Regressão Linear Simples

Capítulo 16, Estatística Básica (Bussab&Morettin, 8a Edição)

10a AULA – 18/05/2015

MAE229 - Ano letivo 2015

Lígia Henriques-Rodrigues

Introdução

A análise de regressão estuda a relação entre uma variável chamada a variável dependente e outras variáveis chamadas variáveis independentes.

A relação entre elas é representada por um modelo matemático, que associa a variável dependente com as variáveis independentes.

Este modelo é designado por modelo de regressão linear simples (MRLS) se define uma relação linear entre a variável dependente e uma variável independente.

Se em vez de uma, forem incorporadas várias variáveis independentes, o modelo passa a denominar-se **modelo de regressão linear múltipla**.

No MRLS vamos estudar a relação linear entre duas variáveis quantitativas.

Exemplos:

- Altura dos pais e altura dos filhos;
- Renda semanal e despesas de consumo;
- Variação dos salários e taxa de desemprego;
- Demanda dos produtos de uma firma e publicidade.

Sob dois pontos de vista:

- Explicitando a forma dessa relação: regressão
- Quantificando a força ou o grau dessa relação: correlação

As técnicas de análise de correlação e regressão estão muito ligadas.

Diagrama de dispersão

Os dados para a análise de regressão e correlação simples são da forma:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$$

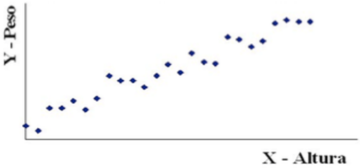
Com base nos dados constrói-se o diagrama de dispersão, que deve exibir uma tendência linear para que se possa usar a regressão linear.

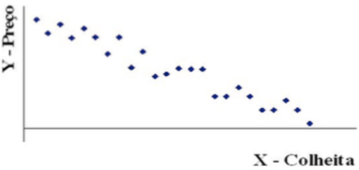
Este diagrama permite decidir empiricamente:

• se um relacionamento linear entre as variáveis X e Y deve ser assumido

• se o grau de relacionamento linear entre as variáveis é forte ou fraco, conforme o modo como se situam os pontos em redor de uma recta imaginária que passa através do enxame de pontos.

Diagramas de dispersão que sugerem uma regressão linear entre as variáveis

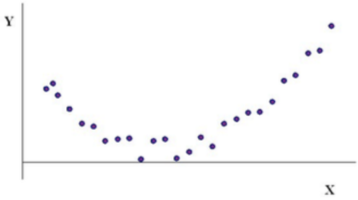


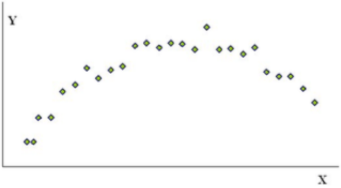


Existência de correlação positiva (em média, quanto maior for a altura maior será o peso)

Existência de correlação negativa (em média, quanto maior for for a colheita menor será o preço

Diagramas de dispersão que sugerem uma regressão não linear entre as variáveis





Nota:

O termo **linear** é usado para indicar que o modelo é linear nos parâmetros da regressão, α e β e não porque Y é função linear dos X's. Por exemplo, uma expressão da forma $E(Y|x) = \alpha + \beta x + \gamma x^2$, é um modelo linear em α, β e γ, mas o modelo $E(Y|x) = \alpha \exp^{\beta x}$, não é um modelo linear em α e β.

Coeficiente de correlação linear

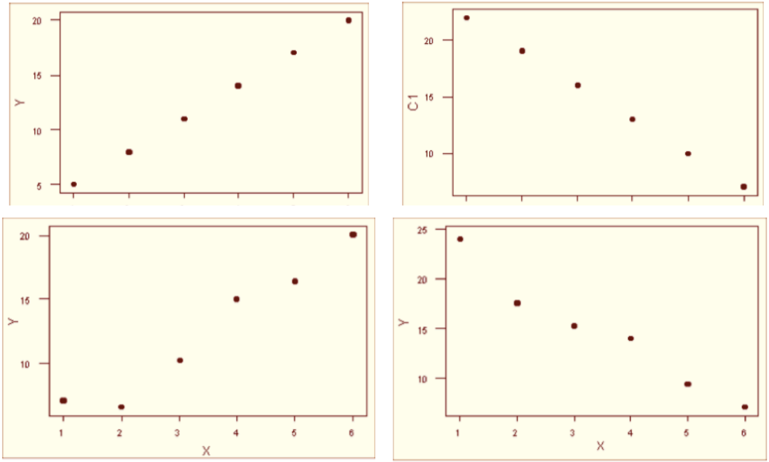
Designamos de coeficiente de correlação linear

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

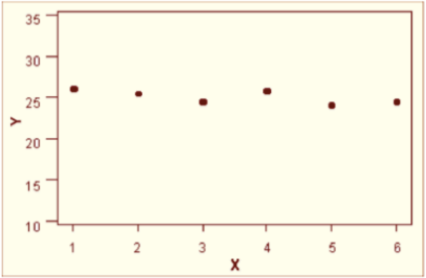
Este coeficiente é uma medida do grau de dependência linear entre as duas variáveis, X e Y.

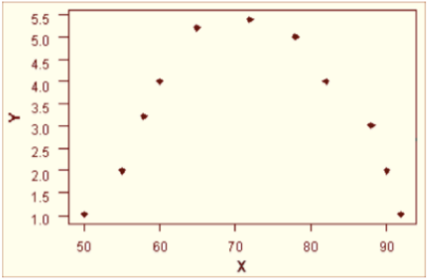
- $-1 \leq r \leq 1$;
- $r = 1$: relação linear perfeita (e positiva) entre X e Y;
- $r = 0$: inexistência de relação linear entre X e Y;
- $r = -1$: relação linear perfeita (e negativa) entre X e Y;
- $r > 0$: relação linear positiva entre X e Y;
- $r < 0$: relação linear negativa entre X e Y.

$r = 1$ $\qquad$ $r = -1$



$0 < r < 1$

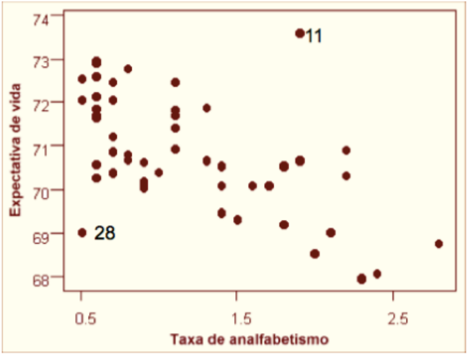

35
30
25
20
15
10
Y
1
2
3
4
5
6
X


5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Y
50
60
70
80
90
X

$$r = 0$$

Exemplo: Considere as duas variáveis abaixo observadas em 50 estados norte-americanos.

Y: expectativa de vida
X: taxa de analfabetismo



Observações: Quanto maior é a taxa de analfabetismo, menor é a expectativa de vida, e observamos ainda a existência de uma tendência linear entre as variáveis.

Exercício: Calcule o coeficiente de correlação entre X e Y, sabendo que:

$$\overline{y} = 70,88; \quad \overline{x} = 1,17; \quad \sum_{i=1}^{n} x_i y_i = 4122,8$$

$$\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 88,247; \quad \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 18,173$$

$$r = \frac{4122,8 - 50 \times 1,17 \times 70,88}{\sqrt{88,247 \times 18,173}} = \frac{-23,68}{40,047} = -0,59$$

O Modelo de regressão linear simples (MRLS)

$$Y = E(Y|X = x) + \epsilon = \alpha + \beta x + \epsilon$$

Y - variável explicada ou dependente (aleatória)

X - variável explicativa ou independente medida sem erro (não aleatória)

α - coeficiente de regressão, que representa o intercepto (parâmetro desconhecido do modelo -> a estimar)

β - coeficiente de regressão, que representa o declive (inclinação) (parâmetro desconhecido do modelo -> a estimar)

ϵ - erro aleatório ou estocástico, onde se procuram incluir todas as influências no comportamento da variável Y que não podem ser explicadas linearmente pelo comportamento da variável X;

Dadas n observações da variável X: $x_1, x_2, \ldots, x_n$, obtemos n v.a.'s $Y_1, Y_2, \ldots, Y_n$ satisfazendo a equação,

$$Y_i = E(Y|X = x_i) + \epsilon_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

Assume-se que as v.a.'s ϵ_i são v.a.'s independentes com média zero, $E(\epsilon_i|x) = 0$, e variância σ^2, $Var(\epsilon_i|x) = \sigma^2$.

Logo,

$$E(Y_i|X = x_i)) = \mu_{Y_i} = \alpha + \beta x_i \quad \text{e} \quad Var(Y_i|X = x_i) = \sigma^2$$

Recolhida uma amostra de n indivíduos, teremos n pares de valores (x_i, y_i), $i = 1, 2, \ldots, n$, que devem satisfazer o seguinte modelo,

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Temos n equações e $n + 2$ incógnitas $(\alpha, \beta, \epsilon_1, \ldots, \epsilon_n)$, por isso precisamos de introduzir um critério que permita encontrar α e β .

Método dos mínimos Quadrados (MMQ)

Encontrar os valores de α e β que minimizam a soma dos quadrados dos erros (ou desvios ou resíduos), dados por

$$\epsilon_i = y_i - (\alpha + \beta x_i)$$

Obtemos então, a quantidade de informação perdida pelo modelo ou soma dos quadrados dos resíduos

$$SQ(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Derivando em relação a α e β obtemos o sistema

$$
\begin{cases}
\left.\dfrac{\partial SQ(\alpha,\beta)}{\partial \alpha}\right|_{\alpha=\hat{\alpha}} = 0 \\[2ex]
\left.\dfrac{\partial SQ(\alpha,\beta)}{\partial \beta}\right|_{\beta=\hat{\beta}} = 0
\end{cases}
\Leftrightarrow
\begin{cases}
\displaystyle\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \\
\displaystyle\sum_{i=1}^{n}x_i(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0
\end{cases}
$$

$$
\Leftrightarrow
\begin{cases}
\displaystyle\sum_{i=1}^{n} y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} x_i \\
\displaystyle\sum_{i=1}^{n} x_i y_i = \hat{\alpha}\sum_{i=1}^{n} x_i + \hat{\beta}\sum_{i=1}^{n} x_i^2
\end{cases}
\Leftrightarrow
\begin{cases}
\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \\
\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}
\end{cases}
,
$$

onde $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ e $\overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$

Reta de regressão estimada

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

Definindo

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \qquad S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

obtemos

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \quad \text{e} \quad \hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

Interpretação das estimativas $\hat{\alpha}$ e $\hat{\beta}$

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

$x = 0$: $\hat{y} = \hat{\alpha}$;

$x \to x + 1$: $\Delta\hat{y} = \hat{\alpha} + \hat{\beta}(x + 1) - (\hat{\alpha} + \hat{\beta}x) = \hat{\beta}$

Logo, $\hat{\alpha}$ é o ponto onde a reta corta o eixo das ordenadas e pode ser interpretável ou não.

$\hat{\beta}$ é o coeficiente angular, e representa o quanto varia a média de Y para um aumento de uma unidade da variável X.

Nota:

Tendo em conta a notação apresentada, o coeficiente de correlação simples pode ser escrito do seguinte modo

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Previsão

Uma aplicação muito importante de um modelo de regressão é a previsão de novas ou futuras observações de Y, $(Y_f(x))$ correspondente a um dado valor da variável explicativa X, x_f, então o estimador será

$$\hat{Y}_f = \hat{y}_f = \hat{\alpha} + \hat{\beta} x_f.$$

Exemplo (pág. 458) Um psicólogo esta investigando a existência de um relação linear entre o tempo que um indivíduo leva a reagir a um estímulo visual (Y) e a respetiva idade (X), para indivíduos com idades compreendidas no intervalo $[20, 40]$. Os resultados observados permitiram obter:

$$n = 20 \quad \sum y_i = 2150 \quad \sum x_i = 600 \quad \sum x_i y_i = 65400$$

$$\overline{y} = 107,50 \quad \overline{x} = 30 \quad \sum x_i^2 = 19000$$

Obtenha a equação do modelo ajustado, interprete as estimativas obtidas e estime o tempo médio de reação para um indivíduo de 25 anos, e para 45 anos?

Resolução:

$S_{xy} = 65400 - 20 \times 30 \times 107,50 = 900$

$S_{xx} = 19000 - 20 \times 30^2 = 1000$

logo

$\hat{\beta} = \frac{900}{1000} = 0,9$

$\hat{\alpha} = 107,50 - 0,9 \times 30 = 80,50$

Equação do modelo ajustado: $\hat{y}_i = 80,50 + 0,9x_i, \ i = 1,2,\ldots 20$

Interpretação: $\hat{\alpha} = 80,50$ – tempo de reação para um recém-nascido (inadequação do modelo)
$\hat{\beta} = 0,9$ – por cada ano de envelhecimento das pessoas, o tempo médio de reação aumenta 0,9 unidades.

Previsão: $\hat{y}(25) = 80,50 + 0,9 \times 25 = 103$
$\hat{y}(45)$ – $45 \notin [20,40]$, logo não é possível determinar $\hat{y}(45)$.

Resíduo

$$e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}x_i$$

• Se os resíduos forem pequenos temos uma indicação de que o modelo está produzindo bons resultados.

Estimador de $\sigma^2 = Var(\epsilon|X)$

Para obtermos um estimador não enviesado de σ^2, analisamos a dispersão em torno da reta de regressão - Variação não explicada/Residual

$$SQRes = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 \quad \text{(Soma dos quadrados dos resíduos).}$$

Como $E(SQRes) = (n-2)\sigma^2$, então um estimador não enviesado de σ^2 é

$$\hat{\sigma}^2 = QMRes = \frac{SQRes}{n-2}$$

Definindo a Variação Total, como sendo a dispersão em torno de $\overline{y}$

$$SQTot = \sum_{i=1}^{n}(y_i - \overline{y})^2 = S_{yy} \quad \text{(Soma de quadrados totais)}$$

Prova-se que:

$$SQRes = S_{yy} - \hat{\beta} S_{xy}$$

Propriedades dos estimadores de mínimos quadrados

Pressupostos do modelo

$$Y_i = E(Y|x_i) + \epsilon_i = \alpha + \beta x_i, \quad i = 1, 2, \ldots, n$$

- A variável explicativa X é controlada pelo experimentador
- O MRLS está especificado de forma correta
- Os erros são não correlacionados
- $E(\epsilon|X) = 0$ e $Var(\epsilon|X) = \sigma^2$
- Os erros têm distribuição normal, isto é, $\epsilon_i \sim N(0, \sigma^2)$ o que implica que $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$.

Estimador $\hat{\beta}$

Prova-se que

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right) \iff \frac{\hat{\beta} - \beta}{\sigma}\sqrt{S_{xx}} \sim N(0,1)$$

Estimador $\hat{\alpha}$

Prova-se que

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum x_i^2}{nS_{xx}}\right) \iff \frac{\hat{\alpha} - \alpha}{\sigma}\sqrt{\frac{nS_{xx}}{\sum x_i^2}} \sim N(0,1)$$

Intervalos de confiança para α e β

Sendo $\hat{\sigma}^2 = QMRes$,

$$(\hat{\beta} - \beta)\sqrt{\frac{S_{xx}}{QMRes}} \sim t(n-2) \quad (*_1)$$

e

$$(\hat{\alpha} - \alpha)\sqrt{\frac{nS_{xx}}{QMRes\sum x_i^2}} \sim t(n-2) \quad (*_2)$$

e tendo em conta que $\sum x_i^2 = S_{xx} + n\overline{x}^2$, obtemos os intervalos de confiança a γ% de confiança para α e β, respectivamente:

$$IC(\alpha;\gamma) = \left(\hat{\alpha} \pm t_\gamma(n-2)\sqrt{QMRes\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}\right),$$

$$IC(\beta;\gamma) = \left(\hat{\beta} \pm t_\gamma(n-2)\sqrt{\frac{QMRes}{S_{xx}}}\right),$$

Voltando ao Exemplo: Obtenha os intervalos de confiança a 95% para os parâmetros da regressão.

$$
\begin{aligned}
IC(\alpha;\gamma) &= \left(80,50 \pm 2,101\sqrt{31,278 \times \left[\frac{1}{20} + \frac{30^2}{1000}\right]}\right) = (69,05; 91,95) \\
IC(\beta;\gamma) &= \left(0,90 \pm 2,101\sqrt{\frac{31,278}{1000}}\right) = (0,60; 1,20)
\end{aligned}
$$

Intervalo de confiança para $E(Y|x)$ e intervalo de predição

O interesse consiste em estimar um intervalo de confiança para

$$E(Y|X = x_i) = \mu(x_i) = \alpha + \beta x_i.$$

Um estimador pontual de $\mu(x_i)$ é

$$\widehat{\mu(x_i)} = \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

Mostra-se que:

$$T = \frac{\widehat{\mu(x_i)} - \mu(x_i)}{\sqrt{QMres\left[\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}\right]}} \sim t(n-2)$$

Intervalo de confiança a $\gamma\%$ para $\mu(x_i)$

$$IC(\mu(x_i);\gamma) = \left(\hat{y}_i \pm t_\gamma(n-2)\sqrt{QMres\left[\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}\right]}\right)$$

Intervalo de predição para uma resposta futura

Vimos que a previsão de novas ou futuras observações de Y, $(Y_f(x))$ correspondente a um dado valor da variável explicativa X, x_f, é uma aplicação muito importante de um MRLS. O estimador de Y_f é então,

$$\hat{Y}_f = \hat{y}_f = \hat{\alpha} + \hat{\beta}x_f$$

e o erro de previsão, $e_f = Y_f - \hat{Y}_f$. Logo, um intervalo de predição com $\gamma\%$ de confiança para uma futura observação é dado por:

$$IP(Y_f;\gamma) = \left(\hat{y}_f \pm t_\gamma(n-2)\sqrt{QMres\left[1 + \frac{1}{n} + \frac{(x_f - \overline{x})^2}{S_{xx}}\right]}\right)$$

Voltando ao Exemplo: Obtenha o intervalo de confiança a 95% para o tempo médio de reação de um paciente com 28 anos e o intervalo de predição a 95% de confiança para as futuras observações.

A estimativa pontual é:

$$\hat{y}(28) = 80,5 + 0,9 \times 28 = 105,7$$

Logo,

$$\begin{aligned} IC(\mu(28);\gamma) &= \left(105,7 \pm 2,101\sqrt{31,278 \times \left[\frac{1}{20} + \frac{(28-30)^2}{1000}\right]}\right) \\ &= (103,0; 108,4) \end{aligned}$$

e

$$\begin{aligned} IP(Y_f;\gamma) &= \left(105,7 \pm 2,101\sqrt{31,278 \times \left[1 + \frac{1}{20} + \frac{(28-30)^2}{1000}\right]}\right) \\ &= (93,6; 117,8) \end{aligned}$$

Teste de Hipóteses para α e β

As estatísticas $(*_1)$ e $(*_2)$ podem também ser utilizadas para realizar testes de hipóteses bilaterais sobre os parâmetros do modelo. Assim,

$$H_0 : \beta = \beta_0 \quad \textit{versus} \quad H_1 : \beta \neq \beta_0$$
$$H_0 : \alpha = \alpha_0 \quad \textit{versus} \quad H_1 : \alpha \neq \alpha_0$$

Teste de significância do MRLS via análise de variância

Neste caso, a hipótese a testar é

$$H_0 : \beta = 0 \quad \textit{versus} \quad H_1 : \beta \neq 0$$

ou seja, as hipóteses a testar são:

H_0: não existe relação linear entre X e Y *versus* H_1: existe relação linear entre X e Y.

Para este teste podemos utilizar as técnicas de análise de variância.

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SQTot}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SQReg}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SQRes}}$$

$$SQTot = SQReg + SQRes,$$

Sendo,

$$SQReg = \hat{\beta}^2 S_{xx} = \hat{\beta} S_{xy} \qquad SQRes = S_{yy} - \hat{\beta} S_{xy}$$

Tabela de ANOVA para o MRLS

F.V.	g.l.	SQ	QM	F
Regressão	1	SQReg	QMReg=SQReg	QMReg/QMRes
Resíduo	$n-2$	SQRes	QMRes=$SQRes/(n-2)$	
Total	$n-1$	SQTot	QMTot=$SQTot/(n-1)$	

Sob a validade de H_0, a estatística

$$F = QMreg/QMRes \sim F_{(1,n-2)},$$

sendo a região crítica

$$RC = (c, +\infty), \quad P(F_{(1,n-2)} > c) = \alpha.$$

Notas:

- Utilizando a estatística $(*_1)$ e a distribuição $t(n-2)$, obteríamos uma região crítica dada por uma reunião de caudas.
- Sob a validade de H_0,

$$(*_1)^2 = \left[(\hat{\beta} - 0)\sqrt{\frac{S_{xx}}{QMRes}}\right]^2 = \frac{\hat{\beta}^2 S_{xx}}{QMRes} = \frac{SQReg}{QMRes} = \frac{QMReg}{QMRes}.$$

Voltando ao Exemplo: Teste a significância do modelo e construa a tabela ANOVA (considere $\alpha = 5\%$).

• Tabela ANOVA

F.V.	g.l.	SQ	QM	F
Regressão	1	810	810	25,90
Resíduo	18	563	31,28	
Total	19	1373	72,26	

• $H_0 : \beta = 0 \quad versus \quad H_1 : \beta \neq 0$
• $F = QMreg/QMRes \sim F_{(1,18)}$
• $RC = (4,41; +\infty)$
• $F_{obs} = 25,90 \in RC$, logo a decisão é a de rejeitar H_0 ao n.s. de 5%, isto é, existem evidências de que existe uma relação linear entre a idade do indivíduo e o tempo de reação a um estímulo visual.

Adequação do MRLS

Análise de resíduos

A análise dos resíduos $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$ é importante para averiguarmos a adequação do ajuste. A construção do gráfico dos resíduos padronizados:

$$\frac{e_i}{\hat{\sigma}^2} = \frac{e_i}{QMRes},$$

dá-nos uma indicação da qualidade do ajuste do modelo. Assim, se os pontos estiverem distribuídos dentro do intervalo $[-2, 2]$, temos indicação de que o modelo está bem ajustado. Se houver pontos acima de 2 ou abaixo de -2, podemos estar na presença de pontos aberrantes.

Coeficiente de determinação

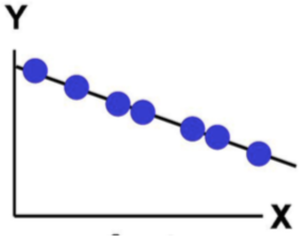
O quociente entre SQReg e SQTot dá-nos uma medida da proporção da variação total que é explicada pelo MRLS. A esta medida dá-se o nome de coeficiente de determinação (r^2),

$$r^2 = \frac{SQReg}{SQTot} = 1 - \frac{SQRes}{SQTot}$$

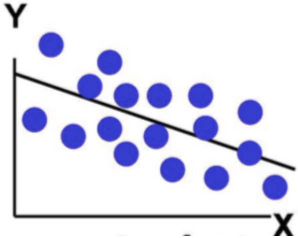
Este coeficiente pode ser utilizado como uma medida da qualidade do ajustamento, ou como medida da confiança depositada na equação de regressão como instrumento de previsão, e representa a porcentagem da variação total que é explicada pelo MRLS. Note-se que o ajustamento será tanto melhor quanto mais pequeno for SQRes (e portanto, maior for SQReg) relativamente a SQTot.

- $0 \leq r^2 \leq 1$;
- $r^2 \approx 0$ – modelo linear muito pouco adequado;
- $r^2 \approx 1$ – modelo linear bastante adequado.

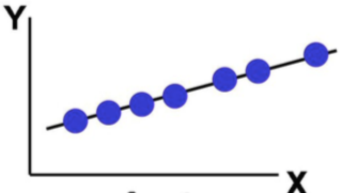
Exemplos



$r^2 = 1$



$0 < r^2 < 1$



$r^2 = 1$

Voltando ao Exemplo: Calcule e interprete o coeficiente de determinação, sabendo que $\sum y_i^2 = 232498$.

- $r^2 = \dfrac{\hat{\beta} S_{xy}}{S_{yy}} = \dfrac{0,9 \times 900}{232498 - 20 \times 107,50^2} = \dfrac{810}{1373} = 0,59 \rightarrow 59\%$

Interpretação: 59% da variação no tempo de reação está relacionada linearmente com a idade do indivíduo, sendo os restantes 41% da variação resultantes de outros fatores não considerados (sexo, acuidade visual,...).

Alguns abusos no modelo de regressão

Seleccção de variável explicativa: É possível desenvolver uma relação estatisticamente significativa entre a variável resposta (Y) e a variável explicativa (X) que não faça sentido na prática.

Extrapolação: A relação linear assumida para as variáveis resposta e explicativa não pode ser estendida para fora do domínio de actuação dos dados observados, a não ser que haja informação adicional sobre a validade do modelo para esse domínio estendido.