4.3 Regression and correlation analysis

4.3.1 Preliminaries

Problem: Develop a statistical test which predicts that the theoretical dependence between x and y,

$$y = ax + b,$$

corresponds to the set of numerical data

$$(x_1, y_1); (x_2, y_2); ...; (x_n, y_n)$$

4.3.2 Linear regression

We assume that x is deterministic while y is random.

- Sample mean:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j, \qquad \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

- Sample variance:

$$s_x^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2, \qquad s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$$

- Sample covariance

$$s_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$$

- Linear regression:

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

Example:

$$(20, 3.1); (30, 4.1); (40, 5.4); (50, 6.7)$$

4.3.3 Linear correlation

We assume that both x and y are random.

- Sample correlation coefficient

$$r = \frac{s_{xy}}{s_x s_y} : \qquad -1 \leq r \leq 1$$

- Linear correlation:

$$\frac{y - \bar{y}}{s_y} = r \frac{x - \bar{x}}{s_x}$$

- When $r = 1$ or $r = -1$, all data points belong to the same line, i.e. random variable y is a deterministic linear function of random variable x.

- When $r = 0$, all data points are uncorrelated, i.e. random variables y and x are independent.

- As $0 < r^2 < 1$, there exists a dependence (correlation) between random variables x and y, but this dependence is not linear and may not be deterministic.

Remark: The values for parameters (a, b) and r of the linear regression,

$$a = \frac{s_{xy}}{s_x^2}, \quad b = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}, \quad r = \frac{s_{xy}}{s_x s_y},$$

serve as point estimates for the random variables (α, β) and ρ. Confidence intervals for (α, β) and ρ around (a, b) and r with a given level of confidence can be found from a normal distribution in an advanced statistical algorithm.