Universidade Federal do Paraná
Departamento de Estatística

Regressão para Dados Binários - Estudo de Dengue

CE225 - Modelos Lineares Generalizados

Francielle Przibiciem de Mattos GRR20124686
Guilherme Henrique Stadler de Mello GRR20124678
Simone Matsubara GRR20124663

Curitiba, 24 de Outubro de 2017

Sumário

1 Introdução

Transmitida pelo mosquito Aedes Aegypti, a dengue é uma doença viral que se espalha rapidamente no mundo. Nos últimos 50 anos, a incidência aumentou 30 vezes, com ampliação da expansão geográfica para novos países e, na presente década, para pequenas cidades e áreas rurais. É estimado que 50 milhões de infecções por dengue ocorram anualmente e que aproximadamente 2,5 bilhões de pessoas morem em países onde a dengue é endêmica.

Este trabalho tem como objetivo apresentar uma análise de regressão para dados binários, em um estudo de dengue, onde se avalia se o indivíduo contraiu ou não a doença recentemente. O objetivo do estudo é tentar prever ou explicar a probabilidade de um indivíduo contrair a doença dadas as variáveis explicativas: idade, nível econômico e setor da cidade. Os dados foram retirados do texto de Modelos de Regressão, do autor Gilberto de Paula, capítulo 3.

2 Material e Métodos

2.1 Material

O estudo foi realizado com 196 indivíduos, onde o entrevistado responde se contraiu (caso=1), ou não contraiu (caso=0) a doença em questão. As variáveis presentes neste estudo são:

idade: Idade (em anos) do entrevistado
necon: Nível sócio-econômico (1:alto, 2:médio, 3:baixo)
setor: Setor da cidade onde mora o entrevistado (1:setor 1, 2:setor 2)
resp: Diagnóstico da doença (1:sim, 0:não).

Tabela 1: Primeiras 6 observações do conjunto de dados.

	idade	necon	setor	resp
1	33	1	1	0
2	35	1	1	0
3	6	1	1	0
4	60	1	1	0
5	18	3	1	1
6	26	3	1	0

2.2 Métodos

Como estamos trabalhando com dados de resposta binária, utilizaremos a distribuição Binomial:

$$y_i|x_i \sim Binomial(m_i, \pi_i)$$

E função de ligação:

$$g(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

No tópico 4, a seguir, ajustaremos o Modelo Binomial com suas funções de ligação, para determinar qual delas se adequa melhor ao nosso problema em questão e a que melhor se ajusta aos nossos dados.

3 Análise Descritiva

A fim de conhecer melhor os dados com os quais estamos estudando, e prever possíveis problemas e sugestões de modelagem, mostraremos a seguir uma análise descritiva e exploratória dos dados.

Observando o comportamento dos dados, podemos dizer que metade dos entrevistados, tem idade igual, ou inferior, à 21 anos; O nível socioecônomico varia nos três níveis, mas com a maior quantidade de entrevistados, concentrada no nível alto da população; E o setor da cidade, que foi dividiva entre setor 1 e setor2, encontra-se com aproximadamente, 60% dos entrevistados localizados no setor 1.

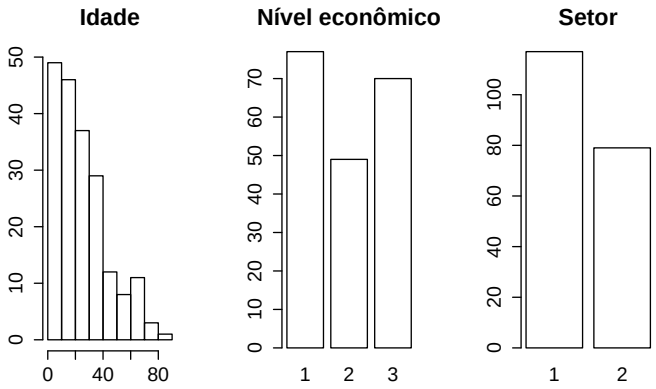

Figura 1: Histogramas

Em números, quanto mais proximo de 1, maior é a correlação que existe entre as variáveis, podendo ser positiva ou negativa. Analisando estas variáveis, nota-se que todas as correlações observadas são fracas, principalmente a da variável nível sócio-econômico com a variável respota.

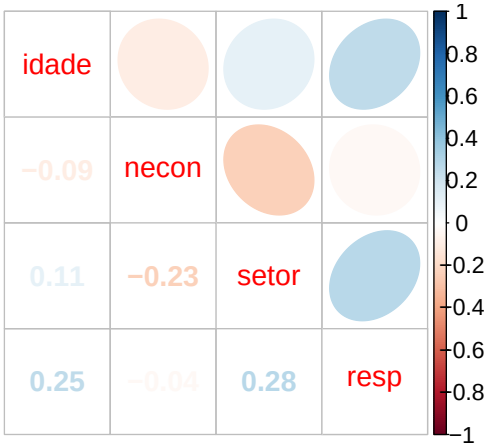

Figura 2: Gráfico de correlações

Por enquanto, manteremos todas elas no estudo, vamos primeiramente ajustar um modelo, considerando as covariáveis de forma aditiva.

4 Ajuste dos Modelos de Regressão

Testaremos nessa seção, as 4 funções de ligações: Logito, Probito, Log-Log e Cauchy, para podermos assim definir, um modelo que melhor explica a obtenção da doença recentemente. Todas estas funções de ligação que serão apresentadas a seguir, seguem a seguinte Distribuição Binomial:

$$y_i|x_i \sim Binomial(\pi_i, 1)$$

4.1 Logito

Começando pela função de ligação logito, com expressão do modelo dada por:

$$ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 resp + \beta_2 idade + \beta_3 necon + \beta_4 setor$$

```
ajuste1 <- glm(resp ~ .,family=binomial(link = (link='logit')),data = dados)
```

4.2 Probito

A segunda função de ligação, é a probito, com expressão do modelo dada por:

$$\phi^{-1}(\pi_i) = \beta_0 + \beta_1 resp + \beta_2 idade + \beta_3 necon + \beta_4 setor$$

```
ajuste2 <- glm(resp ~ .,family=binomial(link = (link='probit')),data = dados)
```

4.3 Complemento Log-Log

A terceira função de ligação, é a Complemento Log-Log, com expressão do modelo dada por:

$$ln[-ln(1-\pi)] = \beta_0 + \beta_1 resp + \beta_2 idade + \beta_3 necon + \beta_4 setor$$

```
ajuste3 <- glm(resp ~ .,family=binomial(link = (link='cloglog')),data = dados)
```

4.4 Cauchy

A última função de ligação que testaremos, é a Cauchy, com expressão do modelo dada por:

$$tan[\pi_i(\mu_i - 0,5)] = \beta_0 + \beta_1 resp + \beta_2 idade + \beta_3 necon + \beta_4 setor$$

```
ajuste4 <- glm(resp ~ .,family=binomial(link = (link='cauchit')),data = dados)
```

5 Escolha do Modelo

Utilizando os modelos ajustados, anteriormente propostos, definiremos qual deles é o mais adequado por meio das medidas do critério de informação AIC e da verossimilhança.

```
##    ajuste      aic    logLik
## 1  logito 219.2635 -105.6318
## 2 probito 218.9677 -105.4839
## 3 cloglog 219.7237 -105.8618
## 4  cauchy 220.3520 -106.1760
```

O modelo escolhido portanto foi o Probito, pois apresentou o menor AIC e a maior verossimilhança. Porém, não é sempre que todas as variáveis influenciam na variável resposta, então reajustaremos o modelo selecionado, deixando apenas as variáveis significativas.

Tabela 2: Estimativas do modelo ajustado

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.2074	0.4636	-4.76	0.0000
idade	0.0168	0.0052	3.25	0.0012
necon	0.0723	0.1193	0.61	0.5447
setor	0.7343	0.2059	3.57	0.0004

Como podemos observar na Tabela 2, usaremos apenas as variáveis idade e setor em nosso modelo final, dado que a variável necon (nível sócio-econômico) não apresentou significancia. E nosso novo modelo é o seguinte:

$$g(\pi_i) = \phi^{-1}(\pi_i) = \beta_0 + \beta_1 resp + \beta_2 idade + \beta_3 setor$$

6 Análise de Diagnóstico

Vamos primeiramente avaliar a qualidade do nosso ajuste.

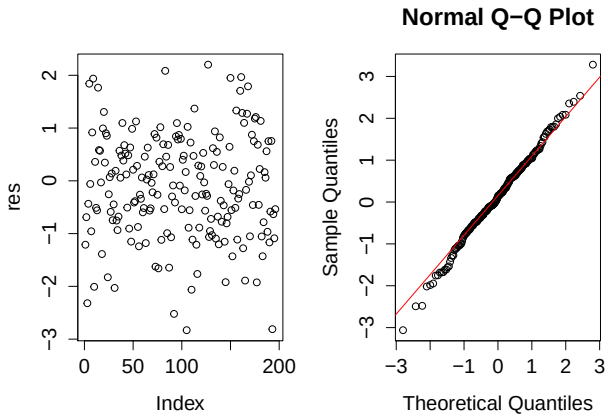

Figura 3: Resíduos Quantílicos Aleatorizados

Como vemos na Figura 3, os resíduos estão dispersos dentro do intervalo aceitável e apresentam uma boa adesividade a distribuição Normal, o que é indício de um bom ajuste.

Em uma análise de resíduos do modelo, foi observado que aparentemente não existem candidatos a outliers e que o modelo atende os pressupostos. Para podermos afirmar com certeza, veremos as medidas influentes, ou seja, algum valor que se destaque em relação aos demais. Como observado na Figura 4, não há candidatos a outliers, nem indicativos de observações influentes, portanto, nenhuma das observações foi retirada do estudo.

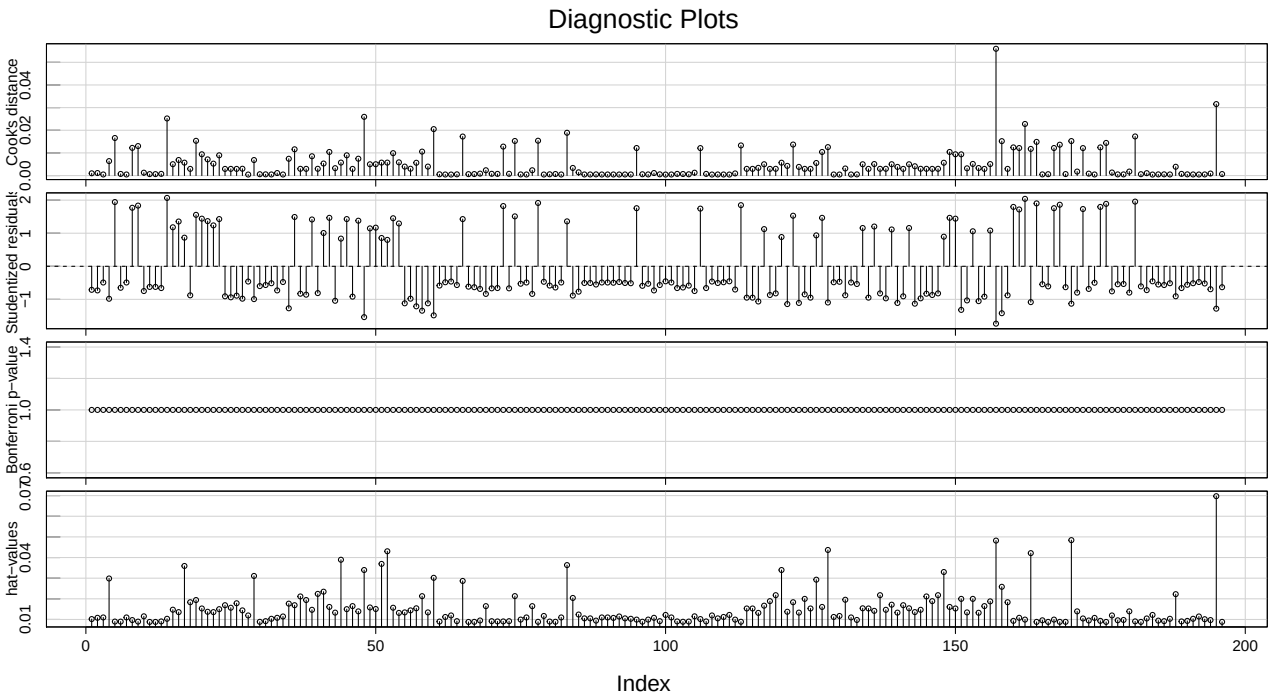

Figura 4: Gráfico de Medidas de Influência

Foi também verificada a adequação do modelo à distribuição Normal, por meio de um envelope simulado, onde os resíduos se apresentaram dispersos dentro do intervalo esperado. Portanto assim, podemos concluir que o ajuste está correto.

7 Predição

Selecionamos aleatóriamente, dois indivíduos do banco de dados, para tentarmos prever a probabilidade destes dois indivíduos contraírem a doença dadas as variáveis explicativas idade e setor.

```
aleat <- sample(1:196, size = 2)
```

Indivíduo 1
Idade: 23
Setor: 1

Indivíduo 2
Idade: 16
Setor: 2

Qual seria a probabilidade destes indivíduos acabarem contraindo a doença (caso=1), tendo estes determinados perfis?

```
predict(ajustec, interval = 'prediction', newdata = perfis, type = 'response')
```

```
##         1         2
## 0.1761262 0.3655618
```

Como temos poucas amostras, não é aconselhado separar a base, portanto não conseguimos testar o poder de predição do modelo. Esta predição feita, é apenas para fins demonstrativos.
Como observado, a probabilidade de o indivíduo 1 contrair a dengue, é baixa, entorno de 17%, enquanto a probabilidade de o indivíduo 2 contrair dengue, é de aproximadamente 36%.