Estatística para Cursos de Engenharia e Informática

Pedro Alberto Barbetta / Marcelo Menezes Reis / Antonio Cezar Bornia

São Paulo: Atlas, 2004

Cap. 11 – Complemento: Regressão Múltipla

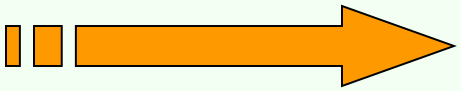
APOIO:
Fundação de Apoio à Pesquisa Científica e Tecnológica do Estado de Santa Catarina (FAPESC)
Departamento de Informática e Estatística – UFSC (INE/CTC/UFSC)

Regressão Múltipla

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

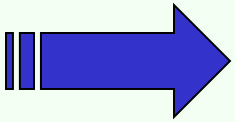
- Conhecer o quanto variações de **X_j** (j = 1,...,k) podem afetar **Y**.

Regressão Múltipla

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

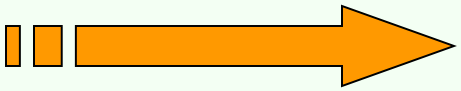
Aplicação na economia:

X_1 = renda
X_2 = taxa de juros
X_3 = poupança



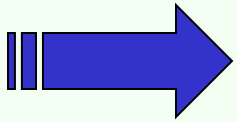
Y = consumo

Regressão Múltipla

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

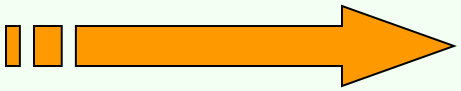
Aplicação no mercado mobiliário (avaliação) :

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



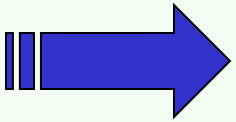
Y = preço do imóvel

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

Modelo de Regressão Múltipla

- $E\{Y\} = f(X_1, X_2, ..., X_k)$

- Linear: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$

  – onde Y, X_1, ..., X_k podem representar as variáveis originais ou transformadas.

  – Admite-se que X_1, ..., X_k são variáveis matemáticas e Y é uma variável aleatória.

Modelo de Regressão Múltipla

- $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$

  - O coeficiente $\boldsymbol{\beta_k}$ 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.
  - O primeiro objetivo é estimar os coeficientes: β_0, β_1, β_2, ..., β_k.

Modelo de Regressão 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}

- $E\{y_i\} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$
- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + e_i$

termo aleatório (erro)

Modelo de Regressão Múltipla
Suposições

- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + e_i$

termo aleatório (erro)

- 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

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

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

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

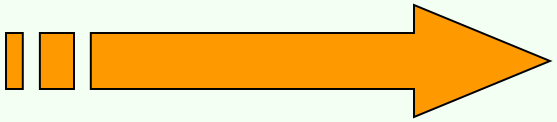
$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

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

Modelos lineares:
Estimador de mínimos quadrados

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

Estimador de mínimos quadrados de β, isto é, o vetor ***b*** que minimiza a função **$L(\beta) = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$** :



$$b = (X'X)^{-1}(X'Y)$$

$$\boldsymbol{b} = (b_0, b_1, \ldots, b_k)'$$

Regressão 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:

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

(estimativa da) variância do erro:

$$s_e^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \frac{SQE}{n-k-1}$$

Medida do Ajuste

Coeficiente de determinação (R^2)

$$R^2 = \frac{\text{Variação explicada}}{\text{Variação total}} = \frac{\Sigma (\hat{y}_i - \bar{y})^2}{\Sigma (y_i - \bar{y})^2}$$

$$0 \leq R^2 \leq 1$$

Coef. de correlação múltiplo (R): coef. de correlação entre y_i e $\hat{y}_i$

Regressão Múltipla: *teste sobre o modelo*

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

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

$$SQE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \qquad SQT = \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2$$

$$SQR = SQT - SQE$$

$$f = \frac{SQR / k}{SQE / n-k-1}$$

Sob H_0 e considerando as suposições do modelo, ***f*** tem distrib. **F** com g.l. ***k*** (no num.) e **(*n*-*k*-1)** (no denom.)

Regressão Múltipla:
teste sobre um particular coeficiente

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \ldots + \beta_k X_k$$

$$H_0: \beta_j = 0$$

$$t = \frac{b_j}{s_e \sqrt{c_{jj}}}$$

onde c_{jj} é o k-ésimo elemento da diag. princ. da matriz $C = (X'X)^{-1}$.

Sob H_0 e considerando as suposições do modelo,

t tem distrib. **t de student** com g.l. = ***n-k-1***

Ex. de regressão múltipla:

O sistema de entrega de um distribuidor de cervejas

Pretende-se predizer o tempo **(y)** requerido para se fazer um lote de entregas. O Eng. de produção encarregado de fazer o estudo sugere que o tempo é influenciado fundamentalmente por dois fatores: o número de entregas **(x_1)** e a distância máxima **(x_2)** que o entregador precisa fazer por viagem.

Sistema de entrega do distribuidor de cerveja

ENTREGAS	DISTANC (X1)	TEMPO (X2)	(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

Ex. de regressão múltipla:

O sistema de entrega de um distribuidor de cervejas

Resumo da regressão

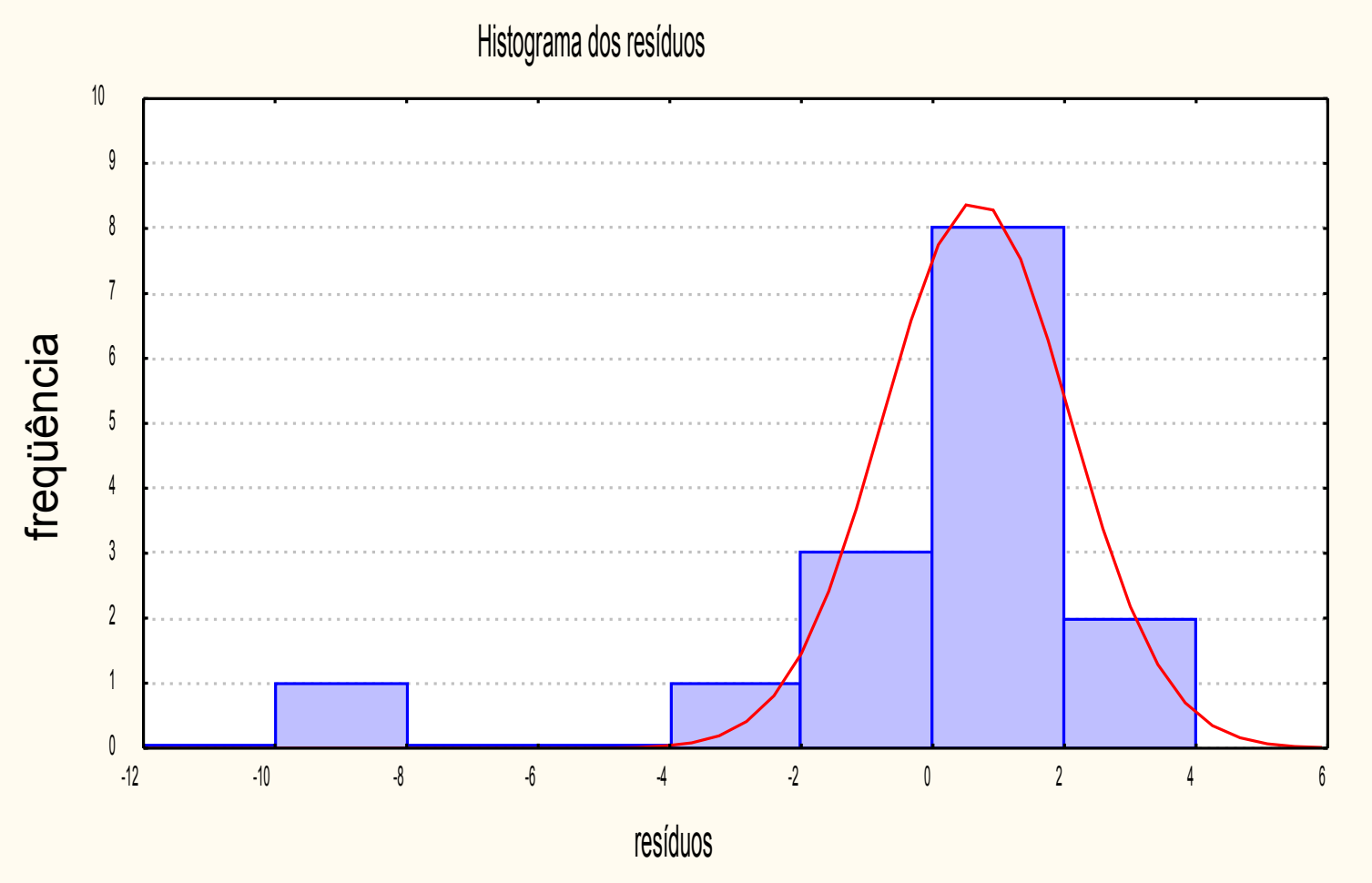
$R^2 = 0{,}736$ $s_e^2 = 3{,}1408$

$F(2,12) = 16{,}795$ $p < 0{,}00033$

	coef.	erro padrão dos coef.	$t_{(12)}$	p
Intercepto	2,311	5,857	0,394	0,700
ENTREGAS	0,877	0,153	5,732	0,000
DISTANC	0,455	0,146	3,106	0,009

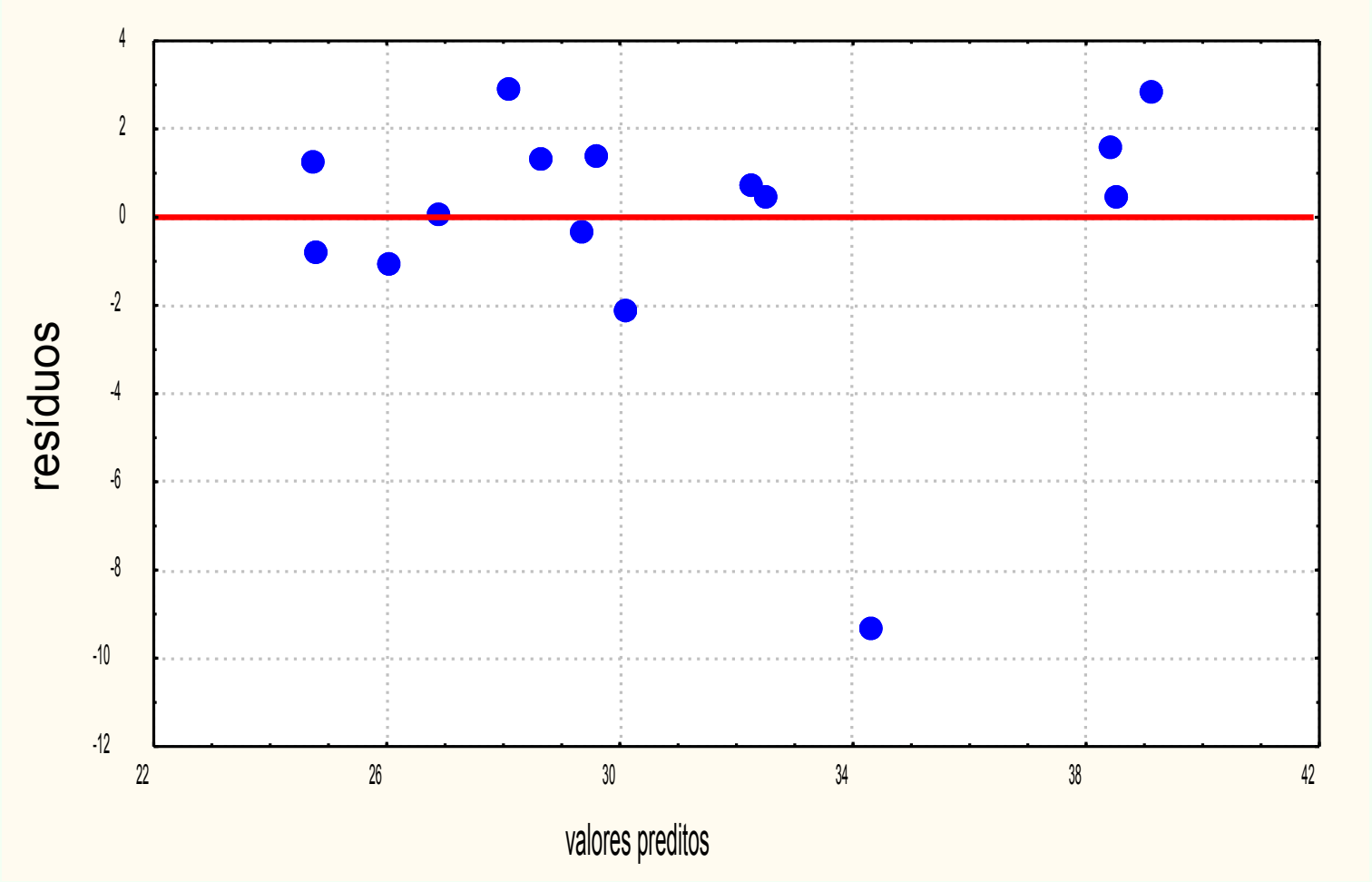
Análise dos resíduos:

O sistema de entrega de um distribuidor de cervejas



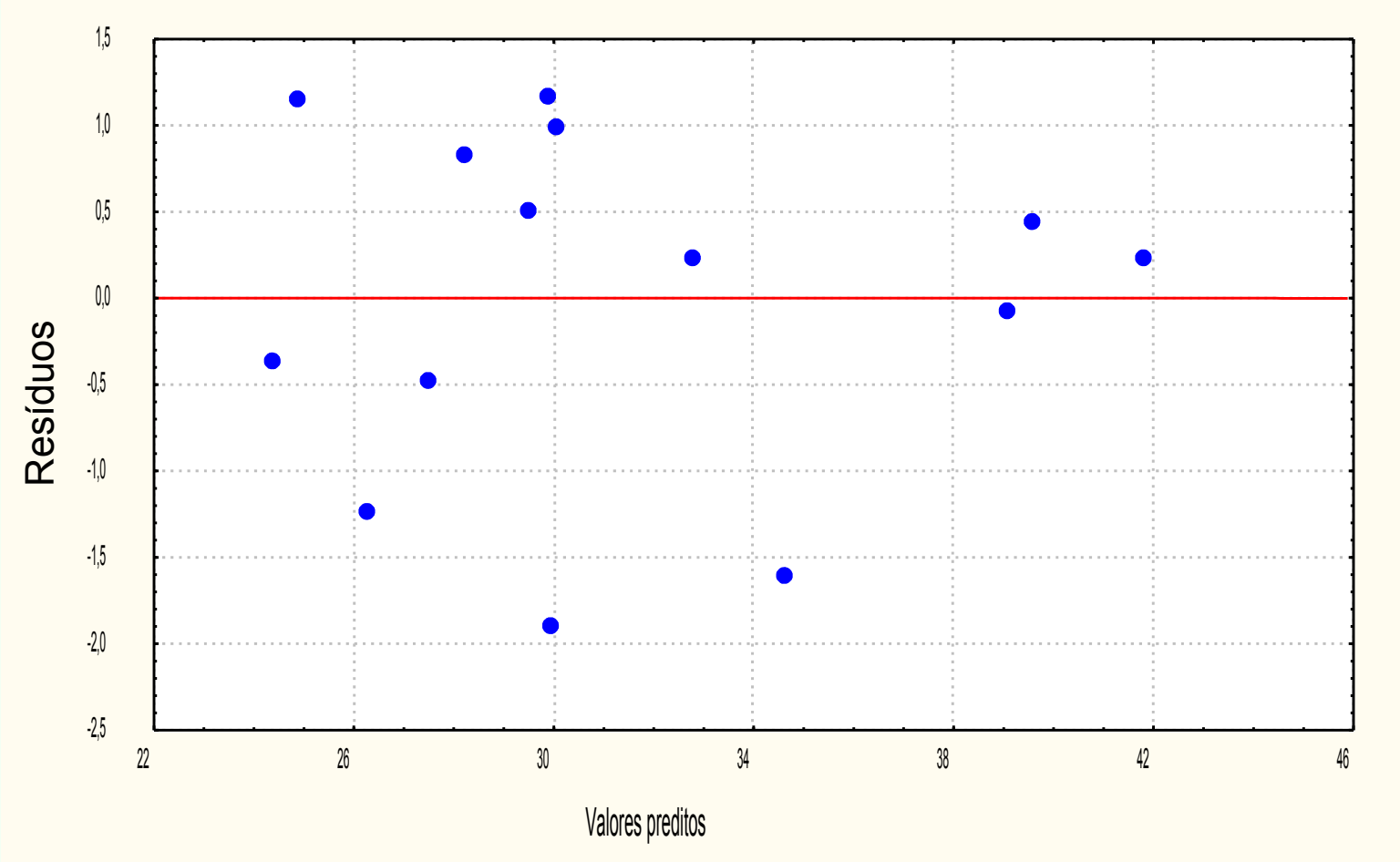
Análise dos resíduos:

O sistema de entrega de um distribuidor de cervejas



Análise dos resíduos: modelo sem o ponto discrepante

O sistema de entrega de um distribuidor de cervejas



Análise dos resíduos: modelo sem o ponto discrepante

O sistema de entrega de um distribuidor de cervejas

Resumo da regressão

$R^2 = 0{,}968$ $\quad s_e^2 = 1{,}0878$

$F(2,11) = 168{,}94$ $\quad p < 0{,}00000$

	Coef.	E.P.	$t_{(11)}$	valor p
Intercepto	2,92	2,03	1,44	0,179
ENTREGAS	1,00	0,05	18,35	0,000
DISTANC	0,38	0,05	7,39	0,000

tempo esperado = 2,92 + 1,00(nº de entregas) + 0,38(distância)

Regressão múltipla: variáveis independentes qualitativas

- Ex. (Chatterjee, Hadi e Price – "Regression Analysis by Example", 2000, p. 124)
- Variável dependente: salários de uma empresa;

- Variáveis independentes:
  - experiência (anos de trabalho na empresa);
  - cargo de gerência (0 = não, 1 = sim);
  - nível educacional (1 = primeiro grau
    2 = segundo grau
    3 = superior)

Regressão múltipla: variáveis independentes qualitativas

- As variáveis qualitativas devem entrar no modelo na forma de variáveis indicadoras (0 - 1);
  - cargo de gerência, **G** (0 = não, 1 = sim)
  - nível educacional, $\mathbf{E_1}$ (1 = primeiro grau 0 = caso contrário)
  - nível educacional, $\mathbf{E_2}$ (1 = segundo grau 0 = caso contrário)
    - $E_1 = 0, E_2 = 0$ ===> superior
    - $E_1 = 1, E_2 = 0$ ===> primeiro grau
    - $E_1 = 0, E_2 = 1$ ===> segundo grau

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: β_2 é o incremento esperado no salário pelo indivíduo ocupar um cargo de gerente.