Estimação de Máxima Verossimilhança em Regressão Logística

Lucas de Oliveira Ferreira de Sales

Universidade Federal do Rio Grande do Norte-UFRN
Centro de Ciências Exatas e da Terra-CCET
Departamento de Estatística-DEST
Programa de Educação Tutorial- PET

04 de abril de 2016

Sumário

Introdução

A epidemiologia é uma área da saúde, na qual, propõe-se a estudar os fatores condicionantes e determinantes para a proliferação e combate de doenças.

O método de máxima verossimilhança e a regressão logística é amplamente utilizado em estudos epidemiologicos, pois é um método simples e de extrema importância para a obtenção (através de bancos de dados) de estimativas dos parâmetros de interesse. Sejam eles taxas de mortalidade devido a alguma epidemia, ou o quanto alguns fatores afetam ou não a causa de alguma doença.

Função de verossimilhança

Segundo Bolfarine e Sandoval (2012), sejam $X_1, \ldots, X_n$ uma amostra aleatória (a.a) de tamanho n da variável aleatória X com função de densidade (ou probabilidade) $f(x|\theta)$ com $\theta \in \Theta$, onde Θ é o espaço paramétrico.
A função de verossimilhança de θ é dada por:

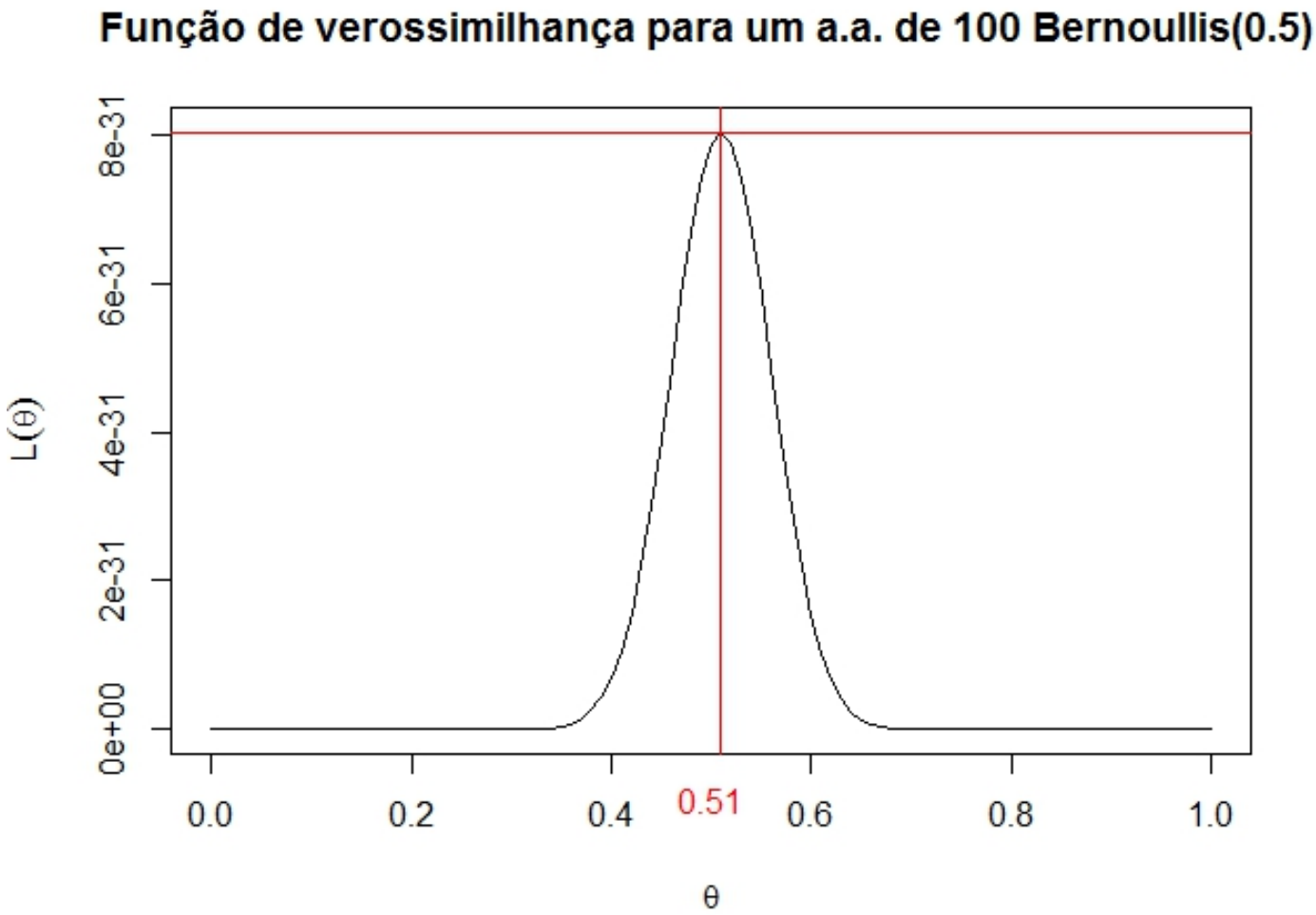
$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta).$$

Estimador e Estimativa de máxima verossimilhança

Segundo Bolfarine e Sandoval (2012), a estimativa de máxima verossimilhança de θ é o valor $\hat{\theta} \in \Theta$ que maximiza a função de verossimilhança $L(\theta|\mathbf{x})$.

Já o EMV é a estatística, na qual, com base na amostra, nos fornece esta estimativa de máxima verossimilhança.

Estimador e Estimativa de máxima verossimilhança



Obtenção do EMV

O EMV pode ser obtido seguindo os passos abaixo:

I. Encontrar a função de verossimilhança

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)$$

Obtenção do EMV

O EMV pode ser obtido seguindo os passos abaixo:

I. Encontrar a função de verossimilhança

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)$$

II. Aplicar o *ln* em $L(\theta|\mathbf{x})$ onde:

$$l(\theta|\mathbf{x}) = ln(L(\theta|\mathbf{x}))$$

Obtenção do EMV

O EMV pode ser obtido seguindo os passos abaixo:

I. Encontrar a função de verossimilhança

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)$$

II. Aplicar o *ln* em $L(\theta|\mathbf{x})$ onde:

$$l(\theta|\mathbf{x}) = ln(L(\theta|\mathbf{x}))$$

III. Encontrar a função escore e igualar a zero:

$$U(\theta) = \frac{\partial l(\theta|\mathbf{x})}{\partial \theta} = 0$$

IV. Verificar se o ponto achado é ponto de máximo:

$$\frac{\partial U(\theta)}{\partial \theta} < 0$$

Propriedades do EMV

O EMV é muito utilizado por possuir uma facilidade computacional relativamente simples e por ter algoritmos já estabelecido para estimar alguns parâmetros por métodos numericos (ex.: Algoritmo de Newton-Raphson). Outro fator que proporciona o grande uso deste estimador são suas propriedades:

- Invariância: Se $\hat{\theta}$ for estimador para θ, então $g(\hat{\theta})$ é um EMV para $g(\theta)$.
- Normalidade assintótica do EMV: Para um tamanho de amostra suficientemente grande temos que:

$$\hat{\theta} \dot{\sim} Normal(\theta, \mathrm{II}_F^{-1}(\theta))$$

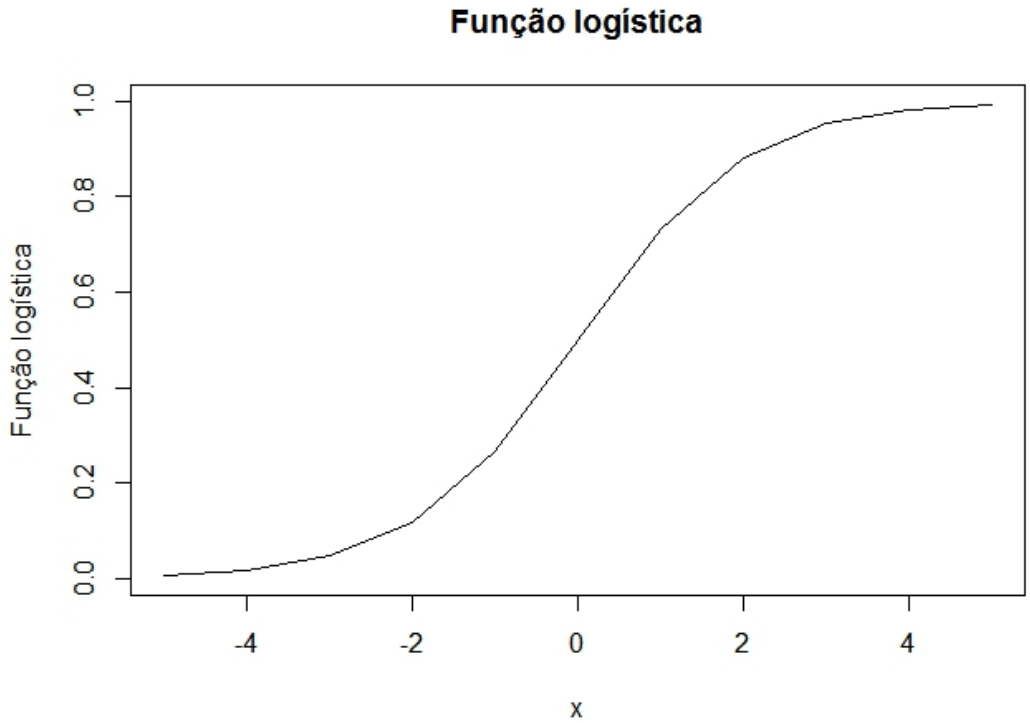
História

Uma explicação para o termo regressão logística é porque o método está baseado na função logística. Que assim como uma probabilidade também está entre 0 e 1;

$$P(x) = \frac{e^x}{1 + e^x} \tag{1}$$

História

A função ou curva logística é uma curva sigmóide que foi batizada por Pierre Fraçois Verhults, que a utilizou para o estudo de taxas do crescimento populacional (Verhulst, PierreFrançois (1845). "Recherches mathématiques sur la loi d'accroissement de la population. Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles 18: 1–42.).



Distribuição logística

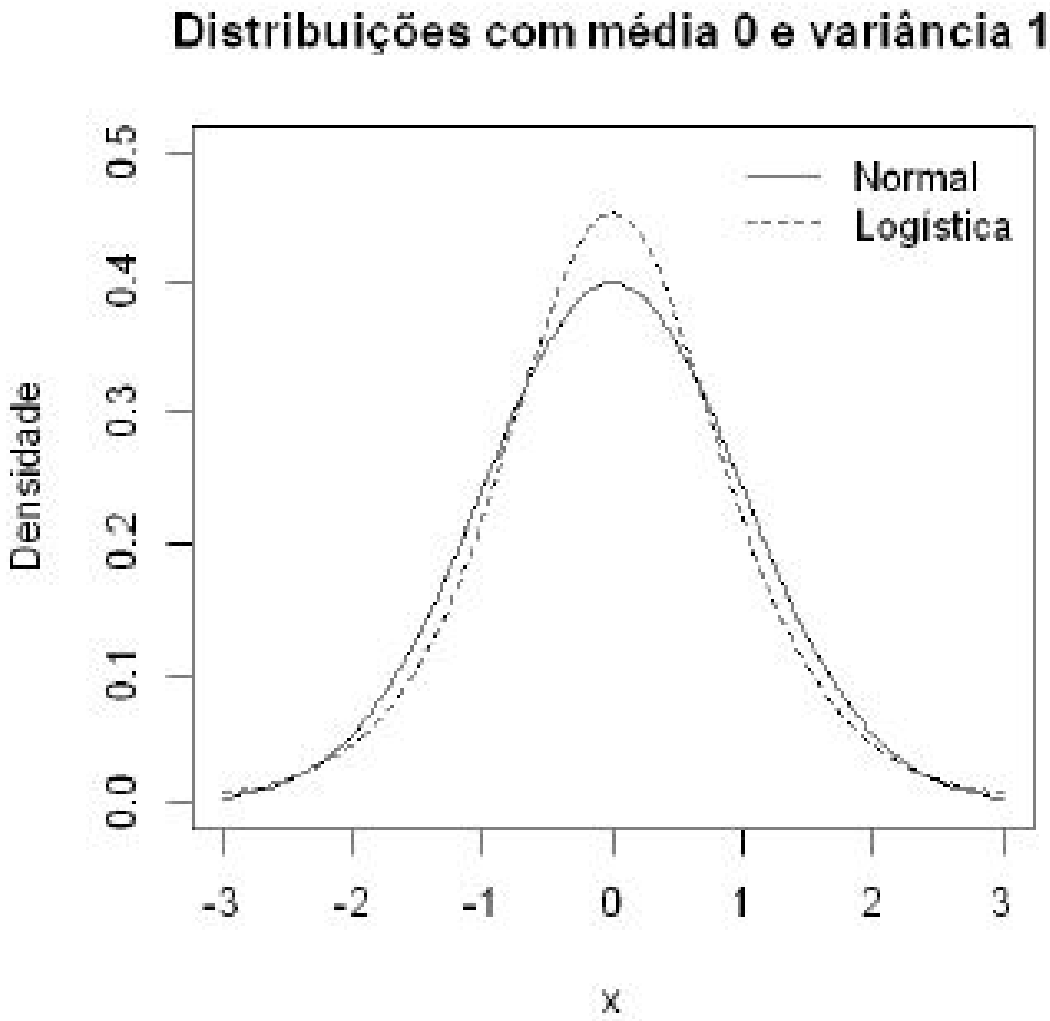
Seja L uma variável aleatória com distribuição logística, então :

$$f_L(l) = \frac{e^{\frac{l}{s}}}{s(1 + e^{\frac{l}{s}})};$$

$$F_L(l) = \frac{e^{l}}{1 + e^{l}}$$

Notemos que $F_L(l)$ é igual a (1). Logo, se $L \sim$ logística$(0, 1)$ sua função de distribuição acumulada é a função logística.

Distribuição Logística



Regressão

Em uma regressão desejamos obter uma determinada informação (variável resposta), a partir de informações que já conhecemos (variáveis regressoras) e que tenham alguma relação coerente com a informação que você deseja obter. Em outros casos podemos análisar a influência das variáveis regressoras na sua resposta e o impacto nela.

Regressão Logística

Quando nossa variável resposta é binária, ou seja, assume apenas 0 ou 1, podemos falar que $Y_i \sim bernoulli(p_i)$ e consideramos o seguinte modelo de regressão:

$$Y_i = E(Y_i) + \epsilon_i$$

Regressão Logística

Quando nossa variável resposta é binária, ou seja, assume apenas 0 ou 1, podemos falar que $Y_i \sim bernoulli(p_i)$ e consideramos o seguinte modelo de regressão:

$$Y_i = E(Y_i) + \epsilon_i$$

Porém como nossa variável Y só assume zero ou um, o erro aleatório (ϵ_i) não tem distribuição normal, mas assim como no modelo linear simples a $E(\epsilon_i) = 0$. Uma suposição que se encaixa melhor para a distribuição dos $\epsilon_i's$ é assumir que os $\epsilon_i's$ tenham distribuição Logistica.

Regressão Logística

Entretanto, a distribuição dos $\epsilon_i's$ acaba dependendo da distribuição Bernoulli dos respectivos $Y_i's$. Assim, é preferível apresentar o modelo de regressão da seguinte forma da seguinte forma:

$$Y_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} + \epsilon_i's,$$

ou seja, $E(Y_i) = p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$.

Relação do EMV com a regressão logística

Normalmente ao fazermos à abordagem citada, desejamos estimar os nossos $p_i's$, os quais, nos dão as probabilidades de sucesso da nossa variável Y, condicionada a determinada característica da variável X.

Relação do EMV com a regressão logística

Normalmente ao fazermos à abordagem citada, desejamos estimar os nossos $p_i's$, os quais, nos dão as probabilidades de sucesso da nossa variável Y, condicionada a determinada característica da variável X.
Porém nossos $p_i's$ são funções de $\beta_0 + \beta_1$, os quais, não conhecemos. Entretanto uma forma muito prática e coerente de se estimar esses parâmetros é utilizando o método da máxima verossimilhança, pois os valores dos parâmetros que máximizam a função de verossimilhança se aproximam muito bem dos valores reais dos parâmetros no nosso caso se aproxima muito bem dos valores reais dos $p_i's$

Relação do EMV com a regressão logística

A seguir será mostrado como a função de verossimilhança pode virar função de β_0 e β_1.

$$f_Y(y) = p_i^y(1-p_i)^{1-y}$$

Relação do EMV com a regressão logística

A seguir será mostrado como a função de verossimilhança pode virar função de β_0 e β_1.

$$f_Y(y) = p_i^y(1-p_i)^{1-y}$$

$$\Rightarrow L(p_i|\mathbf{y}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

Relação do EMV com a regressão logística

A seguir será mostrado como a função de verossimilhança pode virar função de β_0 e β_1.

$$f_Y(y) = p_i^y (1 - p_i)^{1-y}$$

$$\Rightarrow L(p_i|\mathbf{y}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i).$$

Agora aplicaremos o ln em $L(p_i|\mathbf{y})$:

$$l(p_i|\mathbf{y}) = \sum_{i=1}^{n} y_i ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n}(1-p_i)$$

Agora aplicaremos o ln em $L(p_i|\mathbf{y})$:

$$l(p_i|\mathbf{y}) = \sum_{i=1}^{n} y_i ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n}(1-p_i)$$

Devemos lembrar que :

$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \text{ e } (1-p_i) = \frac{1}{1+e^{\beta_0+\beta_1 x_i}}$$

Então:

$$\left(\frac{p_i}{1-p_i}\right) = e^{\beta_0+\beta_1 x_i}$$

$$\Rightarrow ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

Sendo assim, podemos escrever $l(p_i|\mathbf{y})$ como sendo:

$$l(p_i|\mathbf{y}) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} ln(1 + e^{\beta_0+\beta_1 x_i})$$

Podemos observar agora que, $l(p_i|\mathbf{y})=l(\beta_0, \beta_1|\mathbf{y})$. Portanto, pelos motivos já citados podemos utilizar o método da máxima verossimilhança para encontrar β_0 e β_1 tal que nos de uma boa estimativa dos valores verdadeiros das nossas probabilidades de sucesso ($p_i's$).

Exemplo

SITUAÇÃO:
Realizou-se um treinamento com 30 funcionários de um determinado setor de uma empresa. O objetivo do treinamento foi determinar o menor número de horas de treinamento necessários para ocorrência de um número aceitável de erros na montagem.

Exemplo

As tabelas a seguir, contém os dados dos 30 funcionários submetidos ao treinamento.

Funcionário	Horas de treinamento	Peças defeituosas	Peças montadas
1	30.00	2.00	200.00
2	30.00	2.00	200.00
3	30.00	2.00	200.00
4	29.00	2.00	200.00
5	28.00	3.00	200.00
6	27.00	4.00	200.00
7	26.00	5.00	200.00
8	26.00	5.00	200.00
9	25.00	6.00	200.00
10	24.00	6.00	200.00

Exemplo

Funcionário	Horas de treinamento	Peças defeituosas	Peças montadas
11	23.00	8.00	200.00
12	20.00	8.00	200.00
13	20.00	8.00	200.00
14	20.00	8.00	200.00
15	17.00	9.00	200.00
16	17.00	9.00	200.00
17	17.00	10.00	200.00
18	16.00	10.00	200.00
19	15.00	11.00	200.00
20	13.00	11.00	200.00

Exemplo

Funcionário	Horas de treinamento	Peças defeituosas	Peças montadas
21	12.00	12.00	200.00
22	11.00	12.00	200.00
23	11.00	12.00	200.00
24	11.00	13.00	200.00
25	10.00	13.00	200.00
26	10.00	13.00	200.00
27	9.00	13.00	200.00
28	8.00	13.00	200.00
29	8.00	14.00	200.00
30	5.00	14.00	200.00

Exemplo

Para determinar qual a probabilidade de erro, na qual é explicada pelas horas de treinamento, foi realizada uma regressão lógistica conjuntamente com o método da máxima verossimilhança. Onde temos que Horas de treinamento (X) é nossa variável explicativa e a quantidade de peças defeituosas (Y) nossa resposta. E $Y \sim binomial(200, p_i)$.
*OBS: O p_i representa a probabilidade de montar uma peça errada, levando em conta o tempo de treinamento do funcionário.

Exemplo

Após aplicar o método da máxima verossimilhança para encotrar as estimativas de β_0 e β_1, nos deparamos com está situação :

$$U(\beta_0, \beta_1) = \begin{cases} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} m_i \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = 0 \\ \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} m_i x_i \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = 0 \end{cases}$$

Porém estas equações são não-lineares nos parâmetros e para resolvê-las é preciso recorrer a métodos numéricos interativos. Isto mostra que mesmo algébricamente não sendo possível encontrar a emv, através de métodos simples e que não requerem um grande esforço computacional é possível estimar os parâmetros desejados.

Exemplo

O método utilizado para a resolução deste problema foi o método interativo de Newton-Raphson, o qual é um método interativo que apartir de um chute inicial tem o objetivo de estimar as raízes de uma função, ou como neste caso um sistema.
A formula utilizada por este algoritmo para está situação foi :

$$\beta_n = \beta_0 - U'(\beta_0)^{-1} U(\beta_0)$$

Onde β_0 é uma matriz 2_x1 com os chutes iniciais, $U'(\beta_0)$ é a derivada da função escore aplicada no ponto β_0 (que nós da uma matriz 2_x2) e $U(\beta_0)$ é a função escore aplicada no ponto β_0, o que também nos da uma matriz 2_x1.

Exemplo

O chute inicial dado foi:

$$\beta_0 = 0 \text{ e } \beta_1 = 0$$

E após 13 interações obtivemos os seguintes valores estimados:

$$\beta_0 = -2.0065 \text{ e } \beta_1 = -0.663$$

Após os parâmetros estimados, podemos através da regresão logística obter os valores das probabilidade de montar um equipamento errado. Vale relembrar que:

$$p_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

A tabela abaixo, nos da as probabilidades de montar uma peça errada, levando em conta as horas de treinamento:

	Prob. de Erro		Prob. de Erro		Prob. de Erro
1	0.0181	11	0.0284	21	0.0572
2	0.0181	12	0.0345	22	0.0609
3	0.0181	13	0.0345	23	0.0609
4	0.0193	14	0.0345	24	0.0609
5	0.0206	15	0.0417	25	0.0648
6	0.0219	16	0.0417	26	0.0648
7	0.0234	17	0.0417	27	0.0689
8	0.0234	18	0.0445	28	0.0733
9	0.0250	19	0.0474	29	0.0733
10	0.0266	20	0.0537	30	0.0880

Aplicação Epidemiológica

Em muitos casos de diarreia em recém nascidos, a causa pode partir do próprio leite materno. Dependendo da quantidade de antibióticos naturais contidos no leite da mãe o bebê pode ter, ou não, uma resistência maior a uma bactéria chamada de Vibrio cholerae. Onde está é responsável pela colêra, e um de seus principais sintomas é a diarreia.
O objetivo desta aplicação é, com base em uma amostra de 30 bebês diagnósticados com a bactéria, determinar qual a probabilidade de bebês terem ou não o forte sintoma da diarreia, levando em conta a quantidade de antibióticos ingeridos no leite.

Aplicação Epidemiológica

A tabela abaixo relaciona os casos dos bebês que tem ou não diarreia com a quantidade de antibióticos no leite, na qual foi dividida em duas categórias baixo nível de antibióticos e alto nível de antibióticos.

Nível de Antibiótico	Com diarreia(y=1)	Sem diarreia(y=0)	Total
Baixo (x=1)	12	2	14
Alto (x=0)	7	9	16
Total	19	11	30

Aplicação Epidemiológica

Para estimarmos os parâmetros β_0 e β_1 iremos utilizar o software R (R Core Team,2015).

```
####Entrando com as informações presentes na tabela###
y1= rep(c(0,1),c(2,12))
y2=rep(c(0,1),c(9,7))
y= c(y1,y2)
x=rep(c(1,0),c(14,16))
##### Criando uma tabela para verificar os dados####
dados=data.frame("R"=y,"X"=x)
#####Utilizando a função glm #######
ajuste=glm(R~X,dados,family= binomial)
summary(ajuste)
ajuste$coefficients
(Intercept)           X
 -0.2513144   2.0430739
```

Aplicação Epidemiológica

Com auxílio do software, obtivemos as seguintes estimativas:

$$\beta_0 = -0.2513 \text{ e } \beta_1 = 2.0431$$

Apartir dai podemos obter os seguintes valores para p_i:

$$p_0 = 0.4376 \textbf{ e } p_1 = 0.8571$$

Onde p_0 representa a probabilidade do bebê desenvolver o forte sintoma de diarreia, ingerindo o leite materno com uma quantidade alta de antibióticos, e p_1 representa a probabilidade de desenvolver o sintoma, ingerindo o leite materno com uma baixa quantidade de antibióticos.

Conclusão

Podemos observar, que através do método da máxima verossimilhança e utilizando recursos de regressão foi possível solucionar um problema prático. Para a indústria e estudos epidemiológicos, essa prática é muito comum e usual, pois utilizando o conhecimento estátistico agregado a outras áreas acarreta em uma série de soluções para problemas que atingem essas áreas afins.

Referências

- *BOLFARINE, Heleno; SANDOVAL, Mônica Carneiro. Introdução à inferência estatística. SBM, 2001.*
- *COLE, Stephen R.; CHU, Haitao; GREENLAND, Sander. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. American journal of epidemiology, p. kwt245, 2013.*
- *PINHO, André Luis Santos de. Modelos de regressão. 11 abril 2016, 02 maio 2016. Notas de Aula.*
- *R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.*