Continuous Assessment: Multi Linear Regression

1st Aafaq Iqbal khan
National College of Ireland
x20108851

I. INTRODUCTION

In the continuous assessment of the Statistics of Data Analytics we estimated a multiple linear regression model to facilitate the understanding of the relationships between house characteristics and sale prices and used for the prediction of sale prices. I will use R for the statistical analysis of the dataset as R has more applicability in industry, including advanced visualization, data gathering and wrangling tools. The objective of the project is to apply different multi linear regression models to predict the sales price of the houses by determine appropriate parameter selection. For evaluation, I will use R-Square/Adjusted R-Square, and Mean Square Error (MSE) as well as checking the assumptions validation.

II. RESEARCH QUESTION

To estimate a multiple linear regression model to predict the accurate price of the houses and satisfy the all linear regression assumptions.

III. DESCRIPTION OF THE DATA

To answer the research question, we have provided a dataset called HouseDetails.txt. The dataset contains 16 variables as columns, and 1728 observations as rows. The response variable is "price" which is a continuous numeric variables so we can safely use regression algorithm. In dataset, we have 10 numeric variables and 6 categorical or factor variables. The brief description of the variables given below:

- **price:** price (US dollars)
- **lotSize:** size of lot (acres)
- **age:** age of house (years)
- **landValue**: value of land (US dollars)
- **livingArea**: living are (square feet)
- **pctCollege**: percent of neighborhood that graduated college
- **bedrooms**: number of bedrooms
- **firplaces:** number of fireplaces
- **bathrooms**: number of bathrooms (half bathrooms have no shower or tub)
- **rooms**: number of rooms
- **heating**: type of heating system
- **fuel**: fuel used for heating
- **sewer**: type of sewer system
- **waterfront**: whether property includes waterfront
- **newConstruction**: whether the property is a new construction
- **centralAir**: whether the house has central air

A. *Data Processing*

The first step in any data processing is to check whether dataset has any missing values if exists then we have to deal with these missing values because missing values can reduce the accuracy of the predictive model significantly. The HouseDetails.txt is already a clean dataset that has not any missing values.

```
> sum(is.na(df))
[1] 0
>
```

Fig. 1: zero missing values

The four variables i.e. "bedrooms", "fireplaces", "bathrooms", and "rooms" are given as numeric datatypes in original dataset but when I explored the data values they have only definitive distinct values so it will be good if we transform them into factors or level then it will be easy for visualization and better interpretation from the graphs. But for linear regression I will convert back into numeric rather than use as factors.

```
> df$fireplaces <- as.factor(df$fireplaces)
> df$bedrooms <- as.factor(df$bedrooms)
> df$bathrooms <- as.factor(df$bathrooms)
> df$rooms <- as.factor(df$rooms)
> str(df[c("bedrooms", "fireplaces", "bathrooms", "rooms")])
'data.frame':   1728 obs. of  4 variables:
 $ bedrooms  : Factor w/ 7 levels "1","2","3","4",..: 2 3 4 3 2 4 4 4 3 3 ...
 $ fireplaces: Factor w/ 5 levels "0","1","2","3",..: 2 1 2 2 1 2 2 2 1 1 ...
 $ bathrooms : Factor w/ 9 levels "0","1","1.5",..: 2 5 2 3 2 2 3 3 3 3 ...
 $ rooms     : Factor w/ 11 levels "2","3","4","5",..: 4 5 7 4 2 7 7 8 7 5 ...
```

Fig.2 Conversion to factors

Now find the 5 number summary of the numerical predictors only.

```
> summary(df[c("price","lotSize", "age","landValue", "livingArea","pctCollege")])
     price          lotSize            age           landValue      livingArea      pctCollege
 Min.   :  5000   Min.   : 0.0000   Min.   :  0.00   Min.   :   200   Min.   : 616   Min.   :20.00
 1st Qu.:145000   1st Qu.: 0.1700   1st Qu.: 13.00   1st Qu.: 15100   1st Qu.:1300   1st Qu.:52.00
 Median :189900   Median : 0.3700   Median : 19.00   Median : 25000   Median :1634   Median :57.00
 Mean   :211967   Mean   : 0.5002   Mean   : 27.92   Mean   : 34557   Mean   :1755   Mean   :55.57
 3rd Qu.:259000   3rd Qu.: 0.5400   3rd Qu.: 34.00   3rd Qu.: 40200   3rd Qu.:2138   3rd Qu.:64.00
 Max.   :775000   Max.   :12.2000   Max.   :225.00   Max.   :412600   Max.   :5228   Max.   :82.00
```

Fig. 3 Summary of numerical predictors

We can see there are some odd observations in the dataset its better if we remove those at this movement before implementation the linear regression. Like there are 12 observations which have "none" sewer system according to given dataset, 1 observation with "*bathroom*" less than 1, and 2 observations with 0 acers in "*lotSize*" which are not practically feasible so I decided to remove them as they will have less effect on linear models (Fig. 4).

```
df= df[which(df$sewer!="none"),]   #12 observations
df= df[which(df$bathrooms>=1),]    # 1 observations
df= df[which(df$lotsize!=0),]      # 2 observations
```

Fig. 4

B. *Descriptive Statistics*

In this section, I will explain the descriptive statistics of response variable as well as all independent variables. Firstly, look at the distribution of the each variables. We want all variable should be normally distributed. In fig.5, the histogram of the *price*, *age* and *living area* show that they have positively skewed data with skewness value 1.57, 2.49 and 0.90 respectively, the price variable seems to have a good normal distribution with very few high value that might potential outliers. The *lotSize* variable has range from 0 to 12 but having around 95% of the data points within 0 to 3 range so it has a long tail at right with 7.18 skewness.

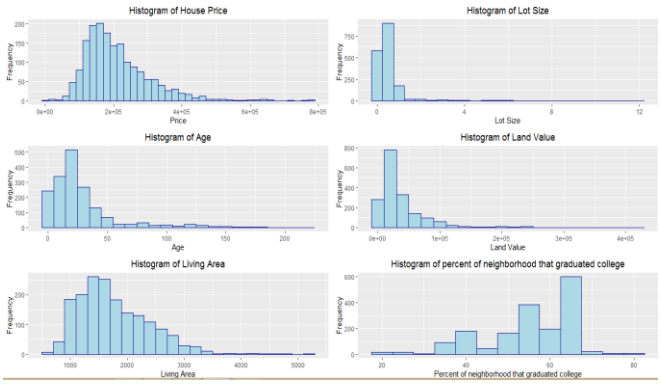


Fig.5 Histograms

The fig.6 shows the distribution of the data of *bedrooms, fireplace, bathrooms,* and *rooms'* variables. There is no issue with *rooms* distribution that having normal distribution. More of the houses have single fireplace rather than multiple fireplaces in the house.

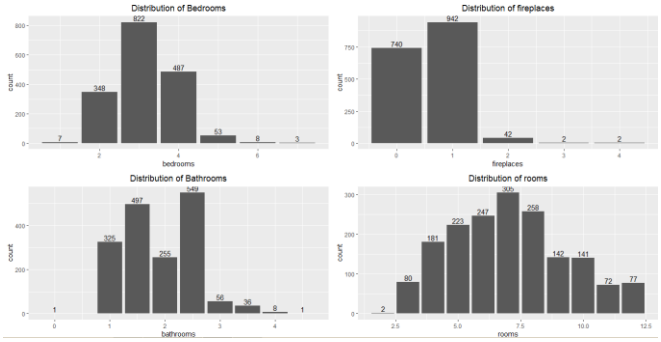


Fig.6

The fig. 7 shows the histograms of the factors variable in the dataset. "*Waterfront"* and "*newConstruction"* have imbalanced distribution of the levels. Whereas the rest of the factor variables are also not equally distributed.

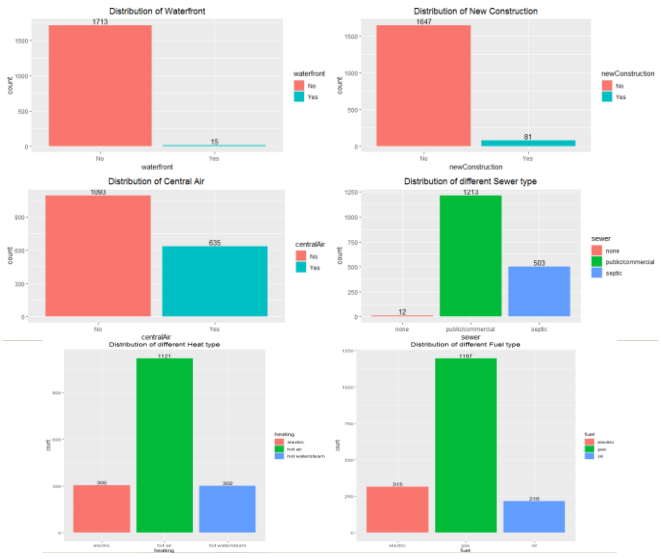


Fig. 7

Secondly, for better regression modelling we should know the relationship between the dependent variable and predictors. We can check the relationship by plotting a series of scatter plot of our response variable against each numeric predictors.

Now let us to plot scatter plot with trend line one by one to understand their relationship with dependent variable.

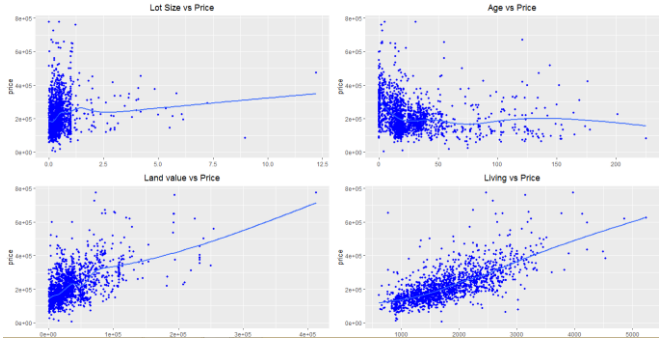


Fig. 8 Scatter plots having price on y-axis

There are linear trend lines in *livingArea and age* which will help us for better model prediction (Fig. 8). There might be a linear relationship in *lotSize* and *landValue* but not much visible in scatter plots due to the concentration of the data points on left side, we will find later correlation value among them for better understanding.

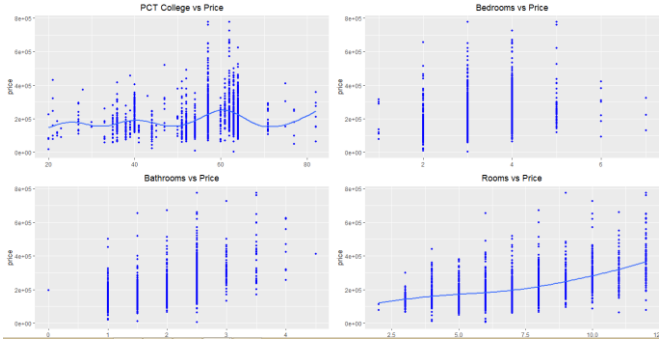


Fig. 9

To plot the *rooms, bedrooms* and *bathrooms* first we have to convert back them into numeric to plot their scatter plot. As there have only few distinct values so their plots don't have much relation or trend with *price* variable. We will see in linear modelling section whether they are significant to predict response variable or not.

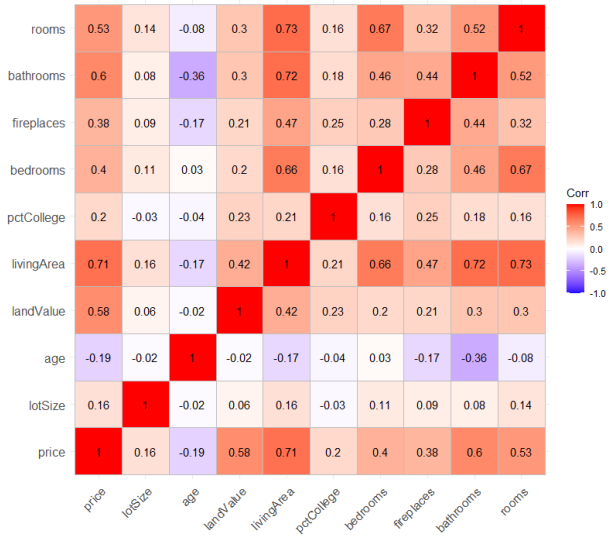


Fig. 10 correlation matrix

In correlation matrix (Fig. 10), we have 4 numeric predictors which have correlation with "*price*" greater than 0.5 that's good for our model and these correlation values also support our findings in scatter plots previously mentioned. If more of

the predictors have a strong correlation with response variable that will be good for better estimation and model will have greater predictive powers. However, we don't want any correlation between the predictors because it cause the problem of collinearity that violates our assumptions. In our dataset, there are few predictors that correlate to each other for example *livingArea*, *rooms*, and *bathrooms* have correlations greater than 0.6 that will cause a problem we will discuss more about it in modelling section.

IV. MODEL BUILDING, DIAGNOSTICS AND ASSUMPTION CHECKING

Linear regression is a powerful machine learning algorithm. It describe the linear relationship between dependent and independent variables. For this project we have more than one explanatory variables so we use multiple linear regression model. My aim for this project to build a final model (Best Linear Unbiased Estimator) that have only significant predictors in regression equation and satisfy every assumptions. In this section, I will get my final linear model iteratively and at each stage I will check the Gauss–Markov as well as additional assumptions.

A. Technique 1: Raw Model

Firstly, I included all the variables to get an idea about the model accuracy. Initially, we have to check whether any of the predictor variables have strong relationship with dependent variable or any of them contributing in the estimation of response variable or not, for this we look at F-statistics of the model.

H_0	The independent variables do not have an influence on the dependent variable
H_1	At least one the independent variable have a significant relationship to the dependent variable.

As, F-statistic of the model is 182.2 on 17 and 1695 DF and significant p-value = 2.2e-16, so we reject NULL hypothesis.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.231e+04  9.965e+03   1.235  0.21692
lotSize                 7.781e+03  2.248e+03   3.462  0.00055 ***
age                    -1.322e+02  5.844e+01  -2.262  0.02380 *
landValue               9.031e-01  4.956e-02  18.223  < 2e-16 ***
livingArea              7.032e+01  4.631e+00  15.184  < 2e-16 ***
pctCollege             -1.104e+02  1.520e+02  -0.726  0.46764
bedrooms               -7.629e+03  2.569e+03  -2.970  0.00302 **
fireplaces              1.549e+03  2.995e+03   0.517  0.60508
bathrooms               2.307e+04  3.392e+03   6.802 1.43e-11 ***
rooms                   2.966e+03  9.640e+02   3.076  0.00213 **
heatinghot air          7.942e+01  1.230e+04   0.006  0.99485
heatinghot water/steam -1.064e+04  1.283e+04  -0.829  0.40705
fuelgas                 1.096e+04  1.212e+04   0.904  0.36616
fueloil                 6.733e+03  1.288e+04   0.523  0.60117
sewerseptic             1.142e+03  3.684e+03   0.310  0.75667
waterfrontYes           1.200e+05  1.592e+04   7.535 7.90e-14 ***
newConstructionYes     -4.501e+04  7.314e+03  -6.154 9.38e-10 ***
centralAirYes           9.444e+03  3.488e+03   2.708  0.00684 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58180 on 1695 degrees of freedom
Multiple R-squared:  0.6464,    Adjusted R-squared:  0.6428
F-statistic: 182.2 on 17 and 1695 DF,  p-value: < 2.2e-16
```

Fig.11 summary of Raw Model

The summary of raw model shows *pctcollege*, *fireplaces*, *heating*, *fuel*, and *sewer* are not significant predictors based on their p-values. Adjusted R-squared is 0.6428 that means our independent variables successfully explained the 64.28% of variance in dependent variable.

B. Technique 2: Manually Iterative Model

In pervious technique, our raw model does not meet the all assumptions and also have non-significant regressors. Now I will add or remove predictor manually until I get a model with all possible significant predictors. But before the removal of the insignificant variables I will try to add few feasible interaction in the model because there is a possibility that those insignificant variables will become significant in the presence of the interactions. Based on the descriptive analysis of the data and some domain knowledge we can say there will be a possible interaction between *livingArea*, *landValue*, and *centralAir* as a house that has more living area is more likely to have more land value and possibly central air system. So I update the raw model by introducing the all possible interaction among these 3 variables.

```
Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                           2.868e+04  1.084e+04   2.646  0.00823 **
lotSize                               6.695e+03  2.226e+03   3.008  0.00267 **
age                                  -1.565e+02  5.783e+01  -2.706  0.00689 **
landValue                             1.238e+00  1.817e-01   6.814 1.32e-11 ***
livingArea                            6.189e+01  6.027e+00  10.269  < 2e-16 ***
pctCollege                           -1.697e+02  1.534e+02  -1.106  0.26877
bedrooms                             -7.779e+03  2.561e+03  -3.038  0.00242 **
fireplaces                            2.169e+03  2.963e+03   0.732  0.46425
bathrooms                             2.214e+04  3.356e+03   6.598 5.56e-11 ***
rooms                                 2.711e+03  9.497e+02   2.855  0.00436 **
heatinghot air                        2.370e+03  1.213e+04   0.195  0.84516
heatinghot water/steam               -5.854e+03  1.270e+04  -0.461  0.64492
fuelgas                               8.783e+03  1.193e+04   0.736  0.46186
fueloil                               4.646e+03  1.269e+04   0.366  0.71428
sewerseptic                           1.960e+03  3.650e+03   0.537  0.59134
waterfrontYes                         1.218e+05  1.571e+04   7.753 1.53e-14 ***
newConstructionYes                   -4.244e+04  7.270e+03  -5.837 6.36e-09 ***
centralAirYes                        -6.326e+04  1.375e+04  -4.600 4.53e-06 ***
landValue:livingArea                 -1.049e-04  8.711e-05  -1.204  0.22872
landValue:centralAirYes               8.563e-02  2.691e-01   0.318  0.75040
livingArea:centralAirYes              4.414e+01  7.224e+00   6.111 1.23e-09 ***
landValue:livingArea:centralAirYes   -1.471e-04  1.138e-04  -1.293  0.19630
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57260 on 1691 degrees of freedom
Multiple R-squared:  0.6582,    Adjusted R-squared:  0.654
F-statistic: 155.1 on 21 and 1691 DF,  p-value: < 2.2e-16
```

Fig. 12 Summary of raw model with interactions

Now I have 12 significant predictors instead of 10 in the case of model without interaction and my adjusted R-Squared also increased to 0.654 that means this model is more accurate to estimate the target variable.

Now we eliminated those predictors that have largest p-values one by one until we have all significant variables in our model. Initially, eliminated the all least significant interactions except "*livingArea:centralAir*" that have significant p-value. Our model still have few independent variables that are not contributing much in the estimation of the response variable so in each iteration I removed the variable with largest p-value I eliminated "*sewer*", "*heating*", "*fireplaces*", "*pctCollege*", and "*fuel*" in sequential order. After that I have a model with all significant predictors.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.908e+04  6.996e+03   4.156 3.40e-05 ***
lotSize                 6.406e+03  2.038e+03   3.144  0.00170 **
age                    -1.411e+02  5.466e+01  -2.580  0.00995 **
landValue               8.609e-01  4.821e-02  17.857  < 2e-16 ***
livingArea              5.910e+01  4.929e+00  11.990  < 2e-16 ***
bedrooms               -6.294e+03  2.520e+03  -2.498  0.01259 *
bathrooms               2.332e+04  3.300e+03   7.066 2.31e-12 ***
rooms                   2.841e+03  9.562e+02   2.972  0.00300 **
waterfrontYes           1.216e+05  1.563e+04   7.782 1.23e-14 ***
newConstructionYes     -4.045e+04  7.067e+03  -5.724 1.22e-08 ***
centralAirYes          -3.723e+04  9.334e+03  -3.988 6.94e-05 ***
livingArea:centralAirYes  2.758e+01  4.847e+00   5.689 1.50e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57750 on 1701 degrees of freedom
Multiple R-squared:  0.6503,    Adjusted R-squared:  0.6481
F-statistic: 287.6 on 11 and 1701 DF,  p-value: < 2.2e-16
```

Fig. 13 Summary of model

Final equation of the raw linear model is

$$\hat{y} = 29076 + 6406.08 * lotSize - 141.05 * age + 0.860933 * landValue + 59.1 * livingArea - 6293 * bedrooms + 23322.61 * bathrooms + 2841.24 * rooms + 121619.4 * waterfrontYes - 40453.84 * newConstructionYes - 37228.95 * centralAirYes + 27.5768 * livingArea{:} centralAirYes$$

For diagnostic and Assumptions Validation we have to satisfy the following assumptions:

1. Linearity:
   - Check: residuals vs fitted values plot. There should not be a relation
   - Solution: transformation
2. Homoscedasticity
   - Check: standardized residuals vs fitted values and ncv Test.
   - Solution: transformation of response variable.
3. Independence of Errors:
   - Check: Durbin–Watson statistic
4. Errors are normally distributed
   - Check: Normal Q-Q plot.
   - Solution: transformation or logged values.
5. Absence of Multicollinearity:
   - Check: VIF test
   - Solution: drop or new combination of variables
6. No influential data points:
   - Check: residuals vs leverage (cook distance)
   - Solution: drop influential data.

Manual Iterative Model Adequacy Check

To check whether our manual iterative model fulfils the assumption or not, we plot the diagnostics plots.

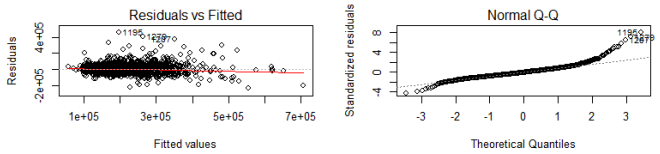


Fig. 14

In fig. 14, the left plot does not seem to have any relation just having few random noise so linearity assumption is valid. However, at the right side normal Q-Q plot shows the residuals are not normally distributed so we can assume that our model will not work well when apply to the population. The possible solution will be a transformation that I will check in next section. The normal Q-Q plot also indicates the potentials outliers with above absolute value of 4.

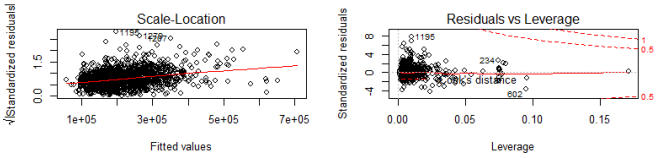


Fig. 15

However, the residual vs leverage plot in fig. 15 shows that there are no influential points since no values exceed 1 so no influential data point's assumption is satisfied.

Whereas, the plot at the left in fig.15 shows that homoscedasticity assumption is not to be going valid because there is some heteroscedasticity problem as there is an upward trend line in plot.

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 418.041, Df = 1, p = < 2.22e-16
```

For confirmation I performed NCV Test and got significant p-value (2.22e-16) that suggest we have not met the constant variance assumption. The proposed solution is transformation of independent variable that I will do in next section.

To check the independence of the error, I performed Durbin–Watson statistic test. There should not any autocorrelation between the errors. Durbin–Watson test informs whether the assumption of independent errors is valid or not.

H_0	There is no autocorrelation among residuals
H_1	Residuals are auto-correlated.

```
> durbinWatsonTest(lm8)
 lag Autocorrelation D-W Statistic p-value
   1       0.1785881      1.641954       0
 Alternative hypothesis: rho != 0
```

Usually, Durbin Watson statistic near to 2 is considered better and for our model it is 1.6415 which is around 2 but look at the p-value we have significant p-value so we have to reject Null hypothesis meaning that errors are dependent. For multicollinearity assumption, a formal test variance inflation factor (VIF) test is preferable. It measures how much common variance exists between predictor and other variables. Ideally, we want zero multicollinearity it means VIF value is 1.

```
    lotSize                 age           landvalue
   1.040956            1.311307            1.361346
livingArea            bedrooms           bathrooms
   4.798505            2.188224            2.415789
       rooms          waterfront    newConstruction
   2.522299            1.017058            1.155686
centralAir livingArea:centralAir
  10.399712           12.884567
```

Fig. 15 VIF test results

Our predictors "*centralAir"* and "*livingArea:centralAir"* have VIF value is greater than 10 which is not suitable meaning that collinearity exists between them so best solution is to drop them. The "*bathrooms", "bedrooms"*, and "*rooms"* have VIF about 2 which is somehow acceptable. This correlation is expected since few of dependent variables are some form of room.

To resolve the above issues, I tried to use two different transformations log, square and root but before that we know there is multicollinearity in errors due the 2 predictors "*centralAir*" and "*livingArea:centralAir*" which have VIF score greater than 10 so we dropped them. The log transformation of "*price*" increased the normality in error but decrease the Adj.R^2. The square root transformation did not improve anything it almost same as the model without any transformation.

	Adj. R2	Linearity	Homoscedasticity	Independence of Error	Normally distributed	Multicollinearity	No influential data points
Manually Iterative Model	0.648	Pass	Fail	Fail	Fail	Fail	Pass
Log(price)	0.571	Pass	Fail	Fail	Pass	Pass	Pass
Sqrt(price)	0.636	Pass	Fail	Fail	Fail	Pass	Pass

Table 1: Summary of Manually Iterative Models

C. *Technique 3: Final Model Selection*

In our manual iterative models there are some issues regarding assumptions as stated above I tried to relsove them by doing different transformations and dropping problematic predictors but still there was not a significant improvemment. Now I will go with automated techniques which automatically select the most appropiate predictors automatically according to pre-defined critera. I used backward elimation technique with Bayesian information criterion (BIC) method. BIC compares different possible models and find the best fit model. The model which was selected by BIC is one that explains the maximum variance in the response variable with minimum possible predictors. As there is a chance that our model can be overfitted so I splited data into training and testing data with 80-20 ratio. I will validate each model with test data based on R^2.

```
> n = length(resid(new))
> new_bic = step(new, direction = "backward",k = log(n), trace = 0)
> names(coef(new_bic))[-1]
[1] "lotSize"            "age"                "landValue"
[4] "livingArea"         "bathrooms"          "waterfrontYes"
[7] "newConstructionYes" "centralAirYes"
```

Fig. 16 possible predictors

The backward elimination method eliminate the insignificant predictors and only left important predictors *"lotSize", "age", "landValue", "livingArea", "bathrooms", "waterfrontYes", "newConstructionYes"*, and *"centralAirYes"*.

The F-statistic is 319 with p-value 2.2e-16 of the model with theses predictors and all independent variables have significant p-values. The adjusted R-squared is 0.6502. Now for assumption diagnostic of the model I will perform different tests and draw plots.

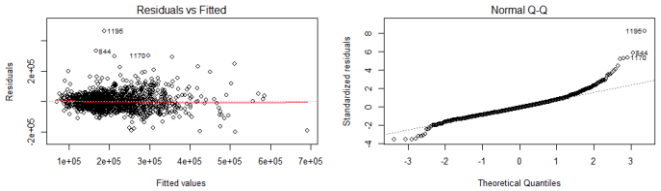
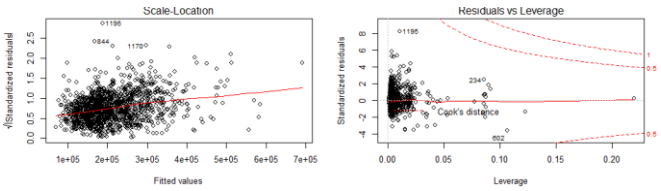


Fig. 17

Again linearity assumption is valid, fitted values and residuals have horizontal straight line but in Normal Q-Q plot there is not an exact 45 degree line although plot shows that model is more close to normal distribution than our previous models but we cannot say our model is general and can be apply perfectly on population.




```
> ncvTest(model_simple)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 224.0519, Df = 1, p = < 2.22e-16
```

Fig. 18

As the graph at the left side have an upward trend line so it means model has heteroscedasticity problem we need to resolve this otherwise standard errors are biased downwards. The NCV test also support this argument as Null hypothesis is rejected so we do not have constant variance in errors. In residual vs leverage plot, no points have a cook distance more than 1 so no influence data point's assumption is valid.

```
> durbinWatsonTest(model_simple)
 lag Autocorrelation D-W Statistic p-value
   1    -0.007794246       2.000362   0.996
 Alternative hypothesis: rho != 0
```

Fig 19 Durbin Watson Test

Again we used Durbin Watson test to check the independence of the error and to detect any autocorrelation between the errors. In test, we want non-significant p-value with D-W statistic near to 2. The model successfully met the independence of the error assumption.

```
vif(model_simple)
      lotSize           age      landValue     livingArea
     1.032912      1.256686      1.322365       2.482665
    bathrooms    waterfront newConstruction     centralAir
     2.414073      1.010383      1.164596       1.232873
```

Fig. 20 VIF test

A formal test, variance inflation factor (VIF) is performed to detect any multicollinearity between a predictor and other variables. Fig. 20 shows multicollinearity assumption is valid.

This model is better because it met most of the assumptions but still homoscedasticity and normally distributed are not valid yet, so I tried two transformation on response variable one is log and other is square root. After careful examination, there is no improvement in model in fact adjusted R-squared slightly decreased. Table. 2 shows the summary.

	Adj. R2	Linearity	Homoscedasticity	Independence of Error	Normally distributed	Multicollinearity	No influential data points
Simple (price)	0.650	Pass	Fail	Pass	Fail	Pass	Pass
Log(price)	0.586	Pass	Fail	Pass	Fail	Pass	Pass
Sqrt(price)	0.643	Pass	Fail	Pass	Fail	Pass	Pass

Table 2: Summary of models

To resolve the homoscedasticity and normally distribution of error violations, we have to consider and think at all the previous models what are the main causes of these violation. In Normal Q-Q plot of each model there were always some extreme data points of errors and due to their presence our models do not have normal distribution of errors even tried different transformation so we have to remove those points that cause these extreme standardized residuals. To identify these point, I used cook distances of each point to detect them and remove them from training dataset. These are very few data points and the removal of them did not affect much on the size of the data set. So, I removed these extreme point and use this approach on each model i.e. simple, log transformation and squared root transformation, among them square root transformation performed well having highest adjusted R^2 0.717 on training dataset and met the all linear regression assumptions.

I quickly discussed the assumption diagnostic analysis of my final model that is with squared root transformation. The square root transformation is only applied on response variable while all the predictors are in original form.

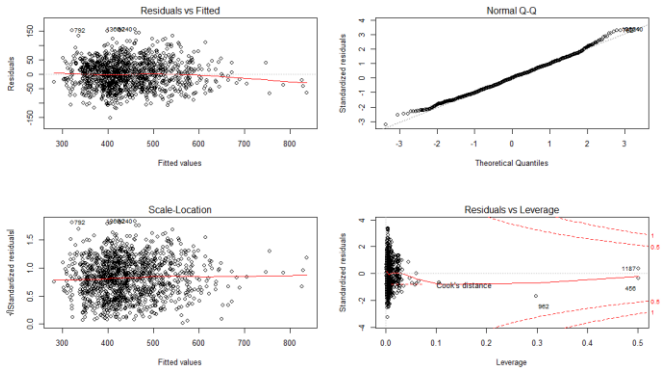


Fig. 21

There is no issue with linearity, and the normal Q-Q plot shows a straight diagonal curve meaning that errors are normally distributed. There is no influence point as all have cook distance less than 1 and no issue with homoscedasticity as standardized residuals does not have any trend, for confirmation I checked NCV test that resulted non-significant p-value. For the validation of independence of errors, Durbin Watson test gave non-significant p-value.

```
> ncvTest(model_sqrt_final)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.235613, Df = 1, p = 0.26632
> durbinWatsonTest(model_sqrt_final)
 lag Autocorrelation D-W Statistic p-value
   1     0.006548511      1.986238    0.79
 Alternative hypothesis: rho != 0
```

Fig. 22 NCV and Durbin Watson Test

V. SUMMARY

To recap, I started from the descriptive analysis of housing dataset, identified the distribution of the attributes and explored their relationship with the "*price*" by plotting scatter plots. Check the dataset for any missing value. In order to build an accurate and generalized predictive linear regression model all independents variables were considered with mainly two types of techniques i.e. custom manually iterative model and automatic predictor's selection technique backward elimination with BIC. In manual predictor's selection model, manually eliminate the predictors with insignificant p-values in each iteration and added few interaction terms based on graphical analysis and domain knowledge. The resultant model failed to meet the most of the assumptions so log and square root transformation are applied on response variable but still result was not satisfactory.

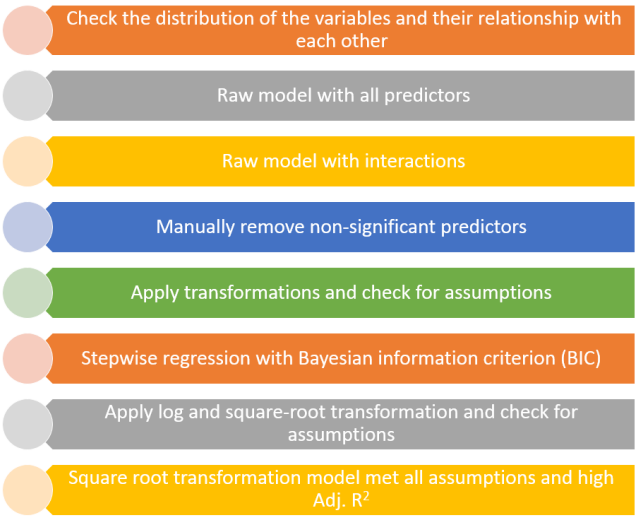


Fig. 22 Procedure Summary

In second approach, stepwise linear regression was used to identify the small set of important independent variables which are sufficient to explain the variance in the "*price*". As in stepwise regression, there is a chance of model overfitting so training-testing splitting was used to validate the model. Updated models were produced with only those predictors suggested by stepwise regression for each case i.e. *"price", "log(price)*", and "*sqrt(price)*". For the validation, these models were applied on test dataset to predict the price of the house and for evaluation R^2 parameter was used on test dataset. The results are summarized in table 3.

	Adj. R^2 (Training)	R^2 (Test Dataset)	Linearity	Homoscedasticity	Independence of Error	Normally distributed	Multicollinearity	No influential data points
Simple	0.718	0.596	Pass	Fail	Pass	Pass	Pass	Pass
log	0.686	0.522	Pass	Fail	Pass	Pass	Pass	Pass
sqrt	0.717	0.621	Pass	Pass	Pass	Pass	Pass	Pass

Table 3

Final equation of the linear model is

$$sqrt(\hat{y}) = 243.89 + 9.99 * lotSize - 0.308 * age + 0.000945 * landValue + 0.0721 * livingArea + 23.87 * bathrooms + 116.45 * waterfrontYes - 39.70 * newConstructionYes - 10.85 * centralAirYes$$

The final models suggested that square root transformation model performed well with high adjusted R^2 and met all the assumptions. The high R^2 on testing data suggests that the final model is not overfitting to the training dataset and generalizes well to an unseen dataset as well as can be applied on population as it meets the normal distribution assumption.

Although there is an ambiguity in linear regression equation that is if all predictors are zero then still the expected price of the house would be $ $(243.89)^2$. How a house can has nothing but still has price of $ $(243.89)^2$, in this condition, the intercept has no real meaning but is still important for price prediction. Ultimately, we determined the model that has high adjusted R^2 on training data, high R^2 on testing data, and satisfied all assumptions.

BIBLIOGRAPHY

[1] Cock, D. D. (2011). Ames, iowa: Alternative to the boston housing data as an end of semester regression project. Journal of Statistics Education, 19(3)

[2] Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL: http://CRAN.R-project.org/doc/Rnews/

[3] Thomas Lumley using Fortran code by Alan Miller (2009). leaps: regression subset selection. R package version 2.9. http://CRAN.R-project.org/package=leaps

[4] H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009