PÓS-GRADUAÇÃO EM 2025
PUC MINAS

Disciplina:

TÉCNICAS DE AMOSTRAGEM E REGRESSÃO LINEAR

Professora: Anaíle Mendes Rabelo

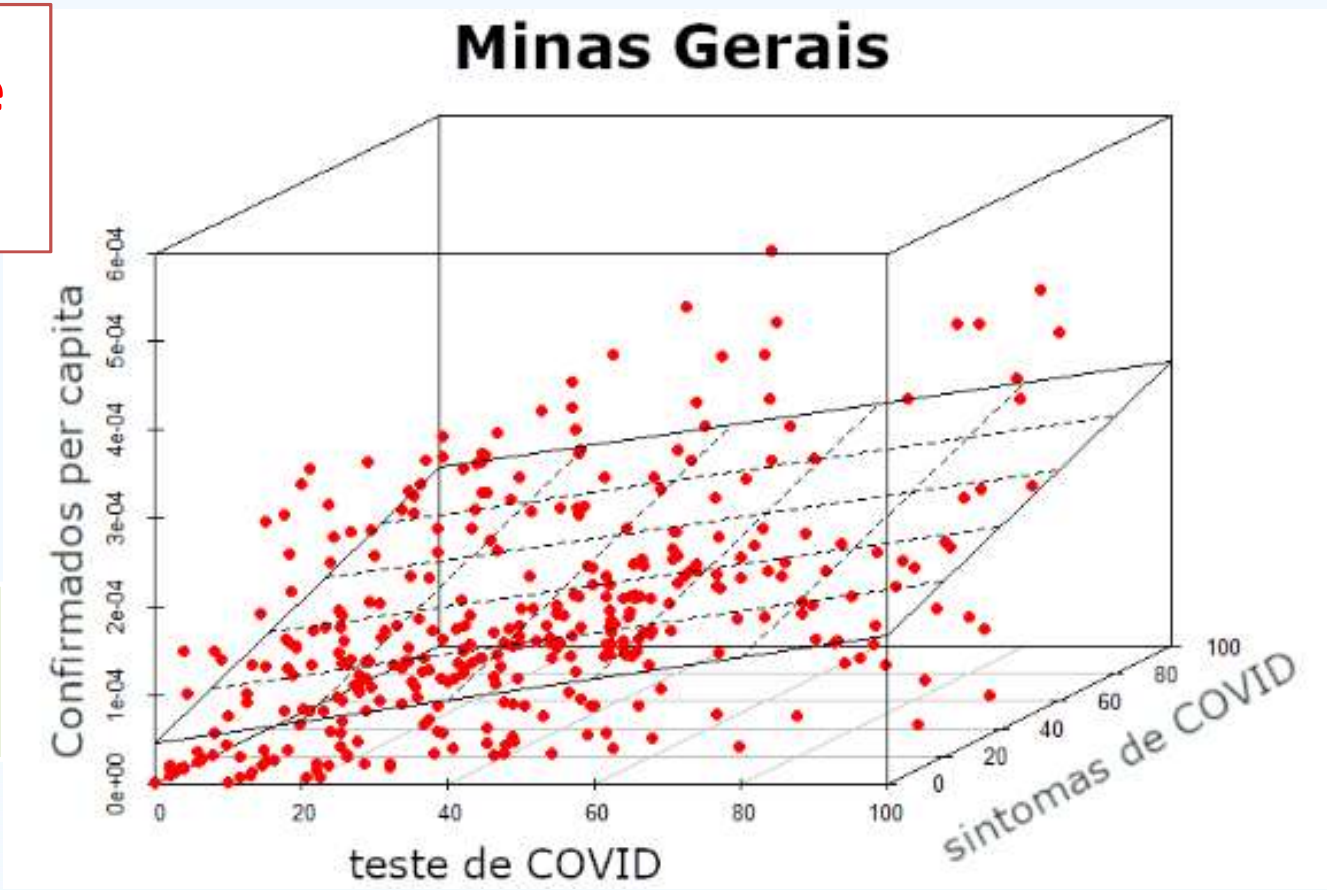
REGRESSÃO LINEAR MÚLTIPLA

PUC Minas

REGRESSÃO LINEAR MÚLTIPLA

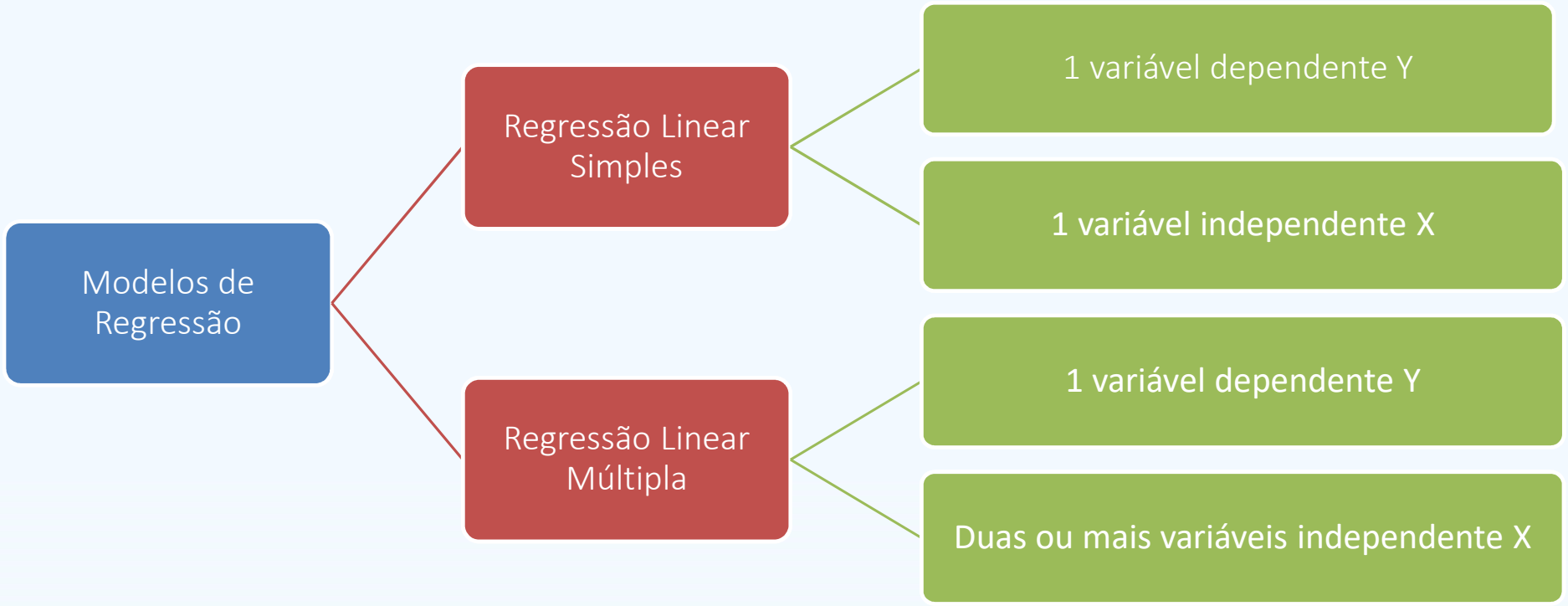
E quando possuímos mais de uma variável independente?

A solução está na : Regressão Linear Múltipla



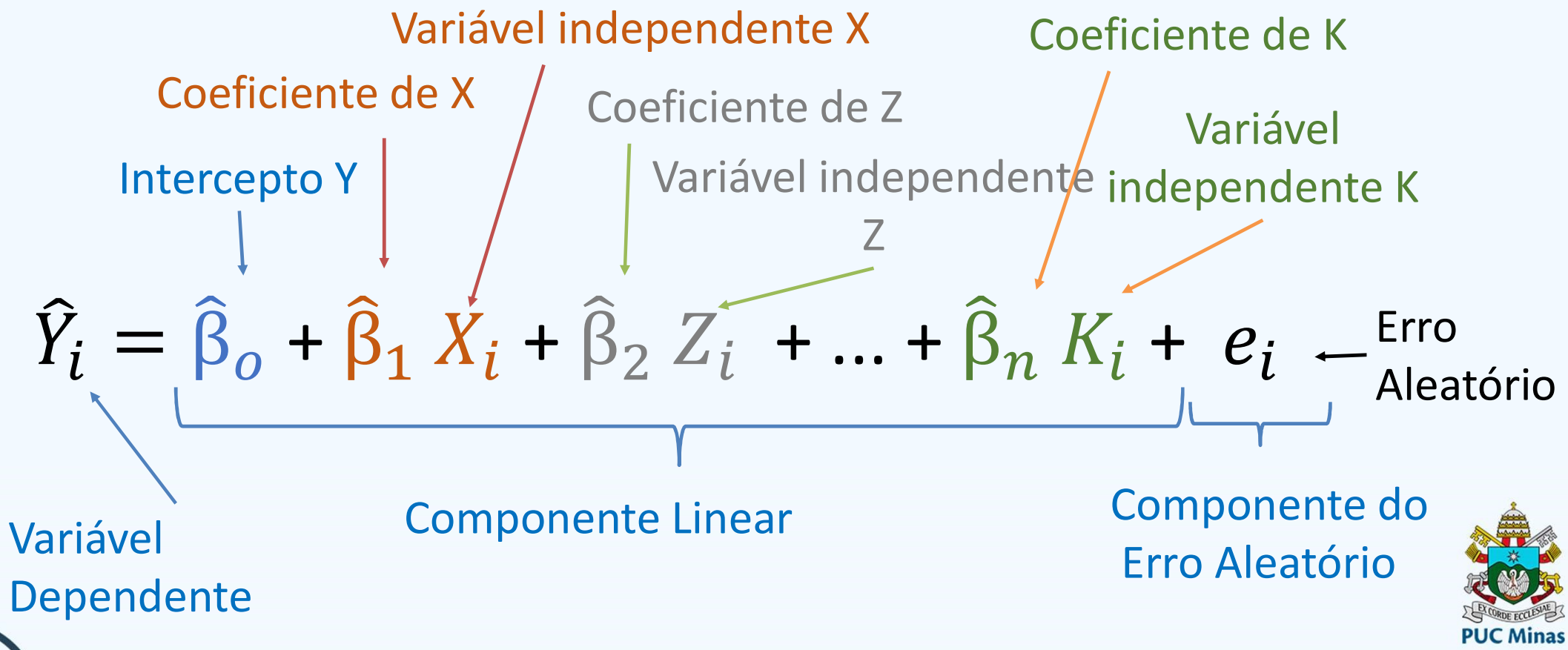
PUC Minas

REGRESSÃO LINEAR MÚLTIPLA



PUC Minas

REGRESSÃO LINEAR MÚLTIPLA



PUC Minas

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

$$Y = X\,B + e$$

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

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

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

$$Y = XB + e$$

$$e = Y - XB$$

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

$$e'e = \begin{bmatrix} e_1 & e_2 & e_3 & e_4 & e_5 & \cdot & \cdot & \cdot & e_n \end{bmatrix} X \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix} = e^2{}_1 + e^2{}_2 + e^2{}_3 + e^2{}_4 + e^2{}_5 + \ldots + e^2{}_n$$

PUC Minas

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

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

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

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

PUC Minas

ESTIMATIVA DOS COEFICIENTES: FORMA MATRICIAL

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

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

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

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

PUC Minas

PREMISSAS DA REGRESSÃO LINEAR MÚLTIPLA

- ❑ Análise de Outiliers de resíduos
- ❑ Homocedasticidade
- ❑ Normalmente distribuído
- ❑ **Ausência de multicolinearidade e autocorrelação**

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

- **Média = 0**
- **Variância constante**
- **Covariância = 0**

PUC Minas

ANÁLISE DE OUTLIERS

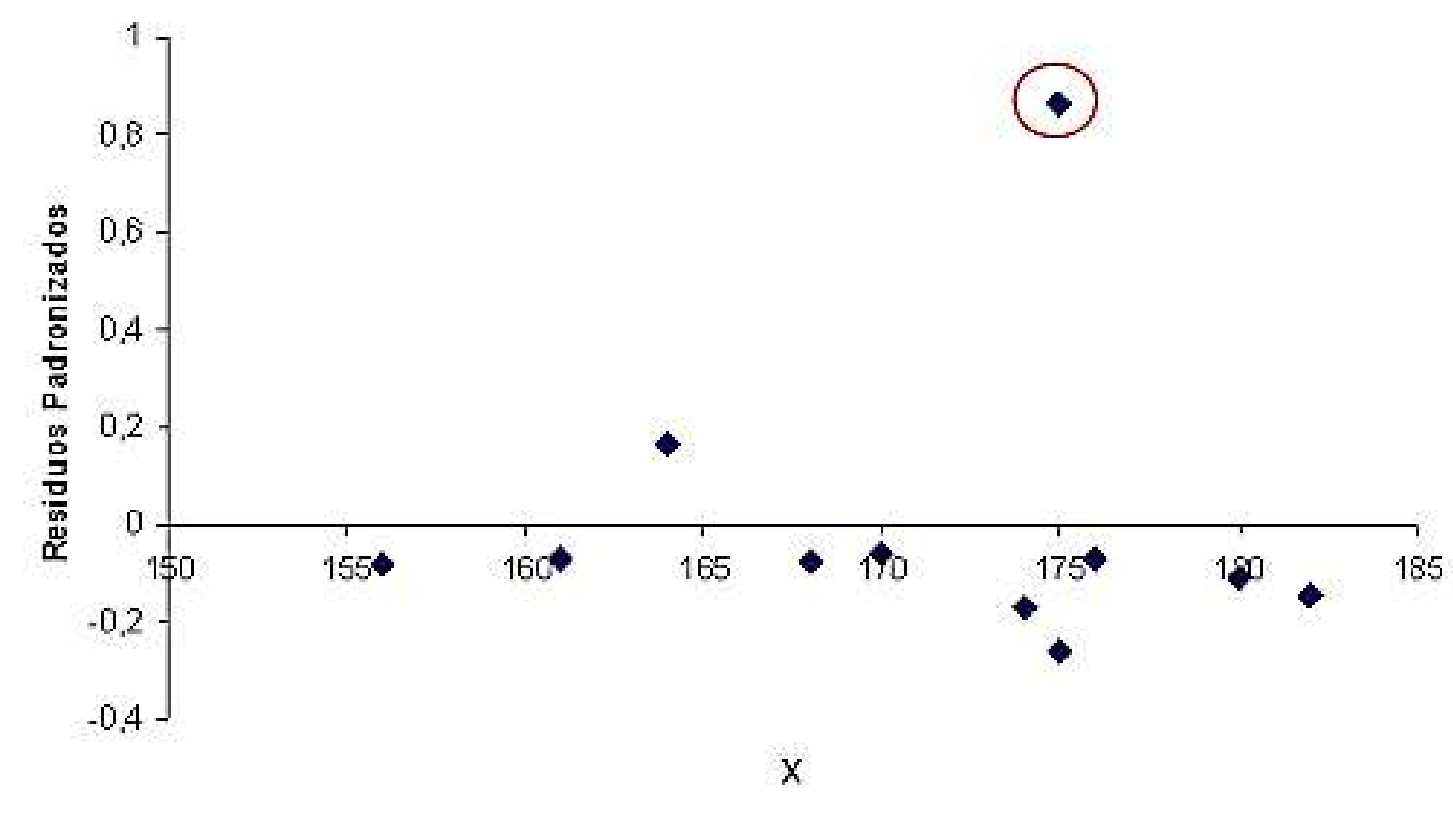


Gráfico de Resíduos padronizados vs Valores ajustados

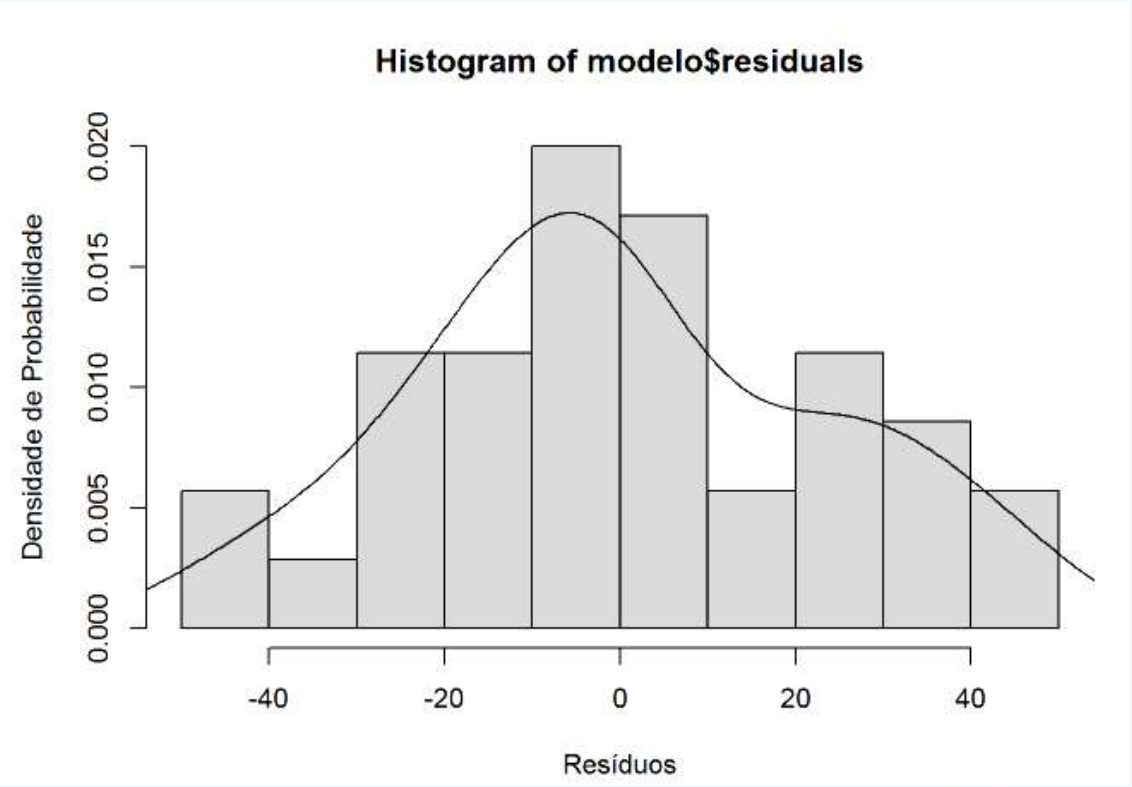
PUC Minas

NORMALIDADE DOS RESÍDUOS

Teste de Shapiro Wilk

H_0 = distribuição normal : p > 0.05

H_1 = distribuição não normal : p <= 0.05



ANÁLISE DA HOMOCEDASTICIDADE DOS RESÍDUOS

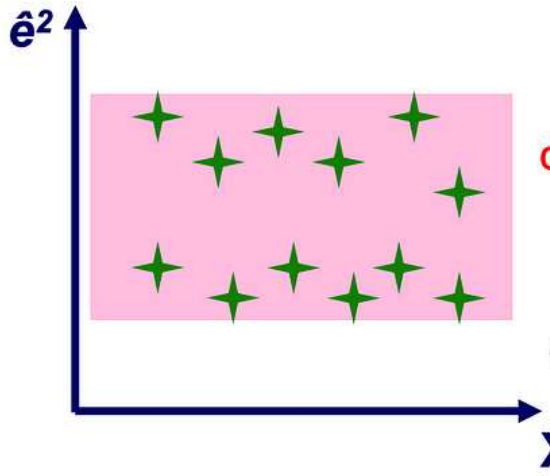
Homocedasticidade: A variância dos erros e, condicionada aos valores das variáveis explanatórias, será constante.

Teste Breusch-Pagan (Homocedasticidade)

Ho = existe homocedasticidade : p > 0.05

Ha = não existe homocedasticidade : p <= 0.05

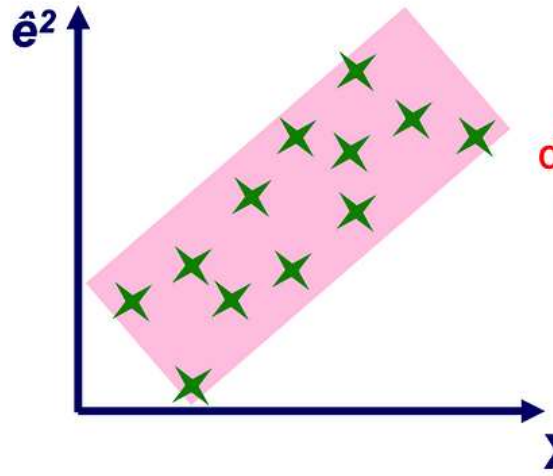
Homocedasticidade



A dispersão dos resíduos é a mesma ao longo de *X*

$$\sigma^2 = constante$$

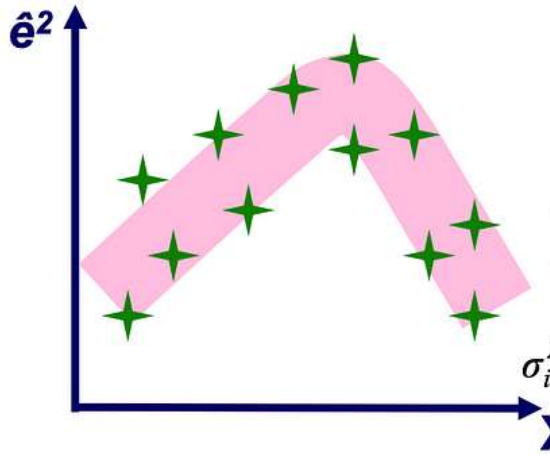
Heterocedasticidade



A dispersão dos resíduos é uma função linear de *X*

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

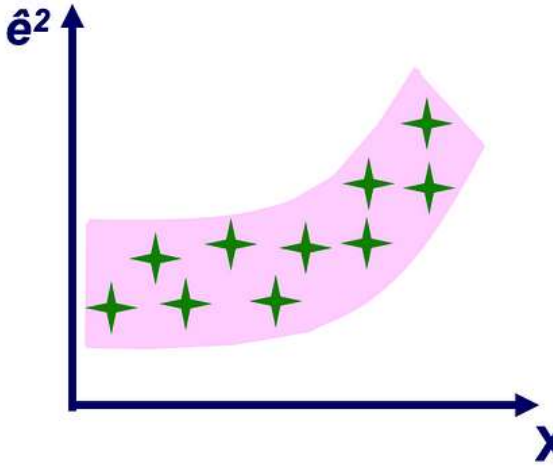
Heterocedasticidade



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

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

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

$$H_0 : p - valor \geq 0{,}05$$
$$H_1 : p - valor < 0{,}05$$

PUC Minas

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.

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

H_1: Pelo menos um β_j é diferente de zero

$$H_0 = F_{Calc} \leq F_{Crítico} \; ou \; p - valor \geq 0{,}05$$
$$H_1 = F_{Calc} > F_{Crítico} \; ou \; p - valor < 0{,}05$$

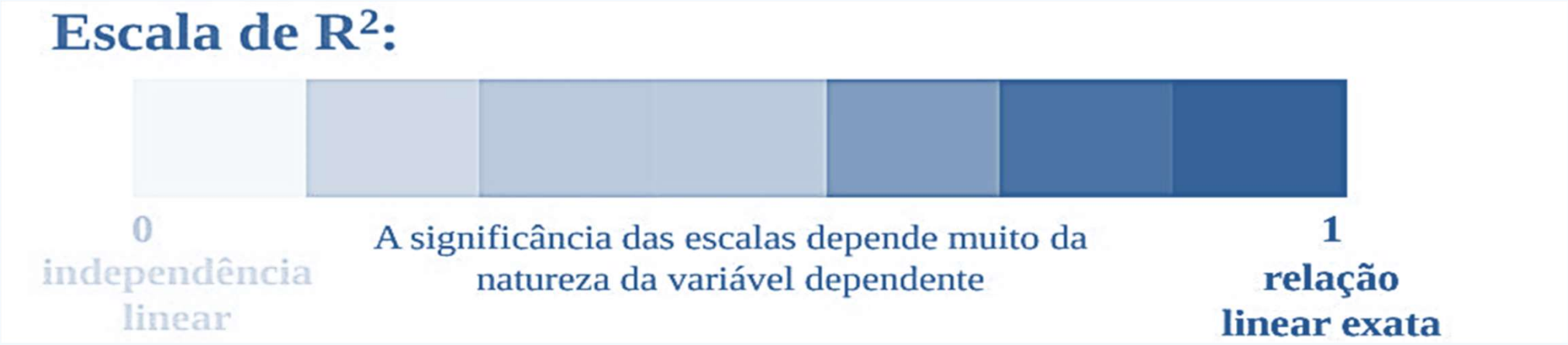
PUC Minas

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 independente do modelo de regressão.

$$R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^{n}(\widehat{Y_i} - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n} \widehat{e_i}^2}{\sum_{i=1}^{n} y_i^2} = \widehat{\beta_1}^2 \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2}$$



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.

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

PUC Minas

MULTICOLINEARIDADE

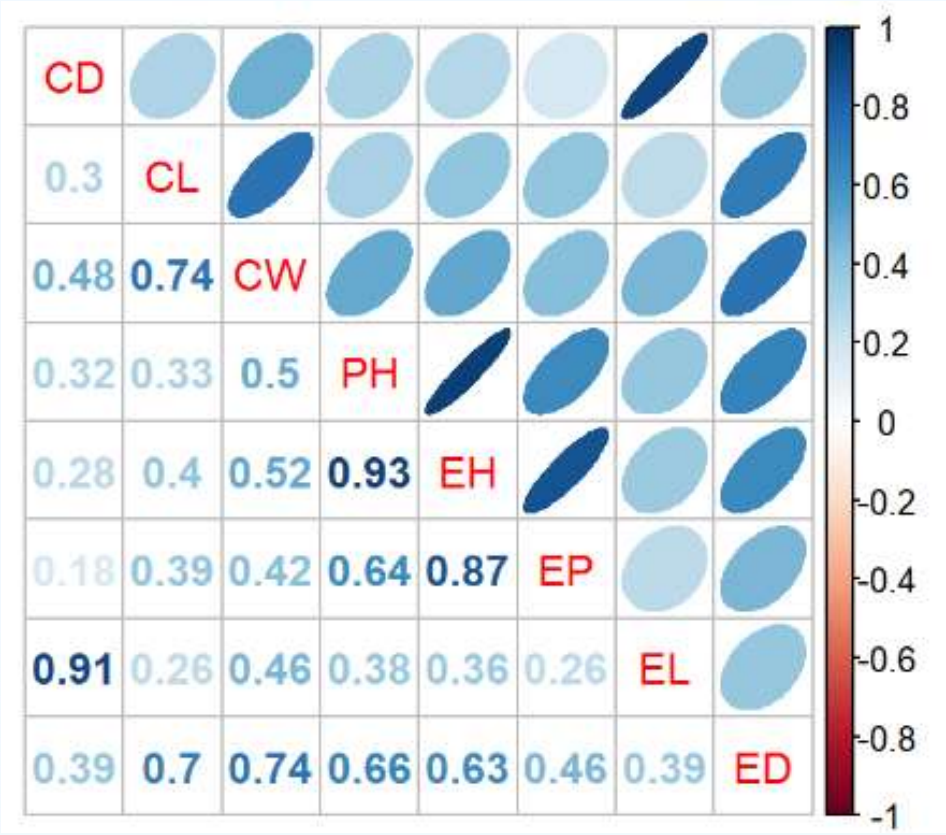
- O valor do fator de inflação da variância (VIF), que mede **quanto a variância do coeficiente estimado para uma variável é inflada** devido à multicolinearidade com as outras variáveis independentes.

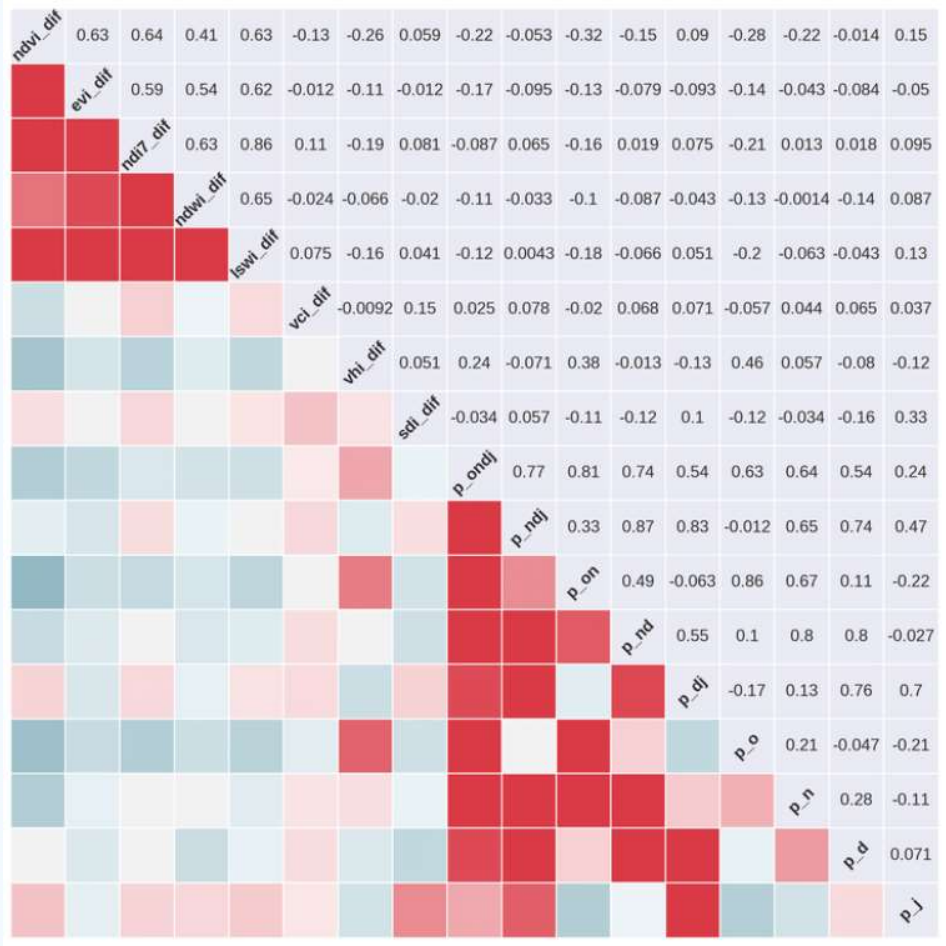
- VIFs maiores que 10 indicam alta multicolinearidade, enquanto valores entre 5 e 10 podem ser preocupantes.

- A maneira mais simples de lidar com a multicolinearidade é excluir a variável multicolinear

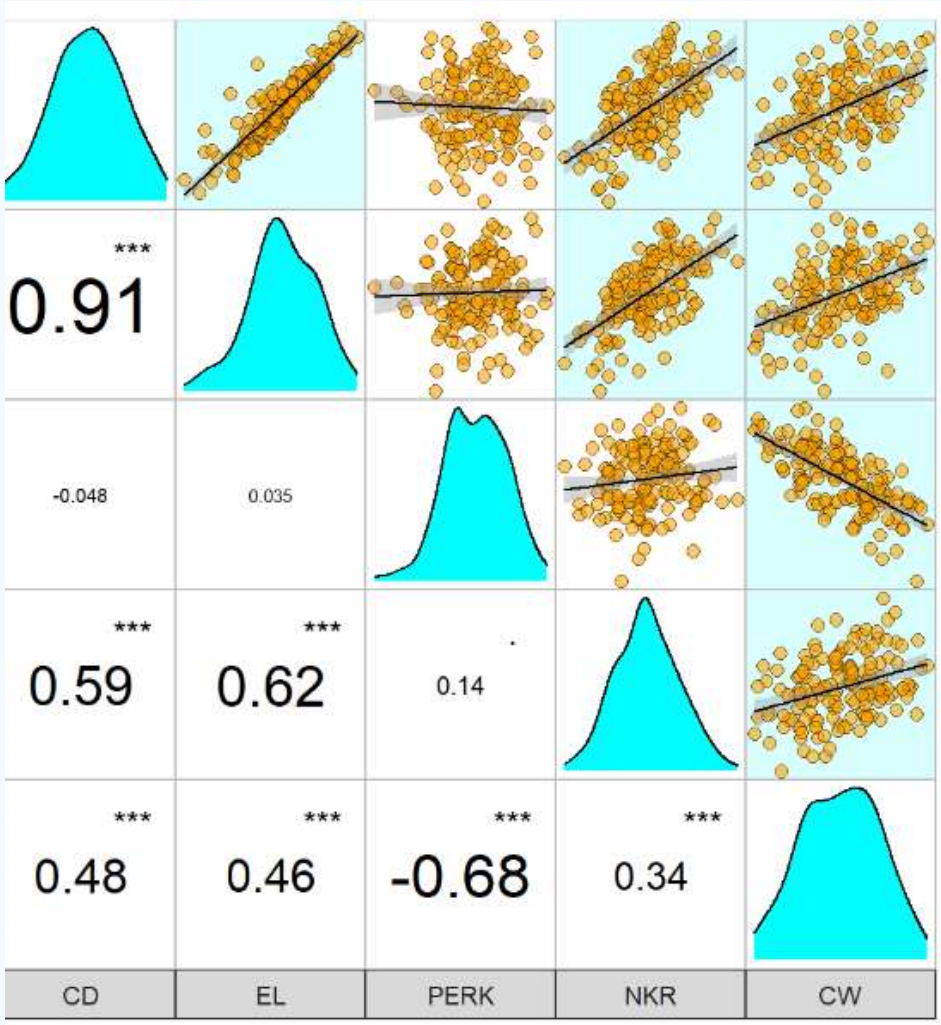
PUC Minas

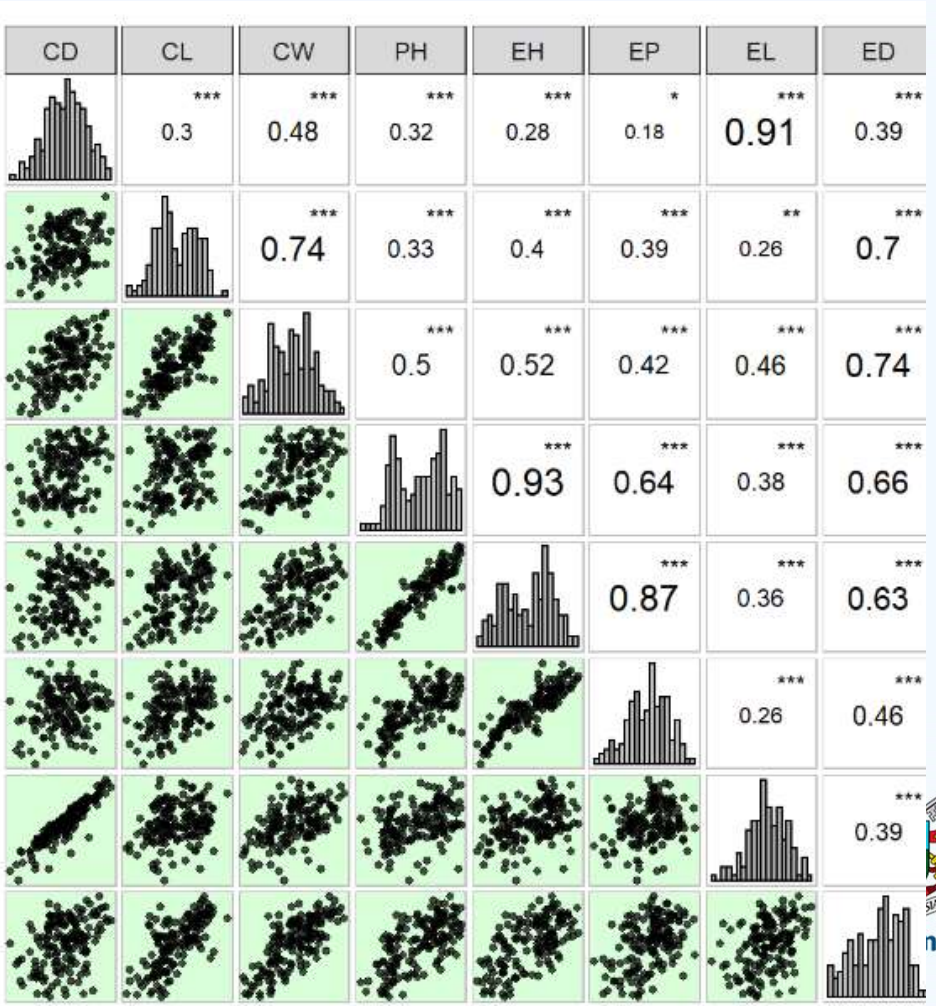
Scatter plot de uma matriz de correlação





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

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
```

Estimador dos coeficientes

P - valor do teste t

R^2 Ajustado

Detalhamento da saída do PYTHON

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

	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	0.863	-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

R^2

R^2 - ajustado

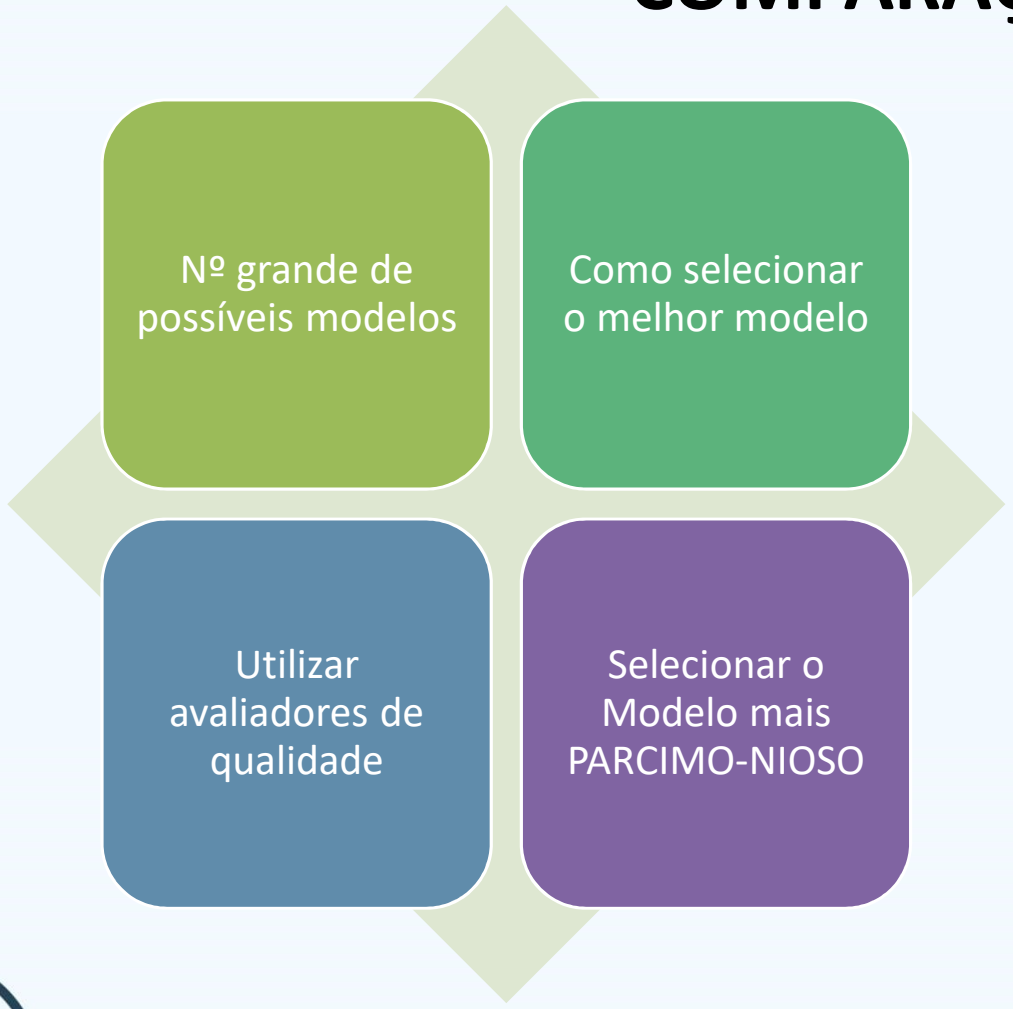
Teste F

Estimadores dos coeficientes

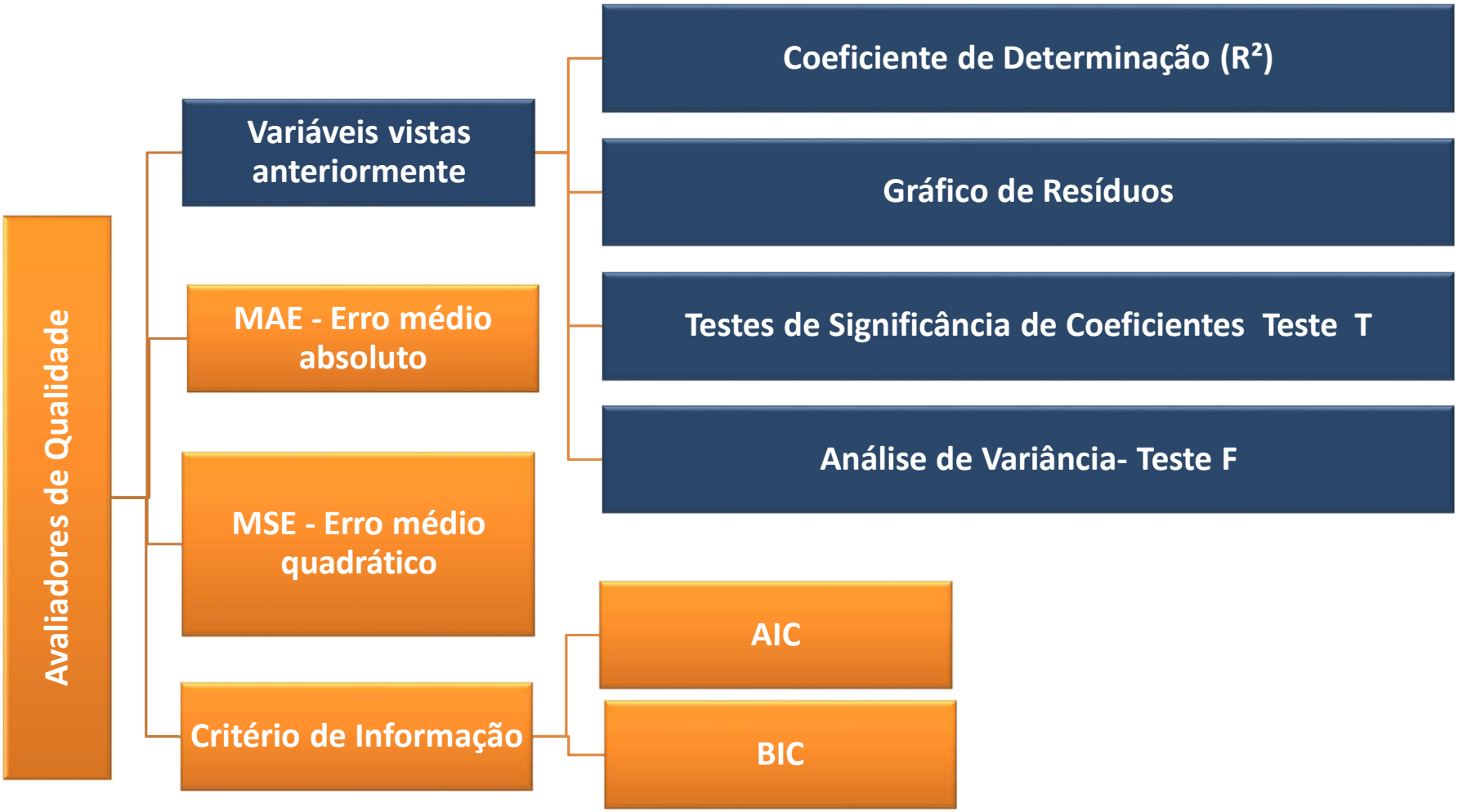
Teste T

COMPARAÇÃO DE MODELOS

COMPARAÇÃO DE MODELOS



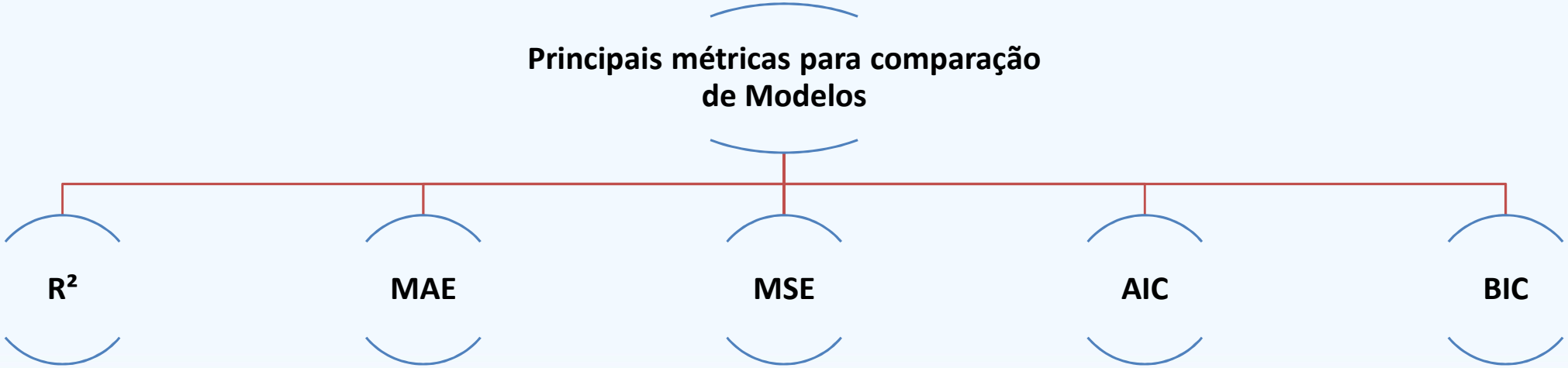
Modelo Parcimonioso é o modelo que **melhor explica o fenomeno que estamos estudando da forma mais simples possível**, menor número de parâmetros.



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**



PUC Minas

CRITÉRIO DE INFORMÇÃO 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

$$AIC = -n \cdot \ln\left({}^{SQR}/_{n}\right) + 2p$$

Nº de parâmetros

PUC Minas

INTERPRETAÇÃO DO AIC

- ↓ AIC, melhor o ajuste do modelo aos dados.

- O termo *-2*ln(SQR)* penaliza a qualidade do ajuste, **favorecendo modelos que explicam melhor a variabilidade dos dados.**

- O termo *2*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 INFORMÇÃO 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

$$BIC = -n \cdot \ln\left({}^{SQR}/_{n}\right) + \ln(n)p$$

P -> Nº de parâmetros

N -> tamanho da amostra

PUC Minas

INTERPRETAÇÃO DO BIC

- ↓ BIC, melhor o ajuste do modelo aos dados.

- **O termo -2 * ln(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.

PUC Minas

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).

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_i|$$

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.

PUC Minas

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.

PUC Minas

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.