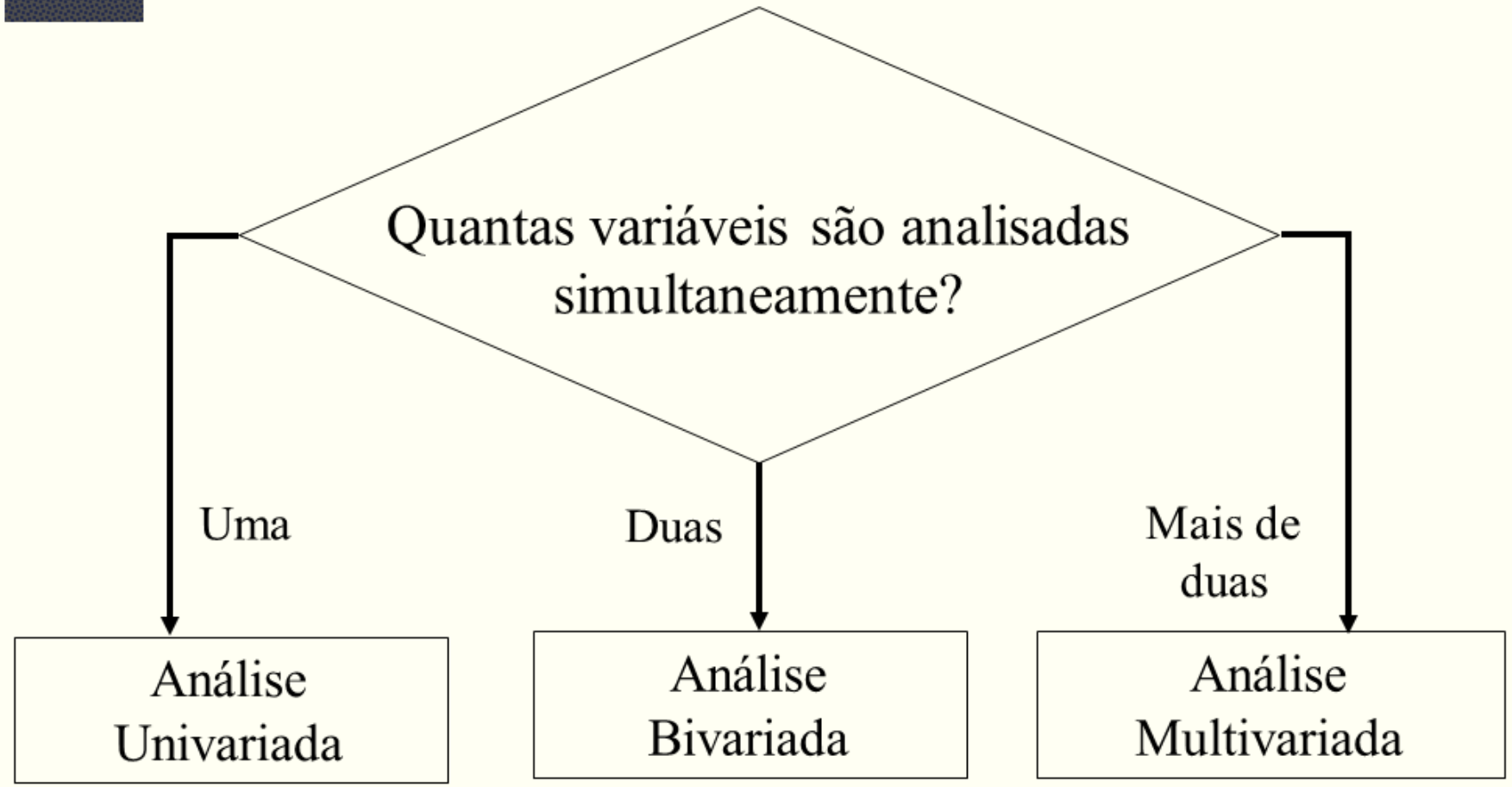
# Análise multivariada de dados

## Regressão linear múltipla

Professora Ana Amélia Benedito Silva

aamelia@usp.br

# Nº de Variáveis X Análise



# Dados Multivariados e Análise Multivariada

- Dados multivariados: quando se registra mais de uma variável aleatória referente a pessoas ou objetos, gerando um vetor de observações.
- Dados multivariados são a regra, não a exceção.

# Dados multivariados e Análise Multivariada

- Nem sempre se conhece quais variáveis são as importantes para explicar um fenômeno, daí a necessidade de não se registrar apenas uma ou duas variáveis.
- Exemplos: psicologia, educação, arqueologia, ciência ambiental, sociologia, economia, transito urbano …

# Dados multivariados e Análise Multivariada

- Frequentemente representados em forma matricial:

| Unit | Variable 1 | ... | Variable $q$ |
|---|---|---|---|
| 1 | $x_{11}$ | ... | $x_{1q}$ |
| ⋮ | ⋮ | ⋮ | ⋮ |
| $n$ | $x_{n1}$ | ... | $x_{nq}$ |

- Matriz de dados multivariados (n x q): **X**
- Dados univariados: $X_1, X_2, \ldots X_q$
- Às vezes faz sentido analisar separadamente as variáveis, mas frequentemente não.

# Principais métodos de análise multivariada

## Métodos de dependência

São métodos de inferência – utilizados quando as amostras são correlacionadas e se deseja testar alguma hipótese específica.

- ANOVA
- Regressão linear múltipla
- Regressão Logística

## Métodos de interdependência

São métodos exploratórios – utilizados para detecção de padrões, caracterizados pela ênfase em técnicas gráficas e de visualização de dados.

- Análise fatorial
- Análise de cluster

# Regressão Linear Múltipla

## técnica de dependência

- Calcula a dependência estatística de uma variável dependente quantitativa em relação a duas ou mais variáveis
- Principais objetivos
  - Encontrar relação causal entre as variáveis
  - Estimar os valores da variável dependente a partir dos valores conhecidos das variáveis independentes

# Regressão Linear Múltipla

## Objetivos

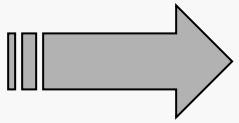
- Predizer valores de uma variável dependente **(Y)** em função de variáveis independentes **($X_1$, $X_2$, ..., $X_k$)**.

- Conhecer o quanto as variações de **($X_1$, $X_2$, ..., $X_k$)** afetam **Y**.

# Regressão Linear Múltipla

- Exemplo de aplicação na educação física

$X_1$ = exercício aeróbico
$X_2$ = calorias ingeridas
$X_3$ = circunferência da cintura



Y = perda de peso

# Regressão Múltipla

$(X_1, X_2, ..., X_k)$  **Y**

Aplicação na economia:

$X_1$ = renda
$X_2$ = taxa de juros
$X_3$ = poupança

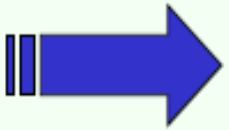
Y = consumo

# Regressão Múltipla

$(X_1, X_2, ..., X_k)$  **Y**

Aplicação no mercado mobiliário (avaliação) :

$X_1$ = área construída
$X_2$ = custo do $m^2$
$X_3$ = localização



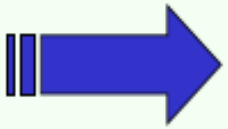
Y = preço do imóvel

# Regressão Múltipla

$(X_1, X_2, ..., X_k)$  **Y**

Aplicação na ciência da computação:

$X_1$ = memória RAM
$X_2$ = sistema operacional
$X_3$ = tipo de processador



Y = tempo de resposta

# Regressão Linear Múltipla

- Exemplo de Regressão linear múltipla no youtube

https://www.youtube.com/watch?v=TLlzToeIpGc&feature=youtu.be

# Modelo de Regressão Linear Múltipla

- Geral: $E(y) = f(X_1, X_2, ..., X_k)$
- Linear: $E(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$

onde $Y, X_1, ..., X_k$ representam as variáveis originais ou transformadas (em log (X), raiz (X), 1/(X), etc)

O coeficiente $\beta_k$ representa a variação esperada de **Y** para cada unidade de variação em $\mathbf{X_k}$ (k = 1, 2, ..., k), considerando as outras variáveis independentes fixas.

# Modelo de Regressão Linear Múltipla

**AMOSTRA:**

| | variáveis | | | | |
|---|---|---|---|---|---|
| obs. | Y | $X_1$ | $X_2$ | ... | $X_k$ |
| 1 | $y_1$ | $x_{11}$ | $x_{12}$ | ... | $x_{1k}$ |
| 2 | $y_2$ | $x_{21}$ | $x_{22}$ | ... | $x_{2k}$ |
| ... | ... | ... | ... | ... | ... |
| n | $y_k$ | $x_{n1}$ | $x_{n2}$ | ... | $x_{nk}$ |

- $E(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$
- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + e_i$

termo aleatório (erro)

# Regressão Linear Múltipla

Equação de regressão ajustada aos dados:

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

Valores preditos:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$$

Resíduos:

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

Suposições: Os erros ($e_i$) são independentes e variam aleatoriamente segundo uma distribuição (normal) com média zero e variância constante.

# Modelos lineares:

## Exemplo com regressãolinear simples

| i | $x_i$ | $y_i$ | $y_i = \beta_0 + \beta_1 x_i + e_i$ |
|---|---|---|---|
| 1 | 20 | 98 | $98 = \beta_0 + \beta_1.20 + e_1$ |
| 2 | 25 | 110 | $110 = \beta_0 + \beta_1.25 + e_2$ |
| 3 | 30 | 112 | $112 = \beta_0 + \beta_1.30 + e_3$ |
| 4 | 35 | 115 | $115 = \beta_0 + \beta_1.35 + e_4$ |
| 5 | 40 | 122 | $122 = \beta_0 + \beta_1.40 + e_5$ |

# Modelos lineares:

## Exemplo com regressão linear simples

$$\mathbf{Y} = \mathbf{X}\,\beta + \varepsilon$$

$$\begin{pmatrix} 98 \\ 110 \\ 112 \\ 115 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 20 \\ 1 & 25 \\ 1 & 30 \\ 1 & 35 \\ 1 & 40 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{pmatrix}$$

# Modelos lineares:
## Exemplo com regressão linear múltipla

| i | $x_{1i}$ | $x_{2i}$ | $y_i$ |
|---|---|---|---|
| 1 | 20 | 70 | 98 |
| 2 | 25 | 68 | 110 |
| 3 | 30 | 83 | 112 |
| 4 | 35 | 77 | 115 |
| 5 | 40 | 65 | 122 |

# Modelos lineares:
## Exemplo com regressão linear múltipla

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} 98 \\ 110 \\ 112 \\ 115 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 20 & 70 \\ 1 & 25 & 68 \\ 1 & 30 & 83 \\ 1 & 35 & 77 \\ 1 & 40 & 65 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{pmatrix}$$

# Regressão Linear Múltipla - Variáveis Dummy

https://www.youtube.com/watch?v=kIHAe0QxOSk&t=68s

https://www.youtube.com/watch?v=01cQm2gUydY

# Exemplo de regressão linear múltipla quando as variáveis independentes são qualitativas

As variáveis qualitativas devem entrar no modelo na forma de variáveis ( 0, 1)

- Experiência, $X_1$
- cargo de gerência, G (0 = não, 1 = sim)
- nível educacional, $E_1$ (1 = primeiro grau, 0 = caso contrário)
- nível educacional, $E_2$ (1 = segundo grau, 0 = caso contrário)

| | $E_1$ | $E_2$ |
|---|---|---|
| Superior | 0 | 0 |
| Primeiro grau | 1 | 0 |
| Segundo grau | 0 | 1 |

# Regressão múltipla: variáveis independentes qualitativas

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 G + \beta_3 E_1 + \beta_4 E_2$$

- O coeficiente de uma variável indicadora indica a variação esperada em Y quando a variável indicadora muda de 0 para 1, mantendo-se as demais variáveis constantes.
  - Ex: $\beta_2$ é o incremento esperado no salário pelo indivíduo ocupar um cargo de gerente.

# Exemplo de regressão linear múltipla
## quando as variáveis independentes são qualitativas

- Variável dependente: IMC
- Variáveis independentes:
  - TR (dobra cutânea triciptal)
  - SOMA_DC (soma das dobras cutâneas)
  - SEXO (0 = feminino, 1= masculino)
  - LOCAL (1, 2, 3)

# Regressão múltipla
## com variáveis independentes qualitativas

**Modelo 1**:

$$E(y) = \beta_0 + \beta_1 Sexo + \beta_2 TR + \beta_3 Soma\_dc$$

- O coeficiente de uma variável independente indica a variação esperada em Y quando a variável independente muda de 0 para 1, mantendo-se as demais variáveis constantes.
  - Ex: $\beta_1$ é o incremento esperado no IMC pelo indivíduo ser do sexo masculino.

# Regressão múltipla:
# com variáveis independentes qualitativas

**Modelo 2:** $E(y) = \beta_0 + \beta_1 Sexo + \beta_2 TR + \beta_3 Soma\_dc + \beta_4 Local_2 + \beta_5 Local_3$

- A variável Local tem 3 categorias e portanto devo criar 2 variáveis dummies:

| | $Local_2$ | $Local_3$ |
|---|---|---|
| Local =2 | 1 | 0 |
| Local = 3 | 0 | 1 |
| Local = 1 (referência) | 0 | 0 |

Variável dummy

# Regressão múltipla:
# com variáveis independentes qualitativas

**Modelo 2:**

$$E(y) = \beta_0 + \beta_1 Sexo + \beta_2 TR + \beta_3 Soma\_dc + \beta_4 Local_2 + \beta_5 Local_3$$

IMC = beta0 + beta1 1 + beta2*TR + beta3*DC + beta4*1 + beta5*0

Se um sujeito for sexo=1 e morar no local=2

- O coeficiente de uma variável independente indica a variação esperada em Y quando a variável independente muda da 1ª categoria para as outras categorias, mantendo-se as demais variáveis constantes.
  - Ex: $\beta_4$ é o incremento esperado no IMC pelo indivíduo ser do Local_2 em relação ao Local_1.
  - Ex: $\beta_5$ é o incremento esperado no IMC pelo indivíduo ser do Local_3 em relação ao Local_1.

# Regressão Linear Múltipla : Seleção de variáveis

- Forward
- Backward
- Stepwise

# MÉTODO FORWARD (passo a frente)

Considera-se inicialmente um modelo de regressão linear simples, usando como variável independente (X), aquela de maior valor da estatística t (ou menor *valor de p*) quando ajustada à variável dependente Y.

- ✓ As etapas se sucedem enquanto houver uma variável com $p \leq 0,05$;
- ✓ Se não houver mais inclusão, o processo é interrompido e as variáveis selecionadas até esta etapa definem o modelo final.

# MÉTODO FORWARD - PROCEDIMENTO

Passo 1

- ajustar todos os modelos de regressão linear simples e escolher a variável candidata com maior valor da estatística t para entrar no modelo, desde que $p \leq \alpha$ (caso $p > \alpha$ o modelo é interrompido);

Passo 2

- para cada variável não pertencente ao modelo do passo 1, escolher a variável candidata que tiver o maior valor da estatística t, desde que $p \leq \alpha$ (caso $p > \alpha$ o modelo é interrompido);

Passo 3

- repetir o processo, até que todas as variáveis que não estão no modelo apresentem um valor de t, tal que o *valor* $p > \alpha$.

# MÉTODO BACKWARD (passo atrás)

- Neste método incorporam-se inicialmente todas as variáveis em um modelo de regressão linear múltipla;
- Percorrem-se etapas, nas quais uma variável por vez pode vir a ser eliminada;
- Se em cada etapa não houver eliminação de alguma variável, o processo é interrompido e as variáveis restantes definem o modelo final.

# MÉTODO BACKWARD - PROCEDIMENTO

Passo 1

- ajustar o modelo completo de *k* variáveis;

Passo 2

- retirar do modelo completo obtido no passo 1, a variável com menor valor da estatística t (ou maior *valor de p*).
- caso todas as variáveis apresentem $p \leq \alpha$ o processo é interrompido e o modelo final é selecionado.

Passo 3

- ajustar o modelo com *k-1* variáveis e voltar ao passo 2.

# MÉTODO STEPWISE (passo a passo)

- Consiste em uma generalização do procedimento Forward;
- Após cada etapa de incorporação de uma variável, temos uma etapa em que uma das variáveis já selecionadas pode ser descartada;
- O procedimento chega ao final quando nenhuma variável é incluída ou descartada.

# MÉTODO STEPWISE - PROCEDIMENTO

Passo 1

ajustar todos os modelos com *m* variáveis (no modelo inicial *m*=1) e escolher a variável candidata com maior valor da estatística t para entrar no modelo, considerando que o *valor de p* $\leq \alpha$ (caso $p>\alpha$ o modelo é interrompido);

Passo 2

para cada variável não pertencente ao modelo do passo 1, ajustar um modelo de regressão considerando no modelo as variáveis que entraram no passo 1 e escolher a variável candidata que tiver o maior valor da estatística t, desde que $p \leq \alpha$ (caso $p>\alpha$ o modelo é interrompido);

Passo 3

verificar se o valor da estatística t das variáveis que estão no modelo apresentam $p \leq \alpha$. Caso uma ou mais variáveis que já estão no modelo apresente $p > \alpha$ , retira-se a variável do modelo que possua o maior *valor de p*.

Passo 4

ajustar o modelo no passo 3, tal que $p \leq \alpha$ para todas as variáveis. Voltar o passo 2 e repetir todo o processo até que todas as variáveis que estão fora do modelo tenham $p > \alpha$.

# Qualidade do ajuste

- Ajustou-se uma equação de regressão entre a variável dependente e as variáveis independentes.
- Qual é a qualidade deste ajuste?
  - Análise de variância do modelo
  - Análise dos resíduos

# Regressão Linear Múltipla

- Teste de hipóteses sobre o modelo: ANOVA

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

$H_1$: nem todos os $\beta_i$ são iguais à 0

# Regressão Linear Múltipla

- Teste de hipóteses sobre um particular coeficiente $\beta_j$: teste t de Student

$t = \beta_j/SE$ sendo **SE** o erro padrão da estimativa $\beta$ j

$H_0: \beta_j = 0$

$H_1: \beta_j \neq 0$

# Qualidade do ajuste

Pode-se mostrar que:

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}_{SSR}$$

$SST \longmapsto$ Soma dos quadrados totais - Variação total

$SSE \longmapsto$ Soma dos quadrados dos resíduos - Variação não explicada

$SSR \longmapsto$ Soma dos quadrados da regressão - Variação explicada

Isto é:

| Variação Total de $Y$ à volta da sua média | = | Variação que o ajustamento não consegue explicar | + | Variação explicada pelo ajustamento |
|---|---|---|---|---|

# Coeficiente de determinação

O quociente entre *SSR* e *SST* dá-nos uma medida da proporção da variação total que é explicada pelo modelo de regressão. A esta medida dá-se o nome de coeficiente de determinação ( $r^2$),

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = \frac{SST}{SST} - \frac{SSE}{SST} = 1 - \frac{SSE}{SST}$$

Note que:

- $0 \leq r^2 \leq 1$;
- $r^2 \cong 1$ (próximo de 1) significa que grande parte da variação de $Y$ é explicada linearmente pelas variáveis independentes;
- $r^2 \cong 0$ (próximo de 0) significa que grande parte da variação de $Y$ não é explicada linearmente pelas variáveis independentes.

# Coeficiente de determinação

Este coeficiente pode ser utilizado como uma medida da qualidade do ajustamento, ou como medida da confiança depositada na equação de regressão como instrumento de previsão:

- $r^2 \cong 0 \longmapsto$ modelo linear muito pouco adequado;
- $r^2 \cong 1 \longmapsto$ modelo linear bastante adequado.

# Análise dos Resíduos

- Resíduo é a diferença R = Y - $\hat{Y}$
- Para verificar a adequação do ajuste deve-se construir o gráfico dos resíduos padronizados : $\frac{R}{s_R}$
- Se os pontos estiverem distribuídos dentro do intervalo [-2;+2], é uma indicação que o modelo está bem ajustado

# Q-Q Plot

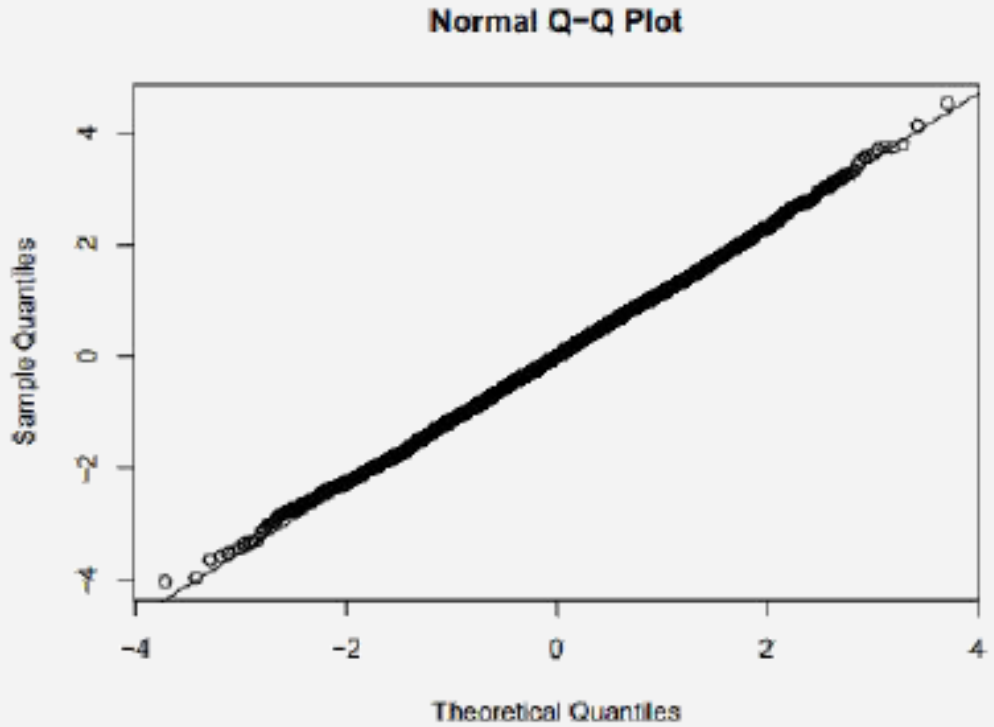
É uma ferramenta gráfica que permite avaliar se um conjunto de dados é uma distribuição normal.

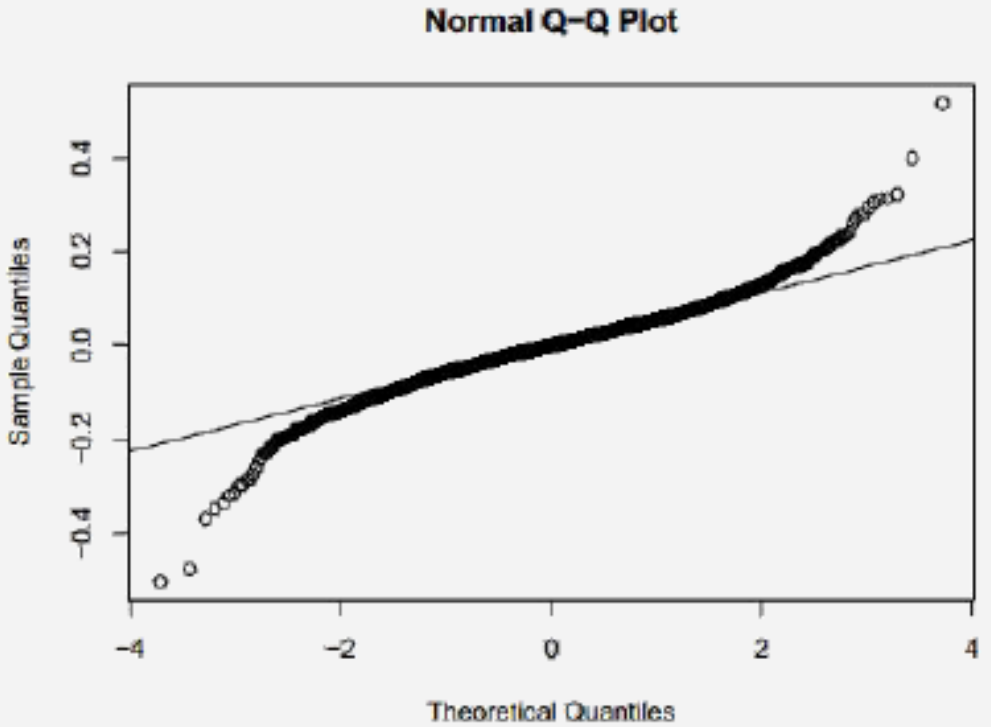
É gerado um gráfico de dispersão criado ao plotar dois conjuntos de quartis um contra ao outro.

Se os dois conjuntos de quartis são da mesma distribuição, então os pontos formarão uma linha.

Os quartis do conjunto de dados em análise é plotado contra quartis calculados de uma distribuição teórica.

# Q-Q Plot


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
−4
−2
0
2
4

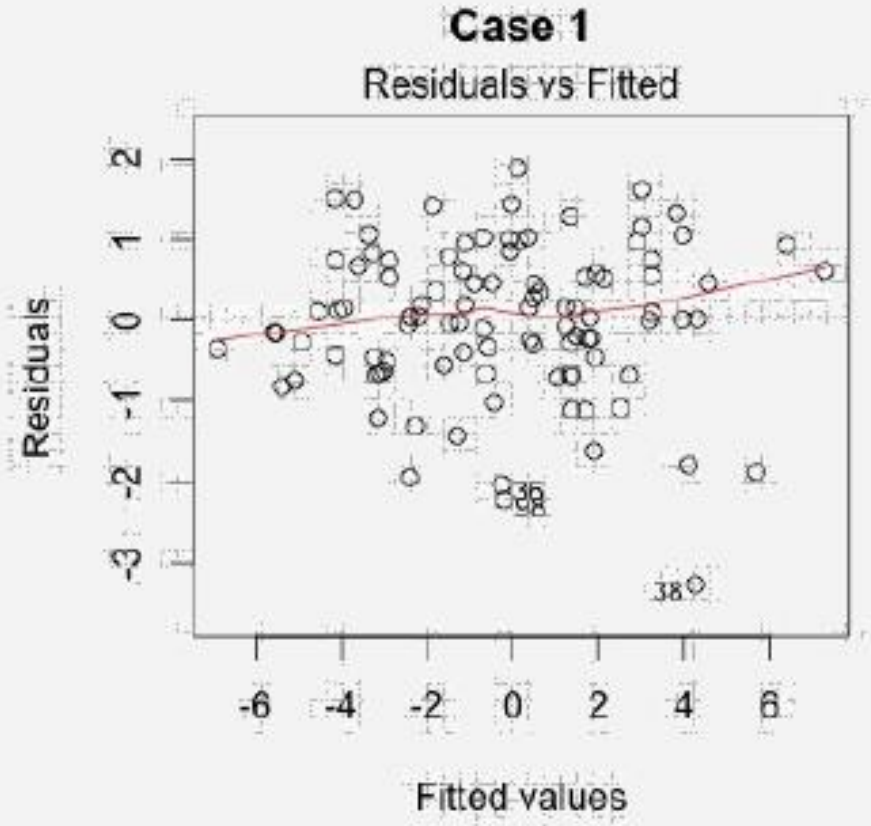

Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
−0.4
−0.2
0.0
0.2
0.4
−4
−2
0
2
4

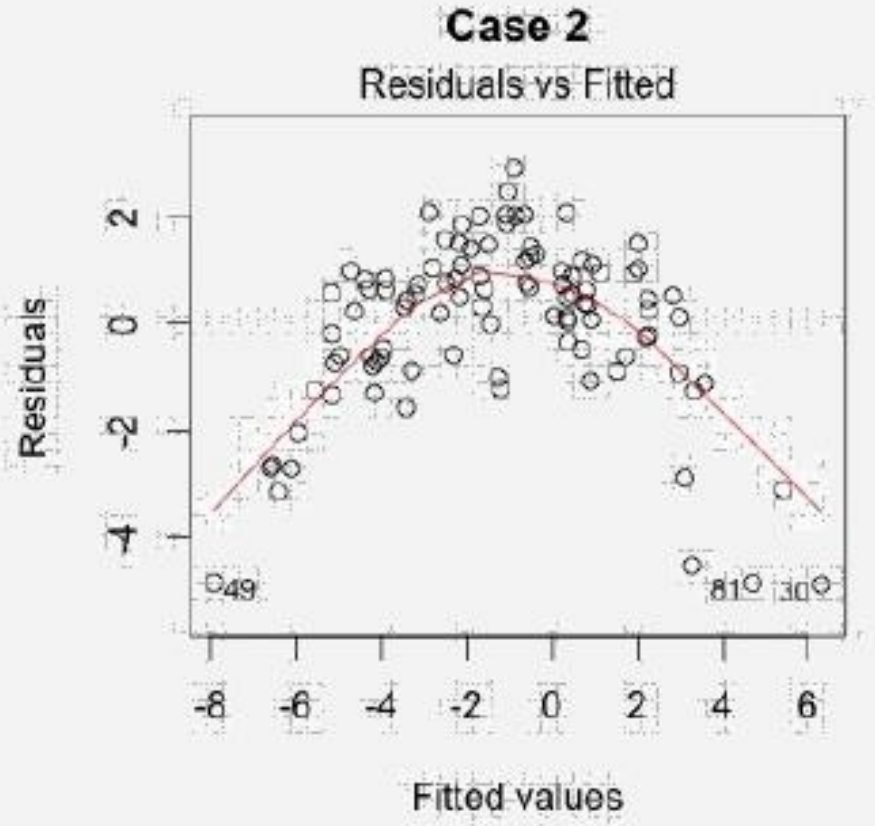
# Residuals VS Fits Plot

É um gráfico de dispersão utilizada para detectar não linearidade e homocedasticidade.

No eixo y estão os resíduos e no eixo x os valores dependentes previstos.

# Residuals VS Fits Plot - Linearidade

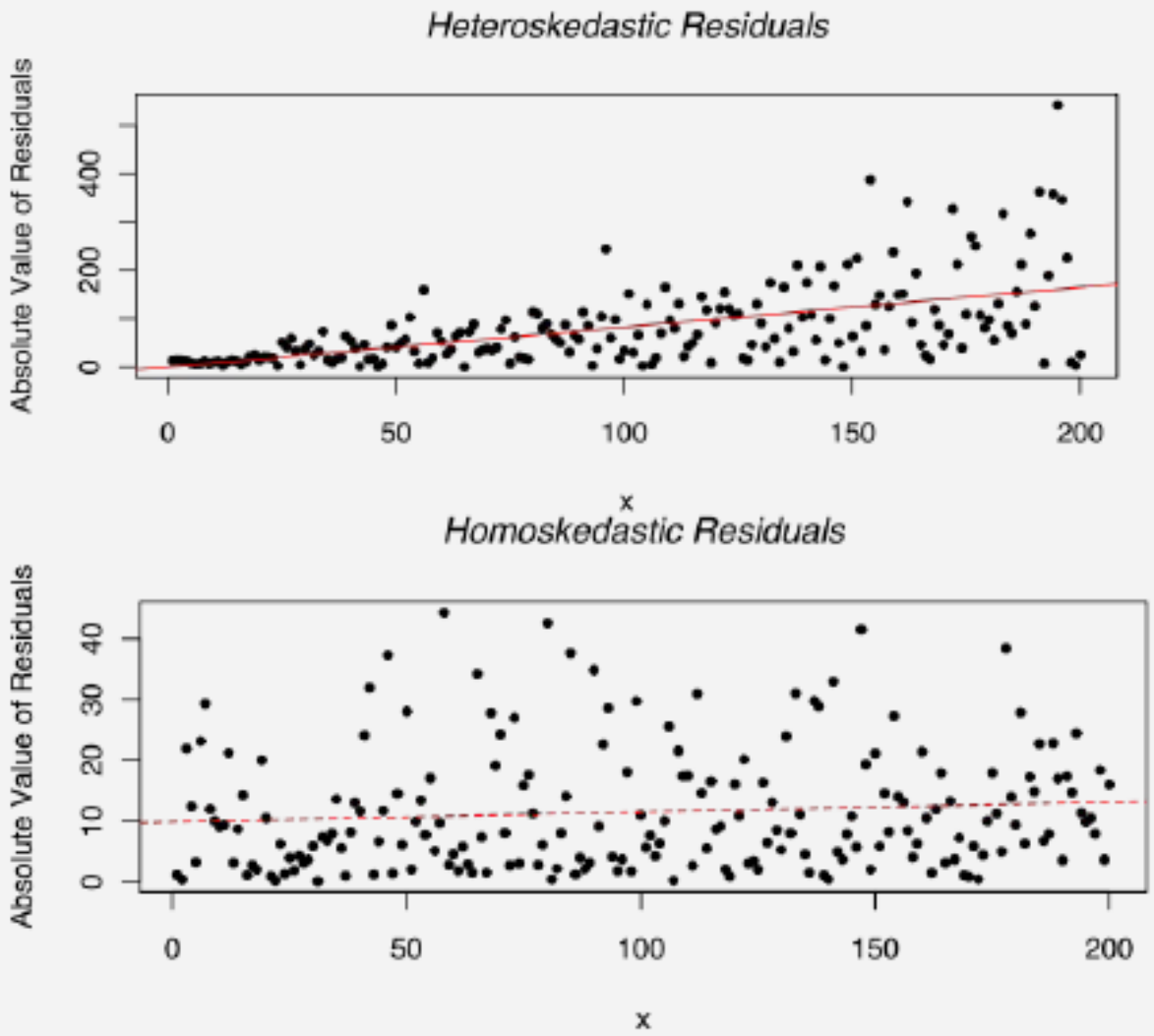


# Residuals VS Fits Plot

É um gráfico de dispersão utilizada para detectar não linearidade e homocedasticidade.

No eixo y estão os resíduos e no eixo x os valores dependentes previstos.




# Residuals VS Fits Plot

# HOMOCEDASTI-CIDADE

# Exemplo: distribuidor de cervejas

Pretende-se predizer o tempo (Y) requerido para se fazer um lote de entregas.

O Engenheiro de Produção encarregado de fazer o estudo sugere que o tempo é influenciado por 2 fatores: o número de entregas ($x_1$) e a distancia máxima ($x_2$) que o entregador precisa fazer por viagem.

# Exemplo: distribuidor de cervejas

| Entregador | $N^0$ entregas ($X_1$) | Distância ($X_2$) | Tempo (min) (Y) |
|---|---|---|---|
| 1 | 10 | 30 | 24 |
| 2 | 15 | 25 | 27 |
| 3 | 10 | 40 | 29 |
| 4 | 20 | 18 | 31 |
| 5 | 25 | 22 | 25 |
| 6 | 18 | 31 | 33 |
| 7 | 12 | 26 | 26 |
| 8 | 14 | 34 | 28 |
| 9 | 16 | 29 | 31 |
| 10 | 22 | 37 | 39 |
| 11 | 24 | 20 | 33 |
| 12 | 17 | 25 | 30 |
| 13 | 13 | 27 | 25 |
| 14 | 30 | 23 | 42 |
| 15 | 24 | 33 | 40 |

# Exemplo: distribuidor de cervejas

## Resultado da regressão

| | Coeficiente | Erro-padrão | t | p-value |
|---|---|---|---|---|
| Intercepto | 2,311 | 5,857 | 0,394 | 0,700 |
| Entregas | 0,877 | 0,153 | 5,732 | 0,000 |
| distancia | 0,455 | 0,146 | 3,106 | 0,009 |

Regression Summary for Dependent Variable: tempo
**$R^2$= 0,737** erro-padrão = 3,1408
**F(2,12)=16,795 p<,00033;**

Tempo = 2,311 + 0,877*entregas + 0,455*distância

| Entregador | $N^0$ entregas | Distância | Tempo (min) | tempo previsto | resíduo | resíduo padronizado |
|---|---|---|---|---|---|---|
| 1 | 10 | 30 | 24 | 24,7 | -0,7 | -0,24 |
| 2 | 15 | 25 | 27 | 26,8 | 0,2 | 0,07 |
| 3 | 10 | 40 | 29 | 29,3 | -0,3 | -0,10 |
| 4 | 20 | 18 | 31 | 28 | 3 | 1,03 |
| 5 | 25 | 22 | 25 | 34,2 | -9,2 | -3,17 |
| 6 | 18 | 31 | 33 | 32,2 | 0,8 | 0,28 |
| 7 | 12 | 26 | 26 | 24,7 | 1,3 | 0,45 |
| 8 | 14 | 34 | 28 | 30,1 | -2,1 | -0,72 |
| 9 | 16 | 29 | 31 | 29,5 | 1,5 | 0,52 |
| 10 | 22 | 37 | 39 | 38,4 | 0,6 | 0,21 |
| 11 | 24 | 20 | 33 | 32,5 | 0,5 | 0,17 |
| 12 | 17 | 25 | 30 | 28,6 | 1,4 | 0,48 |
| 13 | 13 | 27 | 25 | 26 | -1 | -0,34 |
| 14 | 30 | 23 | 42 | 39,1 | 2,9 | 1,00 |
| 15 | 24 | 33 | 40 | 38,4 | 1,6 | 0,55 |
| | | | | Média = | 0,03 | |
| | | | | Desvio-padrão = | 2,90 | |

Equação para o tempo previsto = 2,311 + 0,877*entregas + 0,455*distância

# Exemplo: distribuidor de cervejas
## Análise dos resíduos

# Exemplo: distribuidor de cervejas

## Resultado da regressão excluindo o *outlier*

| | Coeficiente | Erro-padrão | t | p-value |
|---|---|---|---|---|
| Intercepto | 2,915 | 2,030 | 1,43623 | 0,178762 |
| Entregas | 1,003 | 0,055 | 18,35246 | <0,00001 |
| distancia | 0,380 | 0,051 | 7,39343 | <0,0001 |

Regression Summary for Dependent Variable: tempo
**$R^2$=** 0,968 erro-padrão = 1,088
**F(2,11)=** 168,94; p< 0,00000001

$\hat{Y}$ = 2,915 + 1,003*entregas + 0,380*distância

# FIM