Statistics problems: individual contest

May 2019

1. Suppose we have n pairs of observations (X_i, Y_i), $i = 1, \ldots, n$. Suppose we fit a simple linear regression with Y as the response variable and the value of the regression coefficient estimator is 1. What happens if the role of X and Y are switched, i.e., we fit a simple regression with X as the response variable?

2. A math test consists of 10 questions. For each question, one either answers it correctly ($Y = 1$) or incorrectly ($Y = 0$). Thus for a test taker, his/her answers consist of $Y_1, \ldots, Y_{10}$, where Y_i is the answer to the ith question and takes value 1 or 0. Suppose a reasonable statistical model is that for each student, his/her responses to the 10 questions are independent Bernoulli variables with the following specification:

$$P(Y_i = 1) = 1 - P(Y_i = 0) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}}, \quad i = 1, \ldots, 10,$$

where θ is his/her math ability (different students have different θ values) and b_i is the difficulty level for the ith question. The test is designed, of course, to find out the test taker's θ value. This model implies that a person with higher θ value has a larger probability to answer a question correctly, while a more difficult question (larger b value) make the probability of a correct answer smaller. The teacher allocates 10 points equally to each of the 10 questions for the total of 100 points for the test.

Suppose that student A answered two easiest questions (2 smallest b_i values) incorrectly thus scoring 80 out of 100 and that student B answered two most difficult questions (2 largest b_i values) incorrectly thus also scoring 80 out of 100. Student A claims that it is unfair to him (in comparison to student B) because his 8 correct answers are on the more difficult questions. And more difficult questions should worth more points. Do you think student A has a valid point? Do you think the teacher's scoring system is fair? Explain your thinking from the statistical perspective.

3. Two research centers, A and B, collected two separate data sets to study relationship between two variables X and Y. Center A looked at its data, denoted by $(X_1, Y_1), \ldots, (X_m, Y_m)$, and found a positive correlation. Center B also looked at its own data, denoted by $(X_{m+1}, Y_{m+1}), \ldots, (X_{m+n}, Y_{m+n})$, and also found a positive correlation. Now a new researcher pooled the two data set together into a larger one, $(X_1, Y_1), \ldots, (X_{m+n}, Y_{m+n})$. He claims that for the pooled data set, X and Y are negatively correlated. Do you think this is possible? Explain your answer.