ANÁLISE DE DADOS MULTIVARIADOS I - REGRESSÃO

(AULA 10)

Novembro e dezembro de 2018

Reinaldo Soares de Camargo

Modelos Lineares Generalizados

Modelos lineares Generalizados

- Modelos lineares generalizados (MLGs) são definidos por uma distribuição de probabilidade para a variável resposta Y pertencente à família exponencial, um conjunto de variáveis explicativas que podem ser numéricas ou categóricas e uma função de ligação.

- Um dos modelos lineares generalizados mais utilizados na área de saúde é o modelo de regressão logística binária, onde a variável resposta do modelo tem distribuição de Bernoulli (ou Binomial) e a função de ligação é a função logística. Na área de saúde, o referido modelo poderia ser adotado, por exemplo, para estimar a probabilidade do paciente: aderir ao tratamento medicamentoso (adesão=1; não adesão=0); reportar um estado de saúde não bom (não bom=1; bom=0); ter uma determinada doença crônica (ter DC=1; não ter DC=0).

- A função utilizada para ajustar modelos lineares generalizados no R é a função “glm”. Nesta função é necessário especificar as variáveis explicativas e a variável resposta do modelo, a distribuição de probabilidade da variável resposta do modelo (family) e a função de ligação (link) desejada pelo pesquisador. Com a função “glm” é possível obter as estimativas pontuais dos parâmetros do modelo e algumas medidas de qualidade do ajuste (AIC e deviances).

- Após a ajustar o MLG de interesse é necessário utilizar a função “summary” para obter outros resultados do ajuste do modelo além das estimativas pontuais. Entre os resultados obtidos com a função “summary” do RStudio estão: as estimativas pontuais, os erros padrão referentes as estimativas pontuais, os valores observados da estatística de Wald e os p-valores do teste de Wald.

Modelos Lineares Generalizados

- Na área de saúde, os pesquisadores estão mais interessados em analisar as estimativas das medidas de associação (como, por exemplo, a razão de prevalência ou a razão de chance, em inglês *odds ratio*) ao invés das estimativas pontuais dos parâmetros do modelo. Entretanto, estas medidas de associação não fazem parte do conjunto de resultados fornecidos pela função "summary" do RStudio. O exemplo a seguir mostra como ajustar o modelo de regressão logística binária usando a função "glm", e como obter as medidas de razão de chance e seus respectivos intervalos de confiança a partir das saídas fornecidas pelo comando "glm".

- Os dados se referem a um estudo sobre autoavaliação geral de saúde (1=não boa, 0=boa) de n=30 indivíduos com idade variando de 20 a 95 anos. O objetivo do estudo é estudar a relação entre a autoavaliação de saúde (Y) e as seguintes variáveis explicativas: idade(em anos) e renda familiar per capita (1=Mais de 3 s.m, 0= Até 3 s.m=base).

```
modelo1=glm(saude~idade+renda,family=binomial(link="logit"))
summary(modelo1)
```

Modelos Lineares Generalizados

- estimativas pontuais dos parâmetros e os seus erros padrão, os valores observados da estatística de Wald e os p-valores do teste de Wald, entre outras informações.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.93790    1.74439  -1.684  0.09214 .
idade        0.13296    0.05123   2.595  0.00945 **
renda       -3.17898    1.45863  -2.179  0.02930 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.191  on 29  degrees of freedom
Residual deviance: 18.711  on 27  degrees of freedom
AIC: 24.711

Number of Fisher Scoring iterations: 6
```

- **Medidas de associação (razões de chance).** Pode-se demonstrar matematicamente que a razão de chance é o exponencial da estimativa pontual.

```
> OR1=exp(modelo1$coefficients);OR1
(Intercept)       idade       renda
 0.05297680  1.14220209  0.04162821
```

Modelos Lineares Generalizados

- estimativas pontuais dos parâmetros e os seus erros padrão, os valores observados da estatística de Wald e os p-valores do teste de Wald, entre outras informações.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.93790    1.74439  -1.684  0.09214 .
idade        0.13296    0.05123   2.595  0.00945 **
renda       -3.17898    1.45863  -2.179  0.02930 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.191  on 29  degrees of freedom
Residual deviance: 18.711  on 27  degrees of freedom
AIC: 24.711

Number of Fisher Scoring iterations: 6
```

- **Medidas de associação (razões de chance).** Pode-se demonstrar matematicamente que a razão de chance é o exponencial da estimativa pontual.

```
> OR1=exp(modelo1$coefficients);OR1
(Intercept)       idade       renda
 0.05297680  1.14220209  0.04162821
```

Modelos Lineares Generalizados

- Os intervalos de 95% de confiança para os parâmetros do modelo, com base na estatística de Wald.

```
> ICbeta1=confint.default(modelo1,level=0.95);ICbeta1
                  2.5 %     97.5 %
(Intercept) -6.35684588  0.4810436
idade        0.03255546  0.2333606
renda       -6.03783801 -0.3201166
```

- Os intervalos de 95% de confiança para os parâmetros do modelo, com base na estatística de Wald

```
> ICOR1=exp(ICbeta1);ICOR1
                  2.5 %    97.5 %
(Intercept) 0.001734830 1.6177619
idade       1.033091190 1.2628368
renda       0.002386713 0.7260644
```

- Razão de chance e intervalo de confiança

```
> round((cbind(OR1, ICOR1)),3)
              OR1 2.5 % 97.5 %
(Intercept) 0.053 0.002  1.618
idade       1.142 1.033  1.263
renda       0.042 0.002  0.726
```

Modelos Lineares Generalizados

- Os intervalos de 95% de confiança para os parâmetros do modelo, com base na estatística de Wald.

```
> ICbeta1=confint.default(modelo1,level=0.95);ICbeta1
                  2.5 %     97.5 %
(Intercept) -6.35684588  0.4810436
idade        0.03255546  0.2333606
renda       -6.03783801 -0.3201166
```

- Os intervalos de 95% de confiança para os parâmetros do modelo, com base na estatística de Wald

```
> ICOR1=exp(ICbeta1);ICOR1
                  2.5 %    97.5 %
(Intercept) 0.001734830 1.6177619
idade       1.033091190 1.2628368
renda       0.002386713 0.7260644
```

- Razão de chance e intervalo de confiança

```
> round((cbind(OR1, ICOR1)),3)
              OR1 2.5 % 97.5 %
(Intercept) 0.053 0.002  1.618
idade       1.142 1.033  1.263
renda       0.042 0.002  0.726
```

Modelos Lineares Generalizados

```
> round((cbind(OR1, ICOR1)),3)
              OR1 2.5 % 97.5 %
(Intercept) 0.053 0.002  1.618
idade       1.142 1.033  1.263
renda       0.042 0.002  0.726
```

- **Interpretação das razões de chance (*odds ratio*)**
- Tanto a idade quanto a renda familiar per capita estão significativamente relacionadas com a chance de autoavaliação de saúde não boa (OBS: Note que o p-valor é menor que o nível de significância de 5% e o IC para OR não inclui a unidade).
- A chance do indivíduo reportar um estado de saúde não bom aumenta em 14,2% ao aumentar em 1 ano a idade.

  (1-1,142) = 0,142

- Indivíduos com mais de 3 salários mínimos tem uma chance de reportar um estado de saúde não bom 95,8% menor do que os indivíduos que ganham no máximo 3 salários mínimos.

  (1-0,042) = 0,958.

Introdução à Regressão Logística

Regressão com Resposta Binária

- Considere o modelo de regressão tradicional:

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

- 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

- 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 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

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"?

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"?

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, ... , N}

- A função de frequência de uma variável aleatória binomial tem expressão:

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

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

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

- 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}

O Modelo de Regressão Logística

- Conforme vimos acima, a variável de Bernoulli assume valores 0 ou 1, com probabilidade de sucesso Prob[Y = 1] = *p*, e probabilidade de insucesso 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 $\mathrm{Prob}[y_i = 1] = p_i$

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

O Modelo de Regressão Logística

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

- Uma possível alternativa é assumir

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

- O problema da alternativa acima é que $Prob[y_i = 1] = p_i$ 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:

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

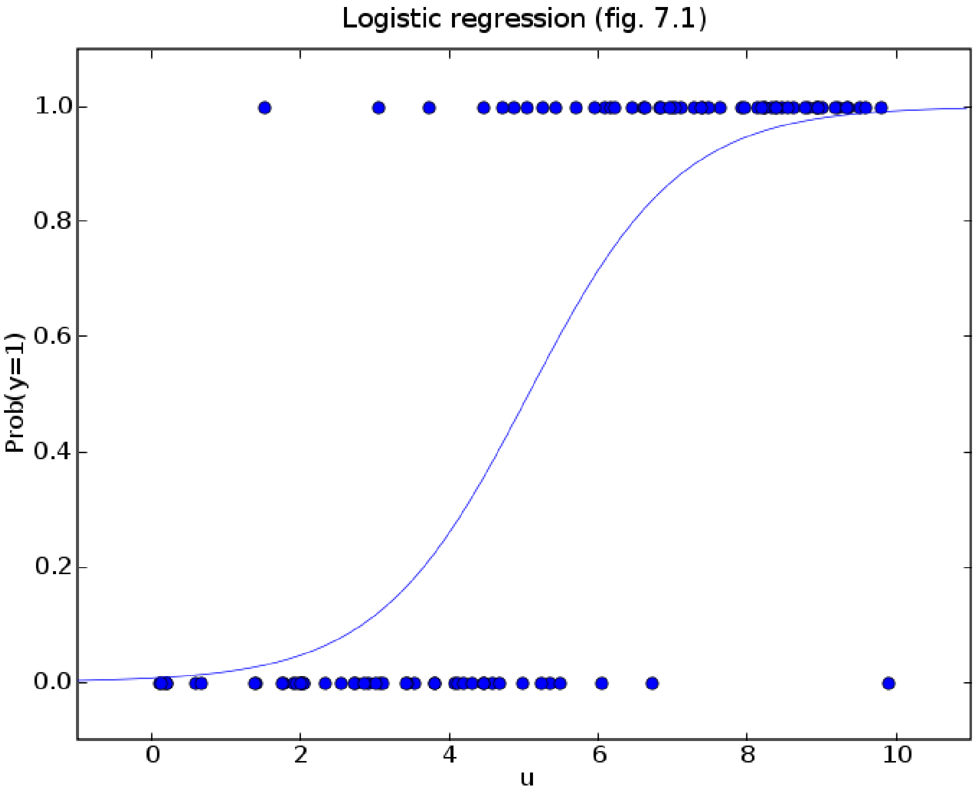
- A fórmula acima implica que as probabilidades p_i vão se situar o intervalor (0,1), como desejado

- Pode-se mostrar que, quando β_1 é positivo, quando x_{1i} aumenta, a probabilidade de "sucesso" também aumenta

O Modelo de Regressão Logística

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

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



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)
```

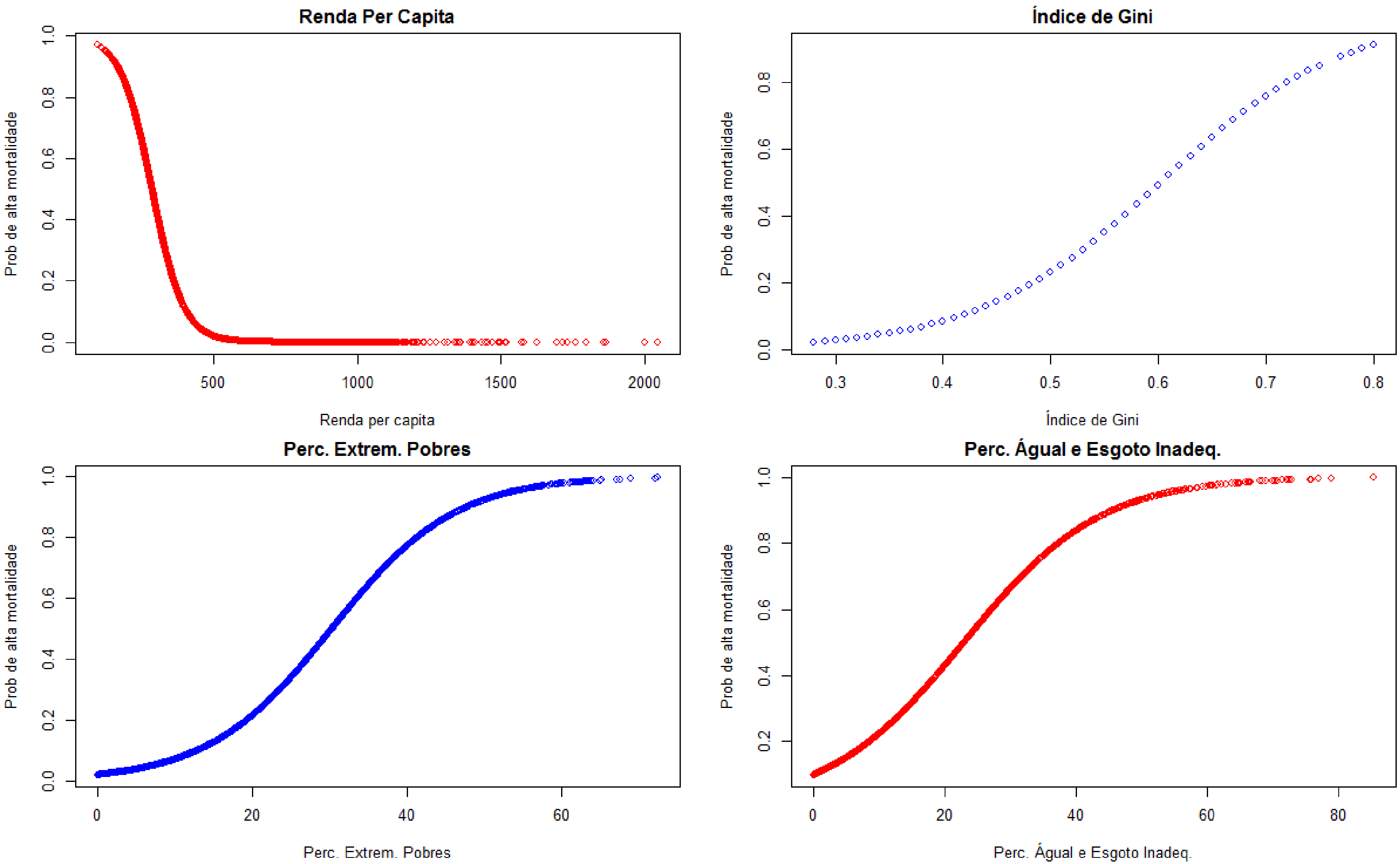
Regressão Logística no R

```
> 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



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

- 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

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

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

- O método de máxima verossilhança 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

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

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

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

- 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

- 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

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

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

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
```

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 excluímos 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

- 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,

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

$$= -[\mathrm{Deviance}_{\mathrm{irrest}} - \mathrm{Deviance}_{\mathrm{rest}}]$$

- Com $\mathrm{Deviance}_{\mathrm{irrest}}$ e $\mathrm{Deviance}_{\mathrm{rest}}$ 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

R^2 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 R^2 da regressão linear

- McFadden's R^2: $R^2{}_{McF} = 1 - \ln(L_M) / \ln(L_0)$, 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-R^2 não são tão simples quando do R^2 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

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
```

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 Palmeiras tem chance de 3 contra 1 de vencer o campeonato paulista. Nesse caso, a probabilidade do o Palmeira ganhar é de 3/(3+1) = 75%

- Por outro lado, dado que o Palmeiras 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

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

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

Interpretação dos Coeficientes da Reg Logística

- As razões de chance (*odds ratio*) são dadas por or = p / (1-p), onde p = probabilidade de sucesso e (1-p) é a probabilidade de fracasso.

- Uma razão de chances de 1 indica que a condição ou evento sob estudo é igualmente provável de ocorrer nos dois grupos. Uma razão de chances maior do que 1 indica que a condição ou evento tem maior probabilidade de ocorrer no primeiro grupo. Finalmente, uma razão de chances menor do que 1 indica que a probabilidade é menor no primeiro grupo do que no segundo.

Interpretação dos Coeficientes da Reg Logística

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

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

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

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

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

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

- 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

- 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_criancas_extrem_pobres
            + perc_criancas_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)))

odds.ratio(mod5.reduzido)   #--- pacote "questionr"
```

Modelos de Regressão

- **4ª Lista de exercícios para entregar em 04/12/2018.**
  - Os exercícios podem ser entregues em grupos de 2 alunos, e o grupo deve submeter o código em R utilizado para responder ao exercício, juntamente com a discussão dos resultados.
  - Utilize a base de dados do IDH brasil 2010 (IDH_Brasil_2010.csv)
  - Rode a regressão logística abaixo:

```
mod5.reduzido <- glm(formula = alta_mort_infantil ~ renda_per_capita
            + indice_gini
            + salario_medio_mensal
            + perc_criancas_extrem_pobres
            + perc_criancas_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)
```

  - Questão 1: Interprete os coeficientes da regressão que apresentem significância estatística;

Modelos de Regressão Logística

- **(continuação):**
  - Questão 2: Refaça a questão 1, considerando o modelo de regressão logística abaixo, entreprete os odds-ratio dos coeficientes que apresentem significância estatística.

```
mod5.reduzido <- glm(formula = alta_mort_infantil ~ renda_per_capita
            + indice_gini
            + salario_medio_mensal
            + perc_criancas_extrem_pobres
            + perc_criancas_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)
```

Questão 3: Com base nos critérios AIC e BIC qual desses modelos seriam selecionados?

Obrigado!