Inferential statistics involves also determining whether a r.v. depends on one ore more numerical or quantative vars.

Predictor vars considered as fixed.

Types of relationships: simple, multiple (number of independent vars)

Types of r.: positive, negative

Assumptions for the Correlation Coefficient

Befor regression analysis we first calculate the linear Correlation Coefficient. Assumptions:

- Sample is random sample
- The values have a joint normal distribution (X, Y normally distributed; and given any specific value of Y the X values are normally distributed)

! next math stat.

? unit assotiations

Tests for Correlation

$H_0: r = 0$ ← correlation coefficient

$H_1: r \neq 0$

$$R(X,Y) = \frac{\sum^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{S_X \, S_Y}$$

for large sample

$$T_R = \frac{R(X,Y)\sqrt{n-2}}{\sqrt{1 - R(X,Y)^2}} \approx t(n-2)$$

t-test for two-sided hypothesis

For small samples we use Fisher's z-test

$$Z = \frac{1}{2} \ln\left(\frac{1 + R(X,Y)}{1 - R(X,Y)}\right)$$

Then under H_0

$$Z \approx N\left(0, \frac{1}{n-3}\right)$$

$$Z\sqrt{n-3} \approx N(0,1)$$

$$|Z\sqrt{n-3}| > \lambda_\alpha \xrightarrow{+} \text{reject } H_0$$

Population linear regression math-ly:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

parameters

random error (residual)

Assumptions of LR

- ✓ Error values (ε) are statistically independent
- ✓ Error values are normally distr. for any given value of X
- ✓ Probability distr. of ε is normal

Estimated Regression Model

values β_0, β_1 are estimated from sample data: $Y_i = \beta_0 + \beta_1 X_i \quad i = 1, \dots, n$

$$\varepsilon_i = Y_i - (\beta_0 - \beta_1 X_i)$$

Least Squares Method

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

ε_i terms have a zero mean

multiply $\bar{X}$

$$\begin{cases} \sum^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \quad | : n \\ \sum^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

$$\bar{y} - \beta_0 - \beta_1 \bar{X} = 0$$

* least square method has no assumption about distr (contrary to MLs and moments) But we still make assumption to calculate a standard error and conf. intr-s.

$$\bar{x} \sum^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \; (1)' \qquad \beta_0 = \bar{y} - \beta_1 \bar{X}$$

$$(2) - (1)' = (2) \qquad = ?$$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum (x_i - \bar{x})(y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i) = 0$$

$$\sum^{n} (x_i - \bar{x})(y_i - \bar{y}) - \beta_1 \sum^{n} (x_i - \bar{x})(x_i - \bar{x})$$

$$\beta_1 = \frac{\sum^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum^{n} (x_i - \bar{x})^2}$$

reestimate regression without β_0
if β_0 is not stat. signif.

Multiple R - correlation between
y_i and $\hat{y}_i$