Linear Regression Tutorials

Exercise 1 *We wish to study the relationship between the father's IQ (X) and the son's IQ (Y). From a random sample of 12 father-son pairs, the IQ of each father and his son has been recorded. We thus have two paired samples of measurements (x_i, y_i):*

Father-son pair (index i)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Father's IQ x_i	*123*	*144*	*105*	*110*	*98*	*138*	*131*	*90*	*119*	*109*	*125*	*100*
Son's IQ y_i	*102*	*138*	*126*	*133*	*95*	*146*	*115*	*100*	*142*	*105*	*130*	*120*

- *Plot the points (x_i, y_i) on an orthogonal coordinate system.*
- *Compute the linear correlation coefficient between X and Y.*
- *Does a linear adjustment seem appropriate? If so, compute the parameters of the regression line of Y on X.*
- *Compute the coordinates of the centroid G of the scatterplot.*
- *On the same graph, draw the regression line and verify that it passes through the point G.*
- *Use this regression line to estimate the son's IQ when the father's IQ is 120.*

The following quantities are provided: $\sum x_i = 1392$, $\sum x_i^2 = 164566$, $\sum y_i = 1452$, $\sum y_i^2 = 179068$, $\sum x_i y_i = 170394$.

Exercise 2 *We study the effect of an antibiotic on a bacterial culture.*

Part A: *Ten test tubes are prepared, each containing equal volumes of bacterial culture. Different quantities X of antibiotic are added, and after incubation, the optical density D is measured. The optical density reflects the bacterial concentration in the culture medium.*

Antibiotic X	*0.2*	*0.2*	*0.4*	*0.4*	*0.6*	*0.6*	*0.8*	*0.8*	*1.0*	*1.0*
Optical density D	*19*	*21*	*35*	*38*	*64*	*66*	*115*	*130*	*200*	*210*

1) Plot the scatter diagram representing the optical density D as a function of the antibiotic concentration X.

2) Compute the linear correlation coefficient between X and D.

3) Does a linear relationship appear to be appropriate?

Part B: *The analysis is repeated by defining $Z = \ln(D)$, where $\ln$ denotes the natural logarithm.*

$Z = \ln(D)$	*2.94*	*3.05*	*3.56*	*3.64*	*4.16*	*4.19*	*4.74*	*4.87*	*5.30*	*5.35*

1) Repeat questions 1, 2, and 3 from Part A.

3) Determine the regression line of Z on X and calculate the coefficient of determination.

4) Deduce an explicit expression for D as a function of X, and justify the model used.

Given data: $\sum x_i = 6$, $\sum x_i^2 = 4.4$, $\sum d_i = 898$, $\sum d_i^2 = 126148$, $\sum x_i d_i = 721.2$, $\sum z_i = 41.79$, $\sum z_i^2 = 181.57$, $\sum x_i z_i = 27.42$.

$\mathscr{B}.\ \mathscr{T}.$

Rappels

- The variance of X is **estimated** (in the case of a sample) by:

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{X})^2.$$

- The **covariance** between X and Y is given by:

$$\mathrm{Cov}(X,Y) = S_{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right).$$

- The **slope** β of the regression line is given by:

$$b = \frac{S_{XY}}{S_X^2}.$$

- The **intercept** is:

$$a = \overline{Y} - b\,\overline{X},$$

  where

$$\overline{Y} = \frac{\sum y_i}{n} \quad \text{and} \quad \overline{X} = \frac{\sum x_i}{n}.$$

- The **correlation coefficient** between X and Y is:

$$r_{XY} = \mathrm{corr}(X,Y) = \frac{S_{XY}}{S_X\, S_Y} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{\left[n\sum x_i^2 - (\sum x_i)^2\right]\left[n\sum y_i^2 - (\sum y_i)^2\right]}}.$$