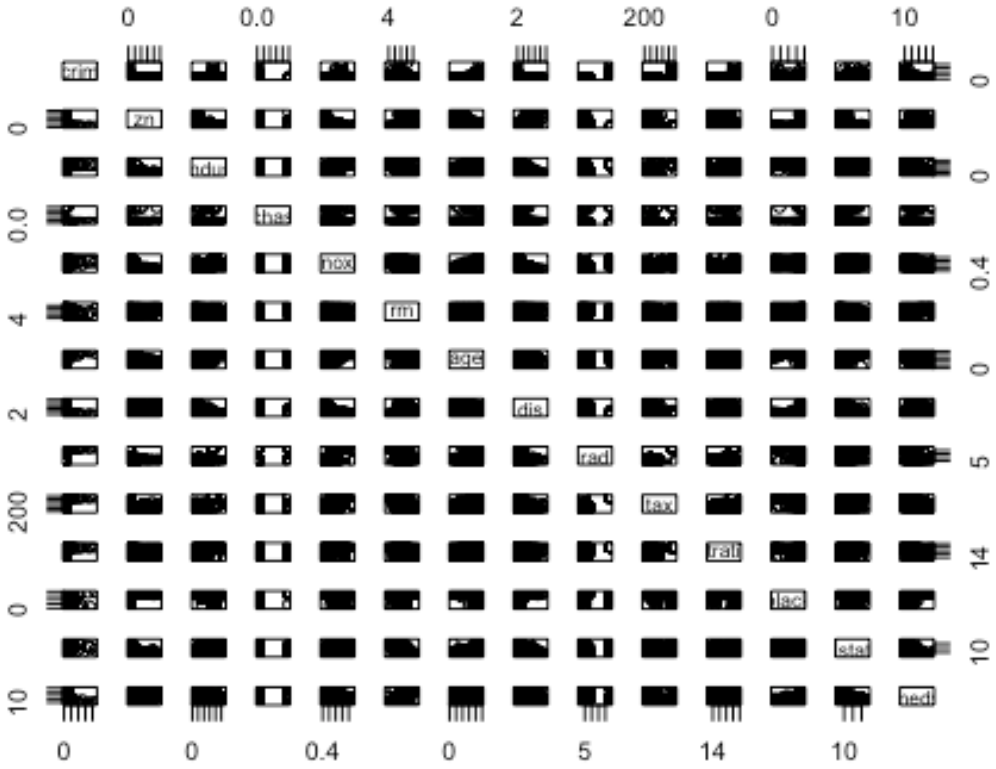
Take Home Exam

Michael Zhang

Book Problems

Chapter 2, #10

a) There are 506 rows and 14 columns. The rows represent observations (housing values in Boston suburbs), and the columns contain features (e.g. median house value, per capita crime rate, etc.).



b) We see a fairly strong negative correlation between lstat and medv. We also see a fairly strong positive correlation between rm and medv and a similarly negative correlation between rm and lstat. Additionally, we see a negative correlation between dis and lstat. There are other correlations between the predictors, but these are some of the strongest that we see from the pairwise scatterplots.

c) Yes, there are predictors associated with crime rate.

- The older the home, the more crime
- The farther away from Boston employment centers, the less crime
- The more accessible to radial highways, the more crime
- The higher the property tax, the more crime
- The higher the student-teacher ratio, the more crime

d) Yes, the median crime rate is 0.26, but the crime rate ranges from 0.01 to 88.98. Most cities have low crime rates, and the 5% of suburbs with the highest crime rate have crime rates larger than 15.78.

Tax rate is similar in that the median tax rate is 330, but there is a big jump between suburbs with lower tax rates and suburbs with tax rates higher than 650.

The pupil-teacher ratio distribution among the suburbs is pretty evenly distributed with the exception of a peak of suburbs with a pupil-teacher ratio of ~20.

e) 35 surburbs out of 506 bound the Charles River

f) The median pupil-teacher ratio is 19.05.

```
subset(Boston, medv == min(Boston$medv))
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio  black
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97
##     lstat medv
## 399 30.59    5
## 406 22.98    5
```

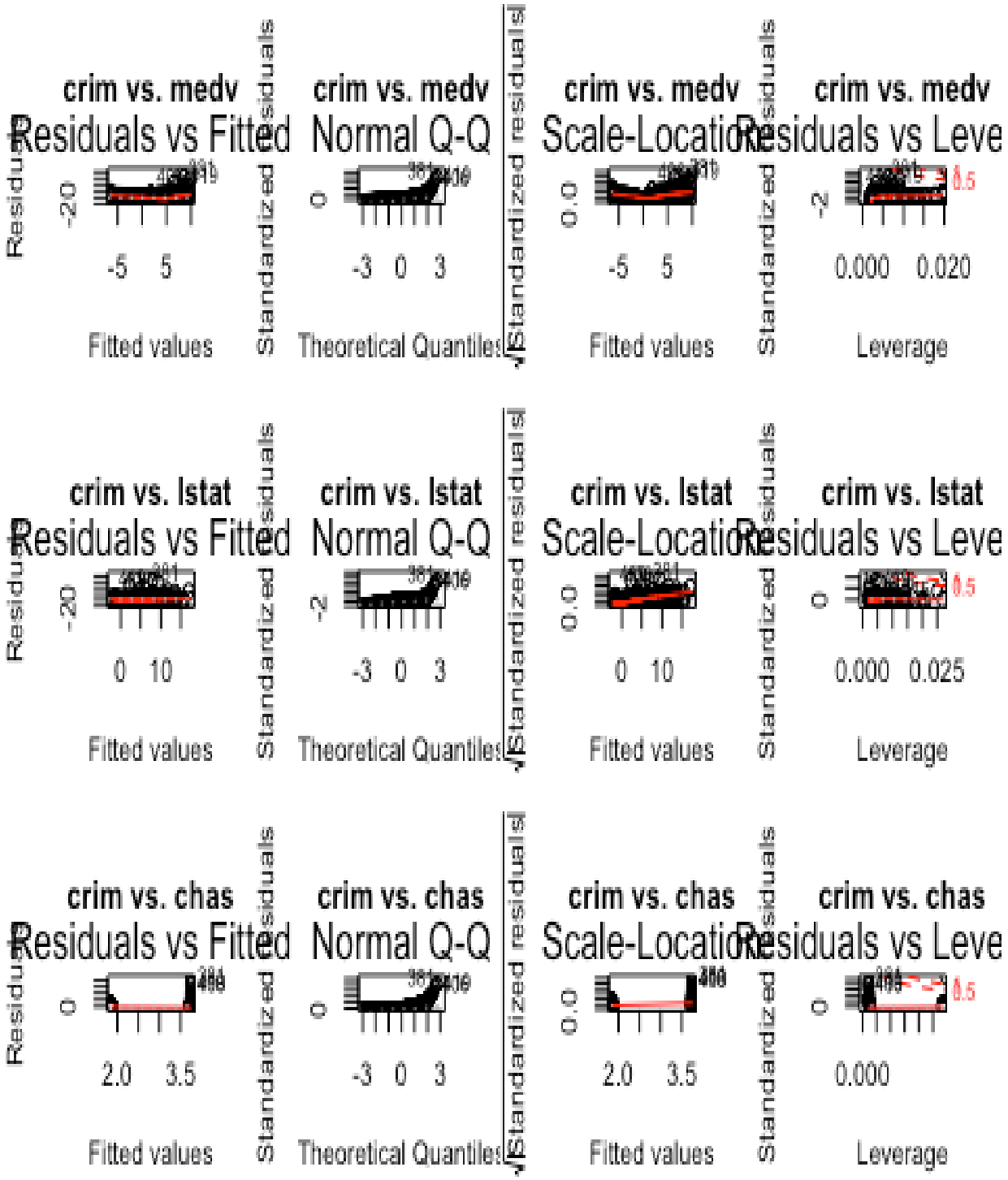
High crime, low proportion of large lots, large proportion of non-retail business, not bound by Charles River, larger nitrogen oxide concentration, oldest homes, closest to employment centers, most accessible to radial highways, higher property taxes, higher student-teacher ratios, higher proportion of blacks, and higher proportion of lower status. Doesn't seem like an optimal place to live.

h) There are 64 suburbs with more than 7 rooms per dwelling, and 13 suburbs with more than 8 rooms per dwelling.

The suburbs with more than 8 rooms per dwelling have less crime, higher median value, and lower lstat.

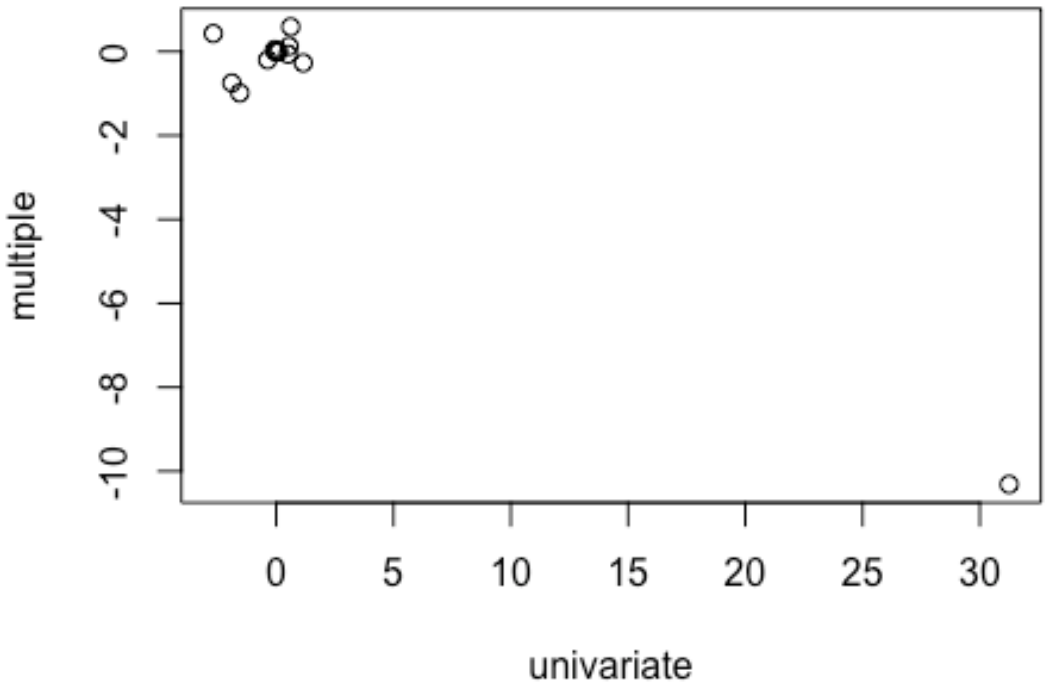
Chapter 3, #15

a) There is a statistically significant association with the response for all predictors except chas.



b) We can reject the null hypothesis for zn, dis, rad, black, and medv

c)



In the univariate model, the coefficient for nox is 31, while in the multiple regression, the coefficient is -10.3. Most of the multiple regeression coefficients are around 0 and slightly negative.

d) Yes, there is evidence of non-linear association. Specifically, we can reject the null hypothesis of a linear association for all predictors except black and chas.

Chapter 6, #9

```
library(ISLR)
attach(College)
set.seed(2015)

train = data.frame(College)
test = data.frame(College)

tr = sample(1:nrow(College), 0.7*nrow(College))

train = train[tr,]
test = test[-tr,]
```

b) Test error: 2,050,905

c) Test error: 2,105,986

d) Test error: 2,031,401 All coefficient estimates are nonzero except Books

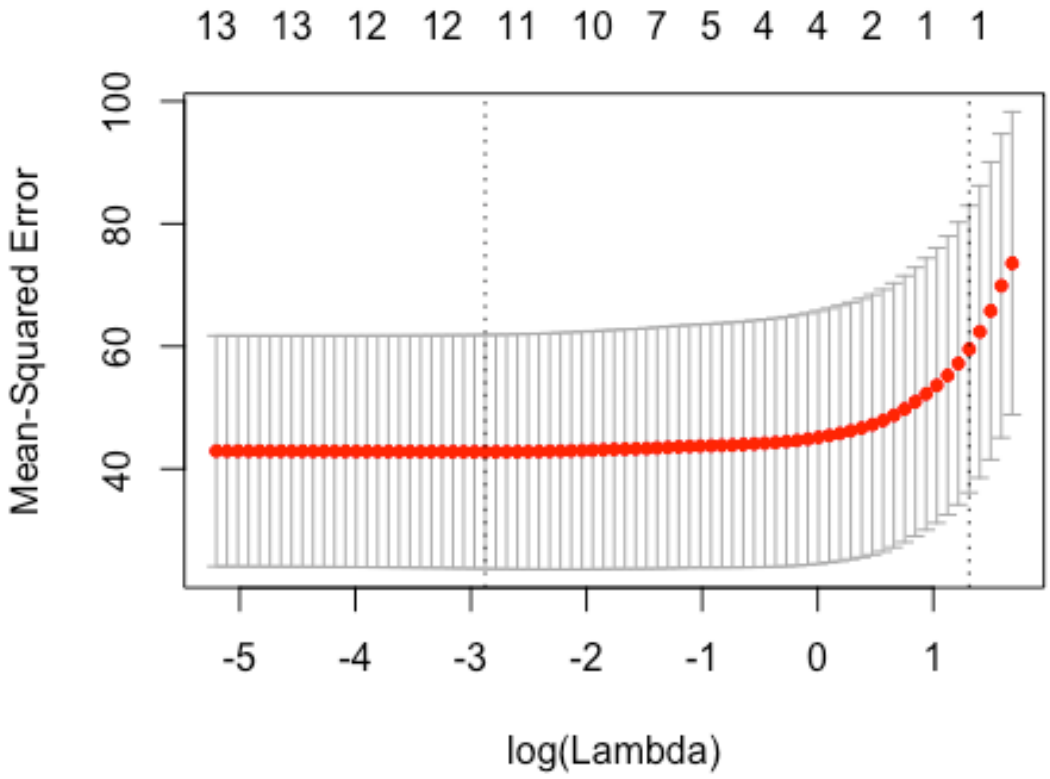
e) Test error: 4,034,479

f) Test error: 2,079,952

g) We can predict the number of college applications received with a fair amount of accuracy (R^2 > 0.9 for all except PCR). Excluding PCR, all the models are pretty similar in the resulting test errors.

Chapter 6, #11

a)

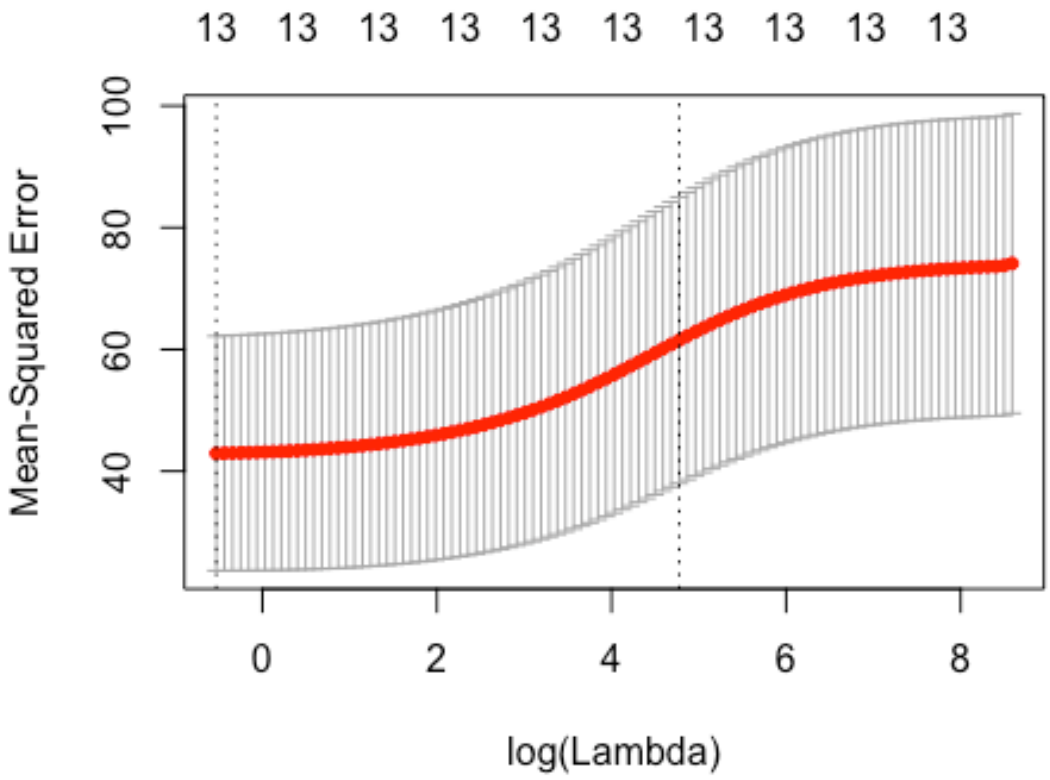



```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                      1
## (Intercept) 1.7799525
## zn          .
## indus       .
```

```
## chas        .
## nox         .
## rm          .
## age         .
## dis         .
## rad         0.1920089
## tax         .
## ptratio     .
## black       .
## lstat       .
## medv        .

## [1] 7.71689
```

Using lasso, we get a test MSE of 7.72, and the only coefficient that seems to matter is the one for rad (accessibility to radial highways).




```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                          1
## (Intercept)  1.984601820
## zn          -0.002712896
## indus        0.023751842
## chas        -0.120843958
## nox          1.499819910
```

```
## rm          -0.118580408
## age          0.005022922
## dis         -0.075115440
## rad          0.034570984
## tax          0.001596306
## ptratio      0.056535699
## black       -0.001981669
## lstat        0.027880296
## medv        -0.018358122

## [1] 7.842953
```

Using ridge regression, we get a test MSE of 7.84, and the most important variable seems to be nox (nitrogen oxides concentration).

b) The results from lasso and ridge are fairly similar and can be thought of as interchangeable in terms of this dataset. I would select lasso based on the relatively lower test error.

c) The lasso model involves all the features in the data. However, the only variable that has an associated coefficient is rad, which means that if we wanted to, we could use this information to run a OLS with just rad against crim.

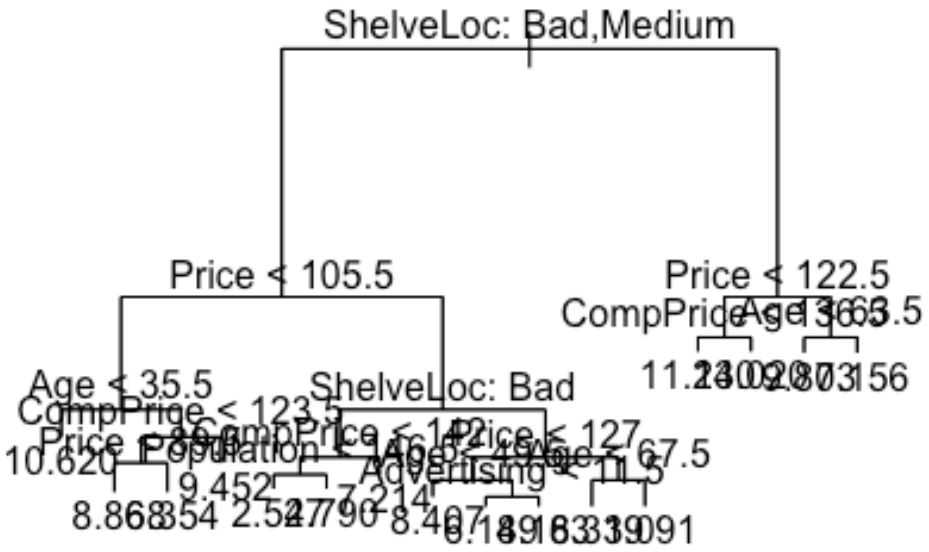
Chapter 8, #8

a)

```
library(ISLR)
attach(Carseats)

set.seed(2015)
train = sample(dim(Carseats)[1], dim(Carseats)[1]*2/3)
Carseats.train = Carseats[train, ]
Carseats.test = Carseats[-train, ]
```

b)




```
## [1] 5.149231
```

We see that ShelveLoc is the first break in the tree, followed by price. The test MSE is 5.15.

c) Using cross-validation, we see that the optimal level of tree complexity is 5. Using this, the test MSE was reduced to 5.00.

d) Using bagging, we obtain a test MSE of 2.74. The importance() function shows us that the 3 most important variables in descending order are ShelveLoc, Price, and CompPrice.

e) Using random forests, we obtain a test MSE of 3.12. The 3 most important variables in descending order are ShelveLoc, Price, and Advertising. Varying m leads to test MSEs in the range of 2.77 and 3.25.

Chapter 8, #11

```
library(ISLR)
train = 1:1000
Caravan$Purchase = ifelse(Caravan$Purchase == "Yes", 1, 0)
Caravan.train = Caravan[train,]
Caravan.test = Caravan[-train,]
```

b) Based on our boosting model, the 3 most important predictors in descending order are PPERSAUT, MKOOPKLA, and MOPLHOOG.

c) Using the boosting model, around 20% of the people predicted to make a purchase actually make one. The boosting results are significantly better than the results obtained from KNN and logistic regression.

Problem 1: Beauty Pays!

1.

```
##
## 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
```

After running a multiple linear regression, we see that holding all other variables constant, we see a positive correlation between beauty score and course evaluation score. Specifically, a 1 point increase in beauty score results in a 0.30 increase in course evalution score. Surprisingly, we see a negative coefficient for female.

2. What Dr. Hamermesh means is that in most problems similar to this one, it is likely that there are several confounding variables. As always, correlation does not equal causation, and we can not say with certainty that the relationship between attractiveness and instructor ratings is not due to an underlying variable.

Problem 2: Housing Price Structure

1.

```
## The following object is masked from Carseats:
##
##     Price

##
## Call:
## lm(formula = Price ~ ., data = midcity.clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27337.3  -6549.5    -41.7   5803.4  27359.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2159.498   8877.810   0.243  0.80823
## Nbhd2       -1560.579   2396.765  -0.651  0.51621
## Nbhd3       20681.037   3148.954   6.568 1.38e-09 ***
## Offers      -8267.488   1084.777  -7.621 6.47e-12 ***
## SqFt           52.994      5.734   9.242 1.10e-15 ***
## BrickYes    17297.350   1981.616   8.729 1.78e-14 ***
## Bedrooms     4246.794   1597.911   2.658  0.00894 **
## Bathrooms    7883.278   2117.035   3.724  0.00030 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10020 on 120 degrees of freedom
## Multiple R-squared:  0.8686, Adjusted R-squared:  0.861
## F-statistic: 113.3 on 7 and 120 DF,  p-value: < 2.2e-16
```

Based on the multiple linear regression, it seems that, everything else being equal, there is a premium for brick houses. We feel strongly certain about this considering that the 95% confidence interval for the coefficient of BrickYes does not include 0. In fact, the coefficient states that on average, a brick house costs $17,297 more than a non-brick house.

2. Yes, there seems to be a premium for houses in neighborhood 3. The 95% confidence interval for the coefficient does not include 0, and on average, houses in neighborhood 3 cost $20,681 more than houses in neighborhood 1 and $22,242 ($20,681 + $1,560) more than houses in neighborhood 2.

3.

```
##
## Call:
## lm(formula = Price ~ ., data = midcity.clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26939.1  -5428.7   -213.9   4519.3  26211.4
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3009.993   8706.264   0.346  0.73016
## Nbhd2           -673.028   2376.477  -0.283  0.77751
## Nbhd3          17241.413   3391.347   5.084 1.39e-06 ***
## Offers         -8401.088   1064.370  -7.893 1.62e-12 ***
## SqFt              54.065      5.636   9.593  < 2e-16 ***
## BrickYes       13826.465   2405.556   5.748 7.11e-08 ***
## Bedrooms        4718.163   1577.613   2.991  0.00338 **
## Bathrooms       6463.365   2154.264   3.000  0.00329 **
## nbhd3_brickyes 10181.577   4165.274   2.444  0.01598 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9817 on 119 degrees of freedom
## Multiple R-squared:  0.8749, Adjusted R-squared:  0.8665
## F-statistic:   104 on 8 and 119 DF,  p-value: < 2.2e-16
```

Yes, the 95% confidence interval for the coefficient of the interaction term between brick houses and neighborhood 3 does not contain 0, meaning that there does seem to be an extra premium for brick houses in neighborhood 3.

4. Based on the multiple linear regression, we see that the t value for the neighborhood 2 coefficient is -0.283, which means that the 95% confidence interval includes 0. Knowing this, it seems reasonable to combine neighborhoods 1 and 2 into a single "older" neighborhood.

Problem 3: What causes what??

1. If you got data from a few different cities and ran the regression of "Crime" on "Police," you would not be able to discern if more crime is causing more cops in the street or if more cops in the street are causing more crime.

2. The researchers from UPENN were able to use the fact that when the terror alert is at "orange," additional police officers are stationed at the National Mall. Because this terror alert has nothing to do with street crime, they could accurately assess the effect of these additional cops on street crime. The results in Table 2 show that in fact, crime does decrease with additional cops, and this is a significant effect.

3. They had to control for METRO ridership because of the potential confounding variable of less people on the streets due to the high terror alert. Because the METRO ridership did not change, they concluded that there was similar availability of potential crime victims on both high-alert days and non-high-alert days.

4. The model results outlined in Table 4 include an interaction variable between "high alert" and the dummy variable for D.C. District 1, and the conclusion is that the additional police had a significant effect on decreasing crime only in District 1 and not in other districts.