

Generalized Linear Models: solução dos exercícios para fixação

Introdução aos Modelos Não-Lineares — Cedeplar-UFMG

Professor: Gilvan Guedes; Monitor: Júlia Calazans

May 15, 2018

1 Exercícios Teóricos

Exercício 1 - Se uma variável aleatória, Z , segue uma distribuição binomial e representa o número de filhos homens em 3 gravidezes bem sucedidas, pede-se:

1. represente a variável aleatória
2. defina o seu espaço amostral e sua cardinalidade
3. calcule a probabilidade associada a cada elemento do espaço amostral
4. estime a probabilidade de um casal ter até 2 filhos homens
5. qual a média desta variável?
6. e a sua variância?

Solução

1. Seja Z uma variável aleatória que representa o número de filhos homens em 3 gravidezes bem sucedidas:

$$Z = \{z_1, \dots, z_4\}$$

Essa variável aleatória pode ser descrita como:

$$Z \sim \text{Binomial}(n, \pi), \text{ com } \pi = 0.5$$

A média e a variância de Z são dadas por:

$$\mu(Z) = n\pi$$

$$\sigma^2(Z) = n\pi(1 - \pi)$$

2. O espaço amostral de Z e sua cardinalidade são dados, respectivamente, por:

$$\Omega(Z) = \{0, 1, 2, 3\}$$

$$n\Omega(Z) = 4$$

3. A fórmula geral para o cálculo da probabilidade por ponto de Z é definida como:

$$P(Z = z) = \binom{n}{z} \pi^z (1 - \pi)^{n-z}$$

Com base em Ω , as probabilidades por ponto são:

$$P(Z = 0) = \binom{3}{0} \times 0.5^0 \times (1 - 0.5)^{3-0} = 0.1250$$

$$P(Z = 1) = \binom{3}{1} \times 0.5^1 \times (1 - 0.5)^{3-1} = 0.3750$$

$$P(Z = 2) = \binom{3}{2} \times 0.5^2 \times (1 - 0.5)^{3-2} = 0.3750$$

$$P(Z = 3) = \binom{3}{3} \times 0.5^3 \times (1 - 0.5)^{3-3} = 0.1250$$

4. A probabilidade de um casal ter até dois filhos homens em 3 gravidezes bem sucedidas será:

$$P(Z \leq 2) = \sum_{x=0}^2 \binom{3}{x} \times 0.5^x \times (1 - 0.5)^{3-x} = 0.8750$$

5. A média de Z será:

$$\mu(Z) = 3 \times 0.5 = 1.5$$

6. A variância, por seu turno, será:

$$\sigma^2(Z) = 3 \times 0.5 \times (1 - 0.5) = 0.75$$

Exercício 2 - A partir de um experimento compreendido por 5 lançamentos consecutivos de uma moeda, pede-se:

1. apresente a fórmula genérica da distribuição binomial
2. calcule a probabilidade de cair 3 caras
3. mostre, mapeando as sequências possíveis, que o resultado desse experimento pode ser dado por:

$$P(A = 3) = \sum P(\{(e_1, \dots, e_N)\}) = \binom{5}{3} \cdot 0.5^3 \times 0.5^{5-3}$$

Dica: lembre-se de que a binomial é um produto de Bernoullis.

Solução

1. Denote Y a variável aleatória que representa o número de caras em 5 lançamentos independentes de uma moeda não viciada. Seu espaço amostral será dado por $\Omega(Y) = \{0, 1, 2, 3, 4, 5\}$. Portanto, a probabilidade associada a cada elemento de Ω é dada por:

$$P(Y = y) = \binom{5}{y} \times \pi^y (1 - \pi)^{5-y} \text{ (Probabilidade por ponto)}$$

$$P(Y \leq y) = \sum_{y=0}^5 \binom{5}{y} \times \pi^y (1 - \pi)^{5-y} \text{ (Probabilidade acumulada)}$$

2. A probabilidade de obter 3 caras em 5 jogadas será:

$$P(Y = 3) = \binom{5}{3} \times 0.5^3 (1 - 0.5)^{5-3} = 0.3125$$

3. Primeiro, vejamos quantas possíveis combinações de 3 caras em 5 jogadas podemos obter, utilizando o *coeficiente binomial*:

$$\binom{5}{3} = 10$$

Portanto, há 10 diferentes sequências de Bernoullis possíveis para se obter 3 caras em 5 lançamentos independentes de moedas não viciadas. Chame c cara e k coroa. Essas sequências são dadas por:

$$\begin{aligned} A(Y = 3|n = 5, t = 0) &= \{k, k, c, c, c\} \\ &= \{k, c, k, c, c\} \\ &= \{k, c, c, k, c\} \\ &= \{k, c, c, c, k\} \\ &= \{c, k, k, c, c\} \\ &= \{c, k, c, k, c\} \\ &= \{c, k, c, c, k\} \\ &= \{c, c, k, k, c\} \\ &= \{c, c, k, c, k\} \\ &= \{c, c, c, k, k\} \end{aligned}$$

onde t corresponde a uma das possíveis sequências em 5 lançamentos independentes, definida em $t = 1, \dots, 10$. Veja que cada realização do lançamento da moeda, representada por uma variável aleatória X com distribuição de Bernoulli, tem duas possibilidades, $\{0, 1\}$. Assim:

$$P(X = x) = \pi^x \times (1 - \pi)^{1-x} = \pi = 0.5$$

para o caso de uma Bernoulli com $\Omega = \{0, 1\}$. Para cada uma das sequências acima, a probabilidade de ocorrência é:

$$\begin{aligned} P(\{k, k, c, c, c\}) &= (1 - 0.5) \times (1 - 0.5) \times 0.5 \times 0.5 \times 0.5 \\ &= (1 - 0.5)^2 \times 0.5^3 \\ &= 0.03125 \end{aligned}$$

Isso vale para qualquer uma das 10 sequências:

$$\begin{aligned} P(\{c, c, c, k, k\}) &= 0.5 \times 0.5 \times 0.5 \times (1 - 0.5) \times (1 - 0.5) \\ &= 0.5^3 \times (1 - 0.5)^2 \\ &= 0.03125 \end{aligned}$$

Como há 10 sequências possíveis com a mesma probabilidade de ocorrência, então:

$$\begin{aligned} P(A = 3) &= \sum P(\{(e_1, \dots, e_5)\}) \\ &= \binom{5}{3} \cdot 0.5^3 \times 0.5^{5-3} \\ &= 10 \times 0.03125 = 0.3125 \end{aligned}$$

Exercício 3 - Com base nos seus conhecimentos sobre a família exponencial da forma

$$f_Y(y; \theta, \Phi) = \exp\{a(\phi)[y\theta - b(\theta)] + c(y, \phi)\}$$

pede-se:

1. mostre que a distribuição binomial pertence à família exponencial. Dica: encontre quem são θ , $b(\theta)$ e $c(y, \phi)$. No caso da Binomial, $a(\phi) = \frac{1}{\phi} = 1$.
2. identifique qual é a função de ligação. Dica: lembre-se que a média μ de um vetor y de um GLM é expressa por uma função conhecida (monótona e diferenciável) de η : $\mu_i = g^{-1}(\eta_i)$, $i = 1, \dots, n$. Nesse caso, $g(\cdot)$ é a função de ligação.
3. A partir de então, calcule a média e a variância de Y . Lembre-se que:

$$\begin{aligned} E(Y) &= \mu = b'(\theta) \\ Var(Y) &= a(\phi)b''(\theta) = \frac{1}{\phi}b''(\theta) \end{aligned}$$

Solução

1. Seja Y uma variável aleatória do tipo:

$$Y \sim Binomial(n\pi)$$

Sua distribuição de probabilidade pode ser descrita como:

$$P(Y = y) = \binom{N}{y} \pi^y (1 - \pi)^{N-y}$$

Para mostrar que a distribuição binomial pertence à família exponencial, basta fazer:

$$\begin{aligned} P(Y = y) &= \binom{N}{y} \pi^y (1 - \pi)^N (1 - \pi)^{-y} \\ &= \binom{N}{y} \left(\frac{\pi}{1 - \pi} \right)^y (1 - \pi)^N \end{aligned}$$

Agora, basta aplicar o logaritmo e a função inversa (exponencial) sobre a equação acima, de modo a identificar os elementos θ , $c(y, \phi)$, $b(\theta)$, $a(\phi)$ e $y\theta$:

$$\begin{aligned} \exp \{ \ln [P(Y = y)] \} &= \exp \left\{ \ln \left[\binom{N}{y} \left(\frac{\pi}{1 - \pi} \right)^y (1 - \pi)^N \right] \right\} \\ &= \exp \left\{ \underbrace{\ln \binom{N}{y}}_{c(y, \phi)} + \underbrace{y \ln \left(\frac{\pi}{1 - \pi} \right)}_{y\theta} + \underbrace{N \ln (1 - \pi)}_{-b(\theta)} \right\} \end{aligned}$$

Veja que a fórmula genérica da densidade para membros da família exponencial, dada por:

$$f_Y(y; \theta, \Phi) = \exp \{ a(\phi) [y\theta - b(\theta)] + c(y, \phi) \}$$

corresponde a:

$$\begin{aligned} \exp \ln [P(Y = y)] &= \exp \{ a(\phi) [y\theta - b(\theta)] + c(y, \phi) \} \\ &= \exp \left\{ 1 \left[y \ln \left(\frac{\pi}{1 - \pi} \right) + N \ln (1 - \pi) \right] + \ln \binom{N}{y} \right\} \end{aligned}$$

em que $a(\phi) = 1$.

2. Como $\mu_i = g^{-1}(\eta_i)$, $i = 1, \dots, n$, em que $g(\cdot)$ é a função de ligação, tem-se que:

$$\theta = g(\cdot) = \ln \left(\frac{\pi}{1 - \pi} \right)$$

3. A média é dada por:

$$\begin{aligned} E(Y) &= \mu = b'(\theta) \\ &= \frac{\partial b(\theta)}{\partial \theta} \end{aligned}$$

Podemos derivar $E(Y)$ diretamente, como se segue. Primeiro, vamos isolar π em θ , aplicando sua função inversa:

$$\begin{aligned}\theta &= \ln\left(\frac{\pi}{1-\pi}\right) \\ \exp\theta &= \frac{\pi}{1-\pi} \\ \pi &= (1-\pi)\exp\theta \\ &= \exp\theta - \pi\exp\theta \\ \exp\theta &= \pi + \pi\exp\theta \\ \pi(1+\exp\theta) &= \exp\theta \\ \pi &= \frac{\exp\theta}{1+\exp\theta}\end{aligned}$$

Partamos de $b(\theta)$:

$$b(\theta) = -n \ln(1 - \pi)$$

Como

$$\exp\theta = \frac{\pi}{1-\pi}$$

é fácil mostrar que:

$$1 - \pi = \frac{1}{1 + \exp\theta}$$

A prova está abaixo:

$$\begin{aligned}\exp\theta &= \frac{\pi}{1-\pi} \\ \exp\theta(1-\pi) &= \pi \\ 1-\pi &= \frac{\pi}{\exp\theta}\end{aligned}$$

Como $\pi = \frac{\exp\theta}{1+\exp\theta}$, então:

$$\begin{aligned}1 - \pi &= \frac{\frac{\exp\theta}{1+\exp\theta}}{\exp\theta} \\ &= \frac{1}{1 + \exp\theta}\end{aligned}$$

Assim:

$$\begin{aligned}b(\theta) &= -n \ln(1 - \pi) \\ &= -n \ln\left(\frac{1}{1 + \exp\theta}\right) \\ &= \underbrace{-n \ln 1}_{n \times 0} + n \ln(1 + \exp\theta) \\ &= n \ln(1 + \exp\theta)\end{aligned}$$

A média de Y agora pode ser obtida facilmente:

$$\begin{aligned}
 E(Y) &= \frac{\partial b(\theta)}{\partial \theta} \\
 &= \frac{\partial n \ln(1 + \exp \theta)}{\partial \theta} \\
 &= \frac{n}{1 + \exp \theta} \exp \theta \\
 &= n \left(\frac{\exp \theta}{1 + \exp \theta} \right) \\
 &= n\pi
 \end{aligned}$$

lembre-se que:

$$\pi = \frac{\exp \theta}{1 + \exp \theta}$$

A variância, por seu turno, é dada por:

$$\sigma^2(Y) = a(\phi)b''(\theta) = \frac{1}{\phi}b''(\theta)$$

Como

$$a(\phi) = \frac{1}{\phi} = 1$$

então:

$$\begin{aligned}
 \sigma^2(Y) &= b''(\theta) \\
 &= \frac{\partial^2 b(\theta)}{\partial^2 \theta} \\
 &= \frac{\partial^2}{\partial^2 \theta} - n \ln(1 + \exp \theta) \\
 &= \frac{\partial}{\partial \theta} n\pi \\
 &= \frac{\partial}{\partial \theta} \frac{n \exp \theta}{1 + \exp \theta} \\
 &= \frac{n \exp \theta (1 + \exp \theta) - n \exp \theta \exp \theta}{(1 + \exp \theta)^2} \\
 &= \frac{n \exp \theta + n \exp \theta^2 - n \exp \theta^2}{(1 + \exp \theta)^2} \\
 &= \frac{n \exp \theta}{(1 + \exp \theta)^2} \\
 &= n \underbrace{\frac{\exp \theta}{1 + \exp \theta}}_{\pi} \underbrace{\frac{1}{1 + \exp \theta}}_{1 - \pi} \\
 &= n\pi(1 - \pi)
 \end{aligned}$$

2 Exercícios de Computador

Todos os exercícios abaixo devem ser feito no *software* R. O `script` deve ser entregue junto com a resolução dos exercícios.

Exercício 1 - Calcule a probabilidade de observar 30 migrantes numa amostra de 100 pessoas, com probabilidade de migrar $p = 0.6$.

Solução no R

```
pbinom(30,100,0.6)-pbinom(30,100,0.6)
```

Caso queira mostrar a notação não científica no R, basta digitar:

```
format(pbinom(30,100,0.6)-pbinom(30,100,0.6),  
scientific=FALSE)
```

O valor obtido deverá ser:

$$\begin{aligned} P(X = 30|N = 100, \pi = 0.6) &= \binom{100}{30} \times 0.6^{30} \times 0.4^{100-30} \\ &= 0.000000000905056 \end{aligned}$$

Exercício 2 - Plote num gráfico a distribuição de probabilidade binomial, com:

$$n = 0, 1, 2, \dots, 10; N = 10; p = 0.1, 0.5, 0.9$$

Solução no R

Populando a matriz da densidade binomial em 10 tentativas e com probabilidade de sucesso variável:

```
binom=matrix(nrow=11,ncol=3,NA)  
w=seq(0.1,0.9,by=0.4)  
for (j in 0:10) {  
  for (k in 1:3) {  
    binom[j+1,k]=dbinom(j,10,w[k])  
  }  
}
```

Calculando os coeficientes de assimetria da distribuição binomial para diferentes probabilidades de sucesso:

```
a=seq(from=0.1,to=0.9,by=0.4)  
b=NA  
for (i in 1:3){  
  b[i]=as.character(  

```

```
round(eval((1-2*a[i])/(sqrt(10*a[i]*(1-a[i])))),
2))
}
```

Salvando as probabilidades de sucesso como texto:

```
atext=NA
for (i in 1:3){
atext[i]=as.character(eval(a[i]))
}
```

Plotando os gráficos:

```
par(mfrow=c(1,3))
for (k in 1:3){
plot(binom[,k],
main=paste('Assimetria :',b[k]),
ylab=paste("Densidade Binomial com p=",atext[k]),
xlab="Trial",
ylab="Densidade Binomial",
col="blue",
pch=16,
cex=1.5)
lines(binom[,k],
col='red',
lwd=1.5)
}
```

Exercício 3 - Com base no exercício anterior, pode-se:

1. calcule o coeficiente de assimetria para cada uma das distribuições (ou seja, para $p = 0.1$, $p = 0.5$ e $p = 0.9$). Para a binomial, $S = \frac{1-2p}{\sqrt{np(1-p)}}$.
2. calcule a variância da distribuição com $p = 0.1$ se a amostra aumentasse para 100 casos e $n = 0, 1, \dots, 100$
3. calcule a média para a distribuição anterior com 100 casos e compare com a média quando a distribuição era baseada em 10 casos. A que conclusão você chegou?

Solução no R

1. Para $\pi = 0.1, 0.5, 0.9$, S será:

```
a=seq(from=0.1,to=0.9,by=0.4)
b=NA
for (i in 1:3){
b[i]=as.character(
```

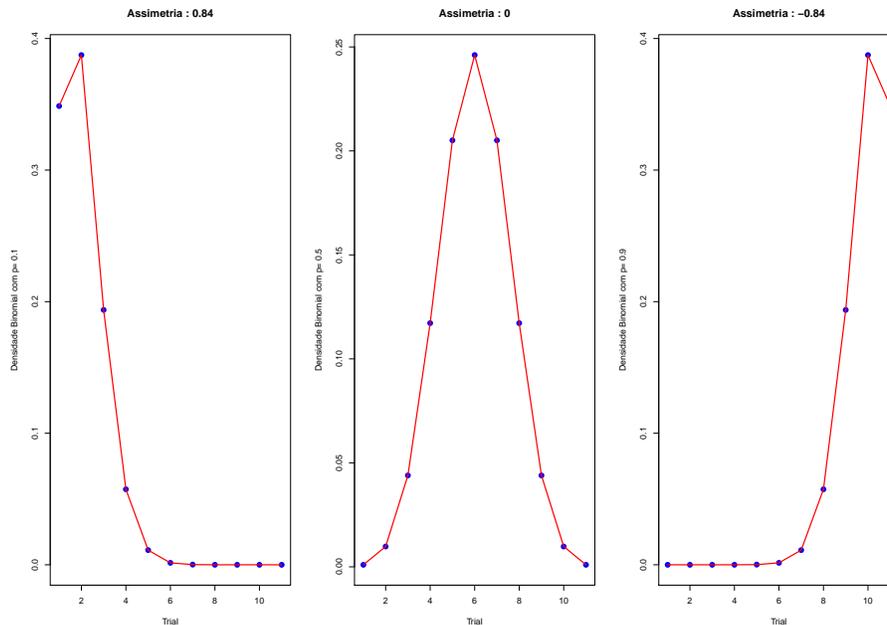


Figure 1: Distribuição de Probabilidade Binomial sob diferentes probabilidades de sucesso

```
round(eval((1-2*a[i])/(sqrt(10*a[i]*(1-a[i])))),
2))
}
```

$S = \{0.84, 0.00, -0.84\}$ para $\pi = \{0.1, 0.5, 0.9\}$, respectivamente.

2. A variância será:

$$\begin{aligned}\sigma^2(X) &= n\pi(1 - \pi) \\ &= 100 \times 0.1 \times (1 - 0.1) \\ &= 9\end{aligned}$$

No R, o cálculo pode ser feito como:

```
100*0.1*(1-0.1)
```

Se quiser simular uma variável que seja uma distribuição binomial com $\pi = 0.1$ com tamanho $N = 100$, a variância tornar-se-á tão mais próxima de 9 quanto maior for o número de replicações do experimento. Veja:

```
set.seed(13052018)
N=c(10,100,1000,10000,100000)
```

```

binom=matrix(nrow=100000,ncol=5,NA)
var=NA
for (i in 1:5){
binom[1:N[i],i]=rbinom(N[i],100,0.1)
var[i]=var(na.omit(binom[,i]))
}
round(var,1)

```

Veja que a variância para $\pi = 0.1$ da distribuição anterior, com $n = 10$, era:

$$\begin{aligned}
\sigma^2(X) &= n\pi(1 - \pi) \\
&= 10 \times 0.1 \times (1 - 0.1) \\
&= 0.9
\end{aligned}$$

3. A média de X quanto $N = 100$ será:

$$\begin{aligned}
\mu(X) &= n\pi \\
&= 100 \times 0.1 \\
&= 10
\end{aligned}$$

A média de X quanto $N = 10$ era:

$$\begin{aligned}
\mu(X) &= n\pi \\
&= 10 \times 0.1 \\
&= 1
\end{aligned}$$

Portanto, o aumento do tamanho da população muda apenas o parâmetro de escala.