Abstract

The real estate market is very important for every country, not merely because it ensures construction structures and infrastructure necessary for life and work, but also because it has a strong, multiple impact on structures and the development of nation's entire economy.

Many real-estate professionals, home buyers are interested in determining how a set of characteristics such as the number of baths and bedrooms, total square feet, and others, affects the values of a ***property***.

The goal of the research is to define the factors that must be taken into account for an analytical forecast of trends in ***Real-estate*** prices.

Introduction

Property characteristics that affect value are called independent or ***predictor variables*** because they help predict what the property is worth. An estimated sales price is called the ***dependent variable*** because it depends on the predictor variables.

One statistical tool used to estimate value is called ***regression analysis***. In a regression model, sales price is "regressed" on a set of property characteristics to determine how much of the variation in the sales prices, are due to the variation in the set of property characteristics.

Regression Modeling Approaches

As Dependent variable is continuous ***Linear regression*** is the one among remaining all types regressions, suitable for Sales price prediction. ***Simple linear regression***, assesses the relationship or association between a single dependent variable, such as a sales price, and a single independent or predictor variable, such as square footage. ***Multiple linear regression*** assesses the relationship or association between a single dependent variable, such as sales price, and multiple independent or predictor variables, such as square footage, lot size and age of the property.

Although simple or ***Simple*** linear regression models are good because of their simplicity, they will not produce as accurate result as a ***multivariable*** regression analysis for a number of reasons. First, the selection of which variable(s) to use might leave out potentially important factors. Second, ***Simple*** regression does not control for the associations among the predictor variables themselves.

Multivariable models will automatically adjust for the interdependency of these two (or more) predictors while simple linear regression models will not. As a consequence, if you use two ***Simple*** models to develop your estimate, you would overestimate the sale price because you would be giving too much weight to the number of bedrooms within the context of the property's square footage. The degree of error introduced by adding together the results of ***Simple*** models increases as you consider more and more predictors. It is much easier to use a ***multivariable*** model to make the correct adjustments and estimate a value without having to do all of the mathematical work of combining separate estimate.

Data Interpretation:

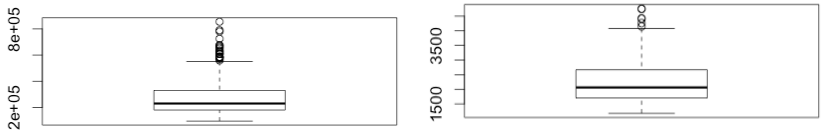
1. There are twelve variables including sales price.
2. Out of 11, four predictor variables are categorical variables. Pool, Quality, Air Conditioner, Adjacent to highway.
3. Year built given in 4-digit format converted to number years by subtracting it from current year, so that it will become a continuous variable to contribute in regression.

4. Data separated into two parts using excel random function.
5. One part contains 400 observations for development and other part has 124 observations for testing.

Variable Number	Variable Name	Description
1	Identification number	1–522
2	Sales price	Sales price of residence (dollars)
3	Finished square feet	Finished area of residence (square feet)
4	Number of bedrooms	Total number of bedrooms in residence
5	Number of bathrooms	Total number of bathrooms in residence
6	Air conditioning	Presence or absence of air conditioning: 1 if yes; 0 otherwise
7	Garage size	Number of cars that garage will hold
8	Pool	Presence or absence of swimming pool: 1 if yes; 0 otherwise
9	Year built	Year property was originally constructed
10	Quality	Index for quality of construction: 1 indicates high quality; 2 indicates medium quality; 3 indicates low quality
11	Style	Qualitative indicator of architectural style
12	Lot size	Lot size (square feet)
13	Adjacent to highway	Presence or absence of adjacency to highway: 1 if yes; 0 otherwise

Regression Analysis:

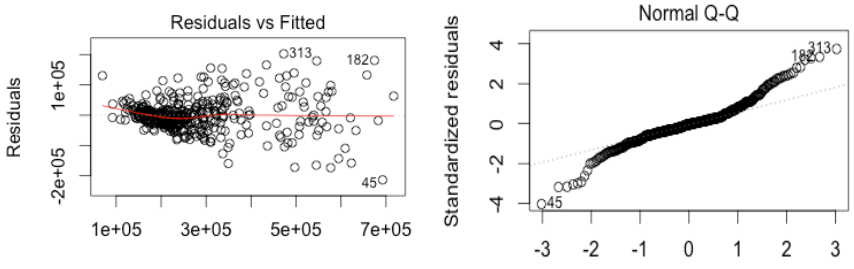
1. The first data analysis includes exploring data through appropriate graphs and calculating various summary statistics.
2. R studio used for analysis. R- Command for linear regression is : fit <- lm(y ~ x1 + x2+x3+ …xn , data =dataset) , Here y- Predictor variable and x1….xn are independent variables which are factors affecting the regression.



3. Fig shows boxplots for Sales price and Square feet variables. Outliers are visible in the plots, which might effect the regression. Finding Influential points and removing is required
4. First started with full model including all indicator variables. Threshold level for removing variables maintained at 0.05 all the time.
5. Bed, garage, pool, highway and AC variables that are not significant for 0.05 cutoff levels are removed one by one. Significance is measured by p values of t - test, which tells us likelihood
   H0: Variable not Significant and including it in regression will not have any affect.
   Ha: Variable is significant with some non-zero coefficient.
   For example, p- value < 0.05 for corresponding t values, we reject null. That means variable is significant.

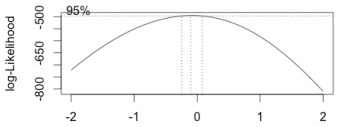
```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.919e+05  2.270e+04   8.454 5.64e-16 ***
WorkData$sqft                1.041e+02  7.690e+00  13.534  < 2e-16 ***
WorkData$bath                9.087e+03  4.322e+03   2.103   0.0361 *
WorkData$year               -1.337e+03  2.054e+02  -6.511 2.30e-10 ***
WorkData$style              -5.639e+03  1.426e+03  -3.955 9.10e-05 ***
factor(WorkData$quality)2 -1.318e+05  1.103e+04 -11.942  < 2e-16 ***
factor(WorkData$quality)3 -1.390e+05  1.466e+04  -9.484  < 2e-16 ***
WorkData$lot                 1.144e+00  2.463e-01   4.646 4.63e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55310 on 391 degrees of freedom
Multiple R-squared:  0.8381,    Adjusted R-squared:  0.8352
F-statistic: 289.1 on 7 and 391 DF,  p-value: < 2.2e-16
```



Transformation

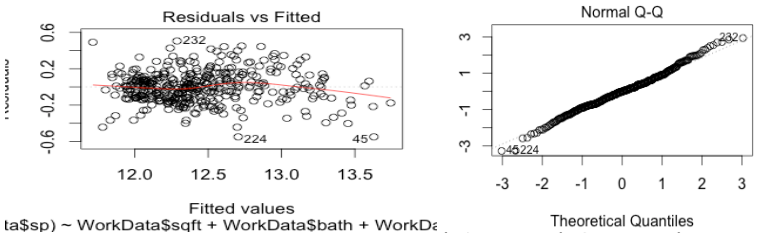
1. Though model suites well with low p-values and good R-square, above Plots show some Heteroscedasticity, and deviation from normality. Outlier presence also observed.
2. First to remove non-constant variance transformation procedure applied. Checked for Boxcox transformation
3. Observed plots after transformation. Results are better with Log transformation rather than power transformation.

i. 

4. Y variable transformed to Log(y) to remove Heteroscedasticity.
5. With Transformation P values improved and also Hetero-scadasticity effect reduced, so does the deviation in QQplot.

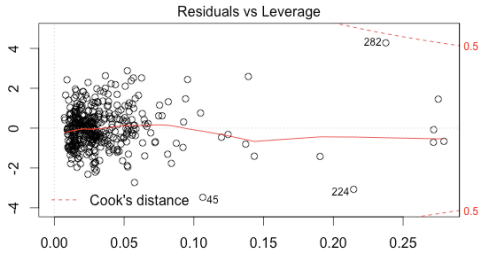
6.
```
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.202e+01  7.187e-02 167.208  < 2e-16 ***
WorkData$sqft                3.128e-04  2.434e-05  12.850  < 2e-16 ***
WorkData$bath                4.811e-02  1.368e-02   3.517 0.000488 ***
WorkData$year               -4.281e-03  6.502e-04  -6.584 1.48e-10 ***
WorkData$style              -1.214e-02  4.514e-03  -2.689 0.007481 **
factor(WorkData$quality)2 -2.702e-01  3.493e-02  -7.737 8.75e-14 ***
factor(WorkData$quality)3 -3.758e-01  4.640e-02  -8.097 7.20e-15 ***
WorkData$lot                 4.265e-06  7.796e-07   5.471 8.02e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1751 on 391 degrees of freedom
Multiple R-squared:  0.8359,    Adjusted R-squared:  0.833
F-statistic: 284.6 on 7 and 391 DF,  p-value: < 2.2e-16
```

7. 

ta$sp) ~ WorkData$sqft + WorkData$bath + WorkDa

8. From the Residual leverage diagram, influential points are clearly visible.



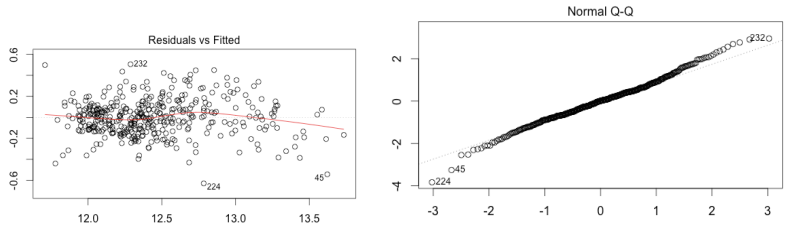
9. Finding the outliers are done by the fallowing command

```
identify(WorkData$sp ,cooks.distance(model ))
with(WorkData,plot(sp ,cooks.distance(model)))
```

10. Influenced outliers are removed and results observed.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.193e+01  6.719e-02 177.506  < 2e-16 ***
mydata$sqft              3.388e-04  2.280e-05  14.863  < 2e-16 ***
mydata$lot               3.384e-06  7.495e-07   4.516 8.45e-06 ***
mydata$bath              6.493e-02  1.313e-02   4.944 1.16e-06 ***
mydata$style            -1.572e-02  4.115e-03  -3.821 0.000155 ***
mydata$year             -4.963e-03  6.068e-04  -8.179 4.41e-15 ***
factor(mydata$quality)2 -2.255e-01  3.258e-02  -6.921 1.93e-11 ***
factor(mydata$quality)3 -2.881e-01  4.386e-02  -6.569 1.69e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1562 on 377 degrees of freedom
Multiple R-squared:  0.8661,    Adjusted R-squared:  0.8636
F-statistic: 348.4 on 7 and 377 DF,  p-value: < 2.2e-16
```



11. R^2 improved and Squared errors reduced.

Interaction Terms

12. After transformation model fitness improved. Interaction is another factor that needs to verified in the model Four interactions are tested with R step function which will quickly eliminate less effective terms. 4 interaction terms are tested which are Years , Style and Quality, Bed and Bath effects on sales price.

```
mydata$sqft              3.389e-04  2.332e-05  14.531  < 2e-16 ***
mydata$bed               6.613e-02  2.377e-02   2.782 0.005672 **
mydata$bath              1.213e-01  3.119e-02   3.890 0.000119 ***
mydata$year             -5.105e-03  6.153e-04  -8.297 2.00e-15 ***
mydata$style            -1.493e-02  4.058e-03  -3.679 0.000269 ***
factor(mydata$quality)2 -2.230e-01  3.253e-02  -6.856 2.97e-11 ***
factor(mydata$quality)3 -2.672e-01  4.427e-02  -6.037 3.81e-09 ***
mydata$lot               4.141e-06  7.740e-07   5.349 1.55e-07 ***
factor(mydata$pool)1     1.035e-01  3.185e-02   3.249 0.001266 **
factor(mydata$ac)1       1.882e-02  2.546e-02   0.739 0.460217
mydata$grg              -1.875e-03  1.519e-02  -0.123 0.901824
factor(mydata$highway)1 -1.742e-01  9.217e-02  -1.891 0.059463 .
mydata$bed:mydata$bath  -1.888e-02  7.953e-03  -2.374 0.018108 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1525 on 371 degrees of freedom
Multiple R-squared:  0.8745,    Adjusted R-squared:  0.8701
F-statistic: 198.8 on 13 and 371 DF,  p-value: < 2.2e-16
```

13.
14. Among all interaction terms, bedroom and bathroom interaction looks effective initially but when trying to reduce the model, again it became insignificant.
15. Finally, same model has been noticed significant. Nested F-test has been done with this model and Full model. Results given below

```
Response: log(mydata$sp)
                       Df Sum Sq Mean Sq  F value    Pr(>F)
mydata$sqft             1 51.258  51.258 2101.188 < 2.2e-16 ***
mydata$lot              1  0.309   0.309   12.646 0.0004242 ***
mydata$bath             1  2.229   2.229   91.351 < 2.2e-16 ***
mydata$style            1  1.163   1.163   47.655 2.164e-11 ***
mydata$year             1  3.326   3.326  136.325 < 2.2e-16 ***
factor(mydata$quality)  2  1.213   0.607   24.864 7.194e-11 ***
Residuals             377  9.197   0.024
---
```

16. Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
Analysis of Variance Table

Model 1: log(mydata$sp) ~ mydata$sqft + mydata$lot + mydata$bath + +mydata$style +
    mydata$year + factor(mydata$quality)
Model 2: log(mydata$sp) ~ mydata$sqft + mydata$lot + mydata$bed * mydata$bath +
    mydata$style + mydata$year + factor(mydata$quality) + mydata$grg +
    factor(mydata$pool) + factor(mydata$highway) + factor(mydata$ac)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    377 9.1969
2    371 8.6232  6   0.57372 4.1139 0.0005168 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

17. With F value 4.1139 with P-value 0.004669 this test shows model is valid when compared with full model.

18. **Final Model:** 1.193 + 3.388*10^-4*SQFT +3.384*10^-6 * LOT +6.493*10^-2 *BATH -1.572*10^-2*STYLE

-4.963 *10-3 *YEAR = LOG[E(Y)] – Quality :1

1.193 + 3.388*10^-4*SQFT +3.384*10^-6 * LOT +6.493*10^-2 *BATH -1.572*10^-2*STYLE

-4.963 *10-3 *YEAR -2.255 *10^-1 *2 = LOG[E(Y)] – Quality :2

1.193 + 3.388*10^-4*SQFT +3.384*10^-6 * LOT +6.493*10^-2 *BATH -1.572*10^-2*STYLE

-4.963 *10-3 *YEAR -2.881*10^-1 *2 = LOG[E(Y)] – Quality :3

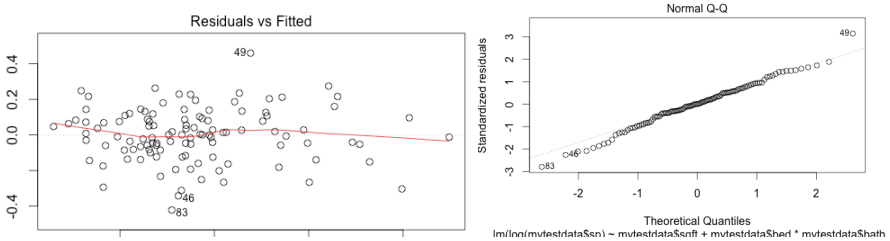
Verifying Model Validity:

- So far we build a model to predict sales price of a property with certain criteria. Now this model needs to be tested either real world data or with test data that has been separated from our complete dataset.

- Test data has ben loaded into R. Reduced model applied on test data. Some outlier presence detected in test data also. After removing outliers similarity between developed model and test model has been observed.

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(testintmodel)
Analysis of Variance Table

Response: log(mytestdata$sp)
                              Df  Sum Sq Mean Sq F value    Pr(>F)
mytestdata$sqft                1 15.3009 15.3009 575.688 < 2.2e-16 ***
mytestdata$bath                1  1.1980  1.1980  45.074 1.028e-09 ***
mytestdata$year                1  0.5976  0.5976  22.485 6.760e-06 ***
mytestdata$style               1  0.5432  0.5432  20.439 1.636e-05 ***
factor(mytestdata$quality)     2  0.9256  0.4628  17.412 3.002e-07 ***
mytestdata$lot                 1  0.2945  0.2945  11.080  0.001208 **
Residuals                    104  2.7642  0.0266
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Finally, with Predict function in R results validated for some values. d<- predict (final.model , newdata)

```
> d<- predict (final.model , newdata)
> d
       1        2        3        4        5        6        7        8        9
12.63796 12.50897 12.12554 12.34707 12.29767 13.16880 11.91488 11.96511 12.03785
      10       11       12       13       14       15       16       17       18
12.25021 12.51874 12.38900 12.26093 11.78116 12.06374 12.05455 12.42969 12.35750
```

The R out put for first point is exp(12.63796) = 308032.3 Actual value 259000.

Inferences:

The model can be used, for example, consider a home with the following features: 3408 square feet, lot size 29307, 4 bathrooms, 6 bedrooms, built in 1960, three-car garage average quality home will cost with in the same neighborhood predicted 308032.3 dollars.

Model	R^2	Residuals
Full model	0.8411	55080
Reduced model	0.8381	55310
After Transformation	0.8359	0.1751
After outlier Removal	0.8661	0.1562

Test Data	0.8772	0.163

Regression will help you identify the features for estimating the value for a given property, possibly including some you might not have thought were important but which turn out to be important. Very un-expecting result would be number bed rooms are not that contributing factor as we thought that would be. There is possibility that model might change because of some redundant data or outliers. And one more area needs to focus to make it good fit was multicollinearity

As we can see R^2 values sufficiently large, which will conclude the model will explain most part. According to the model sales price will not raise amenities like for AC, Pool, garage size, Bed room or if its near to the high way. They mostly look for how old the house will be and quality, style of the house and size (Lot size and Square feet). Bath has some effect on model.