

Introdução ao Software R e à Análise Econométrica

Junho a Setembro de 2018

Alexandre Xavier Ywata Carvalho
Geraldo Sandoval Góes

Introdução à Regressão Logística

Regressão com Resposta Binária

- Considere o modelo de regressão tradicional:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i$$

- Nesse modelo, a variável dependente y_i geralmente é uma variável contínua (renda per capita, taxa de mortalidade etc.)
- Uma das hipóteses básicas comumente encontrada nos livros de estatística é que variável y_i possui distribuição normal; essa hipótese não necessita ser verdadeira, para que possamos utilizar os modelos de regressão linear
- Por outro lado, há diversas situações nas quais seria interessante termos um modelo de regressão adaptado, para diferentes tipos de variável resposta
- Uma dessas situações correspondem aos casos nos quais a variável resposta é uma variável binária
- A variável resposta pode corresponder a, por exemplo: cliente pagou ou não pagou o empréstimo, o curso de pós-graduação foi ou não bem sucedido, o imóvel é alugado ou próprio etc.

Regressão com Resposta Binária

- Considere o modelo de regressão tradicional:
- Nesse modelo, a variável dependente geralmente é uma variável contínua (renda per capita, taxa de mortalidade etc.)
- Uma das hipóteses básicas comumente encontrada nos livros de estatística é que variável possui distribuição normal; essa hipótese não necessita ser verdadeira, para que possamos utilizar os modelos de regressão linear
- Por outro lado, há diversas situações nas quais seria interessante termos um modelo de regressão adaptado, para diferentes tipos de variável resposta
- Uma dessas situações correspondem aos casos nos quais a variável resposta é uma variável binária
- A variável resposta pode corresponder a, por exemplo: cliente pagou ou não pagou o empréstimo, o curso de pós-graduação foi ou não bem sucedido, o imóvel é alugado ou próprio etc.

Regressão com Resposta Binária

- Na prática, precisamos codificar devidamente as duas alternativas para as variáveis resposta
- A codificação mais comum é através da utilização dos valores 0 e 1; por exemplo, 0 corresponde a imóvel alugado e 1 corresponde a imóvel próprio; 0 corresponde a um curso mal sucedido e 1 corresponde a um curso bem sucedido
- Dessa forma, podemos sempre utilizar um template mais geral, com uma variável resposta y_i assumindo valores 0 ou 1 (importante ter claramente na nossa mente o que é o valor 0 e o que é o valor 1)
- Portanto, na nossa tabela de dados, precisamos ter uma coluna, com valores estritamente 0 ou 1, dependendo da categoria da variável resposta
- Em geral, os softwares estatísticos estão preparados para trabalhar com outras categorizações, não somente 0 e 1 apenas. O usuário pode indicar qual a categoria corresponde à situação de “sucesso”
- O termo “sucesso” utilizado nesse caso vem da variável aleatória de Bernoulli

Regressão com Resposta Binária

- Na prática, precisamos codificar devidamente as duas alternativas para as variáveis resposta
- A codificação mais comum é através da utilização dos valores 0 e 1; por exemplo, 0 corresponde a imóvel alugado e 1 corresponde imóvel próprio; 0 corresponde a um curso mal sucedido e 1 corresponde a um curso bem sucedido
- Dessa forma, podemos sempre utilizar um template mais geral, com uma variável resposta assumindo valores 0 ou 1 (importante ter claramente na nossa mente o que é o valor 0 e o que é o valor 1)
- Portanto, na nossa tabela de dados, precisamos ter uma coluna, com valores estritamente 0 ou 1, dependendo da categoria da variável resposta
- Em geral, os softwares estatísticos estão preparados para trabalhar com outras categorizações, não somente 0 e 1 apenas. O usuário pode indicar qual a categoria corresponde à situação de “sucesso”
- O termo “sucesso” utilizado nesse caso vem da variável aleatória de Bernoulli

Regressão com Resposta Binária

Situação do Imóvel	Idade do Chefe	Número de Residentes	Renda Familiar (R\$)	Variável Y (Preencher ...)
Alugado	46	3	2200	
Alugado	50	2	1500	
Próprio	28	4	4600	
Alugado	31	4	2823	
Próprio	63	3	4100	
Próprio	53	2	1200	
Alugado	36	2	7800	
Alugado	51	3	3230	
Próprio	42	6	5622	

Regressão com Resposta Binária

- A variável aleatória de Bernoulli, tradicionalmente vista nos livros de estatística, corresponde a uma variável que assume apenas dois valores, 0 ou 1, sendo que 1 corresponde à situação de “sucesso” e 0 à situação de “insucesso”. Obviamente, esses termos são totalmente ilustrativos
- O importante nessa conceituação é que, atrelado ao evento de sucesso, temos uma probabilidade. Essa probabilidade de sucesso é normalmente representada pela letra p , e está entre 0 e 1
- Um exemplo muito comum da variável de Bernoulli é a variável aleatória associada a jogarmos uma moeda
- Cara corresponde a “sucesso” e tem probabilidade de $p = 50\%$ (assumindo que a moeda é não viciada)
- Seja X então a variável aleatória nesse caso. Sabemos que X assume valores 0 ou 1 (de acordo com a nossa codificação, sendo que escolhemos arbitrariamente que 1 corresponde a “cara” e 0 a “coroa”)
- Lembrando que o espaço amostral S corresponde ao conjunto de valores possíveis de uma variável aleatória. Nesse caso, $S = \{0, 1\}$
- Como podemos modelar então um caso mais geral de jogada de uma moeda N vezes, e contagem do número de vezes que a moeda resultou “cara”?

Variável Aleatória Binomial

- **Variável aleatória binomial** – trata-se de um “template” muito utilizado, para modelar, por exemplo, o número ocorrência de “sucesso” em N tentativas. Por exemplo, em um grupo de 100 pacientes, quantos têm algum tipo de câncer. O número de pacientes com câncer entre os 100 no grupo pode ser modelado por uma variável aleatória binomial.
- O espaço amostral de uma variável aleatória binomial é dado por $S = \{0, 1, 2, 3, 4, \dots, N\}$
- A função de frequência de uma variável aleatória binomial tem expressão:

$$\text{Prob}[X = x] = f(x) = \binom{N}{x} p^x (1 - p)^{N-x}, \quad x = 0, 1, 2, 3, 4, \dots, N$$

- O símbolo $\binom{N}{x}$ corresponde ao número de combinações possíveis de x elementos entre os N totais

$$\binom{N}{x} = \frac{N!}{x! (N - x)!} = \frac{1 \times 2 \times \dots \times (N - 1) \times N}{(1 \times 2 \times \dots \times x) \times (1 \times 2 \times \dots \times (N - x))}$$

- Em geral, N é conhecido e procura-se estimar o parâmetro p com base em uma amostra. O parâmetro p pode ser interpretado como a probabilidade de um indivíduos no grupo ter câncer. Portanto, p varia entre 0 e 1.
- Quando $N = 1$, a variável binomial é chamada variável de Bernoulli, e tem $S = \{0,1\}$

Variável Aleatória Binomial

- **Variável aleatória binomial** – trata-se de um “template” muito utilizado, para modelar, por exemplo, o número ocorrência de “sucesso” em N tentativas. Por exemplo, em um grupo de 100 pacientes, quantos têm algum tipo de câncer. O número de pacientes com câncer entre os 100 no grupo pode ser modelado por uma variável aleatória binomial.
- O espaço amostral de uma variável aleatória binomial é dado por $S = \{0, 1, 2, 3, 4, \dots, N\}$
- A função de frequência de uma variável aleatória binomial tem expressão:
- O símbolo $\binom{N}{x}$ corresponde ao número de combinações possíveis de x elementos entre os N totais
- Em geral, N é conhecido e procura-se estimar o parâmetro p com base em uma amostra. O parâmetro p pode ser interpretado como a probabilidade de um indivíduos no grupo ter câncer. Portanto, p varia entre 0 e 1.
- Quando $N = 1$, a variável binomial é chamada variável de Bernoulli, e tem $S = \{0,1\}$

Variável Aleatória Binomial

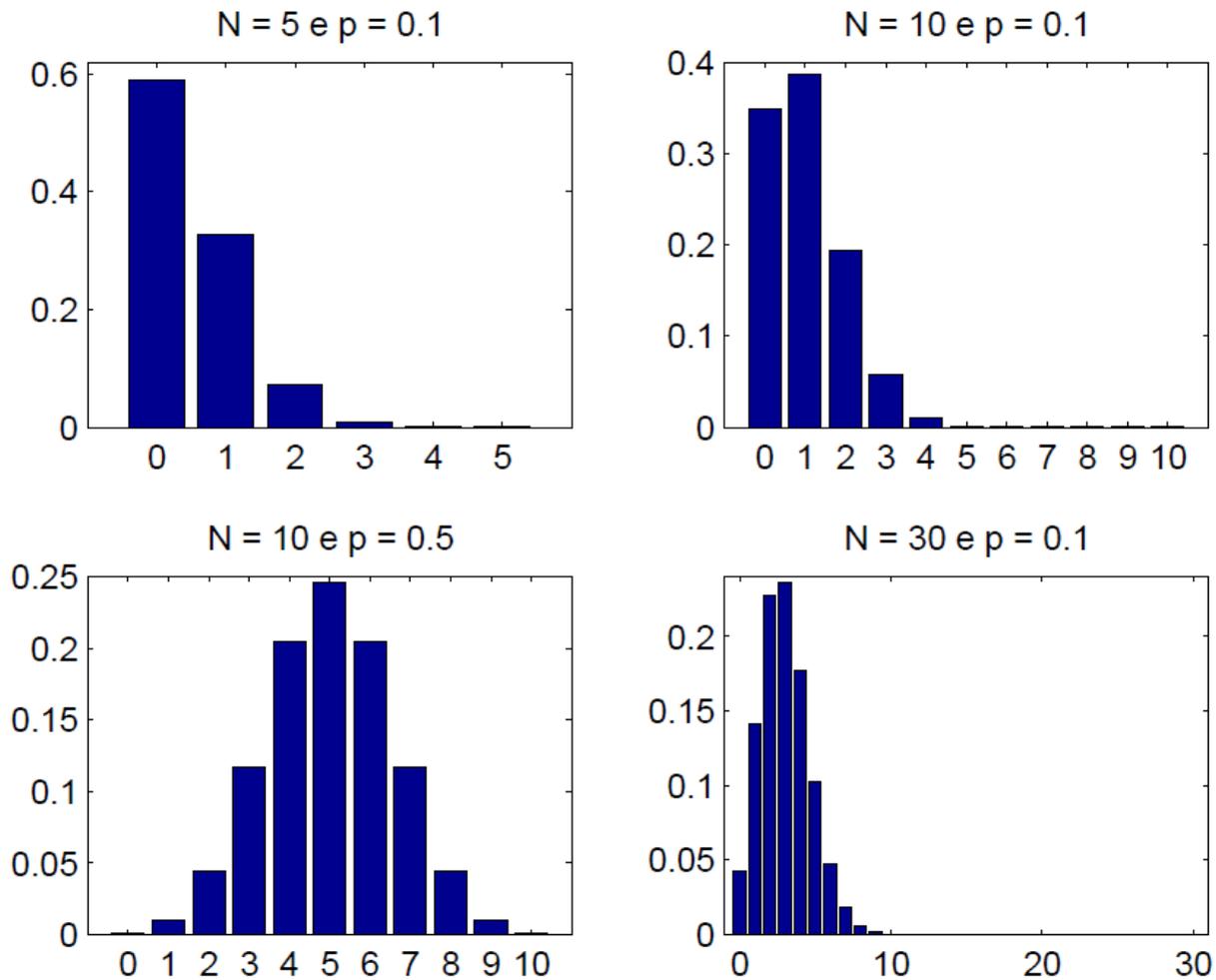


Figura 3.1: Função de frequência para a variável aleatória binomial.

Exercícios

- Exercício. Seja X uma variável aleatória binomial, com parâmetros $N = 10$ e $p = 0.2$. Determine:

- a) O espaço amostral S
- b) $f(0) = \text{Prob}[X = 0]$
- c) $f(1) = \text{Prob}[X = 1]$
- d) $f(10) = \text{Prob}[X = 10]$
- e) $f(6) = \text{Prob}[X = 6]$
- f) A probabilidade de que X seja menor ou igual a 3 ($\text{Prob}[X \leq 3]$)
- g) A probabilidade de que X seja maior ou igual a 6 ($\text{Prob}[X \geq 6]$)

- Para a variável aleatória binomial, algumas funções úteis no R

```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)
```

Simulações com Variáveis Aleatórias

#--- simulações de Monte Carlo com variáveis binomiais (10 tentativas e prob = 0.2)

```
amostra <- rbinom(n=20, size=10, prob=0.2)
amostra
amostra <- rbinom(n=200, size=10, prob=0.2)
amostra
amostra <- rbinom(n=20000, size=10, prob=0.2)
hist(amostra, col = "red")
```

```
media <- mean(amostra)
media
variancia <- var(amostra)
Variancia
```

#--- simulações de Monte Carlo com variáveis de Bernoulli

```
amostra <- rbinom(n=200, size=1, prob=0.2)
amostra
amostra <- rbinom(n=20000, size=1, prob=0.2)
hist(amostra, col = "red")
```

```
media <- mean(amostra)
```

Momentos de Variáveis Aleatórias

- Nos exemplos anteriores, vocês simularam diferentes valores aleatoriamente para algumas das distribuições “template”
- Com base nos valores simulados, vocês encontraram estimativas para as médias e para as variâncias (e conseqüentemente para os desvios padrões) dessas distribuições
- Essas estimativas parecem estáveis quando aumentamos o número de valores simulados
- Quando alteramos os parâmetros livres das variáveis aleatórias simuladas, os valores de estimativas para as médias e variâncias também se alteram
- A pergunta então é:

É possível antecipar quais valores de média e variância (e desvio padrão) serão obtidos?

De outra forma:

Para determinados valores dos parâmetros das distribuições “template”, qual os valores

Momentos de Variáveis Aleatórias

- Os valores das médias e variâncias para as variáveis aleatórias “template”, e outras de forma geral, podem ser obtidas a partir das funções de frequência $f(x) = \text{Prob}[X = x]$
- Definimos como **média**, **valor esperado**, **expectância** ou **primeiro momento** de uma variável aleatória discreta, com função de frequência $f(x)$, o somatório:

$$E[X] = \sum_{x \in S} x \times f(x) = \sum_{x \in S} x \times \text{Prob}[X = x]$$

- O somatório ocorre para todos os valores de x no espaço amostral S
- Exemplo: considere uma variável aleatória discreta com função frequência $f(0) = f(1) = f(2) = 0.2$, e espaço amostral $S = \{0, 1, 2, 3\}$. A média dessa variável aleatória é dada por:

$$E[X] = 0 \times 0.2 + 1 \times 0.2 + 2 \times 0.2 + 3 \times 0.4 = 1.8$$

- Exercício. Seja X uma variável aleatória discreta, com $S = \{1, 2, 3, 4\}$, e função frequência $f(1) = f(2) = f(3) = 0.2$. Encontre a média dessa distribuição.
- Exercício. Seja X uma variável aleatória discreta, com $S = \{2, 4, 6, 8\}$, e função frequência $f(2) = f(4) = f(6) = 0.2$. Encontre a média dessa distribuição.

Momentos de Variáveis Aleatórias

- Os valores das médias e variâncias para as variáveis aleatórias “template”, e outras de forma geral, podem ser obtidas a partir das funções de frequência $f(x) = \text{Prob}[X = x]$
- Definimos como **média**, **valor esperado**, **expectância** ou **primeiro momento** de uma variável aleatória discreta, com função de frequência $f(x)$, o somatório:
 - O somatório ocorre para todos os valores de x no espaço amostral S
 - Exemplo: considere uma variável aleatória discreta com função frequência $f(0) = f(1) = f(2) = 0.2$, e espaço amostral $S = \{0, 1, 2, 3\}$. A média dessa variável aleatória é dada por:

$$E[X] = 0 \times 0.2 + 1 \times 0.2 + 2 \times 0.2 + 3 \times 0.4 = 1.8$$

- Exercício. Seja X uma variável aleatória discreta, com $S = \{1, 2, 3, 4\}$, e função frequência $f(1) = f(2) = f(3) = 0.2$. Encontre a média dessa distribuição.

Momentos de Variáveis Aleatórias

- Definimos como **variância**, ou **segundo momento centrado** de uma variável aleatória discreta, com função de frequência $f(x)$, o somatório:

$$Var[X] = \sum_{x \in S} [x - E[X]]^2 \times f(x) = \sum_{x \in S} [x - E[X]]^2 \times \text{Prob}[X = x]$$

- O somatório ocorre para todos os valores de x no espaço amostral S
- Exemplo: considere uma variável aleatória discreta com função frequência $f(0) = f(1) = f(2) = 0.2$, e espaço amostral $S = \{0, 1, 2, 3\}$. A variância dessa variável aleatória é dada por:

$$E[X] = [0 - 1.8]^2 \times 0.2 + [1 - 1.8]^2 \times 0.2 + [2 - 1.8]^2 \times 0.2 + [3 - 1.8]^2 \times 0.4 = ?$$

- Exercício. Seja X uma variável aleatória discreta, com $S = \{1, 2, 3, 4\}$, e função frequência $f(1) = f(2) = f(3) = 0.2$. Encontre a variância dessa distribuição.
- Exercício. Seja X uma variável aleatória discreta, com $S = \{2, 4, 6, 8\}$, e função frequência $f(2) = f(4) = f(6) = 0.2$. Encontre a variância dessa distribuição.

Momentos de Variáveis Aleatórias

- Definimos como **variância**, ou **segundo momento centrado** de uma variável aleatória discreta, com função de frequência $f(x)$, o somatório:
 - O somatório ocorre para todos os valores de x no espaço amostral S
 - Exemplo: considere uma variável aleatória discreta com função frequência $f(0) = f(1) = f(2) = 0.2$, e espaço amostral $S = \{0, 1, 2, 3\}$. A variância dessa variável aleatória é dada por:
$$E[X] = [0 - 1.8]^2 \times 0.2 + [1 - 1.8]^2 \times 0.2 + [2 - 1.8]^2 \times 0.2 + [3 - 1.8]^2 \times 0.4 = ?$$
- Exercício. Seja X uma variável aleatória discreta, com $S = \{1, 2, 3, 4\}$, e função frequência $f(1) = f(2) = f(3) = 0.2$. Encontre a variância dessa distribuição.
- Exercício. Seja X uma variável aleatória discreta, com $S = \{2, 4, 6, 8\}$, e função frequência $f(2) = f(4) = f(6) = 0.2$. Encontre a variância dessa distribuição.

Momentos de Variáveis Aleatórias

- Para variáveis aleatórias “template” que vimos acima, existem formas bem definidas para as médias e as variâncias
- Para a variável aleatória X de Bernoulli, para a qual $S = \{0,1\}$, considerando-se uma probabilidade de sucesso $p = 0.1$, calcule a média e a variância:

$$E[X] = 1 \times 0.1 + 0 \times 0.9 = 0.1$$

$$\text{Var}[X] = [1 - 0.1]^2 \times 0.1 + [0 - 0.1]^2 \times 0.9 = 0.9^2 \times 0.1 + (-0.1)^2 \times 0.9 = 0.9 \times 0.1$$

- Para a variável aleatória X de Bernoulli, com probabilidade de sucesso $p = 0.5$, calcule a média e a variância:

$$E[X] = 1 \times 0.5 + 0 \times 0.5 = 0.5$$

$$\text{Var}[X] = [1 - 0.5]^2 \times 0.5 + [0 - 0.5]^2 \times 0.5 = 0.5^2 \times 0.5 + (-0.5)^2 \times 0.5 = 0.5 \times 0.5$$

- De maneira mais geral, pode-se mostrar que para uma variável de Bernoulli, com parâmetro p (entre 0 e 1), temos:

$$E[X] = p$$

$$\text{Var}[X] = p \times (1 - p)$$

Momentos de Variáveis Aleatórias

- Conforme discutimos anteriormente, uma variável de Bernoulli corresponde a uma variável aleatória binomial, com $N = 1$
- De maneira mais geral, para uma variável aleatória binomial, com parâmetros N e p , temos a média:

$$E[X] = \sum_{x \in S} x \times f(x) = \sum_{x=0}^N x \times \binom{N}{x} p^x (1-p)^{N-x}$$

- A variância é dada por:

$$Var[X] = \sum_{x \in S} [x - E[X]]^2 \times f(x) = \sum_{x=0}^N [x - E[X]]^2 \times \binom{N}{x} p^x (1-p)^{N-x}$$

- Pode-se mostrar que, para a variável aleatória de binomial, com parâmetros N e p :

$$\begin{aligned} E[X] &= N \times p \\ Var[X] &= N \times p \times (1 - p) \end{aligned}$$

Momentos de Variáveis Aleatórias

- Conforme discutimos anteriormente, uma variável de Bernoulli corresponde a uma variável aleatória binomial, com $N = 1$
- De maneira mais geral, para uma variável aleatória binomial, com parâmetros N e p , temos a média:
- A variância é dada por:
- Pode-se mostrar que, para a variável aleatória de binomial, com parâmetros N e p :

Exercícios

- **Exercício 7 - para entregar em 2 semanas:**

- Como de costume, os exercícios podem ser entregues em grupos de 2 ou três alunos, e o grupo deve submeter o código em R utilizado para responder ao exercício, juntamente com a discussão dos resultados
- Utilize como base o código em R 'Análise_de_Regressão_com_Variáveis_Binárias'

- Questão 1. Seja X uma variável aleatória binomial, com parâmetros $N = 25$ e $p = 0.6$. Utilizando os códigos de demonstração em R, responda:

a) Utilizando 10000 valores gerados aleatoriamente, encontre uma estimativa para a média dessa distribuição

a) Utilizando 10000 valores gerados aleatoriamente, encontre uma estimativa para a variância dessa distribuição

a) Quais os valores “teóricos” da média e da variância para essa distribuição?

a) Os valores “teóricos” correspondem aos valores estimados via simulações de Monte Carlo?

O Modelo de Regressão Logística

- Conforme vimos acima, a variável de Bernoulli assume valores 0 ou 1, com probabilidade de sucesso $\text{Prob}[Y = 1] = p$, e probabilidade de insucesso $\text{Prob}[Y = 0] = 1-p$
- Como adaptar então o conceito de variável de Bernoulli ao conceito de regressão?
- Vamos agora tratar então da chamada regressão logística
- Consideremos então uma base de dados de unidades observacionais (indivíduos, domicílios, municípios, países, cursos de pós-graduação etc.)
- Para cada unidade observacional, temos uma variável y_i assumindo valores 0 ou 1, e temos um conjunto de colunas que podem ser usadas para construirmos variáveis explicativas $x_{1,i}, x_{2,i}, \dots, x_{k,i}$
- A ideia básica da regressão logística é assumir que cada valor individual y_i corresponde a uma variável aleatória de Bernoulli, com probabilidade de sucesso (por exemplo, indivíduo ter câncer – paradoxalmente!) dada por $\text{Prob}[y_i = 1] = p_i$
- **O “pulo do gato” é fazer com que $\text{Prob}[y_i = 1] = p_i$ dependa das variáveis explicativas $x_{1,i}, x_{2,i}, \dots, x_{k,i}$**

O Modelo de Regressão Logística

- Conforme vimos acima, a variável de Bernoulli assume valores 0 ou 1, com probabilidade de sucesso $\text{Prob}[Y = 1] = p$, e probabilidade de insucesso $\text{Prob}[Y = 0] = 1-p$
- Como adaptar então o conceito de variável de Bernoulli ao conceito de regressão?
- Vamos agora tratar então da chamada regressão logística
- Consideremos então uma base de dados de unidades observacionais (indivíduos, domicílios, municípios, países, cursos de pós-graduação etc.)
- Para cada unidade observacional, temos uma variável assumindo valores 0 ou 1, e temos um conjunto de colunas que podem ser usadas para construirmos variáveis explicativas
- A ideia básica da regressão logística é assumir que cada valor individual corresponde a uma variável aleatória de Bernoulli, com probabilidade de sucesso (por exemplo, indivíduo ter câncer – paradoxalmente!) dada por
- **O “pulo do gato” é fazer com que dependa das variáveis explicativas**

O Modelo de Regressão Logística

- O “pulo do gato” é fazer com que $\text{Prob}[y_i = 1] = p_i$ dependa das variáveis explicativas $x_{1,i}, x_{2,i}, \dots, x_{k,i}$

- Uma possível alternativa é assumir

$$\text{Prob}[y_i = 1] = p_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$$

- O problema da alternativa acima é que $\text{Prob}[y_i = 1] = p_i$ tem que estar estritamente no intervalo $[0, 1]$
- O termo $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$, por outro lado, pode assumir valores menores do que 0 ou maiores do que 1
- Modelo de regressão logística:

$$\text{Prob}[y_i = 1] = p_i = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}}$$

- A fórmula acima implica que as probabilidades p_i vão se situar o intervalo $(0,1)$, como desejado
- Pode-se mostrar que, quando β_1 é positivo, quando x_{1i} aumenta, a probabilidade de “sucesso” também aumenta

O Modelo de Regressão Logística

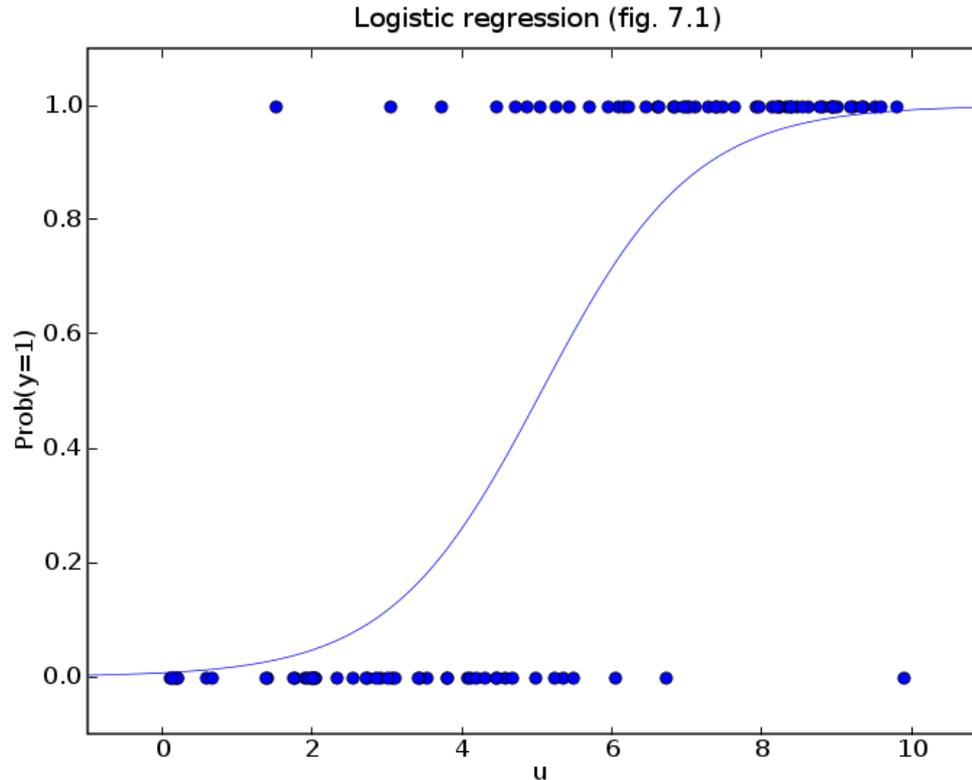
- O “pulo do gato” é fazer com que dependa das variáveis explicativas
- Uma possível alternativa é assumir
- O problema da alternativa acima é que tem que estar estritamente no intervalo $[0, 1]$
- O termo por outro lado, pode assumir valores menores do que 0 ou maiores do que 1
- Modelo de regressão logística:
- A fórmula acima implica que as probabilidades vão se situar o intervalo $(0,1)$, como desejado
- Pode-se mostrar que, quando é positivo, quando aumenta, a probabilidade de “sucesso” também aumenta

O Modelo de Regressão Logística

- Considere um modelo simplificado de regressão logística, no qual temos a probabilidade de sucesso dada por

com

$$\text{Prob}[y_i = 1] = p_i = \frac{e^{\alpha + \beta x_{1i}}}{1 + e^{\alpha + \beta x_{1i}}}, \text{ com } \beta > 0$$



Regressão Logística no R

```
dados3$alta_mort_infantil <- ifelse(dados3$mort_infantil > 24, 1, 0)
```

```
#-----  
#--- rodando uma regressão logística  
#-----
```

```
mod1 <- glm(formula = alta_mort_infantil ~ renda_per_capita,  
            family = binomial(link = "logit"), data = dados3)  
summary(mod1)
```

```
mod2 <- glm(formula = alta_mort_infantil ~ indice_gini,  
            family = binomial(link = "logit"), data = dados3)  
summary(mod2)
```

```
mod3 <- glm(formula = alta_mort_infantil ~ perc_criancas_extrem_pobres,  
            family = binomial(link = "logit"), data = dados3)  
summary(mod3)
```

```
mod4 <- glm(formula = alta_mort_infantil ~ perc_pessoas_dom_agua_estogo_inadequados,  
            family = binomial(link = "logit"), data = dados3)  
summary(mod4)
```

Regressão Logística no R

```
> summary(mod1)
```

Call:

```
glm(formula = alta_mort_infantil ~ renda_per_capita, family = binomial(link = "logit"),  
     data = dados3)
```

Deviance Residuals:

```
   Min     1Q  Median     3Q      Max  
-2.4659 -0.2831 -0.0536 -0.0003  3.1928
```

Coefficients:

```
              Estimate Std. Error z value Pr(>|z|)  
(Intercept)    5.2070406  0.1834282  28.39 <2e-16 ***  
renda_per_capita -0.0182626  0.0006154 -29.68 <2e-16 ***
```

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

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 6163.7 on 5563 degrees of freedom  
Residual deviance: 3042.4 on 5562 degrees of freedom  
AIC: 3046.4
```

Number of Fisher Scoring iterations: 7

Regressão Logística no R

```
> summary(mod4)
```

Call:

```
glm(formula = alta_mort_infantil ~ perc_pessoas_dom_agua_estogo_inadequados,  
     family = binomial(link = "logit"), data = dados3)
```

Deviance Residuals:

```
   Min     1Q  Median     3Q    Max  
-3.4654 -0.5446 -0.4690 -0.4570  2.1500
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.206750	0.051311	-43.01	<2e-16 ***
perc_pessoas_dom_agua_estogo_inadequados	0.096166	0.003172	30.32	<2e-16 ***

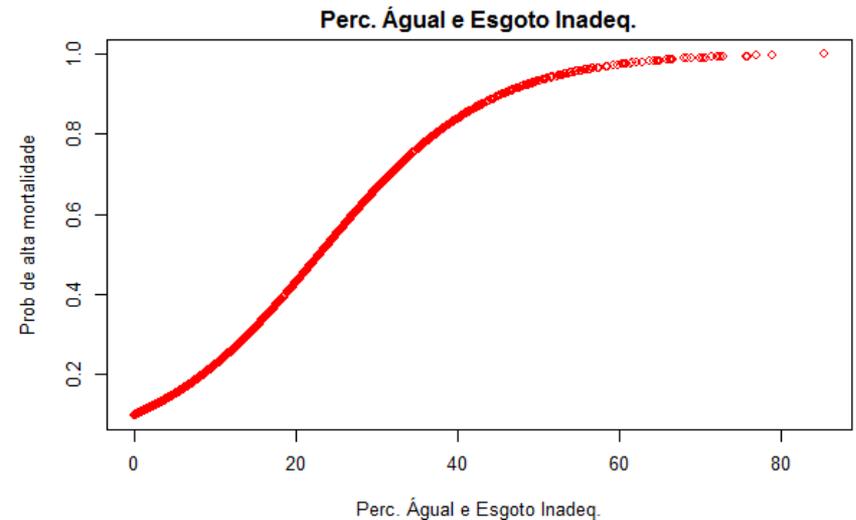
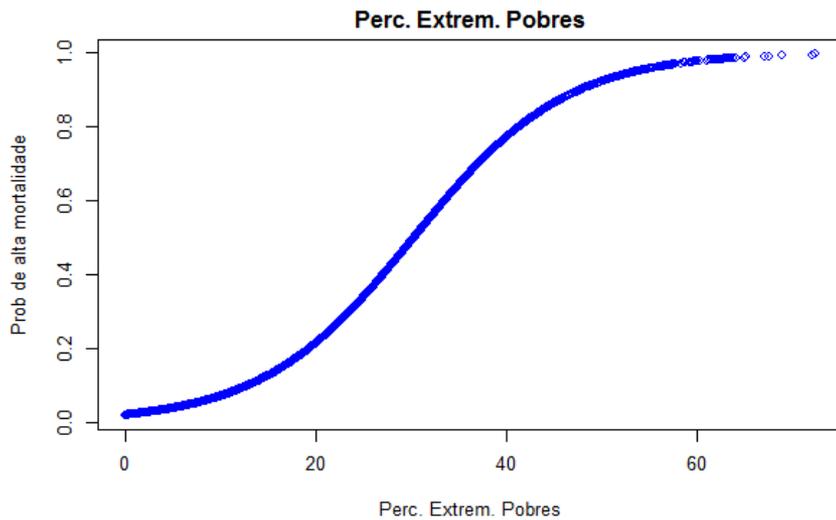
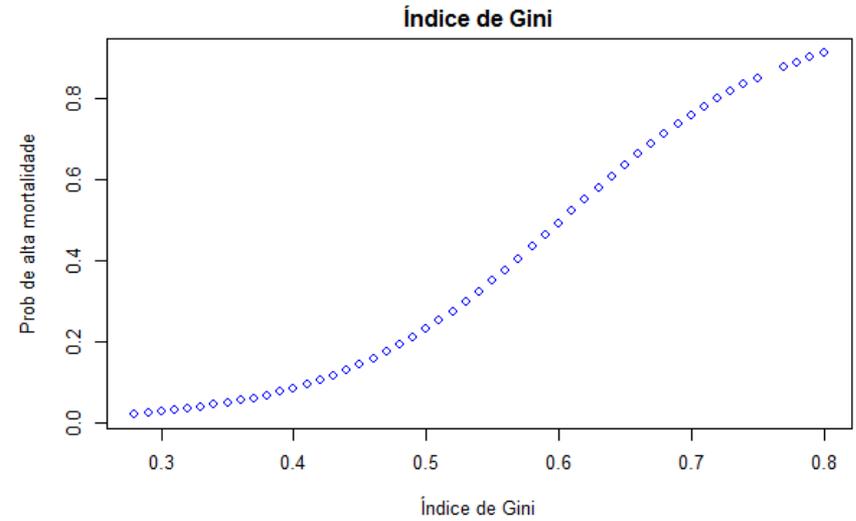
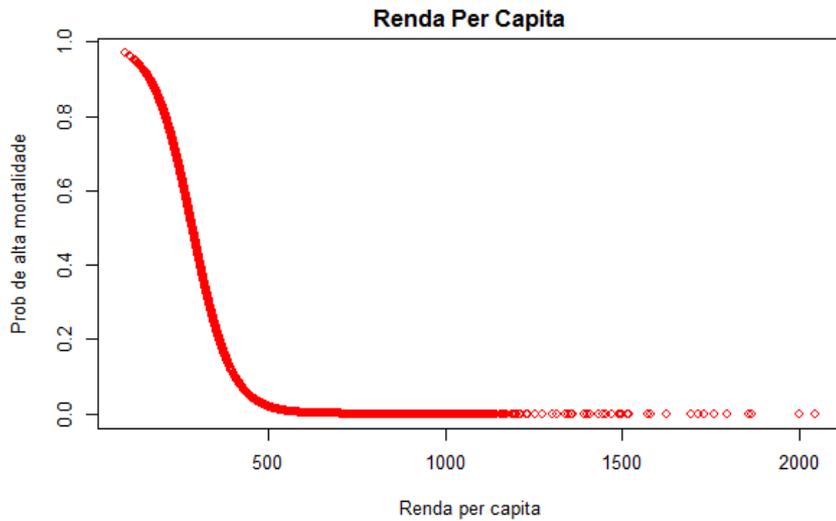
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 6163.7 on 5563 degrees of freedom
Residual deviance: 4819.0 on 5562 degrees of freedom
AIC: 4823

Number of Fisher Scoring iterations: 4

Regressão Logística no R



Exemplo

Table 4. Final logistic regression model of the variable amputation as a function of social and clinical variables

Variable	β	S.E.	Wald		p-value	OR	95% CI
			χ^2	df			
Lack of primary care assistance	1.193	0.584	4.176	1	0.041	3.30	1.05–10.36
Previous amputation	2.390	0.740	10.434	1	0.001	10.91	2.56–46.51
CKD	0.835	0.576	2.102	1	0.147	2.31	0.75–7.12
CAD	1.68	0.689	5.92	1	0.015	5.35	1.38–20.68
AA	2.77	1.07	6.67	1	0.010	15.90	1.95–129.63
Hemoglobin A1C	1.58	0.282	31.46	1	<0.001	4.87	2.80–8.47
Constant	-14.33	2.32	38.18	1	<0.001	--	---

Rcr (Cox and Snell R^2)=0.547; RN (Nagelkerke R^2)=0.749

B: Coefficient of the logistic regression equation to predict the dependent variable using the independent variable.

SE: Standard errors associated with the coefficients.

Wald: Wald chi-squared test to test the null hypothesis that the constant is equal to 0

df: Degree of freedom for the Wald chi-squared test.

Método de Máxima Verossimilhança

- Da mesma forma que no caso da regressão linear, com base em uma amostra de observações, queremos estimar os parâmetros desconhecidos $\beta_0, \beta_1, \beta_2, \dots, \beta_k$
- O método comumente utilizado nesse caso é o método de “máxima verossimilhança”
- Para cada observação i , a probabilidade que observaremos $y_i=1$ é igual a p_i , enquanto a probabilidade de observarmos $y_i=0$ é igual a $(1-p_i)$
- De maneira compacta, podemos dizer que a probabilidade de observar o valor y_i (0 ou 1) é igual a

$$\text{Prob}[y_i] = p_i^{y_i} \times (1 - p_i)^{1-y_i}$$

- De fato, se $y_i=1$, $\text{Prob}[y_i = 1] = p_i^1 \times (1 - p_i)^{1-1} = p_i$
- De fato, se $y_i=0$, $\text{Prob}[y_i = 0] = p_i^0 \times (1 - p_i)^{1-0} = (1 - p_i)$
- A probabilidade de observar toda amostra é dada pelo produto das probabilidades individuais (assumindo que as observações são independentes)

$$\text{Prob}[y_1, \dots, y_n] = \prod_{i=1}^n p_i^{y_i} \times (1 - p_i)^{1-y_i}$$

Método de Máxima Verossimilhança

- Da mesma forma que no caso da regressão linear, com base em uma amostra de observações, queremos estimar os parâmetros desconhecidos
- O método comumente utilizado nesse caso é o método de “máxima verossimilhança”
- Para cada observação i , a probabilidade que observaremos $=1$ é igual a p , enquanto a probabilidade de observarmos $=0$ é igual a $(1-p)$
- De maneira compacta, podemos dizer que a probabilidade de observar o valor $(0$ ou $1)$ é igual a

- De fato, se $=1$,
- De fato, se $=0$,
- A probabilidade de observar toda amostra é dada pelo produto das probabilidades individuais (assumindo que as observações são independentes)

Método de Máxima Verossimilhança

- A função de verossimilhança é justamente a probabilidade de observar o que de fato encontramos na amostra, ou seja $\text{Prob}[y_1, \dots, y_n] = \prod_{i=1}^n p_i^{y_i} \times (1 - p_i)^{1-y_i}$
- Considere então um vetor qualquer de parâmetros $\beta_0, \beta_1, \dots, \beta_k$
- A função de verossimilhança, assumindo que os valores y_i são variáveis de Bernoulli, independentes, é escrita como

$$\begin{aligned} L(\beta_0, \beta_1, \dots, \beta_k) &= \prod_{i=1}^n p_i^{y_i} \times (1 - p_i)^{1-y_i} \\ &= \prod_{i=1}^n \left[\frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}} \right]^{y_i} \times \left[\frac{1}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}} \right]^{1-y_i} \end{aligned}$$

- O método de máxima verossimilhança é comumente empregado para estimar os parâmetros do modelos de regressão de forma geral
- O método consistem em encontrar os valores dos parâmetros $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ para os quais a função $L(\beta_0, \beta_1, \dots, \beta_k)$ atinge um valor máximo
- Note que os valores de $x_{1,i}, x_{2,i}, \dots, x_{k,i}$ e y_i são conhecidos, dado que estamos usando uma amostra disponível

Método de Máxima Verossimilhança

- A função de verossimilhança é justamente a probabilidade de observar o que de fato encontramos na amostra, ou seja
- Considere então um vetor qualquer de parâmetros
- A função de verossimilhança, assumindo que os valores x_i são variáveis de Bernoulli, independentes, é escrita como
- O método de máxima verossimilhança é comumente empregado para estimar os parâmetros do modelos de regressão de forma geral
- O método consistem em encontrar os valores dos parâmetros θ para os quais a função $L(\theta)$ atinge um valor máximo
- Note que os valores de x_i e n são conhecidos, dado que estamos usando uma amostra disponível

Método de Máxima Verossimilhança

- Conforme vimos anteriores, por motivos numéricos e analíticos, trabalhamos com o log da função de verossimilhança, ao invés da função original

$$\log L(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n y_i \log p_i + (1 - y_i) \log(1 - p_i)$$

$$\begin{aligned} & \log L(\beta_0, \beta_1, \dots, \beta_k) \\ &= \sum_{i=1}^n y_i [\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}] - \sum_{i=1}^n \log[1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}] \end{aligned}$$

- Obtemos então os estimadores de máxima verossimilhança para os parâmetros $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ encontrando o máximo da função de log-verossimilhança $\log L(\beta_0, \beta_1, \dots, \beta_k)$
- Uma das formas de se encontrar os máximos de uma função é encontrar as derivadas e igualar as derivadas a zero
- Para o caso da estimação de máxima verossimilhança no caso de regressão linear, a técnica de achar as derivadas e igualar as derivadas a zero implica na fórmula fechada do estimador de mínimos quadrados ordinários
- Para regressão linear, o estimador de máxima verossimilhança é numericamente igual ao estimador de mínimos quadrados ordinários

Método de Máxima Verossimilhança

- Conforme vimos anteriores, por motivos numéricos e analíticos, trabalhamos com o log da função de verossimilhança, ao invés da função original
- Obtemos então os estimadores de máxima verossimilhança para os parâmetros encontrando o máximo da função de log-verossimilhança
- Uma das formas de se encontrar os máximos de uma função é encontrar as derivadas e igualar as derivadas a zero
- Para o caso da estimação de máxima verossimilhança no caso de regressão linear, a técnica de achar as derivadas e igualar as derivadas a zero implica na fórmula fechada do estimador de mínimos quadrados ordinários
- Para regressão linear, o estimador de máxima verossimilhança é numericamente igual ao estimador de mínimos quadrados ordinários

Método de Máxima Verossimilhança

- No caso de regressão logística, não é possível encontrar uma fórmula fechada para o estimador dos parâmetros $\beta_0, \beta_1, \beta_2, \dots, \beta_k$
- Por isso, o R (e outros programas estatísticos) têm que efetuar uma maximização iterativa numérica, quando têm que estimar os parâmetros via máxima verossimilhança
- Considere um modelo simplificado de regressão logística, no qual temos a probabilidade de sucesso dada por

$$\text{Prob}[y_i = 1] = p_i = \frac{e^{\alpha + \beta x_{1i}}}{1 + e^{\alpha + \beta x_{1i}}}$$

- Nesse caso, temos dois parâmetros desconhecidos α e β
- A função de log verossimilhança tem expressão

$$\log L(\alpha, \beta) = \sum_{i=1}^n y_i [\alpha + \beta x_{1i}] - \sum_{i=1}^n \log[1 + e^{\alpha + \beta x_{1i}}]$$

- Maximizando $\log L(\alpha, \beta)$, encontramos os estimadores $\hat{\alpha}$ e $\hat{\beta}$ para os parâmetros α e β

Método de Máxima Verossimilhança

- No caso de regressão logística, não é possível encontrar uma fórmula fechada para o estimador dos parâmetros
- Por isso, o R (e outros programas estatísticos) têm que efetuar uma maximização iterativa numérica, quando têm que estimar os parâmetros via máxima verossimilhança
- Considere um modelo simplificado de regressão logística, no qual temos a probabilidade de sucesso dada por

- Nesse caso, temos dois parâmetros desconhecidos β_0 e β_1
- A função de log verossimilhança tem expressão

- Maximizando $l(\beta_0, \beta_1)$, encontramos os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ para os parâmetros β_0 e β_1

Regressão Logística no R

```
> summary(mod1)
```

Call:

```
glm(formula = alta_mort_infantil ~ renda_per_capita, family = binomial(link = "logit"),  
     data = dados3)
```

Deviance Residuals:

```
   Min     1Q  Median     3Q      Max  
-2.4659 -0.2831 -0.0536 -0.0003  3.1928
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.2070406	0.1834282	28.39	<2e-16 ***
renda_per_capita	-0.0182626	0.0006154	-29.68	<2e-16 ***

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 6163.7 on 5563 degrees of freedom
Residual deviance: 3042.4 on 5562 degrees of freedom
AIC: 3046.4

Number of Fisher Scoring iterations: 7

Regressão Logística no R

```
> summary(mod1)
```

Call:

```
glm(formula = alta_mort_infantil ~ renda_per_capita, family = binomial(link = "logit"),  
     data = dados3)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4659	-0.2831	-0.0536	-0.0003	3.1928

???

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.2070406	0.1834282	28.39	<2e-16 ***
renda_per_capita	-0.0182626	0.0006154	-29.68	<2e-16 ***

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 6163.7 on 5563 degrees of freedom
Residual deviance: 3042.4 on 5562 degrees of freedom
AIC: 3046.4

Number of Fisher Scoring iterations: 7

Matriz de Variância-Covariância

Variâncias na diagonal principal e covariâncias fora da diagonal principal (matriz simétrica)

$$\Sigma = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) & \text{Cov}(X, Z) \\ \text{Cov}(X, Y) & \text{Var}(Y) & \text{Cov}(Y, Z) \\ \text{Cov}(X, Z) & \text{Cov}(Y, Z) & \text{Var}(Z) \end{bmatrix}$$

Matriz de Correlações

Correlação entre as variáveis fora da diagonal principal (matriz simétrica com diagonal principal com todos os elementos iguais a 1)

```
. pwcorr
```

	Happin~s	Exercise	Sleep	Jobsat~n	Pets
Happiness	1.0000				
Exercise	0.6056	1.0000			
Sleep	-0.1952	-0.4974	1.0000		
Jobsatisfac~n	0.8601	0.7312	0.0246	1.0000	
Pets	0.6590	0.7897	-0.4082	0.5847	1.0000

Método de Máxima Verossimilhança

- Função de log verossimilhança tem expressão

$$\log L(\alpha, \beta) = \sum_{i=1}^n y_i[\alpha + \beta x_{1i}] - \sum_{i=1}^n \log[1 + e^{\alpha + \beta x_{1i}}]$$

- A matriz hessiana da função $\log L(\alpha, \beta)$ é dada pelas segundas derivadas da função de log verossimilhança

$$\text{Hessian}(\alpha, \beta) = \begin{bmatrix} \frac{\partial^2}{\partial \alpha^2} \log L(\alpha, \beta) & \frac{\partial^2}{\partial \alpha \partial \beta} \log L(\alpha, \beta) \\ \frac{\partial^2}{\partial \alpha \partial \beta} \log L(\alpha, \beta) & \frac{\partial^2}{\partial \beta^2} \log L(\alpha, \beta) \end{bmatrix}$$

- A partir da matriz hessiana, podemos obter os erros padrões das estimativas dos parâmetros
- A matriz de variância-covariância das estimativas $\hat{\alpha}$ e $\hat{\beta}$ corresponde à inversa da matriz hessiana

$$\Sigma = -[\text{Hessian}(\alpha, \beta)]^{-1}$$

- Os erros padrões correspondem às raízes quadradas da diagonal principal da matriz Σ

Método de Máxima Verossimilhança

- Função de log verossimilhança tem expressão
- A matriz hessiana da função é dada pelas segundas derivadas da função de log verossimilhança

$$\text{Hessian}(\alpha, \beta) = \begin{bmatrix} \frac{\partial^2}{\partial \alpha^2} \log L(\alpha, \beta) & \frac{\partial^2}{\partial \alpha \partial \beta} \log L(\alpha, \beta) \\ \frac{\partial^2}{\partial \alpha \partial \beta} \log L(\alpha, \beta) & \frac{\partial^2}{\partial \beta^2} \log L(\alpha, \beta) \end{bmatrix}$$

- A partir da matriz hessiana, podemos obter os erros padrões das estimativas dos parâmetros
- A matriz de variância-covariância das estimativas $\hat{\alpha}$ e $\hat{\beta}$ corresponde à inversa da matriz hessiana
- Os erros padrões correspondem às raízes quadradas da diagonal principal da matriz

Regressão Logística no R

`vcov(mod1)`

```
> vcov(mod1)
```

	(Intercept)	renda_per_capita
(Intercept)	0.0336458958	-1.094468e-04
renda_per_capita	-0.0001094468	3.787026e-07

`sqrt(diag(vcov(mod1)))`

```
> sqrt(diag(vcov(mod1)))
```

(Intercept)	renda_per_capita
0.1834281762	0.0006153882

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.2070406	0.1834282	28.39	<2e-16 ***
renda_per_capita	-0.0182626	0.0006154	-29.68	<2e-16 ***

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

Para mais detalhes, vide Carvalho, Cajueiro e Camargo, “Introdução aos Métodos Estatísticos para Economia e Finanças”

Algoritmo Fisher-Scoring

- Considere a função de log verossimilhança, para uma regressão mais geral, com expressão

$$\begin{aligned} & \log L(\beta_0, \beta_1, \dots, \beta_k) \\ &= \sum_{i=1}^n y_i [\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}] - \sum_{i=1}^n \log[1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}] \end{aligned}$$

- O estimador de máxima verossimilhança busca encontrar o vector de coeficientes que maximiza a função acima
- Não é possível encontrar uma fórmula fechada para o vetor de coeficientes estimados $\beta_0, \beta_1, \dots, \beta_k$
- Vamos agora ilustrar o processo de estimação via máxima verossimilhança, através de um método iterativo
- Método comumente utilizado é o método de Newton-Raphson. No caso de máxima verossimilhança, utiliza-se o método correlato, conhecido como Fisher-Scoring
- Esses métodos consistem em darmos um “chute” inicial para o vetor $(\beta_0, \beta_1, \dots, \beta_k)$
- Com base nesse chute inicial, a cada iteração, nós encontramos um novo valor do vetor $(\beta_0, \beta_1, \dots, \beta_k)$, cada vez mais próximo do vetor correspondente ao máximo

Algoritmo Fisher-Scoring

- Considere a função de log verossimilhança, para uma regressão mais geral, com expressão
- O estimador de máxima verossimilhança busca encontrar o vetor de coeficientes que maximiza a função acima
- Não é possível encontrar uma fórmula fechada para o vetor de coeficientes estimados
- Vamos agora ilustrar o processo de estimação via máxima verossimilhança, através de um método iterativo
- Método comumente utilizado é o método de Newton-Raphson. No caso de máxima verossimilhança, utiliza-se o método correlato, conhecido como Fisher-Scoring
- Esses métodos consistem em darmos um “chute” inicial para o vetor
- Com base nesse chute inicial, a cada iteração, nós encontramos um novo valor do vetor , cada vez mais próximo do vetor correspondente ao máximo

Algoritmo de Fisher-Scoring

- O algoritmo para quando novas iterações não implicam em mudanças no vetor $(\beta_0, \beta_1, \dots, \beta_k)$; nesse caso, dizemos que o algoritmo convergiu
- O item mais importante dos algoritmos iterativos é o passo de atualização do parâmetro sendo estimado
- Seja então o vetor $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ o vetor de coeficientes de interesse. No caso do algoritmo de Fisher-Scoring, o passo de atualização tem expressão:

$$\beta^{m+1} = \beta^m + H^{-1}(\beta^m) \times V(\beta^m)$$

- β^{m+1} é o novo vetor, atualizado no passo m
- β^m é o vetor do passo anterior
- O vetor $V(\beta^m)$ corresponde ao vetor de primeiras derivadas da função $\log L(\beta_0, \beta_1, \dots, \beta_k)$
- A matriz $H^{-1}(\beta^m)$ corresponde à inversa da matriz hessiana $H(\beta^m)$ da função de log-verossimilhança, conforme vimos anteriormente
- Portanto, é importante calcular o vetor de derivadas $V(\beta^m)$ e a matriz de segundas derivadas $H(\beta^m)$

Algoritmo de Fisher-Scoring

- O algoritmo para quando novas iterações não implicam em mudanças no vetor ; nesse caso, dizemos que o algoritmo convergiu
- O item mais importante dos algoritmos iterativos é o passo de atualização do parâmetro sendo estimado
- Seja então o vetor o vetor de coeficientes de interesse. No caso do algoritmo de Fisher-Scoring, o passo de atualização tem expressão:
 - $\beta^{(m)}$ é o novo vetor, atualizado no passo m
 - $\beta^{(m-1)}$ é o vetor do passo anterior
 - O vetor $\beta^{(m)}$ corresponde ao vetor de primeiras derivadas da função
 - A matriz $H^{(m)}$ corresponde à inversa da matriz hessiana da função de log-verossimilhança, conforme vimos anteriormente
 - Portanto, é importante calcular o vetor de derivadas e a matriz de segundas derivadas

Algoritmo de Fisher-Scoring

- Pode-se mostrar que o vetor de derivadas $V(\beta)$ é dado por:

$$V(\beta) = \begin{bmatrix} \frac{\partial}{\partial \beta_0} \log L(\beta_0, \beta_1, \dots, \beta_k) \\ \frac{\partial}{\partial \beta_1} \log L(\beta_0, \beta_1, \dots, \beta_k) \\ \dots \\ \frac{\partial}{\partial \beta_k} \log L(\beta_0, \beta_1, \dots, \beta_k) \end{bmatrix} = X^T y - X^T p = X^T (y - p)$$

- Onde p é a probabilidade predita,

$$p = \begin{bmatrix} \frac{e^{\beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \dots + \beta_k x_{k1}}}{1 + e^{\beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \dots + \beta_k x_{k1}}} \\ \dots \\ \dots \\ \frac{e^{\beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \dots + \beta_k x_{kn}}}{1 + e^{\beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \dots + \beta_k x_{kn}}} \end{bmatrix}$$

- X^T corresponde à matriz transporte da matriz de desenho X
- Uma fórmula semelhante pode ser derivada para a matriz hessiana $H(\beta^m)$

Algoritmo de Fisher-Scoring

- Pode-se mostrar que o vetor de derivadas é dado por:
- Onde \hat{p}_i é a probabilidade predita,
- \mathbf{J} corresponde à matriz transporte da matriz de desenho
- Uma fórmula semelhante pode ser derivada para a matriz hessiana

Algoritmo de Fisher-Scoring

- Utilizando o R, vamos ilustrar o algoritmo de Fisher-Scoring, com o passo de atualização com expressão:

$$\beta^{m+1} = \beta^m + H^{-1}(\beta^m) \times V(\beta^m)$$

- O algoritmo visa chegar a um ponto β^{m+1} para o qual o vetor de primeiras derivadas $V(\beta^{m+1})$ é igual a zero (ou arbitrariamente próximo a zero)
- O algoritmo para quando encontramos um valor de norma de $V(\beta^{m+1})$ menor do que um valor arbitrário, ou quando β^{m+1} e β^m são muito próximos
- Vamos usar o critério de parada quando $|V(\beta^{m+1})| < 1e-8$, por exemplo
- Podemos contar quantos passos o algoritmo faz até chegar à convergência, com base no nosso critério de parada
- Ao final do nosso processo iterativo, podemos comparar os resultados obtidos com os resultados obtidos utilizando-se a função do pacote em R
 - `summary(modsimul)`

Algoritmo de Fisher-Scoring

- Utilizando o R, vamos ilustrar o algoritmo de Fisher-Scoring, com o passo de atualização com expressão:
- O algoritmo visa chegar a um ponto para o qual o vetor de primeiras derivadas é igual a zero (ou arbitrariamente próximo a zero)
- O algoritmo para quando encontramos um valor de norma de β menor do que um valor arbitrário, ou quando β e β_{ant} são muito próximos
- Vamos usar o critério de parada quando $\|\beta - \beta_{\text{ant}}\| < 1e-8$, por exemplo
- Podemos contar quantos passos o algoritmo faz até chegar à convergência, com base no nosso critério de parada
- Ao final do nosso processo iterativo, podemos comparar os resultados obtidos com os resultados obtidos utilizando-se a função do pacote em R
 - `summary(modsimul)`

Simulações de Monte Carlo

- Vamos agora estudar as propriedades do estimador de máxima verossimilhança via simulações de Monte Carlo
- As simulações de Monte Carlo consistem em repetirmos o processo observacional aleatório, e estimar, para cada amostra gerada, os parâmetros do modelo
- Vamos fazer de conta que conhecemos o parâmetro real do modelo, e vamos gerar amostras com base nesse parâmetro escolhido
- Com base nas amostras geradas, vamos estimar o modelo de regressão logística, e checar se o parâmetro estimado se aproxima do parâmetro ‘real’ (que é desconhecido na prática)
- Podemos então estudar algumas perguntas importantes:
 - O parâmetro estimado se aproxima do parâmetro “real”?
 - Qual a variabilidade das estimativas em torno do parâmetro estimado?
 - Como essa variabilidade se altera quando aumentamos o tamanho da amostra?
 - Qual a distribuição aproximada das estimativas para diferentes amostras?
 - O que significam então os intervalos de confiança?
 - Podemos analisar o comportamento das estatísticas teste?

Teste da Razão de Verossimilhança

- Para testar vários parâmetros ao mesmo tempo, o teste mais comumente empregado é o teste da razão de verossimilhança, ou likelihood ratio test (LRT)
- Vamos supor que queremos testar a hipótese nula conjunta:

$$H_0: \beta_2 = \beta_5 = 0$$

$$H_A: \beta_2 \neq 0 \text{ ou } \beta_5 \neq 0$$

- O teste de razão verossimilhança tem como estatística teste simplesmente a diferença

$$LRT = 2 \times [\log L(\beta) - \log L(\beta | \beta_2 = \beta_5 = 0)]$$

- $\log L(\beta)$ é o log-verossimilhança (no máximo) para o modelo sem restrição
- $\log L(\beta | \beta_2 = \beta_5 = 0)$ é o log-verossimilhança (no máximo), para o modelo com restrição, dada pela hipótese nula. Nesse caso, a restrição corresponde a simplesmente excluirmos as variáveis x_2 e x_5 da regressão
- Qual a distribuição aproximada para essa estatística teste, assumindo que a hipótese nula é verdadeira (ou seja, $\beta_2 = \beta_5 = 0$)

Teste da Razão de Verossimilhança

- Para testar vários parâmetros ao mesmo tempo, o teste mais comumente empregado é o teste da razão de verossimilhança, ou likelihood ratio test (LRT)
- Vamos supor que queremos testar a hipótese nula conjunta:
 - O teste de razão verossimilhança tem como estatística teste simplesmente a diferença
 - é o log-verossimilhança (no máximo) para o modelo sem restrição
 - é o log-verossimilhança (no máximo), para o modelo com restrição, dada pela hipótese nula. Nesse caso, a restrição corresponde a simplesmente excluirmos as variáveis x_1 e x_2 da regressão
 - Qual a distribuição aproximada para essa estatística teste, assumindo que a hipótese nula é verdadeira (ou seja,)

Teste da Razão de Verossimilhança

- Intrinsecamente relacionado à estatística de log-likelihood está a estatística *Deviance*
- Essa estatística é dada pelo output da regressão, e tem expressão

$$Deviance = -2 \times \log L(\beta)$$

- Portanto,

$$\begin{aligned} LRT &= 2 \times [\log L(\beta) - \log L(\beta | \beta_2 = \beta_5 = 0)] \\ &= - [\text{Deviance}_{\text{irrest}} - \text{Deviance}_{\text{rest}}] \end{aligned}$$

- Com $\text{Deviance}_{\text{irrest}}$ e $\text{Deviance}_{\text{rest}}$ correspondendo aos modelos irrestrito e restrito
- Quando a hipótese nula é verdadeira, ou seja, $\beta_2 = \beta_5 = 0$, a estatística teste LRT tem distribuição qui-quadrada, com número de graus de liberdade igual ao número de restrições no modelo
- Para duas restrições, o valor crítico da estatística teste é dado por `valor_critico_5pc <- qchisq(0.95, 2) = 5.991465`, para 5% de probabilidade de erro do tipo I

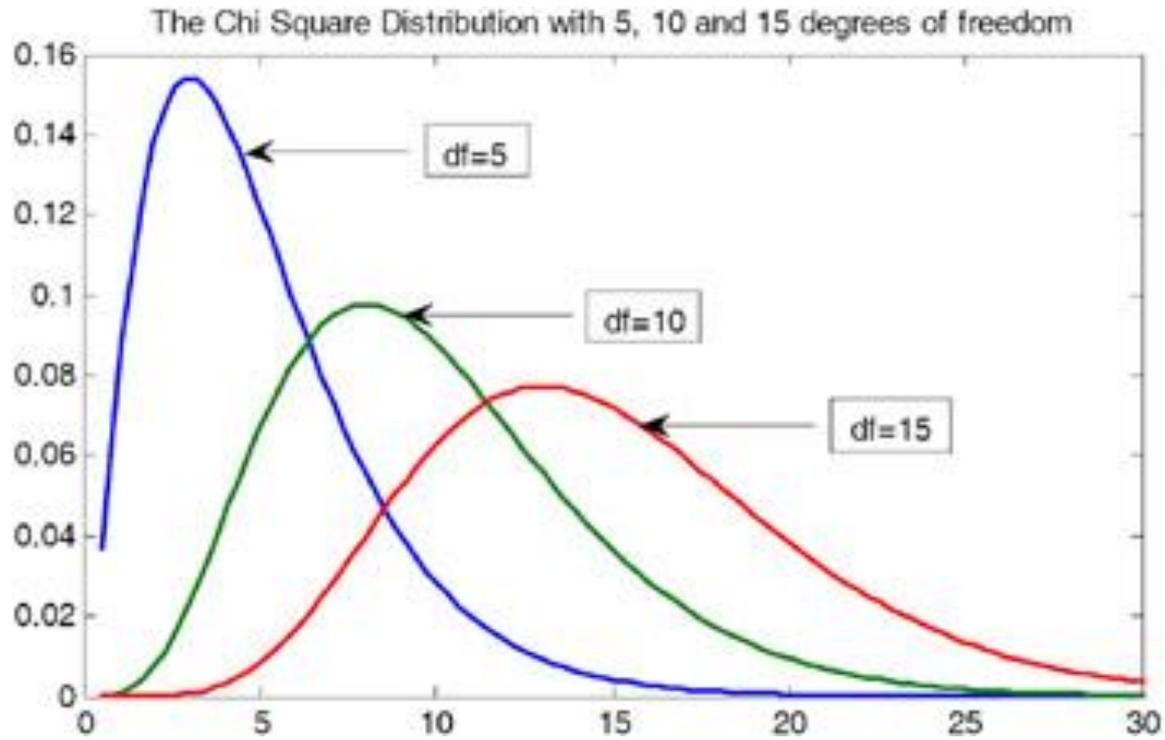
Teste da Razão de Verossimilhança

- Intrinsecamente relacionado à estatística de log-likelihood está a estatística *Deviance*
- Essa estatística é dada pelo output da regressão, e tem expressão

- Portanto,

- Com l_0 e l_1 correspondendo aos modelos irrestrito e restrito
- Quando a hipótese nula é verdadeira, ou seja, H_0 , a estatística teste LRT tem distribuição qui-quadrada, com número de graus de liberdade igual ao número de restrições no modelo
- Para duas restrições, o valor crítico da estatística teste é dado por `valor_critico_5pc <- qchisq(0.95, 2) = 5.991465`, para 5% de probabilidade de erro do tipo I

Teste da Razão de Verossimilhança



R² para Regressão Logística

- Em regressão linear, uma medida comumente utilizada para verificar o ajuste do modelo é o coeficiente de determinação
- No caso de regressão logística, há várias alternativas para o equivalente ao R² da regressão linear
- McFadden's R²: $R^2_{McF} = 1 - \ln(L_M) / \ln(L_0)$, onde $\ln(L_0)$ é função de log-verossimilhança, para um modelo com apenas o intercepto
- Nagelkerke / Cragg & Uhler's:

$$R^2_{C\&U} = \frac{1 - \left[\frac{L_0}{L_M}\right]^{\frac{2}{n}}}{1 - L_0^{2/n}}, \text{ com } 0 \leq R^2_{C\&U} \leq 1$$

- Cox & Snell (maximum likelihood):

$$R^2_{C\&S} = 1 - \left[\frac{L_0}{L_M}\right]^{\frac{2}{n}}$$

- No caso de Cox & Snell, o valor máximo não é 1. A interpretação dos pseudo-R² não são tão simples quando do R² no caso linear

R² para Regressão Logística

- Em regressão linear, uma medida comumente utilizada para verificar o ajuste do modelo é o coeficiente de determinação
- No caso de regressão logística, há várias alternativas para o equivalente ao R² da regressão linear
- McFadden's R²: $R^2_{MCF} = 1 - \ln(L_M) / \ln(L_0)$, onde $\ln(L_0)$ é função de log-verossimilhança, para um modelo com apenas o intercepto
- Nagelkerke / Cragg & Uhler's:
- Cox & Snell (maximum likelihood):
- No caso de Cox & Snell, o valor máximo não é 1. A interpretação dos pseudo-R² não são tão simples quando do R² no caso linear

Seleção de Variáveis

- Da mesma maneira que no caso de regressão linear, podemos usar os indicadores AIC e BIC para seleção de modelos
- Dentre vários modelos, podemos selecionar aquele (ou aqueles) com menor AIC ou BIC
- Critério de Informação de Akaike - AIC

$$AIC = -2 \log L(\beta_0, \beta_1, \dots, \beta_k) + 2 \times p$$

O número p corresponde ao número de parâmetros livres na regressão. No caso da regressão logística, temos: um intercepto e k variáveis preditoras

$$p = 1 + k = 1 + k$$

- Critério de Informação Bayesiano - BIC

$$BIC = -2 \log L(\beta_0, \beta_1, \dots, \beta_k) + \log n \times p$$

- Os termos $[2 \times p]$ e $[\log n \times p]$, no AIC e BIC, correspondem a pênaltis para a inclusão adicional de variáveis
- Portanto, a inclusão de variáveis vai aumentar $\log L(\beta_0, \beta_1, \dots, \beta_k)$, mas aumenta também os pênaltis $[2 \times p]$ e $[\log n \times p]$
- Como de costume, o BIC tende a selecionar modelos mais parcimoniosos

Seleção de Variáveis

- Da mesma maneira que no caso de regressão linear, podemos usar os indicadores AIC e BIC para seleção de modelos
- Dentre vários modelos, podemos selecionar aquele (ou aqueles) com menor AIC ou BIC
- Critério de Informação de Akaike - AIC

O número p corresponde ao número de parâmetros livres na regressão. No caso da regressão logística, temos: um intercepto e k variáveis preditoras

- Critério de Informação Bayesiano - BIC
- Os termos e^{-2} e e^{-2} , no AIC e BIC, correspondem a pênaltis para a inclusão adicional de variáveis
- Portanto, a inclusão de variáveis vai aumentar R^2 , mas aumenta também os pênaltis e^{-2} e e^{-2}
- Como de costume, o BIC tende a selecionar modelos mais parcimoniosos

Regressão Logística no R

```
> summary(mod1)
```

Call:

```
glm(formula = alta_mort_infantil ~ renda_per_capita, family = binomial(link = "logit"),  
     data = dados3)
```

Deviance Residuals:

```
   Min     1Q  Median     3Q      Max  
-2.4659 -0.2831 -0.0536 -0.0003  3.1928
```

Coefficients:

```
                Estimate Std. Error z value Pr(>|z|)  
(Intercept)      5.2070406  0.1834282  28.39 <2e-16 ***  
renda_per_capita -0.0182626  0.0006154 -29.68 <2e-16 ***
```

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 6163.7 on 5563 degrees of freedom
Residual deviance: 3042.4 on 5562 degrees of freedom
AIC: 3046.4

Number of Fisher Scoring iterations: 7

O Modelo de Regressão Logística

- De maneira geral, várias das técnicas que nós vimos em regressão linear se aplicam também à regressão logística
 - Testes de hipótese para parâmetros individuais
 - Intervalos de confiança
 - Seleção de modelos
 - Testes de hipótese para vários parâmetros simultaneamente
- A syntax correspondente em R também é bastante similar ao que vimos para o caso de regressão linear
- Exemplos:

```
confint(mod5)          #--- probabilidade de cobertura de 95%  
confint(mod5, level = 0.9) #--- probabilidade de cobertura de 90%  
confint(mod5, level = 0.8) #--- probabilidade de cobertura de 80%
```

```
AIC(mod5); BIC(mod5)
```

```
anova(mod5.rest, mod5, test='LRT')
```

```
step1 <- step(mod5, direction = "backward")  
step2 <- step(mod5, direction = "forward")  
step3 <- step(mod5, direction = "both")
```

Interpretação dos Coeficientes da Reg Logística

- O método de máxima verossimilhança nos dá os coeficientes estimados para o modelo de regressão logística
- No entanto, precisamos entender qual o significado desses coeficientes. Como interpretá-los? Sabemos interpretar os sinais dos coeficientes, e precisamos agora entender a magnitude
- *Odds ratio* ou “razão de chances”: o Bahia tem chance de 3 contra 1 de vencer o campeonato baiano. Nesse caso, a probabilidade do o Bahia ganhar é de $3/(3+1) = 75\%$
- Por outro lado, dado que o Bahia tem 75% de chance de vencer, a razão de chances é $0.75/0.25 = 3$ contra 1
- Para a regressão logística, a razão de chances de sucesso versus insucesso (1 versus 0) é dada pela razão das probabilidades

$$\frac{p_i}{1 - p_i} = \frac{\frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}}}{1 - \left[\frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}} \right]} = \frac{\frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}}}{\frac{1}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}}}$$

$$\frac{p_i}{1 - p_i} = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}$$

Interpretação dos Coeficientes da Reg Logística

- O método de máxima verossimilhança nos dá os coeficientes estimados para o modelo de regressão logística
- No entanto, precisamos entender qual o significado desses coeficientes. Como interpretá-los? Sabemos interpretar os sinais dos coeficientes, e precisamos agora entender a magnitude
- *Odds ratio* ou “razão de chances”: o Bahia tem chance de 3 contra 1 de vencer o campeonato baiano. Nesse caso, a probabilidade do o Bahia ganhar é de $3/(3+1) = 75\%$
- Por outro lado, dado que o Bahia tem 75% de chance de vencer, a razão de chances é $0.75/0.25 = 3$ contra 1
- Para a regressão logística, a razão de chances de sucesso versus insucesso (1 versus 0) é dada pela razão das probabilidades

Interpretação dos Coeficientes da Reg Logística

- Portanto, para uma regressão logística, a razão de chances para a observação i é dada por

$$r_i = \frac{p_i}{1 - p_i} = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}$$

- Imagine agora que a variável x_{1i} teve um incremento de uma unidade. A nova razão de chances vai ser

$$r_i^* = e^{\beta_0 + \beta_1 [1 + x_{1i}] + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}$$

$$r_i^* = e^{\beta_1} e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}$$

$$r_i^* = e^{\beta_1} r_i$$

- Portanto, e^{β_1} indica o aumento (ou redução) da razão de chances quando aumentamos em uma unidade a variável x_{1i}
- Se x_{1i} for uma variável dummy indicando se o paciente teve um tratamento ou não, o termo e^{β_1} indica o quanto a razão de chances se altera quando o paciente passa pelo tratamento (versus quando ele não passa)
- A maioria dos softwares estatísticos reporta os termos e^{β_1} para todos os coeficientes no modelo. É possível também extrair intervalos de confiança para e^{β_1}

Interpretação dos Coeficientes da Reg Logística

- Portanto, para uma regressão logística, a razão de chances para a observação i é dada por
- Imagine agora que a variável X_i teve um incremento de uma unidade. A nova razão de chances vai ser
- Portanto, e^{β_j} indica o aumento (ou redução) da razão de chances quando aumentamos em uma unidade a variável X_j
- Se X_j for uma variável dummy indicando se o paciente teve um tratamento ou não, o termo e^{β_j} indica o quanto a razão de chances se altera quando o paciente passa pelo tratamento (versus quando ele não passa)
- A maioria dos softwares estatísticos reporta os termos e^{β_j} para todos os coeficientes no modelo. É possível também extrair intervalos de confiança para

Interpretação dos Coeficientes da Reg Logística

- O código abaixo calcula os valores para e^{β_1} , com os respectivos intervalos de confiança, com 95% de probabilidade de cobertura

#--- odds-ratio, com intervalos de confiança de 95%

```
mod5.reduzido <- glm(formula = alta_mort_infantil ~ renda_per_capita
+ indice_gini
+ salario_medio_mensal
+ perc_crianças_extrem_pobres
+ perc_crianças_pobres
+ perc_pessoas_dom_agua_estogo_inadequados
+ perc_pessoas_dom_paredes_inadequadas
+ perc_pop_dom_com_coleta_lixo
+ perc_pop_rural
+ as.factor(Regiao),
family = binomial(link = "logit"), data = dados3)
summary(mod5.reduzido)

data.frame(exp(coef(mod5.reduzido)), exp(confint(mod5.reduzido)))
```

Interpretação dos Coeficientes da Reg Logística

- O código abaixo calcula os valores para β , com os respectivos intervalos de confiança, com 95% de probabilidade de cobertura

#---- odds-ratio, com intervalos de confiança de 95%

```
mod5.reduzido <- glm(formula = alta_mort_infantil ~ renda_per_capita
+ indice_gini
+ salario_medio_mensal
+ perc_crianças_extrem_pobres
+ perc_crianças_pobres
+ perc_pessoas_dom_agua_estogo_inadequados
+ perc_pessoas_dom_paredes_inadequadas
+ perc_pop_dom_com_coleta_lixo
+ perc_pop_rural
+ as.factor(Regiao),
family = binomial(link = "logit"), data = dados3)
summary(mod5.reduzido)

data.frame(exp(coef(mod5.reduzido)), exp(confint(mod5.reduzido)))
```

Classificação com Regressão Logística

- Ao final da estimação da regressão logística, a regressão vai nos fornecer as probabilidades preditas de uma determinada observação ter valor 1 (“sucesso”) ou 0 (“insucesso”)
- No entanto, em muitas situações, gostaríamos de classificar aquela observação como 0 ou 1, com base nas variáveis preditoras
- Por exemplo, com base nas características de um cliente no banco, gostaríamos de classificá-lo como potencial pagador ou potencial inadimplente
- Com base em um valor definido de corte c , uma das formas de fazer isso é através da regra:
 - Caso a probabilidade predita $p_i > c$, então a observação i é classificada na categoria “sucesso”
 - Caso a probabilidade predita $p_i \leq c$, então a observação i é classificada na categoria “insucesso”
- Por exemplo, podemos assumir $c = 0.50$
- Como averiguar quão boa ou ruim é essa regra de classificação? Quais os indicadores usados para fazer essa averiguação?

Classificação com Regressão Logística

- Ao final da estimação da regressão logística, a regressão vai nos fornecer as probabilidades preditas de uma determinada observação ter valor 1 (“sucesso”) ou 0 (“insucesso”)
- No entanto, em muitas situações, gostaríamos de classificar aquela observação como 0 ou 1, com base nas variáveis preditoras
- Por exemplo, com base nas características de um cliente no banco, gostaríamos de classificá-lo como potencial pagador ou potencial inadimplente
- Com base em um valor definido de corte , uma das formas de fazer isso é através da regra:

Caso a probabilidade predita \geq τ , então a observação i é classificada na categoria “sucesso”

Caso a probabilidade predita $<$ τ , então a observação i é classificada na categoria “insucesso”

- Por exemplo, podemos assumir $\tau = 0.50$
- Como averiguar quão boa ou ruim é essa regra de classificação? Quais os indicadores usados para fazer essa averiguação?

Classificação com Regressão Logística

- Normalmente, as medidas de qualidade da classificação estão relacionadas ao grau de acerto das classificações
- Por exemplo, um indicador comumente empregado é a chamada matriz de “confusão” (*confusion matrix*)
- Essa matriz corresponde a uma tabulação cruzada entre a classificação de acordo com o algoritmo e a classificação real observada na amostra

	Classificação 0 observada	Classificação 1 observada
Classificação 0 predita	Verdadeiro negativo	Falso negativo
Classificação 1 predita	Falso positivo	Verdadeiro positivo

- Com base na matriz de confusão, várias métricas de desempenho podem ser calculadas: acurácia, precisão, recall e score F-1

Classificação com Regressão Logística

```
#-----  
#--- classificação com regressão logit  
#-----
```

```
mod6 <- glm(formula = formula(step3),  
            family = binomial(link = "logit"), data = dados3)  
summary(mod6)
```

```
corte <- 0.5
```

```
dados3$pred_alta_mortalidade <- ifelse(mod6$fitted.values > 0.5, 1, 0)
```

```
#--- matriz de comparação da classificação
```

```
table(dados3$pred_alta_mortalidade)  
table(dados3$pred_alta_mortalidade, dados3$alta_mort_infantil)
```

```
> table(dados3$pred_alta_mortalidade, dados3$alta_mort_infantil)
```

	0	1
0	3815	188
1	400	1161

Classificação com Regressão Logística

- Acurácia: corresponde ao percentual de casos que são corretamente classificados

$$\text{Acurácia} = \frac{[\text{positivos verdadeiros} + \text{negativos verdadeiros}]}{n}$$

- Precisão, para a classe C: corresponde ao percentual de observações na classe C que foram corretamente classificadas. Portanto, há um valor de precisão para cada classe (0 ou 1)

$$\text{Precisão}_1 = \frac{[\text{positivos verdadeiros}]}{[\text{positivos verdadeiros} + \text{falsos negativos}]} = \text{Sensitivity}$$

$$\text{Precisão}_0 = \frac{[\text{negativos verdadeiros}]}{[\text{negativos verdadeiros} + \text{falsos positivos}]} = \text{Specificity}$$

- Recall, para a classe C: corresponde ao percentual de previsões da classe C que foram corretamente classificadas. Portanto, há um valor de recall para cada classe (0 ou 1)

$$\text{Recall}_1 = \frac{[\text{positivos verdadeiros}]}{[\text{positivos verdadeiros} + \text{falsos positivos}]}$$

$$\text{Recall}_0 = \frac{[\text{negativos verdadeiros}]}{[\text{negativos verdadeiros} + \text{falsos negativos}]}$$

Classificação com Regressão Logística

- Acurácia: corresponde ao percentual de casos que são corretamente classificados
- Precisão, para a classe C: corresponde ao percentual de observações na classe C que foram corretamente classificadas. Portanto, há um valor de precisão para cada classe (0 ou 1)
- Recall, para a classe C: corresponde ao percentual de previsões da classe C que foram corretamente classificadas. Portanto, há um valor de recall para cada classe (0 ou 1)

Classificação com Regressão Logística

- Score F-1, para a classe C: corresponde a uma combinação entre precisão e recall. Também há um valor de score F-1 para cada classe (0 ou 1)

$$scoreF1_1 = \frac{2 \times Precisão_1 \times Recall_1}{(Precisão_1 + Recall_1)}$$

$$scoreF1_0 = \frac{2 \times Precisão_0 \times Recall_0}{(Precisão_0 + Recall_0)}$$

- Podemos combinar as precisões, os recalls e os scores F-1, para obter medidas gerais, para a classificação como um todo:
 - Para isso, podemos fazer as médias das precisões, dos recalls e dos scores F-1, de todas as classes
- As medidas acima podem ser utilizadas em problemas mais gerais de classificação, nos quais podemos querer classificar em mais de duas classes
- No caso de classificação binária, podemos estar interessados mais diretamente na precisão, no recall e no score F-1 da classe 1 (“sucesso”)

Classificação com Regressão Logística

- Score F-1, para a classe C: corresponde a uma combinação entre precisão e recall. Também há um valor de score F-1 para cada classe (0 ou 1)
- Podemos combinar as precisões, os recalls e os scores F-1, para obter medidas gerais, para a classificação como um todo:
 - Para isso, podemos fazer as médias das precisões, dos recalls e dos scores F-1, de todas as classes
- As medidas acima podem ser utilizadas em problemas mais gerais de classificação, nos quais podemos querer classificar em mais de duas classes
- No caso de classificação binária, podemos estar interessados mais diretamente na precisão, no recall e no score F-1 da classe 1 (“sucesso”)

Classificação com Regressão Logística

```
cmatrix <- table(dados3$pred_alta_mortalidade, dados3$alta_mort_infantil)
cmatrix
```

```
acuracia <- sum(diag(cmatrix))/sum(cmatrix)
acuracia
```

```
precisao <- diag(cmatrix) / colSums(cmatrix)
precisao
```

```
recall <- diag(cmatrix) / rowSums(cmatrix)
recall
```

```
scoreF1 <- 2 * precisao * recall / (precisao + recall)
scoreF1
```

```
resultados.class <- data.frame(precisao, recall, scoreF1)
resultados.class
```

```
macroPrecisao <- mean(precisao)
macroPrecisao
```

```
macroRecall <- mean(recall)
macroRecall
```

```
macroScoreF1 <- mean(scoreF1)
macroScoreF1
```

```
data.frame(macroPrecisao, macroRecall, macroScoreF1)
```

Modelos de Regressão Logística

- **Exercício 8:**
- Utilize como base o código em R 'Análise_de_Regressao_com_Variaveis_Binarias'
 - Questão 1: Considere o modelo de regressão logística abaixo. Com valores de corte $c = 0.3, 0.5$ e 0.7 , encontre as matrizes de "confusão" para a classificação de municípios com alta mortalidade infantil. Para esses mesmos valores de corte, encontre de acurácia, média das precisões, média dos recalls e média dos scores F-1

```
mod5.reduzido <- glm(formula = alta_mort_infantil ~ renda_per_capita
+ indice_gini
+ salario_medio_mensal
+ perc_crianças_extrem_pobres
+ perc_crianças_pobres
+ perc_pessoas_dom_agua_estogo_inadequados
+ perc_pessoas_dom_paredes_inadequadas
+ perc_pop_dom_com_coleta_lixo
+ perc_pop_rural
+ as.factor(Regiao),
family = binomial(link = "logit"), data = dados3)
summary(mod5.reduzido)
```

Modelos de Regressão Logística

- **Exercício 8 (continuação):**
- Utilize como base o código em R 'Análise_de_Regressao_com_Variaveis_Binarias'
 - Questão 2: Refaça a questão 1, considerando o modelo de regressão logística abaixo. Qual dos dois modelos (questão 1 ou questão 2) apresenta melhor critério de acurácia?

```
mod5.reduzido <- glm(formula = alta_mort_infantil ~ renda_per_capita
+ indice_gini
+ salario_medio_mensal
+ perc_crianças_extrem_pobres
+ perc_crianças_pobres
+ perc_pessoas_dom_agua_estogo_inadequados
+ perc_pessoas_dom_paredes_inadequadas
+ perc_pop_dom_com_coleta_lixo
+ perc_pop_rural
+ as.factor(Regiao)
+ as.factor(Regiao)*renda_per_capita,
family = binomial(link = "logit"), data = dados3)
summary(mod5.reduzido)
```

Curva ROC

- Uma das formas de avaliar a performance de classificação a partir de uma regressão logística é utilizando-se a curva ROC (*Receiver Operating Characteristic*)
- Lembrando a tabela de comparação entre observado e predito em classificação:

	Classificação 0 observada	Classificação 1 observada
Classificação 0 predita	Verdadeiro negativo	Falso negativo
Classificação 1 predita	Falso positivo	Verdadeiro positivo

- À medida que aumentamos o valor de corte, nós classificamos menos observações na categoria 1; portanto, aumentamos o número de falsos negativos e reduzimos o número de falsos positivos
- Duas medidas muito utilizadas para fins de classificação e detecção de doenças, por exemplo, é a sensibilidade (*sensitivity*) e a especificidade (*specificity*)

Curva ROC

- Lembrando:

$$\frac{[\text{positivos verdadeiros}]}{[\text{positivos verdadeiros} + \text{falsos negativos}]} = \text{Sensitivity}$$

$$\frac{[\text{negativos verdadeiros}]}{[\text{negativos verdadeiros} + \text{falsos positivos}]} = \text{Specificity}$$

- *Sensitivity* é a capacidade de detectar pacientes com câncer, dentre aqueles que de fato possuem câncer, por exemplo
- *Specificity* é a capacidade de detectar pacientes que não possuem câncer, dentre aqueles que de fato não possuem
- Quando aumentamos o valor de corte, aumentamos a sensibilidade e reduzimos a especificidade
- A curva ROC nos fornece um gráfico da sensibilidade versus a especificidade, quando aumentamos o valor de corte
- A curva sob a curva, conhecida com AUC (area under the curve), é usada como uma medida de qualidade do ajuste da regressão logística

Curva ROC

- Lembrando:
- *Sensitivity* é a capacidade de detectar pacientes com câncer, dentre aqueles que de fato possuem câncer, por exemplo
- *Specificity* é a capacidade de detectar pacientes que não possuem câncer, dentre aqueles que de fato não possuem
- Quando aumentamos o valor de corte, aumentamos a sensibilidade e reduzimos a especificidade
- A curva ROC nos fornece um gráfico da sensibilidade versus a especificidade, quando aumentamos o valor de corte
- A curva sob a curva, conhecida com AUC (area under the curve), é usada como uma medida de qualidade do ajuste da regressão logística

Curva ROC

