ST1131 Assignment II: Statistical Report

Name: Lai Zheyuan | Student ID: A0287898U

Background: In this report, the aim is to explore the dataset about the house selling price in Oregon and purpose a linear regression model for the response variable – house price, to discover which factor(s) may influent the price of a house.

Tools Used: Use **RStudio** to analyze the provided dataset *house_selling_prices_OR.csv*, fit the linear models, and plot scatterplots, boxplots, residual plots, and QQ plots.

Part I – Explore the Variables

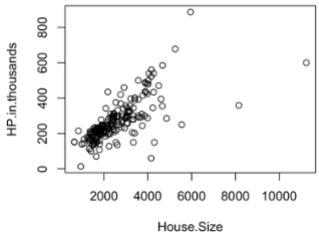
- Summarize the Variables
  - Quantitative Variables

Var/Stats	Min	1st Qu.	Median	Mean	3rd Qu.	Max
HP.in.thousand	13.0	200.0	242.5	267.5	300.0	887.2
House.Size	648	1710	2303	2551	3098	11239
Acres	0.000	0.130	0.190	0.533	0.320	5.820
Lot.Size	0	5663	8276	23217	13939	253519
Bedrooms	0.00	3.00	3.00	3.08	3.00	8.00
T.Bath	0.000	1.500	2.000	2.025	2.500	4.000
Age	1.00	11.75	34.00	34.75	51.00	107.00

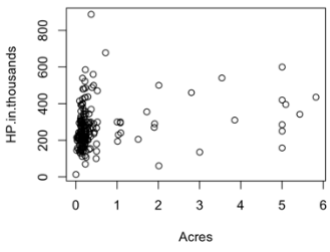
  Key findings: There are outliers in each variable.
  - Categorical Variables
    - *Garage*: (1=yes: 153; 0=no: 47)
    - *Condition*: (0=good: 174; 1=not good: 26)
    - *Age.Category*: (Old: 37; Medium: 72; New: 78; Very Old: 13)

- Check the association between response variable and one explanatory variable.

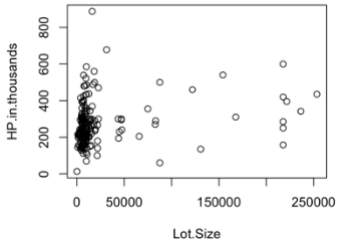
House size and house price



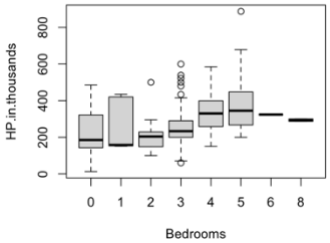
Acres and house price

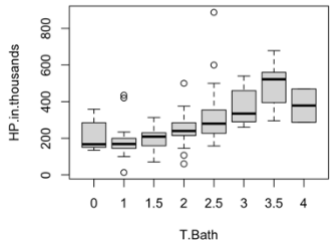
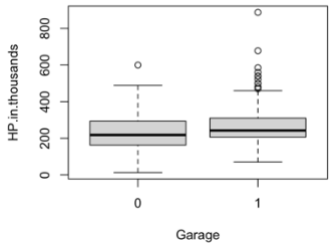
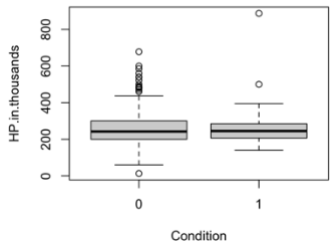
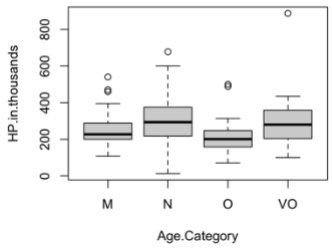


Lot size and house price



Bedrooms and house price



Bathrooms and house price	Garage and house price
	
Condition and house price	**Age category and house price**
	

Key findings: *House.Size* and *Age.Category* can a regressor; *Acres*, *Lot.Size*, *Garage*, *Condition*, and *Age* cannot be a regressor; *Bedrooms* and *T.Bath* may be a regressor.

Part II – Building Model

Model M1: consider *House.Size* and *Age.Category* as regressors, as purposed in Part I.
The model is $Price = 97.37 + 0.06 * size + 29.36 * I(Age.Category = N) - 19.84 * I(Age.Category = O) + 31.30 * I(Age.Category = VO)$.

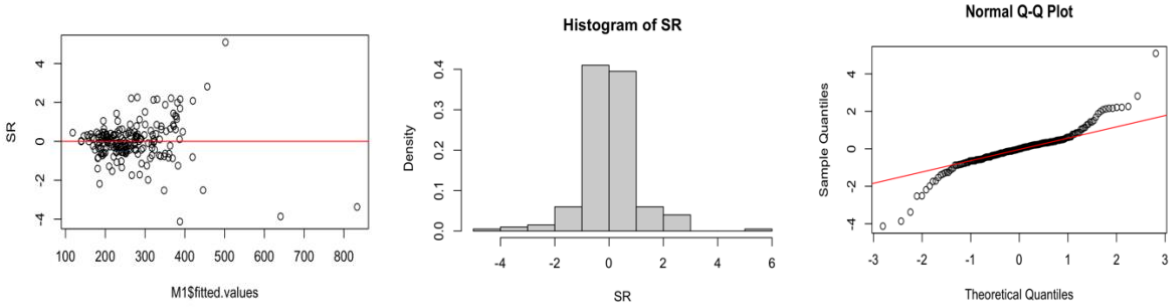
Result of t-test of each regressor:

- p-value of *House.Size* is very small => very significant
- p-value of *Age.CategoryN* = 0.0287 => not very significant
- p-value of *Age.CategoryO* = 0.22 => not significant
- p-value of *Age.CategoryVO* = 0.199 => not significant

So, the regressor *Age.Category* is not significant.

There are several outliers (index = 70, 98, 127, 130) with no influential points.

Check the residual plots:



From Plot 1, the points are not scatter randomly around 0, and some falls outside of (-3, 3). The slight funnel shape implies that the constant variance assumption is somewhat violated.
From Plot 2 and Plot 3, the standard residuals are not distributed normally.

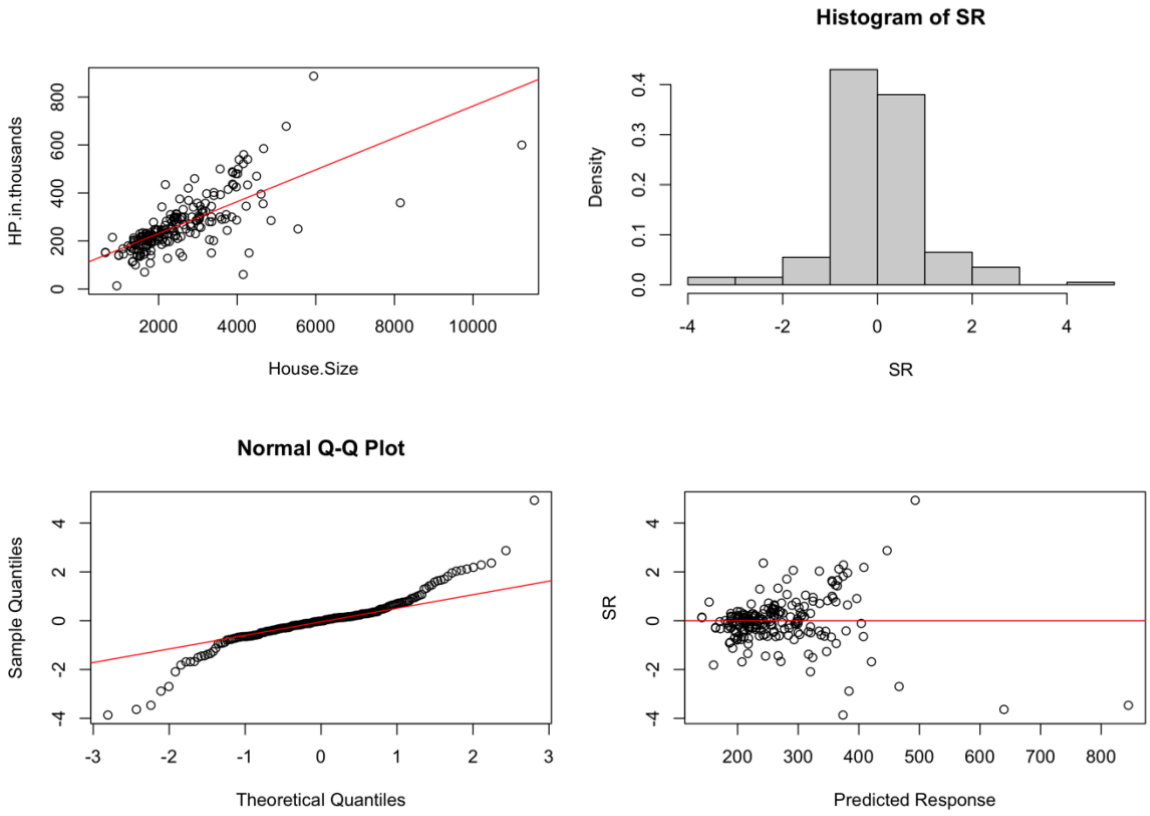
From the above evidence, model M1 is not adequate, we need to change the model by removing *Age.Category* variable, which is not significant from previous t-test.

Model M2: exclude *Age.Category*, use only *House.Size* as regressor.
The model is $Price = 97.997 + 0.066 * Size$

Regressor *House.Size* is significant given the small p-value.

There are several outliers (index = 70, 98, 127, 130) and one influential point (index = 127).

Check the scatterplot and residual plots:

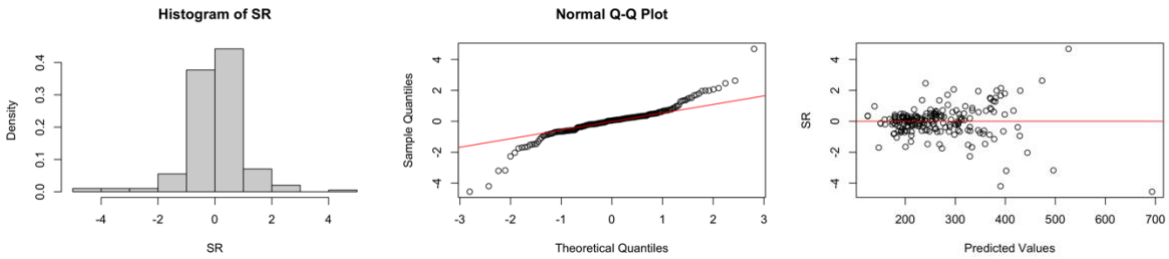


From the plots, we can conclude that the standard residuals are not normally distributed, and the funnel shape in scatterplot and SR against prediction imply non-constant variance.

From the above evidence, M2 is not adequate, need to remove the influential point (index=127).

Model M3: remove the influential point and rebuild the model using *House.Size* as regressor.
The model is $Price = 75.86 + 0.076 * Size$

Check the residual plots:



Results are similar to M2, SRs are not normally distributed, and variance is not constant. From the above evidence, M3 is not adequate, we need to add regressor(s).

Model M4: use *House.Size*, *Bedrooms* and *T.Bath* as regressors.
The model is $Price = 30.32 + 0.06 * Size - 8.54 * Bedrooms + 54.40 * T.Bath$

Result of t-test for each regressor:

- p-value of *House.Size* is very small => very significant
- p-value of *Bedrooms* = 0.1596 => not significant
- p-value of *T.Bath* is very small => very significant

From the above evidence, M4 is not adequate since regressor *Bedrooms* is not significant.

Model M5: Delete *Bedrooms*, use *House.Size* and *T.Bath* as regressors.
The model is $Price = 19.86 + 0.06 * Size + 46.66 * T.Bath$

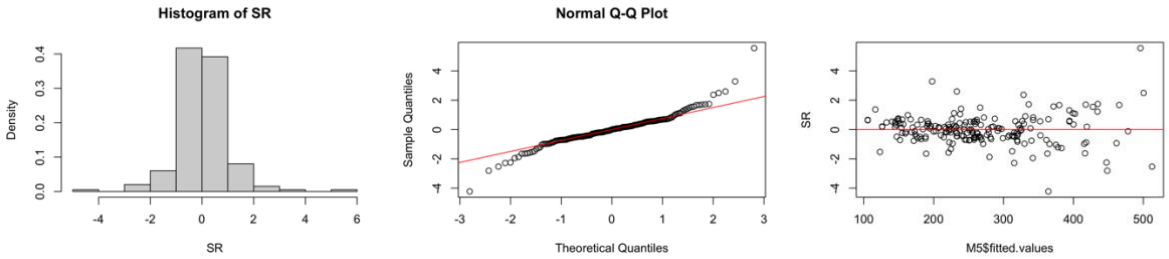
Result of t-test for each regressor:

- p-value of *House.Size* is very small => very significant
- p-value of *T.Bath* is very small => very significant

The regressors in this model are all significant.

There are several outliers (index = 70, 98, 171) with no influential points.

Check the residual plots:



From Plot 1 and Plot 2, standard residuals are nearly normally distributed despite some minor outliers, and Plot 3 implies that the constant variance assumption is satisfied.

Therefore, we conclude that Model M5 is adequate, so M5 is the final model.

Final Model: The final model considers *House.Size* and *T.Bath* as regressors and exclude influential point (index= 127), and the model can be represented by:

$$House.Price = 19.86 + 0.06 * House.Size + 46.66 * T.Bath$$

Model interpretation: Both increase in *House.Size* and *T.Bath* will cause the response variable *HP.in.thousand* to increase linearly.