Stats 10 Lab 3: Linear Regression, Probability, and Sampling

Objectives

1. Understand linear regression in R and verify linear regression assumptions
2. Plotting time series to analyze trends
3. Use R for sampling and simulation

Linear Regression in R

You learned about linear regression in lecture. Now, we will learn how to code it in lab. We will be using the following new commands:

- lm() will create an output containing the equation of the regression line, correlation coefficient, p-values, residuals, and much more. We typically create an object to store the results of the lm() function (see example below).
- abline() is a command to generate a regression line on a plot of the form y = a + bx. It requires arguments for a and b, or linear model results (see example below).

Try running an example of the code on the following page by loading and using NCbirths. We will discuss what each line does in section.

```
## Run the linear model of weight against Mom's age and print a summary
linear_model <- lm(NCbirths$weight ~ NCbirths$Mage)
summary(linear_model)

## Create a plot of the data, and draw the regression line using abline
plot(NCbirths$weight ~ NCbirths$Mage, xlab = "Mom Age", ylab = "Weight",
    main = "Regression of Weight on Mother's Age")
abline(linear_model, col = "red", lwd = 2)

## Create a plot of the residuals to assess regression assumptions
plot(linear_model$residuals ~ NCbirths$Mage, main = "Residuals plot")
## Add a line of y = 0 to help visualize the residuals
abline(a = 0, b = 0, col = "red", lwd = 2)
```

Exercise 1 (The exercise will be included in Part I of Assignment 3.)
We will be working with some soil mining data and are interested in looking at some of the relationships between metal concentrations (in ppm). Download the data 'soil_complete.txt' from the course websit and read it into R. When you read in the data, name your object "soil".

a. Run a linear regression of lead against zinc concentrations (treat lead as the response variable). Use the summary function just like in the example above and paste the output into your report.

b. Plot the lead and zinc data, then use the abline() function to overlay the regression line onto the data.

c. In a separate plot, plot the residuals of the regression from (a), and again use the abline() function to overlay a horizontal line.

Parts d-h can be answered by hand, using a calculator, or any R functions of your choice.

d. Based on the output from (a), what is the equation of the linear regression line?

e. Imagine we have a new data point. We find out that the zinc concentration at this point is 1,000 ppm. What would we expect the lead concentration at this point to be?

f. Imagine two locations (A and B) for which we only observe zinc concentrations. Location A contains 100ppm higher concentration of zinc than location B. How much higher would we expect the lead concentration to be in location A compared to location B?

g. Report the R-squared value and explain in words what it means in context.

h. Comment on whether you believe the three main assumptions (linearity, symmetry, equal variance) for linear regression are met for this data. List any concerns you have.

Exercise 2 (The exercise will be included in Part I of Assignment 3.)
Our next data set is what is known as a time series, or data in time. It contains the measurements via satellite imagery of sea ice extent in millions of square kilometers for each month from 1988 to 2011. Please download the "sea_ice" data from the course website and read it into R. If you have your working directory properly set, you can use the line below:

```
ice <- read.csv("sea_ice.csv", header = TRUE)
```

Note that currently R does not know what class the Date column is. We need to convert the Date column into class "date" using the following line:

```
ice$Date <- as.Date(ice$Date, "%m/%d/%Y")
```

a. Produce a summary of a linear model of sea ice extent against time.

b. Plot the data and overlay the regression line. Does there seem to be a trend in this data?

c. Plot the residuals of the model over time and include a horizontal line. What assumption(s) about the linear model should we be concerned about?

Sampling and simulating in R

We can use the sample() function to sample data from a vector, and the replicate() function to simulate random events many times over. Note that the computer is not truly random, but it is close enough for our purposes to consider it random. We also rely on the set.seed() function to make our "random" results reproducible. Try the following examples in R and see the ## comments for descriptions.

```
## Set seed for reproducibility
set.seed(1335)
## Create a names vector
names = c("Leslie", "Ron", "Andy", "April", "Tom", "Ben", "Jerry")
## Sample 3 of the names with replacement
sample(names, 3, replace = TRUE)
## Sample 3 of the names without replacement
sample(names, 3, replace = FALSE)

## Create a vector from 1 to 10
numbers = 1:10
## Simulate sampling 3 different numbers at random, 10 times
replicate(10, sample(numbers, 3, replace = FALSE))
## One more time, but save as an object
rand_draws = replicate(10, sample(numbers, 3, replace = FALSE))
## Perform analysis on the random draws
colMeans(rand_draws) ## Takes the mean of each sample
colSums(rand_draws) ## Takes the sum of each sample
```

Exercise 3 (The exercise will be included in Part I of Assignment 3.)
One of Adam's favorite casino games is called "Craps". In the first round of this game, two fair 6-sided dice are rolled. If the sum of the two dice equal 7 or 11, Adam doubles his money! If a 2, 3, or 12 are rolled, Adam loses all the money he bets. ☹

a. Based on your lecture notes, what is the chance Adam will double his money in the first round of the game? What is the chance Adam will lose his money in the first round of the game?

b. Let's now approximate the results in (a) by simulation. First, set the seed to 123. Then, create an object that contains 5,000 sample first round Craps outcomes (simulate the sum of 2 dice, 5,000 times). Use the appropriate function to visualize the distribution of these outcomes (*hint: are the outcomes discrete or continuous?*).

c. Imagine these sample results happened in real life for Adam. Using R functions of your choice, calculate the percentage of time Adam doubled his money. Calculate the percentage of time Adam lost his money.

d. Adam winning money and Adam losing money can both be considered events. Are these two events independent, disjoint, or both? Explain why.

e. Quickly mathematically verify by calculator if those events are independent using part (a) and what you learned in lecture. Show work.