Predicting League of Legends match outcome with Bayesian methods

Z. Nagelj

April 24, 2019

1 Uvod

Logistična regresijska se uporablja za analizo povezavo med kategorično odvisno spremenljivko in neodvisnimi spremenljivkami.Polege logistične regresije se za analizo kategoričnih spremenljivk uporablja diskriminantna analiza, ki za razliko od logistične regresije predpostavlja normalno porazdelitev neodvisnih spremenljivk.

2 Logistična regresija

Kot vhodni podatek za logistično regresijo dobimo podatkovni set N točk. Vsaka i-ta točko sestavlja set m-tih neodvisnih spremenljivk in kategorična odvisna spremenljivka Y_i z dvema možnima izidoma.

2.1 Logit in logistična transormacija

Naši kategorični odvisni spremenljivki najprej dodelimo numerične vrednost (0 in 1). Povprečje na vzorcu predstavlja delež ugodnih izidov p, razmerje $p/(1-p)$ pa obeti (odds). Logit transormacijo definiramo kot logaritem obetov (log odds):

$$l = logit(p) = \log \frac{p}{1-p}$$

S pomočjo te transformacije preidemo iz omejene zaloge vrednosti p na intervalu $[0, 1]$, na obete $p/(1-p)$ omejene z zalogov vrednosti $[0, \infty)$ in na koncu na logaritem obetov z zalogo vrednosti $(-\infty, \infty)$. Inverzno transormacijo imenujemo logistična transformacija:

$$p = \text{logistic}(l) = \frac{\exp l}{1 + \exp l}$$

S transformacijo se izognemo problemu omejene zaloge vrednosti odvisne spremenljivke. Potencialno bi lahko izbrali tudi kakšno drugo transormacijo (probit).

2.2 Logistični model

Kategorično odvisno spremenljivko definiramo kot slučajno spremenljivko $Y_i$porazdeljeno po Bernoulliju s pričakovano vrednostjo p_i. Vsak izid je torej določen s svojo neznano verjetnostjo p_i, ki je določena na podlagi neodvisnih spremenljivk.

$$Y_i|x_{1,i},...,x_{m,i} \sim \text{Bernoulli}(p_i)$$
$$E[Y_i|x_{1,i},...,x_{m,i}] = p_i$$
$$P(Y_i = y|x_{1,i},...,x_{m,i}) = p_i^y(1-p_i)^{(1-y)}$$

Ideja je zelo podobna kot pri linearni regressiji, torej verjetnost p_i modeliramo kot linearno kombinacijo neodvisnih spremenljivk. Razlika je v tem, da verjetnosti transformiramo s pomočjo logit funkcije. V modelu nastopa dodaten intercept člen, zato imamo $m+1$ regresijskih koficientov β.

$$\text{logit}(E[Y_i|X_i]) = logit(p_i) = \log\frac{p_i}{1-p_i} = X_i\beta$$

Oziroma:

$$E[Y_i|X_i] = p_i = \text{logit}^{-1}(X_i\beta) = \frac{1}{1+\exp^{-X_i\beta}}$$
$$P[Y_i = y_i|X_i] = p_i^y(1-p_i)^{1-y} = \frac{\exp^{y_iX_i\beta}}{1+\exp^{X_i\beta}}$$

2.3 Določanje vrednosti regresijski koficientov

Regresijski koficienti in verjetnosti p_i so določene optimizacijo, na primer MLE. Za lažjo predstavo si najprej ogledamo MLE za enostaven primer Bernoullija:

2.4 MLE za Bernoulli(p)

Zapišemo enačbo za verjetje in jo logaritmiramo.

$$L = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i} = p^{\sum_{i=1}^{n} Y_i}(1-p)^{n-\sum_{i=1}^{n} Y_i}$$

$$l = \log(L) = \sum_{i=1}^{n} Y_i \log(p) + (n - \sum_{i=1}^{n} Y_i)\log(1-p)$$

Cenilko za $\hat{p}$ določimo s prvim parcialnim odvodom, ki ga enačimo z 0. Za asimptotski interval zaupanja cenilke določimo Fisherjevo informacijsko matriko. Saj populacijske vrednosti p ne poznamo Fisherjevo informacijo določimo z oceno $\hat{p}$.

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
$$I(p) = E[-\frac{\partial^2 l}{\partial p^2}] = \frac{1}{p(1-p)}$$
$$(\hat{var}) = \frac{1}{nI(\hat{p})} = \frac{\hat{p}(1-\hat{p})}{n}$$
$$(\hat{SE}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Za velik n je naša cenilka porazdeljena približno normalno:

$$\hat{p} \sim_{CLI} \text{Normal}(p, \frac{p(1-p)}{n})$$

2.5 MLE za Bernoulli(p_i)

Saj ima pri logisicne regresije vsak izid svojo verjetnost p_i je naša enačba za verjetje naslednja:

$$L = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} = p_i^{\sum_{i=1}^{n} Y_i}(1-p_i)^{n-\sum_{i=1}^{n} Y_i} = (\frac{\exp^{X_i\beta}}{1+\exp^{X_i\beta}})^{\sum_{i=1}^{n} Y_i}(\frac{1}{1+\exp^{X_i\beta}})^{n-\sum_{i=1}^{n} Y_i}$$

$$l = \log(L) = \sum_{i=1}^{n} Y_i \log(\frac{\exp^{X_i\beta}}{1+\exp^{X_i\beta}}) + (n - \sum_{i=1}^{n} Y_i)\log(\frac{1}{1+\exp^{X_i\beta}})$$

Posamezne regresijske koficiente dobimo s parcialnim odvodom logaritma verjetja. Saj rešitev v zaključeni formi ne obstaja uporabimo numeriče metode (npr. Newtnova metoda). Prav tako določimo informacijsko matriko za določanje asimptotske kovariančne matrike in intervalov zaupanja.

3 Statistično testiranje regresijskih koficientov

Za testiranje hipotez neničelnosti regresijski koficientov sta v uporabi Waldov test in test razmerja verjetij.

3.1 Waldov test

Waldov test se uporablja za določanje statistične značilnosti posameznih regresijskih koficientov (podobno kot t-test pri linearni regresiji). Za koficiente pridobljene z MLE je testna statistika naslednja:

$$Z = \frac{\hat{\beta}_i}{\hat{SE}}$$

Ničelelna hipoteza testira $H_0 : \beta_i = 0$. Kot velja za vse cenilke pridobljene po metodi največjega verjetja so te asimptotsko nepristranske (dosledne) in normalno porazdeljene okoli prave vrednosti z varianco $\frac{1}{nI(p)}$.

3.2 Test razmerja verjetij

Pri testu nas zanima logaritem Wilksovega lambda pri dveh različnih modelih, pri čemer je en model gnezden (podmnožica drugega). Primerjali bomo polni model s **k** regresijskimi koficienti in delni model z **m** regresijskimi koficienti, kjer je $\mathbf{m} < \mathbf{k}$. Ničelna hipoteza (delni model) trdi da so testirani regresijski koficienti enaki 0. Alternativna hipoteza (polni model) pa trdi, da so vsi regresijski koficienti različni od 0. Pod ničelno hipotezeo torej testiramo ničelnost k - m β. Naša testrna statistika je:

$$LR = -2\log\Lambda = -2\log\frac{L(H_0)}{L(H_A)} = -2(l(H_0) - l(H_A)) = -2(l(\hat{\beta}^{(0)}) - l(\hat{\beta}))$$

Asimptotsko je ta porazdeljena s χ^2 z k - m stopinjami prostosti. Test razmerja verjetij se od Waldovega testa razlikuje po tem, da je potrebno narediti dva modela pod različnimi hipotezami.

4 Naloga

4.1 Opis

Na vzorcu bolnikov z rakom primerjamo dve vrsti operacije. Zanima nas ali obstaja povezanost s stadijem bolnika in ali ena vrsta operacije povzroča manj zapletov. Cilj naloge je analizirati napako I. reda pri testiranja hipotez na način, da testiramo vsako spremenljivko posebaj.

4.2 Postopek testiranje

Sledili bomo naslednjim korakom:

4.2.1 Stadij

1. Generiranje podatkov pod dano hipotezo
2. Izdelava 4 modelov logistične regresije z informacijo o posameznem stadiju
3. Pridobimo p-vrednosti Waldovega testa vseh 4 modelov z delnimi podatki
4. Izdelava 1 modela logistične regresije z informacijo o vseh stadijh (1 spremenljivka)
5. Pridobimo p-vrednosti testa razmerja verjetij za model z vsemi podatki
6. Analiziramo porazdelitve p-vrednosti in primerjamo deleže zavrnjenih hipotez

4.2.2 Zaplet

1. Generiranje podatkov pod dano hipotezo
2. Izdelava 10 modelov logistične regresije z informacijo o posameznem zapletu
3. Izdelava 1 modela logistične regresije z informacijo o vseh zapletih (10 spremenljivk)
4. Pridobimo p-vrednosti Waldovega testa vseh 11 modelov
5. Analiziramo porazdelitve p-vrednosti in primerjamo deleže zavrnjenih hipotez

4.3 Pričakovani rezultati

Zagotovo, pričakujemo, da bo naveden način testiranja pri stadiju napačen, saj so stadiji med seboj neodvisni, česar v modelih ne zajamemo. Prav tako v model ne vključimo informacije o ostalih stadijih in jih obravnavamo kot enakovredne. Pravilno bi stadij tako, da ga v celoti vključimo v model ter z testom razmerja verjetij testiramo neničelnost regresijskega koficienta.
Pri obravnavi zapletov operacij, ob predpostavki da so si te neodvisni, predvidevam, da tak način testiranja ne bi bil napačen. V realnost pa temu zagotovo ni tako in so posamezni zapleti med seboj povezani, zato bi zagotovo naleteli na enake probleme kot pri testiranju posameznega stadija.

5 Podatki

5.1 Generiranje

Saj je v nalogi določeno, da je vzorec velik 300 pacientov, kjer vsaka polovica prejme en tip operacije, najprej generiramo večji vzorec $k = 10000$ bolnikov iz katerga bomo vzorčili. Pridobiti moramo vrednosti spremenljivke stadij in desetih spremenljivk zaplet.

5.1.1 Stadij

Definiramo populacijske verjenosti za vsak stadij (Tabela 1), ter glede na njih s funkcijo *sample* izžrebamo stadij vsakega bolnika in določimo modelsko matriko. Verjetnosti stadijev se morajo sešteti v 1.

	verjetnost
Stadij 1	0.60
Stadij 2	0.25
Stadij 3	0.10
Stadij 4	0.05

Table 1: Populacijski verjenosti posameznega stadija raka

5.1.2 Zaplet

Definiramo populacijske verjenosti za vsak zaplet (Tabela 2), ter glede na njih s funkcijo *rbinom* izžrebamo ali se je posamezen zaplet zgodil.

	verjetnost
Stadij 1	0.10
Stadij 2	0.12
Stadij 3	0.14
Stadij 4	0.16
Stadij 5	0.18
Stadij 6	0.20
Stadij 7	0.30
Stadij 8	0.40
Stadij 9	0.50
Stadij 10	0.60

Table 2: Populacijski verjenosti posameznega zaplete pri operaciji

5.1.3 Vzorec

Ko imamo določene vrednosti spremenljiv na podlagi definirane hipoteze (oz. regresijskih koficientov) določimo linearne kombinacije spremenljivk ter s pomočjo logit transformacije verjetnost za tip operacije za vsakega izmed k bolnikov. Izbran tip operacij pridobimo iz Bernoullijeve porazdelitve, glede na p_i. Da zadostimo specifikacijam naloge iz vzorca 10000 bolnikov ob vsaki izmed $m = 1000$ ponovitev naključno izžrebamo 150 operacij vsakega tipa.

5.2 Hipoteze

Podatke generiramo glede na 2 hipoteze, ničelno in alternativno. Glede na ničelno hipotezo so vsi regresijski koficienti enaki 0. Vrednosti regresijski koficientov pod alternativno hipotezo so vidni v Tabeli 3.

	stadij	zaplet
beta0	0.00	0.00
beta1	0.10	-3.50
beta2	0.30	3.00
beta3	0.60	-2.50
beta4	-1.20	2.00
beta5		-1.50
beta6		1.00
beta7		-0.80
beta8		0.60
beta9		-0.40
beta10		0.20

Table 3: Vrednosti regresijski koficientov pri HA glede na neodvisno spremenljivko

6 Rezultati

Opazujemo velikost testa (Tabela 4, 6) in moč (Tabela 5, 7). Notacija P(zavrneX) predstavlja delež zavrnjenih ničelnih hipotez glede na tip modela. U v predstavlja model, ko upoštevamao le informacije o posameznem stadiju/zapletu, M pa predstavlja model z vsemi informacijami. Iz simulacije smo določili še verjetnosti, da oba testa zavrneta Hipotezo, ter pogojne verjetnosti glede na kaj prikaže prvi test.

V velikem številu raziskav si na tak način pomagamo pri izbiri neodvisnih spremenljivk. Torej za vsako izmed neodvisnih spremenljivk se izvede regresija in na podlagi tega določi katere spremenljivke bomo vključili v glavni model. Zato nas zanima s kašno verjetnostjo bomo spremenljivko s takšnim načinom pravilno izločili. To nam pove P(zavrneM | zavrneU). Velja seveda tudi obratno.

6.1 Stadij

Pri modeliranju posameznih stadijev, je v resnici naša ničelna hipoteza, da stadij$_{n} nimavplivanatipoperacije, razisk$
blizu 0.

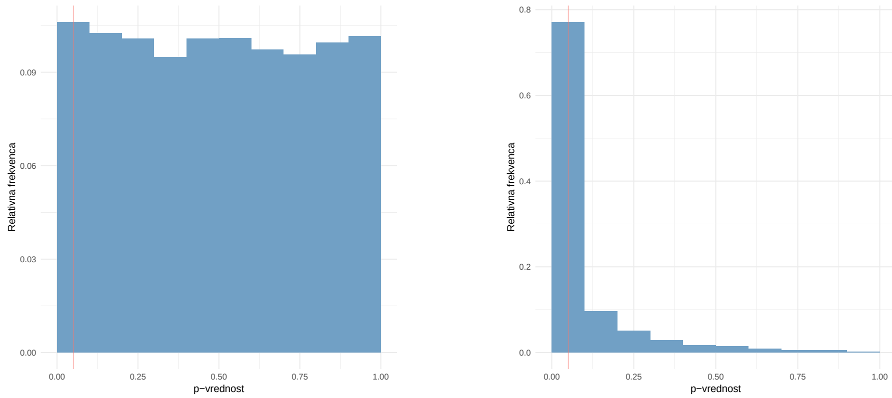
Vidimo, da če gledamo velikosti posameznih testov so vsi približno 0.05. Rezultati nam jasno povejo P(zavrneM | zavrneU), da s takšnim načinom modeliranja, v določenih primerih spremenljivke po krivem izključimo v tudi do 60% primerih.

	S1	S2	S3	S4
P(zavrneU)	0.051	0.053	0.047	0.032
P(zavrneM)	0.054	0.054	0.054	0.054
P(zavrneM | zavrenU)	0.377	0.396	0.389	0.509
P(zavrneU | zavrenM)	0.353	0.390	0.336	0.301

Table 4: Verjetnosti zavrnitev pri H0

	S1	S2	S3	S4
P(zavrneU)	0.071	0.119	0.255	0.603
P(zavrneM)	0.670	0.670	0.670	0.670
P(zavrneM | zavrenU)	0.938	0.931	0.942	0.874
P(zavrneU | zavrenM)	0.100	0.166	0.359	0.787

Table 5: Verjetnosti zavrnitev pri HA



(a) H0 (b) HA

Figure 1: Porazdelitve p-vrednosti pridobljenih z LRT glede na hipotezo

6.2 Zaplet

Pri testiranju z modelom s posameznihm zapletom testiramo ali posamezen zaplet vpliva pri izbiri tipa operacije. Polni model tesitra enako hipotezo, vendar ob upoštevanju drugih zapletov.

V primeru zapletov, ki so neodvisno generirani, vidimo da je težava, ki smo jo izpostavili veliko manj prisotna, oziroma je skladanje z načinom, ko smo upoštevali posamezne zaplete in ko smo upoštevali vse zaplete, veliko večje (85%).

Ko testiramo hipoteze pod alternativno hipotezo opazimo, da je moč testa, ko upoštevamo vse zaplete večja, kot pri modelih s posameznimi zapleti.

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
P(zavrneU)	0.043	0.044	0.053	0.052	0.047	0.050	0.046	0.048	0.051	0.053
P(zavrneM)	0.047	0.049	0.055	0.056	0.050	0.054	0.051	0.051	0.054	0.053
P(zavrneM | zavrenU)	0.866	0.878	0.851	0.851	0.872	0.878	0.858	0.838	0.884	0.856
P(zavrneU | zavrenM)	0.793	0.775	0.817	0.804	0.821	0.815	0.785	0.788	0.824	0.860

Table 6: Verjetnosti zavrnitev pri H0

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
P(zavrneU)	0.936	0.994	0.996	0.979	0.893	0.616	0.551	0.356	0.212	0.116
P(zavrneM)	0.930	0.987	0.999	0.997	0.965	0.786	0.722	0.506	0.289	0.112
P(zavrneM | zavrenU)	0.993	0.992	0.999	0.999	0.991	0.956	0.942	0.881	0.746	0.557
P(zavrneU | zavrenM)	0.999	0.999	0.997	0.980	0.917	0.749	0.718	0.620	0.548	0.576

Table 7: Verjetnosti zavrnitev pri HA

7 Zaključek