Linear Regression Analysis on House Price of Ames

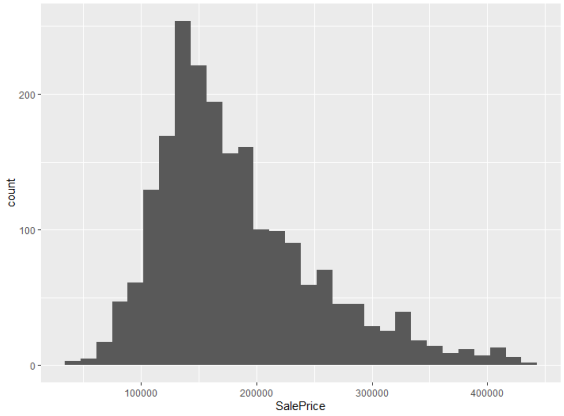
1. Introduction

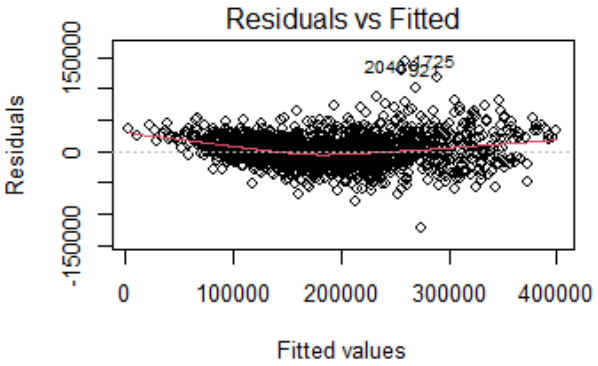
This report will show the process of linear regression modelling and the interpretation of the final model.

2. Modelling Process

2.1 Data Transformation

The density plot of SalePrice is right-skewed and also if I do the linear regression the Residuals vs Fitted plot shows non-constant variance. Thus, I take log transformation in response SalePrice to eliminate the heteroscedasticity.





Total 28 explanatory variables can be used to predict the house prices. I first do the data transformation for different type of variables.

According to variable descriptions, total 16 qualitative variables can be transformed into factors, which can be further verified by summary function.

There are 4 variables are related to rates or measures but cannot be divided into groups of same values. Their orders matter but values cannot be split into certain levels. Thus, they are not transformed into factors.

For other continuous quantitative variables, scatter plots and density plots or histograms are used to show whether it is necessary to do box-cox transformations. Doing log transformation in Age is more meaningful since we compare the percentage change in house price relative to percentage change in age. House age doesn't matter in units since one house can have hundreds of life length, different from people. Therefore, it is better to take log transformation in the age of house.

Log transformation enlarges the differences between small values and heavily reduce the differences between large values. It makes huge differences between large values ambiguous. Square root transformation also does the same thing but remains the relative differences between large values. Scatterplots of LotFrontage, LotArea and LowQualSinSF shows that most of values are around certain values with quite few huge values of huge difference between most of small values. To eliminate the effects caused by these huge values of small quantities, I use log transformation on these variables. In contrast, scatterplots of GrLivArea, BaseLivArea, and GarageArea also shows there are large differences between small values and large values but with larger amounts in both groups, that is to say, the large differences here would not affect the coefficients because there are also many small values and similar number of large values, different from the before case of thousands of small values and few large values. Therefore, square root transformation is preferred. I use log to eliminate the leverage effects of outliers in LotFrontage, LotArea and LowQualSinSF, and use square root transformation to reduce the units of range and make the coefficients more practical significant in GrLivArea, BaseLivArea, and GarageArea.

Table below shows the transformation done in variables.

Response	Log Transformation		SalePrice
Predictors	Factors		Location, Amenities, RoadRail, TwoStory_dum, FlatContour_dum, FlatRoof_dum, Garage_dum, CentralAirNum, KitchenQual_Ex, Zoning_2, Zoning_3, Zoning_4, YrSold_2007, YrSold_2008, YrSold_2009, YrSold_2010
	No change		BedroomAbvGr, Bathrooms, OverallCond, OverallQual, Fireplaces
	Transformation	log	Age, LotFrontage, LotArea, LowQualFinSF
		sqrt	GrLivArea, BaseLivArea, GarageArea

2.2 Collinearity Detection in Variables

To avoid collinearity in model, I first check the correlation coefficients between variables. Since all correlation coefficients are below 0.7, we need further check in the regression model.

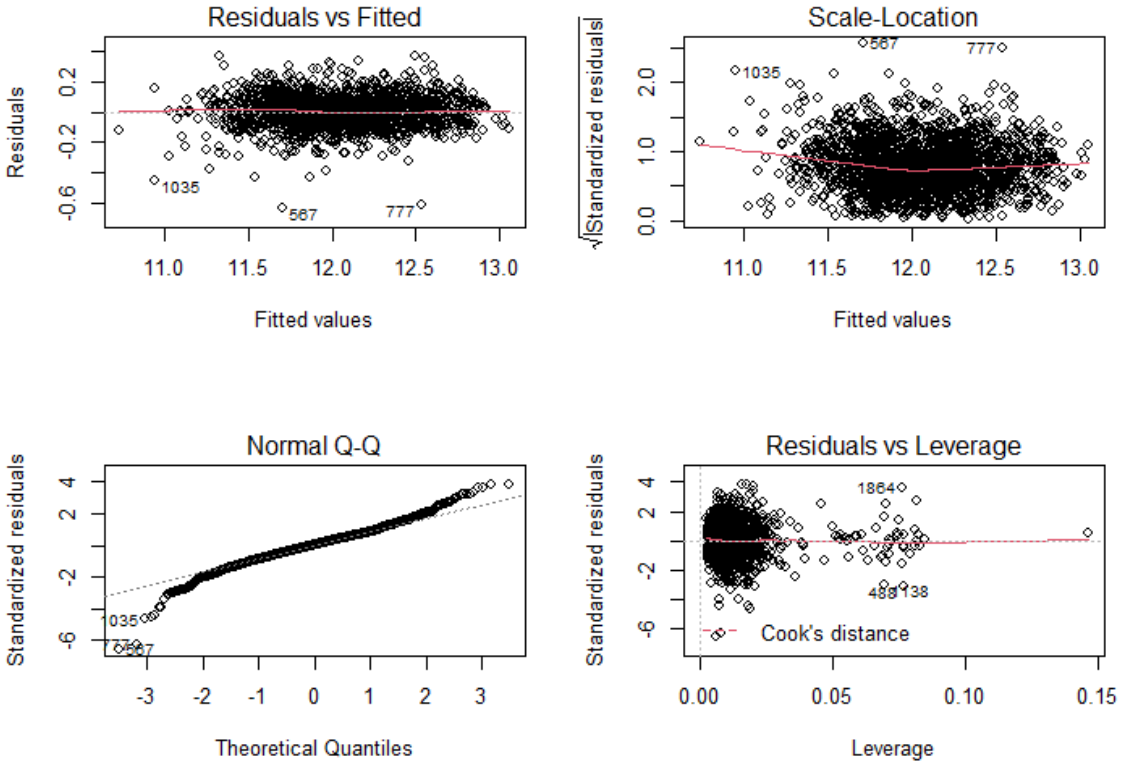
2.3 Model Determination

I use the backward selection to determine the model by first regressing all explanatory variables and then deleting variables of highest p-values until all variables are significant.

The final linear regression model is as below.

$$\begin{aligned}\log(SalePrice) = {} & \log(Age) + sqrt(GrLivArea) + sqrt(BaseLivArea) + Location + RoadRail \\ & + BedroomAbvGr + OverallCond + OverallQual + \log(LotArea) \\ & + TwoStory\_dum + FlatContour\_dum + FlatRoof\_dum + sqrt(GarageArea) \\ & + CentralAirNum + \log(LowQualFinSF) + Fireplaces + KitchenQual\_Ex \\ & + Zoning\_2 + Zoning\_3 + Zoning\_4 + YrSold\_2008 + YrSold\_2010\end{aligned}$$

F statistic is significant and also for t statistics of each variable. Thus, predictors are related to response and there is no collinearity. Adjusted R-squared of 0.9292 shows that the model captures most of the characteristics of sale price and 92.92% of variation in log of sale price can be explained by the model.



The almost horizontal line in Residuals vs Fitted plot shows that there is not obvious trend in residuals thus the variance is constant. The Normal Q-Q plot shows that the residuals have a heavier tailed distribution in small values but it does not cause problems since we have a large sample size. Also, there are no outliers with high leverage in Residuals vs Leverage plot.

3. Interpretation

$$\begin{aligned}\log(SalePrice) = {} & 9.5193137 - 0.0901882\log(Age) + 0.0253959sqrt(GrLivArea) \\ & + 0.0042050sqrt(BaseLivArea) + 0.0655204Location.L + 0.0078998Location.Q \\ & + 0.0031158Location.C - 0.0649119RoadRail - 0.0090965BedroomAbvGr \\ & + 0.0476878OverallCond + 0.0658852OverallQual + 0.0872371\log(LotArea) \\ & - 0.0494014TwoStory\_dum - 0.0264875FlatContour\_dum \\ & + 0.0801957FlatRoof\_dum + 0.0042992sqrt(GarageArea) \\ & + 0.0854610CentralAirNum - 0.0168232\log(LowQualFinSF) \\ & + 0.0345650Fireplaces + 0.0581327KitchenQual\_Ex - 0.0522310Zoning\_2 \\ & + 0.0409902Zoning\_3 + 0.0455631Zoning\_4 + 0.0161815YrSold\_2008 \\ & + 0.0316844YrSold\_2010\end{aligned}$$

Since I do log transformations in most of variables and others are of small values, the coefficients are all relatively small. Log can be interpreted as all else being same, one percent increase in age causes 0.0901882 percentage decrease in house price. Coefficients of factors represent the relative value added to the intercept. All else being the same, sale prices with excellent kitchen quality are 5.81327 percentage higher than those which do not have excellent quality.

4. Prediction

Predicting log transformed response uses following formula,

$$\hat{y} = \exp\left(\widehat{log\, y}\right) * \left(\frac{\sum_{i=1}^{n} \exp \widehat{u_i}}{n}\right),$$

where $\widehat{u_i}$ is the residual.