PUC Minas
Virtual

Interpretação dos Parâmetros Regressão Multinomial

Nesta aula iremos entender sobre a interpretação dos parâmetros de uma regressão multinomial

Regressão Multinominal

- Seja Y uma variável aleatória categórica com J categorias.
- Seja $\pi_j(x) = \Pr(Y = j \mid x), com\ \sum_j \pi_j(x) = 1$
- O modelo compara cada categoria j com uma categoria de referência J, totalizando $\binom{J}{2}$ combinações.

$$\eta = \log\left(\frac{\pi_j(x)}{\pi_k(x)}\right) = \alpha_j + \beta'_j x = \frac{1}{1+e^{-(\alpha_j+\beta' j x)}}$$

SEMPRE IREMOS OLHAR EM RELAÇÃO A UMA CATEGORIA DE REFERÊNCIA

PUC Minas Virtual

Como já vimos, tanto a regressão logística quan a regressão multinomial terá grande aplicação por ser um modelo altamente interpretável.

Essa interpretação sera dada sempre em relação a uma categoria de referencia , pois devemos lembrar que na multinomial iremos combinar as categorias de 2 a 2.

Entao devemos considerar agora que eta sera igual o log da probabilidade de ocorrência da classe j sobre a a probabilidade de ocorrência da classe k que sera = ao nosso componente sistemático

Regressão Multinominal / Binomial

Considerando Y com 3 categorias:

Modelo necessita de 2 funções.

Comparação de categorias:

- Y =0 -> Referencia
- Comparar com Y =1 e Y =2.

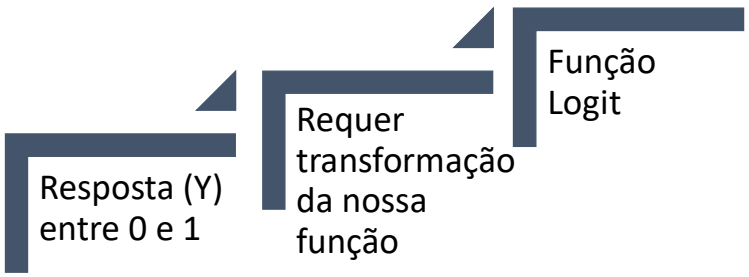
$$g_1(x) = \ln\left(\frac{P(Y=1)|\,x}{P(Y=0)|\,x}\right) = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1p}x_p = x'\beta_1$$

$$g_2(x) = \ln\left(\frac{P(Y=2)|\,x}{P(Y=0)|\,x}\right) = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2p}x_p = x'\beta_1$$

PUC Minas Virtual

Em outras palavras quando temos 3 categorias teremos então que olhar a probabilidade de ocorrência da categoria 1 e 2 em relação a nossa categoria de referencia. Ou seja o modelo nesse caso necessita de duas funções.

Regressão Multinominal



$$E(y) = \frac{e^{x\prime\beta}}{1 + e^{x\prime\beta}} = \frac{1}{1 + e^{-x\prime\beta}}$$

PUC Minas Virtual

Entao temos que a resposta continuara sendo igual a probabilidade do sucesso, ou nesse caso, da probabilidade do numerador em relação a categoria de referencia.]seu valor continua limitante entre 0 e 1.

E por isso faremos da mesma forma a linearização por meio da logit .

Regressão Multinominal

A regressão logística pode ser linearizada:

$$\eta = x'\beta$$

Ser o preditor linear, onde η é definido pela transformação logit.

$$\eta = ln\ \frac{p}{1\ -p}$$

A razão $\frac{p}{1-p}$ é chamada de chance odds .

PUC Minas Virtual

O que estamos falando aqui, é que continuaremos com todo conceito anteriormente trabalhado em relação a trnaformação linear e a chance.

Então queremos linearizar y , ou a prob de encontrar sucesso que é dada por x'B.

Entao a função que lineariza é eta, que é definido por essa transformação ln p/(p-1).

Se vocês relembrarem a aula de conceito do GLM iremos ter vários etas diferentes , para cada glm's diferentes. E todos podem ser vistos como essa linearização para conseguirmos utilizar assim a mesma estrutura linear.

A transformação é necessária para conseguirmos entender e interpretar os coeficientes
Com essa transformação temos a regressão logística expressa de maneira linear.
Fizemos uma linearização

Essa razão é chamado de chance pois é a razão de probabilidade. Razão do sucesso sobre o fracasso

Então fazendo essa transformação conseguimos linearizar a nossa função logística por meio da transformação logit. Onde atribuímos eta (n)
Vamos atribuir o preditor linear = a n, e n é definido pela transformação ln

Estimação de Parâmetros

- A estimação dos parâmetros de $x'_i\beta$ é realizada a partir do método de máxima verossimilhança;
- Como nosso dados seguem a distribuição de Binomial, então a distribuição de probabilidade é dada por:

$$f_i\ (y_i; n; p) = \binom{n}{y} p^y (1\ - p)^{n\ -y}$$

- Logo a função de verossimilhança para v.a. independentes pode ser dada por:

$$\mathrm{L}(y_1, y_2, y_3, \dots, y_n, \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \binom{n}{y} p^y (1\ - p)^{n\ -y}$$

PUC Minas Virtual

A estimação de parâmetros também sera realizada pelo método da máxima versossimilhança.

Porema gora iramos penasr aqui na distribuição binomial.

NA máximo verossimilhança trabalhamos a partir do produtório das função da distribuição de probabilidade de binomial podemos calcular pela função de verossimilhança.

Como os nossos dados seguem a distribuição binomial que pode ser dada pela função de probabilidade .....

Calculando o produtorio da distribuição de binomial fazemos o calculo dos coeficientes computacionalmente

A partir do produtório das função da distribuição de probabilidade de binomial podemos calcular pela função de verossimilhança.

Interpretação dos Parâmetros – ODDS RATIO

Razão entre a probabilidade de um evento ocorrer (p(Y=1)) e a probabilidade de não ocorrer (p(Y=0)).

$$\frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{e^{\hat{\beta}_0}} = e^{\hat{\beta}_1}$$

Chance de sucesso em relação à chance de fracasso.

$$g_1(x) = \ln\left(\frac{P(Y=1)|\,x}{P(Y=0)|\,x}\right) = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1p}x_p = x'\beta_1$$

$$g_2(x) = \ln\left(\frac{P(Y=2)|\,x}{P(Y=0)|\,x}\right) = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2p}x_p = x'\beta_1$$

Interpretar as odds (chances) e os odds ratios (razões de chances) na regressão logística é fundamental para entender o impacto das variáveis independentes nas probabilidades do evento de interesse.

Continuaremos a utilziar da mesma forma o odds ratio, mas devemos lembrar que ele sera sempre interpretada a uma classe de referencia.

Pensando no exemplo do plano de saúde

Ex: Estudo da escolha de um plano de saúde.

Y : Tipos de plano (A,B,C)

X: Idade, tamanho da família,

renda, etc.

Queremos estudar o que leva as pessoas a selecionarem o plano A B C

Entao vamos considerar a renda A e B

Temos como categoria de referencia o plano A

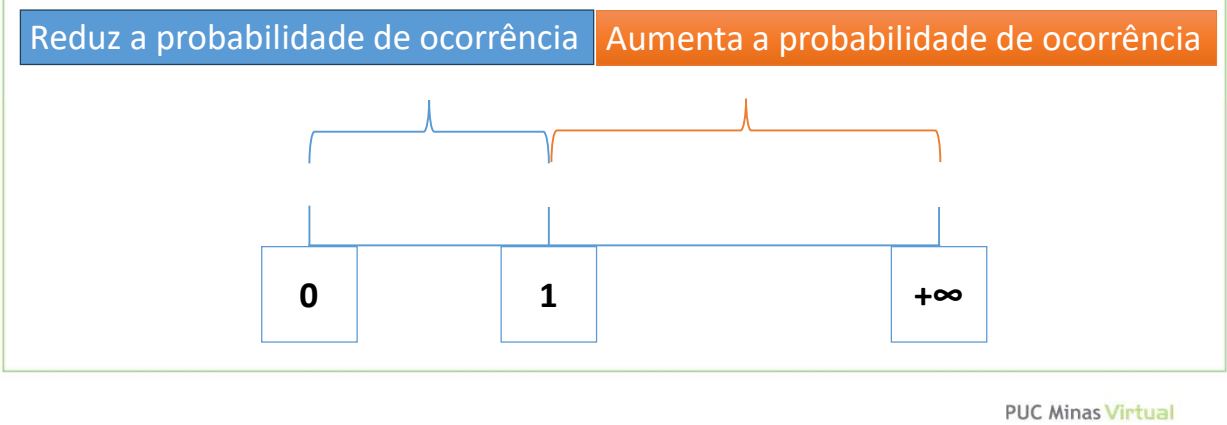
Entao teremos como resposta a probabilidade de selecionar o plano B em relação ao plano a quando temos uma determinada configuraçõ das covariáveis

E em seguida a probabilidade de selecionar o plano C em relação ao plano A.

Observe que a referencia ira depender da resposta que vc quer obter.

Interpretação dos Parâmetros – ODDS RATIO

$$\widehat{O_R} = \frac{odds_{x_i+1}}{odds_{x_i}} = e^{\hat{\beta}_1}$$



PUC Minas Virtual

Entao a interpretação do odds ratio continuara a mesma tmabem, mas olhando sempre as categorias de 2 a 2.

Temos que o e elevado a b1 corresponde ao acréscimo na chance de ocorrer a respos de y = 1. lembrando que nesse caso temso apenas duas possibilidade y =1 e y = 0

Para iniciar temos que ter em mente que vamos ter valores entre 0 e + inf.

Quando esse numero esta entre zero e um, na verdade a presença ou acréscimo desse fator ira diminuir a probabilidade de sucesso.

E quando for maior que 1 irá aumentar a probabilidade de sucesso.

Quando temos o odds ratio =1 temso que o fator não ira influenciar a resposta. No exemplo do covid , vamos imaginar que estamos testando a aqui a covariável da doença pre existente diabetes , teríamos que a razão do odds foi de 1, logo a

pessoa ter ou não diabetes não influenciaria no caso do óbito .

Por exemplo se temos um valor = a 5. vamos dizer que temos 400% de chance de ir a óbito pois 5 -1 = 4 *100 = 400%

E se tivesse mso por exemplo 0,4 – teríamos 60% de chance de sobreviver, funcionaria como um fator de proteção e reduziria a chace de morte. E mais uma vez pegamos 1 – 0,4 = 0,6 *100

Pressupostos da Regressão Multinominal

- A variável reposta precisa ser qualitativa, com mais de duas categorias
- As preditoras podem ser quantitativas ou categóricas
- Ausência de autocorrelação
- Relação linear entre o vetor das variáveis explicativas X e a variável independente Y;
- Ausência de correlação entre os resíduos
- Assume que as observações são independentes
- Ausencia de multicolinearidade

Entao tínhamos vários critérios a serem satisfeitos para que nosso modelos funcionassem bem em cima dos nossos bancos de dados

Os pressupostos garantem que a regressão irá funcionar bem em relação aos nossos dados a regressão multinomial também tem pressupostos

A principal condição é que nossa variavel reposta seja

qualitativa, com mais de duas categorias. Na verdade sera a única difereça em relação ao verificado anteriormente na logistica

As preditoras podem ser quantitativas ou catergoricas (transforamdas em binarias ou dummy)

Assume que as observações sao independentes, ou seja que uma observação nao afeta a outra

PARÂMETROS DOS MODELOS

Verificar a significância das variáveis do modelo

Teste de hipótese para determinar se a variável preditora do modelo é significativamente relacionada com variável resposta do modelo

- Teste de Wald
- Teste de Razão de verossimilhança

E nos testes e validação do modelo teremos exatamente as mesmas coisas, que será padrão para todos GLM's.

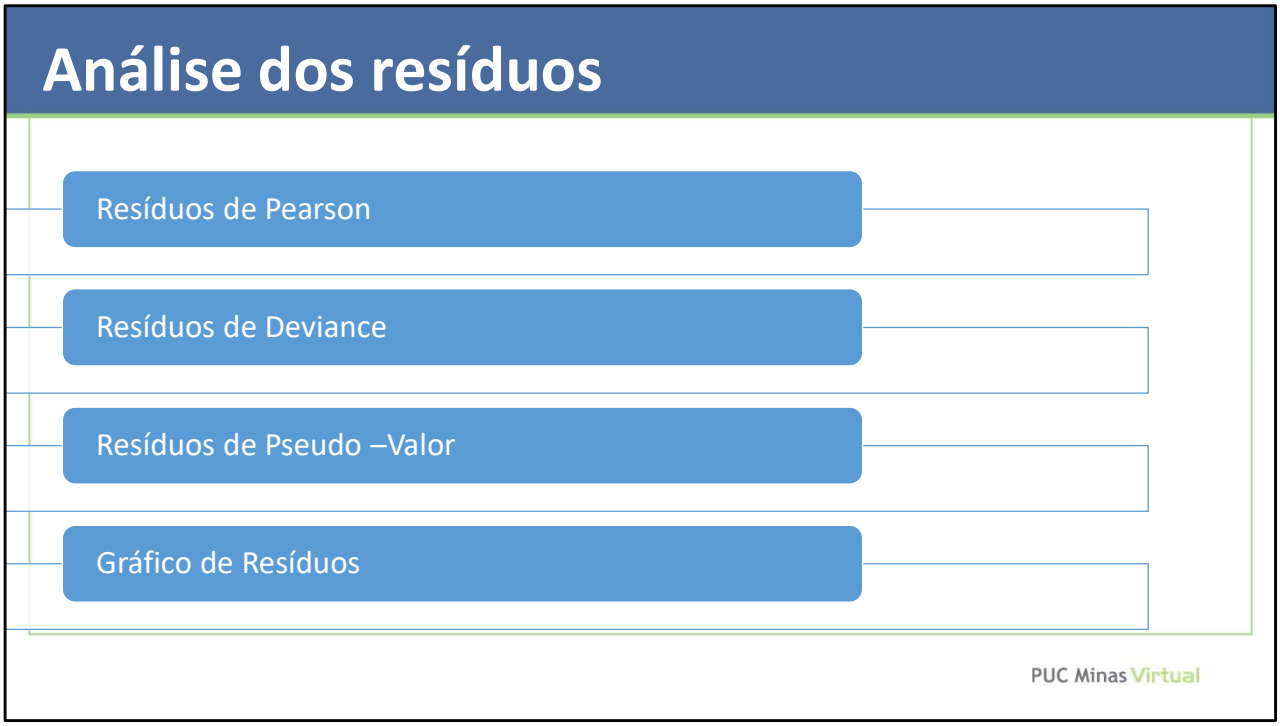
Entao primeiro irmos verificar a significância dos coeficientes estimados, a partir da teste de wald ou da razão da máximo verossimilhança;
Determinei os coeficientes mas será que eles são significativos?

A variável que eu inclui no modelo , é significativa para o modelo?

O modelo que inclui a variável em questão nos diz mais sobre a variável resposta do que o modelo que não inclui essa variável? - > essa variável é importante para o modelo?

Para isso temos dois tipos de este, o teste de razão de verossimilhança e o teste de wald.

Teste de verossimilhança:
Rejeita hipótese nula -> as variáveis são significativas



A avaliação dos resíduos em uma regressão logística é importante para verificar se o modelo está adequadamente ajustado aos dados e se atende às suposições subjacentes. Aqui estão algumas maneiras de avaliar os resíduos em uma regressão logística:

1.Resíduos de Pearson: Os resíduos de Pearson são calculados subtraindo a estimativa da probabilidade observada de ocorrência de um evento da probabilidade real de ocorrência desse evento. Em uma regressão logística, eles são usados para verificar se há heterogeneidade de variância, ou seja, se a variância dos resíduos é constante em toda a faixa de valores previstos. Um gráfico de resíduos de Pearson versus os valores ajustados pode ser usado para detectar essa heterogeneidade.

2.Resíduos de Deviance: Os resíduos de deviance são calculados como a diferença entre a deviance do modelo ajustado e a deviance do modelo saturado (modelo que inclui todas as variáveis independentes). Eles são úteis para avaliar a adequação global do modelo, semelhante ao desvio residual em modelos lineares.

3.Resíduos de Pseudo-Valor: Os resíduos de pseudo-valor são calculados como a diferença entre os valores observados e esperados da variável dependente transformada. Eles são úteis para detectar pontos de influência nos dados.

4.Gráfico de Resíduos: Um gráfico de resíduos, como um gráfico de resíduos versus valores ajustados ou versus variáveis independentes, pode ser útil para detectar

padrões nos resíduos, como não-linearidade ou heterocedasticidade.

5. **Testes de Ajuste do Modelo**: Além de examinar os resíduos individualmente, também é importante considerar testes de ajuste do modelo, como o teste de razão de verossimilhança mencionado anteriormente. Esses testes ajudam a determinar se o modelo se ajusta significativamente melhor aos dados do que um modelo mais simples.

6. **Teste de Influência**: Além disso, é útil realizar testes de influência para identificar pontos de dados que têm um impacto desproporcional no ajuste do modelo. Isso pode incluir o teste de DFBETA para avaliar a influência de cada observação nos parâmetros do modelo.

Ao avaliar os resíduos em uma regressão logística, é importante considerar não apenas as estatísticas numéricas, mas também examinar visualmente os padrões nos gráficos de resíduos e considerar a interpretação substancial dos resultados. Isso ajuda a garantir que o modelo seja apropriado para a análise dos dados em questão.


PUC Minas
Virtual