Análise de Dados Categóricos

Modelo de Regressão Logística

Enrico A. Colosimo/UFMG

http://www.est.ufmg.br/~enricoc/

Departamento de Estatística
Universidade Federal de Minas Gerais

Regressão Logística Binária

1. Característica Básica: RESPOSTA BINÁRIA

2. OBJETIVO:
   - Identificar Fatores de Risco ou Prognóstico;
   - Comparar duas ou mais populações, ajustando por fatores de confusão;
   - Predição.

3. Referências Bibliográficas:
   - Applied Logistic Regression, Hosmer & Lemeshow (2000);
   - Introdução à Análise de Dados Categóricos com Aplicações, Giolo (2017);
   - Modelling Binary Data, Collett (1991);
   - Statistical Methods in cancer Research, vols I (1980) e II (1987), Breslow & Day.

Exemplo

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Probability of Cancer in Pulmonary Nodules Detected on First Screening CT

Annette McWilliams, M.B., Martin C. Tammemagi, Ph.D., John R. Mayo, M.D., Heidi Roberts, M.D., Geoffrey Liu, M.D., Kam Soghrati, M.D., Kazuhiro Yasufuku, M.D., Ph.D., Simon Martel, M.D., Francis Laberge, M.D., Michel Gingras, M.D., Sukhinder Atkar-Khattra, B.Sc., Christine D. Berg, M.D., Ken Evans, M.D., Richard Finley, M.D., John Yee, M.D., John English, M.D., Paola Nasute, M.D., John Goffin, M.D., Serge Puksa, M.D., Lori Stewart, M.D., Scott Tsai, M.D., Michael R. Johnston, M.D., Daria Manos, M.D., Garth Nicholas, M.D., Glenwood D. Goss, M.D., Jean M. Seely, M.D., Kayvan Amjadi, M.D., Alain Tremblay, M.D.C.M., Paul Burrowes, M.D., Paul MacEachern, M.D., Rick Bhatia, M.D., Ming-Sound Tsao, M.D., and Stephen Lam, M.D.

From Vancouver General Hospital (A.M., J.R.M., K.E., R.F., J.Y., J.E., S.L.) and the British Columbia Cancer Agency (A.M., S.A.-K., S.L.), Vancouver, BC; the Department of Community Health Sciences, Brock University, St. Catharines, ON (M.C.T.); University Health Network–Princess Margaret Cancer Centre and Toronto General Hospital, Toronto (H.R., G.L., K.S., K.Y., M.-S.T.); Juravinski Hospital and Cancer Center, Hamilton, ON (J.G., S.P., L.S., S.T.), Ottawa Hospital Cancer Centre, Ottawa (G.N., G.D.G., J.M.S., K.A.); Institut Universitaire de Cardiologie et de Pneumologie de Québec, Quebec City, QC (S.M., F.L., M.G.); Dalhousie University, Halifax, NS (M.R.J., D.M.); University of Calgary, Calgary, AB (A.T., P.B., P.M.); and Memorial University of Newfoundland, St. John's (R.B.) — all in Canada; the National Cancer Institute, National Institutes of Health, Bethesda, MD (C.D.B.); and Hospital Universitario Austral, Pilar, Buenos Aires (P.N.). Address reprint requests to Dr. Lam at the Department of Integrative Oncology, British Columbia Cancer Agency, 675 W. 10th Ave., Vancouver, BC V5Z 1L3, Canada.

N Engl J Med 2013;369:910-9.
DOI: 10.1056/NEJMoa1214726

ABSTRACT

BACKGROUND

Major issues in the implementation of screening for lung cancer by means of low-dose computed tomography (CT) are the definition of a positive result and the management of lung nodules detected on the scans. We conducted a population-based prospective study to determine factors predicting the probability that lung nodules detected on the first screening low-dose CT scans are malignant or will be found to be malignant on follow-up.

METHODS

We analyzed data from two cohorts of participants undergoing low-dose CT screening. The development data set included participants in the Pan-Canadian Early Detection of Lung Cancer Study (PanCan). The validation data set included participants involved in chemoprevention trials at the British Columbia Cancer Agency (BCCA), sponsored by the U.S. National Cancer Institute. The final outcomes of all nodules of any size that were detected on baseline low-dose CT scans were tracked. Parsimonious and fuller multivariable logistic-regression models were prepared to estimate the probability of lung cancer.

RESULTS

In the PanCan data set, 1871 persons had 7008 nodules, of which 102 were malignant, and in the BCCA data set, 1090 persons had 5021 nodules, of which 42 were malignant. Among persons with nodules, the rates of cancer in the two data sets were 5.5% and 3.7%, respectively. Predictors of cancer in the model included older age, female sex, family history of lung cancer, emphysema, larger nodule size, location of the nodule in the upper lobe, part-solid nodule type, lower nodule count, and spiculation. Our final parsimonious and full models showed excellent discrimination and calibration, with areas under the receiver-operating-characteristic curve of more than 0.90, even for nodules that were 10 mm or smaller in the validation set.

CONCLUSIONS

Predictive tools based on patient and nodule characteristics can be used to accurately estimate the probability that lung nodules detected on baseline screening low-dose CT scans are malignant. (Funded by the Terry Fox Research Institute and others; ClinicalTrials.gov number, NCT00751660.)

Tipos de Estudos

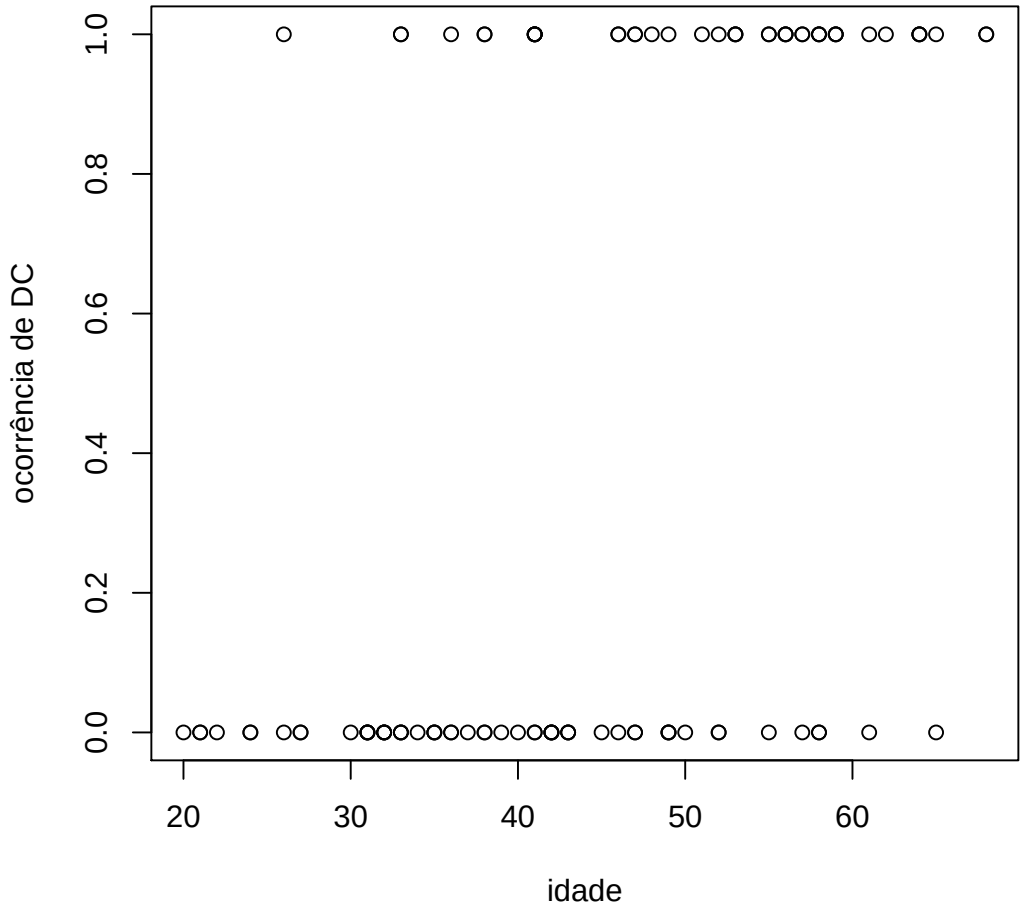
- Estudos Transversais: Regressão Logística usada com frequência.
- Estudos Longitudinais: Regressão Logística pouco ou raramente utilizada nestes desenho.

Exemplo - Texto Profa. Suely - pags. 119-121.

- Uma amostra de 100 pacientes, em que todos tiveram o mesmo período de acompanhamento.
- Resposta: incidência de doença coronariana.
- Resposta para cada indivíduo foi sim (1) ou não (0).
- Covariável de interesse: 8 faixas etárias (idade): 20-29, ..., 60-69.
- Dados aparecem na pag. 120 do livro da Profa. Giolo (2017).
- 43 ocorrências de doença coronariana.

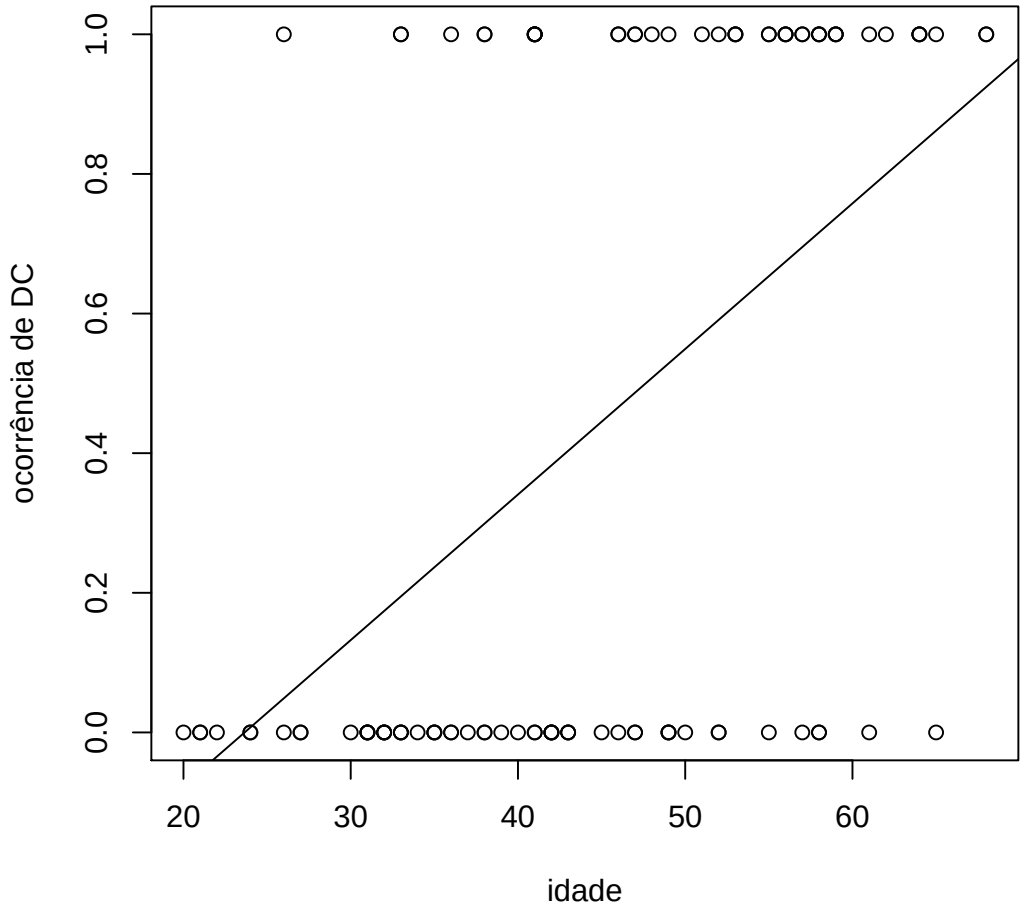
Gráfica de Dispersão

Resposta: 43 casos Covariável: idade contínua.

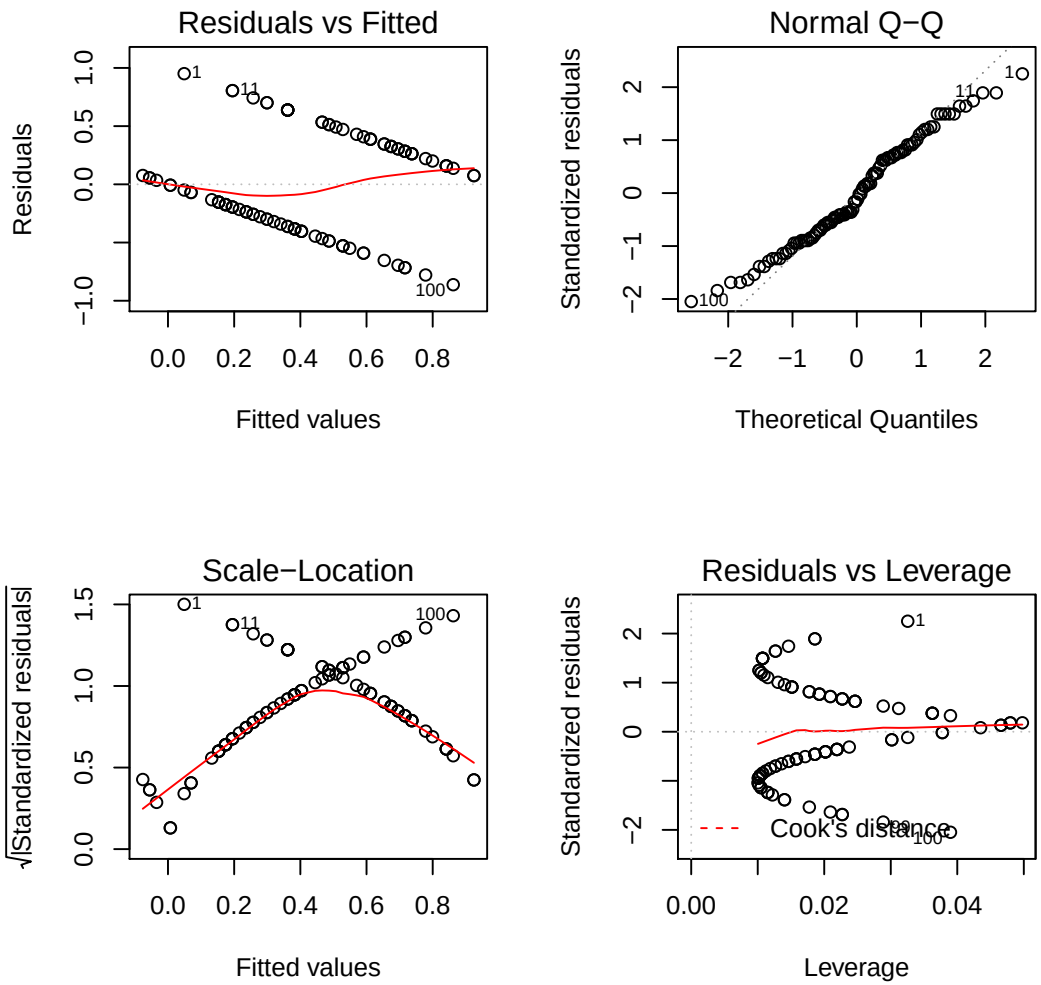


Regressão Linear

Resposta: 43 casos



Regressão Linear - Análise de Resíduos



Regressão Linear - Análise de Resíduos

1. Teste Homocedasticidade: módulo dos resíduos
   valor-p $\approx 0,0158$
2. Teste Normalidade: Shapiro-Wilk
   valor-p $\approx 0,06034$.

Descrever os Dados Agrupados

Faixa Etária	Sim	Não	Prop. DC
20-29 (25)	1	9	0,10
30-34 (32)	2	13	0,13
35-39 (38)	3	9	0,25
40-44 (43)	5	10	0,33
45-49 (47)	6	7	0,46
50-54 (53)	5	3	0,63
55-59 (57)	13	4	0,76
60-69 (65)	8	2	0,80

Entrada dos Dados Grupados

Existem duas formas de entrada dos dados para resposta binária.

- Uma linha para cada indivíduo:

indivíduo	faixa etária	resposta
1	1 (25)	0
.....	..	.
100	5 (47)	1
Total	...	43

Entrada dos Dados

Existem duas formas de entrada dos dados para resposta binária.

- Uma linha para cada combinação de covariáveis.

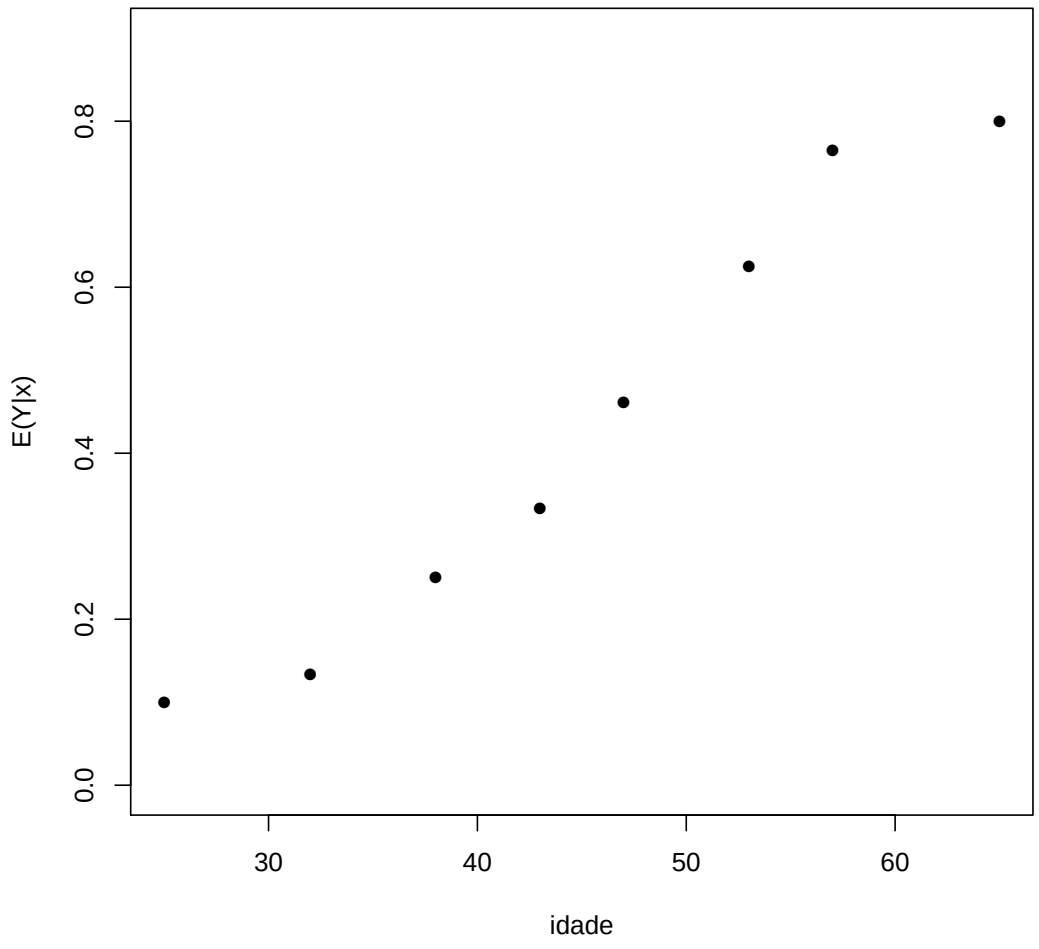
Faixa Etária	Sim	Não
20-29 (25)	1	9
30-34 (32)	2	13
35-39 (38)	3	9
40-44 (43)	5	10
45-49 (47)	6	7
50-54 (53)	5	3
55-59 (57)	13	4
60-69 (65)	8	2

Entrada dos Dados

Existem duas formas de entrada dos dados para resposta binária.

- Na presença de observações Bernoulli, somente é possível entrar com os dados da primeira forma: uma linha para cada indivíduo.
- Este fato sempre ocorre na presença de covariáveis contínuas.
- Quando for possível entrar com os dados das duas formas, **deve-se sempre preferir a segunda: uma linha para cada combinação de covariáveis.**
- Neste último caso, somente é possível realizar o teste de adequação do modelo (será visto adiante) nesta segunda forma de entrar com os dados.

Descrição Gráfica por Faixa Etária



Regressão Logística Binária

Em resumo: Porque não usar o modelo de regressão linear?

$$\pi(x) = E[Y/X] = P[Y = 1/X]$$

Ou seja, nós queremos modelar a probabilidade de ocorrência de um certo evento.

INCONVENIENTES:

1. Y tem uma distribuição binomial;
2. $Var(Y) \propto E[Y]$;
3. $0 \leq P[Y = 1/X] \leq 1$.

MLG: Regressão Logística Binária

1. Y : $0/1$: distribuição Bernoulli (binomial) pertence a família exponencial.
2. $X\beta$: preditor linear.
3. função de ligação logit (canônica).

$$g(E(Y/X)) = g(P(Y = 1/X)) = log\frac{P(Y = 1/X)}{1 - P(Y = 1/X)} = X\beta$$

Regressão Logística Binária

Funções de Distribuições (inverso da função de ligação):

- $\pi(x) = \frac{\exp(x)}{1+\exp(x)}$ (logit, canônica)
- $\pi(x) = \Phi(x)$ (probit)
- $\pi(x) = \exp -(\exp(1-x))$ (complemento log-log)

Caso mais simples: somente uma covariável.

$$E(Y/X) = \pi(x) = P[Y = 1/x] = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Transformação LOGIT - Função de ligação.

$$\text{logit}(\pi(x)) = log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

Regressão Logística Binária

1. Forma do Modelo
   - Logit (Regressão Logística).
   - Probit.
   - Complemento log log.
2. Inferência para β
   - Função de Verossimilhança;
   - Propriedades dos Estimadores;
   - Estatísticas de Teste (Wald e RV)
3. Técnicas de Adequação do modelo.
4. Interpretação do modelo (razão de chances)
5. Aplicações.
6. Extensões do Modelo de Regressão Binária.

REGRESSÃO LOGÍSTICA MÚLTIPLA

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1 + \exp(-\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}$$

$$logit(x) = logit(E(Y/X)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = X\beta$$

EXEMPLO

- Y: mortalidade infantil;
- X_1: educação da mãe;
- X_2: número de uniões da mãe;
- X_3: região geográfica (urbana ou rural);
- X_4: idade da mãe.

Modelo de Regressão Logística Binomial

Considere uma amostra de tamanho n:

$(y_1, x_1), \ldots, (y_n, x_n) \quad y_i : 0, 1$

$y_i = g^{-1}(x_i\beta) + \varepsilon_i; \quad i = 1, 2, ..., n$

$\pi(x_i) = E(y_i|x_i) = P[y_i = 1|x_i] = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$

$x_i\beta = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$

- Função de Ligação - Logit

  $logit(x_i) = log\frac{\pi(x_i)}{1-\pi(x_i)} = x_i\beta$

- Inferência para β

  $L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}$

Modelo de Regressão Logística Binomial

- Função de Log-Verossimilhança

$$l(\beta) = \sum_{i=1}^{n} y_i log(\pi(x_i)) + (1 - y_i) log(1 - \pi(x_i))$$

- Função Escore

$$U(\beta) = \begin{cases} \sum_{i=1}^{n} [y_i - \pi(x_i)] & \text{Para } \beta_0 \\ \sum_{i=1}^{n} x_{ij}(y_i - \pi(x_i)) & \text{Para } \beta_j; \quad j = 1, \ldots, p \end{cases}$$

EMV: Solução de $U(\beta) = 0$

Modelo de Regressão Logística Binomial

- Matriz de Informação (Observada = Fisher)

$$I(\beta) = \begin{cases} -\frac{\partial^2 l(\beta)}{\partial \beta_j^2} = \sum_{i=1}^{n} x_{ij}^2 (\pi(x_i))(1 - \pi(x_i)) \\ -\frac{\partial^2 l(\beta)}{\partial \beta_j \beta_l} = \sum_{i=1}^{n} x_{ij} x_{il} \pi(x_i)(1 - \pi(x_i)) \end{cases}$$

Modelo de Regressão Logística Binomial

Estatísticas Assintóticas

- Wald

  $\hat{\beta} \rightarrow N(\beta, I^{-1}(\beta))$

- TRV

  $TRV = -2log[\frac{L(\beta)}{L(\hat{\beta})}] \sim \chi^2_{p+1}$

Estatísticas relacionadas ao EMV

1. WALD

$$\widehat{\beta} \approx N(\beta, I^{-1}(\beta))$$

ou

$$W = (\widehat{\beta} - \beta)' I(\beta)(\widehat{\beta} - \beta)$$

2. RAZÃO DE VEROSSIMILHANÇA

$$-2\ log(L(\beta)/L(\widehat{\beta})) = 2(l(\widehat{\beta}) - l(\beta))$$

3. Escore (Rao)

$$S = U(\beta)' I^{-1}(\beta) U(\beta)$$

Observações

- Resultados empíricos mostram que a estatística S é a melhor das três seguida pela RV.
- A estatística Escore não depende de $\widehat{\beta}$ (estimador irrestrito).
- A estatística RV não depende de $I(\beta)$: Informação de Fisher.

Resultados Assintóticos

Considere que a dimensão de β é $p+1$. Então:

- as estatísticas podem ser utilizadas para testar hipóteses e construir intervalos de confiança.
- as três estatísticas tem assintoticamente uma distribuição qui-quadrado com $p+1$ (dimensão de β) graus de liberdade.
- $I(\beta)$ deve ser estimada por $I(\widehat{\beta})$;
- frequentemente estamos interessados no teste para um subconjunto de β, $H_0 : \beta_1 = \beta_1^0$ de dimensão $q < p+1$. Neste caso, precisamos encontrar o EMV restrito (sob H_0) $\tilde{\beta}$.

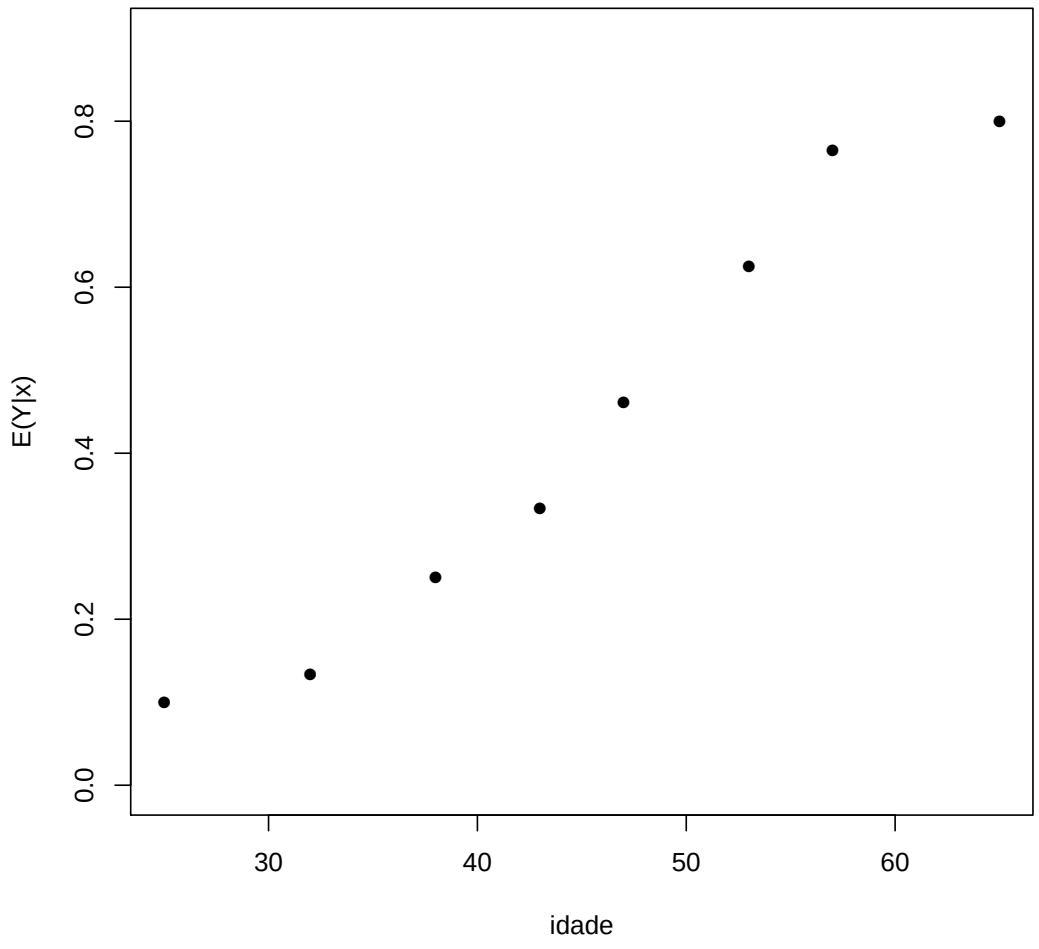
Retornando ao Exemplo - Texto Profa. Suely - pag. 119-121.

- Uma amostra de 100 pacientes, em que todos tiveram o mesmo período de acompanhamento.
- Resposta: incidência de doença coronariana.
- Resposta para cada indivíduo foi sim (1) ou não (0).
- Covariável de interesse: 8 faixas etárias (idade): 20-29, ..., 60-69.
- Dados aparecem na pag. 98 do texto da Profa. Suely.
- 43 ocorrências de doença coronariana.

Banco de Dados

Faixa Etária	Sim	Não
20-29 (25)	1	9
30-34 (32)	2	13
35-39 (38)	3	9
40-44 (43)	5	10
45-49 (47)	6	7
50-54 (53)	5	3
55-59 (57)	13	4
60-69 (65)	8	2

Descrição Gráfica por Faixa Etária



Resultados do Ajuste

```
Coefficients:
             Estimate    Std. Error z value Pr(>|z|)
  (Intercept) -5.12300     1.11111  -4.611 4.01e-06 ***
   idade       0.10578     0.02337   4.527 5.99e-06 ***

Number of Fisher Scoring iterations: 4

> anova(ajust1,test="Chisq")

Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
 NULL                     7    28.7015
 idade  1   28.118         6     0.5838 1.142e-07 ***
```

Resultados do Ajuste

Y: presença ou não de doença coronariana;
X: idade (em anos);
$n = 100$.

Variável	Estimativa	E.P.	Wald
Idade	0,106	0,023	4,53 ($p < 0,001$)
Constante	-5,123	1,11	-4,61 ($p < 0,001$)

$$\widehat{\pi}(x) = \frac{\exp(-5,12 + 0,106 \text{ idade})}{1 + \exp(-5,12 + 0,106 \text{ idade})}$$

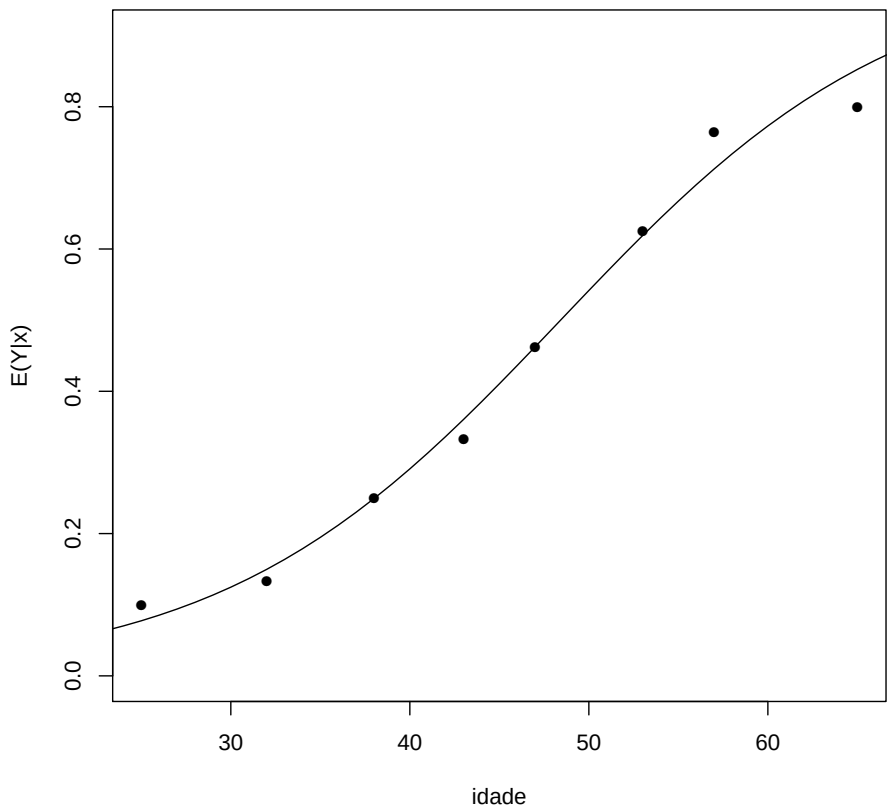
$$\widehat{logit}(x) = -5,12 + 0,106 \text{ idade}$$

Sob $H_0 : \beta_1 = 0$,

TRV = Null Deviance $-$ Residual Deviance $= 28,70 - 0,58 = 28,12. = 5.3^2$

Resultados do Ajuste

Interpretação: Razão de chances = $\exp(0,1058) = 1,11$ (1,06;1,16), isto significa que para o aumento de um ano na idade a chance de doença coronariana aumenta em 11%.



Técnicas de Adequação do Ajuste

- Os resultados do ajuste somente são válidos se o modelo estiver adequado.
- Utilizamos as estatísticas de Pearson e do Desvio para verificar a adequação do modelo ajustado.
- No entanto, estas estatísticas somente têm validade se $N << n$.
  - N: número de possíveis combinações das covariáveis;
  - n: tamanho da amostra
  - No exemplo: $N = 8 << n = 100$
- Na realidade, N não deve aumentar com o aumento de n.
- No caso em que esta suposição é violada, devemos utilizar outro teste de adequação, tipo Hosmer e Lemeshow.

Verificando a Adequação do Ajuste para $N << n$

- Teste para H_0: modelo é adequado.
- Estatística de Pearson;

$$Q_P = \sum_{i=1}^{N} \frac{(y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)}$$

- Estatística do desvio (deviance).

$$Q_D = 2\sum_{i=1}^{N}(y_i \log(y_i/(n_i\widehat{\pi}_i)) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i(1-\widehat{\pi}_i)}))$$

- Q_P e Q_D têm, para grandes amostras, uma distribuição qui-quadrado com N-p-1 graus de liberdade.
- A raiz quadrada dos componentes individuais de Q_p e Q_D são, respectivamente, os resíduos de Pearson e do desvio.

Testes de Adequação do Ajuste

- Desvio:

$$Q_D = 0,5838; \quad valor - p = 0,997 \quad gl : 8 - 2 = 6$$

- Pearson:

$$Q_P = 0,5965; \quad valor - p = 0,996$$

Teste de Hosmer e Lemeshow

- Quando $N \approx n$ os testes de Pearson e do Desvio não podem ser utilizados.
- Hosmer e Lemeshow (1980) propuseram agrupar os dados baseado nas probabilidades estimadas. Usualmente utilizamos no máximo $g = 10$ grupos.
- Ou seja, após ordenarmos as probabilidades estimadas, dividimos em 10 grupos com pontos de cortes nos decis.
- A estatística teste é do tipo Pearson:

$$HL = \sum_{i=1}^{g} \frac{(o_i - n_i\overline{\pi}_i)^2}{n_i(\overline{\pi}_i)(1 - \overline{\pi}_i)}$$

sob H_0 (modelo é adequado), HL tem uma distribuição qui-quadrado com $g - 2$ graus de liberdade.

Observações Importantes

- O gráfico de envelope deve ser utilizado também na avaliação da adequação do modelo.

- A forma funcional de covariáveis contínuas deve ser avaliado através de um gráfico na escala do logit. Ou seja, fazendo um gráfico de

$$logit(x) = \log \frac{P(Y=1)}{1-P(Y=1)} \quad \text{vs} \quad x$$

  Por exemplo: para Idade:
  - Estratificar idade de acordo com o tamanho da amostra;
  - estimar $P(Y=1)$ para cada estrato;
  - fazer o gráfico de logit(x) para cada estrato de idade, use o ponto médio de idade em cada estrato;
  - use o lowess para suavizar o gráfico.

Interpretando os Coeficientes Estimados

1- Regressor Dicotômico

	X=1	X=0
$Y = 1$	$\pi(1) = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$	$\pi(0) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$
$Y = 0$	$1 - \pi(1) = \frac{1}{1+\exp(\beta_0+\beta_1)}$	$1 - \pi(0) = \frac{1}{1+\exp(\beta_0)}$

$$RC = \frac{\pi(1)/1 - \pi(1)}{\pi(0)/1 - \pi(0)} = \exp(\beta_1)$$

Interpretando os Coeficientes Estimados

1- Regressor Dicotômico

EXEMPLO:
Y: doença coronariana

X: sexo

- Feminino: 12 eventos para 33 mulheres;
- Masculino: 30 eventos para 45 homens;
- $\widehat{RC} = \frac{12/21}{30/15} = 0,29$

RESULTADOS: $\widehat{\beta_1} = -1,253$ e portanto $\widehat{RC} = \exp(-1,253) = 0,29$.

INTERPRETAÇÃO: a chance de doença coronariana entre mulheres é cerca de 0,3 vezes à dos homens. Ou, a chance de doença coronariana entre os homens é 3,5 (1/0,29) vezes à das mulheres.

Interpretando os Coeficientes Estimados

2- Regressor Categórico

EXEMPLO:
Y: mortalidade infantil

X: raça da mãe (branca, parda ou preta)

Raça	X_1	X_2
Branca	0	0
Parda	1	0
Preta	0	1

RESULTADOS: $\widehat{\beta_1} = 0,40$, $\widehat{\beta_2} = 1,1$ e portanto
$\widehat{RC_1} = \exp(0,40) = 1,5$ e $\widehat{RC_2} = \exp(1,1) = 3$.

INTERPRETAÇÃO: a chance de mortalidade infantil entre mulheres pardas é cerca de 1,5 vezes à das brancas.

Interpretando os Coeficientes Estimados

3- Regressor Contínuo

Vai depender da forma como o regressor entrou no modelo da unidade de medida.

EXEMPLO:
Y: Mortalidade Infantil
X: idade da mãe (em anos)
RESULTADOS:

$$\widehat{logit}(X) = -1,8 - 0,05X$$

$$\widehat{RC}(1) = \exp(-0,05) = 0,95$$

INTERPRETAÇÃO: Isto indica que a cada aumento de um ano na idade da mãe existe a redução da mortalidade infantil em 5%.

Interpretando os Coeficientes Estimados

INCONVENIENTES Do Regressor Contínuo.

1. **Interpretação em termos de acréscimos de um ano**: pode não ter interesse clínico. Por exemplo, interpretação em termos de acréscimo de quatro anos:

$$\widehat{RC}(4) = \exp(4\widehat{\beta}_1) = 0,82$$

A cada aumento de 4 anos na idade da mãe reduz-se a mortalidade infantil em cerca de 18%.
Um Int. de 95% de confiança é dado por:

$$\exp(c\widehat{\beta}_1 \pm 1,96\ (c)\ E.P.(\widehat{\beta}_1))$$

NO EXEMPLO:

$$\exp(4(-0,05) \pm 1,96(4)(0,015)) \quad \text{ou} \quad (0,73; 0,92).$$

Interpretando os Coeficientes Estimados

INCONVENIENTES Do Regressor Contínuo.

2 O modelo prediz mesma redução redução de mortalidade infantil para:

- comparar uma mãe de 16 anos com outra de 20 anos e
- comparar uma com 36 anos com outra de 40 anos.

Provavelmente, este fato não condiz com a realidade.

Solução: verificar a forma funcional para incluir idade no modelo. Por exemplo, incluir um termo quadrático para idade no preditor linear?

Interpretando os Coeficientes Estimados

4 - Caso Multiparamétrico.

EXEMPLO:

Y: Mortalidade Infantil
X_1: região (urbana e rural)
X_2: educação da mãe (em anos).

	Região	
	rural	urbano
No. mortes	9	4
Média de Ed. Mãe	3	10
No. Crianças	100	100

$$\widehat{RC}(\text{não ajustado por educação}) = \frac{9/91}{4/96} = 2,4$$

Interpretando os Coeficientes Estimados

4 - Caso Multiparamétrico.

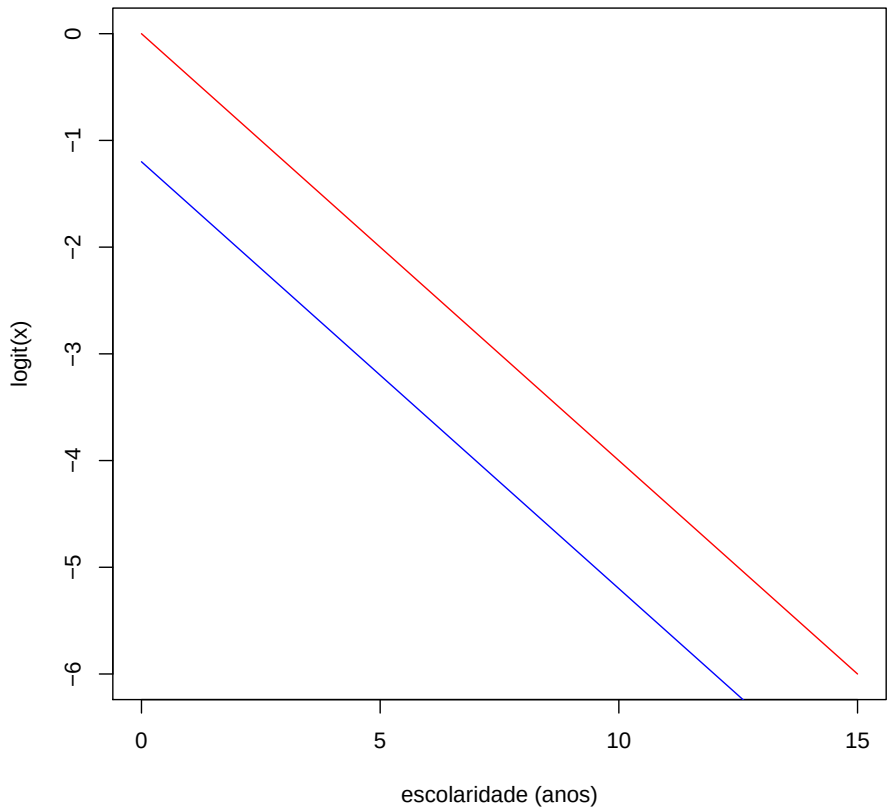
PERGUNTA: Esta diferença é realmente devido aos grupos ou também à educação da mãe?

$$\widehat{RC}(\text{ajustado por educação}) = 1,5$$

Educação da mãe comporta-se como um fator de confusão.

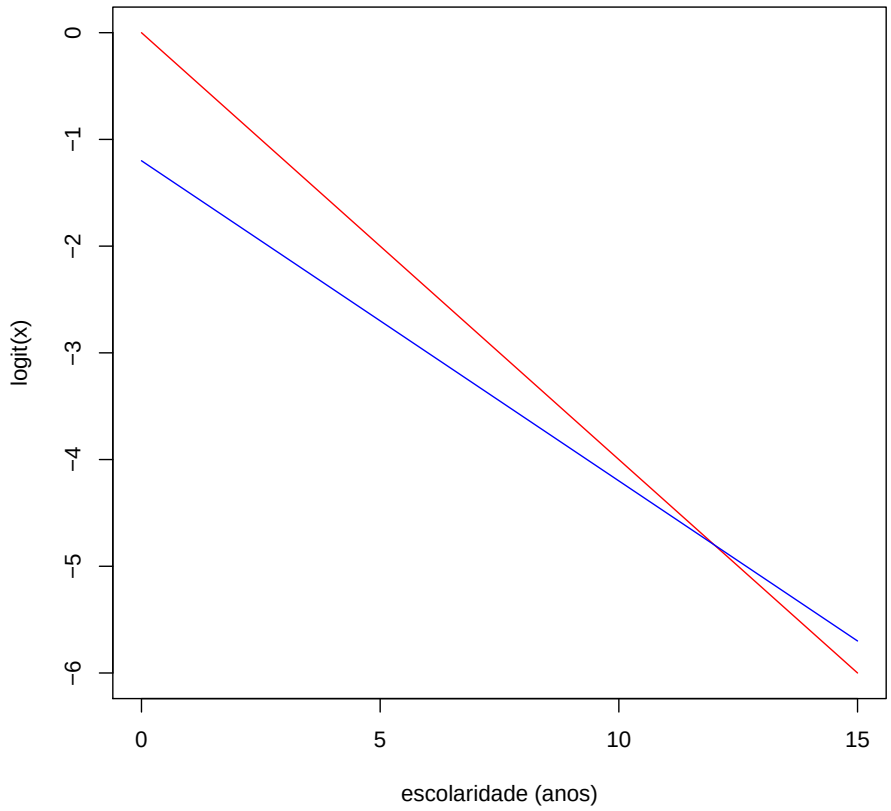
Ideia de Interação

- Sem Interação



- efeito aditivo dos grupos (urbano e rural).

- Com Interação



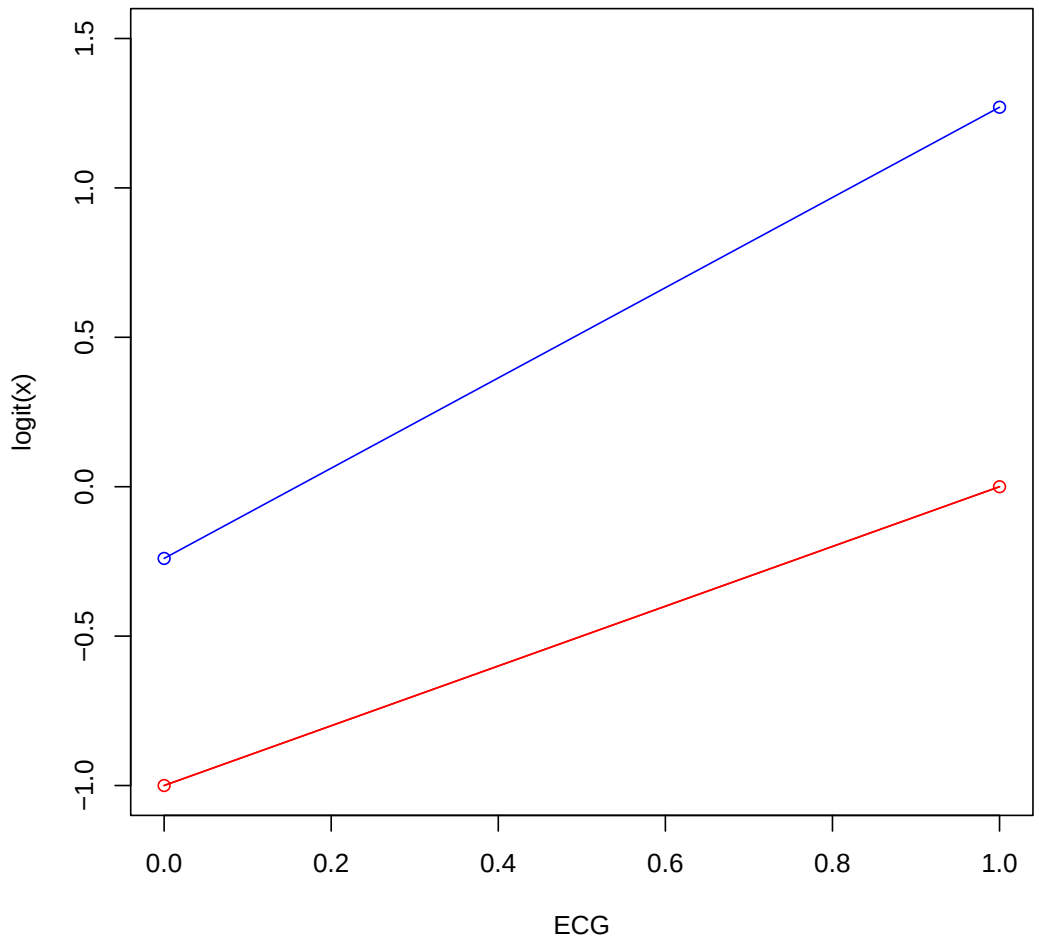
- efeito não aditivo dos grupos (urbano e rural).
- Preditor linear =
  $\beta_0 + \beta_1$ grupo $+ \beta_2$ educação $+ \beta_3$ grupo*educação

Exemplo 1 - seção 7.3.1 - pag. 135.

Y: presença ou não de doença coronariana;
X_1: ECG (0: normal /1: alterado);
X_2: sexo (0: masculino /1: feminino)
$n = 78$.

Covariáveis		Doença		
Gênero	ECG	Sim	Não	Total
Feminino	Normal	4 (27%)	11	15
Feminino	Alterado	8 (44%)	10	18
Masculino	Normal	9 (50%)	9	18
Masculino	alterado	21 (78%)	6	27

Gráfico de Interação



vermelho: feminino e azul: masculino.

Exemplo 1 - seção 7.3.1 - pag. 135.

Variável	Estimativa	E.P.	Wald
sexo	-1,227	0,498	-2,56 ($p = 0,0103$)
ecg	1,055	0,498	2,12 ($p = 0,034$)
Constante	0,102	0,417	0,245 ($p = 0,806$)

$$\widehat{logit}(x) = 0,10 - 1,28 \text{ sexo} + 1,06 \text{ ecg}$$

TRV $(\beta_1 = \beta_2 = 0) = 11,98 - 0,21 = 10,77 \rightarrow$ valor-p $< 0,01$.

Testes de Adequação do Ajuste

- ajust$fitted.values
  1 2 3 4 0.2360103 0.4699914 0.5255469 0.7607465
- ajust$y
  1 2 3 4 0.2666667 0.4444444 0.5000000 0.7777778
- ajust$residuals
  1 2 3 4 0.17002058 -0.10255715 -0.10245520 0.09357272
- Deviance:

$$Q_D = 0.2140933; \quad valor - p = 0.6435778$$

- Pearson:

$$Q_P = 0.2154859; \quad valor - p = 0.6425012$$

Resultados do Ajuste

Interpretação: Razão de chances

- Sexo: RC = $1/\exp(-1,227)$ = 3,6 (1,4; 9,5), isto significa a chance de doença coronariana entre os homens é 3,6 vezes a chance entre as mulheres.

- ECG: RC = $\exp(1,054)$ = 2,9 (1,1; 7,6), isto significa a chance de doença coronariana entre aqueles com ECG alterado é cerca de 3 vezes a chance entre os com ECG normal.

Aplicação: Mini Avaliação Nutricional (MAN) para Idosos

1. Motivação

   - A desnutrição é uma condição que se inicia com o baixo consumo de nutrientes podendo evoluir para estados mais graves;
   - Desafios para os geriatras é identificar os idosos que necessitam de uma intervenção dietética.
   - O ideal seria que uma avaliação nutricional completa mas fica restrita aos seus custos e ao tempo demandado para tal.
   - A Mini Avaliação Nutricional (MAN) foi desenvolvida com o objetivo de proporcionar um rápido diagnóstico do estado nutricional.

2. O escore final da MAN classifica: ≤ 24 - risco de desnutrição, > 24 - bem nutrido.

Aplicação: Mini Avaliação Nutricional (MAN) para Idosos

3. Descrição do Estudo: Um estudo transversal foi conduzido na FM da UFMG com 33 idosos para verificar a relação entre as variáveis bioquímicas (hemoglobina e ferritina) e antropométricas (ângulo de fase e percentual de gordura corporal) e o escore obtido por meio da aplicação da Mini Avaliação Nutricional.

4. Objetivo: avaliar se a MAN seria uma boa ferramenta para predizer alterações bioquímicas e antropométricas características da desnutrição.

5. Covariáveis:

   - Categóricas: gênero;
   - Contínuas: idade (anos), ângulo de fase (o), percentual de gordura (%), hemoglobina (g/dl) e ferritina (ng/ml)

MAN para Idosos: Análise Descritiva

```
>  summary(dados)
   MAN             Sexo        Idade           Angulo           GC
 <24 :15   Feminino :16   Min.   :60.00   Min.   :4.400   Min.   :17.23
 >=24:18   Masculino:17   1st Qu.:68.00   1st Qu.:4.830   1st Qu.:21.44
                          Median :75.00   Median :5.500   Median :27.25
                          Mean   :73.88   Mean   :5.785   Mean   :26.85
                          3rd Qu.:78.00   3rd Qu.:6.160   3rd Qu.:31.50
                          Max.   :90.00   Max.   :8.820   Max.   :35.12
       Hb          Ferritina
 Min.   : 9.3   Min.   : 10.10
 1st Qu.:12.0   1st Qu.: 34.50
 Median :12.7   Median : 62.90
 Mean   :13.0   Mean   : 67.42
 3rd Qu.:14.1   3rd Qu.: 99.40
 Max.   :17.1   Max.   :175.10
```

MAN para Idosos: Multicolinearidade

```
>   cor(dados[3:7])
                Idade       Angulo           GC         Hb  Ferritina
Idade       1.0000000 -0.482143419  0.146190678 -0.3446945 -0.2352079
Angulo     -0.4821434  1.000000000  0.004429318  0.3683612  0.2243951
GC          0.1461907  0.004429318  1.000000000 -0.5056622 -0.1695677
Hb         -0.3446945  0.368361172 -0.505662175  1.0000000  0.2268356
Ferritina -0.2352079  0.224395065 -0.169567725  0.2268356  1.0000000
```

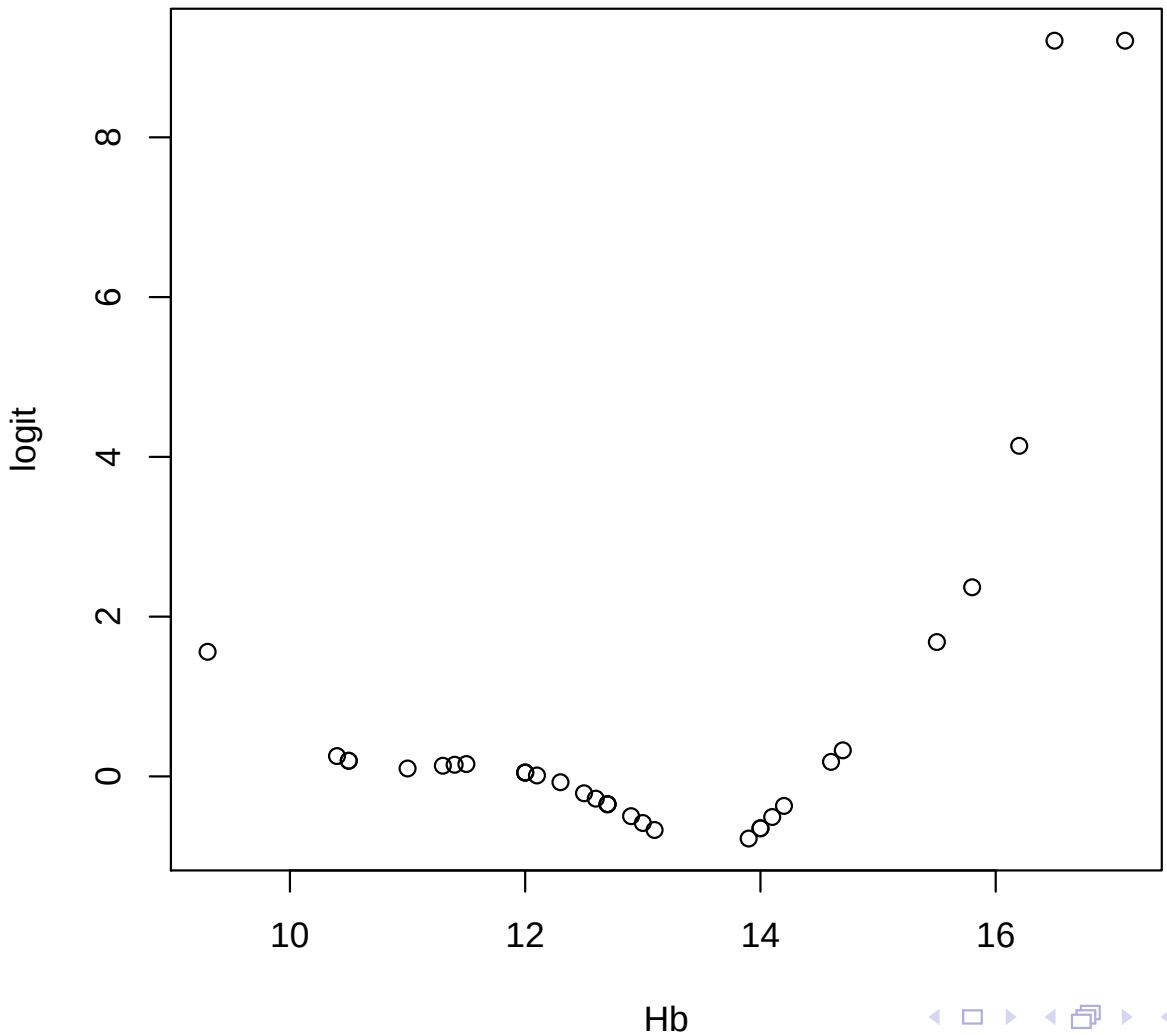
MAN para Idosos: Análise Bivariada

```
> summary(glm(MAN ~ Angulo, family = binomial , data = dados))

Call: glm(formula = MAN ~ Angulo, family = binomial(link = "logit"),
    data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4776  -1.1098   0.4329   1.0747   1.4811

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.8729     2.2198  -1.745   0.0810 .
     Angulo  0.7152      0.3954 1.809    0.0705 .
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45.475  on 32  degrees of freedom
Residual deviance: 40.968  on 31  degrees of freedom AIC: 44.968
```

MAN para Idosos: Análise Bivariada

```
>  anova(ajusteAngulo, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: MAN

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                      32     45.475
Angulo  1 4.5067    31 40.968

---
```

MAN para Idosos: Forma Funcional de Hb



MAN para Idosos: Modelo Final

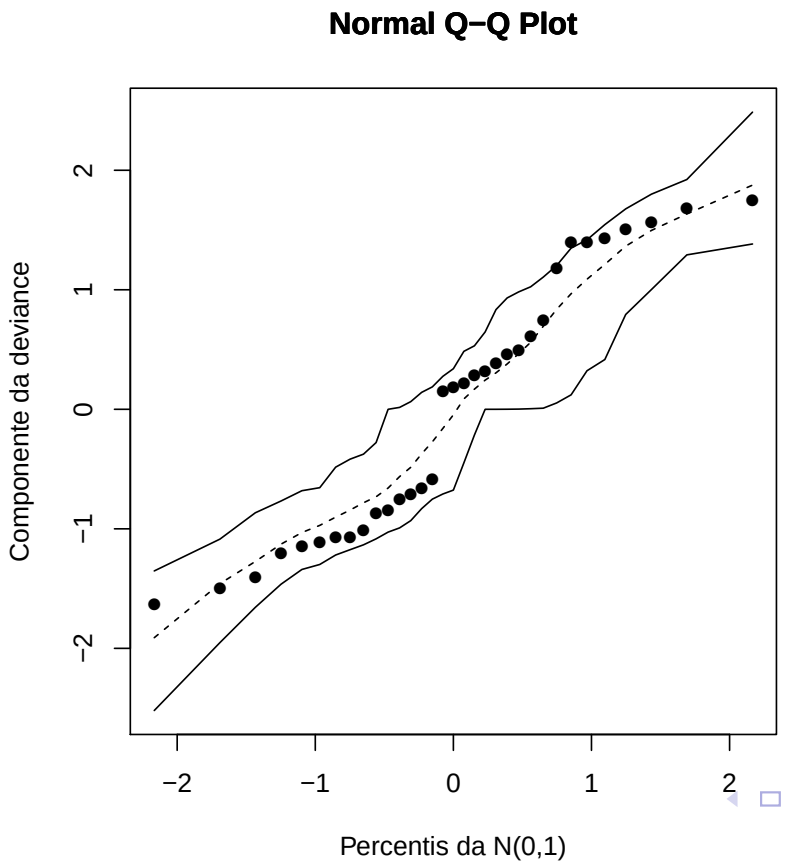
```
>  summary(ajuste4 <- glm(MAN ~ Angulo + Hb + I(Hb^2), family = binomial,

Call: glm(formula = MAN ~ Angulo + Hb + I(Hb^2), family =
binomial(link = "logit"),
    data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4904  -0.9620   0.1757   0.6898   1.6699

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  48.9708    27.0187   1.812   0.0699 .
 Angulo 0.9093 0.5002   1.818                0.0691 .
 Hb           -8.5186 4.3537 -1.957          0.0504 .
 I(Hb^2)       0.3308 0.1713   1.931          0.0534 .
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45.475  on 32  degrees of freedom
Residual deviance: 33.150  on 29  degrees of freedom AIC: 41.15

Number of Fisher Scoring iterations: 6
```

MAN para Idosos: Modelo Final - TRV

```
>  ajuste5<- glm(MAN ~ Angulo, family = binomial(link = "logit"), data = d
>  anova(ajuste5,ajuste4,test="Chisq") # TRV para remoção de Hb
Analysis of Deviance Table

Model 1: MAN ~ Angulo
Model 2: MAN ~ Angulo + Hb + I(Hb^2)

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        31     40.968 2       29     33.150  2   7.8181  0.02006 *


>  ajuste6 <- glm(MAN ~ Hb + I(Hb^2), family = binomial(link = "logit"), d
>  anova(ajuste6, ajuste4,test="Chisq") # TRV para remoção de Angulo
Analysis of Deviance Table

Model 1: MAN ~ Hb + I(Hb^2)
Model 2: MAN ~ Angulo + Hb + I(Hb^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        30      38.14 2       29      33.15  1   4.9904  0.02549 *
---
```

MAN para Idosos: Adequação do modelo

```
>   hosmerlem(y, ajuste4$fitted.values, g = 10) ## Ok
$chisq [1] 9.799756

$p.value [1] 0.2793628

$df [1] 8
```



MAN para Idosos: Interpretação do modelo

- Modelo Final $\rightarrow$ Ângulo + Hb + Hb^2
- Angulo

$$\widehat{RC} = \exp(0.9093) = 2,5$$

A chance de eutrofico (bem nutrido) aumenta em 150% para cada uma unidade de aumento do ângulo de fase.

- Hemoglobina $\rightarrow$ -8,52* Hb + 0,331 Hb^2
- Hb=9/10 $\rightarrow$

$$\widehat{RC} = \exp(-8,52 + 0,331 * (100 - 81)) = 0,11$$

A chance de eutrofico reduz muito ($\exp(0,11)$) ao passar Hb de 9 para 10.
Não faz sentido biológico!!!!

MAN para Idosos: Interpretação do modelo

- Hb=11/12 $\rightarrow$

$$\widehat{\text{RC}} == exp(-8,52 + 0,331(144 - 121)) = 0,40$$

A chance de eutrofico reduz $(\exp(0,4))$ ao passar Hb de 11 para 12.

- Hb=13/14 $\rightarrow$

$$\widehat{\text{RC}} = \exp(-8,52 + 0,331(144 - 121)) = 1,52$$

- Hb=14/15 $\rightarrow \widehat{RC}$ = 2,94.