PS2_q2_Aastha

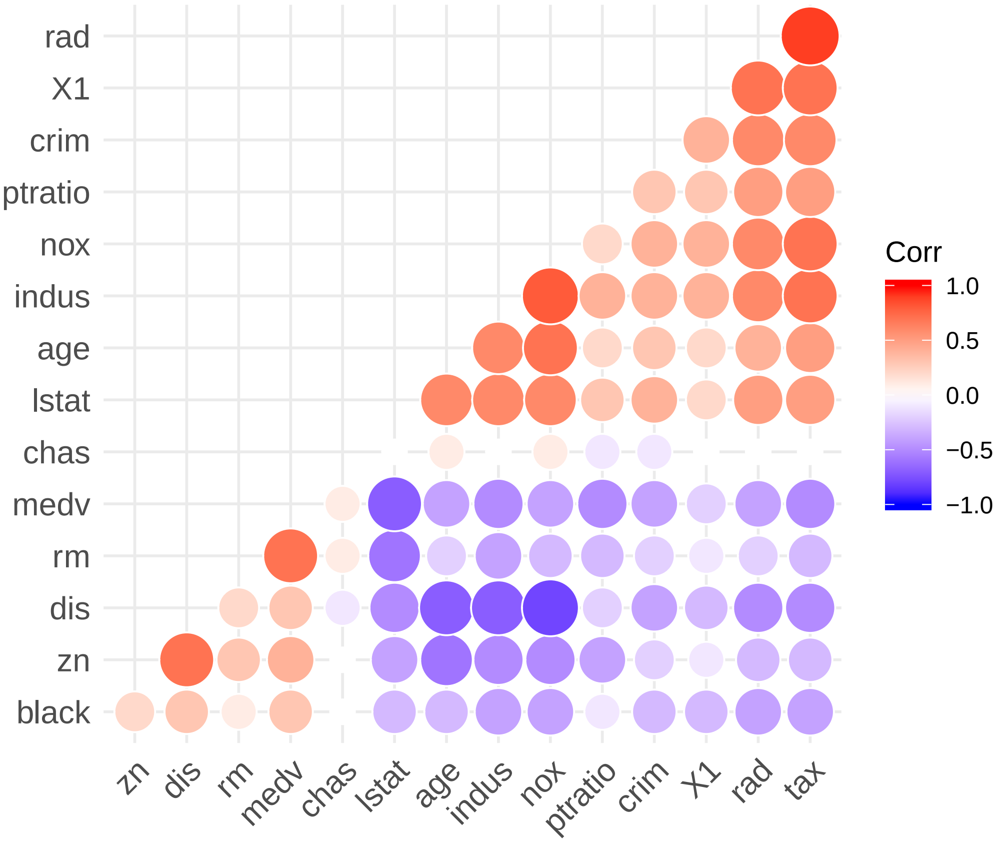
Aastha

2/18/2020

loading and splitting the data

2.1 Reporting correlations of these variables

```
##           X1 crim   zn indus chas  nox   rm  age  dis  rad  tax ptratio black
## X1       1.0  0.4 -0.1   0.4  0.0  0.4 -0.1  0.2 -0.3  0.7  0.7     0.3  -0.3
## crim     0.4  1.0 -0.2   0.4 -0.1  0.4 -0.2  0.3 -0.4  0.6  0.6     0.3  -0.3
## zn      -0.1 -0.2  1.0  -0.5  0.0 -0.5  0.3 -0.6  0.7 -0.3 -0.3    -0.4   0.2
## indus    0.4  0.4 -0.5   1.0  0.0  0.8 -0.4  0.6 -0.7  0.6  0.7     0.4  -0.4
## chas     0.0 -0.1  0.0   0.0  1.0  0.1  0.1  0.1 -0.1  0.0  0.0    -0.1   0.0
## nox      0.4  0.4 -0.5   0.8  0.1  1.0 -0.3  0.7 -0.8  0.6  0.7     0.2  -0.4
## rm      -0.1 -0.2  0.3  -0.4  0.1 -0.3  1.0 -0.2  0.2 -0.2 -0.3    -0.3   0.1
## age      0.2  0.3 -0.6   0.6  0.1  0.7 -0.2  1.0 -0.7  0.4  0.5     0.2  -0.3
## dis     -0.3 -0.4  0.7  -0.7 -0.1 -0.8  0.2 -0.7  1.0 -0.5 -0.5    -0.2   0.3
## rad      0.7  0.6 -0.3   0.6  0.0  0.6 -0.2  0.4 -0.5  1.0  0.9     0.5  -0.4
## tax      0.7  0.6 -0.3   0.7  0.0  0.7 -0.3  0.5 -0.5  0.9  1.0     0.5  -0.4
## ptratio  0.3  0.3 -0.4   0.4 -0.1  0.2 -0.3  0.2 -0.2  0.5  0.5     1.0  -0.1
## black   -0.3 -0.3  0.2  -0.4  0.0 -0.4  0.1 -0.3  0.3 -0.4 -0.4    -0.1   1.0
## lstat    0.2  0.4 -0.4   0.6  0.0  0.6 -0.6  0.6 -0.5  0.5  0.5     0.3  -0.3
## medv    -0.2 -0.4  0.4  -0.5  0.1 -0.4  0.7 -0.4  0.3 -0.4 -0.5    -0.5   0.3
##         lstat medv
## X1        0.2 -0.2
## crim      0.4 -0.4
## zn       -0.4  0.4
## indus     0.6 -0.5
## chas      0.0  0.1
## nox       0.6 -0.4
## rm       -0.6  0.7
## age       0.6 -0.4
## dis      -0.5  0.3
## rad       0.5 -0.4
## tax       0.5 -0.5
## ptratio   0.3 -0.5
## black    -0.3  0.3
## lstat     1.0 -0.7
## medv     -0.7  1.0
```



Yes, there seems to be some variables that are highly correlated with each other. As can be seen in the plot above, only a few variables have zero correlations, while several of them have bright red and blue color on the graph above pointing to high positive or negative correlations. This points out to the fact that we need to be careful in using highly correlated variables in the regression analysis that follows.

##2.2 Estimating the original HR model using the training data projecting median house price onto all of the other variables.

```
##
## Call:
## lm(formula = log(medv) ~ rm^2 + age + dis + rad + tax + ptratio +
##     black + lstat + crim + zn + indus + chas + nox^2, data = train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.714060 -0.096820 -0.016709  0.094901  0.755436
##
## Coefficients:
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  4.1328e+00  2.2417e-01  18.4358 < 2.2e-16 ***
## rm           9.0324e-02  1.8307e-02   4.9338 1.197e-06 ***
## age          7.9909e-07  5.8187e-04   0.0014 0.9989050
## dis         -4.9626e-02  8.9384e-03  -5.5519 5.232e-08 ***
## rad          1.3559e-02  3.1174e-03   4.3494 1.745e-05 ***
## tax         -6.3602e-04  1.7635e-04  -3.6067 0.0003504 ***
## ptratio     -3.9338e-02  5.6075e-03  -7.0152 1.021e-11 ***
## black        3.9680e-01  1.1698e-01   3.3919 0.0007650 ***
## lstat       -2.8947e+00  2.1809e-01 -13.2729 < 2.2e-16 ***
## crim        -1.1130e-02  1.3727e-03  -8.1079 6.734e-15 ***
## zn           1.1146e-03  5.9878e-04   1.8615 0.0634275 .
## indus        3.4779e-03  2.6462e-03   1.3143 0.1895182
## chas         9.1757e-02  3.7670e-02   2.4358 0.0153052 *
## nox         -7.5108e-02  1.6436e-02  -4.5698 6.560e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.18635 on 390 degrees of freedom
## Multiple R-squared:  0.79347,    Adjusted R-squared:  0.78658
## F-statistic: 115.25 on 13 and 390 DF,  p-value: < 2.22e-16
```

After projecting median house price onto all of the other variables we see that, most of the variables that are included in the regression have a significant effect on median house values except 'indus: proportion of non-retail business acres per town.

2.3a Estimating the model using LASSO with largest penalty

```
## [1] 0.01489646
```

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept)  3.539594e+00
## crim        -8.574114e-03
## zn           .
## indus        .
## chas         5.901490e-02
## age          .
## dis         -8.122134e-03
## rad          .
## tax         -9.209853e-06
## ptratio     -2.891983e-02
## black        2.852140e-01
## lstat       -2.683566e+00
## nox2        -1.573316e-03
## rm2          9.445933e-03
```

Note that, the LASSO is not very good at handling variables which show correlation between them and thus can sometimes show very wild behaviors. Notice that 4 variables have been dropped in this regression.

Optimal penalty is 0.024 for lasso regression

2.3b Ridge Regression with largest penalty

```
## [1] 0.1421939
```

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept)  3.5221923172
## crim        -0.0075619638
## zn           0.0003534904
## indus       -0.0013490631
## chas         0.0987322223
## age         -0.0007387262
## dis         -0.0166150373
## rad          0.0003291084
## tax         -0.0001632629
## ptratio     -0.0267846770
## black        0.3597670394
## lstat       -1.8147615176
## nox2        -0.0024152848
## rm2          0.0098574103
```

This model does not drop variables which are highly correlated like in the case of LASSO regression. optimal penalty is 0.17 in this case. The results from ridge regression are as above.

2.4 Expanding the data to contain square term of all variables and estimating using LASSO

```
## 25 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept)  4.052066e+00
## crim        -1.582099e-02
## zn           .
## indus        .
## chas         8.381005e-02
## age          .
## dis         -3.697057e-02
## rad          8.388444e-03
## tax         -3.060564e-04
## ptratio     -3.543675e-02
## black        3.156262e-01
## lstat       -3.979693e+00
## nox2        -4.573840e-03
## rm2          7.815079e-03
## crim2        7.140443e-05
## zn2          6.249731e-06
## indus2       5.942213e-05
## chas2        8.008046e-03
## age2         .
## dis2         .
## rad2         .
## tax2         .
## ptratio2     .
## black2       .
## lstat2       3.588727e+00
```

This time LASSO drops 3 variables (from the original dataset) and drops 6 squared variables in the expanded dataset. Also, note that nox^2 and rm^2 still survive this model with expanded dataset.

2.5 Comparing internal MSE versus test data MSE

I obtain that the LASSO in the extended model performs the best in out of sample prediciton, while the original LASSO model performs the worst in and out of sample in this situation. The original HR model performs good in-sample but not out of sample.RIdge regression does not perform the best either in-sample or out of sample.