Predictive Models - Take Home Exam

Sam Malcolm

7/23/2018

Chapter 2 - #10

(a) How many rows are in this dataset? How many columns? What do the rows and columns represent?

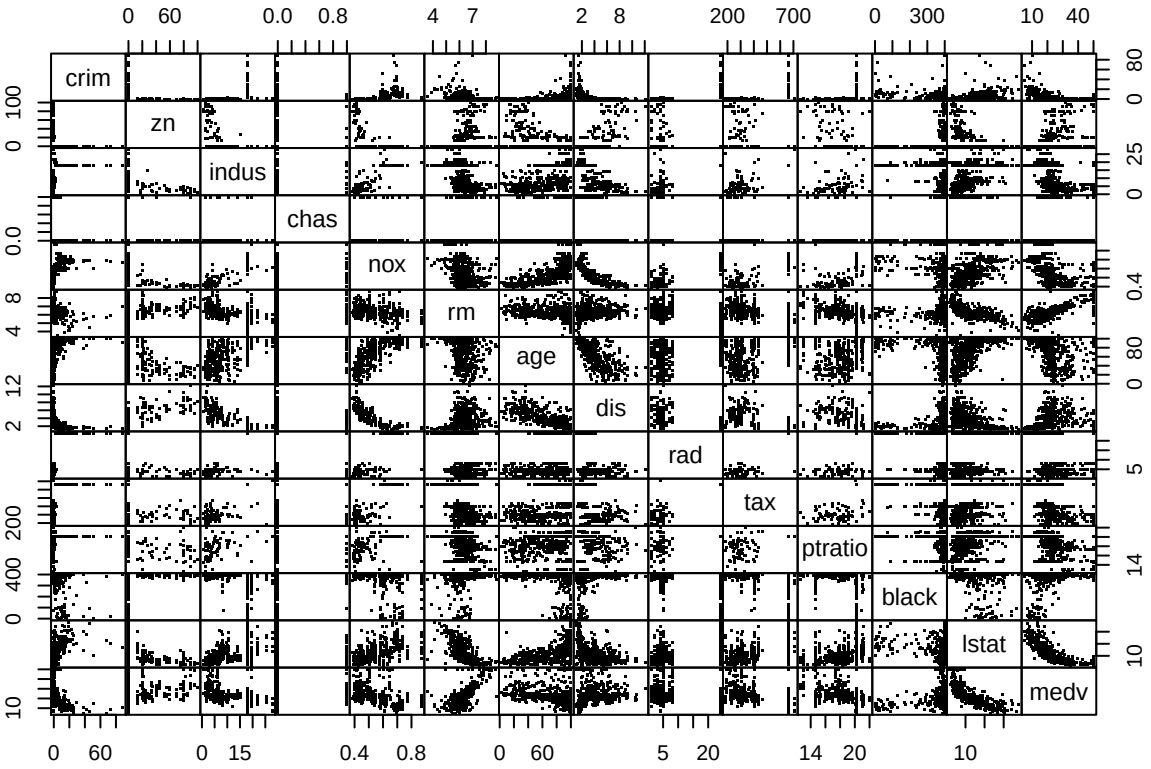
There are 506 rows and 14 columns. The rows represent observations in the sample. In this case, suburbs/neighborhoods in the Boston area. The columns represent different variables (measurements) captured for each observation.

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

(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings

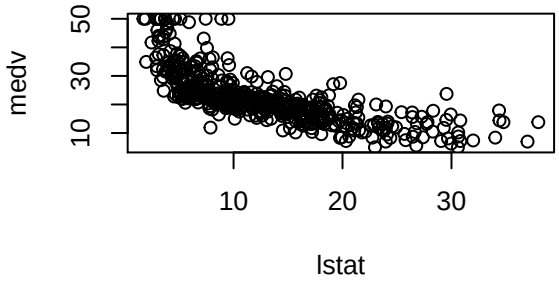
Overall, there are very few strong indications of correlation between variables. Of what we can see, it's easiest to observe linear relationships from this plot. The strongest relationships seem to be between: nox + age, dis; rm + lstat, medv; lstat + medv;

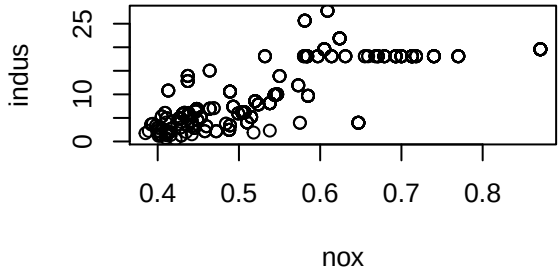
It's a bit easier to distinguish relationships via the correlation matrix. Of those, additional pairings that may be significantly correlated (>|.5|) include: crim + rad, tax; zn + indus, nox, age, dis; indus + nox, age, dis, tax, lstat; nox + rad, tax, lstat; age + dis, tax, lstat; dis + tax, lstat; rad + tax; tax + lstat; ptratio + medv

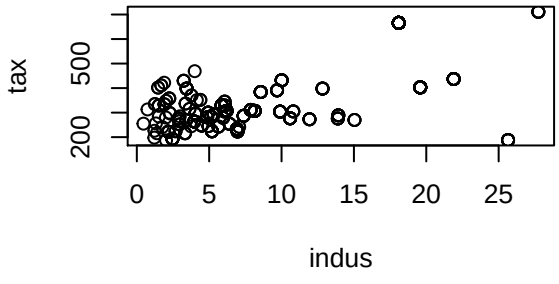


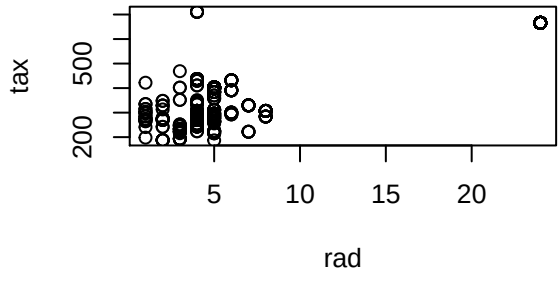
```
##          crim    zn indus  chas   nox    rm   age   dis   rad   tax
## crim     1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35 -0.38  0.63  0.58
## zn      -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57  0.66 -0.31 -0.31
## indus    0.41 -0.53  1.00  0.06  0.76 -0.39  0.64 -0.71  0.60  0.72
## chas    -0.06 -0.04  0.06  1.00  0.09  0.09  0.09 -0.10 -0.01 -0.04
## nox      0.42 -0.52  0.76  0.09  1.00 -0.30  0.73 -0.77  0.61  0.67
## rm      -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24  0.21 -0.21 -0.29
## age      0.35 -0.57  0.64  0.09  0.73 -0.24  1.00 -0.75  0.46  0.51
## dis     -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75  1.00 -0.49 -0.53
## rad      0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46 -0.49  1.00  0.91
## tax      0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91  1.00
## ptratio  0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46
## black   -0.39  0.18 -0.36  0.05 -0.38  0.13 -0.27  0.29 -0.44 -0.44
## lstat    0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60 -0.50  0.49  0.54
## medv    -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38  0.25 -0.38 -0.47
##         ptratio black lstat  medv
## crim       0.29 -0.39  0.46 -0.39
## zn        -0.39  0.18 -0.41  0.36
## indus      0.38 -0.36  0.60 -0.48
## chas      -0.12  0.05 -0.05  0.18
## nox        0.19 -0.38  0.59 -0.43
## rm        -0.36  0.13 -0.61  0.70
## age        0.26 -0.27  0.60 -0.38
## dis       -0.23  0.29 -0.50  0.25
## rad        0.46 -0.44  0.49 -0.38
## tax        0.46 -0.44  0.54 -0.47
## ptratio    1.00 -0.18  0.37 -0.51
## black     -0.18  1.00 -0.37  0.33
## lstat      0.37 -0.37  1.00 -0.74
## medv      -0.51  0.33 -0.74  1.00
```

Here are a few scatterplots of some of the stronger relationships:





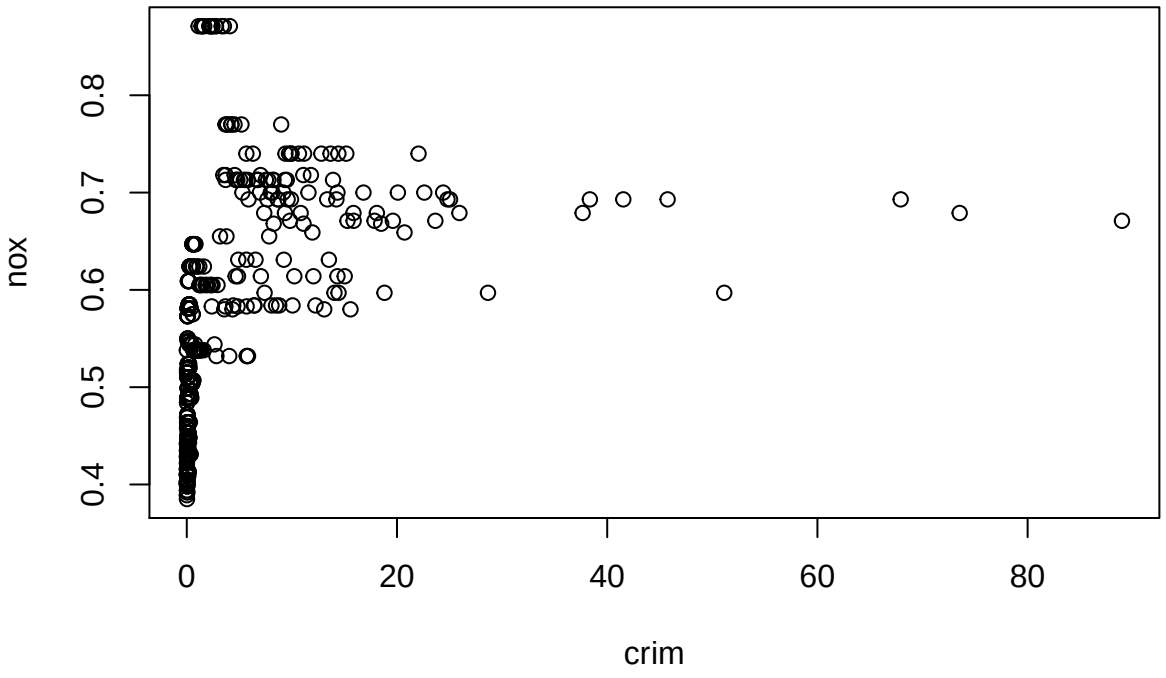


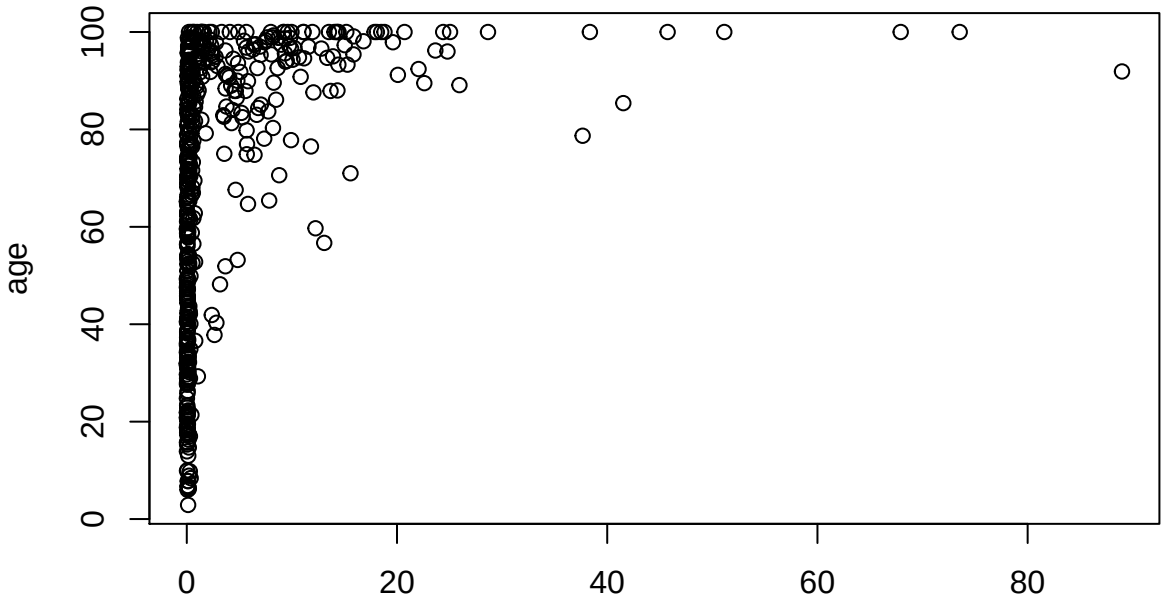


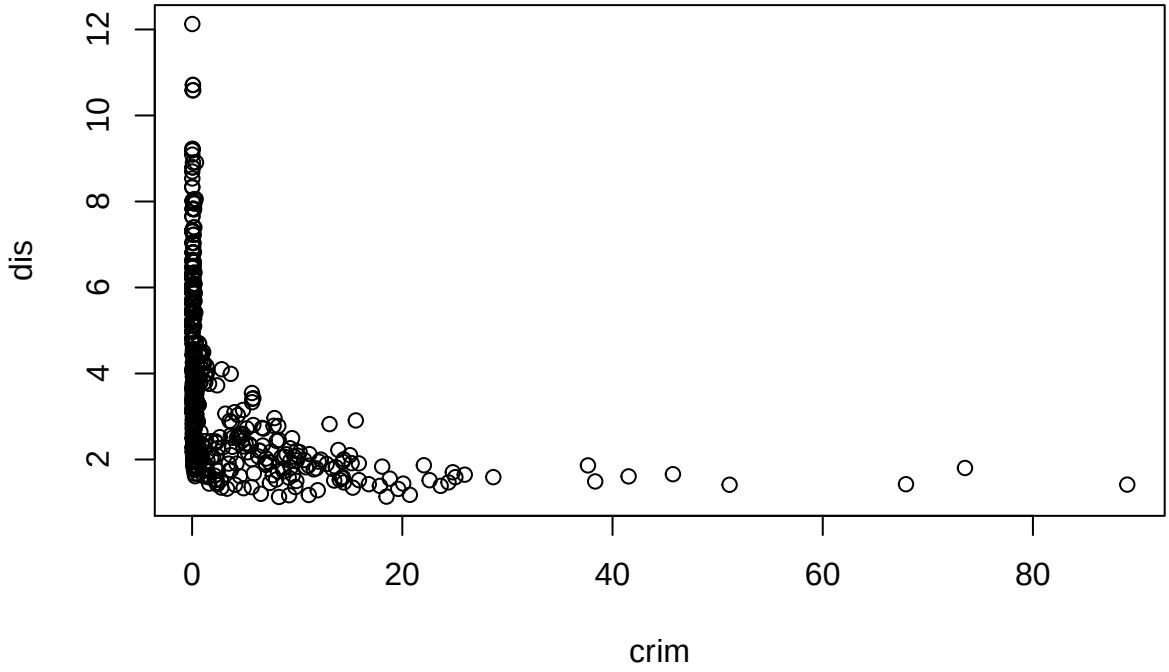
(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

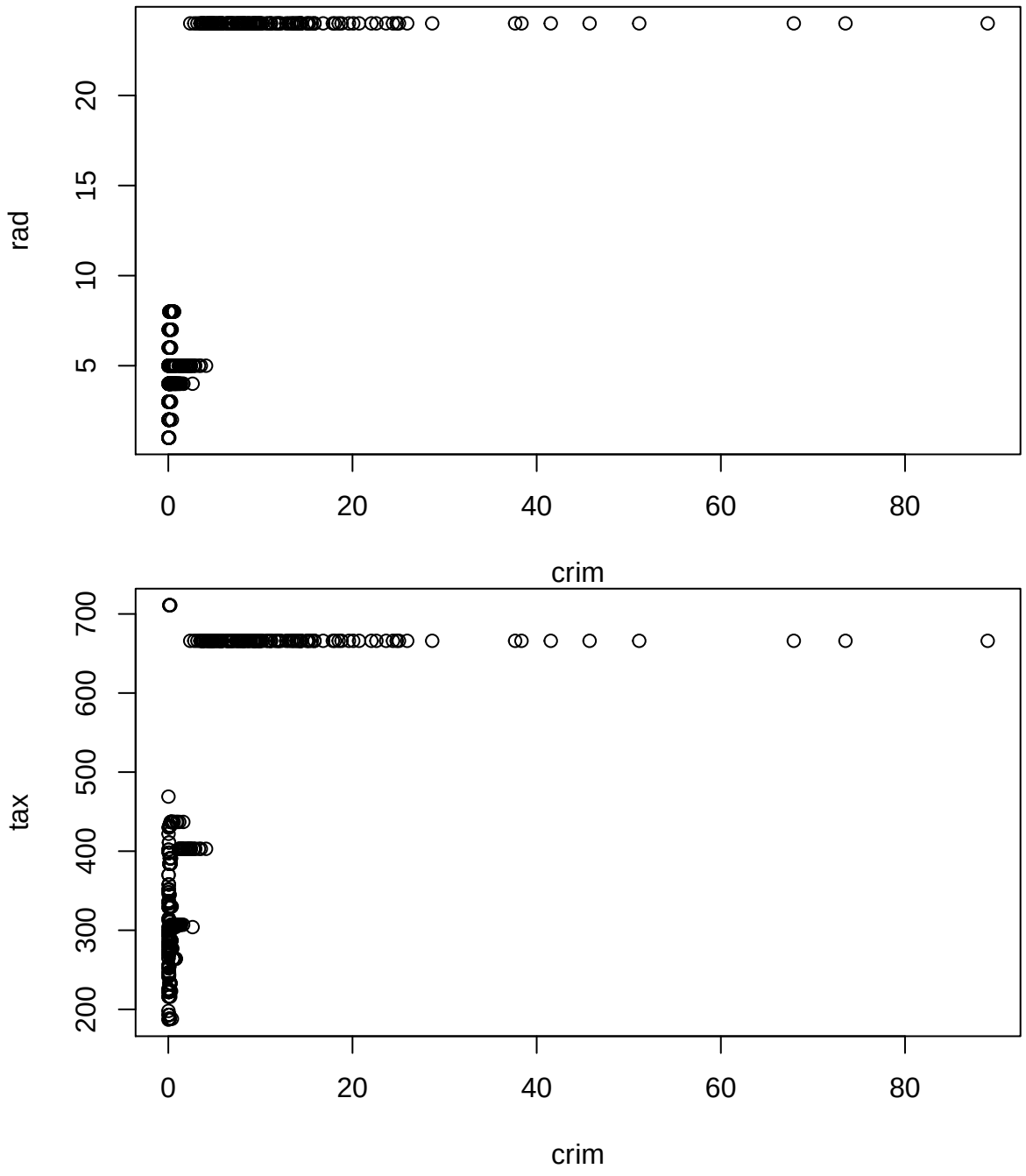
- More crime seems to appear when NO concentration is between .5 and .8
- Some tendency for crime to increase as Age proportion increases
- Less crime at greater distances to employment centres
- More crime with greater accessibility to radial highways
- More crime with greater full-value property-tax rate
- More crime with a greater lower status percent
- More crime with lower median value

In general, it seems to indicate that more crime occurs in older, poorer areas near employment centres and highways. Aka - city centers.

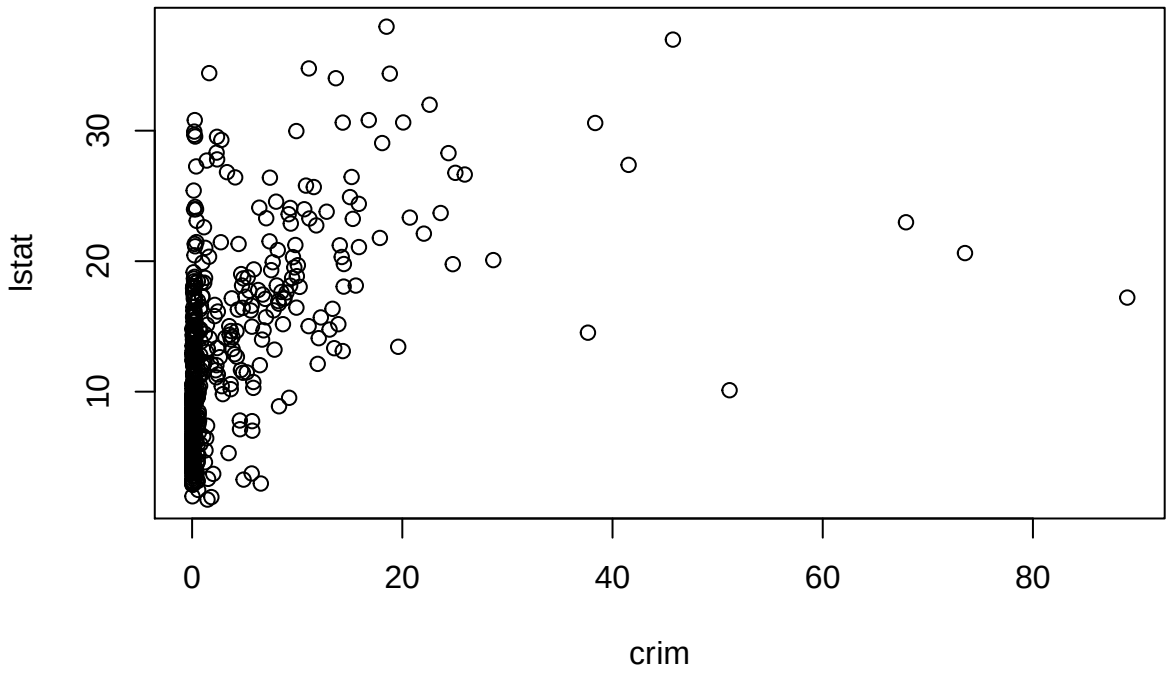



age
0
20
40
60
80
100
0
20
40
60
80
crim


dis
2
4
6
8
10
12
0
20
40
60
80
crim

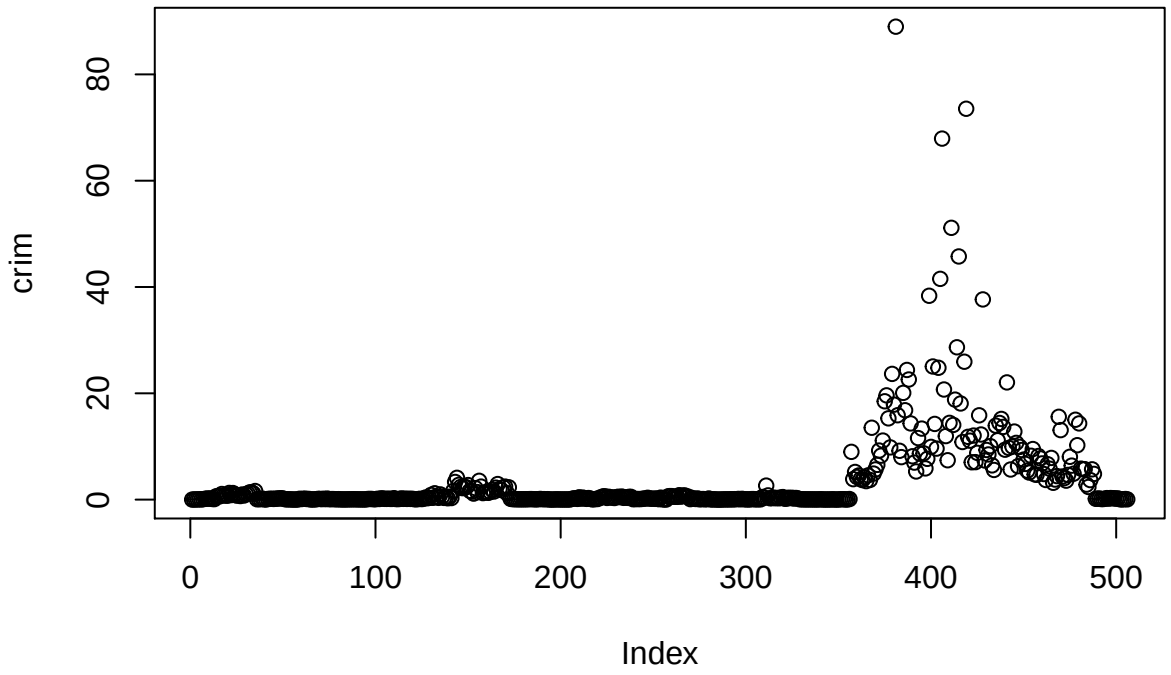

rad
5
10
15
20
0
20
40
60
80
crim

tax
200
300
400
500
600
700
0
20
40
60
80
crim

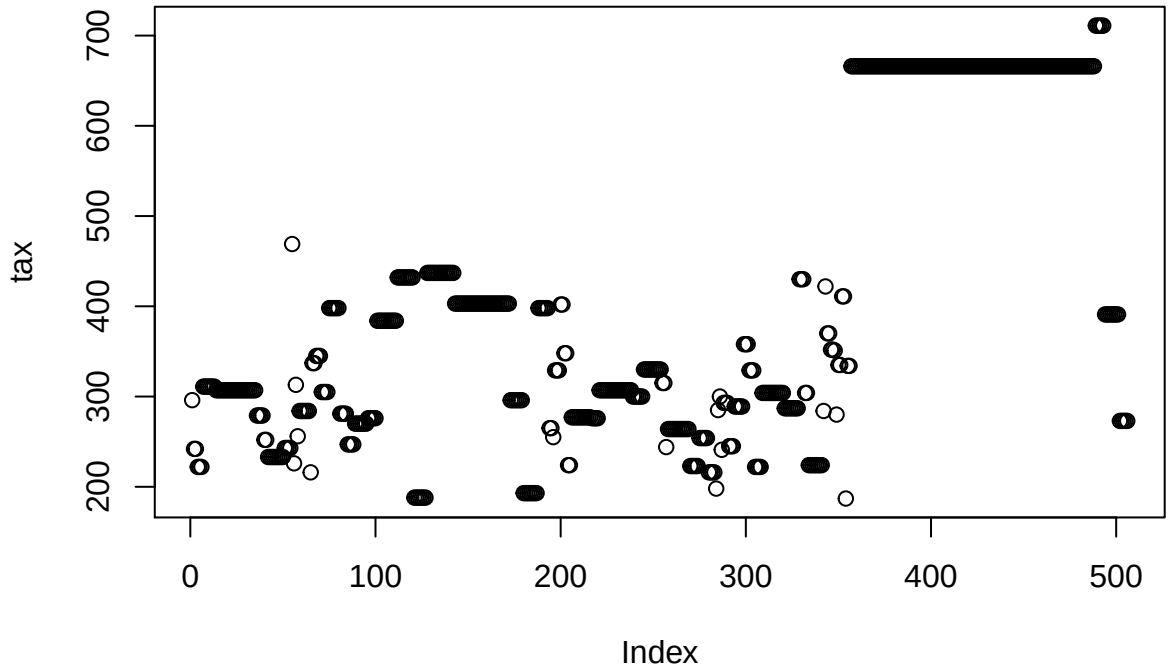


(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

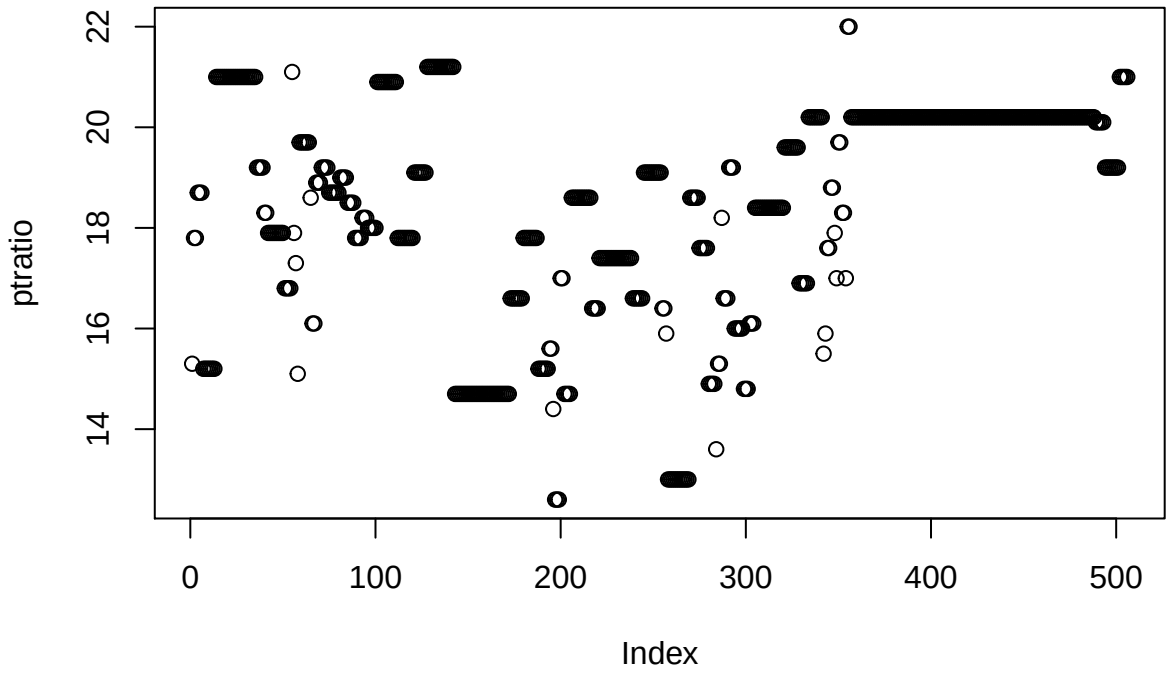
- Certain suburbs do seem to have particularly high rates of crime. There is a cluster between index 350 and 500 or so. The range of per capita crime rate per town in total is 88.97. The great majority of suburbs have crime close to 0.
- Again, there is a cluster between index 350 and 500 or so with particularly high tax rates. The range in total is 524 though most are between 200 and 500. Again, it is the one particular subset with seemingly aberrant values.
- The range of pupil-teacher ratio is a little more evenly spread. There are certain areas with lower and higher values, of course, but no discernible pattern on the whole. However, our favorite index range of 350 to 500 does have consistenly high values. I'm beginning to sense a theme from this question… Overall, the range of values is 9.4.




```
## [1] 88.96988
```




```
## [1] 524
```




```
## [1] 9.4
```

(e) How many of the suburbs in this data set bound the Charles river?

- 35 suburbs bound the Charles river

```
## [1] 35
```

(f) What is the median pupil-teacher ratio among the towns in this data set?

- The median pupil-teacher ratio is 19.05

```
## [1] 19.05
```

(g) Which suburb of Boston has lowest median value of owneroccupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

There are two suburbs that share the lowest median value of owneroccupied homes - those at index 399 and index 406.

In addition to having the lowest median value, they stand out in other ways: - Very high crime rates, age, and lower status percent; - Relatively high in proportion of non-retail business acres, nitrogen oxides concentration, access to radial highways, property tax rate, pupil-teacher ratio, and proportion of black residents; - Close to employment centres

These are most likely in poorer areas near downtown.

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   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
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

```
##              399      406
## crim     38.3518  67.9208
## zn        0.0000   0.0000
## indus    18.1000  18.1000
## chas      0.0000   0.0000
## nox       0.6930   0.6930
## rm        5.4530   5.6830
## age     100.0000 100.0000
## dis       1.4896   1.4254
## rad      24.0000  24.0000
## tax     666.0000 666.0000
## ptratio  20.2000  20.2000
## black   396.9000 384.9700
## lstat    30.5900  22.9800
## medv      5.0000   5.0000
```

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

- 64 suburbs average more than seven rooms per dwelling
- 13 suburbs average more than eight rooms per dwelling
- Of the 13 suburbs that average more than eight rooms per dwelling, we observe relatively lower crime, lower lower status, and higher median value of homes.

```
## [1] 64 14
```

```
## [1] 13 14
```

```
##       crim               zn             indus            chas
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000
```

```
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000
##       nox               rm             age              dis
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907
##       rad              tax           ptratio          black
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9
##      lstat            medv
##  Min.   :2.47   Min.   :21.9
##  1st Qu.:3.32   1st Qu.:41.7
##  Median :4.14   Median :48.3
##  Mean   :4.31   Mean   :44.2
##  3rd Qu.:5.12   3rd Qu.:50.0
##  Max.   :7.44   Max.   :50.0
```

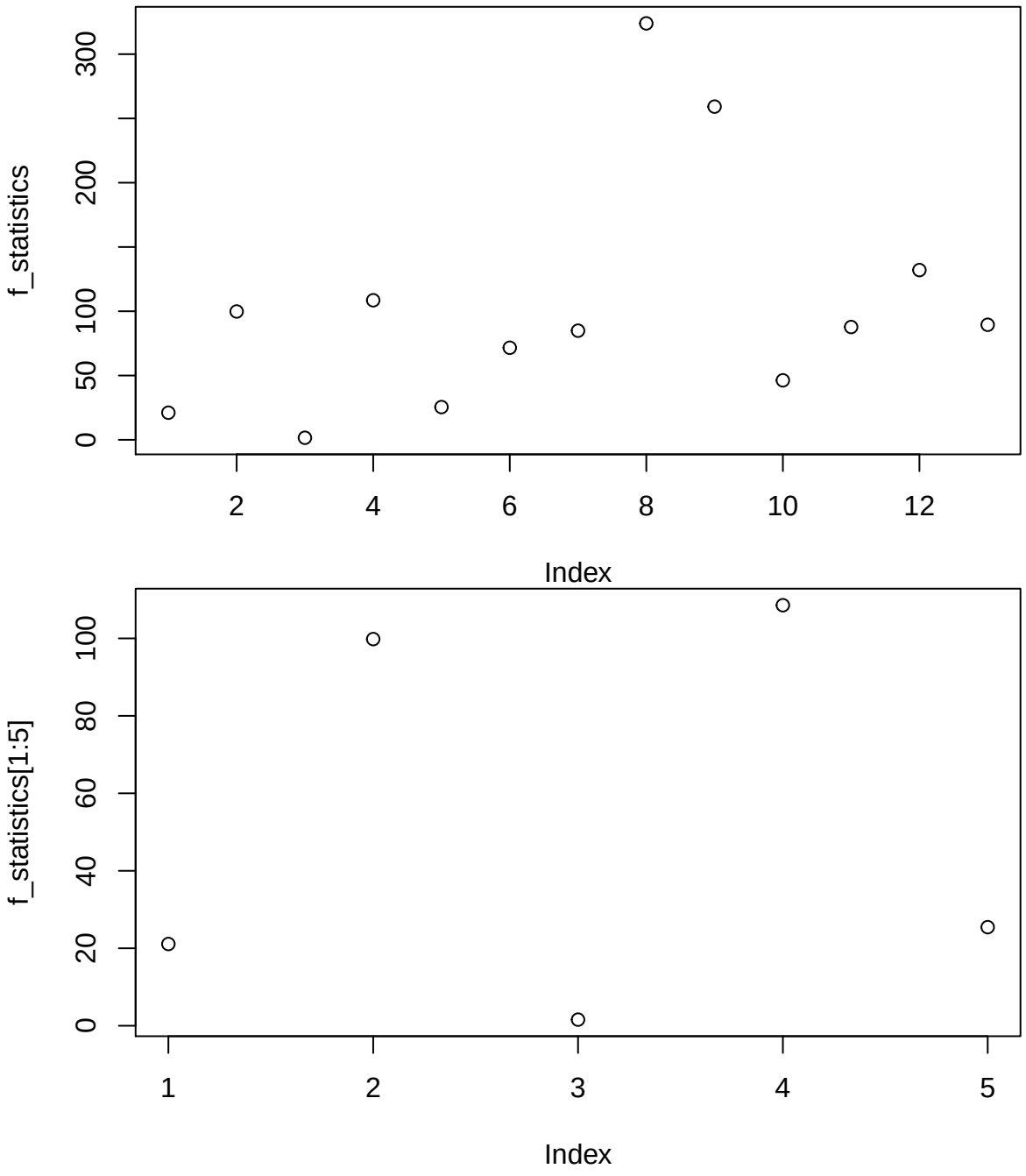
Chapter 3 - #15

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

Statistically significant relationships appear between Per Capita Crime Rate and all variables EXCEPT for whether or not the suburb tract bounds the Charles River.

Looking at the plot of F-Statistics, we can get an overall sense of what variables are most significant. When we slice to just a few, we can get a better sense of the insignificance of crim.chas against some other variables.

```
##      value      value      value      value      value      value
##  21.102782  99.817037   1.579364 108.555329  25.450204  71.619402
##      value      value      value      value      value      value
##  84.887810 323.935172 259.190294  46.259453  87.739763 132.035125
##      value
##  89.486115
```



(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 : j = 0?

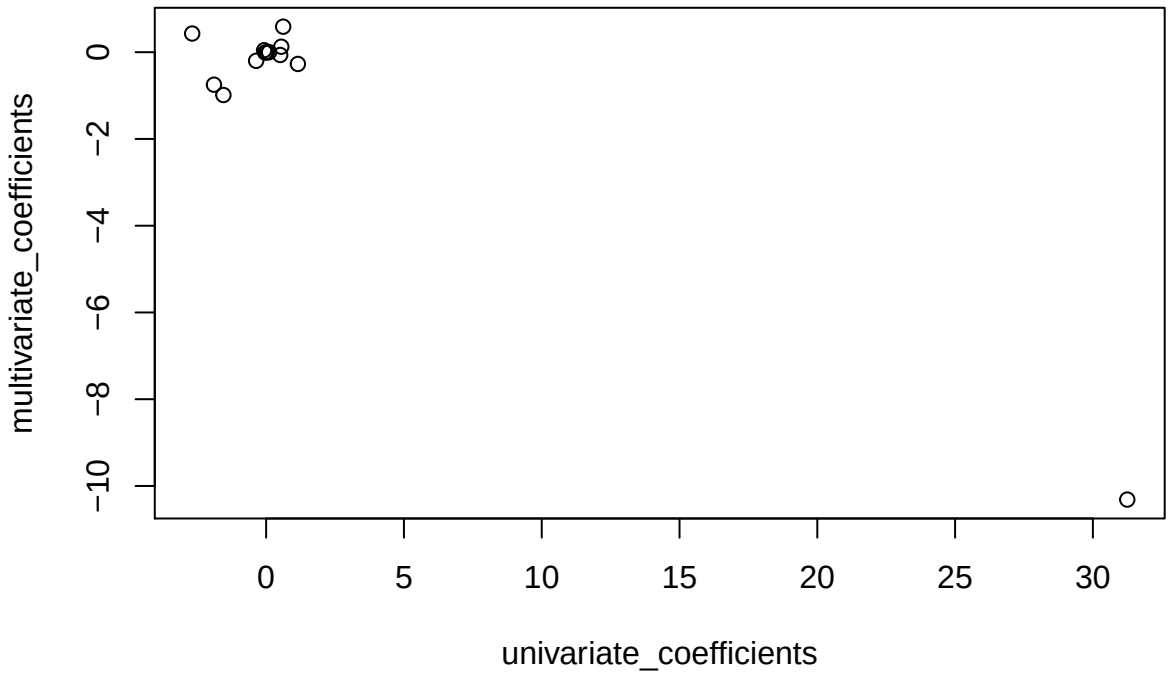
At a 95% confidence level, we can reject the null hypothesis for the zn, dis, rad, black and medv variables. Surprisingly, there are fewer variables to be included in this multiple variable model. This suggests that there is likely some collinearity between variables (which we can confirm with previous analyses).

```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

The most interesting observation is that nox's coefficient in a single model was 31.249 (indicating a positive relationship) while its coefficient in a multivariable model was -10.31 (indicating a negative relationship).



(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form $Y = 0 + 1X + 2X2 + 3X3$.

There is evidence for some degree of non-linear association in all variables but chas and black. It seems strongest (lowest p-values to the third power) with indus, nox, dis, and medv.

Chapter 6 - #9

In this exercise, we will predict the number of applications received using the other variables in the College data set.

(a) Split the data set into a training set and a test set.

(b) Fit a linear model using least squares on the training set, and report the test error obtained.

I obtained a test RMSE of 1218.857.

```
## [1] 1218.857
```

(c) Fit a ridge regression model on the training set, with chosen by cross-validation. Report the test error obtained.

I obtained a test RMSE of 1252.274.

```
## [1] 1252.274
```

(d) Fit a lasso model on the training set, with chosen by crossvalidation. Report the test error obtained, along with the number of non-zero coefficient estimates.

I obtained a test RMSE of 1245.895. There are 15 non-zero coefficient estimates when re-fitting lasso on the full dataset with the best lambda obtained via cross-validation.

```
## [1] 1245.895
```

```
##   (Intercept)    PrivateYes        Accept        Enroll     Top10perc
## -5.787986e+02 -4.373618e+02  1.473020e+00 -2.571285e-01  3.586670e+01
##     Top25perc   F.Undergrad   P.Undergrad      Outstate    Room.Board
## -3.936315e+00  0.000000e+00  2.606761e-02 -6.207031e-02  1.285693e-01
##         Books      Personal           PhD      Terminal     S.F.Ratio
##  0.000000e+00  2.702163e-03 -6.024195e+00 -3.252826e+00  6.024102e+00
##   perc.alumni        Expend     Grad.Rate
## -8.791013e-01  7.079505e-02  5.599698e+00
```

(e) Fit a PCR model on the training set, with M chosen by crossvalidation. Report the test error obtained, along with the value of M selected by cross-validation.

The test RMSE is 1218.857 The value of M selected by cross-validation is 17

```
## Data:    X dimension: 378 17
##  Y dimension: 378 1
## Fit method: svdpc
## Number of components considered: 17
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV            3616     3636     1802     1666     1402     1322     1289
## adjCV         3616     3639     1800     1509     1372     1306     1285
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV        1292     1286     1190      1130      1135      1152      1169
## adjCV     1291     1288     1185      1128      1132      1148      1165
##        14 comps  15 comps  16 comps  17 comps
## CV         1168      1149      1021      1008
## adjCV      1164      1151      1016      1002
##
## TRAINING: % variance explained
##        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X     31.05331    57.31    63.51    69.69    75.52    80.59    84.34
## Apps   0.04783    75.65    83.99    87.23    87.95    88.12    88.19
##       8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
## X       87.67    90.64     93.13     95.13     96.96     98.14     98.93
## Apps    88.46    90.05     90.89     90.90     90.91     90.91     90.98
##       15 comps  16 comps  17 comps
## X        99.44     99.87    100.00
## Apps     91.02     93.25     93.83
```

```
## [1] 1218.857
```

(f) Fit a PLS model on the training set, with M chosen by crossvalidation. Report the test error obtained, along with the value of M selected by cross-validation.

The test RMSE is 1218.175 The value of M selected by cross validation is 14

```
## Data:    X dimension: 378 17
##  Y dimension: 378 1
```

```
## Fit method: kernelpls
## Number of components considered: 17
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV            3616     1583     1161     1138     1093     1087     1043
## adjCV         3616     1580     1142     1130     1088     1071     1035
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV        1026     1022     1009      1019      1017      1015      1011
## adjCV     1019     1015     1003      1012      1010      1009      1005
##        14 comps  15 comps  16 comps  17 comps
## CV         1011      1011      1011      1011
## adjCV      1004      1005      1005      1005
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X       26.06    32.34    58.39    66.18    68.11    72.09    75.53
## Apps    81.56    90.72    91.13    92.11    93.22    93.55    93.62
##       8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
## X       79.34    81.33     84.20     87.58     91.14     93.49     96.28
## Apps    93.66    93.75     93.79     93.82     93.83     93.83     93.83
##       15 comps  16 comps  17 comps
## X        97.18     98.62    100.00
## Apps     93.83     93.83     93.83
```

```
## [1] 1218.175
```

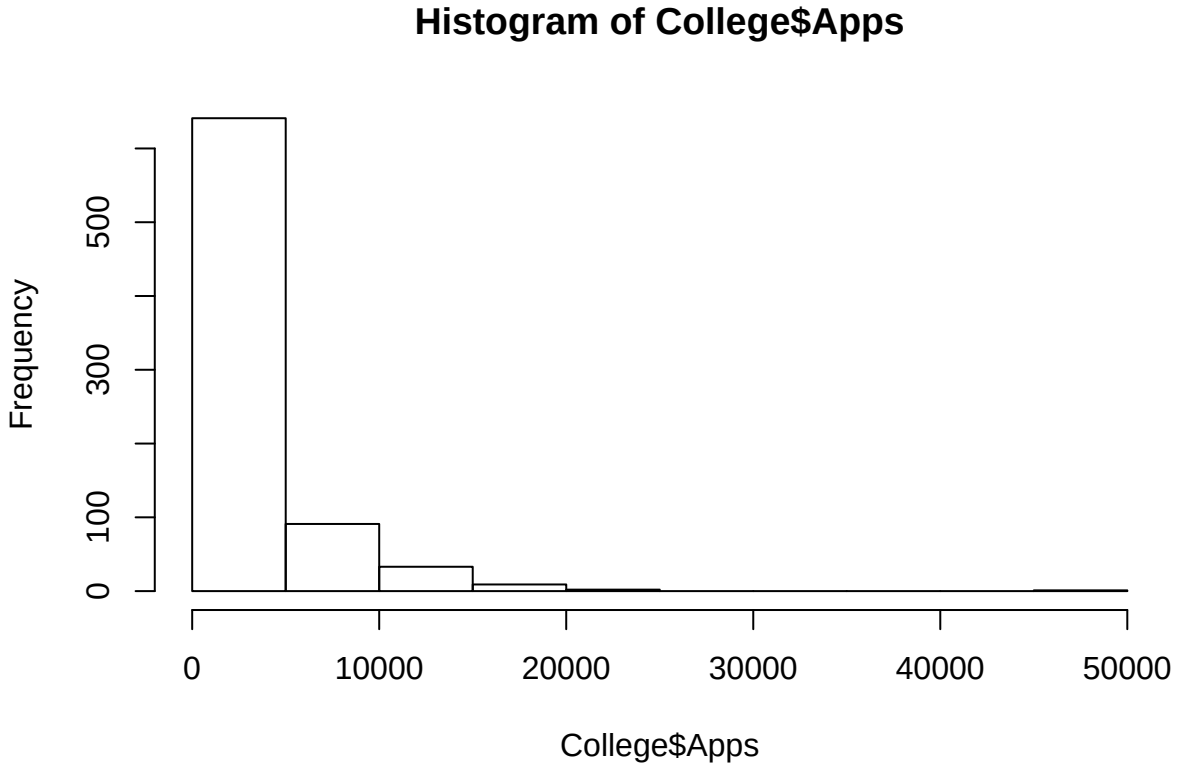
(g) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?

Our best test RMSE score (from the Partial Least Squares Model) was 1218.175. This gives us a 95% confidence interval of +- 2437.5 applications received (a range of 4875 total). This doesn't seem to lend itself to accurate predictions. As you can see from the plot, the distribution has a long right tail. This is driving our estimates higher. 75% of our observations of Applications are below 3624 – which is less than the range of our 95% confidence interval.

Overall, the test errors were very similar amongst all approaches.

```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
```

```
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```



Chapter 6 - #11

We will now try to predict the per capita crime rate in the Boston data set.

(a) Try out some of the regression methods explored in this chapter, such as best subset selection, the lasso, ridge regression, and PCR. Present and discuss results for the approaches that you consider.

1. Best Subset Selection (80/20) - The best cross-validated RMSE is 3.204642 from a 13 variable model
2. Ridge - The best cross-validated RMSE is 3.137949 with a lambda value of 0.2848036
3. Lasso - The best cross-validated RMSE is 3.189885 with a lambda value of 0.0.04641589 with Age as the only coefficient not used
4. PCR - The best cross-validated model had a test RMSE of 3.204642 with M=13
5. PLS - The best cross-validated model had a test RMSE of 3.204685 with M=8

```
## [1] 13
```

```
## [1] 3.204642
```

```
## [1] 3.137949
```

```
## [1] 3.189885
```

```
## (Intercept)          zn       indus        chas         nox          rm
## 13.47632063  0.03744603 -0.07355923 -0.60424048 -7.71567593  0.27610648
##         age         dis
##  0.00000000 -0.83112707
```

```
## Data:    X dimension: 400 13
##  Y dimension: 400 1
## Fit method: svdpc
## Number of components considered: 13
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           9.375    7.869    7.852    7.448    7.439    7.458    7.489
## adjCV        9.375    7.866    7.849    7.441    7.430    7.450    7.478
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       7.491    7.360    7.391     7.400     7.431     7.361     7.306
## adjCV    7.482    7.349    7.378     7.385     7.416     7.344     7.288
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X       47.68    60.60    70.15    76.94    83.33    88.52    91.48
## crim    30.04    30.45    38.52    39.12    39.13    39.40    39.56
##       8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## X       93.92    95.65     97.17     98.52     99.54    100.00
## crim    41.56    41.71     41.80     42.06     43.41     44.39
```

```
## [1] 3.204642
```

```
## Data:    X dimension: 400 13
##  Y dimension: 400 1
## Fit method: kernelpls
## Number of components considered: 13
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           9.375    7.671    7.324    7.306    7.280    7.244    7.249
## adjCV        9.375    7.668    7.318    7.293    7.265    7.231    7.234
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       7.241    7.222    7.227     7.225     7.226     7.226     7.226
## adjCV    7.227    7.209    7.213     7.212     7.213     7.213     7.213
##
```

```
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X       47.21    57.10    63.28    71.97    77.51    80.51    85.01
## crim    33.77    41.05    42.84    43.58    43.94    44.25    44.31
##       8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## X       86.62    89.09     93.71     96.93     98.50    100.00
## crim    44.38    44.39     44.39     44.39     44.39     44.39
```

```
## [1] 3.204685
```

(b) Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, crossvalidation, or some other reasonable alternative, as opposed to using training error.

As a whole, the models all performed with a relatively similar level of accuracy on our test set.

The model that tested best, though, was a Ridge model with a lambda value of 0.2848036 which scored a test RMSE of 3.137949. The coefficients for this model are below.

```
##  (Intercept)           zn        indus         chas          nox
##  7.398081582  0.030551955 -0.077611064 -0.747307109 -4.266487799
##           rm          age          dis          rad          tax
##  0.305104709  0.002430433 -0.632276496  0.390898900  0.004444952
##      ptratio        black        lstat         medv
## -0.104271998 -0.008663639  0.142394169 -0.127299385
```

(c) Does your chosen model involve all of the features in the data set? Why or why not?

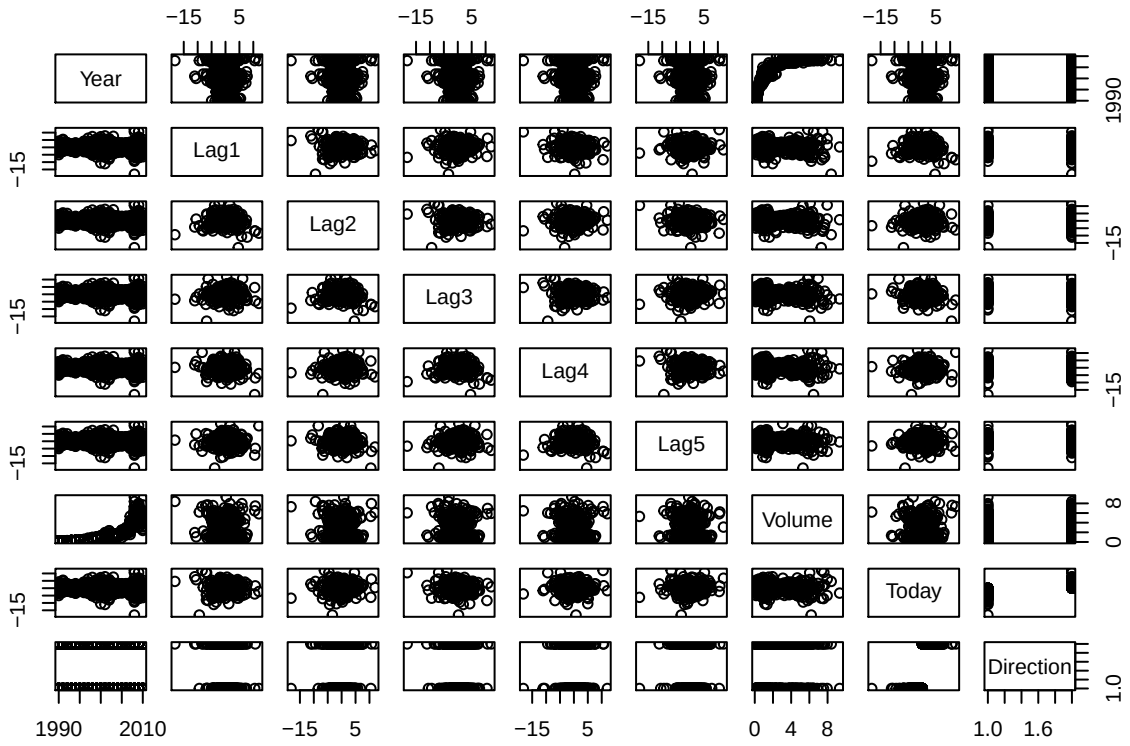
The model with the most accuracy that I've observed does include all of the features in the data set. Even though some collinearity may exist, including those variables may still produce marginal improvements in the model.

However, if I were to choose a model to present to a decision maker, I might go with a simpler model such as a 9 variable linear regression model. This would be relatively more interpretable, and the test RMSE was just .028 higher.

Chapter 4 - #10

(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

The only relationship visible from the pairwise plots appear to be between Year + Volume and between Today + Direction. These make sense. Volume of trading increases as time passes, and "Direction" is fundamentally dependent on Today's value.




```
##                Year         Lag1        Lag2        Lag3         Lag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume        Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -0.072499482 -0.08551314  0.059166717
## Lag3    0.060657175 -0.06928771 -0.071243639
## Lag4   -0.075675027 -0.06107462 -0.007825873
## Lag5    1.000000000 -0.05851741  0.011012698
## Volume -0.058517414  1.00000000 -0.033077783
## Today   0.011012698 -0.03307778  1.000000000
```

(b) Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

Lag2 is the only variable that appears statistically significant.

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

The overall proportion of correct predictions is 56.1%. The logistic regression correctly predicted the market would rise 92% of the time. However, it has a strong tendency to predict "Up", causing it to be very inaccurate when the market drops. It correctly predicted "Down" in only 11.2% of the instances where the market decreased.

```
##         Direction
## glm.pred Down  Up
##     Down   54  48
##     Up    430 557
```

```
## [1] 0.5610652
```

```
## [1] 0.9206612
```

```
## [1] 0.1115702
```

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

The overall proportion of correct predictions on the test data is 62.5%

```
##
## glm.pred.lag2 Down Up
##          Down    9  5
##          Up     34 56
```

```
## [1] 0.625
```

(e) [SKIP]

(f) [SKIP]

(g) Repeat (d) using KNN with K = 1

The overall proportion of correct predictions on the test data is 50%

```
##         test.Direction
## knn.pred Down Up
##     Down   21 30
##     Up     22 31
```

```
## [1] 0.5
```

(h) Which of these methods appears to provide the best results on this data?

The logistic regression with just Lag2 appears best. It has the overall best correct prediction rate, and it has a fewer proportion of incorrect predictions for periods where the Market declines as the logistic model without sacrificing any of the effectiveness at predicting correctly when the market rises.

(i) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

The Lag2 logistic model still appears to be the best.

However, we do find that increasing the K number in KNN can improve accuracy. With the k values I tried, I found the most accurate model to be k = 10. Although that is true only to a point. k = 10 still outperformed k = 25.

```
##
## glm.pred.interaction Down Up
##                 Down    1  1
##                 Up     42 60
```

```
## [1] 0.5865385
```

```
##           test.Direction
## knn.pred.5 Down Up
##       Down   15 20
##       Up     28 41
```

```
## [1] 0.5384615
```

```
##            test.Direction
## knn.pred.10 Down Up
##        Down   20 19
##        Up     23 42
```

```
## [1] 0.5961538
```

```
##            test.Direction
## knn.pred.25 Down Up
##        Down   20 25
##        Up     23 36
```

```
## [1] 0.5384615
```

Chapter 8 - #8

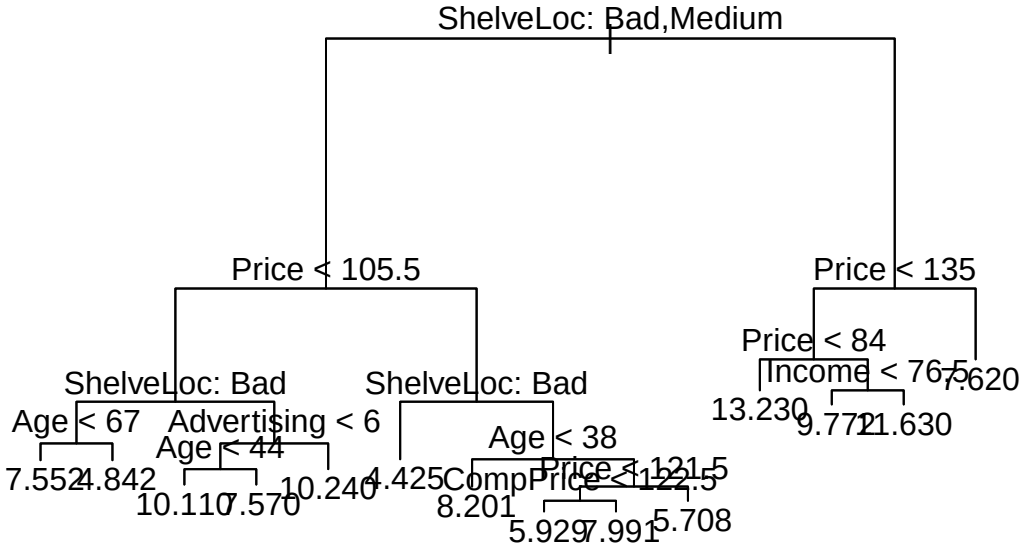
(a) Split the data set into a training set and a test set.

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

After observing the plot, Shelf Location is the most important factor. After that, Price. The tree is a bit complex at first glance, but is not un-interpretable. It has 14 nodes.

I obtained a test MSE of 5.361858

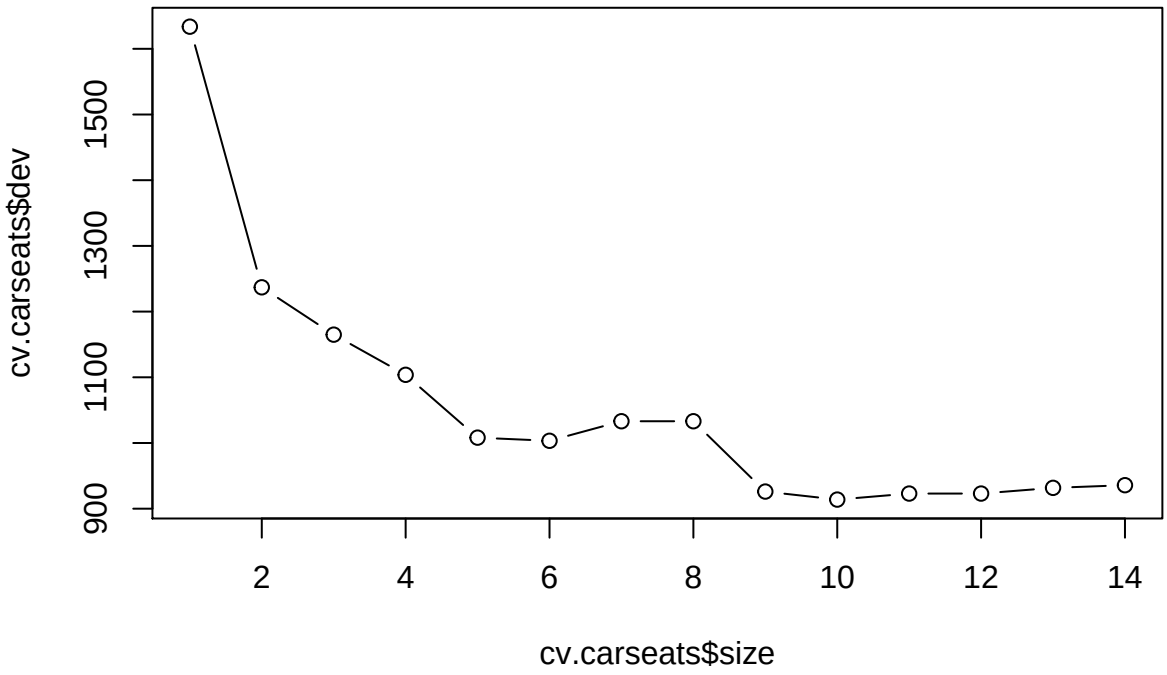
```
##
## Regression tree:
## tree(formula = Sales ~ ., data = Carseats, subset = train)
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Age"         "Advertising" "CompPrice"
## [6] "Income"
## Number of terminal nodes:  14
## Residual mean deviance:  2.438 = 453.5 / 186
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -4.0400 -1.0150  0.1300  0.0000  0.9399  4.0000
```




```
## [1] 5.361858
```

(c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

Pruning the tree to 9 nodes produced a test MSE of 5.141049. This improved the MSE.




```
## [1] 5.141049
```

(d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important.

I obtained a test MSE of 2.752007. This was a remarkable improvement.

The importance() function indicates that Shelf Location and Price are by far the most important variables.

```
## [1] 2.752007
```

```
##                %IncMSE IncNodePurity
## CompPrice   17.7454729    111.252097
## Income       7.5419044     84.916422
## Advertising 13.3191896    102.426761
## Population  -2.0848788     56.979219
## Price       56.5104498    465.903327
## ShelveLoc   62.3351737    531.696010
## Age         16.9778533    129.797312
## Education    0.8643775     39.521837
## Urban       -3.0785913      5.170353
## US           2.4220366     12.067021
```

(e) Use random forests to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

I obtained a test MSE of ~3.283942 with the default value of m=3.

The importance() function indicates once again that Shelf Location and Price are by far the most important variables.

It appears that as we increase m, our test error generally drops. This makes sense. We should expect some added accuracy as complexity increases (but only to a point).

```
## [1] 3.283942
```

```
##                %IncMSE IncNodePurity
## CompPrice    8.9922869     126.72404
## Income       1.9639561     107.60119
## Advertising  6.9957080     116.67982
## Population   0.9323326     112.15858
## Price       35.6381904     366.85253
## ShelveLoc   44.0237104     389.62473
## Age         15.5539420     164.73256
## Education    3.0458011      69.54870
## Urban       -0.5604532      12.30179
## US           3.3605021      19.01148
```

```
##           [,1]
##  [1,] 5.026738
##  [2,] 3.730712
##  [3,] 3.242268
##  [4,] 3.048820
##  [5,] 2.947395
##  [6,] 2.825886
##  [7,] 2.772054
##  [8,] 2.804948
##  [9,] 2.804086
## [10,] 2.824990
```

Chapter 8 - #11

This question uses the Caravan data set.

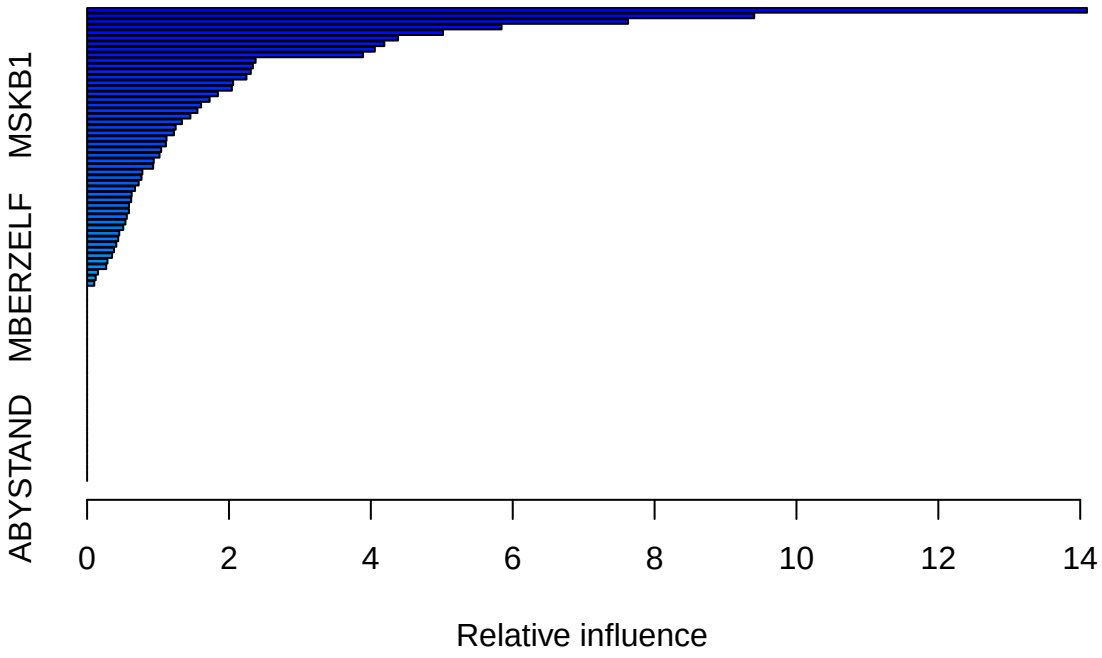
(a) Create a training set consisting of the first 1,000 observations, and a test set consisting of the remaining observations.

(b) Fit a boosting model to the training set with Purchase as the response and the other variables as predictors. Use 1,000 trees, and a shrinkage value of 0.01. Which predictors appear to be the most important?

PPERSAUT (Contribution car policies), MKOOPKLA (Purchasing power class), and MOPLHOOG (High level education) appear to be the most important.

```
## Warning in gbm.fit(x, y, offset = offset, distribution = distribution, w =
## w, : variable 50: PVRAAUT has no variation.
```

```
## Warning in gbm.fit(x, y, offset = offset, distribution = distribution, w =
## w, : variable 71: AVRAAUT has no variation.
```




```
##                var    rel.inf
## PPERSAUT PPERSAUT 14.0960413
## MKOOPKLA MKOOPKLA  9.4049151
## MOPLHOOG MOPLHOOG  7.6259448
## PBRAND     PBRAND  5.8425758
## MBERMIDD MBERMIDD  5.0180029
## MINK3045 MINK3045  4.3821264
## MGODGE     MGODGE  4.1904883
## MOSTYPE   MOSTYPE  4.0560767
## ABRAND     ABRAND  3.8901280
## MAUT1       MAUT1  2.3752928
## MSKA         MSKA  2.3376690
## PWAPART   PWAPART  2.3079791
## MSKC         MSKC  2.2466985
## MBERARBG MBERARBG  2.0585278
## MGODOV     MGODOV  2.0407739
## MAUT2       MAUT2  1.8437774
## MGODPR     MGODPR  1.7277296
## MSKB1       MSKB1  1.6075540
## PBYSTAND PBYSTAND  1.5559578
## MFGEKIND MFGEKIND  1.4562351
## MBERHOOG MBERHOOG  1.3368334
## MRELGE     MRELGE  1.2486231
## MFWEKIND MFWEKIND  1.2245165
## MINKGEM   MINKGEM  1.1195023
## APERSAUT APERSAUT  1.1113014
## MGODRK     MGODRK  1.0442135
## MINK7512 MINK7512  1.0208539
## MAUT0       MAUT0  0.9399860
## MRELOV     MRELOV  0.9313128
## MINKM30   MINKM30  0.7791627
## MSKD         MSKD  0.7669811
## MGEMLEEF MGEMLEEF  0.7271207
```

```
## MHKOOP     MHKOOP  0.6757084
## MRELSA     MRELSA  0.6282492
## MOPLMIDD MOPLMIDD  0.6213145
## MGEMOMV   MGEMOMV  0.5899735
## PMOTSCO   PMOTSCO  0.5897519
## MINK4575 MINK4575  0.5599583
## MBERBOER MBERBOER  0.5409664
## MZPART     MZPART  0.5097208
## MZFONDS   MZFONDS  0.4533915
## MOSHOOFD MOSHOOFD  0.4391056
## PLEVEN     PLEVEN  0.4117381
## MSKB2       MSKB2  0.3798373
## MBERARBO MBERARBO  0.3518454
## MFALLEEN MFALLEEN  0.2871957
## MHHUUR     MHHUUR  0.2697150
## MINK123M MINK123M  0.1540271
## MBERZELF MBERZELF  0.1216540
## MOPLLAAG MOPLLAAG  0.1009455
## MAANTHUI MAANTHUI  0.0000000
## PWABEDR   PWABEDR  0.0000000
## PWALAND   PWALAND  0.0000000
## PBESAUT   PBESAUT  0.0000000
## PVRAAUT   PVRAAUT  0.0000000
## PAANHANG PAANHANG  0.0000000
## PTRACTOR PTRACTOR  0.0000000
## PWERKT     PWERKT  0.0000000
## PBROM       PBROM  0.0000000
## PPERSONG PPERSONG  0.0000000
## PGEZONG   PGEZONG  0.0000000
## PWAOREG   PWAOREG  0.0000000
## PZEILPL   PZEILPL  0.0000000
## PPLEZIER PPLEZIER  0.0000000
## PFIETS     PFIETS  0.0000000
## PINBOED   PINBOED  0.0000000
## AWAPART   AWAPART  0.0000000
## AWABEDR   AWABEDR  0.0000000
## AWALAND   AWALAND  0.0000000
## ABESAUT   ABESAUT  0.0000000
## AMOTSCO   AMOTSCO  0.0000000
## AVRAAUT   AVRAAUT  0.0000000
## AAANHANG AAANHANG  0.0000000
## ATRACTOR ATRACTOR  0.0000000
## AWERKT     AWERKT  0.0000000
## ABROM       ABROM  0.0000000
## ALEVEN     ALEVEN  0.0000000
## APERSONG APERSONG  0.0000000
## AGEZONG   AGEZONG  0.0000000
## AWAOREG   AWAOREG  0.0000000
## AZEILPL   AZEILPL  0.0000000
## APLEZIER APLEZIER  0.0000000
## AFIETS     AFIETS  0.0000000
## AINBOED   AINBOED  0.0000000
## ABYSTAND ABYSTAND  0.0000000
```

(c) Use the boosting model to predict the response on the test data. Predict that a person will make a purchase if the estimated probability of purchase is greater than 20 %. Form a confusion matrix. What fraction of the people predicted to make a purchase do in fact make one? How does this compare with the results obtained from applying KNN or logistic regression to this data set?

21.29% of people predicted to make a purchase did make one.

It outperformed a KNN with k=1 (~9%), k=3 (~3%), and k=5 (~9%).

It performed very similarly to a logistic regression which correctly predicted 20% of purchasers.

```
##
## boost.pred    0    1
##          0 4411  256
##          1  122   33
```

```
## [1] 0.2129032
```

```
##         test.Purchase
## knn.pred    0    1
##        0 4262  263
##        1  271   26
```

```
##          test.Purchase
## knn.pred3    0    1
##         0 4469  281
##         1   64    8
```

```
##         test.Purchase
## knn.pred    0    1
##        0 4262  263
##        1  271   26
```

```
## [1] 0.0899654
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
```

```
##
## log.pred    0    1
##        0 4183  231
##        1  350   58
```

```
## [1] 0.200692
```

Problem 1: Beauty Pays!

1. Using the data, estimate the effect of "beauty" into course ratings. Make sure to think about the potential many "other determinants". Describe your analysis and your conclusions.

Overall, "beauty" is the most influential among the variables recorded. In general, the pattern we see is that a higher Beauty Score does indicate higher Course Evals. In our most accurate model (full multivariable linear regression) with all else equal, a 1 unit increase in BeautyScore is predicted to lead to a 0.30415 increase in Course Evals.

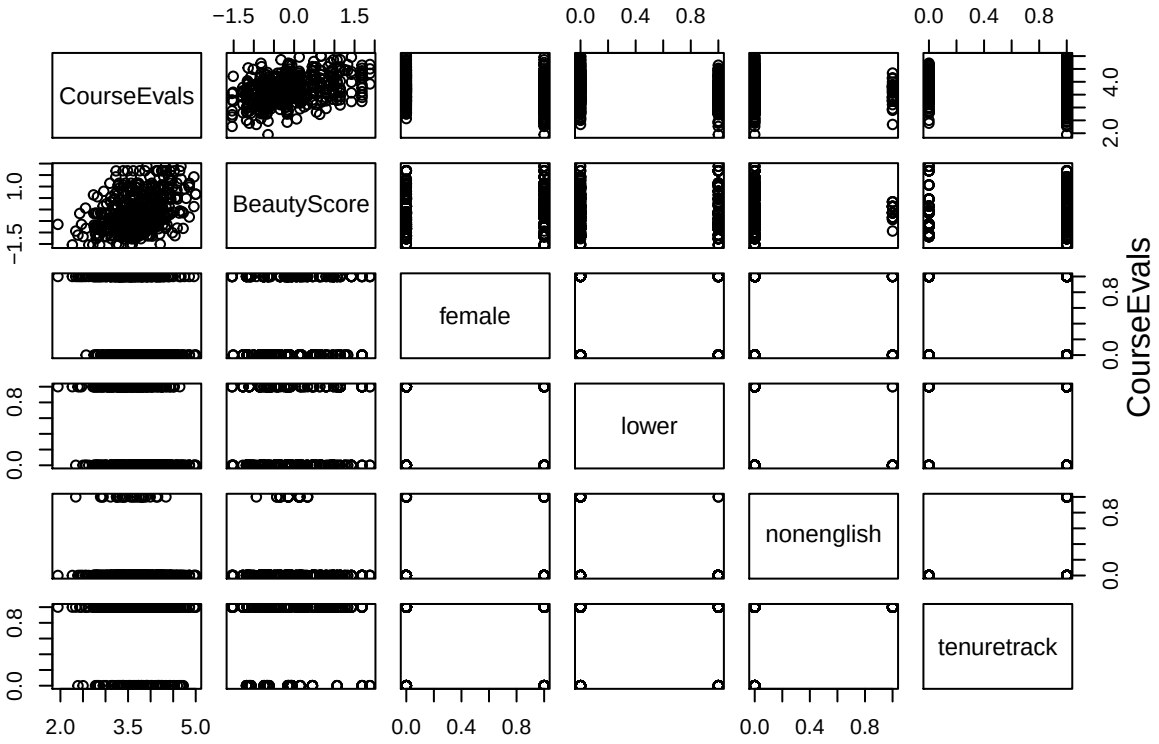
However, the data is fairly spread, so it would be difficult to predict with precision. Test RMSE's in various modeling methods ranged from 0.4477573 to 0.480466. This gives us a 95% confidence interval with a range of ~1.79. The IQR for CourseEvals is just .741.

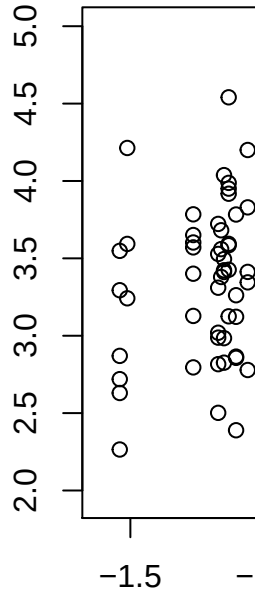
We can also see the effect the other variables have. The presence of Female, Lower, Nonenglish, and Tenure-track generally will indicate a lower CourseEval at the same level of BeautyScore. All of these variables were considered statistically significant in the model. This indicates either some potential bias against these factors or potential contextually related factors that aren't controlled (such as course level and class size) that could be relevant.

EDA

There isn't a particularly strong correlation between any of the variables. The strongest (at 0.41) is between CourseEvals and BeautyScore

```
##   CourseEvals     BeautyScore          female           lower
##  Min.   :1.944   Min.   :-1.53884   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:3.326   1st Qu.:-0.74462   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :3.682   Median :-0.15636   Median :0.0000   Median :0.0000
##  Mean   :3.689   Mean   :-0.08835   Mean   :0.4212   Mean   :0.3391
##  3rd Qu.:4.067   3rd Qu.: 0.45725   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5.000   Max.   : 1.88167   Max.   :1.0000   Max.   :1.0000
##    nonenglish       tenuretrack
##  Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:1.0000
##  Median :0.00000   Median :1.0000
##  Mean   :0.06048   Mean   :0.7797
##  3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.0000
```






```
##              CourseEvals BeautyScore female lower nonenglish tenuretrack
## CourseEvals         1.00        0.41  -0.23 -0.25      -0.08       -0.04
```

```
## BeautyScore        0.41        1.00   0.13  0.03       0.01       -0.02
## female            -0.23        0.13   1.00 -0.06       0.00       -0.07
## lower             -0.25        0.03  -0.06  1.00      -0.14       -0.14
## nonenglish        -0.08        0.01   0.00 -0.14       1.00        0.13
## tenuretrack       -0.04       -0.02  -0.07 -0.14       0.13        1.00
```

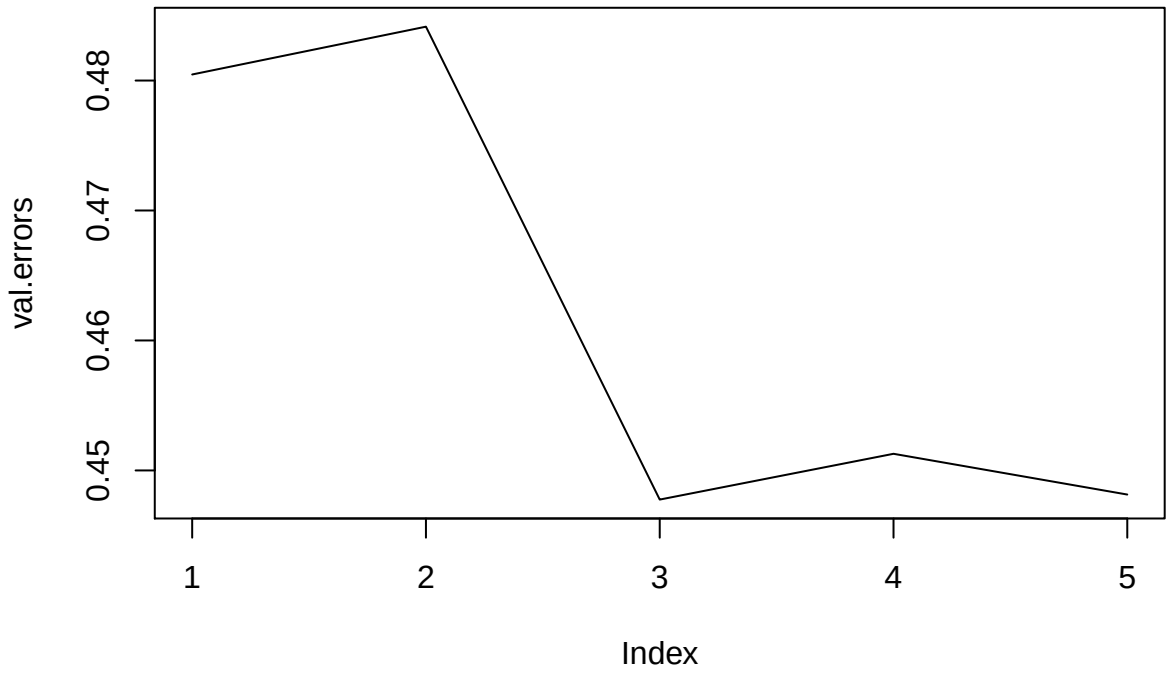
Modeling

```
##
## Call:
## lm(formula = CourseEvals ~ BeautyScore, data = beauty.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60503 -0.33345  0.01421  0.36167  1.21954
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.72806    0.02543 146.607  < 2e-16 ***
## BeautyScore  0.27646    0.03430   8.059 1.09e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4813 on 368 degrees of freedom
## Multiple R-squared:   0.15,  Adjusted R-squared:  0.1477
## F-statistic: 64.95 on 1 and 368 DF,  p-value: 1.085e-14
```

```
## [1] 0.480466
```

```
## Subset selection object
## Call: regsubsets.formula(CourseEvals ~ ., beauty.train)
## 5 Variables  (and intercept)
##             Forced in Forced out
## BeautyScore     FALSE      FALSE
## female          FALSE      FALSE
## lower           FALSE      FALSE
## nonenglish      FALSE      FALSE
## tenuretrack     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          BeautyScore female lower nonenglish tenuretrack
## 1  ( 1 ) "*"         " "    " "   " "         " "
## 2  ( 1 ) "*"         "*"    " "   " "         " "
## 3  ( 1 ) "*"         "*"    "*"   " "         " "
## 4  ( 1 ) "*"         "*"    "*"   "*"         " "
## 5  ( 1 ) "*"         "*"    "*"   "*"         "*"
```

```
## [1] 0.4804660 0.4841433 0.4477573 0.4512796 0.4481454
```




```
##
## Call:
## lm(formula = CourseEvals ~ BeautyScore + female + lower, data = beauty)
##
## Coefficients:
## (Intercept)  BeautyScore       female        lower
##      3.9584       0.3031      -0.3244      -0.3115
```

```
##
## Call:
## lm(formula = CourseEvals ~ ., data = beauty)
##
## Coefficients:
## (Intercept)  BeautyScore       female        lower   nonenglish
##     4.06542      0.30415     -0.33199     -0.34255     -0.25808
## tenuretrack
##    -0.09945
```

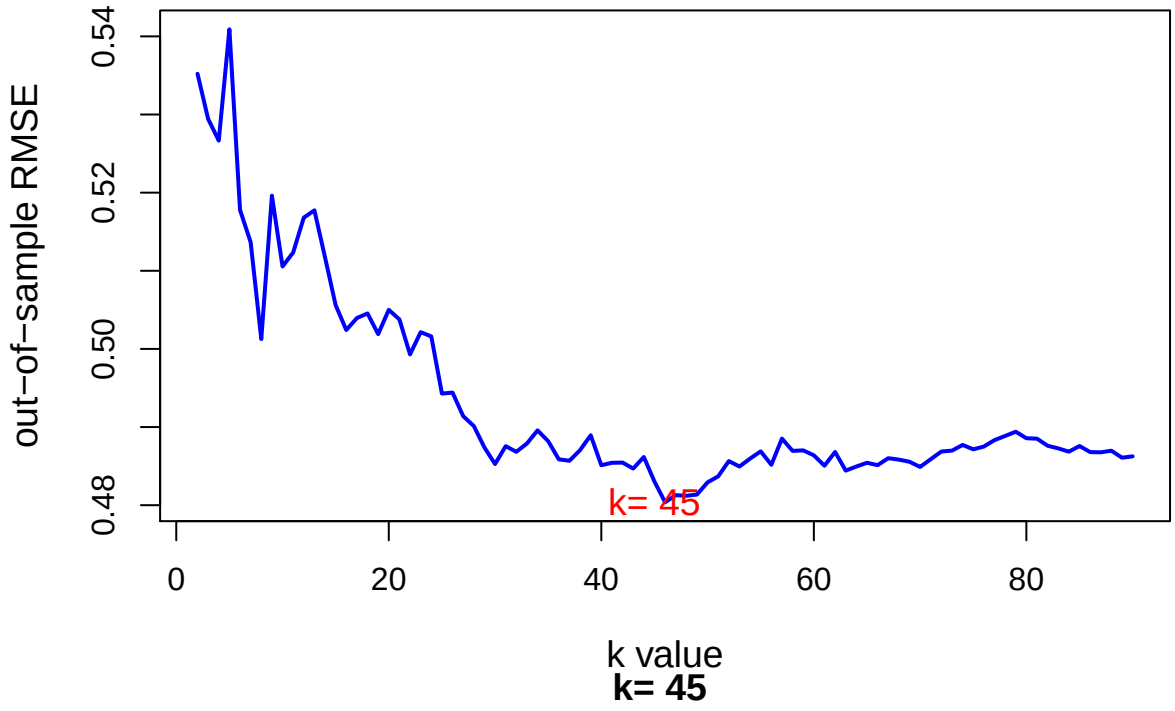
```
##
## Call:
## lm(formula = CourseEvals ~ ., data = beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31385 -0.30202  0.01011  0.29815  1.04929
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.06542    0.05145  79.020  < 2e-16 ***
## BeautyScore  0.30415    0.02543  11.959  < 2e-16 ***
## female      -0.33199    0.04075  -8.146 3.62e-15 ***
## lower       -0.34255    0.04282  -7.999 1.04e-14 ***
## nonenglish  -0.25808    0.08478  -3.044  0.00247 **
```

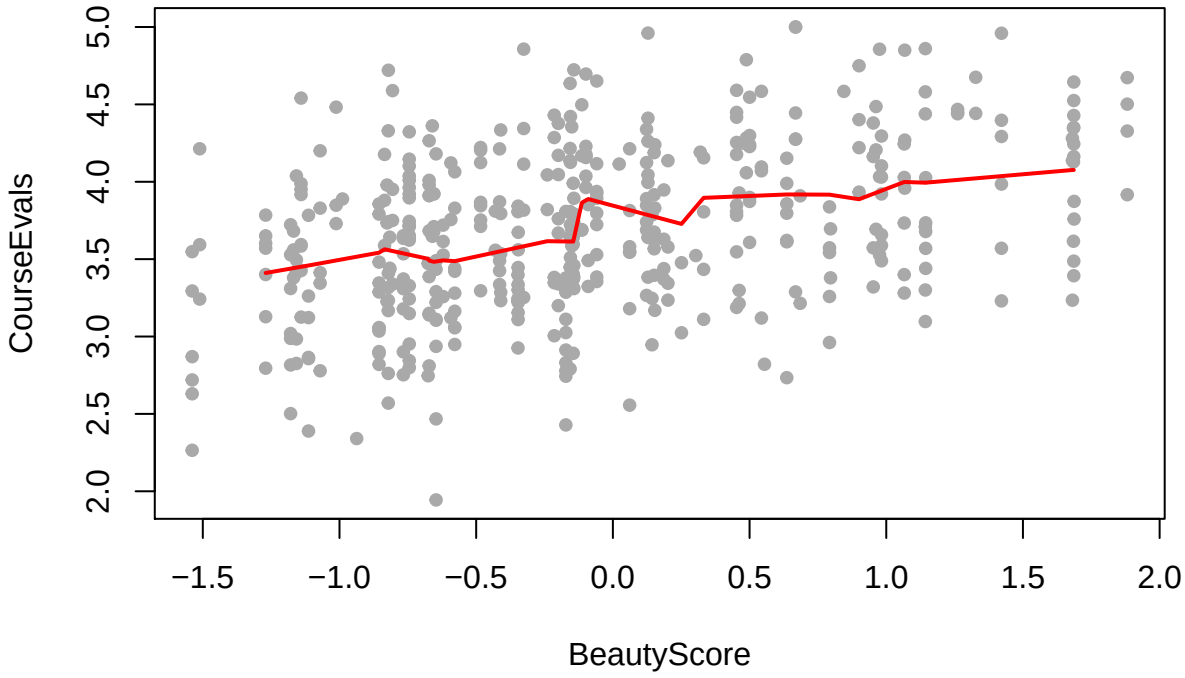
```
## tenuretrack -0.09945    0.04888  -2.035  0.04245 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4273 on 457 degrees of freedom
## Multiple R-squared:  0.3471, Adjusted R-squared:  0.3399
## F-statistic: 48.58 on 5 and 457 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = beauty.train$CourseEvals ~ poly(beauty.train$BeautyScore,
##     4))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61940 -0.33925  0.01795  0.35947  1.21432
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        3.69150    0.02511 147.032  < 2e-16
## poly(beauty.train$BeautyScore, 4)1  3.87881    0.48294   8.032 1.34e-14
## poly(beauty.train$BeautyScore, 4)2 -0.15902    0.48294  -0.329    0.742
## poly(beauty.train$BeautyScore, 4)3  0.11769    0.48294   0.244    0.808
## poly(beauty.train$BeautyScore, 4)4 -0.27462    0.48294  -0.569    0.570
##
## (Intercept)                        ***
## poly(beauty.train$BeautyScore, 4)1 ***
## poly(beauty.train$BeautyScore, 4)2
## poly(beauty.train$BeautyScore, 4)3
## poly(beauty.train$BeautyScore, 4)4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4829 on 365 degrees of freedom
## Multiple R-squared:  0.1512, Adjusted R-squared:  0.1419
## F-statistic: 16.25 on 4 and 365 DF,  p-value: 2.93e-12
```

```
## [1] 45
```

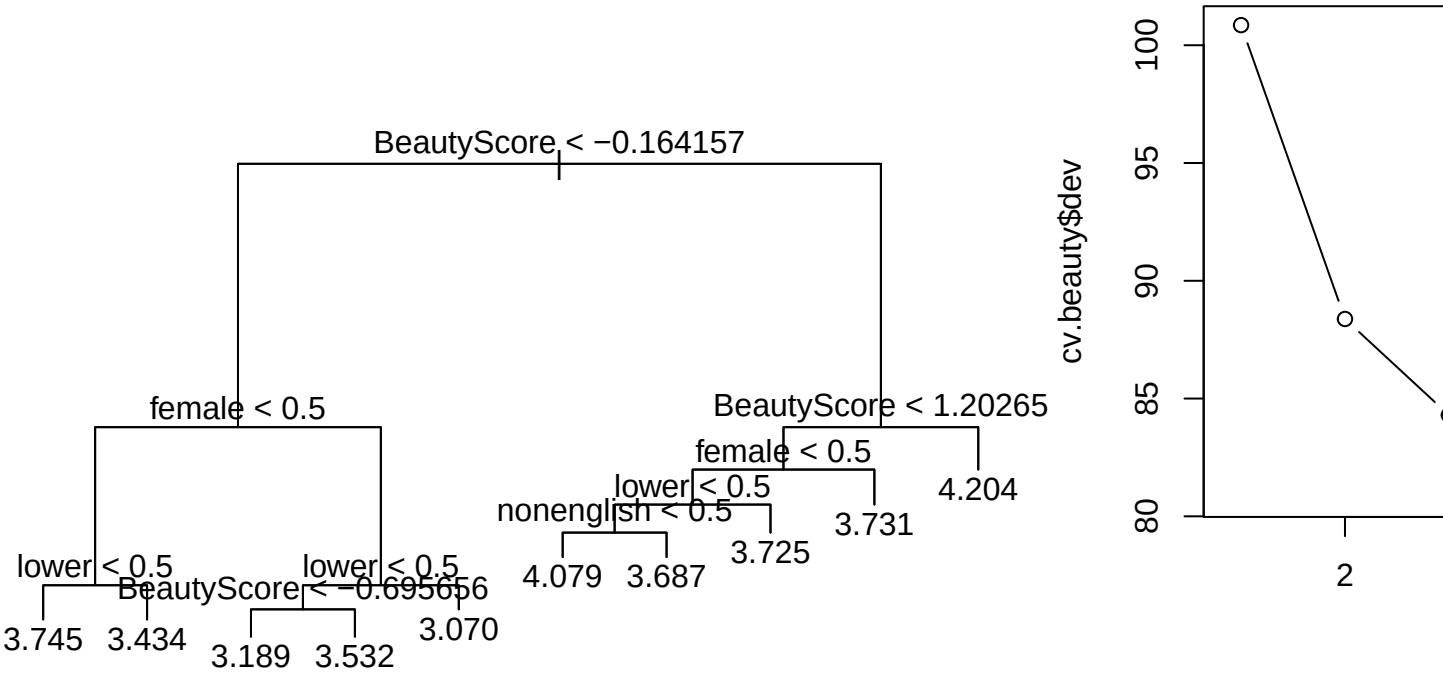
```
## [1] 0.480377
```






```
##
## Regression tree:
## tree(formula = CourseEvals ~ ., data = beauty, subset = train)
## Variables actually used in tree construction:
## [1] "BeautyScore" "female"      "lower"       "nonenglish"
## Number of terminal nodes:  10
## Residual mean deviance:  0.1846 = 66.46 / 360
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
```

```
## -1.25800 -0.31840  0.00277  0.00000  0.27320  1.12600
```




```
## [1] 0.4688049
```

2. In his paper, Dr. Hamermesh has the following sentence: "Disentangling whether this outcome represents productivity or discrimination is, as with the issue generally, probably impossible". Using the concepts we have talked about so far, what does he mean by that?

In an observational study, it is essentially impossible to discern causality without some level of doubt since we are unable to control for all of the factors that go into an outcome. This is, generally, the unreducible error that we must consider when building models. Until we can measure for these instances, Dr. Hamermesh's assertion will stand.

For example, it is conceivable that a Professor that is considered attractive might be a better teacher because of it. They might have a better chance of holding students' attention (see: the Indiana Jones effect). In this case, you would say that the Professor's attractiveness led to an increase in their productivity.

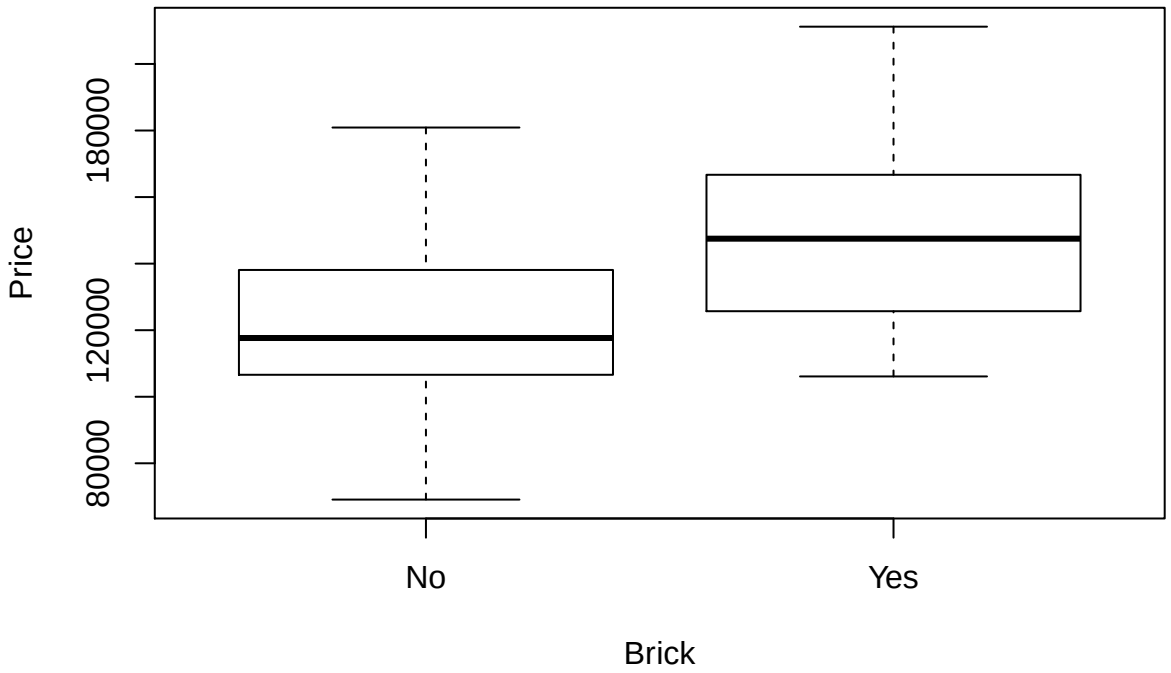
It is also conceivable that a talented Professor would more likely be considered attractive by their students. If you do well in a class, you would feel good about yourself, and aren't we charmed by those who make us feel good about ourselves? Here, attractiveness would instead be a result of the Professor's productivity.

Finally, it's also possible that it truly does represent discrimination. Perhaps people are more inclined to reward those they deem attractive and punish those they consider unattractive.

Problem 2: Housing Price Structure

1. Is there a premium for brick houses everything else being equal?

There is a premium. This is observable in the plot. Fitting a linear model across the full dataset estimates that a brick house will cost an additional ~$25,811 with all else equal.

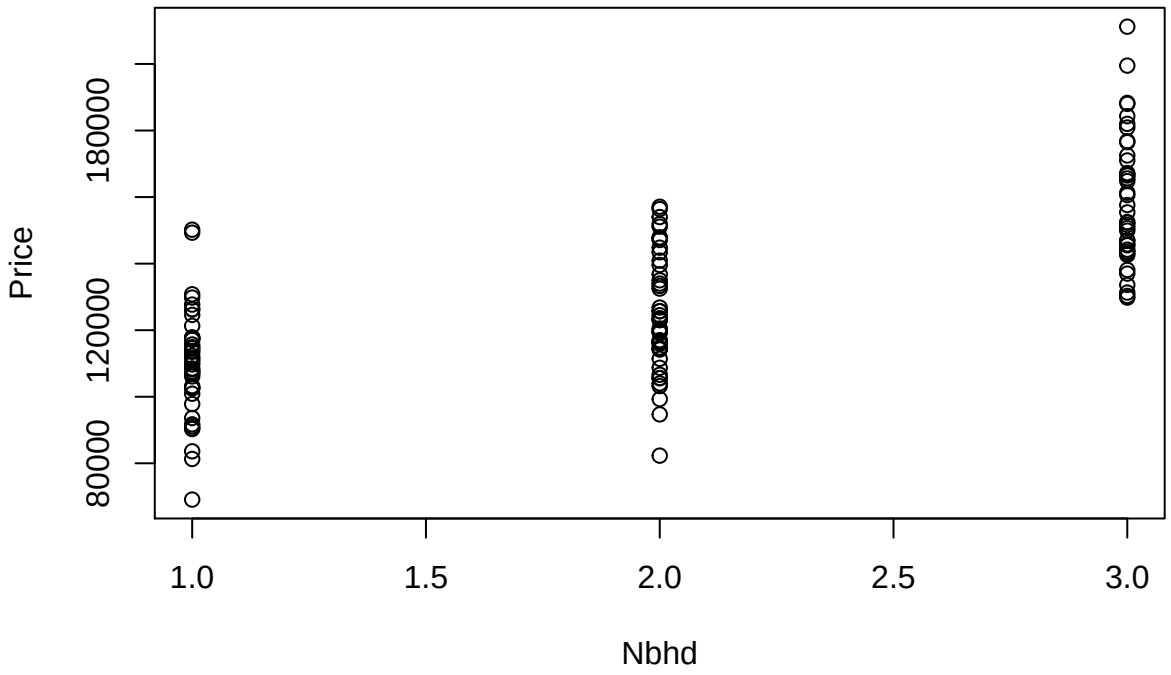



```
## [1] 24200.82
```

```
##
## Call:
## lm(formula = Price ~ Brick, data = midcity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52858 -16758  -3564  18781  63431
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   121958       2594  47.024  < 2e-16 ***
## BrickYes       25811       4528   5.701 8.02e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24050 on 126 degrees of freedom
## Multiple R-squared:  0.205,  Adjusted R-squared:  0.1987
## F-statistic:  32.5 on 1 and 126 DF,  p-value: 8.023e-08
```

2. Is there a premium for houses in neighborhood 3?

There is a premium. This is observable in the plot. Fitting a linear model across the full dataset estimates that a house's price will increase by an additional $24,369 as you move from Neighborhood 1 to Neighborhood 2 to Neighborhood 3.




```
## [1] 21789.68
```

```
##
## Call:
## lm(formula = Price ~ Nbhd, data = midcity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49079 -12773     89  10830  55452
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    82642       4278   19.32   <2e-16 ***
## Nbhd           24369       2019   12.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18370 on 126 degrees of freedom
## Multiple R-squared:  0.5363, Adjusted R-squared:  0.5326
## F-statistic: 145.7 on 1 and 126 DF,  p-value: < 2.2e-16
```

3. Is there an extra premium for brick houses in neighborhood 3?

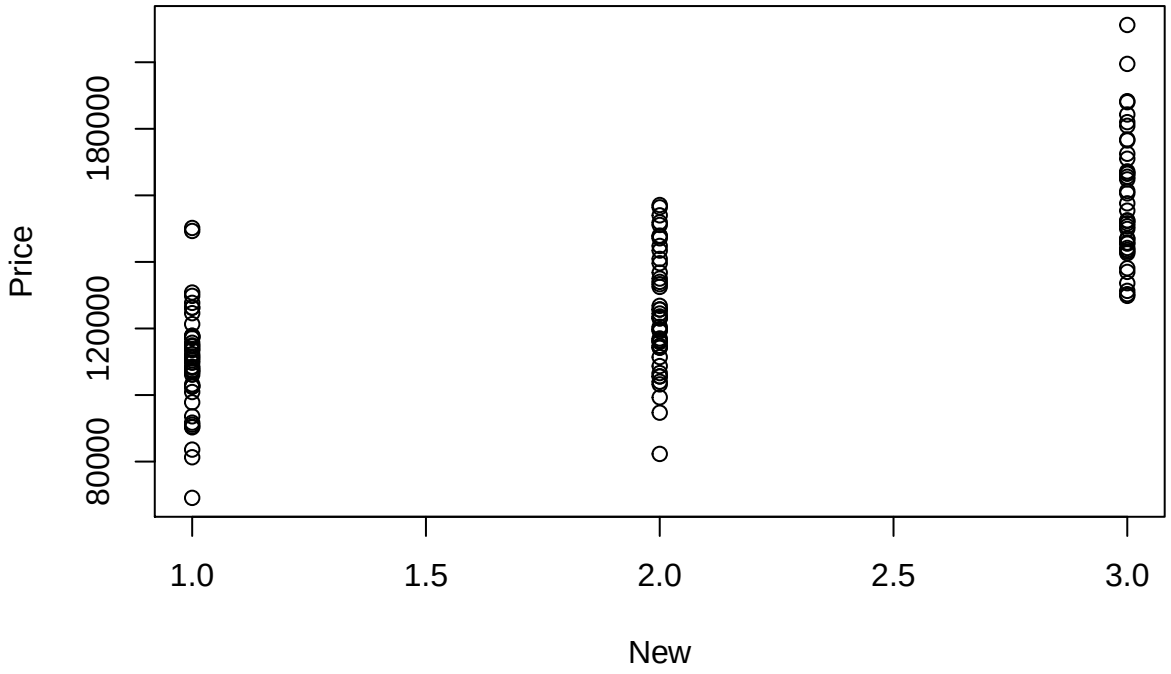
Yes, there is a stasticially significant premium for brick houses.

```
##
## Call:
## lm(formula = Price ~ Brick, data = midcity.3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23600  -9415  -2700   8135  36000
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   148230       3067  48.325  < 2e-16 ***
## BrickYes       26970       4789   5.632 1.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14710 on 37 degrees of freedom
## Multiple R-squared:  0.4616, Adjusted R-squared:  0.447
## F-statistic: 31.72 on 1 and 37 DF,  p-value: 1.981e-06
```

4. For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single "older" neighborhood?

Yes, you can. This actually decreased our test RMSE from 21789.68 when modeling based on neighborhood to 19933.36.




```
## [1] 21789.68
```

```
##
## Call:
## lm(formula = Price ~ New, data = midcity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -48678 -12320  -1786  10368  51905
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   117778       2002   58.83   <2e-16 ***
## New            41517       3627   11.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 18890 on 126 degrees of freedom
## Multiple R-squared:  0.5098, Adjusted R-squared:  0.5059
## F-statistic:   131 on 1 and 126 DF,  p-value: < 2.2e-16
```

Problem 3: What Causes What??

1. Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city).

From a practical sense, it's difficult to change the amount of police that are present in a city. It's potentially dangerous (and a bad PR move) to willingly decrease the amount of police, and it's potentially expensive to increase the amount of police.

Despite that, one might think that they can just compare the crime rates between cities that have different amounts of active police. However, this doesn't account for the potential contextual differences between the populations (how their police are trained and deployed for example). Any correlation between "Crime" and "Police" in these instances may actually be caused by unobserved factors.

2. How were the researchers from UPENN able to isolate this effect? Briefly describe their approach and discuss their result in the "Table 2" below.

The researchers were able to capitalize on the terror alert system in Washington D.C. When the terror alert was higher, there was a policy to increase police in Washington D.C., a potential terrorist target. With this, they were able to measure whether the crime rate changed on days when the terrorist alert level was different and thus the police presence was different.

They found that crime did, in fact, decrease on days when the terrorist alert level was highest.

3. Why did they have to control for METRO ridership? What was that trying to capture?

They established that on days when the terrorist alert level was higher, police presence was higher. However, they theorized that there may also be fewer tourists and potential crime victims when the terrorist level was higher.

To measure activity in the city, they used METRO ridership. If METRO ridership also varied on high alert days, then it would weaken the evidence that the increase in police caused a decrease in crime. It would instead suggest that fewer crimes occurred because there were fewer people in public.

By controlling for METRO ridership, they were able to remove that factor as a potential variable affecting crime rate.

4. In the next page, I am showing you "Table 4" from the research paper. Just focus on the first column of the table. Can you describe the model being estimated here? What is the conclusion?

They have built a multi-variable linear model to attempt to predict the change in crime rate on the National Mall on High-Alert Days.

According to their model, the crime rate decreases by 2.621 units on High Alert days in District 1 (the district containing the National Mall). This coefficient had a relatively small standard error indicating that this coefficient is relatively consistent.

It also asserts that the crime rate decreases by .571 units on High Alert days in other districts. However, this coefficient standard error is relatively much larger indicating that this may not be as consistent. In fact, this variable is not noted as being statistically significant.

Finally, it indicates that crime rate increases by 2.477 units for one unit increase in log(midday ridership) with a relatively modest coefficient standard error.

The conclusion is that lower crime rate is correlated with increased police presence and decreased METRO ridership in a statistically significant manner. However, this is not enough to claim true causality, and there may be other factors to attempt to control before a stronger case can be made. For example, does an increased terrorist threat level affect a population's behavior in other ways? Perhaps it increases national pride and sense of community, thus making people less likely to commit crimes. It's a bit abstract of an assertion, but that is part of what makes determining causality via observational studies so difficult.

Problem 4: BART

Apply BART to the California Housing Data example of Section 4. Does BART outperform RF or Boosting?

My test data showed that RF slightly outperformed BART, both of which greatly outperformed the particular Boosting model I applied.

BART MSE = 0.05701371

RF MSE = 0.05307826

BOOST MSE = 0.1148911

```
##   longitude latitude housingMedianAge population households medianIncome
## 1   -122.23    37.88                41        322        126       8.3252
## 2   -122.22    37.86                21       2401       1138       8.3014
## 3   -122.24    37.85                52        496        177       7.2574
## 4   -122.25    37.85                52        558        219       5.6431
## 5   -122.25    37.85                52        565        259       3.8462
## 6   -122.25    37.85                52        413        193       4.0368
##   AveBedrms AveRooms AveOccupancy        y
## 1 1.0238095 6.984127     2.555556 13.02276
## 2 0.9718805 6.238137     2.109842 12.78968
## 3 1.0734463 8.288136     2.802260 12.77167
## 4 1.0730594 5.817352     2.547945 12.74052
## 5 1.0810811 6.281853     2.181467 12.74315
## 6 1.1036269 4.761658     2.139896 12.50507
```

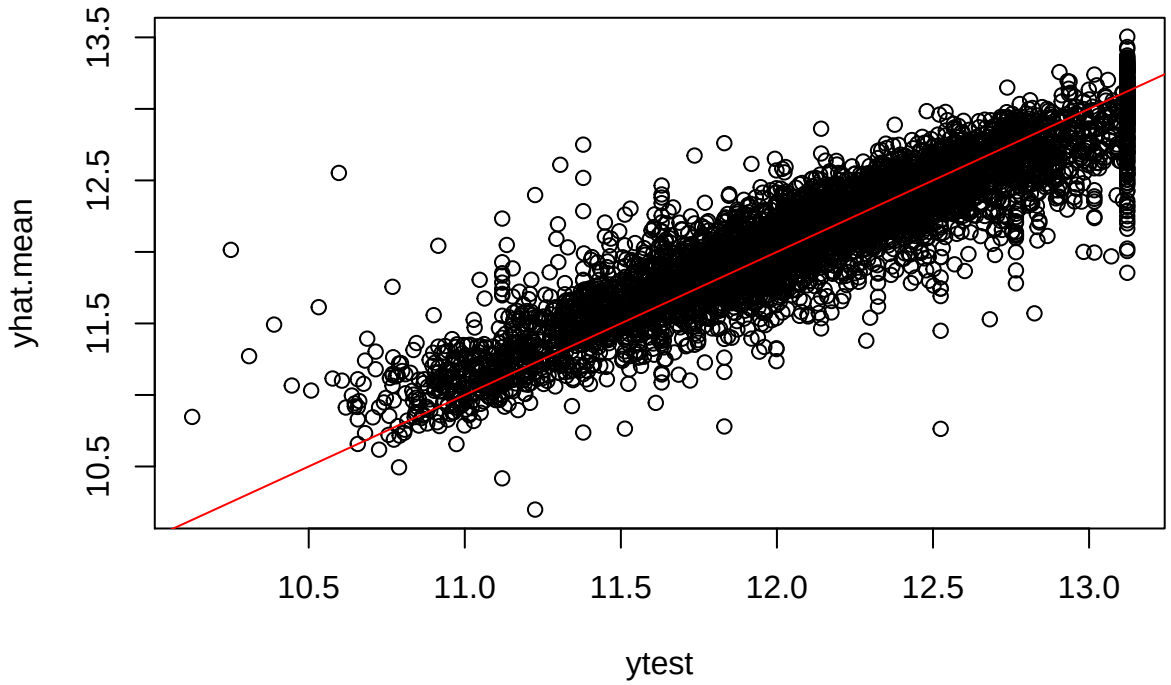
```
## *****Into main of wbart
## *****Data:
## data:n,p,np: 20640, 9, 0
## y1,yn: 0.937880, -0.684008
## x1,x[n*p]: -122.230000, 2.616981
## *****Number of Trees: 200
## *****Number of Cut Points: 100 ... 100
## *****burn and ndpost: 50, 200
## *****Prior:beta,alpha,tau,nu,lambda: 2.000000,0.950000,0.061989,3.000000,0.022423
## *****sigma: 0.339283
## *****w (weights): 1.000000 ... 1.000000
## *****Dirichlet:sparse,theta,omega,a,b,rho,augment: 0,0,1,0.5,1,9,0
## *****nkeeptrain,nkeeptest,nkeeptestme,nkeeptreedraws: 200,200,200,200
## *****printevery: 100
```

```
## *****skiptr,skipte,skipteme,skiptreedraws: 1,1,1,1
##
## MCMC
## done 0 (out of 250)
## done 100 (out of 250)
## done 200 (out of 250)
## time: 30s
## check counts
## trcnt,tecnt,temecnt,treedrawscnt: 200,0,0,200
```

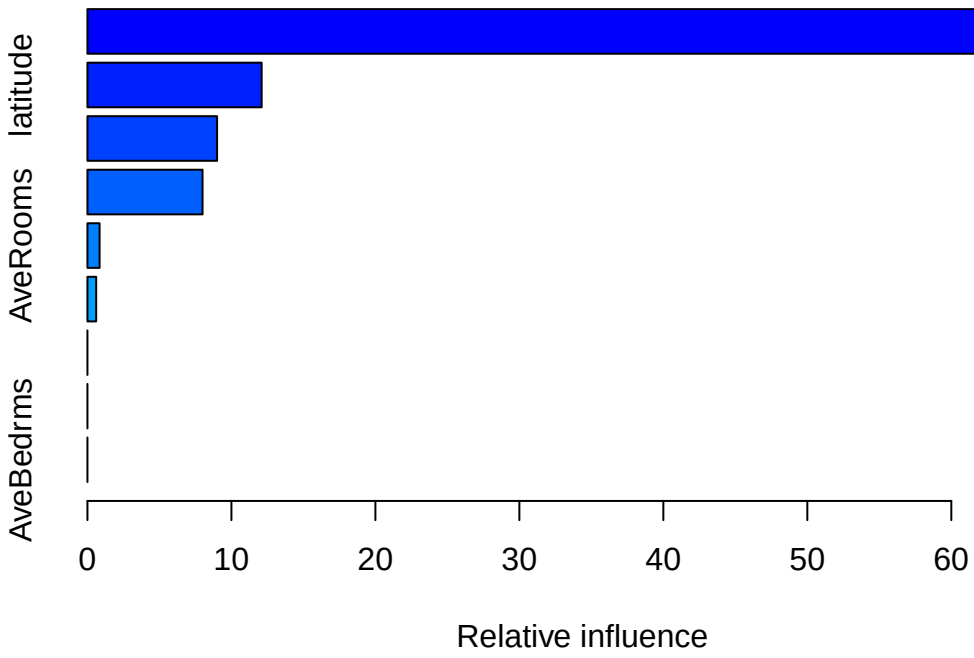
```
## train sample size is  15480  and test sample size is  5160
```

```
## *****Into main of wbart
## *****Data:
## data:n,p,np: 15480, 9, 0
## y1,yn: 0.021755, 0.185572
## x1,x[n*p]: -117.020000, 3.727799
## *****Number of Trees: 200
## *****Number of Cut Points: 100 ... 100
## *****burn and ndpost: 100, 1000
## *****Prior:beta,alpha,tau,nu,lambda: 2.000000,0.950000,0.061989,3.000000,0.022288
## *****sigma: 0.338258
## *****w (weights): 1.000000 ... 1.000000
## *****Dirichlet:sparse,theta,omega,a,b,rho,augment: 0,0,1,0.5,1,9,0
## *****nkeeptrain,nkeeptest,nkeeptestme,nkeeptreedraws: 1000,1000,1000,1000
## *****printevery: 100
## *****skiptr,skipte,skipteme,skiptreedraws: 1,1,1,1
##
## MCMC
## done 0 (out of 1100)
## done 100 (out of 1100)
## done 200 (out of 1100)
## done 300 (out of 1100)
## done 400 (out of 1100)
## done 500 (out of 1100)
## done 600 (out of 1100)
## done 700 (out of 1100)
## done 800 (out of 1100)
## done 900 (out of 1100)
## done 1000 (out of 1100)
## time: 127s
## check counts
## trcnt,tecnt,temecnt,treedrawscnt: 1000,0,0,1000
```

```
## *****In main of C++ for bart prediction
## tc (threadcount): 1
## number of bart draws: 1000
## number of trees in bart sum: 200
## number of x columns: 9
## from x,np,p: 9, 5160
## ***using serial code
```




```
## [1] 0.05701371
## [1] 0.05307826
```




```
##                                 var    rel.inf
## medianIncome         medianIncome 69.4510734
## latitude                 latitude 12.1004008
## AveOccupancy         AveOccupancy  9.0038870
## longitude               longitude  7.9933207
## AveRooms                 AveRooms  0.8430427
## housingMedianAge housingMedianAge  0.6082753
## population             population  0.0000000
```

```
## households             households  0.0000000
## AveBedrms               AveBedrms  0.0000000
```

```
## [1] 0.1148911
```

Problem 5: Neural Nets

Re-run the Boston housing data example using a single layer neural net. Cross validate for a few choices of size and decay parameters.

Two neural nets stand out. One of mid-size and high decay (nn3) performed best. However, one with large size and small decay (nn6) also performed well.

```
## # weights:  76
## initial  value 228897.672216
## iter  10 value 25186.002728
## iter  20 value 23789.619032
## iter  30 value 20693.439604
## iter  40 value 19749.601607
## iter  50 value 18107.696883
## iter  60 value 15915.408219
## iter  70 value 15430.259043
## iter  80 value 15133.716523
## iter  90 value 14511.172351
## iter 100 value 13571.276106
## final  value 13571.276106
## stopped after 100 iterations
```

```
## # weights:  76
## initial  value 248510.241247
## final  value 31963.693715
## converged
```

```
## # weights:  376
## initial  value 204852.565875
## iter  10 value 28185.151137
## iter  20 value 20493.319158
## iter  30 value 14810.674735
## iter  40 value 13088.226933
## iter  50 value 11946.721853
## iter  60 value 10833.400021
## iter  70 value 10537.005790
## iter  80 value 9900.830631
## iter  90 value 9590.063908
## iter 100 value 8451.104386
## final  value 8451.104386
## stopped after 100 iterations
```

```
## # weights:  376
## initial  value 167952.313262
## iter  10 value 25434.547199
## iter  20 value 25239.921318
## iter  30 value 24935.554350
## iter  40 value 24923.813145
## iter  50 value 24908.898094
## iter  60 value 24876.754210
```

```
## iter  70 value 24857.301952
## iter  80 value 24854.798737
## final  value 24854.771665
## converged
```

```
## # weights:  901
## initial  value 217306.581201
## iter  10 value 29454.330038
## iter  20 value 25983.214487
## iter  30 value 24287.654395
## iter  40 value 24158.791671
## iter  50 value 23372.224469
## iter  60 value 21418.770471
## iter  70 value 20518.915730
## iter  80 value 19674.157350
## iter  90 value 19470.182565
## iter 100 value 18978.396500
## final  value 18978.396500
## stopped after 100 iterations
```

```
## # weights:  901
## initial  value 210595.627047
## iter  10 value 25169.454706
## iter  20 value 23011.888983
## iter  30 value 21664.176831
## iter  40 value 15522.609465
## iter  50 value 14746.433057
## iter  60 value 13062.544199
## iter  70 value 10620.779394
## iter  80 value 10189.925246
## iter  90 value 9671.714537
## iter 100 value 9536.293823
## final  value 9536.293823
## stopped after 100 iterations
```

```
## [1] 6.789904
```

```
## [1] 10.09481
```

```
## [1] 5.672912
```

```
## [1] 9.148157
```

```
## [1] 7.967575
```

```
## [1] 6.20886
```

Problem 6: Final Project

1. Describe your contribution to the final group project

Data Cleaning

- Our Active Oil Rigs data was listed in an Excel workbook with a table in a different worksheet for every month from the past ~18 years. I had to collate this data, pulling the relevant info from each of the ~215 sheets, into one table representing average active oil rigs per month.

- Our Cushing Spot Price data was originally in nominal form. I had to convert each month's average price using the appropriate inflation index from that period in order to get the price data in real terms.

EDA

- I conducted some initial Exploratory Data Analysis by creaing scatter plots showing the relationship between each variable and the Spot Price. I also plotted Cushing Inventories vs. Active Oil Rigs to discern some potential collinearity between those variables.

Analysis

- I ran the initial Multi Variable Linear Regression models. First, I ran a Best Subset analysis based on least squares to estimate what variables the most accurate model would contain. I then ran best subsets in k-fold cross validation with a test set to discern the difference in average RMSE between models. Finally, I fit the best model on the full dataset and created a 3D plot to represent the model.
- We all collaborated on interpretations, what our analysis captured, and the many, many ways it could be improved.

Presentation

- I introduced our presentation, and I spoke about my EDA and Multi Variable Linear Regression Analysis.