UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE MATEMÁTICA

DEPARTAMENTO DE ESTATÍSTICA

REGRESSÃO LOGÍSTICA

Autor: ANGELA RADÜNZ

Orientadora: Professora JANDYRA M. G. FACHEL

Monografia apresentada para obtenção do título de Bacharel em Estatística.

Porto Alegre, dezembro de 1992.

AGRADECIMENTOS

Gostaria de agradecer a todos aqueles que me ajudaram, direta ou indiretamente, para a execução deste trabalho.

Agradeço a professora Jandyra M. G. Fachel pela orientação prestada à esta monografia.

Manifesto também meus agradecimentos à professora Vera Beatriz Wald do Departamento de Medicina Animal da Faculdade de Veterinária da UFRGS por ter cedido o material que utilizamos no exemplo prático.

ÍNDICE

1. INTRODUÇÃO

O presente trabalho trata da técnica de Análise de Regressão Logística, que é de grande utilidade e vem sendo usado principalmente na área epidemiológica. Esta técnica é usada quando temos uma variável resposta dicotômica ou binária. Este método procura o melhor ajustamento para os dados em que, primeiramente, são estimados os parâmetros do modelo e, em seguida, testa-se a significância dos mesmos.

A técnica permite, na área epidemiológica, a interpretação das relações fator de exposição-risco obtida através da análise de estudos caso-controle, mas pode ser utilizada em outros estudos. O advento da técnica de Regressão Logística Múltipla utilizando modelagem estatística permitiu a análise de relações mais complexas entre fatores de estudo e a variável dependente.

Os recentes programas computacionais para a técnica fornecem, em geral, os parâmetros estimados do modelo, seus erros padrão, a razão de chances (odds ratio) e seu intervalo de confiança para cada covariável no modelo, estimadores da função de verossimilhança e testes de ajustamento do modelo (goodness-of-fit).

Apresentaremos uma breve exposição sobre a Regressão Logística. No primeiro tópico daremos uma definição sobre o modelo. Após, falaremos sobre os métodos de estimação para os coeficientes do modelo, seu teste de significância e interpretação. Apresentaremos, ainda, uma breve exposição sobre a regressão logística múltipla.

Mostraremos, no capítulo 4, um exemplo aplicativo para melhor compreensão do método.

Este trabalho foi elaborado de uma maneira simples e direta. Ele dá uma base sobre o que é o modelo e suas aplicações, a nível introdutório.

2. MODELO DE REGRESSÃO LOGÍSTICA

2.1. Definição do Modelo de Regressão Logística

A Regressão Logística é um método de regressão apropriado a uma variável resposta binária ou dicotômica (variável dependente) e uma ou mais variáveis explanatórias ou covariáveis (variáveis independentes) que podem ser binárias, categóricas ou contínuas. O objetivo desta análise é encontrar o melhor modelo ou o que melhor se ajusta aos dados. Aplicações na área biológica ou também em outras áreas serão exemplificadas.

Exemplo 2.1.: Apresentamos aqui alguns dados em que se deseja explorar a relação entre a presença ou não de doença coronária e a idade dos indivíduos. Aqui neste caso temos a variável CHD (presença ou ausência de doênça coronária) como variável resposta, binária, com a ausência da doença codificada como 0 (zero) e a presença da doença codificada como 1 (um). Os dados estão dispostos no Figura 2.1. Como podemos observar, é dificil uma análise através destes dados da maneira como estão dispostos, pois sabemos que há uma variabilidade muito grande de CHD em todas as idades. Para remover esta variação utilizou-se um método muito comum: criar intervalos para a variável independente (idade) e calcula-se a média da variável resposta (média de presença de CHD) em cada grupo de idade (ver Tabela 2.1 e Figura 2.2).

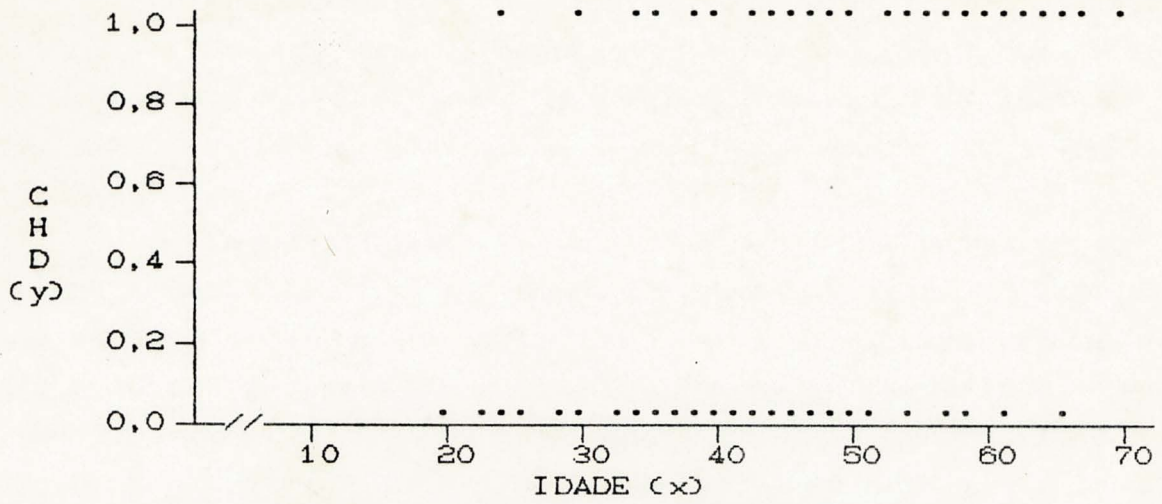

Figura 2.1 - Gráfico de Dispersão de CHD por Idade.

Tabela 2.1 - Tabela de Freqüencias da Ausência e Presença da Doença Coronária por Grupos de Idade.

Grupos de idade	CHD			Média (prop)
	n	Ausência	presença	
20-29	10	9	1	0,10
30-34	15	13	2	0,13
35-39	12	9	3	0,25
40-44	15	10	5	0,33
45-49	13	7	6	0,46
50-54	8	3	5	0,63
55-59	17	4	13	0,76
60-69	10	2	8	0,80
Total	100	57	43	0,43

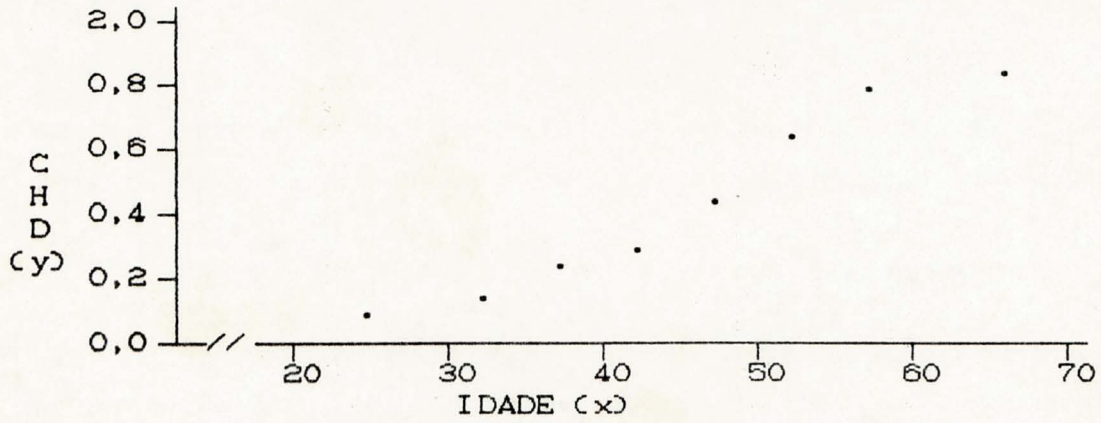

Figura 2.2 - Gráfico de Dispersão da Proporção de CHD pelos Grupos de Idade.

Em 1920 Pearl e Reed usaram uma curva logística na forma de S dada pela função $y=1/(1+ce^{-\beta_1 x})$ para um modelo de crescimento populacional. Mais tarde esta curva começou a ser usada em ensaios biológicos, para uma curva de resposta de dosagem.

Segundo Hosmer & Lemeshow (1989), a primeira diferença entre a Regressão Logística e a Regressão Linear Simples consiste na relação entre a variável resposta e a variável independente. Na análise de regressão é importantea média da variável resposta, dado o valor da variável independente. Esta quantidade é a média condicional $E(Y|x)$, onde Y é a variável resposta e x é a variável independente).

Na Regressão Linear, esta média pode ser expressa pela equação:

$$E(Y/x)=\beta_0+\beta_1 x \qquad -\infty \leq x \leq \infty$$

Na Regressão Logística, a $E(Y|x)$ está estimada pela coluna da média de presença de CHD na tabela 2.1 e é mostrada na Figura 2.2, em que os pontos são muito próximos aos verdadeiros. Para dados dicotômicos a média condicional está entre zero e um ($0 \leq E(Y|x) \leq 1$). A figura 2.2 mostra que $E(Y|x)$ tem um comportamento gradual crescente na forma de S, assemelhando-se a uma distribuição acumulada. Neste caso o modelo a ser usado é a da distribuição Logística.

Na distribuição Logística podemos usar $\pi(x)=E(Y|x)$ e a forma específica do modelo é:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-\beta_0-\beta_1 x}}$$

Podemos usar uma transformação de $\pi(x)$ chamada de <u>transformação logit</u> (logit é abreviatura de <u>logarithmic unit</u>), definida como:

$$g(x) = \text{logit}\,(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0+\beta_1 x \qquad (2.1)$$

A transformação $g(x)$, por ser linear nestes parâmetros, tem propriedades da regressão linear, podendo ser contínua (de $-\infty$ a $+\infty$, dependendo dos valores de x).

A segunda diferença entre a regressão linear e a

regressão logística está na distribuição condicional da variável resposta. Na Regressão Linear as observações são expressas como $y=E(Y|x)+\varepsilon$, onde ε (erro) é o desvio das observações da média condicional. ε tem distribuição Normal com média 0 e variância constante. A distribuição condicional da variável resposta (Y), dado x, será Normal com média $E(Y|x)$ e variância constante. No caso da variável resposta ser dicotômica, expressamos o valor da variável, dado o valor de x, como $y=\pi(x)+\varepsilon$, com ε podendo assumir um ou dois valores possíveis. Se y=1 então $\varepsilon=1-\pi(x)$ com probabilidade $\pi(x)$; se y=0 então $\varepsilon=-\pi(x)$ com probabilidade $1-\pi(x)$. ε tem uma distribuição com média 0 e variância igual a $\pi(x)[1-\pi(x)]$. Isto é, a distribuição condicional da variável resposta segue uma distribuição Binomial com probabilidade dada pela média condicional, $\pi(x)$. Os princípios que orientam a análise de Regressão Linear são os mesmos para a Regressão Logística.

2.2. Estimando os Parâmetros do Modelo de Regressão Logística

Suponha que temos uma amostra de n pares de observações independentes (x_i, y_i), $i=1,2,\ldots,n$, onde y_i é a variável resposta e x_i é a variável independente, ambos do i-ésimo indivíduo. A variável resposta foi condicionada com resposta 0 ou 1, representando ausência ou presença da característica, respectivamente. Para achar um modelo de Regressão Logística, para um conjunto de dados, devemos estimar os valores de β_0 e β_1, os parâmetros não conhecidos no modelo (2.1).

Na regressão Linear o método mais usado para estimar parâmetros não conhecidos é o de mínimos quadrados, onde se acham os valores β_0 e β_1 que miniminizam a soma dos desvios ao quadrado dos valores observados de Y. Infelizmente, o método de mínimos quadrados para Regressão Logística não produz estimadores com propriedades estatísticas desejadas. O método de estimação utilizado em Regressão Logística é o método de Máxima Verossimilhança. Este método estima valores de parâmetros não conhecidos que maximizam a probabilidade de obter o conjunto de dados observados. Para aplicar este método, primeiro construimos a função de verossimilhança, que expressa a probabilidade dos dados observados como uma função dos parâmetros desconhecidos. Os estimadores da Máxima Verossimilhança são escolhidos para serem os valores que maximizam esta função.

Se Y é codificado como 0 ou 1, então $\pi(x)$ é a probabilidade condicional de Y ser igual a 1, dado o valor de x $[P(Y=1|x)]$ e $1-\pi(x)$ é a probabilidade condicional de Y ser igual a zero dado o valor de x $[P(Y=0|x)]$. Para os pares (x_i, y_i), onde $y_i=1$, a contribuição para a função de verossimilhança é $\pi(x_i)$, e para os pares, onde $y_i=0$, a contribuição para a função de verossimilhança é $1-\pi(x_i)$. Uma maneira de expressar a contribuição da função de verossimilhança para o par (x_i, y_i) é o termo:

$$\zeta(x_i) = \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \qquad (2.2)$$

Para observações independentes, a função de verossimilhança é obtida pelo produto dos termos:

$$l(\beta) = \prod_{i=1}^{n} \zeta(x_i) \qquad (2.3)$$

Pelo princípio da Máxima Verossimilhança, a estimativa de β maximiza a expressão da equação (2.3). Se aplicarmos o logarítmo a esta equação, teremos a log verossimilhança:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} \qquad (2.4)$$

Para achar o valor de β que maximize $L(\beta)$, diferenciamos $L(\beta)$ em relação a β_0 e β_1 e usamos:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \qquad (2.5)$$

e

$$\sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0 \qquad (2.6)$$

que são chamadas equações de verossimilhança.

Na regressão linear, as equações de verossimilhança são lineares com parâmetros não conhecidos; na regressão logística, elas são não lineares em β_0 e β_1 e requerem métodos especiais para sua solução. Estes métodos são iterativos por natureza e existem softwares de regressão logística para a sua solução.

O valor de β pelas soluções das equações (2.5) e (2.6), são chamadas de estimativas da máxima verossimilhança e denotadas como $\hat{\beta}$. Este valor dá uma estimativa da probabilidade condicional de Y ser igual a 1, dado que x é igual a x_i.

Para os dados do Exemplo 2.1, usando um pacote de regressão logística, com a variável contínua idade, temos os resultados:

Tabela 2.2 - Resultados do Ajustamento do Modelo de Regressão Logística aos Dados do Exemplo 2.1.

Variável	Coeficiente Estimado	Erro Padrão	Coef./E.P.
IDADE	0,111	0,024	4,61
Constante	-5,310	1,134	-4,68

log-verossimilhança = -53,677

As estimativas de β_0 e β_1 foram: $\hat{\beta}_0$=-5,310 e $\hat{\beta}_1$=0,111. Pelos valores encontrados temos a equação

$$\pi(x) = \frac{e^{-5,31+0,111\ IDADE}}{1+e^{-5,31+0,111\ IDADE}} \qquad (2.7)$$

e o logit estimado

$$\hat{g}(x) = -5,31+0,111\ IDADE$$

A Tabela 2.2 contém os coeficientes estimados, seu erro padrão e a estatística definida pela razão entre o coeficiente e seu erro padrão, que é a estatística de Wald a ser apresentada na próxima seção.

2.3. Testando a significância das variáveis e dos coeficientes do modelo

Depois de estimar os coeficientes para o nosso modelo, o próximo passo consiste na avaliação da significância da variável no modelo. Isto geralmente envolve formulação e teste de hipótese estatística para determinar se as variáveis independentes no modelo são "significativamente" relacionadas com a variável resposta. O método para realizar este teste é geral e difere de um tipo de modelo para outro somente nos detalhes específicos.

A função matemática usada para comparar os valores observados e preditos depende do problema particular. Se os valores preditos da variável no modelo são melhores, ou mais exatos que quando a variável não está no modelo, então dizemos que a variável em questão é "significante".

O método geral para avaliar a significância das variáveis é facilmente ilustrado num modelo de regressão linear, e pode ser utilizado na regressão logística. Uma comparação das duas, salientará as diferenças entre um modelo com variável resposta contínua e com variável resposta dicotômica.

Na regressão linear a avaliação da significância do coeficiente de inclinação é feita através da análise de variância. Nesta análise, calcula-se a soma de quadrados da regressão e a soma de quadrados dos resíduos, para fazer a comparação entre os valores observados e os preditos, através do modelo.

O princípio fundamental da regressão logística é o mesmo: comparar valores observados com valores preditos da variável resposta obtida de modelos com e sem as variáveis independentes em questão. Na regressão logística, a comparação de modelos esperados e observados está baseada na função log-verossimilhança definida na equação 2.4.

A comparação de valores observados e preditos usando a função verossimilhança é baseada na seguinte expressão:

$$D = -2\ln\left[\frac{\text{(verossimilhança do modelo corrente)}}{\text{(verossimilhança do modelo saturado)}}\right] \qquad (2.8)$$

O modelo saturado é um modelo que contém tantos parâmetros quanto pontos (um exemplo de modelo saturado, ajustado por um modelo de regressão linear, é quando temos somente dois pontos, n=2).

A quantidade dentro dos colchetes na expressão acima é chamada razão de verossimilhança. O motivo de usar menos duas vezes este log é matemático e é necessário para obter uma distribuição conhecida e pode ser usado com o propósito de testar hipóteses. Desta maneira o teste é chamado de teste da razão de verossimilhança. Usando as equações 2.4 e 2.8 teremos

$$D = -2 \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i) \ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \qquad (2.9)$$

onde $\hat{\pi}_i = \hat{\pi}(x_i)$. Esta estatística D é denominada "deviance" e tem o mesmo papel que a soma dos quadrados dos desvios na Regressão Linear.

Com o propósito de avaliar a significância de uma variável independente, comparamos o valor de D com e sem a variável independente na equação. A mudança em D para incluir a variável no modelo é obtido como segue:

G = D(modelo sem a variável) - D(modelo com a variável)

Esta estatística tem a mesma função na Regressão Logística que o numerador do teste F parcial na regressão linear.

$$G = -2 \ln\left[\frac{(\text{verossimilhança sem a variável})}{(\text{verossimilhança com a variável})}\right]$$

Para o caso específico de uma variável independente é fácil mostrar que, quando ela não está presente no modelo, o estimador de máxima verossimilhança de β_0 é $\ln(n_1/n_0)$ onde $n_1=\Sigma y_i$ e $n_0=\Sigma(1-y_i)$ e que o valor predito é constante, n_1/n. Neste caso o valor de G é:

$$G = -2 \ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n} \hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{(1-y_i)}}\right]$$

ou,

$$G = 2\left\{\sum_{i=1}^{n}[y_i \ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)]\right\} \quad (2.10)$$

Sob a hipótese de que $\beta_1=0$, a estatística G segue uma distribuição χ^2 com 1 g.l. para amostra suficientemente grande.

Consideremos os dados do exemplo 2.1, e os coeficientes estimados e a log-verossimilhança dados na tabela 2.2. Para estes dados $n_1=43$ e $n_0=57$; calculando G pela equação 2.10 encontramos $G=29,31$ $(p<0,001)$.

Pelo resultado acima, temos que idade é uma variável estatísticamente significante para predizer CHD. Também devemos considerar se a variável é biologicamente importante para o modelo ajustado, assim como a inclusão de outras variáveis potencialmente importantes.

O cálculo do log-verossimilhança e teste da razão de verossimilhança são obtidos em programas computacionais de regressão logística. Com eles é possível verificar a significância da adição de novos termos. No caso de uma variável independente, podemos primeiro ajustar um modelo apenas com o termo constante e, após, podemos ajustar um modelo contendo a variável independente juntamente com o termo constante, este nos dará um novo log-verossimilhança.

Um possível teste de significância para coeficientes é o teste de Wald. Dado por

$$W = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

Ele testa a hipótese do coeficiente β_1 ser igual a zero, e tem uma distribuição Normal padronizada. Outra estatística utilizada para o teste dos coeficientes é o teste "Score" que é fornecido por alguns pacotes computacionais.

2.4. Interpretação dos Coeficientes

A interpretação do modelo ajustado requer que sejamos capazes de fazer inferências práticas em relação aos coeficientes estimados no modelo. A questão é: **Quais os coeficientes do modelo que nos mostram as questões pesquisadas que motivaram o estudo?** Os coeficientes estimados para as variáveis independentes representam a inclinação ou razão de mudança da função da variável dependente por unidade de medida na variável independente. Desta maneira, interpretação envolve duas coisas: decisão sobre a relação entre a variável dependente e a variável independente e definição da unidade de medida da variável independente.

Para os leitores familiarizados com modelos lineares generalizados, função Link, no caso de um modelo de regressão linear é a função identidade (Y=y). No modelo de Regressão Logística a função Link é a transformação logit $g(x) = \ln(\pi(x)/[1-\pi(x)]) = \beta_0+\beta_1 x$.

Para um modelo de regressão linear dizemos que o coeficiente de inclinação, β_1, é dado por $\beta_1 = y(x+1) - y(x)$, onde $y(x) = \beta_0+\beta_1 x$. Neste caso, a interpretação do coeficiente expressa a mudança na escala de medida da variável dependente para uma mudança de uma unidade na variável independente. Por exemplo, se na regressão de peso e altura de meninos adolescentes nós encontramos a inclinação igual a 5, então nós podemos concluir que uma mudança de 1 polegada na altura é associada com uma mudança de 5 libras ($\cong$ 2,5 kg) no peso.

No modelo de regressão logística $\beta_1 = g(x+1)-g(x)$. Isto é, o coeficiente de inclinação representa uma mudança de uma unidade na variável independente x. A interpretação do coeficiente no modelo de regressão logística está relacionada à diferença entre dois logits.

2.4.1. Interpretação dos coeficientes quando a covariável é Dicotômica

Quando a variável é dicotômica, assumimos que x é codificado como zero ou 1. O modelo então tem dois valores de $\pi(x)$ e equivalentemente dois valores de $1-\pi(x)$. Estes valores podem ser convenientemente mostrados numa tabela 2x2 (Tabela 2.3).

Tabela 2.3 - Valores do Modelo de Regressão Logística quando a variável independente é dicotômica.

		Variável Independente X	
		x=1	x=0
Variável Resposta Y	y=1	$\pi(1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$
	y=0	$1-\pi(1)=\frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0)=\frac{1}{1+e^{\beta_0}}$
	Total	1,0	1,0

O odds da resposta entre indivíduos com x=1 é definido como $\pi(1)/[1-\pi(1)]$. Similarmente, o odds da resposta entre indivíduos com x=0 é definido como $\pi(0)/[1-\pi(0)]$. O log do odds é chamado logit e, neste exemplo

$$g(1) = \ln[\pi(1)/[1-\pi(1)]]$$

e

$$g(0) = \ln[\pi(0)/[1-\pi(0)]]$$

O odds ratio, OR, é definido como a razão do odds em x=1 com o odds em x=0, e é dado pela equação

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \qquad (2.11)$$

O log do odds ratio, chamado de log-odds é

$$\ln(OR) = \ln\left[\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}\right] = g(1)-g(0)$$

que é a diferença do logit.

Agora, usando a expressão para o modelo de regressão logística mostrado na tabela 2.3, o odds ratio é

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\left(\frac{1}{1+e^{\beta_0}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Pela regressão logística com variável independente dicotômica

$$OR = e^{\beta_1} \qquad (2.12)$$

e o logit da diferença, ou log odds, é

$$\ln(OR) = \ln(e^{\beta_1}) = \beta_1$$

Esta interpretação dos coeficientes é a razão fundamental da regressão logística ser um instrumento analítico poderoso para pesquisa epidemiológica.

O odds ratio é uma medida de associação de grande uso, especialmente em epidemiologia, indica o quanto uma característica é mais provável (ou não-provável) entre aqueles com x=1 e entre aqueles com x=0. Por exemplo, se y denota a presença ou ausência de câncer no pulmão e se x denota se a pessoa é ou não um fumante, então $\hat{OR} = 2$ indica que câncer no pulmão ocorre duas vezes mais entre fumantes que entre não fumantes no estudo da população.

A interpretação dada pelo odds ratio é baseada no fato em que muitas vezes ela é aproximadamente igual a uma quantidade chamada risco relativo (RR). O RR é igual a razão $\pi(1)/\pi(0)$. Pela equação 2.11 temos que $OR \cong RR$ se $[1-\pi(0)]/[1-\pi(1)] \approx 1$. Esta aproximação se dará quando $\pi(x)$ é pequeno tanto para x=1 e x=0.

Modelo de dose-resposta. Agora, considere o caso onde, x representaria dose, ou seu logarítmo, e y a freqüência relativa de resposta para vários níveis de doses. O modelo é representado por $y=1/(1+ce^{-\beta_1 x})$, onde c e o β_1 são os parâmetros a serem estimados. Se nós definimos logit $y=\log[y/(1-y)]$, segue que o

logit $y=\beta_0+\beta_1 x$, onde $\beta_0=-\log c$. Se y é a probabilidade de doença, dado o valor de x, (isto é, $y=P(D=1|x)$, D como uma variável ter ou não a doença, então o logit y é uma função linear da variável x. Se nós substituirmos x pela variável de exposição E, então teremos logit $P(D=1|E) = \log[P/(1-P)|E] = \beta_0+\beta_1 E$; em outras palavras, o log do odds da doença está modelado por uma função linear da variável E.

$$\text{odds da doença} = \frac{P(D=1/E)}{1-P(D=1/E)} = e^{\beta_0+\beta_1 E}$$

$$P(D=1|E) = \frac{e^{\beta_0+\beta_1 E}}{1+e^{\beta_0+\beta_1 E}} = \left[1 + e^{-(\beta_0+\beta_1 E)}\right]^{-1}$$

$$= \left[1+\exp[-(\beta_0+\beta_1 E)]\right]^{-1},$$

Considere primeiro uma variável independente de exposição binária E, onde

$$E = \begin{cases} 1 & \text{exposto} \\ 0 & \text{não exposto} \end{cases}$$

com,

$$P(D=1|E=0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}, \text{ e } P(D=1|E=1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}},$$

$$OR = \frac{P(D=1|E=1)/P(D=0|E=1)}{P(D=1|E=0)/P(D=0|E=0)} = \frac{e^{\beta_0+\beta_1 .1}}{e^{\beta_0+\beta_1 .0}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}$$

$$= e^{\beta_0+\beta_1-\beta_0} = e^{\beta_1}$$

Assim, $OR=e^{\beta_1}$, e $\beta_1=\log(OR)$. Nossas estimativas foram $\hat{OR}=e^{\hat{\beta}_1}$, ou $\hat{\beta}_1=\log(\hat{OR})$.

Com as variáveis Doença e Exposição dicotômicas, nós estamos na situação de tabela 2x2. Utilizando o método de estimação de máxima verossimilhança, é fácil mostrar que

		E		
		1	0	
Y	1	a	b	a + b
	0	c	d	c + d
		a + c	b + d	n

$e^{\hat{\beta}_1} = \frac{ad}{bc}$, $e^{\hat{\beta}_0} = \frac{a}{b}$ = chance (odds) de ter a doença entre os expostos (ou os que tem o fator de risco)

Exemplo 2.2.: Considere o exemplo (não real) abaixo:

$$Y = \begin{cases} 1 & \text{Sente dores no estômago} \\ 0 & \text{Não sente dores no estômago} \end{cases}$$

$$E = \begin{cases} 1 & \text{Come muito} \\ 0 & \text{Não come muito} \end{cases}$$

Ajuste o modelo: logit $P(Y=1/E) = \beta_0 + \beta_1 E$.

Tabela 2.4 - Resultados do Ajustamento do modelo de regressão logística para o exemplo 2.2.

Variável	Beta	Êrro Padrão	Z
Intercepto	-1,6	0,2	-8,0
E	1,1	0,4	2,8

Nós estimamos o odds ratio para doença

$$OR = e^{\hat{\beta}_1} = e^{1,1} = 3,0.$$

A chance de um indivíduo ter a doença, dado que ele é exposto, é 3 vezes maior que de um indivíduo ter a doença, dado que ele não é exposto.

Nós podemos estimar a probabilidade de doença para cada grupo

exposto

$$P(D=1 \mid E=0) = e^{-1,6}/(1+e^{-1,6}) = 0,17$$

Probabilidade de ter a doença, dado que não é exposto.

e

$$P(D=1 \mid E=1) = \frac{e^{-1,6+1,1}}{1+e^{-1,6+1,1}} = 0,38.$$

Probabilidade de ter a doença, dado que é exposto.

Se $\hat{OR}$ é calculado a partir destas estimativas, $\hat{OR} = \frac{0,38/0,62}{0,17/0,83} = 3,0$, o mesmo resultado que nós já havíamos encontrado previamente.

Para um teste estatístico da hipótese nula que não haja relação doença/exposição, nós formulamos a hipótese nula na forma

Ho: OR=1, ou equivalente, Ho: $\beta_1 = 0$.

O teste estatístico é

$$Z = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)} = \frac{1,1}{0,4} = 2,8.$$

O teste estatístico tem, para n grande, uma distribuição normal padrão, então, neste caso, podemos rejeitar Ho com $p<0,01$.

Para um intervalo de confiança (IC) do OR:

$\beta_1 \pm Z_{1-\alpha/2}\, s.e.(\hat{\beta}_1)$ para um IC a $100(1-\alpha)\%$, ou

$\beta_1 \pm 1,96\, s.e.(\hat{\beta}_1)$ para um a IC 95%.

Então nós simplificamos exponencialmente para ter um IC do OR:

$$\left[e^{\hat{\beta}_1 - Z_{1-\alpha/2}\, s.e.(\hat{\beta}_1)} \;,\; e^{\hat{\beta}_1 + Z_{1-\alpha/2}\, s.e.(\hat{\beta}_1)} \right]$$

ou, escrevendo da forma,

$$e^{\hat{\beta}_1 \pm Z_{1-\alpha/2}\, s.e.(\hat{\beta}_1)}$$

Para o IC de 95% este é

$$e^{\hat{\beta}_1 \pm 1,96\, s.e.(\hat{\beta}_1)}$$

No Exemplo 2.2 que é "sentir dores no estômago", nós temos um IC

de 95% de $\hat{\beta}_1$

$$1,1 \pm 1,96\,(0,4), \text{ ou } (0,3;1,9).$$

Para $OR=e^{\beta_1}$ nós temos

$$(e^{0,3}, e^{1,9}), \text{ ou } (1,3;6,7).$$

Nós podemos usar o IC para o teste de hipótese Ho: OR=1, observando se 1 caiu ou não no IC. Aqui neste caso como o intervalo não contém o 1, então nós podemos rejeitar Ho com $\alpha=0,05$.

2.4.2. Interpretação dos coeficientes quando a covariável é contínua

Quando o modelo de regressão logística contém uma variável independente contínua, a interpretação do coeficiente estimado para esta variável dependerá da unidade de medida da variável.

Como o logit é linear na covariável contínua, x, a equação para o logit é $g(x)=\beta_0+\beta_1 x$. O coeficiente de inclinação, β_1, dada a mudança no log odds para um aumento de uma unidade em x é $\beta_1=g(x+1)-g(x)$, para qualquer valor de x. Geralmente o valor de "1" não será biologicamente muito interessante. Por exemplo, um aumento de 1 ano em idade ou de 1 mm Hg na pressão sistólica pode ser pequena para ser considerado importante. Uma mudança de 10 anos ou 10 mm Hg é considerada mais relevante. Se a amplitude de x é de 0 a 1, então uma mudança de 1 é também grande e uma mudança de 0,01 pode ser mais real. Para se ter uma interpretação para covariável de escala contínua necessitamos desenvolver um método de estimação por ponto e intervalo para uma mudança arbitrária de c unidades na covariável.

O log do odds para uma mudança de c unidades em x é obtida do logit diferença $g(x+c)-g(x) = c\beta_1$ e o odds ratio associado é obtido pela exponencial deste logit diferença, $OR(c) = OR(x+c,x) = \exp(c\beta_1)$. Uma estimativa pode ser obtida substituindo β_1 com o estimador de máxima verossimilhança $\hat{\beta}_1$. Uma estimativa do erro padrão para a estimação do intervalo de confiança é obtido multiplicando o erro padrão estimado de $\hat{\beta}_1$ por c. A estimativa do IC do$OR(c)$ a $100(1-\alpha)\%$ é

$$\exp[c\hat{\beta}_1 \pm z_{1-\alpha/2} c\hat{SE}(\hat{\beta}_1)]$$

Visto que ambas as estimativas dos pontos e pontos finais do intervalo de confiança depende da seleção de c, que deve estar especificado nas tabelas e cálculos.

Como no exemplo 2.1, considere o modelo univariado na tabela 2.2. O resultado estimado do logit da regressão logística entre as variáveis idade e CHD foi $\hat{g}(idade) = -5,310 + 0,111.idade$. O odds ratio estimado para um acréscimo de 10 anos

na idade é $\hat{OR}(10) = \exp(10 \times 0,111) = 3,03$. Indica que para todo acréscimo de 10 anos na idade, o risco de CHD aumenta 3,03 vezes. O IC do OR a 95% de confiança é $\exp(10 \times 0,111 \pm 1,96 \times 10 \times 0,024) = (1,90; 4,86)$.

Resultados similares a este podem ser colocados em tabelas mostrando os resultados do modelo de regressão logística ajustado. A validade, da interpretação às vezes pode ser questionável, como neste exemplo, visto que o risco adicional de CHD para 40 anos de idade comparado com 30 anos pode ser completamente diferente do risco adicional de CHD para 60 anos comparado com 50 anos. Este é um dilema inevitável quando covariáveis contínuas são modeladas linearmente no logit. Se acreditamos que o logit não é linear na covariável, então o uso de variáveis dummy devem ser consideradas. Alternativamente, uso de termos de ordem elevada ($x^2, x^3, \ldots$) ou escala não linear na covariável (log x) podem ser considerados.

A interpretação do coeficiente estimado para uma variável contínua é similar para variaveis de escala binária: uma estimativa da razão log odds. A diferença é que deve se considerar a unidade de medida da variável.

2.4.3. Interpretação dos coeficientes quando a covariável é categórica

Suponha que em vez de duas categorias a variável independente tenha k>2 valores distintos. Cada uma destas variáveis tem um número fixo de respostas discretas e a escala de medida é nominal. Devemos formar um conjunto de variáveis de planejamento "design" para representar as categorias da variável.

A tabela a seguir mostra este método para a variável Raça versus CHD. Nesta situação a variável Raça tem 4 níveis.

Tabela 2.5 - Classificação cruzada de Raça e CHD em 100 Sujeitos.

CHD	Branca	Preta	Hispânica	Outra	Total
Presente	5	20	15	10	50
Ausência	20	10	10	10	50
Total	25	30	25	20	100
$\hat{OR}$	1,0	8,0	6,0	4,0	
IC a 95%		(2,3;27,6)	(1,7;21,3)	(1,1;14,9)	
$\ln(\hat{OR})$	0,0	2,08	1,79	1,39	

O odds ratio é dado para cada raça, usando a raça branca como grupo de referência, com a qual as demais raças são comparados. Um método para especificar as variáveis design define todos os valores iguais a zero para o grupo de referência e o valor das outras variáveis igual a 1 pra a Raça em questão como mostra a tabela a seguir:

Tabela 2.6 - Especificação de Variáveis Design de RAÇA Usando a Raça Branca como Grupo de Referência.

	Variáveis Designadas		
RAÇA	D1	D2	D3
Branca (1)	0	0	0
Preta (2)	1	0	0
Hispânica (3)	0	1	0
Outra (4)	0	0	1

Tabela 2.7 - Resultados do Ajustamento do Modelo de Regressão Logística para os Dados da Tabela 2.5 Usando Variáveis Design da Tabela 2.6.

Variável	Coeficiente Estimado	Erro Padrão	Coef./E.P.	$\hat{OR}$
RAÇA (2)	2,079	0,633	3,29	8,0
RAÇA (3)	1,792	0,646	2,78	6,0
RAÇA (4)	1,386	0,671	2,07	4,0
Constante	-1,386	0,500	-2,77	

Uma comparação de coeficientes estimados da tabela 2.7 com o OR da tabela 2.6 mostra que

$\ln[\hat{OR}(\text{branca, preta})] = \hat{\beta}_{11} = 2,079$,

$\ln[\hat{OR}(\text{hispânica, branca})] = \hat{\beta}_{12} = 1,792$ e

$\ln[\hat{OR}(\text{outra, branca})] = \hat{\beta}_{13} = 1,386$.

No caso univariado as estimativas do erro padrão encontrado na regressão logística são idênticas às estimativas obtidas usando as freqüencias da tabela de contingência.

Encontrando os limites de confiança para o odds ratio e aplicando exponencial a esses limites encontraremos os limites para o odds ratio.

2.5. Caso Multivariado

Foi discutido a interpretação de um coeficiente de regressão logística estimado no caso quando existe uma única variável independente no modelo. Ajustando uma série de modelos univariados raramente obtemos uma análise adequada dos dados em estudo. Geralmente considera-se uma análise multivariada para melhor compreensão dos dados. Um dos objetivos desta análise é ajustar estastisticamente os efeitos estimados de cada variável no modelo pelas diferenças nas distribuições e associações relacionando outras variáveis independentes. Aplicando este conceito para um modelo de regressão logística multivariada, podemos supor que cada coeficiente estimado produz uma estimativa do log odds ajustado para todas outras variáveis incluídas no modelo.

Considere um conjunto de p variáveis independentes que serão denotadas pelo vetor $x'=(x_1,x_2,\ldots,x_p)$. A probabilidade condicional da variável resposta ser presença será denotada por $P(Y=1|x)=\pi(x)$. Então o logit do modelo de regressão logística múltipla é dado pela equação

$$g(x)=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p$$

neste caso

$$\pi(x)=\frac{e^{g(x)}}{1+e^{g(x)}}$$

A situação em que o modelo contém duas ou mais variáveis independentes (uma dicotômica e outra contínua, por exemplo), é freqüentemente encontrado em pesquisa epidemiológica, quando um fator de risco é registrado como estando presente ou ausente, e nós queremos ajustar para uma variável como idade. A situação análoga na regressão linear é chamada análise de covariância.

2.5.1. Confundimento e Interação

O termo confundimento é usado por epidemiologistas para descrever uma covariável que é associada com uma variável resposta de interesse e uma variável independente primária ou fator de risco. Quando ambas associações estão presentes então a relação entre o fator de risco e a variável resposta é dita ser confundida.

Interação pode ter muitas formas diferentes, mas começamos por descrever a situação quando a interação está ausente. Considere um modelo contendo uma variável fator de risco dicotômico e uma covariável contínua. Se a associação entre a covariável (idade) e a variável resposta é a mesma em cada nível do fator de risco (grupo), então não há interação entre a covariável e o fator de risco. Graficamente, a ausência de interação produz um modelo com duas linhas paralelas, para cada nível da variável de fator de risco. Em geral, a ausência de interação é caracterizada pelo modelo que não contém termos envolvendo duas ou mais variáveis.

Quando a interação está presente, a associação entre o fator de risco e a variável resposta diferem, ou dependem de algumas maneiras no nível da covariável. Epidemiologistas usam o termo efeito de modificação para descrever uma variável que interage com um fator de risco.

Quando a interação não é estatisticamente significante e nem "biologicamente" significante, ela pode ser deixada fora do modelo, partindo de logit $P(Y=1) = \beta_0+\beta_R R+\beta_T T$ do exemplo. Se o foco é em T, a questão permanece e tira-se o R. A significância estatística de $\hat{\beta}_R$ não é importante. Se tirando o R muda a estimativa de β_T, ou muda a estimativa de alguns coeficientes do termo de interação contendo T, ou muda a estimativa do intervalo de confiança para um destes coeficientes, então R está agindo como um confundimento na relação Y e T. Ele não envolve um teste estatístico, mas mais do que isso um julgamento de se uma mudança na estimativa do tamanho da relação Y/T é apreciavelmente afetado pela presença de R no modelo.

2.5.2. Exemplo de Caso Multivariado

Aqui será demonstrado o processo de construção do modelo de Regressão Logística num estudo apresentado por Hosmer e Lemeshow (1989) que trata da relação entre baixo peso ao nascer e várias variáveis independentes descritas na Tabela 2.8. Este exemplo será apresentado aqui para ilustrar a técnica com ênfase na interpretação dos coeficientes num modelo multivariado e também para mostrar o problema da interação e confundimento. Na Tabela 2.8 estão descritas as variáveis independentes e seus respectivos códigos.

Tabela 2.8 - Códigos das Variáveis no Conjunto de Dados de Baixo Peso ao Nascer.

Variável	Abrev.
Código de Identificação	ID
Baixo Peso ao Nascer (0 = Peso ao Nascer ≥ 2500g, 1 = Peso ao Nascer < 2500g.)	LOW
Idade da Mãe em anos	AGE
Peso no Último Período Menstrual	LWT
Raça (1 = Branca, 2 = Preta, 3 = Outra)	RACE
Fumante Durante a Gravidez (1 = Sim, 0 = Não)	SMOKE
História de Parto Prematuro (0 = Nenhum, 1 = Um, etc.)	PTL
História de Hipertensão (1 = Sim, 0 = Não)	HT
Presença de Irritação Uterina (1 = Sim, 0 = Não)	UI
Número de Visitas Médicas durante o Primeiro Trimestre (0 = Nenhuma, 1 = Uma, 2 = Duas, etc.)	FTV
Peso ao Nascer em Gramas	BWT

Os resultados do ajustamento de modelos de Regressão Logística Univariado para estes dados são dados na tabela 2.9.

Tabela 2.9 - Modelos de Regressão Logística Univariada para Dados de Baixo Peso ao Nascer.

Variável	$\hat{\beta}$	$\hat{SE}(\hat{\beta})$	$\hat{OR}$	IC a 95%	Log-Veross.	G	p
Constante	-0,790	0,157			-117,34		
AGE	-0,051	0,031	0,60	(0,32;1,11)	-115,96	2,76	0,10
LWT	-0,014	0,006	0,87	(0,77;0,98)	-114,35	5,98	0,02
RACE(1)	0,845	0,463	2,33	(0,94;5,77)	-114,83	5,01	0,08
RACE(2)	0,636	0,347	1,89	(0,96;3,73)			
SMOKE	0,704	0,319	2,02	(1,08;3,78)	-114,90	4,87	0,03
PTL	0,802	0,316	2,23	(1,20;4,14)	-113,95	6,78	0,01
HT	1,214	0,607	3,37	(1,02;11,06)	-115,33	4,02	0,04
FTV	-0,135	0,156	0,87	(0,64;1,19)	-116,95	0,77	0,38

Na tabela 2.9 para cada variável listada apresentamos a seguinte informação: O coeficiente estimado para o modelo de regressão logística univariado contendo apenas esta variável; o erro padrão do coeficiente estimado; o odds ratio (para as variáveis AGE e LWT o odds ratio foi calculado para um acréscimo de dez anos e dez libras de peso respectivamente); o IC de 95% para o odds ratio; o valor da log-verossimilhança para o modelo e a estatística do teste da razão de verossimilhança, G, para a hipótese que o coeficiente β é zero.

Com exceção da variável número de visitas ao médico (FTV), há evidência que cada variável tem alguma associação com a variável resposta, baixo peso ao nascer. Então, baseado nos resultados univariados, nós ajustamos um modelo multivariado com todas as variáveis exceto FTV. Na análise univariada, segundo Hosmer e Lemeshow, 1989, qualquer variável cujo teste univariado tem um valor $p<0,25$ deve ser considerada como candidata para o modelo multivariado junto com todas as variáveis de importância biológica. Os resultados do ajuste do modelo multivariado são dados na tabela 2.10.

Tabela 2.10 - Coeficientes Estimados, Erros Padrão Estimados e Coeficiente do Erro Padrão do Modelo Multivariado Contendo Variáveis Identificadas nas Análises Multivariadas.

Variável	Coeficiente Estimado	Erro Padrão Estimado	Coef./E.P.
AGE	-0,027	0,036	-0,74
LTW	-0,015	0,007	-2,19
RACE(1)	1,263	0,525	2,40
RACE(2)	0,862	0,438	1,97
SMOKE	0,923	0,400	2,31
PTL	0,542	0,345	1,57
HT	1,834	0,690	2,66
UI	0,759	0,459	1,65
Constante	0,464	1,201	0,38

Log-verossimilhança = -100,71

A partir dos resultados mostrados na Tabela 2.10 vemos que todas as variáveis exceto AGE demonstram considerável importância no modelo multivariado. Neste ponto do desenvolvimento do modelo nós devemos tomar uma decisão em relação a variável idade (AGE). Idade é sabido ser uma variável biologicamente importante, embora seu coeficiente não seja estatisticamente significante neste modelo de peso baixo ao nascer. Como sabe-se que idade interage com outras variáveis nós decidimos manter AGE no modelo.

Pelo resultados de uma análise mais detalhada através da análise de quartis da variável LWT (ver Hosmer e Lemeshow, 1989), decidiu-se dicotomizar a variável LWT, criando a variável LWD, da seguinte maneira: LWD = 1 se LWT está no primeiro quartil e zero caso contrário.

Decidiu-se também dicotomizar a variável número de trabalhos de parto prematuro prévios (PTL) pelo fato de que dos 189 sujeitos da amostra, 159 não tinham história a priori, 24 tiveram um parto e 6 tiveram 2 ou 3. Desta forma com a freqüencia é muito baixa no último grupo, criou-se a nova variável PTD com valor de zero se PTL é zero e 1 caso contrário.

Os resultados do ajuste do modelo multivariado com as novas variáveis LWD e PTD são dadas na tabela 2.11.

Tabela 2.11 - Estimativas Para o Modelo Multivariado Contendo LWD e PTD, Variáveis Dicotômicas Criadas a Partir de LWT e PTL.

Variável	Coeficiente Estimado	Erro Padrão Estimado	Coef./E.P.
AGE	-0,046	0,037	-1,25
LWD	0,842	0,405	2,08
RACE(1)	1,073	0,514	2,09
RACE(2)	0,815	0,444	1,84
SMOKE	0,807	0,404	2,00
PTD	1,282	0,461	2,78
HT	1,435	0,647	2,22
UI	0,658	0,466	1,41
Constante	-1,217	0,954	-1,28

Log-verossimilhança = - 98,78

Nós agora veremos a questão das interações entre as variáveis do modelo. Um total de 42 interações poderia ser formado a partir das variáveis no modelo na tabela 2.11. Nestas situações é recomendado que apenas aquelas interações com alguma razão a priori de interesse ou que faça sentido biológico sejam investigadas. No modelo que estamos tratando as variáveis AGE, RACE e SMOKE tem interação potencial com outras variáveis. Então examinaremos interação de cada uma destas variáveis com todas as outras variáveis do modelo. Sabe-se, também, que peso e hipertensão são relacionadas, portanto examinaremos esta interação também. Os resultados de adicionar cada interação ao modelo já ajustado serão apresentados na tabela 2.12. Estamos interessados em ver como a inclusão dos termos de interação alteram as estimativas apresentadas na Tabela 2.11. De maneira geral, uma interação terá de demonstrar no mínimo um nível moderado de significância estatística para esta ocorrência.

Tabela 2.12 - Teste para Possíveis Interações de Interesse Adicionado no Modelo de Efeitos Principais.

Interação	Log-Veross.	G	gl	p-value
Apenas Efeitos Principais*	-98,78			
AGExRACE	-98,53	0,50	2	0,78
AGExSMOKE	-98,51	0,54	1	0,46
AGExHT	-98,39	0,78	1	0,38
AGExUI	-98,76	0,04	1	0,84
AGExLWD	-97,50	2,56	1	0,11
AGExPTD	-98,36	0,84	1	0,36
RACExSMOKE	-97,61	2,34	2	0,31
RACExHT	-98,63	0,30	2	0,86
RACExUI	-97,62	2,32	2	0,31
RACExLWD	-97,08	3,40	2	0,18
RACExPTD	-98,50	0,56	2	0,76
SMOKExHT	-98,71	0,14	1	0,71
SMOKExUI	-98,12	1,32	1	0,25
SMOKExLWD	-97,61	2,34	1	0,13
SMOKExPTD	-98,31	0,94	1	0,33
LWDxHT	-98,22	1,12	1	0,30
AGExLWD + SMOKExLWD	-96,01	5,54	2	0,06

*Modelo de Efeitos Principais da Tabela 2.11.

Na Tabela 2.12 podemos observar que apenas duas interações são de interesse, AGExLWD e SMOKExLWD. Um modelo contendo estas interações foi ajustado. O resultado do ajustamento deste modelo está mostrado na última linha da Tabela 2.12. Este modelo parece dar uma melhora significante sobre o modelo de apenas efeitos principais. A inclusão destes termos de interação no modelo oferece a possibilidade de melhor descrever efeitos de idade, baixo peso da mãe no último período menstrual e se ela é fumante ou não fumante em relação a variável resposta baixo peso ao nascer dos bebês. Os coeficientes estimados para este modelo estão mostrados na Tabela 2.13.

Tabela 2.13 - Estimativas Para o Modelo Multivariado Contendo Efeitos Principais e Interações Significantes.

Variável	Coeficiente Estimado	Erro Padrão Estimado	Coef./E.P.
AGE	-0,084	0,046	-1,84
RACE(1)	1,083	0,519	2,09
RACE(2)	0,760	0,460	1,63
SMOKE	1,153	0,458	2,52
HT	1,359	0,662	2,05
UI	0,728	0,480	1,52
LWD	-1,730	1,868	-0,93
PTD	1,232	0,471	2,61
AGExLWD	0,147	0,083	1,78
SMOKExLWD	-1,407	0,819	-1,72
Constante	-0,512	1,088	-0,47

Log-verossimilhança = -96,01

Os odds ratios estimados e os IC a 95% para as variáveis RACE, HT, UI e PTD estão mostrados na tabela 2.14. Estes intervalos de confiança sugerem que cada uma destas variáveis afetam o baixo peso ao nascer. Embora o intervalo para UI inclui o 1, ele é muito assimétrico para direita sugerindo que a associação de UI sobre peso baixo ao nascer, após controlar para outras variáveis no modelo, possa ser considerável.

Tabela 2.14 - Odds Ratios estimados e IC a 95% para as variáveis RACE, HT, UI e PTD.

Variável	$\hat{OR}$	IC a 95%
RACE(Branca)	1,0	
RACE(Preta)	3,0	(1,1;8,2)
RACE(Outras)	2,1	(0,9;5,3)
HT	3,9	(1,1;14,2)
UI	2,1	(0,8;5,3)
PTD	3,4	(1,4;8,6)

Uma análise mais detalhada da interpretação dos odds ratio para este estudo é apresentada por Hosmer e Lemeshow, 1989. Os autores concluem que, após controlar pela variável SMOKE, baixo peso da mãe é uma variável importante na predição de peso baixo ao nascer da criança. É interessante notar que LWD era a menos significante das variáveis estudadas na Tabela 2.13. Desta forma vemos que estudar os odds ratio dentro das categorias das outras covariáveis auxilia muito a análise final do modelo de Regressão Logística principalmente quando há termos de interação no modelo.

3. PACOTES COMPUTACIONAIS PARA REGRESSÃO LOGÍSTICA

Os pacotes computacionais mais usados em Regressão Logística são os seguintes: SPSS, BMDP e MULTLR.

3.1. SPSS

O SPSS (Statistical Package for the Social Science) é um pacote estatístico muito usado e inclui a técnica de Regressão Logística na versão SPSSPC - 4.0. É um pacote bem desenvolvido e muito prático, em que o usuário especifica um conjunto de comandos relacionados ao arquivo de dados.

O SPSS ajusta um modelo aos dados estimando os coeficientes β, o erro padrão, a estatística de Wald, os graus de liberdade, a significância e a estatística R para cada variável. A estatística Wald é o quadrado da divisão do coeficiente β pelo erro padrão. A estatística R é a correlação parcial da variável dependente com cada uma das variáveis independentes. Fornece ainda um teste de "Goodness-of-fit", para verificar o melhor ajustamento aos dados e o teste de razão verossimilhança. Pode ser incluído termos com interação.

3.2. BMDP

Este pacote estatístico é o mais completo. A cada passo, uma variável contínua ou um conjunto de variáveis design são acrescentadas ou removidas do modelo. Uma regra hierárquica opcional permite uma interação dentro do modelo somente se todas as interações de ordem menor e efeitos principais estão no modelo. A seleção em cada passos está baseado na razão da máxima verossimilhança (MLR) ou na estimação da covariância assintótica (ACE). O ACE é considerado mais rápido (mais econômico) e é recomendado para a análise inicial de grandes problemas. Os resultados, incluem a função log-verossimilhança, a mudança na log-verossimilhança em relação ao passo anterior, e três estatísticas χ^2 "goodness-of-fit"; inclui histogramas de probabilidades preditas para cada grupo e classificações corretas e incorretas usando diferentes pontos de cortes nas probabilidades calculadas. Além disso, para cada padrão distinto de todas variáveis consideradas, ou para as que são realmente selecionadas, as saídas incluem freqüencias de sucessos e falhas, probabilidades preditas, proporção observada, log odds e resíduos padronizados.

3.3. MULTLR

O MULTLR é um conjunto de programas desenvolvidos em linguagem Pascal, para microcomputadores. Ele faz análise de Regressão Logística Múltipla usando estimadores de Máxima Verossimilhança, condicional e incondicional. O MULTLR calcula para cada modelo os parâmetros estimados, erros padrão, o risco relativo multiplicativo e seus intervalos de confiança, estimativas da função log-verossimilhança antes e depois da convergência e o teste "Score" e o teste da razão verossimilhança.

As variáveis fixadas podem ser independentemente fatoradas por dois diferentes métodos: O método usual de contrastar com a primeira linha usada como base (baseline) e outro permitindo contrastar entre sucessivas categorias (preceding).

O usuário pode criar um arquivo dicionário contendo todos os nomes das variáveis, pontos de corte para variáveis design e a escolha dos contrastes (categorias baseline ou preceding).

4. EXEMPLO PRÁTICO

O exemplo que utilizaremos neste capítulo foi gentilmente cedido pela professora Vera Beatriz Wald do Departamento de Medicina Animal da UFRGS. O objetivo aqui é mostrar a utilização da técnica de regressão logística em dados de Reprodução Animal. A análise destes dados bem como a interpretação dos resultados foram realizados pela Professora Vera Beatriz Wald.

Na área de Reprodução Animal são freqüentes observações de variáveis binárias como cio, gestação, morte embrionária e viabilidade de embriões. É de interesse analisar estas variáveis resposta com variáveis de diversos níveis de medida como idade, tempo de ovulação, peso, raça, entre outras.

Os dados utilizados nas análises são provenientes de um estudo do comportamento reprodutivo pós-parto de éguas Puro Sangue Inglês no Rio Grande do Sul, nos anos 1989 à 1991. Foram analisados o efeito de grau de involução uterina pós parto, idade e tratamento com hormônios na prenhez. Os animais foram classificados conforme a idade, em tres faixas etárias: 150 fêmeas com menos de sete anos, 149 fêmeas de sete a dez anos e 147 fêmeas com mais de dez anos. Pelo exame ginecológico o grau de involução uterina foi classificado como "involuído", "involução média" e "sem involução". Conforme o grau de involução e idade, os animais foram divididos em 2 grupos, sendo que em 1 aplicou-se um hormônio após parto e o outro foi considerado como grupo controle. A fecundidade (variável resposta) foi avaliada pelo número de animais gestântes. O delineamento bem como os resultados encontram-se na tabela 4.1.

Tabela 4.1 - Matriz dos dados originais.

IDADE	ÚTERO	HORMÔNIO		CONTROLE		TOTAL	
		n	Prenhez	n	Prenhez	n	Prenhez
1	1	25	22	25	16	50	38
	2	25	22	25	11	50	33
	3	25	23	25	14	50	37
2	1	25	21	25	23	50	44
	2	25	21	25	24	50	45
	3	24	21	25	22	49	43
3	1	25	22	25	20	50	42
	2	24	21	24	16	48	37
	3	25	22	24	16	49	38
TOTAL		223	195	223	162	446	357

Os dados foram processados utilizando-se os programas MULTLR (CAMPOS FILHO & FRANCO, 1989) - Regressão Logística Incondicional - e SPSS-PC (Statistical Packege for the Social Science) - Qui-quadrado.

A tabela abaixo apresenta as variáveis e seus respectivos códigos.

Tabela 4.2 - Códigos das variáveis utilizadas.

Variável	Código	Abreviatura
Gestação	0 - não 1 - sim	PRENHEZ
Idade	1 - <7 anos 2 - 7 a 10 anos 3 - >10 anos	IDADE
Grau de involução uterina	1 - involuido 2 - involução média 3 - não involuido	ÚTERO
Tratamento	0 - controle 1 - com hormônio	TRAT

Os dados foram analisados, inicialmente, através da estatística Qui-quadrado de Pearson, com o bjetivo de verificar as associações significativas entre as variáveis. Conforme SILVA (1990) este procedimento é adequado para avaliação de um sistema casual. Feita essa abordagem verificou-se a associação entre prenhez e idade, tabela 4.3, sendo que pela análise dos resíduos padronizados nas éguas jovens (idade 1) a frequencia de animais não gestantes é maior. O teste do Qui-quadrado comparando a prenhez com grau de involução uterina, tabela 4.4, não mostrou associação entre as variáveis. Enquanto que, de outra forma, o tratamento com hormônio indicou associação positiva com a prenhez, tabela 4.5.

Tabela 4.3 - Prenhez x Idade

	1	2	3	Total
0	42 2,2	17 -2.3	30 0,1	89 20,0%
1	108 -1,1	132 -1,1	117 -0,1	357 80,0%
Total	150 33,6%	149 33,4%	147 33,0%	446 100,0%

Chi-Square	D.F.	Significance	Min E.F.
12,90895	2	0,0016	29,334

Tabela 4.4 - Prenhez x Útero

	1	2	3	Total
0	26 -0,7	33 0,6	30 0,1	89 20,0%
1	124 0,4	115 -0,3	118 -0,0	357 80,0%
Total	150 33,6%	148 33,2%	148 33,2%	446 100,0%

Chi-Square	D.F.	Significance	Min E.F.
1,16300	2	0,5591	29,534

Tabela 4.5 - Prenhez x Tratamento

	1	2	Total
0	61 2,5	28 -2.5	89 20,0%
1	162 -1,2	195 1,2	357 80,0%
Total	223 50,0%	223 50,0%	446 100,0%

Chi-Square	D.F.	Significance	Min E.F.
14,37397	1	0,0002	44,500

Após a análise das direções das associações entre as variáveis, pelo teste do Qui-quadrado, utilizou-se o Modelo de Regressão Logística aos dados.

A fatoração da variável idade, para a regressão logística, foi feita de modo a obter contraste entre as categorias adjacentes, por se tratar de uma variável intervalar. Já na variável grau de involução uterina, com tres categorias, utilizou-se como categoria de referencia a categoria involuído, biologicamente definida como normal.

A Tabela 4.5 apresenta os coeficientes angulares do modelo logístico univariado, o erro padrão estimado, o Teste Wald (Z-Score) e o Teste da razão de verossimilhança (G).

Tabela 4.6 - Modelos de Regressão Logística Univariados.

VARIÁVEL	B	SE(B)	Z-SCORE	P	G	P
CONSTANTE	1,39	0,12				
TRAT	0,96	0,25	3,83	0,0001	15,59	0,0001
IDADE					13,36	0,0013
IDADE(2)	1,11	0,32	3,50	0,0005		
IDADE(3)	-0,69	0,33	-2,09	0,0364		
UTERO					1,17	0,5566
UTERO(2)	-0,31	0,29	-1,07	0,2833		
UTERO(3)	-0,19	0,30	0,65	0,5168		

Pela observação dos testes Wald e G verifica-se que as variáveis idade e tratamento foram significantes (p<0,05).

Conforme MICKEY & GREENLAND (1989) com a utilização de níveis tradicionais de significância, como 0,05, variáveis importantes podem não ser identificadas. Os autores sugerem a utilização de um nível de significância de 0,25 como critério para seleção. Observa-se que na regressão logística modelos uni e multivariado (Tabela 4.7), semelhante ao obtido na análise com o Qui-quadrado, o grau de involução uterina não apresenta associação com a gestação.

Tabela 4.7 - Modelo de Regressão Logística Multivariado.

VARIÁVEL	B	SE(B)	Z-SCORE	P
TRAT	0,99	0,26	3,90	0,0001
UTERO				
UTERO(2)	-0,34	0,30	-1,11	0,2663
UTERO(3)	-0,20	0,31	-0,66	0,5076
IDADE				
IDADE(2)	1,15	0,32	3,58	0,0003
IDADE(3)	-0,72	0,33	-2,15	0,0313
CONSTANTE	0,68	0,28	2,45	0,0144

Ajustou-se novo modelo multivariado (Tabela 8), eliminando a variável útero.

Tabela 4.8 - Modelo de Regressão Logística Multivariado resultante da retirada da variável Útero.

VARIÁVEL	B	SE(B)	Z-SCORE	P
TRAT	0,99	0,26	3,89	0,0001
IDADE				
IDADE(2)	1,15	0,32	3,58	0,0003
IDADE(3)	-0,72	0,33	-2,14	0,0320
CONSTANTE	0,50	0,21	2,37	0,0176

Goodness of fit p=0,098

O teste da razão de verossimilhança entre os modelos das tabelas 4.7 e 4.8 (um teste de significância para o útero), resultou em um valor de G=1,256, p=0,53, demonstrando que a variável útero adiciona muito pouco ao modelo. Também verifica-se que os coeficientes estimados para as outras variáveis são

idênticas nos dois modelos. Assim sendo, conforme HOSMER & LEMERSHOW (1989), há inexistência do fator de confundimento.

Procurando-se verificar prováveis interações entre tratamento hormonal com a faixa etária, ajustou-se novo modelo (Tabela 4.9). Nota-se que há interação entre idade e tratamento, ou seja o efeito do tratamento é condicionado pela idade. Comparando os modelos das Tabelas 4.8 e 4.9, constata-se uma melhora no goodness of fit,sugerindo que o modelo da Tabela 4.9 ajusta-se melhor aos dados.

Tabela 4.9 - Modelo Resultante das Variáveis Selecionadas e suas Interações.

VARIAVEL	B	SE(B)	Z-SCORE	P
TRAT	1,94	0,44	4,40	0,0000
IDADE				
IDADE(2)	2,25	0,48	4,65	0,0000
IDADE(3)	-1,54	0,50	-3,08	0,0020
IDA(2)*TRAT	-2,64	0,69	-3,80	0,0001
IDA(3)*TRAT	1,77	0,69	2,54	0,0108
CONSTANTE	0,19	0,23	0,81	0,4196

Goodness of fit p=0,41

A utilização de modelos de Regressão Logística em estudos de reprodução animal mostrou-se um instrumento muito útil, pois além de mostrar as direções de associação entre as variáveis individualmente, de maneira análoga ao Qui-quadrado, evidencia as interações entre as variáveis. Se estas interações não fossem percebidas teríamos uma visão compartimentada e parcial. Como é o caso do presente estudo, onde a análise univariada não mostra que nos animais de segunda idade, o tratamento com hormônio não é efetivo. A percepção de tal efeito é importante não só estatisticamente, como também sob o ponto de vista do manejo reprodutivo, orientando em qual faixa etária deve-se ministrar o hormônio.

5. REFERÊNCIAS BIBLIOGRAFICAS

AGRESTI, A. (1990). **Categorical Data Analysis**. New York: John Wiley & Sons.

CHAMBLESS, L. E. (1990). Biostatistics for Epidemiologists. Notas de aula do curso Introdução aos Métodos Quantitativos em Saúde. Faculdade de Medicina - UFRGS. Porto Alegre. (não publicado)

HOSMER, D. W. e LEMESHOW, S. (1989). **Applied Logistic Regression**. New York: John Wiley & Sons, Inc.

SPSS Inc. (1989). **SPSS/PC+ Update for V3.0 and V3.1**. SPSS Inc., Chicago.

CAMPOS-FILHO, N. and FRANCO, E. L. (1989). A Microcomputer Program for Multiple Logistic Regression by Unconditional and Conditional Maximum Likelihood Methods. **Amer J. Epidemiol. 129**(2): 439-444.

CAMPOS-FILHO, N. and FRANCO, E. L. (1988). **MULTLR, User's Manual**. Ludwig Institute for Cancer Research. São Paulo.

DIXON, W. J. (1985). **BMDP Statistical Software**. University of California Press, Berkeley.