Applied Regression Final Project

Elda Piedade, Thomas Oswald

May 4th, 2020

Introduction

The objective of this project is to find the best linear regression model to predict the median home value for the houses in the Houston neighborhoods. Data was gathered from 96 zip codes in Houston by utilizing python web scrapping resources to collect data from the Texas Hometown Locator website (owned by HTL, Inc.).With the dataset extracted and cleaned, exploratory data analysis and statistical analysis were performed to understand the relationship between the median home value and other variables, such as diversity index, per capita income, and average household size. Based on the analysis, data was modeled with linear regression.

Importance

1. With a good model for prediction and analysis, individuals in Houston will be able to understand how to price their homes for sale. 2. Understanding how the demographic factors relate to median home value is valuable social knowledge.

Data

Response variable : Median Home Value

Predictors:

* x_1 -Total Population

* x_2 - Diversity Index

* x_3 - Median Household Income

* x_4 - Per Capita Income

* x_5 - Total Housing Units

* x_6 - Average Household Size

* x_7 - Housing affordability Index

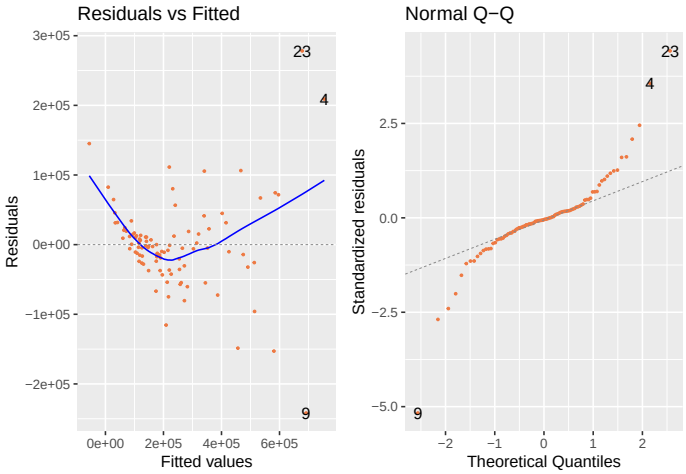
Explained variables:

1. The Diversity Index is a scale of 0 to 100 that represents the likelihood that two persons, chosen at random from the same area, belong to different races or ethnic groups. If an area's entire population belongs to one race and one ethnic group, then the area has zero diversity. An area's diversity index increases to 100 when the population is evenly divided into two or more race/ethnic groups. 2. The Housing Affordability Index base is 100 and represents a balance point where a resident with a median household income can normally qualify to purchase a median price home. Values above 100 indicate increased affordability, while values below 100 indicate decreased affordability. 3. The outcome variable is the median home value (median price home).

Data Loading & Checking full Model Accuracy:

After loeading the data and creating a full Linear Regression (LR) model, we found that the model is not adequate. The residuals have a tunel and bowl shape; the residuals have a heavily-tailed distribution, and the data hasthree possible influencial points. To address this problem we will inspect the data and perform a few transformations.

First model - Residual Plots:

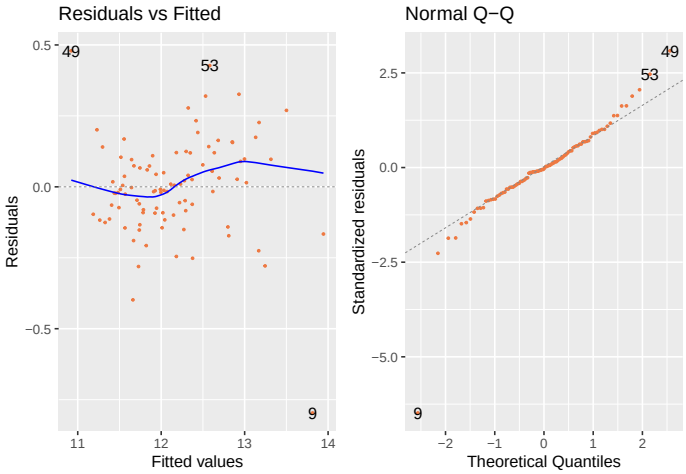


Data Transformation

Constant variance of erros assumptions can often be solved with a response variable transformations. The log transformations performed the best compared to other transformations. It has the most appropriate properties for the normality of residuals and constant variance. The residual plot for the other transformation (reciprocal, square root, reciprocal square root and inverse) did not show much improvement from the original model and were very influenced by possible influencial points.

In conclusion, our best transformation is the " log" transformation. The residual plot does not appear to have any alarming shape, and the the residuals are normally distributed, except for the problematic observations, 9, 49 and 58. Our new transformed model indicate that a linear model provides a decent fit to the data.

Log Transformmed model - Residual Plots:



Full Regression Model Significance Test

Our linear regression model is significant given that the p-value for the F-test is smaller than our level of significance 0.05.This means that at least one regressor has a linear relashionship with the median home value. With the marginal t-test for any individual regressor coefficient, we found that the intercept, total population, median household income, total housing units and housing affordability index are signifacant for predicting log(median home value) - that is, there is evidence that these coefficients are significantly not zero.

In more detail the F-test Hypothesis is : H_0: All coefficients are significantly 0. H_1: At least one coefficient is a significant predictor.

The t-test Hypothesis is : H_0: $B_j = 0$ H_1: B_j != 0

This is a first good step in our analysis and important to keep in mind.

```
##
## Call:
## lm(formula = log(Median.Home.Value) ~ ., data = reduce_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79685 -0.09169 -0.00563  0.09730  0.47930
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.234e+01  2.922e-01  42.240  < 2e-16 ***
## Total.Population             -1.073e-05  4.531e-06  -2.369  0.02003 *
## Diversity.Index1             -2.442e-04  2.137e-03  -0.114  0.90931
## Median.Household.Income       1.481e-05  1.843e-06   8.034 3.95e-12 ***
## Per.Capita.Income            -2.328e-06  3.027e-06  -0.769  0.44398
## Total.Housing.Units           3.206e-05  1.135e-05   2.826  0.00584 **
## Average.Household.Size       -7.059e-02  7.427e-02  -0.950  0.34446
## Housing.Affordability.Index2 -6.052e-03  5.499e-04 -11.005  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.179 on 88 degrees of freedom
## Multiple R-squared:  0.927,  Adjusted R-squared:  0.9212
## F-statistic: 159.6 on 7 and 88 DF,  p-value: < 2.2e-16
```

Evaluating all possible subset regression models

In looking for the "best" model, certain criteria must be met inorder for proper variable selection of the regressor equation. These criteria help us to be able to explain the data in the simpliest way with redundant predictors removed inorder minimize cost and to avoid multi-collinearity in our regression model.

The criteria for our variable selection include: 1) Large R^2 value 2) Maximum Adjusted R^2 value 3) Minimum MSres 4) Minimum Mallow's C_p Statistic value.

Based on the above criteria, the "best" candidate models are:

1) Model 1:

$log(y^{hat}) = (1.136e + 01) + (2.438e - 05)x_4$

2) Model 8:

$log(y^{hat}) = (1.232e+01) + (1.409e-05)x_3 - (7.258e-03)x_7$

3) Model 29:

$log(y^{hat}) = (1.267e+01) + (1.301e-05)x_3 - (1.547e-01)x_6 - (6.197e-03)x_7$

4) Model 64:

$log(y^{hat}) = (1.210e+01) - (1.258e-05)x_1 + (1.382e-05)x_3 + (3.717e-05)x_5 - (6.015e-03)x_7$

5) Model 99:

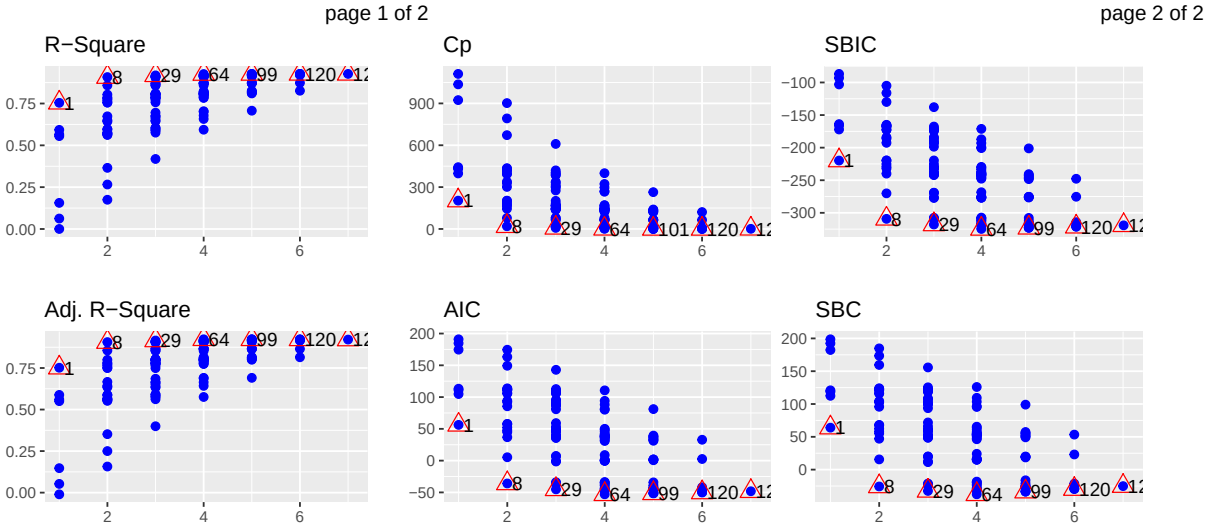
$log(y^{hat}) = (1.221e+01) - (1.036e-05)x_1 + (1.358e-05)x_3 + (3.176e-05)x_5 - (4.477e-02)x_6 - (5.876e-03)x_7$

6) Model 120:

$log(y^{hat}) = (1.232e+01) - (1.084e-05)x_1 + (1.484e-05)x_3 - (2.332e-06)x_4 + (3.223e-05)x_5 - (7.125e-02)x_6 - (6.046e-03)x_7$

7) Model 127:

$log(y^{hat}) = (1.234e+01) - (1.073e-05)x_1 - (2.442e-04)x_2 + (1.481e-05)x_3 - (2.328e-06)x_4 + (3.206e-05)x_5 - (7.059e-02)x_6 - (6.052e-03)x_7$



Evaluation of PRESS and VIF of candidate models:

Once we identified the "best"" candidate models, we compare its predicted residual error sum of squares (PRESS) statistic with other candidate models and selected the model with the smallest value. We also compare candidate models by performing a variance inflation factor (VIF) in order to quantify the severity of multicollinearity in the model. Multicollinearity refers to a situation when two or more explanatory variables in a multiple regression model are highly linearly related. Indicators of multicollinearity can be present in a model with inflatted coefficient estimates.

* The model with the lowest PRESS value is Model 64 however there is evidence of multicollinieariy.

* The model with the second lowest PRESS value is Model 29 and the same model doesnt show any evidence of multicolinearity in the variance inflation factor test of each regressor.

PRESS Statistic:

```
## [1] "Model 1 PRESS: 10.733"
```

```
## [1] "Model 8 PRESS: 4.236"
```

```
## [1] "Model 29 PRESS: 3.814"
```

```
## [1] "Model 64 PRESS: 3.76"
```

```
## [1] "Model 99 PRESS: 3.978"
```

```
## [1] "Model 120 PRESS: 4.97"
```

```
## [1] "Model 127 PRESS: 5.792"
```

Multicolinearity Check:

```
## [1] "Variance Inflation Factor"
```

```
##       x3       x7
## 1.081888 1.081888
```

```
##       x3       x6       x7
## 1.293148 2.136140 1.793435
```

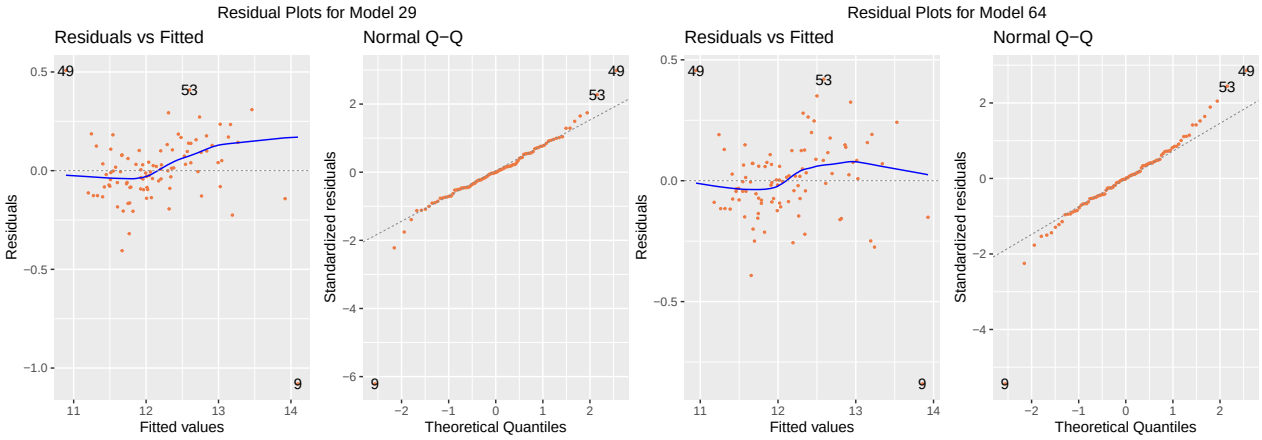
```
##       x1       x3       x5       x7
## 7.539310 1.107578 8.311564 1.636704
```

```
##        x1        x3        x5        x6        x7
## 16.681309  1.388036 16.575790  4.737732  1.960617
```

```
##        x1        x3        x4        x5        x6        x7
## 17.011191  7.448650 14.010788 16.626136  6.039030  2.339524
```

```
##        x1        x2        x3        x4        x5        x6        x7
## 17.742753  1.550252  7.646411 14.012521 16.921586  6.074919  2.360197
```

Residual Plots of best models:



Cross Validation:

The process of cross validtion is the splitting of data into parts for estimation and prediction. After the data is randomly shuffled, the splitting into groups is performed. The process is repeated with the desired iterations and the average cross-validation value is returned. Accordding to Cross-Validation, the best model is Model 29 after splitting our data into 4 groups, with K = 4, and repeating the fold cross-validation with 12 iterarions.

Giving consideration to the R_p^2 statistic, we decided that the best model is Model 29 becuase as we increase regressor the R_p^2 inflates - thats is the reason Model 64 has larger R_p^2.

Model 29:

$log(y^{hat}) = (1.267e+01) + (1.301e-05)MedianHomeValue - (1.547e-01)AverageHouseholdSize - (6.197e-03)HousingAffordabilityIndex$

```
##
## 4-fold CV results:
##   Fit         CV
## 1 LS0 0.08493068
## 2 LS1 0.08249517
##
## Best model:
##    CV
## "LS1"
```

```
## [1] "Model 64 R^2_p"
```

```
## [1] 0.9025903
```

```
## [1] "Model 29 R^2_p"
```

```
## [1] 0.9012032
```

Confidence Interval for coefficients:

In practice it is import to understand how string our estimated coefficients. We have computed the 95% confidence interval for coefficients in the best model. There are presented as:

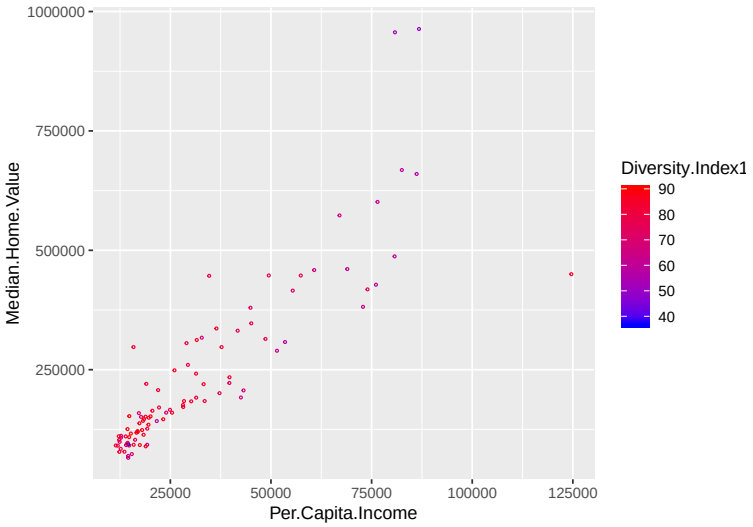
```
##      Coefficients lower                  upper
## [1,] "B1"         "1.1456043757933e-05"  "1.45737086525119e-05"
## [2,] "B2"         "-0.245299237720006"   "-0.064163808649389"
## [3,] "B3"         "-0.00718289254813186" "-0.00521141060612313"
```

```
##
## Call:
## lm(formula = new_y ~ x3 + x6 + x7, data = reduce_dat)
##
## Coefficients:
## (Intercept)           x3           x6           x7
##   1.267e+01    1.301e-05   -1.547e-01   -6.197e-03
```

Exploratory Data Analysis:

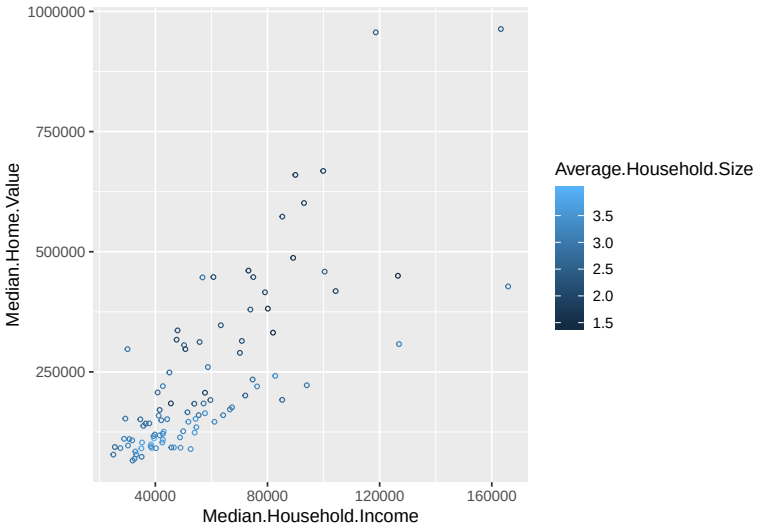
In houston most zip codes have a diversity index close to 80. However we see that neigborhoods with low diversity index tend to be situated in the extremes. The less diverse areas either have a very large per capita income and median home value, or a very per capita income and median home value. Also we see a positive relashionship between both variables.

```
## Loading required package: bitops
```



Total population and median home value are not correlated. However we found some intesresting insights about how Housing affordability index relates to both variables. The housing affordability index is large for houston areas with median home value bellow 250,000.00 dollars.

Median Household income is positively correlated to Median home value. As the Median home value and Median Household income decreases the Average Household Size increases. This means that families who are less wealthy tend to have on average more people residing at their home.



Exploratory Analisys conclusion

Although we are able to see how the variables used for our best models relate to median home value, we must call attention to the fact that our model predicts a transformed version of the median home value; Therefore, the relashionship may be diffent. Using correlation we can have some insight of how the predictors relate to the logged median home value.

```
## [1] "Logged Median value vs median household income 0.77"
```

```
## [1] "Logged Median value vs average household size -0.745"
```

```
## [1] "Logged Median value vs housing affordability index -0.752"
```

It turns out that the correlation of each predictor with the logged median home value coincides with the effects of the coefficients in our best model. Median household income is positively correlated with logged median home value, housing affordability index and average household size are negativelly correlatted with median home value. The coeficients of our model agrees with this observation.

In more detailed: * with all regressors held fixed an increase of median household income by 1 dollar is associated with increase of logged median home value by (1.301e-05) units on average. * with all regressors held fixed an increase of average household size by 1 person is associated with decrease of logged median home value by (1.547e-01) units on average. * with all regressors held fixed an increase of housing affordability index by 1 unit is associated with decrease of logged median home value by (6.197e-03) units on average.

References

Montgomery, Douglas C., and Anne G. Ryan. Introduction to Linear Regression Analysis, Fifth Edition. Wiley, 2013.