HW5_Q3

2022-11-08

Question 3

a

```r
library(MASS)
dim(Boston)
```

```
## [1] 506  14
```

```r
Boston$logcrim = log(Boston$crim) # create log transform of crim
summary(Boston)
```

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv          logcrim       
##  Min.   : 1.73   Min.   : 5.00   Min.   :-5.0640  
##  1st Qu.: 6.95   1st Qu.:17.02   1st Qu.:-2.5005  
##  Median :11.36   Median :21.20   Median :-1.3606  
##  Mean   :12.65   Mean   :22.53   Mean   :-0.7804  
##  3rd Qu.:16.95   3rd Qu.:25.00   3rd Qu.: 1.3021  
##  Max.   :37.97   Max.   :50.00   Max.   : 4.4884
```

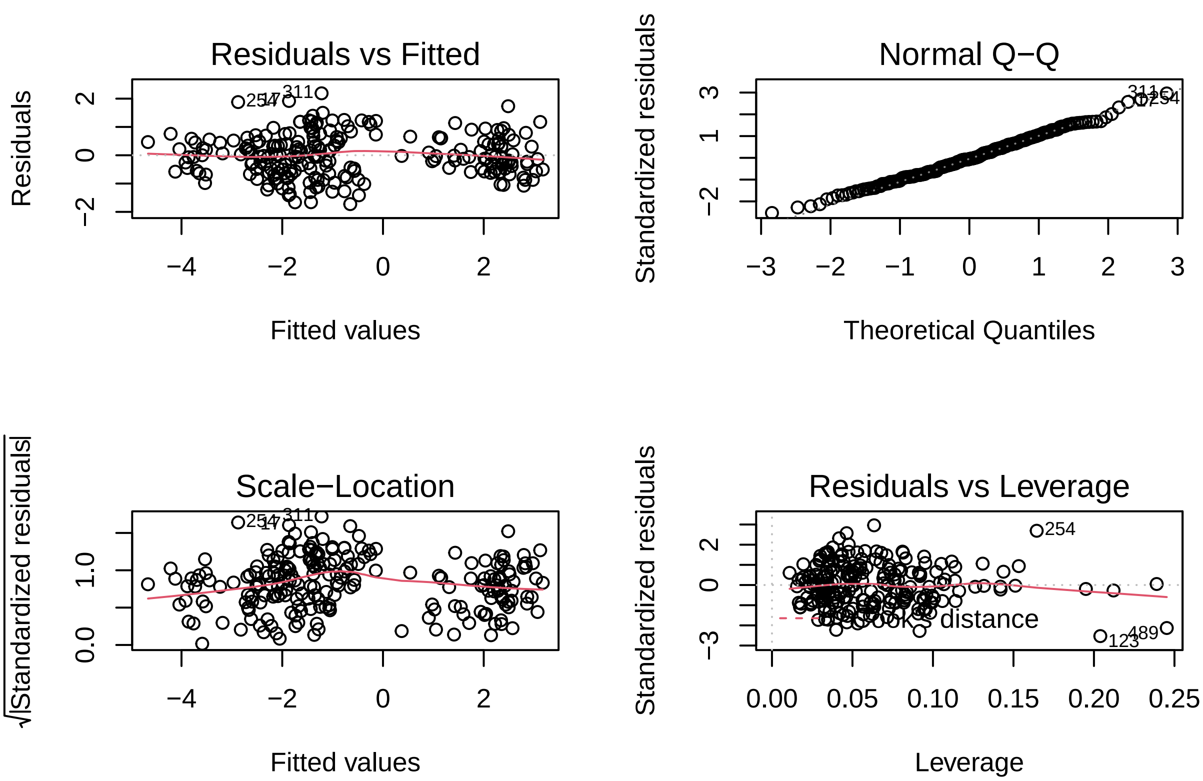
```r
set.seed(12345)
train = runif(nrow(Boston))<.5 # pick train/test split 50% train, 50% test
```

```r
table(train)
```

```
## train
## FALSE  TRUE
##   282   224
```

b

```r
train_set <- Boston[train, ]
test_set <- Boston[!train, ]
m1 <- lm(logcrim ~. - crim, data = train_set)
pred_m1 <- predict(m1, newdata = test_set)
mse <- mean((test_set$logcrim - pred_m1)^2)
par(mfrow = c(2,2))
plot(m1)
```




```r
summary(m1)
```

```
##
```

```
## Call:
## lm(formula = logcrim ~ . - crim, data = train_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72540 -0.55304 -0.04064  0.49671  2.18805
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.0748775  1.3219268  -1.570 0.118017
## zn          -0.0129701  0.0033472  -3.875 0.000143 ***
## indus        0.0229991  0.0171796   1.339 0.182101
## chas        -0.4421855  0.2144270  -2.062 0.040422 *
## nox          3.4151961  0.9629493   3.547 0.000481 ***
## rm          -0.0502194  0.1193062  -0.421 0.674238
## age          0.0062147  0.0031412   1.978 0.049188 *
## dis         -0.0613841  0.0572488  -1.072 0.284845
## rad          0.1310175  0.0183549   7.138 1.52e-11 ***
## tax         -0.0004676  0.0011034  -0.424 0.672141
## ptratio     -0.0420431  0.0345808  -1.216 0.225429
## black       -0.0021630  0.0006748  -3.205 0.001560 **
## lstat        0.0183597  0.0147684   1.243 0.215189
## medv        -0.0088543  0.0124793  -0.710 0.478787
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7638 on 210 degrees of freedom
## Multiple R-squared:  0.8863, Adjusted R-squared:  0.8793
## F-statistic:   126 on 13 and 210 DF,  p-value: < 2.2e-16
```

```
mse
```

```
## [1] 0.7083435
```

```
car::vif(m1)
```

```
##        zn     indus      chas       nox        rm       age       dis       rad
##  2.394561  4.932105  1.103165  4.743268  2.517062  3.235973  4.655808 10.381810
##       tax   ptratio     black     lstat      medv
## 14.073132  2.165536  1.307967  3.731381  4.695713
```

As seen above, the residual vs fitted plot hugs the middle line, which shows that the relationship between the predictors and the dependent variable is indeed linear. However, in the scale-location plot, we can see that the residuals are not randomly distributed. There is a non-constant variance happening, which means that there is heteroskedacticity.

According to VIFs, the full model might have the multicollinearity issue. - In the residuals plots, the variances are not very constant. - The test MSE = 0.7083435. - From the summary, zn, nox, rad, black are the most significant predictors (with large t-statistic or extremely small p-values). chas and age are also relatively important.

c

```r
stepAIC(m1, direction = "backward")
```

```
## Start:  AIC=-107.2
## logcrim ~ (crim + zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat + medv) - crim
##
##           Df Sum of Sq    RSS      AIC
## - rm       1    0.1034 122.60 -109.007
## - tax      1    0.1048 122.60 -109.005
## - medv     1    0.2937 122.79 -108.660
## - dis      1    0.6706 123.17 -107.973
## - ptratio  1    0.8622 123.36 -107.625
## - lstat    1    0.9015 123.40 -107.554
## - indus    1    1.0455 123.54 -107.293
## <none>                 122.50 -107.196
## - age      1    2.2832 124.78 -105.060
## - chas     1    2.4806 124.98 -104.706
## - black    1    5.9926 128.49  -98.498
## - nox      1    7.3372 129.84  -96.166
## - zn       1    8.7587 131.26  -93.727
## - rad      1   29.7208 152.22  -60.538
##
## Step:  AIC=-109.01
## logcrim ~ zn + indus + chas + nox + age + dis + rad + tax + ptratio +
##     black + lstat + medv
##
##           Df Sum of Sq    RSS      AIC
## - tax      1    0.1181 122.72 -110.792
## - medv     1    0.5970 123.20 -109.919
## - dis      1    0.7104 123.31 -109.713
## - ptratio  1    0.8993 123.50 -109.370
## <none>                 122.60 -109.007
## - lstat    1    1.1263 123.73 -108.959
## - indus    1    1.1536 123.75 -108.910
## - age      1    2.1809 124.78 -107.058
## - chas     1    2.5733 125.17 -106.355
## - black    1    5.9700 128.57 -100.357
## - nox      1    7.2814 129.88  -98.084
## - zn       1    9.0354 131.64  -95.079
## - rad      1   29.6177 152.22  -62.538
##
## Step:  AIC=-110.79
## logcrim ~ zn + indus + chas + nox + age + dis + rad + ptratio +
##     black + lstat + medv
##
##           Df Sum of Sq    RSS      AIC
## - medv     1     0.528 123.25 -111.830
## - dis      1     0.653 123.37 -111.603
## - ptratio  1     0.865 123.58 -111.218
## - indus    1     1.079 123.80 -110.831
## <none>                 122.72 -110.792
```

```
## - lstat    1     1.161 123.88 -110.682
## - age      1     2.299 125.02 -108.634
## - chas     1     2.461 125.18 -108.344
## - black    1     5.988 128.71 -102.120
## - nox      1     7.258 129.98  -99.921
## - zn       1    10.203 132.92  -94.903
## - rad      1    97.384 220.10   18.068
##
## Step:  AIC=-111.83
## logcrim ~ zn + indus + chas + nox + age + dis + rad + ptratio +
##     black + lstat
##
##           Df Sum of Sq    RSS      AIC
## - dis      1     0.386 123.63 -113.129
## - ptratio  1     0.530 123.78 -112.869
## <none>                 123.25 -111.830
## - indus    1     1.438 124.69 -111.230
## - age      1     2.249 125.50 -109.779
## - chas     1     2.761 126.01 -108.868
## - lstat    1     3.480 126.73 -107.593
## - black    1     6.078 129.32 -103.047
## - nox      1     8.453 131.70  -98.971
## - zn       1    11.115 134.36  -94.488
## - rad      1    96.855 220.10   16.068
##
## Step:  AIC=-113.13
## logcrim ~ zn + indus + chas + nox + age + rad + ptratio + black +
##     lstat
##
##           Df Sum of Sq    RSS      AIC
## - ptratio  1     0.624 124.26 -114.001
## <none>                 123.63 -113.129
## - indus    1     1.697 125.33 -112.075
## - chas     1     2.732 126.37 -110.232
## - lstat    1     3.253 126.89 -109.311
## - age      1     3.614 127.25 -108.675
## - black    1     6.331 129.97 -103.942
## - nox      1    10.889 134.52  -96.221
## - zn       1    14.414 138.05  -90.426
## - rad      1    97.985 221.62   15.605
##
## Step:  AIC=-114
## logcrim ~ zn + indus + chas + nox + age + rad + black + lstat
##
##         Df Sum of Sq    RSS      AIC
## <none>               124.26 -114.001
## - indus  1     1.527 125.78 -113.265
## - chas   1     2.440 126.70 -111.645
## - lstat  1     2.936 127.19 -110.771
## - age    1     3.690 127.95 -109.447
## - black  1     6.563 130.82 -104.471
## - nox    1    13.923 138.18  -92.212
## - zn     1    14.598 138.85  -91.120
## - rad    1   113.704 237.96   29.543
```

```
##
## Call:
## lm(formula = logcrim ~ zn + indus + chas + nox + age + rad +
##     black + lstat, data = train_set)
##
## Coefficients:
## (Intercept)           zn        indus         chas          nox          age
##   -4.166078    -0.013538     0.023241    -0.424491     4.123156     0.007018
##         rad        black        lstat
##    0.118928    -0.002254     0.025485
```

```r
m2 <- lm(logcrim ~ zn + indus + chas + nox + age + rad + black + lstat, data = train_set)
pred_m2 <- predict(m2, newdata = test_set)
mse_m2 <- mean((test_set$logcrim - pred_m2)^2)
mse_m2
```

```
## [1] 0.7033381
```

The test set MSE is 0.7033381.

d

```r
train_X <- model.matrix(logcrim ~ .-1, data = train_set[, -1])
train_Y <- train_set$logcrim
cv_ridge <- cv.glmnet(train_X, train_Y, alpha = 0)
best_lambda <- cv_ridge$lambda.min


test_X <- model.matrix(logcrim ~ .-1, data = test_set[, -1])
test_Y <- test_set$logcrim
ridge <- glmnet(train_X, train_Y, lambda = best_lambda, alpha = 0)
pred_ridge <- predict(ridge, newx = test_X, s = best_lambda)
mse_ridge <- mean((test_Y - pred_ridge)^2)
mse_ridge
```

```
## [1] 0.7760607
```

The test MSE is 0.7760607. Best lambda is 0.1866301

e

```r
set.seed(1234)
cv_lasso <- cv.glmnet(train_X, train_Y, alpha = 1)
best_lambda <- cv_lasso$lambda.min

lasso <- glmnet(train_X, train_Y, lambda = best_lambda, alpha = 1)
pred_lasso <- predict(lasso, newx = test_X, s = best_lambda)
mse_lasso <- mean((test_Y - pred_lasso)^2)
mse_lasso
```

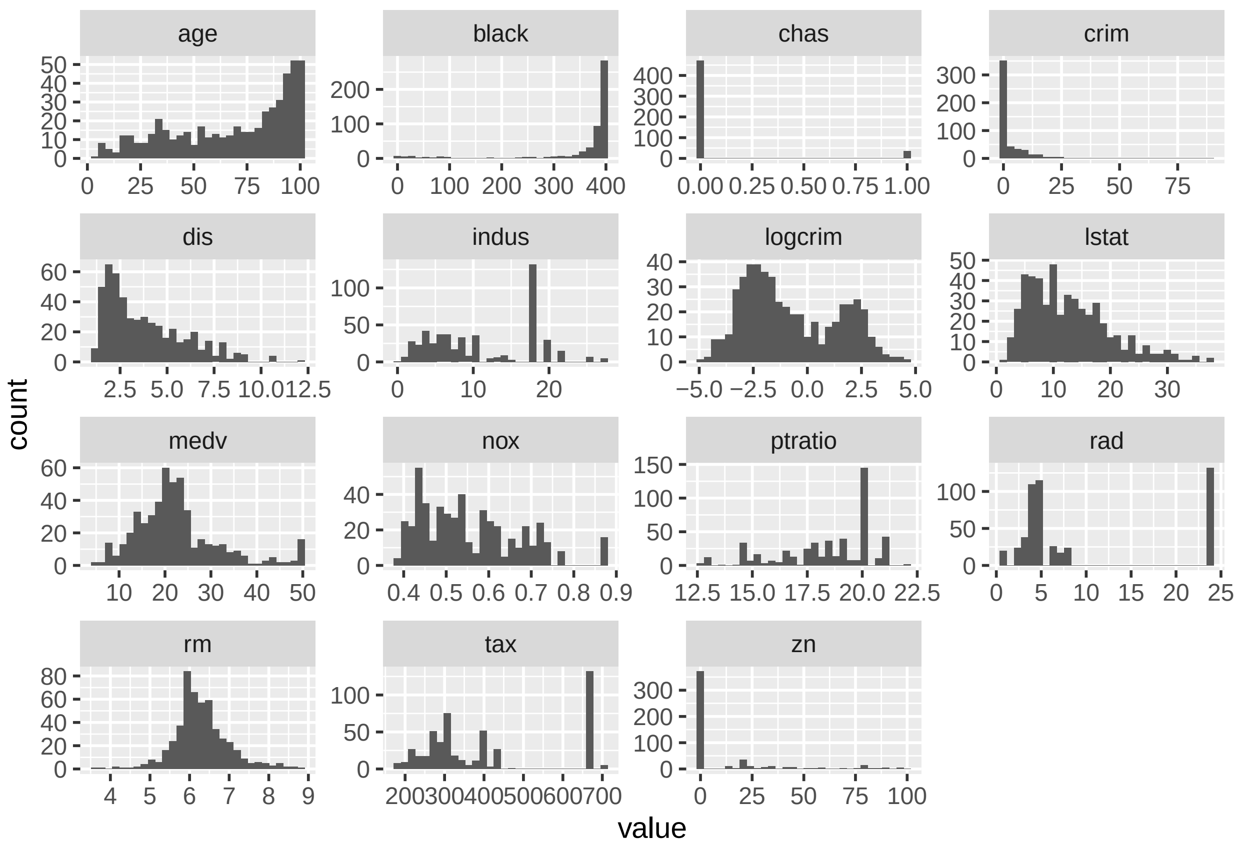
```
## [1] 0.7023476
```

The test MSE is 0.702

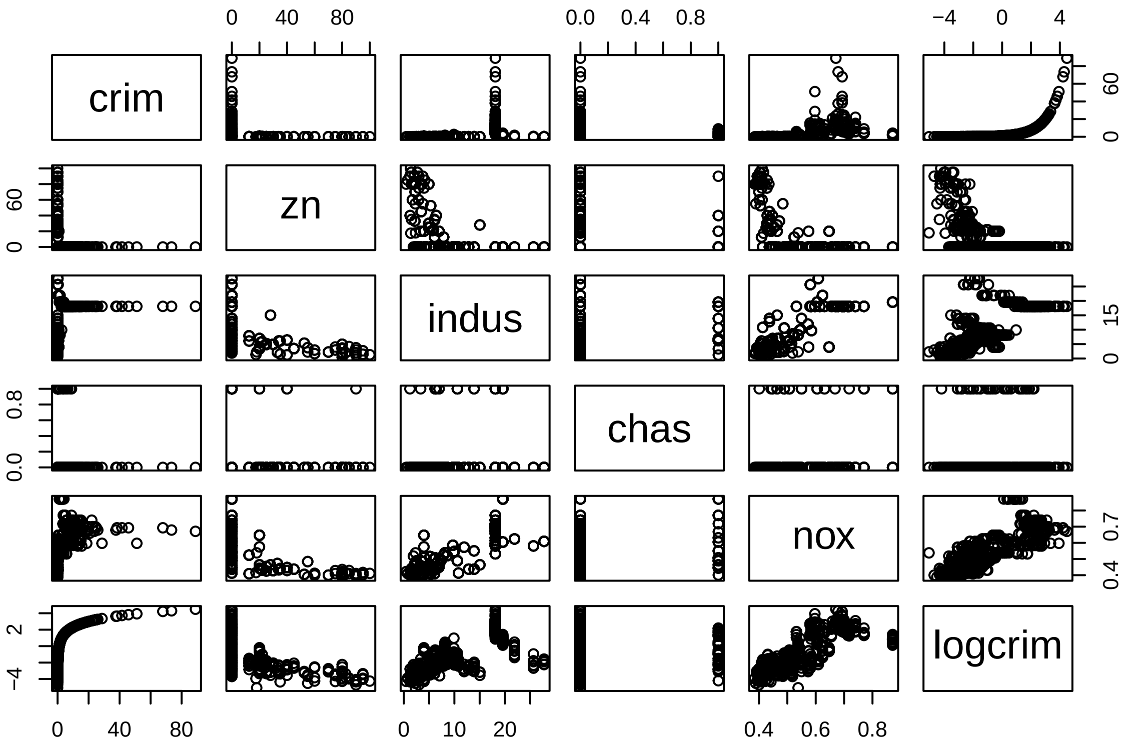
f

```
Boston %>%
keep(is.numeric) %>%
gather() %>%
ggplot(aes(value)) +
facet_wrap(~ key, scales = "free") +
geom_histogram()
```

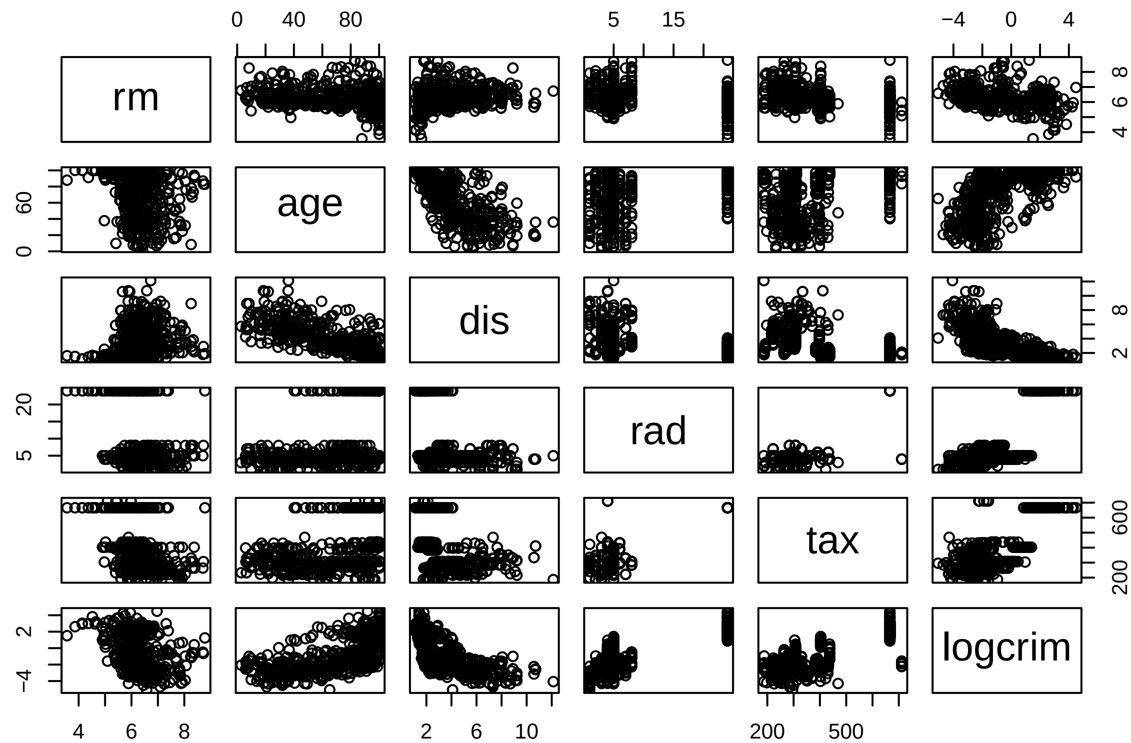
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



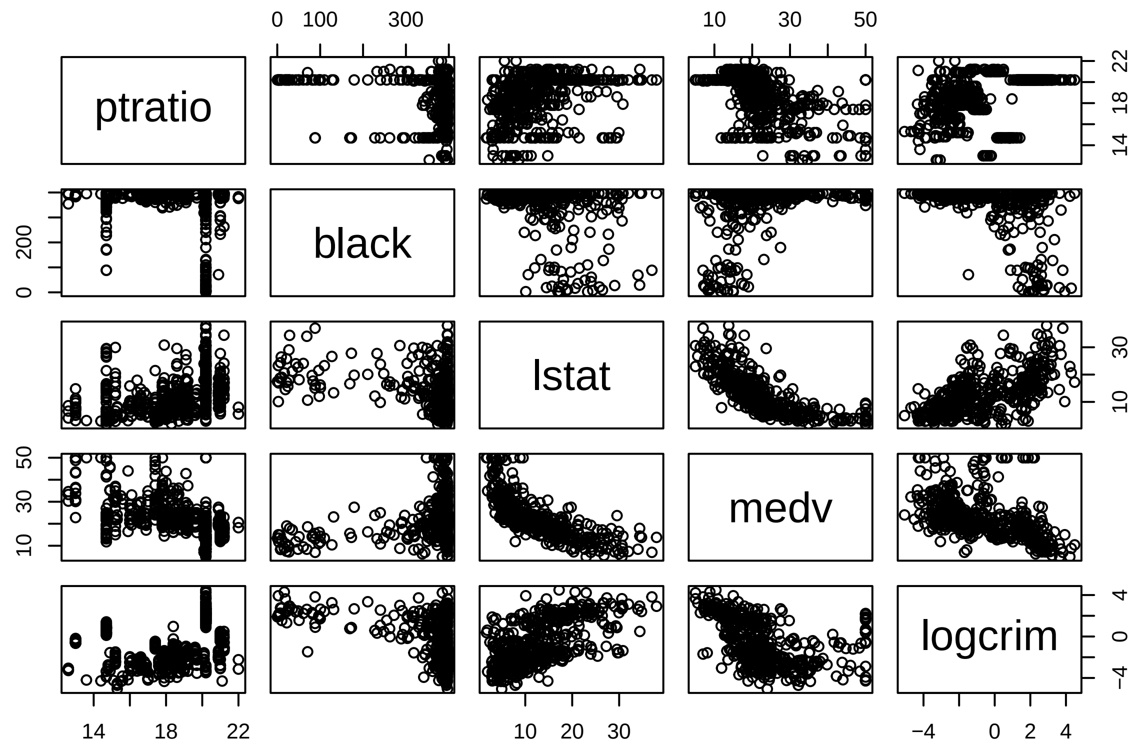

```
pairs(data.frame(Boston[, c(1,2,3,4,5, 15)]))
```




```r
pairs(data.frame(Boston[, c(6,7,8,9,10, 15)]))
```

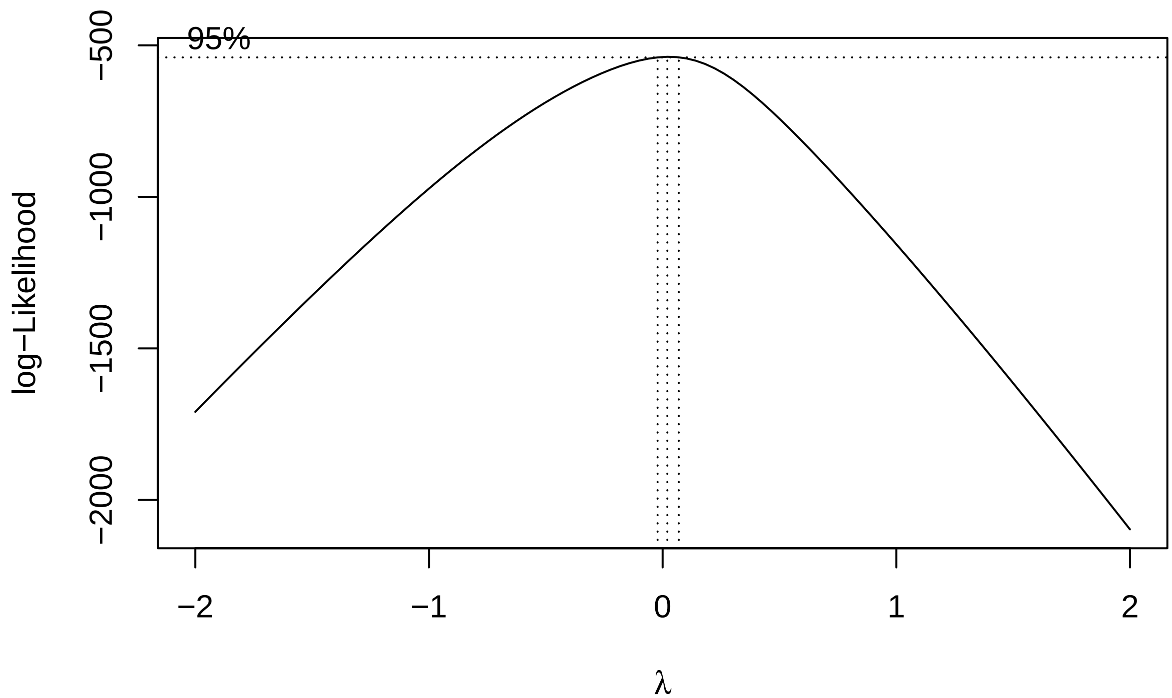



```
pairs(data.frame(Boston[, c(11,12,13,14, 15)]))
```



As seen above, many predictors are not normal. Moreover, when we look at the pairwise scatterplots, we can see there is multicolinearity between the predictors, and there are also some non-linear relationships between the dependent and independent variables. We shall use Box-Cox to fix the non-normality.

```
m <- lm(crim ~ . - logcrim, data = train_set)
boxcox(m)
```



As shown above, the most likely transformation for the y variable, crim, should be log transform. However, we already have a logcrim. So we do not need to do anything really.

```
m_baseline <- lm(cbind(train_set$indus, train_set$nox, train_set$rm, train_set$age, train_set$dis, trai
powerTransform(m_baseline)
```

```
## Estimated transformation parameters
##          Y1          Y2          Y3          Y4          Y5          Y6
##  0.66340928 -1.49518998  1.03704661  1.28392564 -0.01284176  0.24499017
##          Y7          Y8          Y9         Y10         Y11
##  0.74875195  4.77135221  3.90832134  0.06897314  0.62314319
```

```
m_full <- lm(logcrim ~ . + I(indus^0.6) + I(nox ^ -1.5) + I(rm^1.03) + I(age ^ 1.3) + log(dis) + I(rad^

backward_selection <- stepAIC(m_full, direction = "backward")
```

```
## Start:  AIC=-211.61
## logcrim ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
##     tax + ptratio + black + lstat + medv + I(indus^0.6) + I(nox^-1.5) +
##     I(rm^1.03) + I(age^1.3) + log(dis) + I(rad^0.24) + I(tax^0.7) +
##     I(ptratio^4.77) + I(black^4) + log(lstat) + I(medv^0.5)
##
##                   Df Sum of Sq    RSS     AIC
## - I(medv^0.5)      1    0.0000 69.048 -213.61
## - medv             1    0.0009 69.049 -213.61
```

```
## - log(dis)         1    0.0025 69.051 -213.60
## - dis              1    0.0316 69.080 -213.51
## - nox              1    0.0535 69.102 -213.44
## - I(rm^1.03)       1    0.3259 69.374 -212.56
## - rm               1    0.3294 69.378 -212.55
## - age              1    0.3510 69.399 -212.48
## - I(age^1.3)       1    0.5234 69.572 -211.92
## <none>                          69.048 -211.61
## - I(nox^-1.5)      1    0.9617 70.010 -210.51
## - I(rad^0.24)      1    1.2102 70.258 -209.72
## - black            1    1.2960 70.344 -209.45
## - chas             1    1.3355 70.384 -209.32
## - log(lstat)       1    1.3531 70.401 -209.26
## - lstat            1    1.4819 70.530 -208.86
## - I(tax^0.7)       1    1.8331 70.881 -207.74
## - tax              1    1.8644 70.913 -207.65
## - zn               1    2.5746 71.623 -205.41
## - I(black^4)       1    3.3578 72.406 -202.98
## - indus            1    3.7778 72.826 -201.68
## - I(indus^0.6)     1    4.8560 73.904 -198.39
## - I(ptratio^4.77)  1    6.2011 75.249 -194.35
## - ptratio          1    6.9345 75.983 -192.18
## - rad              1    6.9979 76.046 -191.99
## - crim             1   11.5317 80.580 -179.02
##
## Step:  AIC=-213.61
## logcrim ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
##     tax + ptratio + black + lstat + medv + I(indus^0.6) + I(nox^-1.5) +
##     I(rm^1.03) + I(age^1.3) + log(dis) + I(rad^0.24) + I(tax^0.7) +
##     I(ptratio^4.77) + I(black^4) + log(lstat)
##
##                   Df Sum of Sq    RSS     AIC
## - log(dis)         1    0.0025 69.051 -215.60
## - medv             1    0.0178 69.066 -215.56
## - dis              1    0.0318 69.080 -215.51
## - nox              1    0.0536 69.102 -215.44
## - I(rm^1.03)       1    0.3311 69.379 -214.54
## - rm               1    0.3346 69.383 -214.53
## - age              1    0.3511 69.399 -214.48
## - I(age^1.3)       1    0.5238 69.572 -213.92
## <none>                          69.048 -213.61
## - I(nox^-1.5)      1    0.9617 70.010 -212.51
## - I(rad^0.24)      1    1.2108 70.259 -211.72
## - black            1    1.2971 70.345 -211.44
## - chas             1    1.3439 70.392 -211.29
## - log(lstat)       1    1.5368 70.585 -210.68
## - I(tax^0.7)       1    1.8497 70.898 -209.69
## - tax              1    1.8791 70.927 -209.60
## - lstat            1    1.9817 71.030 -209.27
## - zn               1    2.5787 71.627 -207.40
## - I(black^4)       1    3.3584 72.407 -204.97
## - indus            1    3.8030 72.851 -203.60
## - I(indus^0.6)     1    4.9033 73.951 -200.25
## - I(ptratio^4.77)  1    6.2846 75.333 -196.10
```

```
## - rad              1    7.0058 76.054 -193.97
## - ptratio          1    7.0209 76.069 -193.92
## - crim             1   14.0530 83.101 -174.12
##
## Step:  AIC=-215.6
## logcrim ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
##     tax + ptratio + black + lstat + medv + I(indus^0.6) + I(nox^-1.5) +
##     I(rm^1.03) + I(age^1.3) + I(rad^0.24) + I(tax^0.7) + I(ptratio^4.77) +
##     I(black^4) + log(lstat)
##
##                   Df Sum of Sq    RSS     AIC
## - medv             1    0.0168 69.067 -217.55
## - nox              1    0.0591 69.110 -217.41
## - dis              1    0.1242 69.175 -217.20
## - I(rm^1.03)       1    0.3422 69.393 -216.50
## - rm               1    0.3459 69.397 -216.49
## - age              1    0.3678 69.419 -216.41
## - I(age^1.3)       1    0.5547 69.605 -215.81
## <none>                         69.051 -215.60
## - I(nox^-1.5)      1    0.9663 70.017 -214.49
## - I(rad^0.24)      1    1.2420 70.293 -213.61
## - black            1    1.2949 70.346 -213.44
## - chas             1    1.3568 70.407 -213.25
## - log(lstat)       1    1.5551 70.606 -212.62
## - I(tax^0.7)       1    1.8512 70.902 -211.68
## - tax              1    1.8808 70.932 -211.59
## - lstat            1    1.9792 71.030 -211.27
## - zn               1    2.7068 71.757 -208.99
## - I(black^4)       1    3.3566 72.407 -206.97
## - indus            1    4.3004 73.351 -204.07
## - I(indus^0.6)     1    5.4915 74.542 -200.46
## - I(ptratio^4.77)  1    6.3727 75.423 -197.83
## - rad              1    7.0345 76.085 -195.87
## - ptratio          1    7.0921 76.143 -195.70
## - crim             1   15.1326 84.183 -173.22
##
## Step:  AIC=-217.55
## logcrim ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
##     tax + ptratio + black + lstat + I(indus^0.6) + I(nox^-1.5) +
##     I(rm^1.03) + I(age^1.3) + I(rad^0.24) + I(tax^0.7) + I(ptratio^4.77) +
##     I(black^4) + log(lstat)
##
##                   Df Sum of Sq    RSS     AIC
## - nox              1    0.0503 69.118 -219.39
## - dis              1    0.1093 69.177 -219.20
## - I(rm^1.03)       1    0.3324 69.400 -218.47
## - rm               1    0.3362 69.404 -218.46
## - age              1    0.3701 69.438 -218.35
## - I(age^1.3)       1    0.5599 69.627 -217.74
## <none>                         69.067 -217.55
## - I(nox^-1.5)      1    0.9538 70.021 -216.48
## - I(rad^0.24)      1    1.2467 70.314 -215.54
## - black            1    1.2785 70.346 -215.44
## - chas             1    1.3751 70.442 -215.13
```

```
## - log(lstat)       1    1.5859 70.653 -214.47
## - I(tax^0.7)       1    1.8727 70.940 -213.56
## - tax              1    1.8950 70.962 -213.49
## - lstat            1    1.9700 71.037 -213.25
## - zn               1    2.6958 71.763 -210.97
## - I(black^4)       1    3.3488 72.416 -208.94
## - indus            1    4.5428 73.610 -205.28
## - I(indus^0.6)     1    5.7948 74.862 -201.50
## - I(ptratio^4.77)  1    6.3563 75.424 -199.83
## - rad              1    7.0201 76.088 -197.87
## - ptratio          1    7.1092 76.177 -197.60
## - crim             1   16.0598 85.127 -172.72
##
## Step:  AIC=-219.39
## logcrim ~ crim + zn + indus + chas + rm + age + dis + rad + tax +
##     ptratio + black + lstat + I(indus^0.6) + I(nox^-1.5) + I(rm^1.03) +
##     I(age^1.3) + I(rad^0.24) + I(tax^0.7) + I(ptratio^4.77) +
##     I(black^4) + log(lstat)
##
##                   Df Sum of Sq    RSS     AIC
## - dis              1    0.1696 69.287 -220.84
## - age              1    0.3261 69.444 -220.33
## - I(rm^1.03)       1    0.3332 69.451 -220.31
## - rm               1    0.3369 69.455 -220.30
## - I(age^1.3)       1    0.5132 69.631 -219.73
## <none>                          69.118 -219.39
## - I(rad^0.24)      1    1.2041 70.322 -217.52
## - black            1    1.2685 70.386 -217.31
## - chas             1    1.4850 70.603 -216.63
## - log(lstat)       1    1.5427 70.660 -216.44
## - I(tax^0.7)       1    1.8232 70.941 -215.56
## - tax              1    1.8457 70.963 -215.48
## - lstat            1    1.9198 71.037 -215.25
## - zn               1    3.1275 72.245 -211.47
## - I(black^4)       1    3.3517 72.469 -210.78
## - I(nox^-1.5)      1    3.9674 73.085 -208.88
## - indus            1    4.8319 73.950 -206.25
## - I(indus^0.6)     1    6.1141 75.232 -202.40
## - I(ptratio^4.77)  1    7.0996 76.217 -199.49
## - rad              1    7.1399 76.258 -199.37
## - ptratio          1    7.6124 76.730 -197.98
## - crim             1   16.1432 85.261 -174.37
##
## Step:  AIC=-220.84
## logcrim ~ crim + zn + indus + chas + rm + age + rad + tax + ptratio +
##     black + lstat + I(indus^0.6) + I(nox^-1.5) + I(rm^1.03) +
##     I(age^1.3) + I(rad^0.24) + I(tax^0.7) + I(ptratio^4.77) +
##     I(black^4) + log(lstat)
##
##                   Df Sum of Sq    RSS     AIC
## - age              1    0.2842 69.572 -221.92
## - I(rm^1.03)       1    0.3821 69.669 -221.61
## - rm               1    0.3860 69.673 -221.59
## - I(age^1.3)       1    0.4776 69.765 -221.30
```

```
## <none>                         69.287 -220.84
## - black             1    1.2364 70.524 -218.88
## - I(rad^0.24)       1    1.2858 70.573 -218.72
## - chas              1    1.4677 70.755 -218.14
## - log(lstat)        1    1.6098 70.897 -217.69
## - I(tax^0.7)        1    1.7241 71.011 -217.33
## - tax               1    1.7413 71.029 -217.28
## - lstat             1    1.9449 71.232 -216.64
## - I(black^4)        1    3.3138 72.601 -212.37
## - zn                1    3.5174 72.805 -211.75
## - indus             1    4.9053 74.193 -207.52
## - I(indus^0.6)      1    6.2219 75.509 -203.58
## - I(nox^-1.5)       1    6.3181 75.605 -203.29
## - I(ptratio^4.77)   1    6.9305 76.218 -201.48
## - rad               1    7.0734 76.361 -201.06
## - ptratio           1    7.4443 76.732 -199.98
## - crim              1   16.7283 86.016 -174.39
##
## Step:  AIC=-221.92
## logcrim ~ crim + zn + indus + chas + rm + rad + tax + ptratio +
##     black + lstat + I(indus^0.6) + I(nox^-1.5) + I(rm^1.03) +
##     I(age^1.3) + I(rad^0.24) + I(tax^0.7) + I(ptratio^4.77) +
##     I(black^4) + log(lstat)
##
##                   Df Sum of Sq    RSS     AIC
## - I(rm^1.03)        1    0.3703 69.942 -222.73
## - rm                1    0.3740 69.946 -222.72
## <none>                         69.572 -221.92
## - I(rad^0.24)       1    1.3309 70.902 -219.68
## - black             1    1.4371 71.009 -219.34
## - chas              1    1.5219 71.093 -219.07
## - I(tax^0.7)        1    1.8338 71.405 -218.09
## - tax               1    1.8363 71.408 -218.09
## - log(lstat)        1    1.9489 71.520 -217.73
## - I(age^1.3)        1    1.9715 71.543 -217.66
## - lstat             1    2.3750 71.947 -216.40
## - zn                1    3.4190 72.990 -213.18
## - I(black^4)        1    3.5945 73.166 -212.64
## - indus             1    4.9524 74.524 -208.52
## - I(nox^-1.5)       1    6.1391 75.711 -204.98
## - I(indus^0.6)      1    6.3638 75.935 -204.31
## - I(ptratio^4.77)   1    7.0414 76.613 -202.32
## - rad               1    7.0960 76.667 -202.17
## - ptratio           1    7.5329 77.104 -200.89
## - crim              1   17.5561 87.128 -173.52
##
## Step:  AIC=-222.73
## logcrim ~ crim + zn + indus + chas + rm + rad + tax + ptratio +
##     black + lstat + I(indus^0.6) + I(nox^-1.5) + I(age^1.3) +
##     I(rad^0.24) + I(tax^0.7) + I(ptratio^4.77) + I(black^4) +
##     log(lstat)
##
##                   Df Sum of Sq    RSS     AIC
## - rm                1    0.5463 70.488 -222.99
```

```
## <none>                          69.942 -222.73
## - I(rad^0.24)      1    1.3682 71.310 -220.39
## - chas             1    1.3956 71.337 -220.31
## - black            1    1.4474 71.389 -220.14
## - I(tax^0.7)       1    1.7919 71.734 -219.07
## - tax              1    1.7959 71.738 -219.05
## - I(age^1.3)       1    2.0891 72.031 -218.14
## - log(lstat)       1    3.3226 73.264 -214.34
## - lstat            1    3.6812 73.623 -213.24
## - I(black^4)       1    3.7436 73.685 -213.05
## - zn               1    3.7870 73.729 -212.92
## - indus            1    4.8584 74.800 -209.69
## - I(indus^0.6)     1    6.2245 76.166 -205.63
## - I(nox^-1.5)      1    6.3351 76.277 -205.31
## - rad              1    7.1815 77.123 -202.84
## - I(ptratio^4.77)  1    7.4285 77.370 -202.12
## - ptratio          1    8.0228 77.965 -200.41
## - crim             1   17.4919 87.434 -174.73
##
## Step:  AIC=-222.99
## logcrim ~ crim + zn + indus + chas + rad + tax + ptratio + black +
##     lstat + I(indus^0.6) + I(nox^-1.5) + I(age^1.3) + I(rad^0.24) +
##     I(tax^0.7) + I(ptratio^4.77) + I(black^4) + log(lstat)
##
##                   Df Sum of Sq    RSS     AIC
## <none>                          70.488 -222.99
## - I(rad^0.24)      1    1.2152 71.703 -221.16
## - black            1    1.5836 72.072 -220.01
## - chas             1    1.6259 72.114 -219.88
## - I(age^1.3)       1    1.6603 72.148 -219.77
## - I(tax^0.7)       1    1.6631 72.151 -219.77
## - tax              1    1.6650 72.153 -219.76
## - log(lstat)       1    2.7764 73.264 -216.34
## - lstat            1    3.4064 73.895 -214.42
## - zn               1    3.6028 74.091 -213.82
## - I(black^4)       1    3.9846 74.473 -212.67
## - indus            1    5.2094 75.698 -209.02
## - I(nox^-1.5)      1    6.5434 77.031 -205.10
## - I(indus^0.6)     1    6.7467 77.235 -204.51
## - rad              1    6.7712 77.259 -204.44
## - I(ptratio^4.77)  1    7.1540 77.642 -203.34
## - ptratio          1    7.7232 78.211 -201.70
## - crim             1   18.2122 88.700 -173.51
```

```
backward_selection
```

```
##
## Call:
## lm(formula = logcrim ~ crim + zn + indus + chas + rad + tax +
##     ptratio + black + lstat + I(indus^0.6) + I(nox^-1.5) + I(age^1.3) +
##     I(rad^0.24) + I(tax^0.7) + I(ptratio^4.77) + I(black^4) +
##     log(lstat), data = train_set)
##
## Coefficients:
```

```
##     (Intercept)            crim              zn           indus
##       2.807e+00       4.532e-02      -1.020e-02      -3.314e-01
##            chas             rad             tax         ptratio
##      -3.585e-01       2.123e-01      -2.901e-02      -5.189e-01
##           black           lstat     I(indus^0.6)     I(nox^-1.5)
##       2.650e-03       6.555e-02       1.564e+00      -5.871e-01
##      I(age^1.3)    I(rad^0.24)      I(tax^0.7) I(ptratio^4.77)
##       1.245e-03      -1.616e+00       2.368e-01       1.933e-06
##      I(black^4)      log(lstat)
##      -4.909e-11      -7.312e-01
```

```
m_backward_selection <- lm(formula = logcrim ~ crim + zn + indus + chas + rad + tax +
    ptratio + black + lstat + I(indus^0.6) + I(nox^-1.5) + I(age^1.3) +
    I(rad^0.24) + I(tax^0.7) + I(ptratio^4.77) + I(black^4) +
    log(lstat), data = train_set)
pred_backward_selection <- predict(m_backward_selection, newdata = test_set)
mse_backward <- mean((test_set$logcrim - pred_backward_selection)^2)
mse_backward
```

```
## [1] 0.4954733
```

The MSE achieved with backward selection is 0.49

```
train_X <- model.matrix(logcrim ~ . + I(indus^0.6) + I(nox ^ -1.5) + I(rm^1.03) + I(age ^ 1.3) + log(di
train_Y <- train_set$logcrim
cv_ridge <- cv.glmnet(train_X, train_Y, alpha = 0)

test_X <- model.matrix(logcrim ~ . + I(indus^0.6) + I(nox ^ -1.5) + I(rm^1.03) + I(age ^ 1.3) + log(dis
test_Y <- test_set$logcrim
best_lambda <- cv_ridge$lambda.min
m_ridge <- glmnet(train_X, train_Y, lambda = best_lambda, alpha = 0)
pred_ridge <- predict(m_ridge, s = best_lambda, newx = test_X)
mean((test_Y - pred_ridge)^2)
```

```
## [1] 0.5582034
```

The MSE achieved with ridge regression is 0.56

```
set.seed(123)
train_X <- model.matrix(logcrim ~ . + I(indus^0.6) + I(nox ^ -1.5) + I(rm^1.03) + I(age ^ 1.3) + log(di
train_Y <- train_set$logcrim
cv_lasso <- cv.glmnet(train_X, train_Y, alpha = 1)

test_X <- model.matrix(logcrim ~ . + I(indus^0.6) + I(nox ^ -1.5) + I(rm^1.03) + I(age ^ 1.3) + log(dis
test_Y <- test_set$logcrim
best_lambda <- cv_lasso$lambda.min
m_lasso <- glmnet(train_X, train_Y, lambda = best_lambda, alpha = 1)
pred_lasso <- predict(m_lasso, s = best_lambda, newx = test_X)
mean((test_Y - pred_lasso)^2)
```

```
## [1] 0.4753351
```

The MSE achieved with lasso is 0.47

As seen above, transforming the variables so that they are more normal / linear is helpful in reducing the MSE.