CIÊNCIA DE DADOS (BIG DATA)

ANÁLISE ESTATÍSTICA

Professor curador: Mário Olímpio de Menezes

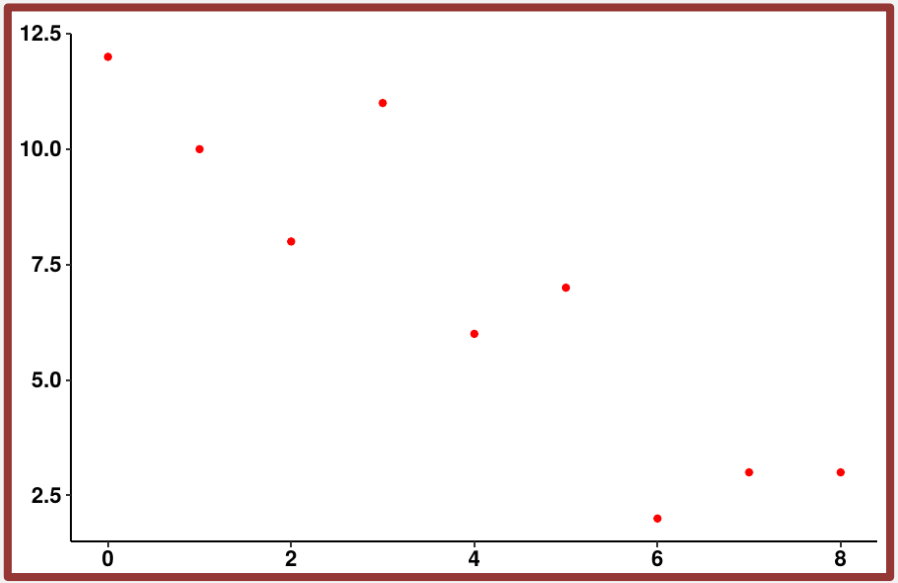
EaD Mackenzie

TRILHA 4
PARTE B – MODELAGEM ESTATÍSTICA

PARTE B – MODELAGEM ESTATÍSTICA

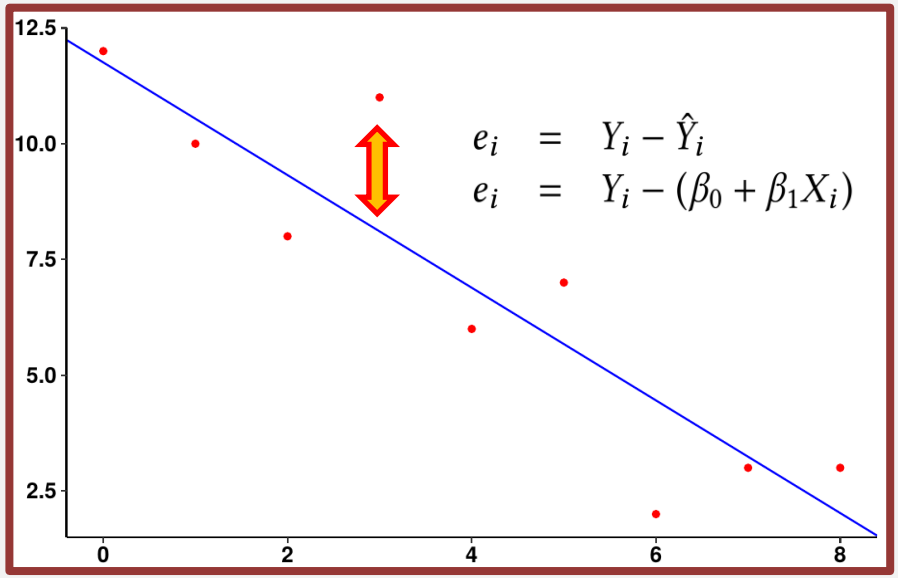
AJUSTANDO UM MODELO LINEAR

- Considere os pontos mostrados ao lado.
- Quando fazemos um ajuste linear, procuramos determinar os parâmetros que farão com que o resíduo seja mínimo.



AJUSTANDO UM MODELO LINEAR

- A figura mostra agora uma reta obtida ajustando-se um modelo linear.
- A diferença entre o ponto e a reta é o resíduo de cada ponto.



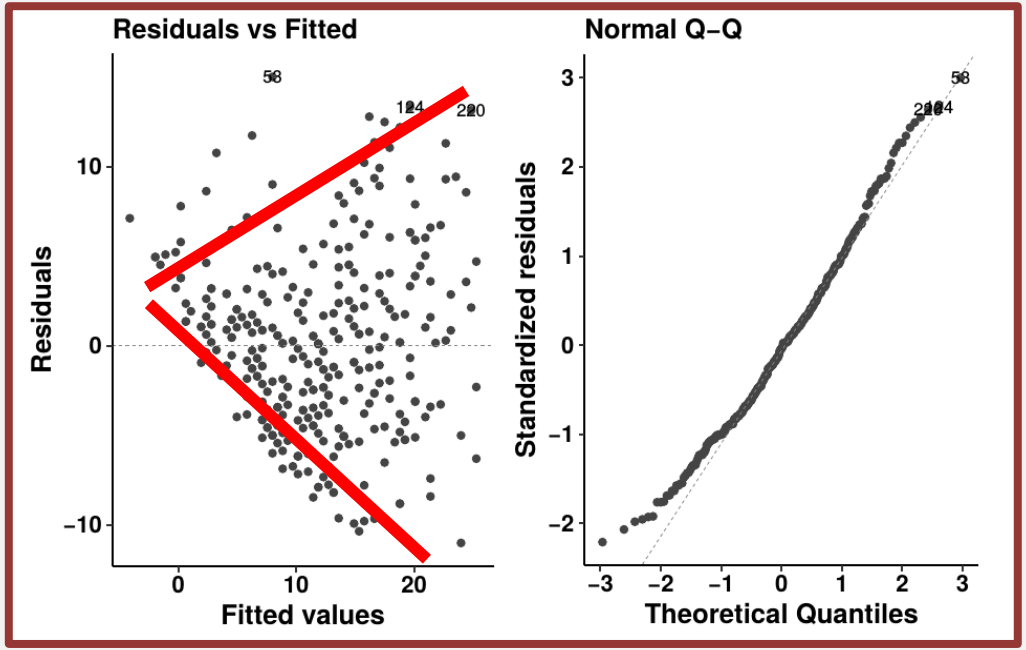
```
Call:
lm(formula = growth ~ tannin, data = reg.data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4556 -0.8889 -0.2389  0.9778  2.8944

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7556     1.0408  11.295 9.54e-06 ***
tannin       -1.2167     0.2186  -5.565 0.000846 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.693 on 7 degrees of freedom
Multiple R-squared:  0.8157,    Adjusted R-squared:  0.7893
F-statistic: 30.97 on 1 and 7 DF,  p-value: 0.0008461
```

p-value's do ajuste

DIAGNÓSTICO DO AJUSTE LINEAR

HIPÓTESES REQUERIDAS

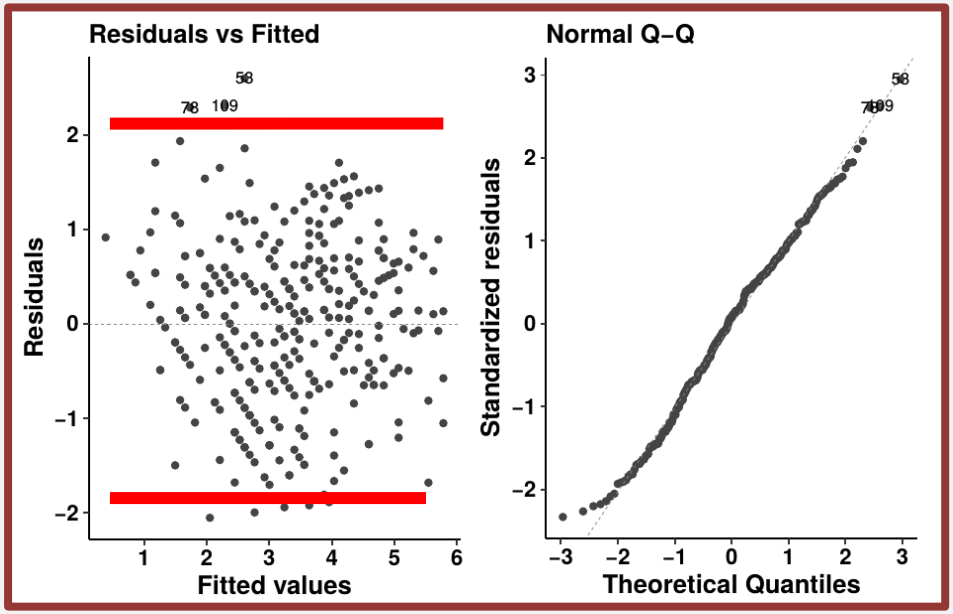
- Normalidade da variável dependente.
- Independência dos valores da variável dependente.
- Linearidade.
- Homocedasticidade – variância de Y é constante.



DIAGNÓSTICO DO AJUSTE LINEAR

HIPÓTESES REQUERIDAS

- Normalidade da variável dependente.
- Independência dos valores da variável dependente.
- Linearidade.
- Homocedasticidade – variância de Y é constante.



ANÁLISE DE VARIÂNCIA – ANOVA

- Quando as variáveis explicativas são categóricas, uma técnica de modelagem possível é a ANÁLISE DE VARIÂNCIA.
- Nesta abordagem, a variável categórica será considerada como o **fator** que provoca alguma resposta diferente na variável reposta.

ANÁLISE DE VARIÂNCIA – ANOVA

- A ideia da ANOVA é verificar se as variações induzidas na variável resposta devem ser atribuídas ao tratamento (fator) ou a algum erro experimental (normal).
- As hipóteses em teste na ANOVA são:
  - H_0 : As médias dos grupos são todas iguais.
  - H_1 : Há, pelo menos, um par de médias diferentes.

PRINCÍPIO DA ANÁLISE DE VARIÂNCIA

- Decomposição da variação total da variável resposta:
  - Parte que pode ser atribuída ao tratamento (variância **entre** as classes).
  - Parte que pode ser atribuída ao erro experimental (variância **dentro** das classes).

ANÁLISE DE VARIÂNCIA – ANOVA

VALOR	TRATAMENTO
X_1	A
X_2	A
X_3	B
X_4	B
X_5	C
X_6	C

Queremos saber se as médias dos valores das amostras de cada tratamento (**A**, **B** e **C**) são iguais ou não.

Ou seja, queremos saber se os tratamentos são realmente significativos.

ANÁLISE DE VARIÂNCIA NO **R**

```
> aovDesc <- aov(Valores ~ Amostras, data = labDescstck)
> summary(aovDesc)
             Df Sum Sq Mean Sq F value Pr(>F)
Amostras      2  12.00   6.000   1.958  0.176
Residuals    15  45.98   3.065
```

p-value > α

Não podemos rejeitar a hipótese nula de que as médias sejam iguais!

ANÁLISE DE VARIÂNCIA NO **R**

- aov(Valores ~ Fator, data = Dados)

```
> aovPrec <- aov(Valores ~ Amostras, data = labPrecstck)
> summary(aovPrec)
            Df Sum Sq Mean Sq F value   Pr(>F)
Amostras     2 12.000   6.000   65.46 3.89e-08 ***
Residuals   15  1.375   0.092
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podemos rejeitar a hipótese nula de que as médias sejam iguais!

ANÁLISE DE VARIÂNCIA NO **R**

Quais são as médias diferentes?

O **R** tem funções para verificarmos por meio da comparação múltipla entre pares:

- `pairwise.t.test`
- `TukeyHSD` – (*Tukey Honest Significant Differences*)

Hipótese nula destes testes é que as médias sejam iguais!


M
Mackenzie