# Simple Linear Regression

To fit the straight line $y = \alpha + \beta x$ to data $(x_i, y_i)$, $i = 1, 2, \ldots n$ by the method of **least squares** the estimates of slope, $\beta$, and intercept, $\alpha$ , are given by:

$$b = \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i \sum y_i\right)}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2} = \frac{S_{xy}}{S_{xx}}, \qquad a = \bar{y} - b\bar{x}$$

If we assume that the $x_i$ are known and that the $y_i$ have normal distributions with means $\alpha + \beta x_i$ , and constant variance $\sigma^2$, written as $y_i \sim N(\alpha + \beta x_i, \sigma^2)$, then if $x_0$ is a fixed value

$$b \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

$$a \sim N\left(\alpha, \sigma^2\left\{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right\}\right)$$

$$a + bx_0 \sim N\left(\alpha + \beta x_0, \sigma^2\left\{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right\}\right)$$

A common alternative is to use $\hat{\alpha}$ for $a$ and $\hat{\beta}$ for $b$.

# Correlation

Given observations $(x_i, y_i)$, $i = 1, 2, \ldots, n$ on two random variables $X$ and $Y$ the **Pearson (product moment)** correlation between them is given by:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i \sum y_i\right)}{\sqrt{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}\sqrt{\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2}}$$

We use $r$ to estimate the correlation, $\rho$ , between $X$ and $Y$. For large $n$, $r$ is approximately $\sim N\left(\rho, \frac{1}{n-2}\right)$. The (**Spearman**) Rank Correlation Coefficient is given by

$$r_S = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where $d_i$ is the difference between the *ranks* of $(x_i, y_i)$, $i = 1, 2, \ldots, n$ . If ranks are tied, see further reading.

---

**Further reading:** Kotz, S., and Johnson,L. (1988) Encyclopedia of Statistical Sciences, Vols.1-9. New York: John Wiley and Sons.