# Advanced Statistical Methods Homework 3

Brandon Hosley

University of Illinois - Springfield

Advanced Statistical Methods Homework 3

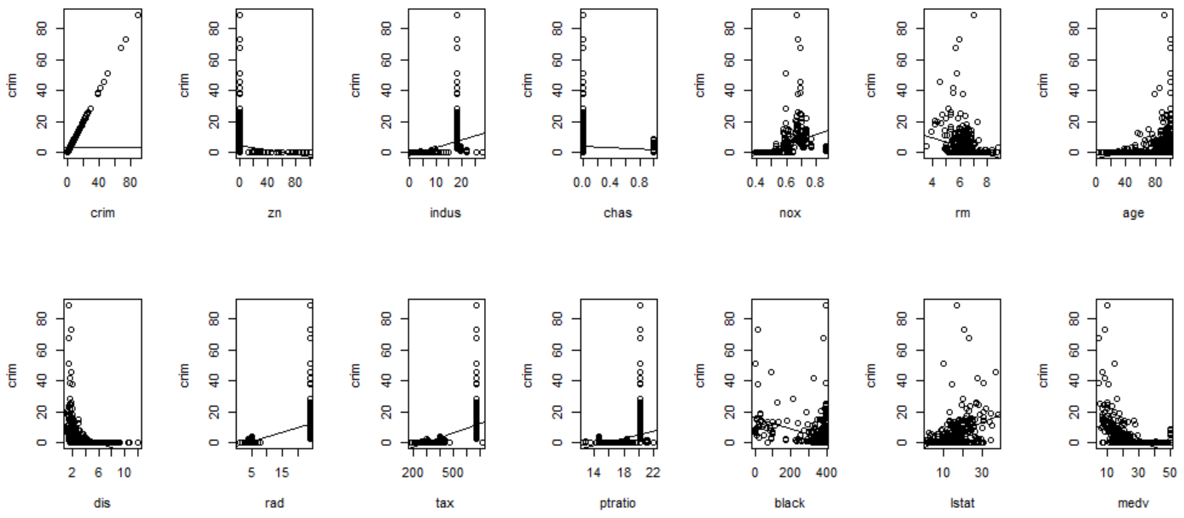
## Introduction to Statistical Learning
## Chapter 3: Problem 15

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.

```
# Load Boston data set
library(MASS)
mount(Boston)
```

### (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.*



```
par(mfrow=c(2,7))
plot(crim,  crim); abline(lm(crim~crim))
plot(zn,    crim); abline(lm(crim~zn))
plot(indus, crim); abline(lm(crim~indus))
plot(chas,  crim); abline(lm(crim~chas))
plot(nox,   crim); abline(lm(crim~nox))
plot(rm,    crim); abline(lm(crim~rm))
plot(age,   crim); abline(lm(crim~age))
plot(dis,   crim); abline(lm(crim~dis))
plot(rad,   crim); abline(lm(crim~rad))
plot(tax,   crim); abline(lm(crim~tax))
plot(ptratio, crim); abline(lm(crim~ptratio))
plot(black, crim); abline(lm(crim~black))
plot(lstat, crim); abline(lm(crim~lstat))
plot(medv,  crim); abline(lm(crim~medv))
```

It appears that only certain predictors have slopes not near-zero or near-infinite: zn, indus, age, rad, lstat, medv.

### (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* $H_0 : \beta_j = 0$*?*

```
summary(lm(crim~., data=Boston))
```

Zn, dis, rad, black, and medv are the predictors that allow rejection of null hypothesis with a fairly high confidence.

### (c)

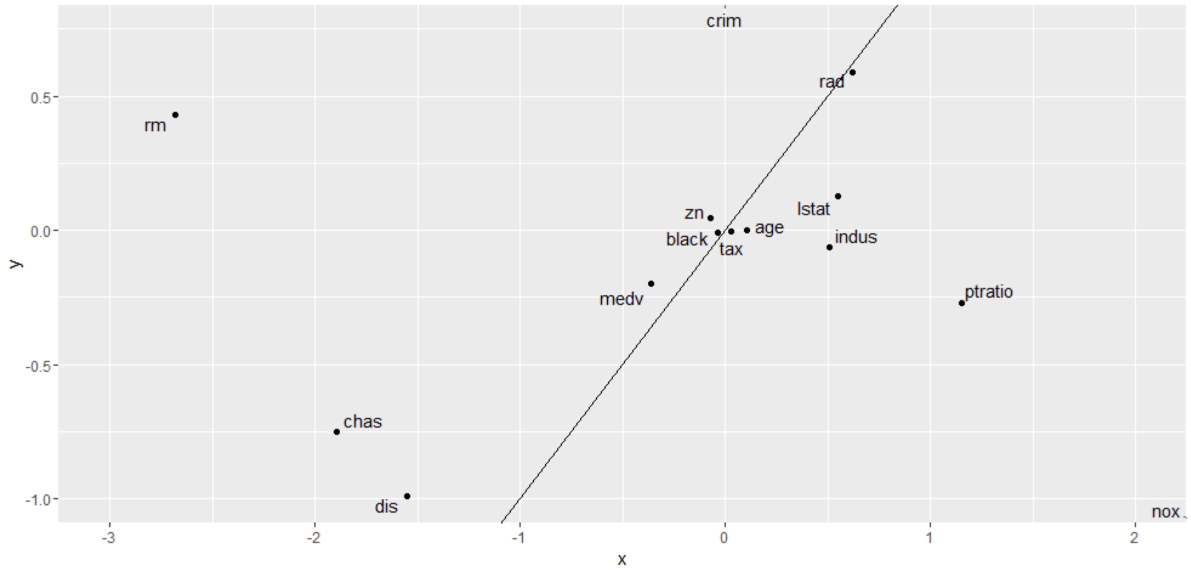
*How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient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

```
predictors <- names(Boston)
y <- coefficients(lm(crim~., data=Boston))
x <- vector(mode="numeric", length=1)
x <- append(x,coefficients(lm(crim~zn))[[2]])
x <- append(x,coefficients(lm(crim~indus))[[2]])
x <- append(x,coefficients(lm(crim~chas))[[2]])
x <- append(x,coefficients(lm(crim~nox))[[2]])
x <- append(x,coefficients(lm(crim~rm))[[2]])
x <- append(x,coefficients(lm(crim~age))[[2]])
x <- append(x,coefficients(lm(crim~dis))[[2]])
x <- append(x,coefficients(lm(crim~rad))[[2]])
x <- append(x,coefficients(lm(crim~tax))[[2]])
x <- append(x,coefficients(lm(crim~ptratio))[[2]])
x <- append(x,coefficients(lm(crim~black))[[2]])
x <- append(x,coefficients(lm(crim~lstat))[[2]])
x <- append(x,coefficients(lm(crim~medv))[[2]])

df <- data.frame(predictors,x,y)

library(ggplot2)
library(ggrepel)
p <- ggplot(df, aes(x,y))
```

```
p <- p + geom_point()
p <- p + geom_text_repel(aes(x,y,label=predictors))
p <- p + coord_cartesian(xlim=c(-3,2),ylim=c(-1,0.75))
p <- p + geom_abline(intercept=0,slope=1)
p
```



This figure shows that for all values other than nox, the coefficient ratios are close to 1. For values above the plotted line the predictors are given a higher coefficient in the multivariate regression model than they would receive on their own.

### (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*

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

```
lm(crim~poly( zn,     3, raw=TRUE))
lm(crim~poly( chas,   3, raw=TRUE))
lm(crim~poly( indus,  3, raw=TRUE))
lm(crim~poly( nox,    3, raw=TRUE))
lm(crim~poly( rm,     3, raw=TRUE))
lm(crim~poly( age,    3, raw=TRUE))
lm(crim~poly( dis,    3, raw=TRUE))
lm(crim~poly( rad,    3, raw=TRUE))
lm(crim~poly( tax,    3, raw=TRUE))
lm(crim~poly(ptratio,3, raw=TRUE))
lm(crim~poly( black,  3, raw=TRUE))
lm(crim~poly( lstat,  3, raw=TRUE))
lm(crim~poly( medv,   3, raw=TRUE))
```

From this information we can determine that certain predictors bear a non-linear relationship to the crime statistic.

dis, medv, nox, ptratio, rm all have significant coefficients associated with their $X^2$ and $X^3$ values.