Especialização em *Data Science* e Estatística Aplicada

Módulo IV - Análise de Dados Categóricos

Prof Dr Márcio Augusto Ferreira Rodrigues

Goiânia, 2025

IME
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
FEN
FACULDADE DE ENFERMAGEM
UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Conteúdo Programático

- Dados categorizados. Tabelas de contingência bidimensionais.
- Tabelas de contingência tridimensionais.
- Testes para dados categorizados: qui-quadrado, exato de Fisher e razão de verossimilhança.
- Modelo de regressão logística binária.
- Regressão Logística Politômica.
- Aplicações no software R.

Conteúdo - Aula 2

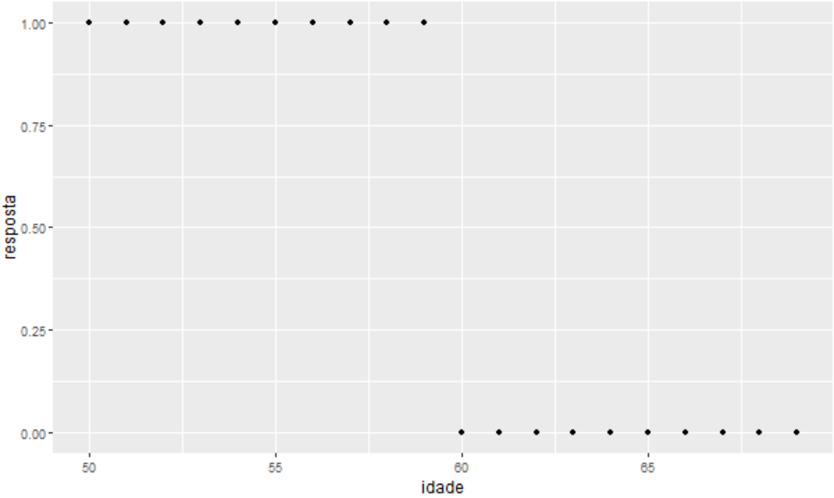
1. Regressão logistica binária
   - Qualidade do Modelo ajustado
   - Diagnóstico em Regressão Logística

2. Regressão Logística Politômica
   - Modelo Logístico Politômico para variável resposta Nominais
   - Modelo Logístico Politômico para variável resposta Ordinais

Regressão logistica binária

1. Característica Básica: Desfecho (varíavel resposta) **Binário**
2. Objetivo:
    - Identificar fatores de risco ou Prognóstico;
    - Comparar grupos, controlando por fatores de confusão;
    - Predição
3. Estudo transversal: Regressão Logística usada com frequência.
4. Estudo Longitudinal: Regressão Logística poucou, ou raramente, utilizada neste desenho.

Regressão logistica binária



Regressão logistica binária

O modelo regressão linear

- Seja Y_i a variável resposta para observações $i = 1, \cdots, n$, e considere a presença de p variáveis explicativas $x_{i1}, \cdots, x_{ip}$.
- Relacionamos as variáveis explicativas com a variável resposta através de um modelo linear:

$$Y_i = \beta_0 + \beta_1 x_{i1}, \cdots, \beta_p x_{ip} + \epsilon_i,$$

  em que ϵ_i são iid com distribuição normal com média 0 e variância σ^2, para $i = 1, \cdots, n$.
- Com isso, Y_i será iid com distribuição normal, para $i = 1, \cdots, n$
- Os $\beta_0, \beta_1, \cdots, \beta_p$ são parâmetros da regressão que quantificam as relações entre as variáveis explicativa e a variável resposta, dadas as outras variáveis do modelo.

Regressão logistica binária

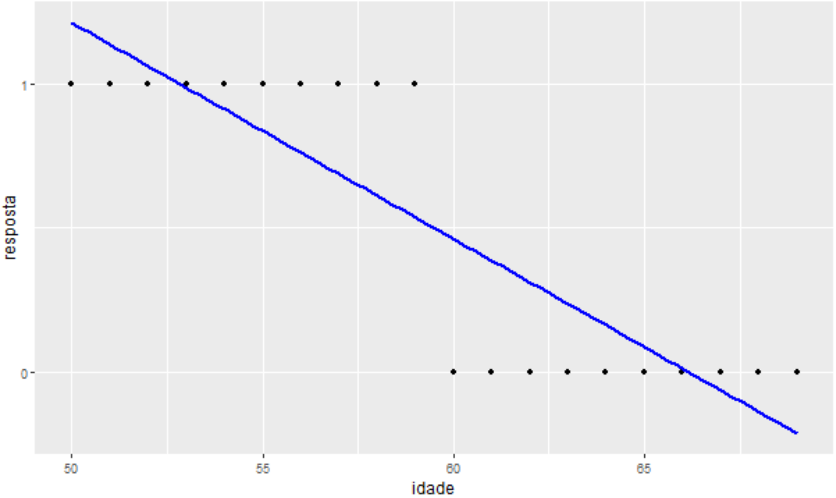
O modelo regressão linear

- Se $\beta_1 = 0 \Rightarrow$ não existe uma relação linear entre a variável x_1 e a variável resposta, dadas as outras variáveis do modelo.
- Se $\beta_1 > 0 \Rightarrow$ existe uma relação positiva e, um aumento na variável explicativa leva a um aumento na variável resposta.
- Se $\beta_1 < 0 \Rightarrow$ existe uma relação negativa e, um aumento na variável explicativa leva a uma diminuição na variável resposta.
- Tomando a esperança de Y_i, também podemos escrever o modelo linear como

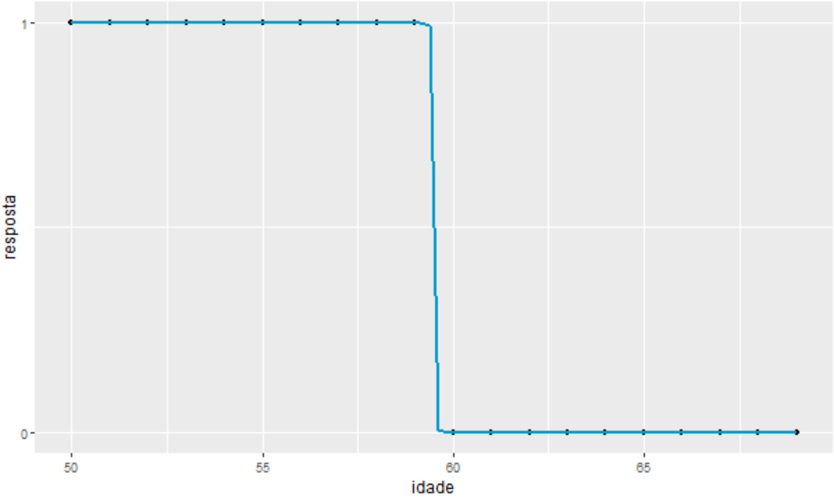
$$E(Y_i) = \beta_0 + \beta_1 x_{i1}, \cdots, \beta_p x_{ip} + \epsilon_i,$$

- Os valores $\beta_0, \beta_1, \cdots, \beta_p$ são estimados usando estimação de mínimos quadrados, por exemplo.

Regressão logistica binária



Regressão logistica binária



Regressão logistica binária

O modelo regressão logística

- Este modelo é uma extensão de regressão linear para o caso em que a variável Y possui distribuição binomial.
- No contexto de repostas binárias, a quantidade que queremos estimar é a probabilidade de sucesso, $\pi = P(Y = 1)$.
- Sejam Y_i variáveis reposta binárias independentes para observações $i = 1, \cdots, n$.
- $Y_1, \cdots, Y_n$ ensaios de Bernoulli independentes, tal que

$$Y_i = \begin{cases} 1, & \text{com probabilidade } \pi_i \\ 0, & \text{com probabilidade } 1 - \pi_i \end{cases}$$

Regressão logistica binária

O modelo regressão logística

- A probabilidade π_i varia de indivíduo para indivíduo
- Queremos que esta variação ocorra em função das covariaveis $\mathbf{X} = (X_1, \cdots, X_p)$, isto é, queremos escrever

$$\pi_i = P(Y = 1|\mathbf{X} = \mathbf{x}_i) = g(\mathbf{x}_i).$$

- Queremos que $g(\mathbf{x}) \to 1$ quando x cresce e que $g(\mathbf{x}) \to 0$ quando x diminui
- Existem algumas escolhas populares para função $g(\mathbf{x})$: logística, probit, log-log complementar.
- A mais usada é a função logística
- Ela possui poucos parâmetros, é simples de entender, é flexível, e ajusta-se facilmente a diferentes contextos.

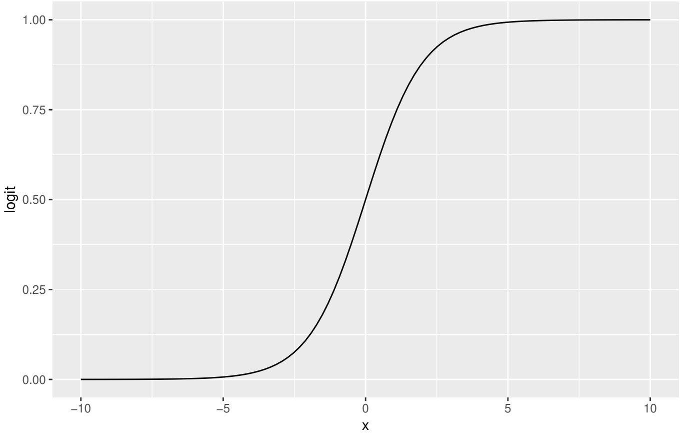
Regressão logistica binária

O modelo regressão logística

- O modelo linear $\pi_i = \beta_0 + \beta_1 x_{i1}, \cdots, \beta_p x_{ip}$ é o mais simples, no entanto este modelo pode levar a valores de π_i menores que 0 ou maiores que 1, dependendo dos valores das variáveis explicativas.
- A função logística é uma transformação que pega qualquer valor entre menos infinito e mais infinito e transforma em um valor entre 0 e 1.

Regressão logistica binária

Função logística



Regressão logistica binária

Regressão logistica binária

- Temos

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_{i1}, \cdots, \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1}, \cdots, \beta_p x_{ip})}$$

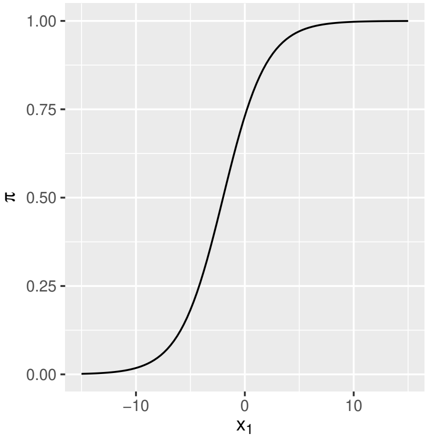
- Consequentemente,

$$1 - \pi(x) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_{i1}, \cdots, \beta_p x_{ip})}$$

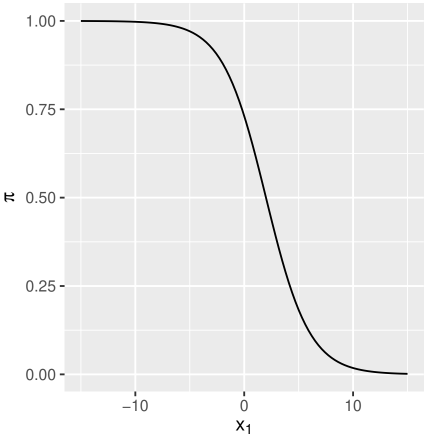
- Observe que com essa formulação, $0 < \pi_i < 1$.
- A regressão logística é um recurso que nos permite **estimar a probabilidade associada à ocorrência de terminado evento**, em face de um conjunto de variáveis explicativas.

Regressão logistica binária

$$\pi = \frac{e^{1+0.5x_1}}{1+e^{1+0.5x_1}}$$



$$\pi = \frac{e^{1-0.5x_1}}{1+e^{1-0.5x_1}}$$



Regressão logistica binária

- Observe que tomando o logaritmo neperiano da razão entre $\pi(x)$ e $1-\pi(x)$ obtem-se um modelo linear, isto é

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_{i1}, \cdots, \beta_p x_{ip}$$

- Tal transformação é denominada logito;
- Como a razão entre $\pi(x)$ e $1-\pi(x)$ define uma chance, tem-se que o logito é o logaritmo de uma chance, logo,

$$\text{chance} = \frac{\pi(x)}{1-\pi(x)} = \exp(\beta_0 + \beta_1 x_{i1}, \cdots, \beta_p x_{ip})$$

Regressão logistica binária

- Uma função monótona e derivável que relaciona a média ao preditor linear é denominada função de ligação (MLG).
- $\ln\left[\dfrac{\pi(x)}{1-\pi(x)}\right]$ é a função de ligação canônica associada ao modelo binomial.
- Geralmente escrevemos o modelo de regressão logistica como

$$logit(\pi) = \beta_0 + \beta_1 x_{i1}, \cdots, \beta_p x_{ip}$$

Regressão Logistica Binária

Estimação dos parâmetros - Um único regressor

MÉTODO DE MÁXIMA VEROSSIMILHANÇA

Considere uma amostra de respostas binárias $y_1, ..., y_n$ independetes e covariáveis $x_1, ..., x_n$. A função de verossimilhança é

$$\begin{aligned}
L(\beta_0, \beta_1) &= \prod_{i=1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)} \right)^{y_i} \left(\frac{1}{1 + exp(\beta_0 + \beta_1 x_i)} \right)^{1-y_i} \\
&= \prod_{i=1}^{n} \frac{exp(\beta_0 + \beta_1 x_i)^{y_i}}{1 + exp(\beta_0 + \beta_1 x_i)}
\end{aligned}$$

Regressão Logistica Binária

Estimação dos parâmetros - Um único regressor

- O estimador de máxima verossimilhança é o valor de (β_0, β_1) que maximiza a função de verossimilhança.
- Função log-verossimilhança

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 x_i) - log(1 + e^{1+exp(\beta_0+\beta_1 x_i)}) \right]$$

- Geralmente, não existe uma solução análitica e portanto é necessário usar métodos numéricos.

Regressão Logistica Binária

Propriedades do EMV

- O EMV tem, para grandes amostras, distribuição normal;

$$\hat{\boldsymbol{\beta}} \rightarrow N(\boldsymbol{\beta}, \mathbf{I}^{-1})$$

- O EMV é consistente;

$$\hat{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta}$$

- Invariância: Se $\hat{\boldsymbol{\beta}}$ é o EMV de $\boldsymbol{\beta}$, então $g(\hat{\boldsymbol{\beta}})$ é o EMV de $g(\boldsymbol{\beta})$

Regressão Logistica Binária

Interpretação dos parâmetros

- Em um modelo de regressão linear em que $E(Y) = \beta_0 + \beta_1 x_1, \cdots, \beta_p x_p$, o parâmetro de regressão β_r é interpretado como a mudança na resposta média para cada aumento de 1 unicade em x_r, mantendo as outras variáveis no modelo constantes.
- Em um modelo de regressão logistica a interpretação dos parâmetros da regressão precisa levar em conta o fato de que eles estão relacionados à probabilidade de sucesso através de

$$logit(\pi) = \beta_0 + \beta_1 x_1, \cdots, \beta_p x_p$$

- Mantendo as outras variáveis constantes, um aumento de 1 unidade em x_r faz com que $logit(\pi)$ mude por β_r.

Regressão Logistica Binária

Interpretação dos parâmetros

- Considerando a chance de um sucesso em um determinado valor de x são

$$Odds_x = \exp(\beta_0 + \beta_1 x)$$

- Se houver um aumento em $c > 0$ unidade em x, as chances de sucesso se tornam

$$Odds_{x+c} = \exp(\beta_0 + \beta_1 x + c).$$

- Para determinar o quanto as chances de sucesso mudaram por esse aumento de c unidades, encontramos a razão de chances:

$$OR = \frac{Odds_{x+c}}{Odds_x} = \frac{\exp(\beta_0 + \beta_1 x + c)}{\exp(\beta_0 + \beta_1 x)} = \exp(c\beta_1).$$

Regressão Logistica Binária

Interpretação dos parâmetros

- Assim, as chances de sucesso mudam em $\exp(c\beta_1)$ vezes para cada c unidades aumentadas em x.
- Portando, $\exp(\beta_1)$ indica o aumento (ou redução) da razão de chances quando aumentamos em uma unidade a variável x
- Se x for uma variável dummy indicando se o paciente teve um tratamento ou não, o termo $\exp(\beta_1)$ indica o quanto a razão de chances se altera quando o paciente passa pelo tratamento versus quando ele não passa.

Regressão Logistica Binária

Significância dos efeitos das variáveis

- Obtidas as estimativas dos parâmetros β_k, faz-se necessário avaliar a adequação do modelo ajustado;
- O princípio em Regressão Logística é o mesmo usado em regressão linear, isto é, compara-se os valores observados da variável resposta com os valores preditos pelos modelos com e sem a variável sob investigação.
- Em regressão linear, esta comparação é feita por meio da análise de variância, em que um valor grande da soma de quadrados devido à regressão sugere que pelo menos uma, ou talvez todas as variáveis explicativas, sejam importantes.
- Em regressão logistica a comparação pode ser feita por meio de testes como o da razão de verossimilhanças (TRV), em que a função de verossimilhança do modelo sem as variáveis (L_S) é comparada com a função de verossimilhança do modelo com as variáveis (L_C.)

Regressão Logistica Binária

Significância dos efeitos das variáveis

Teste da razão de verossimilhança

$$TRV = -2ln\left[\frac{L_S}{L_C}\right] = 2ln(L_C) - 2ln(L_S)$$

Sob a hipótese nula de que os p coeficientes associados às variáveis no modelo não diferem de zero, a estatística TRV segue uma distribuição qui-quadrado com p graus de liberdade.

Portanto, a rejeição de H_0 nos leva a crer que pelo menos um dos p coeficientes difere de zero.

Regressão Logistica Binária

Significância dos efeitos das variáveis

- Em regressão logística, a estatística

$$D = -2ln \left[\frac{\text{verossimilhança do modelo sob estudo}}{\text{verossimilhança do modelo saturado}} \right]$$

  é denominada **deviance**.
- Um modelo saturado é aquele que contém tantos parâmetros quantos dados existirem.
- Assim, a estatística TRV pode ser vista como a diferença entre duas *deviances*, a do modelo sem as variáveis explicativas e a do modelo com tais variáveis, isto é,

Regressão Logistica Binária

Significância dos efeitos das variáveis

$$TRV = \left[-2ln\left(\frac{\text{verossimilhança do modelo sem as variáveis}}{\text{verossimilhança do modelo saturado}}\right)\right] - \left[-2ln\left(\frac{\text{verossimilhança do modelo com as variáveis}}{\text{verossimilhança do modelo saturado}}\right)\right]$$

de modo que

$$TRV = -2ln\left[\frac{L_S}{L_C}\right] = 2ln(L_C) - 2ln(L_S)$$

Regressão Logistica Binária

Significância dos efeitos das variáveis

- Uma tabela de análise de variância similar a obtida em regressão linear pode ser construida em regressão logística.
- Nesse caso, é denominada de análise de *deviance* (ANODEV).
- O objetivo a ANODEV é obter, a partir de uma sequencia de modelos encaixados, os efeitos de fatores, variáveis e suas interações.
- A partir das *deviances* e de suas diferenças, pode-se usar o teste da razão de verossimilhanças para testar a Significância da inclusão de determinads variáveis, bem como suas interações no modelo.
- Assim, pode-se avaliar o quanto da *deviance* associada ao modelo nulo é explicada pela inclusão de termos no modelo.

Regressão Logistica Binária

Significância dos efeitos das variáveis

- Uma alternativa para testar a Significância dos coeficientes seria o teste de Wald (1943), frequentemente utilizado para testar hipóteses relativas a um único parâmetro β_k, que sob a hipótese nula

$$H_0 : \beta_k = 0$$

a estatística para esse teste é

$$W = \frac{(\hat{\beta}_k)^2}{Var(\hat{\beta}_k)}$$

que, sob H_0, segue a distribuição qui-quadrado com 1 grau de liberdade.

Regressão Logistica Binária

Significância dos efeitos das variáveis

- A comparação de modelos pode também ser realizada por meio de critérios que sumarizam o quão próximas as probabilidades preditas pelo modelo tendem a estar das probabilidades verdadeiras.
- Um desses critérios, o de informação de Akaike (AIC), indica o modelo que minimiza

$$AIC = -2(\text{log verossimilhança} - \text{número de parâmentros do modelo})$$

como sendo o que fornece as melhores probabilidades preditas.

Regressão Logistica Binária

Qualidade do Modelo ajustado

- Selecionado o modelo, o próximo passo é avaliar o quão bem ele se ajusta aos dados, isto é, o quão próximos os valores preditos por este modelo se encontram de seus correspondentes valores observados.
- As estatísticas de teste utilizadas para essa finalidade são, em geral, denominadas estatísticas de qualidade do ajuste.
- Duas estatísticas tradicionais de qualidade do ajuste são:
  i) Q_p, a estatística qui-quadrado de Pearson que é baseada nos resíduos de Pearson:

$$Q_p = \sum_{i=1}^{s}\sum_{j=1}^{2}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$$

Regressão Logistica Binária

Qualidade do Modelo ajustado

ii) Q_L, qui-quadrado *deviance* por se basearnos resíduos *deviance*:

$$Q_L = 2\sum_{i=1}^{s}\sum_{j=1}^{2} n_{ij} ln\left(\frac{n_{ij}}{e_{ij}}\right),$$

em que e_{ij} são as quantidades preditas pelo modelo modelo e definidas por

$$e_{ij} = n_{i+}\hat{\pi}(x_i) \qquad j = 1$$

$$e_{ij} = n_{i+}(1 - \hat{\pi}(x_i)) \qquad j = 2$$

Regressão Logistica Binária

Qualidade do Modelo ajustado

- Sob a hipótese nula de que o modelo se ajusta bem aos dados, Q_p e Q_L seguem distribuição aproximadamente qui-quadrado com os graus de liberdade definidos pela diferença entre o número de subpopulações(linhas da tabela de dados) e o número de parâmetros do modelo.

Regressão Logistica Binária

Utilizando o programa R

- Função `glm` (disponível na instalação básica do R)
- Sua sintaxe se assemelha com a da função `lm` (na forma de definir a estrutura de regressão)

  i) com intercepto

  r <= glm $(Y \sim x_1 + x_2 + ... + x_p)$, family=binomial(link="logit"));
  summary(reg)

  ii) Sem intercepto

  r <= glm $(Y \sim -1 + x_1 + x_2 + ... + x_p)$,family=binomial(link="logit"));
  summary(reg)

Regressão Logistica Binária

Diagnóstico em Regressão Logística

- Pregibon (1981) estendeu os métodos de Diagnóstico utilizados em regressão linear para regressão logística fazendo uso das componentes individuais das estatísticas qui-quadrado de Pearson (Q_p) e *deviance* (Q_L).
- Se em uma tabela de Contingência s x 2 existirem n_{i+} indivíduos em cada uma das s linhas, dos quais n_{i1} apresentam a resposta de interesse $(Y = 1)$, define-se o i-ésimo resíduo de pearson por

$$c_i = \frac{n_{i1} - (n_{i+})\hat{p}(\mathbf{x}_i)}{\sqrt{(n_{i+})\hat{p}(\mathbf{x}_i)[1 - \hat{p}(\mathbf{x}_i)]}}, \qquad i = 1, ..., s$$

com $\hat{p}(\mathbf{x}_i)$ a probabilidade $P(y = 1|\mathbf{x}_i)$ predita pelo modelo para a i-ésima linha (subpopulação).

Regressão Logistica Binária

Diagnóstico em Regressão Logística

- Resíduos excedendo os valores $\pm$ 2,5 (ou $\pm$ 3,0) indicam possível falta de ajuste.
- Quanto ao i-ésimo resíduo deviance, este é definido por

$$d_i = \pm \left[2n_{i1} ln \left(\frac{n_{i1}}{e_{i1}} \right) + 2(n_{i+} - n_{i1}) ln \left(\frac{n_{i+} - n_{i1}}{n_{i+} - e_{i1}} \right) \right]^{\frac{1}{2}}$$

para $i = 1, ..., s$, em que $e_{i1} = (n_{i+})\hat{p}(\mathbf{x}_i)$. O sinal de d_i será definido a partir da diferença $(n_{i+} - e_{i1})$.

- A partir da inspeção dos resíduos *deviance* é possível observar a presença de resíduos não usuais (demasiadamente grandes), bem como a presença de valores atípicos (*outliers*) ou, ainda, padrões sistemáticos de variação indicando a escolha de um modelo possivelmente não muito adequado.

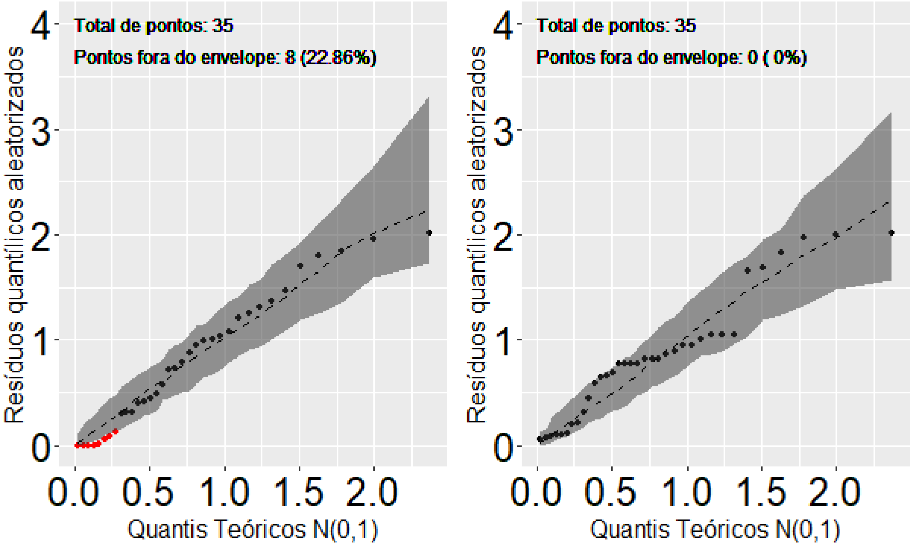
Regressão Logistica Binária

Diagnóstico em Regressão Logística: Métodos auxiliares

Gráfico quantil-quantil com envelope simulado

- Nos casos em que é assumido que a variável resposta segue distribuição normal, é comum que afastamento sérios dessa distribuição sejam verificados por meio do gráfico de probabilidade normal dos resíduos (gráfico quantil-quantil ou Q-Q da normal).
- No contexto de modelos lineares generalizados, em que a distribuições diferentes da normal são consideradas, gráficos similares com envelopes simulados podem ser construídos com os resíduos *deviance*, uma vez que esses resíduos seguem distribuição aproximadamente normal.
- A inclusão do envelope simulado no gráfico Q-Q auxilia a decidir se os pontos diferem significativamente de uma linha reta.
- Para que o modelo ajustado seja considerado satisfatório, é necessário que os resíduos *deviance* estejam dentro do envelope simulado.

Regressão Logistica Binária



Avaliando a Predição do Modelo

Classificação com Regressão logística binária

- Ao ajustarmos um modelo de regressão logística, a modelo vai nos fornecer as probabilidades preditas deu ma determinada observação ter valor 1 (sucesso) ou 0 (fracasso).
- Em muitas situações temos o interesse em classificar as observações em 0 ou 1, com base nas variáveis explicativas.
- Por exemplo com base nas características de um cliente do banco temos o interesse em classifica-lo como potencial pagador ou potencial inadimplente.
- Baseado em um valor pre-definido de corte c, uma forma de fazer essa classificação é através da regra:
    - Caso a probabilidade predita $\pi_i > c$, então a observação i é classificada na categoria sucesso;
    - Caso a probabilidade predita $\pi_i \leq c$, então a observação i é classificada na categoria fracasso

Classificação com Regressão logística binária

- Por exemplo, podemos assumir $c = 0,50$
- Como verificar quão boa ou ruim é essa regra de classificação? Quais os indicadores usados para fazer essa averiguação?
- Geralmente, as medidas de qualidade da classificação estão relacionadas ao grau de acerto das classificações.
- Um indicador comumente empregado é a chamada matriz de confusão
- Essa matriz corresponde a uma tabulação cruzada entre a classificação de acordo com o modelo e a classificação real observada na amostra.

Classificação com Regressão logística binária

Matriz de confusão

Valor Predito	Valor Real	
	0	1
0	Verdadeiro Negativo (VN)	Falso Negativo (FN)
1	Falso Positivo (FP)	Verdadeiro Positivo (VP)

Métricas para avaliar a regra de classificação

- A classificação feita pela regra de decisão (baseada na regressão logística) não é perfeita.
- Ela pode cometer erros: indivíduos que de fato é bom pagador não possuem características x_1 e x_2 típicas de um bom pagador.
- Em consequência, a regra de decisão (que olha apenas os regressores em x) aloca estes indivíduos à classe 0 (inadimplentes)
- Estes são os **falso-negativos**
- Analogamente, vários inadimplentes possuem características típicas de um bom pagador e são alocados pela regra de decisão logística à categoria 1 (bom pagador).
- Estes são os **falso-positivos**

Métricas para avaliar a regra de classificação

- Idealmente, queremos poucos falso-positivos e poucos falso-negativos (ou muitos verdadeiro-positivos e muitos verdadiro-negativos).
- Isto será obtido se tivermos uma pequena probabilidade de ter um falso-positvo (FP) e um falso-negativo (FN).
- No caso de verdadeiro-positivo temos,

$$P(VP) = P(\text{classificado como +}|\text{é +}) = \frac{P(\text{classif + e é +})}{P(\text{é +})}$$

- Esta probabilidade estimada é chamada de sensibilidade e é dada por

$$\textbf{Sensibilidade} = \frac{VP}{VP + FN}$$

Métricas para avaliar a regra de classificação

- Quanto aos verdadeiro-negativos temos,

$$P(VN) = P(\text{classificado como -}|\text{é -}) = \frac{P(\text{classif - e é -})}{P(\text{é -})}$$

- Essa medida é chamada de **especificidade** e é estimada por

$$\textbf{Especificidade} = \frac{VN}{VN + FP}$$

- Observe que $P(VN) + P(FP) = 1$ uma vez que um indivíduo que é negativo, será classificado ou como negativo (corretamente) ou como positivo (incorretamente).
- De modo análogo, $P(VP) + P(FP) = 1$.

Métricas para avaliar a regra de classificação

- **Acurácia:** corresponde ao percentual de casos que são corretamente classificados

$$\textbf{Acurácia} = \frac{VP + VN}{n}$$

- A **Precisão** é dada por

$$\text{Precisão} = P(\text{é } + | \text{classif como } +)$$

- Essa medida é estimada por

$$\text{Precisão} = \frac{VP}{VP + FP}$$

- Observe que a precisão inverte os eventos usados na definição do RECALL (sensibilidade) pois, RECALL é igual a $P(VP) = P(\text{classificado como } + | \text{é } +)$.
- Alta precisão indica que o algoritmo retornou mais resultados relevantes que irrelevantes.

Curva ROC

- Uma das formas de avaliar a performance de classificação a partir de uma regressão logística é utilizando a curva ROC (Receiver Operating Characteristic)
- À medida que aumentamos o valor de corte, aumentamos a sensibilidade e reduzimos a especificidade
- A curva ROC nos fornece um gráfico da sensibilidade (taxa de verdadeiros positivos) versus 1 - especificidade (taxa de falsos positivos), quando aumentamos o valor de corte.
- A área sob a curva, conhecida como AUC (area under the curve), é usada como uma medida de qualidade do ajuste da regressão logística.

Curva ROC

- Usualmente utiliza-se o seguinte critério par classificar o poder discriminatório de um modelo de regressão logística
  - Se AUC = 0,5 o modelo não faz qualquer discriminação entre os indivíduos com e sem a características
  - Se $0,6 \leq AUC \leq 0,7$, o modelo apresenta uma discriminação limitada.
  - Se $0,7 \leq AUC \leq 0,8$, o modelo apresenta uma discriminação aceitável.
  - Se $0,8 \leq AUC \leq 0,9$, o modelo apresenta uma excelente discriminação.
  - Se $AUC \geq 0,9$, o modelo apresenta uma discriminação quase perfeita.

Curva ROC

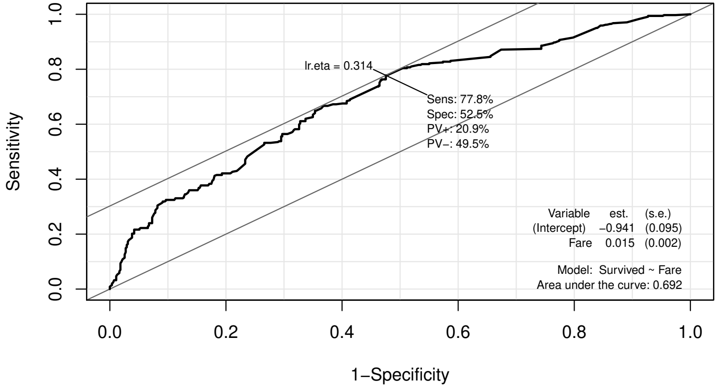


Figura 1: Curva ROC utilizando o pacote Epi do R

Curva ROC

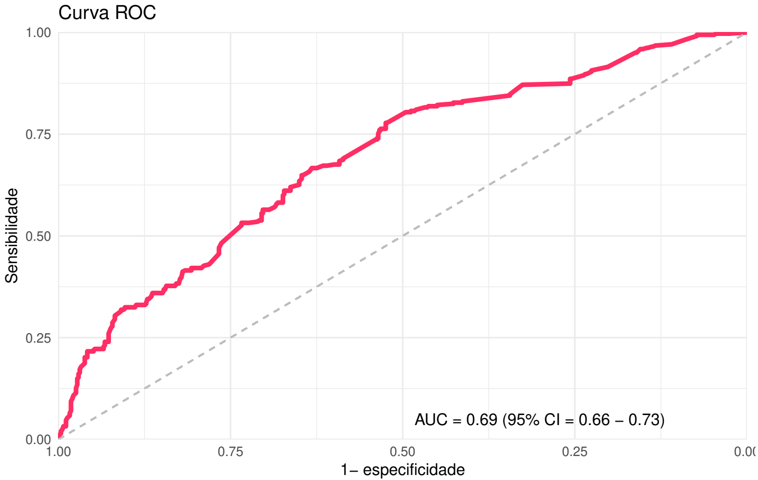


Figura 2: Curva ROC utilizando o pacote pROC do R

Regressão Logística Politômica (Multinomial)

Regressão Logística Multicategórica (Politômica)

Regressão Logística Politômica: a variável resposta com mais de duas categorias.

1. Nominal;
    - Exemplos: escolha da marca de um produto (A,B,C,D e E); escolha do tipo de transporte (carro, ônibus, avião, trem), etc.
    - Modelagem usual: utilizar logit para pares de categorias.
2. Ordinal.
    - Exemplos: qualidade de vida (ruim, razoável, boa, excelete); grau de satisfação (nenhum, muito pouco, pouco, moderado, muito), etc.
    - Modelagem usual: utilizar logit para probabilidades cumulativas.

Regressão Logística Multicategórica (Politômica)

- Nos dois casos o objetivo é modelar a dependência de π_{ij} para o i-ésimo indivíduo na j-ésima categoria,

$$\pi_{ij} = P(Y_i = j), \qquad j = 1, ..., J,$$

em função de um conjunto de covariáveis x (categóricas ou contínuas).

- Os modelos tratam as respostas y, para x fixo, como multinomial.

Regressão Logística Multicategórica (Politômica)

Seja Y a variável resposta com J categorias e seja $\pi_j(\boldsymbol{x}) = P(Y = j|\boldsymbol{x})$, em que $\boldsymbol{x}$ é o conjunto de convariáveis e $\sum_{j=1}^{J} \pi_j(\boldsymbol{x}) = 1$.

Uma extensão da regressão logística pode lidar com a variáveis resposta Politômica:

- Variável resposta (Y) Ordinal $\Longrightarrow$ **Modelo de Odds proporcionais**
- Variável resposta (Y) Nominal $\Longrightarrow$ **Modelo de logitos generalizados**

Modelo Logístico para variável resposta Nominais

- Para variáveis respostas nominais uma extensão do modelo de regressão logística binário fornece um modelo logístico usual para cada par de categorias da resposta.
- Os modelos usam simultaneamente todos os pares de categorias espeficando a chance do resultado em uma categoria em vez da outra.
- Como a resposta tem escala nominal, a ordem das categorias é irrelevante.
- Seja J o número de categorias de Y e seja $\{\pi_1, ..., \pi_J\}$ a probabilidade de resposta, tal que $\sum_{J} \pi_j = 1$.
- Modelos logistico para variável nominal compara cada categoria com uma categoria base. Geralmente a ultima categoria (J) é a base.

Modelo Logístico para variável resposta Nominais

Os logits são:

$$\log\left[\frac{\pi_j}{\pi_J}\right] = \log\left[\frac{P(Y=j)}{P(Y=J)}\right], \qquad j = 1, ..., J-1$$

O modelo logit com um preditor x é :

$$\log\left[\frac{\pi_j}{\pi_J}\right] = \log\left[\frac{P(Y=j)}{P(Y=J)}\right] = \beta_{0j} + \beta_j x, \qquad j = 1, ..., J-1$$

- O modelo tem $J-1$ equações (logits), com parametros independentes para cada.

Modelo Logístico para variável resposta Nominais

- Por exemplo, para uma variável resposta com 3 categorias $(J = 3)$, o modelo usa dois logits

$$\log\left[\frac{\pi_1}{\pi_3}\right] = \log\left[\frac{P(Y=1)}{P(Y=3)}\right] \text{ e } \log\left[\frac{\pi_2}{\pi_3}\right] = \log\left[\frac{P(Y=2)}{P(Y=3)}\right]$$

- Os efeitos variam de acordo com a categoria comparada com a base.

Modelo Logístico para variável resposta Nominais

- Embora o modelo considere $J-1$ logitos dentre todos os C_2^J possíveis pares de categorias, os $J-1$ logitos considerados pelo modelo modelo determinam os logitos para todos os outros pares de categorias.

Seja a e b categorias quaisquer, então

$$\begin{aligned}
\log\left[\frac{\pi_a}{\pi_b}\right] &= \log\left[\frac{\pi_a/\pi_J}{\pi_b/\pi_J}\right] = \log\left[\frac{\pi_a}{\pi_J}\right] - log\left[\frac{\pi_b}{\pi_J}\right] \\
&= (\beta_{0a}+\beta_a x) - (\beta_{0b}+\beta_b x) = (\beta_{0a}-\beta_{0b}) + (\beta_a-\beta_b)x
\end{aligned}$$

- Assim, a equação para as categorias a e b tem a forma $\beta_0+\beta x$, com o intercepto $\beta_0=(\beta_{0a}-\beta_{0b})$ e $\beta=(\beta_a-\beta_b)$

Modelo Logístico para variável resposta Nominais

No caso geral, que temos várias covariáveis, temos

$$\log\left[\frac{\pi_j}{\pi_J}\right] = \log\left[\frac{P(Y=j)}{P(Y=J)}\right] = \beta_{0j} + \boldsymbol{\beta'}_j\boldsymbol{x}, \qquad j = 1, ..., J-1 \tag{1}$$

As probabilidades de resposta π_j

De (5) temos que

$$\frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} = \exp\{\beta_{0j} + \boldsymbol{\beta}'_j\boldsymbol{x}\}, \qquad j = 1, ..., J-1$$

$$\pi_j(\boldsymbol{x}) = \pi_J(\boldsymbol{x})\exp\{\beta_{0j} + \boldsymbol{\beta}'_j\boldsymbol{x}\}, \qquad j = 1, ..., J-1 \tag{2}$$

Modelo Logístico para variável resposta Nominais

As probabilidades de resposta π_j (contiunação)

Como $\sum_{j=1}^{p} \pi_j = 1$ temos que

$$\pi_J(\boldsymbol{x}) = 1 - \sum_{j=1}^{p-1} \pi_J(\boldsymbol{x}) \exp\{\beta_{0j} + \boldsymbol{\beta}_j' \boldsymbol{x}\} \Longrightarrow$$

$$\pi_J(\boldsymbol{x}) + \pi_J(\boldsymbol{x}) \sum_{j=1}^{p-1} \exp\{\beta_{0j} + \boldsymbol{\beta}_j' \boldsymbol{x}\} = 1 \Longrightarrow$$

$$\pi_J(\boldsymbol{x}) = \frac{1}{1 + \sum_{j=1}^{p-1} \exp\{\beta_{0j} + \boldsymbol{\beta}_j' \boldsymbol{x}\}} \tag{3}$$

Modelo Logístico para variável resposta Nominais

As probabilidades de resposta π_j (contiunação)

Substituindo (3) em (2) temos

$$\pi_j(\boldsymbol{x}) = \frac{\exp\{\beta_{0j} + \boldsymbol{\beta}_j' \boldsymbol{x}\}}{1 + \sum_{j=1}^{p-1} \exp\{\beta_{0j} + \boldsymbol{\beta}_j' \boldsymbol{x}\}}, \qquad j = 1, ..., J-1$$

Modelo Logístico para variável resposta Nominais - Exemplo

Considere os dados em que se deseja avaliar se o programa de aprendizado que as crianças preferem estaria associado com a escola e o período escolar.

No R

Modelo Logístico para variável resposta Nominais - Exemplo

Tabela 4 - *Odds* associadas aos logitos 1 e 2.

		logito 1	*logito 2*
Escola	Período	*odds* = π_{hi1}/π_{hi3}	*odds* = π_{hi2}/π_{hi3}
1	Padrão	$\exp\{\beta_{01}+\beta_{31}\}$	$\exp\{\beta_{02}+\beta_{32}\}$
1	Integral	$\exp\{\beta_{01}\}$	$\exp\{\beta_{02}\}$
2	Padrão	$\exp\{\beta_{01}+\beta_{11}+\beta_{31}\}$	$\exp\{\beta_{02}+\beta_{12}+\beta_{32}\}$
2	Integral	$\exp\{\beta_{01}+\beta_{11}\}$	$\exp\{\beta_{02}+\beta_{12}\}$
3	Padrão	$\exp\{\beta_{01}+\beta_{21}+\beta_{31}\}$	$\exp\{\beta_{02}+\beta_{22}+\beta_{32}\}$
3	Integral	$\exp\{\beta_{01}+\beta_{21}\}$	$\exp\{\beta_{02}+\beta_{22}\}$

Modelo Logístico para variável resposta Nominais - Exemplo

	logito 1	logito 2
entre períodos	individual / sala de aula	grupo / sala de aula
$\widehat{OR}_{P/I}$	$e^{\hat{\beta}_{31}} = 2{,}11$	$e^{\hat{\beta}_{32}} = 2{,}10$

Entre aprendizado individual ou em sala de aula

- Odds de preferência pelo "individual"entre alunos do período padrão é $\approx$ o dobro da dos alunos do período integral.

Modelo Logístico para variável resposta Nominais - Exemplo

Entre aprendizado em grupo o em sala de aula

- Odds de preferência pelo "grupo"entre alunos do período padrão é $\approx$ o dobro da dos alunos do período integral.

Entre aprendizado individual ou em grupo

- Odds de preferência entre esses dois métodos de aprendizado não diferiu entre os alunos.

Modelo Logístico para variável resposta Nominais - Exemplo

entre escolas	logito 1 individual / sala de aula	logito 2 grupo / sala de aula
$\widehat{OR}_{2/1}$	$e^{\widehat{\beta}_{11}} = 2{,}95$	$e^{\widehat{\beta}_{12}} = 1{,}19$
$\widehat{OR}_{3/1}$	$e^{\widehat{\beta}_{21}} = 3{,}72$	$e^{\widehat{\beta}_{22}} = 1{,}93$
$\widehat{OR}_{3/2}$	$e^{\widehat{\beta}_{21}-\widehat{\beta}_{11}} = 1{,}26$	$e^{\widehat{\beta}_{22}-\widehat{\beta}_{12}} = 1{,}61$

Entre aprendizado individual ou em sala de aula

- Odds de preferência pelo "individual"entre alunos da escola 2 é $\approx$ 3 vezes a dos alunos da escola 1;
- Odds de preferência pelo "individual"entre alunos da escola 3 é $\approx$ 4 vezes a dos alunos da escola 1;
- Odds de preferência pelo "individual"entre alunos da escola 3 é $\approx$ 1,3 vezes a dos alunos da escola 2;

Modelos de Regressão Logística para variáveis resposta Ordinais

- Como levar em consideração na modelagem o fato das categorias serem ordenadas?
- Quando as categorias de respostas são ordenadas, os logitos podem utilizar a ordenação;
- $P(Y \leq j)$ representa a probabilidade de que a resposta está na categoria j ou abaixo (na categoria 1,2,..., ou j). São chamado de **probabilidade acumulada**;
- Por exemplo, com quatro categorias as probabilidades acumuladas são:
  - $P(Y = 1)$
  - $P(Y \leq 2) = P(Y = 1) + P(Y = 2)$
  - $P(Y \leq 3) = P(Y = 1) + P(Y = 2) + P(Y = 3)$
  - $P(Y \leq 4) = 1$

Modelos de Regressão Logística para variáveis resposta Ordinais

- A ordem na formação das probabilidades acumuladas reflete a ordem na escala da resposta, uma vez que

$$P(Y \leq 1) \leq P(Y \leq 2) \leq P(Y \leq 3) \leq ... \leq P(Y \leq J) = 1$$

- A chance da resposta na categoria j ou abaixo é a razão:

$$\frac{P(Y \leq j)}{P(Y > j)}$$

Modelos de Regressão Logística para variáveis resposta Ordinais

- Por exemplo, quando as chances são iguais a 2,5, a probabilidade da resposta na categoria j ou abaixo é igual a 2,5 vezes a probabilidade da resposta acima da categoria j.
- Cada probabilidade acumulada pode se transformar em uma chance.
- Os logitos das probabilidades acumuladas, chamados de logitos acumulados;
- Um modelo logístico para uma variável resposta ordinal usa os logits das probabilidades acumuladas.

Modelos de Regressão Logística para variáveis resposta Ordinais

Por exemplo, uma variável resposta com $J = 4$ categorias

$$logit[P(Y \leq 1)] = \log\left[\frac{P(Y=1)}{P(Y>1)}\right] = \log\left[\frac{P(Y=1)}{P(Y=2)+P(Y=3)+P(Y=4)}\right]$$

$$logit[P(Y \leq 2)] = \log\left[\frac{P(Y\leq 2)}{P(Y>2)}\right] = \log\left[\frac{P(Y=1)+P(Y=2)}{P(Y=3)+P(Y=4)}\right]$$

$$logit[P(Y \leq 3)] = \log\left[\frac{P(Y\leq 3)}{P(Y>3)}\right] = \log\left[\frac{P(Y=1)+P(Y=2)+P(Y=3)}{P(Y=4)}\right]$$

Modelos de Regressão Logística para variáveis resposta Ordinais

- Como a probabilidade acumulada final necessariamente é igual a 1, nós a excluímos do modelo.
- Cada logit acumulado usa todas as J categorias de respostas.
- Cada logit acumulado considera a resposta como binária avaliando se a resposta está na parte inferior ou superior da escala, onde inferior e superior tem uma definição diferente para cada logit acumulado.

Modelos de Regressão Logística para variáveis resposta Ordinais

Modelo de Odds proporcionais

Um modelo que simultaneamente usa todos os logits comulativos é

$$logit[P(Y \leq j|\boldsymbol{x})] = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x} \tag{4}$$

- Cada logito cumulativo tem seu próprio intercepto.
- Os α_j são incrementos em j, quando $P(Y \leq j|\boldsymbol{x})$ aumenta em j para $\boldsymbol{x}$ fixo.
- Este modelo tem os mesmos efeitos $\boldsymbol{\beta}$ para cada logito.

Modelos de Regressão Logística para variáveis resposta Ordinais

Modelo de Odds proporcionais

- O modelo (4) satisfaz

$$logit[P(Y \leq j|\boldsymbol{x}_1)] - logit[P(Y \leq j|\boldsymbol{x}_2)] = \log\left[\frac{\frac{P(Y \leq j|\boldsymbol{x}_1)}{P(Y > j|\boldsymbol{x}_1)}}{\frac{P(Y \leq j|\boldsymbol{x}_2)}{P(Y > j|\boldsymbol{x}_2)}}\right] = \boldsymbol{\beta}'(\boldsymbol{x}_1 - \boldsymbol{x}_2)$$

- Uma Odds ratio de probabilidades acumuladas é chamada de Odds ratio acumulada.
- A Odds da resposta $\leq j$ em $\boldsymbol{x} = \boldsymbol{x}_1$ é $exp\{\boldsymbol{\beta}'(\boldsymbol{x}_1 - \boldsymbol{x}_2)\}$ vezes a Odds em $\boldsymbol{x} = \boldsymbol{x}_2$.
- A $\log$ da odds acumulada é proporcional a distância entre $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$. A mesma constante de proporcionalidade é aplicada para cada logito.

Modelos de Regressão Logística para variáveis resposta Ordinais

Modelo de Odds proporcionais

- Devido a essa propriedade (4) é conhecido como Modelo de Odds Proporcionais.
- O ajuste do modelo trata as observações como independentes e proveniente de uma distribuição multinomial.
- Na estimativa dos parâmetros usa todas a probabilidades acumuladas de uma só vez, por isso uma única estimativa β para o efeito de x, em vez de três estimativas separadas que obteríamos ajustando o modelo separadamente para cada probabilidade acumulada.
- Se inverter a ordem das categorias de Y, isto é, listando da mais alta para a mais baixa em vez da mais baixa para mais alta, o ajuste do modelo é o mesmo, mas o sinal de β se inverte.

Modelos de Regressão Logística para variáveis resposta Ordinais

Modelo de Odds não proporcionais

Um modelo que simultaneamente usa todos os logits comulativos é

$$logit[P(Y \leq j|\boldsymbol{x})] = \alpha_j + \boldsymbol{\beta_j'}\boldsymbol{x} \tag{5}$$

- Cada logito cumulativo tem seu próprio intercepto.
- Os α_j são incrementos em j, quando $P(Y \leq j|\boldsymbol{x})$ aumenta em j para $\boldsymbol{x}$ fixo.
- $\boldsymbol{\beta}_j = (\beta_{1j}, \cdots, \beta_{pj})$ é o vetor de parâmetros de regressão, de modo que, β_{kj} descreve, para o logito j, o efeito da covariável $k, k = 1, \cdots, p$.
- O modelo (5) assume que os efeitos das covariáveis diferem entre os logitos cumulativos (propriedade de chances não proporcionais).

Modelos de Regressão Logística para variáveis resposta Ordinais

Modelo de Odds não proporcionais

- Agresti(2010) observa que o modelo (5) pode ser mais indicado para dados com poucas covariáveis, sendo todas de natureza categórica, do que para dados com diversas covariáveis, com algumas delas contínuas.
- E necessário testar a hipótese de que os efeitos das covariáveis não diferem entre os logitos, isto é, testar se $\boldsymbol{\beta}_j = \boldsymbol{\beta}$, para $j = 1, \cdots, r-1$;
- Uma estatística de teste que pode ser utilizada é a da razão de verossimilhanças, que sob $H_0 : \boldsymbol{\beta}_j = \boldsymbol{\beta}$, segue distribuição aproximadamente qui-quadrado com os graus de liberdade igual a diferença entre os números de parâmetros dos modelos sob as hipóteses H_0 e $H_1 : \boldsymbol{\beta}_j \neq \boldsymbol{\beta}$.

Modelos de Regressão Logística para variáveis resposta Ordinais

Modelo de Odds não proporcionais

- No caso de a hipótese nula ser rejeitada para todas as covariáveis uma opção é utilizar o modelo logitos cumulativos (5);
- Se a hipótese nula for rejeitada para parte das covariávies, foi proposto o modelo de chances proporcionais parciais.

Modelos de Regressão Logística para variáveis resposta Ordinais - Exemplo

Para abordar essa situação, considere os dados a seguir em que se deseja avaliar se o grau de melhora de pacientes com artrite estaria associado com sexo e tratamento.

		Grau de melhora			
Sexo	Tratamentos	Acentuada	Alguma	Nenhuma	Totais
F	A	16	5	6	27
F	Placebo	6	7	19	32
M	A	5	2	7	14
M	Placebo	1	0	10	11

Modelos de Regressão Logística para variáveis resposta Ordinais - Exemplo

Como $J = 3$ temos dois logitos acumulativos.

O primerio logit é o $log(Odds)$ de melhora acentuada para alguma ou nenhuma melhora

$$\begin{aligned} logit(\theta_{hi1}) &= logit[P(Y \leq 1)] = log\left[\frac{P(Y=1)}{P(Y>1)}\right] = log\left[\frac{\pi_{hi1}}{\pi_{hi2} + \pi_{hi3}}\right] \\ &= log\left[\frac{P(Y=1)}{P(Y=2) + P(Y=3)}\right] = \beta_{01} + \beta_1 x_1 + \beta_2 x_2 \end{aligned}$$

Modelos de Regressão Logística para variáveis resposta Ordinais - Exemplo

O segundo logit é o $log(Odds)$ de melhora acentuada ou alguma melhora para nenhuma melhora

$$
\begin{aligned}
logit(\theta_{hi2}) &= logit[P(Y \leq 2)] = log\left[\frac{P(Y \leq 2)}{P(Y > 2)}\right] = log\left[\frac{\pi_{hi1} + \pi_{hi2}}{\pi_{hi3}}\right] \\
&= log\left[\frac{P(Y = 1) + P(Y = 2)}{P(Y = 3)}\right] = \beta_{02} + \beta_1 x_1 + \beta_2 x_2
\end{aligned}
$$

Modelos de Regressão Logística para variáveis resposta Ordinais - Exemplo

Tabela 4 -*Odds* associadas ao MOP ajustado.

Sexo	Tratamentos	$\pi_{hi1}/(\pi_{hi2}+\pi_{hi3})$	$(\pi_{hi1}+\pi_{hi2})/\pi_{hi3}$
F	A	$\exp\{\beta_{01}+\beta_1+\beta_2\}$	$\exp\{\beta_{02}+\beta_1+\beta_2\}$
F	Placebo	$\exp\{\beta_{01}+\beta_1\}$	$\exp\{\beta_{02}+\beta_1\}$
M	A	$\exp\{\beta_{01}+\beta_2\}$	$\exp\{\beta_{02}+\beta_2\}$
M	Placebo	$\exp\{\beta_{01}\}$	$\exp\{\beta_{02}\}$

Modelos de Regressão Logística para variáveis resposta Ordinais - Exemplo

Melhora acentuada vs alguma ou nenhuma melhora

- a Odds de melhora acentuada entre as mulheres é $\exp(\hat{\beta}_1) \approx 4$ **vezes** a dos homens.
- a Odds de melhora acentuada entre os pacientes sob tratamento A é $\exp(\hat{\beta}_2) \approx 6$ **vezes** a daqueles sob placebo.

Melhora acentuada ou alguma vs nenhuma melhora

- a Odds de melhora acentuada entre as mulheres é $\exp(\hat{\beta}_1) \approx 4$ **vezes** a dos homens.
- a Odds de melhora acentuada entre os pacientes sob tratamento A é $\exp(\hat{\beta}_2) \approx 6$ **vezes** a daqueles sob placebo.

Especialização em *Data Science* e Estatística Aplicada

Módulo IV - Análise de Dados Categóricos

Prof Dr Márcio Augusto Ferreira Rodrigues

marcioaugusto@ufg.br

IME
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

FEN
FACULDADE DE ENFERMAGEM

UFG
UNIVERSIDADE FEDERAL DE GOIÁS