Data Analysis and Linear Models Homework 8

### 1.Introduction

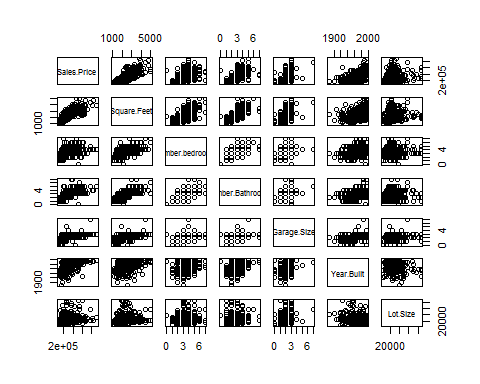
This project is about that using the data of residential sales that occurred during the year 2002 were available from a city in the midwest to bulid a model to predict the sales price.These demographic variable include sales price, style, finished square feet, number of bedrooms, whether the house has a pool, lot size, year built, whether air conditioning is installed and whether or not the lot is adjacent to a highway.

## 'data.frame': 522 obs. of 13 variables:  
## $ Id.Number : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Sales.Price : int 360000 340000 250000 205500 275500 248000 229900 150000 195000 160000 ...  
## $ Square.Feet : int 3032 2058 1780 1638 2196 1966 2216 1597 1622 1976 ...  
## $ Number.bedrooms : int 4 4 4 4 4 4 3 2 3 3 ...  
## $ Number.Bathrooms : int 4 2 3 2 3 3 2 1 2 3 ...  
## $ Air.Condition : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 1 ...  
## $ Garage.Size : int 2 2 2 2 2 5 2 1 2 1 ...  
## $ Pool : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...  
## $ Year.Built : int 1972 1976 1980 1963 1968 1972 1972 1955 1975 1918 ...  
## $ Quality : Factor w/ 3 levels "1","2","3": 2 2 2 2 2 2 2 2 3 3 ...  
## $ Style : Factor w/ 10 levels "1","2","3","4",..: 1 1 1 1 7 1 7 1 1 1 ...  
## $ Lot.Size : int 22221 22912 21345 17342 21786 18902 18639 22112 14321 32358 ...  
## $ Adjacent.to.Highway: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

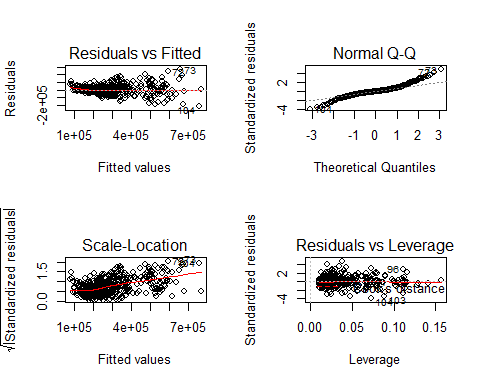
#### 2.Exploratory data analysis

#### (a)

The R function *str* can tell us that ,only *Sales.Price,Square.Feet,Number.bedrooms,Number.Batchrooms,Garage.Size,Lot.Size* are quantitative variables.

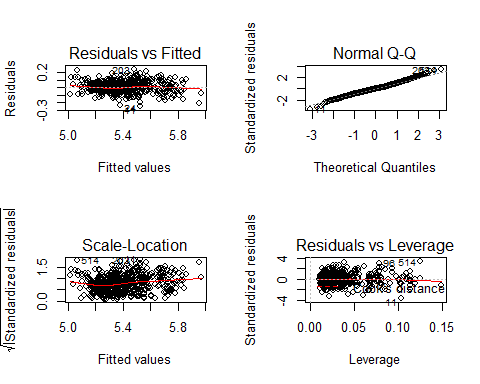
 The scatterplot matrix show us that the response variable *Sales Price* almost have a linear relationship with all other predictive variables,and the relationship across predictors are weak.

#### (b)

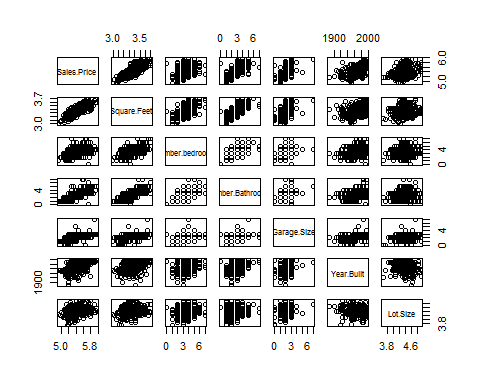


Residual vs Fitted values in plot(1) has a trend of funnel shape,which indicates error terms in model has a nonconstant variables.This is not To meet the equal variance conditions;  
The QQ-plot of residual in plot(2) show us that the error term in model is not to meet the normality condition;  
And the plot(3),(4) tell us that there are outliers in raw dataset.  
So model assumptions are violated.

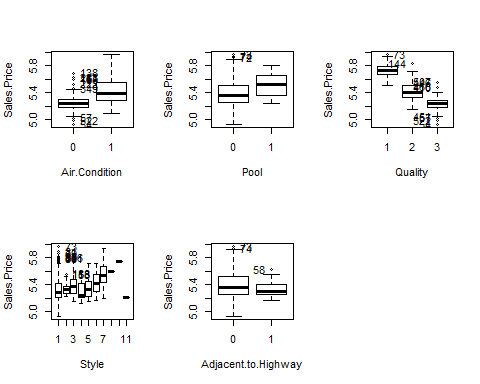
Make a transformation in predictive variables *Square.Feet,Lot.Size* and response variable *Sales.Price* with **log10**

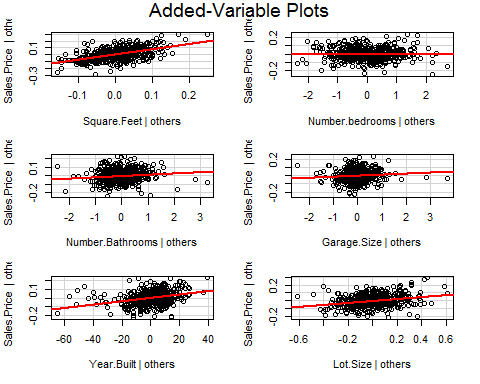
 Afther the data transformation ,we bulid the model again.According to the plot of model.These graph show us that model assumations are not violated.

#### (c)

* i  The linear relationship perform more stronger in scatterplot of response variable and predictive variables.
* ii *Air.Condition,Pool,Quality,Style,Adjacent.to.Highway* are qualitative variables

## Loading required package: car



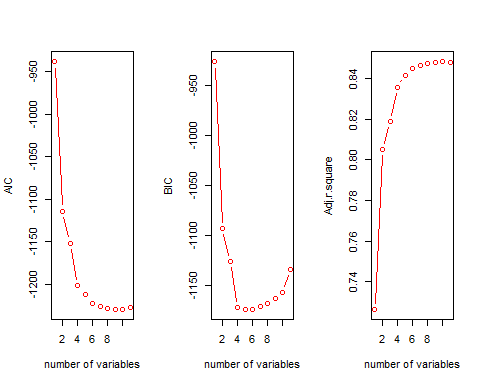
* iii 
* iv

Based on i and iii,the plot show us that *Square.Feet*,*Number.Bathrooms*,*Year.Built* and *Lot.size* has a strong linear correlation with the dependent variable respectively. In ii,boxplot show that *Air.condition*,*Pool*,*Quality* has more bigger difference against response compared to other qualitative variables. So,I decide use *Square.Feet*,*Number.Bathrooms*,*Year.Built*,*Lot.size*,*Air.condition*,*Pool*,*Quality* as predictors in model.The functional form is :

* *Sales.Price*~*Square.Feet*+*Number.Bathrooms*+*Year.Built*+*Air.condition*+*Pool*+*Quality*

### 3.Model building

#### (a)



According to the plot above,when numberof variables increase,the *AIC* decrease,and when number is 9,the *AIC* is not changed again. *BIC* is decremented before 4, increasing after 4,and is minimum at 4.*Adj.r.square* has been increasing,and stop at number is 9 and is not changed again.These tell us that the criteria optimized for models of different sizes.

#### (b)

## Five model variable subsets in AIC:....

## [[1]]  
## [1] "Square.Feet" "Number.Bathrooms" "Garage.Size"   
## [4] "Pool" "Year.Built" "Quality"   
## [7] "Style" "Lot.Size" "Adjacent.to.Highway"  
##   
## [[2]]  
## [1] "Square.Feet" "Number.Bathrooms" "Air.Condition"   
## [4] "Garage.Size" "Pool" "Year.Built"   
## [7] "Quality" "Style" "Lot.Size"   
## [10] "Adjacent.to.Highway"  
##   
## [[3]]  
## [1] "Square.Feet" "Number.Bathrooms" "Garage.Size"   
## [4] "Pool" "Year.Built" "Quality"   
## [7] "Style" "Lot.Size"   
##   
## [[4]]  
## [1] "Square.Feet" "Number.bedrooms" "Number.Bathrooms"   
## [4] "Air.Condition" "Garage.Size" "Pool"   
## [7] "Year.Built" "Quality" "Style"   
## [10] "Lot.Size" "Adjacent.to.Highway"  
##   
## [[5]]  
## [1] "Square.Feet" "Number.Bathrooms" "Pool"   
## [4] "Year.Built" "Quality" "Style"   
## [7] "Lot.Size"

## Five model variable subsets in BIC:....

## [[1]]  
## [1] "Square.Feet" "Number.Bathrooms" "Year.Built"   
## [4] "Quality" "Lot.Size"   
##   
## [[2]]  
## [1] "Square.Feet" "Number.Bathrooms" "Garage.Size"   
## [4] "Year.Built" "Quality" "Lot.Size"   
##   
## [[3]]  
## [1] "Square.Feet" "Year.Built" "Quality" "Lot.Size"   
##   
## [[4]]  
## [1] "Square.Feet" "Number.Bathrooms" "Garage.Size"   
## [4] "Pool" "Year.Built" "Quality"   
## [7] "Lot.Size"   
##   
## [[5]]  
## [1] "Square.Feet" "Number.Bathrooms" "Air.Condition"   
## [4] "Garage.Size" "Pool" "Year.Built"   
## [7] "Quality" "Lot.Size"

## Five model variable subsets in Adj.r.square:....

## [[1]]  
## [1] "Square.Feet" "Number.Bathrooms" "Air.Condition"   
## [4] "Garage.Size" "Pool" "Year.Built"   
## [7] "Quality" "Style" "Lot.Size"   
## [10] "Adjacent.to.Highway"  
##   
## [[2]]  
## [1] "Square.Feet" "Number.Bathrooms" "Garage.Size"   
## [4] "Pool" "Year.Built" "Quality"   
## [7] "Style" "Lot.Size" "Adjacent.to.Highway"  
##   
## [[3]]  
## [1] "Square.Feet" "Number.bedrooms" "Number.Bathrooms"   
## [4] "Air.Condition" "Garage.Size" "Pool"   
## [7] "Year.Built" "Quality" "Style"   
## [10] "Lot.Size" "Adjacent.to.Highway"  
##   
## [[4]]  
## [1] "Square.Feet" "Number.Bathrooms" "Garage.Size"   
## [4] "Pool" "Year.Built" "Quality"   
## [7] "Style" "Lot.Size"   
##   
## [[5]]  
## [1] "Square.Feet" "Number.Bathrooms" "Pool"   
## [4] "Year.Built" "Quality" "Style"   
## [7] "Lot.Size"

According to the result ,there models are generally the similiar.variable *Square.Feet* is consistently present in all the “best” sub-models.

#### (c)

##   
## Call:  
## lm(formula = Sales.Price ~ Square.Feet + Number.Bathrooms + Garage.Size +   
## Pool + Year.Built + Quality + Style + Lot.Size + Adjacent.to.Highway,   
## data = sales.new)  
##   
## Coefficients:  
## (Intercept) Square.Feet Number.Bathrooms   
## -1.441120 0.786642 0.016416   
## Garage.Size Pool1 Year.Built   
## 0.013251 0.028017 0.001907   
## Quality2 Quality3 Style2   
## -0.124295 -0.156612 -0.029761   
## Style3 Style4 Style5   
## -0.008087 0.031273 -0.029378   
## Style6 Style7 Style9   
## -0.004407 -0.048421 -0.053670   
## Style10 Style11 Lot.Size   
## -0.110174 -0.166064 0.122398   
## Adjacent.to.Highway1   
## -0.040835

We can get the stepwise model function from the above result.This model is :

* *Sales.Price*~*Square.Feet* + *Number.Bathrooms* + *Garage.Size* + *Pool* + *Year.Built* + *Quality* + *Style* + *Lot.Size* + *Adjacent.to.Highway*

This model is identified by this procedure the same as any of the one identified in part (b).

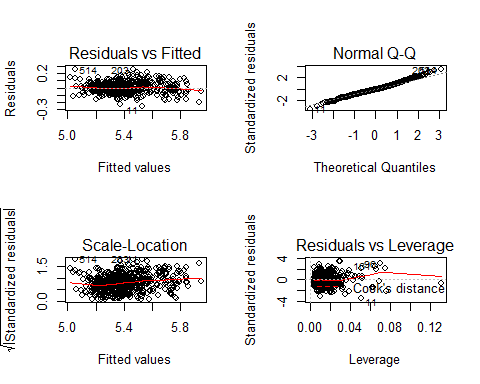
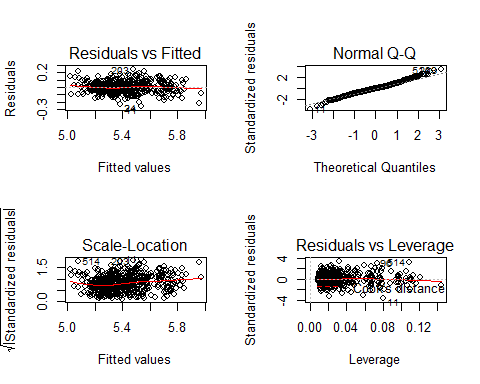
#### (d)

I pick the two model below:

* 1. model.3: *Sales.Price*~*Square.Feet* + *Number.Bathrooms* + *Garage.Size* + *Pool* + *Year.Built* + *Quality* + *Style* + *Lot.Size* + *Adjacent.to.Highway*
  2. model.4: *Sales.Price*~*Square.Feet* + *Number.Bathrooms* + *Garage.Size* +*Year.Built* + *Quality* + *Lot.Size*

### 4.Model Diagnostic

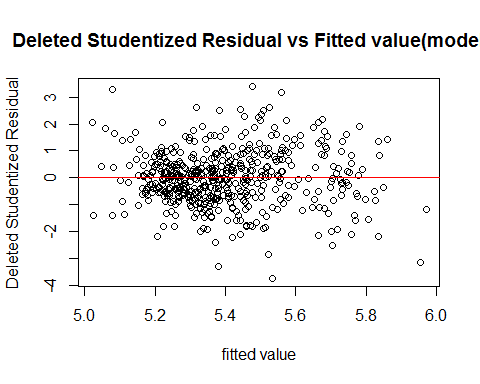
#### (a)



The “traditional” residual plots of the picked model show that,the errom term in two model both meet the cosntannt variance and k normality assumations.

#### (b)

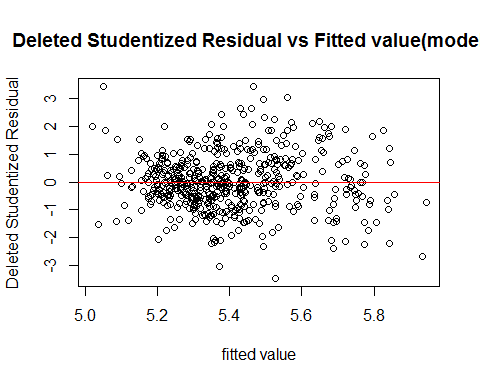
## Loading required package: MASS



## Use the t-test to compare difference between semi-studentized  
## residuals and deleted studentized residual...

##   
## Welch Two Sample t-test  
##   
## data: semi.student.res.3 and stud.re.del.3  
## t = -0.00058339, df = 526.91, p-value = 0.9995  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.159514 1.158825  
## sample estimates:  
## mean of x mean of y   
## 1.772608e-16 3.442411e-04

The *p-value* of the test is 0.9995347,the p-value is less than 0.05,which inicates that there is no difference between them.



## Use the t-test to compare difference between semi-studentized  
## residuals and deleted studentized residual...

##   
## Welch Two Sample t-test  
##   
## data: semi.student.res.4 and stud.re.del.4  
## t = -0.0010062, df = 527.06, p-value = 0.9992  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.141135 1.139967  
## sample estimates:  
## mean of x mean of y   
## -2.350251e-16 5.841749e-04

The *p-value* of the test is 0.9991976,the p-value is less than 0.05,which inicates that there is no difference between them.

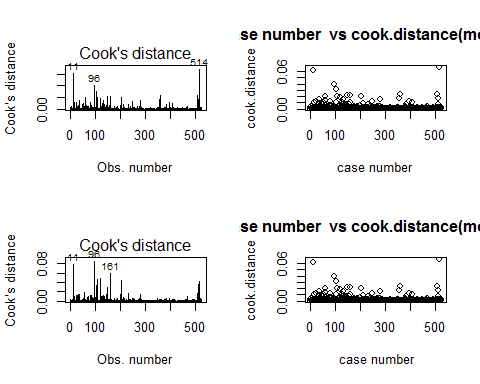
#### (c)

##   
## Call:  
## lm(formula = Sales.Price ~ Square.Feet + Number.Bathrooms + Garage.Size +   
## Pool + Year.Built + Quality + Style + Lot.Size + Adjacent.to.Highway,   
## data = sales.new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.254792 -0.045087 -0.000605 0.040855 0.239287   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.4411199 0.5573499 -2.586 0.0100 \*\*   
## Square.Feet 0.7866422 0.0541325 14.532 < 2e-16 \*\*\*  
## Number.Bathrooms 0.0164158 0.0052027 3.155 0.0017 \*\*   
## Garage.Size 0.0132513 0.0063104 2.100 0.0362 \*   
## Pool1 0.0280167 0.0131992 2.123 0.0343 \*   
## Year.Built 0.0019065 0.0002587 7.370 7.07e-13 \*\*\*  
## Quality2 -0.1242946 0.0127578 -9.743 < 2e-16 \*\*\*  
## Quality3 -0.1566121 0.0177976 -8.800 < 2e-16 \*\*\*  
## Style2 -0.0297611 0.0116109 -2.563 0.0107 \*   
## Style3 -0.0080870 0.0110803 -0.730 0.4658   
## Style4 0.0312727 0.0230773 1.355 0.1760   
## Style5 -0.0293776 0.0189529 -1.550 0.1218   
## Style6 -0.0044071 0.0192089 -0.229 0.8186   
## Style7 -0.0484206 0.0109662 -4.415 1.23e-05 \*\*\*  
## Style9 -0.0536703 0.0740547 -0.725 0.4689   
## Style10 -0.1101744 0.0751305 -1.466 0.1432   
## Style11 -0.1660638 0.0736120 -2.256 0.0245 \*   
## Lot.Size 0.1223978 0.0195487 6.261 8.20e-10 \*\*\*  
## Adjacent.to.Highway1 -0.0408352 0.0225668 -1.810 0.0710 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.07309 on 503 degrees of freedom  
## Multiple R-squared: 0.8533, Adjusted R-squared: 0.848   
## F-statistic: 162.5 on 18 and 503 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = Sales.Price ~ Square.Feet + Number.Bathrooms + Garage.Size +   
## Year.Built + Quality + Lot.Size, data = sales.new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.24899 -0.04396 -0.00373 0.04501 0.25009   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.8319756 0.5534275 -1.503 0.13337   
## Square.Feet 0.7068398 0.0458182 15.427 < 2e-16 \*\*\*  
## Number.Bathrooms 0.0147707 0.0052455 2.816 0.00505 \*\*   
## Garage.Size 0.0154650 0.0064375 2.402 0.01664 \*   
## Year.Built 0.0017084 0.0002525 6.766 3.60e-11 \*\*\*  
## Quality2 -0.1350925 0.0124792 -10.825 < 2e-16 \*\*\*  
## Quality3 -0.1649726 0.0175696 -9.390 < 2e-16 \*\*\*  
## Lot.Size 0.1311253 0.0195232 6.716 4.95e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0751 on 514 degrees of freedom  
## Multiple R-squared: 0.8417, Adjusted R-squared: 0.8395   
## F-statistic: 390.4 on 7 and 514 DF, p-value: < 2.2e-16

In model *Sales.Price*~*Square.Feet* + *Number.Bathrooms* + *Garage.Size* + *Pool* + *Year.Built* + *Quality* + *Style* + *Lot.Size* + *Adjacent.to.Highway*,not all predictors marginally are significant,but in model *Sales.Price*~*Square.Feet* + *Number.Bathrooms* + *Garage.Size* +*Year.Built* + *Quality* + *Lot.Size* ,all the predictors marginally are significant. The first model R.Square is *0.8479999*,and in second model R.Square is *0.8395409*.According to the result ,we know that the R.Square are similiar for these models.

#### (d)

 According to the plot,the cook distance of model.3 show us that ,the observation **11,96,514** are outliers;the cook distance of model.4 show us that ,the observation **11,96,161** are outliers.So we remove the observation:11,96,161,514.

#### (e)

### 5.Model Validation

#### (a)

According to the result we get from the above ,the mean squared prediction error (*MSPR*) is 0.0054524,the estimated *MSE* is 0.0050845 in **model.3**;the mean squared prediction error (*MSPR*) is 0.0057448,the estimated *MSE* is 0.005382 in **model.4**. And the absolute error between *MSPR* and *MSE* in **model.3** is3.679702210^{-4};the absolute error between *MSPR* and *MSE* in **model.4** is 3.627734110^{-4}.And we can known that the comparisonis is smaller in both model,which tell us that these two model both have good goodness of fit,that is the two model's adequacy are both good.

#### (b)

## The confidence intervals of estimated coefficient with origin dataset:

## 2.5 % 97.5 %  
## (Intercept) -2.681012276 -0.481083908  
## Square.Feet 0.671818753 0.881164088  
## Number.Bathrooms 0.009427587 0.029602365  
## Garage.Size 0.003930238 0.028755665  
## Pool1 0.010809509 0.062090359  
## Year.Built 0.001473569 0.002491338  
## Quality2 -0.140760359 -0.090956810  
## Quality3 -0.174113114 -0.104031842  
## Style2 -0.051124569 -0.006137240  
## Style3 -0.026660164 0.015961526  
## Style4 -0.008672378 0.079858074  
## Style5 -0.062067650 0.010690181  
## Style6 -0.040255376 0.033438109  
## Style7 -0.065098693 -0.022792353  
## Style9 -0.184234757 0.099810268  
## Style10 -0.257920361 0.030228158  
## Style11 -0.302045213 -0.019793791  
## Lot.Size 0.083589231 0.159677162  
## Adjacent.to.Highway1 -0.081860116 0.004685170

## The confidence intervals of estimated coefficient with validattion dataset:

## 2.5 % 97.5 %  
## (Intercept) -5.023389393 0.672712066  
## Square.Feet 0.366581049 0.930817169  
## Number.Bathrooms 0.015589192 0.069507200  
## Garage.Size -0.021505253 0.037872649  
## Pool1 0.002210887 0.137749418  
## Year.Built 0.001110153 0.003767863  
## Quality2 -0.147060013 -0.021357383  
## Quality3 -0.161822344 0.014965859  
## Style2 -0.082366792 0.057500740  
## Style3 -0.055175972 0.038035892  
## Style4 -0.027903403 0.143545851  
## Style5 -0.012341184 0.172945086  
## Style6 -0.086066156 0.072081723  
## Style7 -0.094251504 0.005578519  
## Style10 -0.296741683 0.050690875  
## Lot.Size 0.032186652 0.227565973  
## Adjacent.to.Highway1 -0.216229674 0.097118256

The *MSE* in origin *model.3* and *model.3* with validation dataset are **0.0050845** and**0.0056418**,and the R.Square are **0.8545648** and **0.8287445 respectively.We can see that the absolute error of *MSE* is 5.573715310^{-4}**,the absolute error of *R.Square* is **0.0258203**.And the absolute errors of confidence intervals of estimated coefficient in both dataset are small tooOf course,the absolute error of them are small,which indicate that there is litte difference in model built with origin dataset and validation dataset.

## The confidence intervals of estimated coefficient with origin dataset:

## 2.5 % 97.5 %  
## (Intercept) -2.033549868 0.148645926  
## Square.Feet 0.616319430 0.793676086  
## Number.Bathrooms 0.008138316 0.028570768  
## Garage.Size 0.006376675 0.031687440  
## Year.Built 0.001265912 0.002259692  
## Quality2 -0.150391175 -0.101557197  
## Quality3 -0.181764505 -0.112382658  
## Lot.Size 0.088903652 0.164907431

## The confidence intervals of estimated coefficient with validattion dataset:

## 2.5 % 97.5 %  
## (Intercept) -2.9301979306 2.155297345  
## Square.Feet 0.3330822233 0.824379785  
## Number.Bathrooms 0.0111893371 0.065004973  
## Garage.Size -0.0108339116 0.047270007  
## Year.Built 0.0005426365 0.002822541  
## Quality2 -0.1540367117 -0.033466654  
## Quality3 -0.1841178107 -0.010797993  
## Lot.Size 0.0248170301 0.202231183

The *MSE* in origin *model.4* and *model.4* with validation dataset are **0.005382** and **0.0058835**,and the R.Square are **0.8460543** and **0.8214091** respectively.We can see that the absolute error of *MSE* is **5.014994910^{-4}**,the absolute error of *R.Square* is **0.0246452**.And the absolute errors of confidence intervals of estimated coefficient in both dataset are small too.Of course,the absolute error of them are small,which indicate that there is litte difference in model built with origin dataset and validation dataset.

### 6.Conclusion

I pick *model.4* to present to the city tax assessor.In validation part we know that ,the absolute error between origin dataset and validation dataset is small,and the the R.Square is **0.8460543** in origin model.And the change of confidence intervals of estimated coefficient are smaller than *model.3*.In a word,*model.4* is more robustand has a goodness of fit.

##   
## Call:  
## lm(formula = Sales.Price ~ Square.Feet + Number.Bathrooms + Garage.Size +   
## Year.Built + Quality + Lot.Size, data = sales.new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.221022 -0.043316 -0.004856 0.043411 0.253070   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.9424520 0.5553717 -1.697 0.090311 .   
## Square.Feet 0.7049978 0.0451375 15.619 < 2e-16 \*\*\*  
## Number.Bathrooms 0.0183545 0.0052001 3.530 0.000454 \*\*\*  
## Garage.Size 0.0190321 0.0064416 2.955 0.003276 \*\*   
## Year.Built 0.0017628 0.0002529 6.970 9.85e-12 \*\*\*  
## Quality2 -0.1259742 0.0124283 -10.136 < 2e-16 \*\*\*  
## Quality3 -0.1470736 0.0176578 -8.329 7.56e-16 \*\*\*  
## Lot.Size 0.1269055 0.0193431 6.561 1.32e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.07336 on 510 degrees of freedom  
## Multiple R-squared: 0.8481, Adjusted R-squared: 0.8461   
## F-statistic: 406.9 on 7 and 510 DF, p-value: < 2.2e-16

According to the summary fo model,we will interpret some of the regression coefficients.The variable *Square.Feet* represent the finished area of residence,the estmated coefficient is *0.7049978*,which indicates that has a positive effect on *sales price* and when log10 *Square.Feet* add one unit,the log10 *sales price* will add about 0.7 unit if fixed others.And *Number.Bathrooms*,*Garage.Size* have simialiar positive effect on response,if fixed others,when there two add one unit,*sales price* will add about 0.2 unit respectively.And *Lot.Size* also has positive effect and contributes about 0.13 unit to response.But *Quality* tell us that,medium and low quality will habe negative effect on *sales price*,and they contribute about -0.13 and -0.15 unit to response respectively.According to common sense,above explanation are reasonable and consistented with reality.These variables are main effective factors on sales price in reality.And I think *Square.Feet* is the most importanct in determining sales price and it's impact is the biggest.