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:

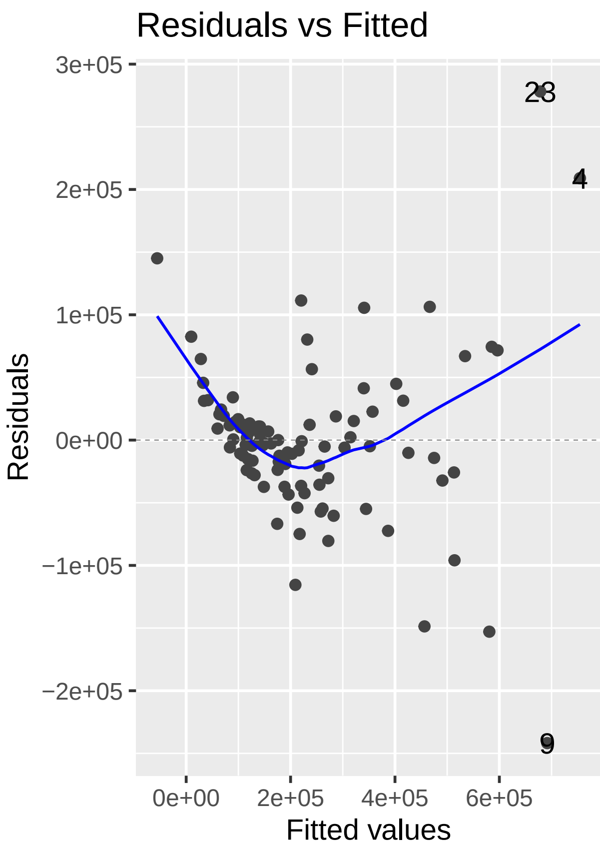
- x_1 -Total Population
- x_2 - Diversity Index
- Median Household Income
- Per Capita Income
- Total Housing Units
- Average Household Size
- Housing affordability Index

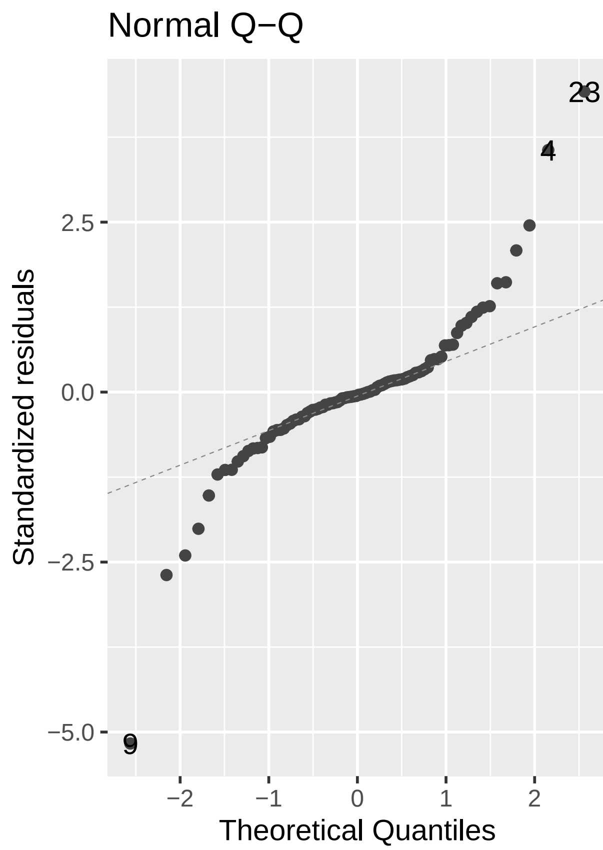
Data Loading & Checking full Model Accuracy:

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 data is heavily-tailed distribution with three possible influencial points. To address this problem we will inspect the dat and perform a few transformations.

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = reduce_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -241788  -25978   -3236   19032  278207
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.905e+05  1.112e+05   1.713 0.090259 .
## x1          -3.782e+00  1.724e+00  -2.193 0.030914 *
## x2          -2.845e+03  8.135e+02  -3.497 0.000740 ***
## x3           2.401e+00  7.015e-01   3.423 0.000943 ***
## x4           2.666e+00  1.152e+00   2.314 0.022996 *
## x5           1.099e+01  4.318e+00   2.546 0.012636 *
## x6           7.181e+04  2.827e+04   2.541 0.012821 *
## x7          -1.348e+03  2.093e+02  -6.439 6.14e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68120 on 88 degrees of freedom
## Multiple R-squared:  0.8633, Adjusted R-squared:  0.8524
## F-statistic: 79.37 on 7 and 88 DF,  p-value: < 2.2e-16
```





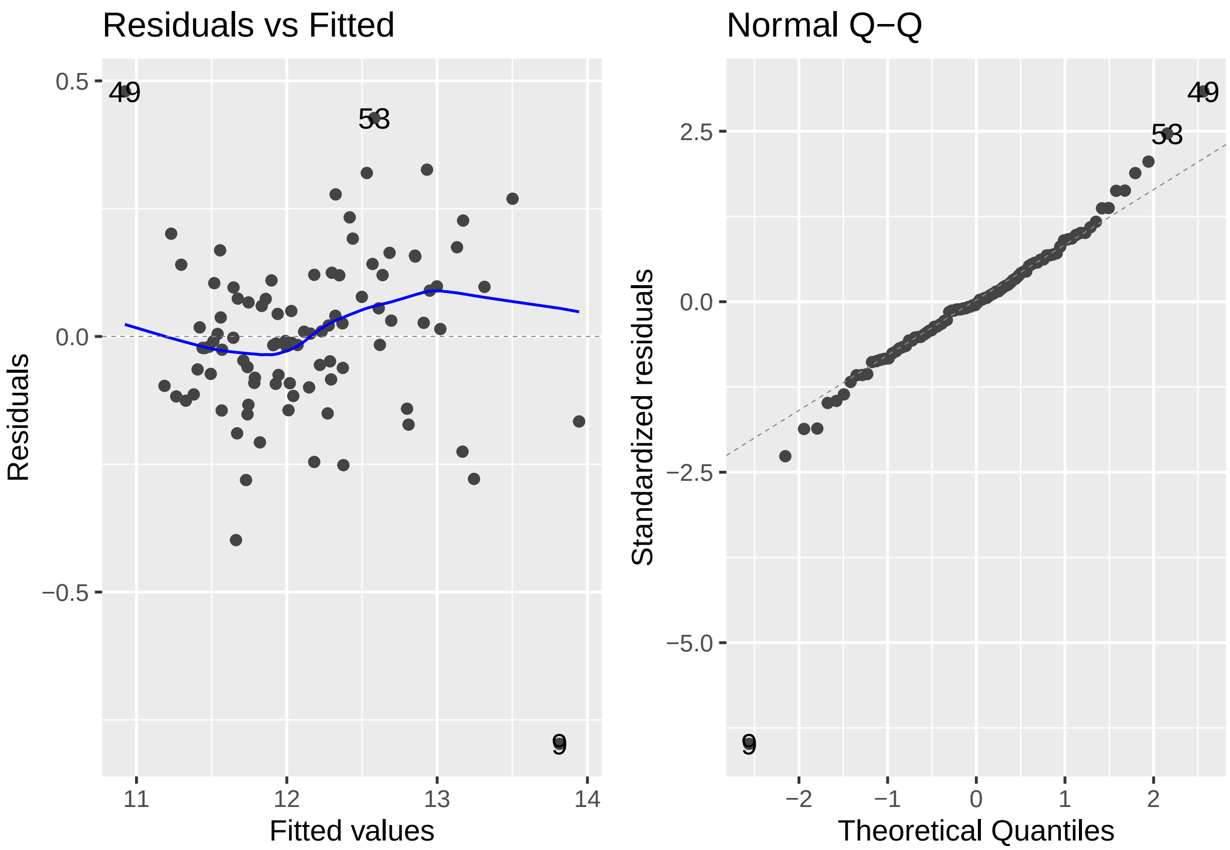
Data Transformation

Constant variance of erros assumptions can often be solve with 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.

```
y_recs <- y^(-1/2)
y_rec <- y^(-1)
```

```r
fit1 <- lm(sqrt(y)~x1+x2+x3+x4+x5+x6+x7,data=reduce_dat)
fit2 <- lm(log(y)~x1+x2+x3+x4+x5+x6+x7,data=reduce_dat)
fit3 <- lm(y_recs ~x1+x2+x3+x4+x5+x6+x7,data=reduce_dat)
fit4 <- lm(y_rec~x1+x2+x3+x4+x5+x6+x7,data=reduce_dat)
autoplot(fit2)[1:2]
```



Full regression model

```r
summary(lm(log(Median.Home.Value) ~., data = reduce_dat))
```

```
##
## 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 ***
```

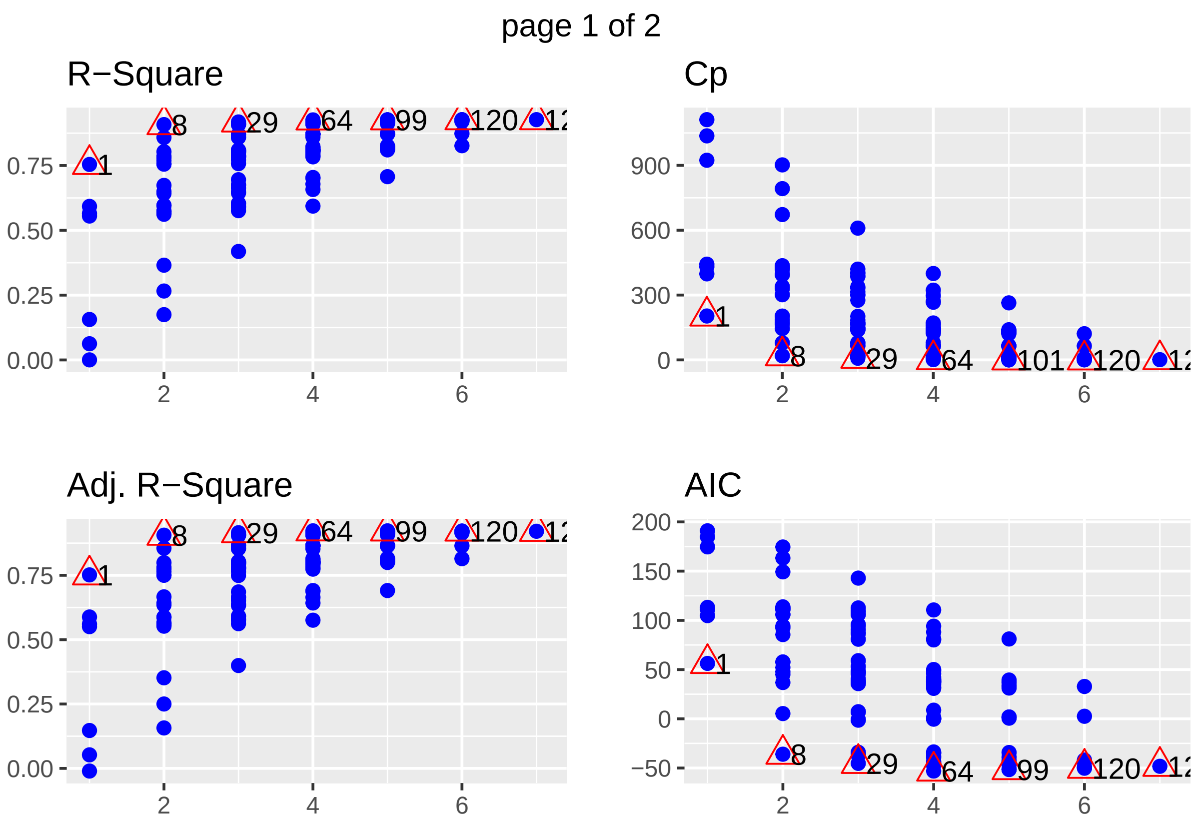
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.179 on 88 degrees of freedom
## Multiple R-squared:  0.927,	Adjusted R-squared:  0.9212
## F-statistic: 159.6 on 7 and 88 DF,  p-value: < 2.2e-16
```

Our linear regression model is significant given that the p-value for the F-test is smaller than our level of significance 0.05. Looking at the regressor individually, we found that the intercept, total population, median household income, total housing units and housing affordability index are signifacnt for predicting log(median home value). In model detail the F-test is perfomed to understand if at least one regressor is not equal to zero. The conclusion is supported by the t-test performed for each regressor. This is a first good step in our analysis and important to keep in mind.

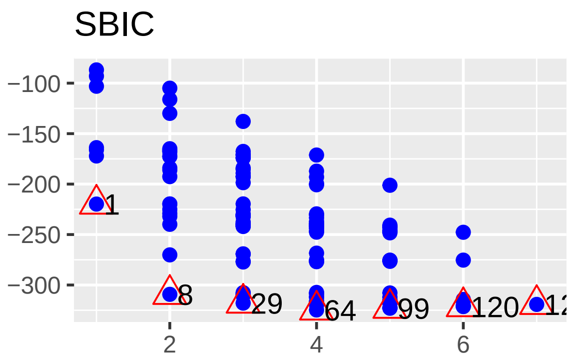
Evaluating all possible subset regression models

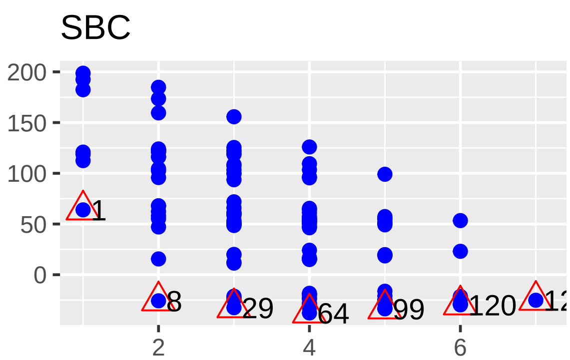
For Variable selection it is a good practice to exclude problematic observations?

```r
model <- lm(log(y)~x1+x2+x3+x4+x5+x6+x7,data=reduce_dat)
#ols_step_all_possible(model)
plot(ols_step_all_possible(model))
```



page 2 of 2





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 Cp Statistic value

Based on the above criteria, the “best” candidate models are:

1) Model 1: y ~ x4
2) Model 8: y ~ x3 + x7
3) Model 24: y ~ x5 + x7
4) Model 64: y ~ x2 + x3 + x6 + x7
5) Model 99: y ~ x1 + x3 + x5 + x6 + x7
6) Model 122:y ~ x1 + x2 + x3 + x4 + x5 + x7
7) Model 127:y ~.

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.

```
fit1 <- lm(log(y)~x4,data=reduce_dat)
fit8 <- lm(log(y)~x3+x7,data=reduce_dat)
fit24 <- lm(log(y)~x5+x7,data=reduce_dat)
fit64 <- lm(log(y)~x2+x3+x6+x7,data=reduce_dat)
fit99 <- lm(log(y)~x1+x3+x5+x6+x7,data=reduce_dat)
fit122 <- lm(log(y)~x1+x2+x3+x4+x5+x7,data=reduce_dat)
fit127 <- lm(log(y)~x1+x2+x3+x4+x5+x6+x7,data=reduce_dat)
```

```r
PRESS(fit1) # lowest PRESS
```

```
## [1] 10.73345
```

```r
PRESS(fit8)
```

```
## [1] 4.235565
```

```r
PRESS(fit24)
```

```
## [1] 18.10438
```

```r
PRESS(fit64)
```

```
## [1] 4.156367
```

```r
PRESS(fit99)
```

```
## [1] 3.97833
```

```r
PRESS(fit122)
```

```
## [1] 5.736032
```

```r
PRESS(fit127)
```

```
## [1] 5.792131
```

```r
vif(fit8)
```

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

```r
vif(fit24)
```

```
##       x5       x7
## 1.116877 1.116877
```

```r
vif(fit64)
```

```
##       x2       x3       x6       x7
## 1.395554 1.417276 2.389292 1.851421
```

```r
vif(fit99)
```

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

```r
vif(fit122)
```

```
##        x1        x2        x3        x4        x5        x7
## 12.273783  1.541093  7.148115 11.000198 11.194655  2.307864
```

```r
vif(fit127)
```

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

Interpretation of PRESS and Vif of candidate models:

The model with the lowest PRESS value is model 99 [log(y) ~ x1 + x3 + x5 + x6 + x7] however there is evidence of multicolliniearity.

The model with the second lowest PRESS value is model 64 [log(y) ~ x2 + x3 + x6 + x7] and the same model doesnt show any evidence of multicolinearity in the variance inflation factor test of each regressor.

Plot of model:

```
autoplot(fit64,size = 0.5, colour = 'sienna2')[1:2]
```