Final Report

Yueqi Chen

06/08/2020

- UW ID: 20737045

1 L.oading data

```
load("final.Rdata")
```

2. Preliminary Analysis

By analysing this dataset, I intend to know what factors will affet the house prices and how much impact they can have so that we can make predictions on unknown house prices with such information.

In my dataset, I have 18 explanatory variables that are considered to be possible to have impact on house prices. They are:

- bathrm: the number of bathrooms in the house;
- hf_bathrm: the number of half bathrooms in the house;
- heat: the heating type in the house;
- ac: whether the house has air conditioning or not;
- rooms: the number of rooms in the house;
- bedrm: the number of bedrooms in the house;
- ayb: the earliest time the main portion of the house was built;
- yr_rmdl: the year when the house structure was remodelled;
- eyb: the year an improvement was built more recent than actual year built;
- stories: the number of stories in primary dwelling in the house;
- saledate: date of most recent sale of the house, in the form of "yyyy-mm-dd 00:00:00";
- gba: gross building area of the house in square feet;
- style: the house style;
- grade: reviews of the house;
- extwall: the material of exterior wall;
- kitchens: the number of kitchens in the house;
- fireplaces: the number of fireplaces in the house;
- landarea: land area of property in square feet.

2.1 summaries of variables

2.1.1 Response Variable (price)

```
hist(dtrain$price, main = "Histogram of House Prices", xlab = "House Price")
```

From the histogram of house price above, we can see that it has a right-skewed distribution. In this case, we know that there is a limit for house price, which causes its peak off center. Apparently, we can understand that a house will be much harder to be sold when house price is (extremely) high and on the other hand, a house is not hard to be sold with a low price. However, mostly houses with low price will have some limitations on themselves, such as limited rooms or no air conditioning. Therefore, people would like to choose houses with proper and affordable price that meet their living demands.

2.1.2 Numerical Variables

```
summary(dtrain[c("bathrm", "hf_bathrm", "rooms", "bedrm", "ayb", "yr_rmdl", "eyb",
                 "stories", "gba", "kitchens", "fireplaces", "landarea")])
```

```
##      bathrm        hf_bathrm          rooms            bedrm
##  Min.   :0.000   Min.   :0.0000   Min.   : 0.000   Min.   :0.000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.: 6.000   1st Qu.:3.000
##  Median :2.000   Median :1.0000   Median : 7.000   Median :3.000
##  Mean   :2.038   Mean   :0.6178   Mean   : 6.849   Mean   :3.395
##  3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.: 8.000   3rd Qu.:4.000
##  Max.   :6.000   Max.   :3.0000   Max.   :19.000   Max.   :8.000
##
##       ayb          yr_rmdl          eyb          stories           gba
##  Min.   :1870   Min.   :1925   Min.   :1928   Min.   :1.000   Min.   : 535
##  1st Qu.:1922   1st Qu.:2004   1st Qu.:1957   1st Qu.:1.500   1st Qu.:1200
##  Median :1929   Median :2010   Median :1964   Median :2.000   Median :1426
##  Mean   :1938   Mean   :2006   Mean   :1967   Mean   :1.824   Mean   :1529
##  3rd Qu.:1947   3rd Qu.:2014   3rd Qu.:1967   3rd Qu.:2.000   3rd Qu.:1759
##  Max.   :2017   Max.   :2018   Max.   :2017   Max.   :9.000   Max.   :5129
##                 NA's   :578                   NA's   :2
##     kitchens       fireplaces        landarea
##  Min.   :0.000   Min.   :0.0000   Min.   :  696
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.: 3739
##  Median :1.000   Median :1.0000   Median : 4776
##  Mean   :1.016   Mean   :0.5756   Mean   : 5009
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.: 6000
##  Max.   :2.000   Max.   :5.0000   Max.   :16098
##
```

The summary above summarizes the numeric variables in the dataset. As we can see, there are some missing values in our dataset that need to be solved. For example, variable "yr_rmdl" has 578 NA's and variable "stories" has 2 NA's. For each of them, I used different ways to fill in the missing values.

```
dtrain$yr_rmdl[is.na(dtrain$yr_rmdl)] <-
  round(mean(dtrain$yr_rmdl, na.rm = TRUE) - mean(dtrain$ayb)) +
  dtrain$ayb[is.na(dtrain$yr_rmdl)]
```

- yr_rmdl: I replaced the missing data (NA) by a short calculation. I find mean of the subtraction between ayb and known yr_rmdl and then add this mean to the ayb to get the unknown yr_rmdl. Since yr_rmdl should happen after (greater) ayb, simply using mean of known yr_rmdl to replace the missing data will cause an unreasonable result that yr_rmdl is smaller than ayb.

```
dtrain$stories[is.na(dtrain$stories)] <- median(dtrain$stories, na.rm = TRUE)
```

- stories: I replaced the missing data (NA) with the median of the known stories values. Since the number of stories have a decimal form of .00, .25, .50, .75. The mean value will not keep in this form. Therefore, rather than using mean, median value is the better choice to replace the missing data.

2.1.3 Categorical Variables

```
summary(factor(dtrain$heat))
```

```
##     Air Exchng  Elec Base Brd     Forced Air Gravity Furnac  Hot Water Rad
##              1              1            631              1            416
##        Ht Pump        No Data   Wall Furnace      Warm Cool Water Base Brd
##             26              1              1            224              1
```

```
summary(factor(dtrain$ac))
```

```
##   N   Y
## 359 944
```

```
summary(factor(dtrain$style))
```

```
##         1 Story   1.5 Story Fin 1.5 Story Unfin         2 Story   2.5 Story Fin
##             230             115               5             832              77
## 2.5 Story Unfin         3 Story         4 Story        Bi-Level         Default
##              22              13               1               1               1
##     Split Foyer     Split Level
##               3               3
```

```
summary(factor(dtrain$grade))
```

```
## Above Average       Average  Fair Quality  Good Quality   Low Quality
##           579           640            15            63             1
##      Superior     Very Good
##             1             4
```

```
summary(factor(dtrain$extwall))
```

```
##          Adobe       Aluminum   Brick Veneer   Brick/Siding    Brick/Stone
##              1             72             14             89              5
##   Brick/Stucco   Common Brick       Concrete Concrete Block     Face Brick
##             10            474              4              1              4
##      Hardboard  Metal Siding        Shingle          Stone   Stone Veneer
##             11              3             70              6              5
##   Stone/Siding   Stone/Stucco         Stucco   Stucco Block   Vinyl Siding
##             16              2             77              2            352
##    Wood Siding
##             85
```

The information above shows that the types of each categorical variable and the number of sold houses under that type. As we can see, most sold houses have heat types of "Forced Air", "Hot Water Rad", or "Warm Cool". Also, people are more likely to buy houses with air conditioning. Houses with two-story are the most popular style among all other 11 styles. Houses with quality of "above average" and "average" are pretty popular and it is not hard to understand since these kinds of houses are cost-effective. The exterior walls built by"common brick" or "vinvl siding" are two most popular types.

2.1.5 New Defined Variable

```
dtrain$saleYear <- as.integer(substr(dtrain$saledate, 1, 4))
```

Since saledate is in the form of a date, "yyyy-mm-dd 00:00:00". I decide to extract the useful information to me. I get the year of it and then change it to the numeric variable.

2.1.4 x-y relationship

```
cor(dtrain$price,
    dtrain[c("bathrm", "hf_bathrm", "rooms", "bedrm", "ayb", "yr_rmdl", "eyb",
             "gba", "kitchens", "fireplaces", "landarea", "saleYear")])
```

```
##         bathrm hf_bathrm     rooms     bedrm         ayb   yr_rmdl       eyb
## [1,] 0.5320137 0.1851741 0.3683644 0.4669599 -0.06812407 0.1089956 0.2327073
##           gba  kitchens fireplaces  landarea  saleYear
## [1,] 0.494829 0.1452957  0.2041408 0.1466834 0.6724294
```

The table above shows the correlation coeffecients between price and a numeric explanatory variable. Price and bathrm/bedrm/gba/saleYear have relatively high positive correlation coefficients, which means they have a relatively strong linear relationship. In other words, the more bathrooms (berooms), the higher the house price. The greater the gross building area, the higher the house price. Or the more recent the sale year, the higher the price. We can consider that these four variables are possibly the factors that can affect the house price.

2.1.5 x-x relationship

```
cor(dtrain[c("bathrm", "hf_bathrm", "rooms", "bedrm", "ayb", "yr_rmdl", "eyb",
             "gba", "kitchens", "fireplaces", "landarea", "saleYear")])
```

```
##                bathrm  hf_bathrm      rooms     bedrm         ayb     yr_rmdl
## bathrm     1.00000000 0.04891745 0.45569726 0.6177090  0.17297846  0.34957633
## hf_bathrm  0.04891745 1.00000000 0.19268416 0.1511272  0.15848815  0.18748963
## rooms      0.45569726 0.19268416 1.00000000 0.6446478  0.09301388  0.17009515
## bedrm      0.61770896 0.15112721 0.64464781 1.0000000  0.10240592  0.23559275
## ayb        0.17297846 0.15848815 0.09301388 0.1024059  1.00000000  0.81867441
## yr_rmdl    0.34957633 0.18748963 0.17009515 0.2355928  0.81867441  1.00000000
## eyb        0.42186646 0.21740675 0.25863139 0.3209646  0.79086871  0.82589709
## gba        0.51322358 0.30150128 0.57172393 0.5824250  0.11379789  0.19546114
## kitchens   0.11993694 0.05392731 0.10913602 0.1628645 -0.06555062 -0.01164842
## fireplaces 0.05531792 0.13781265 0.09370245 0.1026266 -0.02815414 -0.11563713
## landarea   0.11824493 0.07949904 0.22155801 0.1924927 -0.05923628 -0.09103139
## saleYear   0.32569405 0.04805270 0.10022692 0.2353410  0.02852529  0.18490531
##                   eyb        gba    kitchens  fireplaces    landarea
## bathrm     0.42186646 0.51322358  0.11993694  0.05531792  0.11824493
## hf_bathrm  0.21740675 0.30150128  0.05392731  0.13781265  0.07949904
## rooms      0.25863139 0.57172393  0.10913602  0.09370245  0.22155801
## bedrm      0.32096463 0.58242499  0.16286447  0.10262664  0.19249274
```

```
## ayb         0.79086871 0.11379789 -0.06555062 -0.02815414 -0.05923628
## yr_rmdl     0.82589709 0.19546114 -0.01164842 -0.11563713 -0.09103139
## eyb         1.00000000 0.32910171  0.02404333 -0.09022458 -0.02225866
## gba         0.32910171 1.00000000  0.08492081  0.20893033  0.33626049
## kitchens    0.02404333 0.08492081  1.00000000  0.03355654  0.02134784
## fireplaces -0.09022458 0.20893033  0.03355654  1.00000000  0.10849148
## landarea   -0.02225866 0.33626049  0.02134784  0.10849148  1.00000000
## saleYear    0.21595230 0.10589945  0.07221087 -0.06308372 -0.04707365
##               saleYear
## bathrm      0.32569405
## hf_bathrm   0.04805270
## rooms       0.10022692
## bedrm       0.23534096
## ayb         0.02852529
## yr_rmdl     0.18490531
## eyb         0.21595230
## gba         0.10589945
## kitchens    0.07221087
## fireplaces -0.06308372
## landarea   -0.04707365
## saleYear    1.00000000
```

The table above shows the correlation coeffecients between each pair of numeric predictors. Variable "bathrm" and "bedroom" / "rooms" and "bedroom" have relatively high correlation coefficients. Similarly, "yr_rmdl" and "ayb"/"eyb" have high correlation coefficients. Therefore, we may prefer to avoid having two variables with high correlation coefficients in the same model.

3. Model Building

I use the stepwise regression with AIC on the square-root-transformed response, and the final model is:

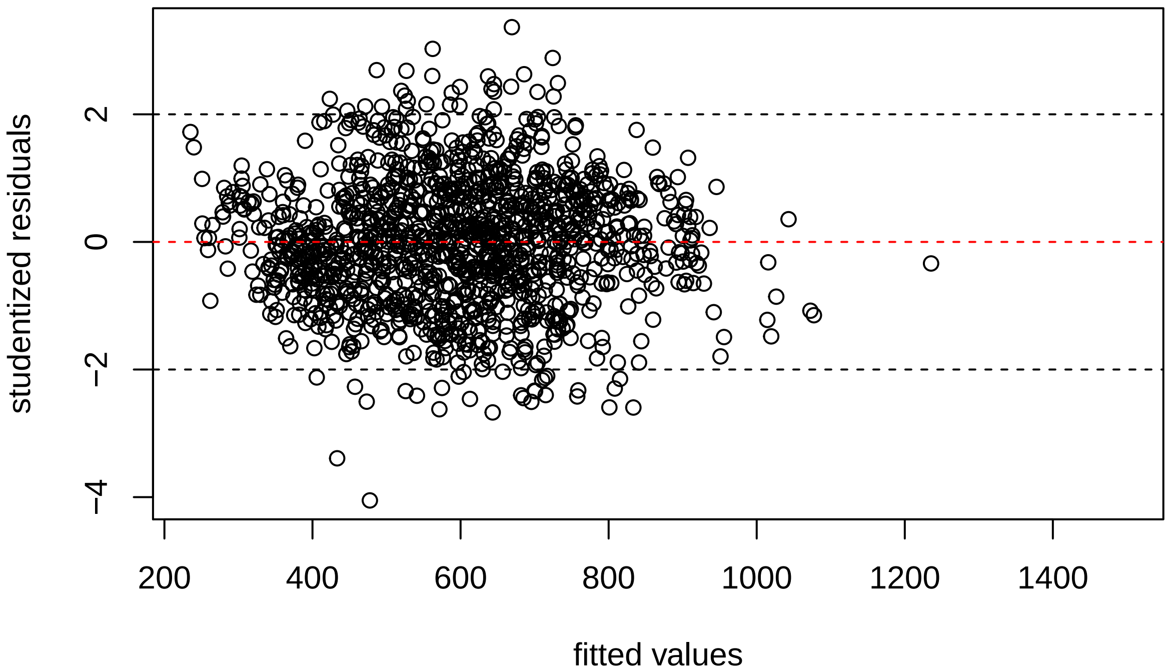
$sqrt(price) \sim saleYear + gba + grade + ayb + bathrm + fireplaces + extwall + eyb + hf\_bathrm + rooms + landarea + kitchens + saleYear: gba + saleYear: grade + saleYear: ayb + saleYear: bathrm + saleYear: eyb + ayb: eyb + fireplaces: eyb$, **where**

$saleYear: gba$ is $saleYear * gba$,
$saleYear: grade$ is $saleYear * grade$,
$saleYear: ayb$ is $saleYear * ayb$,
$saleYear: bathrm$ is $saleYear * bathrm$,
$saleYear: eyb$ is $saleYear * eyb$,
$ayb: eyb$ is $ayb * eyb$,
$fireplaces: eyb$ is $fireplaces * eyb$.

```r
fm <- lm(formula = sqrt(price) ~ saleYear + gba + grade + ayb + bathrm +
           fireplaces + extwall + eyb + hf_bathrm + rooms + landarea +
           kitchens + saleYear:gba + saleYear:ayb + saleYear:bathrm +
           saleYear:eyb + ayb:eyb + fireplaces:eyb, data = dtrain)
```

4. Model Checking

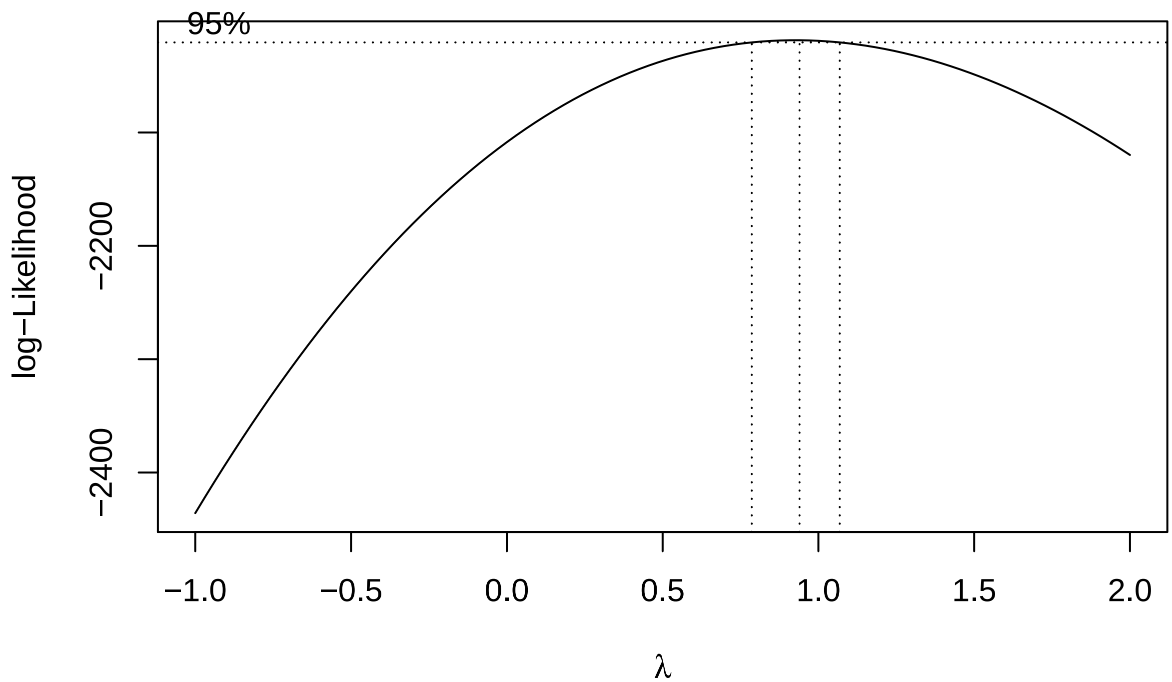
```r
plot(fitted(fm), rstudent(fm), xlab = "fitted values", ylab = "studentized residuals")
abline(a=0, b=0,lty=2, col="red")
abline(a=2, b=0, lty=2)
abline(a=-2, b=0, lty=2)
```



According to the scatter plot, the pattern, especially in the range of 0 and 500000, suggests that the constant variance assumption is violated. Thus, we do transfromation in the following step.

5. Transformation

```r
library(MASS)
boxcox(fm, lambda=seq(-1, 2, 1/20))
```



We can see that the vertex is very close to the point where $\lambda = 0.5$ in the Box-Cox plot above; thus, we pick $\lambda = 0.5$. According to the Box-Cox transformations formula,

$$g(y) = \begin{cases} y^{\lambda}, if \lambda \neq 0 \\ log(y), if \lambda = 0 \end{cases} \quad ,$$

we want to transform our response variable, $price$, to be $price^{0.5} = sqrt(price)$.

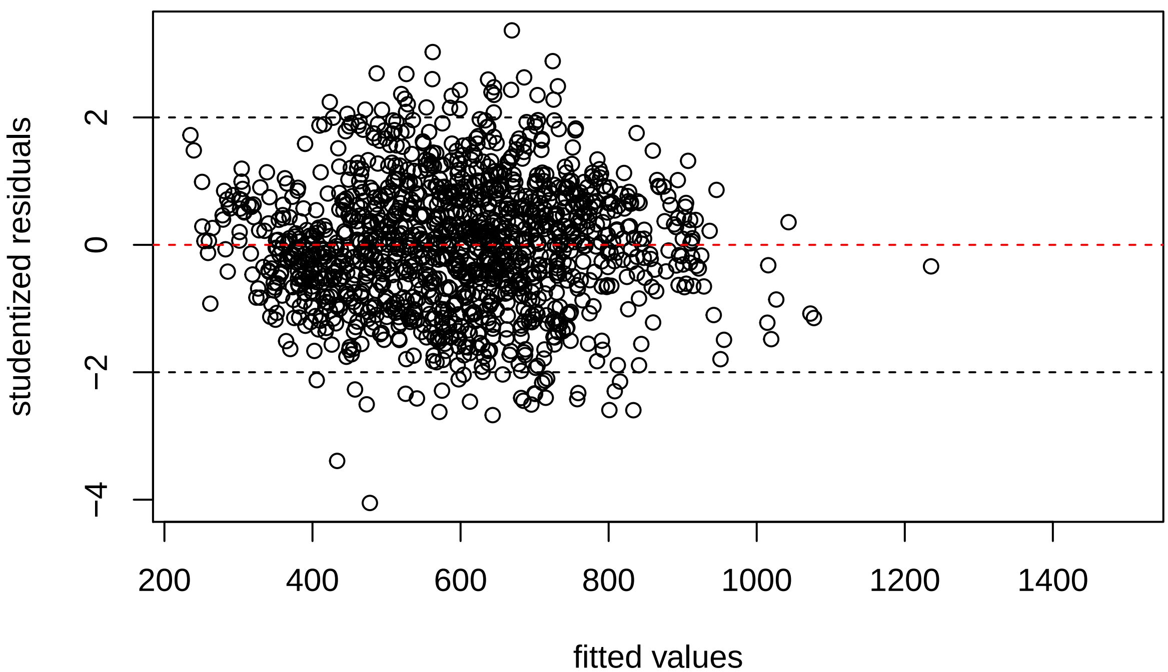
Below is our new model after transformation:

```r
## we choose lambda = 0.5 => sqrt(price)
fm <- lm(formula = sqrt(price) ~ saleYear + gba + grade + ayb + bathrm +
           fireplaces + extwall + eyb + hf_bathrm + rooms + landarea +
           kitchens + saleYear:gba + saleYear:ayb + saleYear:bathrm +
           saleYear:eyb + ayb:eyb + fireplaces:eyb, data = dtrain)
```

6. Model Checking After Transformation

6.1 checking assumptions: mean of zero, constant variance

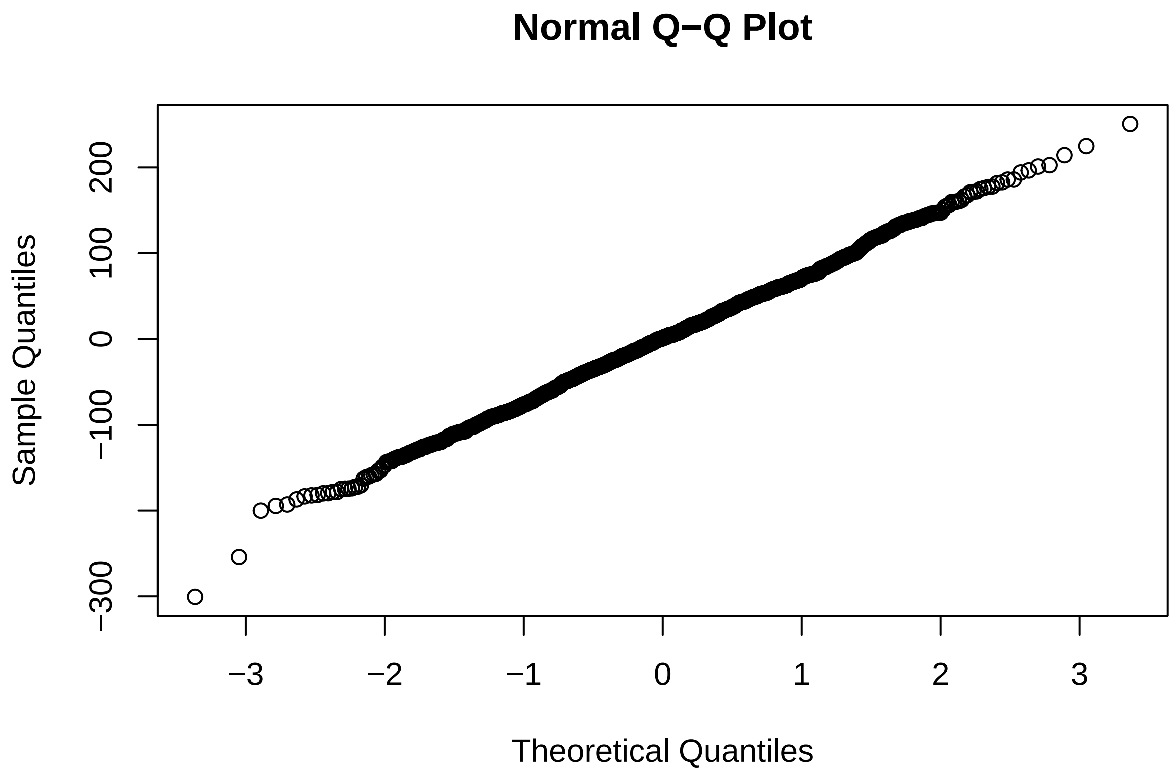
```
plot(fitted(fm), rstudent(fm), xlab = "fitted values", ylab = "studentized residuals")
abline(a=0, b=0,lty=2, col="red")
abline(a=2, b=0, lty=2)
abline(a=-2, b=0, lty=2)
```



According to the scatter plot, the pattern seems much better than the one before transformation. - Assumption of mean of zero holds: Since the studentized residuals lies within a horizontal band around zero and does not exhibit any spcial pattern. Also, approximately 95% of studentized residuals lie within (-2,2) and almost all of them are within (-3,3). - Assumption of constant variance holds: The plot shows that the studentized residuals appear to have constatn variability with respect to the fitted values. Thus, it supports the assumption.

6.2 checking normality assumption

```r
qqnorm(residuals(fm))
```



Assumption of normality holds: According to the Q-Q plot above, it is very similar to the Q-Q plot of a sample from a normal distribution.

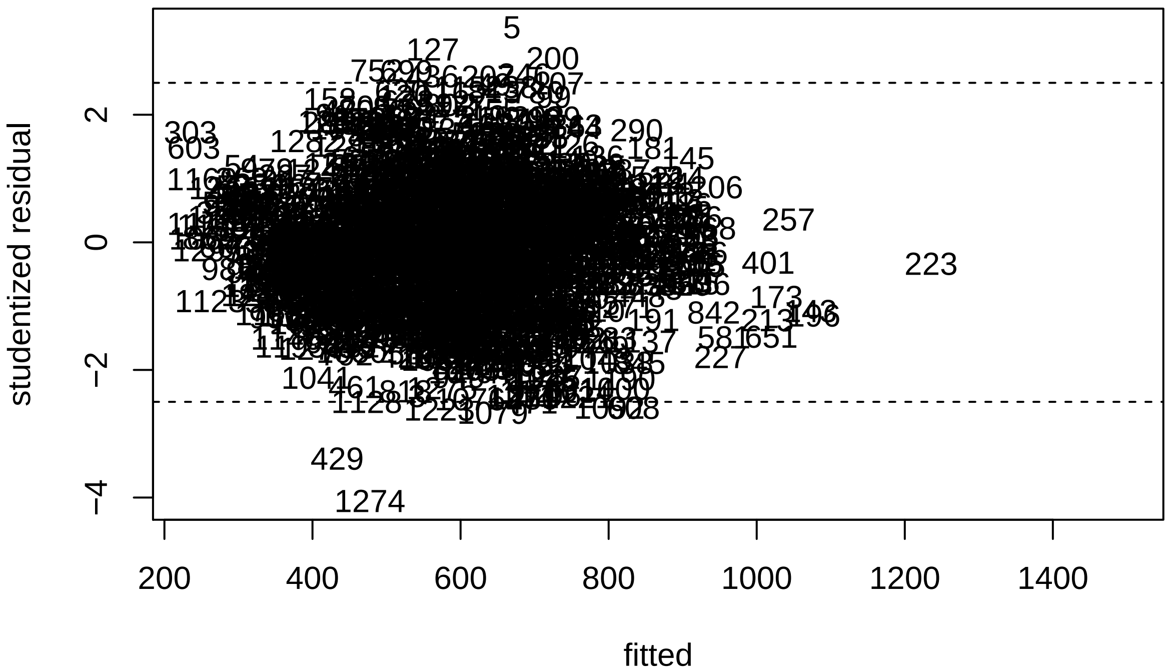
Since all assumptions hold for this new model after transformation, we choose it to be our final model: $sqrt(price) \sim saleYear + gba + grade + ayb + bathrm + fireplaces + extwall + eyb + hf\_bathrm + rooms + landarea + kitchens + saleYear: gba + saleYear: grade + saleYear: ayb + saleYear: bathrm + saleYear: eyb + ayb: eyb + fireplaces: eyb$, **where**

$saleYear: gba$ is $saleYear * gba$,
$saleYear: grade$ is $saleYear * grade$,
$saleYear: ayb$ is $saleYear * ayb$,
$saleYear: bathrm$ is $saleYear * bathrm$,
$saleYear: eyb$ is $saleYear * eyb$,
$ayb: eyb$ is $ayb * eyb$,
$fireplaces: eyb$ is $fireplaces * eyb$.

7. Data Checking

7.1 Outliers in response

```r
plot(fitted(fm), rstudent(fm), type="n", xlab = "fitted", ylab = "studentized residual")
text(fitted(fm), rstudent(fm))
abline(h=c(-2.5, 2.5), lty=2)
```



Large values of studentized residual d_i, where $|d_i| > 2.5$, indicate outliers in y. Thus, possible outliers observing from the plot above are: row 5, row 127, row 429, and row 1274. Let's look at the data in these rows:

```r
dtrain[c(5, 127, 429, 1274), c(4, 5, 7, 9, 12, 15, 20)]
```

```
##      ac rooms  ayb  eyb  price         grade saleYear
## 5     N     4 1920 1964 846300       Average     2016
## 127   Y     8 2007 2010 619600 Above Average     2007
## 429   Y     6 1923 1957  32100 Above Average     2000
## 1274  Y     6 2009 2012  31300 Above Average     2007
```

By refering to the dataset, the price of row 127 is reasonable enough. But row 5 has a really high price as an old house first built in 1920. The improvement of this house was made in 1964 and it has only 4 rooms with no air conditioner. Therefore, the price is unreasonablely high. The high price is possibly because it is a meaningful house; for example, one celebrity used to live there. However, buying such a special house is not a living demand for the majority of people. Since it will influence the prediction result, I decide to remove it.

As my analysis in part 2, houses with quality of "above average" and "average" are pretty popular. Though both row 429 and row 1274 has a very low price of 23100 and 31300 respectively, I have different considerations about these two. Row 429 shows that the house is an old house without imprevement in a very long term. Although it has 6 rooms and air conditioner, the house price is still reasonably low. However, It seems unreasonale for row 1274 as a 6-room house built in 2007 with a grade of "above average". Thus, I decide to keep row 429 and remove row 1274.

Now get my new dataset:

```
dtrain2 <- dtrain[-c(5, 1274),]
fm <- lm(formula = sqrt(price) ~ saleYear + gba + grade + ayb + bathrm +
           fireplaces + extwall + eyb + hf_bathrm + rooms + landarea +
           kitchens + saleYear:gba + saleYear:ayb + saleYear:bathrm +
           saleYear:eyb + ayb:eyb + fireplaces:eyb, data = dtrain2)
```

7.2 Outliers in predictors

```
n <- length(dtrain2$price)
p <- length(dtrain2) - 2
outliers <- c()
hii <- hatvalues(fm)
for (i in 1:n){
    if (hii[i] > 2*(p+1)/n){
        outliers <- c(outliers, i)
    }
}
outliers
```

```
##   [1]    1    2    3    4    5   10   11   12   13   26   31   33   39   45   92
##  [16]   93   96  104  111  115  117  118  120  121  124  126  130  132  137  138
##  [31]  140  142  143  146  147  150  152  156  158  163  164  165  169  172  173
##  [46]  178  180  186  190  191  193  194  195  197  199  200  201  205  206  212
##  [61]  216  219  222  226  233  234  236  242  243  250  251  256  259  260  261
##  [76]  266  283  289  295  297  302  303  310  312  318  322  323  324  325  326
##  [91]  336  338  347  351  357  386  387  399  400  413  421  422  429  436  441
## [106]  444  447  461  465  467  470  480  489  494  506  507  508  511  513  515
## [121]  516  534  535  537  547  548  554  556  564  570  575  579  580  582  598
## [136]  600  602  607  612  616  617  624  632  633  642  645  649  650  661  665
## [151]  666  673  678  686  694  701  702  703  705  706  709  710  712  713  714
## [166]  715  716  722  731  736  740  748  750  754  755  760  779  781  782  792
## [181]  794  798  800  802  813  814  819  832  837  844  857  867  884  885  888
## [196]  898  901  906  908  910  911  938  942  944  963  965  967  974  978  982
## [211]  984  987  992  997 1009 1010 1014 1020 1027 1038 1042 1047 1053 1059 1064
## [226] 1068 1080 1083 1099 1101 1105 1108 1109 1122 1124 1131 1138 1143 1149 1154
## [241] 1156 1167 1168 1173 1178 1181 1182 1185 1186 1190 1192 1199 1200 1203 1204
```

```
## [256] 1205 1209 1214 1219 1222 1224 1227 1230 1231 1237 1240 1243 1246 1248 1255
## [271] 1260 1263 1265 1266 1275 1277 1279 1280 1287 1296
```

By definition, $h_{ii} > 2*(p+1)/n$ represent significant outliers in x. Thus, above are all possible significant outliers in x.

Let's check some of these data:

```
dtrain2[c(965), c(1, 2, 3, 5, 7, 9, 12, 17, 18)]
```

```
##     bathrm hf_bathrm    heat rooms  ayb  eyb  price kitchens fireplaces
## 966      0         0 No Data     0 1941 1928 150300        0          0
```

I decide to remove row 966, because it misses too many values, such as bathrm, hf_bathrm, heat, bedrm, kitchens, and fireplaces. Only the price is known. Also, it probably has recording mistakes as well since the improvement of the house was made even before the house was built. This data will not be helpful to the prediction result.

```
dtrain2[c(142, 222, 715, 881, 1240, 1275), c(1, 5, 6, 13, 19)]
```

```
##      bathrm rooms bedrm  gba landarea
## 143       4    14     5 5129    15000
## 223       6    12     8 3726    10200
## 716       5    19     6 2040     6750
## 882       1     6     3  988      696
## 1241      1     5     2  535     2000
## 1277      3    13     5 1880    16098
```

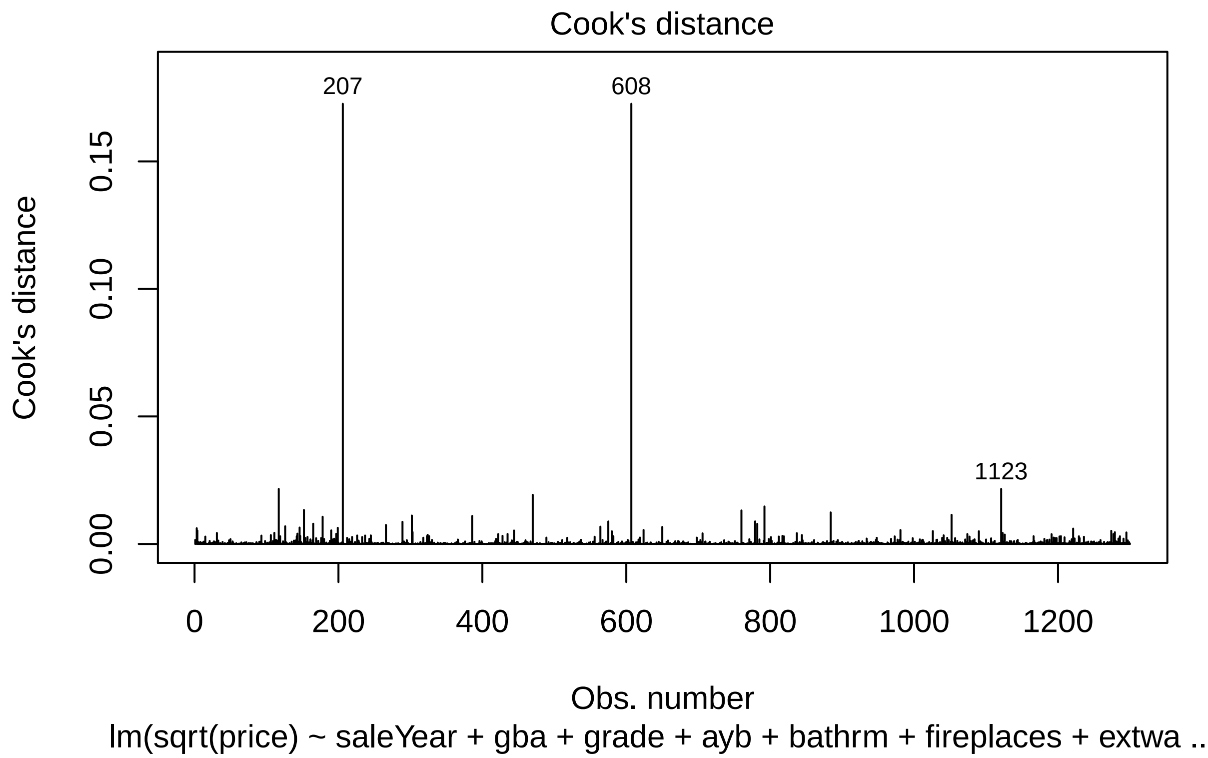
I decide to keep all these rows above. Although they are all "extreme" values in some aspect, they are reasonable. The larger landarea or gba corresponds to the house with more rooms. On the other hand, the smaller landarea or gba corresponds to the house with less rooms.

Now get my new dataset:

```
dtrain2 <- dtrain2[-c(965),]
fm <- lm(formula = sqrt(price) ~ saleYear + gba + grade + ayb + bathrm +
           fireplaces + extwall + eyb + hf_bathrm + rooms + landarea +
           kitchens + saleYear:gba + saleYear:ayb + saleYear:bathrm +
           saleYear:eyb + ayb:eyb + fireplaces:eyb, data = dtrain2)
```

7.3 Infulential cases

```r
plot(fm, which=4)
```




```r
qf(0.5, p+1, n-p-1)
```

```
## [1] 0.9656433
```

According to the Cook's Distance plot, we can see that none of observations have cook's distance greater than $F(0.5, p+1, n-p-1) = 0.9656433$. Thus, no influential points.

8. Summary

```r
summary(fm)
```

```
##
## Call:
## lm(formula = sqrt(price) ~ saleYear + gba + grade + ayb + bathrm +
##     fireplaces + extwall + eyb + hf_bathrm + rooms + landarea +
##     kitchens + saleYear:gba + saleYear:ayb + saleYear:bathrm +
##     saleYear:eyb + ayb:eyb + fireplaces:eyb, data = dtrain2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -254.00  -48.33    0.13   50.55  223.47
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.130e+04  9.521e+04   0.329 0.742440
## saleYear              -8.073e+01  4.895e+01  -1.649 0.099358 .
## gba                   -6.348e+00  1.485e+00  -4.274 2.07e-05 ***
## gradeAverage          -4.409e+01  4.964e+00  -8.883  < 2e-16 ***
## gradeFair Quality     -7.951e+01  2.095e+01  -3.796 0.000154 ***
## gradeGood Quality      5.413e+01  1.053e+01   5.139 3.20e-07 ***
## gradeSuperior          1.234e+02  8.990e+01   1.373 0.169981
## gradeVery Good         5.308e+01  3.831e+01   1.386 0.166144
## ayb                    1.992e+02  4.056e+01   4.912 1.02e-06 ***
## bathrm                -1.564e+03  8.966e+02  -1.744 0.081369 .
## fireplaces            -1.469e+03  5.067e+02  -2.899 0.003803 **
## extwallAluminum        1.542e+01  7.715e+01   0.200 0.841639
## extwallBrick Veneer    1.149e+01  7.954e+01   0.144 0.885169
## extwallBrick/Siding   -2.009e+01  7.717e+01  -0.260 0.794640
## extwallBrick/Stone    -3.460e+01  8.391e+01  -0.412 0.680171
## extwallBrick/Stucco   -4.330e+01  7.991e+01  -0.542 0.588002
## extwallCommon Brick   -4.611e+00  7.659e+01  -0.060 0.951999
## extwallConcrete       -7.064e+01  8.549e+01  -0.826 0.408818
## extwallConcrete Block  2.769e+01  1.072e+02   0.258 0.796227
## extwallFace Brick      4.198e+01  8.578e+01   0.489 0.624634
## extwallHardboard       5.447e+01  7.991e+01   0.682 0.495546
## extwallMetal Siding    1.060e+02  8.808e+01   1.203 0.229172
## extwallShingle         7.585e+00  7.713e+01   0.098 0.921676
## extwallStone          -2.257e+01  8.341e+01  -0.271 0.786775
## extwallStone Veneer    2.427e+01  8.380e+01   0.290 0.772154
## extwallStone/Siding   -1.119e+01  7.896e+01  -0.142 0.887303
## extwallStone/Stucco    6.143e+01  9.405e+01   0.653 0.513759
## extwallStucco         -3.952e+00  7.708e+01  -0.051 0.959117
## extwallStucco Block   -2.912e+00  9.336e+01  -0.031 0.975118
## extwallVinyl Siding   -1.223e+01  7.675e+01  -0.159 0.873450
## extwallWood Siding     1.623e+01  7.697e+01   0.211 0.833008
## eyb                   -1.559e+02  7.069e+01  -2.205 0.027661 *
## hf_bathrm              1.395e+01  4.019e+00   3.470 0.000538 ***
## rooms                  5.844e+00  1.888e+00   3.094 0.002015 **
## landarea               2.346e-03  1.133e-03   2.070 0.038612 *
```

```
## kitchens              1.241e+01  1.464e+01   0.848 0.396794
## saleYear:gba          3.182e-03  7.388e-04   4.308 1.78e-05 ***
## saleYear:ayb         -6.650e-02  1.997e-02  -3.330 0.000892 ***
## saleYear:bathrm       7.902e-01  4.460e-01   1.772 0.076664 .
## saleYear:eyb          1.113e-01  3.539e-02   3.144 0.001706 **
## ayb:eyb              -3.384e-02  6.031e-03  -5.610 2.49e-08 ***
## fireplaces:eyb        7.598e-01  2.576e-01   2.950 0.003239 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.91 on 1258 degrees of freedom
## Multiple R-squared:  0.8071, Adjusted R-squared:  0.8008
## F-statistic: 128.4 on 41 and 1258 DF,  p-value: < 2.2e-16
```

$\beta_0 = 31300$ is the expected house price when the a house first built in year 0 and having improvement made on year 0 has the sale year of 0, the gross building area of 0, the number of bathrooms of 0 , the number of fireplaces of 0, the number of half-bathroom sof 0, the number of fooms of 0, the land area of 0, and the number of kitchens of 0 without presenting a grade, an exterior wall material,

$\beta_1 = -80.73$ is the expected decrease in house price with sale year increased by one unit while the house first built in year 0 has a gross building area of 0 and the number of bathrooms of 0 and other predictors hold constant.

$\beta_2 = -6.348$ is the expected decrease in house price with gross building area increased by one unit while the house has a sale year of 0 and other predictors hold constant.

$\beta_3 = -44.09$ is the expected decrease in house price with the grade changing from "Above Average" to "Average" while other predictors hold constant.

$\beta_4 = -79.51$ is the expected decrease in house price with the grade changing from "Average" to "Fair Quality" while other predictors hold constant.

$\beta_5 = 54.13$ is the expected increase in house price with the grade changing from "Fair Quality" to "Good Quality" while other predictors hold constant.

$\beta_6 = 123.4$ is the expected increase in house price with the grade changing from "Good Quality" to "Superior" while other predictors hold constant.

$\beta_7 = 53.08$ is the expected increase in house price with the grade changing from "Superior" to "Very Good" while other predictors hold constant.

$\beta_8 = 199.2$ is the expected increase in house price with the year when the house first built increased by one unit while the house has a sale year of 0 and the improvement year of 0 and other predictors hold constant.

$\beta_9 = -1564$ is the expected decrease in house price with the number of bathrooms in the house increased by one unit while the house has a sale year of 0 and the improvement year of 0 and other predictors hold constant.

$\beta_{10} = -1469$ is the expected decrease in house price with the number of fireplaces in the house increased by one unit while the improvement year is 0 and other predictors hold constant.

$\beta_{11} = 15.42$ is the expected increase in house price with the exterior wall material changing from "Adobe" to "Aluminum" while other predictors hold constant.

$\beta_{12} = 11.49$ is the expected increase in house price with the exterior wall material changing from "Aluminum" to "Brick Veneer" while other predictors hold constant.

$\beta_{13} = -20.09$ is the expected decrease in house price with the exterior wall material changing from "Brick Veneer" to "Brick/Siding" while other predictors hold constant.

$\beta_{14} = -34.6$ is the expected decrease in house price with the exterior wall material changing from "Brick/Siding" to "Brick/Stone" while other predictors hold constant.

$\beta_{15} = -43.3$ is the expected decrease in house price with the exterior wall material changing from "Brick/Stone" to "Brick/Stucco" while other predictors hold constant.

$\beta_{16} = -4.611$ is the expected decrease in house price with the exterior wall material changing from "Brick/Stucco" to "Common Brick" while other predictors hold constant.

$\beta_{17} = -70.64$ is the expected decrease in house price with the exterior wall material changing from "Common Brick" to "Concrete" while other predictors hold constant.

$\beta_{18} = 27.69$ is the expected increase in house price with the exterior wall material changing from "Concrete" to "Concrete Block" while other predictors hold constant.

$\beta_{19} = 41.98$ is the expected increase in house price with the exterior wall material changing from "Concrete Block" to "Face Brick" while other predictors hold constant.

$\beta_{20} = 54.47$ is the expected increase in house price with the exterior wall material changing from "Face Brick" to "Hardboard" while other predictors hold constant.

$\beta_{21} = 106$ is the expected increase in house price with the exterior wall material changing from "Hardboard" to "Metal Siding" while other predictors hold constant.

$\beta_{22} = 7.585$ is the expected increase in house price with the exterior wall material changing from "Metal Siding" to "Shingle" while other predictors hold constant.

$\beta_{23} = -22.57$ is the expected decrease in house price with the exterior wall material changing from "Shingle" to "Stone" while other predictors hold constant.

$\beta_{24} = 24.27$ is the expected increase in house price with the exterior wall material changing from "Stone" to "Stone Veneer" while other predictors hold constant.

$\beta_{25} = -11.19$ is the expected decrease in house price with the exterior wall material changing from "Stone Veneer" to "Stone/Siding" while other predictors hold constant.

$\beta_{26} = 61.43$ is the expected increase in house price with the exterior wall material changing from "Stone/Siding" to "Stone/Stucco" while other predictors hold constant.

$\beta_{27} = -3.952$ is the expected decrease in house price with the exterior wall material changing from "Stone/Stucco" to "Stucco" while other predictors hold constant.

$\beta_{28} = -2.912$ is the expected decrease in house price with the exterior wall material changing from "Stucco" to "Stucco Block" while other predictors hold constant.

$\beta_{29} = -12.23$ is the expected decrease in house price with the exterior wall material changing from "Stucco Block" to "Vinvl Siding" while other predictors hold constant.

$\beta_{30} = 16.23$ is the expected increase in house price with the exterior wall material changing from "Vinvl Siding" to "Wood Siding" while other predictors hold constant.

$\beta_{31} = -155.9$ is the expected decrease in house price with the improvement year of the house increased by one unit while the house first built in year 0 has a sale year of 0 and the number of fireplaces of 0 and other predictors hold constant.

$\beta_{32} = 13.95$ is the expected increase in house price with the number of half-bathrooms increased by one unit while other predictors hold constant.

$\beta_{33} = 5.844$ is the expected increase in house price with the number of rooms increased by one unit while other predictors hold constant.

$\beta_{34} = 0.002346$ is the expected increase in house price with the land area increased by one unit while other predictors hold constant.

$\beta_{35} = 12.41$ is the expected increase in house price with the number of kitchens increased by one unit while other predictors hold constant.

$\beta_{36} * saleYear + \beta_2 = 0.003182 * saleYear - 6.3481$ is the expected decrease (increase) in house price with the gross building area increased by one unit at different levels of sale year while other predictors hold constant.

$\beta_{37} * saleYear + \beta_{40} * eyb + \beta_8 = 0.003182 * saleYear - 0.03384 * eyb + 199.2$ is the expected decrease (increase) in house price with the year when the house first built increased by one unit at different levels of sale year and improvement year while other predictors hold constant.

$\beta_{38} * saleYear + \beta_9 = 0.7902 * saleYear - 1564$ is the expected decrease (increase) in house price with the nubmer of bathrooms increased by one unit at different levels of sale year while other predictors hold constant.

$\beta_{39} * saleYear + \beta_{40} * ayb + \beta_{41} * fireplaces + \beta_{31} = 0.003182 * saleYear - 0.03384 * ayb + 0.7958 * fireplaces - 155.9$ is the expected decrease (increase) in house price with the improvement year increased by one unit at different levels of sale year, building year, and the number of fireplaces while other predictors hold constant.

$\beta_{40} = -0.03384$ has been discussed above when discuss $\beta 37$ and $\beta 39$.

$\beta_{41} * eyb + \beta_9 = 0.7958 * eyb - 1469$ is the expected decrease (increase) in house price with the nubmer of fireplaces increased by one unit at different levels of improvement year while other predictors hold constant.

==

In summary, house prices are affected by a large amount of factors. The dataset we have just provides us a few possible factors that can probably influence the house prices. Besides these, the infrastructure such as hospitals, schools, and transportation system, the surrounding environment, and etc. can probably influence the house prices as well.

In the interest of this dataset, I conclude that the sale year, the gross building area, the grade, the exterior wall material, the year when the building was first built, the year when the house was recently improved, theland area, and the number of bathrooms, fireplaces, half-bathrooms, rooms, and kitchens, are the factors that can affect the house prices. It is not hard to understand.

The real estate market has its house prices fluctuation due to many factors, such as policies or inflation, which is not our research direction at this point. But this is a fact that, house prices can be relatively high at some period time compared to other periods. Therefore, the year when the house is sold will play an important role in the house prices prediction.

Secondly, people are interested in when the house is built, when the house is recently improved, and the grade of the house. Because the house condition will be a main factor that can affect the house price. An old house without recent improvement can indicate a lot of problems.

Thirdly, the exterior wall material is also a factor that people may pay attention to because this indicates the safety of the house. Especially in some cities that have typhoon or earthquakes, the degree of the stability of the house is important.

Finally, the number of rooms, bathrooms, half-bathrooms, kitchens, and fireplaces will be taken into consideration. Apparently, this is the first plan that a family or a person will have to make. This is the basic need when people plan to buy a house.

Therefore, my final model includes all factors mentioned above and my prediction is based on the model that I construct.