House Features and Residential Home Sales Prices

Candace McKeag
PSTAT 126

Introduction

The Real Estate Sales data set consists of 12 variables, all of which describe various characteristics of a house and its surrounding property. Using the data on 521 home sales during the year 2002, we wish to predict the sales price in dollars of a residence using the characteristics in the data set. Our interest is in investigating if certain predictors are truly significant in predicting the sales price using various methods, and in the accuracy of our model. We have found that the sales price of a residence depends on many different characteristics of the home and its surrounding property, including garage size, number of bedrooms, and more. The extent to which some of these characteristics affect the sales price also depends on other associated characteristics. We will demonstrate what criteria we have produced our model based upon, how we fine-tuned the model, and how we used the model to answer our questions of interest.

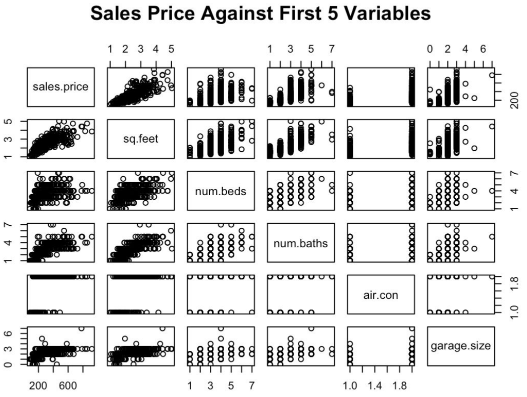
Questions of Interest

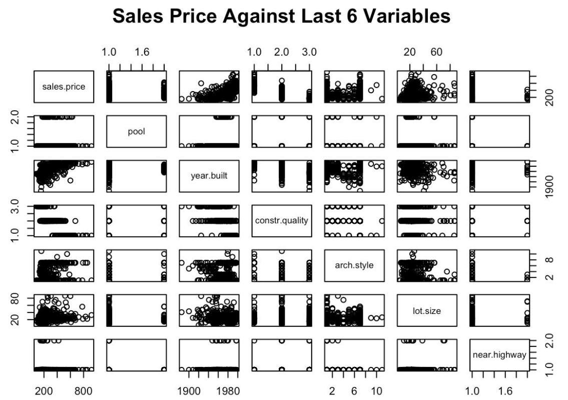
1. *Does the regression model contain at least one predictor useful in predicting the sales price?*
    a. To answer this question, we will test if all slope parameters equal zero by conducting a overall F test.
2. *If any, what effect does adjacency to highway have on the predicted sales price, holding all other factors constant?*
    a. Because highway adjacency is a binary indicator predictor, the extent to which it affects sales price depends on the estimated coefficient of the variable. To find this value, we will need to conduct a t-test for slope parameter on the highway predictor.
3. *Using the adjusted R^2 as the criterion, which predictors would be used in the "best" model? Why?*
    a. To answer this question, we will need to find the adjusted R^2 values for each subset of predictors, and for each quantity of parameters in a model, and conclude what model would be best at predicting sales price.
4. *Using the predictors found previously, how much variance in our response (sales price) can be explained by variance in our predictors?*
    a. To answer this question, we will need to find the R^2 value of the best model.
5. *Are there any outliers that are influential?*
    a. To answer this question, we will search for outliers using the externally studentized residuals method, and we will also use Cook's Distance Measure to find any potentially influential points. We will then compare the model with and without that data point to see if it has significantly affected any part of the results.

Building the Model:

1. *<u>Creating scatterplot matrices</u>*: Scatterplot matrices allow us to consider the relationship between the response and each of the predictors, and also how the predictors are related

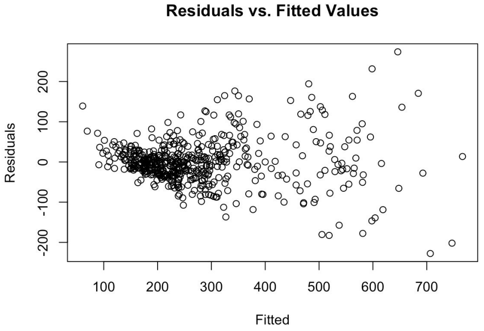
among each other. We can see that there are some obvious linear relationships between the candidate predictors and the response. There are also linear relationships between some of the candidate predictors, so adding interaction terms may be necessary.



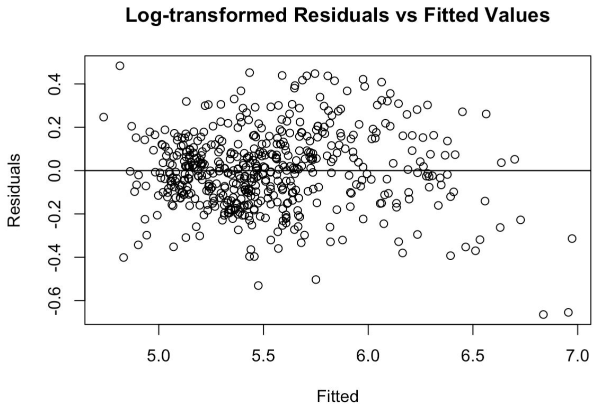


There seems to be a substantial linear relationship between sales.price and sq.feet, num.beds, num.baths, garage.size, year.built, constr.quality, and lot.size. There also seems to be significant interaction between sq.feet and num.beds, sq.feet and garage.size, and sq.feet and num.baths. However, we need to check the p-values before we decide which predictors/interaction terms are truly significant.

2. *Model Selection*: The first step we can take in creating our model is deciding what size model is the best for our data. We used adjusted R² as criteria by choosing the model size with the highest adjusted R² for the best model accuracy. Using this as criteria helps because it does not necessarily increase as more predictors are added--only if those predictors are significant. We found that the model with 8 predictors is the best.
3. *Best Subsets Regression Procedure:* To choose which predictors to use in the model, we followed the procedure of best subsets regression with adjusted R² as criteria. The model we chose is the subset of predictors that is the best at fitting the data. We found that the best model includes *square feet*, *construction quality*, *architectural style*, *year built*, *lot size*, *garage size*, *number of bedrooms*, and *highway adjacency* in that order.
4. *Residual Analysis*: We check the residuals vs fitted plot to see if the residuals are linear, independent, and have equal variance.

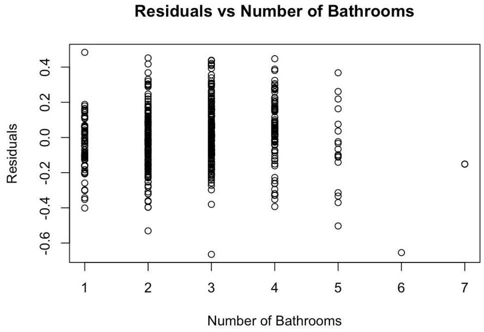


This residuals vs fitted plot clearly exhibits nonconstant variance. There is a fanning pattern from left to right. We need to log-transform the Y values to fix it.



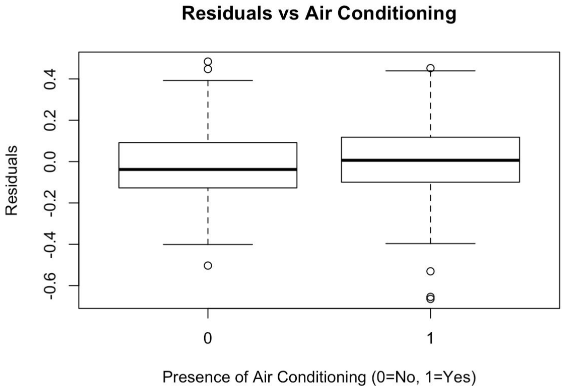
The new log-transformed residuals vs fitted plot looks much better than the previous. The closer together points have been spread out, and the spread out points have been brought closer together. The residuals bounce around randomly around the 0 line and roughly form a horizontal band. We should also plot these residuals against the predictors that have been omitted to ensure that there is no variability left unexplained by the variables.

-Residuals vs Number of Bathrooms:



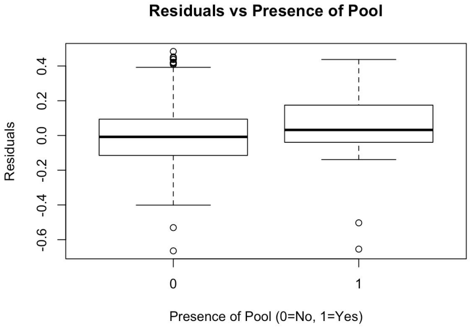
There does not seem to be any sign of a linear relationship between the number of bathrooms and the residuals of the model.

-Residuals vs Air Conditioning:



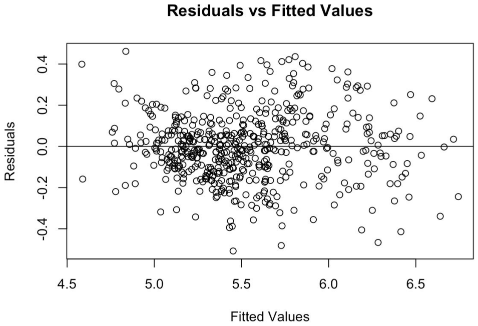
The residuals seem to hover around the 0 line, so no sign of a pattern.

-Residuals vs Presence of Pool:

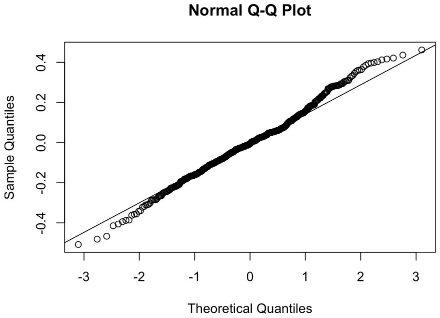


Similar to the previous plot, the residuals stay approximately around 0, so there is no strong correlation between presence of a pool and the residuals.

5. *Adding Interaction Terms:* Looking at the t tests for β_i=0 where i corresponds to each predictor, we see that some of the predictors have p values greater than 0.05. We can attempt to explain more variation by adding interaction terms. To find the appropriate terms, we use the add1() function to compare the current model with a model that contains a test interaction term. If adding that term yields a model with a higher adjusted R^2 and predictors that are significant, we add the interaction term to the model. We should test interaction terms of variables that seemed to be associated in the scatterplot matrices. After this process of trial and error, we find that adding interaction terms between square feet and garage size and square feet and number of bedrooms increases the adjusted R^2 and leads us to a model with only significant predictors (all p-values < 0.05).
6. *Final Residual Analysis*: We return for a final analysis of residuals to ensure that our model satisfies the conditions of linearity, equal variance, independence, and normality.



The residual vs fitted plot shows equal variance and linearity.



The Normal Quantile-Quantile plot shows the residuals are normal. (We do not know the order in which this data was taken, so we cannot plot a residuals vs order plot to check independence.)

7. *Testing for Influential Points*: The first method we will use to test if there are any outliers in the data set is externally studentized residuals. We find that there is only one outlier: the 24th observation. We also use Cook's Distance measure to see if there are any potentially influential based on that criteria. However, under Cook's there are no viable potentially influential points. So, we move onto testing if the 24th data point is influential. We exclude the point from the analysis, and compare the fitted model with 24 and without 24. Although excluding this point increases the adjusted R^2 by 0.0022, it does not significantly affect any coefficient estimates or test conclusions, so we cannot justifiably remove any data points.
8. *Final and Best Model:* We can finally conclude that the best model is:

```
SalesPrice=-4.224+0.67sq.feet-0.293constr.quality1-0.367constr.quality2-0.015ar
ch.style+0.004year.built+0.005lot.size+0.11garage.size+0.158num.beds-0.11near.h
              ighway-0.061sq.feet*num.beds-0.035sq.feet*garage.size
Where constr.quality1=0, constr.quality2=0 when Quality=1
constr.quality1=1, constr.quality2=0 when Quality=2
constr.quality1=0, constr.quality2=1 when Quality=3
```

This model was achieved through best subsets regression, and mainly using adjusted R^2 as criteria. By finding the model with the highest adjusted R^2, we choose the model with terms that best fit the line, while also adjusting for the number of terms in the model. By adding useful predictors, we have achieved a high adjusted R^2 and optimal variation explained by only the independent variables and interaction terms that actually affect the response.

Results and Interpretation

1. *Does the regression model contain at least one predictor useful in predicting the sales price?*

   After conducting an overall F-Test with H_0: β_1, β_2, β_3, β_4, β_5, β_6, β_7, β_8, β_9=0 (the slope coefficients we found for our best model) and H_A: $\beta_i \neq 0$ for at least one value of i (1-9), the ANOVA table returns a test statistic (F*) of 282.31, and a p-value of 2.2e-16. With a p-value approximately equal to zero, we reject our null hypothesis and have sufficient evidence to conclude that at least one of the slope parameters is not equal to zero, and therefore the model contains at least one predictor significantly useful in predicting sales price.
2. *If any, what effect does adjacency to highway have on the estimated sales price, holding all other factors constant?*

   If all other factors are held constant, when the house is adjacent to a highway it results in an estimated average increase of 0.11 dollars in home sales price.
3. *Using the adjusted R^2 as the criterion, which predictors would be used in the "best" model? Why?*

   By using the summary function, we were able to find the adjusted R^2 values for models with varying numbers of predictors. The 'which' section of the best subsets regression function allows us to see which specific predictors should be used in each model. From this, we found that the best model for our data uses 8 predictors (adjusted R^2=0.7872412). These 8 predictors are the finished square feet, construction quality, architectural style, year built, lot size, garage size, number of bedrooms, and adjacency to highway.

4. *Using the predictors found previously, how much variance in our response (sales price) can be explained by variance in our predictors?*

   After finding the best predictors for our model and adding necessary interaction terms, our final model has an R^2 value of 0.8495, meaning that 84.95% of variance in sales price can be explained by variance in our predictors.

5. *Are there any outliers that are influential?*

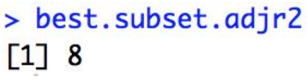
   Using the externally studentized residuals method, we only found one outlier in our data (at point 24). After testing the model with and without the point, we found that it is not significantly influential, as it does not seriously alter any coefficient estimates, hypothesis test conclusions, or the adjusted R^2. We also used the Cook's Distance measure to find any potentially influential points, but none were found according to this criteria. Although observation 24 is an outlier, it is not a result of procedural errors nor does it invalidate the measurement. Since it is representative of the intended study population, we should not delete any data point without a good, objective reason, just because it does not fit our pre-conceived regression model.

Conclusion

Based on our final model, we have found that the sales price of a residence in a midwestern city in the year 2002 was affected by many different characteristics of the home and surrounding area. The factor that most significantly affects the sales price is the finished square feet of the home. This is understandable, since usually a large home is going to sell at a significantly higher price than a small home. It also makes sense that the next most significant factor in the selling price of a home is the number of bedrooms, since a larger, more expensive house will likely have more bedrooms than a smaller, cheaper house. The main message that we can conclude from this analysis is that the selling price of a house is a complicated sum, to which many factors contribute. However, there are most likely even more factors that contribute which are not included as variables in the Real Estate Sales data set. For example: the location of the house, nearby school quality, crime rate of the neighborhood, etc. Though this data set does have a large sample, it is only from one location: a midwestern city. The elements that affect how much a house sells for is likely to differ depending on city, region, state, and country due to difference in economies, cultural preferences, and demand. The results from this regression analysis can be generalized and applied to most midwestern cities, but a different model should be formulated based on new data for other areas such as large cities, and more diverse variables should be taken into account.

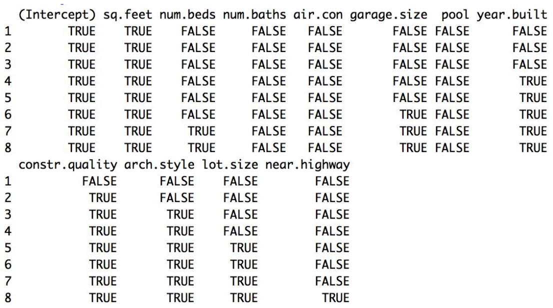
Appendix

```
# import data set realestate
library(readr)
realestate <- read_delim("~/Downloads/realestate.txt",
                         "\t", escape_double = FALSE, trim_ws = TRUE)
View(realestate)
# assign values to variables
attach(realestate)
sales.price=SalePrice
sq.feet=SqFeet
num.beds=Beds
num.baths=Baths
air.con=factor(Air)
garage.size=Garage
pool=factor(Pool)
year.built=Year
constr.quality=factor(Quality)
arch.style=Style
lot.size=Lot
near.highway=factor(Highway)
# form scatterplot matrices to find linear patterns
pairs(~sales.price+sq.feet+num.beds+num.baths+air.con+garage.size,main="Sales Price Against
First 5 Variables")
pairs(~sales.price+pool+year.built+constr.quality+arch.style+lot.size+near.highway,main="Sales
Price Against Last 6 Variables")
# use best subsets regression
library(leaps)
mod=regsubsets(cbind(sq.feet,num.beds,num.baths,air.con,garage.size,pool,year.built,constr.qual
ity,arch.style,lot.size,near.highway),sales.price)
summary.mod=summary(mod)
best.subset.adjr2<-which.max(summary.mod$adjr2)
Best.subset.adjr2
```



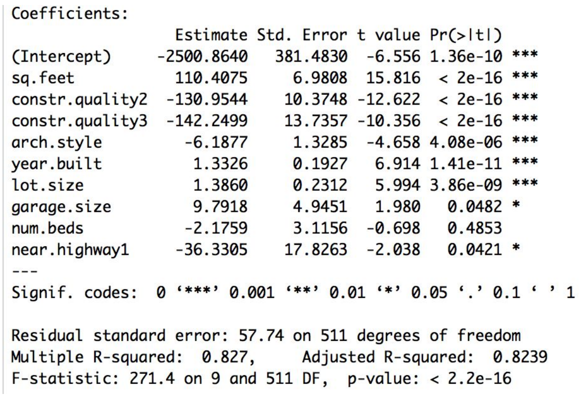
```
> best.subset.adjr2
[1] 8
```

```
# best model has 8 predictors
summary.mod$which
```



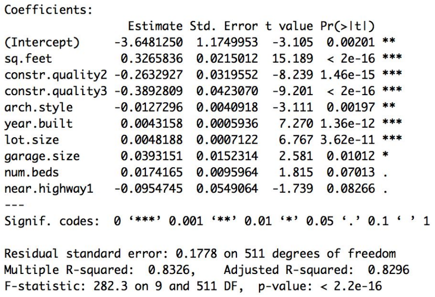
```
  (Intercept) sq.feet num.beds num.baths air.con garage.size  pool year.built
1        TRUE    TRUE    FALSE     FALSE   FALSE       FALSE FALSE      FALSE
2        TRUE    TRUE    FALSE     FALSE   FALSE       FALSE FALSE      FALSE
3        TRUE    TRUE    FALSE     FALSE   FALSE       FALSE FALSE      FALSE
4        TRUE    TRUE    FALSE     FALSE   FALSE       FALSE FALSE       TRUE
5        TRUE    TRUE    FALSE     FALSE   FALSE       FALSE FALSE       TRUE
6        TRUE    TRUE    FALSE     FALSE   FALSE        TRUE FALSE       TRUE
7        TRUE    TRUE     TRUE     FALSE   FALSE        TRUE FALSE       TRUE
8        TRUE    TRUE     TRUE     FALSE   FALSE        TRUE FALSE       TRUE
  constr.quality arch.style lot.size near.highway
1          FALSE      FALSE    FALSE        FALSE
2           TRUE      FALSE    FALSE        FALSE
3           TRUE       TRUE    FALSE        FALSE
4           TRUE       TRUE    FALSE        FALSE
5           TRUE       TRUE     TRUE        FALSE
6           TRUE       TRUE     TRUE        FALSE
7           TRUE       TRUE     TRUE        FALSE
8           TRUE       TRUE     TRUE         TRUE
```

```
# shows us which predictors are best
best.model=lm(sales.price~sq.feet+constr.quality+arch.style+year.built+lot.size+garage.size+num.beds+near.highway)
summary(best.model)
```



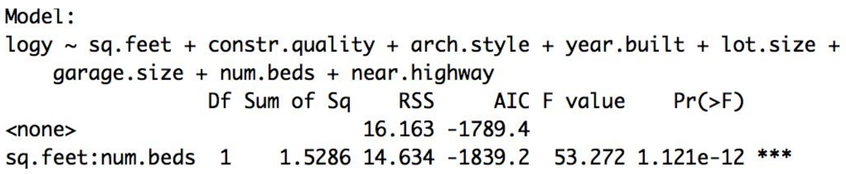
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -2500.8640   381.4830  -6.556 1.36e-10 ***
sq.feet           110.4075     6.9808  15.816  < 2e-16 ***
constr.quality2  -130.9544    10.3748 -12.622  < 2e-16 ***
constr.quality3  -142.2499    13.7357 -10.356  < 2e-16 ***
arch.style         -6.1877     1.3285  -4.658 4.08e-06 ***
year.built          1.3326     0.1927   6.914 1.41e-11 ***
lot.size            1.3860     0.2312   5.994 3.86e-09 ***
garage.size         9.7918     4.9451   1.980   0.0482 *  
num.beds           -2.1759     3.1156  -0.698   0.4853    
near.highway1     -36.3305    17.8263  -2.038   0.0421 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57.74 on 511 degrees of freedom
Multiple R-squared:  0.827,	Adjusted R-squared:  0.8239 
F-statistic: 271.4 on 9 and 511 DF,  p-value: < 2.2e-16
```

```
yhat=fitted(best.model)
resid=sales.price-yhat
plot(yhat,resid,main="Residuals vs. Fitted Values",ylab='Residuals',xlab='Fitted')
# this residual vs fitted plot shows nonconstant variance! need to transform Y vals
logy=log(y)
log.model=lm(logy~sq.feet+constr.quality+arch.style+year.built+lot.size+garage.size+num.beds+near.highway)
log.yhat=fitted(log.model)
log.resid=logy-log.yhat
plot(log.yhat,log.resid,main='Log-transformed Residuals vs Fitted Values',xlab='Fitted',ylab='Residuals')
abline(0,0)
# looks better
# plot residuals against omitted predictors to ensure no linearity pattern
plot(num.baths,log.resid,xlab='Number of Bathrooms',ylab='Residuals',main='Residuals vs Number of Bathrooms')
plot(air.con,log.resid,xlab='Presence of Air Conditioning (0=No, 1=Yes)',ylab='Residuals',main='Residuals vs Air Conditioning')
plot(pool,log.resid,xlab='Presence of Pool (0=No, 1=Yes)',ylab='Residuals',main="Residuals vs Presence of Pool")
# no linear patterns
summary(log.model)
```



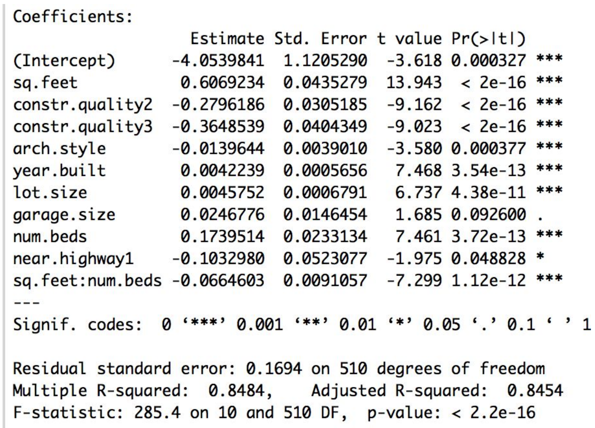
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -3.6481250  1.1749953  -3.105  0.00201 ** 
sq.feet          0.3265836  0.0215012  15.189  < 2e-16 ***
constr.quality2 -0.2632927  0.0319552  -8.239 1.46e-15 ***
constr.quality3 -0.3892809  0.0423070  -9.201  < 2e-16 ***
arch.style      -0.0127296  0.0040918  -3.111  0.00197 ** 
year.built       0.0043158  0.0005936   7.270 1.36e-12 ***
lot.size         0.0048188  0.0007122   6.767 3.62e-11 ***
garage.size      0.0393151  0.0152314   2.581  0.01012 *  
num.beds         0.0174165  0.0095964   1.815  0.07013 .  
near.highway1   -0.0954745  0.0549064  -1.739  0.08266 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1778 on 511 degrees of freedom
Multiple R-squared:  0.8326,	Adjusted R-squared:  0.8296 
F-statistic: 282.3 on 9 and 511 DF,  p-value: < 2.2e-16
```

```
# some p values greater than 0.05, so we add interaction terms
add1(log.model,~.sq.feet*num.beds,test='F',scope=~sq.feet+constr.quality+arch.style+year.built
+lot.size+garage.size+num.beds+near.highway+sq.feet*num.beds)
```



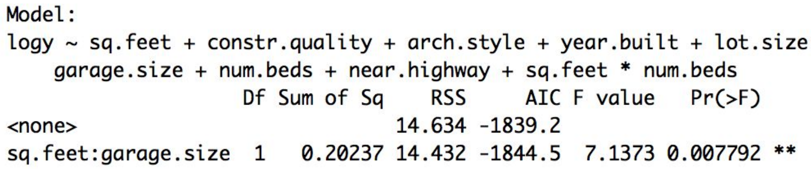
```
Model:
logy ~ sq.feet + constr.quality + arch.style + year.built + lot.size +
    garage.size + num.beds + near.highway
                 Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                        16.163 -1789.4
sq.feet:num.beds  1    1.5286 14.634 -1839.2  53.272 1.121e-12 ***
```

```
# p value is significantly lower than 0.05 so we add it to the model
log.model=lm(logy~sq.feet+constr.quality+arch.style+year.built+lot.size+garage.size+num.beds
+near.highway+sq.feet*num.beds)
summary(log.model)
```



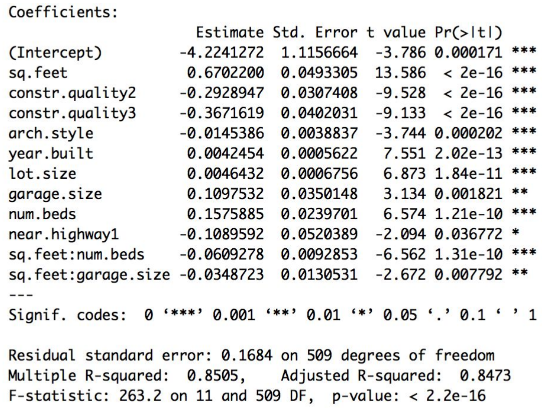
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4.0539841  1.1205290  -3.618 0.000327 ***
sq.feet          0.6069234  0.0435279  13.943  < 2e-16 ***
constr.quality2 -0.2796186  0.0305185  -9.162  < 2e-16 ***
constr.quality3 -0.3648539  0.0404349  -9.023  < 2e-16 ***
arch.style      -0.0139644  0.0039010  -3.580 0.000377 ***
year.built       0.0042239  0.0005656   7.468 3.54e-13 ***
lot.size         0.0045752  0.0006791   6.737 4.38e-11 ***
garage.size      0.0246776  0.0146454   1.685 0.092600 .
num.beds         0.1739514  0.0233134   7.461 3.72e-13 ***
near.highway1   -0.1032980  0.0523077  -1.975 0.048828 *
sq.feet:num.beds -0.0664603  0.0091057  -7.299 1.12e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1694 on 510 degrees of freedom
Multiple R-squared:  0.8484,    Adjusted R-squared:  0.8454
F-statistic: 285.4 on 10 and 510 DF,  p-value: < 2.2e-16
```

```
# adjusted r2 has increased significantly. lets try to lower the p value of garage.size
add1(log.model,~.sq.feet*garage.size,test='F',scope=~sq.feet+constr.quality+arch.style+year.buil
t+lot.size+garage.size+num.beds+near.highway+sq.feet*num.beds+sq.feet*garage.size)
```



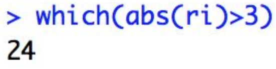
```
Model:
logy ~ sq.feet + constr.quality + arch.style + year.built + lot.size +
    garage.size + num.beds + near.highway + sq.feet * num.beds
                    Df Sum of Sq    RSS     AIC F value   Pr(>F)
<none>                           14.634 -1839.2
sq.feet:garage.size  1   0.20237 14.432 -1844.5  7.1373 0.007792 **
```

```
# p value is significantly less than 0.05 so we add garage.size*sq.feet to the model
log.model=lm(logy~sq.feet+constr.quality+arch.style+year.built+lot.size+garage.size+num.beds
+near.highway+sq.feet*num.beds+sq.feet*garage.size)
summary(log.model)
```



```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -4.2241272  1.1156664  -3.786 0.000171 ***
sq.feet             0.6702200  0.0493305  13.586  < 2e-16 ***
constr.quality2    -0.2928947  0.0307408  -9.528  < 2e-16 ***
constr.quality3    -0.3671619  0.0402031  -9.133  < 2e-16 ***
arch.style         -0.0145386  0.0038837  -3.744 0.000202 ***
year.built          0.0042454  0.0005622   7.551 2.02e-13 ***
lot.size            0.0046432  0.0006756   6.873 1.84e-11 ***
garage.size         0.1097532  0.0350148   3.134 0.001821 ** 
num.beds            0.1575885  0.0239701   6.574 1.21e-10 ***
near.highway1      -0.1089592  0.0520389  -2.094 0.036772 *  
sq.feet:num.beds   -0.0609278  0.0092853  -6.562 1.31e-10 ***
sq.feet:garage.size -0.0348723  0.0130531  -2.672 0.007792 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1684 on 509 degrees of freedom
Multiple R-squared:  0.8505,	Adjusted R-squared:  0.8473 
F-statistic: 263.2 on 11 and 509 DF,  p-value: < 2.2e-16
```

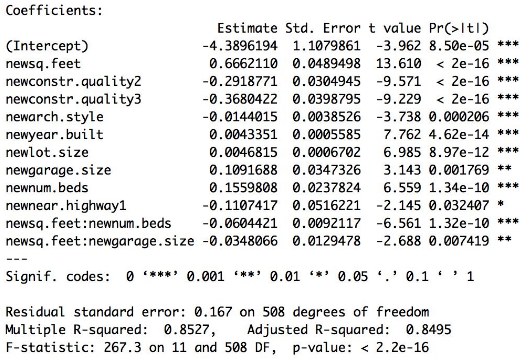
```
# all p values are less than 0.05 and r2 is high! this is an appropriate regression model.
log.yhat2=fitted(log.model)
log.resid2=logy-log.yhat2
plot(log.yhat2,log.resid2,xlab='Fitted Values',ylab='Residuals',main='Residuals vs Fitted Values')
abline(0,0)
# residuals vs fitted plot looks linear, equal variance, no outliers.
qqnorm(log.resid2)
qqline(log.resid2)
# residuals follow normal distribution
# use externally studentized residuals to find any outliers
ri=rstudent(log.model)
which(abs(ri)>3)
```



```
> which(abs(ri)>3)
24
```

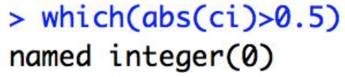
```
# 24th data point is the only outlier according to externally studentized residuals
# remove 24th point and compare with full data to see if influential
newrealestate<-realestate[-c(24),]
attach(newrealestate)
newsales.price=SalePrice
newsq.feet=SqFeet
newnum.beds=Beds
newnum.baths=Baths
newair.con=factor(Air)
newgarage.size=Garage
newpool=factor(Pool)
newyear.built=Year
newconstr.quality=factor(Quality)
newarch.style=Style
newlot.size=Lot
newnear.highway=factor(Highway)
```

```
log.model=lm(log(newsales.price)~newsq.feet+newconstr.quality+newarch.style+newyear.built+newlot.size+newgarage.size+newnum.beds+newnear.highway+newsq.feet*newnum.beds+newsq.feet*newgarage.size)
summary(log.model)
```



```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -4.3896194  1.1079861  -3.962 8.50e-05 ***
newsq.feet                0.6662110  0.0489498  13.610  < 2e-16 ***
newconstr.quality2       -0.2918771  0.0304945  -9.571  < 2e-16 ***
newconstr.quality3       -0.3680422  0.0398795  -9.229  < 2e-16 ***
newarch.style            -0.0144015  0.0038526  -3.738 0.000206 ***
newyear.built             0.0043351  0.0005585   7.762 4.62e-14 ***
newlot.size               0.0046815  0.0006702   6.985 8.97e-12 ***
newgarage.size            0.1091688  0.0347326   3.143 0.001769 ** 
newnum.beds               0.1559808  0.0237824   6.559 1.34e-10 ***
newnear.highway1         -0.1107417  0.0516221  -2.145 0.032407 *  
newsq.feet:newnum.beds   -0.0604421  0.0092117  -6.561 1.32e-10 ***
newsq.feet:newgarage.size -0.0348066  0.0129478  -2.688 0.007419 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.167 on 508 degrees of freedom
Multiple R-squared:  0.8527,	Adjusted R-squared:  0.8495 
F-statistic: 267.3 on 11 and 508 DF,  p-value: < 2.2e-16
```

```
# point 24 is an outlier, but not influential. does not significantly change any p values, estimates, or adjusted r2
# use cook's distance as another measure to find outliers
ci=cooks.distance(log.model)
which(abs(ci)>0.5)
```



```
> which(abs(ci)>0.5)
named integer(0)
```

```
# according to cook's distance, there are no outliers potential outliers
```