ISLR

Chapter 3 Ex. 15

Tatiana Romanchishina

15. This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.
(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results.

1.
```
===============================================
                 Dependent variable:
                 ---------------------------
                         crim
-----------------------------------------------
zn                       -0.074***
                         (0.016)

Constant                 4.454***
                         (0.417)

-----------------------------------------------
Observations              506
R2                       0.040
Adjusted R2                      0.038
Residual Std. Error      8.435 (df = 504)
F Statistic      21.103*** (df = 1; 504)
===============================================
Note:            *p<0.1; **p<0.05; ***p<0.01
```

This model is showing a significant association between crime rate and proportion of residential land zoned for lots over 25,000 sq.ft.

2.
```
===============================================
                 Dependent variable:
                 ---------------------------
                         crim
-----------------------------------------------
indus                    0.510***
                         (0.051)
```

```
Constant                 -2.064***
                         (0.667)

-----------------------------------------------
Observations               506
R2                       0.165
Adjusted R2                     0.164
Residual Std. Error      7.866 (df = 504)
F Statistic       99.817*** (df = 1; 504)
================================================
Note:             *p<0.1; **p<0.05; ***p<0.01
```

This model is showing a significant association between crime rate and proportion of non-retail business acres per town.

3.

```
================================================
                  Dependent variable:
                  ---------------------------
                         crim
-----------------------------------------------
chas                     -1.893
                         (1.506)

Constant                 3.744***
                         (0.396)

-----------------------------------------------
Observations               506
R2                       0.003
Adjusted R2                     0.001
Residual Std. Error      8.597 (df = 504)
F Statistic              1.579 (df = 1; 504)
================================================
Note:             *p<0.1; **p<0.05; ***p<0.01
```

This model shows that there is no significant relationship between crime rate and Charles River dummy variable.

4.

```
================================================
                  Dependent variable:
                  ---------------------------
                         crim
-----------------------------------------------
nox                      31.249***
```

```
                     (2.999)

Constant             -13.720***
                     (1.699)

-----------------------------------------------
Observations          506
R2                   0.177
Adjusted R2                  0.176
Residual Std. Error  7.810 (df = 504)
F Statistic       108.555*** (df = 1; 504)
===================================================
Note:             *p<0.1; **p<0.05; ***p<0.01
```

This model shows that there is a significant relationship between crime rate and nitrogen oxides concentration.

5.

```
===================================================
              Dependent variable:
              ---------------------------
                     crim
-----------------------------------------------
rm                   -2.684***
                     (0.532)

Constant             20.482***
                     (3.364)

-----------------------------------------------
Observations          506
R2                   0.048
Adjusted R2                  0.046
Residual Std. Error  8.401 (df = 504)
F Statistic       25.450*** (df = 1; 504)
===================================================
Note:             *p<0.1; **p<0.05; ***p<0.01
```

This model displays a significant relationship between crime rate and the number of rooms per dwelling.

6.

```
===================================================
              Dependent variable:
              ---------------------------
                     crim
-----------------------------------------------
```

```
age                      0.108***
                         (0.013)

Constant                 -3.778***
                         (0.944)

-----------------------------------------------
Observations              506
R2                       0.124
Adjusted R2                        0.123
Residual Std. Error      8.057 (df = 504)
F Statistic      71.619*** (df = 1; 504)
===============================================
Note:            *p<0.1; **p<0.05; ***p<0.01
```

This model displays a significant relationship between crime rate and the proportion of owner-occupied units built prior to 1940.

7.

```
===============================================
                 Dependent variable:
                 ---------------------------
                         crim
-----------------------------------------------
dis                      -1.551***
                         (0.168)

Constant                 9.499***
                         (0.730)

-----------------------------------------------
Observations              506
R2                       0.144
Adjusted R2                        0.142
Residual Std. Error      7.965 (df = 504)
F Statistic      84.888*** (df = 1; 504)
===============================================
Note:            *p<0.1; **p<0.05; ***p<0.01
```

This model displays a significant relationship between crime rate and the weighted mean of distances to five Boston employment centres.

8.

```
===============================================
                 Dependent variable:
                 ---------------------------
                         crim
```

```
---------------------------------------------
rad                      0.618***
                         (0.034)

Constant                 -2.287***
                         (0.443)

---------------------------------------------
Observations              506
R2                       0.391
Adjusted R2                       0.390
Residual Std. Error     6.718 (df = 504)
F Statistic     323.935*** (df = 1; 504)
=============================================
Note:           *p<0.1; **p<0.05; ***p<0.01
```

This model displays a significant relationship between crime rate and the index of accessibility to radial highways.

9.

```
=============================================
                Dependent variable:
                ---------------------------
                         crim
---------------------------------------------
tax                      0.030***
                         (0.002)

Constant                 -8.528***
                         (0.816)

---------------------------------------------
Observations              506
R2                       0.340
Adjusted R2                       0.338
Residual Std. Error     6.997 (df = 504)
F Statistic     259.190*** (df = 1; 504)
=============================================
Note:           *p<0.1; **p<0.05; ***p<0.01
```

This model displays a significant relationship between crime rate and the full-value property-tax rate per \$10,000.

10.
```
================================================
                   Dependent variable:
                   ---------------------------
                            crim
-----------------------------------------------
ptratio                     1.152***
                            (0.169)

Constant                    -17.647***
                            (3.147)

-----------------------------------------------
Observations                 506
R2                          0.084
Adjusted R2                          0.082
Residual Std. Error     8.240 (df = 504)
F Statistic        46.259*** (df = 1; 504)
================================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```
This model displays a significant relationship between crime rate and the pupil-teacher ratio by town.

11.
```
================================================
                   Dependent variable:
                   ---------------------------
                            crim
-----------------------------------------------
black                       -0.036***
                            (0.004)

Constant                    16.554***
                            (1.426)

-----------------------------------------------
Observations                 506
R2                          0.148
Adjusted R2                          0.147
Residual Std. Error     7.946 (df = 504)
F Statistic        87.740*** (df = 1; 504)
================================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```
This model displays a significant relationship between crime rate and the proportion of blacks by town.

12.

```
==================================================
                    Dependent variable:
                    ---------------------------
                              crim
-----------------------------------------------
lstat                         0.549***
                              (0.048)

Constant                      -3.331***
                              (0.694)

-----------------------------------------------
Observations                   506
R2                            0.208
Adjusted R2                             0.206
Residual Std. Error     7.664 (df = 504)
F Statistic        132.035*** (df = 1; 504)
==================================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

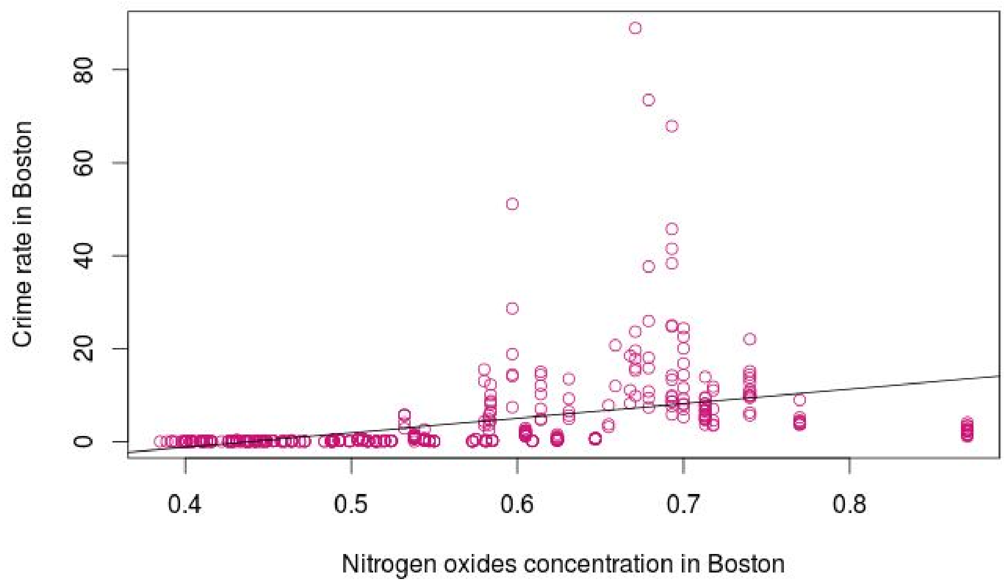
This model displays a significant relationship between crime rate and the lower status of the population.

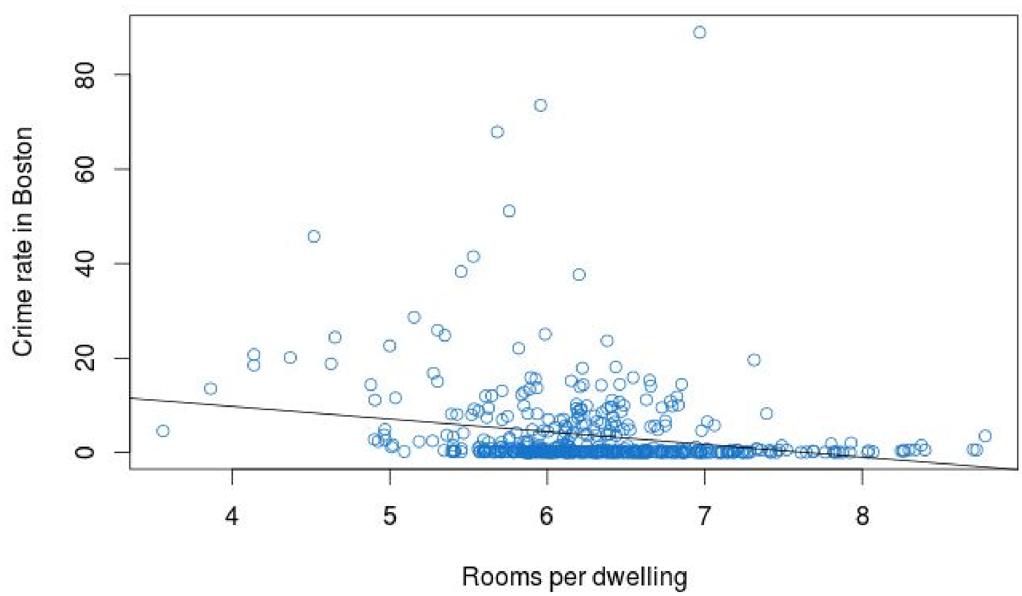
13.

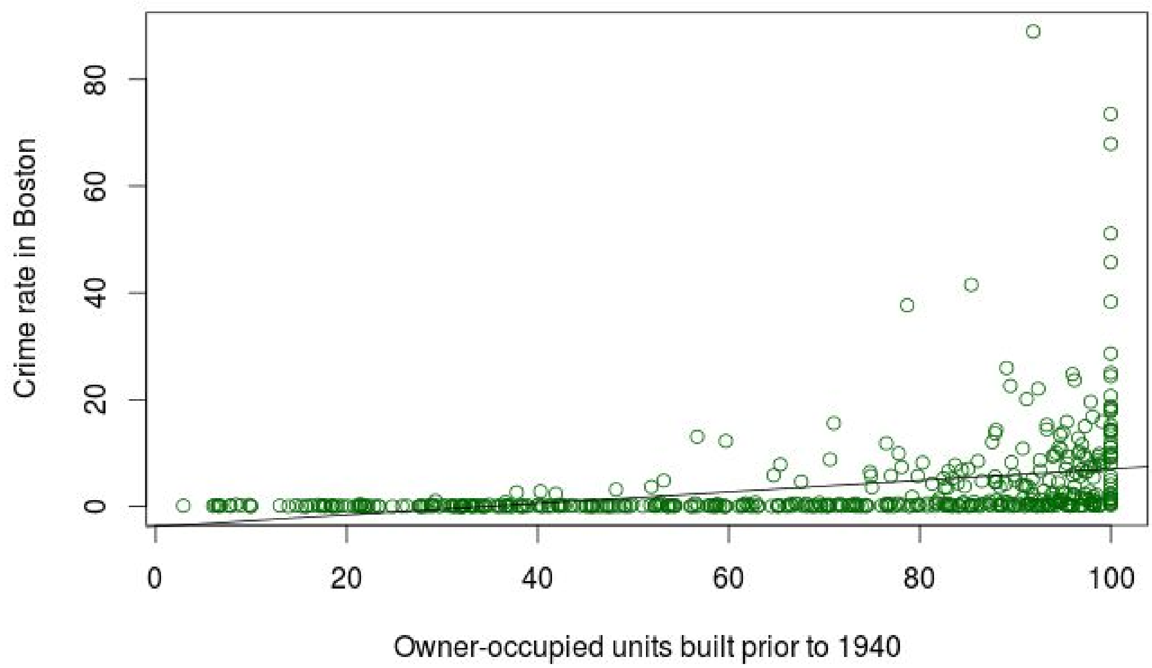
```
==================================================
                    Dependent variable:
                    ---------------------------
                              crim
-----------------------------------------------
medv                          -0.363***
                              (0.038)

Constant                      11.797***
                              (0.934)

-----------------------------------------------
Observations                   506
R2                            0.151
Adjusted R2                             0.149
Residual Std. Error     7.934 (df = 504)
F Statistic        89.486*** (df = 1; 504)
==================================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

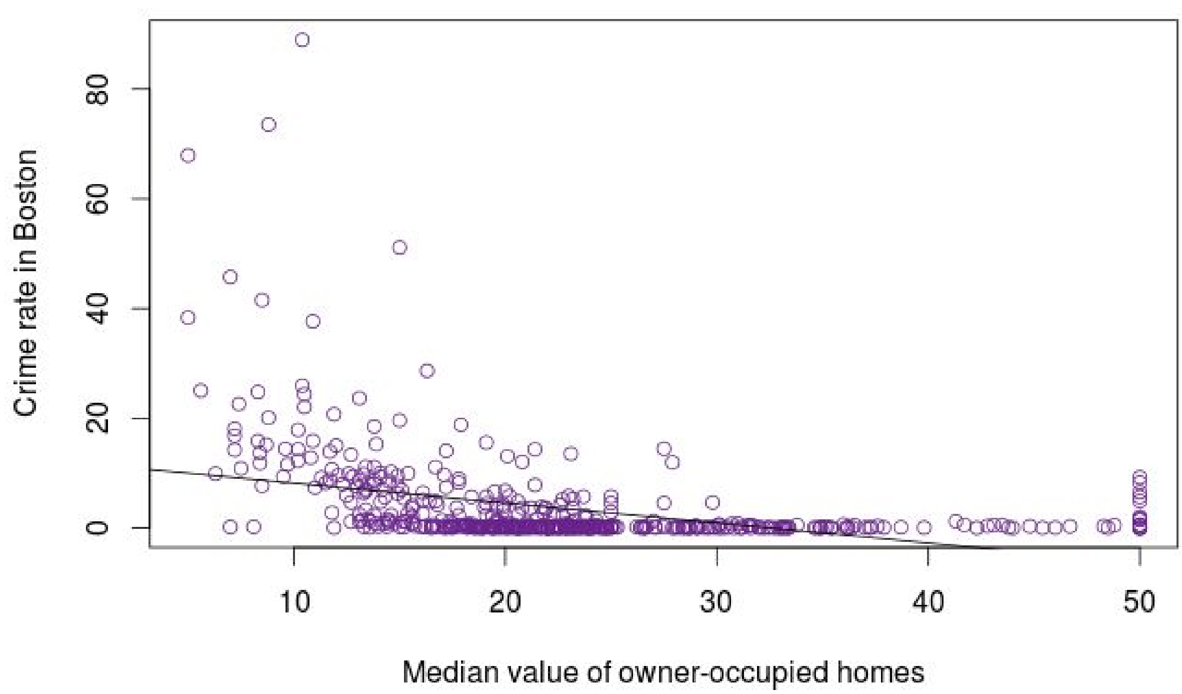
This model displays a significant relationship between crime rate and the median value of owner-occupied homes.

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.
All the predictors except for chas showed a significant association when considered with the response individually. However the R^2 value is very low in a lot of cases, which may mean that the relationships are not that significant. Below are some plots demonstrating the relationships:






Crime rate in Boston
0
20
40
60
80
0
20
40
60
80
100
Owner-occupied units built prior to 1940


Crime rate in Boston
0
20
40
60
80
10
20
30
40
50
Median value of owner-occupied homes

(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results.

```
> library(stargazer)
> lm.all <- lm(crim ~ ., data = Boston)
## using stargazer for pretty summaries
> stargazer(lm.all, type = "text")
================================================
                    Dependent variable:
                    ---------------------------
                            crim
------------------------------------------------
zn                          0.045**
                            (0.019)

indus                       -0.064
                            (0.083)

chas                        -0.749
                            (1.180)

nox                         -10.314*
                            (5.276)

rm                          0.430
                            (0.613)

age                         0.001
                            (0.018)

dis                         -0.987***
                            (0.282)

rad                         0.588***
                            (0.088)

tax                         -0.004
                            (0.005)

ptratio                     -0.271
                            (0.186)

black                       -0.008**
                            (0.004)
```

```
lstat                    0.126*
                         (0.076)

medv                     -0.199***
                         (0.061)

Constant                 17.033**
                         (7.235)

-----------------------------------------------
Observations               506
R2                       0.454
Adjusted R2                        0.440
Residual Std. Error     6.439 (df = 492)
F Statistic      31.470*** (df = 13; 492)
===============================================
Note:            *p<0.1; **p<0.05; ***p<0.01
```

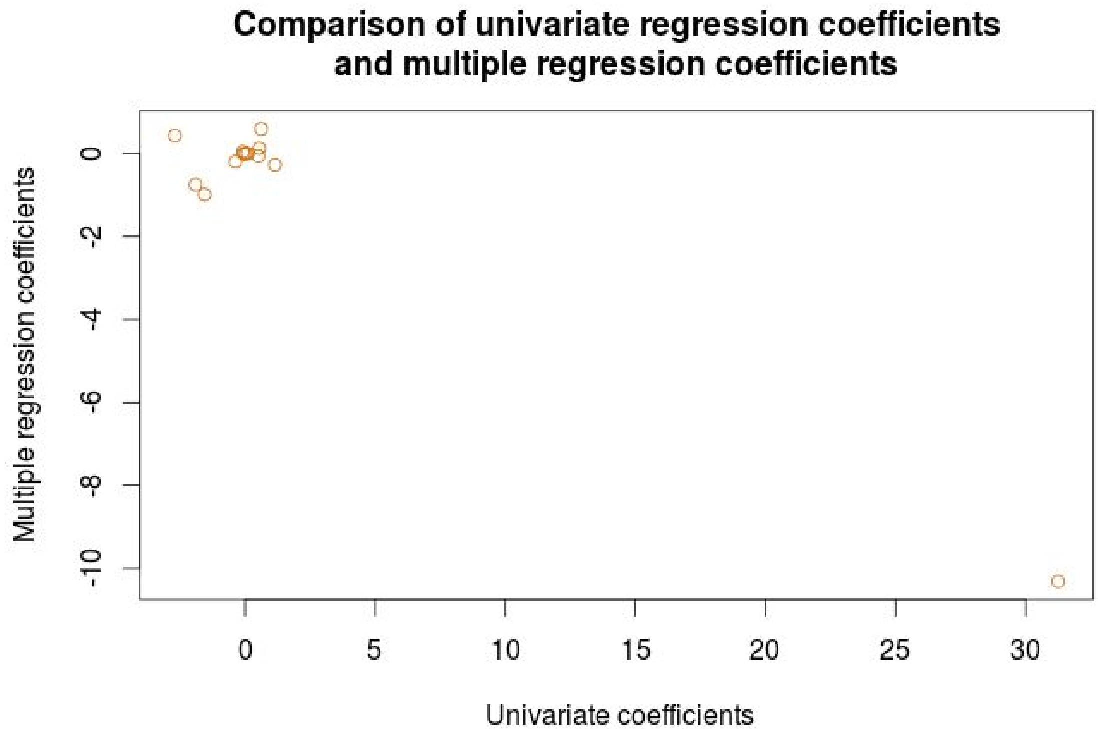
For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

If we decide to keep the variables with the p-value < .01, then we can reject the null hypothesis for dis, rad, medv.

If we decide to keep the variables with the p-value < .05, then we can reject the null hypothesis for the same three variables - dis, rad, medv, - and also for zn and black.

(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.

(See next page)



(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 $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.

1.

```
================================================
                    Dependent variable:
                    ---------------------------
                              crim
------------------------------------------------
zn                            -0.332***
                              (0.110)

I(zn^2)                       0.006*
                              (0.004)

I(zn^3)                       -0.00004
                              (0.00003)

Constant                      4.846***
```

```
                    (0.433)

-----------------------------------------------
Observations              506
R2                        0.058
Adjusted R2                      0.053
Residual Std. Error     8.372 (df = 502)
F Statistic        10.349*** (df = 3; 502)
===============================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

2.

```
===============================================
                    Dependent variable:
                    ---------------------------
                          crim
-----------------------------------------------
indus                     -1.965***
                          (0.482)

I(indus^2)                0.252***
                          (0.039)

I(indus^3)                -0.007***
                          (0.001)

Constant                  3.663**
                          (1.574)

-----------------------------------------------
Observations              506
R2                        0.260
Adjusted R2                      0.255
Residual Std. Error     7.423 (df = 502)
F Statistic        58.688*** (df = 3; 502)
===============================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

3.

	Dependent variable: crim
chas	-1.893
	(1.506)
I(chas2)	
I(chas3)	
Constant	3.744***
	(0.396)
Observations	506
R2	0.003
Adjusted R2	0.001
Residual Std. Error	8.597 (df = 504)
F Statistic	1.579 (df = 1; 504)
Note:	*p<0.1; **p<0.05; ***p<0.01

4.

	Dependent variable: crim
nox	-1,279.371***
	(170.397)
I(nox2)	2,248.544***
	(279.899)
I(nox3)	-1,245.703***
	(149.282)
Constant	233.087***

```
                         (33.643)

-----------------------------------------------
Observations              506
R2                       0.297
Adjusted R2                       0.293
Residual Std. Error     7.234 (df = 502)
F Statistic        70.687*** (df = 3; 502)
===================================================
Note:             *p<0.1; **p<0.05; ***p<0.01
```

5.

```
===================================================
                  Dependent variable:
                  ---------------------------
                         crim
-----------------------------------------------
rm                       -39.150
                         (31.311)

I(rm2)                   4.551
                         (5.010)

I(rm3)                   -0.174
                         (0.264)

Constant                 112.625*
                         (64.517)

-----------------------------------------------
Observations              506
R2                       0.068
Adjusted R2                       0.062
Residual Std. Error     8.330 (df = 502)
F Statistic        12.168*** (df = 3; 502)
===================================================
Note:             *p<0.1; **p<0.05; ***p<0.01
```

6.

```
================================================
                    Dependent variable:
                    ---------------------------
                          crim
------------------------------------------------
age                       0.274
                          (0.186)

I(age2)                   -0.007**
                          (0.004)

I(age3)                   0.0001***
                          (0.00002)

Constant                  -2.549
                          (2.769)

------------------------------------------------
Observations               506
R2                        0.174
Adjusted R2                         0.169
Residual Std. Error       7.840 (df = 502)
F Statistic         35.306*** (df = 3; 502)
================================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

7.

```
================================================
                    Dependent variable:
                    ---------------------------
                          crim
------------------------------------------------
dis                       -15.554***
                          (1.736)

I(dis2)                   2.452***
                          (0.346)

I(dis3)                   -0.119***
                          (0.020)

Constant                  30.048***
                          (2.446)
```

Observations	506
R^2	0.278
Adjusted R^2	0.274
Residual Std. Error	7.331 (df = 502)
F Statistic	64.374*** (df = 3; 502)
Note:	*p<0.1; **p<0.05; ***p<0.01

8.

	Dependent variable:
	crim
rad	0.513
	(1.044)
I(rad2)	-0.075
	(0.149)
I(rad3)	0.003
	(0.005)
Constant	-0.606
	(2.050)
Observations	506
R^2	0.400
Adjusted R^2	0.396
Residual Std. Error	6.682 (df = 502)
F Statistic	111.573*** (df = 3; 502)
Note:	*p<0.1; **p<0.05; ***p<0.01

9.

```
===============================================
                    Dependent variable:
                    ---------------------------
                              crim
-----------------------------------------------
tax                           -0.153
                              (0.096)

I(tax2)                       0.0004
                              (0.0002)

I(tax3)                       -0.00000
                              (0.00000)

Constant                      19.184
                              (11.796)

-----------------------------------------------
Observations                   506
R2                            0.369
Adjusted R2                             0.365
Residual Std. Error           6.854 (df = 502)
F Statistic         97.805*** (df = 3; 502)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

10.

```
===============================================
                    Dependent variable:
                    ---------------------------
                              crim
-----------------------------------------------
ptratio                       -82.361***
                              (27.644)

I(ptratio2)                   4.635***
                              (1.608)

I(ptratio3)                   -0.085***
                              (0.031)
```

```
Constant              477.184***
                      (156.795)

---------------------------------------------
Observations            506
R2                    0.114
Adjusted R2                   0.108
Residual Std. Error   8.122 (df = 502)
F Statistic       21.484*** (df = 3; 502)
=================================================
Note:             *p<0.1; **p<0.05; ***p<0.01
```

11.

```
=================================================
                  Dependent variable:
                  ---------------------------
                      crim
---------------------------------------------
black                 -0.084
                      (0.056)

I(black2)             0.0002
                      (0.0003)

I(black3)             -0.00000
                      (0.00000)

Constant              18.264***
                      (2.305)

---------------------------------------------
Observations                  506
R2                    0.150
Adjusted R2                   0.145
Residual Std. Error   7.955 (df = 502)
F Statistic       29.492*** (df = 3; 502)
=================================================
Note:             *p<0.1; **p<0.05; ***p<0.01
```

12.

```
==============================================
                    Dependent variable:
                    ---------------------------
                              crim
-----------------------------------------------
lstat                         -0.449
                              (0.465)

I(lstat2)                     0.056*
                              (0.030)

I(lstat3)                     -0.001
                              (0.001)

Constant                      1.201
                              (2.029)

-----------------------------------------------
Observations                            506
R2                            0.218
Adjusted R2                             0.213
Residual Std. Error           7.629 (df = 502)
F Statistic         46.629*** (df = 3; 502)
==============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

13.

```
==============================================
                    Dependent variable:
                    ---------------------------
                              crim
-----------------------------------------------
medv                          -5.095***
                              (0.434)

I(medv2)                      0.155***
                              (0.017)

I(medv3)                      -0.001***
                              (0.0002)

Constant                      53.166***
```

```
                    (3.356)

-----------------------------------------------
Observations                      506
R2                         0.420
Adjusted R2                       0.417
Residual Std. Error     6.569 (df = 502)
F Statistic        121.272*** (df = 3; 502)
===============================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

indus, nox, age, dis, ptratio and medv seem to have a non-linear association with the response (crim).