

Disciplina:

Técnicas de Amostragem e Modelos de Regressão

Professora: Anaíle Mendes Rabelo

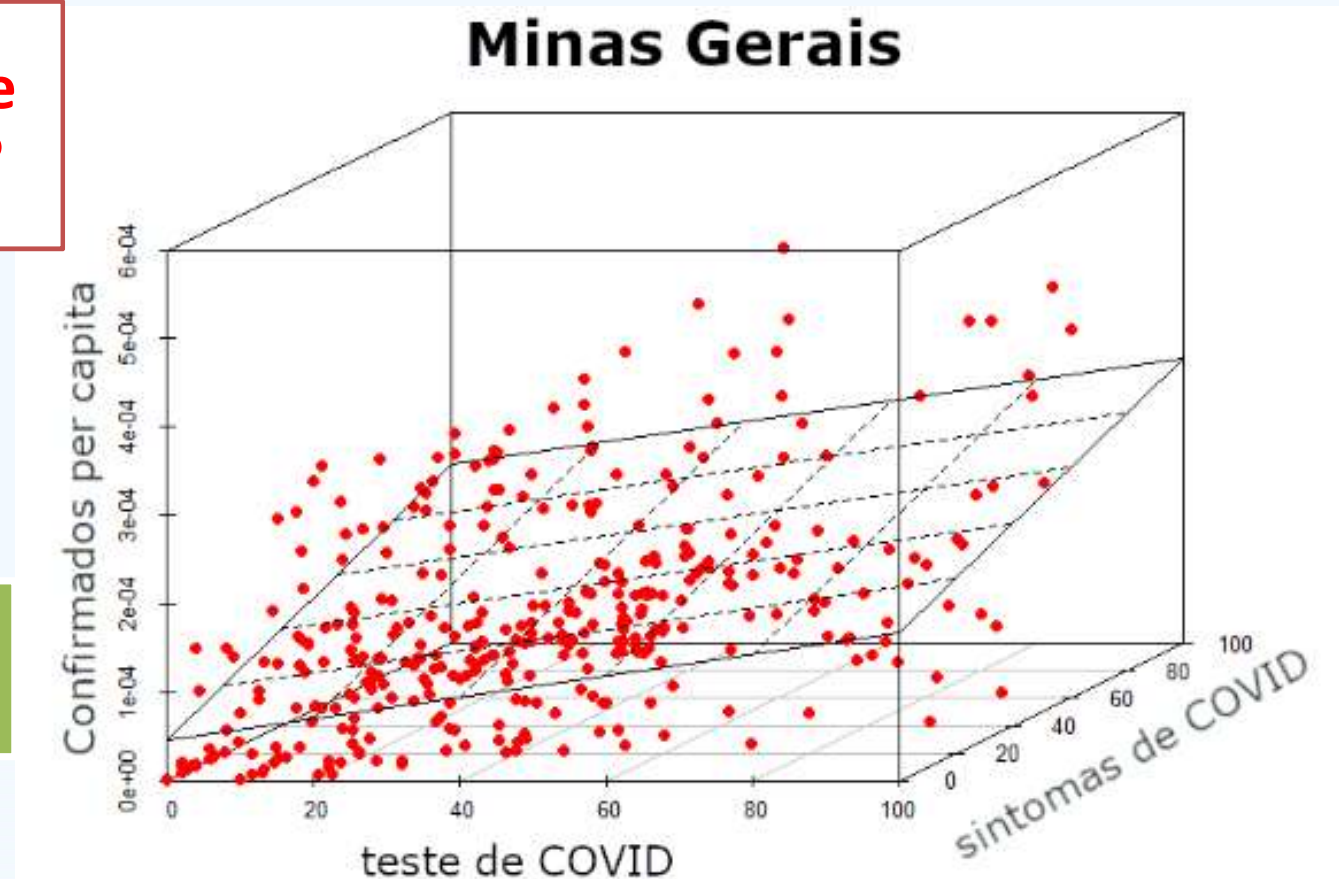
REGRESSÃO LINEAR MÚLTIPLA

REGRESSÃO LINEAR MÚLTIPLA

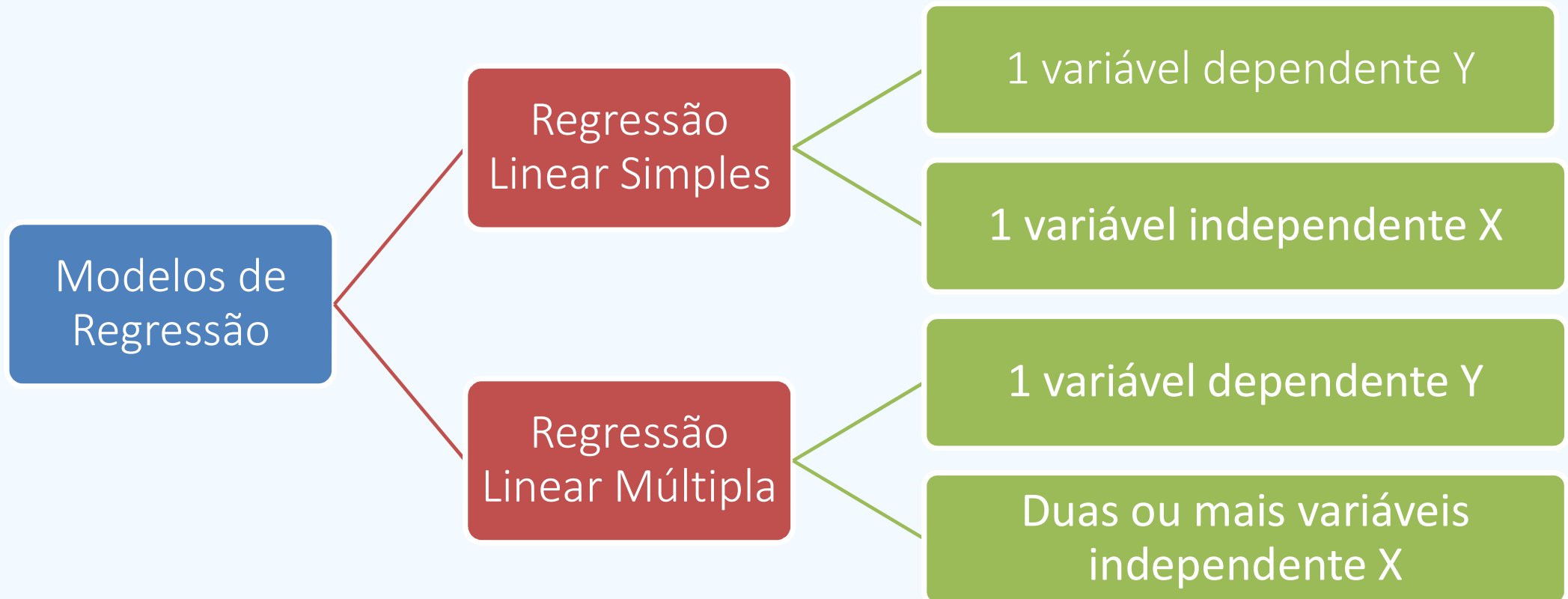
E quando possuímos mais de uma variável independente?



A solução está na :
Regressão Linear Múltipla



REGRESSÃO LINEAR MÚLTIPLA



REGRESSÃO LINEAR MÚLTIPLA

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \dots + \hat{\beta}_n K_i + e_i$$

Diagram illustrating the components of the Multiple Linear Regression equation:

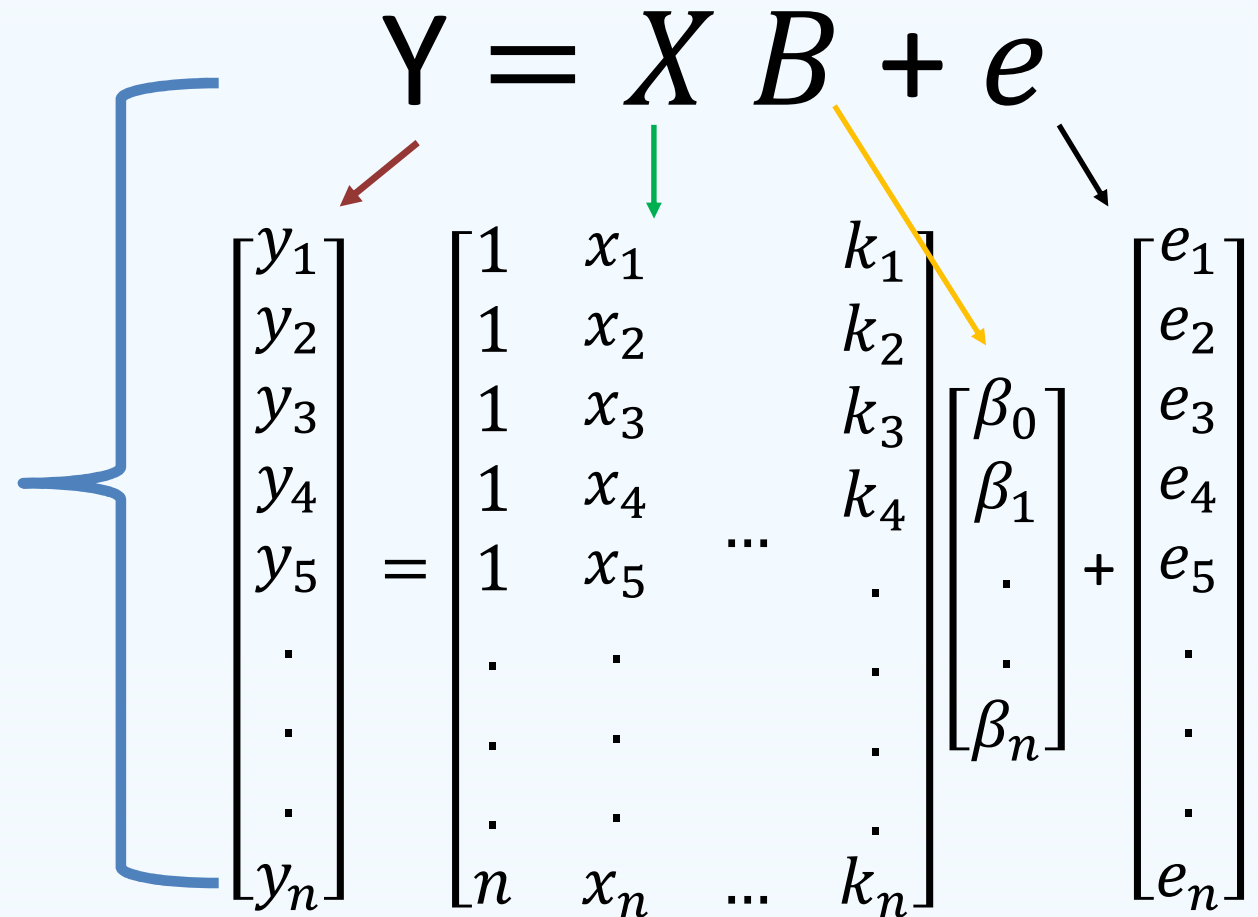
- \hat{Y}_i : Variável Dependente (Dependent Variable)
- $\hat{\beta}_0$: Intercepto Y (Y Intercept)
- $\hat{\beta}_1$: Coeficiente de X (Coefficient of X)
- X_i : Variável independente X (Independent Variable X)
- $\hat{\beta}_2$: Coeficiente de Z (Coefficient of Z)
- Z_i : Variável independente Z (Independent Variable Z)
- $\hat{\beta}_n$: Coeficiente de K (Coefficient of K)
- K_i : Variável independente K (Independent Variable K)
- e_i : Erro Aleatório (Random Error)

The equation is structured into two main components:

- Componente Linear**: The sum of the intercept and the products of coefficients and independent variables ($\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \dots + \hat{\beta}_n K_i$).
- Componente do Erro Aleatório**: The random error term (e_i).

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \dots + \hat{\beta}_n K_i + e_i$$

$$Y = X B + e$$


$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \vdots \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & & k_1 \\ 1 & x_2 & & k_2 \\ 1 & x_3 & & k_3 \\ 1 & x_4 & & k_4 \\ 1 & x_5 & \dots & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ n & x_n & \dots & k_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

$$Y = XB + e$$

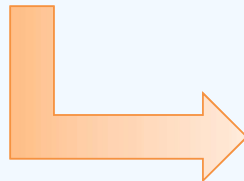
$$e = Y - XB$$

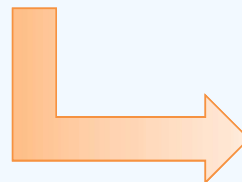
$$SQR = e'e = (Y - XB)'(Y - XB)$$

$$e'e = \begin{bmatrix} e_1 & e_2 & e_3 & e_4 & e_5 & \cdot & \cdot & \cdot & e_n \end{bmatrix} \mathbf{X} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix} = e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2 + \dots + e_n^2$$

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

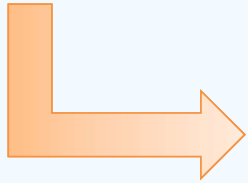
$$SQR = e'e = (Y - XB)'(Y - XB)$$


$$= Y'Y - 2BX'Y + B'X'XB$$

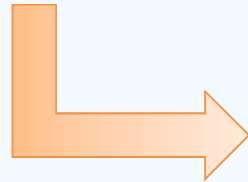

$$\frac{\partial SQR}{\partial B} = -2X'Y + 2B'X'X \equiv 0$$

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

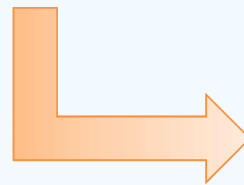
$$-2X'Y + 2B'X'X = 0$$



$$2B'X'X = 2X'Y$$



$$B'X'X = X'Y$$



$$\hat{B} = (X'X)^{-1}X'Y$$

PREMISSAS DA REGRESSÃO LINEAR MÚLTIPLA

☐ Análise de Outliers de resíduos

☐ Homocedasticidade

☐ Normalmente distribuído

☐ Ausência de multicolinearidade e autocorrelação

$$e_i = y_i - \hat{y}_i$$

Média = 0

Variância constante

Covariância = 0

ANÁLISE DE OUTLIERS

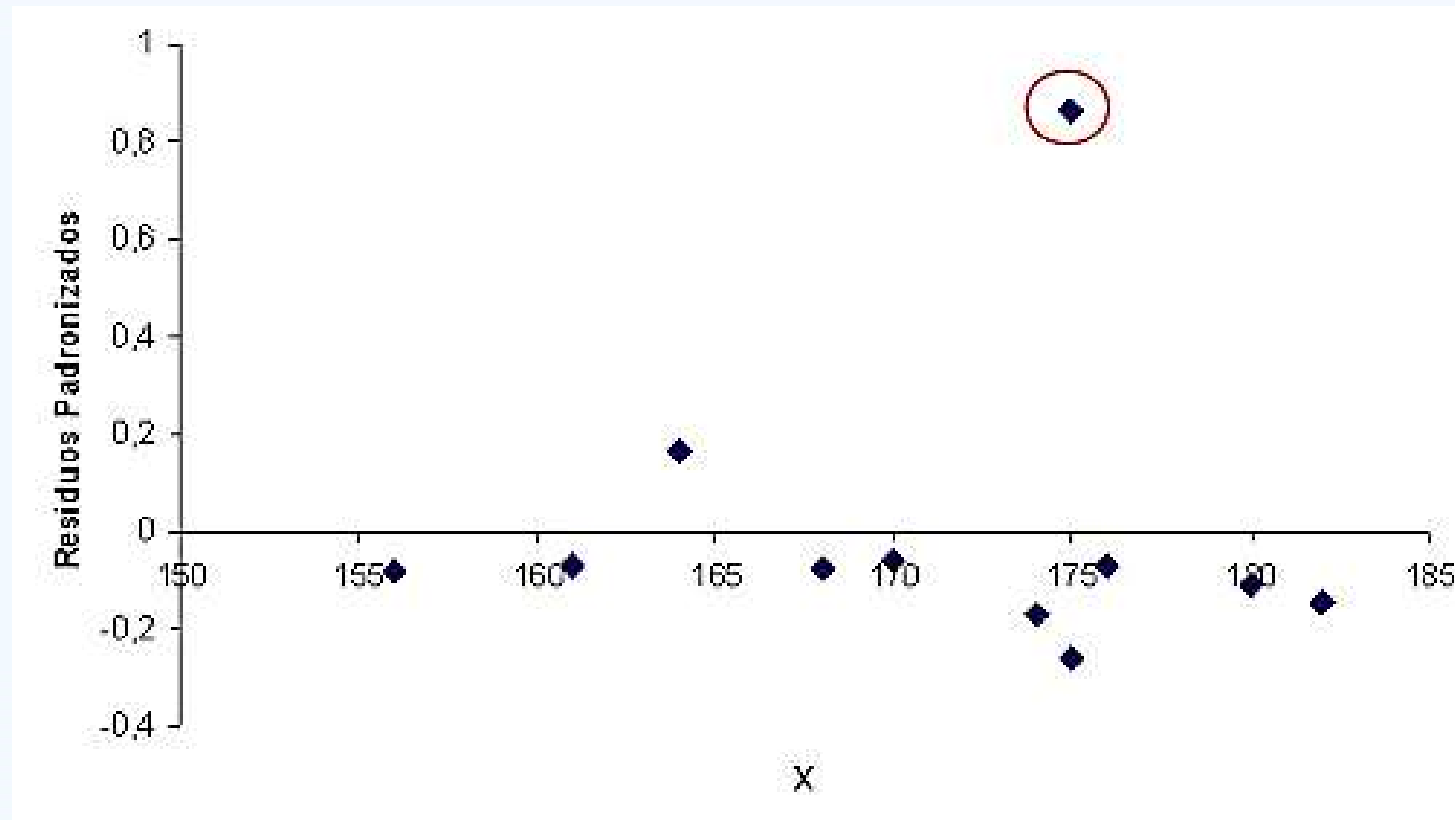


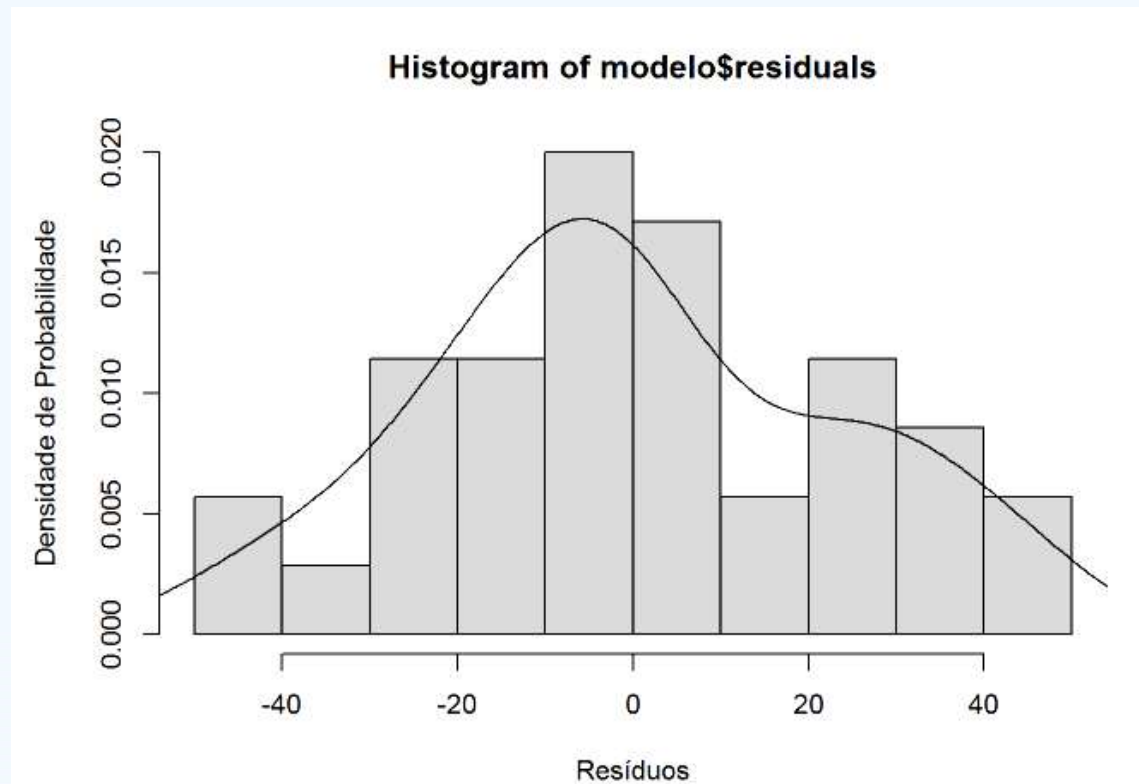
Gráfico de Resíduos padronizados vs Valores ajustados

NORMALIDADE DOS RESÍDUOS

Teste de Shapiro Wilk

H_0 = distribuição normal : $p > 0.05$

H_1 = distribuição não normal : $p \leq 0.05$



ANÁLISE DA HOMOCEDASTICIDADE DOS RESÍDUOS

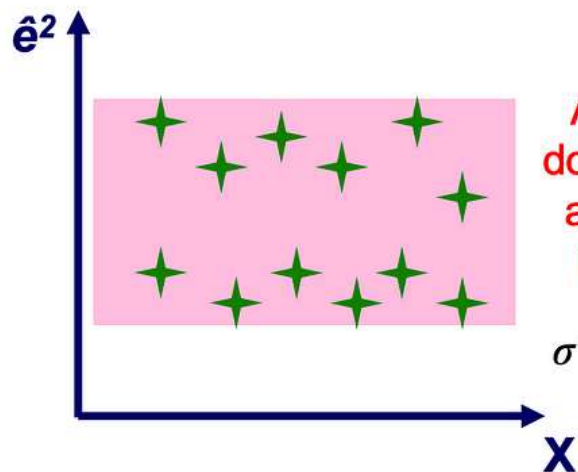
Homocedasticidade: A variância dos erros e , condicionada aos valores das variáveis explanatórias, será constante.

Teste Breusch-Pagan (Homocedasticidade)

H_0 = existe homocedasticidade : $p > 0.05$

H_a = não existe homocedasticidade : $p \leq 0.05$

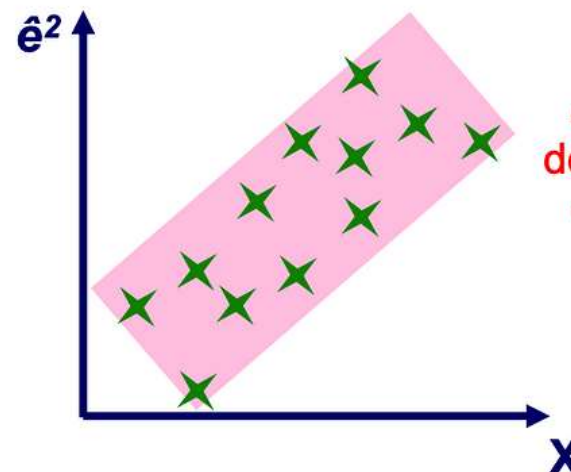
Homocedasticidade



A dispersão dos resíduos é a mesma ao longo de X

$$\sigma^2 = \text{constante}$$

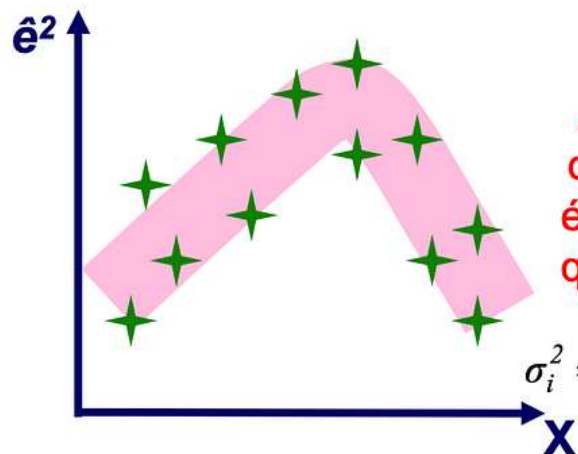
Heterocedasticidade



A dispersão dos resíduos é uma função linear de X

$$\sigma_i^2 = \sigma^2 X_i$$

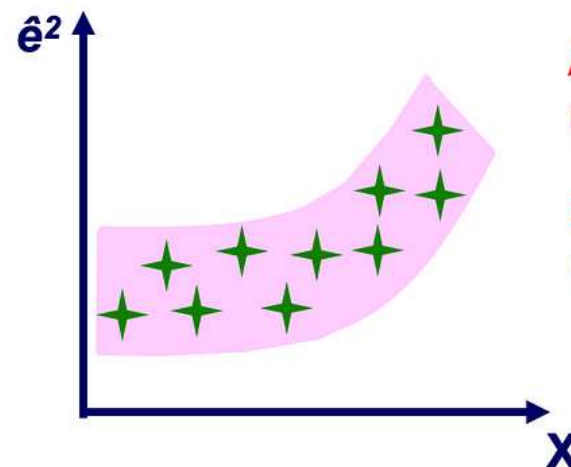
Heterocedasticidade



A dispersão dos resíduos é uma função quadrática de X

$$\sigma_i^2 = \alpha_1 X_i + \alpha_2 X_i^2$$

Heterocedasticidade



A dispersão dos resíduos cresce de maneira quadrática com os valores de X

$$\sigma_i^2 = \sigma^2 X_i^2$$

TESTE - T

Avaliando a significância de cada parâmetro β do modelo

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

$$H_0: p - \text{valor} \geq 0,05$$

$$H_1: p - \text{valor} < 0,05$$

TESTE - F

Avalia a significância global de um modelo de regressão linear, ou seja, para testar se pelo menos uma das variáveis independentes tem um efeito significativo sobre a variável dependente.

O teste F é comumente usado em modelos de regressão múltipla.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

H_1 : Pelo menos um β_j é diferente de zero

$$H_0 = F_{Calc} \leq F_{Crítico} \text{ ou } p - \text{valor} \geq 0,05$$

$$H_1 = F_{Calc} > F_{Crítico} \text{ ou } p - \text{valor} < 0,05$$

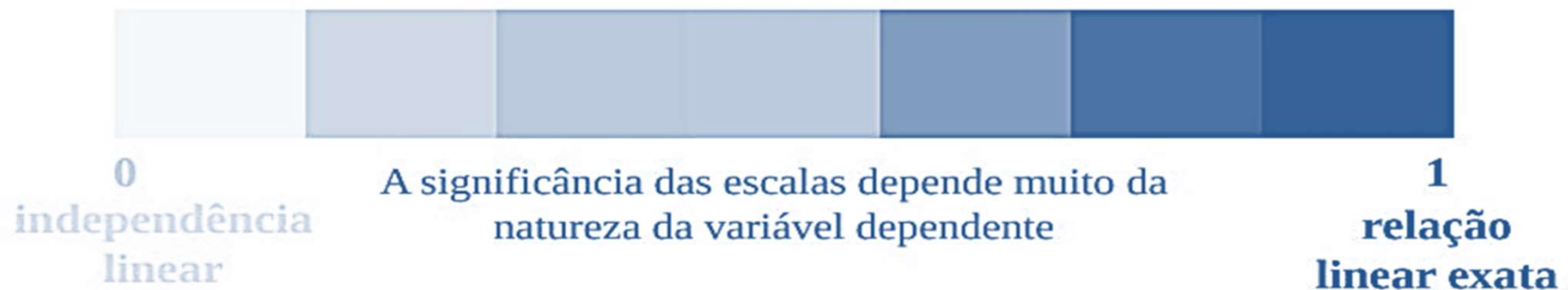
COEFICIENTE DE DETERMINAÇÃO (R^2)

Como avaliar o modelo?

O **coeficiente de determinação (R^2)** estima a proporção da variabilidade da variável dependente (Y) que é explicada pelas(s) variáveis independentes do modelo de regressão.

$$R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n y_i^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n y_i^2}$$

Escala de R^2 :



COEFICIENTE DE DETERMINAÇÃO (R^2)

R^2 x R^2 ajustado

R^2 é a proporção da variabilidade total da variável dependente explicada pela regressão

R^2 ajustado leva em conta o número de variáveis independentes no modelo e penaliza o modelo por incluir variáveis irrelevantes.

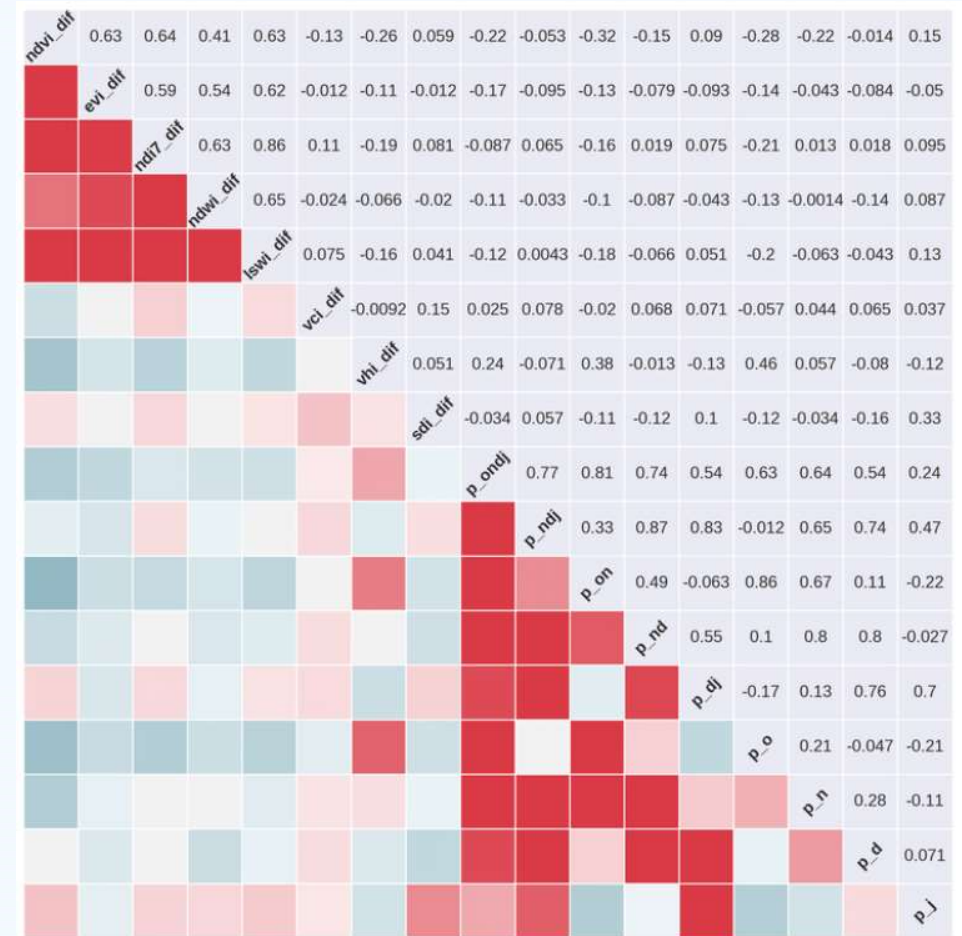
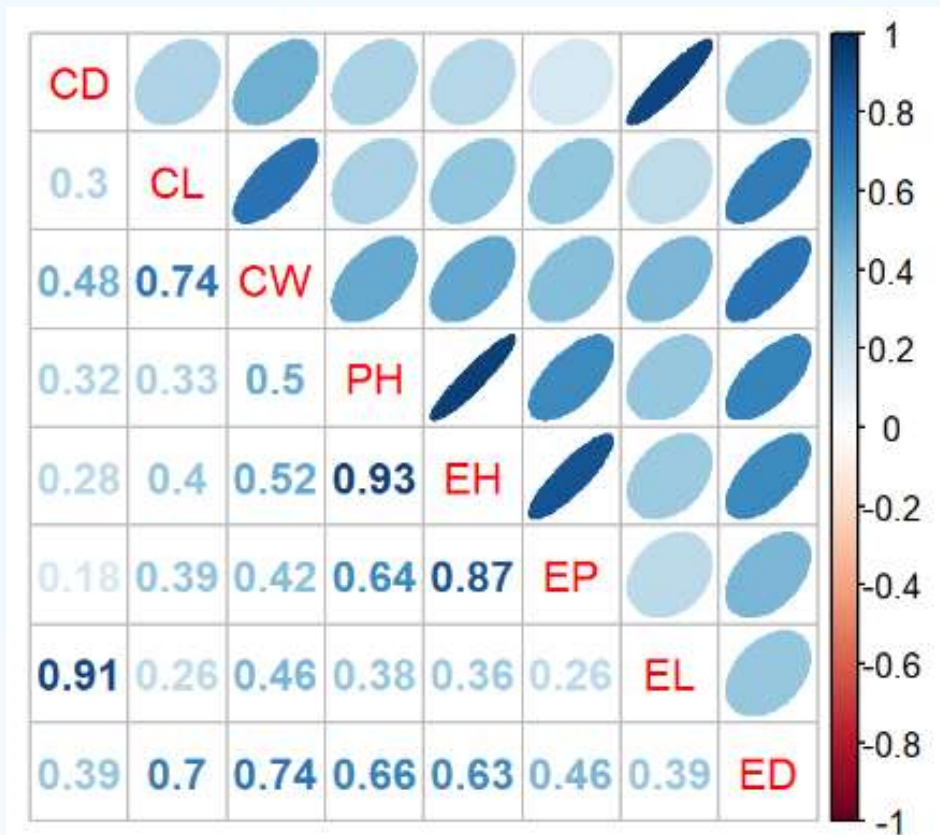
MULTICOLINEARIDADE

- **Preditores correlacionados com outros preditores**, 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

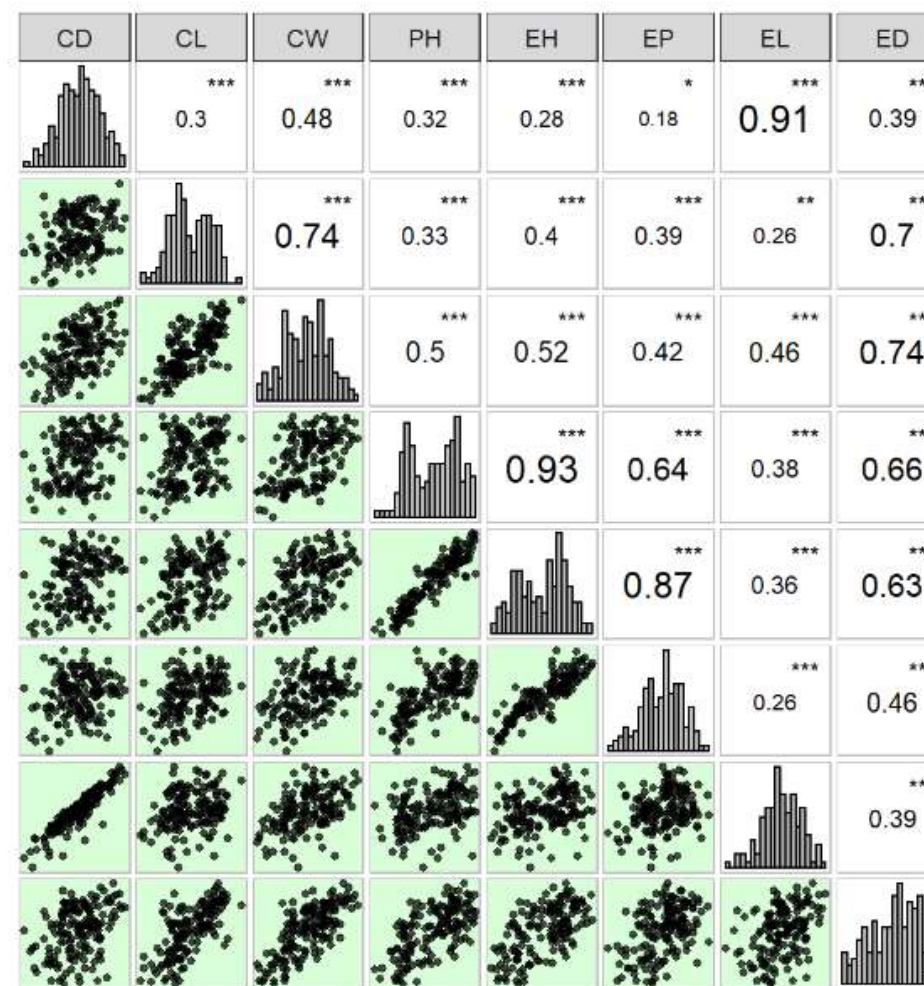
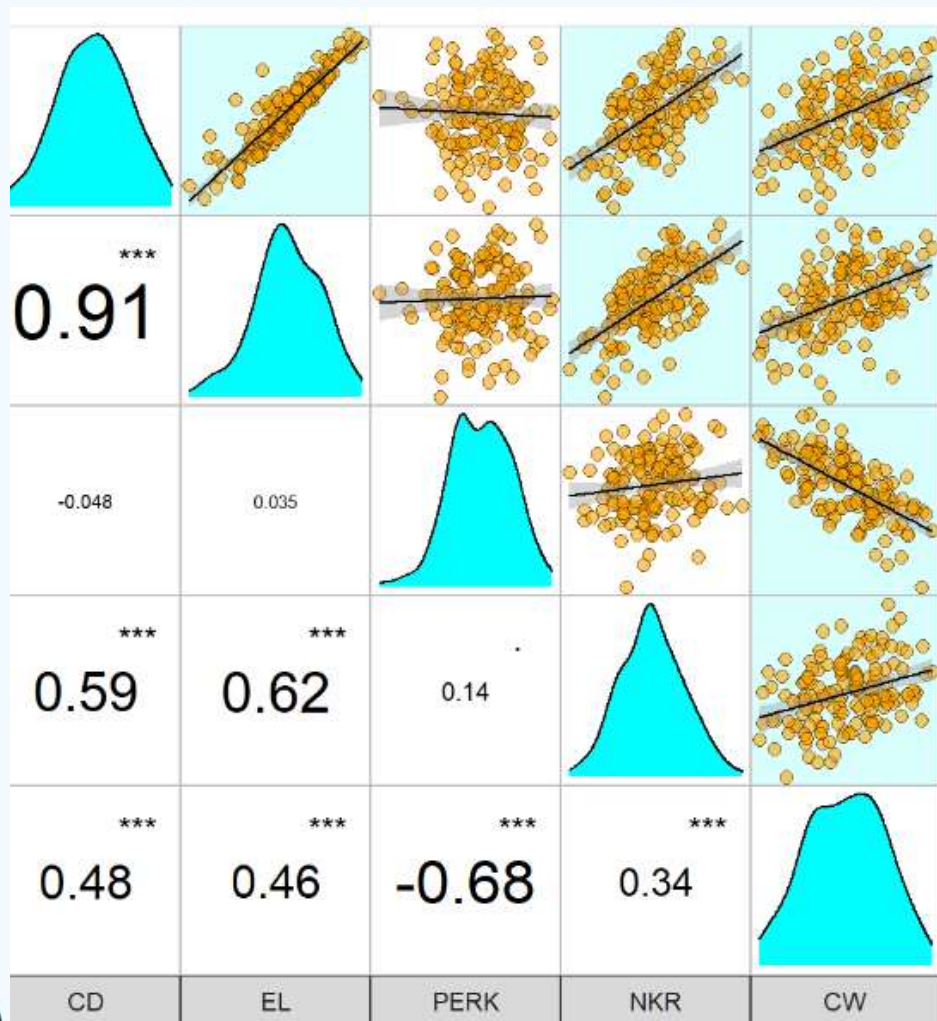
MULTICOLINEARIDADE

- 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

Scatter plot de uma matriz de correlação



Scatter plot de uma matriz de correlação



Detalhamento da saída do R

```
Call:
lm(formula = fat ~ . - id_pizza, data = dados)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-11.2688  -2.2798   0.0192   2.1821   6.7930
```

P - valor do teste t

Estimador dos coeficientes

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.22820    0.67148   10.77  < 2e-16 ***
sodium        21.11931    0.62813   33.62  < 2e-16 ***
carb         -0.04968    0.01290   -3.85 0.000144 ***
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.156 on 297 degrees of freedom
```

```
Multiple R-squared:  0.8772,    Adjusted R-squared:  0.8764
```

```
F-statistic: 1061 on 2 and 297 DF,  p-value: < 2.2e-16
```

R² Ajustado

Detalhamento da saída do PYTHON

Dep. Variable:	Consumo de cerveja (litros)	R-squared:	0.723	R^2
Model:	OLS	Adj. R-squared:	0.719	R^2 - ajustado
Method:	Least Squares	F-statistic:	187.1	
Date:	Mon, 28 Jun 2021	Prob (F-statistic):	1.19e-97	Teste F

	coef	std err	t	P> t	[0.025	0.975]
const	6.4447	0.845	7.627	0.000	4.783	8.107
Temperatura Media (C)	0.0308	0.188	0.164	0.870	-0.339	0.401
Temperatura Minima (C)	-0.0190	0.110	-0.172	0.883	-0.236	0.198
Temperatura Maxima (C)	0.6560	0.095	6.895	0.000	0.469	0.843
Precipitacao (mm)	-0.0575	0.010	-5.726	0.000	-0.077	-0.038
Final de Semana	5.1832	0.271	19.126	0.000	4.650	5.716

Estimadores dos coeficientes

Teste T

COMPARAÇÃO DE MODELOS

COMPARAÇÃO DE MODELOS

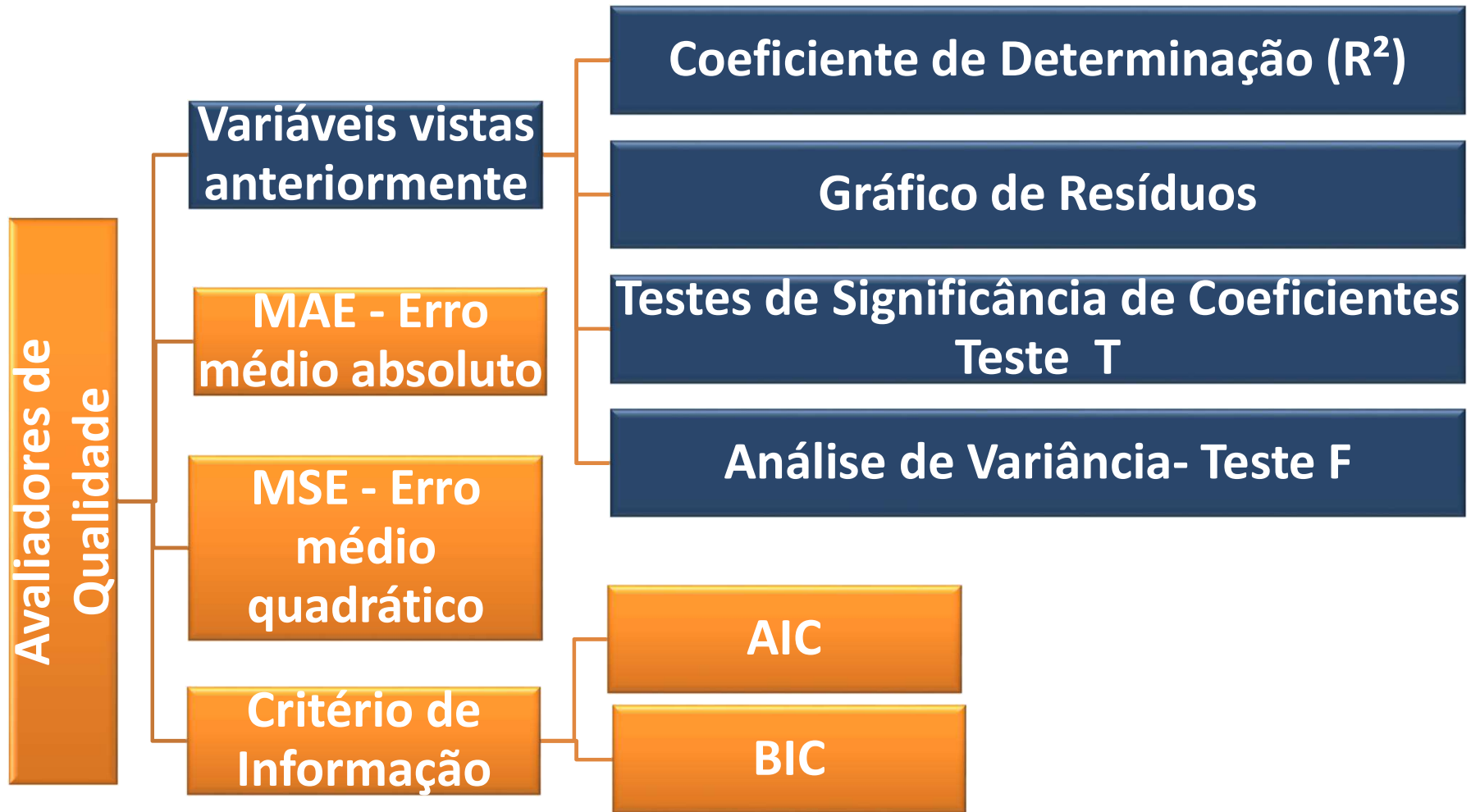
Nº grande
de
possíveis
modelos

Como
selecionar
o melhor
modelo

Utilizar
avaliadores
de
qualidade

Selecionar
o Modelo
mais
PARCIMO-
NIO SO

Modelo Parcimonioso é o modelo que **melhor explica o fenómeno que estamos estudando da forma mais simples possível**, menor número de parâmetros.

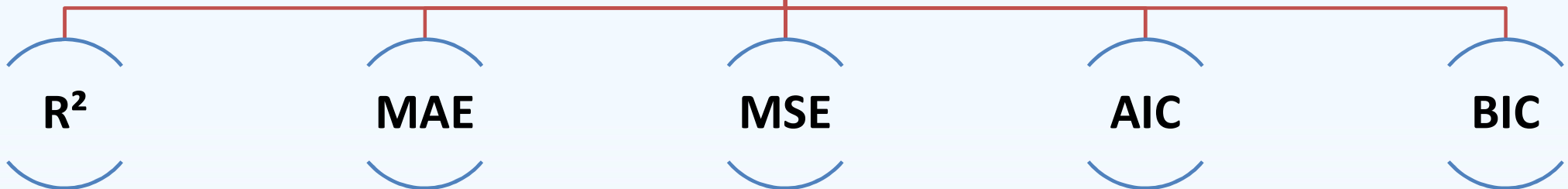


Usados para **medir o quão bem um modelo se ajusta aos dados observados**.
Ajudam a determinar a **qualidade** geral do ajuste do modelo, sua **capacidade de generalização** e o equilíbrio entre ajuste e complexidade.

COMPARAÇÃO DE MODELOS

Utilizamos os avaliadores de qualidade de modelos para comparar os modelos e entender qual modelo **explica** melhor o fenômeno de maneira mais **parcimoniosa**

Principais métricas para comparação de Modelos




CRITÉRIO DE INFORMAÇÃO DE AKAIKE - AIC

Métrica usada para **comparar diferentes modelos estatísticos, especialmente em contextos de seleção de modelos**. Foi proposto por Akaike em 1973 e é uma **ferramenta valiosa para equilibrar a qualidade do ajuste do modelo e sua complexidade**.

Logaritmo natural da Soma dos quadrados dos resíduos

$$AIC = -n \cdot \ln \left(\overbrace{SQR/n} \right) + \underbrace{2p}_{\text{Nº de parâmetros}}$$

INTERPRETAÇÃO DO AIC

-  AIC, melhor o ajuste do modelo aos dados.
- O termo $-2 \cdot \ln(SQR)$ penaliza a qualidade do ajuste, **favorecendo modelos que explicam melhor a variabilidade dos dados.**
- O termo $2 \cdot p$ **penaliza a complexidade do modelo**, favorecendo modelos mais simples com menos parâmetros.

O objetivo é selecionar o modelo com o menor AIC, pois ele equilibra eficazmente a precisão do ajuste e a parcimônia do modelo. No entanto, o AIC não fornece uma medida absoluta de ajuste, apenas comparações relativas entre modelos.

CRITÉRIO DE INFORMAÇÃO BAYESIANO - BIC

Também conhecido como Critério de Schwarz, é uma métrica semelhante ao AIC, mas incorpora uma **penalização mais forte para modelos mais complexos**. Ele foi proposto por Gideon E. Schwarz em 1978 e é particularmente útil quando se deseja **evitar o sobreajuste**.

Logaritmo natural da Soma dos quadrados dos resíduos


$$BIC = -n \cdot \ln(SQR/n) + \ln(n)p$$

AIC
BIC

P -> Nº de parâmetros

N -> tamanho da amostra

INTERPRETAÇÃO DO BIC

-  BIC, melhor o ajuste do modelo aos dados.
- O termo $-2 * \ln(\text{SQR})$ penaliza a qualidade do ajuste.
- O termo $p * \ln(n)$ penaliza a complexidade do modelo, sendo a penalização mais forte do que no AIC.

O BIC tende a favorecer modelos mais simples do que o AIC, especialmente quando o tamanho da amostra é pequeno. Portanto, o BIC é uma escolha apropriada quando se deseja evitar a inclusão de variáveis desnecessárias e manter um modelo mais parcimonioso.

COMPARANDO O AIC E O BIC

- ❑ Ambos, o AIC e o BIC, são ferramentas valiosas para seleção de modelos e ajudam a encontrar um equilíbrio entre ajuste e complexidade.
- ❑ A escolha entre eles dependerá do objetivo específico da análise e das preferências do pesquisador:
 - O AIC pode ser preferível quando se busca um ajuste mais preciso.
 - O BIC é mais útil quando a simplicidade do modelo é priorizada para evitar o sobreajuste.
 - Ambos os critérios proporcionam insights importantes na tomada de decisões sobre seleção de modelos.

MAE – ERRO MÉDIO ABSOLUTO

Métrica de avaliação de modelos de regressão que **mede a média das diferenças absolutas** entre as previsões do modelo e os valores reais (observados).

Onde:

- n -> número de observações.
- y_i -> valor real da variável dependente.
- \hat{y}_i -> valor previsto pelo modelo.

INTERPRETAÇÃO DO MAE

Ele representa a **média das diferenças absolutas entre as previsões e os valores reais**, sem levar em conta a direção (positiva ou negativa) das diferenças.

Quanto  o valor do MAE, melhor o ajuste do modelo.

MSE – ERRO MÉDIO QUADRÁTICO

Métrica de avaliação de modelos de regressão que **mede a média dos quadrados das diferenças** entre as previsões do modelo e os valores reais.

$$MSE = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_i)^2$$

Onde:

- n -> número de observações.
- y_i -> valor real da variável dependente.
- \hat{y}_i -> valor previsto pelo modelo.

INTERPRETAÇÃO DO MSE

O MSE penaliza erros maiores mais fortemente do que erros menores, devido à natureza dos quadrados. Assim como o MAE, quanto menor o valor do MSE, melhor o ajuste do modelo.

TRANSFORMAÇÃO DE VARIÁVEIS

INTRODUÇÃO

Financeiros

Econômicos

Comportamento do consumo no varejo

Industria

Saúde

Redes Sociais

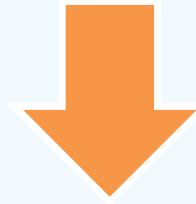
Segurança publica

INTRODUÇÃO

Já vimos que:

- A modelagem preditiva é essencial para análise de dados em diversas áreas.
- Abordagem teórica e estatística sólida é crucial para modelos precisos.

Responder perguntas a partir do fenômeno a ser estudado com as variáveis que eventualmente relacionam com esse fenômeno



Variáveis?
Podemos escolher os tipos que convém aos modelos?

INTRODUÇÃO

Notas de Alunos do
ENEM

- Faixa de Renda
- Escolaridade dos Pais
- Quantidade de horas de estudo no ano anterior ao vestibular

Análise de crédito

- Faixa de Renda
- Gastos mensais
- Perfil do cliente:
 - Gênero
 - Idade ...

INTRODUÇÃO



Modelo Preditivo Básico

- Estrutura do modelo:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + e_i$$

- Variáveis preditoras (X) influenciam a variável resposta (Y).

Como podemos inserir importantes variáveis qualitativas em modelos preditivos

INCORPORANDO VARIÁVEIS QUALITATIVAS

- ✓ Na modelagem é comum encontrarmos diversos tipos de **modelos** que têm como **pré-requisito que todas as variáveis sejam numéricas**
- ✓ Desafios ao **lidar com variáveis qualitativas** importantes para o modelo.
- ✓ Uso de **variáveis dummy** como solução para inserção de variáveis qualitativas.

O QUE SÃO VARIÁVEIS DUMMY ?

Variáveis Dummy são **variáveis dicotômicas** ou binárias (0 ou 1) criadas para representar uma variável com duas ou mais categorias.



VARIÁVEIS DUMMY PARA DUAS CATEGORIAS

SEXO		SEXO_Feminino
Feminino	→	1
Masculino	→	0
Feminino	→	1
Feminino	→	1
Masculino	→	0
Feminino	→	1

Precisa da variável : SEXO_Masculino ?

VARIÁVEIS DUMMY PARA DUAS CATEGORIAS



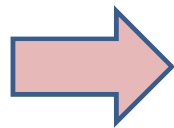
Como ficaria o Modelo de regressão ?

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{DUMMY} + e_i$$

Onde:

- $\hat{\beta}_0$ representa o valor médio da variável resposta, para as observações que apresentam um valor 0 da variável x.
- A soma $\hat{\beta}_0 + \hat{\beta}_1$, representa o valor médio de y para as observações quando $x = 1$

UF
MG
SP
RJ
ES
MG
ES
SP
RJ
RJ
ES
MG
ES



UF_MG
1
2
3
4
1
0
2
3
3
4
1
0

**EXEMPLO DE VARIÁVEIS
DUMMY PARA MAIS DE DUAS
CATEGORIAS**

ERRADO!!!!

VARIÁVEIS DUMMY PARA MAIS DE DUAS CATEGORIAS

Por que está errado?

Dois valores próximos serão mais parecidos que valores distantes;

- MG(1) está mais próximo de SP(2) do que ES(4)

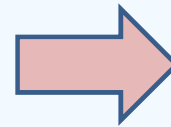
Por conta da diferença dos valores alguns atributos podem ter **maior peso** que os outros;

- ES -> Peso 4
- MG -> Peso 1

VARIÁVEIS DUMMY PARA MAIS DE DUAS CATEGORIAS

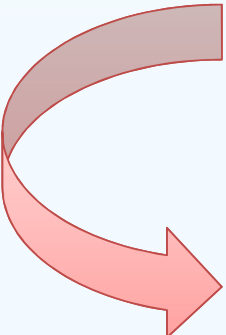
N-1 variáveis!

UF
MG
SP
RJ
ES
MG
ES
SP
RJ
RJ
ES
MG
ES



UF_MG	UF_SP	UF_RJ
1	0	0
0	1	0
0	0	1
0	0	0
1	0	0
0	0	0
0	1	0
0	0	1
0	0	1
0	0	0
1	0	0
0	0	0

VARIÁVEIS DUMMY PARA MAIS DE DUAS CATEGORIAS



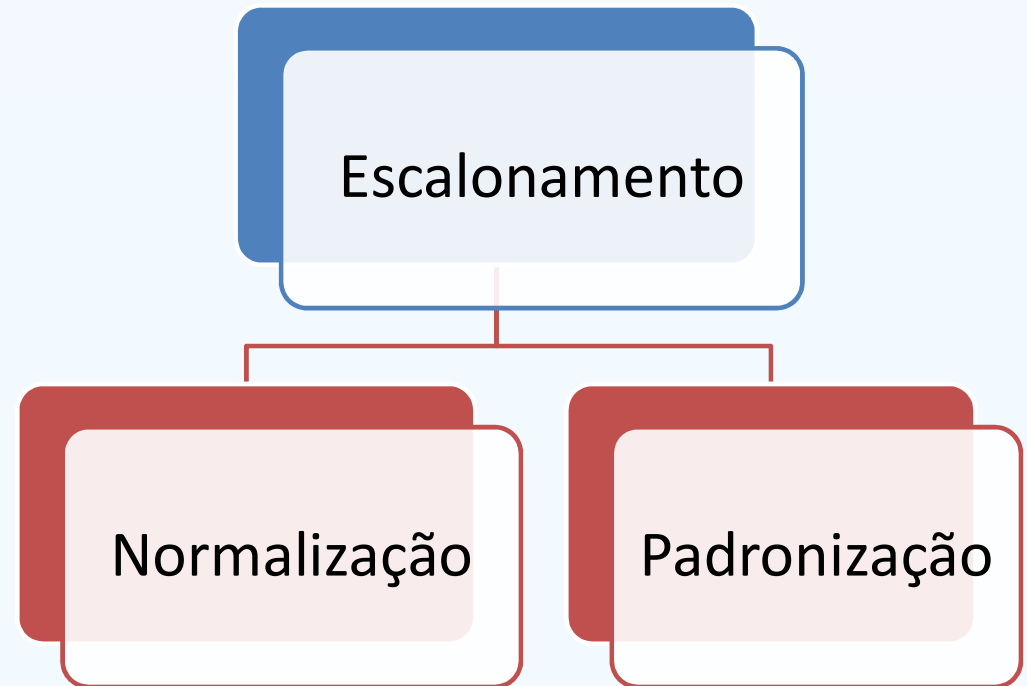
Como ficaria o Modelo de regressão ?

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{DUMMY_MG} + \hat{\beta}_2 X_{DUMMY_SP} + \hat{\beta}_3 X_{DUMMY_RJ} + e_i$$

- $\hat{\beta}_0$ representa o valor médio da variável resposta, para o ES
- A soma $\hat{\beta}_0 + \hat{\beta}_1$, representa o valor médio de y para MG
- A soma $\hat{\beta}_0 + \hat{\beta}_2$, , representa o valor médio de y para SP
- A soma $\hat{\beta}_0 + \hat{\beta}_3$, , representa o valor médio de y para RJ

ESCALONAMENTO

- É uma das transformações mais importantes ;
- A maioria dos algoritmos não funcionam bem quando **atributos numéricos de entrada tem escalas muito diferentes**
- Geralmente não é necessário escalonar os valores alvos ou target (Y)



NORMALIZAÇÃO: ESCALA MIN - MAX

- Os valores são deslocados e redimensionados para que **variem entre 0 e 1**

$$X_{normalizado} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- No *sklearn* possui um transformador chamado ***MinMaxScaler***
- No *R* o *preprocess* do *caret*, *method = range*

PADRONIZAÇÃO

- Muito utilizada por ser bem mais sensível aos outliers.

$$Z = \frac{x - \mu}{\sigma}$$

- No **sklearn** possui um transformador chamado **StandardScaler**
- No R **scale**

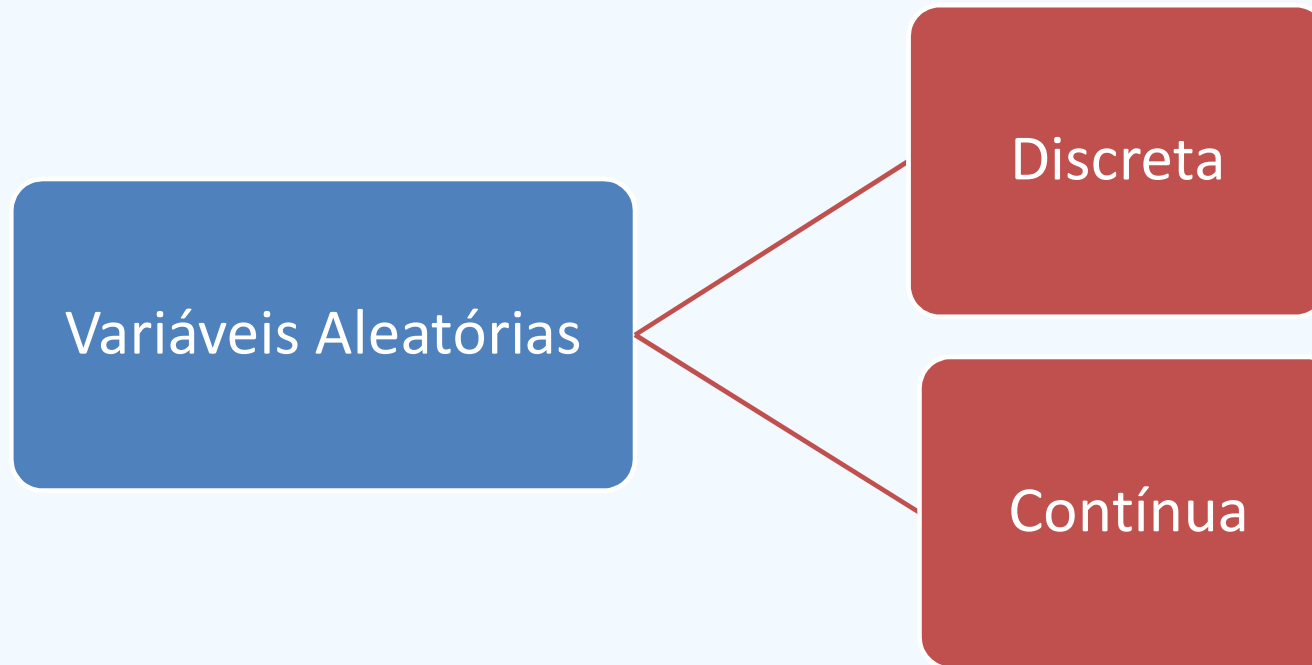
Introdução aos Modelos Lineares Generalizados - GLM

Como selecionar o modelo correto ?



Conhecemos o fenômeno aleatório gerador dos dados

Classificação de Variáveis Aleatórias



Variáveis discretas → suporte em um conjunto de valores enumeráveis (finitos ou infinitos)

Variáveis contínuas → suporte em um conjunto não enumerável de valores

Transformação de Box - Cox

Transforma o valor observado y , desde que esse valor seja positivo em:

$$z = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{para } \lambda \neq 0 \\ \log y, & \text{para } \lambda = 0 \end{cases}$$

Em que λ é uma constante conhecida.

Tem o objetivo de produzir aproximadamente a normalidade, variabilidade constante e a linearidade.

Essa média é igual a um η , onde ele será a combinação linear de nossas covariáveis.

$$E(Z) = \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Transformação de Box - Cox

As suposição não são atendidas.

A transformação NÃO será capaz de ajustar os dados

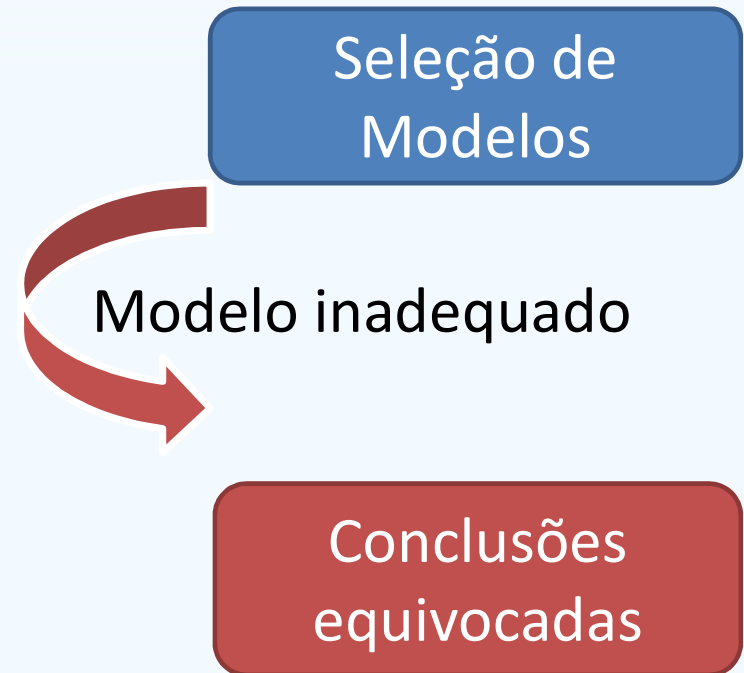
As suposição são atendidas

Temos um intervalo de valores de λ nas quais essas suposição é garantida

Antes dos computadores, os modelos de regressão eram aplicados mesmo quando a **distribuição da variável resposta não era normal**. Por exemplo, ao lidar com proporções, a distribuição normal era usada e transformações eram aplicadas para tentar ajustar os dados à normalidade.

Introdução

- ❑ A seleção de modelos é o procedimento mais importante na modelagem;
- ❑ A partir do modelo que realizaremos a inferência e predição
- ❑ Modelos parcimoniosos



Recapitulando o Modelos Lineares Clássicos

$$Y = \mu + \varepsilon$$

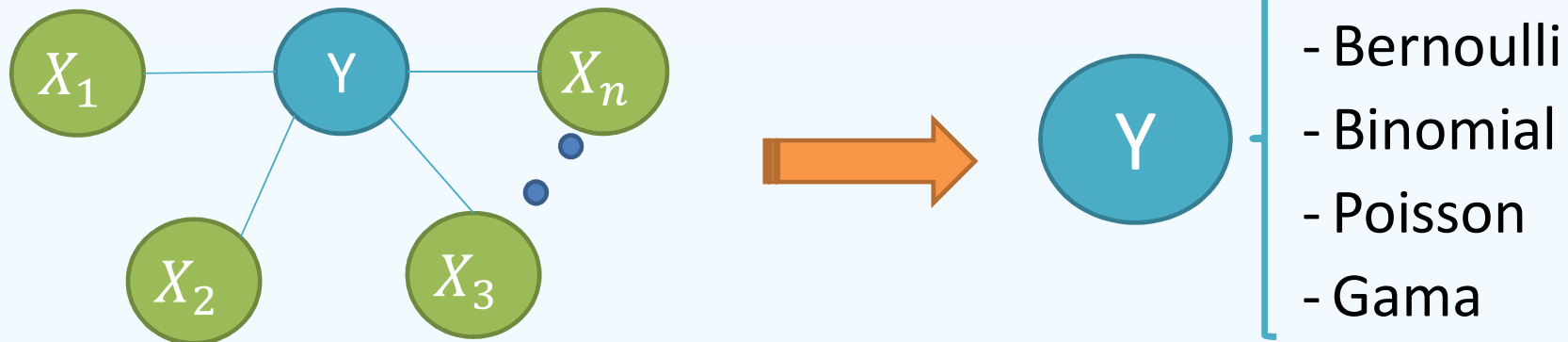
Pressupostos:

- $Y \sim N(\mu, \sigma^2)$;
- Aditividade do modelo;
- Homogeniedade das variâncias residuais;
- $e \sim N(0, \sigma^2)$;
- Resíduos independentes;

Onde:

- **Y**: vetor da variável resposta
- **$\mu = E(Y) = X\beta$** , **X**: Matriz de Dimensões
- **β** : Vetor de coeficientes
- **ε** : Componente aleatório

Fenômeno a ser estudado



Diagnóstico sobre Y

- Fundamental importância
- Elaborado antes da criação dos modelos
- Auxilia na escolha do melhor modelo

Origem dos GLM's

Criação de modelos para trabalhar com problemas específicos;

Modelos para dados qualitativos , qualitativos. Mesmo não seguindo distribuição Normal.

Pontos de convergência entre os modelos

Os modelos poderiam ser generalizados se utilizássemos uma função de ligação.

Nelder & Wedderburn (1975) propuseram uma teoria unificadora da modelagem estatística a que deram o nome de modelos lineares generalizados (GLM), como uma extensão dos modelos lineares classe

Surgimento dos GLM's

- Nelder & Wedderburn mostraram , que a maioria dos problemas estatísticos resolvidos por modelos de regressão podem ser generalizados pelo denominado “modelo linear generalizado”.
- Esses modelos envolvem uma variável resposta univariada , variáveis explicativas e uma amostra aleatória de n observações sendo que:

COMPONENTES DE UM GLM

1° Componente

- Componente aleatório. Variável resposta (Y)
- segue uma distribuição que pertence à família exponencial.

2° Componente

- Componente sistemático
- É a parte fixa do modelo - as covariáveis (estrutura linear do nosso modelo)

3° Componente

- vai ligar o componente aleatório com o componente sistemático.
- Chamaremos de função de ligação(*link function*)

Identificando os componentes em um modelo clássico

$$Y = \mu + \varepsilon$$

Onde:

- **Y**: vetor da variável resposta
- **$\mu = E(Y) = X\beta$** , o componente sistemático
- **X**: Matriz de Dimensões
- **β** : Vetor de coeficientes
- **E**: Componente aleatório
- $Y \sim N(\mu, \sigma^2)$;

- Identificando a função de ligação : $g(\mu) = XB$
- No caso do modelo clássico a função de $g(\mu) = \mu$ (identidade)

DEFININDO O GLM ALGEBRICAMENTE

$$\eta = \alpha + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_n X_{ni} + e_i$$

Onde:

η : função de ligação canônica

α : representa a constante

$\hat{\beta}_1$ a $\hat{\beta}_k$: coeficientes de cada uma das variáveis independentes $X_{1i}, X_{2i}, \dots, X_{ni}$.

Modelo	Característica de Y	Distribuição	Função de ligação Canônica (η)
Linear	Quantitativa	Normal	\hat{Y}
Logística binária	Qualitativa com 2 categorias (Dummy)	Bernoulli	$\ln\left(\frac{p}{1-p}\right)$
Logística Multinomial	Qualitativa com 3 ou mais categorias	Binomial	$\ln\left(\frac{p_m}{1-p_m}\right)$
Poisson	Quantitativa com dados de contagem	Poisson	$\ln(\lambda)$
Binomial Negativa	Quantitativa com dados de contagem	Poisson-gama	$\ln(\lambda)$

Especificações dos modelos lineares generalizados

Linear:

$$\hat{Y} = \alpha + \hat{\beta}_1 X_{1i} + \hat{\beta}_1 X_{2i} + \dots + \hat{\beta}_n X_{nn} + e_i$$

**Logística
Binária:**

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \hat{\beta}_1 X_{1i} + \hat{\beta}_1 X_{2i} + \dots + \hat{\beta}_n X_{nn}$$

**Logística
Multinomial:**

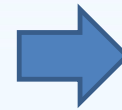
$$\ln\left(\frac{p_m}{1-p_m}\right) = \alpha + \hat{\beta}_1 X_{1i} + \hat{\beta}_1 X_{2i} + \dots + \hat{\beta}_n X_{nn}$$

**Dados de
Contagem:**

$$\ln(\lambda) = \alpha + \hat{\beta}_1 X_{1i} + \hat{\beta}_1 X_{2i} + \dots + \hat{\beta}_n X_{nn}$$

Família exponencial

As distribuições pertencentes a Família exponencial já são conhecidas



Quais distribuições poderemos aplicar aqui o GLM

Família
exponencial de
distribuições

Discreta

Contínuas

Independente do tipo dos dados Estaremos Aplicando o **Modelo linear Generalizado**

Não é necessário trabalhar com transformações, podemos utilizar apenas o componente aleatório e aplicar o GLM.

Família exponencial uniparamétrica:

É caracterizada por uma função de probabilidade ou densidade que depende de um parâmetro desconhecido $\theta \in \Theta$, especificada na forma:

$$f(\mathbf{x}; \theta) = h(\mathbf{x}) \cdot \exp[\eta(\theta) \cdot \mathbf{t}(\mathbf{x}) - \mathbf{b}(\theta)]$$

em que as funções $\eta(\theta)$, $\mathbf{t}(\mathbf{x})$ e $h(\mathbf{x})$ assumem valores no subconjunto dos reais e não são únicos.

Temos que pelo teorema da fatoração de Neyman, a estatística $T(\mathbf{X})$ é suficiente para θ . Ou seja, a partir dessa estatística podemos construir estimadores ótimos (máxima verossimilhança)

Exemplo

A distribuição de Rayleigh, usada para análise de dados contínuos positivos, tem função densidade ($x > 0, \theta > 0$) dada por :

$$f(x; \theta) = \frac{x}{\theta^2} \cdot \exp\left(-\frac{x^2}{2\theta^2}\right)$$

Essa função depende da família exponencial?

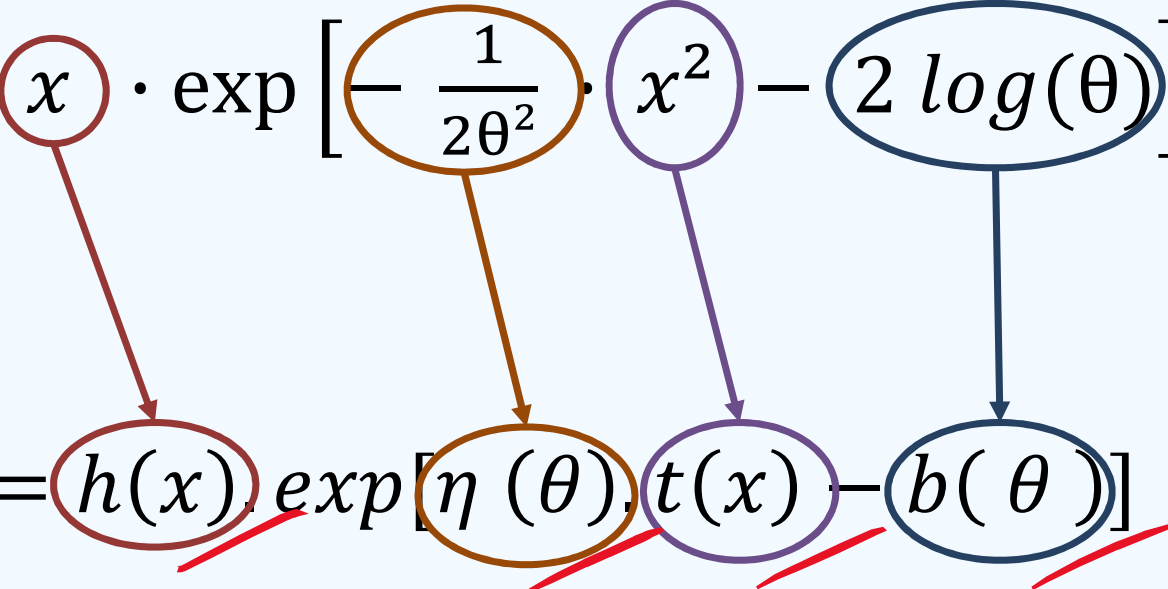
$$\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{h}(\mathbf{x}) \cdot \exp[\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \mathbf{t}(\mathbf{x}) - \mathbf{b}(\boldsymbol{\theta})]$$

Observa-se que:

$$f(x; \theta) = x \cdot \exp\left[-\frac{1}{2\theta^2} \cdot x^2 - 2 \log(\theta)\right]$$

Exemplo

$$f(x; \theta) = x \cdot \exp \left[-\frac{1}{2\theta^2} \cdot x^2 - 2 \log(\theta) \right]$$

$$f(x; \theta) = \cancel{h(x)} \cdot \exp[\cancel{\eta(\theta)} \cdot \cancel{t(x)} - \cancel{b(\theta)}]$$


Portanto a distribuição de Rayleigh pertence a família exponencial.

Distribuição Conjunta

Sejam X_1, X_2, \dots, X_n variáveis i.i.d que seguem distribuição pertencente a família exponencial. A distribuição conjunta de X_1, X_2, \dots, X_n é dada por:

$$f(x_1, x_2, \dots, x_n; \theta) = [\prod_{i=1}^n h(x_i)] \cdot \exp[\eta(\theta) \sum_{i=1}^n t(x_i) - n b(\theta)]$$

- ❑ A distribuição conjunta de X_1, X_2, \dots, X_n também é um modelo que pertence à família exponencial.
- ❑ A estatística suficiente de um modelo da família exponencial tem distribuição, também, pertencente à família exponencial.

Por exemplo se $X_1, X_2, \dots, X_n \sim \text{Poisson}(\theta)$ então a estatística suficiente T segue distribuição de Poisson, isto é, $T(X_i) \sim \text{Poisson}(n\theta)$

Família exponencial na forma canônica

Caso especial da família exponencial:

$$f(x; \theta) = h(x) \cdot \exp[\eta(\theta) \cdot t(x) - b(\theta)]$$

Onde:

$$\eta(\theta) = \theta \text{ e } t(x) = x \quad \longrightarrow \quad \text{Função identidade}$$

A família exponencial na forma canônica é definida considerando que as funções $\eta(\theta)$ e $t(x)$ são iguais à funções identidade, na forma:

$$f(x; \theta) = h(x) \cdot \exp[\theta x - b(\theta)]$$

Definição de um GLM

Definição de um GLM



Quem são os componentes e como conseguimos identificá-los

Um modelo linear generalizado é definido pela especificação de três componentes, o componente aleatório, sistemático e uma função de ligação.

Componentes de um GLM

Seja Y uma variável aleatória associada a um conjunto de variáveis explanatórias X_1, \dots, X_p . Para uma amostra aleatória de tamanho n , (y_i, x_i) em que $x_i = (x_{i1}, \dots, x_{ip})^T$ é o vetor coluna de variáveis explanatórias, o GLM envolve os três componentes:

1º - Componente aleatório: $Y \sim \text{família} - \text{exponencial}(\theta_i, \phi)$

$$f(x; \theta; \phi) = \exp\{\phi^{-1}[y\theta - b(\theta)] + c(y, \phi)\}$$

em que $\phi > 0$ é um parâmetro de dispersão e o θ_i o parâmetro canônico

Propriedades do componente aleatório

Temos que:

$$E(Y_i) = \mu_i = b'(\theta_i)$$

$$\text{Var}(Y_i) = \phi b''(\theta_i) = \phi V(\mu_i)$$

Em que **$V(\mu_i)$** é a **função de variância** que depende somente da média μ_i .

2º Componente sistemático: É o preditor linear do modelo, em que são inseridas as covariáveis por meio de uma combinação linear de parâmetros isto é:

$$\eta_i = \sum_{r=1}^p x_{ir} \beta_r = x_i^T \beta$$

Parâmetro de interesse será os β -> é ele que queremos estimar.

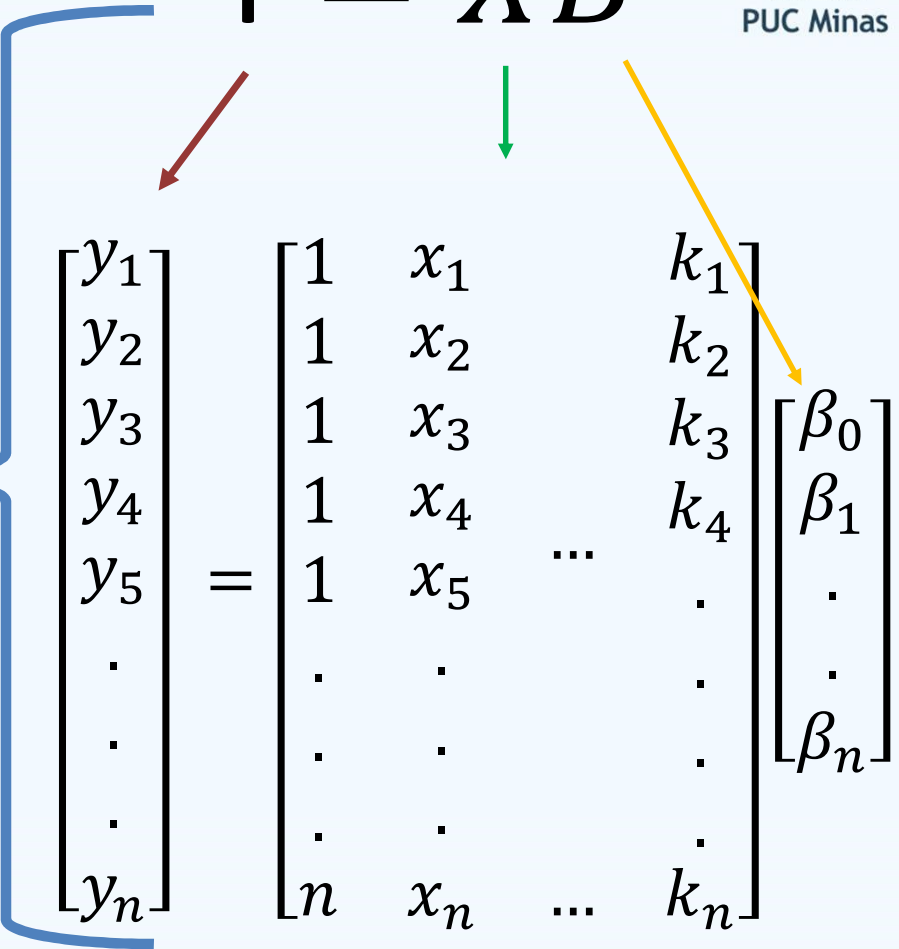
Ou

$$\eta_i = X\beta$$

Em que $\mathbf{X} = (x_1, \dots, x_n)^T$ é a matriz do modelo de $\beta = (\beta_1, \dots, \beta_p)^T$ é o vetor de parâmetros desconhecidos e $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$ é o preditor linear

$$Y = XB$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \dots + \hat{\beta}_n K_i$$


$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \vdots \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & & k_1 \\ 1 & x_2 & & k_2 \\ 1 & x_3 & & k_3 \\ 1 & x_4 & & k_4 \\ 1 & x_5 & \dots & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ n & x_n & \dots & k_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_n \end{bmatrix}$$

3º Função de ligação: é uma função que relaciona o componente aleatório ao componente sistemático, ou seja, vincula a média ao preditor linear, isto é:

$$\eta_i = g(\mu_i)$$

Sendo $g(.)$ uma função monótona e diferenciável



```
graph TD; A[Componente Aleatório] --- B[Função de ligação]; B --- C[Componente Sistemático];
```

Componente Aleatório

Função de ligação

Componente Sistemático

Seja o valor esperado $E(Y_i | x_{i1}, \dots, x_{ip})$ para $i = 1, \dots, n$.

Então, temos que (definição de GLM):

Função da
média

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

ou

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})$$

Média em termos da inversa $g(\cdot)$, aplicada ao preditor linear.