# CMSC 478 — Spring 2017 — C. S. Marron
# Lab 2: Simple and Multiple Regression

## Collinearity in Simulated Data

**Exercise 1:** Perform the following commands in R:

```
> set.seed(1)
> x1 = runif(100)
> x2 = 0.5*x1 + rnorm(100)/10
> y = 2 + 2*x1 + 0.3*x2 + rnorm(100)
```

The last line corresponds to creating a linear model in which `y` is a function of `x1` and `x2`.

a. Write out the form of the linear model.

b. What are the regression coefficients?

**Exercise 2:** What is the correlation between `x1` and `x2`? Create a scatterplot displaying the relationship between the two variables.

**Exercise 3:** Fit a least-squares regression to predict `y` using `x1` and `x2`. Describe the results obtained.

a. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$?

b. How do these relate to the true $\beta_0$, $\beta_1$, and $\beta_2$?

c. Can you reject the null hypothesis $H_0 : \beta_1 = 0$? Can you reject the null hypothesis $H_0 : \beta_2 = 0$?

**Exercise 4:** Now you will fit two least-squares regressions, one using only `x1` as a predictor, the other using only `x2`.

a. Fit a least-squares regression to predict `y` using only `x1`. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

b. Fit a least-squares regression to predict `y` using only `x2`. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

c. Explain why your results for this question do *not* contradict the results from Exercise 3.

## Multiple Linear Regression

For the following exercises, you will use the Boston data set, which is part of the MASS library. Load the library with the command

```
> library(MASS)
```

The variables in the Boston data set are described in the appendix.

**Exercise 5:** Fit a multiple regression model to predict `crim` using all other variables as predictors. Describe your results. For which predictors can we reject the hypothesis $H_0 : \beta_j = 0$?

**Exercise 6:** For each of the predictors `zn`, `indus`, `nox`, and `medv`, fit a simple linear regression model to predict the response. Describe your results.

a. In which models is there a statistically significant association between the predictor and the response?

b. Explain any discrepancies between the simple linear regression models and the results from Exercise 5.

**Exercise 7:** Is there evidence of non-linear association between any of the predictors `zn`, `indus`, `nox`, or `medv` and the response? To answer this question, for each predictor, fit a model of the form

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

# Appendix: Boston Data Set

**crim**
: per capita crime rate by town.

**zn**
: proportion of residential land zoned for lots over 25,000 sq.ft.

**indus**
: proportion of non-retail business acres per town.

**chas**
: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

**nox**
: nitrogen oxides concentration (parts per 10 million).

**rm**
: average number of rooms per dwelling.

**age**
: proportion of owner-occupied units built prior to 1940.

**dis**
: weighted mean of distances to five Boston employment centres.

**rad**

index of accessibility to radial highways.

**tax**
full-value property-tax rate per $10,000.

**ptratio**
pupil-teacher ratio by town.

**black**
1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

**lstat**
lower status of the population (percent).

**medv**
median value of owner-occupied homes in $1000s.

[Top]