Disciplina:

MODELOS ESTATÍSTICOS

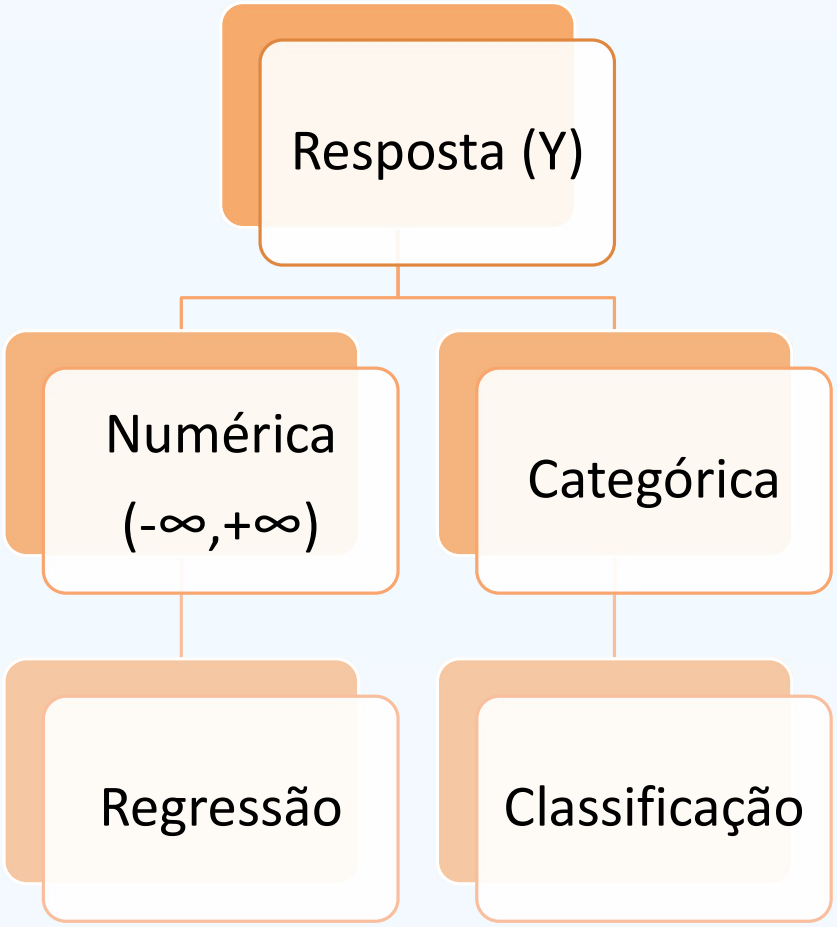
Professor: Anaíle Mendes Rabelo

PUC Minas

Regressão Logística

PUC Minas

Modelos de Machine Learning



PUC Minas

Regressão Logística

Variáveis Categóricas

Binárias

Estudar a probabilidade de ocorrência

Obtemos uma resposta numérica (probabilidade)

Probabilidade Valores estão entre 0 e 1

PUC Minas

Quando utilizar a Regressão logística

VARIÁVEL DEPENDENTE COM DISTRIBUIÇÃO BINOMIAL

REPROVAÇÃO NO TESTE DE HOMOCEDASTICIDADE

RESÍIDUOS NÃO TEM DISTRIBUIÇÃO NORMAL

PUC Minas

Regressão logística

- Técnica de classificação usada para prever uma resposta qualitativa

Exemplos

- Fraudes – Probabilidade da transação ser fraudulenta ou não. Classificação usada para prever uma resposta qualitativa (binária)
- Marketplace - probabilidade do cliente comprar ou não a mercadoria

PUC Minas

Pressupostos da Regressão Logistica

A variável reposta precisa ser qualitativa, dicotômica ou binaria(modelo tradicional)

As preditoras podem ser quantitativas ou categóricas (transformadas em binarias ou Dummy)

Assume que as observações são independentes, que uma não afeta a outra

Por que não uma regressão linear?

Problema da ordenação:

Não podemos utilizar uma regressão linear para prever eventos categóricos.

Ex: Condição médica dos pacientes:
- AVC
- Parada Cardíaca
- Overdose

Podemos simplesmente ordenar e realizar a regressão linear?

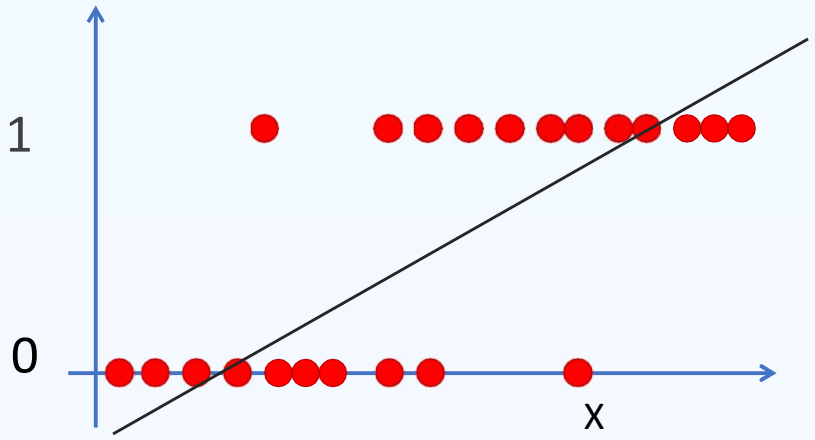
1 = AVC
2 = Parada Cardíaca
3 = Overdose

Devemos utilizar as variáveis dummy e transformar as variáveis.

Regressão Logística - Definição Teórica

Permite **estimar a probabilidade** associada à **ocorrência de determinado evento** em vista de um conjunto de variáveis preditoras.

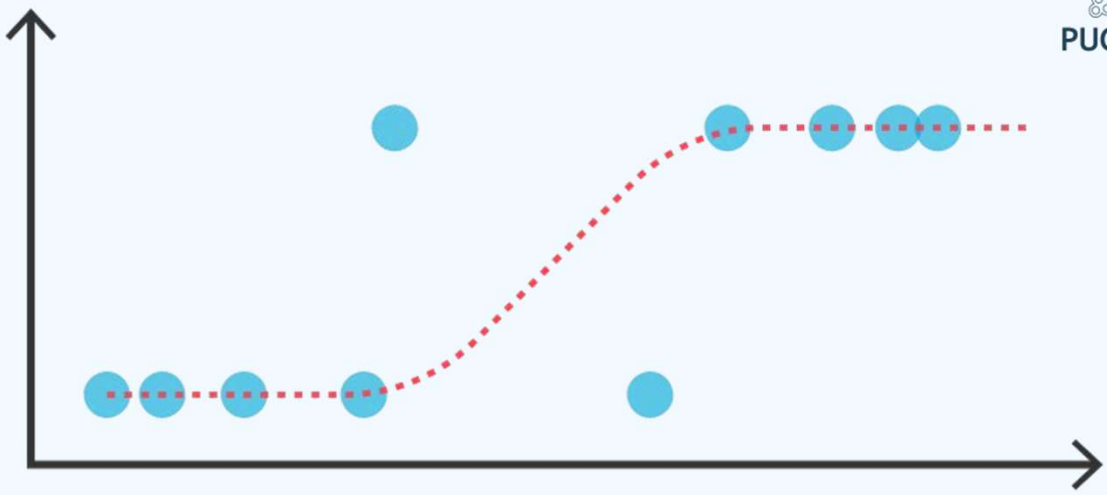
- Probabilidade de sucesso (1)
- Probabilidade de fracasso (0)



Ao interpretar Y como probabilidade, temos que realizar transformações para que a resposta de nossa regressão esteja entre 0 e 1

PUC Minas

Regressão Logística



Função Logística

- Retorna os valores entre 0 e 1
- Tem formato de “S”

Suponha que o modelo tenha a seguinte forma:

$$\mathrm{Y} = X'B + e$$

Em que $X' = [1, X_{i1}, X_{i2}, \dots, X_{in}]$, $\mathrm{B} = [\beta_0, \beta_1, \beta_2, \dots, \beta_n]$, e a variável resposta entre 0 e 1.

Assumimos que a variável reposta é uma variável aleatória de Bernoulli, com função de Probabilidade:

y_i	Probabilidade
1	P(y_i = 1) = p_i
0	P(y_i = 0) = 1 - p_i

Recapitulando - Distribuição de Bernoulli

A variável aleatória Y tem distribuição de Bernoulli se apresenta apenas **dois resultados possíveis**, representados por 0 (fracasso ou negativo) e 1 (sucesso ou positivo). O parâmetro $0 < p < 1$ é a **probabilidade de sucesso**. Dessa forma, a função de probabilidade é

$$p(y) = \begin{cases} 1-p & \text{se "fracasso" ou } y = 0 \\ p & \text{se "sucesso" ou } y = 1, \end{cases}$$
$$= p^y \cdot (1-p)^{1-y}, \quad y \in \{0, 1\}.$$

Denotamos por $Y \sim \text{Ber}(p)$.

Com isso, Y apresenta:

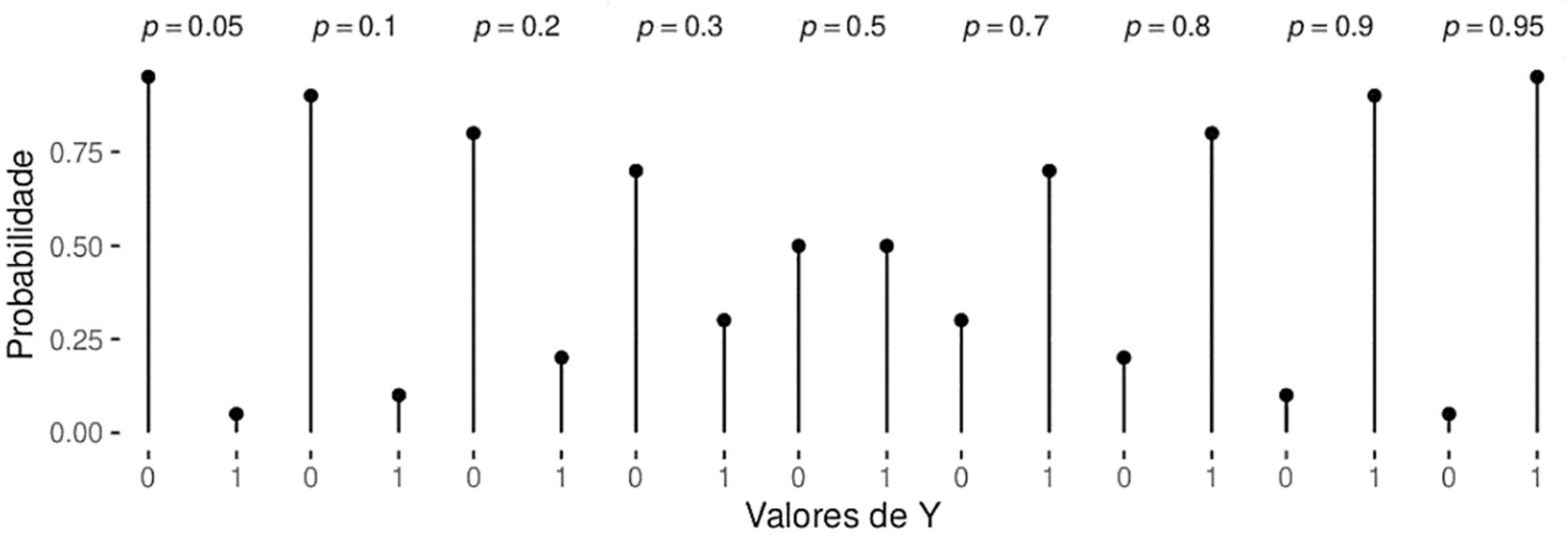
- $\mu = \text{E}(Y) = p$.
- $\sigma^2 = \text{V}(Y) = p \cdot (1-p)$.

Distribuição de Bernoulli

Onde,

P -> Probabilidade de sucesso

1 – p -> Probabilidade de fracasso



PUC Minas

Regressão Logistica

- Uma vez que E(y_i) = 1(p_i) + 0(1- p_i) = p_i

- Temos que:

$$E(y_i) = x'_i\beta = p_i$$

Logo, a **resposta encontrada na regressão logística** sempre será a **probabilidade de sucesso (1)**

PUC Minas

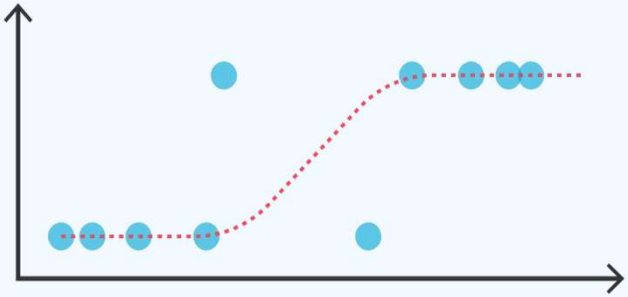
Regressão logística

Resposta (Y) entre 0 e 1

Requer transformação da nossa função

Função Logit

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



Conseguimos transformar essa linha em forma de s, em uma linha reta (linearização), para que possa ser possível o cálculo dos coeficientes?

Regressão Logística

A regressão logística pode ser linearizada:

$$\eta = x'\beta$$

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

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

Essa transformação é frequentemente chamada de **transformação logit** da probabilidade p e a razão $\frac{p}{p-1}$ é chamada de chance odds .

Função ligação que associa os valores esperados da função resposta aos preditores lineares do modelo.

PUC Minas

Regressão Linear – Analisando o erro

- Como temos uma resposta binária (0 e 1), temos que os termos de erro só podem ter dois valores:

$$\varepsilon_i = 1 - x'_i\beta \,, \qquad y_i = 1$$
$$\varepsilon_i = -x'_i\beta \,, \qquad y_i = 0$$

Logo os **erros não podem ser normais**, e a **variância não é constante**.

$$E(\sigma^2_{yi}) = E\{y_i - E(y_i)\} = (1- p_i)^2 p_i + (0- p_i)^2(1- p_i) = p_i(1- p_i)$$
$$\sigma^2_{yi} = E(y_i)[1 - E(y_i)]$$

PUC Minas

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 Bernoulli, então a distribuição de probabilidade é dada por:

$$f_i\ (y_i) = p_i^{y_i}[1\ - E(p_i)]^{1\ - y_i}\ , \qquad \text{i = 1, 2, 3, ..., n}$$

- E cada observação assume o valor de 0 e 1.
- Logo a função de verossimilhança para v.a. independentes pode ser dada por:

$$\text{L}(y_1, y_2, y_3, \dots, y_n, \beta) = \prod_{i=1}^{n} f_i(y_i) = = \prod_{i=1}^{n} p_i^{y_i}[1\ - E(p_i)]^{1- y_i}$$

Estimação de Parâmetros

A forma geral de um modelo de regressão logística é

$$y_i = \mathrm{E}(y_i) + \varepsilon_i$$

Em que as observações são variáveis aleatórias independentes de Bernoulli com valores esperados

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

PUC Minas

Interpretação dos Parâmetros

- Considere o caso em que o preditor linear tem apenas uma variável preditora, de forma que o valor ajustado do preditor linear em um valor particular de x, xi é:

$$\hat{\eta}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 \, x_i$$

O valor ajustado em x_i +1 é:

$$\hat{\eta}(x_i + 1) = \hat{\beta}_0 + \hat{\beta}_1 \, (x_i + 1)$$
$$\hat{\eta}(x_i + 1) = \hat{\beta}_0 + \hat{\beta}_1 \, (x_i) + \hat{\beta}_1$$

- E a diferença dos valores previstos é:

$$\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) = \hat{\beta}_1$$

$$\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) = \ln\left(odds_{x_i+1}\right) - \ln\left(odds_{x_i}\right) = ln\frac{odds_{x_i+1}}{odds_{x_i}} = \hat{\beta}_1$$

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

- Para as variáveis preditora binárias, podemos realizar a seguinte análise:

$$ln\frac{p}{1-p} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Para x_i = 0, temos :

$$ln\frac{p_0}{1-p_0} = \hat{\beta}_0$$

Para x_i = 1, temos :

$$ln\frac{p_1}{1-p_1} = \hat{\beta}_0 + \hat{\beta}_1$$

Logo:

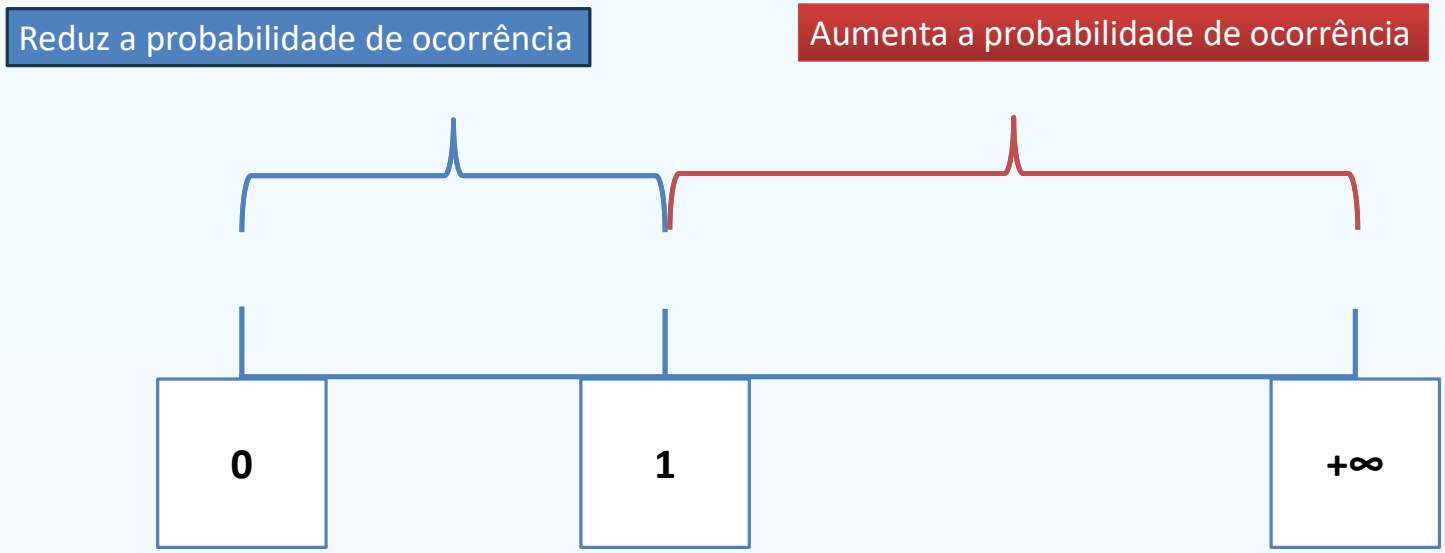
$$ln\frac{p_1}{1-p_1} - ln\frac{p_0}{1-p_0} = \hat{\beta}_0 + \hat{\beta}_1 - \hat{\beta}_0$$

$$ln\frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \hat{\beta}_1 \rightarrow \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = e^{\hat{\beta}_1}$$

PUC Minas

ODDS

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



Exemplo de Interpretação:

Suponha que você esteja estudando a probabilidade de uma pessoa comprar um produto online com base em duas variáveis: idade e gênero. Após ajustar um modelo de regressão logística, você obtém os seguintes resultados:

- Para a variável idade, o odds ratio é 1.05.
- Para a variável gênero (sendo 1 para masculino e 0 para feminino), o odds ratio é 0.8.

Interpretação:

- Para a idade: A cada aumento de uma unidade na idade, as chances de comprar o produto aumentam em 5%.
- Para o gênero: As chances de comprar o produto são 20% menores para homens em comparação com mulheres.

Tipos de resíduos nos GLMs

PUC Minas

Tipos de Resíduos nos GLMs

O **resíduo de Pearson** é definido como (r_i^P) é definido como:

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}},$$

Componente da estatística de Pearson generalizada X_p^2.

Desvantagem: Muito assimétrico para modelos não normais

PUC Minas

Resíduo de Pearson Padronizado

O resíduo de Pearson tem uma versão padronizada, com média 0 e variância aproximadamente 1, definido por:

$$r_i^{P\prime} = \frac{y_i - \widehat{\mu_i}}{\sqrt{\phi V(\widehat{\mu_i})(1 - \widehat{h_{ii}})}},$$

Em que h_{ii} é o i-ésimo elemento da diagonal da matriz H (de pesos e observações)

Os resíduos de Pearson padronizados apresentam propriedade razoáveis de segunda ordem, mas podem ter distribuições muito divergentes da normal

Resíduo Componente da Deviance

São as raízes quadradas dos componentes da *deviance* com sinal igual a $y_i - \widehat{\mu_i}$. Sabe-se que a *deviance* é dada por:

$$D_p = 2\sum_{i=1}^{n}\left[y_i(\widetilde{\theta_i} - \widehat{\theta_i})\right] + b(\widehat{\theta_i}) - b(\widetilde{\theta_i})$$

Então temos:

$$r_i^D = sinal\ (y_i - \widehat{\mu_i}) \cdot \sqrt{2y_i\left(\widetilde{\theta_i} - \widehat{\theta_i}\right) + b\left(\widehat{\theta_i}\right) - b(\widetilde{\theta_i})}$$

O resíduo r_i^D representa a distância da observação y_i ao valor ajustado $\widehat{\mu_i}$, medida na escala da verossimilhança.

PUC Minas

Resíduo Componente da Deviance

Quando r_i^D ⬆ temos que a i-ésima observação está mal ajustada pelo modelo

O resíduo componente da *deviance* possui uma versão padronizada que é dada por:

$$r_i^{D\prime} = \frac{r_i^D}{\sqrt{\phi\,(1-\widehat{h_{ii}})}},$$

Em que h_{ii} é o i-ésimo elemento da diagonal da matriz H(pesos e observações)

Resíduo Quantílico

O **Resíduo Quantílico** aleatorizado pode ser definido por:

$$r_i^q = \phi^{-1}\left[F\left(y_i;\ \mu_i;\ \phi\right)\right]$$

De forma que se o modelo estiver corretamente especificados, estes resíduos seguiram uma distribuição normal padrão.

Se y_i fo uma v.a. discreta, então $F\left(y_i;\ \mu_i;\ \phi\right)$ é uma com fda com “saltos” em cada valor de y_i.

PUC Minas

Análise de Resíduos

Os resíduos de Perason e componente da Deviance geralmente **não possuem boas aproximações com a distribuição normal**, ainda que o **modelo ajustado esteja correto**.

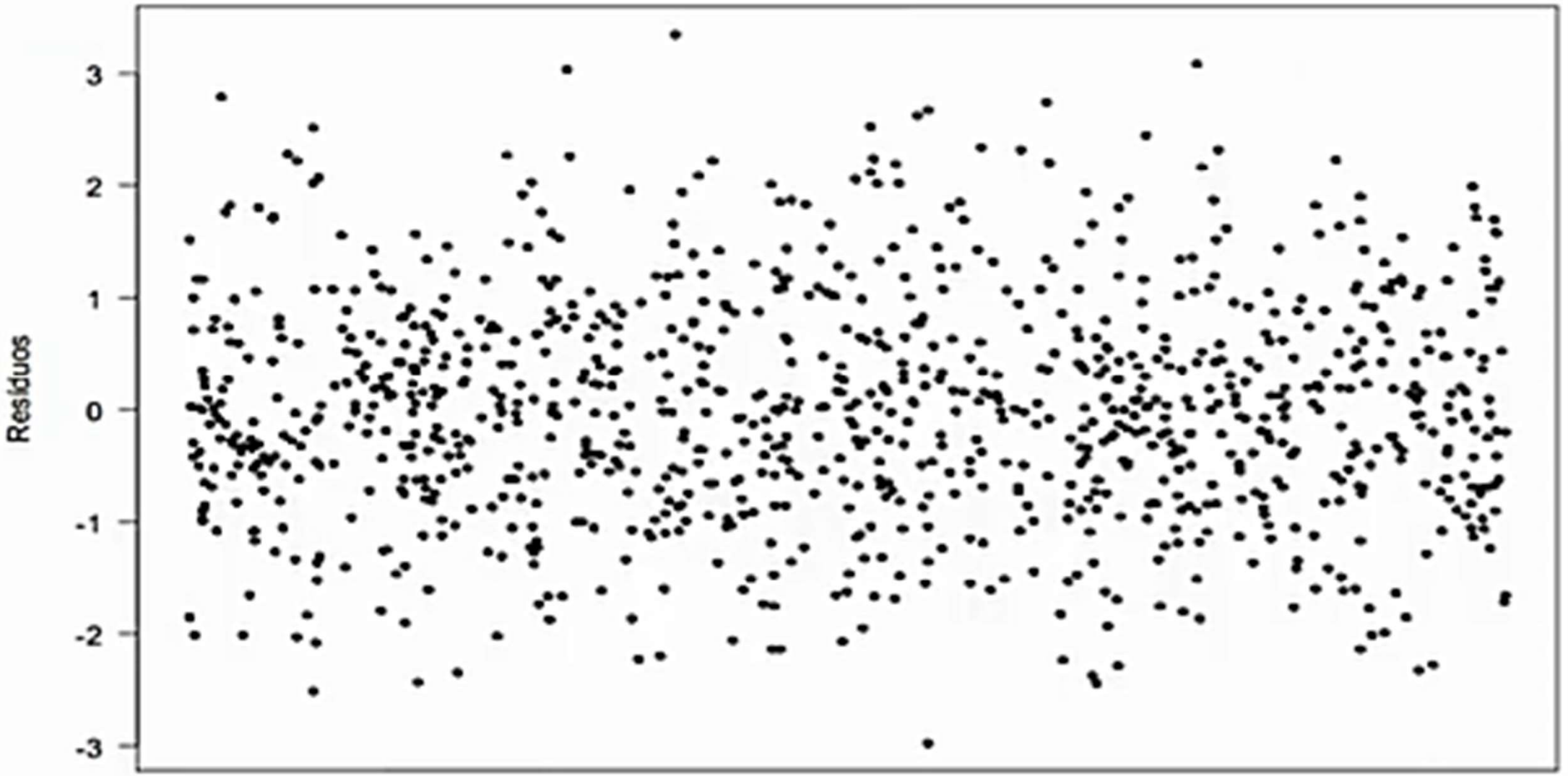
A avaliação da qualidade do ajuste baseado em gráficos de probabilidade (QQ-plot), podem não ser adequados.

Um tipo de resíduo que, por construção, tem **distribuição normal caso o modelo ajustado esteja correto é o resíduo quantílico aleatorizado**

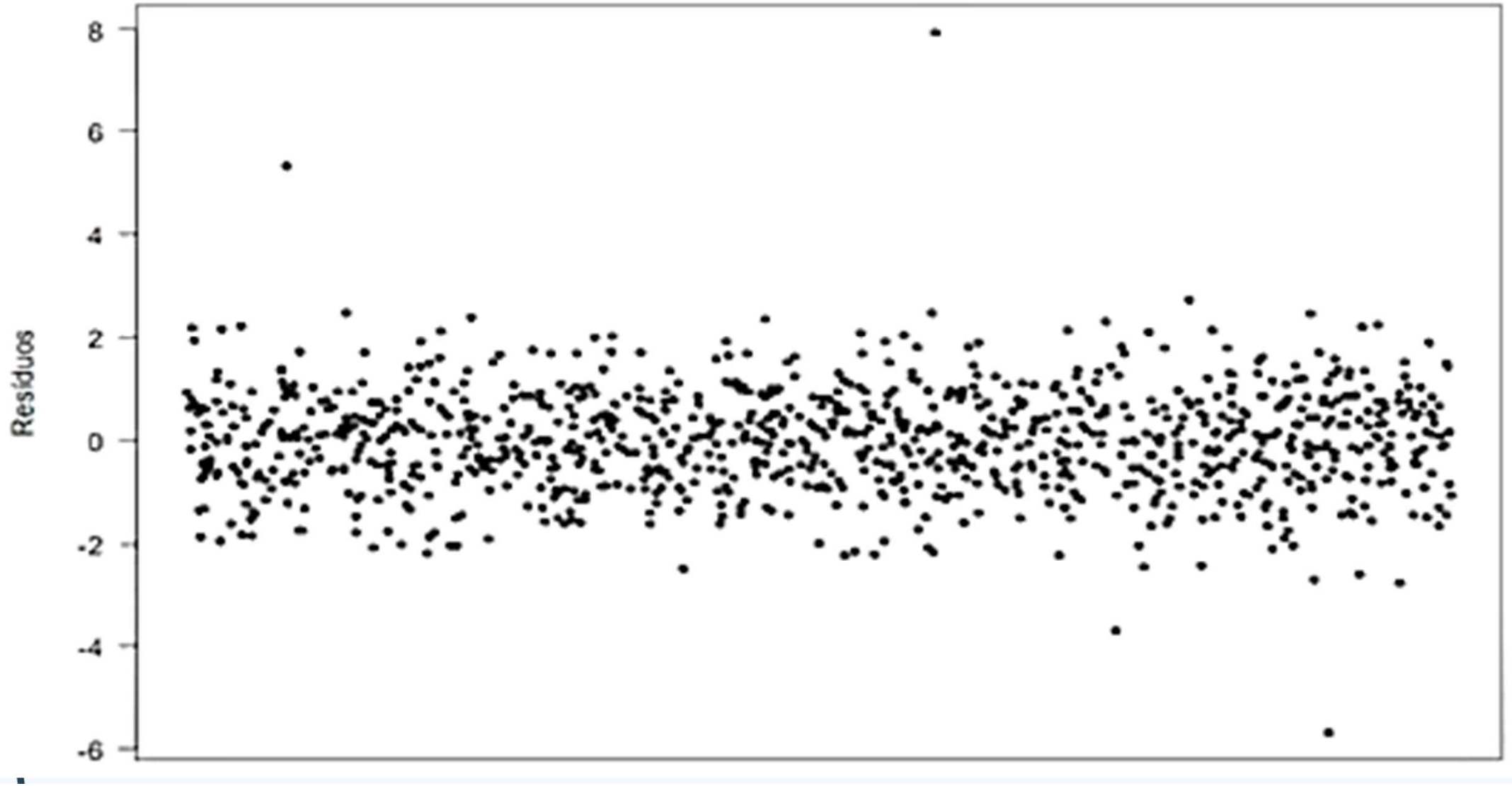
Análise de Resíduos

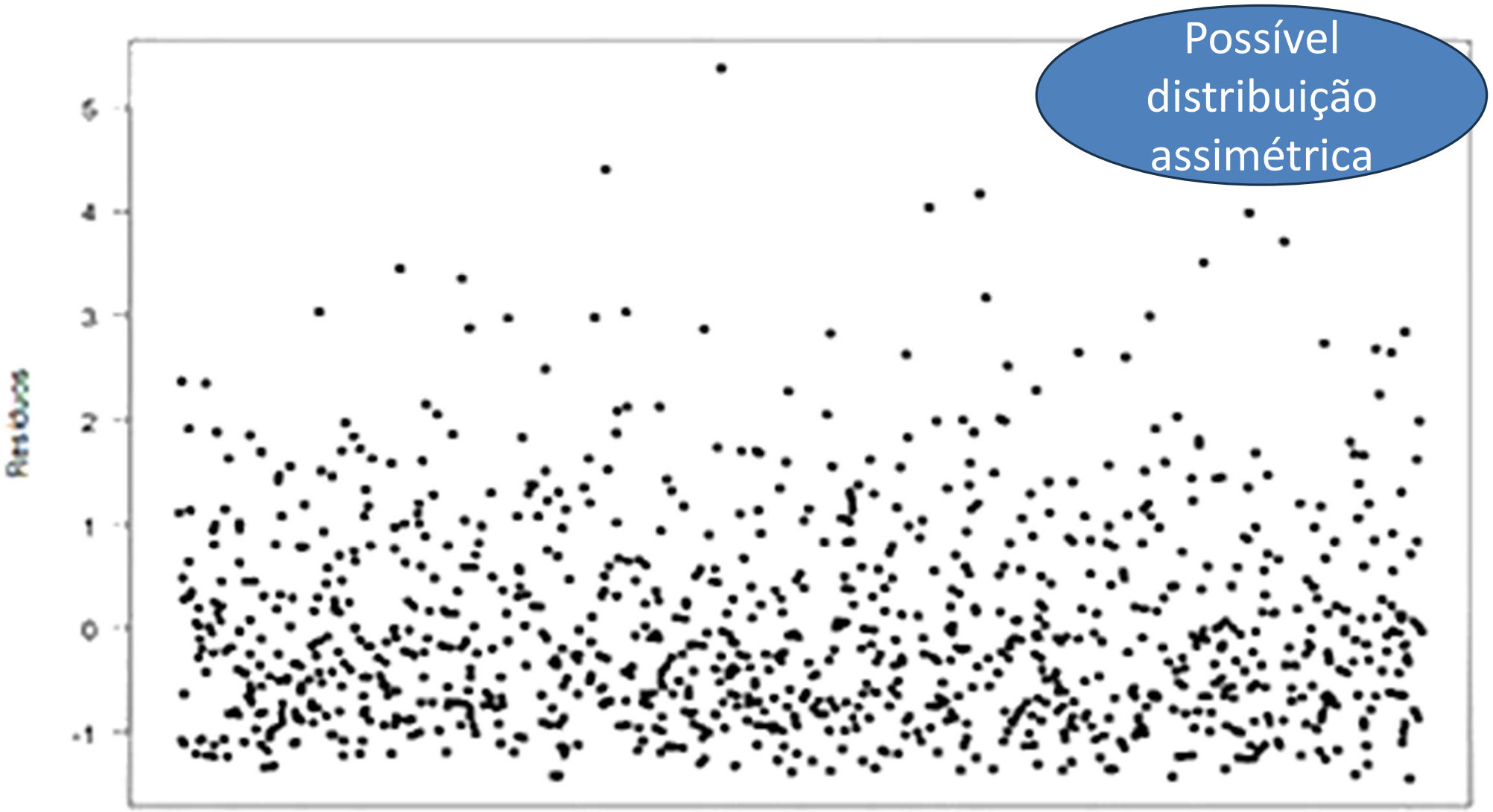
- **Resíduos vs valores ajustados:** Para um modelo bem ajustado, deve-se observar a dispersão aleatória dos pontos, centrada em zero, com média e variâncias constantes e sem valores extremos.
- **Resíduos vs variáveis incluídas no modelo:** padrões não aleatórios indicam que a variável não está inserida corretamente no modelo;
- **Resíduos vs variáveis não incluídas no modelo:** Padrões não aleatórios sinalizam a necessidade (e a forma) de inclusão da variável no modelo;
- **Resíduos vs ordem de coleta de dados:** Padrões não aleatórios indicam dependência das observações gerada pela ordem de coleta (no tempo , no espaço, ...)

Resíduos vs Valores ajustados

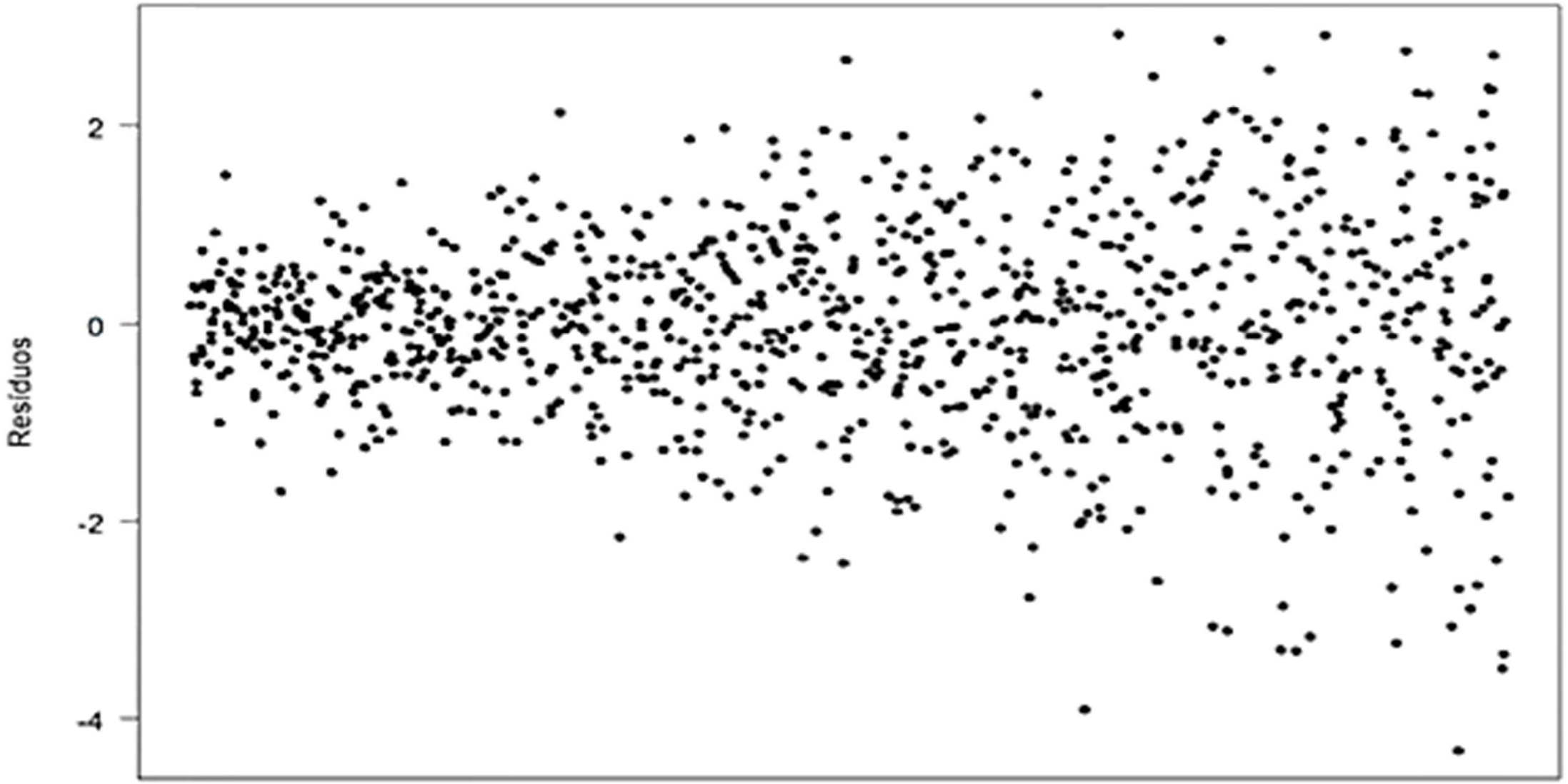


Resíduos vs Valores ajustados

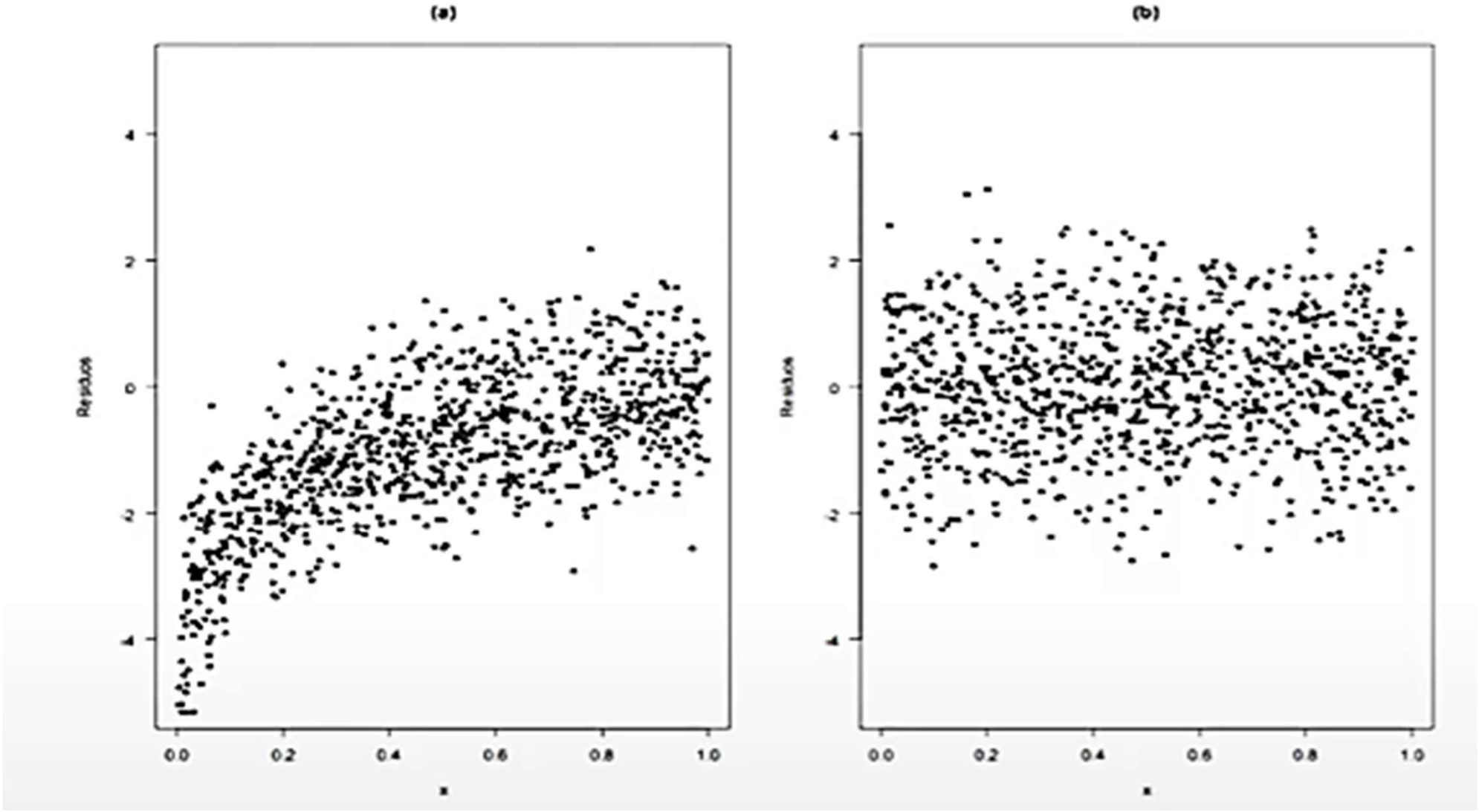



Possível distribuição assimétrica
6
4
3
2
1
0
-1

Resíduos vs Valores ajustados



Resíduos vs Variáveis incluídas no modelo



PUC Minas

Multicolinearidade

- **Preditoras correlacionadas com outras preditoras**, resulta quando você tem fatores que são, de certa forma, um pouco **redundantes**.
- Ou seja, quando **duas ou mais variáveis independentes em um modelo de regressão encontram-se altamente correlacionadas**
- Examinar a matriz de correlação das variáveis independentes.
  - 0,70 Altamente correlacionadas
  - 0,80 Alerta

- O valor do fator de inflação da variância (VIF), **que mede quanto a variância do coeficiente estimado para uma variável é inflada devido à multicolinearidade** com as outras variáveis independentes.
- VIFs maiores que 10 indicam alta multicolinearidade, enquanto valores entre 5 e 10 podem ser preocupantes.
- A maneira mais simples de lidar com a multicolinearidade é **excluir a variável multicolinear**

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

PUC Minas

Teste de Razão de Verossimilhança

- Compara valores observados x preditos , com e sem determinadas variáveis.
- A comparação dos valores observados x valores preditos é baseada na log verossimilhança
- Pode-se pensar que o valor observado de Y, é um valor predito resultante de um modelo saturado.

$$\mathbf{D = -\ 2\ ln}\left(\frac{\boldsymbol{Verossimilhança\ modelo\ ajustado}}{\boldsymbol{Verossimilhança\ modelo\ saturado}}\right)$$

Devaince: para a reg. Logística irá desempenhar o mesmo papel que o SQR na reg. Linear.

Modelo Saturado -> Modelo que se ajusta perfeitamente os dados

PUC Minas

Teste de Razão de Verossimilhança

- Utilizando a *deviance* para comparação de modelos que não sejam saturados.

G = D(modelo sem a variável) – D(modelo com a variável)

Modelo saturado é o mesmo para os 2.

$$G = -2 \ln \left(\frac{Verossimilhança\ sem\ a\ variável}{Verossimilhança\ com\ a\ variável} \right)$$

PUC Minas

Teste de Razão de Verossimilhança

H_0 : *O modelo nulo (mais simples) é verdadeiro, ou seja, a inclusão das variáveis adicionais no modelo completo não melhora significativamente o ajuste do modelo.*

H_1 *: O modelo completo é significativamente melhor que o modelo nulo.*

$\boldsymbol{H_0}$: ***Verossimilhança do modelo Nulo = Verossimilhança do Modelo Completo***

$\boldsymbol{H_1}$: ***Verossimilhança do modelo Completo > Verossimilhança do Modelo Nulo***

PUC Minas

Teste Wald

Obtido por comparação entre a estimativa de máxima verossimilhança do parâmetro $\hat{\beta}_j$ *e a estimativa de seu erro padrão.*

$$H_0 : \hat{\beta}_1 = 0$$
$$H_1 : \hat{\beta}_1 \neq 0$$

$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_i$$

A estatística do Teste Wald para a regressão logística é dada por:

$$W_j = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

Se não rejeitarmos H_0 , temos que a variável X não explica a variável resposta.

Medidas de qualidade do ajuste do modelo

- Para analisar o desempenho geral do modelo ajustado podemos utilizar vários tipos de Testes de qualidade de ajuste.

- Teste que necessitam dados replicados (múltiplas observações com os mesmos valores para todos os preditores):

  - χ^2 de Pearson
  - Deviance

- O teste hosmer-lemeshow é útil para conjuntos de dados não replicados ou que contém apenas algumas observações replicadas.
  - As observações são agrupadas com base em suas probabilidades estimadas.

PUC Minas

Deviance

- Pequenos valores de *Deviance* (ou elevado valor p) implicam que o modelo fornece um ajuste satisfatório aos dados , enquanto grandes valores de *deviance* implicam que o modelo atual não é adequado.

- Podemos dividir o *deviance* pelo graus de liberdade.
  - Se $\frac{D}{n-k} \gg 1 \rightarrow O\ Modelo\ não\ é\ adequado\ aos\ dados$
  - Onde $n - k$ é o gruas de liberdade. k é o número de parâmetro do modelo

  - D é dado por:

$$D = 2 \sum_{i=1}^{n} \left[y_i \ln \left(\frac{y_i}{n_i \hat{p}_i} \right) + (n_i - y_i) \ln \left(\frac{n_i - y_i}{n_i (1 - \hat{p}_i)} \right) \right]$$

PUC Minas

χ^2 de Pearson

- Compara as probabilidade de sucesso e fracasso observadas e esperadas em cada grupo de observações
  - Nº esperado de sucesso : $n_i\hat{p}_i$
  - Nº esperado de fracasso: $n_i(1-\hat{p}_i)$

- A estatística de de Pearson é dada por:

$$\chi^2_{n-k} = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}$$

- **Valores pequenos da estatística de teste ou um grande valor de p , implica que o modelo fornece um ajuste satisfatório aos dados**

Exemplos