Statistična analiza

LINEARNA REGRESIJA

Večkrat moramo skozi izmerjene točke aproksimirati premico ($y=bx+a$). Denimo, da imamo N točk $T_i(x_i,y_i)$. Postopek imenujemo linearna regresija, kjer minimiziramo odstopanje merskih točk od premice

$$\sum_{i=1}^{N}(y_i - bx_i - a)^2 = \mathrm{f}(a,b) = \min$$

Konstanti b in a določimo z odvodi. Veljati mora

$$\frac{\partial \mathrm{f}(a,b)}{\partial a} = 0, \frac{\partial \mathrm{f}(a,b)}{\partial b} = 0$$

Dobimo enačbo za naklon in odsek

$$b = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \overline{x}^2}$$

$$a = \overline{y} - b\overline{x}$$

kjer so količine

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i\, y_i$$

$$\overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

$$\overline{y^2} = \frac{1}{N}\sum_{i=1}^{N} y_i^2$$

Določimo še regresijski koeficient, ki nam pove, kako dobro linearna funkcija opiše merjene točke. Izračunamo ga kot

$$r = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sqrt{\overline{x^2} - \overline{x}^2}\sqrt{\overline{y^2} - \overline{y}^2}}$$

Nato izračunamo kvadrat odstopanja merskih točk od točk, ki jih napove premica (SSE) oziroma nepojasnjeno odstopanje

$$\widehat{y_i} = bx_i + a$$

$$\sum_{i=1}^{N}(\widehat{y_i} - y_i)^2 = SSE$$

in nato standardno napako regresije

$$s_{y/x} = \sqrt{\frac{SSE}{N-2}}$$

Tu delimo z $(N-2)$, ker so to preostale prostostne stopnje, dve odstranimo z linearno regresijo, kjer določimo dva parametra (a, b). Napaka naklona je enaka

$$s_b = \frac{s_{y/x}}{\sqrt{N\left(\overline{x^2} - \overline{x}^2\right)}}$$

ter napaka odseka

$$s_a = s_{y/x}\sqrt{\frac{1}{N} + \frac{\overline{x}^2}{N\left(\overline{x^2} - \overline{x}^2\right)}}$$

Poglejmo si še nekaj količin. Odstopanje točk, ki jih napove premica, glede na povprečno vrednost, oziroma odstopanje, ki jih pojasni regresija

$$\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2 = SSR$$

Celotno odstopanje y-ov je enako

$$SST = \sum_{i=1}^{N}(y_i - \overline{y})^2 = SSR + SSE$$

Kvadrat korelacijskega koeficienta lahko izračunamo tudi iz odstopanj in sicer kot razmerje med pojasnjenim odstopanjem z linearno regresijo in celotnim odstopanjem

$$r^2 = \frac{SSR}{SST}$$

Fischerjeva F statistka nam pove, če je linearna regresija primerna za opis. Faktor izračunamo kot

$$F = \frac{SSR/1}{SSE/(N-2)}$$

Če je ta faktor večji od kritičnega faktorja, potem je linearna regresija boljši približek, kot da y-ni niso odvisni od x-ov.

Pri računanju si lahko pomagamo še z naslednjimi zvezami

$$SSE = \sum_{i=1}^{N}(\hat{y}_i - y_i)^2 = \sum_{i=1}^{N}(y_i - bx_i - a)^2 = \sum_{i=1}^{N}(y_i - bx_i - \overline{y} + b\overline{x})^2$$
$$= N(\overline{y^2} - \overline{y}^2 - b(\overline{xy} - \overline{x}\cdot\overline{y}))$$
$$SST = \sum_{i=1}^{N}(y_i - \overline{y})^2 = N(\overline{y^2} - \overline{y}^2)$$
$$SSR = \sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2 = Nb(\overline{xy} - \overline{x}\cdot\overline{y})$$