UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

CARACTERIZAÇÃO DE MODELOS DE REGRESSÃO PARA ANÁLISE DE DADOS CATEGÓRICOS:

Um Estudo da Coleta do Exame Preventivo de Colo Uterino

AUTORA: Marilei Bender Xavier

ORIENTADORA: Profª. Márcia Echeveste

BANCA EXAMINADORA: Profª. Drª. Jandira Fachel

MONOGRAFIA APRESENTADA PARA A OBTENÇÃO
DO TÍTULO DE BACHAREL EM ESTATÍSTICA

PORTO ALEGRE, MARÇO DE 2003.

Índice

CAPÍTULO 1

1. COMENTARIOS INICIAIS

Grande parte de dados observados em pesquisa de ciências sociais e biomédicas resultam em medidas categóricas. Durante algum tempo, esses dados eram tratados de forma descritiva ou como uma tabela, onde dados cruzados eram analisados. Outros métodos de análise em pesquisa aplicada tornaram-se comum em anos recentes, devido na maior parte, pela disponibilidade de softwares que se desenvolveram em análises para o propósito de dados categóricos. Variáveis de natureza categórica podem ser tratadas através de modelos de regressão, dependendo da escala de medida, das características de classificação destas variáveis em cada experimento. Em particular, é preciso lidar atentamente com o uso de modelos aplicados a cada grupo de variáveis, sejam elas categóricas ou não.

Os métodos e as técnicas estatísticas de análise de dados categóricos vêm se desenvolvendo nos últimos 25 anos. Variáveis categóricas (ou qualitativas) são aquelas que podem ser separadas em diferentes categorias e que se distinguem por alguma classe não-numérica. Esta definição diferencia variável categórica das variáveis contínuas, onde essa última pode assumir uma infinidade de valores.

Antes de classificar os tipos de variáveis é importante diferenciar uma variável explanatória ou explicativa (independente) de uma variável resposta (dependente). Variável explanatória é aquela usada como preditora da variável resposta. Assim, objetiva-se revelar o efeito da variável explanatória sobre a variável resposta.

Os tipos de valores assumidos pelas variáveis resposta e pelas variáveis explanatórias podem ser: (1) variável do tipo dicotômica ou binária; (2) variável politômica nominal; (3) variável politômica ordinal e (4) variável contínua. Variáveis com os valores do tipo (1), (2) e (3) são ditas categóricas, sendo que as variáveis contínuas podem ser categorizadas. A categorização é freqüentemente necessária na prática, quando a pesquisa busca a relação entre uma variável categórica e uma variável contínua.

A variedade de dados em ciências sociais, biológicas e da área médica pode ser apresentada na forma de tabelas cruzadas, comumente referidas como tabelas de contingência. Quando uma população é classificada dentro de muitas categorias, há possibilidade de contar o número de indivíduos em cada categoria, ou seja, se os elementos desta população estão classificados em duas partições probabilísticas, uma com "i" elementos numa das categorias e outra com "j" elementos. Desta forma, a hipótese de que as duas partições são independentes, isto é, dado a informação de que um elemento da amostra pertence a uma determinada categoria de uma delas, não modifica, nem aumenta a probabilidade dela pertencer a determinada categoria da outra variável. Neste caso, uma tabela de contingência 2 X 2 é a forma mais simples de análise, sendo que os dados em cada célula são representados pela freqüência (Fachel, 1999, notas de aula)

O teste qui-quadrado (χ^2) pode ser usado para avaliar a associação entre duas variáveis qualitativas, contudo para descrever a relação entre a variável resposta e estimar o efeito de uma ou mais variáveis explanatórias, são os modelos de regressão que realizam essa análise.

Os modelos de regressão descrevem o relacionamento entre a variável resposta e uma ou mais variáveis explanatórias. Os modelos de regressão mais conhecidos são usados com variáveis resposta contínuas (dependente) e explanatórias (independente).

A maioria dos métodos estatísticos pressupõe dados contínuos. Entretanto, há situações onde as respostas são binárias do tipo "sucesso ou falha". Quando as respostas são medidas por dados binários, devem ser tratadas como dados categóricos. Quando as variáveis explanatórias não são contínuas, a variável do modelo é aplicada para distinguir as diferenças entre grupos (Gujarati, 1995).

Enfim, quando as variáveis resposta são categóricas, os modelos de regressão tradicionais não são adequados. Na regressão quantitativa, modela-se o comportamento da média da variável resposta. No caso de variáveis binárias, modela-se a proporção de sucesso, isto é, quanto o percentual de sucessos depende das variáveis explicativas. De acordo com as características dos dados na pesquisa (contínuo/categórico) e a classificação das variáveis (explanatória/resposta), tem-se uma orientação do método estatístico apropriado.

Para modelar variáveis respostas categóricas em função de variáveis explicativas quantitativas ou qualitativas, os modelos de regressão mais utilizados são os modelos de regressão logística, o modelo Logit, loglinear e o modelo Probit.

Em dados categóricos, o uso da regressão logística tem por objetivo analisar os fatores que são prognósticos para a ocorrência de um evento, por exemplo, em um estudo médico, a sobrevivência ou não de um paciente. A identificação das variáveis que estão relacionadas à probabilidade da sobrevivência ou da morte é de interesse prático e científico. Nesta situação, pode-se fazer uso da regressão logística para determinar as variáveis que estão diretamente relacionadas à variável dependente.

Desta forma, obtém-se uma resposta às perguntas, quais variáveis independentes ou explicativas possuem maior peso em determinar a sobrevivência ou a morte de um paciente.

Nesta mesma linha, há os modelos Logit, loglinear e Probit os quais permitem uma mistura de variáveis independentes categóricas e contínuas, tanto qualitativas quanto quantitativas, relacionadas a uma variável dependente categórica com duas ou mais categorias (nominal ou ordinal). No modelo Probit a variável resposta pode ser quantitativa.

O modelo logit pode ser considerado um caso particular dos modelos loglineares. O modelo logit distingue entre uma única variável resposta e várias variáveis explicativas, ao passo que modelos loglineares não distinguem entre variáveis explicativas e variáveis respostas. Visto que, em muitas aplicações práticas há apenas uma variável resposta e muitas explicativas, o modelo logit é considerado mais prático do que o modelo loglinear. Desta forma, em geral, a ênfase maior é dada no estudo de desenvolvimento em modelos de regressão logística e logit, principalmente na área biomédica. Nesta monografia os modelos apresentados serão o modelo re regressão logística, modelo logit e modelo probit. Estes modelos conduzem, geralmente às mesmas conclusões para os mesmos dados.

Na aplicação destes modelos, existe uma certa confusão por parte de pesquisadores, sobretudo das áreas biomédicas e humanas, no uso de modelos de regressão com dados categóricos, principalmente em determinar as variáveis que irão medir esse fenômeno e na natureza deste relacionamento. Isto é, pode-se ter variáveis respostas dicotômicas, politômicas e variáveis explicativas categóricas dicotômicas ou não, o que dificultam a escolha de um modelo apropriado ou o entendimento de como funcionam esses modelos.

A combinação do tipo de variável resposta e do tipo das variáveis explicativas pode gerar muitas combinações. As possibilidades (situações que podem ser encontradas) para as variáveis independentes (explicativas ou explanatórias), e as variáveis dependentes (resposta), podem ser vistas conforme a Tabela 1.

Tabela 1. Situações para as variáveis dependentes e independentes

Variável resposta ou dependente	Variável explanatórias ou independente		
	Dicotômica	Politômica	Contínua
Dicotômica	(a)	(b)	(c)
Politômica	(d)	(e)	(f)
Contínua	(g)	(h)	(i)

Esta tabela resulta em 9 possibilidades de combinações entre as variáveis resposta e as variáveis explicativas, considerando ainda somente o caso univariado. Estas combinações podem resultar em diferentes alternativas de modelos de regressão que se adaptam às variáveis em questão. Sendo que em muitos casos, um mesmo modelo pode atender a duas ou mais possibilidades destas combinações.

Neste contexto, falta uma orientação aos pesquisadores com o intuito de explicitar o que cada método propõe-se a estimar e como escolher o modelo mais adequado.

1.1 JUSTIFICATIVA DA MONOGRAFIA

Muitas das escolhas feitas por indivíduos e empresas são respostas do tipo binária (sim-não). Uma empresa decide ou não anunciar seus produtos na Internet, um professor precisa decidir entre dois métodos de instrução, ou ainda, um médico decide operar ou não seu paciente.

A estatística interessa-se por explicar por que se fazem tais escolhas e descobrir que fatores entram no processo de decisão. Deseja-se também saber em quanto cada fator afeta o resultado. Tais questões levam-nos ao problema da construção de um modelo, onde a variável resposta, que descreve uma escolha, é uma variável dependente binária. Esse fato afeta as escolhas de um modelo estatístico.

Considerando modelos de regressão linear tradicionais, do tipo $Y_i = \beta_i + \beta_2 X_i + u_i$ que expressa Y_i dicotômico, como uma função linear de uma ou mais variáveis explicativas X_i, são chamados de modelos de probabilidade linear (MPL). Esse modelo é falho e raramente aplica-se na prática, pois apresenta diversas limitações, a saber: (1) não normalidade do termo de erro; (2) heteroscedasticidade e (3) possibilidade da probabilidade estimada ficar fora do limite entre 0 e 1. Mesmo que esses problemas sejam resolvidos, o MPL supõe que as probabilidades condicionais aumentam linearmente com os valores das variáveis explicativas. Mais provavelmente, as probabilidades tenderão a diminuir enquanto os valores das variáveis explicativas aumentam ou diminuem indefinidamente.

Portanto, necessita-se de um modelo de probabilidade que tenha o aspecto de "S" da função distribuição acumulada (f.d.a). Embora a escolha da f.d.a seja ampla, na prática são escolhidas as f.d.a's logística e normal. A primeira dá origem ao modelo Logit e a segunda, ao modelo Probit.

Tanto o modelo Logit quanto o modelo Probit asseguram que as probabilidades estimadas se situem no limite 0-1, onde as probabilidades se relacionam não linearmente com as variáveis explicativas. Vários pesquisadores fizeram ampliações dos modelos Probit e Logit, incluindo Probit bivariado, Probit ordenado. Essa discussão mais detalhada pode ser encontrada no livro de Maddala (1985).

No entanto, apesar da gama de aplicações dos modelos para variáveis categóricas, os pesquisadores não encontram um roteiro para encontrar o modelo certo. Vide algumas organizações realizadas para encontrar a melhor técnica estatística para variáveis quantitativas (Anderson,1996).

Com a ajuda de um material elucidativo e de fácil aplicação, muitos pesquisadores das áreas humanas podem aplicar erradamente modelos e, tomar decisões sem o conhecimento de aspectos importantes de suposições e validação dos modelos.

Por essa razão esta monografia propõe-se a: (i) Apresentar uma descrição dos modelos de regressão Logística, Logit e Probit, através de uma revisão teórica sobre esses assuntos; (ii) Criar um fluxograma que auxilia o pesquisador a encontrar o melhor modelo para seus dados e; (iii) Apresentar a aplicação do fluxograma para um estudo de caso na área médica, podendo ser expandido para outras áreas com estudos similares.

1.2 TEMA E OBJETIVOS

O tema desta monografia é a discussão de modelos de regressão para dados categóricos, na apresentação dos modelos de Regressão Logística, Modelo Logit e Modelo Probit.

Este estudo faz um levantamento bibliográfico para a caracterização de modelos de análise de dados categóricos, considerando os modelos de Regressão Logística, Logit e Probit.

A partir desta caracterização, como objetivo secundário, apresenta-se um fluxograma, indicando qual modelo mais adequado, a partir da natureza das variáveis de estudo. Para exemplificar a aplicação do modelo de análise, será apresentado um estudo prático na área da medicina aplicado no município de Turiaçu, no estado do Maranhão. O estudo deste caso conduzirá a escolha e o detalhamento maior da análise estatística empregada para sua análise.

Este trabalho delimita-se no estudo de três modelos de regressão em variáveis categóricas, modelo de regressão logística, modelo Logit e modelo Probit, apresentando as características de tais modelos, envolvendo o tipo de variável em estudo.

O estudo de caso será um exemplo de uma situação possível utilizando o direcionamento do fluxograma construído, na aplicação de um modelo de análise (Logística, Logit ou Probit) proposto.

Não será realizado neste trabalho de monografia, a aplicação destes modelos para cada tipo de situação envolvendo outros exemplos de variáveis resposta e as variáveis explicativas.

1.3 ESTRUTURA DO TRABALHO

No capitulo 2 é desenvolvido uma introdução na caracterização de modelos de análise de dados categóricos, considerando os modelos de regressão logística, Logit e Probit. Nesta abordagem, justifica-se a importância destes modelos e os objetivos das estimativas. Neste capítulo, também é feita a revisão bibliográfica contendo os seguintes tópicos: (i) análise de dados categóricos; (ii) modelos de análise para dados categóricos (iii) modelo logístico; modelo Logit e modelo Probit. Para cada um dos itens apresentados na seção 2, serão discutidos a interpretação dos modelos, o modelo estatístico e as suposições. Ao final deste capítulo será apresentada uma proposta de mapeamento, através de um fluxograma, auxiliando no direcionamento da escolha do modelo de análise, em função do tipo de variável em estudo, baseado no uso das técnicas discutidas neste capítulo.

No capítulo 3 será apresentado o desenvolvimento de um estudo de caso, descrevendo seus objetivos, levantamento dos dados e aplicação de um dos modelos estudados para análise dos dados, utilizando como ferramenta de auxílio o fluxograma proposto no capítulo 3. Será feita a apresentação dos resultados obtidos a partir da utilização do fluxograma, interpretações, validação do modelo escolhido e ajustes.

No capítulo 4 serão tecidas as considerações finais da realização do trabalho exposto.

CAPÍTULO 2

2. REVISÃO BIBLIOGRÁFICA

A análise de regressão para dados categóricos ocupa-se com o desenvolvimento de modelos que atendam as características da variável dependente e das variáveis independentes. Como lidar com modelos que combinam a mistura de classificações das variáveis, bem como a aplicação a cada caso é de interesse da pesquisa em estatística.

O desenvolvimento de métodos de dados categóricos foi estimulado pela necessidade de uma metodologia estatística nas áreas de ciências sociais e biomédicas. A estatística desenvolveu modelos de regressão para dados categóricos, devida a necessidade de encontrar respostas para análises envolvendo variáveis com muitas categorias. Os níveis de medida das variáveis podem ser classificadas de acordo com o Quadro 1.

Quadro 1. Classificação dos níveis de medida das variáveis para dados categóricos

Variável	Níveis de Medida	Natureza
Qualitativas	Nominal	Dicotômica Politômica
	Ordinal	Politômica
Quantitativas	Discreta	
	Contínuas	Razão Intervalar

Variáveis categóricas são aquelas nas quais a escala de medida consiste na classificação de objetos em categorias. Uma variável categórica, onde os níveis não têm uma ordem natural é chamada de nominal.

Entretanto, muitas variáveis categóricas possuem níveis ordenados, tais como, diagnóstico de cura de um tumor (certo, provável, duvidoso, sem cura), como exemplo.

As variáveis também podem ser classificadas como contínuas ou discretas, de acordo com os valores que elas podem assumir. Variáveis continuas podem assumir uma infinidade de valores, como por exemplo "medidas de um teste", enquanto que variáveis discretas podem assumir valores de número inteiro.

2.1 MODELOS DE ANÁLISE DE DADOS CATEGÓRICOS

Modelos de regressão que envolvem variáveis resposta dicotômicas ou politômicas são aplicáveis em vários campos. Considerando alguns exemplos: um certo medicamento ou é eficaz na cura de uma doença ou não é; suponha que se deseja estudar a situação sindical de professores universitários como um função de diversas variáveis quantitativas e qualitativas. Enfim, para lidar com tais modelos e como estimá-los, serão considerados três modelos: Modelo de Regressão Logística, Modelo Logit e o Modelo Probit.

Antes de discutir os três modelos acima citados, deve-se fazer algumas considerações a respeito do Modelo de Probabilidade Linear (MPL), que expressa Y dicotômico como função linear de uma ou mais variáveis explicativas X.

Considera-se o modelo de probabilidade linear (MPL) apresentado na equação 2.1.

$$Y_i = B_1 + B_2 X_i + u \qquad \text{(Eq.2.1)}$$

onde

Y_i variável dependente do modelo, Y = 1 se o evento ocorrer, Y = 0 se o evento não ocorrer.

B_1 é o termo constante

B_2coefficiente(s) da variável(s) independente

X_i é a variável(s) independente

U é o termo do erro.

$E(Y_i / X_i)$ é a esperança condicional de Y_i dado X_i, que pode ser interpretada como a probabilidade condicional de que o evento ocorrerá dado X_i, ou seja, $P(Y_i=1/X_i)$.

Admitindo que $E(u_i) = 0$, obtém-se

$$E(Y_i / X_i) = B_1 + B_2 X_i \qquad \text{(eq.2.2)}$$

Indicando por P_i = probabilidade de que $Y_i = 1$ (ou seja, de que o evento ocorra) e $1 - P_i$ = probabilidade de que $Y_i = 0$ (ou seja, de que o evento não ocorra), a variável Y_i tem a seguinte distribuição:

Y_i	Probabilidade
0	$1 - P_i$
1	P_i
Total	1

Portanto, pela definição de esperança matemática, obtém-se

$$E(Yi) = 0\ (1 - P_i) + 1\ (P_i) = P_i \qquad \text{(eq. 2.3)}$$

Comparando (eq.2.2) com (eq. 2.3), pode-se equacionar

$$E(Y_i / X_i) = B_1 + B_2 X_i = P_i \qquad \text{(eq. 2.4)}$$

ou seja, a esperança condicional do modelo (eq. 2.1) pode ser interpretada como a probabilidade condicional de Y_i.

Como a probabilidade P_i deve se situar entre 0 e 1, tem-se a restrição

$$0 \leq E(Y_i / X_i) \leq 1 \qquad \text{(eq. 2.5)}$$

ou seja, a esperança condicional (ou probabilidade condicional) assume valores entre 0 e 1.

As probabilidades preditas do modelo são geralmente onde ocorrem os problemas. Há três problemas ao usar o MPL que afetam as probabilidades preditas: (1) Não normalidade dos residuos; (2) variância heteroscedástica dos u_i; (3) Não satisfação de $0 \leq E(Y_i / X) \leq 1$

Quanto ao item (1), não-normalidade dos resíduos u_i:

P_i faz exame em somente dois valores. Como $u_i = Y_i - B_1 - B_2X_i$, então, quando

$$Y_i = 1 \qquad u_i = 1 - B_1 - B_2X_i$$

e quando

$$Y_i = 0\ , \qquad u_i = -B_1 - B_2X_i \qquad \text{(eq.2.6)}$$

Desta forma, não se pode supor que u_i seja distribuído normalmente, pois, na verdade, segue a distribuição binomial. Porém, o não atendimento da hipótese de normalidade pode não ser tão crítico como parece. Conforme aumenta, indefinidamente, o tamanho da amostra, pode-se mostrar que os estimadores de Mínimos Quadrados Ordinais (MQO) tendem, em geral, a se distribuir normalmente.

A prova baseia-se no teorema do limite central e pode ser encontrada em Malinvaud, E., citado em 1966. Portanto, em amostras grandes, a inferência estatística do MPL seguirá o método usual dos MQO sob a hipótese de normalidade.

Quanto ao item (2), variâncias Heterocedásticas dos u_i:

Os u_is da (Eqs.2.2.6) têm a seguinte distribuição de probabilidade:

Y_i	U_i	Probabilidade
0	$-B_1 - B_2X_i$	$1 - P_i$
1	$1 - B_1 - B_2X_i$	P_i
Total		1

Por definição

$$Var(u_i) = E[u_i - E(u_i)]^2 = E(u_i^2) \text{ para } E(u_i) = 0 \text{ por hipótese.}$$

Assim, usando a distribuição de probabilidade de u_i anterior, obtém-se

$$Var(u_i) = E(u_i^2) = (-B_1 - B_2X_i)^2(1 - P_i) + (1 - B_1 - B_2X_i)^2(P_i)$$

$$= (-B_1 - B_2X_i)^2(1 - B_1 - B_2X_i) + (1 - B_1 - B_2X_i)^2(B_1 + B_2X_i) \quad \text{(eq.2.7)}$$

$$= (B_1 + B_2X_i)(1 - B_1 - B_2X_i)$$

ou

$$Var(u_i) = E(Y_i/X_i)[1-E(Y_i/X_i)]$$
$$= P_i(1-P_i) \qquad \text{(eq.2.8)}$$

Na expressão (eq.2.8) $E(Y_i/X_i) = B_1 + B_2X_i = P_i$. A (Eq.2.8) mostra que a variância de u_i é heteroscedástica, pois depende da esperança condicional de Y, que, naturalmente, depende do valor assumido por X. Portanto, a variância de u_i depende de X e, deste modo não é homoscedástica.

Entretanto, o problema da heteroscedasticidade não é insuperável, um meio de resolvê-lo, segundo Gujarati (1995) pois se pode transformar os dados dividindo ambos os lados do modelo (Eq.2.2.1) por

$$\sqrt{E(Y_i/X_i)[1-E(Y_i/X_i)]} = \sqrt{P_i(1-P)_i} = \sqrt{w_i}$$

então,

$$\frac{Y_i}{\sqrt{w_i}} = \frac{B_i}{\sqrt{w_i}} + B_2\frac{X_i}{\sqrt{w_i}} + \frac{u_i}{\sqrt{w_i}} \qquad \text{(eq.2.9)}$$

O termo de erro em (Eq.2.9) será agora homocedático. Portanto, pode-se passar para a estimativa por MQO.

Em relação à (3) não satisfação de $0 \le E(Y_i/X) \le 1$*:*

Uma vez que, o modelo de probabilidade linear, $E(Y_i/X)$ mede a probabilidade condicional de o evento Y ocorrer, dado X, ela deve, necessariamente, se situar entre 0 e 1. Embora isto seja verdadeiro a priori, não há garantia alguma de que $\hat{Y}_i$, os estimadores de $E(Y_i/X)$ irão atender a essa restrição. Este é o problema real com a estimativa de MQO do MPL.

Como discutiu-se, o uso do MPL para dados categóricos qualitativos apresenta diversos problemas, apesar de alguns serem superáveis. Porém, o problema fundamental do MPL é que ele supõe que $P_i = E(Y = 1/X)$ aumenta linearmente com X, quando na verdade, o efeito marginal de X permanece constante do início ao fim.

Desta forma, precisa-se de um modelo (de probabilidade) que tenha as seguintes características: (1) conforme X_i aumenta, $P_i = E(Y = 1/X)$ também aumenta, porém, nunca sai fora do intervalo entre 0 e 1; (2) a relação entre Pi e X_i deve ser não linear, ou seja, aproxima-se de zero mais lentamente conforme X_i fica menor e aproxima-se de 1 mais lentamente conforme X_i fica maior.

Na verdade, o modelo que se deseja parece com a curva sigmóide ou em forma de "S", que se assemelha muito com a Função Distribuição Acumulada (f.d.a.) de uma variável aleatória. Desta forma, pode-se facilmente utilizar a f.d.a. para modelar regressões em que a variável de resposta seja dicotômica, assumindo valores 0-1. Embora todas as f.d.a.s tenham a forma de "S", para cada variável aleatória há uma única f.d.a.. Por razões históricas e práticas, as f.d.a.s geralmente escolhidas para representar os modelos de resposta entre 0 e 1 são: (1) a Logística e (2) a Normal. A Logística dá origem ao modelo Logit e a Normal dá origem ao modelo Probit. Tanto o modelo Logit quanto o modelo Probit asseguram que as probabilidades estimadas se situem no limite 0-1 e que elas se relacionem não linearmente com as variáveis explanatórias.

2.1.1 O MODELO DE REGRESSÃO LOGÍSTICA

Regressão logística trata de uma técnica estatística que procura descrever as relações entre uma variável categórica dependente e uma ou mais variáveis explicativas (Everitt 1992).

O modelo de regressão logística binária trabalha com a variável resposta Y dicotômica (codificada 0;1), sendo que as variáveis independentes $X_i^{'s}$ possuem a liberdade de serem, qualitativas, quantitativas uma mistura de ambas. Quando a variável resposta possui mais do que duas categorias, (Tabachnick, 1996) a análise é chamada de multinomial, ou regressão logística politômica. Neste caso, haverá mais do que um modelo de equação de regressão. Então, a variável resposta com mais do que duas categorias poderá ser classificada em nominal (sem ordem), ou ordinal (com ordem).

A regressão logística permite prever a razão de chances, ou a relação, entre uma variável dependente dicotômica e um grupo de variáveis independentes, descrevendo tal relação. Pode-se prever as probabilidades de um evento Y=1, presença de alguma característica, contra Y=0, ausência de característica, representando a proporção de vezes que Y assume o valor 1 dentro das categorias da variável independente X.

Situações que envolvem variável de resposta dicotômica são comuns em aplicações na área médica. A variável resposta pode ser um paciente que sobrevive ou não cinco anos após o tratamento para o câncer, ou ainda, um paciente tem uma reação adversa a uma certa droga. Enfim, há o interesse em encontrar uma combinação apropriada do relacionamento de variáveis independentes para explicar o resultado binário em questão.

O modelo de regressão logística é usado quando se espera que a distribuição de respostas da variável dependente Y seja não linear com uma ou mais variáveis independentes X. Devido ao modelo produzido pela regressão logística ser não linear, a equação usada para descrever as respostas é um pouco mais complexa (Everitti 1992).

Considera-se um conjunto de K variáveis independentes, denotado pelo vetor $X' = (x_1, x_2, \ldots, x_K)$, onde a probabilidade condicional da variável resposta Y é denotada por $P(Y = 1/X) = \pi(x)$.

Então, o Logit do modelo de regressão é dado pela equação (Hosmer, e Lemeshow, 1989)

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots . \beta_K X_K \qquad \text{(eq.2.10)}$$

Em cada caso

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad , \qquad \text{(eq 2.11)}$$

que é chamada de função de regressão logística.

O modelo descreve como a proporção de sucessos (Y=1) é influenciada pelas variáveis explicativas. A proporção esperada de sucessos (Y=1) é denotada por $\pi = E(Y)$. π também representa a probabilidade que um indivíduo, aleatoriamente, escolhido possua a característica de interesse, baseado nas variáveis explicativas do modelo.

Por razões de simetria, a regressão logística expressa os coeficientes como Logits. Um Logit é o logarítimo natural de uma relação de probabilidade e contém exatamente a mesma informação que as relações das probabilidades. Logits são medidas da força do relacionamento entre as variáveis e, como são simétricos podem ser comparados.

A equação de regressão linear fornece o Logit ou o log de chances (TabachnicK, 1996):

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)=g(x) \qquad \text{(eq.2.12)}$$

Isto é, o modelo de regressão logística é o logaritmo natural da probabilidade de ocorrência num grupo, dividido pela probabilidade de ocorrência no outro grupo.

Conforme a (eq.2.12) o quociente $\frac{\pi}{1-\pi}$ é denominado odds, que representa quantas vezes o sucesso é mais provável que o fracasso.

Teste dos coeficientes e ajuste do modelo

Para testar hipóteses que lidam com modelos de regressão não linear, pode-se utilizar o Teste da Razão de Verossimilhança e o Teste Wald. Ambos são comparados à Distribuição Qui-Quadrado com graus de liberdade dependendo dos níveis de cada variável.

Os coeficientes da regressão logística e seus erros padrão envolvem cálculos, nos quais os valores são representados usando o método de máxima verossimilhança. Estes valores são usados para avaliar o ajuste de um ou mais modelos.

Teste de Wald

O teste de Wald é obtido pela comparação da máxima verossimilhança estimada do parâmetro β_i, dividido pelo erro padrão.

$$W = \frac{\hat{B}_1}{\hat{SE}(\hat{B}_1)} \qquad \text{(eq.2.13)}$$

O resultado desta razão é a estatística Z, representando a distribuição normal padrão, sob a hipótese nula $H_0 : \beta_i = 0$, ou seja, X não tem efeito na probabilidade π de que Y=1, sendo Y independente de X.

Deve-se ressaltar que o teste de Wald é uma aproximação, sendo mais apropriado para amostras grandes.

Diversos autores identificaram problemas com o uso da estatística de Wald. (Menard,1995) adverte que para coeficientes grandes, o erro padrão aumenta, reduzindo o valor da estatística de Wald (Qui-quadrado). Segundo (Agresti,1996), indica que o teste da Razão de Verossimilhança é mais confiável para tamanhos de amostra pequenos do que o teste de Wald.

Teste da Razão de Verossimilhança

O Teste da Razão de Verossimilhança compara a razão de verossimilhança de dois modelos. Utiliza a relação do valor da função de máxima verossimilhança para o modelo saturado e o valor da função de máxima verossimilhança para o modelo reduzido (corrente).

Através do teste, verifica-se o ajuste do modelo, onde o log da verossimilhança é calculado, baseado na soma das probabilidades associadas com o valor predito de cada caso da variável resposta observado. A equação é representada por:

$$\log - \textit{verossimilhança} = \sum_{i=1}^{N} \left[Y_i \ln(\hat{Y}_i) + (1 - Y_i) \ln(1 - \hat{Y}_i) \right] = D \qquad \text{(eq.2.14)}$$

onde $\hat{Y}_i$, i=1,0, é o valor da probabilidade predita de Y_i, para cada caso, Y=1 e Y=0.

Para avaliar a significância das variáveis independentes compara-se a estatística D com e sem as variáveis independentes e, pode-se escrever:

G = D (modelo reduzido) - D(modelo saturado) (eq.2.15)

O primeiro modelo, chamado reduzido, contém apenas a constante do modelo, assumindo que os β_i=0. O segundo modelo, chamado de saturado, contém todas as variáveis independentes do modelo.

A (eq.2.15) também pode ser expressada, segundo Hosmer & Lemeshow (1989), por:

$$G = -2\ln\left\{\frac{verossimilhança\ do\ \mathrm{mod}elo\ atual}{verossimilhança\ do\ \mathrm{mod}elo\ saturado}\right\} \qquad \text{(eq.2.16)}$$

Interpretação dos Coeficientes do Modelo de Regressão Logística

A interpretação dos coeficientes envolve dois aspectos: (1) Determinação da função de relação entre a variável dependente e a variável independente. (2) Mudança no logaritmo do odds em relação à alteração de uma unidade na variável independente, quando esta é quantitativa (Hosmer & Lemeshow ,1989).

O modelo logístico pode ser escrito em termos da chance de um evento ocorrer. A razão de chance de um evento ocorrer é definido como a razão da probabilidade do evento ocorrer pela probabilidade do evento não ocorrer.

O Logit, como é chamado, pode ser escrito em termos do log de odds (chances), no modelo logístico:

$$\ln\left(\frac{\Pr ob(evento)}{\Pr ob(não\ \ evento)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k \qquad \text{(eq.2.17)}$$

Pela (eq.2.17) é mais fácil pensar na razão de chances do que no log de chances e, desta forma, a equação logística pode ser escrita em termos de odds como:

$$\ln\left(\frac{\Pr ob(evento)}{\Pr ob(não\ \ evento)}\right) = e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k} \ldots e^{\beta_k X_k} \qquad \text{(eq.2.18)}$$

Se B_i é positivo, significa que a probablidade de ocorrência do evento aumenta, quando a variável independente X aumenta. Se B_i é negativo, a probabilidade de ocorrência do evento diminui, quando a variável independente X diminui (Echeveste e Nodari, 2000)

Limitações da Análise de Regressão Logística

Segundo (Tabachnick e Fidell1996), a regressão logística é relativamente livre de restrições, com capacidade para analisar uma mistura de todos os tipos de variáveis (contínuas, discretas e dicotômicas). A variedade e complexidade que os dados podem ser analisados é bastante ilimitada.

2.1.2 O MODELO LOGIT

Na regressão logística, quando todas as variáveis explicativas são categóricas e a variável resposta Y é dicotômica, chama-se modelo Logit. O modelo logit se diferencia do modelo loglinear pelo fato de não assumir interação entre as variáveis explicativas com a variável resposta, ou seja, assume que Y é associado com cada X_1, X_2, X_3, variáveis explicaticas mas os efeitos médios de cada variável explicativa em Y são os mesmos para cada combinação de níveis das outras variáveis. Portanto, o modelo logit não assume a interação (YX_1X_2) nem a interação ($YX_1X_2X_3$), que no caso de modelo loglinear estas interações são consideradas.

O modelo Logit tem resultados idênticos aos da regressão logística. e a variável resposta pode ter mais do que duas categorias. Neste caso, o modelo logit é uma generalização de casos para variável resposta binária.

Em variáveis categóricas, em que a variável resposta Y possui apenas duas categorias, representadas por 1 e 0, a distribuição de Bernoulli para variáveis aleatórias binárias especifica a probabilidade $P(Y=1)=\pi$ e $P(Y=0)=1-\pi$.

Sendo $\pi = E(Y)$, quando Y_i tem distribuição Bernoulli com parâmetro π_i, a função massa de probabilidade é dada por (Agresti, 1990)

$$f(y_i;\pi_i)=\pi_i^{y_i}(1-\pi_i)^{1-y_i}=(1-\pi_i)[\pi_i/(1-\pi_i)]^{y_i}$$

$$= \quad (1-\pi_i)\exp\left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right)\right] \qquad \text{(eq.2.19)}$$

para $y_i = 0$ e 1. Esta é a distribuição da família exponencial.

O parâmetro $L(\pi)=\log[\pi/(1-\pi)]$ é o log odds da resposta 1, chamado de Logit de π.

O modelo Logit, na verdade é originado da função distribuição logística acumulada e é representado por

$$L(\pi) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 X_i \qquad \text{(eq.2.20)}$$

ou seja, $L(\pi)$ é o log da razão de probabilidades.

Também pode-se escrever

$$1-\pi_i = \frac{1}{1+e^{g(x)}} \qquad \text{(eq.2.21)}$$

,

Portanto,

$$\frac{\pi_i}{1-\pi_i} = \frac{1+e^{g(x)}}{1+e^{-g(x)}} = e^{g(x)} \qquad \text{(eq.2.22)}$$

Assim, $\pi_i /(1-\pi_i)$ é a razão de probabilidades de um evento ocorrer, Y = 1, e de um evento não ocorrer, Y = 0.

π_i é a probabilidade do evento ocorrer

1- π_i é a probabilidade do evento não ocorrer

Supondo um modelo com dois fatores, ou seja, duas variáveis independentes A e B, para a variável resposta Y dicotômica, denotado por I o número de categorias de A e J o número de categorias de B. Então, $\pi_{k/ij}$ é a probabilidade de resposta k, quando a variável A corresponde à categoria i e a variável B corresponde à categoria j.

Desta forma, $\pi_{1/ij} + \pi_{2/ij} = 1$ e o modelo Logit é dado por

$$\log\left(\frac{\pi_{1/ij}}{\pi_{2/ij}}\right) = \alpha + B_i^A + B_j^B \qquad \text{(eq.2.23)}$$

onde

α é a média dos Logits

B_i^A é o efeito da variável A através das categorias I na variável resposta Y.

B_j^B é o efeito da variável B através da categoria J na variável resposta Y.

Interpretação dos coeficientes do modelo Logit

Na interpretação do modelo Logit, β_2 é a inclinação e mede a variação em $L(\pi)$ para uma mudança unitária em X, ou seja, a chance em log em favor de Y = 1, varia conforme X varia em uma unidade. O intercepto β_1 é o valor da chance em log em favor de Y = 1 se X for igual a zero. Como uma interpretação de um intercepto, esta pode não ter significado físico.

Os Logits contém as mesmas informações que as relações de probabilidades, ou seja, eles medem a força do relacionamento entre as variáveis.

Um Logit positivo significa que a variável independente X tem o efeito de aumentar as probabilidades da variável dependente Y, quando esta assume o valor 1 ou 0, sendo que um Logit negativo significa que a variável independente tem o efeito de diminuir as probabilidades da variável dependente Y, quando esta assume o valor 1 ou 0.

Segundo (Gujarati 1995), O modelo Logit supõe que o log da razão de probabilidades se relaciona linearmente com X_i e com os parâmetros. Conforme π vai de 0 a 1, o Logit L vai de $-\infty$ a $+\infty$. Portanto, as probabilidades se situam entre 0 e 1 e os Logits não se restringem a esses limites.

Estimativa do Modelo Logit

A função resposta individual para cada para cada combinação entre as variáveis explicativas pode ser definido como:

$$L_i = F(\beta x_i) = \frac{1}{1+e^{-\beta x_j}} \qquad \text{(eq.2.24)}$$

onde

L_i é a verdadeira probabilidade para a variável resposta Y = 1 ou Y = 0. Esta relação pode ser usada para definir a função de verossimilhança (L) das observações.

$$L(y_1,\ldots,y_n;x_1,\ldots,x_n) = \prod_{i-1}^{n} F(\beta x_i)^{y_i}\left[1-F(\beta x_i)\right]^{(1-y_i)}$$

onde y_i = 0, 1, é a variável resposta

x_i = vetor de observações das variáveis independentes.

Teste de Hipótese e Ajuste do Modelo

No ajuste de modelos Logit, normalmente condiciona-se a contagem de célula na marginal da tabela, na forma de combinação de categorias das variáveis independentes. Trata-se a marginal A-B-C da tabela como fixa e assume-se quatro amostras binomiais independentes para a variável resposta Y. Isto, considerando um modelo com a variável resposta dicotômica e três variáveis independentes (Andersen, 1997).

Quadro 2. Representação de tabela cruzada com 3 variáveis independentes categóricas, cada uma com 2 níveis e uma variável resposta dicotômica.

Variáveis independentes			Variável Resposta Y		Totais Marginais
A	B	C	1	0	
1	1	1	f_{1111}	f_{1110}	$n_{111/1}$
	2	2	f_{1221}	f_{1220}	$n_{122/0}$
2	1	1	f_{2111}	f_{2110}	$n_{211/1}$
	2	2	f_{2221}	f_{2220}	$n_{222/0}$

Considerando o Quadro 2, com três variáveis independentes, A, B, e C, descrevendo a influência na variável resposta Y. O modelo Logit saturado assume a forma

$$\log it(\pi_{1/ijk}) = \alpha + \beta_i^A + \beta_j^B + \beta_k^C + \beta_{ij}^{AB} + \beta_{ik}^{AC} + \beta_{jk}^{BC} + \beta_{ijk}^{ABC} \qquad \text{(eq.2.25)}$$

Este modelo pode ser simplificado se a soma das interações, envolvendo mais do que uma variável independente, puder ser retirada do ,odelo. Neste caso, a hipótese de interesse é

$$H_0 : \beta_{ij}^{AB} = \beta_{ik}^{AC} = \beta_{jk}^{BC} = \beta_{ijk}^{ABC} = 0 \qquad \text{para todo i,j e k.} \qquad \text{(eq.2.26)}$$

Se H_0 é aceito, o modelo Logit é reduzido para

$$\log it(\pi_{1/ijk}) = \alpha + \beta_i^A + \beta_j^B + \beta_k^C \qquad \text{(eq.2.27)}$$

Se o efeito da variável C é nulo, $\beta_k^C = 0$, esta não contribui para descrever a variação na variável resposta Y. Assim, dado o modelo da (eq.2.25), pode-se evoluir a contribuição de cada variável independente através da comparação de três hipóteses paralelas

$$H_1 : \beta_k^C = 0, \quad k = 1, \ldots, K$$
$$H_2 : \beta_j^B = 0, \quad j = 1, \ldots, J$$
$$H_3 : \beta_i^A = 0, \quad i = 1, \ldots, I$$

Usando seus níveis de significância no teste contra H_0, compara-se

$$H_{(2)} : \beta_j^B = \beta_k^C = 0$$
$$H_{(3)} : \beta_i^A = \beta_j^B = \beta_k^C$$

Modelo Logit quando a variável resposta é ordinal

Quando a variável resposta é ordinal com mais do que duas categorias (c≥2), existe muitas maneiras de formar Logits. Para uma variável resposta politômica, com probabilidades (πi,..., πc), a combinação de níveis da variável explanatória pode ser representada da forma condicional

$$\ln\left[\frac{(\pi_j/(\pi_j+\pi_k))}{(\pi_k/(\pi_j+\pi_k))}\right] = \ln\left(\frac{\pi_j}{\pi_k}\right) \qquad \text{(eq.2.28)}$$

Este é o log de odds da classificação na categoria j comparado com a categoria k, dado uma observação abaixo das categorias básicas. O Logit pode ser derivado em vários tamanhos básicos, c - 1 .

Se

$$L_j = \ln(\pi_j / \pi_c) \qquad j = 1, \ldots, c-1, \qquad \text{(eq.2.29)}$$

Então,

$$\ln(\pi_j / \pi_k) = L_j - L_k \qquad \text{para } 1 \le j \le k \le c-1. \qquad \text{(eq.2.30)}$$

A formação dos Logits, quando a variável resposta é ordinal, é feita usando apenas duas categorias como foi feita na (eq.2.28). Os Logits são formados por grupos de categorias que são próximas dentro de uma escala ordinal, ou seja, como se fossem classificados ou categorizados perante à ordem. Dentro dessa categorização ordenada, desenvolveu-se três tipos de Logits: (1) Logits cumulativo; (2) Logit da razão contínua e (3) Logit de categoria-adjacente. Cabe ressaltar que, quando c = 2, esses três tipos de Logits são simplificados para o "Logit padrão" ($\pi 1/\pi 2$), (Agresti, 1984).

Neste estudo, não será desenvolvido estes modelos Logits, já que fogem do objetivo principal deste trabalho, pois são classificações de modelos para um caso específico de reclassificação categórica ordenada da variável resposta. Entretanto, dentro da idéia em que a variável resposta é ordinal, será discutido o modelo simples de Logit, para os casos em que a variável explanatória é ordinal ou nominal, fazendo uso do modelo Logit cumulativo.

Conforme foi desenvolvido acima, para cada tipo de Logit aplicado a c categorias da variável resposta, pode-se formar c-1 Logits e, desta forma, pode-se fazer análise incorporando c-1 Logits dentro de um modelo simples.

Modelo Logit para variável resposta e explanatória ordinal

Supondo a variável resposta Y e variável explanatória X, ambas ordinais, o modelo Logit padrão para a variável resposta dicotômica, é representado por (Agresti, 1984):

$$L_{j(i)} = \alpha_j + \beta_j(u_i - \overline{u}), \qquad 1 \le i \le r, \qquad 1 \le j \le c-1 \qquad \text{(eq.2.31)}$$

A variável resposta Y , sendo composta de duas categorias, considera a primeira categoria representada por j e a segunda por c-j. Usou-se c-j para generalizar os casos em que Y possui mais do duas categorias.

O modelo da (eq.2.31) supõe $\beta_1 = \ldots = \beta_{c-1}$. Para c-1 Logits em cada linha, tem-se um total de r(c-1) Logits. A interpretação dos parâmetros do modelo é simples, $\sum_i L_{j(i)} = r\alpha_j$, α_j é a média dos r Logits que são formados quando o ponto de corte representa a categoria j. Para cada i, $L_{1(i)} \ge L_{2(i)} \ge \ldots \ge L_{c-1,(i)}$, sendo que α_j decresce a cada linha. Cada c-1 Logits são linearmente relacionados com X, com β assumindo a mesma inclinação em todos os Logits. Se $\beta = 0$, então, para todo j, o Logit é o mesmo em cada linha, implica que X e Y são independentes. Se $\beta > 0$, o Logit aumenta como X aumenta. Isso implica que a distribuições condicionais de Y são maiores para valores maiores de X.

O parâmetro β também pode ser interpretado através do log da razão de chances, $L_{j(b)} - L_{j(a)}$, neste caso, a e b, tem-se apenas duas linhas para a variável dicotômica e o ponto de corte representando a categoria j. Para o modelo (eq.2.31) ,

$$L_{j(b)} - L_{j(a)} = \beta(u_b - u_a)$$

A estimação do modelo pode ser solucionada através da máxima verossimilhança e a estimativa do Logit $\hat{L}_{j(i)} = \hat{\alpha}_j + \hat{\beta}_j(u_i - \bar{u})$, produz a estimativa da probabilidade cumulativa $\hat{F}_{j(i)} = \left[\exp(\hat{L}_{j(i)} + 1\right]^{-1}$.

Então, a estimativa da probabilidade condicional na linha é $\hat{F}_{j(i)} - \hat{F}_{j-1,(i)}$, j = 1,..., c, onde $\hat{F}_{0(i)} = 0$

Modelo Logit para variável resposta ordinal e explanatória nominal

Segundo Agresti, 1984 o modelo envolvendo a variável resposta ordinal e a variável explanatória nominal, a associação de termos que são adicionados para o modelo básico independente, leva a forma de efeito de linha. Considerando o modelo cumulativo tendo efeito de linha, sendo cada linha identificada por c-1 Logits. O modelo independente é expresso por

$$L_{j(i)=\alpha_j}, \quad 1 \le i \le r, \quad 1 \le j \le c-1$$

Adicionando os efeitos de linha para os níveis da variável nominal, tem-se

$$L_{j(i)} = \alpha_j + \tau_i, \quad 1 \le i \le r, \quad 1 \le j \le c-1 \ , \quad \text{onde } \sum \tau_i = 0.$$

Da mesma forma que no modelo da (eq.2.29) α_j representa a média através das r linhas dos valores do Logit cumulativo. O efeito de linha nos parâmetros especifica a natureza da associação. A diferença nos Logits, para cada par de linhas a e b é $L_{j(b)} - L_{j(a)} = \tau_b - \tau_a$, sendo constante para todos c-1 Logits. Se $\tau_b > \tau_a$, então a distribuição condicional é estocasticamente maior na linha b do que na linha a.

O teste de independência condicional para o modelo $L_{j(i)} = \alpha_j + \tau_i$ em que $\tau_1 = \ldots = \tau_r = 0$, é G2(I/R) = G2(I) - G2(R) , onde(I) é a estimativa de freqüência esperada para o modelo Independente e (R) é o Logit cumulativo do efeito de linha do modelo. Este teste tem assintoticamente uma distribuição qui-quadrado sob H_0 com r-1 graus de liberdade.

2.1.3 O MODELO PROBIT

Segundo a citação de Goldberger (1964) na análise de modelo Probit é assumido que há uma variável resposta subjacente, Y_i^*, que na prática é não observável, sendo o que se observa é a variável dicotômica Y_i. Esta idéia, será discutida através de um exemplo de (McFadden, 1973), onde o modelo Probit baseia-se na teoria da utilidade, ou na perspectiva de escolha racional no comportamento. Considerando um exemplo para desenvolver a idéia do modelo Probit, suponha que a decisão da i-ésima família de possuir um carro ("sucesso") ou não ("fracasso") depende de um índice de utilidade não observável I_i. Este índice é determinado por uma ou mais variáveis explicativas ou independentes, como por exemplo, a renda X_i, de tal modo que, quanto maior o valor do índice I_i, maior a probabilidade de a família possuir um carro. Expressa-se I_i como

$$I_i = \beta_1 + \beta_2 X_i \qquad \text{(eq.2.32)}$$

Seja Y = 1 se a família possui um carro e Y = 0 caso contrário. Supondo que para cada família exista um nível crítico ou limiar do índice, I_i^*, de forma que, se Ii superar I_i^*, a família possuirá um carro, caso contrário, não possuirá. O I_i^* não é observável e nem I_i, porém, admitindo que eles se distribuem normalmente com a mesma média e variância, é possível estimar os parâmetros do índice dado em (eq.2.32) e obter alguma informação sobre o índice não observável (Gujarati, 1995).

Pela hipótese de normalidade, a probabilidade de I_i^* ser menor ou igual a I_i, pode ser calculada da f.d.a.normal padronizada

$$P_i = \Pr(Y = 1) = \Pr(I_i^* \leq I_i) = F(I_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{T_i} e^{-t^2/2} dt$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_i + \beta_2 X_i} e^{-t^2/2} dt \qquad \text{(eq.2.33)}$$

onde t é uma variável normal padronizada, ou seja, $t \sim N(0,1)$. P_i representa a probabilidade de ocorrer um evento e, pelo exemplo, a probabilidade de uma família possuir um carro é medida pela área da curva normal padrão de $-\infty$ a I_i.

Para obter informação sobre I_i, o índice de utilidade , e de β_1 e β_2, pega-se o inverso da (eq.2.33) e obtêm-se

$$I_i = F^{-1}(I_i) = F^{-1}(P_i) = \beta_1 + \beta_2 X_i \qquad \text{(eq.2.34)}$$

onde F^{-1} é o inverso da f.d.a. normal.

A análise Probit também é usada para analisar dados tipo "dose-resposta " em aplicações biomédicas. Suponha que se deseja saber qual a dosagem média requisitada de determinado medicamento para prevenir ataques do coração na metade de uma população de vítimas. Assim, foca-se a proporção de casos em duas ou mais categorias da variável dependente, onde esta é predita por muitas variáveis que são contínuas ou codificadas para tornarem-se dicotômicas.

A análise Probit usa a transformação Probit quando cada proporção observada é substituída pelo valor da curva normal padronizada (valor Z), pelo qual a proporção observada é estabelecida.

Interpretação dos Coeficientes do Modelo Probit

Os coeficientes no modelo Probit correspondem aos coeficientes em regressão logística e aos coeficientes do modelo Logit. Os coeficientes diferem na magnitude. O coeficiente é quanto de diferença faz na mudança em uma unidade nos termos da probabilidade normal cumulativa da variável resposta. A média do coeficiente Probit mede o efeito da mudança em uma unidade na probabilidade da variável resposta. Esta média depende da categoria da variável independente. Portanto, para calcular o efeito dos coeficiente do modelo Probit, é necessário escolher alguma categoria da variável independente como ponto de referência e em particular, o padrão de referência é quando todas as variáveis independentes são suas médias amostrais.

Estimativa do Modelo Probit

Considerando que se tem dados somente da variável independente X_i e da situação da variável resposta Y =1 ou Y = 0, para ajustar o modelo Probit, obter o índice I_i e para estimar β_1 e β_2, procede-se seguindo os passos:

1. A partir dos dados agrupados, estimar P_i como no caso do modelo Logit.
2. Dado Pi obter I_i da f.d.a. normal padrão.
3. Usar o estimado $I_i = \hat{I}_i$ obtido no passo 2.
4. Adicionar 5 aos I_i estimados para convertê-los em Probits, o seja, Probit = I_i +5 e usar os Probits obtidos como as variáveis dependentes na regressão $I_i = \beta_1 + \beta_2 X_i + u_i$.

Obs.: Adiciona-se 5 aos I_i porque I_i será negativo quando $P_i < 0,5$.

Na linguagem da análise de Probit, o índice de utilidade não observável I_i é conhecido como desvio equivalente normal (d. e. n.), ou normit.

Para resumir os modelos abordados nesta monografia, construiu-se um quadro, Quadro Geral, referente às características de cada modelo de Regressão Logística, Logit e Probit.

Quadro Geral das Características Relacionadas a Cada Modelo de Regressão Logística, Logit e Probit

Modelo	*Descrição*	*Representação*	*Estimativa Modelo*	*Teste dos Coeficientes e Hipóteses nula*	*Ajuste do Modelo*
Logística	A variável resposta é binária e as variáveis independentes são qualitativas quantitativas uma mistas de ambas.	$\pi(x)=\frac{e^{g(x)}}{1+e^{g(x)}}$	Máxima Verossimilhança	Teste da Razão de Verossimilhança ou Teste de Wald H_0: $\beta_i = 0$	Através da tabela classificatória
Logit	A variável resposta pode ser, dicotômica ou politômica nominais ou ordinais e as variáveis independentes são qualitativas.	Resposta dicotômica $\log\left(\frac{\pi_{1/ij}}{\pi_{2/ij}}\right)=\alpha+B_i^A+B_j^B$ Resposta ordinal $L_{j(i)}=\alpha_j+\beta_j(u_i-\bar{u}),$	Máxima verossimilhança	Teste da Razão de Verossimilhança para cada variável explicativa $H_0 : \beta_1^X = \ldots = \beta_i^X$	Comparação entre o modelo reduzido e o modelo saturado
Probit	A variável resposta é medida em doses, geralmente é uma resposta subjacente não observável, as variáveis independentes são contínuas ou discretas.	$\Pr(Y=k)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\beta_i+\beta_2 X_i} e^{-t^2/2}dt$	Máxima Verossimilhança	$I_i = \beta_1 + \beta_2 X_i + u_i .$	Através da freqüência relativa (medida empírica da probabilidade)

2.2 PROPOSTA DE FLUXOGRAMA

A discussão desenvolvida no Capítulo 2 apresentou características peculiares de cada um dos modelos de regressão logística, logit e probit. Baseado na teoria de diversos autores, realizou-se um apanhado geral de cada modelo, com o intuito de diferenciar a aplicação destes, a medida em que as variáveis vão se inserindo no processo de pesquisa, conforme sua classificação, como qualitativa ou quantitativa.

O sucesso de qualquer análise estatística depende, muitas vezes, da disponibilidade do modelo estatístico se incorporar "corretamente" aos dados, de acordo com os tipos de variáveis em estudo. No caso de um ou mais modelos se ajustarem aos dados, surge o problema da escolha do "melhor". Esta escolha pode estar fundamentada em estudos teóricos.

O propósito do fluxograma construído é elucidar a escolha de modelos de regressão, considerando o modelo logístico, logit e probit, em relação às variáveis dependentes e independentes.

A construção do fluxograma apresentado na Figura 2.1 foi fundamentada através do estudo de cada modelo, associado com as características das variáveis dependentes e independentes que entram no processo da análise estatística. A caracterização das variáveis (resposta/explicativa) em qualitativa ou quantitativa foi vital na desenvoltura do processo de construção do mapeamento, conduzindo ao caminho que leva ao encontro dos modelos Logístico, logit ou probit. O caminho percorrido ao destino do modelo é relacionado com o percurso dado pela conjunção das variáveis resposta e explicativas.

Ao final da classificação das variáveis resposta e explicativas, é o momento de escolher qual o modelo apropriado. Como este direcionamento está relacionado com ambas variáveis resposta e explicativa, usou-se três caminhos apresentados na Figura 2.1, em forma de (linhas) de diferentes formatos, "linha contínua", "linha pontilhada" e "linha alternada contínua pontilhada", representando a conjunção das variáveis que levam ao mesmo modelo.

Um exemplo de utilização do modelo segue com os passos utilizados para uso do fluxograma: suponha um estudo onde a variável resposta é Situação de uma Família, codificada em *Y = 1, se a família possui uma casa; Y = 0, se a família não possui uma casa.* A variável explicativa é Renda Familiar.

Iniciando a classificação pela variável resposta, seguem-se os passos:

1. **A variável é Qualitativa ou Quantitativa ?**

   A variável resposta "Situação de uma Família" é Qualitativa.
2. **A variável resposta é Dicotômica ou Politômica ?**

   Tem-se uma variável codificada (0,1), Dicotômica.
3. **É nominal ou ordinal?**

   Neste exemplo a variável resposta é nominal.
4. **Qual a classificação da variável explicativa?**

   Neste passo, deve-se fazer a classificação da variável explicativa e esperar pelo caminho que conduzirá até o modelo, de forma que seja o mesmo (relacionado com a linha) da variável resposta.
   Continuando a classificação, neste exemplo para a variável explicativa:
5. **A variável explicativa é Qualitativa ou Quantitativa ?**

   A variável Renda Familiar é uma variável Quantitativa.
6. **A variável explicativa é Discreta ou Contínua?**

   A variável Renda Familiar é uma variável Contínua
7. **Qual o modelo proposto?.**

   Escolha do modelo.
   O modelo escolhido para este caso, foi o modelo Logístico.

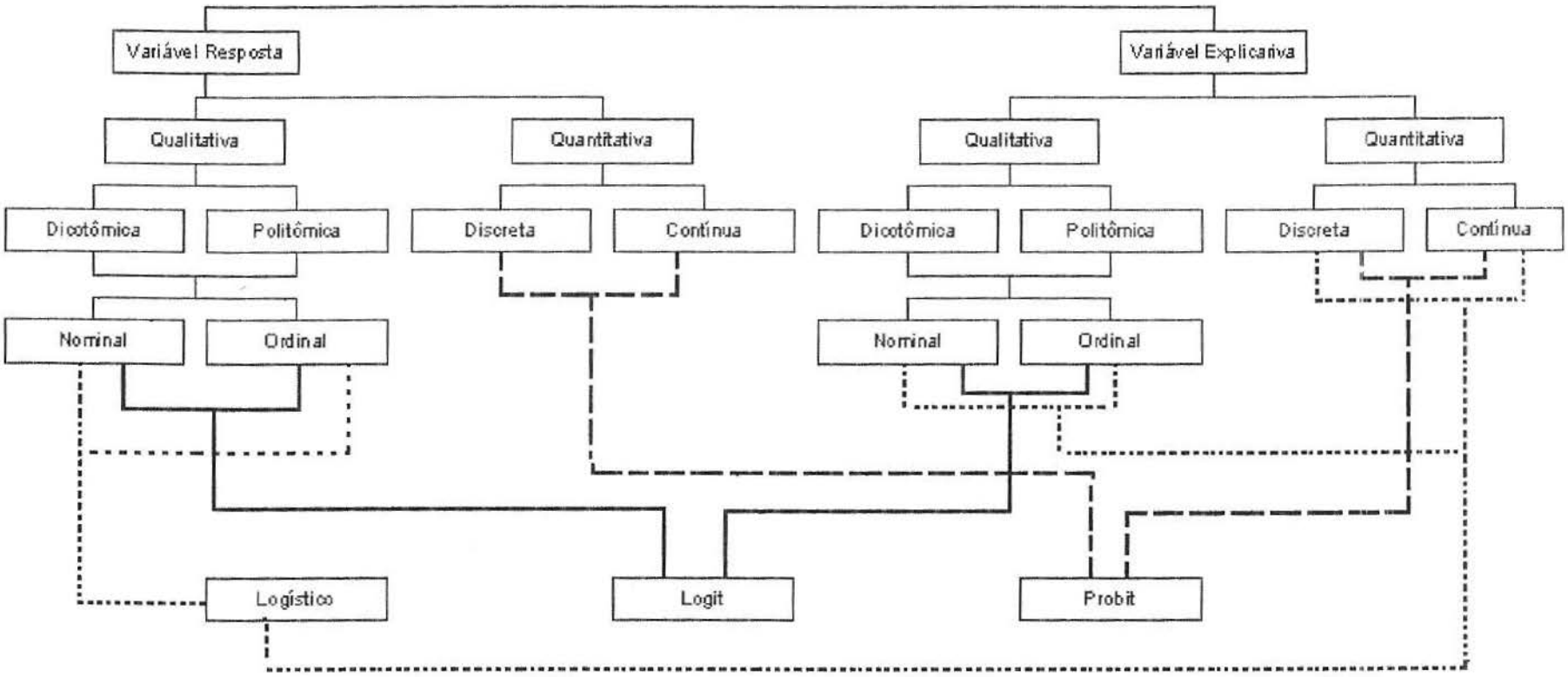

Figura 2.1 Fluxograma para escolha do modelo de Regressão através da classificação das variáveis resposta e das variáveis explicativa

CAPÍTULO 3

3. UM ESTUDO DA COLETA DO EXAME PREVENTIVO DE COLO UTERINO

3.1. DESCRIÇÃO GERAL DA PESQUISA

O estudo apresentado nesta monografia está baseado numa pesquisa realizada no município de Turiaçu localizado na zona Oeste do Maranhão. A população em questão é composta de 31.289 habitantes está distribuída numa área territorial de 2.326,39 km quadrados, de difícil acesso, com características geográficas de Pré-Amazônia, onde o deslocamento em muitos povoados é feito por via marítima e/ou através das matas. Não existe uma linha regular de transporte para os outros municípios ou entre os povoados. O deslocamento das pessoas é realizado nas camionetas "Pau-de-Arara". As atividades econômicas desenvolvidas são a pesca, cultura do abacaxi, a caça, a pesca e a agricultura de subsistência (IBGE, 2000).

Este trabalho utiliza dados obtidos através de uma pesquisa realizada pelo grupo PITS (Programa de Interiorização em Saúde (PITS) composto de três médicos e seis enfermeiros, de Turiaçu. Esse grupo viabilizou, dentre as atividades desenvolvidas no programa, uma proposta de trabalho direcionada para prevenção do câncer de colo de útero e de mamas. Foi realizado o exame clinico das mamas e orientação para o auto-exame, e o rastreamento do câncer de colo uterino, através do exame de Papanicolaou.

Diante deste contexto, uma pesquisa foi conduzida através do preenchimento do formulário fornecido pelo Ministério da Saúde, Requisição de Exame Citopatológico Colo Uterino (Ver anexo).

Os dados analisados foram produzidos através das coletas de exame preventivo de colo uterino ocorridas *no período de 22 de outubro de 2001 a 13 de maio de 2002* no município de Turiaçu, Maranhão.

O Instituto Brasileiro de Geografia e Estatística estimou para o município de Turiaçu no ano 2000 uma população de 31.289 habitantes, sendo 15.033 do sexo feminino, desta, 6.835 na faixa etária de 15 a 49 anos. A população feminina estava distribuída em 4.254 moradoras na zona urbana e 10.799 na zona rural.

Os objetivos iniciais desta pesquisa foram a análise dos dados sócio-culturais obtidos durante a realização das coletas dos exames de Papanicolaou; descrição dos principais padrões citopatológicos dessa população com ênfase em suas prevalências, comparativamente aos encontrados e/ou estimados para o Estado do Maranhão, por faixa etária, nível de escolaridade, hábitos e área geográfica de residência e; Examinar o papel de alguns fatores de risco, citados na literatura, na determinação da doença neste grupo;

Para este estudo, foi utilizado a ficha padrão do Ministério da Saúde para requisição de exames citopatológico. As variáveis do estudo foram: alguns dados sócio econômicos (número de prontuário, data da coleta, idade, escolaridade e local de residência), os dados da anamnese. (uso de métodos anticoncepcionais, primeiro exame ou não), adequabilidade do material colhido e resultados citopatológicos dos laudos (dados apresentados pelo citologista na avaliação do esfregaço). Os dados foram colhidos através de perguntas diretas às pacientes quando do preenchimento da requisição do exame citopatológico e dos resultados dos laudos registrados no verso desta mesma ficha.

As mulheres, após a divulgação, procuraram espontaneamente, o serviço de coleta de preventivos realizado por uma médica e cinco enfermeiros das equipes do PITS.

Para a coleta dos exames, montou-se um consultório ginecológico no ambulatório da Unidade Mista na sede do município e equipou-se uma unidade móvel (um ônibus), estruturado com um consultório ginecológico, ar condicionado, luz, água, pia e estufa para esterilização de material. Foi utilizada também, por um período de cinco dias a unidade móvel (trailer) da Gerência de Qualidade de Vida, equipado inclusive com colposcópio e eletro-cautério de alta freqüência, porém não foi possível a utilização do colposcópio pela médica, pois o mesmo estava instalado de forma inadequada, não permitindo, pela sua distancia em relação a mesa ginecológica, uma localização satisfatória do colo uterino.

A divulgação foi feita junto à comunidade previamente à chegada da equipe. Foram coletados exames na Unidade Mista, na zona urbana e nos povoados de: Porto Santo, Canarinho, Colônia Amélia, Nova Caxias, Santa Rosa, Pedro Bom de Dança, Banta, Estrela Divina, São Romão, Cutia e Antônio Dino.

As mulheres da região das ilhas, deslocaram-se de barco até a sede do município onde foram atendidas.

As coletas foram realizadas em todos os dias da semana, inclusive em alguns sábados, domingos e feriados, com o objetivo de disponibilizar o exame a todas as mulheres, incluindo as que trabalhavam fora ou na roça. A amostra coletada foi de 1213 observações.

Maiores detalhes sobre a pesquisa são encontrados em ELLWANGER, M. I. V., HOPPE, L. C., RODRIGUES, M. G. V., MORGADO, W. ESTUDO DA PREVENÇÃO DE CÂNCER DE COLO UTERINO, Monografia apresentada por curso de Especialização em Saúde da Família. São Luís, 2002.

3.2 OBJETIVOS DA ANALISE ESTATÍSTICA

O estudo desta monografia, utilizando o estudo de caso descrito acima, objetiva fazer uma análise mais detalhada, em relação a algumas variáveis. Portanto, fazendo uso do fluxograma apresentado no capítulo 3, desenvolve-se a análise estatística procurando responder as questões deste estudo de caso.

Os autores da pesquisa permaneceram com a seguinte questão: se a variável "Escolaridade" e "Área residencial" são fatores determinantes em relação à variável "Fez exame alguma vez" (sim, não). Esta questão não foi respondida na ocasião do estudo,

3.3. DADOS DO EXPERIMENTO

O banco de dados da pesquisa foi transportado para o software SPSS, no qual se processou todas as análises estatísticas.

Inicialmente, fez-se uma análise exploratória, apenas para visualizar a o comportamento das variáveis de interesse (Quadro 1).

Na seqüência, procedeu-se o Teste Qui-Quadrado, com a análise de resíduos ajustados para detectar a influência das variáveis **Escolaridade** e **Área residencial** sob a variável **Fez exame alguma vez**, bem como verificar a significância ou não das relações encontradas.

Através da análise de resíduos, é possível detectar a significância dentro de cada categoria das variáveis em questão.

Após a análise inicial, classificaram-se as variáveis explicativas e a variável resposta, recorrendo, então, ao auxílio do fluxograma proposto. Assim, procede-se a escolha do modelo mais indicado para se proceder às análises.

Quadro 1. Descrição da codificação das variáveis consideradas no estudo

Variáveis Explicativas	Códigos	Label
Escolaridade	1=Analfabeto 2=1° grau incompleto 3=1° grau completo 4=2° grau completo	Esc
Área Residencial	1=Urbana 2=Rural 3=Praia	Área
Variável Resposta	Y=0 Y=1	Não fez exame Fez exame alguma vez

3.4 APRESENTAÇAO DOS RESULTADOS

Resultados da Análise Exploratória

A Tabela 3.1 representa a saída do SPSS, na análise cruzada, primeiramente, pelas variáveis Fez exame alguma vez e Escolaridade

Tabela 3.1 Classificação cruzada de " Fez exame alguma vez" e "Escolaridade"

Escolaridade		Fez exame alguma vez		Total
		Não	Sim	
Analfabeto	Freqüência	143	56	199
	Resíduo Ajustado	2,1	-2,1	
1º Grau incompleto	Freqüência	479	236	715
	Resíduo Ajustado	1,5	-1,5	
1º Grau completo	Freqüência	87	33	120
	Resíduo Ajustado	1,8	-1,8	
2º Grau completo	Freqüência	40	73	113
	Resíduo Ajustado	-7,0	7,0	
	Count	749	398	1147

Tabela 3.2 Resultados do teste Qui-Quadrado

Teste	Valor	gl	Significância
Pearson Qui-quadrado	52,018	3	0,000
Razão de verossimilhança	49,424	3	0,000
Total	1147		

Através do resultado do teste Qui-Quadrado (Tabela 3.2), sob a hipótese nula (H_0: As variáveis são independentes), a decisão estatística é a rejeição de H_0, já que o valor calculado, 52,018, é maior do que o tabelado (*p-value* é menor do que 0,001), ou seja existe associação significativa entre Escolaridade e Fez exame alguma vez.

Em relação à análise de resíduos (Tabela 3.1), verifica-se que existe uma freqüência maior que a esperada de mulheres analfabetas que não fizerem exame.

Dentre as mulheres com primeiro grau completo (mesma situação para o primeiro grau incompleto) não há uma associação significativa entre ter realizado ou não o exame. Já dentre as mulheres que possuem o segundo grau completo, há uma proporção maior que a esperada de mulheres que fizeram exame alguma vez, comparando com os demais níveis de escolaridade.

A Tabela 3.3 apresenta a saída do SPSS, na análise cruzada pelas variáveis **Fez exame alguma vez** e **Área residencial.**

Tabela 3.3 Classificação cruzada de "Fez exame alguma vez" e "Área residencial"

Área residencial		Fez exame alguma vez		Total
		Não	Sim	
Urbana	Freqüência	188	195	383
	Resíduo Ajustado	-8,2	8,2	
Rural	Freqüência	561	204	765
	Resíduo Ajustado	7,4	-7,4	
Praia	Freqüência	47	18	65
	Resíduo Ajustado	1,2	-1,2	
	Total	796	417	1213

Tabela 3.4 Teste Qui-Quadrado

Teste	Valor	gl	Significância
Pearson Qui-quadrado	67,875	2	0,000
Razão de verossimilhança	66,355	2	0,000
Total	1213		

Na Tabela 3.4 , pelo resultado do teste Qui-Quadrado sob a hipótese nula (H0: As variáveis são independentes), a decisão estatística é a rejeição de H0, já que o valor calculado, 67,875, é maior do que o tabelado (*p-value* é menor do que 0,001), ou seja existe associação significativa entre **Área residencial** e **Fez exame alguma vez.**

Em relação à análise de resíduos ajustados, verifica-se que dentre as mulheres provenientes da área rural há uma associação significativa com a categoria "não fizeram exame alguma vez".

Em relação àquelas que moram em área urbana, existe uma proporção maior de mulheres que já realizaram o exame em relação às mulheres provenientes das outras áreas consideradas neste estudo.

Tabela 3.5 Classificação cruzada das variáveis "Escolaridade" e "Área residencial"

Escolaridade		Área residencial			Total
		Urbana	Rural	Praia	
Analfabeto	Freqüência	33	162	4	199
	Resíduo Ajustado	-5,1	5,8	-2,0	
1º Grau incompleto	Freqüência	205	464	46	715
	Resíduo Ajustado	-2,9	1,4	3,3	
1º Grau completo	Freqüência	40	77	3	120
	Resíduo Ajustado	0,4	0,2	-1,2	
2º Grau completo	Freqüência	87	24	2	113
	Resíduo Ajustado	10,9	-9,8	-1,6	
Total		365	727	55	1147

Tabela 3.6 Teste Qui-Quadrado

Teste	Valor	gl	Significância
Pearson Qui-quadrado	142,119	6	0,000
Razão de verossimilhança	136,253	6	0,000
Total	1147		

Na Tabela 3.6., pelo resultado do teste Qui-Quadrado sob a hipótese nula (H0: As variáveis são independentes), a decisão estatística é a rejeição de H0, já que o valor calculado, 142,119, é maior do que o tabelado (*p value* é menor do que 0,001), ou seja existe associação significativa entre a **área residencial** e **escolaridade**.

Em relação à análise de resíduos ajustados (Tabela 3.5.), verifica-se que dentre as mulheres moram na área rural há uma associação significativa em relação às que são analfabetas, ou seja, entre mulheres provenientes da área rural, mais do que o esperado são analfabetas. Em relação àquelas que moram em Área urbana, existe uma proporção maior de mulheres que possuem o 2º grau completo em relação às mulheres provenientes das outras áreas consideradas neste estudo.

Essa tabela apenas relaciona as variáveis explicativas, sendo, às vezes, importante para análises e considerações do pesquisador.

Tabela 3.7 Cruzamento das variáveis área residencial, escolaridade e Fez exame alguma vez.

Área residencial	Escolaridade	Fez exame alguma vez		Total	% Total Geral
		Não	Sim		
Urbana	Analfabeto	15	18	33	2,9
	1º Grau incompleto.	109	96	205	17,9
	1º Grau completo	25	15	40	3,5
	2º Grau completo	28	59	87	7,6
Rural	Analfabeto	125	37	162	14,1
	1º Grau incompleto	337	127	464	40,5
	1º Grau completo	59	18	77	6,7
	2º Grau completo	12	12	24	2,1
Praia	Analfabeto	3	1	4	0,3
	1º Grau incompleto.	33	13	46	4
	1º Grau completo	3	0	3	0,3
	2º Grau completo	0	2	2	0,2
	Total	749	398	1147	100

A Tabela 3.7. estima que a proporção de mulheres com 1º grau incompleto, em relação ao total geral, representada em cada área residencial é maior em relação às demais categorias de escolaridade. Verifica-se que 17,9% são da área urbana, 40,5 % são provenientes da área rural e 4% são provenientes da área praia. Evidencia-se também que 14,1% das mulheres são analfabetas e provenientes da área rural.

Escolha do modelo de análise

Considerando as seguintes variáveis em questão: X_1: escolaridade, X_2: área residencial e Y: Fez exame alguma vez. Conforme o objetivo de pesquisa, deseja-se verificar se as variáveis escolaridade e área residencial são fatores predominantes para escolha da resposta "Fez exame alguma vez". Neste caso, tem-se claramente uma variável resposta dicotômica, "Fez exame alguma vez", codificada por (1-sim; 0-não), com duas variáveis explicativas categóricas: " escolaridade" e "área residencial".

As variáveis consideradas no estudo podem ser classificas como segue:

- Variável Resposta "Fez exame alguma vez": Qualitativa, Dicotômica, Nominal
- Variável Explicativa "Escolaridade": Qualitativa, Politômica, Ordinal
- Variável Explicativa "Área Residencial": Qualitativa, Politômica, Nominal

Fazendo uso do fluxograma do capítulo anterior (Figura 2.1), segue-se os caminhos desenvolvidos até se chegar a um dos três modelos abordados, considerando as características da variável resposta e a natureza das variáveis explicativas. Conforme foi discutido anteriormente, o modelo de regressão logística pode ser usado quando temos variáveis explicativas qualitativas ou quantitativas.

Neste estudo, a análise será feita de acordo com o modelo de regressão logística, sendo posteriormente feita uma análise usando o modelo logit para dados categóricos, de forma a explorar as probabilidades associadas a cada combinação entre as variáveis Escolaridade e Área residencial. A análise dos dados é feita no software SPSS a partir dos comandos para a regressão logística.

Análise dos dados usando o modelo de Regressão Logística obtido através do software SPSS

A amostra total contém 1213 observações, para o estudo foram consideradas 1147 observações, devido a variável independente, Escolaridade, possuir 66 observações ignoradas (missing), representando 5,4 % das 1213 observações, sendo que 96,4% foram incluídas na amostra. A Tabela 3.8. representa estes resultados.

Tabela 3.8 Histórico das observações obtidas do Estudo de Caso do Exame Preventivo de Colo Uterino

Casos Selecionados	N	Percentagem
Observações incluídas na Análise	1147	94,6
Observações "missing"	66	5,4
Total	1213	100

Verificando a validade do modelo de regressão estimado:

A Tabela 3.9. apresenta a validação da escolha do modelo através do teste da razão de verossimilhança e da tabela classificatória interpretados na seqüência.

Tabela 3.9 Histórico da interação do ajuste do modelo incluindo somente a constante

-2Log Verossimilhança (-2LL)	Coeficientes
	Constante
1481,032	-0,612

A estatística da razão de verossimilhança (-2LL) para o modelo do estudo de caso, somente com a constante resultou em 1481,032. Essa estatística avalia o quão bem o modelo estimado se ajusta aos dados.

Conforme a Tabela 3.9 , para o modelo logístico que contém somente a constante β_0, o valor de (-2 LL) é 1481,032 e para o modelo que contém todas as variáveis, Área residencial e Escolaridade, o valor de (-2 LL) é 1392,553. Conforme Tabela 3.11. é representada a diferença entre estes dois valores, resultando a estatística Qui-Quadrado = 88,373. Este resultado mostra que variáveis Escolaridade e Área residencial foram significativas para explicar o modelo.

Tabela 3.11 Teste dos Coeficientes do Modelo

	Qui-Quadrado	gl	Significância
Passo	88,373	5	0,000
Bloco	88,373	5	0,000
Modelo	88,373	5	0,000

Através da classificação da Tabela 3.12 para a variável "Fez exame alguma vez", incluindo todos os casos, verifica-se que há mais coerência entre as que nunca fizeram exame, pois 96,3% são corretamente classificados pelo modelo. Daquelas que já fizeram exame alguma vez, 15,3% são corretamente classificados pelo modelo e no geral, 68,2% foram classificadas corretamente pelo modelo predito, Área residencial e Escolaridade, na variável resposta "Fez exame alguma vez".

Tabela 3.12 Classificação da variável "Fez exame alguma vez", após a inclusão da variável Escolaridade e a variável Área residencial.

Observado		Predito Fez exame alguma vez		Percentagem
		Não	Sim	
Fez exame alguma vez	Não	721	28	96,3
	Sim	337	61	15,3
Percentagem sobre todos os casos				68,2

Quando as variáveis explicativas são quantitativas é possível substituí-las diretamente no modelo de regressão estimado. No entanto, quando há mais de duas categorias, cria-se uma nova variável para representar essas categorias. A nova variável será uma variável dummy que assumirá o valor 1 se a categoria estiver presente e 0 caso contrário. A variável codificada representará o efeito da variável sobre uma categoria de referência. O valor 1 é atribuído à categoria em questão. A categoria referência recebe o valor 0 para todas as variáveis arbitrárias.

Com variáveis categóricas a interpretação somente pode ser feita sobre o efeito de uma categoria particular em comparação com alguma outra categoria escolhida como referência.

As categorias de referência para as variáveis Área residencial e Escolaridade, foram indicadas em relação à primeira categoria para serem comparadas. Para Área residencial, a categoria referencial é Área Urbana e para Escolaridade, a categoria referencial é Analfabeto. A codificação destas variáveis em variáveis indicadoras aparece na Tabela 3.14. .

Tabela 3.14 Codificação das variáveis Escolaridade e Área residencial segundo a categoria de referência.

Variáveis Independentes	Categorias	Freqüência	Código dos Parâmetros		
			(1)	(2)	(3)
Escolaridade	Analfabeto	199	,000	,000	,000
	1º Grau incompleto	715	1,000	,000	,000
	1º Grau completo	120	,000	1,000	,000
	2º Grau completo	113	,000	,000	1,000
Área residencial	Urbana	365	,000	,000	
	Rural	727	1,000	,000	
	Praias	55	,000	1,000	

Tabela 3.15 Resultado da razão de chances Exp(B) para todas a variável Área Rural. (categoria de referência: Urbana.)

Variáveis	B	Erro	Wald	gl	Sig.	Exp(B)
AREA RESIDENCIAL			39,209	2	0,000	
(1)Rural	-0,884	0,142	38,590	1	0,000	0,413
(2)Praia	-0,799	0,320	6,248	1	0,012	0,450

A interpretação de Exp(β) (Tabela 3.15), para a variável Área residencial na categoria Rural, representa que há mais chance de fazer exame provindo da área urbana do que provindo da área rural 2,421 vezes; para a categoria Praia, Exp(β) representa que há 2,222 vezes mais chance de fazer exame sendo da área urbana do que provindo da área praia.

Tabela 3.16 Resultado da razão de chances Exp(B) para a variável Escolaridade. (categoria de referência: Analfabeto)

Variável	B	Erro	Wald	gl	Sig.	Exp(B)
ESCOLARIDADE			22,140	3	0,000	
(1)1° grau Incompleto	0,116	0,181	0,412	1	0,521	1,123
(2)1° grau Completo	-0,197	0,264	0,556	1	0,456	0,821
(3)2° grau Completo	1,041	0,267	15,201	1	0,000	2,832
Constante	-,222	0,195	1,296	1	0,255	0,801

Para a variável Escolaridade, na categoria 2° grau completo, correspondendo, Exp(β) representa uma estimativa de 2,832 vezes mais chance de fazer exame tendo o 2° grau completo em relação às que são analfabetas (Tabela 3.16.).

Análise do modelo Logit para dados categóricos

Quando há interesse nas probabilidades associadas com cada combinação de níveis das variáveis explanatórias relacionadas com a variável resposta, procede-se à análise recomendada pelo modelo logit, baseado nas freqüências de cada linha da tabela cruzada, que gera as probabilidades. Essa análise é apresentada na Tabela 3.17.

A Tabela 3.17 apresenta as probabilidades, os valores de Odds e Logit para a variável "fez exame alguma vez", dentre os níveis das variáveis explicativas escolaridade e área residencial.

Nesta tabela, aparecem as 12 probabilidades e, ou os 12 valores de Odds, das combinações entre os níveis das variáveis Escolaridade e Área residencial, considerando o modelo estimado para a variável dependente "Fez exame alguma vez" e as variáveis explanatórias "Escolaridade" e "Área residencial".

A probabilidade estimada de uma mulher fazer o exame preventivo de colo uterino sendo "analfabeta e provindo da área urbana" é de 0,545, sendo que a probabilidade dela não fazer o exame é 0,455.

Para as mulheres que possuem o 2º grau completo e provém da área urbana, a probabilidade estimada de fazer o exame é 0,678.

A probabilidade estimada para uma mulher não fazer o exame sendo analfabeta e provindo de uma área rural é estimada em 0,772, sendo que a chance dela fazer o exame, nesta situação é estimada em 0,228.

Tabela 3.17 Probabilidades para variável "Fez exame alguma vez".

	Área Urbana		Fazer Exame	Não Fazer Exame		
Escolaridade	**Analfabeto**	Freq	P_1	$1 - P_1$	odds	Logit
y = 0	não	15	0,545	0,455	1,2	0,18
y = 1	sim	18				
	Total	33				
Escolaridade	**1°Incomp**	Freq	P_1	$1 - P_1$	odds	Logit
y = 0	não	109	0,468	0,531	0,880	-0,126
y = 1	sim	96				
	Total	205				
Escolaridade	**1°Comple**	Freq	P_1	$1 - P_1$	odds	Logit
y = 0	não	25	0,375	0,625	0,6	-0,511
y = 1	sim	15				
	Total	40				
Escolaridade	**2°Completo**	Freq	P_1	$1 - P_1$	odds	Logit
y = 0	não	28	0,678	0,329	2,107	0,745
y = 1	sim	59				
	Total	87				
	Área Rural					
Escolaridade	**Analfabeto**	Freq	P_1	$1 - P_1$	odds	Logit
y = 0	não	125	0,228	0,772	0,296	-1,217
y = 1	sim	37				
		162				
Escolaridade	**1° Incompleto**	Freq	P_1	$1 - P_1$	odds	Logit
y = 0	não	337	0,274	0,726	0,376	-0,975
y = 1	sim	127				
	Total	464				
Escolaridade	**1° Completo**	Freq	P_1	$1 - P_1$	odds	Logit
y = 0	Não	59	0,234	0,766	0,305	-1,187
y = 1	Sim	18				
	Total	77				
Escolaridade	**2° Completo**	Freq	P_1	$1 - P_1$	odds	Logit
y = 0	não	12	0,5	0,5	1	0
y = 1	sim	12				
	Total	24				
	Área Praia					
Escolaridade	**Analfabeto**	Freq	P_1	$1 - P_1$	odds	Logit
y = 0	não	3	0,25	0,75	0,333	-1,098
y = 1	sim	1				
	Total	4				
Escolaridade	**1° Incompleto**	Freq	P_1	$1 - P_1$	odds	Logit
y = 0	não	33	0,283	0,717	0,393	-0,931
y = 1	sim	13				
	Total	46				
Escolaridade	**1° Completo**	Freq	P_1	$1 - P_1$	odds	Logit
y = 0	não	3	0	1	0	-
y = 1	sim	0				
	Total	3				
Escolaridade	**2° Completo**	Freq	P_1	$1 - P_1$	odds	Logit
y = 0	não	0	1	0	-	-
y = 1	sim	2				
	Total	2				

Corroborando com as análises anteriores, A análise deste estudo revela que as mulheres com baixa escolaridade são menos propícias a realizar o exame preventivo de colo uterino. Evidenciou-se que mulheres provindo de uma área urbana possuem maior nível de escolaridade, e desta forma, são mais propícias à realizar o exame, embora no modelo multivariado, área residencial é ajustado à escolaridade dada e vice-versa.

A Figura 4.1 apresenta um gráfico com as probabilidades de realizar o exame (eixo das ordenadas) para cada nível das variáveis explicativas (eixo das abcissas). O gráfico não indica haver interação entre as variáveis, isto é, mulheres provenientes da área urbana tendem a realizar o exame em maior proporção que as mulheres da área rural, independente da escolaridade.

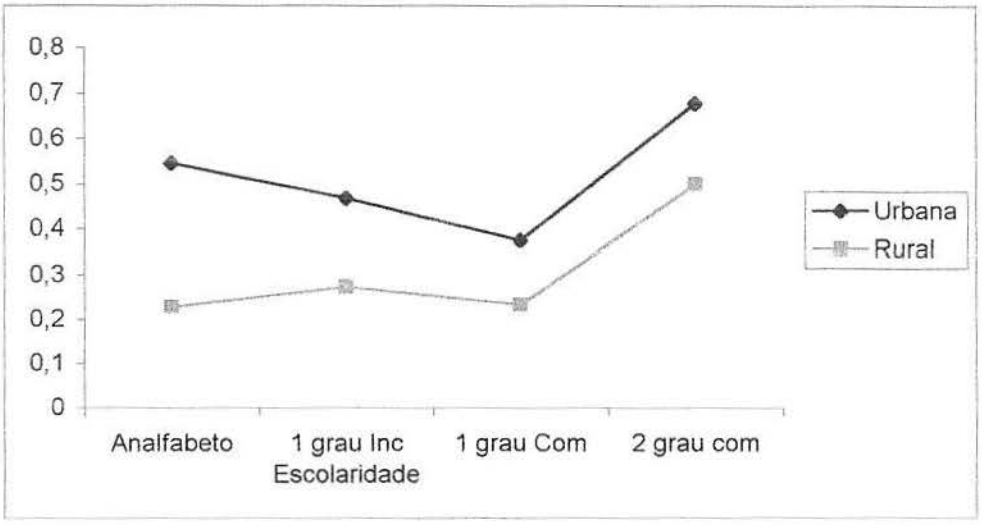

Figura 4.1.Comparação das probabilidades de realizar o exame entre as variáveis explicativas

Da mesma forma, mulheres com 2 grau completo tendem a fazer o exame em maior proporção que as de escolaridade inferior. Considerando este exemplo, notou-se que os analfabetos da área urbana apresentaram maior proporção de realizarem o exame que os analfabetos da área rural. Em relação à área urbana, esse percentual foi maior que as mulheres de primeiro grau completo. Uma possível justificativa para isso, segundo a opinião da autora da monografia é a realização de programas específicos desenvolvidos nas áreas urbanas para atendimento de mulheres de baixas rendas, que são, na maioria analfabetas na região de realização deste estudo.

CAPÍTULO 4

4. CONSIDERAÇÕES FINAIS

Variáveis categóricas (dicotômicas e politômicas) na classificação de dados dividem uma amostra em vários subgrupos com base em suas qualidades ou atributos (sexo, estado civil, raça, etc) e implicitamente permitem utilizar regressões individuais para cada subgrupo. A aplicação de modelos de regressão para variáveis qualitativas deve ser manipulada com cautela.

Modelos de regressão descrevem o relacionamento entre a variável resposta e uma ou mais variáveis explanatórias, assumindo, geralmente, que a variável resposta é contínua. Entretanto, quando a variável resposta é dicotômica, os modelos de regressão devem ser tratados para dados categóricos.

Quando a variável dependente é dicotômica leva-nos a pensar, imediatamente, no modelo de regressão logística. Porém, deve-se olhar atentamente para as variáveis independentes e, então ajustar aos dados o modelo mais adequado.

Para variáveis resposta binária o modelo de regressão logística descreve como a probabilidade de uma particular categoria, Y=1 ou Y=0, depende das mudanças nos valores da variável explanatória. Após o ajuste do modelo pode-se prever probabilidade de vários grupos em função das variáveis explicativas.

O modelo logit, no qual todas as variáveis explicativas são categóricas é um caso especial de modelos loglineares. O modelo logit contém as mesmas informações do modelo loglinear, porém sem a interação das variáveis explicativas com a variável resposta.

Na formulação de modelos logits para variáveis ordinais, há três tipos de logit que são formados por grupos de categorias que são próximas à escala ordinal. Esses logits categorizam a ordem da variável e são chamados de Logit cumulativo, Logit da razão contínua e Logit de categoria adjacente. Como extensão da regressão logística, o modelo logit cumulativo acumula probabilidades para a variável resposta ordinal.

Os modelos que utilizam a ordem natural de variáveis resultam em mais poder de inferência (Agresti e Finlay, 1994).

O modelo probit é mais utilizado quando a variável resposta e explicativa são ambas contínuas. Além disso, a variável resposta também pode ser dicotômica, havendo uma variável resposta subjacente.

Para exemplificar o uso destes modelos, propôs-se a organização dos mesmos num quadro geral e num fluxograma que pretende contribuir para orientar pesquisadores em diversas áreas a utilizar esses modelos. Para ilustrar a aplicação, discute-se um estudo de caso real, analisando as variáveis e a interpretação dos resultados estimados pelo modelo escolhido.

A análise detalhada em relação à classificação categórica das variáveis envolvidas no processo é importante no momento da escolha do modelo. Conforme foi discutido, cada modelo tem suas peculiaridades em relação à variável dependente e independente, sendo muitas vezes a variável dependente quem decide, em última análise, a escolha por este ou aquele modelo. Também em outras ocasiões a variável independente é quem determinará a decisão do modelo.

Pode haver, ainda, mais do que um modelo de regressão que pode ser usado para os mesmos dados. Através dos caminhos percorridos pelo fluxograma da Figura 2.1, percebe-se que o modelo Logit também poderia ser utilizado para os dados do estudo de caso aqui analisado, em função das variáveis em questão.

O bom senso do pesquisador, aliado com a análise exploratória de classificação das variáveis, bem como o entrelaçamento entre variáveis dependentes e independentes, será o fator de decisão na boa escolha do modelo de estimação. Este fato refletirá maior segurança na análise dos resultados.

Como objetivo principal desta monografia, realizou-se uma revisão bibliográfica onde foram apresentados conceitos básicos de cada modelo, constituindo o mínimo de conhecimento para aplicação em estudos práticos. Para ilustrar, a partir do referencial teórico, aplicou-se um modelo de análise de regressão a um estudo de caso. No estudo de caso, os passos foram exemplificados no desenvolvimento do processo de análise dos dados. Como objetivo secundário, propôs-se orientar a escolha do modelo que envolvem dados categóricos, baseado nas características das variáveis resposta e explicativas ativa. Para tanto, construiu-se um fluxograma como forma de orientação.

Ao final da pesquisa desenvolvida nessa monografia, algumas constatações podem ser reportadas: como recomendação a trabalhos futuros, sugere-se um estudo mais detalhado, considerando os modelos loglineares, com ênfase nos modelos logits. Recomenda-se ainda uma abordagem mais minuciosa e substancial, relacionando as diferenças dos resultados das estimativas de cada modelo, para os mesmos dados, visto que, esta questão ficou pendente e desejosa por parte da autora. Com essa análise é possível aperfeiçoar o fluxograma e o quadro comparativo, produtos da proposta original desta monografia.

ANEXOS

MINISTÉRIO DA SAÚDE

REQUISIÇÃO DE EXAME CITOPATOLÓGICO - COLO DO ÚTERO

Viva Mulher - Programa Nacional de Controle do Câncer do Colo do Útero e de Mama

UF | Cartão SUS | Código da Unidade de Saúde

Unidade de Saúde

Município | Prontuário

Informações Pessoais

ATENÇÃO: Não serão processados os exames que não tiverem o nome, idade, endereço e nome da mãe da paciente preenchidos

Nome completo da mulher

Nome completo da mãe

Apelido da mulher

Identidade | Órgão emissor | UF | CNPF (CPF)

Data de nascimento / / | Idade

Dados residenciais

Logradouro

Número | Complemento

Bairro | UF

Município

CEP – | DDD | Telefone –

Ponto de referência

ESCOLARIDADE: ☐ Analfabeta ☐ 1º grau incompleto ☐ 1º grau completo ☐ 2º grau completo ☐ 3º grau completo

Dados da Anamnese

1. Fez o exame preventivo (Papanicolaou) alguma vez?
   - ☐ Sim. Quando fez o último exame?
     Ano
   - ☐ Não ☐ Não sabe

2. Usa DIU? ☐ Sim ☐ Não ☐ Não sabe

3. Está grávida? ☐ Sim ☐ Não ☐ Não sabe

4. Usa pílula anticoncepcional?
   ☐ Sim ☐ Não ☐ Não sabe

5. Usa hormônio / remédio para tratar a menopausa?
   ☐ Sim ☐ Não ☐ Não sabe

6. Já fez tratamento por radioterapia?
   ☐ Sim ☐ Não ☐ Não sabe

7. Data da última menstruação / regra:
   / / ☐ Não sabe / não lembra

8. Tem ou teve algum sangramento após relações sexuais?
   (não considerar a primeira relação sexual na vida)
   - ☐ Sim
   - ☐ Não / não sabe / não lembra

9. Tem ou teve algum sangramento após a menopausa?
   (não considerar o(s) sangramento(s) na vigência de reposição hormonal)
   - ☐ Sim
   - ☐ Não / não sabe / não lembra/ não está na menopausa

Exame Clínico

10. Inspeção do colo
    - ☐ Normal
    - ☐ Ausente (anomalias congênitas ou retirado cirurgicamente)
    - ☐ Alterado
    - ☐ Colo não visualizado

11. Sinais sugestivos de doenças sexualmente transmissíveis?
    - ☐ Sim
    - ☐ Não

Data da coleta / / | Coletor

Unidade de Saúde

Nome completo da mulher

REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A. (1984) **Analysis of Ordinal Categorical data**. Florida, Ed. John Wiley & Sons.

AGRESTI, A. (1990) **Categorical Data Analysis**. Florida,. Ed. John Wiley & Sons.

AGRESTI, A. e FINLAY, B. (1997) **Statistical Methods for the Social Sciences**. 3 ed, Florida, Ed. Wiley & Sons.

ANDERSEN, E. B. (1997) **Introduction to the Statistical Analisys of Categorical Data**. Heidelberg, Ed. Springer.

ECHEVESTE, M. E. e NODARI (2000) **Comparação do modelo de Regressão Logística através das rotinas do Software ALOGIT e do Software Estatístico SPSS**. Escola de Engenharia da UFRGS, Material desenvolvido durante curso de doutorado PPGEP/UFRGS.

ELLWANGER, M. I. V. , HOPPE, L.C., RODRIGUES, M. G, V., MORGADO, W. (2002) **Estudo da prevenção de cancêr de colo uterino.** Monografia apresentada por curso de Especialização em Saúde da Família. São Luís, Maranhão.

EVERITT, B. S. (1992) **The analysis of contingency tables**. 2 ed., London, Ed. Chapman and Hall.

FACHEL, J. G., (1999), Notas de Aula, Departamento de Estatística, Instituto de Matemática, UFRGS. Porto Alegre.

FIENBERG, S. E. (1980) **The Analysis of Cross-Classified Categorical Data**. 2 ed., London, Ed. The MIT Press

FLATH,D. AND LEONARD **Journal of Marketing Research**. A Comparison of Two Logit Models in the Analysis of Qualitative Marketing Data, Vol. XVI (November 1979), 533-8

GUJARATI, D. N .(1995) **Econometria Básica**. 3 ed., São Paulo, Editora Afiliada.

GOODMAN, L. A. (1978) **Analyzing Qualitative/Categorical Data, Log-Linear Models and Latent Structure Analysis**. University of Chicago, Ed. ABT BOOKS

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

MADALLA G.S. (1985) **Limited-dependent and qualitative variables in econometrics**. New York, Ed. Cambridge University Press.

McFadden, D. (1978) **Model choose of the position residential**. Registro 673 pp. da pesquisa do transporte 72-77. National Academy of Sciences.

SPSS Inc. (1993) **SPSS for Windows Base System User's Guide** Release 6.0. Chicago.

SPSS Professional Statistics 7.5, Copyright (1997) by SPSS inc.

TABACHNICK, B. G. e. FIDELL, L. S. (1996) **Using Multivariate Statistics**. 4 ed., California State University, Northridge.