Info 573: Data Science I: Theoretical Foundations
Multiple linear regression exercise
Instructor: Ben Althouse; Contact: bma85@uw.edu

Here you will practice multiple linear regression. Work with a partner on this exercise, we will regroup and go over the answers in class.

The data we are examining is housing values in suburbs of Boston and we wish to examine the various factors influencing the value of homes. Data come from: Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. *J. Environ. Economics and Management* **5**, 81–102.

Dataset description: Table below gives the variables in the dataset.

home.value	median value of owner-occupied homes in $1000s.
NO.concentration	nitrogen oxides concentration (parts per 10 million)
distance.to.work	weighted mean of distances to five Boston employment centers
student.teacher.ratio	Student-teacher ratio by town

Load the data: Download the data from Canvas and use the command `load("BostonData.Rdat")`. Don't forget to set your working directory!

Research questions: we wish to examine the effects of NOx, distance to employment, and education quality (as assessed through the student-teacher ratio) on home value. We will test them individually and then adjust for all factors.

0) Plot a scatterplot matrix of the data and describe what you see. Are any variables tightly related?

```
plot(boston)
```

1) Fit single linear regressions with home.value as the outcome and each of the predictors:

```
mod1 <- lm(home.value ~ NO.concentration, data = boston)
summary(mod1)

mod2 <- lm(home.value ~ distance.to.work, data = boston)
summary(mod2)

mod3 <- lm(home.value ~ student.teacher.ratio, data = boston)
summary(mod3)
```

What are the associations of NOx, distance to employment, and education with home value?

2) Compare the adjusted r^2 values for each model. Which predictor explains the data the best?

3) Run a multiple linear regression with home.value as the outcome and the other three variables as the predictors:

```
mod.full <- lm(home.value ~ distance.to.work + NO.concentration +
student.teacher.ratio, data = boston)
summary(mod.full)
```

Again interpret the associations. Do you see anything surprising (hint: maybe distance to work)? What could explain this discrepancy? Remember you are now adjusting each predictor for the other values.

4) Compare the adjusted r^2 value for this multivariate model to the single linear regression values, which model fits the data the best?

5) Predict and find the prediction interval the median home value of a home 3 km from work, with a NOx concentration of 0.35, and a student-teacher ratio of 10.

```
predict(mod.full, newdata=data.frame("distance.to.work" = 3,
"NO.concentration" = 0.35, "student.teacher.ratio" = 10),
interval="prediction")
```