Linearna regresija

Vesna Zupanc

23 1 2021

Teoretični del

Naloga je osredotočena na linearno regresijo, ki spada pod posplošene linearne modele. V tem poglavju bomo predstavili, kaj so linearni in posplošeni linearni modeli ter metode, s pomočjo katerih lahko ocenjujemo regresijske koeficiente modela.

Linearna regresija

Linearna regresija je statistični model, ki ga v najbolj enostavni obliki lahko zapišemo kot:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

kjer so ϵ_i med seboj neodvisne slučajne spremenljivke, x_i pa dane vrednosti. Velja $\epsilon_i \sim N(0, \sigma^2)$ za vsak i in tako $Y_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$. Model lahko razširimo na več linearnih parametrov:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

kjer so ϵ_i neodvisne enako porazdeljene slučajne spremenljivke in $1 \leq i \leq n$ (Pfajfar, 2018).

Lahko ga zapišemo tudi v matrični obliki:

$$Y = X\beta + \epsilon.$$

Med temeljne predpostavke regresijskega modela spada predpostavka, da med neodvisnimi spremenljivkami ni popolne kolinearnosti ali multikolinearnosti. Najbolj tipičen vzrok za kršenje te predpostavke je, da smo v model kot neodvisni vključili dve spremenljivki, med katerima obstaja močna linearna povezanost. Do multikolinearnosti pride tudi, če v model vključimo več spremenljivk kot je velikost vzorca. Z naraščanjem multikolinearnosti narašča tudi varianca ocen regresijskih koeficientov, kar pa vpliva tudi na značilnost samih koeficientov, ki so zato pogosto neznačilni. Ne vpliva pa na vrednost R^2. Prvi indikator multikolinearnosti je torej visoka vrednost R^2 pri statistični neznačilnosti večine ocenjenih regresijskih koeficientov (Pfajfar, 2018). Še pred izdelavo modela pa lahko pogledamo korelacijske koeficiente med pojasnjevalnimi spremenljivkami, ki ne smejo biti močno povezane. Kot je bilo že omenjeno, močno povezane spremenljivke povzročajo multikolinearnost. Lahko pa jo odkrivamo tudi z uporabo ANOVE, izračunom variančnega inflacijskega faktorja in še z drugimi pristopi (Bhandari, 2020).

Opazovanja so med seboj neodvisna. V primeru kršenja te predpostavke je smiselno uporabiti posplošene linearne modele, običajno longitudinalni (vzdolžni) model. Vse predpostavke linearnega regresijskega modela so navedene v naslednjem razdelku.

Linearni modeli

Cilj linearne regresije je oceniti vpliv neodvisnih spremenljivk na odvisno spremenljivko. Ker spada pod posplošene linearne modele, si poglejmo še nekaj teorije o linearih modelih.

Normalni linearni model

Normalni linearni model zapišemo kot

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i,$$

kjer velja $\epsilon_i \sim N(0, \sigma^2)$.

Parametre v normalnem linearnem modelu lahko ocenjujemo z metodo najmanjših kvadratov ali pa z metodo največjega verjetja, kjer maksimiziramo log-verjetnostno funkcijo. V našem primeru bomo za ocenjevanje parametrov pri linearni regresiji uporabljali funkcijo `lm`, ki v splošnem uporablja osnovno metodo najmanjših kvadratov (MNK ali angleško OLS), zato v je v nadaljevanju podrobneje predstavljena samo ta metoda.

Metodo najmanjših kvadratov je pri 16 letih odkril nemški matematik Carl F. Gauss. Zaradi svojih lastnosti je najbolj razširjena metoda ocenjevanja regresijskih koeficientov (Pfajfar, 2018, str.53). Pri MNK na primeru osnovnega regresijskega modela velikosti $p = 1$ iščemo β_0 in $\beta 1$ tako, da bo vsota kvadratov ostankov najmanjša možna. Pri danih (x_i, y_i) torej iščemo

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Za razumevanje oznak v predpostavkah metode ločimo dva modela, in sicer linearni vzorčni regresijski model $y_i = b_1 + b_2 x_i + e_i$ in linearni populacijski regresijski model $y = \beta_1 + \beta_2 x_i + u_i$. Pfajfar (2018) navaja naslednje predpostavke metode najmanjših kvadratov:

- linearnost regresijskega modela: $y = \beta_1 + \beta_2 x_i + u_i$
- ničelna povprečna vrednost u_i: $E(u_i) = 0$
- homoskedastičnost: $Var(u_i) = E(u_i^2) = \sigma^2$
- odsotnost avtokorelacije: $cov(e_i, e_j | x_i, x_j) = 0$ za vsak $i \neq j$
- nekoreliranost med pojasnjevalnimi spremenljivkami in slučajno spremenljivko u: $Cov(x_2, u) = Cov(x_3, u) = ... = Cov(x_k, u) = 0$
- število opazovanj mora presegati število ocenjenih parametrov oz. pojasnjevalnih spremenljivk: $n > k$
- $Var(X)$ je končno pozitivno število
- pravilno specificiran regresijski model: vključene vse relevantne pojasnjevalne spremenljivke in izbrana ustrezna funkcijska oblika modela
- odsotnost popolne multikolinearnosti: $\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_k X_k = 0$
- slučajna spremenljivka u je normalno porazdeljena: $u_i \sim N(0, \sigma_u^2)$. Posledično je pogojna porazdelitev odvisne spremenljivke y tudi normalna in sicer $N(\beta_1 x_{1i} + ... + \beta_k x_{ki}, \sigma_u^2)$

Normalni linearni model ima zaradi potrebnih predpostavk kar nekaj omejitev. Kadar ni izpolnjena predpostavka o normalni porazdelitvi ostankov in odvisne spremenljivke, si lahko pomagamo z raznimi transformacijami, s katerimi preoblikujemo odvisno spremenljivko tako, da s transformacijo dobimo normalno porazdelitev $g(y_i) \sim N(x_i^T \beta, \sigma^2)$, ali pa si pomagamo s posplošenimi linearnimi modeli.

Posplošeni linearni modeli

Pri posplošenih linearnih modelih za razliko od normalnega modela ne potrebujemo izpolnjenih predpostavk o normalnosti in homoskedastičnosti. Dovoljene so porazdelitve odvisne spremenljivke iz različnih družin. Regresijski koeficienti so običajno ocenjeni po metodi največjega verjetja. Modeli so sestavljeni iz več komponent. Pri modeliranju s posplošenimi linearnimi modeli moramo najprej izbrati porazdelitveno družino. V tej seminarski nalogi si bomo pogledali enoparametrične verjetnostne porazdelitve, ki pripadajo eksponentni družini, saj bomo le-te uporabili v nadaljevanju.

Komponente posplošenih linearnih modelov so tako:

- Slučajna komponenta: y_i, $i = 1, \ldots, n$, so neodvisne spremenljivke z gostoto iz eksponentne družine:

$$f(y|\theta, \phi) = exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right\},$$

  kjer je $\phi > 0$ disperzijski parameter in $b(.)$, $c(.)$ znani funkciji, θ pa je naravni (kanonični) parameter.

- Sistematična komponenta: $\eta_i = \eta_i(\beta) = x_i^T\beta$ je linearni prediktor, kjer je β vektor neznanih regresijskih parametrov.

- Parametrične link komponente: Link fukncija opisuje kako je povprečje $E[Y_i] = \mu_i$ odvisno od linearnega prediktorja:

$$g(\mu_i) = \eta_i$$

  oz. naravna (kanonična) link funcija, če velja $\theta = \eta$.

- Funkcija variance: opisuje kako je varianca $Var[Y_i]$ odvisna od povprečja $Var[Y_i] = \theta V(\mu)$

Za reševanje problema se uporablja metoda največjega verjetja, kjer zapišemo log-verjetje:

$$\ell(\mu|y) = \sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right)$$

Cenilko $\hat{\mu}$ pridobimo z reševanjem t.i. "score" funkcije:

$$s(\mu) = \frac{\partial}{\partial\mu}\ell(\mu|y) = \frac{\partial}{\partial\theta}\ell(\mu|y)\frac{\partial\theta}{\partial\mu} = \left(\frac{y_1 - \mu_1}{\phi V(\mu_1)}, \ldots, \frac{y_n - \mu_n}{\phi V(\mu_n)}\right)$$

.

Ker velja $\mu = \mu(\beta)$, je score funkcija za parameter β enaka:

$$s(\beta) = \frac{\partial}{\partial\beta}\ell(\beta|y) = \frac{\partial}{\partial\theta}\ell(\mu|y)\frac{\partial\theta}{\partial\mu}\frac{\partial\mu}{\partial\eta}\frac{\partial\eta}{\partial\beta} = \sum_{i=1}^{n}\frac{y_i - \mu_i}{\phi V(\mu_i)}\frac{1}{g'(\mu_i)}x_i$$

.

Pogoj, ki ga dobimo je:

$$\sum_{i=1}^{n}\frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)}\frac{1}{g'(\hat{\mu}_i)}x_i = 0$$

,

ki je neodvisen od disperzijskega parametra in velja $g(\hat{\mu}_i) = x^T\hat{\beta}$.

Splošna metoda za rešitev enačbe je iterativni algoritem Fisherjeva metoda točkovanja (ang. Fisher's Method of Scoring). V metodi je v k-ti iteraciji nova ocena $\beta^{(k+1)}$ pridobljena iz prejšne ocene z:

$$\beta(k+1) = \beta^{(k)} + s(\beta^{(k)})\left[E\left[\frac{\partial s(\beta)}{\partial\beta}\right]\Bigg|_{\beta=\beta^{(k)}}\right]^{-1}.$$

Pokazati se da, da je lahko ta iteracija preoblikovana v:

$$\beta^{(k+1)} = (X^TW^{(k)}X)^{-1}X^TW^{(k)}z^{(k)}$$

,

kjer sta vektor z in matrika uteži W definirana kot:

$$z_i = g(\mu_i) + g'(\mu_i)(y_i - \mu_i)$$

$$w_i = \frac{1}{V(\mu_i)(g'(\mu_i))^2}$$

Tako oceno $\hat{\beta}$ izračunamo iterativno z uteženimi najmanjšimi kvadrati (IWLS) z naslednjimi koraki:

1. določimo začetne vrednosti $\mu_i^{(0)}$
2. izračunamo $z_i^{(k)}$ in uteži $w_i^{(k)}$
3. izračunamo $\beta^{(k+1)}$ z uteženimi najmanjšimi kvadrati
4. ponavljamo koraka 2 in 3, dokler ne dosežemo konvergence