STAC67 Case Study: A model for predicting housing values in Boston

Revon Villava-Rayen (1002987501)

Sizhe Gao ()

Alexander Nguyen (1001522257)

Gopi Santhiran (1001371534)

Yu Heng Wu (1003475330)

Group 13

March 20 2019

# Abstract

Being able to predict housing values is a great boon for home buyers, sellers, and investors alike. However, there are many factors that go into a house’s value. Using R statistical analysis, this study aims to examine and determine which model best represents the correlation between housing prices and these influential factors.

# Background and Significance

Whether you are a student, someone dipping into the housing market or otherwise, the purchase or selling of a house is a massive financial burden. Will the house you’re selling be at its top value? Will the house you’re buying be at a low enough discount? Particularly in Boston, which currently faces a “growing housing challenge” (Bluestone, 2018), to be able to predict housing values is a great boon. As such we propose the following regression model to predict housing values in Boston, given 506 observations over 13 predictor variables.

# Exploratory Data Analysis

This Data set contains 506 observations on 13 predictors variables, which are per capita crime rate by town, proportion of residential land zoned for lots over 25,000 sq. ft., proportion of non-retail business acres per town Charles River dummy variable, nitric oxide concentration (parts per 10 million), average number of rooms per dwelling proportion of owner occupied units built prior to 1940, weighted distances to five Boston employment centres index of accessibility to radial highways, full-value property-tax rate per 10,000, pupil-teacher ratio by town, 1000(B - 0.63)^2 where B is the proportion of African Americans by town, a numeric vector of percentage values of lower status population.

Because of the rather long names, we will assign shorter names in capitals listed below:

### CRIME - per capita crime rate by town

The crime rate, measured by number of crimes per 100,000 of the population of the suburb observed. The mean and mode of this crime rate is 3.613 and 0.01501 respectively

### ZL - proportion of residential land zoned for lots over 25,000 sq. ft.

This is the measurement of land set aside for residential buildings that is over 25,000 sq ft. The mean and mode of this is 11.363 and 0 and respectively.

### NR\_PROP - proportion of non-retail business acres per town

This is the measurement of the amount of non-retail land in the town. The mean and mode is 11.136 and 18.1 respectively.

### CHR\_V - Charles River dummy variable

The Charles River is a 129 km long river in eastern Massachueetts. This is a binary variable and 6.9% of houses live near this river.

### NOX - nitric oxide concentration (parts per 10 million)

Refers to the concentration in parts per 10 million of nitric oxide. Short-term exposure to nitric oxide can inflict irritation to the respiratory system, and long term exposure can cause asthma and other respiratory infections (National Library of Medicine, 2017), suggesting that a large concentration of nitric oxide can adversely affect housing value. The mean and mode of nitric oxide concentration are 0.554 and 0.538 respectively.

### ROOM - average number of rooms per dwelling

This predictor measures the average room per dwelling in the observed suburb. The mean and mode of the number of rooms is 6.284 and 5.713 respectively.

### AGE - proportion of owner occupied units built prior to 1940

The mean and mode of the proportion of owner occupied units built prior to 1940 is 68.574 and 100 respectively.

### DIS - weighted distances to five Boston employment centres

The mean and mode weighted distances to five Boston employment centres is 3.795 and 3.4952 respectively.

### HWY - index of accessibility to radial highways

Radial highways are high-capacity urban roads leading to or from an urban center. The mean and mode of its index of accessibility is 9.549 and 24.

### TAX - full-value property-tax rate per 10,000

The full-value property-tax rate, measured per 10,000, meaning the amount of tax paid on a cap of $10000. The mean and mode of this tax rate is 408.237 and 666 respectively.

### PT\_RATIO - pupil-teacher ratio by town

This ratio is the total number of pupils enrolled at a level of education divided by the number of teachers at that same level. “It is generally assumed that a low pupil-teacher ratio signifies smaller classes, … teacher to pay more attention to individual student”(UNESCO, 2018) and vice versa for a high pupil-teacher ratio. The mean and mode is 18.455 and 20.2 respectively.

### B - 1000(B - 0.63)^2 where B is the proportion of African Americans by town

1000(B - 0.63)^2 is the scaling of the proportion of African Americans by town to match the scale of the other measurements. B represents the proportion of African American living in the town. The mean and mode of the proportion of African Americans by town is 356.674 and 396.9 respectively.

### L\_PER - a numeric vector of percentage values of lower status population

This predictor is the percentage of lower status population in each observation. The mean and mode is 12.653 and 8.05.

# Model

# Co-Lineartiy

## CRIME ZL NR\_PROP CHR\_V NOX ROOM AGE DIS HWY  
## CRIME 1.000 -0.200 0.407 -0.056 0.421 -0.219 0.353 -0.380 0.626  
## ZL -0.200 1.000 -0.534 -0.043 -0.517 0.312 -0.570 0.664 -0.312  
## NR\_PROP 0.407 -0.534 1.000 0.063 0.764 -0.392 0.645 -0.708 0.595  
## CHR\_V -0.056 -0.043 0.063 1.000 0.091 0.091 0.087 -0.099 -0.007  
## NOX 0.421 -0.517 0.764 0.091 1.000 -0.302 0.731 -0.769 0.611  
## ROOM -0.219 0.312 -0.392 0.091 -0.302 1.000 -0.240 0.205 -0.210  
## AGE 0.353 -0.570 0.645 0.087 0.731 -0.240 1.000 -0.748 0.456  
## DIS -0.380 0.664 -0.708 -0.099 -0.769 0.205 -0.748 1.000 -0.495  
## HWY 0.626 -0.312 0.595 -0.007 0.611 -0.210 0.456 -0.495 1.000  
## TAX 0.583 -0.315 0.721 -0.036 0.668 -0.292 0.506 -0.534 0.910  
## PT\_RATIO 0.290 -0.392 0.383 -0.122 0.189 -0.356 0.262 -0.232 0.465  
## B -0.385 0.176 -0.357 0.049 -0.380 0.128 -0.274 0.292 -0.444  
## L\_PER 0.456 -0.413 0.604 -0.054 0.591 -0.614 0.602 -0.497 0.489  
## MEDV -0.388 0.360 -0.484 0.175 -0.427 0.695 -0.377 0.250 -0.382  
## TAX PT\_RATIO B L\_PER MEDV  
## CRIME 0.583 0.290 -0.385 0.456 -0.388  
## ZL -0.315 -0.392 0.176 -0.413 0.360  
## NR\_PROP 0.721 0.383 -0.357 0.604 -0.484  
## CHR\_V -0.036 -0.122 0.049 -0.054 0.175  
## NOX 0.668 0.189 -0.380 0.591 -0.427  
## ROOM -0.292 -0.356 0.128 -0.614 0.695  
## AGE 0.506 0.262 -0.274 0.602 -0.377  
## DIS -0.534 -0.232 0.292 -0.497 0.250  
## HWY 0.910 0.465 -0.444 0.489 -0.382  
## TAX 1.000 0.461 -0.442 0.544 -0.469  
## PT\_RATIO 0.461 1.000 -0.177 0.374 -0.508  
## B -0.442 -0.177 1.000 -0.366 0.333  
## L\_PER 0.544 0.374 -0.366 1.000 -0.738  
## MEDV -0.469 -0.508 0.333 -0.738 1.000

The R code shown below displays the step forward and backwards process as well as the final model we will be using.

##   
## Call:  
## lm(formula = MEDV ~ CRIME + CHR\_V + ROOM + AGE + DIS + PT\_RATIO +   
## B, data = mod\_housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.810 -1.895 -0.407 1.657 12.441   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -28.739866 3.573034 -8.044 1.40e-14 \*\*\*  
## CRIME 0.250559 0.380749 0.658 0.510931   
## CHR\_V 1.172626 0.656461 1.786 0.074929 .   
## ROOM 10.112359 0.285453 35.426 < 2e-16 \*\*\*  
## AGE -0.063802 0.008536 -7.475 6.41e-13 \*\*\*  
## DIS -0.724717 0.116426 -6.225 1.40e-09 \*\*\*  
## PT\_RATIO -0.623436 0.089051 -7.001 1.33e-11 \*\*\*  
## B 0.018392 0.004974 3.697 0.000253 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.205 on 346 degrees of freedom  
## Multiple R-squared: 0.8585, Adjusted R-squared: 0.8557   
## F-statistic: 300 on 7 and 346 DF, p-value: < 2.2e-16

## [1] "So from our outlier Test with t-crit value of 3.925756, we see that only outlying data point is 162."

## rstudent unadjusted p-value Bonferonni p  
## 162 4.116193 4.842e-05 0.017141

## 1 2 3 4 5 6 7 8 9 10 11 12   
## -1.26 -0.95 0.64 0.92 1.57 1.37 0.56 2.23 0.83 0.02 -2.35 -0.30   
## 13 14 15 16 17 18 19 20 21 22 23 24   
## 0.22 0.57 -0.20 0.62 0.78 -0.17 1.98 0.44 -0.01 0.55 -1.46 -0.36   
## 25 26 27 28 29 30 31 32 33 34 35 36   
## -0.34 0.43 0.43 -0.64 -1.23 -1.17 -0.49 -1.42 -0.61 -0.58 -1.23 -0.55   
## 37 38 39 40 41 42 43 44 45 46 47 48   
## 0.14 0.13 0.83 0.37 0.25 -1.53 -0.02 -0.36 -0.47 -0.24 -0.22 -0.99   
## 49 50 51 52 53 54 55 56 57 58 59 60   
## 0.79 0.92 -0.47 -0.51 -1.00 0.14 1.50 0.48 0.46 0.24 0.67 0.17   
## 61 62 63 64 65 66 67 68 69 70 71 72   
## 0.94 -0.25 -0.27 -0.33 1.57 -1.02 -0.47 0.82 0.51 0.59 -0.81 0.03   
## 73 74 75 76 77 78 79 80 81 82 83 84   
## -0.25 -0.54 -0.43 -0.57 -0.57 -0.41 -0.08 0.14 -0.31 -0.72 -0.04 0.01   
## 85 86 87 88 89 90 91 92 93 94 95 96   
## -0.50 -0.40 0.13 -0.46 -2.34 -1.19 -1.11 -1.18 -1.51 -0.54 -1.28 0.19   
## 97 98 99 100 101 102 103 104 105 106 107 108   
## -0.56 -0.84 1.06 -0.77 0.72 0.08 0.60 0.01 0.14 1.04 1.09 0.28   
## 109 110 111 112 113 114 115 116 117 118 119 120   
## -0.80 -0.17 0.19 -1.16 0.23 -0.28 -1.23 0.21 -0.10 -0.20 0.89 0.59   
## 121 122 123 124 125 126 127 128 129 130 131 132   
## 0.20 -0.54 -0.13 -0.61 -0.45 -0.17 -0.45 0.81 -0.93 0.13 -0.60 -0.42   
## 133 134 135 136 137 138 139 140 141 142 143 144   
## 0.88 1.27 1.00 -0.55 0.50 -1.32 -0.53 -0.18 -1.28 2.25 -1.50 -1.01   
## 145 146 147 148 149 150 151 152 153 154 155 156   
## 0.14 -1.70 0.33 1.25 1.67 -0.79 -0.18 1.67 1.15 0.84 -1.59 -2.17   
## 157 158 159 160 161 162 163 164 165 166 167 168   
## 0.91 3.14 0.19 -0.78 0.25 4.12 2.77 1.09 -0.05 0.41 2.67 1.16   
## 169 170 171 172 173 174 175 176 177 178 179 180   
## -0.69 -1.66 -0.76 -1.05 1.75 -0.88 0.31 -0.06 -0.02 -0.27 -0.25 1.24   
## 181 182 183 184 185 186 187 188 189 190 191 192   
## 0.13 3.66 1.40 1.54 2.74 1.86 2.97 0.19 0.18 0.06 1.89 0.45   
## 193 194 195 196 197 198 199 200 201 202 203 204   
## 0.92 -0.47 -0.33 2.16 -1.04 -1.08 -0.39 1.47 0.35 -0.21 0.91 1.41   
## 205 206 207 208 209 210 211 212 213 214 215 216   
## 1.32 0.12 -0.07 1.50 0.77 2.12 0.58 1.51 0.59 0.61 2.59 0.26   
## 217 218 219 220 221 222 223 224 225 226 227 228   
## 0.12 0.25 -0.09 -0.99 -1.68 -0.48 -1.33 0.56 0.30 0.64 -1.26 -0.48   
## 229 230 231 232 233 234 235 236 237 238 239 240   
## 2.01 0.40 0.77 -1.19 -0.90 1.67 -0.22 0.32 -1.07 -1.13 -1.38 -1.53   
## 241 242 243 244 245 246 247 248 249 250 251 252   
## -2.57 -0.55 -0.55 -1.05 1.57 1.84 1.35 0.56 0.73 -0.26 -0.36 -0.14   
## 253 254 255 256 257 258 259 260 261 262 263 264   
## 0.30 0.54 0.16 0.43 1.88 -0.28 -0.39 -0.88 -0.87 1.33 0.57 -2.09   
## 265 266 267 268 269 270 271 272 273 274 275 276   
## 0.08 0.14 -1.17 0.77 0.91 -0.50 -0.13 -0.52 -1.15 -1.43 -0.14 -0.13   
## 277 278 279 280 281 282 283 284 285 286 287 288   
## -0.89 0.08 -0.02 0.15 1.23 0.28 1.68 1.31 -0.53 -1.56 -0.36 -0.45   
## 289 290 291 292 293 294 295 296 297 298 299 300   
## -0.85 -0.93 -1.25 0.79 -0.76 -0.50 -0.45 -0.40 -0.11 0.38 -1.12 -1.27   
## 301 302 303 304 305 306 307 308 309 310 311 312   
## -1.64 -2.17 -0.73 -0.17 0.15 -0.38 -1.17 -1.10 -1.47 -0.18 0.07 -0.86   
## 313 314 315 316 317 318 319 320 321 322 323 324   
## -0.42 -0.53 -0.56 -0.37 -0.16 0.66 -0.53 -0.35 -0.35 -0.42 -0.35 0.59   
## 325 326 327 328 329 330 331 332 333 334 335 336   
## -0.16 -0.47 -0.45 0.44 0.01 -0.50 -0.32 -0.48 -0.76 -0.23 -0.66 0.10   
## 337 338 339 340 341 342 343 344 345 346 347 348   
## 0.10 0.04 -0.42 -0.57 -0.44 -0.11 -2.20 -0.93 0.70 0.16 0.64 -0.53   
## 349 350 351 352 353 354   
## -0.92 0.21 0.46 0.95 0.95 1.30

## [1] "Below is the T crit value: "

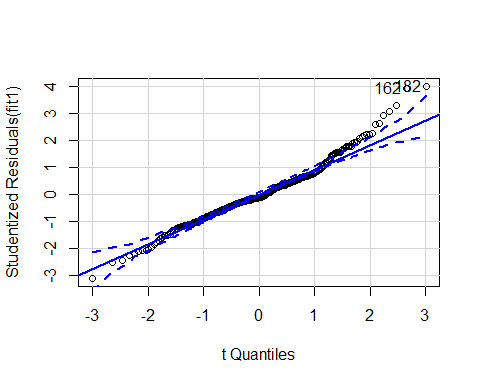
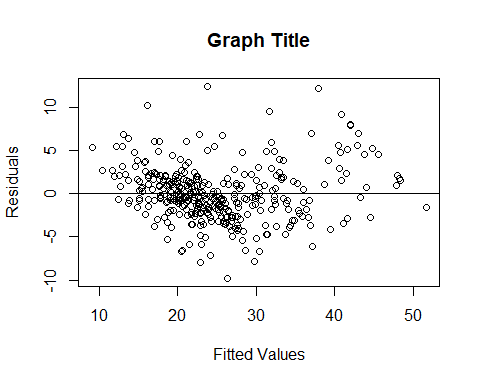
## [1] 3.925756

## 162   
## 162

## [1] "Model Validation"

## [1] 10.27365

## [1] 44.31514



## [1] 162 182

# Conclusions

The goal of this case study was to find and develop a model based on the dataset provided that can predict Boston housing prices so that prospective home buyers can look for an optimal time to purchase property. Our final model is MEDV ~ …. (WRITE OUT THE MODEL), which eliminates many variables with high collinearity, providing us with the lowest AIC alongside retaining a high r^2 value with the original model.

This model does have limitations, it has a small sample size as it only contains 506 records. For reference in the month of November 2017 alone, there were 2,429 single-family homes listed on the market (Woods, E.). The model also lacks other predictor variables that affect the price of a home, such as property taxes, management fees, mortgage rates, hydro and electrical costs, access to public transit to name a few. The data also does not separate houses of different types, as the there is a price jump from Condos to Houses despite both having the same number of bedrooms. A possible area of future study may be expanding the sample size or focusing on a single property type to obtain a more accurate prediction model.

# Citations

Bluestone, B. & Huessy, J. (2017). The Greater Boston Housing Report Card 2017 Ideas from the Urban Core - Responsive Development - as a Model for Regional Growth, 68

National Library of Medicine. (2017). Nitrogen Oxides: Your Environment, Your Health | National Library of Medicine. Retrieved from <https://toxtown.nlm.nih.gov/chemicals-and-contaminants/nitrogen-oxides>

UNESCO. (2018). Pupil-teacher ratio (PTR). Retrived from <http://uis.unesco.org/node/334770>

Woods, E. (2018). Greater Boston saw record median home prices in November <http://realestate.boston.com/buying/2018/12/28/greater-boston-record-november-home-sale-prices/>