STOR 455 Homework #3

40 points - Due 2/21 at 5:00pm

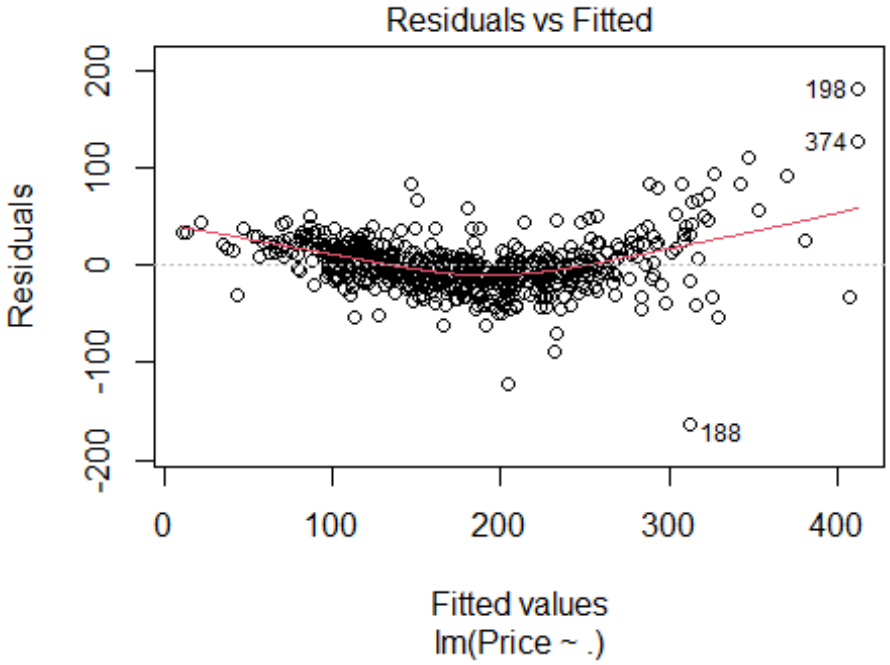
```
library(readr)
library(car)

## Loading required package: carData

library(carData)
library(corrplot)

## corrplot 0.92 loaded

library(leaps)

AmesTrain15 <- read_csv('AmesTrain15.csv')

## Rows: 600 Columns: 42

## -- Column specification ----------------------------------------
------------
## Delimiter: ","
## chr (15): LotConfig, HouseStyle, ExteriorQ, ExteriorC, Foundation,
BasementH...
## dbl (27): Order, Price, LotFrontage, LotArea, Quality, Condition,
YearBuilt,...

##
## i Use `spec()` to retrieve the full column specification for this
data.
## i Specify the column types or set `show_col_types = FALSE` to quiet
this message.

source("https://raw.githubusercontent.com/JA-
McLean/STOR455/master/scripts/ShowSubsets.R")
```

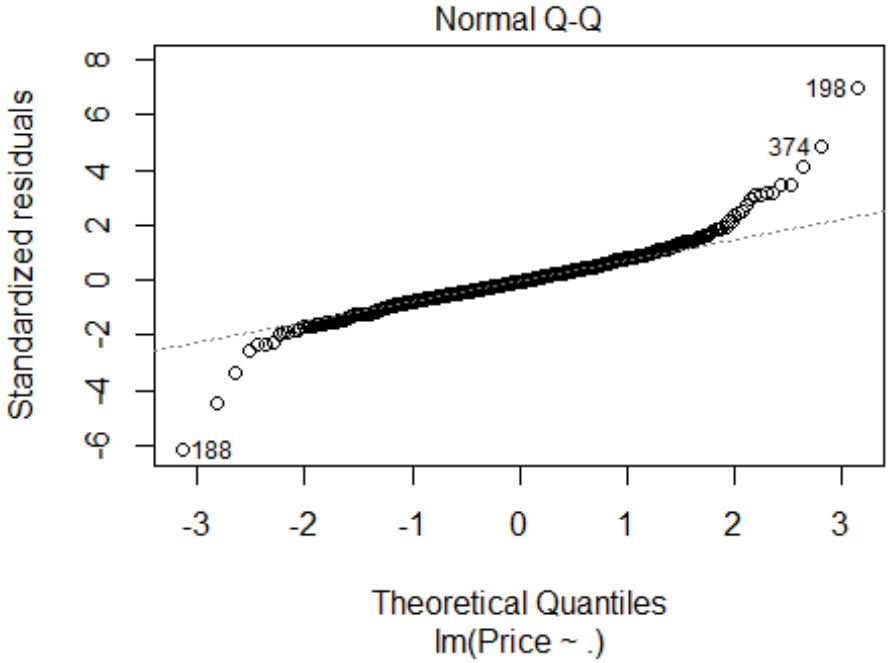
Part 1. Build an initial "basic" model

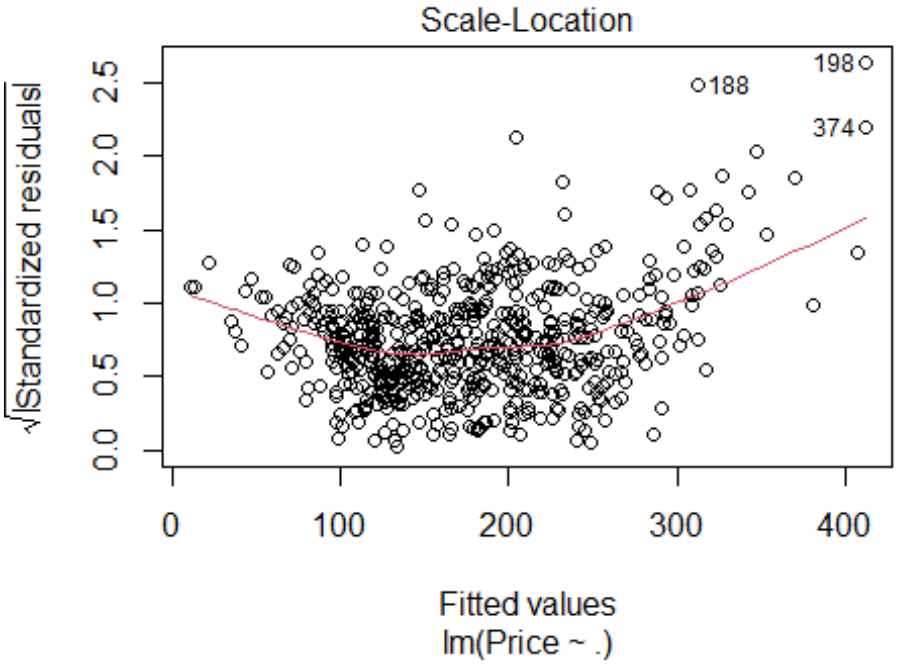
mod1: the linear relationship is overall good. The P value of some predictors are large, which are not good for being predictors or because of multicollinearity. The VIF of some variables are lager than 5.

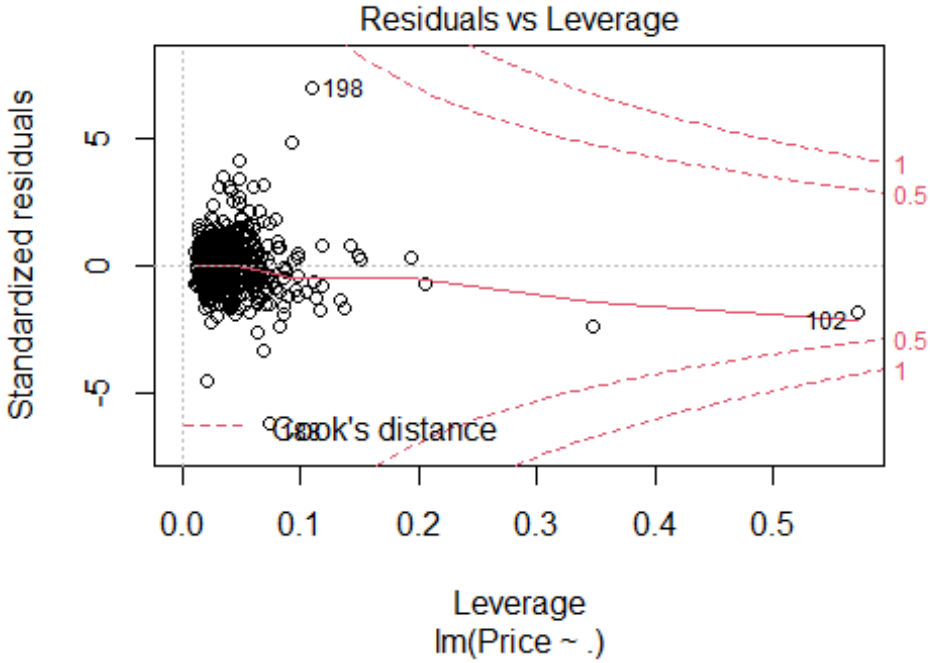
```
library(dplyr)

##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

#remove the categorical variables
newdata=select_if(AmesTrain15,is.numeric)
#remove the Order variable and predictors that are exactly related
newdata1=subset(newdata,select =-c(Order,BasementSF,GroundSF))

mod1 = lm(Price~., data=newdata1)
plot(mod1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Price ~ .)
198
374
188
200
100
0
-100
-200
0
100
200
300
400


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Price ~ .)
198
374
188
8
6
4
2
0
-2
-4
-6
-3
-2
-1
0
1
2
3






```
summary(mod1)

##
## Call:
```

```
## lm(formula = Price ~ ., data = newdata1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -162.465  -14.700   -1.206   12.526  180.199
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.450e+03  1.696e+02  -8.548  < 2e-16 ***
## LotFrontage      1.011e-01  3.777e-02   2.678 0.007617 **
## LotArea          7.212e-04  1.049e-04   6.878 1.59e-11 ***
## Quality          1.334e+01  1.366e+00   9.763  < 2e-16 ***
## Condition        4.400e+00  1.216e+00   3.618 0.000323 ***
## YearBuilt        4.126e-01  6.975e-02   5.916 5.67e-09 ***
## YearRemodel      2.819e-01  8.239e-02   3.421 0.000667 ***
## BasementFinSF    3.224e-02  5.369e-03   6.005 3.39e-09 ***
## BasementUnFinSF  1.443e-02  4.768e-03   3.027 0.002583 **
## FirstSF          6.450e-02  7.666e-03   8.413 3.16e-16 ***
## SecondSF         4.927e-02  6.224e-03   7.917 1.26e-14 ***
## BasementFBath    3.569e+00  3.178e+00   1.123 0.261892
## BasementHBath   -3.492e-01  5.200e+00  -0.067 0.946488
## FullBath         1.304e+00  3.303e+00   0.395 0.693112
## HalfBath         3.150e+00  3.328e+00   0.947 0.344271
## Bedroom         -4.010e+00  2.225e+00  -1.802 0.071996 .
## TotalRooms       1.211e+00  1.492e+00   0.812 0.417213
## Fireplaces       3.797e+00  2.211e+00   1.717 0.086499 .
## GarageCars       7.756e-01  3.603e+00   0.215 0.829629
## GarageSF         3.485e-02  1.193e-02   2.922 0.003619 **
## WoodDeckSF       6.779e-03  8.982e-03   0.755 0.450737
## OpenPorchSF      2.357e-02  1.992e-02   1.183 0.237152
## EnclosedPorchSF  4.911e-02  2.249e-02   2.184 0.029382 *
## ScreenPorchSF    6.079e-02  1.808e-02   3.363 0.000822 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.48 on 576 degrees of freedom
## Multiple R-squared:  0.8642, Adjusted R-squared:  0.8588
## F-statistic: 159.4 on 23 and 576 DF,  p-value: < 2.2e-16

vif(mod1)

##     LotFrontage         LotArea         Quality       Condition
YearBuilt
##        1.138633        1.153816        2.831580        1.483626
3.755010
##     YearRemodel   BasementFinSF BasementUnFinSF         FirstSF
SecondSF
##        2.410622        3.843505        3.669310        5.364219
5.686039
##   BasementFBath   BasementHBath        FullBath        HalfBath
```

```
Bedroom
##        2.153095        1.196368        2.667441        2.249405
2.355825
##      TotalRooms      Fireplaces      GarageCars        GarageSF
WoodDeckSF
##        4.046101        1.506368        6.328314        5.530520
1.138736
##     OpenPorchSF EnclosedPorchSF   ScreenPorchSF
##        1.320660        1.302683        1.118973
```

backward method

```
MSE = (summary(mod1)$sigma)^2
backward_mod = step(mod1, scale=MSE,trace=FALSE)
summary(backward_mod)

##
## Call:
## lm(formula = Price ~ LotFrontage + LotArea + Quality + Condition +
##     YearBuilt + YearRemodel + BasementFinSF + BasementUnFinSF +
##     FirstSF + SecondSF + Bedroom + Fireplaces + GarageSF +
EnclosedPorchSF +
##     ScreenPorchSF, data = newdata1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -163.707  -15.529   -1.181   12.460  182.143
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.541e+03  1.478e+02 -10.426  < 2e-16 ***
## LotFrontage      1.049e-01  3.711e-02   2.828 0.004851 **
## LotArea          7.098e-04  1.034e-04   6.862 1.74e-11 ***
## Quality          1.338e+01  1.356e+00   9.865  < 2e-16 ***
## Condition        4.302e+00  1.188e+00   3.622 0.000318 ***
## YearBuilt        4.340e-01  6.360e-02   6.824 2.22e-11 ***
## YearRemodel      3.087e-01  7.794e-02   3.961 8.37e-05 ***
## BasementFinSF    3.397e-02  5.025e-03   6.760 3.34e-11 ***
## BasementUnFinSF  1.311e-02  4.612e-03   2.843 0.004632 **
## FirstSF          7.081e-02  6.676e-03  10.607  < 2e-16 ***
## SecondSF         5.619e-02  4.373e-03  12.850  < 2e-16 ***
## Bedroom         -3.549e+00  1.942e+00  -1.827 0.068144 .
## Fireplaces       3.934e+00  2.182e+00   1.803 0.071897 .
## GarageSF         3.602e-02  6.895e-03   5.223 2.45e-07 ***
## EnclosedPorchSF  5.098e-02  2.238e-02   2.278 0.023064 *
## ScreenPorchSF    5.883e-02  1.782e-02   3.302 0.001017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 27.42 on 584 degrees of freedom
## Multiple R-squared:  0.863,  Adjusted R-squared:  0.8595
## F-statistic: 245.2 on 15 and 584 DF,  p-value: < 2.2e-16
```

stepwise method

Both of the backward and stepwise methods have the same adjusted R square. Compared with mod1, the model improves from 0.8588 to 0.8595. The predictor of 'Bedroom' and 'Fireplaces' have p values greater than 0.05, which are not significant at a 5% level. There is no multicollinearity, all VIF is smaller than 5. I choose it as my basic model.

```
none = lm(Price~1, data=newdata1)
stepwise_mod = step(none, scope=list(upper=mod1),
scale=MSE,trace=FALSE)
summary(stepwise_mod)

##
## Call:
## lm(formula = Price ~ Quality + FirstSF + SecondSF + YearBuilt +
##     BasementFinSF + GarageSF + LotArea + YearRemodel + Condition +
##     ScreenPorchSF + LotFrontage + BasementUnFinSF + EnclosedPorchSF
+
##     Bedroom + Fireplaces, data = newdata1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -163.707  -15.529   -1.181   12.460  182.143
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.541e+03  1.478e+02 -10.426  < 2e-16 ***
## Quality          1.338e+01  1.356e+00   9.865  < 2e-16 ***
## FirstSF          7.081e-02  6.676e-03  10.607  < 2e-16 ***
## SecondSF         5.619e-02  4.373e-03  12.850  < 2e-16 ***
## YearBuilt        4.340e-01  6.360e-02   6.824 2.22e-11 ***
## BasementFinSF    3.397e-02  5.025e-03   6.760 3.34e-11 ***
## GarageSF         3.602e-02  6.895e-03   5.223 2.45e-07 ***
## LotArea          7.098e-04  1.034e-04   6.862 1.74e-11 ***
## YearRemodel      3.087e-01  7.794e-02   3.961 8.37e-05 ***
## Condition        4.302e+00  1.188e+00   3.622 0.000318 ***
## ScreenPorchSF    5.883e-02  1.782e-02   3.302 0.001017 **
## LotFrontage      1.049e-01  3.711e-02   2.828 0.004851 **
## BasementUnFinSF  1.311e-02  4.612e-03   2.843 0.004632 **
## EnclosedPorchSF  5.098e-02  2.238e-02   2.278 0.023064 *
## Bedroom         -3.549e+00  1.942e+00  -1.827 0.068144 .
## Fireplaces       3.934e+00  2.182e+00   1.803 0.071897 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 27.42 on 584 degrees of freedom
## Multiple R-squared:  0.863,  Adjusted R-squared:  0.8595
## F-statistic: 245.2 on 15 and 584 DF,  p-value: < 2.2e-16

vif(stepwise_mod)

##         Quality         FirstSF        SecondSF       YearBuilt
BasementFinSF
##        2.803892        4.085771        2.819819        3.136790
3.382215
##        GarageSF         LotArea     YearRemodel       Condition
ScreenPorchSF
##        1.856329        1.127769        2.166455        1.421239
1.091640
##     LotFrontage BasementUnFinSF EnclosedPorchSF         Bedroom
Fireplaces
##        1.104353        3.449046        1.295556        1.803771
1.473448
```

part 2: Residual analysis

Linearity is ok, since the residuals vs fitted plot is described by a horizontal line.

qq plot: a little bit of a compacted on both sides of this because the ends points go down below the line a little bit.

But overall, it is good without clearly huge skews.

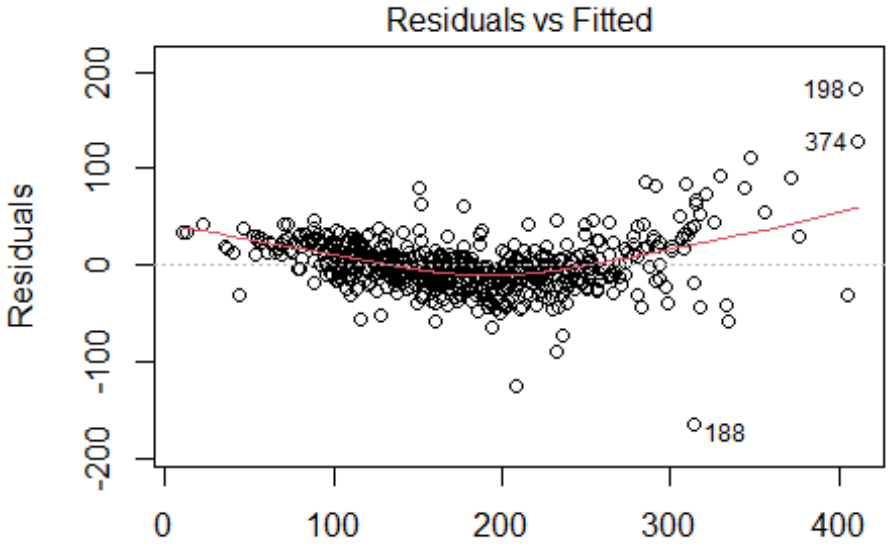
The 12 houses indexes listed are outliers, they have studentized residuals larger than 3 or or smaller than -3.

All cook's distance are smaller than 0.5, so there is no data points have a drastic effect on the whole model.

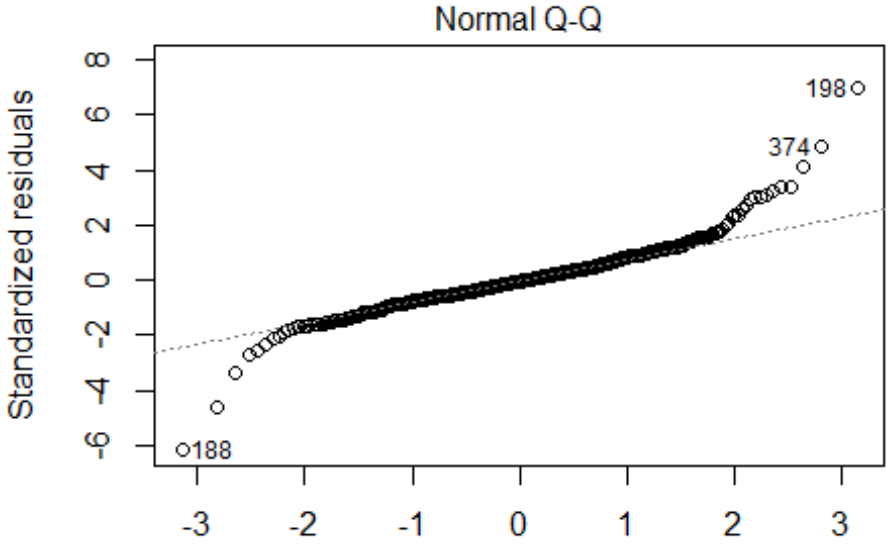
Overall, the result of residual analysis is good.

I made an adjustment based on P value above, which deletes the predictor of 'Bedroom' and 'Fireplaces', the adjusted R is smaller, which is 0.8582.

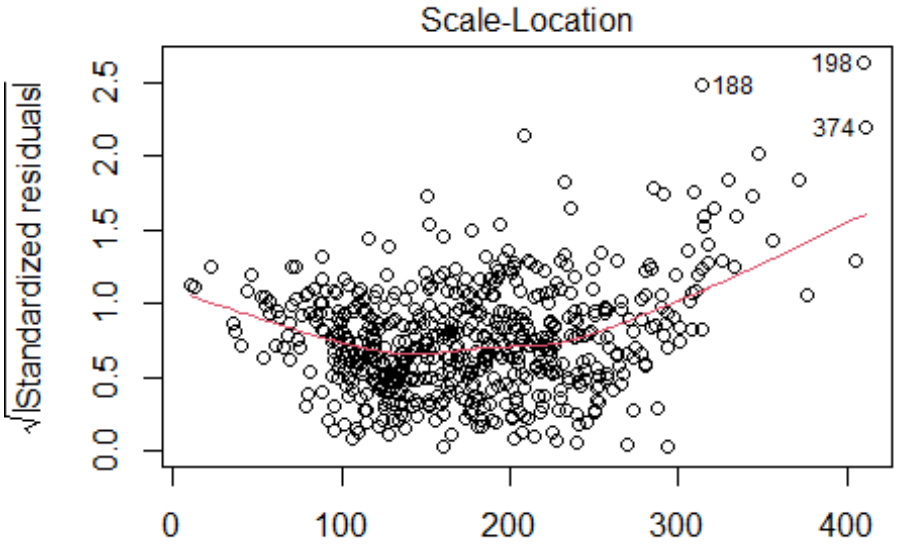
```
plot(stepwise_mod)
```



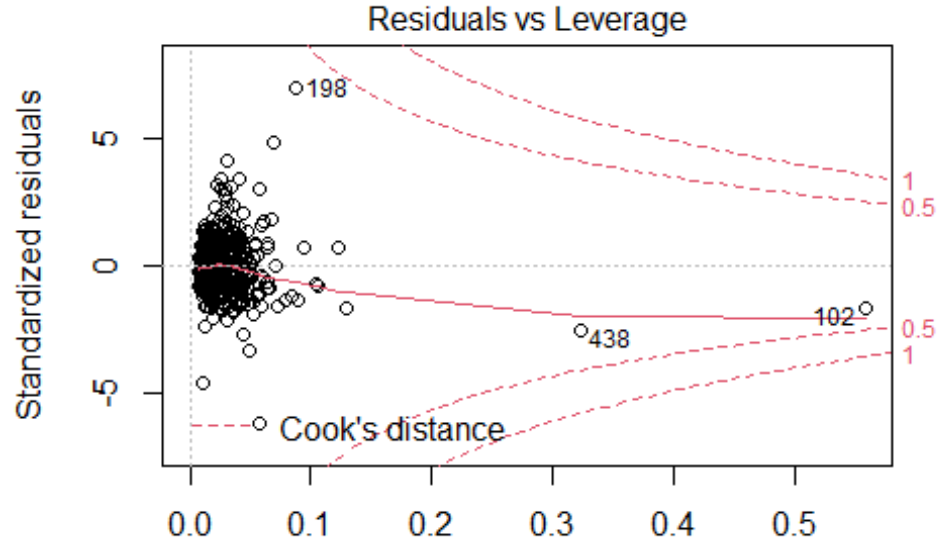
Price ~ Quality + FirstSF + SecondSF + YearBuilt + BasementFinSF +



Price ~ Quality + FirstSF + SecondSF + YearBuilt + BasementFinSF +



Price ~ Quality + FirstSF + SecondSF + YearBuilt + BasementFinSF +



Price ~ Quality + FirstSF + SecondSF + YearBuilt + BasementFinSF +

```
head(sort(abs(rstudent(stepwise_mod)), decreasing=TRUE), 15)
```

```
##       198       188       374       204        62        70       581
268
```

```
## 7.257870 6.355659 4.909422 4.681470 4.146571 3.437666 3.397817
3.370246
##      202      537      535      386      572      228      203
## 3.226921 3.116382 3.043575 3.012107 2.993612 2.748336 2.717243

head(sort(cooks.distance(stepwise_mod), decreasing=TRUE), n=5)

##       198       102       438       188       374
## 0.2915410 0.2152589 0.1902544 0.1449232 0.1070727

adjustmod = lm(formula = Price ~ Quality + FirstSF + SecondSF +
YearBuilt +
    BasementFinSF + GarageSF + LotArea + YearRemodel + Condition +
    ScreenPorchSF + LotFrontage + BasementUnFinSF + EnclosedPorchSF,
data = newdata1)
summary(adjustmod)

##
## Call:
## lm(formula = Price ~ Quality + FirstSF + SecondSF + YearBuilt +
##     BasementFinSF + GarageSF + LotArea + YearRemodel + Condition +
##     ScreenPorchSF + LotFrontage + BasementUnFinSF + EnclosedPorchSF,
##     data = newdata1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -161.815  -15.967   -2.116   12.778  179.683
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.559e+03  1.468e+02 -10.613  < 2e-16 ***
## Quality          1.406e+01  1.337e+00  10.516  < 2e-16 ***
## FirstSF          6.954e-02  6.018e-03  11.555  < 2e-16 ***
## SecondSF         5.281e-02  3.326e-03  15.878  < 2e-16 ***
## YearBuilt        4.321e-01  6.355e-02   6.799 2.60e-11 ***
## BasementFinSF    3.441e-02  5.038e-03   6.830 2.13e-11 ***
## GarageSF         3.875e-02  6.795e-03   5.702 1.87e-08 ***
## LotArea          7.254e-04  1.034e-04   7.017 6.28e-12 ***
## YearRemodel      3.143e-01  7.645e-02   4.111 4.51e-05 ***
## Condition        4.323e+00  1.182e+00   3.657 0.000278 ***
## ScreenPorchSF    6.340e-02  1.779e-02   3.564 0.000394 ***
## LotFrontage      1.002e-01  3.720e-02   2.693 0.007277 **
## BasementUnFinSF  1.149e-02  4.585e-03   2.507 0.012442 *
## EnclosedPorchSF  5.577e-02  2.239e-02   2.490 0.013041 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.53 on 586 degrees of freedom
## Multiple R-squared:  0.8613, Adjusted R-squared:  0.8582
## F-statistic: 279.9 on 13 and 586 DF,  p-value: < 2.2e-16
```

Part 3 & Part 4: please see the group work.

Part 5: Final model

(The fancier final model is put in group work.)

A 95% prediction interval for the Price of this house is [228.7178, 348.0183]

```
newHouse=data.frame(LotFrontage=400,
LotArea=21540,
Quality=7,
Condition=5,
YearBuilt=1983,
YearRemodel=1999,
BasementUnFinSF=757,
BasementFinSF=0,
BasementSF=757,
FirstSF=1485,
SecondSF=947,
BasementFBath=0,
Bedroom=4,
Fireplaces=1,
GarageCars=2,
GarageSF=588,
EnclosedPorchSF=0,
ScreenPorchSF=0)

predict.lm(stepwise_mod, newHouse, interval="prediction")
```

```
##        fit      lwr      upr
## 1 288.3681 228.7178 348.0183
```