Modelos de Regressão Linear Simples - Erro Puro e Falta de Ajuste

Erica Castilho Rodrigues

28 de Setembro de 2016

Erro Puro

Teste F da Falta de Ajuste

- Existem dois motivos pelos quais os pontos observados podem não cair na reta ajustada:
  - o modelo não descreve bem os dados (falta de ajuste);
  - existe uma variação aleatória em torno da reta (erro puro).
- Se grande parte do erro é devido a falta de ajuste:
  - devemos reformular o modelo.

As análises apresentadas aqui só podem ser feitas se tivermos mais de um valor da variável resposta para cada valor da explicativa.

- Essas repetições devem ser medidas em unidades amostrais diferentes.
- Não pode ser a mesma unidade medida várias vezes

- Considere que uma variável resposta Y.
- Seja X uma variável explicativa.
- Coletamos uma amostra de tamanho n.
- Dentro dessa amostra, alguns valores de X são repetidos.
- Temos m valores de X distintos, com $m < n$

$$X_1, X_2, \ldots, X_m \, .$$

- Vamos chamar de n_i o número de vezes que i-ésimo X_i aparece

$$X_1 \rightarrow n_1 \text{ observações}$$

$$X_2 \rightarrow n_2 \text{ observações}$$

onde $\sum_{i=1}^{m} n_i = m$.

- Veja um exemplo a seguir.

X_i	Y_i
2	39
2	35
3	40
4	45
4	46
4	50

- Temos que

$$n_1 =$$

- Veja um exemplo a seguir.

X_i	Y_i
2	39
2	35
3	40
4	45
4	46
4	50

- Temos que

$$n_1 = 2 \quad n_2 = 1 \quad n_3 = 3$$

Erro Puro

- Variabilidade que permanece no Y mesmo quando o valor de X é fixado.
- Variabilidade nos valores de Y entre indivíduos com o mesmo valor de X.

- Para cada valor de X_i, podemos associar uma média dos Y's

$$Y_1, Y_2, \ldots, Y_{n_1} \rightarrow n_1 \text{ observações de } X_1 \rightarrow \text{ média } \overline{Y}_1$$

$$Y_1, Y_2, \ldots, Y_{n_2} \rightarrow n_2 \text{ observações de } X_2 \rightarrow \text{ média } \overline{Y}_2$$

$$\vdots$$

$$Y_1, Y_2, \ldots, Y_{n_m} \rightarrow n_m \text{ observações de } X_m \rightarrow \text{ média } \overline{Y}_m$$

Decomposição da Soma de Quadrados dos Resíduos

- Quando ajustamos o modelo

$$Y = \beta_0 + \beta_1 X + \epsilon$$

e obtemos a reta ajustada

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

todos indivíduos com o mesmo valor $X = X_j$ tem o mesmo valor estimado $\hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j$.

- Só teremos $\hat{Y}_k \neq \hat{Y}_l$ se $X_k \neq X_l$.

- A soma dos quadrados dos erros podem ser agrupadas pelos valores repetidos de X

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n_1}(e_i's \text{ de } X_1)^2 + \sum_{i=1}^{n_2}(e_i's \text{ de } X_2)^2 + \cdots + \sum_{i=1}^{n_m}(e_i's \text{ de } X_m)^2$$

$$= \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_j)^2$$

$$= \underbrace{\sum_{j=1}^{m} n_j(\hat{Y}_j - \overline{Y}_j)^2}_{\substack{\text{soma de quadrados} \\ \text{da falta de ajuste}}} + \underbrace{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2}_{\substack{\text{soma de quadrados} \\ \text{do erro puro}}}$$

- Vejamos porque essa decomposição é veradeira.
- Temos que

$$Y_{ij} - \hat{Y}_j = (Y_{ij} - \overline{Y}_j) - (\hat{Y}_j - \overline{Y}_j)\,.$$

- Elevando ao quadrado

$$(Y_{ij} - \hat{Y}_j)^2 =$$

- Vejamos porque essa decomposição é veradeira.
- Temos que

$$Y_{ij} - \hat{Y}_j = (Y_{ij} - \overline{Y}_j) - (\hat{Y}_j - \overline{Y}_j)\,.$$

- Elevando ao quadrado

$$(Y_{ij} - \hat{Y}_j)^2 = (Y_{ij} - \overline{Y}_j)^2 - 2(\hat{Y}_j - \overline{Y}_j)(Y_{ij} - \overline{Y}_j) + (\hat{Y}_j - \overline{Y}_j)^2$$

- Somando em j e em i ficamos com

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_j)^2 =$$

- Vejamos porque essa decomposição é veradeira.
- Temos que

$$Y_{ij} - \hat{Y}_j = (Y_{ij} - \overline{Y}_j) - (\hat{Y}_j - \overline{Y}_j)\,.$$

- Elevando ao quadrado

$$(Y_{ij} - \hat{Y}_j)^2 = (Y_{ij} - \overline{Y}_j)^2 - 2(\hat{Y}_j - \overline{Y}_j)(Y_{ij} - \overline{Y}_j) + (\hat{Y}_j - \overline{Y}_j)^2$$

- Somando em j e em i ficamos com

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_j)^2 = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2$$

$$-2\sum_{j=1}^{m}\sum_{i=1}^{n_j}(\hat{Y}_j - \overline{Y}_j)(Y_{ij} - \overline{Y}_j) + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(\hat{Y}_j - \overline{Y}_j)^2$$

- Observe que o termo $(\hat{Y}_j - \overline{Y}_j)$ é constante em i logo

$$\sum_{i=1}^{n_j} (\hat{Y}_j - \overline{Y}_j)^2 = n_j (\hat{Y}_j - \overline{Y}_j)^2 \, .$$

- Então a decomposição fica

$$\sum_{j=1}^{m} \sum_{i=1}^{n_j} (Y_{ij} - \hat{Y}_j)^2 = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_j)^2$$

$$-2 \sum_{j=1}^{m} n_j (\hat{Y}_j - \overline{Y}_j) \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_j) + \sum_{j=1}^{m} n_j (\hat{Y}_j - \overline{Y}_j)^2$$

- Vamos mostrar agora que

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)n_j(\hat{Y}_j-\overline{Y}_j)=0$$

- Temos que

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)n_j(\hat{Y}_j-\overline{Y}_j)=\sum_{j=1}^{m}n_j(\hat{Y}_j-\overline{Y}_j)\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)$$

mas $\overline{Y}_j=\frac{\sum_{i=1}^{n_j}Y_{ij}}{n_j}$ e portanto

$$\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)=0$$

isso implica que

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)n_j(\hat{Y}_j-\overline{Y}_j)=0\,.$$

- Vamos usar a seguinte notação.
- SQFA $\Rightarrow$ soma de quadrados da falta de ajuste.
- SQEP $\Rightarrow$ soma de quadrados do erro puro.
- Vimos que o Coeficiente de Determinação é dado por

$$R^2 = \frac{SQR}{SQT} \quad \text{mas} \quad \max R^2 = \frac{SQT - SQEP}{SQT}$$

  ou seja, na verdade, só poderá ser 1 se $SQEP = 0$.
- Portanto o verdadeiro valor do coeficiente de determinação é

$$R^2_{real} = \frac{R^2}{\max R^2} .$$

- Vejamos quantos graus de liberade têm cada uma das componentes.
- O termo

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_j)^2$$

tem $n - 2$ graus de liberbade pois precisamos estimar $\hat{\beta}_0$ e $\hat{\beta}_1$

- Cada termo da soma

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2$$

tem $n_j - 1$ graus de liberdade pois $\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j) = 0$.

- Então o total de graus de liberdade é

$$\sum_{j=1}^{m}(n_j - 1) =$$

- Vejamos quantos graus de liberade têm cada uma das componentes.
- O termo

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_j)^2$$

tem $n - 2$ graus de liberbade pois precisamos estimar $\hat{\beta}_0$ e $\hat{\beta}_1$

- Cada termo da soma

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2$$

tem $n_j - 1$ graus de liberdade pois $\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j) = 0$.

- Então o total de graus de liberdade é

$$\sum_{j=1}^{m}(n_j - 1) = \sum_{j=1}^{m} n_j - \sum_{j=1}^{m} 1$$

- Vejamos quantos graus de liberade têm cada uma das componentes.
- O termo

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_j)^2$$

tem $n-2$ graus de liberbade pois precisamos estimar $\hat{\beta}_0$ e $\hat{\beta}_1$

- Cada termo da soma

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2$$

tem $n_j - 1$ graus de liberdade pois $\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j) = 0$.

- Então o total de graus de liberdade é

$$\sum_{j=1}^{m}(n_j - 1) = \sum_{j=1}^{m} n_j - \sum_{j=1}^{m} 1 = n - m$$

- O número de graus de liberade do termo

$$\sum_{j=1}^{m} n_j(\hat{Y}_j - \overline{Y}_j)^2$$

é dado pela subtração dos outros dois

- O número de graus de liberade do termo

$$\sum_{j=1}^{m} n_j(\hat{Y}_j - \overline{Y}_j)^2$$

é dado pela subtração dos outros dois

$$(n-2) - (n-m) =$$

- O número de graus de liberade do termo

$$\sum_{j=1}^{m} n_j (\hat{Y}_j - \overline{Y}_j)^2$$

é dado pela subtração dos outros dois

$$(n-2) - (n-m) = m-2\,.$$

- Então os graus de liberdade de cada uma das parcelas ficam

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_j)^2 = \sum_{j=1}^{m} n_j (\hat{Y}_j - \overline{Y}_j)^2 + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2$$

$$(n-1) = (m-2) + (n-m)$$

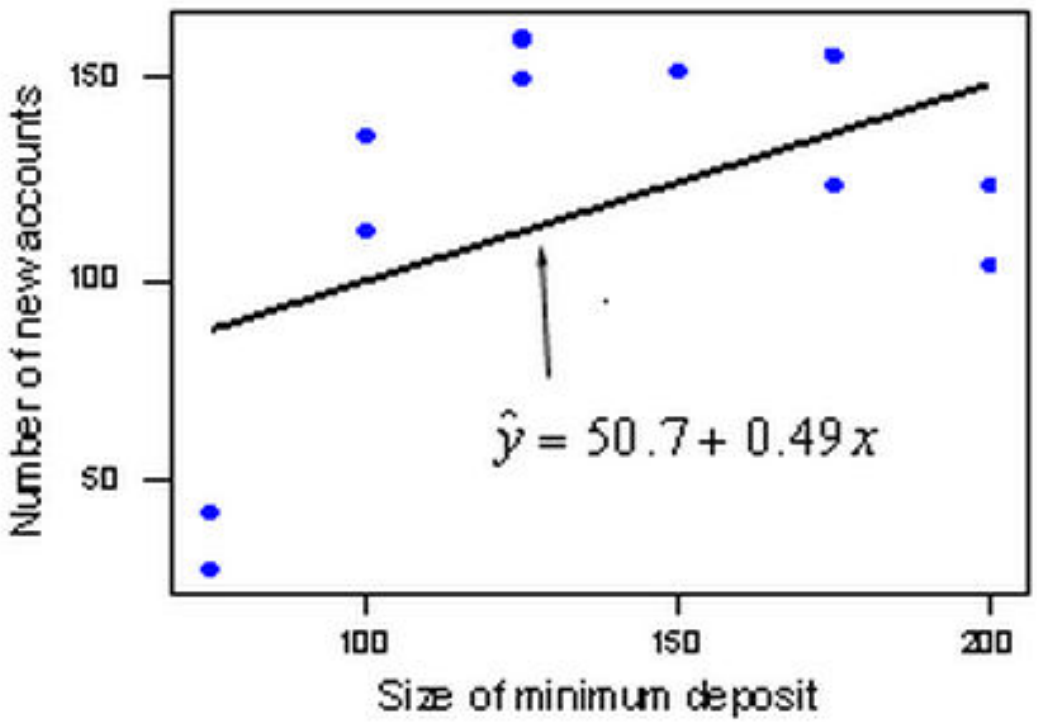
- A Tabela ANOVA fica da seguinte maneira

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Estatística F
Regressão	1	SQR	$QMR = SQR/1$	$\frac{QMR}{S^2}$
Residual	$n-2$	SQE	$QME = \frac{SQE}{(n-2)}$	
(Falta de Ajuste)	(m-2)	$(SQFA)$	$QMFA = \frac{SQFA}{m-2}$	$\frac{QMFA}{S_e^2}$
(Erro Puro)	(n-m)	$(SQEP)$	$S_e^2 = \frac{SQEP}{n-m}$	
Total	$n-1$	SQT		

Tabela: Tabela ANOVA

Exemplo:

- Vamos considerar duas variáveis.
- A figura abaixo apresenta o gráfico de dipersão e a reta ajustada.



Exemplo: (continuação)

- Se o modelo está bem ajustado:
  - a média de Y para um valor fixo de X deve ficar próxima do valor predito.
- Essa distância é dada pela soma da falta de ajuste:

$$\underbrace{\sum_{j=1}^{m} n_j(\hat{Y}_j - \overline{Y}_j)^2}_{\text{soma de quadrados da falta de ajuste}}$$

- Para esses dados, a soma é igual a

$$\sum_{j=1}^{m} n_j(\hat{Y}_j - \overline{Y}_j)^2 = 13594$$

Exemplo: (continuação)

- O restante da variação de Y é causada por erro aleatório

$$\underbrace{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)^2}_{\text{soma de quadrados do erro puro}}$$

- Para esses dados, a soma é igual a

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)^2 = 1148$$

Exemplo: (continuação)

- A decomposição da soma de quadrados fica

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\hat{Y}_j)^2 = \underbrace{\sum_{j=1}^{m} n_j(\hat{Y}_j-\overline{Y}_j)^2}_{\substack{\text{soma de quadrados}\\ \text{da falta de ajuste}}} + \underbrace{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)^2}_{\substack{\text{soma de quadrados}\\ \text{do erro puro}}}$$

$$14742 = 13594 + 1148$$

- Qual a conclusão?

Exemplo: (continuação)

- A decomposição da soma de quadrados fica

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\hat{Y}_j)^2 = \underbrace{\sum_{j=1}^{m} n_j(\hat{Y}_j-\overline{Y}_j)^2}_{\text{soma de quadrados da falta de ajuste}} + \underbrace{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)^2}_{\text{soma de quadrados do erro puro}}$$

$$14742 = 13594 + 1148$$

- Qual a conclusão?
- A maior parte da variabilidade é devido a falta de ajuste.

Exemplo: (continuação)

- A decomposição da soma de quadrados fica

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\hat{Y}_j)^2 = \underbrace{\sum_{j=1}^{m} n_j(\hat{Y}_j-\overline{Y}_j)^2}_{\text{soma de quadrados da falta de ajuste}} + \underbrace{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)^2}_{\text{soma de quadrados do erro puro}}$$

$$14742 = 13594 + 1148$$

- Qual a conclusão?
- A maior parte da variabilidade é devido a falta de ajuste.
- O modelo não está bem ajustado.

Exemplo: (continuação)

- A decomposição da soma de quadrados fica

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\hat{Y}_j)^2 = \underbrace{\sum_{j=1}^{m} n_j(\hat{Y}_j-\overline{Y}_j)^2}_{\substack{\text{soma de quadrados}\\ \text{da falta de ajuste}}} + \underbrace{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)^2}_{\substack{\text{soma de quadrados}\\ \text{do erro puro}}}$$

$$14742 = 13594 + 1148$$

- Qual a conclusão?
- A maior parte da variabilidade é devido a falta de ajuste.
- O modelo não está bem ajustado.
- Obsevamos isso pelo gráfico, a relação parece não ser linear.

Teste F da Falta de Ajuste

- Vejamos como usar essas informaçõe para testar se o modelo está bem ajustado.
- Queremos testar as seguintes hipóteses:

H_0 : o modelo linear é adequado (não há falta de ajuste)

H_1 : o modelo linear não é adequado (há falta de ajuste)

- A estatística de teste é dada por

$$F = \frac{QMFA}{S_e^2}$$

que sob H_0 tem distribuição $F_{m-2,n-m}$.

- Devemos rejeitar H_0 para valores altos ou baixos de F?

- Devemos rejeitar H_0 para valores altos ou baixos de F?
- Altos.
- Se F é grande, $QMFA$ é grande, há falta de ajuste.

Exemplo:

- Considere os dados apresentados na tabela a seguir.

Time Order	Y	X	Time Order	Y	X	Time Order	Y	X
12	2.3	1.3	19	1.7	3.7	3	3.5	5.3
23	1.8	1.3	20	2.8	4.0	6	2.8	5.3
7	2.8	2.0	5	2.8	4.0	10	2.1	5.3
8	1.5	2.0	2	2.2	4.0	4	3.4	5.7
17	2.2	2.7	21	3.2	4.7	9	3.2	6.0
22	3.8	3.3	15	1.9	4.7	13	3.0	6.0
1	1.8	3.3	18	1.8	5.0	14	3.0	6.3
11	3.7	3.7				16	5.9	6.7

- O modelo ajustado é dado por

$$Y_i = 1,426 + 0,316X_i + \epsilon_i$$

onde $\epsilon_i \sim^{iid} N(0, \sigma^2)$.

Exemplo: (continuação)

- A tabela ANOVA é apresentada a seguir.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Estatística F
Regressão	1	$SQR = 5.499$	$QMR = 5499$	$\frac{QMR}{S^2} = 7,56$
Residual	21	$SQE = 15.287$	$QME = 0.728$	
Total	22	$SQT = 2.0777$		

Tabela: Tabela ANOVA

- O valor crítico da Tabela F com $\alpha = 0,05$ é $F_{1,21} = 4,325$

Exemplo: (continuação)

- A tabela ANOVA é apresentada a seguir.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Estatística F
Regressão	1	$SQR = 5.499$	$QMR = 5499$	$\frac{QMR}{S^2} = 7,56$
Residual	21	$SQE = 15.287$	$QME = 0.728$	
Total	22	$SQT = 2.0777$		

Tabela: Tabela ANOVA

- O valor crítico da Tabela F com $\alpha = 0,05$ é $F_{1,21} = 4,325$ (lembre que esse teste é unilateral!)
- Conlusão:

Exemplo: (continuação)

- A tabela ANOVA é apresentada a seguir.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Estatística F
Regressão	1	$SQR = 5.499$	$QMR = 5499$	$\frac{QMR}{S^2} = 7,56$
Residual	21	$SQE = 15.287$	$QME = 0.728$	
Total	22	$SQT = 2.0777$		

Tabela: Tabela ANOVA

- O valor crítico da Tabela F com $\alpha = 0,05$ é $F_{1,21} = 4,325$ (lembre que esse teste é unilateral!)
- Conlusão: Rejeitamos a hipótese de $\beta_1 = 0$.

Exemplo: (continuação)

- Vamos agora encontrar o valor do erro puro e falta de ajuste.
- Por exemplo para $X = 1.3$ temos que

$$\overline{Y}_1 =$$

Exemplo: (continuação)

- Vamos agora encontrar o valor do erro puro e falta de ajuste.
- Por exemplo para $X = 1.3$ temos que

$$\overline{Y}_1 = \frac{(2.3 + 1.8)^2}{2} = 2.05\,.$$

- Logo

$$\sum_{i=1}^{2} (Y_{i1} - \overline{Y}_1) =$$

Exemplo: (continuação)

- Vamos agora encontrar o valor do erro puro e falta de ajuste.
- Por exemplo para $X = 1.3$ temos que

$$\overline{Y}_1 = \frac{(2.3 + 1.8)^2}{2} = 2.05\,.$$

- Logo

$$\sum_{i=1}^{2}(Y_{i1} - \overline{Y}_1) = (2.3 - 2.05)^2 + (1.8 - 2.05)^2 = 0,125$$

Exemplo: (continuação)

- Repetindo essa conta para todos valores distintos de X obtemos os seguintes resultados:

Level of X	$\Sigma_{u=1}^{n}(Y_{ju} - \overline{Y}_j)^2$	df
1.3	0.125	1
2.0	0.845	1
3.3	2.000	1
3.7	2.000	1
4.0	0.240	2
4.7	0.845	1
5.3	0.980	2
6.0	0.020	1
Totals	7.055	10

Exemplo: (continuação)

- A Tabela ANOVA fica na forma:

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Estatística F
Regressão	1	5.499	$QMR = 5499$	$\frac{QMR}{S^2} = 7,56$
Residual	21	15.287	$QME = 0.728$	
(Falta de Ajuste)	11	8.233	$QMFA = 0.748$	$\frac{QMFA}{S_e^2} = 1.061$
(Erro Puro)	10	7.055	$S_e^2 = 0.706$	
Total	22	20.777		

Tabela: Tabela ANOVA com cálculo da Falta de Ajuste.

Exemplo: (continuação)

- Vamos testar as hipóteses:

  H_0 : o modelo linear é adequado (não há falta de ajuste)

  H_1 : o modelo linear não é adequado (há falta de ajuste)

- O valor observado para estatística de teste é

$$F = \frac{QMFA}{S_e^2} =$$

Exemplo: (continuação)

- Vamos testar as hipóteses:

  H_0 : o modelo linear é adequado (não há falta de ajuste)

  H_1 : o modelo linear não é adequado (há falta de ajuste)

- O valor observado para estatística de teste é

$$F = \frac{QMFA}{S_e^2} = 1.061$$

sob H_0, $F \sim$

Exemplo: (continuação)

- Vamos testar as hipóteses:

  H_0 : o modelo linear é adequado (não há falta de ajuste)

  H_1 : o modelo linear não é adequado (há falta de ajuste)

- O valor observado para estatística de teste é

$$F = \frac{QMFA}{S_e^2} = 1.061$$

  sob H_0, $F \sim F_{11,10}$.

- Usando $\alpha = 5\%$ da tabela temos que $F_{11,10} = 2,854$.
- A região crítica é dada por

Exemplo: (continuação)

- Vamos testar as hipóteses:

  H_0 : o modelo linear é adequado (não há falta de ajuste)

  H_1 : o modelo linear não é adequado (há falta de ajuste)

- O valor observado para estatística de teste é

$$F = \frac{QMFA}{S_e^2} = 1.061$$

  sob H_0, $F \sim F_{11,10}$.

- Usando $\alpha = 5\%$ da tabela temos que $F_{11,10} = 2,854$.
- A região crítica é dada por

$$F > 2,854\,.$$

Exemplo: (continuação)

- Rejeitamos ou não H_0?

Exemplo: (continuação)

- Rejeitamos ou não H_0? Como 1.061< 2.854, não rejeitamos H_0.
- Conclusão: com 5% de significância temos evidência de que o modelo linear é adequado nesse caso, ou seja, **não há falta de ajuste**.
- Vamos agora calcular o Coeficiente de Determinação Real.
- Temos que

$$R^2 = \frac{SQR}{SQT} = \frac{5.499}{20.77} = 0,2674$$

$$\max R^2 = \frac{SQT - SQEP}{SQT} = \frac{20.777 - 7.055}{20777} = 0.6604$$

Exemplo: (continuação)

- O Coeficiente de Determinação Real é dado por:

$$R^2_{real} = \frac{R^2}{\max R^2} =$$

Exemplo: (continuação)

- O Coeficiente de Determinação Real é dado por:

$$R^2_{real} = \frac{R^2}{\max R^2} = \frac{0,2674}{0.6604} = 0,4049$$

- Conclusão:

Exemplo: (continuação)

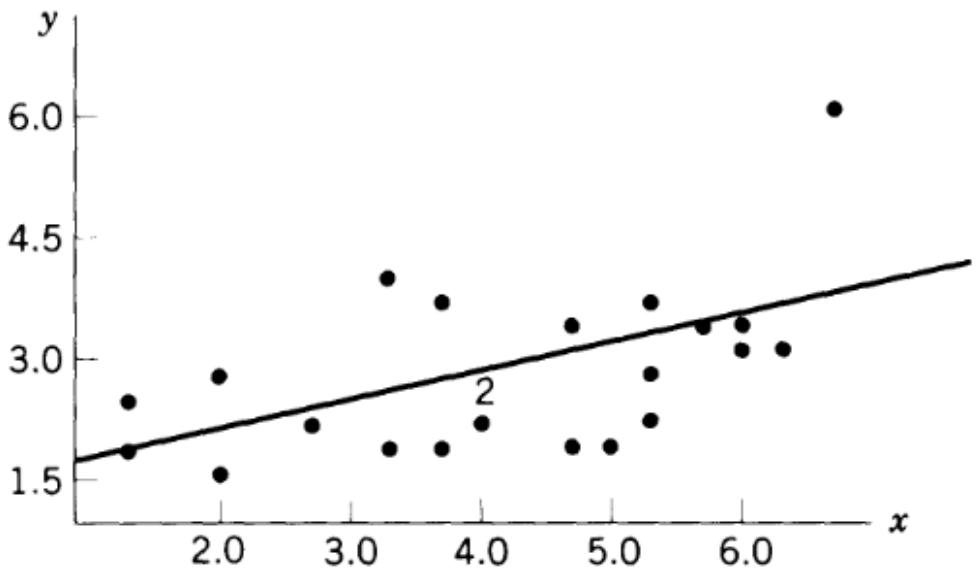
- O Coeficiente de Determinação Real é dado por:

$$R^2_{real} = \frac{R^2}{\max R^2} = \frac{0,2674}{0.6604} = 0,4049$$

- Conclusão:40,49% da variabilidade total dos dados pode ser explicada pelo modelo de regressão.
- Esse valor da uma idéia melhor do que foi alcançado pelo modelo dentro do que era possível.

Exemplo: (continuação)

- A figura a seguir mostra os dados coletados e a reta ajustada.



- Observe que:

variação em torno da reta $\approx$ variação do Y para cada valor fixo de X

Exemplo: (continuação)

- Isso foi comprovado pelo teste de falta de ajuste.
- A variabilidade em torno da reta reflete a variabilidade intrínseca aos dados.

Exemplo:

- Foram analisados dados de 15 árvores.
- As variáveis observadas foram:
  - altura e diâmetro da árvore.
- Vamos considerar

$$Y = \{\text{altura da árvore}\}$$

$$X = \{\text{diâmetro da árvore}\}$$

- Foram considerados 5 diâmetros distintos.
- Paraca cada valor de diâmetro foram registradas as alturas de 3 árvores.
- Qual valor de m?

Exemplo:

- Foram analisados dados de 15 árvores.
- As variáveis observadas foram:
  - altura e diâmetro da árvore.
- Vamos considerar

$$Y = \{\text{altura da árvore}\}$$

$$X = \{\text{diâmetro da árvore}\}$$

- Foram considerados 5 diâmetros distintos.
- Paraca cada valor de diâmetro foram registradas as alturas de 3 árvores.
- Qual valor de m?5
- Qual valore de n_1, n_2, n_3?

Exemplo:

- Foram analisados dados de 15 árvores.
- As variáveis observadas foram:
  - altura e diâmetro da árvore.
- Vamos considerar

$$Y = \{\text{altura da árvore}\}$$

$$X = \{\text{diâmetro da árvore}\}$$

- Foram considerados 5 diâmetros distintos.
- Paraca cada valor de diâmetro foram registradas as alturas de 3 árvores.
- Qual valor de m?5
- Qual valore de n_1, n_2, n_3? 3
- Qual valor de n?

Exemplo:

- Foram analisados dados de 15 árvores.
- As variáveis observadas foram:
  - altura e diâmetro da árvore.
- Vamos considerar

$$Y = \{\text{altura da árvore}\}$$

$$X = \{\text{diâmetro da árvore}\}$$

- Foram considerados 5 diâmetros distintos.
- Paraca cada valor de diâmetro foram registradas as alturas de 3 árvores.
- Qual valor de m?5
- Qual valore de n_1, n_2, n_3? 3
- Qual valor de n? 15

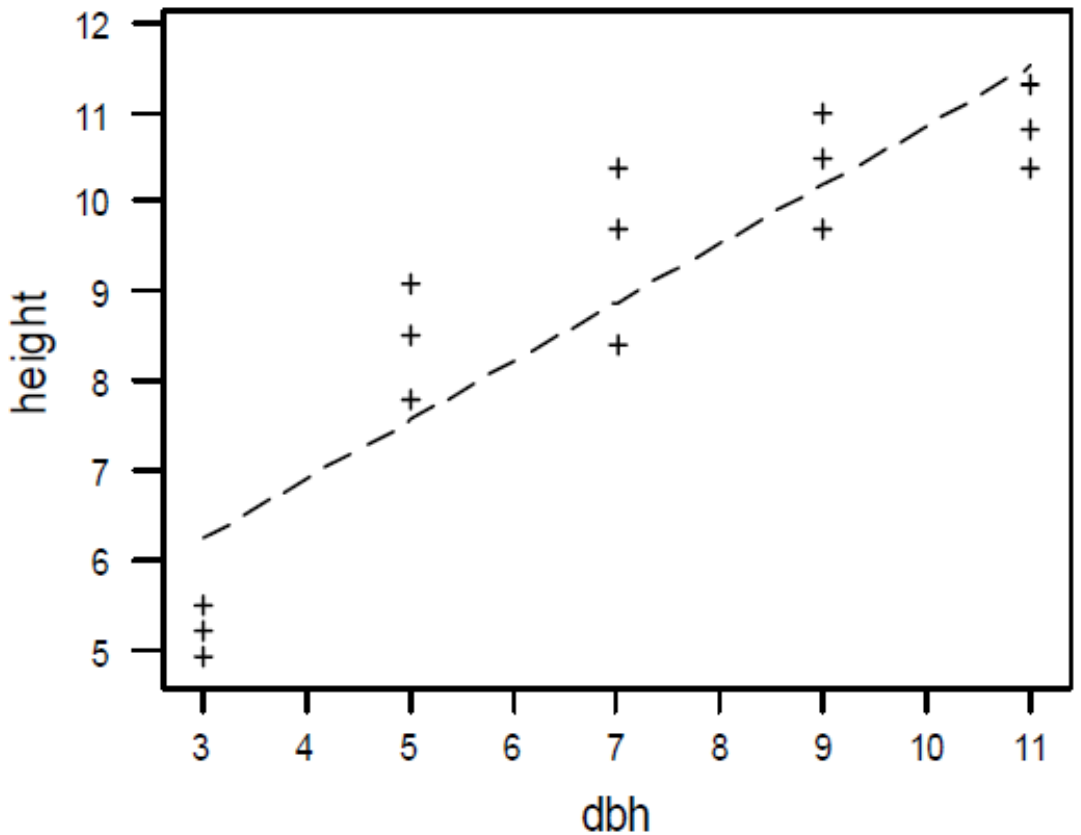
Exemplo: (continuação)

- A figura a seguir apresenta os dados coletados.

X_j	Y_{ij}	$\overline{Y}_j$	e_{ij}	$\hat{Y}_j$
3	4.9	5.200	-1.3400	6.24
3	5.5	5.200	-0.7400	6.24
3	5.2	5.200	-1.0400	6.24
5	7.8	8.467	0.2400	7.56
5	9.1	8.467	1.5400	7.56
5	8.5	8.467	0.9400	7.56
7	8.4	9.500	-0.4800	8.88
7	9.7	9.500	0.8200	8.88
7	10.4	9.500	1.5200	8.88
9	9.7	10.400	-0.5000	10.20
9	10.5	10.400	0.3000	10.20
9	11.0	10.400	0.8000	10.20
11	10.4	10.833	-1.1200	11.52
11	10.8	10.833	-0.7200	11.52
11	11.3	10.833	-0.2200	11.52

Exemplo: (continuação)

- A figura a seguir mostra o gráfico de dispersão dos dados.



Exemplo: (continuação)

- O modelo ajustado foi o seguinte

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- A Tabela ANOVA é apresentada a seguir

```
Analysis of Variance Table

Response: Y
             Df Sum Sq Mean Sq  F value    Pr(>F)
X             1 52.272  52.272 120.0735 6.832e-07 ***
Residuals    13 12.752   0.981
 Lack of fit  3  8.399   2.800   6.4308   0.01061 *
 Pure Error  10  4.353   0.435
```

Exemplo: (continuação)

- Quais conclusões podem ser retiradas a partir dessa tabela?

Exemplo: (continuação)

- Quais conclusões podem ser retiradas a partir dessa tabela?
- Para testar as hipóteses

$$H_0 : \beta_1 = 0 \quad H_1 : \beta \not= 0$$

Exemplo: (continuação)

- Quais conclusões podem ser retiradas a partir dessa tabela?
- Para testar as hipóteses

$$H_0 : \beta_1 = 0 \quad H_1 : \beta \neq 0$$

devemos rejeitar H_0.

- Conclusão: com 5% de significância há evidências de que o diâmetro da árvore é significativo para explicar sua altura.

Exemplo: (continuação)

- Vamos testar agora falta de ajuste.
- As hipóteses a serem testadas são

$$H_0: \text{ o modelo não possui falta de ajuste}$$

$$H_1: \text{ o modelo não possui falta de ajuste}$$

- Rejeitamos ou não H_0?

Exemplo: (continuação)

- Vamos testar agora falta de ajuste.
- As hipóteses a serem testadas são

$$H_0: \text{ o modelo não possui falta de ajuste}$$

$$H_1: \text{ o modelo não possui falta de ajuste}$$

- Rejeitamos ou não H_0? Rejeitamos.
- Conclusão:

Exemplo: (continuação)

- Vamos testar agora falta de ajuste.
- As hipóteses a serem testadas são

$$H_0 : \text{ o modelo não possui falta de ajuste}$$

$$H_1 : \text{ o modelo não possui falta de ajuste}$$

- Rejeitamos ou não H_0? Rejeitamos.
- Conclusão: O modelo linear não parece ser adequado nesse caso.