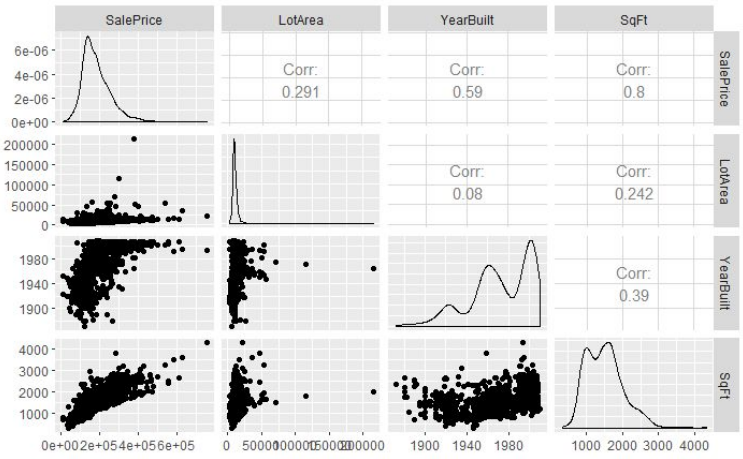
STAT 3301 Real Estate Analysis Project Report

Author: Shijie Qu, Yifan Song

In this project, we are doing analysis on a Iowa real estate data set to build a model for estimating the sale price of a housing.

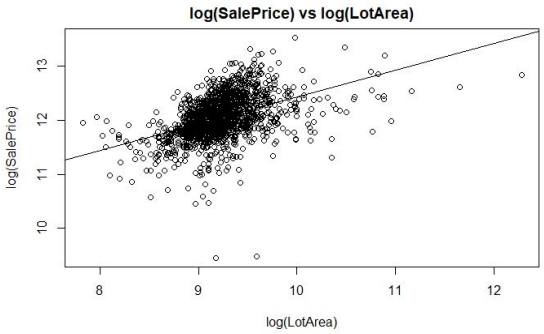
AR-1

We first made a pairwise scatterplots to get the relationship between SalePrice, LotArea, YearBuilt and SqFt. Below is the result:

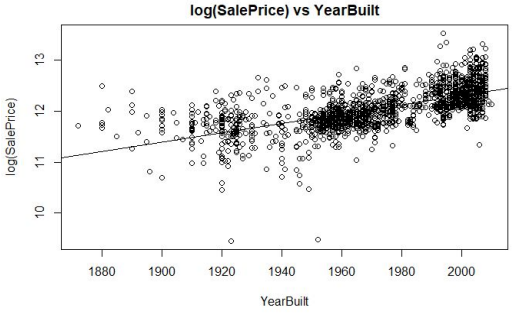


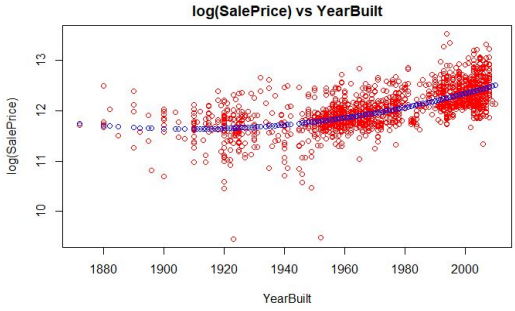
From the above, we can see a strong positive association between SqFt and SalePrice from the plot with a correlation of 0.8; YearBuilt and SalePrice also have a relatively strong positive correlation but might need some transformation according to the plot; all other relations seem to have very moderate association so we should use some transformations to find the relationship between LotArea and SalePrice.

To find any transformations for building the model, we first access LotArea and SalePrice where a transformation is necessary. As the values range over more than one order of magnitude and are positive, we start with a log transformation. After trying taking logarithm on either/both side, we found that the best choice is to take logarithm on both LotArea and SalePrice. The plot below shows the new relation with a R^2 of 0.20. Though it's still not very correlated, it's better compare to the original data.

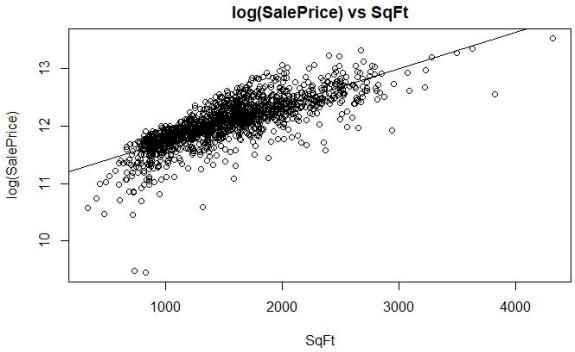


For YearBuilt and SalePrice, we still start with the log transformation and decide to take logarithm in the SalePrice side. The new model has a R^2 of 0.42 which is better than the original one. An interesting observation is that the linear relation fits where after the year of 1940 and the whole plot seems to have a quadratic relation. So we also tried to make a quadratic model, shown in the second graph, which has a higher R^2 of 0.45. But as it's only a minor improve, we still choose the log model to make the whole analysis consistent and easy to interpret.





To keep consistent on the log transformation on SalePrice, we also make the model between log(SalePrice) and SqFt. As the model also has a good fit and an almost same R^2 with the original one, we decide to use this new model.



AR-2

From the discussion in AR-1, we made a linear model for log(SalePrice) with predictors SqFt, log(LotArea) and YearBuilt. Here is the summary of model:

```
Call:
lm(formula = log(SalePrice) ~ SqFt + log(LotArea) + YearBuilt,
    data = realEst_data)

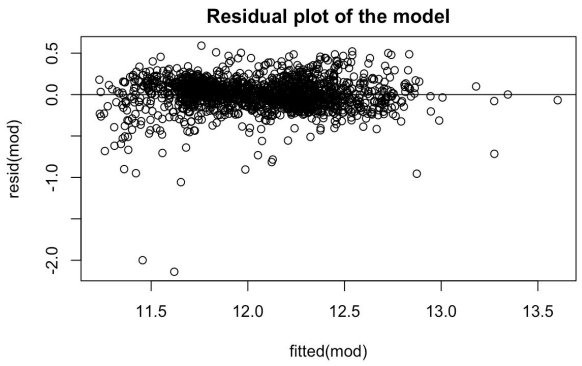
Residuals:
     Min       1Q   Median       3Q      Max
-2.13909 -0.09405  0.01214  0.10990  0.59032

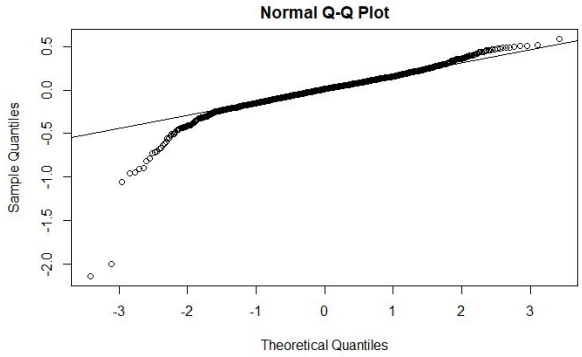
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.982e-01  3.710e-01  -1.613    0.107
SqFt         4.756e-04  1.114e-05  42.691   <2e-16 ***
log(LotArea) 1.273e-01  1.476e-02   8.623   <2e-16 ***
YearBuilt    5.455e-03  1.851e-04  29.473   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1947 on 1594 degrees of freedom
Multiple R-squared:  0.7697,    Adjusted R-squared:  0.7692
F-statistic:  1775 on 3 and 1594 DF,  p-value: < 2.2e-16
```

According to the p-value and R^2, we can say that this is a relatively good fitted model. But we are also curious in whether the interactions should be included in the model. The potential interactions we worked on are the three interactions between each two predictors. We first tried both AIC and BIC to compare the main effect model and the mode include one of the interactions. The results showed that all three interactions is in the “middle” that AIC differences indicate to include the interaction while BIC differences indicate to keep the main effect model. We then run the F-test and tried to include the interaction for a new model. However, none of these new models gave a better fit compared to the main effect model; some even increased the original predictor’s p-value to a high level. Thus, we were not including any interactions and kept the original model. (Detailed process and results in R-code and appendix1)

In the next step, we made the residual plot and Normal Q-Q plot for our model. From the residual plot, we can see that the points are evenly distributed across the zero-line without any significant pattern which demonstrates a good fit for the model. Some outliers exist in the left-bottom of the graph, these points may represent some extremely bad house with low price. In the Q-Q plot, the points also have a very good fit with the line except the left tail part. The reason should be the same for some cheap housings. Overall, the residual and Q-Q plot both indicate a good fit for our model.





For square footage interpretation: With LotArea and YearBuilt fixed, one unit increase of SqFt, would result a 4.756e-04 unit increase in log(SalePrice), in other word, 100*(e^(4.756e-04)-1)=0.05 percent increase of SalePrice.

AR-3

In this part, before actually adding variables, we first decided the type of variables, in particular whether continuous or factor for Bedrooms, Rooms, FullBath and HalfBath. Take Bedrooms as an example, we made two models: one of them is the original model plus Bedrooms as numerical variable and the other one is the original model plus Bedrooms as factor variable. As we are clear which variable to test, we used BIC for both models and checked their difference. The model with factor Bedroom has a lower BIC therefore we decided to make it a factor variable. We had the same process on other three variables and decided the following: Rooms-factor, FullBath-continuous, HalfBath-continuous. (Detailed process and results in R-code and appendix2)

Then, we thought there would be an overlap meaning between Rooms and other specific room variable (Bedrooms+FullBath+HalfBath). So we are curious if we need both of them. We then create three models: the first one including both(all) variables, the second one including only bedrooms and baths, and the third one including only rooms. We also run BIC test on them and the result showed that the best model is the second one without room variable. (Detailed process and results in R-code and appendix3)

At last, we ran the BIC stepwise search both from null and full. The two methods removed the FullBath and gave the same result, our final model. (Detailed process and results in R-code and appendix4)

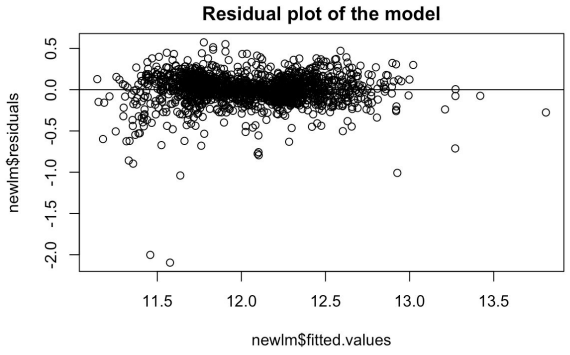
```
Call:
lm(formula = log(SalePrice) ~ SqFt + YearBuilt + Style + Bedroom_Fact +
    log(LotArea) + HalfBath, data = realEst_data)

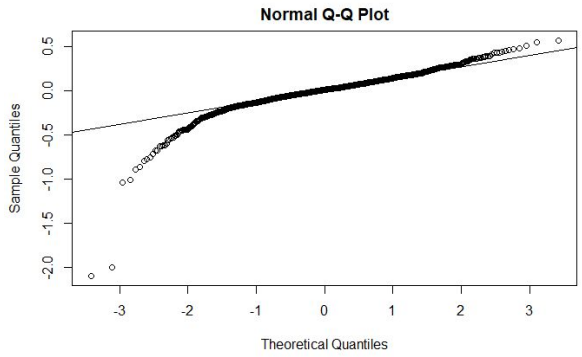
Residuals:
     Min       1Q   Median       3Q      Max
-2.09532 -0.07741  0.01213  0.09846  0.57354

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.132e-01  4.054e-01   1.759  0.07871 .
SqFt           6.313e-04  1.541e-05  40.973  < 2e-16 ***
YearBuilt      5.100e-03  1.918e-04  26.596  < 2e-16 ***
Style2Story   -8.338e-02  1.593e-02  -5.236 1.86e-07 ***
Bedroom_Fact1 -3.338e-01  1.119e-01  -2.983  0.00290 **
Bedroom_Fact2 -3.010e-01  1.077e-01  -2.796  0.00524 **
Bedroom_Fact3 -3.202e-01  1.071e-01  -2.988  0.00285 **
Bedroom_Fact4 -4.343e-01  1.079e-01  -4.023 6.02e-05 ***
Bedroom_Fact5 -6.260e-01  1.165e-01  -5.374 8.83e-08 ***
log(LotArea)   7.776e-02  1.476e-02   5.267 1.58e-07 ***
HalfBath      -5.508e-02  1.392e-02  -3.956 7.95e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1829 on 1587 degrees of freedom
Multiple R-squared:  0.7977,    Adjusted R-squared:  0.7964
F-statistic: 625.8 on 10 and 1587 DF,  p-value: < 2.2e-16
```

With the same process in AR-2, we made the residual plot and Q-Q plot for our new model. Both the plots are similar with the plots in AR-2. In the residual plot,the points are evenly distributed across the zero-line without significant pattern, while in Q-Q plot, the points have a very good fit with the line. The outliers and left tail still exist because of some cheap housings. But overall, the model still has a good fit.





For square footage interpretation: With Style, Bedrooms, Halfbath, LotArea and YearBuilt fixed, one unit increase of SqFt, would result a 6.313e-04 unit increase in log(SalePrice), in other word, 100*(e^(6.313e-04)-1)=0.06 percent increase of SalePrice.

AR-4

- Detailed information of the house:

Suppose the square footage of the given house above ground is 1995 square feet with the entire lot size be 15500 square feet. The house will be a 2-story house and is relatively new, say built in 2013. There are 10 rooms in total for the entire house, with 3 bedrooms and 4 bedrooms. Among the 4 bathrooms, 2 are half bath and 2 are full bath. All the rooms in the house are above grade.

- Our model:

As mentioned above, the number of full bathrooms and the number of total rooms is not included as predictors in our model. Therefore, our final model is the following:

$$\begin{aligned}\log(SalePrice) &= 6.313 * 10^{-4} SqFt + 5.1 * 10^{-3} YearBuilt - 8.338 * 10^{-2} Style2Story \\ &- 0.3338 BedroomFact1 - 0.301 BedroomFact2 - 0.3202 BedroomFact3 \\ &- 0.4343 BedroomFact4 - 0.626 BedroomFact5 + 7.776 \\ &* 10^{-2} \log(LotArea) - 5.508 * 10^{-2} HalfBath + 0.7132\end{aligned}$$

- Prediction using the model:

By plugging in the needed predictor values (SqFt = 1995, YearBuilt = 2013, Style = '2Story', Bedroom_Fact = '3', LotArea = 15500, HalfBath = 2), we got the fitted value of log(SalePrice). Then, we take the exponent of it to acquire the fitted value of SalePrice, which is 261847.4 dollars in our case.

The 95% prediction interval of this one yields a range between 12.11508 dollars and 12.83595 dollars for log(SalePrice). So when we take the exponent, we get (182605.8,375475.8) for the SalePrice of our house. Therefore, given the information of the house above, the sale price of the house is predicted to be between 182605.8 and 375475.8 dollars. (95% PI) (Results see figure 5 in appendix)

Appendix

1.

```
##Check interaction
AIC(lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+SqFt:log(LotArea),data = realEst_data)) - AIC(mod)
BIC(lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+SqFt:log(LotArea),data = realEst_data)) - BIC(mod)
AIC(lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+SqFt:YearBuilt,data = realEst_data)) - AIC(mod)
BIC(lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+SqFt:YearBuilt,data = realEst_data)) - BIC(mod)
AIC(lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+log(LotArea):YearBuilt,data = realEst_data)) - AIC(mod)
BIC(lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+log(LotArea):YearBuilt,data = realEst_data)) - BIC(mod)
anova(mod, lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+SqFt:log(LotArea),data = realEst_data))
anova(mod, lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+SqFt:YearBuilt,data = realEst_data))
anova(mod, lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+log(LotArea):YearBuilt,data = realEst_data))
```

```
[1] -3.145902
[1] 2.230606
[1] -1.323251
[1] 4.053257
[1] -5.017692
[1] 0.3588163
Analysis of Variance Table

Model 1: log(SalePrice) ~ SqFt + log(LotArea) + YearBuilt
Model 2: log(SalePrice) ~ SqFt + log(LotArea) + YearBuilt + SqFt:log(LotArea)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   1594 60.428
2   1593 60.234  1   0.19428 5.1381 0.02354 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Model 1: log(SalePrice) ~ SqFt + log(LotArea) + YearBuilt
Model 2: log(SalePrice) ~ SqFt + log(LotArea) + YearBuilt + SqFt:YearBuilt
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   1594 60.428
2   1593 60.303  1   0.12554 3.3163 0.06878 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Model 1: log(SalePrice) ~ SqFt + log(LotArea) + YearBuilt
Model 2: log(SalePrice) ~ SqFt + log(LotArea) + YearBuilt + log(LotArea):YearBuilt
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   1594 60.428
2   1593 60.163  1   0.26479 7.0111 0.00818 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2.

```
# try to use see whether use factored variable or continuous variable
HalfBath_Fact = as.factor(realEst_data$HalfBath)
FullBath_Fact = as.factor(realEst_data$FullBath)
Bedroom_Fact = as.factor(realEst_data$Bedrooms)
Room_Fact = as.factor(realEst_data$Rooms)
# decide whether combine half&full bath
# decide rooms to be continuous or factored variable
roomFact = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+Room_Fact+Style,data = realEst_data)
roomCont = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+Rooms+Style,data = realEst_data)
diffRoom = BIC(roomFact) - BIC(roomCont)
diffRoom
# decide HalfBath to be continuous or factored variable
HalfBathFact = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+HalfBath_Fact+Style,data = realEst_data)
HalfBathCont = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+HalfBath+Style,data = realEst_data)
diffHalfBath = BIC(HalfBathFact) - BIC(HalfBathCont)
diffHalfBath
# decide FullBath to be continuous or factored variable
FullBathFact = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+FullBath_Fact+Style,data = realEst_data)
FullBathCont = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+FullBath+Style,data = realEst_data)
diffFullBath = BIC(FullBathFact) - BIC(FullBathCont)
diffFullBath
# decide Bedrooms to be continuous or factored variable
BedroomsFact = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+Bedroom_Fact+Style,data = realEst_data)
BedroomsCont = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+Bedrooms+Style,data = realEst_data)
diffBedrooms = BIC(BedroomsFact) - BIC(BedroomsCont)
diffBedrooms

```

```
[1] -3.789575
[1] 1.704016
[1] 5.270383
[1] -17.15674
```

3.

```
```{r}
# decide whether should replace Rooms as a predictor
WithRooms = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+HalfBath+FullBath+Bedroom_Fact+Room_Fact+Style,data = realEst_data)
NoRooms = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+HalfBath+FullBath+Bedroom_Fact+Style,data = realEst_data)
diffR = BIC(WithRooms) - BIC(NoRooms)
diffR
# decide whether should replace other type of rooms (keep only with the # of total rooms)
WithOtherRooms = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+HalfBath+FullBath+Bedroom_Fact+Room_Fact+Style,data = realEst_data)
NoOtherRooms = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+Room_Fact+Style,data = realEst_data)
diffO = BIC(WithOtherRooms) - BIC(NoOtherRooms)
diffO

# decide whether to combine half&full bath to bath
bath = realEst_data$FullBath+0.5*realEst_data$HalfBath
combineBath = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+bath+Bedroom_Fact+Style,data = realEst_data)
seperateBath = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+HalfBath+FullBath+Bedroom_Fact+Style,data = realEst_data)
diffR = BIC(combineBath) - BIC(seperateBath)
diffR

## therefore, use factored value bedroom and continuous for full bath and half bath & without rooms
```
```

```
[1] 16.9608
[1] -40.37383
[1] 6.922164
```

4.

```
# stepwise regression by AIC
null = lm(log(SalePrice)~1, data = realEst_data )
full = lm(log(SalePrice)~SqFt+log(LotArea)+YearBuilt+HalfBath+FullBath+Bedroom_Fact+Style,data = realEst_data)
library(MASS)
stepAIC(full,scope = list(lower=null,upper=full),direction="both",k=log(dim(realEst_data)[1]));
```

```
Start:  AIC=-5354.04
log(SalePrice) ~ SqFt + log(LotArea) + YearBuilt + HalfBath +
    FullBath + Bedroom_Fact + Style

               Df Sum of Sq    RSS     AIC
- FullBath      1     0.047 53.066 -5360.0
<none>                     53.019 -5354.0
- HalfBath      1     0.563 53.582 -5344.5
- Style         1     0.731 53.750 -5339.5
- log(LotArea)  1     0.930 53.949 -5333.6
- Bedroom_Fact  5     2.960 55.979 -5304.1
- YearBuilt     1    17.834 70.853 -4898.1
- SqFt          1    41.923 94.942 -4430.4

Step:  AIC=-5360.02
log(SalePrice) ~ SqFt + log(LotArea) + YearBuilt + HalfBath +
    Bedroom_Fact + Style

               Df Sum of Sq     RSS     AIC
<none>                      53.066 -5360.0
+ FullBath      1     0.047  53.019 -5354.0
- HalfBath      1     0.523  53.589 -5351.7
- Style         1     0.917  53.982 -5340.0
- log(LotArea)  1     0.928  53.993 -5339.7
- Bedroom_Fact  5     2.973  56.038 -5309.8
- YearBuilt     1    23.653  76.718 -4778.4
- SqFt          1    56.134 109.199 -4214.2

Call:
lm(formula = log(SalePrice) ~ SqFt + log(LotArea) + YearBuilt +
    HalfBath + Bedroom_Fact + Style, data = realEst_data)

Coefficients:
  (Intercept)           SqFt   log(LotArea)      YearBuilt       HalfBath  Bedroom_Fact1  Bedroom_Fact2  Bedroom_Fact3
    0.7131947      0.0006313      0.0777602      0.0051000     -0.0550769     -0.3338386     -0.3010446     -0.3201789
Bedroom_Fact4  Bedroom_Fact5    Style2Story
   -0.4342703     -0.6260150     -0.0833815
```

```
newlm = stepAIC(null,scope = list(upper=full),direction="both",k=log(dim(realEst_data)[1]));
newlm$terms
```

```
Start:  AIC=-2880.04
log(SalePrice) ~ 1

               Df Sum of Sq     RSS     AIC
+ SqFt          1   164.135  98.199 -4442.9
+ FullBath      1   118.872 143.461 -3837.1
+ YearBuilt     1   109.033 153.301 -3731.1
+ log(LotArea)  1    51.374 210.960 -3220.9
+ Bedroom_Fact  5    41.459 220.875 -3118.0
+ HalfBath      1    35.422 226.912 -3104.5
+ Style         1    27.494 234.840 -3049.6
<none>                      262.334 -2880.0

Step:  AIC=-4442.89
log(SalePrice) ~ SqFt

               Df Sum of Sq     RSS     AIC
+ YearBuilt     1    34.952  63.247 -5138.6
+ Bedroom_Fact  5    12.473  85.726 -4623.1
+ Style         1     9.652  88.547 -4600.8
+ FullBath      1     7.083  91.116 -4555.1
+ log(LotArea)  1     4.839  93.360 -4516.3
+ HalfBath      1     1.175  97.023 -4454.8
<none>                      98.199 -4442.9
- SqFt          1   164.135 262.334 -2880.0

Step:  AIC=-5138.55
log(SalePrice) ~ SqFt + YearBuilt

               Df Sum of Sq     RSS     AIC
+ Style         1     5.604  57.643 -5279.4
+ HalfBath      1     3.844  59.403 -5231.4
+ Bedroom_Fact  5     4.566  58.681 -5221.4
+ log(LotArea)  1     2.819  60.428 -5204.0
<none>                      63.247 -5138.6
+ FullBath      1     0.012  63.235 -5131.5
- YearBuilt     1    34.952  98.199 -4442.9
- SqFt          1    90.054 153.301 -3731.1

Step:  AIC=-5279.43
log(SalePrice) ~ SqFt + YearBuilt + Style

               Df Sum of Sq     RSS     AIC
+ Bedroom_Fact  5     3.218  54.425 -5334.4
+ log(LotArea)  1     1.004  56.639 -5300.1
+ HalfBath      1     0.462  57.181 -5284.0
```

```
+ Style         1     9.652  88.547 -4600.8
+ FullBath      1     7.083  91.116 -4555.1
+ log(LotArea)  1     4.839  93.360 -4516.3
+ HalfBath      1     1.175  97.023 -4454.8
<none>                       98.199 -4442.9
- SqFt          1   164.135 262.334 -2880.0

Step:  AIC=-5138.55
log(SalePrice) ~ SqFt + YearBuilt

               Df Sum of Sq     RSS     AIC
+ Style         1     5.604  57.643 -5279.4
+ HalfBath      1     3.844  59.403 -5231.4
+ Bedroom_Fact  5     4.566  58.681 -5221.4
+ log(LotArea)  1     2.819  60.428 -5204.0
<none>                       63.247 -5138.6
+ FullBath      1     0.012  63.235 -5131.5
- YearBuilt     1    34.952  98.199 -4442.9
- SqFt          1    90.054 153.301 -3731.1

Step:  AIC=-5279.43
log(SalePrice) ~ SqFt + YearBuilt + Style

               Df Sum of Sq     RSS     AIC
+ Bedroom_Fact  5     3.218  54.425 -5334.4
+ log(LotArea)  1     1.004  56.639 -5300.1
+ HalfBath      1     0.462  57.181 -5284.9
<none>                       57.643 -5279.4
+ FullBath      1     0.002  57.641 -5272.1
- Style         1     5.604  63.247 -5138.6
- YearBuilt     1    30.904  88.547 -4600.8
- SqFt          1    81.135 138.778 -3882.8

Step:  AIC=-5334.36
log(SalePrice) ~ SqFt + YearBuilt + Style + Bedroom_Fact

               Df Sum of Sq     RSS     AIC
+ log(LotArea)  1     0.836  53.589 -5351.7
+ HalfBath      1     0.432  53.993 -5339.7
<none>                       54.425 -5334.4
+ FullBath      1     0.004  54.421 -5327.1
- Bedroom_Fact  5     3.218  57.643 -5279.4
- Style         1     4.256  58.681 -5221.4
- YearBuilt     1    23.726  78.151 -4763.5
- SqFt          1    74.431 128.856 -3964.5

Step:  AIC=-5351.71
log(SalePrice) ~ SqFt + YearBuilt + Style + Bedroom_Fact + log(LotArea)

               Df Sum of Sq     RSS     AIC
+ HalfBath      1     0.523  53.066 -5360.0
<none>                       53.589 -5351.7
+ FullBath      1     0.007  53.582 -5344.5
- log(LotArea)  1     0.836  54.425 -5334.4
- Bedroom_Fact  5     3.050  56.639 -5300.1
- Style         1     2.862  56.451 -5275.9
```

```
- Bedroom_Fact  5     3.050  56.639 -5300.1
- Style         1     2.862  56.451 -5275.9
- YearBuilt     1    23.422  77.011 -4779.6
- SqFt          1    56.001 109.590 -4215.9

Step:  AIC=-5360.02
log(SalePrice) ~ SqFt + YearBuilt + Style + Bedroom_Fact + log(LotArea) +
    HalfBath

               Df Sum of Sq     RSS     AIC
<none>                      53.066 -5360.0
+ FullBath      1     0.047  53.019 -5354.0
- HalfBath      1     0.523  53.589 -5351.7
- Style         1     0.917  53.982 -5340.0
- log(LotArea)  1     0.928  53.993 -5339.7
- Bedroom_Fact  5     2.973  56.038 -5309.8
- YearBuilt     1    23.653  76.718 -4778.4
- SqFt          1    56.134 109.199 -4214.2
log(SalePrice) ~ SqFt + YearBuilt + Style + Bedroom_Fact + log(LotArea) +
    HalfBath
attr(,"variables")
list(log(SalePrice), SqFt, YearBuilt, Style, Bedroom_Fact, log(LotArea),
    HalfBath)
attr(,"factors")
               SqFt YearBuilt Style Bedroom_Fact log(LotArea) HalfBath
log(SalePrice)    0         0     0            0            0        0
SqFt              1         0     0            0            0        0
YearBuilt         0         1     0            0            0        0
Style             0         0     1            0            0        0
Bedroom_Fact      0         0     0            1            0        0
log(LotArea)      0         0     0            0            1        0
HalfBath          0         0     0            0            0        1
attr(,"term.labels")
[1] "SqFt"         "YearBuilt"    "Style"        "Bedroom_Fact" "log(LotArea)" "HalfBath"
attr(,"order")
[1] 1 1 1 1 1 1
attr(,"intercept")
[1] 1
attr(,"response")
[1] 1
attr(,".Environment")
<environment: R_GlobalEnv>
attr(,"predvars")
list(log(SalePrice), SqFt, YearBuilt, Style, Bedroom_Fact, log(LotArea),
    HalfBath)
attr(,"dataClasses")
log(SalePrice)          SqFt      YearBuilt         Style   Bedroom_Fact   log(LotArea)       HalfBath
     "numeric"     "numeric"     "numeric"      "factor"       "factor"     "numeric"      "numeric"
```

5. Fitted value of AR-4, and 95% PI

```{r}
# 95% prediction interval for log(salePrize)
logPI = predict(newlm, newdata=d, interval="prediction", level=.95)
logPI
realPI = exp(logPI)
realPI

```

```
       fit      lwr      upr
1 12.47552 12.11508 12.83595
       fit      lwr      upr
1 261847.4 182605.8 375475.8
```