Regresijska analiza

Pogojno povprečje

- (X, Y) statistični vektor.
- Kako je porazdeljena spremenljivka Y ob pogoju, da spremenljivka X zavzame neko fiksno vrednost x?

  Krajše: $\boxed{Y|X = x}$ =?
- Kakšno je povprečje Y ob pogoju, da X zavzame neko fiksno vrednost x (pogojno povprečje ali regresija)?

  Krajše: $\boxed{E(Y|X = x)}$ =?
- PRIMER:
  - X meri dolžino repa v cm.
  - Y meri dolžino kočnika v cm.
  - Zanima nas $Y|X = 5$.
    Porazdelitev dolžine kočnika pri dolžini repa 5 cm.
  - Zanima nas $E(Y|X = 5)$.
    Povprečna dolžina kočnika pri dolžini repa 5 cm.

Regresijska krivulja

- Ko vrednost X-a spreminjamo, se spreminja tudi pogojno povprečje $E(Y|X = x)$. Torej je $E(Y|X = x)$ funkcija spremenljivke x: $f(x) = E(Y|X = x)$.
- Regresijska krivulja je krivulja, ki jo opiše funkcija $f(x) = E(Y|X = x)$.

Regresijska premica

- $(X, Y) \sim N(\mu, \nu, \sigma, \tau, \rho)$ – statistični vektor, dvorazsežna normalna porazdelitev.

  VELJA:
- $Y|X = x \sim N\left(\nu + \frac{\rho\tau}{\sigma}(x - \mu), \tau\sqrt{1 - \rho^2}\right) = N(\mu_p, \sigma_p)$.
- Regresijska krivulja je v tem primeru regresijska premica:
  $y = E(Y|X = x) = \beta x + \alpha$, kjer je
  $\beta = \frac{\rho\tau}{\sigma}$, $\alpha = \nu - \beta\mu$.
- β je regresijski koeficient – smerni koeficient regresijske premice.
- Vse pogojne porazdelitve imajo enak standardni odklon.
- Težišče razsevnega diagrama je v (μ, ν).
- X in Y nista linearno funkcijsko povezani.
- Pogojno povprečje Y je linearno funkcijsko povezano z X.

Cenilke

- Naj bo $X \sim N(\mu, \sigma)$, $Y \sim N(\nu, \tau)$, ρ Pearsonov korelacijski koeficient in naj bodo (x_i, y_i) podatki vzorca velikosti n.
- Cenilka za populacijski regresijski koeficient $\beta = \frac{\rho\tau}{\sigma}$ je vzorčni regresijski koeficient:

  $$B = r\frac{S_Y}{S_X} = \frac{S_{XY}}{S_X^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}.$$

- Cenilka za $\alpha = \nu - \beta\mu$ je $A = \overline{Y} - B\overline{X}$.

Premica najboljšega prileganja

- Regresijska premica je premica, ki se podatkom $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ v ravnini najbolj tesno prilega.
- Normalno porazdeljeni (zvezni) spremenljivki X in Y za $\rho \neq \pm 1$ nista linearno funkcijsko povezani. Če bi bili, bi vsaki vrednosti x_i ustrezala natanko ena vrednost $\hat{y}_i$ na regresijski premici.
- Regresijska premica je določena tako, da je pri njej, v primerjavi z drugimi premicami, vsota kvadratov odklonov $\sum_{i=1}^n (\hat{y}_i - y_i)^2$ najmanjša (metoda najmanjših kvadratov).
- MOTIVACIJA: Na podlagi znanih podatkov določimo model, ki ob podani telesni višini omogoči izračun predvidene telesne teže. Pri tem želimo, da bo naša napoved v povprečju točna - uporabimo regresijsko premico.

Linearni regresijski model

- Enostavna linearna regresija - študij odvisnosti med eno odvisno in eno neodvisno spremenljivko.
- Vrednost odvisne spremenljivke Y želimo izraziti s pomočjo neodvisne spremenljivke X v obliki linearne zveze:

  $Y = a + bX + U$ ali $Y = \hat{Y} + U$, kjer je $\hat{Y} = a + bX$.

  Predvidevamo linearni vpliv X na Y.

  $U = Y - \hat{Y}$ meri ostale vplive (spremenljivka odklona), ki nastane zaradi slučajnih vplivov ali zaradi tega, ker v model niso vključene vse spremenljivke, ki vplivajo na Y.

Predpostavke linearnega modela

- $E(U) = 0$ (sicer lahko popravimo a).
  Linearni model je v povprečju točen.
- Pogojne porazdelitve $Y|X = x$ so normalne in imajo enak standardni odklon: $\sigma(Y|X = x) = \sigma(U)$ (homoskedastičnost).
- Porazdelitve Y pri neodvisnih vrednostih X so neodvisne.

  Pri teh predpostavkah je $\boxed{\hat{Y} = a + bX}$ regresijska premica.

  Torej, $\boxed{b = r\frac{S_Y}{S_X} = \frac{S_{XY}}{S_X^2}}$ in $\boxed{a = \overline{Y} - b\overline{X}}$.
- Označimo: $\hat{y}_i = a + bx_i$.

  $u_i = y_i - \hat{y}_i$ je ostanek, residual.

  Ta razlika pripada spremenljivki U.

Determinacijski koeficient

Spomnimo se analize variance:

- Skupna variabilnost Y je vsota pojasnjene variabilnosti spremenljivke $\hat{Y} = a + bX$ in nepojasnjene variabilnosti spremenljivke U. Torej:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

- Kvocient pojasnjene in celotne variance je enak r^2 (r je vzorčni korelacijski koeficient). Rečemo mu determinacijski koeficient.
- Determinacijski koeficient r^2 pove, kolikšen delež variance spremenljivke Y pojasni spremenljivka X (za primer s telesno višino in težo je $r^2 = 0.7^2 = 0.49$).