

UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna matematika – 1. stopnja

Mitja Mandić

**Iterativne numerične metode v posplošenih linearnih
modelih**

Delo diplomskega seminarja

Mentor: izred. prof. dr. Jaka Smrekar

Ljubljana, 2021

KAZALO

1. Uvod	4
2. Posplošeni linearni modeli	4
2.1. Sestavni deli posplošenega linearnega modela	4
2.2. Linearna regresija	5
2.3. Poissonova regresija	5
2.4. Logistična regresija	5
2.5. Probit regresija	6
2.6. Ocenjevanje parametrov	6
3. Numerične metode	6
3.1. Newton – Raphsonova metoda	6
3.2. Fisher-scoring algoritem	6
4. Primeri	6
4.1. Ocenjevanje parametrov v logističnem modelu	6
4.2. Ocenjevanje parametrov v probit modelu	6
Slovar strokovnih izrazov	6
Literatura	6

Iterativne numerične metode v posplošenih linearnih modelih

POVZETEK

V povzetku na kratko opiši vsebinske rezultate dela. Sem ne sodi razlaga organizacije dela – v katerem poglavju/razdelku je kaj, pač pa le opis vsebine.

Iterative numerical methods in generalized linear models

ABSTRACT

Math. Subj. Class. (2010): navedi vsaj eno klasifikacijsko oznako – dostopne so na www.ams.org/mathscinet/msc/msc2010.html

Ključne besede: navedi nekaj ključnih pojmov, ki nastopajo v delu

Keywords: angleški prevod ključnih besed

1. UVOD

Posplošeni linearni modeli so ključen del statistične analize. Pomagajo nam bolje razumeti relacije med rezultati meritev in s temi izsledki predvideti trende v prihodnosti. Modeli morajo biti kar se da enostavni, ampak hkrati zagotavljati določeno natančnost. Vsa teorija nam pa v praksi ne koristi, če konkretnih števil ne znamo izračunati.

V dobi ogromne množice podatkov je računska učinkovitost ključni del obdelave. Tu nam pomagajo numerične metode.

V delu bom najprej predstavil posplošene linearne modele in najpriljubljenejše numerične metode, ki se uporabljajo za njihovo obdelavo. Teorijo bom osvetlil tudi s praktičnimi primeri.

2. POSPLOŠENI LINEARNI MODELI

2.1. Sestavni deli posplošenega linearnega modela. Vsak posplošeni linearni model sestavljajo trije deli: *slučajni del* je slučajna spremenljivka Y in njena porazdelitev, *sistematični del* predstavlja relacijo med pojasnjevalnimi spremenljivkami, *povezovalna funkcija* pa transformira Y , da se ta bolje prilega podatkom.

2.1.1. Slučajni del. *Slučajni del* privzame porazdelitev slučajnega vektorja Y , pri čemer privzemamo tudi neodvisnost komponent. Porazdelitev Y privzemamo odvisno od podatkov; mnogokrat je „binarna“, torej ima dve možni vrednosti - „uspeh“ ali „neuspeh“. Splošneje je lahko izid tudi število uspehov v fiksnem številu poskusov. V takih primerih privzamemo binomsko porazdelitev. Y nam lahko meri tudi števne podatke, naprimer koliko zabav je obiskal študent v preteklem mesecu. Seveda pa lahko Y predstavlja tudi zvezne podatke, v tem primeru lahko privzamemo normalno porazdelitev (ali pa kakšno drugo zvezno porazdelitev).

2.1.2. Sistematični del. *Sistematična komponenta* posplošenega linearnega modela poda relacije med pojasnjevalnimi spremenljivkami $x_{i,j}$. Te nastopajo linearno, torej je sistematični del enak

$$\beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \dots + x_{i,p}\beta_p$$

2.1.3. Povezovalna funkcija. Tretji del posplošenega linearnega modela je *povezovalna funkcija*, ta nam poda funkcijo $g(\cdot)$ med slučajno komponento in sistematičnim delom. Če označimo $\mu = E(Y)$, je

$$g(\mu) = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \dots + x_{i,p}\beta_p$$

Najenostavnejša taka funkcija je kar identiteta, torej $g(\mu) = \mu$. Ta nam torej da linearno povezavo med pojasnjevalnimi spremenljivkami in pričakovano vrednostjo naših slučajnih spremenljivki. To je ena od oblik regresije za zvezne podatke.

Mnogokrat pa linearna relacija ni primerna - fiksna sprememba pojasnjevalnih spremenljivk ima lahko večji vpliv, če je pričakovana vrednost bližje 0, kot če je bližje 1. Recimo, da je π verjetnost, da bo oseba kupila nov avto, ko je njen dohodek enak x . Sprememba v dohodku za 10.000€ ima manjši vpliv, če je dohodek 1.000.000€, kot če je 50.000€.

Takrat je smiselno uporabiti kakšno drugo povezovalno funkcijo, ki dopušča tudi nelinearne kombinacije pojasnjevalnih spremenljivk. Naprimer, $g(\mu) = \log(\mu)$ modelira pričakovano vrednost logaritma. Smiselno jo je uporabiti, če pričakovana

vrednost ne more zavzeti negativnih vrednosti. Takemu modelu rečemo *log-linear* model.

Spet druga povezovalna funkcija je $\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, ki nam modelira logaritem deležev - smiselno jo je uporabiti, ko μ ne zavzame vrednosti izven $(0, 1)$, torej ko imamo opravka z verjetnostmi. Takemu modelu rečemo logistični model.

2.2. Linearna regresija. Linearna regresija je najenostavnejši primer posplošnega linearnega modela. Enostavno jo lahko zapišemo kot: $Y = X\beta + \varepsilon$ kjer je Y proučevan slučajni vektor, X je matrika pojasnjevalnih slučajnih spremenljivk, β je vektor koeficientov, ki jih želimo oceniti, ε pa slučajna spremenljivka, ki predstavlja napako - pri računanju, meritvah Privzemimo, da je $E(\varepsilon) = 0$. Iz tega sledi $\mu = E(Y) = X\beta$. Model torej pričakovano vrednost slučajne spremenljivke predstavi kot linearno funkcijo pojasnjevalnih spremenljivk. Parametre β ocenimo z metodo najmanjših kvadratov in ob predpostavki polnega ranga za matriko X dobimo $\hat{\beta} = (X^T X)^{-1} X^T Y$.

2.3. Poissonova regresija.

2.4. Logistična regresija. Logistična regresija se uporablja za določanje deležev oziroma računanje verjetnosti. V pošttev pride, ko imamo odgovore tipa uspeh-neuspeh oziroma govorimo o prisotnosti ali odsotnosti neke lastnosti.

Spomnimo se Binomske porazdelitve $Y_i \sim B(n_i, p_i)$. Ta pravi, da je

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

Pričakovana vrednost in varianca sta odvisni le od p_i , in sta enaki $E(Y_i) = p_i$ in $Var(Y_i) = p_i(1 - p_i)$. Poglejmo si sedaj podrobneje *logit* transformacijo. Če se spomnemo, želimo določiti verjetnost nekega dogodka pri danih podatkih. Ob uporabi identitente transformacije se nam kaj hitro lahko zgodi, da za posamezne verjetnosti dobimo vrednosti izven intervala $[0, 1]$. Ta problem bomo rešili v dveh korakih.

Najprej uvedimo

$$\text{delež}_i = \frac{p_i}{1 - p_i}$$

kjer se premaknemo iz verjetnosti v *delež* – verjetnost dogodka proti verjetnosti, da se ne bo zgodil. Če je p_i enak $\frac{1}{2}$, bo delež enak 1. Vidimo, da so deleži vedno pozitivni in niso omejeni navzgor

V naslednjem koraku pa pogledimo logaritem deležev ali logit verjetnosti

$$\eta_i = \text{logit}(p_i) = \log \frac{p_i}{1 - p_i}$$

s tem pa si odstranimo tudi omejitev navzdol. Opazimo še, da če je $p_i = \frac{1}{2}$, je delež enak 1 in je logaritem 0. Kot funkcija p , je logit strogo naraščajoča, torej imamo inverz. Običajno ga imenujemo *antilogit*, izrazimo ga z:

$$p_i = \text{logit}(\eta_i) = \frac{\exp \eta_i}{1 + \exp \eta_i}$$

Vse skupaj nam da *logistični model*, ki za slučajni del vzame binomsko porazdelitev.

2.4.1. *Ocenjevanje parametrov.* Imamo binomske slučajne spremenljivke in imamo povezovalno funkcijo, $\text{logit} p_i = X\beta$, kjer so β neznani parametri. V naslednjem razdelku si bomo ogledali kako zanje izpeljemo enačbe verjetja, ki jih nato uporabimo v numeričnih algoritmihi.

Kot v vsakem posplošenem linearnem modelu tudi v tem predpostavimo neodvisnost komponent slučajnega vektorja Y zato

$$\begin{aligned} P(Y = \vec{y}) &= \prod_{i=1}^n P(Y_i = y_i) \\ &= \prod_{i=1}^n \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} \end{aligned}$$

Naprej si oglejmo logaritemsko funkcijo verjetja. V nadaljnjem računanju bom izpuščal binomski simbol na začetku - je samo konstanta, ki na končen rezultat nima vpliva. Po prejšnjih oznakah je torej

$$\begin{aligned} \ell(p_i) &= \log \left\{ \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{n_i - y_i} \right\} \\ &= \sum_{i=1}^n \{ y_i \log p_i + (n_i - y_i) \log(1 - p_i) \} \\ (1) \quad &= \sum_{i=1}^n \left\{ n_i \log 1 - p_i + y_i \log \left(\frac{p_i}{1 - p_i} \right) \right\} \end{aligned}$$

Po predpostavki logističnega modela je

$$\text{logit}(p_i) = \log \left(\frac{p_i}{1 - p_i} \right) = \beta_0 + x_{i1}\beta_1 + \dots + x_{ir}\beta_r = x_i^\top \beta$$

2.5. Probit regresija.

2.6. Ocenjevanje parametrov.

3. NUMERIČNE METODE

3.1. Newton – Raphsonova metoda.

3.2. Fisher-scoring algoritem.

4. PRIMERI

4.1. Ocenjevanje parametrov v logističnem modelu.

4.2. Ocenjevanje parametrov v probit modelu.

SLOVAR STROKOVNIH IZRAZOV

LITERATURA

- [1] Alan Agresti. *An introduction to categorical data analysis*. John Wiley & Sons, 2007.
- [2] Erik B. Erhard. Logistic regression and nr. https://statacumen.com/teach/SC1/SC1_11_LogisticRegression.pdf.
- [3] Germán Rodríguez. Lecture notes on generalized linear models. <https://data.princeton.edu/wws509/notes/>, 2007.