

Lab 2: Regression – due 4/29 in class

The aim of this lab is to familiarize you with standard linear regression calculations and to help clarify the difference between linear and nonlinear regression. Our data are a classic set: the heights of 1,078 pairs of fathers and sons, published by Pearson and Lee in 1903. The data were published rounded to the nearest inch, so uniform random noise has been added to get continuous data, see `/class/data/s131TPS/pearson.dat`.

This data set is used extensively in Freedman, Pisani and Purves “Statistics”, 3rd edition, see esp. chapters 8-12.

1. Draw a scatter plot of the data with father’s height on the horizontal and son’s height on the vertical axis.
2. Calculate the least squares regression line of sons’ heights on fathers’ heights. Then calculate the other least squares regression line: of fathers’ heights on sons’ heights.
3. Give the interpretation of and the relation between the slopes of the two regression lines in 2?
4. Explain the phenomenon of “regression to the mean” in terms of this data set.
5. Estimate the sampling error associated with your least squares estimates of the slope and intercept of the regression lines.
6. Show how to use one of these regression lines to predict a son’s height from his father’s height, and roughly what error do you expect to make, on average. Explain carefully.
7. Graph of averages. Calculate the average height of all sons whose fathers were, say, 64 inches, rounded to the nearest inch. Do this for all rounded fathers’ heights from 59” to 75” and plot these averages against the corresponding rounded fathers’ height. Comment on the plot.
8. Fit a straight line by least squares to the data in question 7. How does this relate to the regression line calculated in 2 above.
9. Graph of SDs. Repeat 7 but this time calculate and plot the SD of the sons’ heights corresponding to fathers’ with heights rounded to a given inch.

Now we will make a new (artificial) data set and repeat some of the preceding calculations with it. For each pair (father’s height, son’s height), replace son’s height by

$$\text{new son's height} = 60 + 0.18(\text{son's height} - 68)^2$$

10. With this new data set, repeat 1, 2, 6, 7, 8 and 9 above. For each of these parts, comment on the difference between the results here and the earlier ones.