Sectional Project 1

Group 3

2/11/2021

Introduction to the Boston Housing Dataset

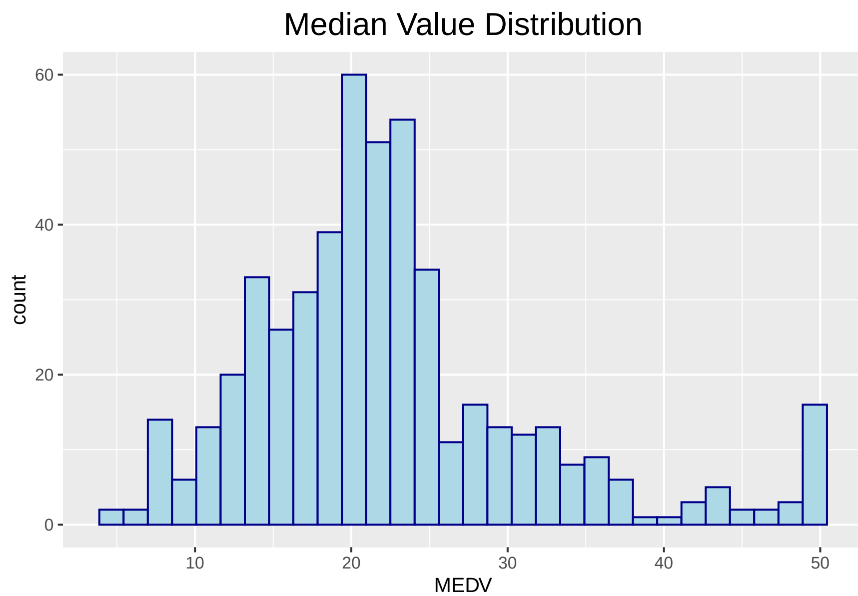
The Boston Housing dataset considers housing values and their associated properties in suburbs of Boston, Massachusetts. The dataset contains 506 observations and 14 attributes. We acquired the dataset from the Machine Learning Database (MLDB), found here. In particular, we are interested in constructing a model through regression techniques to gain insight on housing values. As such, we will use the 13 features to model 'MEDV', the median value of owner occupied homes (in $1,000s).

The data is displayed as follows.

```
##      CRIM ZN INDUS CHAS   NOX    RM  AGE    DIS RAD TAX PTRATIO      B LSTAT
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   MEDV
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

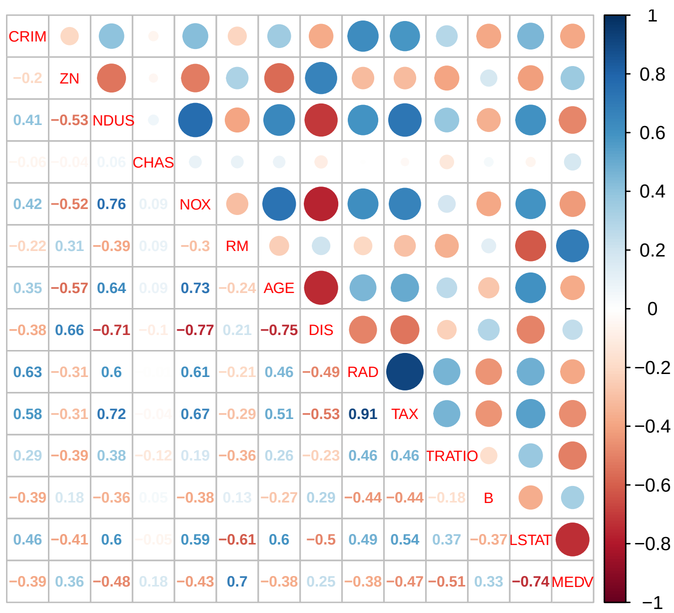
Exploratory Data Analysis

MEDV Distribution



The histogram demonstrates the values are not uniformly distributed. Rather, they follow a mostly normal distributions, with some outliers at the tail.

Correlation Matrix



While we produce many correlation values, we are firstly interested in how each attribute correlates to MEDV. This is represented by the bottom row or last column. We can immediately see the binary CHAS attribute does not correlate strongly with MEDV. However, it can be seen that RM (0.7) and LSTAT (-0.74) correlate with MEDV stronger than other attributes. Furthermore, the correlation between RM and LSTAT is -0.61. Since they do not correlate very strongly with one another, we can select both as predictor attributes without too much concern of collinearity for their case. The greatest correlation is between RAD and TAX of 0.91. Including both of these may raise some concerns regarding the minimal collinearity assumption of linear regression.

Example EDA subsection title 3

Modelling and Regression

MedV takes the value for Y, along 13 feature attributes of the dataset, in the form of $Y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$.

Multiple Linear Regression

This is very subject to change. I chose to simply split the data in a 80/20 split. If you think we should consider another split, or more, please do so. Furthermore, this is simply a start, and have yet to articulate everything. Please feel free to make changes.

Naively consider most attributes at onset.

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + INDUS + NOX + RM + AGE + DIS +
##     RAD + TAX + PTRATIO + B + LSTAT, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8959  -2.3701  -0.4062   1.3002  26.1609
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.009016   5.565276   6.291 8.48e-10 ***
## CRIM         -0.131286   0.028286  -4.641 4.73e-06 ***
## ZN            0.046919   0.013308   3.526 0.000473 ***
## INDUS         0.063575   0.059568   1.067 0.286507
## NOX         -14.612755   4.049855  -3.608 0.000348 ***
## RM            3.806793   0.461507   8.249 2.47e-15 ***
## AGE           0.001903   0.013782   0.138 0.890239
## DIS          -1.301183   0.203133  -6.406 4.31e-10 ***
## RAD           0.436588   0.061875   7.056 7.85e-12 ***
## TAX          -0.018724   0.003437  -5.447 9.06e-08 ***
## PTRATIO      -0.903870   0.130627  -6.920 1.86e-11 ***
## B             0.007184   0.002936   2.447 0.014852 *
## LSTAT        -0.539285   0.056712  -9.509  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.301 on 391 degrees of freedom
## Multiple R-squared:  0.7441, Adjusted R-squared:  0.7363
## F-statistic: 94.76 on 12 and 391 DF,  p-value: < 2.2e-16
```

Some initial insight is that LSTAT and RM indeed were strong predictors. Removing INDUS and AGE, every attribute becomes a significant predictor with the possible exception of B, depending on alpha. Let's consider what happens if we remove RAD, which varies strongly with TAX.

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + NOX + RM + DIS + TAX + PTRATIO +
##     B + LSTAT, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.437  -2.810  -0.353   1.582  27.899
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.450e+01  5.611e+00   4.366 1.62e-05 ***
## CRIM        -7.700e-02  2.871e-02  -2.682  0.00762 **
## ZN           4.011e-02  1.396e-02   2.873  0.00428 **
## NOX         -1.028e+01  3.899e+00  -2.636  0.00871 **
## RM           4.287e+00  4.708e-01   9.106  < 2e-16 ***
## DIS         -1.216e+00  1.987e-01  -6.119 2.27e-09 ***
## TAX         -3.327e-04  2.205e-03  -0.151  0.88013
## PTRATIO     -7.813e-01  1.345e-01  -5.810 1.29e-08 ***
## B            6.784e-03  3.099e-03   2.189  0.02917 *
## LSTAT       -5.418e-01  5.405e-02 -10.025  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.553 on 394 degrees of freedom
## Multiple R-squared:  0.7111, Adjusted R-squared:  0.7045
## F-statistic: 107.7 on 9 and 394 DF,  p-value: < 2.2e-16
```

We can see that without RAD, TAX is no longer a strong predictor. As such, TAX adds predictive value in relation to RAD. The next model removes TAX and adds RAD back in.

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + NOX + RM + DIS + RAD + PTRATIO +
##     B + LSTAT, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9970  -2.7405  -0.4075   1.3989  26.1356
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.489991   5.649347   5.574 4.62e-08 ***
## CRIM         -0.126894   0.029208  -4.344 1.78e-05 ***
## ZN            0.035439   0.013533   2.619  0.00917 **
## NOX         -17.651654   3.744914  -4.714 3.38e-06 ***
## RM            4.070530   0.462800   8.795  < 2e-16 ***
## DIS          -1.260201   0.194395  -6.483 2.70e-10 ***
## RAD           0.174630   0.040533   4.308 2.08e-05 ***
## PTRATIO      -0.964906   0.131718  -7.326 1.36e-12 ***
## B             0.008187   0.003026   2.706  0.00711 **
## LSTAT        -0.543257   0.052773 -10.294  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.45 on 394 degrees of freedom
## Multiple R-squared:  0.7241, Adjusted R-squared:  0.7178
## F-statistic: 114.9 on 9 and 394 DF,  p-value: < 2.2e-16
```

MSE for model 1, 2, and 3.

```
## [1] 23.18303
```

```
## [1] 22.6486
```

```
## [1] 23.02184
```

Lasso with parameter tuning will give us further insight into paramater selection. This section can/should be expanded/refined.

Ridge Regression

Lasso Regression

K-Fold Cross Validation

Citations