Homework6

2024-11-09

Question 1

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

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

```
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
```

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

```
table(train)
```

Part a

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

```
prop.table(table(train))
```

```
## train
##     FALSE      TRUE
## 0.5573123 0.4426877
```

Part B Residuals vs Fitted: There doesn't appear to be a strong non-linear pattern, which suggests that a linear model is appropriate.

Q-Q Plot: Residuals seem to follow a normal distribution fairly closely, most points lie along the line. However, there are a few points at the extremes that deviate slightly.

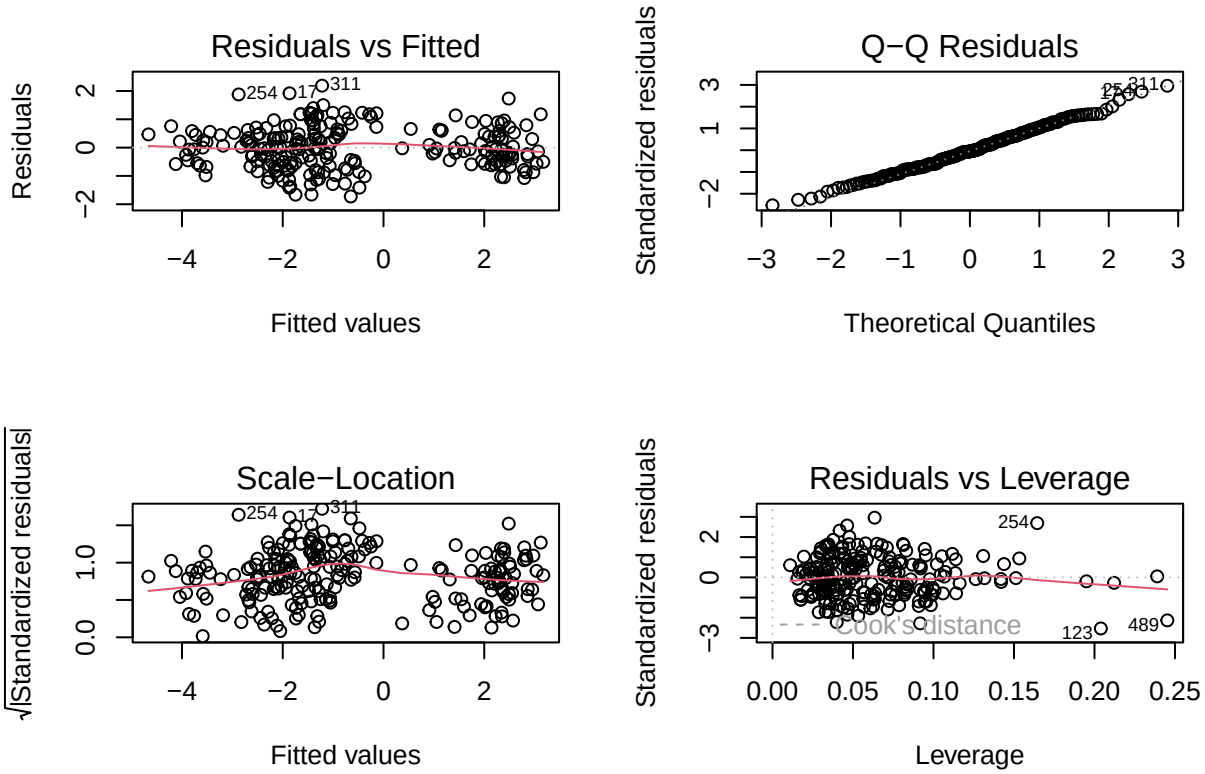
Scale-Location: The red line is fairly flat, which suggests that the variance of residuals is approximately constant, though there might be slight heteroscedasticity.

Residuals vs Leverage: A few high-leverage points appear in the plot, but they don't seem to have an excessively large influence (Cook's distance), so it's worth considering them.

```
fit <- lm(logcrim ~ . - crim, data=Boston, subset=train)

test_pred <- predict(fit, newdata=Boston[!train,])
test_mse <- mean((Boston$logcrim[!train] - test_pred)^2)

par(mfrow=c(2,2))
plot(fit)
```



Part C

```r
backward_model <- step(fit, 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
```

```
backward_pred <- predict(backward_model, newdata=Boston[!train,])
backward_mse <- mean((Boston$logcrim[!train] - backward_pred)^2)
```

```r
print(backward_mse)
```

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

```r
library(glmnet)
```

Part D

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-8
```

```r
# Prepare data for glmnet (requires matrix form)
x_train <- model.matrix(logcrim ~ . - crim, data=Boston[train,])
y_train <- Boston$logcrim[train]
x_test <- model.matrix(logcrim ~ . - crim, data=Boston[!train,])

# Fit ridge regression with cross-validation
ridge_cv <- cv.glmnet(x_train, y_train, alpha=0)

# Get optimal lambda
best_lambda_ridge <- ridge_cv$lambda.min

# Predict on test set
ridge_pred <- predict(ridge_cv, s=best_lambda_ridge, newx=x_test)
ridge_mse <- mean((Boston$logcrim[!train] - ridge_pred)^2)
print(ridge_mse)
```

```
## [1] 0.7767352
```

```r
# Fit lasso regression with cross-validation
lasso_cv <- cv.glmnet(x_train, y_train, alpha=1)

# Get optimal lambda
best_lambda_lasso <- lasso_cv$lambda.min

# Predict on test set
lasso_pred <- predict(lasso_cv, s=best_lambda_lasso, newx=x_test)
lasso_mse <- mean((Boston$logcrim[!train] - lasso_pred)^2)
print(lasso_mse)
```

Part E

```
## [1] 0.7030857
```

Part F Which transformations are important, by coming into the stepwise and/or lasso models?

Overall, log transformations (log_tax, log_rad, log_reflected_age) and non-linear transformations (sqrt_lstat, nox^3) emerge as impactful across stepwise and both Lasso, Stepwise models, revealing their importance in capturing non-linear relationships.

log_tax: This transformation has substantial coefficients in both models: Stepwise: 8.1218 Lasso: 5.1866 log_rad: Another impactful transformation with strong coefficients in both models: Stepwise: -1.2037 Lasso: -0.5105 sqrt_lstat: Although this transformation was only selected by the stepwise model with a coefficient of -1.1964, it suggests that lstat benefits from a non-linear transformation to better capture its effect.

```r
Boston$log_tax <- log(Boston$tax)  # Right skewed
Boston$log_rad <- log(Boston$rad + 1)
Boston$sqrt_zn <- sqrt(Boston$zn + 1)
Boston$sqrt_lstat <- sqrt(Boston$lstat)

max_age <- max(Boston$age) # Change skew to do log transformation
Boston$log_reflected_age <- log(max_age - Boston$age + 1)

# Cubic transformations
Boston$nox3 <- Boston$nox^3
Boston$rm3 <- Boston$rm^3

# Interaction
Boston$sqrt_lstat_rm <- Boston$sqrt_lstat * Boston$rm
```

```r
full_model_transformed <- lm(logcrim ~ . - crim, data=Boston, subset=train)
backward_model_transformed <- step(full_model_transformed, direction="backward")
```

```
## Start:  AIC=-146.69
## logcrim ~ (crim + zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat + medv + log_tax + log_rad +
##     sqrt_zn + sqrt_lstat + log_reflected_age + nox3 + rm3 + sqrt_lstat_rm) -
##     crim
##
##                      Df Sum of Sq     RSS     AIC
## - sqrt_zn             1    0.0033  95.619 -148.69
## - sqrt_lstat_rm       1    0.0116  95.627 -148.67
## - age                 1    0.0479  95.664 -148.58
## - dis                 1    0.1547  95.770 -148.33
## - sqrt_lstat          1    0.2871  95.903 -148.02
## - rm                  1    0.3114  95.927 -147.97
## - medv                1    0.4632  96.079 -147.61
## - ptratio             1    0.5697  96.186 -147.36
## - rm3                 1    0.5825  96.198 -147.33
## - zn                  1    0.8523  96.468 -146.71
## <none>                             95.616 -146.69
## - lstat               1    1.2760  96.892 -145.72
## - chas                1    1.5015  97.117 -145.20
## - indus               1    2.3625  97.978 -143.23
## - log_reflected_age   1    2.9561  98.572 -141.87
## - log_rad             1    3.6731  99.289 -140.25
## - nox3                1    4.0306  99.646 -139.44
## - black               1    4.5970 100.213 -138.17
## - nox                 1    5.9526 101.568 -135.16
## - log_tax             1   14.2739 109.890 -117.53
## - tax                 1   14.4605 110.076 -117.15
## - rad                 1   20.0147 115.630 -106.12
##
## Step:  AIC=-148.69
## logcrim ~ zn + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + black + lstat + medv + log_tax + log_rad + sqrt_lstat +
##     log_reflected_age + nox3 + rm3 + sqrt_lstat_rm
##
##                      Df Sum of Sq     RSS     AIC
```

```
## - sqrt_lstat_rm      1    0.0122  95.631 -150.66
## - age                1    0.0511  95.670 -150.57
## - dis                1    0.1527  95.772 -150.33
## - sqrt_lstat         1    0.2887  95.908 -150.01
## - rm                 1    0.3159  95.935 -149.95
## - medv               1    0.4767  96.096 -149.57
## - rm3                1    0.5926  96.212 -149.30
## - ptratio            1    0.6241  96.243 -149.23
## <none>                            95.619 -148.69
## - lstat              1    1.2744  96.893 -147.72
## - chas               1    1.5299  97.149 -147.13
## - indus              1    2.3593  97.978 -145.22
## - log_reflected_age  1    3.0052  98.624 -143.75
## - log_rad            1    3.9590  99.578 -141.60
## - nox3               1    4.0301  99.649 -141.44
## - black              1    4.5937 100.213 -140.17
## - nox                1    5.9503 101.569 -137.16
## - zn                 1    7.1110 102.730 -134.62
## - log_tax            1   14.3996 110.019 -119.26
## - tax                1   14.4819 110.101 -119.10
## - rad                1   20.4895 116.109 -107.19
##
## Step:  AIC=-150.66
## logcrim ~ zn + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + black + lstat + medv + log_tax + log_rad + sqrt_lstat +
##     log_reflected_age + nox3 + rm3
##
##                     Df Sum of Sq     RSS     AIC
## - age                1    0.0669  95.698 -152.50
## - dis                1    0.1645  95.796 -152.27
## - medv               1    0.5816  96.213 -151.30
## - ptratio            1    0.6290  96.260 -151.19
## <none>                            95.631 -150.66
## - rm3                1    1.0205  96.652 -150.28
## - rm                 1    1.4768  97.108 -149.22
## - chas               1    1.5349  97.166 -149.09
## - sqrt_lstat         1    2.2285  97.860 -147.50
## - indus              1    2.3481  97.979 -147.22
## - lstat              1    2.6613  98.293 -146.51
## - log_reflected_age  1    3.1368  98.768 -145.43
## - log_rad            1    3.9496  99.581 -143.59
## - nox3               1    4.1361  99.767 -143.17
## - black              1    4.7190 100.350 -141.87
## - nox                1    6.0519 101.683 -138.91
## - zn                 1    7.1012 102.732 -136.61
## - log_tax            1   14.3874 110.019 -121.26
## - tax                1   14.4815 110.113 -121.07
## - rad                1   20.5066 116.138 -109.14
##
## Step:  AIC=-152.5
## logcrim ~ zn + indus + chas + nox + rm + dis + rad + tax + ptratio +
##     black + lstat + medv + log_tax + log_rad + sqrt_lstat + log_reflected_age +
##     nox3 + rm3
##
```

```
##                      Df Sum of Sq     RSS     AIC
## - dis                 1    0.1417  95.840 -154.17
## - medv                1    0.5843  96.282 -153.14
## - ptratio             1    0.6116  96.310 -153.07
## <none>                             95.698 -152.50
## - rm3                 1    1.0213  96.719 -152.12
## - rm                  1    1.5246  97.223 -150.96
## - chas                1    1.5876  97.286 -150.81
## - indus               1    2.3919  98.090 -148.97
## - sqrt_lstat          1    2.5961  98.294 -148.50
## - lstat               1    2.9664  98.665 -147.66
## - log_rad             1    4.0657  99.764 -145.18
## - nox3                1    4.1006  99.799 -145.10
## - black               1    4.7033 100.401 -143.75
## - log_reflected_age   1    5.1206 100.819 -142.82
## - nox                 1    6.0762 101.774 -140.71
## - zn                  1    7.0533 102.751 -138.57
## - tax                 1   14.6458 110.344 -122.60
## - log_tax             1   14.6667 110.365 -122.56
## - rad                 1   20.7845 116.483 -110.47
##
## Step:  AIC=-154.17
## logcrim ~ zn + indus + chas + nox + rm + rad + tax + ptratio +
##     black + lstat + medv + log_tax + log_rad + sqrt_lstat + log_reflected_age +
##     nox3 + rm3
##
##                      Df Sum of Sq     RSS     AIC
## - medv                1    0.4840  96.324 -155.04
## - ptratio             1    0.6591  96.499 -154.63
## <none>                             95.840 -154.17
## - rm3                 1    0.9857  96.826 -153.88
## - rm                  1    1.4942  97.334 -152.70
## - chas                1    1.5122  97.352 -152.66
## - indus               1    2.4604  98.300 -150.49
## - sqrt_lstat          1    2.6545  98.494 -150.05
## - lstat               1    3.0715  98.911 -149.10
## - log_rad             1    4.1746 100.014 -146.62
## - black               1    4.8124 100.652 -145.19
## - log_reflected_age   1    5.6901 101.530 -143.25
## - nox3                1    5.7953 101.635 -143.02
## - zn                  1    8.4704 104.310 -137.20
## - nox                 1    9.3798 105.220 -135.25
## - tax                 1   14.5185 110.358 -124.57
## - log_tax             1   14.5280 110.368 -124.55
## - rad                 1   20.6866 116.526 -112.39
##
## Step:  AIC=-155.04
## logcrim ~ zn + indus + chas + nox + rm + rad + tax + ptratio +
##     black + lstat + log_tax + log_rad + sqrt_lstat + log_reflected_age +
##     nox3 + rm3
##
##                      Df Sum of Sq     RSS     AIC
## - ptratio             1    0.3680  96.692 -156.19
## - rm3                 1    0.6409  96.965 -155.56
```

```
## <none>                           96.324 -155.04
## - rm                1    1.1665  97.490 -154.34
## - chas              1    1.6182  97.942 -153.31
## - sqrt_lstat        1    2.2646  98.588 -151.84
## - indus             1    2.4400  98.764 -151.44
## - lstat             1    2.8078  99.132 -150.60
## - log_rad           1    4.2960 100.620 -147.27
## - black             1    4.9970 101.321 -145.71
## - nox3              1    5.4801 101.804 -144.65
## - log_reflected_age 1    5.6179 101.942 -144.34
## - zn                1    8.3865 104.710 -138.34
## - nox               1    9.1088 105.433 -136.80
## - tax               1   15.7732 112.097 -123.07
## - log_tax           1   16.2303 112.554 -122.16
## - rad               1   21.2428 117.567 -112.40
##
## Step:  AIC=-156.19
## logcrim ~ zn + indus + chas + nox + rm + rad + tax + black +
##     lstat + log_tax + log_rad + sqrt_lstat + log_reflected_age +
##     nox3 + rm3
##
##                     Df Sum of Sq     RSS     AIC
## - rm3                1    0.7468  97.439 -156.46
## <none>                            96.692 -156.19
## - rm                 1    1.2666  97.958 -155.27
## - chas               1    1.4999  98.192 -154.74
## - sqrt_lstat         1    2.3325  99.024 -152.85
## - indus              1    2.3841  99.076 -152.73
## - lstat              1    2.8558  99.548 -151.67
## - log_rad            1    4.2021 100.894 -148.66
## - black              1    5.1497 101.841 -146.56
## - nox3               1    5.2120 101.904 -146.43
## - log_reflected_age  1    5.7246 102.416 -145.30
## - zn                 1    8.0885 104.780 -140.19
## - nox                1    8.9654 105.657 -138.32
## - tax                1   16.1047 112.797 -123.68
## - log_tax            1   16.6120 113.304 -122.67
## - rad                1   20.9186 117.610 -114.32
##
## Step:  AIC=-156.46
## logcrim ~ zn + indus + chas + nox + rm + rad + tax + black +
##     lstat + log_tax + log_rad + sqrt_lstat + log_reflected_age +
##     nox3
##
##                     Df Sum of Sq     RSS     AIC
## <none>                            97.439 -156.46
## - chas               1    1.2785  98.717 -155.54
## - rm                 1    1.3329  98.771 -155.42
## - indus              1    2.2008  99.639 -153.46
## - sqrt_lstat         1    4.2209 101.659 -148.96
## - log_rad            1    4.2927 101.731 -148.81
## - lstat              1    4.9196 102.358 -147.43
## - black              1    5.5546 102.993 -146.04
## - nox3               1    5.6082 103.047 -145.93
```

```
## - log_reflected_age  1    5.9817 103.420 -145.12
## - zn                 1    8.4811 105.920 -139.77
## - nox                1    9.7404 107.179 -137.12
## - tax                1   15.8922 113.331 -124.62
## - log_tax            1   16.4107 113.849 -123.60
## - rad                1   20.9057 118.344 -114.92
```

```
backward_pred_transformed <- predict(backward_model_transformed, newdata=Boston[!train,])
backward_mse_transformed <- mean((Boston$logcrim[!train] - backward_pred_transformed)^2)
print(backward_mse_transformed)
```

```
## [1] 0.6559852
```

```
x_train_transformed <- model.matrix(logcrim ~ . - crim, data=Boston[train,])
x_test_transformed <- model.matrix(logcrim ~ . - crim, data=Boston[!train,])

# Ridge
ridge_cv_transformed <- cv.glmnet(x_train_transformed, y_train, alpha=0)
ridge_pred_transformed <- predict(ridge_cv_transformed, s=ridge_cv_transformed$lambda.min, newx=x_test_
ridge_mse_transformed <- mean((Boston$logcrim[!train] - ridge_pred_transformed)^2)
print(ridge_mse_transformed)
```

```
## [1] 0.6589886
```

```
# Lasso
lasso_cv_transformed <- cv.glmnet(x_train_transformed, y_train, alpha=1)
lasso_pred_transformed <- predict(lasso_cv_transformed, s=lasso_cv_transformed$lambda.min, newx=x_test_
lasso_mse_transformed <- mean((Boston$logcrim[!train] - lasso_pred_transformed)^2)
print(lasso_mse_transformed)
```

```
## [1] 0.5849276
```

```
print(coef(backward_model_transformed))
```

```
##       (Intercept)                zn             indus              chas
##     -41.282291007      -0.012625935       0.033759614      -0.319621061
##               nox                rm               rad               tax
##      13.647318361      -0.170280572       0.321811440      -0.024705797
##             black             lstat           log_tax           log_rad
##      -0.002090561       0.169045761       8.121837796      -1.203692642
##        sqrt_lstat log_reflected_age              nox3
##      -1.196394932      -0.196625808      -7.795277336
```

```
print(coef(lasso_cv_transformed, s=lasso_cv_transformed$lambda.min))
```

```
## 23 x 1 sparse Matrix of class "dgCMatrix"
##                              s1
## (Intercept)       -27.537428917
## (Intercept)         .
## zn                  -0.011766470
## indus                0.027786318
## chas                -0.352021644
## nox                 10.406069729
## rm                  -0.071431161
## age                 -0.001774994
## dis                 -0.039556290
## rad                  0.230920332
## tax                 -0.015848744
```

```
## ptratio            -0.035128822
## black              -0.002088398
## lstat               0.073985311
## medv               -0.015604132
## log_tax             5.186593067
## log_rad            -0.510538594
## sqrt_zn            -0.008693414
## sqrt_lstat          .
## log_reflected_age  -0.214648441
## nox3               -5.785671014
## rm3                 0.002052797
## sqrt_lstat_rm      -0.083239484
```