Let $(X_1, Y_2), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample from a population with a linear regression equation:

$$\mu_{Y|x} = \alpha + \beta x.$$

The **least squares** estimators for β and α are

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

and

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2,$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right),$$

and

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2.$$

In **normal regression analysis** we assume that the conditional distribution of Y given x is normal (and that the regression equation is still linear):

$$w(y|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-(\alpha+\beta x)}{\sigma}\right)^2}$$

Under these assumptions the maximum likelihood estimators for α and β are the same as above and the maximum likelihood estimator for σ is

$$\hat{\sigma} = \sqrt{\frac{S_{yy} - \hat{\beta}S_{xy}}{n}}.$$

This leads us to the test statistic

$$t = \frac{\hat{\beta} - \beta}{\hat{\sigma}}\sqrt{\frac{(n-2)S_{xx}}{n}}.$$

which is the value of a random variable having a t distribution and $n-2$ degrees of freedom. The associated $100(1-\alpha)\%$ confidence interval for β is

$$\hat{\beta} \pm t_{\frac{\alpha}{2}, n-2}\hat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}}.$$

In **normal correlation analysis** we assume that X and Y have a bivariate normal distribution (see section 6.7 of the book). Under these assumptions the maximum likelihood estimators for the means are

$$\hat{\mu}_x = \overline{x} \text{ and } \hat{\mu}_y = \overline{y}.$$

The maximum likelihood estimators for the variances are

$$\hat{\sigma}_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} \text{ and } \hat{\sigma}_y = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}.$$

The maximum likelihood estimator for the correlation coefficient is the **sample correlation coefficient**

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

Out test statistic here is

$$z = \frac{\sqrt{n-3}}{2}\ln\left[\frac{(1+r)(1-\rho)}{(1-r)(1+\rho)}\right]$$

which is the value of an approximately standard normal random variable.

1. Let (x_1, y_1, Z_1), $(x_2, y_2, y_2), \ldots,$ (x_n, y_n, z_n) be the values of a random sample from a population with linear regression equation $\mu_{Z|x,y} = \alpha + \beta_1 x + \beta_2 y$. Find a system of equations whose solution $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ comprise the least squares estimators for α, β_1, and β_2.

2. The following data taken from a random sample of 10 students consist of scores on a placement exam (x), number of hours studied for their final exam (y), and their score on the final exam (z).

x	y	z
112	5	79
126	13	97
100	3	51
114	7	65
112	11	82
121	9	93
110	8	81
103	4	38
111	6	60
124	2	86

Assuming that the regression is linear, calculate the least squares estimators $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ and predict the final exam score of a student who scored a 100 on the placement exam and studied for 10 hours.