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 were 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.

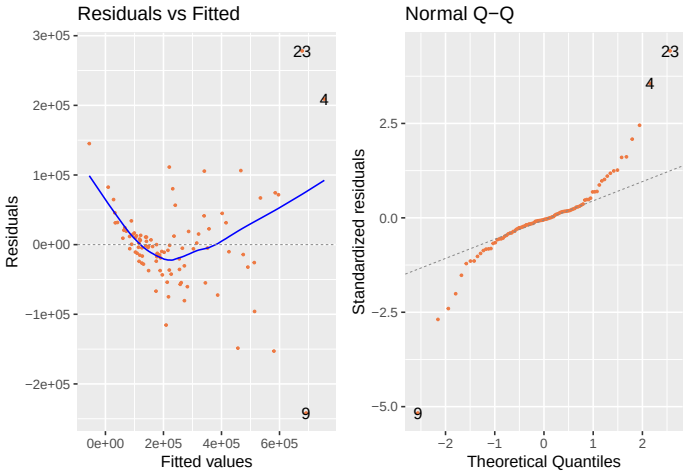
2. The Housing Affordability Index base is 100 and represents a balance point where a resident with a median household income can normally qualify to purchase a median price home. Values above 100 indicate increased affordability, while values below 100 indicate decreased affordability.

3. The outcome variable is the median home value (median price home).

Data Loading & Checking full Model Accuracy:

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

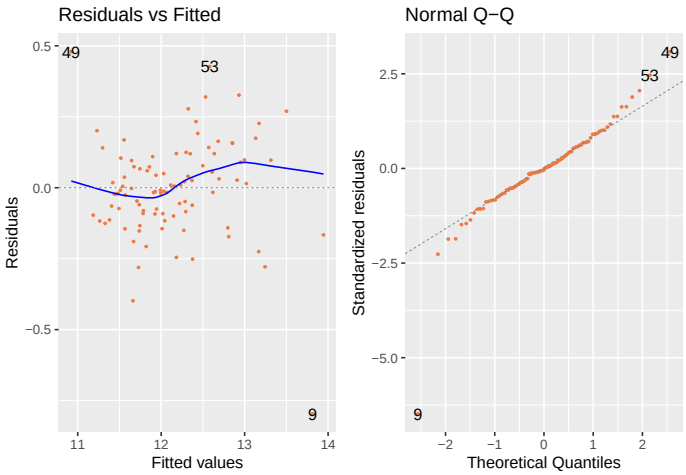
First model - Residual Plots:



Data Transformation

The constant variance of error 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. In conclusion, our best transformation is the "log" transformation. The residual plot does not appear to have any alarming shape, and the residuals are normally distributed, except for the problematic observations, 9, 49, and 53. Our new transformed model indicates that a linear model provides a decent fit to the data.

Log Transformed model - Residual Plots:



Influencial Points :

An observation with a large distance between itself and the center of the x-space as well as a large residual is likely to be influential. An influential point should only be discarded if there is an error in recording a measured value, the sample point is invalid or the observation is not part of the population that was intended to be sampled. An influential point should not be discarded if the point is a valid observation.

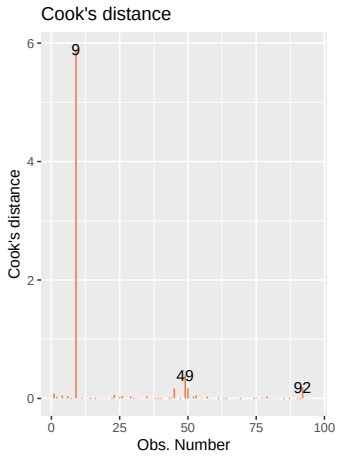
Cook's distance defines influence as a combination of leverage and residual size.A rule thumb is that an observation has high influence if Cook's distance exceeds 4/(n-p -1) where n is the number of observations and p the number of predictor variables (Kassambara, et al.).

In our case, we have identified three influential points, observation 9, 49, and 92. These observations are above Cook's distance of 0.04545455.

1. With regards to observation 9, it has an error with the Housing Affordability index of zero. This is highly unlikely because the affordability of zero would not allow people to buy houses in the area. We will delete this observation and continue evaluating possible models.

2. Concerning observation 49, it is located in northeast Houston. It has a population smaller than the surrounding zip codes and has a median income lower than 20K. Part of this zip code is industrial and undeveloped which explains the lower population and median income.

3. With regards to observation 92, it is located in the energy corridor of Houston, off I-10 West, the population is significantly less than the neighboring zip codes and has a median housing income greater than 165K. Half of this zip code included the George Bush Park which explains the lower population and median income.



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 relationship with the median home value. With the marginal t-test for any individual regressor coefficient, we found that the intercept, diversity index, median household income, per capita income, and housing affordability index are significant for predicting log(median home value) - that is, there is evidence that these coefficients are significantly not zero.

** Deleting observation nine changed the significance of each predictor. Before deleting this point, diversity index and per capita income were deemed insignificant. **

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 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
## -0.37739 -0.06778 -0.00442  0.06235  0.47123
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.191e+01  2.179e-01  54.648  < 2e-16 ***
## Total.Population              1.963e-07  3.514e-06   0.056 0.955571
## Diversity.Index1              5.116e-03  1.666e-03   3.071 0.002849 **
## Median.Household.Income       9.488e-06  1.467e-06   6.469 5.58e-09 ***
## Per.Capita.Income             9.960e-06  2.596e-06   3.836 0.000236 ***
## Total.Housing.Units          -1.144e-06  9.050e-06  -0.126 0.899718
## Average.Household.Size       -5.336e-02  5.402e-02  -0.988 0.325999
## Housing.Affordability.Index2 -6.200e-03  4.001e-04 -15.497  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1301 on 87 degrees of freedom
## Multiple R-squared:  0.9611, Adjusted R-squared:  0.9579
## F-statistic: 306.9 on 7 and 87 DF,  p-value: < 2.2e-16
```

Evaluating all possible subset regression models

In looking for the "best" model, certain criteria must be met for proper variable selection of the regressor equation. These criteria help us to be able to explain the data easily with redundant predictors removed to minimize cost and to avoid multicollinearity 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: $log(y^{hat}) = B_0 + B_4x_4$

2) Model 8: $log(y^{hat}) = B_0 + B_3x_3 + B_7x_7$

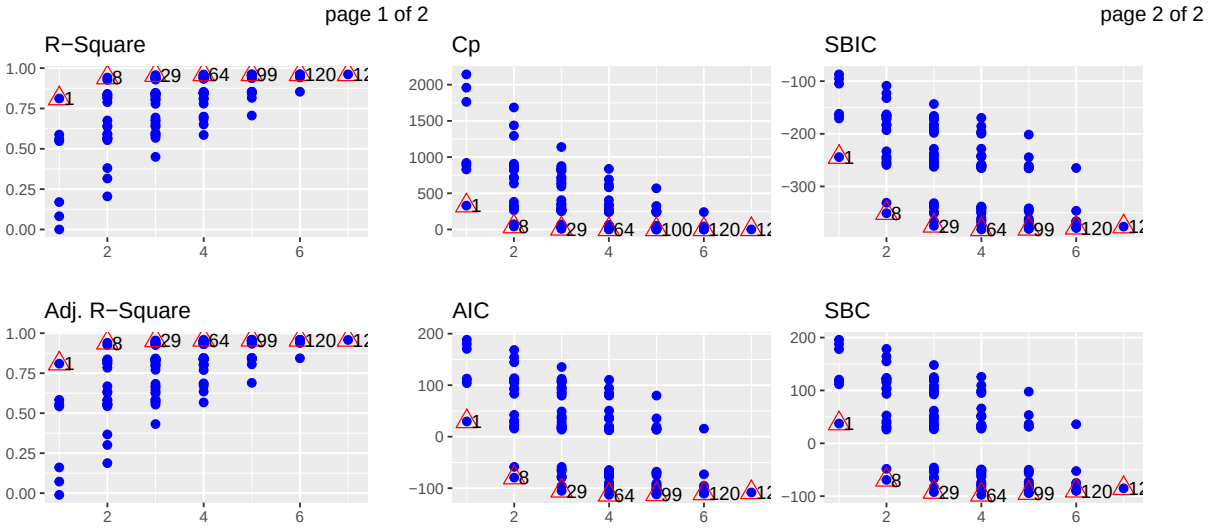
3) Model 29: $log(y^{hat}) = B_0 + B_3x_3 + B_4x_4 + B_7x_7$

4) Model 64: $log(y^{hat}) = B_0 + B_2x_2 + B_3x_3 + B_4x_4 + B_7x_7$

5) Model 99: $log(y^{hat}) = B_0 + B_2x_2 + B_3x_3 + B_4x_4 + B_6x_6 + B_7x_7$

6) Model 120: $log(y^{hat}) = B_0 + B_2x_2 + B_3x_3 + B_4x_4 + B_5x_5 + B_6x_6 + B_7x_7$

7) Model 127: $log(y^{hat}) = B_0 + B_1x_1 + B_2x_2 + B_3x_3 + B_4x_4 + B_5x_5 + B_6x_6 + B_7x_7$



Evaluation of PRESS and VIF of candidate models:

Once we identified the "best" candidate models, we compared their 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) 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 inflated coefficient estimates. VIF values larger than 10 are and an indicator that multicollinearity is present.

* The model with the lowest PRESS value is Model 64. No multicollinearity is found.

* The model with the second-lowest PRESS value is Model 99. However, this model has multicollinearity which is a property that can impact the ability to estimate the regression coefficients. For this reason, we selected Model 29 as our second-best model.

* Model 29 has the third-lowest PRESS value and has no multicollinearity problem.

* Also, the normal probability plot of the residuals in Model 29 has a more normally distributed appearance.

PRESS Statistic:

```
## [1] "Model 1 PRESS: 7.441"
## [1] "Model 8 PRESS: 2.665"
## [1] "Model 29 PRESS: 1.993"
## [1] "Model 64 PRESS: 1.813"
## [1] "Model 99 PRESS: 1.829"
```

```
## [1] "Model 120 PRESS: 1.861"
```

```
## [1] "Model 127 PRESS: 1.929"
```

Multicolinearity Check:

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

```
##       x3       x7
## 1.051021 1.051021
```

```
##       x3       x4       x7
## 5.126836 6.557243 1.775324
```

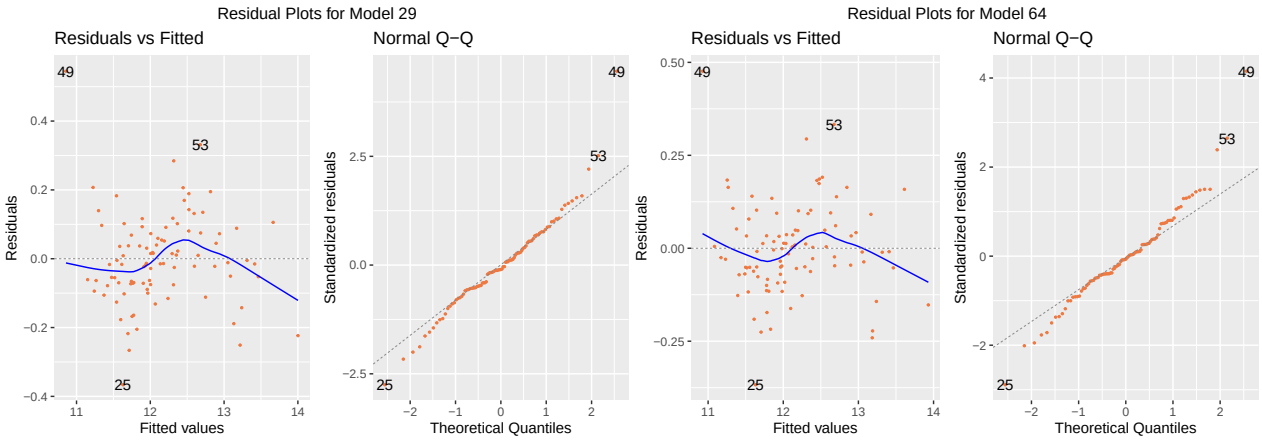
```
##       x2       x3       x4       x7
## 1.573076 5.224476 7.838956 1.845244
```

```
##        x2        x3        x4        x6        x7
##  1.579973  7.284604 14.543112  4.215907  1.939392
```

```
##        x2        x3        x4        x5        x6        x7
##  1.765718  7.425128 14.663463  1.376987  4.270442  2.186054
```

```
##        x1        x2        x3        x4        x5        x6        x7
## 19.369752  1.770449  8.560659 16.062823 19.557791  5.752153  2.186286
```

Residual Plots of best models:



Cross Validation:

The process of cross-validation 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. According to Cross-Validation, after splitting our data into 4 groups, with K = 4, and repeating the fold cross-validation with 12 iterations, the best model is Model 64.

Considering the R_p^2 statistic, we decided that the best model is Model 64 because it has the largest R_p^2 compared to Model 29. Although as we increase regressors the R_p^2 inflates, the cross-validation supports the evidence that the best model is Model 64.

Model 64:

$log(y^{hat}) = (1.176e+01) + (4.860e-03)DiversityIndex + (8.682e-06)MedianHomeValue + (1.181e-05)PerCapitaIncome - (6.254e-03)HousingAffordabilityIndex$

```
##
## 4-fold CV results:
##   Fit         CV
## 1 LS0 0.06943265
## 2 LS1 0.07147628
##
## Best model:
##    CV
## "LS0"

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

## [1] 0.9520864

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

## [1] 0.9473271
```

Confidence Interval for coefficients:

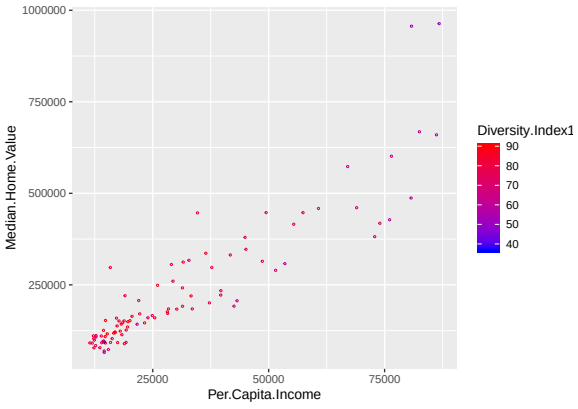
In practice, it is important to understand how strong our estimated coefficients are. We have computed the 95% confidence interval for coefficients in the best model. They are presented as:

```
##      Coef lower                  upper
## [1,] "B2" "0.00177016857008981"  "0.00794929966213829"
## [2,] "B3" "6.42807097332075e-06" "1.09367864765044e-05"
## [3,] "B4" "8.23746779102487e-06" "1.53742010561545e-05"
## [4,] "B7" "-0.00697719412306869" "-0.00553092992540379"
```

Exploratory Data Analysis:

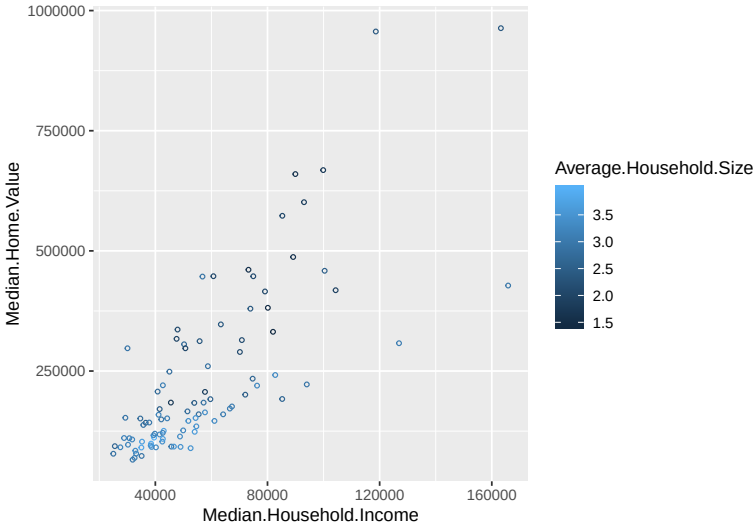
In Houston, most zip codes have a diversity index close to 80. However, we see that neighborhoods with low diversity index tend to be situated in the extremes. The less diverse areas either have a very large per capita income and median home value, or a very small per capita income and median home value.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

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



The total population and median home value are not correlated. However, we found some interesting insights into how the Housing affordability index relates to both variables. The housing affordability index is large for Houston areas with a median home value below 250,000.00 dollars. The total population has virtually no effect on housing affordability index or median home value.

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



Exploratory Analisys conclusion

Although we can 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 relationship may be different. Using correlation, we can have some insight into how the predictors relate to the logged median home value.

```
## [1] "Logged Median value vs Diversity Index  -0.412"
```

```
## [1] "Logged Median value vs median household income 0.767"
## [1] "Logged Median value vs Per capita Income  0.901"
## [1] "Logged Median value vs housing affordability index -0.749"
```

It turns out that the correlation of almost all predictors with the logged median home value coincides with the effects of the coefficients in our best model. Median household income and Per capita income are positively correlated with logged median home value, and these predictors have positive coefficients in the model. The housing affordability index is negatively correlated with median home value and has a negative coefficient in the model. However, the diversity index is negatively correlated with the median home value but has a positive coefficient in the model.

In more detailed:

* With all regressors held fixed, an increase of median household income by 1 dollar is associated with an increase of logged median home value by (8.682e-06) units on average.

* With all regressors held fixed an increase of per capita income by 1 dollar is associated with an increase of logged median home value by (1.181e-05) units on average.

* With all regressors held fixed an increase of diversity index by 1 unit is associated with an increase of logged median home value by (4.860e-03) units on average.

* With all regressors held fixed an increase of housing affordability index by 1 unit is associated with a decrease of logged median home value by (6.254e-03) units on average.

Conclusion

Our objective was to find the best linear regression model to predict the median home value of houses in the Houston neighborhoods. With the use of diagnostics, Data Transformation, and variable selection, we were able to create a subset regression model from the data collected from the Texas Gazetteer. Our best model is based on the logged response variable of Median Home Value, with Median Home Income, Per Capita Income, Diversity Index, and Housing Affordability Index as the regressors. Furthermore, we performed exploratory data analysis to better inform ourselves about the data. Furthermore, by identifying influential points we were able to identify an error within our data. In consequence, we were able to improve the performance of our model. When the regression was done including observation 9 we obtained larger PRESS statistics and the significance of the predictors was not in accordance with what we found by visualizing the data. When we modeled our data without observation 9, the PRESS values were smaller and the significance of the predictors resulted to be in accordance with our exploratory data analysis.

References

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

Kassambara, et al. "Linear Regression Assumptions and Diagnostics in R: Essentials." STHDA, 11 Mar. 2018, www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/.

"ZIP Code 77050." ZIP Code 77050 Map, Demographics, More for Houston, TX, www.unitedstateszipcodes.org/77050/.

"ZIP Code 77094." ZIP Code 77094 Map, Demographics, More for Houston, TX, www.unitedstateszipcodes.org/77094/.

"Houston County TX Data ." Houston County TX Data & Peer Group Rankings, texas.hometownlocator.com/tx/houston/.

"Texas Gazetteer." Maps, Data, Photos for 4,743 Locations, texas.hometownlocator.com/.