StatR 101: Fall 2012
Homework 5
Eli Gurarie
Due Tuesday, October 30, 6:00 pm

Instructions: Please submit a SINGLE DOCUMENT with all the R code, short answers and figures. Upload the completed homework assignment into the course webpage drop-box. This assignment relies heavily (to the point of repeating) the calculations in the Week 5 in-class lab.

1. **Summarizing summary statistics**

(a) Create a function called `SummaryStatsA(x,y)` which takes two vectors and computes the following summary statistics: sample mean and standard deviation of x and y, correlation coefficient r, intercept and slope parameters a and b, the three sums of squares (SS_{total}, SS_{model} and $SS_{residual}$), and the coefficient of determination r^2. Do not use any specialized R functions other than basic arithmetic functions and `sum` or `length`. Use the template below:

```
getSummaryStatisticsByHand <- function(x,y)
{
  n <- length(x)
  x.bar <- sum(x)/n
  y.bar <- sum(y)/n

  # etc...

  return(c(x.bar = x.bar, y.bar = y.bar,
         x.sd = x.sd, y.sd = y.sd,
         r = r, a = a, b = b,
         SS.total = SS.total, SS.model = SS.model, SS.residual = SS.residual,
         r2 = r2))
}
```

Note that you need to name the variables in the output vector in order to identify them later when you run the function. Note also that you will need to create a vector of residuals to obtain some of these variables.

(b) Test the output of this function by plotting and drawing a regression line against any paired set of data you like. This can include some of the data we have seen (e.g. pup sizes, country statistics, iris data, but not Galton's parent-child data) or any other dataset you find or are interested in.

(c) Create a second function called `SummaryStatsB(x,y)` which produces the exact same output but uses various R shortcuts such as `sd()`, `cor()`, and most crucially, `lm()`. Confirm that this function gives the same results as `SummaryStatsA()`.

2. **Anscombe's Quartet** One of the datasets that comes with the `R` base package is called `anscombe`. This is a data frame which consists of four sets of paired data with some surprising properties carefully assembled by Francis Anscombe[1].

(a) Load the data into `R` (by typing "`data(anscombe)`" directly into R) and (without plotting the data) use either of your summary statistics functions from problem 1 to obtain and tabulate the summary statistics for all four data sets.

(b) Now plot all four data sets (on a single plot), with regression lines. Comment on whether you think there is a relationship between the variables in these data sets, whether the linear regression is a valid model, and why or why not.

(c) What do you think Dr. Anscombe's message was when he generated (with, I imagine, considerable effort) these sets of numbers?

Bonus Problem: Fit a quadratic function to the four data sets, plot the fitted line, and report the coefficients and the r^2 value (in a table).

[1] Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27, 1721.