ISTA 116: Lab Assignment #4 (50 pts)

Due Friday, October 21 by 5 P.M.

Problem 1:

(50 pts)

The `Galton86b.csv` data set (on d2l) produced one of the first, if not the very first, regression line(s) ever. It contains two variables, collected from a study of peas. The first, `parent.diameter` is the diameter of each of 7 "parent" peas (in mm). The second, `mean.progeny.diam`, is the mean diameter of all the children of each parent.

a. (3 pts) Compute Pearson's r for this data.

b. (5 pts) Use the `lm()` function in `R` to find the least squares regression line to predict `mean.progeny.diam` from `parent.diameter`. Write out the prediction equation in the form: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

c. (3 pts) Produce a scatterplot of the data, and overlay the regression line.

d. (4 pts) Suppose you find a pea and measure its diameter to be 16.2 mm. How wide would you expect its progeny to be, on average?

e. (12 pts) Check the results of `lm()` by carrying out the computations "by hand" (you can use `R`'s arithmetic functions to speed up the process: e.g., $x - mean(x)$ to compute all the x deviations in one line). Report each of the following components, and show your code and/or calculations.

i. The sum of the squared deviations for `parent.diameter`.

ii. The Sum of the Cross Products.

iii. The slope, $\hat{\beta}_1$.

iv. The intercept, $\hat{\beta}_0$.

f. (4 pts) Produce a plot of the residuals as a function of the *predicted* (i.e., $\hat{y}$) values (you can use the `residuals()` and `fitted.values()` functions on your regression object), and a second with the residuals as a function of the *predictor* (i.e., x) values. Do you see any pattern?

g. (3 pts) Compute the mean and variance of the residuals. These can be interpreted as measures of bias and average squared prediction error (respectively) for your regression model if it were applied to the actual y data (can you see why?)

h. (4 pts) Suppose you didn't have access to `parent.diameter`, but you still wanted to make a prediction for `mean.progeny.diam`. What would your "bias" and "average squared prediction error" be, as defined in part (g), if you just used the sample mean as your prediction every time?

i. (2 pts) By what percent is the average squared prediction error reduced by using the `parent.diameter`?

j. (2 pts) Compute r^2, the square of Pearson's correlation coefficient. Comment on its relationship to the previous question.

k. (4 pts) Compute z-scores for `parent.diameter` and for `mean.progeny.diam`. Compute Pearson's r between these two sets of z-scores. Compare your result to the correlation in the raw data, computed in part (a).

l. (4 pts) Now use `lm()` to compute regression coefficients for the z-scores. Comment on the result as it relates to your previous findings.