TITULO DA TESE

COMPARAÇÃO DO DESEMPENHO DE TESTES PARA O RISCO RELATIVO

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pelo Sr. TEREZA NADYA LIMA DOS SANTOS

e aprovada pela Comissão Julgadora.

Campinas, 21 de março de 1989

Prof. Dr. _____
Orientador

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do Título de Mestre em ESTATISTICA

Ao Isaac

A meus filhos

A Mônica

Agradeço a Isaac pelo seu apoio, a Mônica pelo seu carinho e abnegação para comigo e a meu orientador Norberto Dachs sempre atencioso e amigo durante todo o período em que trabalhei nesta Dissertação.

INTRODUÇÃO

Estudos estratificados de casos e controles nos quais os dados, ao final da amostragem, são tabelas 2x2, são estudos dos quais estamos interessados em conhecer o grau de associação entre um resultado e um fator que, desconfiamos, tenha influência sobre esse resultado. Por causa dessa influência, tal fator é chamado comumente de "fator de risco" e é de grande interesse testar a significância do grau de associação entre o mesmo e o resultado, ou seja, a significância do risco relativo.

O risco relativo é uma medida bem conhecida, particularmente pelos profissionais em Estatística e Medicina como uma medida do grau de associação entre um fator de risco e um resultado, que pode ser uma doença. Usualmente nos estudos citados no parágrafo anterior o risco relativo é substituido por uma outra medida de associação, a razão de produtos cruzados, a qual, além de ser uma boa aproximação para ele, tem a vantagem de ter uma fórmula bem mais simples do que a fórmula do risco relativo.

Se o estudo estratificado é de casos e controles, existem dois procedimentos pelos quais podemos testar a significância da razão de produtos cruzados em tabelas 2x2xm, quando em uma etapa anterior do estudo, aceitamos a hipótese de que há uma razão de produtos cruzados constante, comum a todos os estratos. Um deles é o procedimento de Mantel-Haenszel, frequentemente usado para estimar uma razão de produtos cruzados comum e testar sua significância. O outro procedimento é o de máxima verossimilhança. Esse consiste, inicialmente, em ajustar um modelo de regressão logística linear na variável de risco e nas possíveis variáveis de confundimento Nesse primeiro passo do procedimento, propomos um

modelo para o problema em questão, compreendendo que nesse modelo o parâmetro que é o coeficiente da variável de risco é o logarítimo natural da razão de produtos cruzados ("odds ratio"), suposta comum a todos os estratos. No segundo passo, o teste da razão de verossimilhança é usado para testar a significância do risco relativo, o que é a mesma coisa que testar a hipótese de igualdade a zero do coeficiente da variável de risco no modelo de regressão logística. Dessa forma, a hipótese do teste de Mantel-Haenszel e a hipótese do teste da razão de verossimilhança são equivalentes e é de interesse comparar os poderes desses dois testes.

O principal interesse desse trabalho é comparar o poder do teste de Mantel_Haenszel com o poder do teste equivalente usando regressão logística, respondendo a questão "que procedimento tem maior poder nas mesmas condições? ". Aquele que tiver maior poder será preferido para testar a significância da associação entre o fator de risco e o resultado. Outro interesse desse trabalho, embora não seja o principal, é estudar a adequação da aproximação de Muñoz e Rosner (1984) para o poder do teste de Mantel-Haenszel, em situações desbalanceadas. O que ocorre é que essa aproximação se baseia na distribuição normal. Para situações desbalanceadas, a distribuição dos dados nas tabelas 2x2 de um estudo de casos e controles é assimétrica e uma aproximação normal pode não ser ideal.

No Capítulo 1, apresentamos um resumo sobre os métodos de amostragem que geram dados na forma de tabelas 2x2 e uma discussão sobre a importância do risco relativo e como a razão de produtos cruzados é uma boa aproximação para ele em estudos com dados na forma acima.

Nós discutimos no Capítulo 2, os dois procedimentos que pretendemos comparar. Discutimos com detalhes os dois modelos de regressão logística que formam a base para o teste da razão de verossimilhança e a equivalência entre o teste de Mantel-Haenszel

e o teste de igualdade a zero do coeficiente da variável de risco em um desses modelos.

Na impossibilidade de fazermos um estudo comparativo entre o poder dos dois testes, de forma analítica, faremos estudos de Monte Carlo para gerar amostras de estudos de casos e controles, assim como para obter as estimativas e os testes necessários. Utilizamos também esses estudos de Monte Carlo para estudarmos a adequação da aproximação de Muñoz Rosner. Apresentamos esses estudos de Monte Carlo, detalhes, no Capítulo 3.

No Capítulo 4, nós comentamos sobre um engano no nível de significância adotado por Muñoz e Rosner e concluimos que os obtidos pelos mesmos são equivalentes aos resultados de testes bilaterais a um nível de 0.10. Prosseguindo nesse capítulo, utilizamos os resultados obtidos pelo Monte Carlo para discutirmos a adequação da aproximação de Muñoz e Rosner. Mostramos que em algumas das situações mais desbalanceadas, em que as proporções de casos e de expostos nos dois estratos se distanciam de 0.5, essa aproximação superestima o poder do teste de Mantel-Haenszel. Finalmente observamos pelos resultados do Monte Carlo que, também para situações o mais desbalanceadas possíveis, a regressão logística tem maior poder do que o teste de Mantel-Haenszel quando o estudo tem um número de estratos igual a dois. Observamos também que para três estratos, os dois testes têm o mesmo poder.

CAPÍTULO 1

MÉTODOS DE AMOSTRAGEM QUE GERAM TABELAS 2x2xM E A RAZAO DE PRODUTOS CRUZADOS PARA APROXIMAR O RISCO RELATIVO

1.1. MÉTODOS DE AMOSTRAGEM QUE GERAM TABELAS 2X2XM

Em muitos estudos, a estrutura dos dados no final da amostragem toma a forma de uma coleção de tabelas 2x2. É o que ocorre, por exemplo quando é necessário estratificar os resultados de acordo com os valores de uma variável de confundimento, de modo que cada uma das tabelas da coleção contém informação sobre os casos e controles para um particular nível da variável de confundimento. Por exemplo, para testar a conjectura de que o uso de anticoncepcionais orais leva a um maior risco de câncer de mama, podemos estratificar o estudo por faixas (ou níveis) de idade, desde que seja conhecido que a proporção de usuários correntes desses anticoncepcionais decresce com a idade enquanto que a taxa de incidência de câncer de mama cresce com a idade, como em Muñoz e Rosner (1984). Em estudos como esse as mulheres podem ser classificadas em m faixas de idade e a cada faixa de idade corresponde uma tabela 2x2. As informações contidas em cada tabela são os números observados de mulheres que contraíram a doença e que são ou foram durante algum tempo usuárias de anticoncepcionais, de mulheres doentes não usuárias, de mulheres não doentes usuárias e de mulheres não doentes não usuárias. Cada mulher doente nós classificamos como um caso e cada não doente nós classificamos como um controle. Na conjectura colocada no início

desse exemplo, anticoncepcionais orais são um fator que pode levar a um maior risco de câncer de mama. Na literatura é comum denominar um fator com essa característica, de fator de exposição ou fator de risco e os indivíduos para os quais se verifica a presença de tal fator são ditos expostos. Os indivíduos para os quais se verifica a ausência do fator são ditos não expostos. Assim ao final da amostragem os dados têm a estrutura da Tabela 1.1.1.

Tabela 1.1.1. Estrutura dos dados de um estudo estratificado com fator de risco e resultados dicotômicos

	estrato 1				estrato m	
	casos	controles			casos	controles
expostos	a_1	b_1	....	expostos	a_m	b_m
não expostos	c_1	d_1		não expostos	c_m	d_m

A estratificação da população em estudo é também, muitas vezes, feita com base em combinações de níveis de várias variáveis de confundimento.

Existem, basicamente, três métodos de amostragem que geram tabelas 2x2. No primeiro deles fixamos o número total N de indivíduos a serem amostrados e observamos como resultado aleatório os números de indivíduos que são casos expostos, casos não expostos, controles expostos e controles não expostos. Desse modo só o total da única tabela resultante do estudo é fixado. Os demais valores que compõem a tabela, inclusive os totais marginais de linhas e colunas, são aleatórios e a distribuição de probabilidade dos dados é multinomial. Em estudos estratificados, além de fixarmos o número total N de indivíduos a serem amostrados, fixamos também o número de indivíduos a serem

amostrados no i-ésimo estrato, N_i, dado por uma fração t_i de N. Assim, $N_i = t_i N$. Obviamente, se temos m estratos, $\sum_{i=1}^{m} t_i = 1$, para que a soma dos N_i's seja igual a N. Em cada estrato, a distribuição dos dados é multinomial.

O segundo método é o que aparece em estudos de coorte. Nesse método de amostragem nós fixamos o número total de indivíduos a serem amostrados; fixamos a fração t_i para dar $N_i = t_i N$, o número de indivíduos a serem amostrados no i-ésimo estrato, i=1,...,m e também fixamos N_{1i}, o número de indivíduos expostos a serem amostrados no i-ésimo estrato através de uma fração r_i de N_i. Assim, $N_{1i} = r_i N_i = r_i t_i N$. Quando fixamos N_i e N_{1i}, o número de indivíduos não expostos a serem amostrados, N_{2i}, também fica fixado, através de $N_{2i} = (1 - r_i) t_i N$, i=1,...,m. Em seguida, retiramos uma amostra de tamanho N_{1i} de expostos e uma amostra de tamanho N_{2i} de não expostos e observamos como resultado (aleatório) quantos indivíduos são casos e quantos são não casos (controles) em cada uma dessas amostras. As entradas da i-ésima tabela ficam como na Tabela 1.1.2 abaixo.

Tabela 1.1.2. Entradas da i-ésima tabela oriunda de um estudo coorte estratificado

	expostos	não expostos	
casos	a_i	b_i	M_i
controles	c_i	d_i	$N_i - M_i$
	N_{1i}	N_{2i}	

Resulta desse esquema de amostragem que a distribuição dos dados na i-ésima tabela é obtida de duas binomiais, $B(N_{1i}, p_{1i})$ e $B(N_{2i}, p_{2i})$, onde p_{1i} e p_{2i} são, respectivamente, a probabilidade

de que um indivíduo do i-ésimo estrato seja um caso dado que foi exposto e a probabilidade de que um indivíduo do i-ésimo estrato seja um caso dado que foi não exposto.

O terceiro método de amostragem é o que serve de base para estudos de casos e controles. Neste método de amostragem, fixamos N, o tamanho total da amostra; fixamos N_i, o número total de indivíduos a serem amostrados no i-ésimo estrato ,através de uma fração t_i de N e fixamos também M_i, o número de indivíduos que são casos a serem amostrados através de uma fração s_i de N_i. Desse modo, ficamos com $N_i = t_i N$ e $M_i = s_i N_i = s_i t_i N$, $i=1,\ldots,m$. O número de indivíduos que são controles a serem amostrados fica fixado através de $N_i - M_i = (1-s_i)t_i N$, $i=1,\ldots,m$. Então retiramos uma amostra de tamanho M_i de casos e uma amostra de tamanho $N_i - M_i$ de controles em cada estrato, e observamos como resultado aleatório seu passado em relação à exposição ao fator de risco em estudo, isto é, observamos como resultado aleatório se cada indivíduo é classificado como exposto ou não exposto. As entradas da i-ésima tabela resultante desse delineamento ficam como na Tabela 1.1.3 abaixo.

Tabela 1.1.3. Entradas da i-ésima tabela oriunda de um estudo de casos e controles estratificado

	casos	controles	
expostos	a_i	b_i	N_{1i}
não expostos	c_i	d_i	N_{2i}
	M_i	$N_i - M_i$	N_i

Nesta tabela, a_i é o número de casos expostos, b_i é o número de controles expostos, c_i é o número de casos não expostos e d_i é

o número de controles não expostos, observados na amostra do i-ésimo estrato, i=1,...,m. A distribuição dos dados no i-ésimo estrato é obtida de duas binomiais, $B(M_i, p_{1i})$ e $B(N_i - M_i, p_{2i})$, onde p_{1i} e p_{2i} são, respectivamente, a probabilidade de que um indivíduo tenha estado exposto ao fator de risco para o grupo de casos e a probabilidade de que um indivíduo tenha estado exposto ao fator de risco para o grupo de controles.

1.2. O RISCO RELATIVO E A RAZÃO DE PRODUTOS CRUZADOS COMO APROXIMAÇÃO PARA O RISCO RELATIVO

O risco relativo é bem conhecido entre profissionais em Estatística e em Medicina como uma medida de associação entre um fator de risco e uma resultado, em geral indesejado, que pode ser uma doença ou morte. Se nós consideramos o problema de medir o grau de associação entre um fator de risco e essa doença, podemos definir vários tipos de riscos relativos. Por exemplo, o risco relativo de contrair uma doença, ou seja, de vir a ter a doença no decorrer de um dado período de tempo ou de desenvolver a mesma doença nesse período. Os estudos retrospectivos geralmente fornecem somente estimativas do risco relativo de ter a doença durante um intervalo de tempo especificado (J. Cornfield, 1956). Esse tipo de risco relativo é também denominado de prevalência relativa. Se denotarmos por P a proporção da população no grupo de casos (coeficiente de prevalência), por p_1 a proporção do grupo de casos que caem numa categoria x do fator de risco e por p_2 a proporção do grupo de controles que caem nessa mesma categoria x do fator de risco, então o coeficiente de prevalência para a doença para pessoas pertencentes à categoria x do fator de risco é

$$\frac{p_1 P}{p_1 P + p_2(1-P)} \qquad (1.2.1)$$

Para indivíduos não pertencentes à categoria x a taxa de prevalência é

$$\frac{(1-p_1)P}{(1-p_1)P + (1-p_2)(1-P)} \qquad (1.2.2)$$

e a prevalência relativa é

$$\frac{p_1}{(1-p_1)} \cdot \frac{(1-p_1)P + (1-p_2)(1-P)}{p_1 P + p_2(1-P)} \qquad (1.2.3)$$

Em muitas situações, P é suficientemente pequena com relação a p_2 e a $(1-p_2)$. Desse modo, (1.2.3) pode ser bem aproximada pela razão de produtos cruzados

$$\exp(\gamma) = \frac{p_1 / (1-p_1)}{p_2 / (1-p_2)} \qquad (1.2.4)$$

onde γ é o logaritmo natural da razão de produtos cruzados. A razão de produtos cruzados foi originariamente proposta por Cornfield (1956) como uma medida do grau de associação entre um fator de risco e um resultado. Ela é uma forma de estimar a prevalência relativa e fornece uma boa aproximação para o risco relativo, também proposto por Cornfield, quando a incidência do resultado é pequena, tipicamente inferior a 10%. Cornfield se refere a (1.2.4) como o risco relativo de um indivíduo ter uma doença, no caso em que desejamos estudar a associação entre um fator de risco (ou fator de exposição) e a doença. Se $p_1=p_2$,

$\exp(\gamma)=1$, indicando falta de associação entre o fator de risco e o resultado de interesse. Da maneira como definimos p_1 e p_2 anteriormente, essas são, respectivamente, a probabilidade de exposição para os indivíduos que são casos e a probabilidade de exposição para os indivíduos que são controles. Em estudos de coorte, define-se probabilidades p_1 e p_2, respectivamente, como a probabilidade de que um indivíduo exposto venha a ter a doença e a probabilidade de que um indivíduo não exposto venha a ter a doença. Existindo um único fator de risco dicotômico, a razão de produtos cruzados para estudos de coorte é idêntica à razão de produtos cruzados de estudos de casos e controles. Para ver isso, consideremos as Tabelas 1.2.1 e 1.2.2 a seguir. A Tabela 1.2.1 se refere a estudos de coorte e a Tabela 1.2.2 a estudos de casos e controles.

Tabela 1.2.1. Proporções de expostos e de não expostos num estudo de coorte

	expostos	não expostos
casos	p_1	p_2
controles	$1-p_1$	$1-p_2$

Tabela 1.2.2. Proporções de casos e de controles num estudo de casos e controles

	casos	controles
expostos	p_1	p_2
não expostos	$1-p_1$	$1-p_2$

Na Tabela 1.2.1, p_1 estima a probabilidade de caso dado que houve exposição e p_2 estima a probabilidade de caso dado não exposição. Na Tabela 1.2.2, p_1 estima a probabilidade de exposição dado caso e p_2 a probabilidade de exposição dado não caso (controle). Mas, como pode ser visto em Breslow e Day (1980, Capítulo 2), em qualquer um dos dois tipos de estudo, a razão de produtos cruzados é dada pela expressão (1.2.4).

Se a estratificação que serviu de base para um estudo de coorte em uma população é a mesma estratificação que serviu de base para um estudo de casos e controles, na mesma população, no i-ésimo estrato definimos p_{1i}, p_{2i}, $1-p_{1i}$ e $1-p_{2i}$ para estudos de coorte da mesma maneira que definimos, respectivamente, p_1, p_2, $1-p_1$ e $1-p_2$, na Tabela 1.2.1, isto é, p_{1i} estima a probabilidade de caso dado exposição e p_{2i} estima a probabilidade de caso dado não exposição no i-ésimo estrato. Para estudos de casos e controles, definimos p_{1i}, p_{2i}, $1-p_{1i}$ e $1-p_{2i}$ da mesma maneira que definimos, respectivamente, p_1, p_2, $1-p_1$ e $1-p_2$, na Tabela 1.2.2. Assim, no i-ésimo estrato, seja de um estudo de coorte ou seja de um estudo de casos e controles, temos uma tabela que fornece a razão de produtos cruzados

$$\exp(\gamma) = \frac{p_{1i}(1-p_{2i})}{p_{2i}(1-p_{1i})}, \quad i=1,\ldots,m \qquad (1.2.5)$$

Portanto, a identidade da razão de produtos cruzados para delineamentos de estudos de coorte e de casos e controles feitos na mesma população, continua valendo também para estudos estratificados. Então quando fazemos inferências sobre o risco relativo com base em tabelas 2x2 usando a razão de produtos cruzados em estudos estratificados, não faz diferença se os totais marginais N_{1i} e N_{2i} é que são fixados, como em delineamentos coorte, ou se os totais marginais M_i e N_i-M_i é que são fixados, como nos estudos de casos e controles, e essas inferências podem ser feitas aplicando o mesmo conjunto de cálculos que seriam aplicados a dados de um estudo coorte da mesma população.

CAPÍTULO 2

O PROCEDIMENTO DE MANTEL-HAENSZEL E A REGRESSÃO LOGÍSTICA PARA ESTUDOS DE CASOS E CONTROLES ESTRATIFICADOS

2.1. O PROCEDIMENTO DE MANTEL-HAENSZEL

Consideraremos os dados obtidos de um estudo de casos e controles estratificado para investigar o risco associado com algum fator de risco de interesse . Com a suposição de que a razão de produtos cruzados (risco relativo) é a mesma nas m tabelas oriundas desse estudo, a hipótese testada no procedimento de Mantel-Haenszel é a de que a razão de produtos cruzados comum é um, ou equivalentemente,que seu logaritmo natural é zero. Essa é a hipótese de não associação entre o fator e a resposta. Com as entradas da i-ésima tabela como visto na Tabela 1.1.3, num teste bilateral de nível α do tipo

$$Ho : \exp(\gamma) = 1, \qquad (2.1.1)$$

segundo o procedimento de Mantel-Haenszel, rejeitamos Ho se observamos

$$\frac{\left(\left| \sum_{i=1}^{m} a_i - \sum_{i=1}^{m} \frac{M_i N_{1i}}{N_i} \right| - \frac{1}{2} \right)^2}{\sum_{i=1}^{m} \frac{M_i (N_i - M_i) N_{1i} N_{2i}}{N_i^2 (N_i - 1)}} > Z^2_{\alpha/2} \qquad (2.1.2)$$

onde $Z_{\alpha/2}$ é o $\alpha/2$ percentil superior da distribuição normal

padrão, isto é, $\Phi(Z_{\alpha/2})=1-\alpha/2$ e a_i é vista como uma realização da variável aleatória A_i com distribuição hipergeométrica não central ou "extendida". Essa é também a distribuição de qualquer uma das caselas das tabelas 2x2 quando se supõe que os totais marginais são fixados.

Hipergeométrica extendida é o nome que Harkness (1965) deu à distribuição condicional da variável aleatória binomial X_1 com parâmetros n_1 e p_1 dado que $X_1+X_2 = m$, onde X_2 é também binomial, mas com parâmetros n_2 e p_2. Em um delineamento de casos e controles estratificado gerando m tabelas como a Tabela 1.1.3, X_1 tem distribuição binomial com parâmetros M_i e p_{1i} e X_2 tem distribuição binomial com parâmetros (N_i-M_i) e p_{2i} na i-ésima tabela, para $i=1,\ldots,m$. Desse modo, a função de probabilidade de A_i é dada por

$$P_\gamma(A_i=a_i) = \frac{\binom{N_{1i}}{a_i}\binom{N_{2i}}{M_i-a_i}\left(\exp(\gamma)\right)^{a_i}}{\sum_{x=a}^{b}\binom{N_{1i}}{x}\binom{N_{2i}}{M_i-x}\left(\exp(\gamma)\right)^{a_i}} \qquad (2.1.3)$$

onde

$$a = \max(0, M_i-N_{2i}) \le a_i \le \min(M_i, N_{1i}) = b. \qquad (2.1.4)$$

Quando $\gamma=0$, implicando que $\exp(\gamma)=1$, (2.1.3) se reduz à função de probabilidade de uma variável aleatória com distribuição de probabilidade hipergeométrica ordinária ou central, como chamam alguns autores:

$$P_{\gamma=0}(A_i = a_i) = \frac{\binom{N_{1i}}{a_i}\binom{N_{2i}}{M_i-a_i}}{\binom{N_i}{M_i}}.$$

Nesse caso, a média e a variância de A_i, as quais denotamos por $\mu_i(\gamma)$ e $\sigma_i^2(\gamma)$, respectivamente, tornam-se

$$\mu_i(\gamma) = \frac{M_i N_{1i}}{N_i} = \frac{t_i r_i N \; t_i s_i N}{t_i N} = N t_i r_i s_i$$

e

$$\sigma_i^2(\gamma) = \frac{M_i(N_i - M_i) N_{1i} N_{2i}}{N_i^2(N_i - 1)}$$

$$= N r_i (1 - r_i) s_i (1 - s_i) t_i \left\{ \frac{t_i N}{t_i N - 1} \right\}$$

Fazendo $A. = \sum_{i=1}^{m} A_i$, temos que a média de A.,

$$\mu(\gamma) = N \sum_{i=1}^{m} t_i r_i s_i$$

é o segundo termo dentro do módulo no numerador de (2.1.2), o que pela independência dos A_i, a variância de A. é dada por

$$\sigma^2(\gamma) = N \sum_{i=1}^{m} r_i (1-r_i) s_i (1-s_i) t_i \left\{ \frac{t_i N}{t_i N - 1} \right\}$$

e é o denominador de (2.1.2). Assim, a desigualdade (2.1.2) pode ser escrita na forma

$$\frac{\left(\left| a. - N \sum_{i=1}^{m} t_i r_i s_i \right| - \frac{1}{2} \right)^2}{N \sum_{i=1}^{m} r_i (1-r_i) s_i (1-s_i) t_i \left\{ \frac{t_i}{t_i N - 1} \right\}} > Z^2_{\alpha/2} \qquad (2.1.5)$$

onde a. é uma realização da variável aleatória A.. A distribuição de A. é uma m-convolução das m variáveis aleatórias A_i ,para $i=1,\ldots,m$.

A estatística no lado esquerdo de (2.1.2) (ou (2.1.5)) tem distribuição qui-quadrado com 1 grau de liberdade. O fator de correção para continuidade 1/2 inserido na estatística leva em consideração o fato de que uma distribuição contínua está sendo usada para representar a distribuição discreta de A.. Há estudos sobre o efeito dessa correção feitos por Pearson(1947), Mote,Pavate e Anderson(1958) e Placket(1964). Baseados nesses estudos e em suas próprias análises, Grizzle(1967) e Conover(1968,1974) recomendam que essa correção não seja aplicada. Eles dão como uma razão para isso uma aparente diminuição do nível de significância quando a correção é usada. Essa diminuição resulta em uma redução no poder do teste, isto é, uma redução na probabilidade de detectar uma real associação entre o resultado (caso ou controle) e a exposição. Nesse trabalho não usamos a correção.

Sem a correção de continuidade, como o fazem Muñoz e Rosner (1984), a inequação (2.1.5) para um teste da hipótese de que γ é igual a zero contra a hipótese alternativa unilateral de que γ é maior do que zero fica

$$\frac{(a_{\cdot} - N\sum_{i=1}^{m} t_i r_i s_i)}{\left[N\sum_{i=1}^{m} t_i r_i (1-r_i) s_i (1-s_i)(t_i N/t_i N) \right]^{1/2}} > Z_{\alpha}, \qquad (2.1.6)$$

onde Z_{α} é tal que $\Phi(Z_{\alpha})=1-\alpha$ e $a_{\cdot} = \sum_{i=1}^{m} a_i$.

2.2 APROXIMAÇÕES PARA O PODER DO TESTE DE MANTEL-HAENSZEL

2.2.1. Poder do Teste de Mantel - Haenszel

Para testarmos a hipótese H_0 de (2.1.1) contra alguma

hipótese alternativa, com o uso da distribuição (2.1.3), a nossa atenção se concentra na estatística suficiente A.. Assim, para testar H_0 contra, digamos, $H_1: \exp(\gamma) > 1$, a região crítica de um teste uniformemente mais poderoso consistiria de valores a. na cauda superior da distribuição de A.(D. R. Cox, 1970, cap. 4). Denotando por Π a função poder de um tal teste unilateral com nível de significância α, temos

$$\begin{aligned} \alpha &= \sup_{H_0} \Pi(\gamma) = \Pi(0) \\ &= P(\text{rejeitar } H_0 / \gamma = 0) \\ &= P_{\gamma=0}(A. > c_n(\alpha)), \end{aligned} \qquad (2.2.1.1)$$

onde $c_n(\alpha)$ é tal que $P_{\gamma=0}(A. > c_n(\alpha)) = \alpha$. A expressão (2.2.1.1) é (aproximadamente) equivalente a

$$\alpha = P\left[\frac{A. - \mu_{A.}(0)}{\sqrt{\sigma^2_{A.}(0)}} > \frac{c_n(\alpha) - \mu_{A.}(0)}{\sqrt{\sigma^2_{A.}(0)}} \right]$$

$$= P\left[Z > \frac{c_n(\alpha) - \mu_{A.}(0)}{\sqrt{\sigma^2_{A.}(0)}} \right].$$

Fazendo

$$z_\alpha = \frac{c_n(\alpha) - \mu_{A.}(0)}{\sqrt{\sigma^2_{A.}(0)}},$$

temos que

$$\alpha = P(Z > z_\alpha),$$

onde z_α é o α-percentil superior da distribuição normal padrão, e encontramos que

$$c_n(\alpha) = z_\alpha \sqrt{\sigma^2_{A.}(0)} + \mu_{A.}(0).$$

Então para $\gamma>0$ ou $\exp(\gamma)>1$ (na hipótese alternativa), o poder do teste de Mantel - Haenszel unilateral com nível de significância α é dado aproximadamente por

$$\Pi_\gamma(\alpha) = P\left[\frac{A. - \mu_{A.}(\gamma)}{\sqrt{\sigma^2_{A.}(\gamma)}} > \frac{z_\alpha \sqrt{\sigma^2_{A.}(0)} + \mu_{A.}(0) - \mu_{A.}(\gamma)}{\sqrt{\sigma^2_{A.}(\gamma)}} \right]$$

$$= \phi\left[\frac{\mu_{A.}(\gamma) - \mu_{A.}(0) - z_\alpha \sqrt{\sigma^2_{A.}(0)}}{\sqrt{\sigma^2_{A.}(\gamma)}} \right]. \qquad (2.2.1.2)$$

2.2.2. A Aproximação de Cornfield e a Aproximação de Muñoz e Rosner para o Poder do Teste de Mantel-Haenszel

Muñoz e Rosner(1984) provam que a primeira derivada de $\mu_i(\gamma)$ é igual $\sigma^2_i(\gamma)$, para $i=1,\ldots,m$. Isso mostra que a média da distribuição de A_i é função crescente de sua variância, e portanto, uma não pode ser escrita independentemente da outra. Essa é uma dificuldade que aparece no cálculo do poder exato do teste de Mantel-Haenszel. Mas esse poder pode ser aproximado usando as aproximações para a média e a variância da distribuição de A_i sugeridas por Cornfield(1956) e por Muñoz e Rosner (1984).

Cornfield aproximou a média $\mu_i(\gamma)$ pela solução permissível $\mu_i(\gamma)_c$ da equação quadrática

$$\frac{\mu_i(\gamma)_c \left[N_{zi} - M_i + \mu_i(\gamma)_c \right]}{\left[N_{zi} - \mu_i(\gamma)_c \right] \left[M_i - \mu_i(\gamma)_c \right]} = \exp(\gamma) \qquad (2.2.2.1)$$

a condição de permissibilidade sendo a condição (2.1.4), e aproximou $\sigma_i^2(\gamma)$ por

$$\sigma_i^2(\gamma)_c = \left[\frac{1}{\mu_i(\gamma)_c} + \frac{1}{N_{zi} - \mu_i(\gamma)_c} + \frac{1}{M_i - \mu_i(\gamma)_c} + \frac{1}{N_{zi} - M_i + \mu_i(\gamma)_c} \right]^{-1} \qquad (2.2.2.2)$$

Essas aproximações preservam a relação $\sigma_i^2(\gamma)_c = d\mu_i(\gamma)/d\gamma$ (Muñoz e Rosner,1984). A solução de (2.2.2.1) é

$$\mu_i(\gamma)_c = N\, t_i\, r_i\, s_i\, h(r_i, s_i, \gamma), \qquad (2.2.2.3)$$

onde

$$h(r_i, s_i, \gamma) = (2 r_i s_i)^{-1} \left\{ \left[(\exp(\gamma) - 1)^{-1} + r_i + s_i \right] - \left\{ \left[(\exp(\gamma) - 1)^{-1} + r_i + s_i \right]^2 - 4\, r_i s_i\, (\exp(\gamma) - 1)^{-1} \exp(\gamma) \right\}^{\frac{1}{2}} \right\}$$

para $\gamma > 0$, e (2.2.2.2) fica

$$\sigma_i^2(\gamma) = N\, t_i r_i s_i h'(r_i, r_i, \gamma), \qquad (2.2.2.4)$$

onde $h'(r_i, s_i, \gamma)$ é a primeira derivada de $h(r_i, s_i, \gamma)$ com respeito

a γ . $h(r_i,s_i,0)=1$, já que $\mu_i(0)=Nt_ir_is_i$. Então substituindo (2.2.2.3) e (2.2.2.4) em (2.2.1.2), encontramos $\Pi_\gamma(\alpha)_c$, a aproximação de Cornfield para o poder do teste de Mantel-Haenszel:

$$\Pi_\gamma(\alpha)_c =$$

$$\phi\left(\frac{\left[N^{\frac{1}{2}}\sum_{i=1}^{m}t_ir_is_i(h(r_i,s_i,\gamma)-1)-z_\alpha\left\{\sum_{i=1}^{m}t_ir_is_i(1-r_i)(1-s_i^2)\right\}^{\frac{1}{2}}\right]}{\left\{\sum_{i=1}^{m}t_ir_is_ih'(r_i,s_i,\gamma)\right\}^{1/2}}\right) \quad (2.2.2.5)$$

A exatidão dessa aproximação foi estudada por Levin (1982), para o caso particular de um mesmo número de indivíduos amostrados em cada um dos m estratos (isto é, $t_i=1/m$, $i=1,\ldots,m$) e um mesmo número de indivíduos expostos e não expostos em cada um dos m estratos ($r_i=1/2, i=1,\ldots,m$), mostrando mostrando a adequação dessa aproximação nesses casos.

Muñoz e Rosner (1984) propuseram uma aproximação computacionalmente bem mais simples do que a de Cornfield para o poder do teste de Mantel-Haenszel. Ela é, na verdade, uma aproximação para a aproximação de Cornfield. Eles aproximaram a função h em (2.2.2.5) por seu desenvolvimento em série de Taylor em torno de zero. Considerando que

$$h(r_i,s_i,0)=1,$$

$$h'(r_i,s_i,0)=(1-r_i)(1-s_i)$$

e que

$$h''(r_i,s_i,0)=(1-r_i)(1-s_i)(1-2r_i)(1-2s_i),$$

o desenvolvimento de h é

$$h_{c*}(\gamma)= 1+(1-r_i)(1-s_i)\gamma + 0.5(1-r_i)(1-s_i)(1-2r_i)(1-2s_i)\gamma^2 ,$$

Então

$$h'_{c*}(\gamma)=(1-r_i)(1-s_i) + (1-r_i)(1-s_i)(1-2r_i)(1-s_i)\gamma$$

e a aproximação mais simples é

$$\Pi_\gamma(\alpha)_{c*} =$$

$$\phi\left[\frac{N^{\frac{1}{2}}\sum_{i=1}^{m} t_i r_i s_i (h_{c*}(r_i,s_i,\gamma)-1)-z_\alpha\left\{\sum_{i=1}^{m} t_i r_i s_i (1-r_i)(1-s_i)\right\}^{\frac{1}{2}}}{\left\{\sum_{i=1}^{m} t_i r_i s_i h'_{c*}(r_i,s_i,\gamma)\right\}^{1/2}}\right] \quad (2.2.2.6)$$

ou

$$\Pi_\gamma(\alpha)_{c*} = \phi\left[\frac{\left[N^{\frac{1}{2}}(\gamma B_1+\frac{1}{2}\gamma^2 B_2)-z_\alpha B_1^{1/2}\right]}{(B_1+\gamma B_2)^{1/2}}\right] \quad (2.2.2.7)$$

onde

$$B_1=\sum_{i=1}^{m} t_i r_i s_i (1-r_i)(1-s_i)$$

e

$$B_2=\sum_{i=1}^{m} t_i r_i s_i (1-r_i)(1-s_i)(1-2r_i)(1-2s_i)$$

2.3. REGRESSÃO LOGÍSTICA

2.3.1. O Modelo de Regressão Logística Linear

A teoria de regressão logística foi originalmente desenvolvida para estudos de coorte e a discussão feita nesse parágrafo a eles se refere.

Nosso caso de interesse é aquele em que há um resultado dicotômico cujos níveis, caso e controle, são representados pelos valores da variável aleatória binária Y. Y=1 se o indivíduo é classificado como caso, e Y=0 se o indivíduo é classificado como controle. Denotando por p a probabilidade de um indivíduo ser um

caso, o logito, ou transformada logito de p (p=P(Y=1)) é definida por

$$\lambda = \ln(p/(1-p)), \qquad (2.3.1.1)$$

onde a quantidade p/(1-p) é denominada "odds" de caso. O nome regressão logística deriva de se expressar λ como uma função de variáveis regressoras que correspondem aos fatores de risco. Assim, se temos um fator de risco, podemos modelar λ como

$$\lambda = \alpha + \gamma X , \qquad (2.3.1.2)$$

ou seja, uma função linear de uma constante α e da variável X, a qual podemos denominar de variável risco, pois corresponde ao fator de risco. (2.3.1.2) define um modelo de regressão logistica linear ou modelo logístico linear.

Se o fator de risco é dicotômico, a variável X em (2.3.1.2) é binária com valor zero se o fator está ausente no indivíduo, isto é,se o indivíduo é classificado como não exposto ao fator de risco, e com valor 1 se o indivíduo é classificado como exposto ao fator de risco. Então, de acordo com (2.3.1.2), podemos dizer que p é a probabilidade de Y=1 dado X e conforme X tome valor 0 ou 1, o logito de p é

$$\lambda^0 = \ln\left(\frac{p_0}{1-p_0}\right) = \alpha , \quad \text{onde } p_0 = P(Y=1/X=0)$$

ou (2.3.1.3)

$$\lambda^1 = \ln\left(\frac{p_1}{1-p_1}\right) = \alpha + \gamma X , \text{ onde } p_1 = P(Y=1/X=1)$$

Quando há um fator de confundimento, uma variável correspondente a esse fator deve ser incluída no modelo. Então, o modelo logístico torna-se

$$\lambda = \alpha + \gamma X + \beta Z, \qquad (2.3.1.4)$$

e Z se denomina variável de confundimento.

Quando há múltiplos fatores de confundimento no estudo, m, digamos, o modelo logístico deve incluir m variáveis de confundimento, cada uma correspondendo a um dos fatores de confundimento. Nesse caso,o modelo logístico torna-se (Anderson et al,1980)

$$\lambda = \alpha + \gamma X + \sum_{k=1}^{m} \beta_k Z_k \ . \qquad (2.3.1.5)$$

Quando consideramos uma única variável de confundimento categórica com mais de duas categorias e o estudo é estratificado segundo essas categorias, estamos em uma situação que nos leva a adotar um modelo semelhante ao modelo que é usado no caso em que há múltiplas variáveis de confundimento, mas no qual as variáveis Z_k são variáveis indicadoras (ou dummy) usadas para representar as várias categorias do único fator de confundimento. Mas se existir m categorias do fator de confundimento, apenas m-1 variáveis indicadoras dessas categorias devem ser incluidas no modelo (Anderson et al,1980). Devemos notar que essas variáveis são indicadoras de estrato e são definidas por

$$Z_k = \begin{cases} 1, \text{ se o indivíduo está na categoria k do fator de confundimento.} \\ 0, \text{ se o indivíduo não está na categoria k do fator de confundimento.} \end{cases}$$

Como incluimos (m-1) variáveis indicadoras de categorias no modelo, uma categoria fica sem a correspondente variável indicadora.

No modelo logístico, o coeficiente β_k correspondendo a Z_k é então o acréscimo no logaritmo natural de "odds" de caso dado que o indivíduo está na categoria k e não na categoria j do fator de confundimento. A numeração das categorias do fator de confundimento é arbitrária, de modo que, estatisticamente, qualquer categoria pode ser especificada como a categoria sem a

correspondente variável indicadora. Escolhemos o estrato m para ser o estrato sem a correspondente variável Z. Então se p_i é a probabilidade de um indivíduo ser um caso dado que ele pertence ao estrato i, o logito de p_i é definido por

$$\lambda_i = \ln (p_i / (1-p_i)), \quad i=1,\ldots,m , \qquad (2.3.1.6)$$

e o modelo logístico linear fica

$$\lambda_i = \alpha + \gamma X + \sum_{k=1}^{m-1} \beta_k Z_k, \quad i=1,\ldots,m, \qquad (2.3.1.7)$$

uma função de uma constante α, da variável X, e das (m-1) variáveis indicadoras de estratos.

Quando m é igual a dois, o modelo (2.3.1.7) se torna

$$\lambda_i = \alpha + \gamma X + \beta_1 Z_1 ,$$

confundindo-se com o modelo (2.3.1.4). É facil verificar que outra maneira de definir (2.3.1.7) é por

$$p_i = \frac{\exp (\alpha + \gamma X + \sum_{k=1}^{m-1} \beta_k Z_k)}{1 + \exp (\alpha + \gamma X + \sum_{k=1}^{m-1} \beta_k Z_k)} , \quad i=1,\ldots,m \qquad (2.3.1.8)$$

É comum encontrarmos na literatura para a situação estratificada o modelo

$$\lambda_i = \alpha_i + \gamma X , \quad i=1,\ldots,m , \qquad (2.3.1.9)$$

um modelo de posto completo, no qual os α_i's são constantes expressando a variação na frequência de casos de um estrato para outro e X é a variável risco. Nesse modelo, α_i=1 se a observação é

feita no estrato i e α_i=0 se a observação não é feita no estrato i. Mas há perfeita equivalência entre o modelo (2.3.1.7) e o modelo (2.3.1.9). Para ver isso, notamos por exemplo, que a matriz do modelo (2.3.1.7) para dois estratos é

$$U=(U_1,U_2,U_3)=\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

onde U_1 é o vetor coluna de constantes, U_2 é o vetor coluna de valores de X e U_3 é o vetor coluna de valores de Z, indicando que no primeiro estrato Z é igual a um e no segundo estrato Z é igual a zero. Para o modelo (2.3.1.9), a matriz do modelo é

$$W=(W_1,W_2,W_3)=\begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

onde $W_1' = (1,1,1,1,0,0,0,0)$ e $W_2' = (0,0,0,0,1,1,1,1)$, indicando os valores das variáveis indicadoras para os estratos 1 e 2, respectivamente, e o vetor $W_3' = (1,1,0,0,1,1,0,0)$ é o vetor de valores da variável risco. Agora então vemos que $U_1=W_1+W_2$, $U_2=W_3$, $U_3=W_2$ e que $W_1=U_1-U_3$, $W_2=U_3$ e $W_3=U_2$. Ou seja, os vetores coluna de U são combinações lineares dos vetores coluna de W, e vice-versa. Assim, o modelo (2.3.1.7) é equivalente ao modelo (2.3.1.9). Modelamos o logito de p_i com o modelo (2.3.1.7).

2.3.2. Interpretação dos Parâmetros do Modelo Logístico e a Equivalência entre o Parâmetro γ e a Razão de Produtos Cruzados

No modelo (2.3.1.7), o valor do parâmetro α depende de como as variáveis estão definidas. Como consequência disso, o valor de α ou de sua estimativa, $\hat{\alpha}$, geralmente não é interpretável. O coeficiente β_k correspondendo a Z_k é o acréscimo no logito de p_i dado que o indivíduo está na categoria k do fator estratificador. O parâmetro γ é o acréscimo no logito de p_i dado que o indivíduo está na categoria x do fator de risco, representada por um dos valores da variável binária X (X=0 se o indivíduo é classificado como não exposto ao fator de risco e X=1 se o indivíduo é classificado como exposto ao fator de risco). Em γ reside nosso interesse, pois ele é facilmente identificado como o logaritmo natural da razão de produtos cruzados. (2.3.1.7) e (2.3.1.8) nos permitem ver isso. Conforme X tome valor 0 ou 1 em (2.3.1.7),

$$\lambda_i^c = \ln\left[\frac{P(Y=1/X=0,Z_1,\ldots,Z_m)}{1-P(Y=1/X=0,Z_1,\ldots,Z_m)}\right] = \ln\left(\frac{p_{ic}}{1-p_{io}}\right)$$

ou

$$\lambda_i^1 = \ln\left[\frac{P(Y=1/X=0,Z_2,\ldots,Z_k)}{1-P(Y=1/X=1,Z_2,\ldots,Z_k)}\right] = \ln\left[\frac{p_{i1}}{1-p_{i1}}\right]$$

Por subtração encontramos

$$\gamma = \lambda_i^1 - \lambda_i^0 = \ln\left(\frac{p_{i1}}{1-p_{i1}}\right) - \ln\left(\frac{p_{ic}}{1-p_{io}}\right) =$$

$$= \ln\left[\frac{p_{i1}(1-p_{ic})}{(1-p_{i1})p_{io}}\right],$$

que nós reconhecemos como o logaritmo natural da razão de produtos cruzados da i-ésima tabela 2x2 de um estudo de coorte ou

caso-controle estratificado.

O modelo (2.3.1.7) supõe que o efeito da variável risco é o mesmo para todos os estratos, ou seja, é γ para todos os estratos. Isso equivale à suposição de que o risco relativo é constante, comum a todos os estratos, pois γ sendo o logaritmo natural da razão de produtos cruzados aproxima bem o logaritmo do risco relativo. O teste da hipótese da igualdade do parâmetro γ desse modelo a zero ou da igualdade de $\exp(\gamma)$ a um é equivalente ao teste da hipótese do teste de Mantel-Haenszel, essa equivalência sendo no sentido de que ambos são testes da significância do risco relativo. Portanto, a hipótese da igualdade de γ a zero ou de $\exp(\gamma)$ a um equivale à hipótese do teste de Mantel-Haenszel (o risco relativo comum a todos os estratos é um).

Uma vantagem potencial do uso de regressão logística para testar a significância do risco relativo comum a todos os estratos, em vez de usar o procedimento de Mantel-Haenszel, é que com a regressão logística, temos um modelo para os dados.

2.3.3. Adaptação do Modelo Logístico Linear a Estudos de Casos e Controles

O presente trabalho se refere a estudos de casos e controles. No entanto, lembrando que no início da seção 2.3.1 definimos p como a probabilidade de um indivíduo ser um caso , sabemos que em (2.3.1.2) ou (2.3.1.4) estamos modelando a probabilidade de caso dado exposição. Em estudos de casos e controles é claro que seria apropriado modelar a probabilidade de exposição dado o resultado, mas as mesmas técnicas, utilizando a mesma probabilidade de caso dado exposição dos estudos de coorte se aplicam a estudos de casos controles (Anderson-1972, Breslow e Powers-1978 e Mantel-1973). Em delineamentos de casos e controles o fato de um indivíduo estar ou

não exposto ao fator de risco é observado como um resultado aleatório, ao contrário do que ocorre em delineamentos de coorte. Mas o modelo logístico para probabilidades de resultado (caso ou controle) aleatório, que tem uma interpretação simples em termos de risco relativo pode ser adaptado e usado quando a amostra é obtida com base em um delineamento de casos e controles.

Primeiro, como explicado no parágrafo 1.2, quando falamos sobre a razão de produtos cruzados para tabelas 2x2, inferências sobre o risco relativo para estudos restropectivos são feitas aplicando os mesmos cálculos que seriam aplicados a dados de um estudo prospectivo da mesma população. O modelo logístico tem também essa propriedade. Uma breve demonstração disso baseada em Mantel (1973) e Siegel e Greenhouse (1973) pode ser encontrada em Breslow e Day (1980, capítulo 6) para o caso não estratificado, considerando o modelo (2.3.1.2). Para situações estratificadas, os cálculos e os resultados são análogos. A demonstração de Breslow e Day considera uma variável indicadora Z que denota se alguém é ou não amostrado e define

$$\pi_1 = P(z=1/y=1)$$

como a probabilidade de se incluir uma pessoa doente no estudo como um caso e

$$\pi_0 = P(z=1/y=0)$$

como a probabilidade de se incluir uma pessoa livre de doença no estudo como um controle. π_0 e π_1 são probabilidades de seleção. A probabilidade condicional de que Y=1 (a pessoa é doente), dado que a pessoa é classificada como exposta às variáveis risco $X_1,\ldots,X_p$, envolvidas no estudo e que ela foi amostrada para o estudo de casos e controles pode ser calculada usando o teorema de Bayes. Quando temos uma única variável risco, X, calculamos

$$P(y=1/z=1,X) = \frac{P(z=1/y=1,X)P(y=1/X)}{P(z=1/y=0,X)P(y=0/X)+P(z=1/y=1,X)P(y=1/X)}$$

$$= \frac{\pi_1 \exp(\alpha+\gamma X)}{\pi_0+\pi_1 \exp(\alpha+\gamma X)} = \frac{\exp(\alpha^*+\gamma X)}{1+\exp(\alpha^*+\gamma X)},$$

onde $\alpha^* = \alpha + \ln(\pi_1/\pi_0)$. Portanto, as probabilidades de doença para aqueles na amostra continuam a ser dadas pelo modelo logístico com precisamente o mesmo parâmetro γ (o mesmo do modelo logístico quando a amostra era de um estudo de coorte), apenas com uma constante α diferente. Nós devemos notar também nessa demonstração, que há a suposição implícita de que as probabilidades de seleção dependem somente da doença estar ou não presente , e não dependem da presença ou ausência do fator de risco. Por isso, $P(z=1/Y,X)=P(z=1/Y)$.

Em situações estratificadas, as probabilidades de seleção consideradas para fazer a demonstração podem variar de estrato para estrato, mas o modelo de coorte (o modelo (2.3.1.7)) continua a valer.

Um outro fator que devemos considerar para fazer a desejável adaptação, diz respeito ao procedimento de máxima verossimilhança, o qual utilizamos para obter estimativas e fazer os testes dos parâmetros do modelo (2.3.1.2). Em estudos de casos e controles, as probabilidades a ser modeladas seriam de X dado Y. Para cada indivíduo essa probabilidade pode ser expressa pela probabilidade condicional

$$P(X/Y) = \frac{P(Y/X)\,P(X)}{P(Y)} \qquad (2.3.3.1)$$

onde $P(Y/X)$ é especificada pelo modelo logístico (2.3.1.2).

Tendo m casos e n-m controles, a função de verossimilhança dos dados é dada por

$$L(\alpha,\beta_1,\ldots,\beta_r,\underset{\sim}{y},\underset{\sim}{x}) =$$

$$= \prod_{i=1}^{m} P(X=1/y_i=1) \prod_{j=1}^{n-m} P(X_j=1/y_j=0) \qquad (2.3.3.2)$$

onde $\underset{\sim}{y}$ e $\underset{\sim}{x}$ são respectivamente, os vetores de observações das variáveis X e Y.

Assumindo que a distribuição marginal de X não depende de γ, a estimativa de máxima verossimilhança de γ usando a verossimilhança condicional (2.3.3.2) é idêntica àquela baseada em P(Y=1/X), a qual é especificada apenas pelo modelo logístico (2.3.1.2). Além disso, os erros padrões e a covariância gerados por (2.3.3.1) e (2.3.1.2) são próximos (Anderson et al, 1980 e Prentice e Pyke, 1979).

Do exposto acima, conclui-se que os métodos de estimação baseados em (2.3.1.2) e (2.3.3.1) produzem os mesmos resultados numéricos. Isso justifica o uso do modelo para estudos de coorte aos dados de estudos de casos e controles. Assim, para estudos estratificados usamos o modelo (2.3.1.7)

2.3.4. Estimação de Máxima Verossimilhança dos Parâmetros do Modelo Logístico Linear

A metodologia em uso corrente para solucionar os problemas de estimar os parâmetros do modelo (2.3.1.2) e as probabilidades (2.3.1.4) e testar a significância do parâmetro γ, é a de máxima verossimilhança. O teste mais comumente usado para avaliar a significância da contribuição de variáveis em modelos logísticos é o teste da razão de verossimilhança.

Numa situação em que a única variável de confundimento é a dicotômica (temos dois estratos), para testar a significância do parâmetro γ, o teste da razão de verossimilhança requer ajustar o modelo $\lambda_i = \alpha + \gamma X + \beta_i Z$ duas vezes, uma vez para estimar todos os

três parâmetros e uma segunda vez para estimar α e β_1 quando γ é restrito a ser zero (Anderson et al, 1980, capítulo 9). O caso em que a única variável de confundimento possui várias categorias, havendo portanto,mais de dois estratos), é considerado como um caso de múltiplas variáveis de confundimento. Essas variáveis aparecem no modelo (2.3.1.7) como variáveis indicadoras de estrato. Nesse caso, a hipótese nula é

$$K_o: \gamma = 0 \qquad (2.3.4.1)$$

(N.E.Day e D.P. Byar). Aceitar essa hipótese equivale a concluir que a contribuição da variável risco para explicar o logito de p_i, ou equivalentemente, para explicar p_i, é não significante. Isto equivale ainda a validar o modelo

$$\lambda_i = \alpha + \sum_{k=1}^{m-1} \beta_k Z_k , \qquad (2.3.4.2)$$

o qual passaremos a chamar de modelo 1. Escrever (2.2.4.2) equivale a escrever

$$p_i = \frac{\exp(\alpha+\sum_{k=1}^{m-1} \beta_k Z_k)}{1+\exp(\alpha + \sum_{k=1}^{m-1} \beta_k Z_k)} \qquad (2.3.4.3)$$

O modelo (2.3.4.2) é o modelo no qual a suposição é de que p_i não depende da variável risco. Quando ajustamos o modelo (2.3.1.7), estamos adicionando a (2.3.4.2) a variável X. Portanto o teste de K_o é um teste da significância da contribuição da variável adicional X para explicar o "odds" de caso. Além do teste da razão de verossimilhança, um outro teste que pode ser usado para avaliar a significância da contribuição de variáveis em modelos logísticos é o da estatística escore, a qual é baseada no valor da primeira derivada do logaritmo natural da função de verossimilhança calculada nos valores definidos na hipótese nula

para os parâmetros a ser testados. A utilização de qualquer um desses dois tipos de testes para testar K_o passa pela estimação de máxima verossimilhança de todos os parâmetros do modelo (2.3.1.7), o qual passamos agora a chamar de modelo 2, e pela estimação dos parâmetros $\alpha,\beta_1,\ldots,\beta_{(m-1)}$ do modelo 1 ,o que equivale a estimar $\alpha,\beta_1,\ldots,\beta_{m-1}$ quando γ é restrito a ser zero. Fica entendido que α e $\beta_1,\ldots,\beta_{m-1}$ são estimados duas vezes, uma vez no modelo 1 e uma vez no modelo 2.

O procedimento de máxima verossimilhança se inicia com uma expressão para a função de verossimilhança dos dados. Se temos uma amostra de tamanho total n, m estratos, a amostra no i-ésimo estrato tendo tamanho n_i e y_{ij} é o valor de Y observado para o j-ésimo indivíduo no i-ésimo estrato, a função de verossimilhança tem um termo P_i se o valor de Y observado é 1 e um termo $1-P_i$ se o valor de Y observado é zero, onde P_i é dado por (2.3.4.3) ou por (2.3.1.8), conforme o modelo que esteja sendo considerado. A função de verossimilhança geral é

$$L(\alpha,\beta_1,\ldots,\beta_{m-1},\underset{\sim}{y},\underset{\sim}{z}) = \prod_{i=1}^{m}\prod_{j=1}^{n_i} P_i^{y_{ji}}(1-P_i)^{1-y_{ji}},$$

onde $\underset{\sim}{y}$ e $\underset{\sim}{z}$ são, respectivamente, os vetores das n observações nas variáveis Y e Z ao longo dos m estratos e $\underset{\sim}{y}=(y_{11},\ldots,y_{n_m m})$ e $\underset{\sim}{z}=(z_1,\ldots,z_m)$.

Trabalhando com o modelo 1,a verossimilhança fica

$$L(\alpha,\beta_1,\beta_{m-1},\underset{\sim}{y},\underset{\sim}{z}) =$$

$$= \prod_{i=1}^{m} \prod_{j=1}^{n_i} \left[\frac{\exp\left(\alpha+\sum_{k=1}^{m-1} \beta_k Z_k\right)}{1+\exp\left(\alpha+\sum_{k=1}^{m-1} \beta_k Z_k\right)} \right]^{y_{ji}} \left[1 - \frac{\exp\left(\alpha+\sum_{k=1}^{m-1} \beta_k Z_k\right)}{1+\exp\left(\alpha+\sum_{k=1}^{m} \beta_k Z_k\right)} \right]^{1-y_{ji}} =$$

$$= \frac{\exp\left\{\sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ji} \left[\alpha+\sum_{k=1}^{m-1} \beta_k Z_k\right]\right\}}{\prod_{i=1}^{m} \prod_{j=1}^{n_i} \left\{1+\exp\left[\alpha+\sum_{k=1}^{m-1} \beta_k Z_k\right]\right\}} , \qquad (2.3.4.4)$$

O logaritmo natural dessa função é

$$\ln L(\alpha,\beta_1,\ldots,\beta_{m-1},\underset{\sim}{y},\underset{\sim}{z}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ji} \left[\alpha + \sum_{k=1}^{m-1} \beta_k Z_k\right]$$

$$- \sum_{i=1}^{m} \sum_{j=1}^{n_i} \ln\left[1+\exp\left[\alpha + \sum_{k=1}^{m-1} \beta_k Z_k\right]\right] \qquad (2.3.4.5)$$

Suas primeiras derivadas parciais com respeito aos parâmetros $\alpha,\beta_1,\ldots,\beta_{m-1}$ igualadas a zero, mostram que as estimativas de máxima verossimilhança de $\alpha,\beta_1,\ldots,\beta_{m-1}$, $\hat{\alpha},\hat{\beta}_1,\ldots,\hat{\beta}_{m-1}$, satisfazem o sistema de equações (2.3.4.6) abaixo

$$\frac{\partial \ln L}{\partial \alpha} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[y_{ji} - \frac{\exp\left[\hat{\alpha} + \sum_{k=1}^{m-1} \hat{\beta}_k Z_{kj}\right]}{1+\exp\left[\hat{\alpha} + \sum_{k=1}^{m-1} \hat{\beta}_k Z_{kj}\right]} \right] =$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} [y_{ji} - \hat{P}_i] = 0$$

(2.3.4.6)

$$\frac{\partial \ln L}{\partial \beta_l} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} z_l \left[y_{ji} - \frac{\exp\left(\hat{\alpha} + \sum_{k=1}^{m-1} \hat{\beta}_k Z_k\right)}{1+\exp\left(\hat{\alpha} + \sum_{k=1}^{m-1} \hat{\beta}_k Z_k\right)} \right]$$

$$= \sum_{j=1}^{n_l} z_l [y_{jl} - \hat{P}_l] = 0, \quad l=1,\ldots,m-1.$$

O sistema (2.3.4.6) pode ser escrito na forma (2.3.4.7), a qual é semelhante à forma de um sistema linear.

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} [y_{ji} - \hat{P}_i] = 0$$

(2.3.4.7)

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} Z_k [y_{ji} - \hat{P}_i] = 0$$

No entanto, o sistema (2.3.4.7) é não linear nos parâmetros. Assim, o método que usamos para encontrar as estimativas dos parâmetros foi o método numérico iterativo de Newton-Raphson.

O vetor de primeiras derivadas do logaritmo da função de verossimilhança com respeito aos parâmetros de um modelo de regressão logística é bem conhecido como ESCORE, comumente denotado por S e tem p elementos, onde p é o número de parâmetros no modelo. No caso do modelo 1, S tem dimensão p=m, (m é o número de estratos). A negativa da matriz de segundas derivadas parciais do logaritmo da função de verossimilhança com respeito aos parâmetros de um modelo logístico é conhecida como MATRIZ DE INFORMAÇÃO, comumente denotada por I. I tem dimensão pxp e sendo a

negativa da matriz de segundas derivadas parciais, é simétrica. No caso do modelo 1, p=m e I tem a forma

$$I = \begin{bmatrix} \partial^2 \ln L/\partial\alpha^2 & \partial^2 \ln L/\partial\alpha\partial\beta_1 & \cdots & \partial^2 \ln L/\partial\alpha\partial\beta_{m-1} \\ & \partial^2 \ln L/\partial\beta_1^2 & & \\ & & \ddots & \\ & & & \partial^2 \ln L/\partial\beta_{m-1}^2 \end{bmatrix}$$

onde

$$-\frac{\partial^2 \ln L}{\partial \alpha^2} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[\frac{\exp(\alpha+\beta_k Z_k)}{(1+\exp(\alpha+\beta_k Z_k))} \right] = \sum_{i=1}^{r} \sum_{j=1}^{n_i} P_i(1-P_i)$$

$$-\frac{\partial^2 \ln L}{\partial\alpha\ \partial\beta_k} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} z_k \left[\frac{\exp(\alpha+\beta_k Z_k)}{(1+\exp(\alpha+\beta_k Z_k))^2} \right] = \sum_{j=1}^{n_k} z_k P_k(1-P_k),$$

$k=1,\ldots,m-1.$

$$-\frac{\partial^2 \ln L}{\partial\ \beta_k^2} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} z_k \left[\frac{\exp(\alpha+\beta_k Z_k)}{(1+\exp(\alpha+\beta_k Z_k))^2} \right], \quad k=1,\ldots,m-1.$$

A inversa de I calculada nas estimativas de máxima verossimilhança dos parâmetros é a matriz de variâncias e covariâncias dos parâmetros estimados. Nós a denotamos no caso do modelo 1, como $\mathrm{cov}(\hat{\beta}) = [I(\hat{\beta})]^{-1}$, onde $\hat{\beta} = (\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_{m-1})$.

O Método de Newton-Raphson é um dos métodos mais frequentemente usados para encontrar raízes em sistemas de equações não lineares. Para se ter uma idéia do método, pode-se observar a Figura 2.1 abaixo

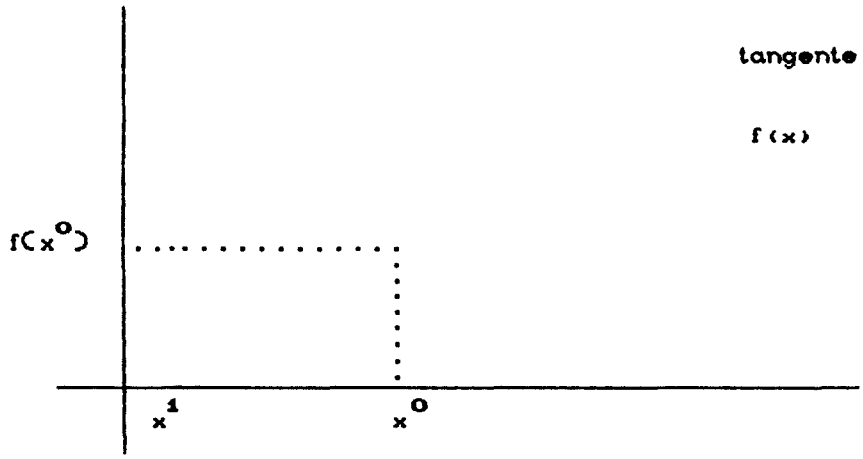


Figura 2.1

Partindo-se de uma estimativa inicial, x^0, da raiz de f(x) prolonga-se a tangente nesse ponto até que essa tangente intercepte o eixo das abcissas; o ponto x^1, onde a tangente intercepta o eixo das abcissas é tomado como a próxima aproximação para a raíz de f(x). O processo continua até que se obtenha um x^* que anule f ou que torne tão próxima de zero quanto se queira especificar.

Para resolver os sistema (2.3.4.6), inicia-se com um vetor inicial $\hat{\underset{\sim}{\beta}}^0 = (\alpha, \beta_1, \ldots, \beta_{m-1})$. Prolonga-se a tangente a $\partial \ln L(\alpha, \beta_1, \ldots \beta_{m-1}, \underset{\sim}{y}, \underset{\sim}{z})$ nesse ponto $\hat{\underset{\sim}{\beta}}^0$. A tangente é dada por $\partial^2 \ln L$, também calculada em $\hat{\underset{\sim}{\beta}}^0$. O processo de Newton-Raphson consiste em fazer na (t+1)-ésima iteração

$$\hat{\underset{\sim}{\beta}}^{t+1} = \hat{\underset{\sim}{\beta}}^t + [I(\hat{\underset{\sim}{\beta}}^t)]^{-1} S(\hat{\underset{\sim}{\beta}}^t), \qquad (2.3.4.8)$$

onde $\hat{\underset{\sim}{\beta}}^t$ é a estimativa de $\underset{\sim}{\beta}$ na t-ésima iteração, e I e S são calculadas nessa estimativa. Fazemos sucessivas iterações, em cada uma delas, determinando um $\hat{\underset{\sim}{\beta}}$. As primeiras derivadas de L se anulam no ponto de máximo e em cada passo do processo iterativo as segundas derivadas parciais são usadas para calcular as sucessivas

tangentes. No final do processo, as estimativas $\hat{\alpha}, \hat{\beta}_1, \ldots \hat{\beta}_{m-1}$ são substituídas em (2.3.4.3) para encontrar $\hat{P}_i$ quando x=0 e quando x=1.

Trabalhando com o modelo 2, a função de verossimilhança tem um termo P_i se o valor de Y observado é zero, onde P_i é dada por (2.3.1.7). Assim, a função de verossimilhança depois de observados os valores de x e y, e assumindo que as observações são independentes dentro dos estratos e entre os estratos é

$$L(\underset{\sim}{\beta}, \underset{\sim}{y}, \underset{\sim}{x}, \underset{\sim}{z}) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} P_i^{y_{ij}} (1-P_i)^{1-y_{ji}} =$$

$$= \frac{\exp\left\{ \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[\alpha + \gamma x_{ji} + \sum_{k=1}^{m-1} \beta_k Z_k \right] y_{ji} \right\}}{\prod_{i=1}^{m} \prod_{j=1}^{n_i} \left\{ 1 + \exp\left[\alpha + \gamma x_{ji} + \sum_{k=1}^{m-1} \beta_k Z_k \right] \right\}}, \qquad (2.3.4.9)$$

onde os vetores $\underset{\sim}{\beta}=(\alpha, \gamma, \beta_1, \ldots, \beta_{m-1})'$, $\underset{\sim}{y}=(y_{11}, \ldots, y_{n_1 1}, y_{12}, \ldots, y_{n_m m})$, $\underset{\sim}{x}=(x_{11}, \ldots, x_{n_1 1}, x_{12}, \ldots, x_{n_m m})$ e $\underset{\sim}{z}$, o qual é um vetor (m-1)x1 com elementos $z_1, \ldots, z_{m-1}$, são, respectivamente, o vetor de parâmetros do modelo e os vetores dos valores observados nas variáveis Y, X e Z. Se a observação é feita no estrato k, $Z_k=1$ e os demais elementos do vetor são zeros. O logaritmo natural da função de verossimilhança para o modelo 2 é

$$\ln L(\alpha, \gamma, \beta_1, \ldots, \beta_{m-1}, \underset{\sim}{y}, \underset{\sim}{x}, \underset{\sim}{z}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[\alpha + \gamma x_{ji} + \sum_{k=1}^{m-1} \beta_k Z_k \right] y_{ji}$$

$$- \sum_{i=1}^{m} \sum_{j=1}^{n_i} \ln(1+\exp(\alpha + \gamma x_{ji} + \sum_{k=1}^{m-1} \beta_k z_k)). \qquad (2.3.4.10)$$

Suas primeiras derivadas parciais igualadas a zero mostram que as estimativas dos parâmetros, $\hat{\alpha}$, $\hat{\gamma}$, $\hat{\beta}_1$, ..., $\hat{\beta}_{m-1}$ satisfazem o

sistema de equações (2.3.4.11) abaixo

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[y_{ji} - \frac{\exp\left\{ \hat{\alpha} + \hat{\gamma}\, x_{ji} + \sum_{k=1}^{m-1} \hat{\beta}_k \right\}}{1 + \exp\left\{ \hat{\alpha} + \hat{\gamma}\, x_{ji} + \hat{\beta}_k z_k \right\}} \right] = 0$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ji}\left[y_{ji} - \frac{\exp\left\{ \hat{\alpha} + \hat{\gamma}\, x_{ji} + \sum_{k=1}^{m-1} \hat{\beta}_k \right\}}{1 + \exp\left\{ \hat{\alpha} + \hat{\gamma}\, x_{ji} + \hat{\beta}_k z_k \right\}} \right] = 0 \qquad (2.3.4.11)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} z_l\left[y_{ji} - \frac{\exp\left\{ \hat{\alpha} + \hat{\gamma}\, x_{ji} + \sum_{k=1}^{m-1} \hat{\beta}_k \right\}}{1 + \exp\left\{ \hat{\alpha} + \hat{\gamma}\, x_{ji} + \hat{\beta}_k z_k \right\}} \right] = 0, l = 1, \ldots, m-1$$

O sistema (2.3.4.11) pode ser escrito como

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ji} - \hat{P}_i \right) = 0 \ ,$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ji}\left(y_{ji} - \hat{P}_i \right) = 0 \ , \qquad (2.3.4.12)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} z_l\left(y_{ji} - \hat{P}_i \right) = 0 \ .$$

Embora (2.3.4.12) pareça linear, não é linear nos parâmetros. Então, novamente, o método de Newton-Raphson é usado para solucionar (2.3.4.12) (equivalentemente, (2.3.4.11)), encontrando as estimativas de máxima verossimilhança dos parâmetros do modelo 2 quando o método converge.

A matriz de informação para o modelo 2 tem elementos

$$\frac{\partial^2 \ln L}{\partial \alpha^2} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[\frac{\exp\left(\alpha + \gamma x_{ji} + \sum_{k=1}^{m-1} \beta_k \right)}{\left(1 + \exp\left(\alpha + \gamma x_{ji} + \beta_k z_k \right) \right)^2} \right],$$

$$\frac{\partial^2 \ln L}{\partial \alpha \partial \gamma} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ji} \left[\frac{\exp\left(\alpha + \gamma x_{ji} + \sum_{k=1}^{m-1} \beta_k \right)}{\left(1 + \exp\left(\alpha + \gamma x_{ji} + \beta_k z_k \right) \right)^2} \right],$$

$$-\frac{\partial^2 \ln L}{\partial \alpha \partial \beta_l} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} z_l \left[\frac{\exp\left(\alpha + \gamma x_{ji} + \sum_{k=1}^{m-1} \beta_k \right)}{\left(1 + \exp\left(\alpha + \gamma x_{ji} + \beta_k z_k \right) \right)^2} \right],$$

para $l=1, \ldots, m-1$,

$$\frac{\partial^2 \ln L}{\partial \gamma^2} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ji}^2 \left[\frac{\exp\left(\alpha + \gamma x_{ji} + \sum_{k=1}^{m-1} \beta_k \right)}{\left(1 + \exp\left(\alpha + \gamma x_{ji} + \beta_k z_k \right) \right)^2} \right],$$

$$\frac{\partial^2 \ln L}{\partial \gamma \partial \beta_l} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ji} z_l \left[\frac{\exp\left(\alpha + \gamma x_{ji} + \sum_{k=1}^{m-1} \beta_k \right)}{\left(1 + \exp\left(\alpha + \gamma x_{ji} + \beta_k z_k \right) \right)^2} \right],$$

para $l = 1, \ldots, m-1$,

$$-\frac{\partial^2 \ln L}{\partial \beta_l^2} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} z_l^2 \left[\frac{\exp\left(\alpha + \gamma x_{ji} + \sum_{k=1}^{m-1} \beta_k \right)}{\left(1 + \exp\left(\alpha + \gamma x_{ji} + \beta_k z_k \right) \right)^2} \right],$$

para $l=1, \ldots, m-1$,

e

$$-\frac{\partial^2 \ln L}{\partial \beta_q \partial \beta_l} = 0, \text{ para } l=1, \ldots, m-1, \; l \neq q.$$

Essa matriz de informação é simétrica e tem a forma

$$I = \begin{bmatrix} \partial^2 \ln L/\partial \alpha^2 & \partial^2 \ln L/\partial \alpha \partial \gamma & \cdots & \partial^2 \ln L/\partial \alpha \partial \beta_{m-1} \\ & \partial^2 \ln L/\partial \gamma^2 & & \partial^2 \ln L/\partial \gamma \partial \beta_{m-1} \\ & & \ddots & \\ & & & \partial^2 \ln L/\partial \beta_{m-1}^2 \end{bmatrix}$$

Em cada passo do processo iterativo, I^{-1} é calculada e usada para determinar a próxima aproximação para o vetor $\underset{\sim}{\beta}=(\alpha,\gamma,\beta_1,\ldots,\beta_{m-1})$. Ao final do processo, substituimos a estimativa $\hat{\underset{\sim}{\beta}}=(\hat{\alpha},\hat{\gamma},\ldots,\hat{\beta}_{m-1})$ em (2.3.1.7) duas vezes,uma vez para calcular $\hat{P}_i$ quando x=0 e outra vez para calcular $\hat{P}_i$ quando x=1.

2.3.5. O Teste da Razão de Verossimilhança para o Risco Relativo

O teste da razão de verossimilhança para testar a hipótese K_o de (2.3.4.1) é realizado simplesmente comparando a diferença entre as estatísticas de verossimilhança dos dois modelos com o valor tabelado de uma variável aleatória com distribuição qui-quadrado

com um número de graus de liberdade igual ao número de restrições em K_0. O número de restrições em K_0 é o número de parâmetros sendo testados a zero, que é também o número de variáveis que adicionamas ao modelo 1 para obtermos o modelo 2. Assim, esse número de graus de liberdade é um.

A estatística de verossimilhança para um modelo de regressão logística é dada por

$$G = -2 \ln L(\hat{\beta}, y, x, z),$$

onde $\hat{\beta}$ é o vetor de parâmetros estimados do modelo e y, x e z são, respectivamente, os vetores de observaçõews de casos e de controles, da variável risco e da variável estratificadora.

A estatística de verossimilhança do modelo 1 é dada por

$$G_1 = -2 \ln L(\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_{m-1}, y, x, z) =$$

$$= -2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ y_{ji} \ln \hat{P}_i + (1-y_{ji}) \ln(1-\hat{P}_i) \right\}$$

onde $\hat{P}_i$ é a estimativa de P_i dada pelo modelo 1 e para o modelo 2, a estatística de verossimilhança é dada por

$$G_2 = -2 \ln L(\hat{\alpha}, \hat{\gamma}, \hat{\beta}_1, \ldots, \hat{\beta}_{m-1}, y, x, z) =$$

$$= -2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ y_{ji} \ln \hat{\hat{P}}_i + (1-y_{ji}) \ln(1-\hat{\hat{P}}_i) \right\}$$

onde $\hat{\hat{P}}_i$ é a estimativa de P_i dada pelo modelo 2. Note-se, então, que a diferença entre G_1 e G_2 é que em G_1, $\hat{P}_i$ não é função da variável risco X, enquanto que em G_2, $\hat{\hat{P}}_i$ o é. Tanto em uma estatística como em outra, as probabilidades estimadas são calculadas nas estimativas de máxima verossimilhança dos parâmetros dos correspondentes modelos.

O modelo 2 é menos restritivo do que o modelo 1, uma vez que inclui uma variável a mais. Assim, G_1 é maior do que G_2. A diferença G_1-G_2 é a estatística do teste da razão de verossimilhança. para nossa hipótese K_0. Se essa estatística tiver um valor maior do que o valor tabelado de uma variável qui-quadrado com um grau de liberdade ao nível de significância especificado, então rejeitamos K_0 a esse nível de significância.

O teste da razão de verossimilhança é o teste mais comumente usado para análise de dados de estudos de casos e controles e é o teste que usamos neste trabalho.

CAPÍTULO 3

ESTUDOS DE MONTE CARLO

Neste capítulo apresentamos uma breve discussão sobre as vantagens possíveis do uso de cada um dos testes apresentados: o teste de Mantel-Haenszel e seu teste equivalente quando usamos regressão logística. Podemos conduzir a discussão em duas direções, a primeira sobre vantagens intrínsecas no uso de uma das metodologias sobre a outra, e a segunda, em situações idênticas, sobre qual dos testes apresenta maior poder.

Sob o primeiro aspecto é inegável que o uso da regressão logística apresenta vantagens pois, além de podermos testar a hipótese de interesse, terminamos o processo com um modelo para o problema. Muitas vezes a existência desse modelo é que permite uma melhor compreensão do problema em estudo.

Sobre o poder do teste de Mantel-Haenszel alguns trabalhos já foram feitos, no sentido de obter aproximações. B. Levin (1982) estudou a exatidão da aproximação de Cornfield (1956) (2.2.2.5) baseada nas aproximações (2.2.2.3) para a média e (2.2.2.4) para a variância da distribuição hipergeométrica extendida, para casos bem balanceados ($t_1 = t_2 = \cdots = t_m = r_1 = r_2 = \cdots = r_m$, $s_1 = s_2 = \cdots = s_m = 0.5$), concluindo que para esses casos a aproximação é boa. Muñoz e Rosner (1984) desenvolveram a aproximação computacionalmente mais simples (2.2.2.7). Na tabela 3.1 apresentamos os valores fornecidos pelas aproximacções de Cornfield (1956) e de Muñoz e Rosner (1984), reproduzidas deste último trabalho em que podemos ver que sua aproximação produz resultados muito próximos da outra (a máxima diferença apresentada

Tabela 3.1. Aproximações de Cornfield e de Muñoz e Rosner para o poder do teste de Mantel-Haenszel para duas tabelas 2x2, $\alpha = 0.10$, $t_1 = t_2 = 0.5$, $r_1 + r_2 = 1 = s_1 + s_2$ e $N = 208$[1].

r_1	s_1								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	(*).328	.410	.444	.446	.422	.374	.304	.215	.116
	(#).325	.416	.457	.464	.442	.392	.316	.215	.103
0.2		.547	.614	.637	.627	.583	.503	.381	.215
		.555	.623	.648	.638	.596	.515	.388	.215
0.3			.698	.734	.734	.702	.630	.503	.304
			.706	.740	.740	.709	.640	.515	.316
0.4				.777	.785	.762	.702	.583	.374
				.781	.789	.767	.709	.596	.392
0.5					.800	.785	.734	.627	.422
					.803	.789	.740	.638	.442
0.6						.777	.734	.637	.446
						.781	.740	.648	.464
0.7							.698	.614	.444
							.706	.623	.457
0.8								.547	.410
								.555	.416
0.9									.328
									.325

(1) Tamanho de amostra estimado para razão de produtos cruzados igual a 2.0 e poder igual 0.80, usando o chi quadrado para a tabela de totais.

(*) Aproximação de Cornfield.

(#) Aproximação de Muñoz e Rosner.

é 0.02), de modo que podemos concluir que para situações balanceadas, a aproximação de Muñoz e Rosner,fornece valores próximos dos reais.

Restam portanto dois problemas a serem estudados e discutidos: A adequação da aproximação de Muñöz e Rosner para situações não balanceadas e fazermos uma comparação do poder dos dois testes sob as mesmas condições. Com relação ao primeiro problema, o que ocorre é que a aproximação de Muñoz e Rosner se baseia na distribuição normal. Para situações desbalanceadas, próximas aos cantos de uma tabela de poder como a Tabela 3.1, a distribuição hipergeométrica é assimétrica e uma aproximação normal pode não ser ideal; com relação ao segundo problema, é de interesse a comparação já que os dois testes são equivalentes. Para esses fins fizemos estudos de Monte Carlo devido à impossibilidade de realizar o trabalho de forma analítica.

Em nossos estudos de Monte Carlo, inicialmente fizemos simulações de dados para gerar tabelas 2x2xm como derivadas de estudos de casos e controles, procurando nos determos nas situações em que m é igual a 2 e 3. Fizemos cada conjunto de simulações fixando um conjunto particular de valores para as frações t_i's, r_i's e s_i's, i=1,...,m, e um conjunto de razões de produtos cruzados,uma para cada estrato. Para alguns conjuntos de simulações fixamos uma razão de produtos cruzados igual a dois para todos os estratos. e para outros, fixamos razões de produtos cruzados distintas para todos os estratos, mas de modo que a média dessas razões de produtos cruzados fosse igual a dois. Depois de fixadas essas quantidades, em cada conjunto de simulações fizemos um número conveniente de replicações, gerando em cada replicação uma tabela 2x2xm. Estamos interessados em estimar e comparar duas proporções: os dois poderes. Esse número conveniente de replicações está relacionado com a amplitude desejada para os intervalos de confiança para essas proporções. Denotando por pmh e prl as estimativas dadas pelo Monte Carlo para, respectivamente, o poder do teste de Mantel-Haenszel e o poder do teste da razão de verossimilhança para a regressão logística, essas estimativas são dadas serão dadas por

$$pmh = \frac{nrmh}{numrep} \quad e \quad prl = \frac{nrrl}{numrep}$$

onde numrep é o número de replicações feitas no estudo de Monte Carlo, nrmh é o número de vezes que a hipótese H_0 foi rejeitada ao longo das numrep replicações e nrrl é o número de vezes que a hipótese K_0 foi rejeitada ao longo das numrep replicações. Usando o mesmo critério dado por Muñoz e Rosner, poderíamos dizer que se ao final das numrep replicações pmh e prl não diferem por mais de 0.02, então os dois poderes são iguais. Preferimos, no entanto, nos basearmos nos intervalos de confiança de 95% para as duas proporções.

Intervalos de confiança de 95% para as verdadeiras proporções são, aproximadamente,

$$pmh \pm 2.00 \sqrt{\frac{pmh(1 - pmh)}{numrep}}$$

para o caso do procedimento de Mantel-Haenszel e

$$prl \pm 2.00 \sqrt{\frac{prl(1 - prl)}{numrep}}$$

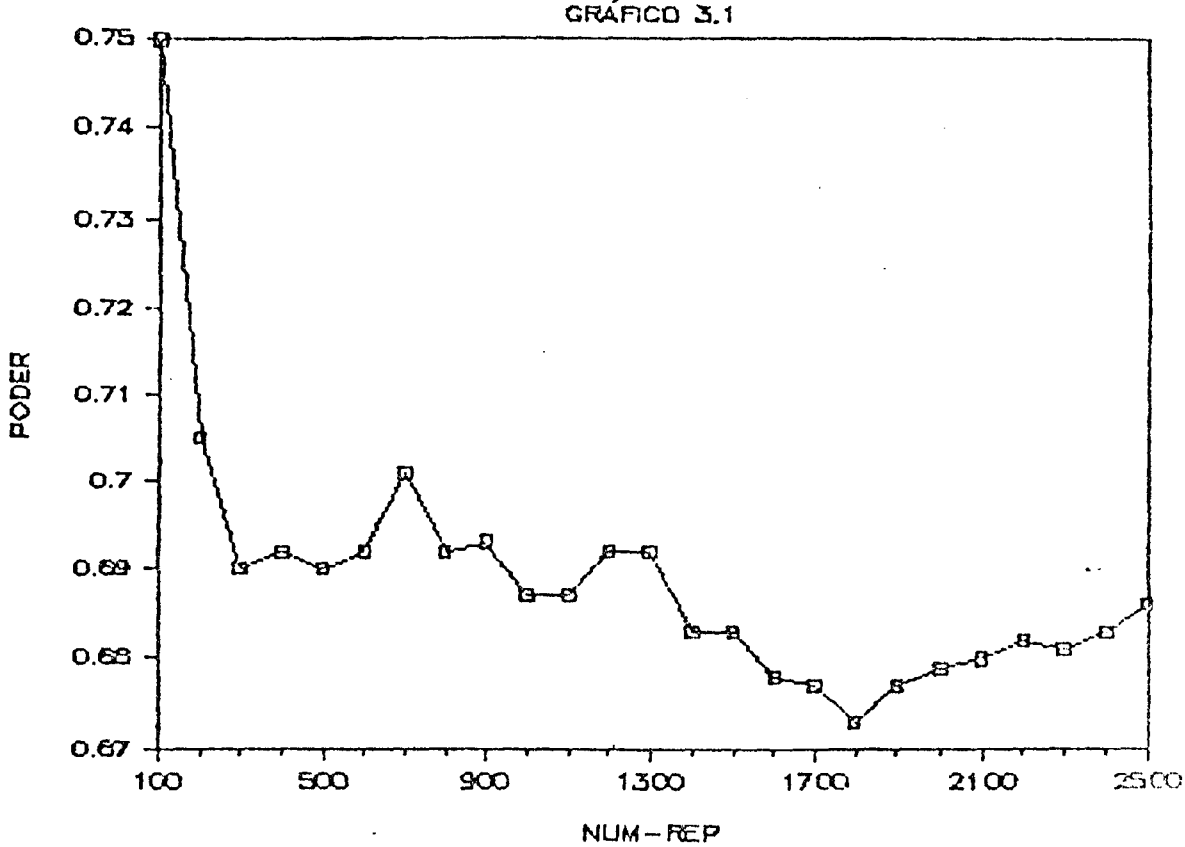
para o caso da regressão logística. Se para uma determinada situação (uma particular distribuição dos t_i's, r_i's e s_i's) os intervalos de confiança para as duas proporções se interceptam, não podemos concluir que os dois poderes são diferentes.

Para as situações em que o número de estratos m é igual a dois, as simulações são feitas em 2500 replicações para que os intervalos acima tenham, ambos, amplitude máxima igual a 0.04, dando uma amplitude de 0.02 de cada lado do intervalo. Os gráficos 3.1, 3.2 e 3.3 são resultantes de um conjunto de simulações feitas

em 2500 replicações para a situação quase perfeitamente balanceada com dois estratos,na qual $t_1=t_2=0.5$, $r_1=0.4$, $r_2=0.6$, $s_1=0.6$, $s_2=0.4$ e razões de produtos cruzados, respectivamente, 1.5 no primeiro estrato e 2.5 no segundo estrato. Esses gráficos representam, respectivamente, uma curva do poder do teste de Mantel-Haenszel, uma curva para o poder da regressão logística e uma curva para a diferença prl - pmh em valor absoluto. Os pontos que compõem os três gráficos são os resultados das simulações ao final de um número de replicações que cresce de cem em cem. Por exemplo, os pontos que compõem os gráficos 3.1 são as proporções de rejeição da hipótese H_o de (2.1.1) ao final das cem primeiras replicações, ao final das duzentas primeiras replicações, e assim por diante. No gráfico 3.1, podemos observar que a aproximação para o poder do teste de Mantel-Haenszel dá um salto passando de 0.75 para 0.71,daí para 0.69 , e após pequenas flutuações parece se estabilizar em torno de 0.68. No gráfico 3.2, podemos observar que a aproximação para o poder do teste da razão de verossimilhança para a regressão logística de início é igual a do poder do teste de Mantel-Haenszel, mas em seguida se estabiliza em torno de 0.69. O comportamento das diferenças absolutas entre as duas aproximações, apresentado no Gráfico 3.3, sugere que essas diferenças se estabilizam em torno de 0.01. A maior diferença absoluta não alcança 0.02. Pelo critério utilizado por Muñoz e Rosner para a Tabela 3.1, ao olharmos o Gráfico 3.3, concluimos que nessa situação os dois poderes são iguais. Usando os intervalos de confiança de amplitude 0.02 de cada lado, chegamos à mesma conclusão. No entanto, o uso, desses intervalos nos permite concluir que as duas proporções são diferentes somente quando a diferença entre entre elas for superior a 0.04, pois só a partir desse valor para a diferença é que os intervalos de confiança não se interceptam. Isso nos dá uma margem de segurança para concluirmos se os dois poderes são diferentes ou não maior do que a margem de segurança que temos com o critério da diferença

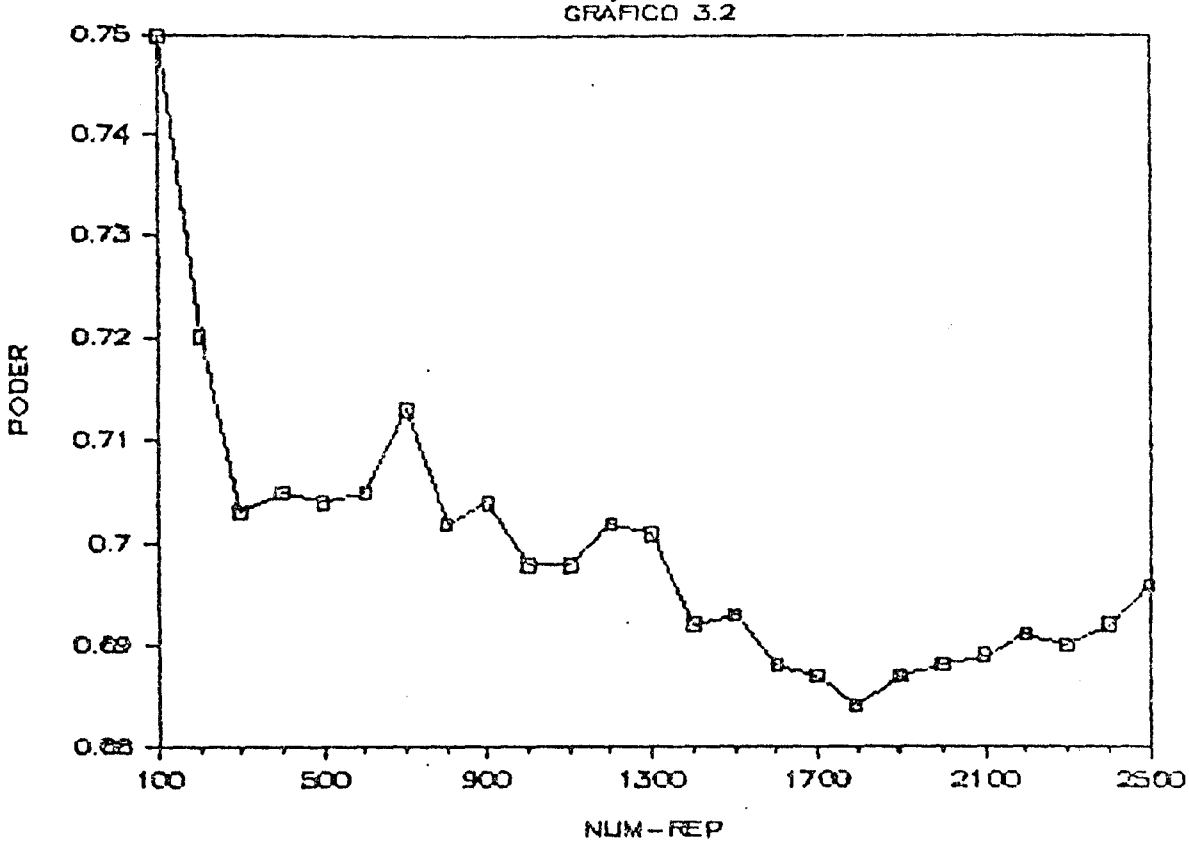
PODER DE MANTEL–HAENZSEL

GRÁFICO 3.1



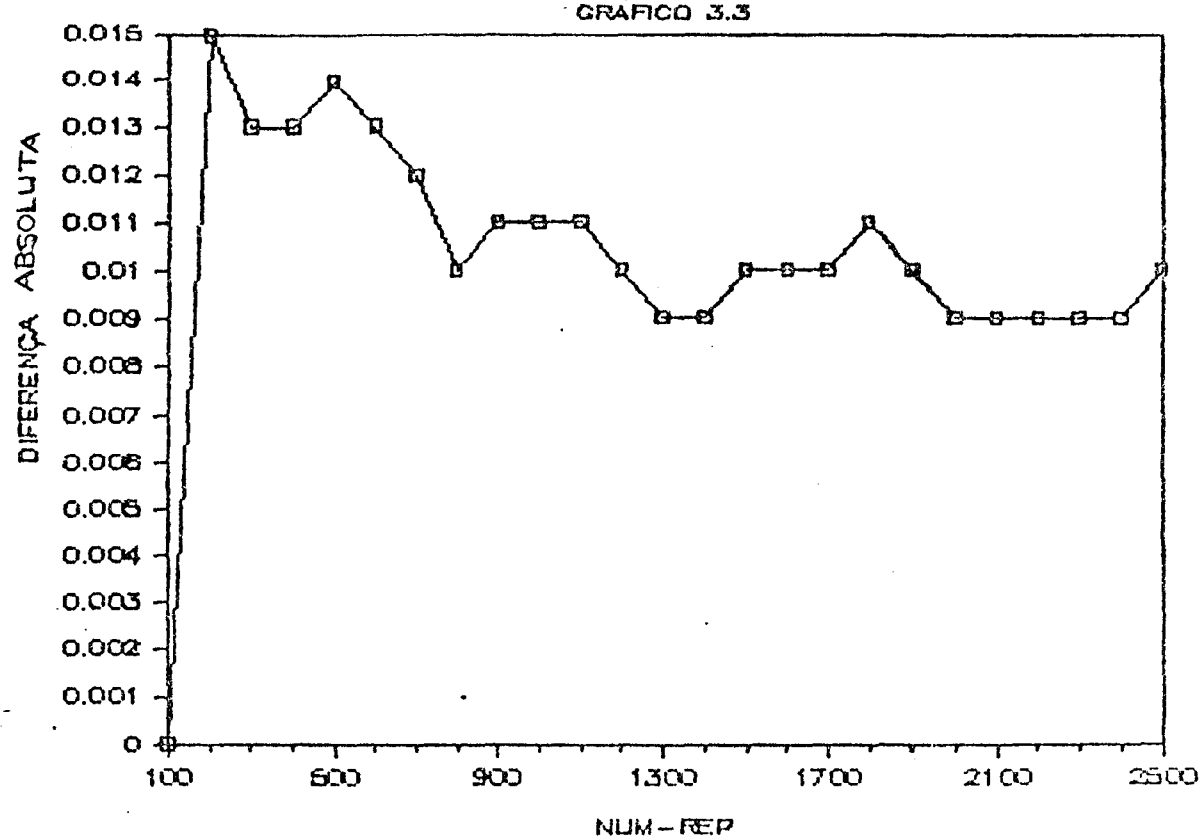
PODER DA REGR. LOGÍSTICA

GRÁFICO 3.2



DIFERENCA ENTRE OS PODERES

GRÁFICO 3.3



absoluta igual a 0.02.. As situações em que m é igual a três, foram feitas em 1000 replicações, dando intervalos de 95% de confiança com amplitude máxima igual a 0.06, o que significa uma amplitude de 0.03 de cada lado do intervalo.

No caso do procedimento de Mantel-Haenszel, nós ressaltamos que utilizamos uma expressão para fazer o teste e comparar com a aproximação para o poder da regressão logística e utilizamos outra expressão para fazer o teste e comparar com os resultados de Muñoz e Rosner. Dessa forma, atendemos aos dois interesses desse trabalho. Para a comparação com a regressão logística,em cada replicação fazemos o teste baseado na estatística (2.1.5) sem a correção de continuidade. Ao final das numrep replicações, a aproximação para o poder do teste é a proporção de vezes que a hipótese H_o foi rejeitada. Usamos a estatística (2.1.6) para obter resultados para comparar com aqueles de Muñoz e Rosner.

No caso da regressão logística, tomamos a aproximação para o poder como a proporção de vezes que a hipótese K_o foi rejeitada.

No final das numrep replicações para cada distribuição de t_i's, r_i's, s_i's e razões de produtos cruzados especificadas, comparamos o poder obtido para o teste de Mantel-Haenszel com o poder obtido para a regressão logística.

Na verdade, o primeiro problema a ser resolvido para fazermos o Monte Carlo foi o da geração de tabelas 2x2xm, e a maneira como o solucionamos merece uma discussão especial. Convém lembrarmos a forma da Tabela 1.1.3 representando a i-ésima tabela resultante de um estudo de casos e controles estratificado. Um programa denominado Gera-tab* gera na i-ésima tabela duas amostras binomiais de um estudo de casos e controles, utilizando um gerador

(*) O leitor interessado em conhecer o programa poderá se comunicar com Departamento de Estatística da Universidade Federal da Bahia, Salvador _ Bahia.

de amostras binomiais. Cada conjunto dos vetores $t=(t_1,\ldots,t_m)$ e $s=(s_1,\ldots,s_m)$ foi considerado por nós, respectivamente, como um caso da distribuição de proporções de indivíduos nos m estratos e um caso da distribuição de casos e de controles nesses m estratos. Fixamos os vetores t e s e a razão de produtos cruzados e fixamos também o vetor $r=(r_1,\ldots,r_m)$. Quando fixamos as frações r_i, não o fizemos para que a soma a_i+b_i dos elementos da primeira linha da tabela fosse sempre igual a $r_i n_i$; o fizemos para obter um valor inicial a para a_i usando a aproximação de Cornfield para $E(A_i)$, e por diferença um valor inicial $b= r_i n_i - a$ para b_i. Uma das variáveis aleatórias binomiais da tabela tem parâmetros $s_i n_i$ e p_{1i} e a outra tem parâmetros $(1-s_i)n_i$ e p_{2i}. Se $s_i \leq 0.5$, p_{1i} é calculado como

$$p_{1i} = \frac{a}{r_i n_i} \qquad (3.1)$$

enquanto p_{2i} é calculado como

$$p_{2i} = \frac{(1-s_i)n_i - b}{(1-s_i)n_i} \qquad (3.2)$$

Nesse caso, o programa Gera-tab gera um valor de uma variável aleatória binomial com parâmetros $s_i n_i$ e p_{1i} como sendo o valor de a_i e um valor de uma outra variável aleatória binomial com parâmetros $(1-s_i)n_i$ e p_{2i} como o valor de d_i. c_i é então calculado como $s_i n_i - a_i$ e b_i é calculado como $(1-s_i)n_i - d_i$. Desse modo a soma $a_i + b_i$ é aleatória e sua média é $r_i n_i$. Se $s_i > 0.5$, p_{1i} é calculado como

$$p_{1i} = \frac{s_i n_i - a}{s_i n_i} \qquad (3.3)$$

e p_{2i} é calculado como

$$p_{2i} = \frac{b}{(1-s_i)n_i} \qquad (3.4)$$

Nesse caso, o programa Gera-tab gera o valor de c_i como o valor de uma variável aleatória binomial com parâmetros $s_i n_i$ e p_{1i} (p_{1i} dado por (3.3)) e gera o valor de b_i como o valor de uma variável aleatória com parâmetros $(1-s_i)n_i$ e p_{2i} (p_{2i} dado por (3.4)). a_i é então calculado como $s_i n_i - c_i$ e d_i é calculado como $(1-s_i)n_i - b_i$. E novamente a soma $a_i + b_i$ é aleatória com média $r_i n_i$.

Para solucionar o problema de, baseado nos dados gerados pelo programa Gera_tab, fazer o teste de Mantel_Haenszel, o ajuste dos modelos de regressão logística e o teste da razão de verossimilhança e estimar o poder dos dois testes, nós fizemos o programa Comparação(*).

(*) Mesma nota de rodapé da página 50.

CAPÍTULO 4

RESULTADOS DOS ESTUDOS DE MONTE CARLO

Inicialmente, temos a comentar que embora em seu artigo Muñoz e Rosner (1984) apresentem a Tabela 3.1 com o título que afirma os testes unilaterais terem sido feitos a um nível de significância 0.05, constatamos que esses resultados equivalem a um nível de significância bilateral de 0.10. Verificamos isso utilizando o programa Comparação, usando os valores críticos adequados a cada nível de significância unilateral especificado na FUNCTION Mantel-Haenszel, a qual faz parte do programa. Primeiro, para um nível de significância unilateral α=0.05, seguindo Muñoz e Rosner usamos como o valor crítico para o teste $Z_{0.05}$=1.645 na inequação (2.1.6) e sob a hipótese nula (2.1.1), obtivemos pelo Monte Carlo um poder igual a 0.10 quando deveríamos obter um poder 0.05, igual ao nível de significância unilateral adotado. Em seguida, fizemos simulações para obter estimativas para o poder do teste de Mantel_Haenszel a um nível de significância unilateral de 0.10 para uma parte das situações apresentadas por Muñoz e Rosner, usando (2.1.6). Os resultados para 2500 replicações estão na Tabela 4.1. Os resultados da Tabela 4.2 são os resultados do Monte Carlo para testes bilaterais com nível de significância 0.10 usando (2.1.5) e para exatamente as mesmas situações da Tabela 4.1. Os resultados dessa última tabela para o teste de Mantel_Haenszel são os mesmos da Tabela 4.1, a menos de uma diferença máxima 0.01 entre os valores das duas tabelas. Isso

significa que os resultados obtidos por Muñoz e Rosner são equivalentes a resultados de testes bilaterais com nível de significância 0.10.

Tabela 4.1. Resultados da simulação para aproximar o poder do teste de Mantel-Haenszel para a hipótese alternativa unilateral de que γ é maior do que zero no caso de dois estratos com razões de produtos cruzados ambas iguais a dois, $t_1 = t_2 = 0.5$ $r_1 + r_2 = 1 = s_1 + s_2$, $\alpha = 0.10$, $N = 208$.

r_1	s_1				
	0.2	0.4	0.5	0.6	0.8
0.2	0.59	0.67	0.62	0.54	0.36
0.4		0.79	0.78	0.74	0.55
0.5			0.81	0.78	0.64
0.6				0.78	0.67
0.8					0.58

Em concordância com o procedimento da razão de verossimilhança, no qual testamos a hipótese K_o de (2.3.4.1) versus a hipótese alternativa bilateral $K_1: \beta_1 \neq 0$, comparamos os dois procedimentos em questão fazendo testes bilaterais através da estatística (2.1.5).

Outro problema a ser discutido é a adequação da aproximação de Muñoz e Rosner para situações desbalanceadas. Comparando os resultados da tabela de poder 4.1, gerada pelo Monte Carlo, com os resultados de Muñoz e Rosner apresentados na Tabela 3.1, observamos que a maior diferença entre os valores correspondentes

Tabela 4.2. Resultados da simulação para dois estratos, com razões de produtos cruzados ambas iguais a dois, $t_1 = t_2 = 0.5$, $r_1 + r_2 = 1 = s_1 + s_2$, $\alpha = 0.10$, $N = 208$.

r_1	s_1				
	0.2	0.4	0.5	0.6	0.8
0.2	(*) 0.59	0.67	0.62	0.53	0.34
	(#) 0.58	0.67	0.63	0.55	0.40
0.4		0.78	0.78	0.74	0.55
		0.79	0.78	0.75	0.56
0.5			0.81	0.78	0.63
			0.81	0.79	0.64
0.6				0.78	0.67
				0.78	0.67
0.8					0.58
					0.57

(*) Aproximação para o poder do teste de Mantel-Haenszel.
(#) Aproximação para o poder do teste baseado na regressão logística.

a cada situação nas duas tabelas é 0.06, o que ocorre na situação em que r_1 é igual a 0.2 e s_1 é igual a 0.6. Intervalos de confiança de 95% com amplitude 0.02 de cada lado para o poder do teste de Mantel-Haenszel nessa situação baseados em cada uma das duas tabelas não se interceptam. A rigor, na situação em que r_1 é igual a 0.2 e s_1 é igual a 0.8, os intervalos de confiança de 95% com amplitude 0.02 de cada lado não se interceptam. Idem, na situação em que r_1 é igual a 0.4 e s_1 é igual a 0.8. Nessas situações podemos concluir que os dois poderes diferem. Devemos observar também que a aproximação de Muñoz e Rosner é maior do que a aproximação dada pelo Monte Carlo nessas três situações. Testes unilaterais com nível de significância 0.05 para a hipótese de que

a aproximação de Muñoz e Rosner é maior do que a aproximação pelo Monte Carlo nos levam a rejeitar essa hipótese em cada uma das três situações acima, nos sugerindo que a aproximação de Muñoz e Rosner superestima o poder do teste de Mantel Haenszel nas situações desbalanceadas.

Finalmente, fizemos comparações entre as estimativas para o poder dos dois testes em diversas tabelas de poder obtidas pelo Monte Carlo no intuito de saber se há diferença significante entre os dois poderes. A Tabela 4.3 apresenta os resultados para a situação desbalanceada em que a proporção t_1 é igual a 0.3 e a proporção t_2 é igual a 0.7. Observando as tabelas 4.2 e 4.3, percebemos que há notadamente um padrão seguido em cada uma dessas tabelas. Em ambas, as aproximações para o poder dos dois testes são iguais, exceto na situação quase extrema em que a proporção r_1 é igual a 0.2 e a proporção s_1 é iguala 0.8. Essa é uma situação em que há uma proporção muito grande de casos em um dos estratos, e, como $s_1+s_2=1$, a proporção de casos no outro estrato é muito pequena; e além disso, a proporção de expostos é muito pequena no estrato em que a proporção de casos é muito grande, e é muito grande no estrato em que a proporção de casos é muito pequena. O que observamos das duas tabelas acima , é que, nessa situação, a diferença entre as duas estimativas é da ordem de 0.06 em ambas as tabelas, Isso implica que intervalos de confiança com amplitude 0.04 para os dois poderes não se interceptam. De fato, podemos ver pela Tabela 4.2, que os intervalos de confiança de 95% para os dois poderes, baseados nas estimativas dadas pelo Monte Carlo nessa situação desbalanceada são, aproximadamente, (0.34-0.02,0.34+0.02)=(0.32,0.36) para o poder do teste de Mantel-Haenszel e (0.40-0.02,0.40+0.02) = (0.38,0.42) para o poder

Tabela 4.3. Resultados da simulação para dois estratos, com razões de produtos cruzados ambas iguais a dois, t_1=0.3, t_2=0.7, $r_1 + r_2 = 1 = s_1 + s_2$, α=0.10, N = 208.

r_1	s_1				
	0.2	0.4	0.6	0.6	0.8
0.2	(*) 0.50	0.57	0.63	0.65	0.47
	(#) 0.48	0.57	0.63	0.67	0.53
0.4		0.70	0.74	0.76	0.56
		0.70	0.74	0.77	0.58
0.5			0.81	0.82	0.63
			0.81	0.82	0.64
0.6				0.83	0.67
				0.83	0.67
0.8					0.51
					0.49

(*) Aproximação para o poder do teste de Mantel-Haenszel.

(#) Aproximação para o poder do teste baseado na regressão logística.

da regressão logística. Esses dois intervalos não se interceptam, o que nos leva a concluir que os dois poderes são diferentes nessa situação. O mesmo podemos concluir da Tabela 4.3. Em particular, para a situação desbalanceada discutida acima, os intervalos com 95% de confiança são (0.47-0.02,0.47+0.02)=(0.45,0.49) para o poder do teste de Mantel-Haenszel e (0.53-0.02,0.53+0.02) = (0.51,0.55) para o poder da regressão logística.

A Tabela 4.4 apresenta os resultados para um outro conjunto de simulações em que as proporções t_1 e t_2 são iguais a 0.5, mas a razão de produtos cruzados são diferentes, embora em média seja igual a dois. As mesmas observações do parágrafo anterior são válidos para os resultados dessa tabela. Os intervalos com 95% de

confiança na situação desbalanceada discutida são (0.30-0.02,0.30+0.02)=(0.28,0.32) para o poder de Mantel-Haenszel e (0.35-0.02,0.35+0.02)=(0.33,0.37) para a regressão logística. Também aqui os intervalos não se interceptam e a conclusão é que os dois poderes são diferentes nessa situação.

Não é apenas o fato de que os dois poderes são diferentes na situação desbalanceada discutida nos dois parágrafos anteriores o. que podemos perceber. Percebemos também dessas discussões que os limites inferiores dos intervalos de confiança para o poder da regressão logística foram sempre maiores do que os limites superiores dos intervalos de confiança para o poder do teste de Mantel-Haenszel. Um teste de nível 0.05 para a hipótese H_1: prl = pmh versus a hipótese H_2: prl > pmh leva à rejeição de H_1 quando baseado em cada uma das tabelas 4.2, 4.3 e 4.4. A nossa conclusão é que o poder da regressão logística é maior do que o poder do teste de Mantel-Haenszel na situação quase extrema discutida acima e que os dois poderes são iguais para outras situações, quer os t_i's sejam balanceados ($t_1=t_2=0.5$) ou não. Resumindo esse parágrafo, podemos dizer que o poder da regressão logística é sempre igual ou maior do que o poder do teste de Mantel_Haenszel.

Observamos ainda que, na mesma situação quase extrema discutida nos parágrafos anteriores, quando mudamos da Tabela 4.2 para a Tabela 4.3, o que é feito desbalanceando os t_i's, ambas as estimativas crescem significantemente. Para o poder do teste de

Tabela 4.4. Resultados da simulação para dois estratos, com razão de produtos cruzados igual a 1.5 no estrato 1 e 2.5 no estrato 2, $t_1 = t_2 = 0.5$, $r_1 + r_2 = 1 = s_1 + s_2$, $\alpha = 0.10$, N=208.

r_1	s_1				
	0.2	0.4	0.5	0.6	0.8
0.2	(*) 0.56	0.64	0.58	0.49	0.30
	(#) 0.54	0.65	0.59	0.51	0.35
0.4		0.76	0.75	0.69	
		0.76	0.75	0.69	
0.5			0.78	0.69	0.60
			0.78	0.70	0.60
0.6				0.75	0.64
				0.75	0.64
0.8					0.56
					0.54

(*) Aproximação para o poder do teste de Mantel-Haenszel.

(#) Aproximação parao poder do teste baseado na regressão logística.

Mantel-Haenszel a estimativa passa de 0.34 para 0.47. A diferença entre essas duas estimativas é 0.13. Um intervalo de confiança de 95% para a verdadeira diferença só incluiria o zero se tivesse amplitude pelo menos 0.13 de cada lado. Um teste de nível 0.05, nos leva a concluir que o poder do teste de Mantel-Haenszel é maior para t_i's desbalanceados como os da Tabela 4.3 do que para o caso de t_i's bem balanceados, como na Tabela 4.2. As mesmas observações podemos fazer para a regressão logística. Concluimos que estando numa situação extrema em que amostramos uma proporção muito grande de casos em um estrato que tem uma proporção pequena de expostos e amostramos uma proporção muito pequena de casos no outro estrato, e é sabido que a proporção de expostos é muito grande no estrato que tem poucos casos é muito pequena no estrato

que tem muitos casos, obteremos maior poder, tanto para o teste de Mantel-Haenszel como para o teste equivalente usando regressão logística, se, amostrarmos mais no estrato onde a proporção de expostos é maior. Isso significa que nessa situação extrema podemos obter maior poder amostrando mais no estrato no qual há mais expostos e menos casos.

Para três estratos, procuramos fazer simulações tentando obter algumas posições da tabela de poder, assim como o fizemos para dois estratos. Para três estratos, uma tabela de poder como as tabelas apresentadas até aqui, contém muito mais posições do que essas últimas, desde que tal tabela deve conter todas as combinações dos elementos dos vetores $r=(r_1,r_2,r_3)$ e $s=(s_1,s_2,s_3)$. As Tabelas 4.5 e 4.6 apresentam os resultados para dois conjuntos de simulações que diferem no fato de que as posições dos t_i's são invertidas. O que podemos perceber é que não há diferença entre os dois poderes, mesmo no canto superior direito da tabela de poder,como acontece em estudos com dois estratos. Os intervalos de confiança de 95% para os poderes se interceptam em todas as situações consideradas, balanceadas ou não. Uma informação adicional é a de que para cada distribuição de r_i's, e s_i's considerada, quando passamos da distribuição dos t's da Tabela 4.5 para a distribuição dos t's da Tabela 4.6, tanto o teste de Mantel-Haenszel como a regressão logística sofrem uma perda de poder de cerca de 10%, sendo excessão o canto superior esquerdo para o qual a perda é de 0.18. Um terceiro conjunto de simulações foi realizado para uma situação o mais balanceada possível quando

Tabela 4.5 . Resultados da simulação para três estratos com razões de produtos cruzados 1.5 no primeiro estrato, 2.0 no segundo estrato e 2.5 no terceiro estrato, t_1=0.2, t_2=0.3, t_3=0.5, α = 0.10, e N=208.

r_1			s_1	s_2	s_3				
r_2 / r_3	0.2	0.2	0.2	0.5	0.5	0.5	0.8	0.8	0.8
0.2	(*)	0.66						0.50	
0.2	(≠)	0.64						0.56	
0.2									
0.5					0.85				
0.5					0.85				
0.5									
0.8			0.49						
0.8			0.54						
0.8									

(*) Aproximação para o poder do teste de Mantel-Haenszel.

(≠) Aproximação para o poder da regressão logística.

o número de estratos é três. Nesse caso consideramos todos t_i's iguais a 0.33, todos os r_i's e s_i's iguais a 0.5 e todas as razões de produtos cruzados iguais a dois. Também nesse caso podemos concluir que os dois poderes são iguais, uma vez que as estimativas para os dois poderes são, ambas 0.70. Isso significa uma perda de poder de cerca de 10% em relação ao caso de dois estratos, no qual a situação mais balanceada possível com razão de produtos cruzados igual a dois para os dois estratos apresenta poder 0.81 para os dois testes.

Tabela 4.6 . Resultados da simulação para três estratos com razões de produtos cruzados 1.5 no primeiro estrato, 2.0 no segundo estrato e 2.5 no terceiro estrato, t_1=0.5, t_2=0.3, t_3=0.2, α 0.10, e N=208.

r_1			s_1	s_2	s_3				
r_2 / r_3	0.2	0.2	0.2	0.5	0.5	0.5	0.8	0.8	0.8
0.2	(*)	0.48						0.41	
0.2	(≠)	0.47						0.46	
0.2									
0.5					0.71				
0.5					0.71				
0.5									
0.8			0.37						
0.8			0.43						
0.8									

(*) Aproximação para o poder do teste de Mantel-Haenszel.

(≠) Aproximação para o poder da regressão logística.

CAPÍTULO 5

CONCLUSÕES

O primeiro problema a que nos propusemos estudar, foi a adequação da aproximação de Muñoz e Rosner para o poder do teste de Mantel-Haenszel para situações não balanceadas. Das observações feitas no Capítulo 4 quando comparamos os resultados originados do trabalho dos dois autores com os resultados originados pelo estudo de Monte Carlo, chegamos à conclusão de que a aproximação de Muñoz e Rosner superestima o poder do teste de Mantel-Haenszel nas situações em que as proporções de casos e expostos são desbalanceadas, se distanciando de 0.5. Nas situações balanceadas e próximas a isso a aproximação é boa e pode ser usada, o que leva o pesquisador a uma maior facilidade computacional no cálculo do poder do teste de Mantel-Haenszel.

Para estudos estratificados com apenas dois estratos, o poder do teste para a significância do risco relativo comum aos dois estratos usando regressão logística é maior do que o poder do teste de Mantel-Haenszel para a mesma hipótese quando estamos numa situação em que as proporções de casos e expostos são muito diferentes uma da outra dentro dos estratos e entre estratos, como é o caso do canto superior direito da tabela de poder. Nessa situação, concluimos que a regressão logística é mais vantajosa do que o procedimento de Mantel-Haenszel, quer os t_i's sejam

balanceados ou não. Nas demais situações estudadas, os dois poderes são iguais e nós concluimos que as duas metodologias são equivalentes em termos de poder.

Ainda para o caso de dois estratos, pudemos concluir que nas situações muito desbalanceadas discutidas no parágrafo anterior, podemos obter maior poder, tanto para o teste de Mantel_Haenszel como para a regressão logística, se amostrarmos mais no estrato no qual há mais expostos.

Para três estratos, da discussão do último parágrafo do Capítulo 4, concluimos efetivamente que para estudos estratificados com um número de estratos igual a três obteremos o mesmo poder, quer usemos o procedimento de Mantel-Haenszel ou a regressão logística.

REFERÊNCIAS BIBLIOGRÁFICAS

Anderson,S., Auquier,A., Hauck,W.W., Oakes,David, Vandaele,W., Weisberg,H.I. with contributions from Bryk,A.S. and Kleinman,J. (1980). Statistical Methods for Comparative Studies. *John Wiley & Sons*.

Breslow,N.E. and Day,N.E. (1980). Statistical Methods in Cancer Research. Vol. 1- The Analysis of Case-control Studies. *IACR Scientific Publication* N. 32.

Conover,W.J. (1968). Uses and Abuses of the Continuity Correction. *Biometrics*, 24, 1028.

Conover,W.J. (1974). Some Reasons for not Using the Yates Continuity Correction on 2x2 Contingency Tables. (With comments). *J. Am. Stat. Assoc.*, 69, 374-382.

Cornfield,J. (1956). A Statistical Problem Arising from Retrospective Studies. *Proceedings of the Third Simposium on Mathematical Statistics and Probabilidade, Vol. IV, J. Neyman (ed.) 135-148. Berkeley: University of California Press.*

Cox,D.R. (1970). The Analysis of Binary Data. *London: Methuen.*

Day,N.E. and Byar,D.P. (1979). Testing Hypoteses in Case-control Studies - Equivalence of Mantel-Haenszel Statistics and Logit Score Tests. *Biometrics*, 35, 623-630.

Fleiss,J.L. (1981). Statistical Methods for Rates and Proportions. *John Wiley & Sons.*

Grizzle,J.E. (1967). Continuity Correction in the χ^2-test for 2x2 Tables. *Am. Stat.*, 21 (october), 28-32.

Harkness,W.L. (1965). Properties of the Extended Hypergeometric Distribution. *Annals of Mathematical Statistics*, **46**, 938-945.

Levin,B. (1982). On the Accuracy of a Normal Approximation to the Power of the Mantel-Haenszel Procedure. *Journal of Statistical Computing and Simulation*, **14**, 201-218.

Mantel,N. (1973). Synthetic Retrospective Studies and Related Topics. *Biometrics*, **29**, 479-486.

Mantel,N. and Haenszel,W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute*, **22**, 719-748.

Mote,V.L., Pavate,M.V. e Anderson,R.L. (1958). Some Studies in the Analysis of Categorical Data. *Biometrics*, **14**, 572-573.

Muñoz,A and Rosner,B. (1984). Power and Sample Size for a Colletion to 2x2 Tables. *Biometrics*, **40**, 995-1004.

Pearson,E.S. (1947). The Choice of Statistical Tests Illustratedon the Interpretation of Data Classed in a 2x2 Table. *Biometrika*, **34**, 139-167.

Plackett,R.L. (1964). The Continuity Correction in 2x2 Tables. *Biometrika*, **51**, 327-337.

Siegel, D. G. and Greenhouse, S. W. (1973). Multiple Relative Risk Functions in Case-control Studies. *American Journal of Epidemiology* Vol. 97, n. 5.