# RESOLUÇÃO ASSISTIDA DE PROBLEMAS
# CAPÍTULO 14 – REGRESSÃO
# PROBLEMA 14.1

*APOIOS À RESOLUÇÃO*

Alínea (i)

Apoio 1 (apenas o resultado)

$Y = 34 + 2.288 \cdot (X - 16) + E$ . ■

Apoio 2 (sugestão)

O gráfico ($X$, $Y$) sugere que a relação $Y = f(X)$ pode ser aproximada por uma relação linear. Estime os parâmetros $\alpha$ e $\beta$ e verifique se $\beta = 0$ (versus $\beta > 0$). ■

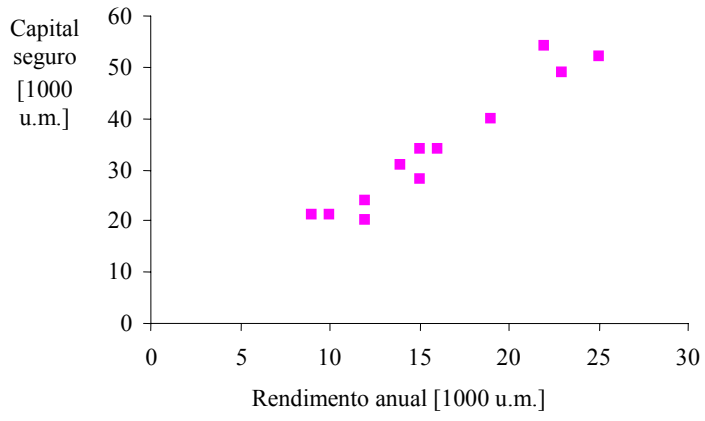
Apoio 3 (resolução completa)

Sejam,

$X$: rendimento anual [1000 u.m.].

$Y$: capital seguro [1000 u.m.].

Na figura seguinte apresenta-se o gráfico ($X$, $Y$) construído com os dados fornecidos




O gráfico sugere que $Y = f(X)$ pode ser aproximada por uma relação linear, cujo modelo pode ser descrito por:

$Y_n = \alpha + \beta \cdot (X_n - \overline{X}) + E_n$,

admitindo-se que os erros $E_n$ são independentes e Normais, com média 0 e variância $\sigma^2$.

Naquela expressão:

$n$: índice denotando as observações do par de variáveis $X$ e $Y$ $(n = 1, ..., 12)$

$(X_n, Y_n)$: $n$-ésima observação do par $(X, Y)$

$\alpha, \beta$: parâmetros fixos (ambos desconhecidos, a estimar) da relação linear entre $X$ e $Y$

$E_n$: erro aleatório associado ao valor observado $Y_n$

Os parâmetros $\alpha$ e $\beta$ podem ser estimados recorrendo ao método dos mínimos quadrados. A minimização da função *Soma dos Erros Quadráticos* (SEQ) conduz aos seguintes estimadores:

$$A = \frac{1}{N} \cdot \sum_n Y_n = \overline{Y} \text{ (estimador de } \alpha)$$

$$B = \frac{S_{XY}}{S_{XX}} \text{ (estimador de } \beta),$$

onde

$$S_{XY} = \sum_n \left[ \left( X_n - \overline{X} \right) \cdot \left( Y_n - \overline{Y} \right) \right]$$

e

$$S_{XX} = \sum_n \left[ \left( X_n - \overline{X} \right)^2 \right].$$

Considerando os valores fornecidos no enunciado, obtêm-se as estimativas seguintes:

$$\overline{x} = 16$$

$$s_{XY} = \sum \left( x_n - \overline{x} \right) \cdot \left( y_n - \overline{y} \right) = (14-16) \cdot (31-34) + \cdots + (16-16) \cdot (34-34) = 682$$

$$s_{XX} = \sum \left( x_n - \overline{x} \right)^2 = (14-16)^2 + (19-16)^2 + \cdots + (16-16)^2 = 298$$

$$s_{YY} = \sum \left( y_n - \overline{y} \right)^2 = (31-34)^2 + (40-34)^2 + \cdots + (34-34)^2 = 1664$$

$$a = \hat{\alpha} = \overline{y} = 34$$

$$b = \hat{\beta} = \frac{s_{XY}}{s_{XX}} = \frac{682}{298} = 2.288.$$

Na relação (linear) entre $X$ e $Y$ estima-se então que $b = \hat{\beta} = 2.288$ O teste mais importante será verificar a hipótese nula de que $\beta = 0$.

No caso de a hipótese alternativa ser formulada como $\beta \neq 0$ fará sentido recorrer ao teste $F$ da análise de variância. O procedimento do teste ANOVA tem a estrutura seguinte:

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0 .$$

A estatística de teste

$$ET = \frac{DQMDR}{DQMR}$$

seguirá uma distribuição $F_{1,10}$ se $H_0$ for verdadeira.

Neste caso, a tabela ANOVA resulta:

| *Fontes de Variação* | *Variações (V)* | *GL* | *DQM* |
|---|---|---|---|
| Devido à Regressão (DR) | $b \cdot S_{XY} = 2.289 \cdot 682 = 1561$ | 1 | 1561 |
| Residual (R) | $1664 - 1561 = 103$ | 10 | 10.3 |
| TOTAL (T) | $S_{YY} = 1664$ | 11 | |

Como

$$ET = \frac{DQMDR}{DQMR} = \frac{1561}{10.3} = 151.55 \quad >> \quad F_{1,10}(\alpha = 0.05) = 4.96 ,$$

a hipótese nula $\beta = 0$ é rejeitada ao nível de significância de 5% (com um valor de prova quase nulo).

No entanto, o teste a efectuar deveria ser unilateral, uma vez que parece ser apenas plausível que *Y* cresça com *X*, ou seja, que $\beta > 0$. Nesta situação não se deve recorrer o teste ANOVA (que é sempre bilateral) mas, preferencialmente, ao teste unilateral:

$$H_0 : \beta = 0$$

$$H_1 : \beta > 0 ,$$

utilizando a estatística

$$ET = \frac{B - \beta_0}{\frac{S}{\sqrt{S_{XX}}}} .$$

Quando $H_0$ é verdadeira *ET* segue uma distribuição $t_{N-2}$.

Para efectuar este teste falta determinar o valor de *S*. Ora,

$$s^2 = DQMR = \frac{s_{YY} - b \cdot s_{XY}}{n-2} = \frac{1664 - 2.288 \cdot 682}{10} = 10.318 .$$

Assim,

$$s = \sqrt{10.318} = 3.212 .$$

Substituindo os valores amostrais na expressão da estatística de teste vem:

$$ET = \frac{2.288 - 0}{\frac{3.212}{\sqrt{298}}} = 12.299 >> t_{10}(\alpha = 0.05) = 1.81,$$

pelo que se rejeita $H_0$, igualmente com um valor de prova $p \approx 0$.

A relação linear estimada entre *X* e *Y* será então

$$Y = 34 + 2.288 \cdot (X - 16) + E,$$

com

$$\hat{\sigma}_E = 3.21.$$

Uma vez que $\beta > 0$, isto é, que há uma relação linear significativa entre as variáveis *X* e *Y*, faz sentido calcular o respectivo coeficiente de determinação

$$R^2_{XY} = \frac{B^2 \cdot S_{XX}}{S_{YY}}.$$

Substituindo, resulta

$$r^2_{XY} = 0.938.$$

Recorde-se que $R^2_{XY}$ representa a proporção de variação de *Y* que é explicada pela regressão. Consequentemente, estima-se que 93.8% da dispersão que se observa nos montantes investidos em seguros de vida se deve à relação linear que existe entre esta variável e o rendimento anual das famílias. ■

Alínea (ii)

Apoio 1 (apenas o resultado)

30.08%. ■

Apoio 2 (sugestão)

O cálculo da probabilidade pretendida requer que se padronize a variável *Y*. A estimativa do valor esperado de $Y|X = 20$ corresponde ao valor sobra recta estimada fazendo *X* = 20, ou seja, a $\hat{\mu}_Y|x = 20$. Para determinar $P(Y > 45 | X = 20)$, será necessário dividir a variável $(Y - \hat{\mu}_Y)$ pela estimativa do seu desvio padrão. Note que $\hat{\mu}_Y$ (que, para simplificar, se denota por $\hat{Y}$) é a melhor previsão disponível de $Y|X$ e que o erro ($\delta$) que se comete nesta previsão é dado pela diferença $\delta = (Y - \hat{Y})$. Recorra então à expressão que permite calcular a variância de $\delta$. ■

Apoio 3 (resolução completa)

Em termos simbólicos a questão formula-se do seguinte modo:

$P(Y > 45 \mid X = 20) = ?$

Ora, para $x = 20$ e recorrendo á expressão da relação linear estimada entre $Y$ e $X$, vem

$\hat{\mu}_Y = \hat{Y} = 34 + 2.288 \cdot (20 - 16) = 43.152$ .

Note-se que $\hat{\mu}_Y$ (que, para simplificar, se denota por $\hat{Y}$ ) é a melhor previsão disponível de $Y|X$ e que o erro ($\delta$) que se comete nesta previsão é dado pela diferença $\delta = (Y - \hat{Y})$. Então,

$$P(Y > 45 \mid X = 20) = P\left( \frac{Y - \hat{Y}}{\sqrt{\hat{\sigma}^2_{(Y-\hat{Y})}}} > \frac{45 - 43.152}{\sqrt{\hat{\sigma}^2_{(Y-\hat{Y})}}} \right).$$

Uma vez que se admitiu a Normalidade dos erros $E_n$, as variáveis $Y_n$ serão ainda Normais (cuja variância $\sigma^2$ é estimada por *DQMR*). Do mesmo modo, os erros de previsão $\delta = Y - \hat{Y}$ também seguirão distribuições Normais. Nestas condições, no problema em causa, a variável padronizada

$$\frac{Y - \hat{Y}}{\sqrt{\hat{\sigma}^2_{(Y-\hat{Y})}}}$$

comportar-se-á como uma distribuição $t_{10}$ (note-se que $\hat{\sigma}^2_{(Y-\hat{Y})}$ é calculado com os mesmos graus de liberdade que $\hat{\sigma}^2 = s^2 = DQMR$ ).

A variância dos erros de previsão é dada por:

$$\mathrm{Var}(Y - \hat{Y}) = \left[ 1 + \frac{1}{N} + \frac{(X - \overline{X})^2}{S_{XX}} \right] \cdot \sigma^2 .$$

Para os dados do problema, a estimativa daquela variância é

$\hat{\sigma}^2_{(Y-\hat{Y})} = \left[ 1 + \frac{1}{12} + \frac{(20-16)^2}{298} \right] \cdot 10.3 = 11.732$ (ou $\hat{\sigma}_{(Y-\hat{Y})} = \sqrt{11.732} = 3.425$ ).

Substituindo, resulta finalmente

$$P(Y > 45 \mid X = 20) = P\left( t_{10} > \frac{45 - 43.152}{3.425} \right) = P(t_{10} > 0.539) = 30.08\% .$$ ■

## PROBLEMA 14.2

*APOIOS À RESOLUÇÃO*

Alínea (i)

Apoio 1 (apenas o resultado)

$Y = 270.(5) - 3.38 \cdot (X - 2001) + E$ . ■

Apoio 2 (sugestão)

O gráfico (*X*, *Y*) sugere que a relação $Y = f(X)$ pode ser aproximada por uma relação linear. Estime os parâmetros $\alpha$ e $\beta$ e verifique se $\beta \neq 0$ ou, melhor, se $\beta < 0$. ■

Apoio 3 (resolução completa)

Denote-se por *X* a variável que representa o ano e por *Y* o volume de produção.

Nestas condições,

$\bar{x} = 2001$

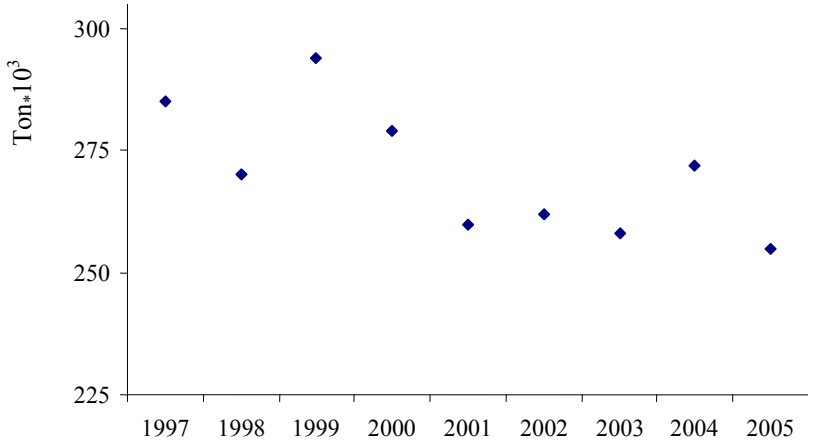
$\bar{y} = 270.(5)$

$s_{XX} = 60$

$s_{YY} = 1416.(2)$

$s_{XY} = -203.$

Na Figura seguinte apresenta-se o gráfico (*X*, *Y*) construído a partir dos dados disponíveis.



O gráfico sugere que a relação $Y = f(X)$ pode ser aproximada por uma relação linear, cujo modelo é descrito por:

$$Y_n = \alpha + \beta \cdot (X_n - \overline{X}) + E_n,$$

com

$$E_n \sim IN(0, \sigma^2).$$

No caso vertente,

$$a = \hat{\alpha} = \bar{y} = 270.(5)$$

e

$$b = \hat{\beta} = \frac{s_{XY}}{s_{XX}} = -\frac{203}{60} = -3.38.$$

Proceder-se-á a um unilateral a $\beta$ uma vez que apenas admitirá que $Y$ decresce com $X$, ou seja, que $\beta < 0$ (deve assim usar-se o teste $t$ em vez do teste ANOVA). As hipóteses em apreço são

$$H_0 : \beta = 0$$

$$H_1 : \beta < 0,$$

e a estatística de teste

$$ET = \frac{B - \beta_0}{\frac{S}{\sqrt{S_{XX}}}}$$

seguirá uma distribuição $t_7$ se $H_0$ for verdadeira.

O estimador de $\sigma$ é dado por:

$$S = \sqrt{\frac{S_{YY} - B \cdot S_{XY}}{N - 2}}.$$

Substituindo, obtém-se a estimativa

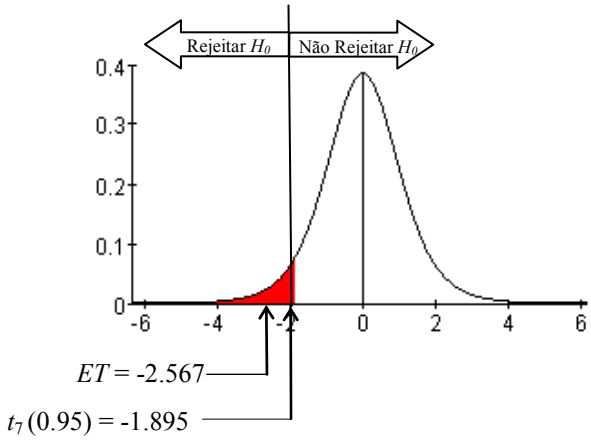
$$s = \sqrt{\frac{s_{YY} - b \cdot s_{XY}}{n - 2}} = \sqrt{\frac{1416.(2) + 3.38 \cdot (-203)}{7}} = \sqrt{104.3} = 10.2.$$

Como

$$ET = \frac{-3.38}{\frac{10.2}{\sqrt{60}}} = -2.567 \; < \; t_7(\alpha = 0.95) = -1.8946,$$

a hipótese nula é rejeitada com o valor de prova $p$ = 1.85%

Na figura seguinte ilustra-se este teste.




A relação linear estimada é então

$Y = 270.(5) - 3.38 \cdot (X - 2001) + E$,

com

$\hat{\sigma}_E = 10.2$.

O coeficiente de determinação (estimado) é:

$$r_{XY}^2 = \frac{b^2 \cdot s_{XX}}{s_{YY}} = \frac{(-3.38)^2 \cdot 60}{1416.2} = 0.484 .$$

Alínea (ii)

Apoio 1 (apenas o resultado)

45.9%. ■

Apoio 2 (sugestão)

Para cada valor de $X$, a melhor previsão de $Y$ obtém-se recorrendo à recta estimada. O erro, $\delta$, que se comete na previsão é dado pela diferença $\delta = Y - \hat{\mu}_Y$ (ou $\delta = Y - \hat{Y}$, já que $\hat{Y} = \hat{\mu}_Y | X$). Procure estimar a variância do erro de previsão, uma vez que, para calcular a probabilidade pretendida, deverá padronizar a variável $Y$. ■

Apoio 3 (resolução completa)

Sejam $Y_{05}$ e $Y_{06}$ os volumes de produção de trigo em 2005 e 2006, respectivamente. Simbolicamente, a questão formula-se do seguinte modo:

$P(Y_{06} > Y_{05}) = ?$

Em 2005 a produção foi, de facto, $Y_{05} = 255$. Para 2006, pode estimar-se (ou prever-se) o respectivo valor esperado $\hat{\mu}_Y | x = 2006$ (que, para simplificar, se denotará por $\hat{Y}_{06}$).

Assim,

$$P(Y_{06} > 255) = P(Y_{06} - \hat{Y}_{06} > 255 - \hat{Y}_{06}) = P\left( \frac{Y_{06} - \hat{Y}_{06}}{\sqrt{\text{Var}(Y_{06} - \hat{Y}_{06})}} > \frac{255 - \hat{Y}_{06}}{\sqrt{\text{Var}(Y_{06} - \hat{Y}_{06})}} \right).$$

A variância do denominador é, afinal, a variância do erro de previsão:

$$\text{Var}(Y_{06} - \hat{Y}_{06}) = \left[ 1 + \frac{1}{N} + \frac{(X - \overline{X})^2}{S_{XX}} \right] \cdot \sigma^2 .$$

A sua estimativa é

$$\hat{\sigma}^2_{(Y_{06} - \hat{Y}_{06})} = \left[ 1 + \frac{1}{9} + \frac{(2006 - 2001)^2}{60} \right] \cdot 104.3 = 159.35$$

(note-se que esta estimativa é obtida com os mesmos graus de liberdade que $\hat{\sigma}^2 = s^2 = DQMR$).

Por sua vez,

$\overline{Y}_{06} = 270.(5) - 3.38 \cdot (2006 - 2001) = 253.655$.

Substituindo, resulta finalmente

$$P(Y_{06} > 255) = P\left( \frac{255 - \hat{Y}_{06}}{\hat{\sigma}_{(Y_{06} - \hat{Y}_{06})}} = t_7 > \frac{255 - 253.655}{\sqrt{159.35}} \right) = P(t_7 > 0.106) = 45.9\% . \blacksquare$$

Alínea (iii)

Apoio 1 (apenas o resultado)

41.12%. ■

Apoio 2 (sugestão)

Considere o modelo de regressão linear simples

$Y_n = \alpha + \beta \cdot (X_n - \overline{X}) + E_n$.

Nestas condições, a previsão de $Y_n$ é

$\hat{Y}_n = \hat{\alpha} + \hat{\beta} \cdot (X_n - \overline{X})$.

Tente mostrar que

$$\text{Var}(\hat{Y}_{n_2} - \hat{Y}_{n_1}) = (n_2 - n_1)^2 \cdot \text{Var}(\hat{\beta}).$$

Esta expressão é necessária para estimar a variância da variável $\left[\left(Y_{n_2} - Y_{n_1}\right) - \left(\hat{Y}_{n_2} - \hat{Y}_{n_1}\right)\right]$ ■

Apoio 3 (resolução completa)

O volume de produção de trigo em 2006 e 2007 denota-se por $Y_{06}$ e $Y_{07}$. Nestas condições, o problema que é colocado especifica-se do seguinte modo:

$$P\left(Y_{07} > Y_{06}\right) = P\left[\left(Y_{07} - Y_{06}\right) > 0\right] = ?$$

As estimativas dos valores esperados (ou, previsões) $\hat{\mu}_Y \left| x = 2006 \right.$ e $\hat{\mu}_Y \left| x = 2007 \right.$ denotar-se-ão, respectivamente, por $\hat{Y}_{06}$ e $\hat{Y}_{07}$. Registe-se ainda que, de acordo com os pressupostos do modelo de regressão adoptado, se admite que as variáveis $Y_{06}$, $Y_{07}$, $\hat{Y}_{06}$ e $\hat{Y}_{07}$ seguem distribuições Normais.

Assim, a probabilidade $P\left[\left(Y_{07} - Y_{06}\right) > 0\right]$ será obtida a partir de:

$$P\left( \frac{\left(Y_{07} - Y_{06}\right) - \left(\hat{Y}_{07} - \hat{Y}_{06}\right)}{\sqrt{\mathrm{Var}\left[\left(Y_{07} - Y_{06}\right) - \left(\hat{Y}_{07} - \hat{Y}_{06}\right)\right]}} > \frac{0 - \left(\hat{Y}_{07} - \hat{Y}_{06}\right)}{\sqrt{\mathrm{Var}\left[\left(Y_{07} - Y_{06}\right) - \left(\hat{Y}_{07} - \hat{Y}_{06}\right)\right]}} \right).$$

O numerador do segundo membro desta expressão é dado por

$$-\hat{Y}_{07} + \hat{Y}_{06} = \hat{Y}_{06} - \hat{Y}_{07} = \left[a + b \cdot \left(2006 - 2001\right)\right] - \left[a + b \cdot \left(2007 - 2001\right)\right],$$

ou seja

$$\hat{Y}_{06} - \hat{Y}_{07} = b \cdot \left(2006 - 2007\right) = -3.38 \cdot -1 = 3.38.$$

Resta agora calcular o denominador, $\mathrm{Var}\left[\left(Y_{07} - Y_{06}\right) - \left(\hat{Y}_{07} - \hat{Y}_{06}\right)\right]$. Ora,

$$\mathrm{Var}\left[\left(Y_{07} - Y_{06}\right) - \left(\hat{Y}_{07} - \hat{Y}_{06}\right)\right] = \mathrm{Var}\left(Y_{07} - Y_{06}\right) + \mathrm{Var}\left(\hat{Y}_{07} - \hat{Y}_{06}\right).$$

O primeiro termo, $\mathrm{Var}\left(Y_{07} - Y_{06}\right)$, pode ser estimado por $\hat{\sigma}^2 + \hat{\sigma}^2 = 2 \cdot DQMR$. O segundo termo, $\mathrm{Var}\left(\hat{Y}_{07} - \hat{Y}_{06}\right)$, estima-se a partir da expressão de $\left(\hat{Y}_{07} - \hat{Y}_{06}\right)$ deduzida anteriormente.

Assim,

$$\mathrm{Var}\left(\hat{Y}_{07} - \hat{Y}_{06}\right) = \mathrm{Var}\left[b \cdot \left(2006 - 2007\right)\right] = \left(2006 - 2007\right)^2 \cdot \mathrm{Var}\left(b\right).$$

Ou seja,

$$\mathrm{Var}\left(\hat{Y}_{07} - \hat{Y}_{06}\right) = 1^2 \cdot \frac{\sigma^2}{s_{XX}}.$$

A estimativa de $\mathrm{Var}\left(\hat{Y}_{07} - \hat{Y}_{06}\right)$ é então

$$\hat{\sigma}^2_{\left(\hat{Y}_{07} - \hat{Y}_{06}\right)} = \frac{DQMR}{s_{XX}} = \frac{DQMR}{60} = \frac{104.3}{60}.$$

Substituindo, resulta

$$P\left(\frac{(Y_{07}-Y_{06})-(\hat{Y}_{07}-\hat{Y}_{06})}{\sqrt{DQMR+DQMR+\frac{DQMR}{60}}}>\frac{0-3.38}{\sqrt{2\cdot 104.3+\frac{104.3}{60}}}\right)=$$

$$=P\left(t_7>\frac{3.38}{10.2\cdot\sqrt{2+\frac{1}{60}}}\right)=P(t_7>0.233)=41.12\%\,.\ \blacksquare$$

## PROBLEMA 14.3

*APOIOS À RESOLUÇÃO*

Alínea (i)

Apoio 1 (apenas o resultado)

Ao nível de significância de 5%, pode afirmar-se que o IPC oficial subestima o IPC correcto. Por outro lado, não há evidência de que o IPC oficial cresça mais depressa (ou mais devagar) do que o IPC correcto. ■

Apoio 2 (sugestão)

Considere o modelo linear $Y_n=\alpha+\beta\cdot(X_n-\overline{X})+E_n$. Se o IPC oficial estiver bem calculado, será $\alpha=\overline{Y}$ e $\beta=1$. Teste ambas as hipóteses. ■

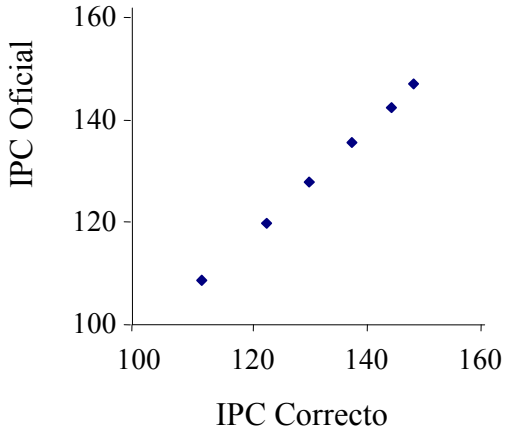
Apoio 3 (resolução completa)

Sejam,

*X*: IPC correcto.

*Y*: IPC oficial.

Na Figura seguinte apresenta-se o gráfico (*X*, *Y*) construído a partir dos dados fornecidos.




Pela observação da figura, o modelo linear $Y_n = \alpha + \beta \cdot \left(X_n - \overline{X}\right) + E_n$ parece perfeitamente adequado.

Assim,

$a = \hat{\alpha} = \overline{y} = 130.3$

e

$b = \hat{\beta} = \frac{s_{XY}}{s_{XX}} = \frac{1012.12}{987.54} = 1.025$.

Comece-se por testar $\alpha$. Aquilo que se pretende é verificar se $\alpha = \overline{Y}$. De facto assim será se o IPC oficial não subavaliar nem sobreavaliar o IPC correcto.

$H_0 : \alpha = \overline{y} = 133.4$

$H_1 : \alpha \neq 133.4$.

A estatística de teste

$$ET = \frac{A - \alpha}{\frac{S}{\sqrt{N}}},$$

seguirá uma distribuição $t_4$ ($t_{N-2}$) se $H_0$ for verdadeira.

Substituindo vem

$$ET = \frac{A - 133.4}{\frac{S}{\sqrt{6}}},$$

onde

$$S^2 = \frac{S_{YY} - B \cdot S_{XY}}{N - 2}.$$

Ora,

$$s^2 = \frac{1037.88 - 1.025 \cdot 1012.12}{4} = 0.1143,$$

pelo que resulta

$$ET = \frac{130.3 - 133.4}{\sqrt{\frac{0.1143}{6}}} = -22.47.$$

Como

$$|ET| = 22.47 >> t_4(\alpha = 0.025) = 2.776$$

A hipótese nula é rejeitada. Isto é, o IPC oficial subestima o IPC correcto.

Teste-se agora o coeficiente angular da recta de regressão, $\beta$. Se o IPC oficial espelhar devidamente o IPC correcto, a hipótese nula será $\beta = 1$.

$$H_0 : \beta = 1$$

$$H_1 : \beta \neq 1.$$

A estatística de teste

$$ET = \frac{B - \beta}{\frac{S}{\sqrt{S_{XX}}}},$$

seguirá uma distribuição $t_4$ ($t_{N-2}$) se $H_0$ for verdadeira.

Substituindo, vem

$$ET = \frac{1.025 - 1}{\frac{0.337}{\sqrt{987.54}}} = 2.074.$$

Como

$$ET = 2.074 < t_4(\alpha = 0.025) = 2.776,$$

não se rejeita $H_0$. Ou seja, ao nível de significância de 5%, não há evidência de que o IPC oficial cresça mais depressa (ou mais devagar) do que o IPC correcto.

■

## PROBLEMA 14.4

*APOIOS À RESOLUÇÃO*

Alínea (i)

Apoio 1 (apenas o resultado)

O parâmetro $\beta$ é significativamente diferente de zero (valor de prova $p \approx 0$). A recta estimada é $y = 43 - 0.3 \cdot x$. ■

Apoio 2 (sugestão)

Efectue um teste ANOVA para verificar se $\beta \neq 0$. ■

Apoio 3 (resolução completa)

Sejam,

*X*: densidade do tráfego, expressa em número de veículos por quilómetro

*Y*: velocidade média dos veículos, expressa em km/hora.

Adopte-se um modelo de regressão linear simples:

$Y = \alpha + \beta \cdot \left(X - \overline{X}\right) + E$,

no qual se admite que os eros são $IN\left(0, \sigma^2\right)$.

Nestas condições,

$a = \hat{\alpha} = \overline{y} = 25$

e

$$b = \hat{\beta} = \frac{s_{XY}}{s_{XX}} = -\frac{1500}{5000} = -0.3 .$$

O procedimento do teste ANOVA tem a estrutura seguinte:

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

recorrendo à estatística de teste:

$$ET = \frac{DQMDR}{DQMR} .$$

Neste caso, se $H_0$ for verdadeira *ET* seguirá uma distribuição $F_{1,18}$.

A tabela ANOVA correspondente é apresentada seguidamente.

| *Fontes de Variação* | *Variações (V)* | *GL* | *DQM* |
|---|---|---|---|
| Devido à Regressão (DR) | $(-0.3)^2 \cdot 5000 = 450$ | 1 | 450 |
| Residual (R) | 630–450=180 | 18 | 10 |
| Total (T) | $s_{YY} = 630$ | 19 | |

Como

$$ET = \frac{DQMDR}{DQMR} = \frac{450}{10} = 45 \quad >> \quad F_{1,18}(\alpha = 0.05) = 4.41,$$

a hipótese nula $H_0\ (\beta = 0)$ é rejeitada ao nível de significância de 5% (com um valor de prova quase nulo).

Nestas condições pode adoptar-se a recta estimada

$y = a + b \cdot (x - \bar{x}) = 25 - 0.3 \cdot (x - 60) = 43 - 0.3 \cdot x$ . ■

Alínea (ii)

Apoio 1 (apenas o resultado)

$[15.13, 28.87]$ ■

Apoio 2 (sugestão)

Para cada valor de *X*, a melhor previsão de *Y* obtém-se recorrendo à recta estimada. O erro que se comete na previsão, $\delta$, é dado pela diferença $\delta = Y - \hat{Y}$. Inicialmente procure estimar a variância do erro de previsão para, em seguida, calcular o intervalo de previsão pretendido. ■

Apoio 3 (resolução completa)

Para $X = 70$, vem

$\hat{y} = 43 - 0.3 \cdot x = 43 - 0.3 \cdot 70 = 43 - 21 = 22$ [km/hora].

Denote-se o erro de previsão por $\delta = Y - \hat{Y}$. De acordo com os pressupostos do modelo de regressão linear, tanto *Y* como $\hat{Y}$ são variáveis Normais (e independentes). Nestas condições, o intervalo de confiança a $(1-\alpha)\cdot 100\%$ para *Y* é dado por

$\hat{Y} \pm t_{GL}(\alpha = 0.025) \cdot \hat{\sigma}_\delta$ ,

sendo a variância dos erros de previsão, $\mathrm{Var}(\delta = Y - \hat{Y})$, calculada por:

$$\mathrm{Var}(\delta) = \mathrm{Var}(Y) + \mathrm{Var}(\hat{Y}) = \left[1 + \frac{1}{N} + \frac{(X - \bar{X})^2}{S_{XX}}\right] \cdot \sigma^2 .$$

O intervalo de previsão a 95% é então

$$\hat{Y} \pm t_{18}(\alpha = 0.025) \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{N} + \frac{(X - \overline{X})^2}{S_{XX}}} .$$

Substituindo,

$$22 \pm 2.1014 \cdot \sqrt{DQMR} \cdot \sqrt{1 + \frac{1}{20} + \frac{100}{5000}}$$

$$\equiv 22 \pm 2.1014 \cdot 3.162 \cdot 1.0344$$

$$\equiv 22 \pm 6.87 .$$

Finalmente, para uma densidade de tráfego de 70 veículos por quilómetro, a velocidade média situar-se-á no intervalo $[15.13, 28.87]$ (com 95% de confiança). ■

Alínea (iii)

Apoio 1 (apenas o resultado)

$x_{op} = 71.67$ e $\hat{\sigma}_{x_{op}} = 6.32$ . ■

Apoio 2 (sugestão)

Determine um ponto estacionário de $Z = X \cdot \hat{Y}$ e verifique se o valor em causa corresponde a um ponto óptimo. ■

Apoio 3 (resolução completa)

O volume, *V*, de tráfego do túnel (em veículos por hora) é dado por $V = X \cdot Y$ . Então,

$$\hat{v} = x \cdot \hat{y} = x \cdot [a + b \cdot (x - \overline{x})] = a \cdot x + b \cdot x^2 - b \cdot \overline{x} \cdot x .$$

O valor $x_{op}$ (densidade do tráfego) que optimiza aquela função é obtido por:

$$\frac{dv}{dx} = a + 2 \cdot b \cdot x - b \cdot \overline{x} = 0 ,$$

$$x_{op} = \frac{b \cdot \overline{x} - a}{2 \cdot b} = \frac{\overline{x}}{2} - \frac{a}{2 \cdot b} = \frac{60}{2} + \frac{25}{2 \cdot 0.3} = 71.67 .$$

Este valor corresponde a um ponto óptimo, uma vez que

$$\frac{d^2 v}{dx^2} = 2 \cdot b < 0 .$$

Estima-se em seguida o valor do desvio padrão do estimador da densidade óptima, $\hat{\sigma}_{x_{op}}$ . Ora

$$\mathrm{Var}(x_{op}) = \mathrm{Var}\left(\frac{\bar{x}}{2} - \frac{a}{2\cdot b}\right) = \mathrm{Var}\left(\frac{a}{2\cdot b}\right).$$

Assim,

$$f(a,b) = \frac{a}{2\cdot b},$$

$$\frac{df}{da} = \frac{1}{2\cdot b} \text{ e } \frac{df}{db} = \frac{-a}{2\cdot b^2}.$$

Do mesmo modo,

$$\left[\frac{\partial f}{\partial a}\,\middle|_{\,\hat{a},\hat{b}}\right]^2 = \left(\frac{1}{2\cdot(-0.3)}\right)^2 = 1.667^2 = 2.778$$

e

$$\left[\frac{\partial f}{\partial b}\,\middle|_{\,\hat{a},\hat{b}}\right]^2 = \left(\frac{-25}{2\cdot 0.09}\right)^2 = 19290.$$

Admitindo que $f(a,b)$ pode ser aproximada por uma relação linear, vem

$$\mathrm{Var}\left(\frac{2}{2\cdot b}\right) \approx \left[\frac{\partial f}{\partial a}\,\middle|_{\,\hat{a},\hat{b}}\right]^2 \cdot \mathrm{Var}(a) + \left[\frac{\partial f}{\partial b}\,\middle|_{\,\hat{a},\hat{b}}\right]^2 \cdot \mathrm{Var}(b).$$

Como

$$\hat{\mathrm{var}}(a) = \frac{\hat{\sigma}^2}{N} = \frac{10}{20} = 0.5 \text{ e } \hat{\mathrm{var}}(b) = \frac{\hat{\sigma}^2}{s_{XX}} = \frac{10}{5000} = 0.002,$$

resulta

$$\hat{\mathrm{var}}\left(\frac{2}{2\cdot b}\right) \approx 2.778^2 \cdot 0.5 + 19290^2 \cdot 0.002 = 39.97.$$

Finalmente,

$$\hat{\sigma}_{x_{op}} = \sqrt{\mathrm{Var}\left(\frac{a}{2\cdot b}\right)} = \sqrt{39.97} = 6.32. \blacksquare$$

## PROBLEMA 14.5

*APOIOS À RESOLUÇÃO*

Alínea (i)

Apoio 1 (apenas o resultado)

A manutenção deve ser programada em função do número de dias decorridos e não do número de horas de voo. ■

Apoio 2 (sugestão)

Para análise do modelo de regressão linear múltipla, recorra ao método progressivo (ou *forward*) para a selecção das variáveis independentes (regressores). ■

Apoio 3 (resolução completa)

Na análise do modelo de regressão linear múltipla, recorrer-se-á ao método progressivo (ou *forward*) para a selecção das variáveis independentes (regressores). Este método envolverá os seguintes passos:

(i) Ajustar dois modelos de regressão linear simples (um para cada um dos regressores, $X_1$ e $X_2$). O regressor que, isoladamente, explicar a maior proporção significativa da variável dependente ($Y$), será incluído no modelo.

(ii) Construir um modelo de regressão linear dupla que associe, como variáveis independentes, o regressor seleccionado no passo (i) e o outro regressor potencial ainda não incluído. Se este segundo regressor potencial não explicar uma proporção adicional significativa será excluído do modelo. Neste caso, o modelo proposto será o modelo de regressão linear simples seleccionado no passo anterior.

Passo (i)

Inicie-se este procedimento avaliando o modelo de regressão linear simples para a variável independente $X_1$:

$$Y = \alpha + \beta \cdot \left(X_1 - \overline{X}_1\right) + E\,,$$

com

$$a = \hat{\alpha} = \overline{y} = 2$$

e

$$b = \hat{\beta} = \frac{s_{X_1 Y}}{s_{X_1 X_1}} = -\frac{120}{1500} = -0.08\,.$$

O procedimento do teste ANOVA tem a estrutura seguinte:

$H_0 : \beta = 0$

$H_1 : \beta \neq 0$.

A estatística de teste

$$ET = \frac{DQMDR}{DQMR}$$

seguirá uma distribuição $F_{1,18}$ se $H_0$ for verdadeira.

A tabela ANOVA correspondente é

| *Fontes de Variação* | *Variações (V)* | *GL* | *DQM* |
|---|---|---|---|
| Devido à Regressão ($X_1$) | $0.08 \cdot 120 = 9.6$ | 1 | 9.6 |
| Residual ($X_1$) | 30.0–9.6 = 20.4 | 18 | 1.133 |
| Total | $s_{YY} = 30.0$ | 19 | |

Como

$$\frac{DQMDR}{DQMR} = \frac{9.6}{1.133} = 8.47 \quad > \quad F_{1,18}(\alpha = 0.05) = 4.41$$

a hipótese nula $H_0$ $(\beta = 0)$ é rejeitada.

Constrói-se de seguida o modelo de regressão linear simples $Y = f(X_2)$. Seja,

$Y = \alpha + \beta_2 \cdot (X_2 - \overline{X}_2) + E$.

Assim,

$a = \hat{\alpha} = \overline{y} = 2$

e

$$b = \hat{\beta} = \frac{s_{X_2 Y}}{s_{YY}} = \frac{60}{300} = 0.2 .$$

Neste caso, a tabela ANOVA é

| *Fontes de Variação* | *Variações (V)* | *GL* | *DQM* |
|---|---|---|---|
| Devido à Regressão ($X_2$) | $0.2 \cdot 60 = 12.0$ | 1 | 12.0 |
| Residual ($X_2$) | 30.0–12.0 = 18.0 | 18 | 1.0 |
| Total | $s_{YY} = 30.0$ | 19 | |

Como

$$\frac{DQMDR}{DQMR} = \frac{12.0}{1.0} = 12.0 \quad > \quad F_{1,18}(\alpha = 0.05) = 4.41$$

a hipótese nula $H_0\ (\beta = 0)$ deve ser rejeitada.

Passo (ii)

Embora ambos os modelos expliquem uma proporção significativa da variação total, a proporção explicada é maior no segundo. Assim, a variável $X_2$ passa a estar incluída no modelo, havendo que determinar se há vantagem ou não em acrescentar-lhe $X_1$. Verifique-se então o modelo de regressão linear dupla $Y = f(X_1, X_2)$:

$$Y = \alpha + \beta_1 \cdot (X_1 - \overline{X}_1) + \beta_2 \cdot (X_2 - \overline{X}_2) + E\,,$$

com

$$a = \hat{\alpha} = \overline{y} = 2\,.$$

As estimativas $b_1 = \hat{\beta}_1$ e $b_2 = \hat{\beta}_2$ são obtidas resolvendo o sistema de equações

$$\begin{cases} b_1 \cdot s_{X_1X_1} + b_2 \cdot s_{X_1X_2} = s_{X_1Y} \\ b_1 \cdot s_{X_2X_1} + b_2 \cdot s_{X_2X_2} = s_{X_2Y} \end{cases}$$

Substituindo,

$$\begin{cases} 1500 \cdot b_1 + 500 \cdot b_2 = 120 \\ 500 \cdot b_1 + 300 \cdot b_2 = 60 \end{cases} \equiv \begin{cases} 1500 \cdot b_1 + 500 \cdot b_2 = 120 \\ -1500 \cdot b_1 - 900 \cdot b_2 = -180 \end{cases}$$

$$-400 \cdot b_2 = -60 \quad \Rightarrow \quad b_2 = \frac{60}{400} = 0.15$$

$$-1500 \cdot b_1 - 500 \cdot 0.15 = 120 \quad \Rightarrow \quad b_1 = \frac{45}{1500} = 0.03\,.$$

A tabela ANOVA para modelo de regressão linear dupla é então:

| *Fontes de Variação* | *Variações (V)* | *GL* | *DQM* |
|---|---|---|---|
| Devido à Regressão ($X_2$, $X_1$) | $0.03 \cdot 120 + 0.15 \cdot 60 = 12.6$ | *2* | |
| $DR(X_2)$ | (12.0) | (1) | |
| $DR(X_1 \mid X_2)$ | (12.6 – 12.0 = 0.6) | (1) | 0.6 |
| Residual ($X_2$, $X_2$) | 30.0–12.6=17.4 | 17 | 1.023 |
| Total | $s_{YY} = 30.0$ | 19 | |

Como,

$$\frac{DQMDR(X_1 \mid X_2)}{DQMR} = \frac{0.6}{1.023} = 0.586 < F_{1,17}(\alpha = 0.05) = 4.45$$

a hipótese nula $H_0\ (\beta_1 = 0)$ não é rejeitada.

Assim, ao nível de significância de 5%, concluiu-se que não valerá a pena incluir a variável $X_1$ no modelo de regressão. Neste caso, $Y$ será função apenas do número de dias. Consequentemente, a manutenção deve ser programada em função do número de dias decorridos e não do número de horas de voo.

Finalmente, o modelo sugerido é o seguinte:

$Y = \alpha + \beta \cdot \left(X_2 - \overline{X}_2\right) + E$,

com

$a = \hat{\alpha} = \overline{y} = 2$

e

$b = \hat{\beta} = \frac{s_{X_2 Y}}{s_{YY}} = -\frac{60}{300} = -0.2$.

Nestas condições resulta $\overline{x}_2 = 10$ e $s^2 = \hat{\sigma}^2 = 1.0$.

Alínea (ii)

Apoio 1 (apenas o resultado)

3.82%. ■

Apoio 2 (sugestão)

Para cada valor de *X*, a melhor previsão de *Y* obtém-se recorrendo à recta estimada. O erro que se comete na previsão, $\delta$, é dado pela diferença $\delta = Y - \hat{Y}$. Inicialmente procure estimar a variância do erro de previsão para, em seguida, calcular o intervalo de previsão pretendido. ■

Apoio 3 (resolução completa)

Faça-se *x* = 15 (valor que corresponde a regular o aparelho logo que decorrerem 15 dias após a intervenção anterior). Neste caso

$\hat{y} = 2 + 0.2 \cdot (15 - 10) = 3$.

Padronizando a variável $\delta = Y - \hat{Y}$ vem,

$$\frac{Y - \hat{Y}}{S \cdot \sqrt{1 + \frac{1}{N} + \frac{\left(X_2 - \overline{X}_2\right)^2}{S_{X_2 X_2}}}},$$

que, atendendo à Normalidade de $\delta = Y - \hat{Y}$ seguirá uma distribuição $t_{18}$ (note-se que 18 corresponde ao número de graus de liberdade utilizado no cálculo de $S_{X_2 X_2}$). Substituindo,

$$\frac{Y - 3}{1 \cdot \sqrt{1 + \frac{1}{20} + \frac{25}{300}}} = \frac{Y - 3}{1.065}.$$

Ora,

$$P(Y>5)=P\left(\frac{Y-3}{1.065}=t_{18}\ >\ \frac{Y-3}{1.065}=1.88\right)$$

ou, finalmente

$P(Y>5)=0.0382$. ■

## PROBLEMA 14.6

*APOIOS À RESOLUÇÃO*

Apoio 1 (apenas o resultado)

Sem catalisador: $r_0=75.75+1.482\cdot(t-24.25)$

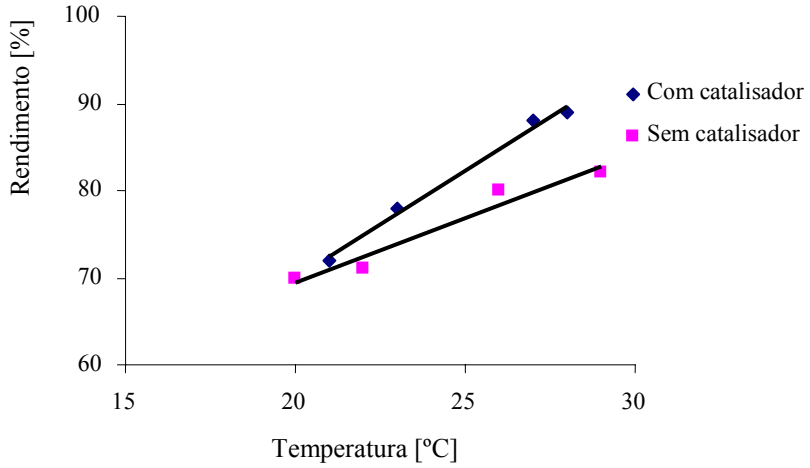
Com catalisador: $r_1=81.75+2.465\cdot(t-24.75)$. ■

Apoio 2 (sugestão)

Deverá testar se a diferença dos coeficientes angulares de dois modelos independentes - um para os rendimentos sem presença do catalizador e outro para rendimentos na presença de catalisador - é ou não significativa. Se não o for, a presença ou não do catalisador no processo químico é um factor qualitativo pode ser representada por uma variável muda (do tipo 0 ou 1). ■

Apoio 3 (resolução completa)

Na Figura seguinte representa-se a variável *R* (rendimento do processo químico) em função da temperatura de reacção (variável *T*), apresentando as duas rectas ajustadas para as situações em que o catalizador está e não está presente.




O gráfico sugere que a presença do catalisador pode contribuir para o aumento do efeito da temperatura sobre o rendimento.

Em primeiro lugar estima-se o modelo de regressão linear simples, que modela a variável *R* em função de *T*, quando na reacção não está presente o catalizador. Nestas condições, obtêm-se os seguintes resultados:

$\bar{t} = 24.25 \qquad \bar{r} = 75.75$

$s_{TR} = 72.25 \quad s_{TT} = 48.75 \quad s_{RR} = 112.75$

$a_0 = 75.75$

e

$$b_0 = \frac{72.25}{48.75} = 1.482 .$$

Repete-se o exercício utilizando os resultados das situações em que o catalizador está presente:

$\bar{t} = 24.75 \qquad \bar{r} = 81.75$

$s_{TR} = 80.75 \quad s_{TT} = 32.75 \quad s_{RR} = 200.75$

$a_1 = 81.75$

e

$$b_1 = \frac{80.75}{32.75} = 2.465 .$$

Admitindo que $\sigma^2 = \sigma_0^2 = \sigma_1^2$, a variância comum $\sigma^2$ pode ser estimada recorrendo à expressão

$$S^2 = \frac{1}{N_0 + N_1 - 4} \cdot \left[ (N_0 - 2) \cdot S_0^2 + (N_1 - 2) \cdot S_1^2 \right],$$

onde

$$S_0^2 = \hat{\sigma}_0^2 = \frac{1}{N_0 - 2} \cdot \sum_n \hat{E}_{0n}^2$$

e

$$S_1^2 = \hat{\sigma}_1^2 = \frac{1}{N_1 - 2} \cdot \sum_n \hat{E}_{1n}^2 .$$

Fazendo os cálculos,

$s_0^2 = (112.75 - 1.482 \cdot 72.25)/2 = 2.836$ ,

$s_1^2 = (200.75 - 2.465 \cdot 80.75)/2 = 0.824$

e

$$s^2 = \frac{1}{4} \cdot (2 \cdot 2.836 + 2 \cdot 0.824) = 1.830 = 1.353^2 .$$

Para testar se a diferença entre $\beta_0$ e $\beta_1$ é significativa, recorre-se a um teste unilateral com a estrutura seguinte:

$$H_0 : \beta_0 = \beta_1$$

$$H_1 : \beta_0 < \beta_1$$

A estatística de teste

$$ET = \frac{B_1 - B_0}{S \cdot \sqrt{\frac{1}{S_{TT_0}^2 + S_{TT_1}^2}}}$$

segue uma distribuição $t_{(N_0+N_1-4)}$ (no caso em análise, $t_4$), se $H_0$ for verdadeira.

Como

$$ET = \frac{2.465 - 1.482}{1.353 \cdot \sqrt{\frac{1}{48.75 + 32.75}}} = 59.21 >> t_4(0.025) = 2.776$$

a hipótese nula é rejeitada (com um valor de prova praticamente nulo).

Nestas condições, devem considerar-se dois modelos independentes:

- um para os rendimentos sem catalizador, cuja relação é estimada por $r_0 = 75.75 + 1.482 \cdot (t - 24.25)$
- outro para os rendimentos na presença do catalizador, cuja relação é estimada por $r_1 = 81.75 + 2.465 \cdot (t - 24.75)$.

Note-se que se $H_0$ não tivesse sido rejeitada, poderia fazer sentido adoptar um modelo de regressão único com dois regressores: a variável *T* (temperatura) e a variável muda *Z*, que tomaria os seguintes valores:

$$Z : \begin{cases} 0, \text{se o catalizador não estiver presente} \\ 1, \text{caso contrário} \end{cases} . \blacksquare$$

## PROBLEMA 14.7

*APOIOS À RESOLUÇÃO*

Apoio 1 (apenas o resultado)

Previsão pontual: 101936

Intervalo de previsão: [73203, 138137]. ■

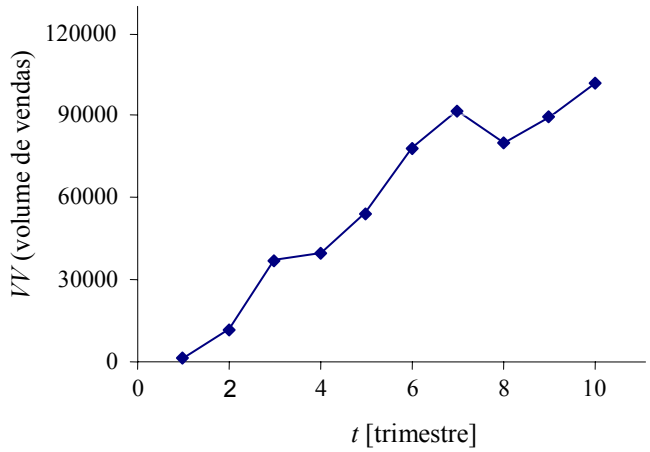
Apoio 2 (sugestão)

Pela análise do gráfico que representa o volume de vendas em função do trimestre, aceita-se que a relação possa apresentar um padrão de «curva S». Esta relação pode ser linearizada recorrendo-se, simultaneamente, à transformação

inversa da variável independente e à transformação logarítmica da variável dependente. ■

Apoio 3 (resolução completa)

Na figura seguinte representa-se graficamente o volume de vendas em função do trimestre.




O gráfico sugere que a relação $VV = f(t)$ apresenta um padrão de «curva S», ou seja uma relação do tipo:

$VV_n = e^{\alpha' + \beta / t_n + E_n}$ (com $\alpha' > 0$ e $\beta < 0$).

Esta relação pode ser linearizada recorrendo-se, simultaneamente, à transformação inversa da variável independente e à transformação logarítmica da variável dependente:

$X = 1/t$

$Y = \ln(VV)$.

Na tabela seguinte apresentam-se os cálculos correspondentes, efectuados a partir dos dados fornecidos.

| $t$ | $vv$ | $x = 1/t$ | $y = \ln(vv)$ |
|---|---|---|---|
| 1 | 1206 | 1.000 | 7.0953 |
| 2 | 11356 | 0.500 | 9.3375 |
| 3 | 36888 | 0.333 | 10.5156 |
| 4 | 39636 | 0.250 | 10.5875 |
| 5 | 53887 | 0.200 | 10.8946 |
| 6 | 77806 | 0.167 | 11.2620 |
| 7 | 91627 | 0.143 | 11.4255 |
| 8 | 80147 | 0.125 | 11.2916 |
| 9 | 89646 | 0.111 | 11.4036 |
| 10 | 101677 | 0.100 | 11.5295 |

O modelo linearizado é então

$$Y_n = \alpha + \beta \cdot \left(X_n - \overline{X}\right) + E_n ,$$

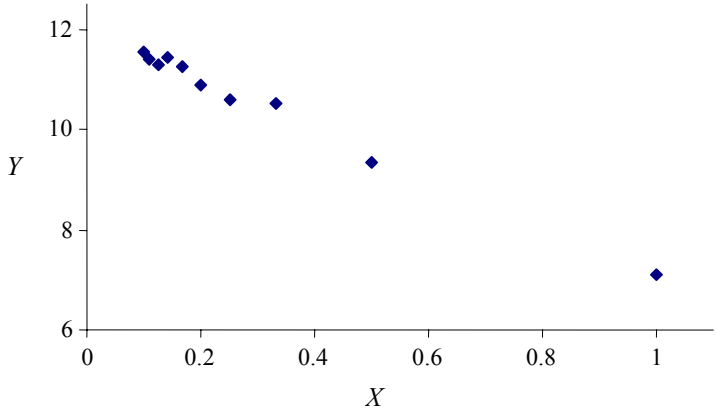
admitindo-se que os erros $E_n$ são independentes e Normais, com média zero e variância $\sigma^2$.

Para este modelo obtêm-se as seguintes estatísticas:

$\overline{x} = 0.2929 \qquad \overline{y} = 10.5343$

$s_{XX} = 0.6919 \qquad s_{YY} = 17.0355 \qquad s_{XY} = -3.4188 .$

Na figura seguinte representam-se graficamente as observações do modelo linearizado.




Por observação desta figura, pode afirmar-se que o modelo obtido por transformação de variáveis parece ser efectivamente linear. Nesse caso:

$a = \hat{\alpha} = \overline{y} = 10.5343$

e

$$b = \hat{\beta} = \frac{s_{XY}}{s_{XX}} = -\frac{3.4188}{0.6919} = -4.9412 .$$

A tabela ANOVA correspondente é:

| *Fontes de Variação* | *Variações (V)* | *GL* | *DQM* |
|---|---|---|---|
| Devido à Regressão (DR) | $(-4.9412)\cdot(-3.4188) = 16.893$ | 1 | 16.893 |
| Residual (R) | 17.036 - 16.893 = 0.143 | 8 | 0.0178 |
| Total (T) | $s_{YY} = 17.036$ | 9 | |

O procedimento do teste ANOVA tem a estrutura seguinte:

$H_0 : \beta = 0$

$H_1 : \beta \neq 0 .$

A estatística de teste

$$ET = \frac{DQMDR}{DQMR}$$

seguirá uma distribuição $F_{1,8}$, se $H_0$ for verdadeira.

Como

$$ET = \frac{DQMDR}{DQMR} = \frac{16.893}{0.0178} = 949 \quad > \quad F_{1,8}(\alpha = 0.05) = 5.32$$

a hipótese nula é rejeitada.

Para $t = 11$ (ou seja, para $x = 1/11 = 0.0909$), o intervalo de previsão de $Y$ a 95%, centrado na estimativa pontual

$$\hat{y} = 10.534 - 4.941 \cdot (0.0909 - 0.2929) = 11.5321,$$

é:

$$\hat{Y} \pm t_8(\alpha = 0.025) \cdot S \cdot \sqrt{1 + \frac{1}{N} + \frac{(X - \overline{X})^2}{S_{XX}}}\,.$$

Como

$$s^2 = DQMR = 0.0178 \quad \Rightarrow \quad s = 0.1334,$$

vem

$$11.5321 \pm 2.306 \cdot 0.1334 \cdot \sqrt{1 + \frac{1}{10} + \frac{(0.0909 - 0.2929)^2}{0.6919}}$$

$$\equiv 11.5321 \pm 0.3312,$$

pelo que o intervalo pretendido para $Y|x = 11$ é

[11.201, 11.863].

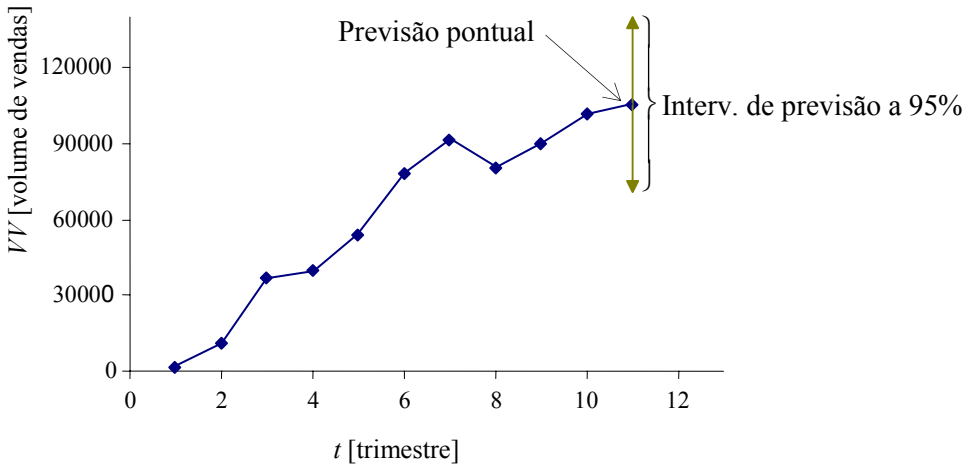
Calcule-se agora a o valor previsto para o volume de vendas no trimestre 11 ($vv_{11}$). A previsão pontual é:

$$\hat{vv}_{11} = e^{11.5321} = 101936.$$

Por sua vez, o intervalo de previsão a 95% resulta:

$$\left[e^{11.201}, e^{11.863}\right] \equiv [73203, 141917].$$

Na figura seguinte representam-se graficamente estes resultados.




■

## PROBLEMA 14.8

*APOIOS À RESOLUÇÃO*

Alínea (i)

Apoio 1 (apenas o resultado)

$\hat{p} = e^{-0.010-0.157 \cdot I}$ . ■

Apoio 2 (sugestão)

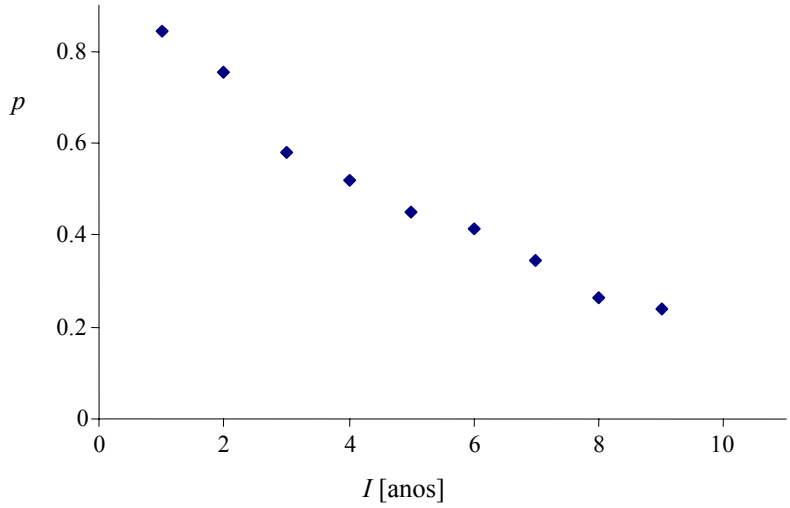
Pela análise do gráfico $p = f(I)$ verifica-se que esta relação apresenta um padrão exponencial. Esta relação pode ser linearizada recorrendo-se à transformação logarítmica da variável dependente. ■

Apoio 3 (resolução completa)

Na figura seguinte representam-se graficamente os dados relativos à variável

$$p = \frac{PVU}{PVN}$$

Em função da idade do veículo em anos ($I$).




O gráfico sugere que a relação $p = f(I)$ apresenta um padrão exponencial, cujo modelo é dado por

$p_n = e^{\alpha' + \beta \cdot I_n + E_n}$ (com $\alpha' > 0$ e $\beta < 0$).

Esta relação pode ser linearizada recorrendo-se à transformação logarítmica da variável dependente:

$Y = \ln(p)$.

Na tabela seguinte apresentam-se os dados amostrais transformados.

| $i$ | $p$ | $y = \ln(p)$ |
|---|---|---|
| 1 | 0.843 | -0.171 |
| 2 | 0.753 | -0.284 |
| 3 | 0.580 | -0.545 |
| 4 | 0.520 | -0.654 |
| 5 | 0.452 | -0.794 |
| 6 | 0.414 | -0.882 |
| 7 | 0.346 | -1.061 |
| 8 | 0.264 | -1.332 |
| 9 | 0.241 | -1.423 |

O modelo linearizado é então

$Y_n = \alpha + \beta \cdot (I_n - \bar{I}) + E_n$

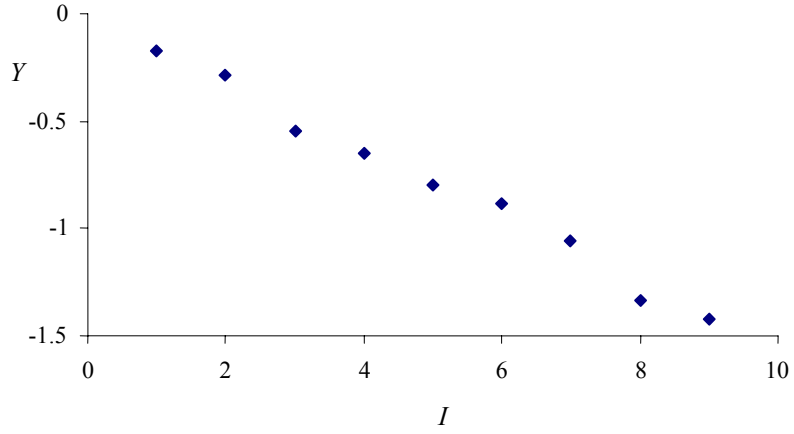
admitindo-se que os erros $E_n$ são independentes e Normais, com média zero e variância $\sigma^2$.

Para este modelo obtêm-se as seguintes estatísticas:

$\bar{i} = 5$ $\quad \bar{y} = -0.794$

$s_{II} = 60$ $\quad s_{IY} = -9.414$ $\quad s_{YY} = 1.495$

Na figura seguinte representam-se graficamente as observações do modelo linearizado.




Por observação desta figura, pode afirmar-se que o modelo obtido por transformação de variáveis parece ser efectivamente linear. Assim:

$a = \hat{\alpha} = \bar{y} = -0.794$

e

$b = \hat{\beta} = \frac{s_{IY}}{s_{II}} = -\frac{-9.414}{60} = -0.157$ .

A tabela ANOVA correspondente é:

| *Fontes de Variação* | *Variações (V)* | *GL* | *DQM* |
|---|---|---|---|
| Devido à Regressão (DR) | 1.478 | 1 | 1.478 |
| Residual (R) | 0.017 | 7 | 0.00243 |
| Total (T) | $s_{YY} = 1.495$ | 8 | |

O procedimento do teste ANOVA tem a estrutura seguinte:

$H_0 : \beta = 0$

$H_1 : \beta \neq 0$ .

A estatística de teste

$$ET = \frac{DQMDR}{DQMR}$$

seguirá uma distribuição $F_{1,7}$, se $H_0$ for verdadeira.

Como

$$\frac{DQMDR}{DQMR} = \frac{1.478}{0.00243} = 608.23 \quad > \quad F_{1,7}(\alpha = 0.05) = 5.59$$

a hipótese nula é rejeitada, com valor de prova quase nulo.

A relação estimada será então

$\hat{Y} = \ln p = -0.794 - 0.157 \cdot (I - 5)$.

Por uma questão de precaução, proceder-se-á verificação da validade desta relação. Seja $I = 0$. Nesta situação o *PVU* (preço do veículo usado) e o *PVN* (preço do veículo novo) devem ser praticamente iguais, pelo que $p$ deverá tomar um valor muito próximo da unidade. Ora

$\hat{Y} = -0.794 + 0.784 = -0.010$

e

$\hat{p} = e^{-0.01} = 0.99 \approx 1$.

Considera-se assim verificada a relação linear estimada. Consequentemente,

$\hat{p} = e^{-0.794 - 0.157 \cdot (I - 5)}$

ou, finalmente, a função que melhor traduz a relação entre $p$ e $I$ é:

$\hat{p} = e^{-0.010 - 0.157 \cdot I}$. ■

Alínea (ii)

Apoio 1 (apenas o resultado)

19.0%. ■

Apoio 2 (sugestão)

Para cada valor de $I$, a melhor previsão de $Y = \ln p$ obtém-se recorrendo à recta estimada. O erro que se comete na previsão é dado pela diferença $\delta = Y - \hat{Y}$. Inicialmente procure estimar a variância do erro de previsão para, em seguida, calcular o intervalo de previsão pretendido. ■

Apoio 3 (resolução completa)

A probabilidade de $p$ ultrapassar 0.65 é

$P(p > 0.65) = P[Y = \ln p > \ln(0.65)]$.

Ou seja, o que se pretende é determinar o valor de

$P(Y > -0.431)$.

Tomado a variável $\delta = Y - \hat{Y}$, vem:

$$P\left( \frac{Y-\hat{Y}}{S\cdot\sqrt{1+\frac{1}{N}+\frac{\left(I-\bar{I}\right)^2}{S_{II}}}} > \frac{-0.431-\hat{Y}}{S\cdot\sqrt{1+\frac{1}{N}+\frac{\left(I-\bar{I}\right)^2}{S_{II}}}} \right).$$

Recorde-se que, de acordo com os pressupostos do modelo de regressão, tanto $Y$ como $\hat{Y}$ serão Normais. Nestas condições, a variável

$$\frac{-0.431-\hat{Y}}{S\cdot\sqrt{1+\frac{1}{N}+\frac{\left(I-\bar{I}\right)^2}{S_{II}}}}$$

seguirá uma distribuição $t_{N-2}$. Trata-se então de calcular

$$P\left( t_{N-2} > \frac{-0.431-\hat{Y}}{S\cdot\sqrt{1+\frac{1}{N}+\frac{\left(I-\bar{I}\right)^2}{S_{II}}}} \right)$$

para um veículo com 3 anos de idade (ou seja, para $I = 3$). Nessas condições, vem:

$\hat{y} = -0.794 - 0.157\cdot(3-5) = -0.480$.

Por sua vez,

$s^2 = DQMR = 0.00243$ (ou $s = 0.0483$).

Assim, substituindo, resulta

$$P(p > 0.65) = P\left( t_7 > \frac{-0.431+0.480}{0.0483\cdot\sqrt{1+1/9+4/60}} = 0.935 \right)$$

e

$P(t_7 > 0.935) = 19.04\%$. ■