LI WILES
DEPARTMENT OF MATHEMATICS

STATISTICAL PROJECT@ ILLINOIS STATE UNIVERSITY

HOUSE PRICE IN KING COUNTY, WASHINGTON

Photo source:http://www.estately.com/

AGENDA

- Data
- Objective
- Model Building
  - Normality & Constant variance
  - Multicollinearity
  - Transformation
  - Subset Selection
- Model Result & Thoughts

DATA

- Data source: Kaggle.com
- Data size: 21,613 observations
- Target Variable: Sales Price of a House
- Predictor Variables: 20

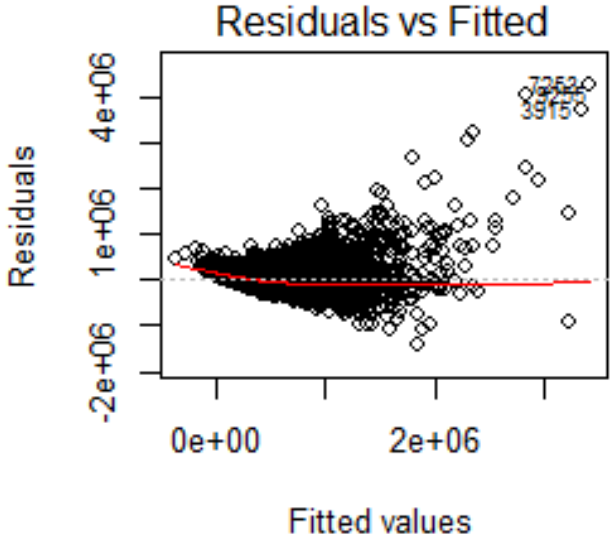
  Including:

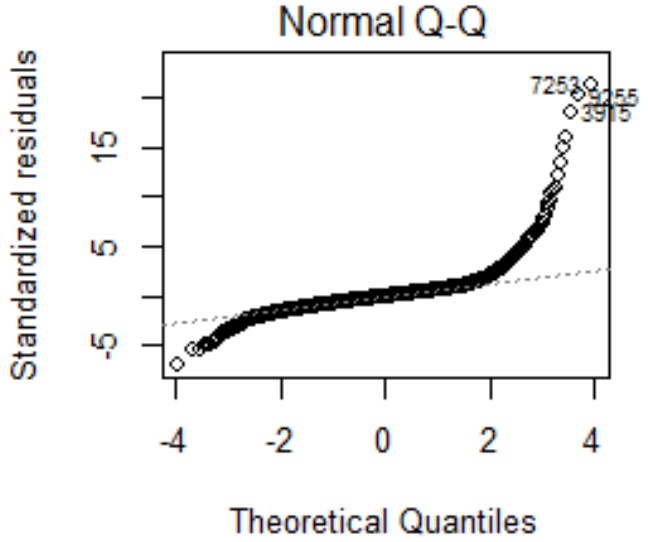
- house purchase date, # of bedrooms, house square footage, lot size, view, condition, grade, year the house's built, latitude, Average house square footage of the 15 closest neighbors, zip code, sqft of basement etc…..

OBJECTIVE

- The main purpose is building a regression model to predict house price in King County, Washington. Relations between house price and house features will be discussed. Potential business use for the model is to help individuals to evaluate whether a house is priced fairly and be used as reference in flipping houses.

NORMALITY & CONSTANT VARIANCE






```
> shapiro.test(res1)
Error in shapiro.test(res1) : sample size must be between 3 and 5000
```

```
        studentized Breusch-Pagan test

data:  m1
BP = 1993.3, df = 20, p-value < 2.2e-16
```

*The residual plot displays megaphone shape which indicates the need for a curvilinear regression function;

*Some outliers are found in the residual plot;

*Shapiro Wilk normality test failed because of data size;

*Normal Probability plot shows the error term distribution is symmetrical with heavy tails;

*BP test also shows the error terms variance is not constant.

CHECK FOR MULTICOLLINEARITY

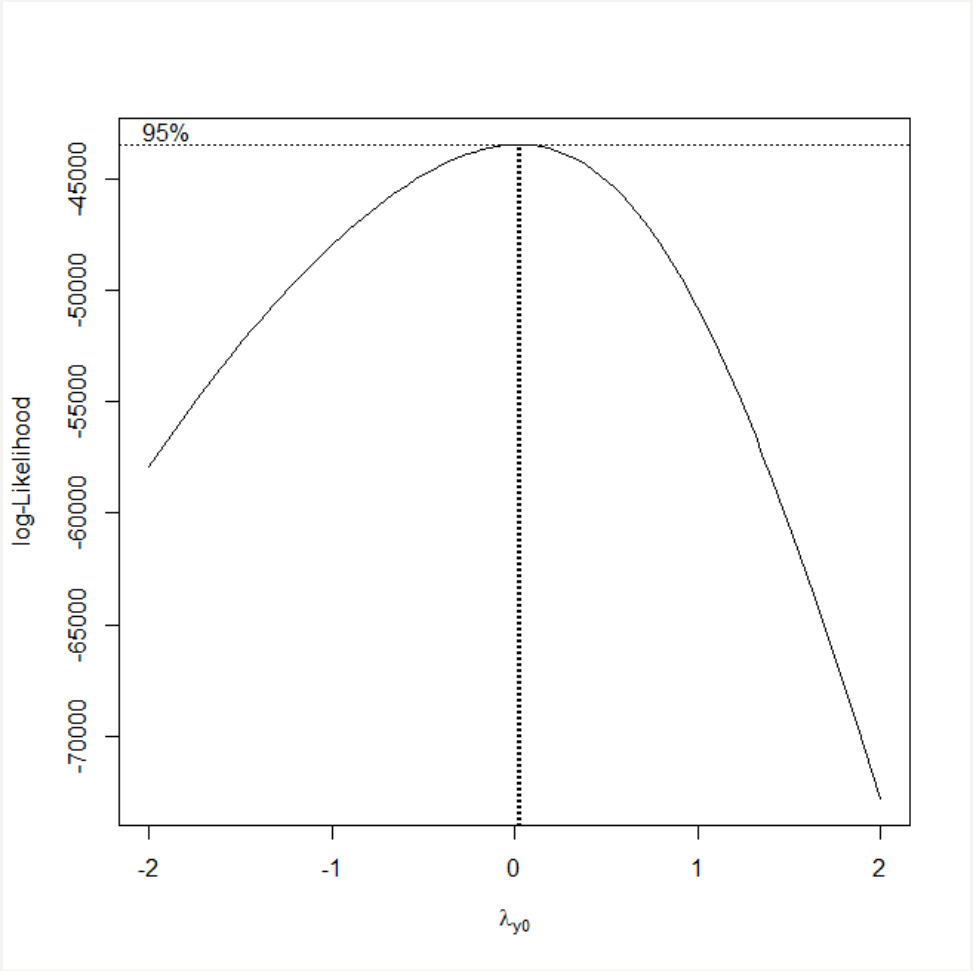
```
> vif(m1)
Error in vif.default(m1) : there are aliased coefficients in the model
```

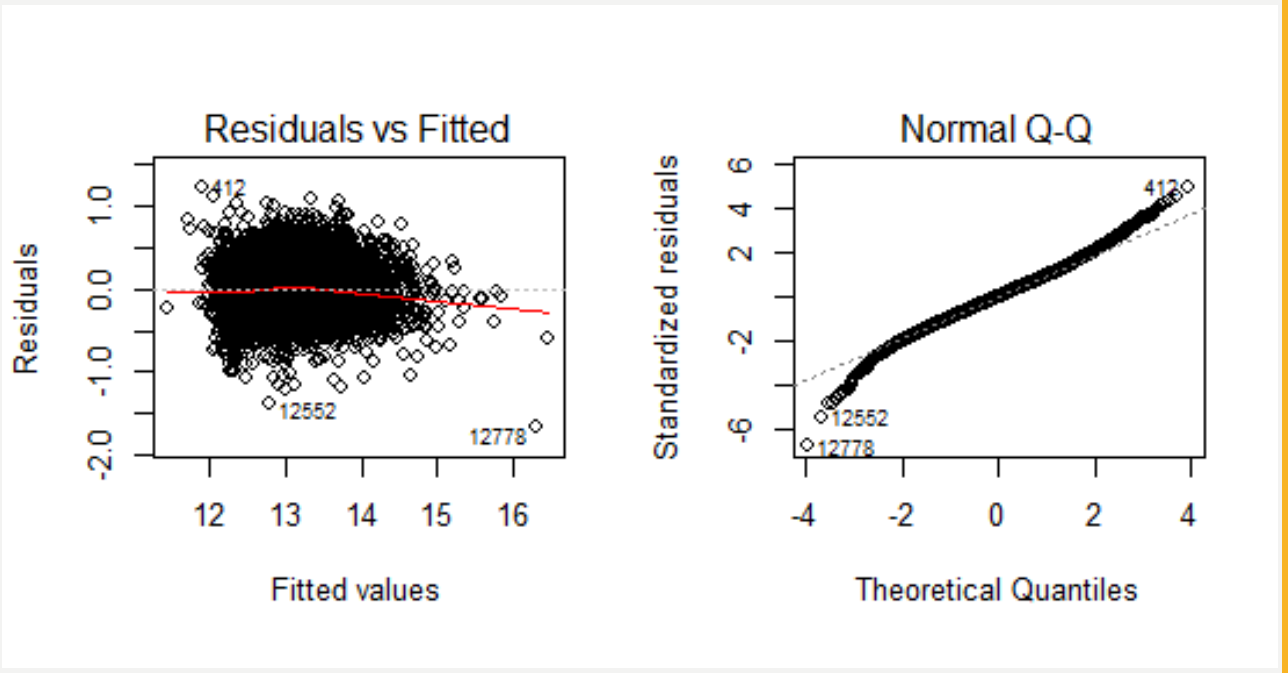
```
> vif(m12)
      x1       x2       x3       x4       x5       x6       x7       x8       x9      x10      x11
1.030384 1.006586 1.676313 3.306999 8.216722 2.026912 1.973160 1.199201 1.429005 1.212217 3.432794
     x12      x14      x15      x16      x17      x18      x19      x20
6.895501 2.282289 1.156915 1.652186 1.177158 1.828878 3.047269 2.092964
```

```
> vif(m13)
      x1       x2       x3       x4       x5       x6       x7       x8       x9      x10      x11
1.030202 1.006545 1.675572 3.231565 4.920241 2.023944 1.517816 1.196914 1.389905 1.204714 3.344222
     x14      x15      x16      x17      x18      x19      x20
2.282265 1.156709 1.651912 1.164065 1.782621 2.987139 2.092308
```

Error showed up indicating there are aliased coefficients in the model. Since X13 (Square feet above the ground) can be explained by square feet of living and grade, I decided to drop X13. Checking VIF values again, X5(Square feet of living) and X12(square feet above the ground) have the highest VIF values. X12 is removed because X5 has higher correlation with house price. After dropping X12, VIF values remain values between 1 and 5, which is acceptable in this project.

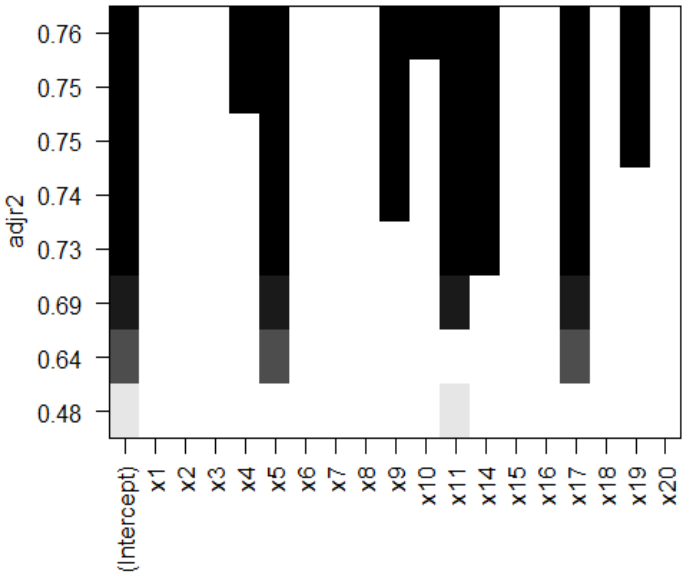
TRANSFORMATION

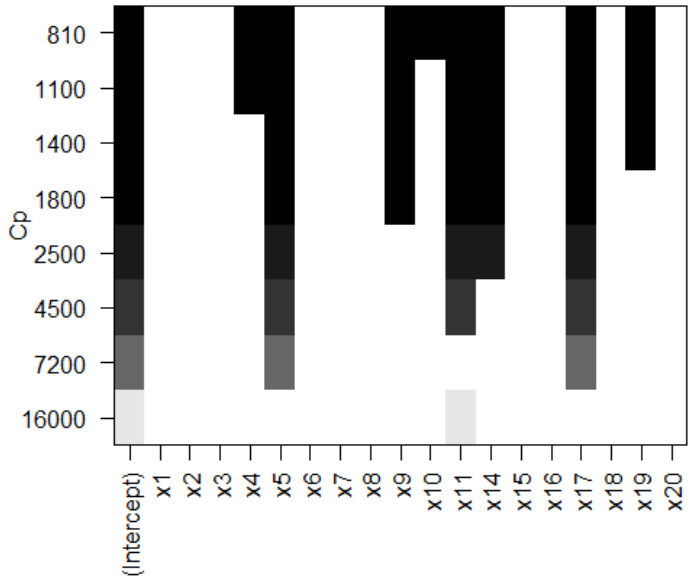


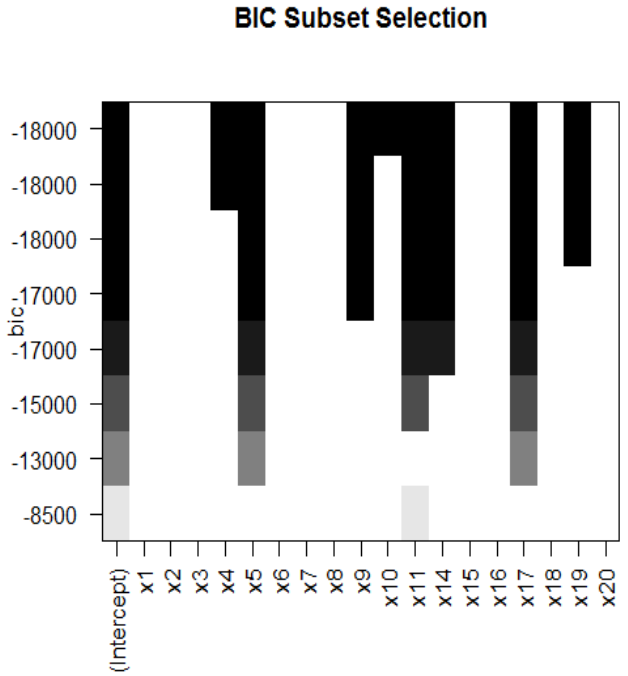


According to Box-Cox approach, logarithmic transformation Y' = log (Y) is the best to use. Heavy tailed issue is alleviated.

BEST SUBSET SELECTION







BEST SUBSET SELECTION

- Tentative Model 1:
- $\log(\hat{Y}) = \beta 0 + \beta 4X4 + \beta 5X5 + \beta 9X9 + \beta 10X10 + \beta 11X11 + \beta 14X14 + \beta 17X17 + \beta 19X19$

- Tentative Model 2:
- $\log(\hat{Y}) = \beta 0 + \beta 4X4 + \beta 5X5 + \beta 9X9 + \beta 11X11 + \beta 14X14 + \beta 17X17 + \beta 19X19$

- Tentative Model 3:
- $\log(\hat{Y}) = \beta 0 + \beta 5X5 + \beta 9X9 + \beta 11X11 + \beta 14X14 + \beta 17X17 + \beta 19X19$

BEST SUBSET SELECTION

	R_a^2	AIC	BIC	MSPR	MSE	MSPR-MSE
Model 1	0.7651	-23614.09	-23546.86	0.06500461	0.07	-0.004995
Model 2	0.7619	-23497.28	-23437.52	0.065904	0.07	-0.004096
Model 3	0.7563	-23297.61	-23245.32	0.06745934	0.07	-0.002541

Comparing R_a^2, AIC, BIC, MSPR, it can be concluded that Model 1 is the best model because it has the smallest values.

MODEL RESULT& THOUGHTS

Price = e ^ (-12.92 + 1.17 (# of bathrooms) + 0.025(Square footage of Living Space)+ 1.057(View)+ 0.758(Condition)+0.628(Grade)-0.162(Year Built)+1.364(Latitude)+0.059(Average house square footage of the 15 closest neighbors))

In King County, Washington, house closer to the north (with higher latitude) is more expensive. The better condition a house holds, more expensive it will be. More views a house gets, more pricy it will be.

Ways to improve model:

- Missing important variables?
- Outliers issue?
- Constant variance issue?