

Негативна биномна и Пуасонова регресија

5.час

30. март 2016.

Негативна биномна расподела

Нека је вероватноћа догађаја A („успеха“) у сваком експерименту једнака p и нека се експерименти изводе независно један од другог, под истим условима, све док се догађај A не оствари r пута.

Посматрамо следеће случајне променљиве:

- X – редни број експеримента у коме се r -ти пут остварио догађај A
- Y – број реализација догађаја $\bar{A}$ („неуспеха“) док се није r -ти пут остварио догађај A .

Имамо следећу везу:

$$Y + r = X.$$

Случајна променљива X представља време чекања r -тог успеха.

Негативна биномна расподела

Специјално, када је $r = 1$, представља случајну променљиву са геометријском расподелом.

За случајну променљиву Y се каже да има негативну биномну расподелу са реалним параметрима r и p , $r > 0$ и $p \in [0, 1]$ и користи се ознака $Y \sim \mathbf{NB}(r, p)$.

Дакле, негативна биномна расподела представља број неуспешних реализација неког експеримента, док се није десило r успеха.

Њен закон расподеле је:

$$P\{Y = k\} = \binom{k+r-1}{r-1} p^r (1-p)^{k-r}, \quad k = 0, 1, 2, \dots$$

Негативна биномна расподела

Математичко очекивање и дисперзија случајне променљиве $Y \sim \mathbf{NB}(r, p)$ су:

$$EY = \frac{r(1-p)}{p}, \quad DY = \frac{r(1-p)}{p^2}.$$

Назив негативна биномна расподела се узима због:

$$P\{Y = k\} = \binom{-r}{k} p^r (-q)^k, \quad q = 1 - p$$

и чињенице да је $\binom{-r}{k} (-q)^k$ сабирак у биномном развоју (са негативним експонентом) за $(1 - q)^{-r}$.

Особине негативне биномне расподеле

Нека случајне променљиве Y и Z имају негативне бинмне расподеле са истим параметром, редом $\mathbf{NB}(r_1, p)$ и $\mathbf{NB}(r_2, p)$. Ако су Y и Z независне случајне променљиве, тада случајна променљива $Y + Z$ има негативну биномну расподелу $\mathbf{NB}(r_1 + r_2, p)$.

Негативна биномна расподела припада експоненцијалној фамилији расподела:

$$f(y, p, r) = \exp \left\{ y \ln(1 - p) + r \ln p + \ln \binom{y + r - 1}{r - 1} \right\}.$$

Веза између негативне биномне и Пуасонове расподеле

Пуасонова случајна променљива представља број догађаја који се остваре у задатом временском интервалу, ако се догађаји остварују независно од времена реализовања последњег догађаја. Негативна биномна расподела је уопштен случај Пуасонове расподеле, у ком стопа интензитета није више константна, већ се претпоставља да прати гама расподелу. Ова чињеница указује на то да очекивање није константно током временског периода.

Нека Y има Пуасонову расподелу са параметром Λ . Параметар Λ је случајна променљива која има двопараметарску гама расподелу са параметрима n и β ($n > 0, \beta > 0$). Нека је функција густине гама расподеле означена са $G(\lambda)$ и дата следећом једнакости:

$$G(\lambda) = \frac{\lambda^{n-1} e^{-\frac{\lambda}{\beta}}}{\beta^n \Gamma(n)}, \quad n > 0, \beta > 0.$$

Веза између негативне биномне и Пуасонове расподеле

$$\begin{aligned} P\{Y = k\} &= \int_0^{\infty} P\{Y = k \mid \Lambda = \lambda\} G(\lambda) d\lambda \\ &= \int_0^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} \cdot \frac{\lambda^{n-1} e^{-\frac{\lambda}{\beta}}}{\beta^n \Gamma(n)} d\lambda = \frac{1}{k!} \cdot \frac{1}{\beta^n \Gamma(n)} \int_0^{\infty} e^{-\lambda\left(1+\frac{1}{\beta}\right)} \lambda^{k+n-1} d\lambda \\ &= \frac{\Gamma(k+n)}{k! \Gamma(n)} \cdot \frac{\beta^k}{(1+\beta)^{k+n}} = \frac{(k+n-1)!}{k! (n-1)!} \left(\frac{\beta}{1+\beta}\right)^k \left(\frac{1}{1+\beta}\right)^n \\ &= \binom{k+n-1}{k} \left(\frac{\beta}{1+\beta}\right)^k \left(\frac{1}{1+\beta}\right)^n. \end{aligned}$$

За $p = \frac{1}{1+\beta}$, на основу добијене једнакости, може се уочити да добијена функција расподеле вероватноће случајне променљиве Y је једнака функцији расподеле вероватноће негативне биномне расподеле, па је Y негативна биномна случајна променљива.

Пуасонова регресија

Пуасонова регресија заснива се на Пуасоновој расподели вероватноћа:

$$P\{Y = y\} = \frac{e^{-\mu} \mu^y}{y!},$$

где је параметар μ очекивана вредност случајне променљиве Y .

Модел Пуасонове регресије је облика:

$$\ln \mu = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p,$$

где су $x_1, \dots, x_p$ дати предиктори (обично подељени на групе), а $\beta_0, \beta_1, \dots, \beta_p$ су регресиони коефицијенти које треба оценити.

Пуасонова регресија

Користећи чињеницу да је $\mu = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$, расподелу за Y можемо записати као:

$$p(y_i) = \frac{e^{-e^{x_i \beta}} (e^{x_i \beta})^{y_i}}{y_i!} = \frac{e^{y_i x_i \beta - e^{x_i \beta}}}{y_i!} = e^{y_i x_i \beta - e^{x_i \beta} - \ln(y_i!)},$$

за $i = 1, 2, \dots, n$.

Параметар β оцењујемо користећи метод максималне веродостојности.

Прекомерна дисперзија

Прекомерна дисперзија (overdispersion) у Пуасоновом моделу јавља се када је дисперзија зависне променљиве већа од очекиване вредности.

Прекомерна дисперзија представља проблем приликом моделовања јер може утицати на то да стандардне грешке оцена параметара модела буду мање него што заправо јесу, односно да буду потцењене. На пример, променљива у моделу може бити проглашена значајном када заправо није значајан предиктор. До прекорачења дисперзије долази ако:

- у моделу изоставимо важан предиктор
- предиктори треба да буду трансформисани
- је претпостављена лоша линеарна веза између зависне променљиве и функције предиктора

Прекомерну дисперзију можемо уочити помоћу коефицијената као што су Пирсон хи-квадрат и прилагођени Пирсон хи-квадрат.

Ако је у моделу присутна прекомерна дисперзија, она може смањити моделовањем негативном биномном регресијом уместо Пуасоновом.

Моделовање негативне биномне регресије

Ова форма је заснована на комбинацији Пуасонове и гама расподеле.

У Пуасон-гама моделу, зависна променљива y_i је из Пуасонове расподеле са очекивањем λ_i где се претпоставља да грешка модела има гама расподелу.

Као што и сам назив говори, Пуасон-гама је комбинација две расподеле. Ова комбинација расподела је направљена да објасни чињеницу велике дисперзије која се обично јавља код дискретних или бројачких база. Постала је веома популарна јер има сличну форму као и негативна биномна регресија.

Моделовање негативне биномне регресије

Негативна биномна регресија је врста уопштеног линеарног модела у коме зависна променљива Y броји колико пута се десио неки догађај. Погодна параметризација негативне биномне расподеле је:

$$P\{Y = y\} = \frac{\Gamma(y + \frac{1}{\alpha})}{\Gamma(y + 1)\Gamma(\frac{1}{\alpha})} \left(\frac{1}{1 + \alpha\mu} \right)^{1/\alpha} \left(\frac{\alpha\mu}{1 + \alpha\mu} \right)^y,$$

где је $\mu > 0$ математичко очекивање од Y и $\alpha > 0$ параметар прекомерне дисперзије. Ова параметризација је комбинација Пуасонове и гама расподеле.

Y има негативну биномну расподелу, са параметрима $p = \frac{1}{1+\alpha\mu}$ и $r = \frac{1}{\alpha}$ и представља број неуспеха неког догађаја пре него што се десио $(1/\alpha)$ -ти успех, мада се не захтева да $1/\alpha$ буде цео број.

Очекивање и дисперзија случајне величине Y биће:

$$EY = \mu, \quad DY = \mu + \alpha\mu^2.$$

Моделовање негативне биномне регресије

Модел негативне биномне регресије је облика:

$$\ln \mu = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p,$$

где су $x_1, x_2, \dots, x_p$ дати предиктори, а $\beta_0, \beta_1, \dots, \beta_p$ су регресиони коефицијенти које треба оценити.

Вредности α и β за које се постиже максимум функције $\ln L(\alpha, \beta)$ биће оцене максимале веродостојности које тражимо.

Оцене α и β биће решења нелинеарног система једначина који се решава неком од нумеричких метода.

Моделовање негативне биномне регресије за стопе

Ако Y пребројава догађаје током одређеног временског интервала t , тада посматрана стопа $\frac{Y}{t}$ може бити моделована коришћењем претходног модела са мањим подешавањима. Напоменимо да t може бити посматрано као простор или као део популације, што такође доводи то чињенице да разматрамо стопу $\frac{Y}{t}$.

Како је $E(\frac{Y}{t}) = \frac{\mu}{t}$, претходни модел трансформисаћемо на следећи начин:

$$\ln \frac{\mu}{t} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p,$$

што, такође, може бити записано у облику:

$$\ln \mu = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \ln t,$$

Сада у функцији веродостојности замењујемо μ са $e^{x_i \beta + \ln t}$ и даље поступамо као и у претходном моделу, максимизирајући нову функцију веродостојности како бисмо оценили параметре α и β .

Пирсон хи-квадрат

Корисна статистика која мери квалитет модела је Пирсон хи-квадрат и рачуна се на следећи начин:

$$\sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\sqrt{D(y_i)}}.$$

Ако су очекивање и дисперзија добро одређени, онда је

$$E \left[\sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\sqrt{D(y_i)}} \right] = n.$$

Вредности ближе n (величини узорка) сугеришу боље слагање модела са подацима.

Прилагођени Пирсон хи-квадрат

Прилагођени Пирсон хи-квадрат дефинисан је као Пирсон хи-квадрат коефицијент подељен степеном слободе.

Ако је овај коефицијент већи од 1, значи да је дошло до прекорачења дисперзије.

Мања прекорачења дисперзије су од мањег значаја, међутим, ако је прилагођени Пирсон хи-квадрат коефицијент већи од 1,25 за узорак умереног обима, тада је потребно извршити корекцију модела.

У моделу са великим бројем опсервација је дошло до прекорачења дисперзије ако је овај коефицијент већи од 1,05.

Пуасонов модел је изведен из функције Пуасонове расподеле, док је модел негативне биномне регресије изведен из функције расподеле која је мешавина Пуасонове и гама расподеле.

Пуасонов модел је основни када је реч о моделовању целобројне променљиве. Ипак некад се до бољих резултата долази са моделима негативне биномне регресије.

Као што знамо очекивање и дисперзија Пуасонове расподеле су једнаки, и то је претпоставка која мора бити испуњена приликом моделирања. Такође, вредности које узима категоричка променљива морају бити међусобно независне. Када ове претпоставке нису испуњене готово увек долази до прекомерне дисперзије и тада се углавном користи модел негативне биномне регресије.

Пример

У овом примеру анализираћемо базу која садржи податке 1316 путника са Титаника. Забележени су пол (мушки или женски), године (дете или одрасли), класа (1, 2 или 3) и да ли је путник преживео или не.

Циљ истраживања је да се процени ризик преживљавања. Ако мало боље погледамо базу, видећемо да је преживео мањи број деце из прве класе, него из друге две. Поставља се питање, да ли су деца из прве класе била изложена већем ризику. Одговор је наравно, негативан. На Титанику је било мање деце из прве класе, тако да је самим тим и број преживелих мањи. Због тога, нећемо правити модел који броји преживеле (Y), већ желимо да моделујемо пропорцију ($\frac{Y}{cases}$) преживелих, што представља стопу преживелих.

Најпре, морамо прилагодити нашу базу тако да садржи податке о броју преживелих у зависности којој класи, годишту и полу припадају путници. Дакле, база `titanic` може се поделити у 12 група.

Пример

```
attach(titanic)
summary(titanic)
```

```
cases<-rep(0,12)
for(i in 1:24){
  if (Age[i]=="Child" && Sex[i]=="Female" && Class[i]=="1st") cases[1] <- cases[1]+Freq[i]
  else if (Age[i]=="Child" && Sex[i]=="Female" && Class[i]=="2nd") cases[2] <- cases[2]+Freq[i]
  else if (Age[i]=="Child" && Sex[i]=="Female" && Class[i]=="3rd") cases[3] <- cases[3]+Freq[i]
  else if (Age[i]=="Child" && Sex[i]=="Male" && Class[i]=="1st") cases[4] <- cases[4]+Freq[i]
  else if (Age[i]=="Child" && Sex[i]=="Male" && Class[i]=="2nd") cases[5] <- cases[5]+Freq[i]
  else if (Age[i]=="Child" && Sex[i]=="Male" && Class[i]=="3rd") cases[6] <- cases[6]+Freq[i]
  else if (Age[i]=="Adult" && Sex[i]=="Female" && Class[i]=="1st") cases[7] <- cases[7]+Freq[i]
  else if (Age[i]=="Adult" && Sex[i]=="Female" && Class[i]=="2nd") cases[8] <- cases[8]+Freq[i]
  else if (Age[i]=="Adult" && Sex[i]=="Female" && Class[i]=="3rd") cases[9] <- cases[9]+Freq[i]
  else if (Age[i]=="Adult" && Sex[i]=="Male" && Class[i]=="1st") cases[10] <- cases[10]+Freq[i]
  else if (Age[i]=="Adult" && Sex[i]=="Male" && Class[i]=="2nd") cases[11] <- cases[11]+Freq[i]
  else if (Age[i]=="Adult" && Sex[i]=="Male" && Class[i]=="3rd") cases[12] <- cases[12]+Freq[i]
}

cases<-c(cases[1], cases[2], cases[3], . . . , cases[10], cases[11], cases[12])
cases
```

Пример

```
> A <- c(0,0,0,0,0,0,1,1,1,1,1,1)
> S <- c(0,0,0,1,1,1,0,0,0,1,1,1)
> Cl <- c(1,2,3,1,2,3,1,2,3,1,2,3)

> titanic1 <- data.frame(survive=survive, cases=cases, age=A, sex=S, class=Cl)
> titanic1
```

	survive	cases	age	sex	class
1	1	1	0	0	1
2	13	13	0	0	2
3	14	31	0	0	3
4	5	5	0	1	1
5	11	11	0	1	2
6	13	48	0	1	3
7	140	144	1	0	1
8	80	93	1	0	2
9	76	165	1	0	3
10	57	175	1	1	1
11	14	168	1	1	2
12	75	462	1	1	3

Пример

Сада можемо прећи на прављење модела. Најпре ћемо стопу преживелих моделовати Пуасоновом регресијом.

```
> lncases<-log(cases)
> titanic_model<-glm(survive ~ age + sex + factor(class) + offset(lncases), family
= poisson, data=titanic1)
> summary(titanic_model)
```

Call:

```
glm(formula = survive ~ age + sex + factor(class) + offset(lncases),
    family = poisson, data = titanic1)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.1793	-0.4240	0.1258	1.0760	2.9891

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.48446	0.15961	3.035	0.002403 **
age	-0.48297	0.14562	-3.317	0.000911 ***
sex	-1.16566	0.09502	-12.267	< 2e-16 ***
factor(class)2	-0.37829	0.11758	-3.217	0.001294 **
factor(class)3	-0.76908	0.10704	-7.185	6.74e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 291.608 on 11 degrees of freedom

Residual deviance: 38.304 on 7 degrees of freedom

AIC: 107.06

Number of Fisher Scoring iterations: 4

Пример

Коефицијенти модела су:

```
> exp(coef(titanic_model))
```

(Intercept)	age	sex	factor(class)2	factor(class)3
1.6232966	0.6169489	0.3117178	0.6850337	0.4634390

```
> (pr<-sum(residuals(titanic_model, type='pearson')^2))
```

```
[1] 39.06073
```

```
> (prdisp<-pr/titanic_model$df.residual)
```

```
[1] 5.580104
```

Прилагођена Пирсон хи-квадрат статистика и у овом примеру говори да је дошло до прекорачења дисперзије, тако да ћемо сада направити модел користећи негативну биномну регресију.

Пример

```
> library(MASS)
> titanic_nb2<-glm.nb(survive ~ age + sex + factor(class) + offset(lncases),
  data=titanic1)
> summary(titanic_nb2)
```

Call:

```
glm.nb(formula = survive ~ age + sex + factor(class) + offset(lncases),
  data = titanic1, init.theta = 9.612198415, link = log)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-2.43070	-0.57657	0.00483	0.58594	1.67408

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.6134	0.3141	1.953	0.05085 .
age	-0.6700	0.2472	-2.711	0.00671 **
sex	-0.9801	0.2301	-4.259	2.06e-05 ***
factor(class)2	-0.3746	0.2967	-1.262	0.20680
factor(class)3	-0.9071	0.2906	-3.122	0.00180 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(9.6122) family taken to be 1)

Null deviance: 41.065 on 11 degrees of freedom

Residual deviance: 12.479 on 7 degrees of freedom

AIC: 99.434

Number of Fisher Scoring iterations: 1

Theta: 9.61

Std. Err.: 5.92

2 x log-likelihood: -87.434

Пример

Коефицијенти модела су:

```
> exp(coef(titanic_nb2))
```

(Intercept)	age	sex	factor(class)2	factor(class)3
1.8466548	0.5116907	0.3752548	0.6875551	0.4037074

```
> (pr<-sum(residuals(titanic_nb2, type='pearson')^2))
```

```
[1] 11.07147
```

```
> (prdisp<-pr/titanic_nb2$df.residual)
```

```
[1] 1.581639
```

Видимо да је сада Пирсон хи-квадрат статистика приближно једнака обиму узорка (у овом случају 12), а прилагођени Пирсон хи-квадрат је приближно један.

Како предиктор који означава другу класу путника није значајан, уклонимо га из модела.

Пример

```
> titanic_nb2f <- glm.nb(survive ~ age + sex + I(class=='3') + offset(lncases),
  data = titanic1)
> summary(titanic_nb2f)
```

```
Call:
glm.nb(formula = survive ~ age + sex + I(class == "3") + offset(lncases),
  data = titanic1, init.theta = 7.466423156, link = log)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6581	-0.4643	0.1603	0.5620	1.1775

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.3651	0.2694	1.355	0.175402
age	-0.6143	0.2660	-2.310	0.020906 *
sex	-0.9171	0.2530	-3.625	0.000289 ***
I(class == "3")TRUE	-0.7298	0.2626	-2.779	0.005447 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(7.4664) family taken to be 1)

Null deviance: 33.567 on 11 degrees of freedom

Residual deviance: 11.713 on 8 degrees of freedom

AIC: 98.741

Number of Fisher Scoring iterations: 1

Theta: 7.47

Std. Err.: 4.09

2 x log-likelihood: -88.741

Пример

Коефицијенти модела су:

```
> exp(coef(titanic_nb2f))
```

(Intercept)	age	sex	I(class == '3')
1.4405991	0.5410341	0.3996861	0.4819995

```
> (pr<-sum(residuals(titanic_nb2f, type='pearson')^2))
```

```
[1] 8.686216
```

```
> (prdisp<-pr/titanic_nb2f$df.residual)
```

```
[1] 1.085777
```

Сада су сви предиктори значајни, а AIC је смањена на 98,741. Такође, прилагођени Пирсон хи-квадрат је сада још ближи јединици.

Дакле, стопу преживелих путника на Титанику можемо представити помоћу формуле:

$$\frac{Survive}{cases} = e^{1.44 + 0.54 \cdot Age + 0.40 \cdot Sex + 0.48 \cdot factor(Class=3)}.$$

Задатак

1. За подакте у бази eba1977 (пакет ISwR) направити одговарајући модел који представља утицај старости и града на стопу случајева оболелих од рака плућа.