UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE MATEMATICA

DEPARTAMENTO DE ESTATISTICA

REGRESSAO LOGISTICA POLITOMICA NOMINAL E ORDINAL

AUTORA: PATRICIA C. KLARMANN

ORIENTADOR: PROF. ALVARO VIGO

CO-ORIENTADOR: PROFA. JANDYRA M. G. FACHEL

MONOGRAFIA APRESENTADA PARA OBTENÇAO DO TITULO
DE BACHAREL EM ESTATISTICA

PORTO ALEGRE, DEZEMBRO DE 1993.

AGRADECIMENTOS

Ao prof. e amigo Álvaro Vigo pela incansável e valiosa orientação.

À profª e amiga Jandyra Fachel por todo apoio e amizade no decorrer do curso.

A todos os professores do departamento de estatística da UFRGS pelos ensinamentos.

Aos colegas e amigos pela convivência e amizade.

Aos professores Fernando Barros e César Victora, do Centro de Pesquisas Epidemiológicas da Faculdade de Medicina da UFPEL por ter cedido os dados utilizados no Exemplo 2.

"RARAMENTE, OU NUNCA, OS DADOS ESTATÍSTICOS FALAM POR SI MESMOS. SÓ HABILMENTE COLETADOS E CRITICAMENTE INTERPRETADOS PODEM SER EXTREMAMENTE ÚTEIS".

F. GONÇALVES

ÍNDICE

CAPITULO 1: INTRODUÇÃO

Toda civilização está repleta de idéias sobre números e medidas. As ciências físicas foram as primeiras a sentir necessidade de expressões quantitativas. O desejo de precisão leva a exprimir quantativamente os fatos relativos às características de interesse. Um dos principais objetivos da ciência é identificar no meio da complexidade do mundo exterior o funcionamento daquilo que podemos chamar de lei (a realidade inatingível). A teoria estatística, ou simplesmente Estatística, designa a exposição de métodos estatísticos que são métodos especialmente adaptados à elucidação de dados quantitativos sujeitos a influência de uma multiplicidade de causas. Estes métodos possibilitam a tomada de decisões acertadas, face às incertezas. O presente trabalho trata de um método estatístico utilizado para descrever as relações entre uma variável categórica dependente e uma ou mais variáveis explicativas, chamado de regressão logística. Antes de apresentar esta técnica de análise estatística, é necessário, entanto, introduzir alguns conceitos e definições relativos aos tipos de variáveis que frequentemente são utilizadas. Posteriormente, também será brevemente descrita a estrutura desta monografia.

1.1. VARIÁVEIS CATEGÓRICAS

Um sistema de medida é um procedimento através do qual podemos atribuir números ou outros rótulos a indivíduos (pessoas, objetos ou eventos), de acordo com uma regra determinada. A regra usualmente especifica as categorias de um atributo ou algum aspecto quantitativo de uma observação, definindo assim uma <u>escala de medida</u>. As escalas de medida são classificadas como <u>nominais</u>, <u>ordinais</u>, de <u>intervalo</u> ou de <u>razão</u>, podendo medir variáveis discretas ou contínuas, veja Cureton (1978, p.764). A principal característica da escala nominal é que os números possuem apenas a função de classificar os elementos em categorias exclusivas e exaustivas. Na escala ordinal, por sua vez, os elementos não apenas diferem de categoria, mas estas são ordenadas segundo algum critério. Em uma escala de intervalo as diferentes categorias, além de estarem ordenadas, possuem suas distâncias numericamente determinadas e o ponto zero é relativo, isto é, não indica ausência total do atributo que está sendo medido. Por fim, a escala de razão difere da intervalar por apresentar um ponto zero verdadeiro ou absoluto, indicando a completa ausência da característica. Estas diferentes escalas de medida podem medir variáveis discretas ou contínuas. Variáveis cuja escala de medida consiste de um conjunto de categorias disjuntas são chamadas de <u>variáveis categóricas</u>. As variáveis categóricas compreendem, portanto, a escala nominal e ordinal. Elas surgem com grande frequência em diversas áreas do conhecimento, tais como educação, ciências sociais, medicina, etc. Por exemplo, o estado de evolução de uma doença pode ser medido como "doença progressiva", "remissão parcial" ou "remissão completa".

Existem diversos tipos de variáveis categóricas, sendo distinguíveis basicamente pela escala de medida utilizada. As variáveis categóricas nominais caracterizam-se pelo fato de que não existe uma ordem natural dos seus níveis ou categorias. Exemplos de variáveis nominais são raça (branca, negra, outras) e estado civil (casado, solteiro, divorciado, viúvo, desquitado). Para as variáveis nominais, a ordem em que aparecem as categorias deveria ser irrelevante na análise estatística; isto é, diferentes permutações na ordem deveriam conduzir aos mesmos resultados.

No entanto, em muitas variáveis categóricas existe uma ordem natural dos seus níveis, mas as distâncias absolutas entre eles são desconhecidas ou sequer estão definidas. Essas variáveis são chamadas de categóricas ordenadas. O exemplo anterior, relativo ao estado de evolução de uma doença, constitui uma aplicação na área médica. Outros exemplos são classe social (baixa, média, alta) e grau de concordância (regular, bom, ótimo). Variáveis contínuas medidas através de escores ou postos (rank, em inglês) também são tratadas como categóricas ordenadas.

Diversos métodos estatísticos para análise da dados categóricos são encontrados na literatura, veja Agresti (1990), Everitt (1992) e Bishop, Frenberg and Holland (1975). Por exemplo, para descrever as relações entre duas variáveis categóricas, métodos estatísticos tradicionais como o χ^2 de Pearson, teste exato de Fischer, teste de Mcnemar ou Teste U de Mann-Witney poderiam ser empregados, de acordo com a situação específica. Entretanto, nenhum destes métodos consegue descrever a magnitude destas relações, que muitas vezes é o principal objetivo do investigador. Para tanto são necessárias técnicas estatísticas mais elaboradas, dentre as quais encontra-se o método de regressão logística, tema central deste trabalho. A seguir comentaremos brevemente a estrutura da monografia.

1.2. ESTRUTURA DA MONOGRAFIA

Nos estudos epidemiológicos, por exemplo, com relativa frequência a resposta de interesse não pode ser medida quantitativamente, assumindo valores de acordo com uma escala nominal ou ordinal. Quando a resposta assume apenas dois atributos, tais como presença e ausência da doença, ela é denominada dicotômica ou binária. Por outro lado, se os indivíduos observados podem ser classificados segundo duas ou mais categorias, então a variável resposta é chamada de politômica. Como vimos, ela pode ser nominal ou ordinal, de acordo com a presença ou não de uma ordem natural de seus níveis.

Nesta monografia vamos apresentar uma técnica para análise estatística denominada regressão logística. A regressão logística é um método estatístico relativamente novo. Talvez por este motivo não existe na literatura um texto que englobe todos os seus aspectos. Este trabalho pretende ser um texto simples com o objetivo principal de apresentar o método de <u>regressão logística politômica</u>, aglutinando idéias de alguns autores. Espera-se que ele sirva como uma fonte inicial para o leitor, com algum conhecimento em regressão logística tradicional, que queira estudar a regressão logística politômica. Serão apresentados os modelos de regressão logística politômica nominal e ordinal bem como dois exemplos para ilustrar estes métodos. Apresentaremos ainda um capítulo sobre recursos computacionais onde será descrito os procedimentos disponíveis nos pacotes estatísticos STATA e SAS para o ajuste destes modelos.

CAPÍTULO 2: REGRESSÃO LOGÍSTICA

2.1. CONCEITOS BÁSICOS

Um método estatístico muito utilizado para descrever as relações entre uma variável dependente e uma ou mais variáveis independentes é a regressão linear. Este método é usualmente empregado quando a variável dependente é contínua. Quando a variável dependente se apresenta de forma categórica, um método estatístico frequentemente empregado é o de regressão logística. Por analogia comparamos a regressão logística com a regressão linear.

Segundo Hosmer & Lemeshow (1989, p.5) uma diferença entre esses métodos de regressão está na relação da variável dependente com as variáveis explicativas. Na regressão linear as relações entre a variável dependente Y (também chamada de variável resposta) e um conjunto de variáveis explicativas x(também chamadas de covariáveis) podem ser expressas através de uma equação linear dos parâmetros. Por exemplo, no caso onde temos apenas uma covariável, esta relação pode ser descrita por

$$E\,(Y|x) = \beta_0 + \beta_1 x$$

onde β_0 é o intercepto e β_1 o coeficiente angular da reta.

Na regressão logística a quantidade $E(Y|x)$ não pode ser expressa por uma função linear, pois a variável resposta é categórica. /Mariana pg. 8 Para intoduzir este modelo usaremos o caso onde temos uma variável resposta binária (indicando, por exemplo, ausência ou presença de um sintoma, de uma doença, ou outra característica qualquer) e uma covariável. Por convenção, usaremos o código 1 (isto é, $Y=1$) para indicar a presença da característica de interesse e o código 0 ($Y=0$) para indicar a ausência. Neste modelo as relações entre a variável resposta e a covariável são descritas através de uma função das probabilidades de uma resposta $Y=1$, dado os valores da covariável X. / Para ilustrar, vejamos o caso onde a variável resposta Y seja a presença ou ausência de câncer e a covariável x a idade do indivíduo. Para tentar buscar as relações entre a idade e o fato de ser ou não portador de câncer podemos usar a regressão logística. A idéia é categorizar a covariável idade e para cada categoria calcular a proporção de indivíduos que são portadores de câncer. Fazendo o gráfico com a proporção de indivíduos portadores de câncer contra os pontos médios dos intervalos das categorias de idade veremos que ele se comporta em forma de S. Isto sugere que ajustemos a $E(Y|x)$ através de uma função logística. /* H e L pg. 6 Denotando $\pi(x)=E(Y|x)=P(Y=1|x)$ temos que

$$\pi(x)=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}.$$

A transformação logit neste modelo especifica que

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

onde $g(x)$ é chamada de <u>função logit</u>. Verifica-se que a função logit é linear nos parâmetros β_0 e β_1. Portanto, o problema na regressão logística é a estimação dos parâmetros β_0 e β_1.

Convém salientar que na regressão logística a variável resposta pode ser categórica nominal ou ordinal e apresentar duas ou mais categorias. As covariáveis podem ser apresentadas em qualquer escala de medida. Chamaremos de regressão logística tradicional aquela onde a variável resposta pode assumir valores em apenas duas categorias; de regressão logística politômica nominal quando a variável resposta for nominal com mais de duas categorias e regressão logística politômica ordinal quando a variável resposta for ordinal.

Neste trabalho não nos preocuparemos com o detalhamento da regressão logística tradicional. Ela é apresentada, por exemplo, em Hosmer & Lemeshow (1989) e Randunz (1992). Na Seção 2.2 vamos apresentar a regressão logística politômica nominal e no capítulo 3 um modelo de regressão para uma variável resposta categórica ordenada.

2.2. REGRESSÃO LOGÍSTICA POLITÔMICA NOMINAL

Tradicionalmente a regressão logística é utilizada para descrever as relações entre uma variável resposta binária e um conjunto de covariáveis. Porém, a regressão logística também se aplica ao caso onde a variável resposta possua mais de duas categorias. Sendo assim, regressão logística politômica é um modelo de regressão caracterizado pelo fato de que a variável resposta Y apresenta mais de duas categorias. Discutiremos neste capítulo o modelo utilizado quando as categorias da variável resposta se apresentam segundo uma escala nominal. Por simplicidade, nos restringiremos ao caso onde a variável resposta possui apenas $\kappa=3$ categorias. Para um modelo onde a variável resposta apresenta mais de três categorias o estudo é análogo.

Vamos assumir que as categorias da variável resposta Y sejam codificadas por 0,1 e 2. Neste caso o modelo de regressão logística politômico terá duas funções logit. Uma para comparar uma resposta na categoria $Y=1$ com $Y=0$ e outra para comparar $Y=2$ com $Y=0$. Fazendo a diferença destas duas funções logit obtemos a comparação $Y=1$ com $Y=2$. Estamos considerando a categoria $Y=0$ como categoria de referência, embora qualquer uma possa ser utilizada para essa finalidade. Normalmente utilizamos como categoria de referência aquela de maior interesse. No caso geral, podemos dizer que o modelo de regressão logística politômica é construído através do ajuste de $\kappa-1$ modelos de regressão logística tradicional; ou seja, através de $\kappa-1$ comparações com a categoria de referência. Essas comparações geram as funções logit.

Para desenvolver este método usaremos a notação empregada por Hosmer & Lemeshow (1989). Seja $x=(x_0,x_1,...,x_p)$ o vetor com as p covariáveis utilizadas no modelo. Sendo assim, a dimensão do vetor x é $p+1$, onde $x_0=1$. As duas funções logit para este caso são definidas como segue. A primeira,

$$g_1(x_i)= \ln \left[\frac{P(Y=1 \mid x_i)}{P(Y=0 \mid x_i)} \right] = \beta_{10}+\beta_{11}x_1+\beta_{12}x_2+...+\beta_{1p}x_p \ ,$$

representa um modelo de regressão logística tradicional utilizado para comparar uma resposta na categoria Y=1 com uma resposta na categoria de referência (isto é, Y=0).A segunda função logit,

$$g_2(x_i) = \ln \left[\frac{P(Y=2 \mid x_i)}{P(Y=0 \mid x_i)} \right] = \beta_{20}+\beta_{21}x_1+\beta_{22}x_2+...+\beta_{2p}x_p \ ,$$

representa um modelo de regressão logística tradicional utilizado para comparar uma resposta na categoria Y=2 com uma resposta na categoria de referência (Y=0).

Quando o vetor de covariáveis assume o valor x_i, para $i=1,...,n$ onde n é o número de combinações possíveis para o vetor de covariáveis x, as probabilidades condicionais para cada categoria da variável resposta são definidas por

$$P(Y=0|x_i) = \frac{1}{1 + \exp[g_1(x_i)] + \exp[g_2(x_i)]} = \pi_0(x_i)$$

$$P(Y=1|x_i) = \frac{\exp[g_1(x_i)]}{1 + \exp[g_1(x_i)] + \exp[g_2(x_i)]} = \pi_1(x_i)$$

$$P(Y=2|x_i) = \frac{\exp[g_2(x_i)]}{1 + \exp[g_1(x_i)] + \exp[g_2(x_i)]} = \pi_2(x_i).$$

Com o objetivo de estimar os coeficientes β_{ij}, para $j=1,...,K$ onde K é o número de categorias da variável resposta, das funções logit usamos, assim como na regressão logística tradicional, o método de máxima verossimilhança. Para tanto calcularemos a função de verossimilhança do modelo.

Para clarear a função de máxima verossimilhança, criaremos três variáveis binárias (Y_0,Y_1,Y_2) codificadas com 0 e 1. Elas são definidas da seguinte maneira: se $Y=0$ então$Y_0=1$, $Y_1=0$ e $Y_2=0$; se $Y=1$ então $Y_0=0$, $Y_1=1$ e $Y_2=0$; e se $Y=2$ então $Y_0=0,Y_1=0$ e $Y_2=1$.

A função de verossimilhança para uma amostra de n observações independentes é dada por

$$l(\underset{\sim}{\beta}) = \prod_{i=1}^{n} \left[\pi_0(x_i)^{Y_{0i}} \ \pi_1(x_i)^{Y_{1i}} \ \pi_2(x_i)^{Y_{2i}} \right] .$$

Para facilitar os cálculos usamos o logaritmo da função de verossimilhança e o fato de que $\sum Y_{ij}=1$. Então,

$$L(\underset{\sim}{\beta})=\sum_{i=1}^{n}\left\{ Y_{1i}g_1(x_i) + Y_{2i}g_2(x_i) - \ln\left[1 + \exp[g_1(x_i)] + \exp[g_2(x_i)]\right]\right\}.$$

Calculando a primeira derivada parcial dessa função em relação a cada um dos 2(p+1) parâmetros desconhecidos e igualando estas derivadas a zero, obtemos as equações de verossimilhança. A solução dessas equações nos fornecem os estimadores de máxima verossimilhança $\hat{\beta}_{ij}$. Como estas equações são funções não lineares nos parâmetros, é necessário o uso de métodos iterativos para resolvê-las . Diversos métodos iterativos são citados na literatura e utilizados nos pacotes estatísticos. Por exemplo, o pacote estatístico STATA utiliza o método iterativo de Newton-Raphson. Para maiores detalhes desses procedimentos iterativos veja, por exemplo, Agresti (1990, p. 445) ou Dobson (1983).

Calculando a segunda derivada parcial do logaritmo da função de verossimilhança obtemos uma matriz de dimensão 2(p+1) x 2(p+1) chamada de Matriz de Informação e denotada por $I(\underset{\sim}{\beta})$. A matriz assintótica de covariância dos parâmetros β_i denotada por $\sum(\underset{\sim}{\beta})$ é definida pela inversa da matriz de informação, isto é,

$$I(\underset{\sim}{\beta}) = \left[\frac{-\,\delta^2 L(\beta)}{\delta\,\underset{\sim}{\beta}\,\delta\,\underset{\sim}{\beta}^t}\right] \text{ e, assim, } \left[I(\underset{\sim}{\beta})\right]^{-1} = \sum(\underset{\sim}{\beta})\,.$$

A estimativa da matriz de informação é obtida pela substituição dos parâmetros desconhecidos β_{ij} pelos respectivos valores estimados $\hat{\beta}_{ij}$.

2.2.1. AVALIANDO A SIGNIFICÂNCIA DO MODELO

Uma etapa muito importante na modelagem estatística é avaliar se o modelo é adequado para fazer inferências. Alguns aspectos serão descritos nesta seção.

AJUSTE DO MODELO: Dizemos que um modelo estatístico para descrever as relações entre uma variável resposta e uma ou mais covariáveis está bem ajustado se as diferenças entre os valores observados e os preditos pelo modelo é mínima. Portanto, é fundamental a avaliação do ajuste do modelo. No modelo de regressão logística politômica este princípio baseia-se na comparação dos valores observados da variável resposta com os respectivos valores preditos pelo modelo ajustado. Primeiramente, esta comparação é feita para cada uma das funções logit, de maneira análoga ao da regressão logística tradicional.Depois integramos estes resultados apenas de forma descritiva. Alguns métodos para verificar o ajuste do modelo são citados na literatura: o teste do χ^2 de Pearson baseado nos resíduos, teste de Hosmer & Lemeshow e teste da razão de verossimilhança. Neste trabalho daremos uma atenção especial ao teste da razão de verossimilhança. Para maiores detalhes, veja

Hosmer & Lemeshow (1989).

O teste da razão de verossimilhança é usado para testar a hipótese nula de que o modelo está bem ajustado. A estatística de teste, chamada de Deviance e denotada por D, baseia-se na comparação dos logaritmos das funções de verossimilhança, dada por

$$D = -2 \ln \left[\frac{\text{Verossimilhança do modelo corrente}}{\text{Verossimilhança do modelo saturado}} \right].$$

A estatística D é comparada com a distribuição de referência de χ^2 de Pearson. O número de graus de liberdade (gl) é dado pelo número de combinações (das categorias das covariáveis) possíveis menos o número de parâmetros do modelo ajustado, ou seja, $[n - (p+1)]$. Se o valor da estatística D for menor que o valor do χ^2 de Pearson, ao nível de significância α, não rejeitamos a hipótese nula e dizemos que há evidências de que o modelo está bem ajustado.

TESTE DE SIGNIFICANCIA DAS COVARIÁVEIS: Não basta apenas sabermos se o modelo está ou não bem ajustado. Ao encontrarmos um modelo com um bom ajuste devemos também testar se uma determinada covariável produz um impacto significante na resposta. Isto pode ser útil para simplificar o modelo, ou seja, as covariáveis que forem detectadas como não significantes podem ser eliminadas. Em geral, é preferível um modelo mais simples que mantenha um bom ajuste, pois temos a vantagem de que num modelo simplificado a interpretação é mais prática. Pelo menos três testes para verificar a significância das covariáveis são citados na literatura: teste da razão de verossimilhança, teste de Wald e teste dos escores, veja Hosmer & Lemeshow (1989).

Neste caso, o teste da razão de verossimilhança é usado para testar a hipótese nula de que o efeito provocado na variável resposta pelas covariáveis do modelo é nulo. A estatística de teste usada é a G, baseada na comparação da estística D do modelo sem as covariáveis com o modelo com as covariáveis. Cabe salientar que este teste pode ser feito para testar apenas uma covariável isoladamente ou qualquer subconjunto delas.

$$G = D(\text{modelo sem as covariáveis}) - D(\text{modelo com as covariáveis}).$$

Observando que a verossimilhança do modelo saturado é a mesma para os dois modelos que estão sendo comparados podemos expressar a estatística G por

$$G = -2 \ln \left[\frac{\text{Verossimilhança do modelo sem as covariáveis}}{\text{Verossimilhança do modelo com as covariáveis}} \right].$$

A estatística G é comparada com a distribuição de referência de χ^2 de Pearson. O número de graus de liberdade é dado pelo número de parâmetros estimados no modelo sem as covariáveis menos o número de parâmetros estimados no modelo com as covariáveis. Se a estatística G for maior que o valor do χ^2 de Pearson, ao nível de significância α, então rejeitamos a hipótese nula e dizemos que há evidências de que o efeito da covariável não é nulo, ou seja, que a covariável produz um impacto significante.

Outro método para testar a significância das covariáveis é o teste de Wald. Este método testa a hipótese nula de que o coeficiente de regressão estimado é igual a zero. A estatística de Wald baseia-se na comparação do estimador de máxima verossimilhança pelo seu respectivo erro padrão. Esta estatística é expressa por

$$W = \frac{\hat{\beta}_{ij}}{\widehat{SE}\,(\hat{\beta}_{ij})}.$$

A estatística W é comparada com a distribuição de referência t de Student com graus de liberdade dado por: número de indivíduos menos 1. Para amostras grandes podemos usar a aproximação pela distribuição normal. Se o valor de W for maior que o valor de t ou Z (dependendo do tamanho da amostra), rejeitamos a hipótese nula e dizemos que há evidências de que o coeficiente de regressão é diferente de zero, ou seja, que a covariável é significante. Este teste é usado para verificar a significância de cada covariável isoladamente.

Para ilustrar este modelo usaremos o exemplo a seguir.

EXEMPLO 1: Este exemplo trata de um estudo para avaliar os fatores associados com o conhecimento, atitude e comportamento das mulheres com respeito a mamografia. Os dados apresentados por Hosmer & Lemeshow (1989) referem-se a uma amostra de 412 mulheres. Contudo, no Apêndice 6 do livro só se encontram os dados relativos a 375 mulheres. Dessa maneira, o modelo será ajustado aos dados disponíveis. Salienta-se que os dados apresentados para a covariável DETC estão trocados. Usaremos que o código 1 apresentado no apêndice 6 equivale ao 3 e vice-versa. Na tabela abaixo são definidas as variáveis coletadas bem como as categorias consideradas com seus respectivos códigos.

TABELA 2.1 - Variáveis utilizadas no estudo sobre Mamografia

Variável	Código	Categorias
Número do sujeito.	ID	1-375
Experiência de mamografia.	ME	0-Nunca 1-A mais de um ano 2-No último ano
"Você não precisa mamografia a menos que desenvolva o sintoma".	SYMPT	1-Concordo fortemente 2-Concordo 3-Discordo 4-Discordo fortemente
Benefícios observados da mamografia.	PB	5-20
Mãe/Irmã com história de câncer na mama.	HIST	0-Não 1-Sim
"Alguém lhe ensinou como examinar suas mamas?"	BSE	0=Não 1=Sim
O quanto e provável que a mamografia detecte um caso de câncer de mama	DETC	1-Não é provável 2-Parcialmente provável 3-Muito Provável

FONTE: Hosmer e Lemeschow (1989, p.279).

Para o ajuste dos modelos de regressão logística politômica a serem apresentados, utilizaremos o pacote estatístico STATA. A variável resposta é Experiência de mamografia, codificada como ME. As outras variáveis serão consideradas explicativas. Os dois primeiros ajustes a serem apresentados neste capítulo servirão apenas para ilustrar os casos especificados. Portanto, não nos preocuparemos em verificar a significância do ajuste do modelo e sua interpretação. Faremos uma análise mais detalhada no caso com mais de uma covariável simultâneamente.

2.2.2. RESPOSTA NOMINAL COM TRÊS CATEGORIAS E UMA COVARIÁVEL CATEGÓRICA BINÁRIA

Neste caso, o ajuste do modelo permite a estimação de dois coeficientes β_{10} e β_{20} e de dois coeficientes de regressão β_{11} e β_{21}, um para cada função logit. Cada um dos coeficientes de regressão estimados é igual ao logaritmo do odds ratio (denotado por ψ_j) dado pela respectiva tabela cruzada. Assim, podemos estimar o odds ratio da comparação da variável resposta (categoria de referência com uma das outras categorias) com a covariável binária calculando a exponencial do coeficiente de regressão $\hat{\beta}_{ij}$ estimado nesta comparação. A estimação do odds ratio da comparação de duas categorias quaisquer da variável resposta (exceto a de referência) é dado pela exponencial da diferença dos dois coeficientes $\hat{\beta}_{ij}$ estimados.

Genericamente, o odds ratio do cruzamento da variável resposta (categoria j e a categoria 0 de referência) pela covariável (categoria a e categoria b) é dado por

$$\psi_{jo}(a,b) = \frac{\dfrac{P(Y=j \mid x=a)}{P(Y=0 \mid x=a)}}{\dfrac{P(Y=j \mid x=b)}{P(Y=0 \mid x=b)}}.$$

Para ilustrar este modelo usaremos, no caso do Exemplo 1, o cruzamento da variável resposta ME com a covariável HIST, que é mostrado na tabela abaixo.

TABELA 2.2 - DADOS REFERENTES AO CRUZAMENTO DA VAR. ME POR HIST

HIST / ME	0	1	Total	ψ_{jo}
0	200	13	213	1,00
1	80	18	98	3,46
2	55	9	64	2,52
Total	335	40	375	

A razão de chances (odds ratio, em inglês) para a comparação da covariável HIST com as categorias 0 e 1 da variável resposta ME, denotada por (ψ_1) é definida por

$$\psi_{10}(1,0) = \psi_1 = \frac{\dfrac{P(Y=1\,|\,x=1)}{P(Y=0\,|\,x=1)}}{\dfrac{P(Y=1\,|\,x=0)}{P(Y=0\,|\,x=0)}} = \frac{18 \times 200}{80 \times 13} = 3{,}46.$$

Analogamente, o odds ratio da comparação da covariável HIST com as categorias 0 e 2 da variável resposta ME, denotada por (ψ_2) é definida por

$$\psi_{20}(1,0) = \psi_2 = \frac{\dfrac{P(Y=2\,|\,x=1)}{P(Y=0\,|\,x=1)}}{\dfrac{P(Y=2\,|\,x=0)}{P(Y=0\,|\,x=0)}} = \frac{9 \times 200}{55 \times 13} = 2{,}52.$$

A tabela abaixo mostra o ajuste do modelo de regressão logística politômica para esta situação.

TABELA 2.9 - RESULTADO DO AJUSTE DO MODELO DE REGR. LOGÍSTICA P/ OS DADOS DA TAB. 2.2

Logit	ME	β_{ij}	SE	Wald
1				
	hist	1,2417	0,3873	3,206
	cons	-.9163	0,1323	-6,927
2				
	hist	.9233	0,4596	2,009
	cons	-1.2910	0,1523	-8,479

Categoria de referência: ME=0

O logit 1 representa o modelo

$$\ln \left[\frac{P(ME=1\,|\,HIST)}{P(ME=0\,|\,HIST)}\right] = -0{,}92 + 1{,}24\ HIST,$$

enquanto o logit 2 representa o modelo

$$\ln \left[\frac{P(ME=2\,|\,HIST)}{P(ME=0\,|\,HIST)}\right] = -1{,}29 + 0{,}92\ HIST.$$

Com os resultados da tabela 2.3 podemos mostrar a relação dos coeficientes β_{ij} com os odds ratio ψ_{ij} descrita anteriormente. O Quadro abaixo ilustra esta relação.

j	β_{1j}	$\exp \beta_{1j} = \hat{\psi}_j$	ψ_j
1	1,241713	3,46	3,46
2	0,923259	2,52	2,52

Para estimar o odds ratio da comparação da covariável HIST com as categorias 1 e 2 da variável resposta ME, basta usar a relação descrita anteriormente, isto é,

$$\psi_{12} = \exp\,(\beta_{12} - \beta_{11}) = (0{,}923259 - 1{,}241713) = 0{,}73.$$

Com os resultados já obtidos podemos, ainda, calcular um intervalo de confiança para cada $\hat{\beta}_{ij}$, dado por $\hat{\beta}_{ij}$ ± Z SE, onde Z é o valor associado da tabela da distribuição normal padrão e SE é a estimativa do erro padrão de $\hat{\beta}_{ij}$. Assim, o intervalo com 95% de confiança para β_{12} é dado por

$$1{,}241713 \pm 1{,}96\ (0{,}387271) = (0{,}4827;2{,}0008),$$

e o intervalo com 95% de confiança para β_{12} é dado por

$$0{,}923259 \pm 1{,}96\ (0{,}4595821) = (0{,}0225;1{,}8240).$$

Para obter o intervalo de confiança do odds ratio, basta calcular a exponencial dos extremos dos intervalos acima, ou seja, exp(0,4827;2,0008)=(1,62;7,39) é o intervalo com 95% de confiança para ψ_1 e exp(0,0225;1,8240)=(1,02;6,19) é o intervalo com 95% de confiança para ψ_2.

2.2.3. RESPOSTA NOMINAL COM TRÊS CATEGORIAS E UMA COVARIÁVEL COM TRÊS CATEGORIAS

No caso onde a covariável também é politômica, uma das categorias deve ser considerada como de referência. Por exemplo, seja a covariável x_1 com três categorias denotadas por 1,2 e 3. Se considerarmos como categoria de referência a categoria

1, teremos duas comparações (categoria 1 versus categoria 2 e categoria 1 versus categoria 3) para cada uma das funções logit do modelo.

Para realizarmos estas comparações é necessário criar variáveis auxiliares, chamadas de variáveis de delineamento. As variáveis de delineamento são variáveis binárias, sendo necessárias tantas quanto o número de categorias da covariável. No caso onde a covariável possui três categorias criamos três variáveis de delineamento. A primeira assumindo valor 1 quando o indivíduo assumiu categoria 1 da covariável e 0 caso contrário. A segunda assumindo valor 1 quando o indivíduo possui o atributo referente a categoria 2 e 0 caso contrário. Por fim, a terceira assume o valor 1 quando o indivíduo apresenta o atributo relativo a terceira categoria e 0 caso contrário. Uma delas é considerada como de referência e as demais como covariáveis.

Genericamente, estimamos $(K-1)\prod_{i=1}^{P}(s_i-1)$ coeficientes $\beta_j(r)$, $j=1,2,\ldots,K-1;$ $r=1,2,\ldots s_i$ e $i=1,\ldots,P$, onde K é o número de categorias da variável resposta e s_i é o número de categorias da i-ésima covariável.

A estimação dos parâmetros do modelo é realizada de forma análoga ao caso com uma covariável binária. Também são válidas as mesmas relações entre os coeficientes de regressão e os odds ratio $\psi_j(a,b)$, bem como os intervalos de confiança para $\hat{\beta}_j(r)$ e para os $\psi_j(a,b)$.

Para ilustrar o ajuste do modelo nesta situação, considere, para os dados do Exemplo 1, o cruzamento da variável resposta ME com a covariável DETC, exibido na tabela abaixo.

Tabela 2.4 - Dados do cruzamento da variavel ME por DETC

ME \ DETC	1	2	3	Total
0	13	69	131	213
1	1	12	85	98
2	4	14	46	64
Total	18	95	262	375

Consideramos como categoria de referência da covariável DETC a categoria 1. Como existem $k=3$ categorias da variável resposta e $s_i=3$ categorias da covariável teremos $(3-1)(3-1) = 4$ comparações a serem feitas. Calculando o odds ratio para cada uma dessas comparações obtemos:

$$\psi_{1,0}(2,1) = \frac{12 \times 13}{69 \times 1} = 2{,}26$$

$$\psi_{1,0}(3,1) = \frac{85 \times 13}{131 \times 1} = 8{,}43$$

$$\psi_{2,0}(2,1) = \frac{14 \times 13}{69 \times 4} = 0{,}66$$

$$\psi_{2,0}(3,1) = \frac{46 \times 13}{131 \times 4} = 1{,}14 .$$

Para simplificar a notação usaremos $\psi_{1,0}(2,1) = \psi_{j0}(r,1) = \psi_j(r) = \psi_{12}$ para representar o odds ratio da comparação da categoria j e a categoria de referência da variável resposta com a categoria r e a categoria de referência da variável explicativa DETC.

A tabela abaixo mostra o ajuste do modelo de regressão logítica politômica para esta comparação.

TABELA 2.5 - RESULTADO DO AJUSTE DO MODELO DE REGR. LOGÍSTICA P/ OS DADOS DA TAB. 2.4

Logit	ME	β_{ij}	SE	Wald
1				
	detc2	.8157	1.0838	.753
	detc3	2.1324	1.0470	2.037
	cons	-2.5649	1.0377	-2.472
2				
	detc2	-.4164	.6425	-.648
	detc3	.1321	.5969	.221
	cons	-1.1786	.5718	-2.061

Categoria de referência: ME=0

O logit 1 representa o modelo

$$\ln\left[\frac{P(ME=1\,|\,DETC)}{P(ME=0\,|\,DETC)}\right] = -2{,}5649 + 0{,}8157\ DETC2 + 2{,}1324\ DETC3,$$

enquanto o logit 2 representa o modelo

$$\ln \left[\frac{P(ME=1 \mid DETC)}{P(ME=0 \mid DETC)} \right] = -2{,}5649 + 0{,}8157\ DETC2 + 2{,}1324\ DETC3.$$

Observe que, como a covariável DETC possui mais de duas categorias foi necessário criar variáveis de delineamento. Sendo assim, por exemplo, DETC1 assume os valores 1 para os indivíduos que apresentam o atributo associado a categoria 1 desta covariável e 0 para os indivíduos que não possuem o atributo da categoria 1. As variáveis de delineamento DETC2 e DETC3 são construídas de maneira análoga.

Comparando os resultados da tabela 2.5 podemos mostrar a relação dos coeficientes de regressão $\beta_j(r)$ com os valores observados $\psi_j(r)$. O quadro abaixo ilustra esta relação.

j	r	$\hat{\beta}_j(r) = \psi_j(r)$	exp $\beta_j(r)$	$\psi_j(r)$
1	2	0,815749	2,26	2,26
	3	2,132403	8,43	8,43
2	2	-0,416394	0,66	0,66
	3	0,132099	1,14	1,14

Para estimar os odds ratio das comparações não apresentadas no ajuste basta utilizarmos a relação descrita no caso em que a covariável é dicotômica, como segue:

$$\psi_{0,1}(2,3) = \exp[\beta_{01}(1,2) - \beta_{01}(1,3)] = \exp(0,8157 - 2,1324) = 0,27$$

$$\psi_{0,2}(2,3) = \exp[\beta_{02}(1,2) - \beta_{02}(1,3)] = \exp(-0,4164 - 0,1321) = 0,58$$

$$\psi_{1,2}(1,3) = \exp[\beta_{02}(1,3) - \beta_{01}(1,3)] = \exp(0,1321 - 2,1324) = 0,13$$

$$\psi_{1,2}(1,2) = \exp[\beta_{02}(1,2) - \beta_{01}(1,2)] = \exp(-0,4164 - 0,8157) = 3,43$$

$$\psi_{1,2}(2,3) = \exp[\beta_{02}(2,3) - \beta_{01}(2,3)] = \exp(-0,5485 + 1,3166 = 2,16.$$

Cabe observar que $\psi_{jj'}(a,b)$ representa o odds ratio da comparação das categorias j e j' da variável resposta com as categorias a e b da covariável.

2.2.4. Regressão logística politômica com mais de uma covariável

Para o caso de um modelo com mais de uma variável explicativa, as estratégias para seleção das covariáveis utilizadas para ajustar o modelo e para avaliar sua adequacidade são análogas aquelas descritas na regressão logística tradicional, veja Hosmer & Lemeshow (1989, p.226). Para ilustrar este caso usaremos os dados do Exemplo 1.

Inicialmente, por considerar que todas as covariáveis citadas causam algum impacto na resposta, iremos ajustar um modelo com todas elas. A tabela abaixo mostra os resultados desse ajuste.

TABELA 2.6 - RESULTADO DO AJUSTE DO MODELO DE REGR. LOGÍSTICA P/ OS DADOS DO EXEMPLO 1

Logit	ME	β_{ij}	SE	Wald
1				
	detc2	.1439	1.1598	.124
	detc3	.9534	1.1267	.846
	bse	1.2084	.5352	2.258
	hist	1.3702	.4502	3.044
	pb	-.2232	.0790	-2.826
	sympt2	.1569	.9324	.168
	sympt3	1.8380	.7894	2.328
	sympt4	2.3733	.7849	3.024
	cons	-2.8580	1.5450	-1.850
2				
	detc2	-.9274	.7152	-1.297
	detc3	-.7065	.6883	-1.026
	bse	.8933	.5201	1.718
	hist	.9395	.4873	1.928
	pb	-.1584	.0817	-1.939
	sympt2	-.0407	.6889	-0.059
	sympt3	.9223	.5958	1.548
	sympt4	1.0928	.6058	1.804
	cons	-.8582	1.1444	-.750

Categoria de referência: ME=0

Com os dados acima, podemos, através do teste da razão de verossimilhança, verificar se os coeficientes de regressão estimados $\hat{\beta}_{ij}$ são conjuntamente diferentes de zero. Observa-se que a estatística G=96,69 comparada com o valor da estatística χ^2 de Pearson com 16 gl atingiu um nível de significância inferior a 10^{-6}, o que conduz a rejeição da hipótese nula. Isso sugere que pelo menos um dos coeficientes de regressão β_{ij} é estatisticamente diferente de zero. A estatística de Wald pode ser usada para testar se individualmente cada um dos β_{ij} é diferente de zero. Assim, constatamos que há evidências de que os coeficientes de regressão associados às covariáveis SYMPT2, DETC2 e DETC3 da função logit 1 e da função logit 2 são estatisticamente iguais a zero. No entanto, os coeficientes associados \`as covariáveis SYMPT3 e SYMPT4 da função logit 1, são diferentes de zero ao nível de significância de 2%. Isso nos leva a recomendar uma recategorização dessa covariável. O sinal e a magnitude dos coeficientes estimados associados \`as categorias da covariável SYMPT (as covariáveis de delineamento) sugerem que as diferentes intensidades de concordância e discordância aparentemente provocam o mesmo tipo de impacto na variável resposta. Diante disso, a recategorização aconselhável é juntar SYMPT1 com SYMPT2 e SYMPT3 com SYMPT4. A nova covariável SYMPT, chamada de SYMPD, assume o valor 0 para indicar concordo (englobando concordo fortemente e concordo) e 1 para indicar discordo (englobando discordo fortemente e discordo). A tabela abaixo mostra os resultados do ajuste com a covariável SYMPT recodificada e as demais covariáveis inalteradas.

TABELA 2.7 - RESULTADO DO AJUSTE DO MODELO DE REGR. LOGISTICA P/ OS DADOS DO EXEMPLO 1 COM A COVARIAVEL SYMPT RECATAGORIZADA

Logit	ME	β_{ij}	SE	Wald
1				
	detc2	.2080	1.1586	.180
	detc3	1.0232	1.1253	.909
	bse	1.1629	.5317	2.187
	hist	1.2940	.4474	2.893
	pb	-.2560	.0757	-3.379
	sympd	1.9818	.4614	4.295
	cons	-2.5254	1.4439	-1.749
2				
	detc2	-.9159	.7152	-1.281
	detc3	-.6913	.6876	-1.005
	bse	.8818	.5193	1.698
	hist	.9258	.4838	1.914
	pb	-.1677	.0788	-2.129
	sympd	1.0143	.3752	2.704
	cons	-.8109	1.0956	-.740

Categoria de referência: ME=0

O valor observado da estatística G é 93,50, sugerindo que pelo menos um dos coeficientes de regressão é estatisticamente diferente de zero. Através da estatística de Wald verifica-se que a recategorização da covariável sympt foi conveniente, pois SYMPD produz certo impacto na variável resposta. Este ajuste é apresentado por Hosmer e Lemeshow (1989) como um modelo final para descrever as relações entre a realização de mamografia e as covariáveis. No entanto, a estatística de Wald sugere que os coeficientes de regressão associados as covariáveis de delineamento DETC2 e DETC3 não são estatisticamente diferentes de zero. Comparando o segundo ajuste com um modelo sem a covariável DETC, através do teste da razão de verossimilhança, obtemos o valor da estatística de teste G

$$G = -2 \ln \left[\frac{\text{Verossimilhança do modelo sem DETC}}{\text{Verossimilhança do modelo com DETC}} \right]$$

$$G = -2 \; (-322{,}17 + 318{,}39) = 7{,}55.$$

Como $7{,}55 < 9{,}48$ (valor do X^2 com $14-10=4$ gl e $\alpha=5\%$) aceitamos a hipótese de que os coeficientes da covariável DETC são estatisticamente iguais a zero. Com base neste resultado podemos dizer que a permanência da covariável DETC no modelo é questionável. Do ponto de vista estatístico seria aconselhável a sua remoção. No entanto, Hosmer & Lemeshow (1989) argumentam que esta covariável estaria agindo como um fator de confundimento da associação de SYMPD com a variável resposta, veja Hosmer & Lemeshow (1989, p.230). Sendo assim apresentamos este último ajuste como um ajuste final para o modelo em estudo.

No modelo final, o logit 1 representa

$$\ln\left[\frac{P\ (ME=1 \mid SYMPD,\ PB,\ HIST,\ BSE,\ DETC2,\ DETC3\ )}{1\ -\ P(ME=1 \mid SYMPD,\ PB,\ HIST,\ BSE,\ DETC2,\ DETC3\ )}\right]$$

$$= -2{,}52 + 1{,}98\ SYMPD - 0{,}26\ PB + 1{,}29\ HIST + 1{,}16\ BSE$$

$$+ 0{,}21\ DETC2 + 1{,}02\ DETC3.$$

e o logit 2 representa:

$$\ln\left[\frac{P\ (ME=2 \mid SYMPD,\ PB,\ HIST,\ BSE,\ DETC2,\ DETC3\ )}{1\ -\ P(ME=2 \mid SYMPD,\ PB,\ HIST,\ BSE,\ DETC2,\ DETC3\ )}\right]$$

$$= -0.81 + 1{,}01\ SYMPD -0{,}17\ PB + 0{,}93\ HIST + 0{,}88\ BSE$$

$$-0{,}92\ DETC2 -0{,}69\ DETC3.$$

Para uma interpretação quantitativa do modelo podemos usar a relação existente entre os coeficientes de regressão estimados e os respectivos odds ratio. Por exemplo, a comparação da variável resposta com a covariável SYMPD é como segue.

A chance relativa de um indivíduo ter feito mamografia a mais de um ano dado que ele discorda da afirmação "não necessito realizar mamografia a menos que desenvolva o sintoma" é exp(2,1217)=8,34 vezes maior do que a chance relativa do indivíduo nunca ter feito mamografia dado que ele concorda com a afirmação acima, mantendo constante as demais covariáveis consideradas no modelo. A interpretação quantitativa para os outros coeficientes pode ser feita de maneira análoga. A interpretação qualitativa é que as covariáveis consideradas no ajuste final influenciam na experiência de uma mulher com a mamografia.

CAPÍTULO 3 - REGRESSÃO LOGÍSTICA POLITÔMICA ORDINAL

3.1. CONCEITOS BÁSICOS

Regressão logística politômica ordinal é caracterizada pelo fato da variável resposta Y se apresentar de forma categórica ordinal com 3 ou mais categorias. Teoricamente esses modelos de regressão se baseiam na possibilidade de existir uma resposta variável, latente e contínua, usualmente não observável. Em outras palavras, os dados observados através da variável resposta Y são uma categorização dessa variável contínua subjacente. Convém salientar que sua inexistência não invalida tais modelos, mas se ela existe a interpretação dos parâmetros se torna mais clara, veja McCullagh(1980).

Seja Z a variável latente com função de distribuição $F_{\eta}(z) = F(z-\eta)$ onde η é um parâmetro de posição. Seja x o vetor de covariáveis e $\eta(x)=\underset{\sim}{\beta}^t x$. Então a distribuição condicional de $Z|x$ é dada por $F(Z - \underset{\sim}{\beta}^t x)$. Como a variável Z não pode ser observada diretamente, usamos uma variável resposta Y que assume valores $Y=j$ se e somente se $Z \in (\theta_{j-1}, \theta_j)$. As probabilidades acumuladas são definidas por

$$\gamma_j(x_i) = P[\, Y \leq j \mid x_i \,] = F(\theta_j - \underset{\sim}{\beta}^t x_i).$$

Qualquer distribuição unimodal simétrica pode ser utilizada para a variável latente Z, produzindo resultados similares. Contudo, a suposição da distribuição logística $F(z) = \dfrac{e^z}{1 + e^z}$ tem se mostrado bastante adequada, principalmente pela facilidade de cálculo. Nesta situação o uso da função de ligação $Logit\{\gamma_j(x_i)\}$ conduz ao modelo de <u>odds proporcionais</u> proposto por McCullagh (1980). Por outro lado, na existência de razões que levem a acreditar que a distribuição subjacente é assimétrica, as funções de ligação <u>log-log</u> ou <u>complementar log-log</u> podem ser usadas, veja Hastie, Botha Schitzler (1989). Essas funções de ligação induzem a suposição de distribuições assimétricas, conduzindo a modelos diferentes.

Nesta monografia abordaremos apenas o modelo de odds proporcionais, utilizado para descrever as relações entre uma variável resposta categórica ordenada e um conjunto de covariáveis. Maiores detalhes sobre esta classe de modelos de regressão podem ser obtidos em Agresti (1990), McCullagh & Nelder (1989) e Vigo (1994).

3.2. MODELO DE ODDS PROPORCIONAIS

Consideremos Y a variável resposta com K categorias ordenadas. Seja $x = (x_1, x_2, \dots, x_p)$ o vetor das P covariáveis a serem utilizadas no modelo. Seja, ainda, $x_i = (x_{i1}, \dots, x_{iP})$ o vetor que representa o i-ésimo valor do vetor de covariáveis, para $i=1,2,\dots,n$, onde n o número de combinações possíveis entre os diferentes níveis dessas covariáveis. Podemos chamar cada x_i de uma subpopulação e observar, pela própria definição de variável categórica vista anteriormente, que são conjuntos disjuntos, ou seja, subpopulações independentes entre si. Podemos representar as frequências observadas de cada vetor x_i para as diferentes categorias da variável resposta através da tabela abaixo:

TABELA 3.1 - FREQUENCIAS DE RESPOSTAS NAS CATEGORIAS DA VARIAVEL Y, PARA OS DISTINTOS VALORES DO VETOR X

Valor de X	Categoria de Resposta			
	1	2	. . .	K
x_1	y_{11}	y_{12}		y_{1K}
x_2	y_{21}	y_{22}		y_{2K}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
x_n	y_{n2}	y_{n2}		y_{nK}

Para cada valor y_{ij}, $i=1,\dots,n$ e $j=1,\dots,K$, podemos associar um valor para $\pi_j(x_i) = P(Y=j \mid x_i)$ que é a probabilidade de se observar uma resposta na categoria j quando o vetor de covariáveis assume o valor x_i. Por simplicidade de notação usaremos que $\pi_j(x_i)=\pi_{ij}$. Portanto, as probabilidades acumuladas são dadas por

$$\gamma_{ij} = \gamma_j(x_i) = P[\,Y \leq j \mid x_i\,] = \pi_{i1} + \pi_{i2} + \ldots + \pi_{ij}.$$

Supondo que a distribuição da variável contínua subjacente é a logística (que é dada por $F(z) = \dfrac{e^z}{1+e^z}$), temos que a probabilidade de observar uma resposta em uma categoria menor ou igual a j é dada por

$$\gamma_{ij} = P[\,Y \leq j \mid x_i\,] = F(\theta_j - \underset{\sim}{\beta}^t x_i) = \frac{\exp\left\{\theta_j - \underset{\sim}{\beta}^t x_i\right\}}{1 + \exp\left\{\theta_j - \underset{\sim}{\beta}^t x_i\right\}}$$

(expressão 3.1)

para todo $j=1,\ldots,K-1$, onde os θ_j são os pontos de corte desconhecidos da variável Z, que satisfazem $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_{K-1}$; $\theta_0 \equiv -\infty$ e $\theta_K \equiv +\infty$. $\underset{\sim}{\beta}$ é o vetor de parâmetros que iremos estimar com o ajuste do modelo e representam os coeficientes de regressão associados às covariáveis.

A transformação logit do modelo especifica que

$$\text{logit } \gamma_j(x_i) = \text{logit } \gamma_{ij} = \ln\left[\frac{\gamma_{ij}}{1-\gamma_{ij}}\right].$$

Utilizando a expressão 3.1 para substituir γ_{ij} na fórmula acima, temos que

$$\operatorname{logit} \gamma_{ij} = \ln \left[\frac{\dfrac{\exp(\theta_j - \underset{\sim}{\beta}^t x_i)}{1 + \exp(\theta_j - \underset{\sim}{\beta}^t x_i)}}{1 - \dfrac{\exp(\theta_j - \underset{\sim}{\beta}^t x_i)}{1 + \exp(\theta_j - \beta^t x_i)}} \right]$$

$$= \theta_j - \underset{\sim}{\beta}^t x_i.$$

Portanto, o modelo de odds proporcionais especifica que logit $\gamma_{ij} = \theta_j - \underset{\sim}{\beta}^t x_i$. O ajuste deste modelo compreende a estimação dos vetores dos parâmetros desconhecidos $\underset{\sim}{\theta} = [\theta_1, \theta_2, \ldots, \theta_{K-1}]^t$ e $\underset{\sim}{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^t$, os quais podem ser representados pelo vetor $\underset{\sim}{\beta}^* = [\theta_1, \ldots, \theta_{K-1}, \beta_1, \ldots, \beta_p]^t$. Para tanto, é necessário determinar a função de verossimilhança do modelo.

Voltando a tabela 3.1, o vetor $y_i = (y_{i1}, y_{i2}, \ldots, y_{iK})^t$, para $i=1,2,\ldots,n$, contém as frequências observadas da i-ésima subpopulação e $\underset{\sim}{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK})^t$ é o vetor de probabilidades associado a essas frequências. Supondo que cada vetor y_i segue uma distribuição multinomial, ou seja,

$$P\left[Y_{i1}=y_{i1},\dots,Y_{iK}=y_{iK} \right] = \frac{n_i!}{y_{i1}!\, y_{i2}! \dots y_{iK}!} \prod_{j=1}^{K} \left[\pi_{ij}^{\,y_{ij}} \right]$$

com as restrições $\sum_{j=1}^{K} y_{ij}=n_i$ e $\sum_{j=1}^{K} \pi_{ij}=1$ e como os vetores y_i são independentes, então a função de verossimilhança é dada por

$$l(\underset{\sim}{\pi};Y) = \prod_{i=1}^{n} \frac{n_i!}{y_{i1}!\, y_{i2}! \dots y_{iK}!} \prod_{j=1}^{K} \left[\pi_{ij}^{\,y_{ij}} \right].$$

Para a estimação dos parâmetros desconhecidos, devemos derivar a função de verossimilhança em relação ao vetor β^*. Para facilitar os cálculos fazemos o logaritmo da função de verossimilhança dado por

$$\ln l(\underset{\sim}{\pi};Y) = \ln \left[\prod_{i=1}^{n} \frac{n_i!}{y_{i1}!\, y_{i2}! \dots y_{iK}!} \right] + \ln \left[\prod_{j=1}^{K} \left[\pi_{ij}^{\,y_{ij}} \right] \right].$$

Dado que $\ln \left[\prod_{i=1}^{n} \frac{n_i!}{y_{i1}!\, y_{i2}! \dots y_{iK}!} \right]$ é uma constante e que a derivada de uma constante é zero, podemos representar o logaritmo da função de verossimilhança por

$$L(\underset{\sim}{\pi};y) = \sum_{j=1}^{K} y_{ij} \ln[\pi_{ij}] = y_i^t \ln[\underset{\sim}{\pi}_i].$$

Calculando a primeira derivada parcial dessa função em relação ao vetor β^* e igualando a zero obtemos as equações de verossimilhança. Resolvendo essas equações obtemos os estimadores de máxima verossimilhança dos parâmetros θ_j e β_i. Como as equações de verossimilhança são funções não lineares dos parâmetros $\underset{\sim}{\theta}$ e $\underset{\sim}{\beta}$ o cálculo desses estimadores é feito através de métodos iterativos. Neste trabalho utilizaremos o método iterativo de Newton-Rapson que é o utilizado no pacote estatístico STATA. No procedimento LOGISTIC do pacote estatístico SAS, o método iterativo utilizado é o de mínimos quadrados iterativamente reponderados (Iterative Reweighted Least Square - IRLS, em inglês), veja SAS Institute Inc. (1989, p.1071).

Analogamente ao modelo de regressão logística politômica nominal calculamos a segunda derivada parcial do logaritmo da função de verossimilhança para obter a matriz de informação $I(\underset{\sim}{\beta})$ de dimensão $(K-1+P)$ por $(K-1+P)$. A inversa da matriz de informação, denotada por $\sum(I)$, é a matriz assintótica de covariâncias dos $(K-1+P)$ parâmetros desconhecidos. O estimador dessa matriz pode ser expresso por

$$\hat{\Sigma}(I) = \left[\hat{I}(\hat{\beta})\right]^{-1}.$$

3.3. AVALIANDO A SIGNIFICÂNCIA DO MODELO

Assim como no modelo de regressão logística politômica nominal uma etapa importante na modelagem é a verificação do ajuste do modelo bem como a avaliação da significância das covariáveis. Os testes e estatísticas utilizados são os mesmos descritos na seção 2.2.1. Para o caso do modelo de odds proporcionais cabe, ainda, salientar a estatística utilizada para testar a hipótese nula de que a suposição de linhas paralelas é adequada. Este teste é baseado em escores e a estatística de teste utilizada tem uma distribuição assintótica de χ^2 de Pearson com $p\ (k - 2)$ graus de liberdade. É impotante aceitarmos a suposição de linhas paralelas, pois só assim é que poderemos realizar as interpretações descritas neste capítulo.

Para ilustrar o modelo, utilizaremos o exemplo seguinte.

EXEMPLO 2: Este exemplo trata de um estudo para avaliar os fatores associados ao baixo peso ao nascer. Os dados referem-se a uma amostra de 5.939 nascimentos ocorridos na cidade de Pelotas(RS). Sabe-se que a variável peso ao nascer é uma variável contínua. Tradicionalmente os estudos sobre baixo peso ao nascer utiliza a variável peso ao nascer como dicotômica, ou seja, categorizando-a como baixo peso ou não. Porém, clinicamente, sabe-se que a delimitação para o corte de baixo peso é questionável e que alto peso também pode ser considerado como fator de risco. Sendo assim, neste estudo, trataremos esta variável com 5 categorias para especificar melhor as relações entre os fatores associados ao baixo peso ao nascer. Observa-se, ainda, que esta variável deve ser tratada como uma variável categórica ordinal. Na tabela abaixo são apresentadas as variáveis investigadas com suas respectivas categorias e códigos.

TABELA 3.2 - VARIÁVEIS UTILIZADAS NO ESTUDO SOBRE BAIXO PESO AO NASCER.

Variável	Código	Categ
Idade da mãe	AGE	13-46
Renda familiar	INCOME	0: 1 sm ou menos 1: 1,1 - 3 sm 2: 3,1 - 6 sm 3: 6,1 - 10 sm 4: + 10 sm
Nível de educação da mãe	EDUCAT	0-22
Raça da mãe	RACE	0: Negra 1: Branca
Hábito de fumar da mãe durante a gravidez	SMOKING	0: Não fumou 1: Pouco em parte 2: Pouco em toda 3: Muito em parte 4: Muito em toda
Número de idas ao médico durante a gravidez	ANTENAT	0-27
Condição de moradia dos pais	COHABIT	0: Moram separados 1: Moram juntos
Tipo de parto	TYPDELIV	0: Cesária 1: Vaginal 2: Vag. Forceps 3: Vag. Vacuum 4: Vag. Induced 5: Vag. Ind. Forc. 6: Vag. Analgesia 7: Vag.Analg.Forc.
Peso ao nascer	BWGR	0: Menos de 2000g 1: 2000 - 2499g 2: 2500 - 3000g 3: 3000 - 3500g 4: 3500g ou mais

Para o ajuste dos modelos de odds proporcionais a serem apresentados neste capítulo utilizaremos o pacote estatístico STATA. A variável resposta é a BWGR (peso ao nascer). As demais variáveis serão consideradas como covariáveis.

3.4. RESPOSTA ORDINAL COM CINCO CATEGORIAS E UMA COVARIÁVEL CATEGÓRICA BINÁRIA

Neste caso o ajuste do modelo permite a estimação de um coeficiente de regressão β_1 e quatro pontos de corte $\theta_1 \leq \theta_2 \leq \theta_3 \leq \theta_4$. Estaremos modelando a seguinte função:

$$\ln \frac{\gamma_{ij}}{1-\gamma_{ij}} = \theta_j - \beta_1 x_i \quad , \text{ para } j=1,2,3,4; \text{ e } x_1=0 \text{ e } x_2=1.$$

Observa-se que a função $\ln \frac{\gamma_{ij}}{1-\gamma_{ij}}$ é linear nos parâmetros θ_j e β_1, portanto, com o ajuste do modelo estamos ajustando 4 retas que são paralelas pelo fato de que o coeficiente de regressão β_1 é o mesmo, quando o modelo de odds proporcionais é adequado.

Para ilustrar o ajuste do modelo nesta situação, considere, para os dados do Exemplo 2 o cruzamento da variável resposta BWGR com a covariável RACE, exibido na tabela abaixo:

TABELA 3.3 - DADOS REFERENTES AO CRUZAMENTO DA VAR. BWGR POR RACE.

BWGR \ RACE	0	1	Total
0	50	154	204
1	78	297	375
2	293	1099	1392
3	373	1834	2207
4	277	1484	1761
Total	1071	4868	5939

A tabela abaixo mostra o ajuste do modelo de odds proporcionais.

TABELA 3.4 - RESULTADO DO AJUSTE DO MODELO DE ODDS PROPORCIONAIS PARA OD DADOS DA TAB. 3.3.

bwgr	β_{ij}	SE	Wald
race	0.2845	0.0615	4.628
θ_1	-3.1087	0.0863	
θ_2	-1.9970	0.0656	
θ_3	-.4679	0.0570	
θ_4	1.1002	0.0586	

O modelo ajustado é

$$\ln\left[\frac{P[BWGR \leq j \mid RACE]}{1 - P[BWGR \leq j \mid RACE]}\right] = \theta_j - 0{,}2845\ RACE.$$

Usando a expressão 3.1 podemos calcular os γ_{ij}. Por exemplo:

$$\gamma_{12} = P[BWGR \leq 2 \mid RACE = 1]$$

$$= \frac{\exp[\theta_2 - \beta_1\ RACE]}{1 + \exp[\theta_2 - \beta_1\ RACE]} = \frac{\exp[-1{,}997 - 0{,}2845(1)]}{1 + \exp[-1{,}997 - 0{,}2845(1)]} = 0{,}09\,.$$

Para calcularmos a probabilidade de observar uma resposta na categoria j, usamos a seguinte relação:

$$P\left[Y = j \mid x_i\right] = P\left[Y \leq j \mid x_i\right] - P\left[Y \leq j-1 \mid x_i\right]$$

$$= \gamma_{ij} - \gamma_{i(j-1)}\,.$$

Por exemplo:

$$P\left[BWGR = 2 \mid RACE = 1\right]$$

$$= P\left[BWGR \leq 2 \mid RACE = 1\right] - P\left[BWGR \leq 1 \mid RACE = 1\right]$$

$$= \gamma_{12} - \gamma_{11} = 0{,}09 - 0{,}03 = 0{,}07\,.$$

A tabela abaixo mostra os valores das probabilidades observadas e das probabilidades preditas pelo modelo de odds proporcionais.

TABELA 3.5 - PROBABILIDADES OBSERVADAS E PREDITAS PELO MODELO DE ODDS PROPORCIONAIS P/ AS CATEG. DE RESPOSTAS DO PESO AO NASCER

RACE / BWGR	Observadas		Preditas	
	0		0	1
0	0,05	0,03	0,04	0,03
1	0,07	0,06	0,08	0,06
2	0,27	0,23	0,26	0,23
3	0,35	0,38	0,37	0,37
4	0,26	0,30	0,25	0,31

Analisando a tabela acima verificamos que os valores observados de cada casela são similares aos valores preditos pelo ajuste do modelo. Isto indica que o modelo de odds proporcionais parece ser adequado para descrever as relações entre raça e baixo peso ao nascer. Cabe salientar que, a primeira vista, a covariável RACE pouco deveria influenciar no baixo peso. Contudo, se analisarmos o fato de que a raça está intimamente ligada com fatores sócio-econômicos e estes sim parecem influenciar o baixo peso, veremos que tem certa lógica a covariável RACE ser significante para explicar o baixo peso.

Uma interpretação quantitativa para o modelo de odds proporcionais pode ser descrita por um quociente de chances dado por

$$\frac{\dfrac{\gamma_{ij}}{1-\gamma_{ij}}}{\dfrac{\gamma_{i'j}}{1-\gamma_{i'j}}} = \frac{\exp\,[\theta_j - \underset{\sim}{\beta}^t\, x_i]}{\exp\,[\theta_j - \underset{\sim}{\beta}^t\, x_{i'}]} = \exp\,[\underset{\sim}{\beta}^t (x_{i'} - x_i)]$$

(expressão 3.2)

para $i \neq i'$. Observe que este quociente independe de θ_j.

Para o modelo que estamos discutindo este quociente é dado por:

$$\frac{\dfrac{\gamma_{01}}{1-\gamma_{01}}}{\dfrac{\gamma_{11}}{1-\gamma_{11}}} = \exp\,[\ \beta_1 (x_0 - x_1)\ ] = \exp\,[0{,}2844\ (0-1)] = 0{,}75\ .$$

Isso significa que a chance relativa de um recém-nascido pesar menos de 2000g ($j=1$) para um indivíduo da raça branca ($x_2=1$) é 0,75 da chance de um indivíduo da raça negra ($x_1=0$). Da mesma maneira, a chance relativa de um recém-nascido pesar menos de 2500g ($j=2$) para um indivíduo da raça branca é 0,75 da chance de um indivíduo da raça negra. E assim para todas as outras categorias j da variável resposta.

Observe ainda, que no cálculo da razão de chances, consideramos o valor atribuído para cada uma das categorias da covariável. Neste caso, onde a covariável é binária, isto não é problema. Dependendo dos valores dados para as categorias das covariáveis teremos um valor para os parâmetros estimados $\beta_1, \theta_1, \theta_2, \theta_3, \theta_4$, mas o valor para a distribuição de probabilidades bem como para a razão de chances permanecem constantes. Isto indica que o ajuste do modelo de odds proporcionais quando as covariáveis são binárias é independente do valor atribuído para cada uma das categorias dessas covariáveis.

A Figura 3.1 mostra uma interpretação visual para o ajuste do modelo de odds proporcionais. O valor em retângulo é a razão de chances calculada anteriormente.

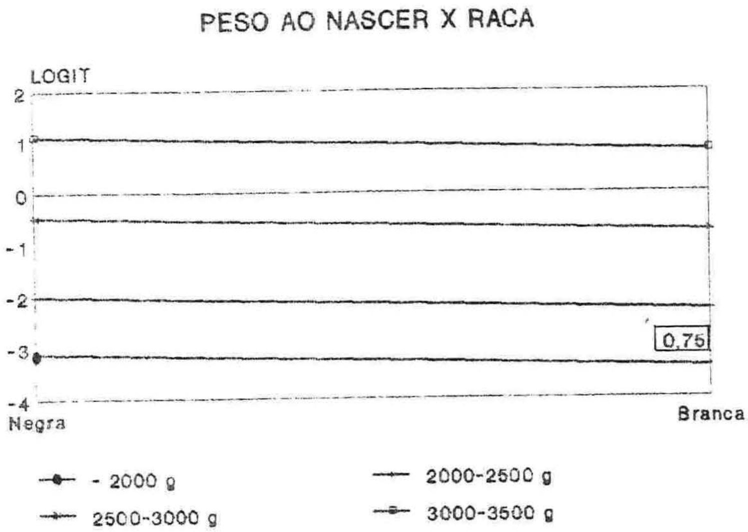

Figura 3.1 - Gráfico do ajuste do modelo de odds proporcionais para as variáveis peso ao nascer versus raça.

3.5. RESPOSTA ORDINAL COM CINCO CATEGORIAS E UMA COVARIÁVEL COM CINCO CATEGORIAS

No caso onde a covariável também é politômica, o ajuste do modelo é analogo ao caso onde a covariável é binária, ou seja,ele permite a estimação de um coeficiente de regressão β_1 e quatro pontos de corte $\theta_1 \leq \theta_2 \leq \theta_3 \leq \theta_4$. Estaremos modelando a seguinte função

$$\ln \frac{\gamma_{ij}}{1 - \gamma_{ij}} = \theta_j - \beta_1 x_i \quad , \text{ para } j=1,2,3,4; \text{ e } i=0,1,2,3,4.$$

Observa-se que, apesar da covariável apresentar mais de duas categorias, continuamos ajustando quatro retas paralelas. Para ilustrar o ajuste do modelo nesta situação, considere, para os dados do exemplo 2, o cruzamento da variável resposta BWGR com a covariável SMOKING, exibido na tabela abaixo.

TABELA 3.6 - DADOS REFERENTES AO CRUZAMENTO DA VARIAVEL BWGR POR SMOKING

SMOKING / BWGR	0	1	2	3	4	Total
0	110	17	43	4	30	204
1	199	26	97	3	50	375
2	762	107	371	8	144	1392
3	1447	177	402	15	166	2207
4	1298	113	259	12	79	1761
Total	3816	440	1172	42	469	5939

A tabela abaixo mostra os resultados do ajuste do modelo de odds proporcionais.

TABELA 3.7 - RESULTADO DO AJUSTE DO MODELO DE ODDS PROPORCIONAIS PARA OD DADOS DA TAB. 3.6

bwgr	β_{ij}	SE	Wald
SMOKING	-.2455	0.0192	-12.814
θ_1	-3.5823	0.0744	
θ_2	-2.4643	0.0483	
θ_3	-.9123	0.0326	
θ_4	.6858	0.0316	

O modelo ajustado especifica que

$$\ln\left[\frac{P\,[BWGR \le j \mid SMOKING]}{1 - P\,[BWGR \le j \mid SMOKING]}\right] = \theta_j + 0{,}2455\ SMOKING\ .$$

Usando a expressão 3.1 podemos calcular os γ_{ij}. Por exemplo

$$\gamma_{23} = P\,[BWGR \le 3 \mid SMOKING = 2] = \frac{\exp\,[\theta_3 - \beta_1\ SMOKING]}{1 + \exp\,[\theta_3 - \beta_1\ SMOKING]}$$

$$= \frac{\exp\,[-0{,}9123 + 0{,}2455(2)]}{1 + \exp\,[-0.9123 + 0{,}2455(2)]} = 0{,}40\ .$$

Para calcularmos a probabilidade de observar uma resposta em uma categoria j usamos a mesma relação descrita no caso anterior. Aplicando a expressão 3.2 temos, por exemplo

$P\,[BWGR = 3 \mid SMOKING = 2] =$

$P\,[BWGR \le 3 \mid SMOKING = 2] - P\,[BWGR \le 2 \mid SMOKING = 2] =$

$\gamma_{23} - \gamma_{22} = 0{,}40 - 0{,}12 = 0{,}28\ .$

A tabela abaixo mostra os valores das probabilidades observadas e das preditas pelo modelo de odds proporcionais.

TABELA 3.8 - PROBABILIDADES OBSERVADAS E PREDITAS PELO MODELO DE ODDS PROPORCIONAIS PARA AS CATEGORIAS DE RESPOSTAS DO PESO AO NASCER

BWGR \ SMOKING	Observados					Preditos				
	0	1	2	3	4	0	1	2	3	4
0	0,03	0,04	0,04	0,09	0,06	0,03	0,03	0,04	0,05	0,07
1	0,05	0,06	0,08	0,07	0,11	0,05	0,06	0,08	0,10	0,11
2	0,20	0,24	0,32	0,19	0,31	0,21	0,25	0,28	0,31	0,34
3	0,38	0,40	0,34	0,36	0,35	0,37	0,38	0,36	0,35	0,32
4	0,34	0,26	0,22	0,29	0,17	0,34	0,28	0,24	0,19	0,16

Analisando a tabela acima observamos que a distribuição de probabilidades preditas apresenta valores similares aos da distribuição de probabilidades observadas o que indica que o modelo de odds proporcionais parece ser adequado para descrever as relações entre o hábito de fumar da mãe e o baixo peso ao nascer.

Para uma interpretação quantitativa usamos o quociente de chances (expressão 3.2) apresentado no caso anterior. Note que neste caso a covariável apresenta 5 categorias e assim podemos calcular $\binom{5}{2} = \frac{5!}{2!\,(5-2)!} = 10$ quocientes de chances. Por simplicidade de análise optamos por apresentar apenas os resultados obtidos quando comparamos a categoria 0 (não fumou) com cada uma das demais categorias da covariável. Estas comparações são apresentadas abaixo

$$\exp[\beta_1 (x_0 - x_1)] = \exp[-0{,}2455\ (0-1)] = 1{,}28$$

$$\exp[\beta_1 (x_0 - x_2)] = \exp[-0{,}2455\ (0-2)] = 1{,}64$$

$$\exp[\beta_1 (x_0 - x_3)] = \exp[-0{,}2455\ (0-3)] = 2{,}08$$

$$\exp[\beta_1 (x_0 - x_4)] = \exp[-0{,}2455\ (0-4)] = 2{,}70 .$$

Estes valores representam, por exemplo, a chance relativa de um recém-nascido pesar menos de 3000g para um indivíduo cuja mãe fumou pouco em parte da gravidez é 1,28 vezes maior do que a chance relativa de um indivíduo cuja mãe não fumou durante a gravidez. As outras interpretações são análogas.

Neste caso, o fato de considerarmos o valor dado as categorias da covariável para o cálculo da razão de chances, nos leva a um problema. Quando atribuímos valores equidistantes para as categorias, parece que não importa a magnitude deles de tal forma que os valores para as razões de chances serão constantes. Agora, se atribuírmos valores não equidistantes, as razões de chances variam conforme os valores dados. Quando a covariável é originalmente contínua, uma possível maneira de contornar este problema é atribuir o valor de uma média ponderada entre os valores originais e suas respectivas frequências, para cada intervalo. Quando a covariável é nominal, a atribuição de valores é mais questionável. Devemos, portanto, usar o bom senso e o conhecimento prévio na área específica do estudo para determinar estes valores. Apesar de considerar não ser o mais conveniente, usamos, neste exemplo, valores equidistantes para as categorias da covariável SMOKING.

A Figura 3.2 mostra uma interpretação visual para o modelo de odds proporcionais. Os valores nos retângulos representam as razões de chances calculadas anteriormente.

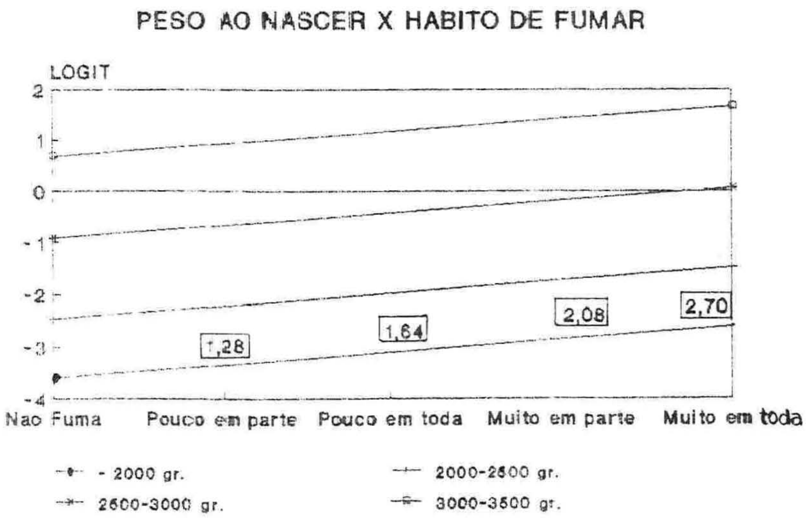

Figura 3.2 - Gráfico do ajuste do modelo de odds proporcionais para as variáveis bwgr e smoking.

3.6 - REGRESSÃO LOGÍSTICA POLITÔMICA ORDINAL COM MAIS DE UMA COVARIÁVEL SIMULTÂNEAMENTE.

Para exemplificar este caso usaremos os dados do Exemplo 2. Primeiramente, é importante fazermos uma análise exploratória dos dados. Observe que uma melhor interpretação para o modelo de odds proporcionais é possível quando as covariáveis em questão não apresentam muitas categorias. Por este motivo resolvemos categorizar as covariáveis AGE e ANTENAT. O processo de categorização é bastante delicado, pois devemos ter o cuidado para não perdermos informação com esta categorização. Uma maneira informal para verificar isso é primeiro ajustar um modelo com os valores originais das covariáveis e em seguida com as covariáveis codificadas. Se os valores para as estimativas de máxima verossimilhança não forem muito distintos, podemos dizer que a categorização parece ter sido adequada. No nosso exemplo resolvemos, por razões clínicas, fazer a seguinte categorização

AGE:	13-15=15	ANTENAT:	1-6=4
	16-15=21		7-11=8
	26-32=29		12-27=13
	33-39=35		
	40-46=41		

O valor dado para cada categoria é a média dos valores originais ponderados pelas frequências observadas.

Verificamos que o primeiro ajuste (veja os resultados no Apêndice) feito com os dados originais e o segundo realizado com as covariáveis AGE e ANTENAT categorizadas e as demais inalteradas, produziram resultados com diferenças mínimas. Sendo assim continuamos a modelagem com as covariáveis categorizadas.

O valor calculado para a estatística G com as oito cavariáveis deste exemplo é 322,76, com um nível de significância associado inferior a 10^{-6}. Isto sugere que pelo menos um dos coeficientes de regressão é estatisticamente diferente de zero. O próximo passo, então, é verificar cada covariável isoladamente.

Através da estatística de Wald verificou-se que as covariáveis EDUCAT e TYPDELIV não são significantes. Portanto, estas duas covariáveis foram eliminadas do modelo.

Um terceiro ajuste foi feito no STATA, agora apenas com as seis covariáveis restantes. A estatística G calculada para este ajuste foi de 318,92 com nível de significância associado inferior a 10^{-6} sugerindo que pelo menos um dos coeficientes de regressão estimados é estatisticamente diferente de zero. Através da estatística de Wald verificou-se que cada covariável isoladamente deve produzir algum impacto significante na variável resposta. Este deveria, portanto, ser apresentado como um ajuste final para o modelo. No entanto, ainda não testamos a suposição de que o modelo é de linhas paralelas. Como o STATA não fornece um teste para esta hipótese, o modelo também foi ajustado através do procedimento LOGISTIC do SAS. Os resultados obtidos para o ajuste foram os mesmos daqueles apresentados pelo STATA. Porém, no teste baseado em escores, utilizado para testar a suposição de linhas paralelas, o valor observado da estatística χ^2 foi de 66,36, que comparada com o valor da estatística χ^2 de Pearson com 15 gl atinge um nível de significância de 0,0001. Isto sugere que a suposição de linhas paralelas seja rejeitada. Isto já era esperado, pois o modelo de odds proporcionais parece não ser adequado quando o ajuste possui muitas covariáveis com um número grande de categorias. A estratégia agora é categorizar algumas covariáveis aglutinando categorias ou eliminar mais alguma covariável do modelo. Neste momento a volta aos dados é imprecindível e a interpretação clínica fundamental. Optamos por retirar mais alguma covariável do modelo. Salientamos que outras opções são igualmente possíveis. Para escolher qual covariável eliminar no modelo partimos da suposição intuitiva de que as covariáveis COHABIT, INCOME e ANTENAT estão intimamente relacionadas e que para a explicação da variável resposta bastaria a influência de uma delas. Optamos por ficar com aquela que possuía o menor número de categorias que é a COHABIT. Portanto, eliminamos as covariáveis INCOME e ANTENAT. Fizemos, novamente no SAS, um quarto ajuste com a covariável AGE categorizada e as covariáveis SMOKING e COHABIT. O valor da estatística para verificar a adequacidade do modelo de odds proporcionais foi 15,09, com nível de significância associado de 0,09. Isto sugere que o modelo de odds proporcionais parece adequado para descrever as relações entre baixo peso ao nascer com idade, hábito de fumar e situação de moradia da mãe. A tabela abaixo mostra os resultados deste último ajuste, apresentado aqui como um ajuste final para estes dados.

TABELA 3.9 - RESULTADO DO AJUSTE DO MODELO DE ODDS PROPORCIONAIS PARA OS DADOS DO EXEMPLO 2

bwgr	β_{ij}	SE	Wald
age	0.0220	0.0042	5,251
smoking	-.2333	0.0192	-12.129
cohabit	.3943	0.0872	4.525
θ_1	-2.6570	0.1448	
θ_2	-1.5371	0.1337	
θ_3	.0222	0.1302	
θ_4	1.6324	0.1322	

Uma interpretação quantitativa pode ser obtida através do cálculo do quociente de chances descrito na expressão 3.2. Por exemplo,

$$\exp [\ \underset{\sim}{\beta}^t\ (x_{i'}-x_i)] = \exp [\ 0,02(3-5) - 0,23\ (0-4) + 0,39\ (1-0)]$$

$$= \exp [\ -0,04 + 0,92 + 0,39\ ] = 3,56.$$

Isto significa que a chance de um indivíduo pesar menos de 2000g para um indivíduo cuja mãe tem idade entre 40 e 46 anos, fumou muito durante toda a gravidez e não mora com o pai é 3,56 vezes maior do que para um indivíduo cuja mãe tem idade entre

26 e 32 anos, não fumou durante a gravidez e mora com o pai. A interpretação para as outras categorias da variável resposta é a mesma. A interpretação para as outras combinações das covariáveis pode ser construídas de maneira análoga. Salientamos que o número dessas combinações é bem alto e que, normalmente, só fazemos a interpretação quantitativa para àquelas mais importantes do ponto de vista clínico.

CAPÍTULO 4: RECURSOS COMPUTACIONAIS

Normalmente, um grande obstáculo encontrado na análise estatística de modelos mais elaborados são recursos computacionais. O objetivo deste capítulo é o de apresentar alguns procedimentos para o ajuste dos modelos de regressão logística encontrados em softwares estatísticos. Pacotes estatísticos como o SPSS, BMDP, MULTILR, SAS e STATA podem ser utilizados para o ajuste de modelos de regressão logística. Descreveremos com maiores detalhes os procedimentos para regressão logística politômica apresentados pelo STATA que foi o pacote utilizado para os ajustes apresentados neste trabalho. Citaremos, também, alguns aspectos do procedimento LOGISTIC do software SAS que também utilizaremos para ajustar o modelo de odds proporcionais.

4.1. O PACOTE STATA

O STATA é um pacote estatístico bastante completo. Dentre outras análises ele pode ser utilizado para o ajuste dos modelos de regressão logística. Primeiramente apresentaremos a descrição de alguns comandos. Em seguida, a relação dos comandos utilizados para o ajuste de modelos com mais de uma covariável explicativa, bem como as respectivas saídas com a especificação dos valores apresentados. Maiores detalhes podem ser obtidos em Computing Research Center (1992).

4.1.1. Comandos do STATA

No STATA podemos entrar com os dados de duas maneiras básicas: importando dados de um determinado arquivo ou digitando no próprio pacote.

Para importar os dados de um determinado arquivo, este deve estar em ASCII com uma coluna em branco separando cada variável. Este arquivo deve ter, obrigatoriamente, a terminação .raw. *Usamos o seguinte comando:*

.infile <u>nome das var. conforme aparecem no arquivo</u> **using** a:exe.raw

Para entrarmos diretamente com os dados existem duas maneiras. A primeira é quando temos apenas duas variáveis categóricas e queremos entrar com os dados na forma de uma tabela de contingência 2x2. Neste caso usamos o seguinte comando:

.input <u>*nome das duas var.*</u> ***pop***

No lugar das variáveis colocamos as diferentes combinações das categorias e no lugar do pop a frequência correspondente a cada uma. Observe que o nome dado para a variável referente às frequências necessariamente deve ser pop. Quando fizermos qualquer análise para estes dados devemos colocar no comando, logo após o nome das variáveis, a expressão [freq=pop].

A segunda maneira é quando queremos entrar com os valores absolutos. Neste caso podemos ter mais de duas variáveis. O comando a ser usado é:

.*input* *nome das variáveis*

Em ambos os casos, para indicar que não existem mais dados a serem digitados escrevemos a expressão end.

Em qualquer momento do trabalho podemos querer gravar o arquivo de dados. Seja porque digitamos os dados ou porque recodificamos o arquivo ASCII já existente ou até mesmo para simplificar o trabalho tendo o arquivo na linguagem do STATA. O comando a ser usado é:

. ***save using*** *a:exemplo*

Não é necessário colocar terminação no nome do arquivo, pois automaticamente ele adquire a terminação .dta que indica que o arquivo está na linguagem do STATA.

Para gravar um arquivo com os comandos e as saídas que foram sendo processadas devemos, antes de iniciar o trabalho, dar o respectivo comando, a saber:

.log using a:exemplo.log

A terminação .log não é obrigatória podendo ser qualquer outra.Ela apenas uniformiza indicando que todos os arquivos .log são arquivos com comandos e saídas do STATA. Com estes arquivos podemos entrar em um editor de textos.

Através dos comandos MLOGIT e OLOGIT podemos ajustar, respectivamente, modelos de regressão logística politômica nominal e modelos de odds proporcionais. Em ambos, os parâmetros das funções logit são estimados pelo método de máxima verossimilhança sendo que o processo iterativo utilizado é o de Newton-Raphson. A parametrização adotada é logit $\gamma_{ij} = \theta_j - \underset{\sim}{\beta}^t x_i$. Como opção do comando MLOGIT podemos especificar a categoria de referência para a variável resposta através do subcomando "base" (o default é considerar como categoria de referência a de maior frequência) e o nível de significância para o intervalo de confiança dos coeficientes de regressão através do subcomando "level" (o default é 0,05). A sintaxe do comando é:

.mlogit *var. dependente* *var. independentes*, ***base ( ), level ( )***

.ologit *var. dependente* *var. independentes*

O comando OLOGIT permite também utilizar métodos automáticos para a seleção de modelos. Neste caso devemos especificar se queremos que seja Backward ou Forward (o default é Backward). Como opção deste comando podemos especificar o nível de significância para que a variável entre no modelo através do subcomando "pe" (o default é 0,2) e também o nível de significância para que uma variável seja removida do modelo através do subcomando "pr" (o default é 0,4). A sintaxe para este comando é:

.***swologit*** *var. depend.* *var. independentes,* ***pe*** *( ),* ***pr*** *( )* *forward*

Para criar as variáveis de delineamento relacionadas com uma covariável com mais de duas categorias deve ser utilizado o comando TABULATE-GENERATE. Ele deve ser usado antes do comando MLOGIT. A sintaxe deste comando é:

.***tabulate*** *EXEMPLO,* ***generate*** *(EXEMPLO)*

Ele cria as covariáveis EXEMPLO1,...,EXEMPLOn onde n é o número de categorias da covariável EXEMPLO.

O comando OLOGITP é utilizado para o cálculo das probabilidades preditas pelo modelo de odds proporcionais. A sintaxe deste comando é

. ***ologitp*** *nome das categorias da variável resposta*

Para cada cada categoria ele cria uma variável que fornece o valor predito para cada um dos indivíduos. Para ver estes valores basta mandar listar estas novas variáveis conjuntamente com as covariáveis envolvidas no modelo. Note que, para cada combinação das categorias das covariáveis teremos um valor predito para cada uma das novas variáveis. A sintaxe do comando para listar é

.list _nome das cat. da var. resposta_ _nome das covariáveis_

4.1.2. EXEMPLOS DE PROGRAMAS NO STATA

Abaixo, estão relacionados, os camandos usados no ajuste do modelo de regressão logística politômica nominal com mais de uma covariável, para o Exemplo 1. A relação das respectivas saídas encontram-se no Apêndice.

```
.infile OBS ME SYMPT PB HIST BSE DETC using a:mamogra.raw
.tabulate DETC, generate (DETC)
.tabulate SYMPT, generate (SYMPT)
.mlogit ME DETC2 DETC3 BSE HIST PB SYMPT2 SYMPT3 SYMPT4
.recode SYMPT 1/2=0 3/4=1
.rename SYMPT SYMPD
.mlogit ME DETC2 DETC3 BSE HIST PB SYMPD
```

Abaixo, estão relacionados os comandos utilizados para o ajuste do modelo de odds proporcionais para o caso multivariado do Exemplo2. As respectivas saídas encontram-se no Apêndice.

```
.infile obs age cohabit educat income race smoking antenat typdeliv bwgr
.ologit bwgr age cohabit educat income race smoking antenat typdeliv
.recode age 13/15=15 16/25=21 26/32=29 33/29=35 40/46=41
.recode antenat 1/6=4 7/11=8 12/27=13
.ologit bwgr age cohabit educat income race smoking antenat typdeliv
.ologit bwgr age cohabit income smoking antenat
.ologit bwgr age cohabit smoking
.ologitp peso0 peso1 peso2 peso3 peso4
.list peso0 peso1 peso2 peso3 peso4 age cohabit smoking
```

4.1.3. EXEMPLO DE SAÍDAS DO STATA COM A ESPECIFICAÇÃO DOS RESULTADOS APRESENTADOS

XEMPLO DE SAÍDA DO PROCEDIMENTO MLOGIT DO STATA

```
logit me detc2 detc3

teration 0:  Log Likelihood = -365.1467 (1)
teration 1:  Log Likelihood =-353.85622
teration 2:  Log Likelihood =-353.36584
teration 3:  Log Likelihood =-353.34908
teration 4:  Log Likelihood =  -353.349

ultinomial regression                         Number of obs =     375 (3)
                                              chi2(4)       =   23.60 (4)
                                              Prob > chi2   =  0.0001 (5)
og Likelihood =    -353.349 (2)               Pseudo R2     =  0.0323
------------------------------------------------------------------------------
     me |  β̂ij Coef. (6)  Std. Err. (7)    t (8)   P>|t| (9)   [95% Conf. Interval] (10)
                           SE             W
--------+---------------------------------------------------------------------
        |
  detc2 (14)|  .8157495   1.083824      0.753   0.452     -1.315496   2.946995
  detc3 |  2.132403   1.047018      2.037   0.042      .0735337   4.191273
  _cons | -2.564949   1.037713     -2.472   0.014     -4.605522  -.5243771
--------+---------------------------------------------------------------------
        |
  detc2 | -.4163942   .6425308     -0.648   0.517     -1.679876   .8470872
) detc3 |  .1320991   .5969064      0.221   0.825     -1.041666   1.305864
  _cons | -1.178655   .5717732     -2.061   0.040     -2.302998  -.0543125
------------------------------------------------------------------------------
Outcome me==0 is the comparison group)
         (13)
```

) Fornece o valor do logaritmo da função de verossimilhaça para ada uma das iterações indicadas

2) Indica o último valor

3) Número total de indivíduos

4) Valor calculado para a estatística G= -2 (L1 - L0) onde
L0: logaritmo da verossimilhança do modelo com as covariáveis
L1: logaritmo da verossimilhança do modelo sem as covariáveis

5) Nível de significância associado ao valor da estatística G

6) Fornece o valor dos coeficientes de regressão estimados

(7) Fornece o valor do erro padrão dos coeficientes de regressão estimados

(8) Fornece o valor da estatística W de Wald

(9) Nível de significância associado ao valor da estatística W

(10) Intervalo de confiança para os coeficientes de regressão estimados

(11) Valores para a primeira função logit

(12) Valores para a segunda função logit

(13) Categoria da variável resposta considerada como de referência

(14) Variáveis de delineamento

EXEMPLO DE SAÍDA DO PROCEDIMENTO OLOGIT DO STATA

```
logit bwgr race

teration 0:  Log Likelihood =-8068.6205
teration 1:  Log Likelihood =-8057.9252
teration 2:  Log Likelihood =-8057.9223

rdered Logit Estimates                          Number of obs =    5939
                                                chi2(1)       =   21.40
                                                Prob > chi2   =  0.0000
og Likelihood = -8057.9223                      Pseudo R2     =  0.0013
```

bwgr	$\hat{\beta}_{ij}$ Coef.	Std. Err. SE	t W	P>|t|	[95% Conf.	Interval]
race	.284459	.0614638	4.628	0.000	.1639677	.4049503
_cut1	-3.108692	.0862595 (15)		(Ancillary parameters)		
_cut2	-1.996989	.0656028				
_cut3	-.4678662	.0570108				
_cut4	1.100216	.0585927				

(15) Valores dos parâmetros de corte estimados.

OBS: os demais resultados são os mesmos apresentados no procedimento MLOGIT

4.2. O PACOTE SAS

Podemos dizer que, atualmente, o SAS talvez seja um dos pacotes estatísticos mais completo e importante. Na versão 6.04 é possível, através do procedimento LOGISTIC e mediante o método de máxima verossimilhança ajustar modelos de regressão logística linear para dados com respota binária ou categórica ordenada. As funções de ligação logit, normit, e complementar log-log estão disponíveis neste procedimento. Para ajustarmos o modelo de odds proporcionais basta optarmos pela função de ligação logit. Como opção tem-se os métodos de seleção automática de modelos Backward,Forward e Stepwise. O método iterativo utilizado para o cálculo das estimativas de máxima verossimilhança é o IRLS (Mínimos quadrados iterativamente reponderados, em português). Devemos observar que a parametrização adotada tem a forma logit $(\gamma_{ij})=\theta_j+\underset{\sim}{\beta}^t x_i$ e que a adotada nesta monografia é logit $(\gamma_{ij})=\theta_j-\underset{\sim}{\beta}^t x_i$. Apresentaremos a seguir uma breve descrição da saída default deste procedimento. Para maiores detalhes veja SAS Institute Inc (1989, p.1071).

Na saída default, primeiramente são apresentadas estatísticas descritivas (média,desvio padrão, mínimo e máximo) de cada covariável. Em seguida os resultados para os diversos passos do método iterativo; o teste de escores para a adequacidade do modelo de odds proporcionais; o valor calculado para três critérios (AIC, SC , -2log L) de avaliação do ajuste do modelo com os respectivos níveis de significância associados, onde salientamos o que -2log L é a estatística G descrita na seção 2.2.1; e finalmente, o quadro para a análise dos estimadores de máxima verossimilhança, onde é fornecido o valor estimado, o desvio padrão e a estatística de Wald de cada estimador com o respectivo nível de significância associado.

4.2.1. Exemplo de programa no SAS

Abaixo estão relacionados os comandos utilizados para o ajuste do modelo de odds proporcionais com mais de uma covariável, para o Exemplo 2.

```
data paty;
infile 'a:paty.dat';
input obs age cohabit educat income race smoking antenat typdeliv
      bwgr;
proc logistic data=paty order=data;
     model bwgr=age cohabit income smoking antenat/ itprint;
run;
```

Observe que o comando <u>order=paty</u> foi utilizado porque o arquivo de dados estava em ordem crescente pela variável resposta BWGR.

CAPITULO 5: CONCLUSÃO

Nesta monografia foram abordados alguns aspectos sobre uma técnica estatística que pode ser utilizada para descrever as relações entre uma variável resposta e uma ou mais variáveis explicativas, chamada de regressão logística. É uma técnica recente que atualmente está sendo muito usada na área biomédica, principalmente em estudos epidemiológicos. Esse tipo de estudo é importante, pois mais do que nunca vemos a necessidade de técnicas estatísticas mais elaboradas para tratar de dados categóricos nominais e ordinais. Verificamos, através do Exemplo 1, que a regressão logística politômica nominal parece adequada para descrever as relações entre as covariáveis em estudo e a experiência com mamografia. Através do Exemplo 2 constatamos que a regressão logística politômica ordinal é uma técnica adequada para descrever as relações entre as variáveis explicativas e o baixo peso ao nascer. Salientamos que a regressão logística politômica é uma técnica bastante sofisticada e o trabalho de modelagem requer estudos adicionais, tais como técnicas de diagnóstico e análise de resíduos. Uma conclusão importante que podemos obter com o ajuste destes modelos é sobre direção e magnitude dos efeitos das variáveis explicativas sobre a resposta.

O processo de modelagem da regressão logística é bastante trabalhoso, porém com o uso de pacotes estatísticos ele fica bem mais prático. Constatamos que tanto o pacote estatístico STATA quanto o SAS são excelentes ferramentas para o ajuste destes modelos. Destacamos o fato de que o procedimento LOGISTIC do SAS apresenta mais recursos para a escolha e avaliação do ajuste do modelo de odds proporcionais. Uma limitação do STATA é que para o ajuste do modelo saturado (necessário para a avaliação do ajuste do modelo) é necessário uma quantidade de memória que não é compatível com qualquer micro. Uma limitação comum aos dois pacotes é a ausência de procedimentos para a análise de resíduos e diagnóstico, úteis para o refinamento dos modelos.

Por fim, podemos dizer que a regressão logística é uma técnica muito útil que necessita de maiores estudos assim como a investigação de sua aplicabilidade em outras áreas, como por exemplo em experimentos industriais. Outro aspecto para estudos futuros é o caso onde a variável resposta for composta por casos raros.

A P Ê N D I C E S

AJUSTE DO MODELO DE REGRESSÃO LOGÍSTICA POLITÔMICA NOMINAL PARA AS VARIÁVEIS HIST E ME DO EXEMPLO 1 (PROCEDIMENTO MLOGIT DO STATA).

```
mlogit me hist

Iteration 0:  Log Likelihood = -365.1467
Iteration 1:  Log Likelihood =-359.56929
Iteration 2:  Log Likelihood =-359.51398
Iteration 3:  Log Likelihood =-359.51392

Multinomial regression                          Number of obs =     375
                                                chi2(2)       =   11.27
                                                Prob > chi2   =  0.0036
Log Likelihood = -359.51392                     Pseudo R2     =  0.0154

------------------------------------------------------------------------------
      me |      Coef.   Std. Err.        t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
         |
    hist |   1.241713   .3872709      3.206   0.001       .4801919   2.003234
   _cons |  -.9162907   .1322876     -6.927   0.000      -1.176418  -.6561633
---------+--------------------------------------------------------------------
         |
    hist |   .9232594   .4595821      2.009   0.045       .0195469   1.826972
   _cons |  -1.290984   .1522558     -8.479   0.000      -1.590377  -.9915917
------------------------------------------------------------------------------
(Outcome me==0 is the comparison group)
```

JUSTE DO MODELO DE REGRESSÃO LOGÍSTICA POLITÔMICA NOMINAL PARA AS VARIÁVEIS ETC E ME DO EXEMPLO 1 (PROCEDIMENTO MLOGIT DO STATA).

```
logit me detc2 detc3

teration 0:   Log Likelihood = -365.1467
teration 1:   Log Likelihood =-353.85622
teration 2:   Log Likelihood =-353.36584
teration 3:   Log Likelihood =-353.34908
teration 4:   Log Likelihood =  -353.349

ultinomial regression                              Number of obs =     375
                                                   chi2(4)       =   23.60
                                                   Prob > chi2   =  0.0001
og Likelihood =   -353.349                         Pseudo R2     =  0.0323

------------------------------------------------------------------------------
      me |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
         |
   detc2 |   .8157495   1.083824     0.753   0.452      -1.315496    2.946995
   detc3 |   2.132403   1.047018     2.037   0.042       .0735337    4.191273
   _cons |  -2.564949   1.037713    -2.472   0.014      -4.605522   -.5243771
---------+--------------------------------------------------------------------
         |
   detc2 |  -.4163942   .6425308    -0.648   0.517      -1.679876    .8470872
   detc3 |   .1320991   .5969064     0.221   0.825      -1.041666    1.305864
   _cons |  -1.178655   .5717732    -2.061   0.040      -2.302998   -.0543125
------------------------------------------------------------------------------
Outcome me==0 is the comparison group)
```

JUSTE DO MODELO DE REGRESSÃO LOGÍSTICA POLITÔMICA NOMINAL COM MAIS DE UMA OVARIÁVEL SIMULTÂNEAMENTE PARA OS DADOS DO EXEMPLO 1 (PROCEDIMENTO MLOGIT DO TATA).

```
logit me detc2 detc3 bse hist pb symp2 symp3 symp4

teration 0:  Log Likelihood = -365.1467
teration 1:  Log Likelihood =-320.60421
teration 2:  Log Likelihood =  -317.042
teration 3:  Log Likelihood =-316.80426
teration 4:  Log Likelihood =-316.80184
teration 5:  Log Likelihood =-316.80184

ultinomial regression                          Number of obs =     375
                                               chi2(16)      =  .96.69
                                               Prob > chi2   = 0.0000
og Likelihood = -316.80184                     Pseudo R2     = 0.1324
```

me	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
detc2	.1438801	1.15984	0.124	0.901	-2.137098	2.424858
detc3	.9533716	1.12672	0.846	0.398	-1.262472	3.169215
bse	1.208437	.535249	2.258	0.025	.1557998	2.261075
hist	1.370253	.4501726	3.044	0.003	.4849301	2.255577
pb	-.2232325	.0790048	-2.826	0.005	-.3786057	-.0678593
sympt2	.1568709	.9323956	0.168	0.866	-1.676807	1.990549
sympt3	1.837993	.7894126	2.328	0.020	.2855098	3.390477
sympt4	2.373288	.7849259	3.024	0.003	.8296285	3.916948
_cons	-2.857967	1.545035	-1.850	0.065	-5.89648	.1805456
detc2	-.9273621	.7152562	-1.297	0.196	-2.334007	.479283
detc3	-.7064893	.6883453	-1.026	0.305	-2.060211	.647232
bse	.893279	.5200801	1.718	0.087	-.1295269	1.916085
hist	.9395071	.4873322	1.928	0.055	-.0188956	1.89791
pb	-.1584003	.0817065	-1.939	0.053	-.3190868	.0022862
sympt2	-.0407409	.6889202	-0.059	0.953	-1.395593	1.314111
sympt3	.922274	.5958229	1.548	0.123	-.2494899	2.094038
sympt4	1.092843	.6058355	1.804	0.072	-.0986122	2.284298
_cons	-.8582264	1.144444	-0.750	0.454	-3.108926	1.392473

)utcome me==0 is the comparison group)

RECODIFICANDO A COVARIÁVEL SYMPT

```
ecode sympt 1/2=0 3/4=1
375 changes made)

ab sympt
```

sympt	Freq.	Percent	Cum.
0	101	26.93	26.93
1	274	73.07	100.00
Total	375	100.00	

```
logit me detc2 detc3 bse hist pb sympt

teration 0:  Log Likelihood = -365.1467
teration 1:  Log Likelihood =-322.22347
teration 2:  Log Likelihood =-318.64519
teration 3:  Log Likelihood =-318.39687
teration 4:  Log Likelihood =-318.39424
teration 5:  Log Likelihood =-318.39424

ultinomial regression                             Number of obs =     375
                                                  chi2(12)      =   93.50
                                                  Prob > chi2   =  0.0000
og Likelihood = -318.39424                        Pseudo R2     =  0.1280
```

me	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
detc2	.2080191	1.158582	0.180	0.858	-2.070399	2.486437
detc3	1.023161	1.125341	0.909	0.364	-1.189886	3.236208
bse	1.16292	.531718	2.187	0.029	.1172664	2.208574
hist	1.294065	.4473807	2.893	0.004	.4142649	2.173864
pb	-.2559549	.0757462	-3.379	0.001	-.404914	-.1069957
sympt	1.981809	.4614431	4.295	0.000	1.074355	2.889264
_cons	-2.525357	1.443859	-1.749	0.081	-5.364788	.3140748
detc2	-.9159057	.7152321	-1.281	0.201	-2.32245	.490639
detc3	-.6913199	.6876081	-1.005	0.315	-2.04354	.6609006
bse	.8818943	.519356	1.698	0.090	-.1394488	1.903237
hist	.9258337	.4837624	1.914	0.056	-.0255127	1.87718
pb	-.1677395	.0788058	-2.129	0.034	-.3227156	-.0127635
sympt	1.014274	.375162	2.704	0.007	.276497	1.752052
_cons	-.8109393	1.095623	-0.740	0.460	-2.965545	1.343666

```
Outcome me==0 is the comparison group)
```

bserve que neste último ajuste a covariável SYMPT é a SYMPD referida no exto.

AJUSTE DO MODELO DE ODDS PROPORCIONAIS COM AS VARIÁVEIS RACE E BWGR DO EXEMPLO 2 (PROCEDIMENTO OLOGIT DO STATA).

```
ologit bwgr race

Iteration 0:  Log Likelihood =-8068.6205
Iteration 1:  Log Likelihood =-8057.9252
Iteration 2:  Log Likelihood =-8057.9223

Ordered Logit Estimates                           Number of obs =     5939
                                                  chi2(1)       =    21.40
                                                  Prob > chi2   =   0.0000
Log Likelihood = -8057.9223                       Pseudo R2     =   0.0013

------------------------------------------------------------------------------
    bwgr |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    race |    .284459   .0614638      4.628   0.000       .1639677    .4049503
---------+--------------------------------------------------------------------
   _cut1 |  -3.108692   .0862595           (Ancillary parameters)
   _cut2 |  -1.996989   .0656028
   _cut3 |  -.4678662   .0570108
   _cut4 |   1.100216   .0585927
------------------------------------------------------------------------------
```

JUSTE DO MODELO DE ODDS PROPORCIONAIS PARA AS VARIÁVEIS SMOKING E BWGR DO EXEMPLO 2 (PROCEDIMENTO OLOGIT DO STATA).

```
logit bwgr smoking

teration 0:   Log Likelihood =-8068.6205
teration 1:   Log Likelihood =-7985.9935
teration 2:   Log Likelihood =-7985.8142

rdered Logit Estimates                                Number of obs =    5939
                                                      chi2(1)       =  165.61
                                                      Prob > chi2   =  0.0000
og Likelihood = -7985.8142                            Pseudo R2     =  0.0103

------------------------------------------------------------------------------
    bwgr |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 smoking |  -.2455054   .0191588    -12.814   0.000      -.2830637   -.2079471
---------+--------------------------------------------------------------------
   _cut1 |  -3.582261   .0743739             (Ancillary parameters)
   _cut2 |  -2.464302    .048307
   _cut3 |  -.9122932   .0326353
   _cut4 |   .6857869   .0315639
------------------------------------------------------------------------------
```

JUSTE DO MODELO DE ODDS PROPORCIONAIS COM MAIS DE UMA COVARIÁVEL IMULTÂNEAMENTE PARA OS DADOS DO EXEMPLO 2

JUSTE COM TODAS AS COVARIÁVEIS DO EXEMPLO COM SEUS DADOS ORIGINAIS PROCEDIMENTO OLOGIT DO STATA)

```
logit bwgr age-antenat typdeliv

teration 0:  Log Likelihood =-8068.6205
teration 1:  Log Likelihood =-7898.0643
teration 2:  Log Likelihood = -7897.336
teration 3:  Log Likelihood =-7897.3358

rdered Logit Estimates                              Number of obs =    5939
                                                    chi2(8)       =  342.57
                                                    Prob > chi2   =  0.0000
og Likelihood = -7897.3358                          Pseudo R2     =  0.0212

------------------------------------------------------------------------------
    bwgr |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .0180743   .0040806      4.429   0.000       .0100748    .0260738
 cohabit |   .1925496     .08958      2.149   0.032       .0169403     .368159
  educat |    .001141   .0075592      0.151   0.880      -.0136777    .0159598
  income |   .1104341   .0312541      3.533   0.000       .0491646    .1717036
    race |   .1063682    .063907      1.664   0.096      -.0189128    .2316491
 smoking |  -.2069768   .0194123    -10.662   0.000       -.245032   -.1689216
 antenat |   .0570388   .0085341      6.684   0.000       .0403089    .0737687
ypdeliv |   .0332389   .0316186      1.051   0.293       -.028745    .0952228
---------+--------------------------------------------------------------------
   _cut1 |  -2.314434   .1573004           (Ancillary parameters)
   _cut2 |  -1.186331   .1472918
   _cut3 |   .3915211   .1448583
   _cut4 |   2.024189   .1473698
------------------------------------------------------------------------------
```

:ATEGORIZANDO AS COVARIÁVEIS AGE E ANTENAT

```
ecode age 13/15=15 16/25=21 26/32=29 33/39=35 40/46=41
5087 changes made)


ab age
        age|      Freq.     Percent        Cum.
-----------+-----------------------------------
         15 |        64        1.08        1.08
         21 |      3048       51.32       52.40
         29 |      1896       31.92       84.32
         35 |       785       13.22       97.54
         41 |       146        2.46      100.00
-----------+-----------------------------------
      Total |      5939      100.00


ecode antenat 1/6=4 7/11=8 12/27=13
4288 changes made)


ab antenat
    antenat|      Freq.     Percent        Cum.
-----------+-----------------------------------
          0 |       300        5.05        5.05
          4 |      2395       40.33       45.38
          8 |      2910       49.00       94.38
         13 |       334        5.62      100.00
-----------+-----------------------------------
      Total |      5939      100.00
```

```
ologit bwgr age-antenat typdeliv

Iteration 0:  Log Likelihood =-8068.6205
Iteration 1:  Log Likelihood = -7907.872
Iteration 2:  Log Likelihood =-7907.2407
Iteration 3:  Log Likelihood =-7907.2406

Ordered Logit Estimates                             Number of obs =   5939
                                                    chi2(8)       = 322.76
                                                    Prob > chi2   = 0.0000
Log Likelihood = -7907.2406                         Pseudo R2     = 0.0200

------------------------------------------------------------------------------
    bwgr |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .0162668   .0043084      3.776   0.000       .0078208    .0247128
 cohabit |   .2228019   .0893781      2.493   0.013       .0475882    .3980155
  educat |   .0032579   .0075339      0.432   0.665      -.0115113    .0180272
  income |   .1223728   .0311689      3.926   0.000       .0612703    .1834753
    race |   .1085462   .0638093      1.701   0.089      -.0165432    .2336357
 smoking |  -.2097986   .0194047    -10.812   0.000       -.247839   -.1717583
 antenat |   .0535192   .0092796      5.767   0.000       .0353278    .0717105
typdeliv |    .026684   .0315518      0.846   0.398      -.0351689     .088537
---------+--------------------------------------------------------------------
   _cut1 |  -2.348029   .1636523              (Ancillary parameters)
   _cut2 |  -1.221223   .1540373
   _cut3 |   .3532774   .1516077
   _cut4 |    1.98204   .1539172
```

```
LIMINANDO AS COVARIÁVEIS EDUCAT TYPDELIV RACE

logit bwgr age cohabit income smoking antenat

teration 0:   Log Likelihood =-8068.6205
teration 1:   Log Likelihood =-7909.7761
teration 2:   Log Likelihood = -7909.161
teration 3:   Log Likelihood =-7909.1609

rdered Logit Estimates                                 Number of obs =    5939
                                                       chi2(5)       =  318.92
                                                       Prob > chi2   =  0.0000
og Likelihood = -7909.1609                             Pseudo R2     =  0.0198

------------------------------------------------------------------------------
    bwgr |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .0154129   .0042602      3.618   0.000       .0070613    .0237644
 cohabit |   .2288058    .088838      2.576   0.010       .0546509    .4029607
  income |   .1377037   .0248619      5.539   0.000       .0889655     .186442
 smoking |  -.2104446   .0193943    -10.851   0.000      -.2484646   -.1724247
 antenat |    .054668   .0091283      5.989   0.000       .0367732    .0725629
---------+--------------------------------------------------------------------
   _cut1 |  -2.470018   .1487551          (Ancillary parameters)
   _cut2 |  -1.343225    .138105
   _cut3 |   .2303481   .1351079
   _cut4 |   1.858095    .137363
------------------------------------------------------------------------------
```

AJUSTE COM AS COVARIÁVEIS AGE SMOKING COHABIT INCOME ANTENAT (PROCEDIMENTO LOGISTIC DO SAS)

The LOGISTIC Procedure

Data Set: WORK.PATY
Response Variable: BWGR
Response Levels: 5
Number of Observations: 5939
Link Function: Logit

Response Profile

Ordered Value	BWGR	Count
1	0	204
2	1	375
3	2	1392
4	3	2207
5	4	1761

Simple Statistics for Explanatory Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
AGE	25.831453	5.763883	15.0000	41.0000
SMOKING	0.805860	1.241553	0.0000	4.0000
COHABIT	0.918505	0.273617	0.0000	1.0000

Maximum Likelihood Iterative Phase

Iter	Step	-2 Log L	INTERCP1 AGE	INTERCP2 SMOKING	INTERCP3 COHABIT	INTERCP4
0	INITIAL	16137	-3.336223	-2.225417	-0.699721	0.863951
			0	0	0	
1	IRLS	15920	-2.622773	-1.511966	0.013729	1.577401
			-0.021266	0.229807	-0.380314	
2	IRLS	15916	-2.656445	-1.537228	0.021078	1.631013
			-0.021962	0.233120	-0.394685	
3	IRLS	15916	-2.657005	-1.537112	0.022164	1.632458
			-0.022021	0.233257	-0.394484	
4	IRLS	15916	-2.656991	-1.537099	0.022174	1.632473
			-0.022022	0.233256	-0.394477	
5	IRLS	15916	-2.656991	-1.537098	0.022175	1.632473
			-0.022022	0.233256	-0.394477	

Last Change in -2 Log L: 1.0355325E-7

Last Evaluation of Gradient

INTERCP1	INTERCP2	INTERCP3	INTERCP4	AGE	SMOKING	COHABIT
4.682E-6	-0.000021	-0.000077	-0.000011	-0.003906	-0.000085	-0.000086

Score Test for the Proportional Odds Assumption

Chi-Square = 15.0955 with 9 DF (p=0.0883)

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	16145.241	15930.334	.
SC	16171.998	15977.159	.
-2 LOG L	16137.241	15916.334	220.907 with 3 DF (p=0.0001)
Score	.	.	216.594 with 3 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCP1	-2.6570	0.1444	338.4712	0.0001	.
INTERCP2	-1.5371	0.1333	132.9954	0.0001	.
INTERCP3	0.0222	0.1299	0.0292	0.8644	.
INTERCP4	1.6325	0.1317	153.6484	0.0001	.
AGE	-0.0220	0.00417	27.8441	0.0001	-0.069983
SMOKING	0.2333	0.0192	147.2265	0.0001	0.159665
COHABIT	-0.3945	0.0869	20.5989	0.0001	-0.059508

Association of Predicted Probabilities and Observed Responses

Concordant	= 50.9%	Somers' D	= 0.162
Discordant	= 34.7%	Gamma	= 0.189
Tied	= 14.4%	Tau-a	= 0.116
(12589923 pairs)		c	= 0.581

AJUSTE COM AS COVARIÁVEIS AGE SMOKING COHABIT (PROCEDIMENTO LOGISTIC DO SAS)

Data Set: WORK.PATY
Response Variable: BWGR
Response Levels: 5
Number of Observations: 5939
Link Function: Logit

Response Profile

Ordered Value	BWGR	Count
1	0	204
2	1	375
3	2	1392
4	3	2207
5	4	1761

Simple Statistics for Explanatory Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
AGE	25.831453	5.763883	15.0000	41.0000
SMOKING	0.805860	1.241553	0.0000	4.0000
COHABIT	0.918505	0.273617	0.0000	1.0000
INCOME	1.259303	1.049040	0.0000	4.0000
ANTENAT	6.264018	2.842326	0.0000	13.0000

Maximum Likelihood Iterative Phase

Iter	Step	-2 Log L	INTERCP1 AGE ANTENAT	INTERCP2 SMOKING	INTERCP3 COHABIT	INTERCP4 INCOME
0	INITIAL	16137	-3.336223	-2.225417	-0.699721	0.863951
			0	0	0	0
			0			
1	IRLS	15825	-2.427039	-1.316233	0.209463	1.773135
			-0.014697	0.204826	-0.217523	-0.134631
			-0.051926			
2	IRLS	15818	-2.467936	-1.342528	0.229139	1.856391
			-0.015330	0.210141	-0.228811	-0.137487
			-0.054811			
3	IRLS	15818	-2.470025	-1.343226	0.230344	1.858080
			-0.015412	0.210448	-0.228807	-0.137701
			-0.054668			
4	IRLS	15818	-2.470019	-1.343226	0.230348	1.858095
			-0.015413	0.210445	-0.228806	-0.137704
			-0.054668			

Last Change in -2 Log L: 0.0040584726

Last Evaluation of Gradient

INTERCP1	INTERCP2	INTERCP3	INTERCP4	AGE
0.0012012334	-0.006523592	-0.021302828	0.0010185088	-0.691776631

Last Evaluation of Gradient

SMOKING	COHABIT	INCOME	ANTENAT
-0.031321766	-0.023498433	-0.036006442	-0.162443473

Score Test for the Proportional Odds Assumption

Chi-Square = 66.3615 with 15 DF (p=0.0001)

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	16145.241	15836.322	.
SC	16171.998	15896.525	.
-2 LOG L	16137.241	15818.322	318.919 with 5 DF (p=0.0001)
Score	.	.	309.502 with 5 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCP1	-2.4700	0.1482	277.8981	0.0001	.
INTERCP2	-1.3432	0.1374	95.5607	0.0001	.
INTERCP3	0.2303	0.1344	2.9371	0.0866	.
INTERCP4	1.8581	0.1367	184.8353	0.0001	.
AGE	-0.0154	0.00423	13.2512	0.0003	-0.048979
SMOKING	0.2104	0.0194	117.7972	0.0001	0.144050
COHABIT	-0.2288	0.0886	6.6755	0.0098	-0.034516
INCOME	-0.1377	0.0249	30.5313	0.0001	-0.079643
ANTENAT	-0.0547	0.00910	36.1023	0.0001	-0.085668

Association of Predicted Probabilities and Observed Responses

Concordant	= 57.2%	Somers' D	= 0.201
Discordant	= 37.1%	Gamma	= 0.214
Tied	= 5.7%	Tau-a	= 0.144
(12589923 pairs)		c	= 0.601

AJUSTE COM AS COVARIÁVEIS AGE SMOKING COHABIT (PROCEDIMENTO OLOGIT DO STATA)

```
ologit bwgr age smoking cohabit

Iteration 0:  Log Likelihood =-8068.6205
Iteration 1:  Log Likelihood =-7958.4835
Iteration 2:  Log Likelihood = -7958.167
Iteration 3:  Log Likelihood = -7958.167

Ordered Logit Estimates                               Number of obs =    5939
                                                      chi2(3)       = 220.91
                                                      Prob > chi2   = 0.0000
Log Likelihood =  -7958.167                           Pseudo R2     = 0.0137

------------------------------------------------------------------------------
    bwgr |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .0220224    .004194     5.251   0.000       .0138006    .0302442
 smoking |  -.2332563    .019232   -12.129   0.000       -.270958   -.1955545
 cohabit |   .3944772   .0871828     4.525   0.000       .2235672    .5653872
---------+--------------------------------------------------------------------
   _cut1 |  -2.656991   .1448349             (Ancillary parameters)
   _cut2 |  -1.537098   .1336775
   _cut3 |   .0221748   .1302194
   _cut4 |   1.632474   .1321571
------------------------------------------------------------------------------
```

REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A.(1990). **Categorical Data Analysis**. New York, Wiley.

BISHOP, Y.V.V.; FIENBERG, S.E. and HOLLAND, P.W. (1975). **Discrete Multivariate Analysis**. Cambridge, MIT Press.

COMPUTING RESOURCE CENTER (1992). **Stata Reference Manual**: Release 3. 5th edition. Santa Monica, CA. Vol. 1,2 e 3.

CURETON, E.E.(1978). Psychometrics. Em: KRUSKAL, W.H. & TANUR, J.M. (Editores). **International Encyclopedia of Statistics**, p.764-782. New York, The Free Press.

DOBSON, A.J.(1983). **An Introduction to Statistical Modelling**. London, Chapman and Hall.

EVERITT, B.S.(1992). **The Analysis of Contingency Tables**. Second Edition. London, Chapman and Hall.

HASTIE, T.; BOTHA, J.L. and SCHNITZLER, C.M.(1989). Regression with an ordered categorical response. **Statistics in Medicine**. 8:785-794.

HOSMER, D. W. Jr. & LEMESHOW, S.(1989). **Applied Logistic Regression**. New York, Wiley.

McCULLAGH, P.(1980). Regression models for ordinal data. **J. R. Statist. Soc. B**. 42(2):109-142.

McCULLAGH, P. & NELDER, J.A.(1989). **Generalized Linear Models**. Second Edition. London, Chapman and Hall.

RADUNZ, A.(1992). **Regressão Logística**. Monografia de conclusão do Bacharelado em Estatística da UFRGS.

SAS Institute Inc. (1989). **SAS/STAT User's Guide**. Version 6. Fourth Edition, Vol. 1,2. Cary, NC: SAS Institute, Inc.

VIGO, A.(1994). **Análise de Experimentos Industriais com Respostas Categóricas Ordenadas: o Método de Taguchi e o Modelo de McCullagh**. Dissertação de Mestrado em Estatística - UNICAMP (em fase de conclusão).