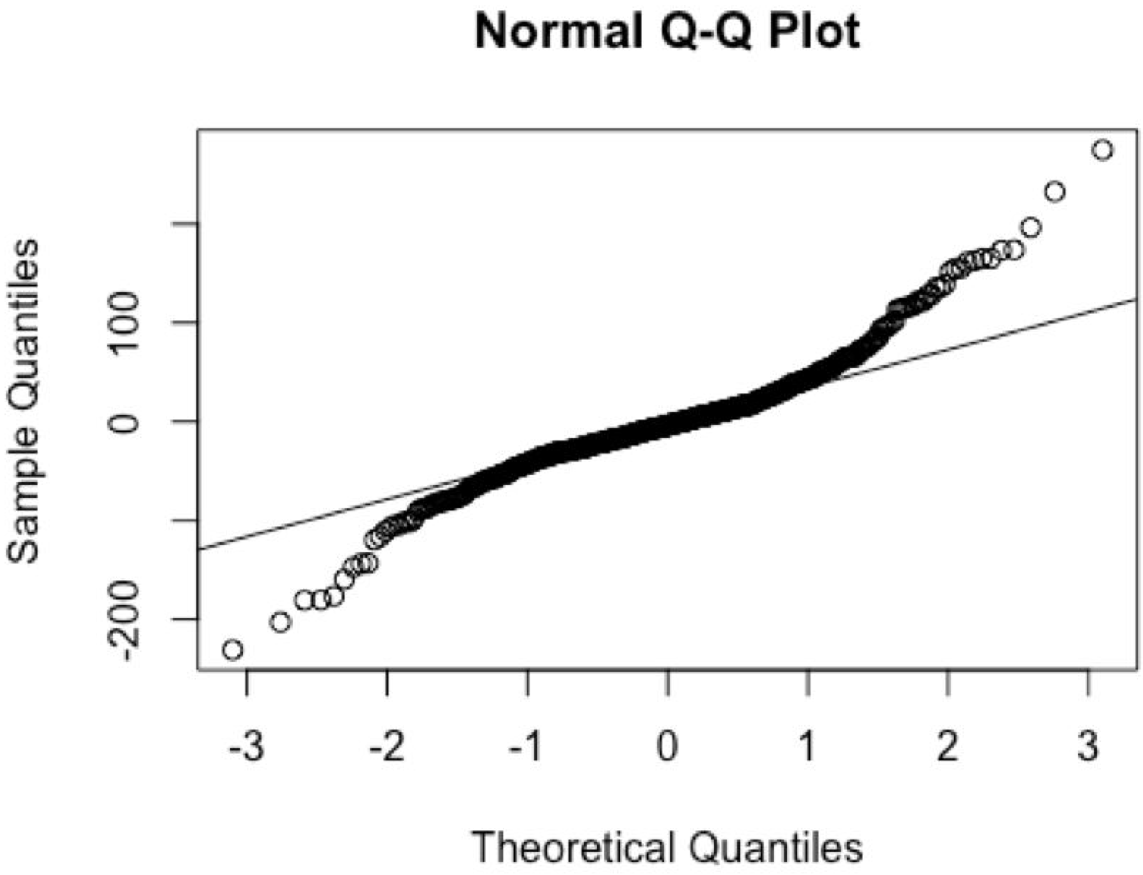
```
hist(e)
```


Histogram of e
Frequency
0
50
100
150
200
-200
-100
0
100
200
300
e

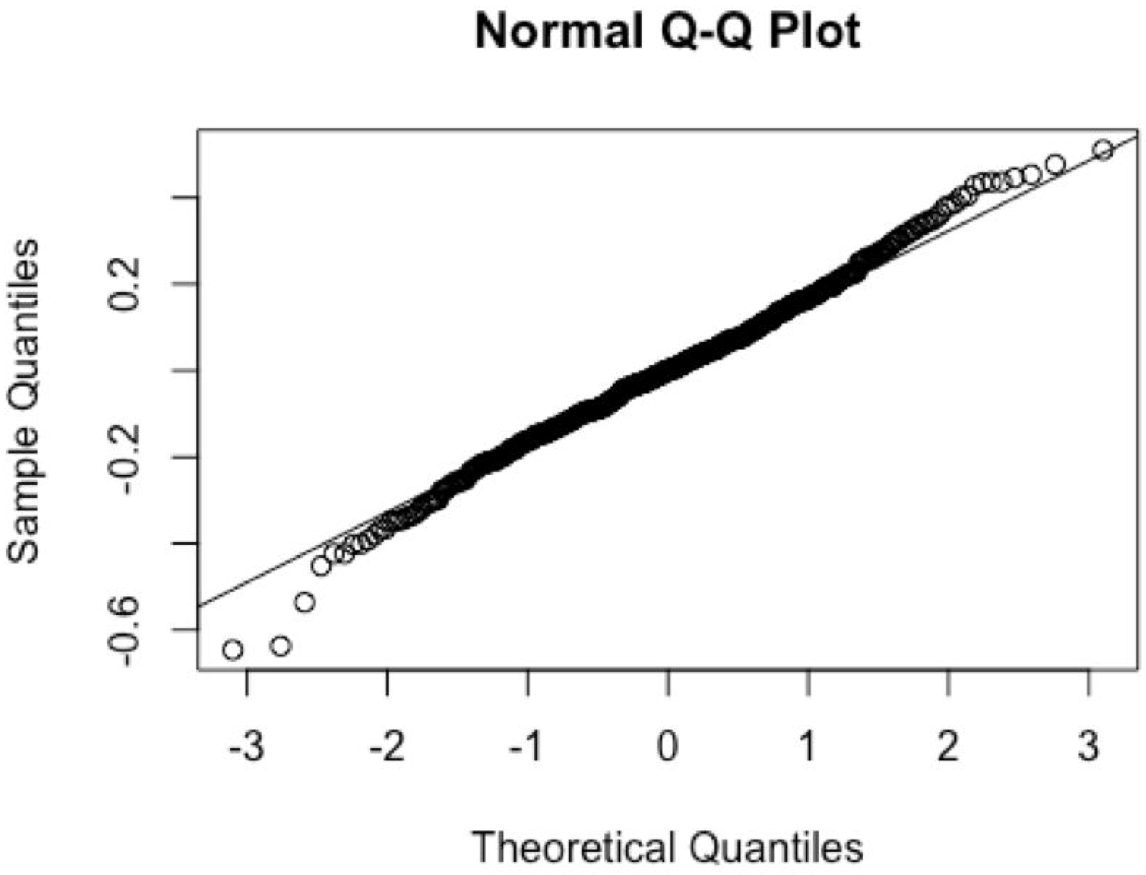
```
hist(e2)
```



```
qqnorm(e)
qqline(e)
```



```
qqnorm(e2)
qqline(e2)
```



```
#potential outlier on the left but we can ignore because of the
symmetrical pattern of the rest of the histogram
```