Introdução à Regressão Logística

Parte 3 – **Inferência para Regressão Logística**

Unidade III

Inferência para Regressão Logística

- Os procedimentos de inferência dependem de tamanhos de amostra grandes.
- Para amostras grandes, sob condições geralmente aplicáveis, os estimadores de máxima verossimilhança para regressão logística são aproximadamente normalmente distribuídos, com pouco ou nenhum viés, e com aproximadamente variâncias e covariâncias correspondentes que são funções das derivadas parciais de segunda ordem do logaritmo da função de verossimilhança.

Inferência para Regressão Logística

- Especificamente, seja ***G*** a matriz de derivadas parciais de segunda ordem do log da função de verossimilhança, sendo as derivadas tomadas em relação aos parâmetros:

$$\underset{p\times p}{\mathbf{G}} = [g_{ij}] \qquad i = 0, 1, \ldots, p-1;\ j = 0, 1, \ldots, p-1$$

Sendo

$$g_{00} = \frac{\partial^2 \log_e L(\boldsymbol{\beta})}{\partial \beta_0^2}$$

$$g_{01} = \frac{\partial^2 \log_e L(\boldsymbol{\beta})}{\partial \beta_0 \partial \beta_1}$$

etc.

Inferência para Regressão Logística

- **Intervalos de Confiança para β_1**

$$b_1 \pm z_{\alpha/2}\,(\text{ASE}).$$

- **Intervalo de Confiança para $e^{\beta 1}$**

(efeito multiplicativo sobre a "odds" do aumento de uma unidade em X)

➔ aplicar e^x aos limites do intervalo para β_1

Inferência para Regressão Logística

- **Intervalos de Confiança para β_1**

$$b_1 \pm z_{\alpha/2}\,(\text{ASE}).$$

- **Intervalo de Confiança para $e^{\beta 1}$**

(efeito multiplicativo sobre a "odds" do aumento de uma unidade em X)

➔ aplicar e^x aos limites do intervalo para β_1

Testes de Significância

$H_0)$ $\beta_1 = 0$
$H_1)$ $\beta_1 \neq 0$

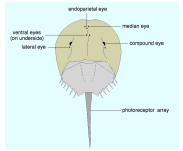
- Teste "Z"

$$Z = \frac{b_1}{ASE} \sim N(0,1)$$

Teste de Wald:

$$Z^2 = \left(\frac{b_1}{ASE}\right)^2 \sim \chi^2 \text{ com 1 g.l.}$$

PROBLEMA



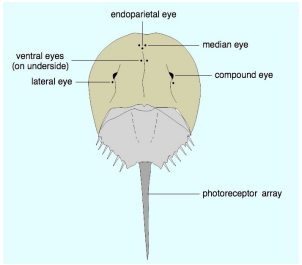
- *Os dados analisados neste exemplo foram extraídos de um estudo de ninhadas de caranguejo do tipo ferradura realizado por J. Brockmann, a ser impresso em Ethology (1996).*

- Neste estudo cada fêmea de caranguejo ferradura tinha um caranguejo macho ligado a ela em seu ninho. O estudo investiga fatores que afetam a existência de qualquer outros machos, chamados *satélites*, morando perto da fêmea. Pensou-se que um dos possíveis fatores seria a *largura da carapaça da fêmea de caranguejo.*

PROBLEMA

Para tal foi selecionada uma amostra de 173 fêmeas e as seguintes características foram investigadas para cada uma delas:

- X - largura da carapaça da fêmea *em cm;*
- Y - *se a fêmea tem pelo menos um satélite (1 – sim e 0- não)*



Number of Observations Read	173
Number of Observations Used	173

Response Profile		
Ordered Value	y	Total Frequency
1	1	111
2	0	62

Probability modeled is y='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	227.759	198.453
SC	230.912	204.759
-2 Log L	225.759	194.453

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	31.3059	1	<.0001
Score	27.8752	1	<.0001
Wald	23.8872	1	<.0001

Estimated Covariance Matrix		
Parameter	Intercept	x
Intercept	6.910227	-0.26685
x	-0.26685	0.01035

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-12.3508	2.6287	22.0749	<.0001
x	1	0.4972	0.1017	23.8872	<.0001

Teste da Razão de Verossimilhança
Formulação Geral

- A forma geral da *estatística do teste da razão de verossimilhança* para uma hipótese nula H_o e uma hipótese alternativa H_1 é

$$G^2 = -2\,(L_0 - L_1)$$

onde

- L_0 - o máximo do logaritmo da função de verossimilhança supondo que H_o é verdadeira
- L_1 - o máximo do logaritmo da função de verossimilhança supondo que H_1 é verdadeira

Sob H_o $G^2 \sim \chi^2$ com v g.l.

onde o número de *graus de liberdade* é a diferença entre o número de parâmetros definido para cada hipótese

Teste da Razão de Verossimilhança
para testar a significância da variável explicativa

A estatística do teste da razão de máxima verossimilhança compara:

- L_0 - o máximo do log da função de verossimilhança quando $\beta_1 = 0$ (isto é, quando $\pi(x)$ é forçado a ser idêntico para todos os valores de x)

com

- L_1 - o máximo do log da função de verossimilhança para β_1 irrestrito (isto é, H_1 é verdadeira).

A estatística do teste é definida por:

$$G^2 = -2\,(L_0 - L_1) \sim \chi^2 \text{ com } 1 \text{ g.l.}$$

Number of Observations Read	173
Number of Observations Used	173

Response Profile		
Ordered Value	y	Total Frequency
1	1	111
2	0	62

Probability modeled is y='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	227.759	198.453
SC	230.912	204.759
-2 Log L	225.759	194.453

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	31.3059	1	<.0001
Score	27.8752	1	<.0001
Wald	23.8872	1	<.0001

Estimated Covariance Matrix		
Parameter	Intercept	x
Intercept	6.910227	-0.26685
x	-0.26685	0.01035

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-12.3508	2.6287	22.0749	<.0001
x	1	0.4972	0.1017	23.8872	<.0001

UnB – IE
Departamento de Estatística

Estimação da Probabilidade de sucesso $\pi(x)$

- A probabilidade estimada de sucesso, (isto é, de que Y=1), para um valor fixado x de X é dada por:

$$\hat{\pi}_i(x_i) = = \frac{\exp(b_0 + b_1 x_i)}{1 + \exp(b_0 + b_1 x_i)}$$

- **Intervalo de Confiança (1-α) para o valor do logito $\pi * (x)$:**

$$(b_0 + b_1 x_i) \pm z_{a/2}\,(\text{ASE})$$

onde erro padrão de estimativa (ASE) é a raiz quadrada de:

$$V(b_0 + b_1 x_i) = V(b_0) + x_i^2 V(b_1) + 2x_i \operatorname{cov}(b_0, b_1)$$

- Estime com 95% de confiança a probabilidade de uma fêmea de caranguejo ferradura com largura de carapaça de 26 cm ter satélite.

Number of Observations Read	173
Number of Observations Used	173

Response Profile		
Ordered Value	y	Total Frequency
1	1	111
2	0	62

Probability modeled is y='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	227.759	198.453
SC	230.912	204.759
-2 Log L	225.759	194.453

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	31.3059	1	<.0001
Score	27.8752	1	<.0001
Wald	23.8872	1	<.0001

Estimated Covariance Matrix		
Parameter	Intercept	x
Intercept	6.910227	-0.26685
x	-0.26685	0.01035

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-12.3508	2.6287	22.0749	<.0001
x	1	0.4972	0.1017	23.8872	<.0001

Avaliação do Modelo

- **Testes de Adequabilidade de Ajustamento:**

H_0) O Modelo de Regressão Logística ajusta-se aos dados.
H_1) O Modelo não se ajusta

ou de forma equivalente:

$$H_0)\ \pi_i = \pi(x_i)$$
$$H_0)\ \pi_i \in (0,1), \qquad i = 1, 2, \ldots, c.$$

- Para cada conjunto distinto i de valores das variáveis explicativas (i =1,2,..., c):

  f_{1i} - número de sucessos (Y=1) para n_i repetições do i-ésimo conjunto de valores das variáveis explicativas.

  $\hat{\pi}_i$ - probabilidade de sucesso predita pelo modelo ajustado.

Então $fe_{1i} = n_i \hat{\pi}_i$ é o número de sucessos ajustado ou esperado.

Consequentemente,

$$fe_{2i} = n_i (1 - \hat{\pi}_i) = n_i - fe_{1i}$$

é o número de insucessos (Y=0) ajustado ou esperado.

Estatística X^2 de Pearson

- Y_{ij} observações independentes e observações repetidas de um ou mais níveis da variável explicativa.

A estatística do teste é

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=0}^{1} \frac{(f_{ji} - fe_{ji})^2}{fe_{ji}}$$

que sob a hipótese H_0, tem distribuição aproximadamente Qui-quadrado com $c - k$ graus de liberdade quando n grande e $k < c$.

k - número de parâmetros do modelo

c - número conjuntos de valores distintos das variáveis explicativas

Estatística X^2 de Pearson

- É recomendável que as frequências esperadas sejam maiores ou igual a 5, e nunca inferiores a 1.
- O teste detecta grandes afastamentos da função resposta logística, mas não é tão sensível para detectar pequenos afastamentos.

Estatística G^2 do Teste da Razão de Verossimilhança

- **Modelo Completo:**

$$E(Y_{ij}) = \pi_j \qquad j = 1,2,\ldots,c$$

onde π_j é a probabilidade de sucesso para cada valor distinto de X, j=1,2,...,c.

O modelo completo no caso da regressão logística é usualmente chamado de ***modelo saturado***.

- **Modelo Restrito (sob H_0):**

$$E[Y_{ij}] = \pi(x) = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}$$

Estatística do Teste:

Considerando a estatística do teste da razão de verossimilhança:

$$G^2 = -2\,ln\left(\frac{L(R)}{L(F)}\right) = -2\left[ln\big(L(R)\big) - ln\big(L(F)\big)\right]$$

- As estimativas de máxima verossimilhança para as ***c*** probabilidades π_j, modelo completo são dadas pelas proporções amostrais:

$$p_j = \frac{Y_{.j}}{n_j} \qquad j = 1,2,\ldots,c$$

- $\hat{\pi}_j$ *estimativa de* π_j *pelo modelo reduzido para cada Xj, j=1,2,..., c,*

Assim a estatística do teste é dada por:

$$G^2 = -2\sum_{j=1}^{c}\left[Y_j\,ln\left(\frac{\hat{\pi}_j}{p_j}\right) + \left(n_j - Y_j\right)ln\left(\frac{1-\hat{\pi}_j}{1-p_j}\right)\right]$$

ou então,

$$G^2 = -2\sum_{j=1}^{c}\left[Y_j\,ln\left(\frac{\hat{Y}_j}{Y_j}\right) + \left(n_j - Y_j\right)ln\left(\frac{n_j-\hat{Y}_j}{n_j-Y_j}\right)\right]$$

que sob a hipótese H_0, tem distribuição aproximadamente Qui-quadrado com $c-k$ graus de liberdade quando n_j for grande e $k < c$

k - número de parâmetros do modelo

c - número conjuntos de valores distintos das variáveis explicativas

Teste de Hosmer e Lameshow

- Construir as classes com base nos percentis das probabilidades estimadas:

  O uso de c =10 grupos resulta no primeiro grupo formado pelos pares de freqüências observadas e estimadas com *n/10* menores valores de probabilidades estimadas, isto é, tendo o primeiro decil de probabilidades preditas; o segundo grupo tendo o segundo decil de probabilidades preditas, e assim por diante.

- Construir as classes com base em valores fixados das probabilidades estimadas:

  O uso de c =10 grupos resulta em pontos de corte definidos nos valores *k/10, k=1, 2, ..., 9* e os grupos contêm todos os pares de freqüências observadas e estimadas com valores de probabilidades estimadas entre pontos de corte adjacentes.

Partition for the Hosmer and Lemeshow Test

		y = 1		y = 0	
Group	Total	Observed	Expected	Observed	Expected
1	19	5	5.39	14	13.61
2	18	8	7.62	10	10.38
3	17	11	8.62	6	8.38
4	17	8	9.92	9	7.08
5	16	11	10.10	5	5.90
6	18	11	12.30	7	5.70
7	16	12	12.06	4	3.94
8	16	12	12.90	4	3.10
9	16	13	13.69	3	2.31
10	20	20	18.41	0	1.59

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
5.2465	8	0.7309