MLR Lab Report II
Xinyi (Anny) Cui

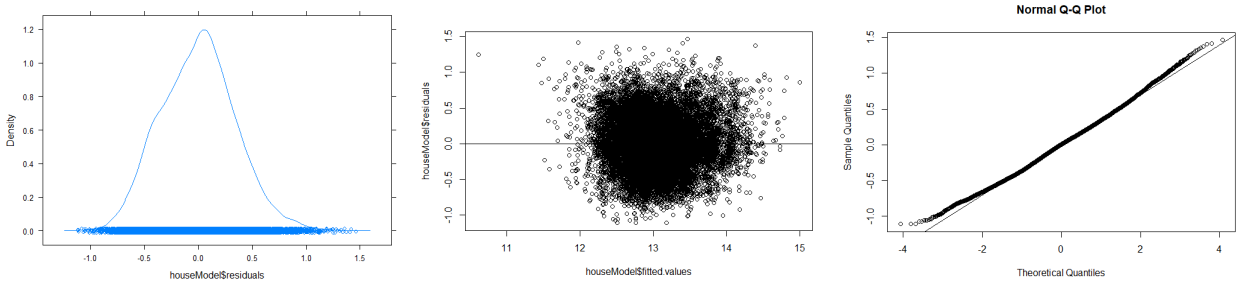
Dear Customer,

By collecting the data of houses sold between May 2014 and May 2015 in King County, Washington, I build a multiple linear regression model to evaluate your house price. There are 21 variables and we need select some useful ones as explanatory variables. After plotting their density plots, most variables display a kind of skewness, which means they need transformation. I want to do log transform to the data but try to avoid some zero values because log(0) provides no meaning. Also, I do not want to remove any data values if I can. Therefore, I decide to use squared footage of the apartments interior living space, which represents the space the buyers can really use for living. The condition of comments to see how physical goodness of the house.

The grade to see how good the building construction and design of the house is. The data of price and squared living need to be transformed so the form of the model should be $log(price) = \beta_0 + \beta_1 log(sqft\_living) + \beta_2 condition + \beta_3 grade + \epsilon$ and then the fitted log model is:

$$log(price) = 8.030256 + 0.400053 log(sqft_living) + 0.100995 condition + 0.215844 grade$$

This model provides a p-value smaller than $2.2e-16$, which means there exist a real trend between the response variable and explanatory variables. The F-statistic is 9230 and is better than the constant one. The only proble is the R^2 value which is 0.5617. This means only 56.17% variation in logPrice can be explained by the multiple linear regression model. However, there are over 20000 observations in the dataset. It is hard to improve the value of coefficient of determination or we need to adjust selected explanatory variable. 0.5617 is a moderate value, not too weak to use. Therefore, the model is reasonable to use. The check on linearity of this model is also shown below:



The normal distribution in density plot, the zero mean value in the graph of fitted values v.s. residuals, and the linear trendline in normal quantile graph display the lineariry of this model. Then, we can check for the variance inflation factor, the VIF for logSqft_living is 2.255702, that of condition is 1.029695, and that of grade is 2.298585. These values are small and smaller than 5, but still larger than 1, so these variables are moderately correlated and none of them needs to be removed. Thus, we can convert coefficients back to get the price model:

$$price = 3072.528569 \times 1.491904 sqft\_living \times 1.106271^{condition} \times 1.240908^{grade}$$

By using this model, we can test with an example of a house with $sqft\_living = 2000$, $condition = 3$, $grade = 7$. Then, the fitted value of price is 394318.6, and it provides a prediction inverval of $[199057, 781118.5]$, which is rational based on the given dataset. Therefore, this model can be used for you to predict your house sold price.

Sincerely,
Anny