Linear Models Activity: Linear Model Fitting

Suppose we have bivariate data from two variables X and Y and based on the scatterplot and the correlation (which is usually denoted r) we wish to fit a linear model to the data. Denote the mean of the X and Y values by $\bar{x}$ and $\bar{y}$ respectively, the standard deviations being s_X and s_Y. The line that minimises the sum of the squares of vertical distances from the line is sometimes called the *regression line* of Y on X and can be written

$$y = a + bx$$

where

$$b = \frac{r s_Y}{s_X}$$

is the slope and the intercept term is

$$a = \bar{y} - b\bar{x}.$$

1. *Exploring the model fitting:* Here we explore some features of the least squares line fit.

   (a) What does the *sign* of the correlation between the X and Y values tell you about the slope of the fitted line? Explain your answer.

   (b) If the standard deviation of the X values is equal to the standard deviation of the Y values, what can you say about the line fitted?

   (c) Suppose we standardise the X values and standardise the Y values. What can you say about the line fitted to the standardised data?

   (d) The *predicted (* or *fitted) values* lie on the line, so for a given value of X, x say, the predicted Y value can be written

   $$\hat{y} = a + bx.$$

   What is the fitted value for Y when $X = \bar{x}$? Show your working clearly.

   (e) The applet available at

   `https://www.geogebra.org/m/ZWSy5SxE`

asks you to guess the least squares fit for a given scatterplot. Notice how the applet shows you not only the regression line but also the "squared errors" for the fit. Play the game enough times that for three games in a row you estimate both the intercept and slope to within 0.5.

2. *Least squares fitting:* Computing the slope and intercept for the regression line given bivariate data is not difficult but tedious if done using a pocket calculator. Finding the correlation requires the most steps. These days we have software at our disposal so do not need to do computations "by hand". It is useful to be able to do the model fitting using software. Some options are listed below:

   (a) *R:* The command `lm`, for "linear model", will create a linear model "object" based on bivariate (or more generally, multivariate) data.

   (b) *Microsoft Excel:* Available to students here, MS Excel enables users to fit a line by least squares via the `LINEST` function.

   (c) *Online applets, such as the ISI suite:* The applets at `http://www.isi-stats.com /isi/applets.html` include one that allows the user to enter bivariate data and perform computations. Click "Correlation/Regression". The "Enter Data" window permits you to submit data in pairs in the form (Explanatory, Response). Click "Use Data" and then select "Show Regression Line" to see the line fitted.

   The following are mean values for pitch (Y, in Hz) for boys of different ages (X, in years) (based on data given in Baken and Orlikoff, 2000, Table 6-5A):

Age:	4	5	6	8	9	11
Mean pitch:	283	252	240	213	219	204

   Using software of your choice, create a graphic to display the data above and then find the regression line for the data. Check whether the point $(\bar{x}, \bar{y})$ lies on your line.
   Baken, Ronald J., and Robert F. Orlikoff. (2000): *Clinical Measurement of Speech and Voice.* San Diego: Singular.