3 Linearna regresija

3.1 Linearna regresija

Zanima nas povezanost števila ur učenja na teden z rezultatom na izpitu iz statistike. Vzemimo, da vemo, da se rezultat na izpitu v populaciji porazdeljuje pogojno normalno: $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$.

- Iz spodnjega izpisa preberite ocene populacijskih parametrov. Interpretirajte rezultate, katere ničelne domneve so testirane in kako?

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-20.683  -4.746   2.844   4.512  14.693

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.2049     7.5172   2.555 0.033921 *
x             3.6850     0.6217   5.927 0.000351 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.44 on 8 degrees of freedom
Multiple R-squared: 0.8145,     Adjusted R-squared: 0.7913
F-statistic: 35.13 on 1 and 8 DF,  p-value: 0.0003508
```

- Kako bi ničelno domnevo $H_0 : \beta_1 = 0$ preverili s posplošenim testom razmerja verjetij?
  Namig: Kjer je le mogoče, uporabite rezultate iz prejšnje naloge

Predlogi za vaje v R-u:

- Naj bo X enakomerno porazdeljena spremenljivka (med 0 in 20, zaokrožena navzdol), $\beta_0 = 15$, $\beta_1 = 4$, $\sigma = 10$. Generirajte vzorec velikosti 10, narišite podatke in vrišite populacijsko ter ocenjeno vrednost premice.

- Izračunajte posplošeni test razmerja verjetij v R-u

5.2 Matrično računanje

Vrednosti neodvisnih spremenljivk združimo v matriko X (design matrix), vrednosti odvisne spremenljivke ter koeficientov predstavljajo vektorja Y in β:

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}; \; Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Matrika X je dimenzije $n \times (p+1)$, kjer je p število spremenljivk. Če naš model ne bi vseboval konstante, bi prvi stolpec X izpustili.

- Zapišite vsoto vrednosti $\sum_{i=1}^{n} Y_i^2$ v matrični obliki.
- Kaj dobimo, če matrično pomnožimo $X\beta$?
- V matrični obliki oceno koeficientov po metodi najmanjših kvadratov (= po metodi največjega verjetja) zapišemo kot $\widehat{\beta} = (X^T X)^{-1} X^T Y$. Pokažite, da za $p = 1$ dobite oceni:

$$\widehat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

- Izpeljite oceno po metodi najmanjših kvadratov še v matrični obliki. Pri tem boste potrebovali naslednje formule za matrično računanje:

$$\begin{aligned} (A+B)^T &= A^T + B^T; \;\; (A^T)^T = A; \;\; (AB)^T = B^T A^T; \\ \frac{\partial \beta^T A}{\partial \beta} &= A; \;\; \frac{\partial \beta^T A^T A \beta}{\partial \beta} = 2A^T A\beta \end{aligned}$$

Namig: Kaj minimiziramo? Kako zapišemo vsoto kvadriranih ostankov v matrični obliki?

- Izpeljite formulo za standardno napako ocenjenih koeficientov v matrični obliki. Intuitivno razložite od česa je odvisna standardna napaka koeficienta β_1 (za $p = 1$). Uporabite formulo: $\text{var}(cY) = c\,\text{var}Y c^T$.
- Kako bi izračunali interval zaupanja za napovedano premico v našem primeru ($p = 1$)? Kako bo tak interval izgledal na sliki?
- Recimo, da nas zanima, kako sta število ur učenja in spol (0=ženski, 1=moški) povezana z rezultatom na izpitu iz statistike. Predpostavimo model, ki vključuje interakcijo. Kako bi preverili, ali je število ur učenja pri moških povezano z rezultatom na izpitu?

Predlogi za vaje v R-u:

- V R-u v matrični obliki zapišite podatke, izračunajte oceno po metodi najmanjših kvadratov ter jo primerjajte z izpisom R-ove funkcije `lm`.
- Ocenite tudi standardno napako ter jo primerjajte z izpisom
- Narišite sliko napovedane premice ter ji dodajte interval zaupanja.

5.3 Predpostavke linearne regresije

Z osnovnim modelom linearne regresije naredimo štiri predpostavke:

- Ostanki so okrog premice porazdeljeni normalno
- Varianca ostankov ni odvisna od vrednosti neodvisne spremenljivke (homoskedastičnost)
- Ostanki so med seboj neodvisni.
- Povezanost med X in Y je linearna

Kaj se zgodi z ocenami koeficientov, njihovo pričakovano vrednostjo, standardno napako in z intervali zaupanja, če je katera izmed prvih treh predpostavk kršena?

- Kaj se spremeni v izpeljavah, če ostanki okrog premice niso porazdeljeni normalno?
- Recimo, da je varianca ostankov odvisna od x.
- Kaj pa če ostanki med seboj niso neodvisni?