Enorazsežna (enostavna) linearna regresija

Iščemo premico $y = \hat{k}x + \hat{n}$, ki najbolje aproksimira podatke $(x_1, y_1), \ldots, (x_n, y_n)$. Metoda najmanjših kvadratov: vsota kvadratov vertikalnih odstopanj od iskane **regresijske premice** je najmanjša. Pri napovedih naredimo napako ϵ s povprečno vrednostjo $\mu(\epsilon) = 0$ in neznano varianco. Če pišemo

$$y_i = \hat{y}_i + \epsilon_i = (\hat{k}x_i + \hat{n}) + \epsilon_i,$$

kjer je y_i dejanska vrednost, $\hat{y}_i$ pa napovedana vrednost, regresijsko premico po metodi najmanjših kvadratov določimo tako, da je

$$SS_E = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{najmanjša.}$$

Vertikalno razliko oziroma odklon $\epsilon_i = y_i - \hat{y}_i$ imenujemo tudi **residual**.

Preverjanje hipotez

Naj bo $S_{XX} = \sum_{i=1}^{n} (x_i - \overline{x})^2$. Pri preverjanju hipoteze $H_0 : k = k_0$ o dejanskem smernem koeficientu premice vzamemo t-porazdeljeno slučajno spremenljivko

$$T_0 = \frac{\hat{k} - k_0}{\frac{SS_E}{(n-2)S_{XX}}}$$

z $n - 2$ prostostnimi stopnjami.

Poseben primer $H_0 : k = 0, \quad H_1 : k \neq 0$ preverja **moč** oziroma **signifikanco regresije** (signifikanco lahko sicer preverjamo tudi z analizo variance ANOVA, kjer bomo imeli F-porazdelitev). Če ničelne hipoteze $H_0 : k = 0$ ne moremo zavrniti, potem je slučajna spremenljivka Y neodvisna od X ali pa odvisnost ni linearna. Če jo zavrnemo, je Y linearno odvisna od X ali pa odvisnost ni linearna, vendar linearni člen delno vpliva na odvisnost.

Intrinzična linearna odvisnost

Včasih je iz grafičnega prikaza (npr. razpršenega diagrama) razvidno, da odvisnost med dvema spremenljivkama ni linearna. Ko lahko z ustrezno transformacijo vseeno pridemo do linearne odvisnosti, takšne nelinearne modele imenujemo **intrinzično linearni**. Primeri intrinzično linearnih funkcij:

$$\begin{aligned} Y = \beta_0 e^{\beta_1 X} &\Rightarrow \ln Y = \ln \beta_0 + \beta_1 X, \\ Y = \beta_0 + \beta_1 \frac{1}{X} &\Rightarrow Y = \beta_0 + \beta_1 Z. \end{aligned}$$

Večrazsežna (multipla) linearna regresija (MLR)

Iščemo večrazsežno linearno funkcijo $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_n x_n$, ki najbolje aproksimira podatke $\{(x_{1i}, \ldots, x_{ni}, y_i)\}_i$. Metoda najmanjših kvadratov: vsota kvadratov odmikov od

iskane regresijske funkcije je najmanjša. Spet pišemo

$$y_i = \hat{y}_i + \epsilon_i = (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_n x_{ni}) + \epsilon_i,$$

kjer je y_i dejanska vrednost, $\hat{y}_i$ pa napovedana vrednost, in regresijsko funkcijo po metodi najmanjših kvadratov določimo tako kot pri enostavni linearni regresiji.

Preverjanje ustreznosti regresijskega modela

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

celotna varianca — pojasnjena varianca — nepojasnjena varianca

Delež pojasnjene variance (SS_R) v celotni varianci (SS_T) imenujemo **determinacijski koeficient**:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Determinacijski koeficient je enak kvadratu korelacijskega koeficienta. Velja $0 \leq R^2 \leq 1$. Manjši residuali (napake) dajo večji R^2. Pri uporabi R^2 je treba biti pazljiv. Če n podatkov aproksimiramo s polinomom stopnje $n-1$, je $R = 1$. Če v model dodamo novo neodvisno spremenljivko, se R^2 poveča, kar pa ne pomeni nujno, da je nov model boljši od prejšnjega. R^2 je lahko velik tudi, če sta Y in X nelinearno odvisni. Včasih zato uporabljamo **prilagojen koeficient determinacije**, ki oceno prilagodi na osnovi velikosti vzorca in števila neodvisnih spremenljivk.

Ustreznost regresijskega modela temelji na upravičenosti predpostavk. Linearna regresija predpostavlja, da je odvisnost v modelu linearna (npr. grafični prikaz odvisnosti, korelacijski koeficienti, R^2, signifikanca regresije). Pri preverjanju hipotez in računanju intervalov zaupanja pa predpostavljamo, da so napake ϵ normalno porazdeljene s povprečjem 0 in konstantno varianco (npr. grafični prikaz residualov, testiranje normalnosti).

V programskem paketu R imamo na voljo naslednje ukaze za linearno regresijo po metodi najmanjših kvadratov.

- `fit = lsfit(X, y)`, kjer je X matrika vrednosti neodvisnih spremenljivk (vsak stolpec pripada eni neodvisni spremenljivki), y pa vektor vrednosti odvisne spremenljivke. Izračunane parametre modela dobimo npr. kot `fit$coef`. Z `ls.print(fit)` dobimo osnovne statistične parametre regresije, vključno s standardno napako, t-vrednostjo in P-vrednostjo za regresijske koeficiente.
- `fit = lm(y ~ x1 + x2 + ... + xn)`, kjer so $x_1, ..., x_n$ vektorji vrednosti neodvisnih spremenljivk, y pa vektor vrednosti odvisne spremenljivke. Izračunane parametre modela dobimo npr. kot `fit$coef`. S `summary(fit)` dobimo osnovne statistične parametre regresije.

Uporabna statistika 2013/2014

VAJA 13: Linearna regresija

1. V grafičnem vmesniku programskega paketa R odprite novo skriptno datoteko.

2. V datotekah *padavine2011.txt* in *padavine2011.ods* so podatki o količini padavin v letu 2011 za glavne meteorološke postaje v Sloveniji. Podani sta nadmorska višina (m) in letna količina padavin (mm). Podatki kažejo, da se s povečevanjem nadmorske višine povečuje tudi količina padavin.

   a.) Podatke shranite v dva vektorja, v enem naj bodo podatki o nadmorski višini, v drugem pa podatki o letni količini padavin.

   b.) Podatke prikažite z razpršenim diagramom. Vsaka meteorološka postaja naj bo grafično predstavljena z eno točko, na abscisni osi naj bo nadmorska višina, na ordinatni osi pa letna količina padavin. Ali slika nakazuje linearno odvisnost?

   c.) Poiščite enačbo premice, ki se danim točkam glede na kriterij najmanjših kvadratov najbolje prilega (regresijsko premico). Kakšen je njen smerni koeficient?

   d.) Regresijsko premico dodajte na zgornji graf, na katerem je vsaka meteorološka postaja prikazana z eno točko. Sliko shranite pod imenom *padavine2011-regresija.jpg*.

   e.) Napake linearnega modela prikažite s histogramom. Nanj dodajte tudi normalno krivuljo. Histogram shranite pod imenom *padavine2011-hist.jpg*.

   f.) Izračunajte determinacijski koeficient R^2 (delež pojasnjene variance v celotni varianci) in korelacijski koeficient obeh spremenljivk. Kaj opazite?

   g.) Določite P-vrednost testiranja hipoteze o signifikanci regresije. Ali lahko trdite, da je model ustrezen?

   R: a.) - b.) - c.) $y = 0.306 \cdot x + 756.646$, $k = 0.3065432$ d.) - e.) - f.) $R^2 = 0.64182$, $\rho = 0.8011$ g.) $P = 0.000996$, linearna odvisnost je statistično značilna

3. Na strani 474 v 3. izdaji knjige *Applied Statistics and probability for Engineers* je tabela s podatki o proizvodnji vetrne elektrarne v odvisnosti od hitrosti vetra.

   a.) Podatke shranite v dva vektorja, v enem naj bodo podatki o hitrosti vetra, v drugem pa podatki o proizvodnji vetrnice.

   b.) Podatke prikažite grafično. Kakšno odvisnost nakazuje slika? Ali je proizvodnja vetrne elektrarne navzgor omejena?

   c.) Poiščite in prikažite regresijsko premico. Sliko shranite pod imenom *vetrnica-regresija-1.pdf*. Izračunajte tudi determinacijski koeficient R^2 in P-vrednost pri signifikanci regresije.

   d.) Podatke za vetrnico modelirajte z dvema intrinzično linearnima funkcijama. Katera od njiju je za to bolj primerna? Oba modela tudi zapišite.

   e.) Na zgornjem grafu prikažite še vaš najboljši intrinzično linearni model. Sliko shranite pod imenom *vetrnica-regresija-2.pdf*.

R: a.) - b.) logaritemsko oziroma eksponentno, da c.) $Y = 0.13 + 0.24X$, $R^2 = 0.8745$, $P = 7.55 \cdot 10^{-12}$ d.) eksponentni: $lnY = -0.94 + 0.21X$ ali $Y = e^{-0.94} \cdot e^{0.21X}$ ($R^2 = 0.6036$); $1/X$ model: $Y = 2.98 - 6.93 \cdot \frac{1}{X}$ ($R^2 = 0.98$), najprimernejši je zadnji model e.) -

4. V datoteki *vpis.csv* so podatki o vpisu na neko univerzo za zadnjih 29 let: podani so stopnja brezposelnosti (BREZPOSELNI), število spomladanskih maturantov (MATURANTI) in dohodek na prebivalca (DOHODEK) v neki ameriški zvezni državi, napovedali bi radi jesenski vpis na univerzo (VPIS).

   a.) Podatke shranite v matriko.

   b.) Podatke prikažite grafično – na treh vzporednih slikah: vpis v odvisnosti od stopnje brezposelnosti, vpis v odvisnosti od števila maturantov, vpis v odvisnosti od dohodka. Na sliki naj bodo poleg točk, ki odvisnost opisujejo, tudi regresijske premice, ki se točkam glede na kriterij najmanjših kvadratov najbolje prilegajo. Sliko shranite pod imenom *vpis-odvisnosti.pdf*. Ali slike nakazujejo linearno odvisnost?

   c.) Analizirajte korelacijske koeficiente med spremenljivkami. Kateri dve slučajni spremenljivki korelirata najmočneje? Katera spremenljivka najbolj vpliva na vpis?

   d.) Z multiplo linearno regresijo poiščite model za napovedovanje vpisa na univerzo v odvisnosti od št. maturantov in dohodka. Zapišite enačbo za napovedovanje vpisa in z njeno pomočjo napovejte vpis na univerzo, če veste, da je spomladi maturiralo 16000 dijakov in je povprečni bruto dohodek na prebivalca 3200 dolarjev. Kaj pa, če dohodek pade na 3000 dolarjev?

   e.) Z multiplo linearno regresijo poiščite model za napovedovanje vpisa na univerzo v odvisnosti od stopnje brezposelnosti, št. maturantov in dohodka. Zapišite enačbo za napovedovanje vpisa in z njeno pomočjo napovejte vpis na univerzo, če veste, da je stopnja brezposelnosti enaka 9%, da je spomladi maturiralo 16000 dijakov in je povprečni bruto dohodek na prebivalca 3000 dolarjev. Ta odgovor primerjajte z napovedjo modela z dvema parametroma.

   f.) Analizirajte izpis funkcije `summary(ime_modela)` za zadnji regresijski model.

   g.) Napake linearnih modelov prikažite na dveh vzporednih histogramih. Sliko shranite pod imenom *vpis-hist.pdf*.

   a.) - b.) - c.) dohodek ($\rho = 0.95$) d.) VPIS $= -6457.26 + 0.38 \cdot$ MATURANTI $+$ $4.73 \cdot$ DOHODEK, napoved=14734 in 13787 e.) VPIS $= -9153.25 + 450.12 \cdot$ BREZPOSELNI $+ 0.41 \cdot$ MATURANTI $+ 4.27 \cdot$ DOHODEK, napoved=14226 f.) - g.) -

5. Vsebino skriptne datoteke shranite pod imenom 'vaja13.r'.