ASSIGNMENT #4: DUE FRIDAY, OCTOBER 24

Both classes: All problems I - IV

I. Usually you will be expected to use software to calculate regression equations, etc. in this class. However, to help understand what is going on in the software, you should do one example "by hand." That is the purpose of this problem. You may use a calculator, or (better yet) spreadsheet software such as Excel (or even worksheet-type statistical software such as Minitab, using only worksheet operations) for calculations (but *not* for graphs), *but be sure to follow instructions and show intermediate results as explained below*.

The extinct animal *Archaeopteryx* is of interest in paleontology since it has some similarities to reptiles and some to birds. As of the date these data were obtained, six fossil specimens were known, and only five of these included both the femur and the humerus bones. The data on the lengths of these bones for these five specimens are:

Femur	38	56	59	64	74
Humerus	41	63	70	72	84

Since the specimens differ noticeably in size, one might reasonably doubt that they in fact belong to different species. However, if the lengths of the femur and humerus bones exhibit a linear relationship, then it is more plausible that the specimens indeed belong to the same species and differ in size simply because they died at different ages.

1. Plot the data by hand (*not* by computer). Does the relationship look linear?

2. Calculate the least squares regression equation and the correlation coefficient "by hand" in the sense that you show each step indicated below. As mentioned above, you may do these calculations by calculator or (preferably) column by column on the computer. However, be sure to show each of the following:

a. The mean of the femur lengths.

b. The mean of the humerus lengths.

[Continued next page]

c. A table including each of the following columns (where x = femur length, y = humerus length):

i. A column showing the x_I's
ii. A column showing the y_I's
iii. A column whose sum is SXX
iv. A column whose sum is SXY
v. A column whose sum is SYY
vi. Any other columns which were intermediate steps in calculating any of the columns (iii) – (v).

d. Labels for columns (iii)-(vi) showing what formulas they represent – for example, $x_i (x_i - \overline{x})$.

e. The sums of each of columns (iii) – (v).

f. The calculations of $\hat{\eta}_1$ and $\hat{\eta}_0$ from what precedes

g. The equation of the least squares regression line.

h. The calculation of the correlation coefficient r from the results above.

3. Plot the regression line *by hand* on your scatterplot. *Explain your method* of plotting the line.

4. Add four more columns:

a. one giving the fitted values for the x_i's
b. one giving the residuals $\hat{e}_i$
c. one giving $x_i\hat{e}_i$ (i.e., residual times x value)
d. one giving the squares of the residuals

5. Calculate the sums of the last three columns and verify that the first two of these three sums are (up to rounding error) zero. What is the last sum called?

7. Use your results above to verify equation (6.14) (up to rounding error) for these data.

8. Add the fitted value for x = 64 and the residual for the fourth observation (64,72) to your plot and label them clearly.

[More problems on next page]

II. Suppose we have data (x_i, y_i) and we use it to find the least squares regression line $\hat{y} = \hat{\eta}_0 + \hat{\eta}_1 x$ and variance estimate $\hat{\sigma}^2$. Now suppose we transform the data by defining

$y_i' = a + by_i$ and $x_i' = c + dx_i$

for certain constants a, b, c, and d (with both b and d non-zero). Then we use the transformed data (x_i', y_i') to find the least squares regression line $\hat{y}' = \hat{\gamma}_0 + \hat{\gamma}_1 x'$ and variance estimate $\hat{\tau}^2$.

a. *Use the formulas* for finding SXX, SXY, SYY, RSS, the regression coefficients, the variance estimate, and r^2 to answer the following questions, being sure to give the evidence supporting your answers:

i) What is the relationship between SXX and SX'X'?

ii) What is the relationship between SXY and SX'Y'?

iii) What is the relationship between SYY and SY'Y'?

iv) What is the relationship between RSS calculated using the original data (x_i, y_i), and RSS calculated from the transformed data (x_i', y_i')?

v) What is the relationship between $\hat{\eta}_1$ and $\hat{\gamma}_1$?

vi) What is the relationship between $\hat{\eta}_0$ and $\hat{\gamma}_0$?

vii) What is the relationship between $\hat{\sigma}^2$ and $\hat{\tau}^2$?

viii) What is the relationship between r^2 calculated from the original data (x_i, y_i), and r^2 calculated from the transformed data (x_i', y_i')?

b. Using a data set of your choosing, transform the data (using a = 2, b = 3, c = 4, d = 5) and run two regressions (one on the original data and one on the transformed data) to check your conclusions in part (a).

III. 5.4 (Note: The last paragraph on p. 93 is simply a comment related to Problem 5.4; it is not part of the problem.)

IV. 6.5.1 (Note: Although the problem refers to the data set Forbes1.lsp, for this part of the problem it will be most convenient to use the data set Forbes.lsp, which contains only Forbes' original data. Case number 11 is the same in both data sets.

Note: *The next assignment (due November 7) will be the mid-term exam - it will not be dropped.*