9 Regressão linear simples

José Luis Duarte Ribeiro
Carla ten Caten

COMENTÁRIOS INICIAIS

Em muitos problemas há duas ou mais variáveis que são relacionadas e pode ser importante modelar essa relação. Por exemplo, a resistência à abrasão de um composto de borracha pode depender da quantidade de óleo adicionada à mistura. Assim, é possível construir um modelo relacionando resistência à abrasão com quantidade de óleo, e então pode-se usar esse modelo para fins de otimização e controle de processo.

Outro exemplo, as vendas de um produto podem estar relacionadas ao valor gasto em marketing com esse produto. Assim, é possível construir um modelo relacionando vendas à gastos com marketing, e então pode-se usar esse modelo para fins previsão de vendas.

Em geral vamos supor que há uma variável dependente (ou variável de resposta) Y que depende de k variáveis independentes (ou variáveis regressoras) $X_1, ..., X_k$. A relação entre essas variáveis será descrita por um modelo matemático, chamado modelo de regressão, o qual é definido (ajustado) a um conjunto de dados.

Algumas vezes a relação funcional entre Y e $X_1, ..., X_k$ é conhecida exatamente. Outras vezes o pesquisador deverá buscar o modelo apropriado testando diferentes funções. Modelos polinomiais são largamente utilizados como uma função aproximada da verdadeira relação entre Y e X, e por isso serão descritos no capítulo 10.

Modelos de regressão são usados com freqüência na análise de dados provenientes de experimentos não planejados (observações de um fenômeno não controlado ou dados históricos).

Mas a análise de regressão também é muito útil no caso de experimentos planejados que incluem fatores a níveis contínuos. Nesse caso a análise de variância é usada para identificar os fatores significativos, e a seguir a análise de regressão é usada para construir um modelo que incorpore esses fatores.

CORRELAÇÃO

Para uma amostra de n pares de valores (x,y) o coeficiente de correlação $\underline{r}$ fornece uma medida da relação linear que existe entre duas variáveis aleatórias X e Y.

Figura 39 - Gráfico de dispersão

O valor de r é calculado como:

Eq 160: $$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}}$$

Desvio-padrão de X

Eq 161: $$S_{XX} = \sum x_i^2 - \left(\sum x_i\right)^2 / n$$

Desvio-padrão de Y

Eq 162: $$S_{YY} = \sum y_i^2 - \left(\sum y_i\right)^2 / n$$

Covariância de X,Y:

Eq 163: $$S_{XY} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right) / n$$

Para uma interpretação adequada do coeficiente de correlação, X e Y deveriam ser variáveis aleatórias, ao contrário do que acontece nos problemas de regressão, onde Y é aleatória, mas X é considerada uma variável fixa.

Mesmo assim, é prática comum calcular r em quase todos os casos, isto é, com X aleatória ou não. O coeficiente de correlação linear "r" mede a intensidade da relação linear entre duas variáveis

Pode ser demonstrado que $-1 \le r \le 1$, onde $r = +1$ ou $r = -1$ correspondem ao caso de uma relação linear perfeita entre X e Y, enquanto que $r = 0$ indica nenhuma relação, ou seja:

valores de "r" próximos de +1 indicam uma forte correlação positiva entre x e y

valores de "r" próximos de -1 indicam uma forte correlação negativa entre x e y

valores de "r" próximos de 0 indicam uma fraca correlação entre x e y

Deve-se ter em conta que r é uma medida da relação *linear* entre as duas variáveis e não tem sentido quando a relação é não linear. Além disso, o pesquisador deve ter em mente que a existência de uma correlação entre duas variáveis não implica necessariamente na

existência de um relacionamento de causa e efeito entre elas.

Exemplo 9.1

Após uma regulagem eletrônica um veículo apresenta um rendimento ideal no que tange a rendimento de combustível. Contudo, com o passar do tempo esse rendimento vai se degradando. Os dados que aparecem na Tabela 11 representam o rendimento medido mês a mês após a regulagem. Calcule o coeficiente de correlação.

X: meses após a regulagem	1	2	3	4	5	6
Y: rendimento	10,7	10,9	10,8	9,3	9,5	10,4
X: meses após a regulagem	7	8	9	10	11	12
Y: rendimento	9,0	9,3	7,6	7,6	7,9	7,7

Tabela 11 - Valores de rendimento de combustível

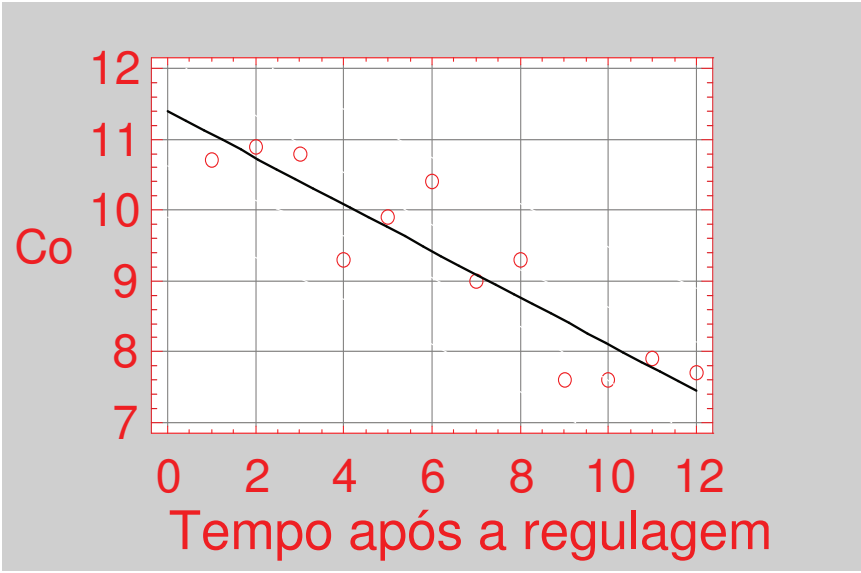


Figura 40 - Valores observados do rendimento em função do tempo após a regulagem.

Para o exemplo do rendimento de combustível, teríamos:

Cálculos iniciais

Meses(X)	Rendimento(Y)	X^2	Y^2	X*Y
1	10,7	1	114,49	10,7
2	10,9	4	118,81	21,8
3	10,8	9	116,64	32,4
4	9,3	16	86,49	37,2
5	9,5	25	90,25	47,5
6	10,4	36	108,16	62,4
7	9	49	81	63
8	9,3	64	86,49	74,4
9	7,6	81	57,76	68,4
10	7,6	100	57,76	76
11	7,9	121	62,41	86,9
12	7,7	144	59,29	92,4
78	**110,7**	**650**	**1039,55**	**673,1**
6,5	**9,225**			

$\Sigma x_i = \ 78{,}00; \Sigma x_i^2 = \ 650{,}00; \ \overline{X} = 6{,}50$

$\Sigma y_i = 110{,}70; \Sigma y_i^2 = 1039{,}55; \overline{Y} = 9{,}225$

Desvio-padrão de X

$$S_{XX} = \sum x_i^2 - \left(\sum x_i\right)^2 \Big/ n = 650 - (78)^2 / 12 = 143{,}00$$

Desvio-padrão de Y

$$S_{YY} = \sum y_i^2 - \left(\sum y_i\right)^2 \Big/ n = 1039{,}55 - (110{,}70)^2 / 12 = 18{,}34$$

Covariância de X,Y:

$$S_{XY} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right) \Big/ n = 673{,}1 - (78 \times 110{,}70) / 12 = -46{,}4$$

Coeficiente de correlação

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{-46{,}45}{\sqrt{143{,}00 \text{ x } 18{,}34}} = -0{,}907$$

Interpretação:

Existe uma correlação linear inversa na amostra entre meses após a regulagem e rendimento. A intensidade desta correlação é forte.

TESTE DE HIPÓTESE PARA O COEFICIENTE DE CORRELAÇÃO

A hipótese da existência de uma relação entre X e Y, pode ser formulada usando-se:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

onde a letra ρ é usada para representar o valor populacional do coeficiente de correlação. Pode ser demonstrado que o valor de t pode

ser calculado usando:

Eq 164: $$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Assim a hipótese da existência de uma relação entre X e Y pode ser verificada diretamente a partir do valor amostral do coeficiente de correlação. Como sempre a hipótese nula será rejeitada se o valor calculado for maior que o tabelado, ou seja, se:

Eq 165: $$|t| > t_{\alpha/2,n-2}$$

Para o exemplo em estudo tem-se:

$$t = \frac{-0{,}907\sqrt{12-2}}{\sqrt{1-(-0{,}907)^2}} = |-6{,}82| > t_{0{,}025;10} = 2{,}228 \Rightarrow \text{rejeita - se } H_0,$$

ou seja, descarta-se a hipótese nula e conclui-se que existe correlação entre as variáveis estudadas.

REGRESSÃO LINEAR SIMPLES

A regressão linear simples estima uma equação matemática (ou modelo) que dado o valor de *X* (variável independente), prevê o valor de *Y* (variável dependente). É dito relação linear simples, pois supõe-se tendência linear entre as variáveis e simples por ser uma única variável independente

Seja que existam dados coletados (pares de valores) associando uma variável de resposta Y (variável dependente) com uma variável regressora X (variável independente). E suponha que a relação entre Y e X seja aproximadamente linear. Então o valor esperado de Y para cada valor de X virá dado por:

Eq 166: $$E(Y/X) = \beta_0 + \beta_1 X$$

onde os parâmetros da relação linear, β_0 e β_1, são desconhecidos. Vamos supor que cada observação Y possa ser descrita pelo modelo:

Eq 167 $$Y = \beta_0 + \beta_1 X + \varepsilon$$

onde ε é o erro aleatório, com média 0 e variância σ^2. A Eq 167 é chamada de modelo de regressão linear simples. Nesta equação, o coeficiente β_0 é a *interseção* (valor de Y para $X = 0$) enquanto que β_1 é a *inclinação* da reta, que pode ser positiva, negativa ou nula. A inclinação da reta representa o quanto *Y* varia para cada unidade da variável *X*.

Se há n pares de dados $(y_1, x_1), ..., (y_n, x_n)$ é possível estimar os parâmetros β_0 e β_1 usando o método dos Mínimos Quadrados, o qual busca minimizar:

Eq 168: $$L = \Sigma (y_i - b_0 - b_1 x_i)^2$$

onde b_0 e b_1 são estimativas amostrais de β_0 e β_1. O uso do método

conduz as seguintes estimativas:

Eq 169: b1 = *SXY / SXX*

Eq 170: $b_0 = \overline{Y} - b_1 \overline{X}$

Exemplo 9.2

Usando os dados do problema do rendimento de combustível, obtenha as estimativas para os parâmetros b_0 e b_1 e a equação da reta de regressão.

Cálculos iniciais

$\Sigma x_i = 78{,}00 \quad \Sigma x_i^2 = 650{,}00 \quad \overline{X} = 6{,}50$

$\Sigma y_i = 110{,}70 \quad \Sigma y_i^2 = 1039{,}55 \quad \overline{Y} = 9{,}225$

$$S_{XX} = \sum x_i^2 - \left(\sum x_i\right)^2 / n = 143{,}00$$

$$S_{YY} = \sum y_i^2 - \left(\sum y_i\right)^2 / n = 18{,}34$$

$$S_{XY} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n = -46{,}45$$

Estimativa dos parâmetros:

$b_1 = S_{XY} / S_{XX} = -46{,}45 / 143{,}00 = -0{,}325$

$b_0 = \overline{Y} - b_1 \overline{X} = 9{,}225 - (-0{,}325)\ 6{,}50 = 11{,}34$

Equação de regressão

$Y = 11{,}34 - 0{,}325\ X$

RELAÇÃO ENTRE O COEFICIENTE DE CORRELAÇÃO E A REGRESSÃO

O valor de r é um valor sem dimensão, que apenas fornece uma idéia da relação linear entre duas variáveis. No caso de regressão, além de se ter uma idéia da relação entre as duas variáveis, também se encontra uma equação que pode ser usada para fornecer estimativas.

Pode ser demonstrado que existe a seguinte relação:

Eq 171: $$S^2 = \frac{n-1}{n-2}\left(1 - r^2\right)S_y^2$$

onde S^2 é a variância dos desvios em relação ao modelo, e S_y^2 é a variância dos valores de Y. Se n é grande, temos:

Eq 172: $$S^2 \cong \left(1 - r^2\right)S_y^2$$

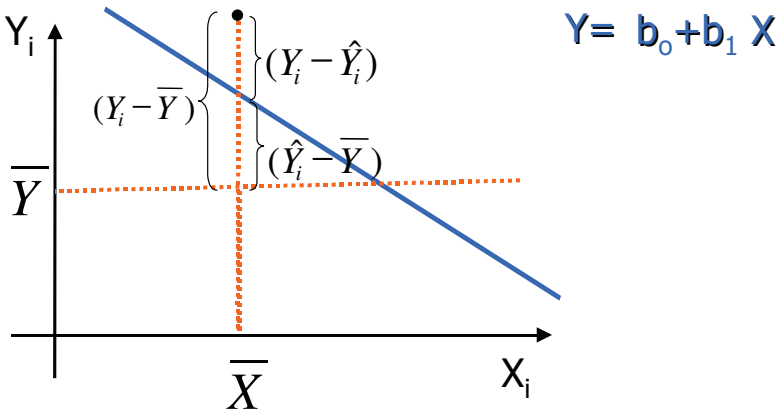


Figura 41 - Decomposição dos resíduos

Nessa forma observamos que r^2 *equivale a proporção da variabilidade dos valores de* Y *que pode ser atribuída à regressão com a variável* X.

r^2 é conhecido como coeficiente de Determinação. Para o exemplo analisado resultou $r = (-0{,}907)^2 = 0{,}82$, ou seja, 82% da variabilidade nos resultados de rendimento de combustível pode ser devida ao tempo decorrido após a regulagem e 18% da variabilidade total é devido a outros fatores que não foram investigados.

Também pode ser demonstrado que:

Eq 173: $r = b_1 S_X / S_Y$

Assim, dado um conjunto de pares (x,y), conhecida a inclinação b_1, é possível calcular o coeficiente de correlação r, ou vice-versa.

VARIÂNCIA DOS ESTIMADORES

Para verificar a precisão das estimativas, determinar intervalos de confiança e testar hipóteses é importante conhecer a variância dos estimadores. Pode ser demonstrado que uma estimativa da variância residual, σ^2, vem dada por

Eq 174: $S^2 = SQR / (n-2)$

onde:

Eq 175: $$SQR = \sum [y_i - (b_0 + b_1 x_i)]^2 = S_{YY} - b_1 S_{XY}$$

E a partir de σ^2 obtém-se as estimativas das variâncias de b_1 e b_0:

Eq 176: $$S^2_{b1} = S^2 / S_{XX}$$

Eq 177: $$S^2_{b0} = S^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right]$$

INTERVALOS DE

Como os resíduos de Y supostamente seguem a distribuição Normal, e

CONFIANÇA E TESTES DE HIPÓTESE

como os valores de b_0 e b_1 são funções lineares de Y, é possível demonstrar que:

$$b_0 \rightarrow N\left(\beta_0, \sigma^2_{b0}\right)$$
$$b_1 \rightarrow N\left(\beta_1, \sigma^2_{b1}\right)$$

Esses resultados podem ser usados em testes de hipótese. Por exemplo, se a hipótese é:

$$H_0 : \beta_1 = \beta_{10}$$
$$H_1 : \beta_1 \neq \beta_{10}$$

então calcula-se:

Eq 178: $Z = (b_1 - \beta_{10}) / \sigma_{b1}$

e, para um nível de probabilidade α, H_0 será rejeitada se resultar $|Z| > Z_{\alpha/2}$. Como em geral a variância S^2 não é conhecida, usa-se:

Eq 179: $t = (b_1 - \beta_{10}) / S_{b1}$

e nesse caso H_0 é rejeitada se $|t| > t_{\alpha/2,n-2}$. O intervalo de confiança para β_1 virá dado por

Eq 180: $$b_1 - t_{\alpha/2} S_{b1} < \beta_1 < b_1 + t_{\alpha/2} S_{b1}$$

Uma hipótese testada com freqüência é:

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

Isto é, testa-se se a inclinação é igual a zero, o que equivale a testar se existe uma relação entre Y e X. Usando a eq. (2) tem-se:

Eq 181: $t = b_1 / S_{b1}$

o qual deve ser comparado com o valor tabelado $t_{\alpha/2,n-2}$. Como sempre, H_0 será rejeitado se $|t| > t_{\alpha/2,n-2}$.

Exemplo 9.3

Usando os dados do problema do rendimento de combustível, obtenha as estimativas para a variância residual e para a variância dos parâmetros b_0 e b_1. Construa um intervalo de confiança para a inclinação b_1 e verifique a hipótese $H_0 : \beta_1 = 0$.

Estimativa das variâncias

$$SQR = S_{YY} - b_1 S_{XY} = 3{,}24$$

$$S^2 = SQR/(n-2) = 0{,}324 \quad ; \quad S = 0{,}569$$

$$S_{b1}^2 = S^2 / S_{XX} = 0{,}00227 \quad ; \quad S_{b1} = 0{,}0476$$

$$S_{b0}^2 = S^2\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right) = 0{,}123 \quad ; \quad S_{b0} = 0{,}351$$

Intervalo de confiança para $\mathbf{b_1}$

$t_{0,025;10} = 2{,}228$

$- 0{,}325 - 2{,}228\ (0{,}0476) < \beta_1 < - 0{,}325 + 2{,}228\ (0{,}0476)$

$- 0{,}431 < \beta_1 < - 0{,}219$

Como esse intervalo não inclui o zero, a hipótese $\beta 1 = 0$ é rejeitada, ou seja, existe uma relação entre o rendimento de combustível e o tempo decorrido após a regulagem.

PREVISÃO DE VALORES DE Y

A análise de regressão produz uma relação entre as variáveis consideradas, a qual pode ser usada para prever valores de *Y*.

Dado um certo valor de $X = x_0$, há dois tipos de previsão: previsão de um valor médio de *Y* e previsão de um valor individual de *Y*. Nos dois casos a estimativa pontual de *Y* é a mesma, mas a amplitude do intervalo de confiança é diferente. O intervalo de confiança é mais amplo para o caso de previsões de valores individuais.

Previsão de um valor médio de Y

A variância dos valores preditos irá depender não somente de S^2, mas também do valor de x_0. Isso acontece porque as previsões são mais precisas quando $x_0 \sim \overline{X}$ e menos precisas quando x_0 aproxima-se dos extremos investigados.

Pode ser demonstrado que a variância da previsão de um valor médio de *Y* vem dada por:

Eq 182:
$$S_{\overline{Y}p}^2 = S^2\left[\frac{1}{n} + \frac{(x_0 - \overline{X})^2}{S_{XX}}\right]$$

Como pode ser visto, a variância da previsão é mínima quando $\boldsymbol{x_0} = \overline{X}$ e aumenta quando $\boldsymbol{x_0}$ afasta-se de $\overline{X}$. Assim, o intervalo de confiança para a previsão de um valor médio virá dado por:

Eq 183:
$$\mu_Y = (b_0 + b_1 X_0) \pm t_{\alpha/2\ ;\ n-2}\left(S_{\overline{Y}\ p}\right)$$

Previsão de um valor individual de Y

A variância da previsão de valores individuais de *Y* segue o mesmo comportamento observado para os valores médios. Contudo, a variância é maior no caso de valores individuais.

Pode ser demonstrado que a variância da previsão de um valor individual de *Y* vem dada por:

Eq 184: $$S^2_{Yp} = S^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{X})^2}{S_{XX}}\right]$$

De modo que o intervalo de confiança para a previsão de um valor individual de Y é:

Eq 185: $$Y = (b_0 + b_1 X_0) \pm t_{\alpha/2\,;\,n-2}\left(S_{Y\,p}\right)$$

Exemplo 9.4

Usando os dados do problema do rendimento de combustível, obtenha os intervalos de confiança de 95% para a previsão de um valor médio e um valor individual de Y para um tempo $x_0 = 8$ meses.

$$(b_0 + b_1 x_0) = 8{,}74; \frac{(x_0-\overline{X})^2}{S_{XX}} = 0{,}0157$$

$$S^2_{\overline{Y}\,p} = 0{,}324\left[\frac{1}{12}+0{,}0157\right] = 0{,}0321 \quad ; \quad S_{\overline{Y}\,p} = 0{,}179$$

$$S^2_{Y\,p} = 0{,}324\left[1+\frac{1}{12}+0{,}0157\right] = 0{,}356 \quad ; \quad S_{Y\,p} = 0{,}597$$

Valor médio para $x_0 = 8$

$\mu_Y = 8{,}74 \pm 2{,}228 \,.\, (0{,}179)$

$\mu_Y = 8{,}74 \pm 0{,}399$

Valor individual para $x_0 = 8$

$Y = 8{,}74 \pm 2{,}228 \,.\, (0{,}597)$

$Y = 8{,}74 \pm 1{,}33$

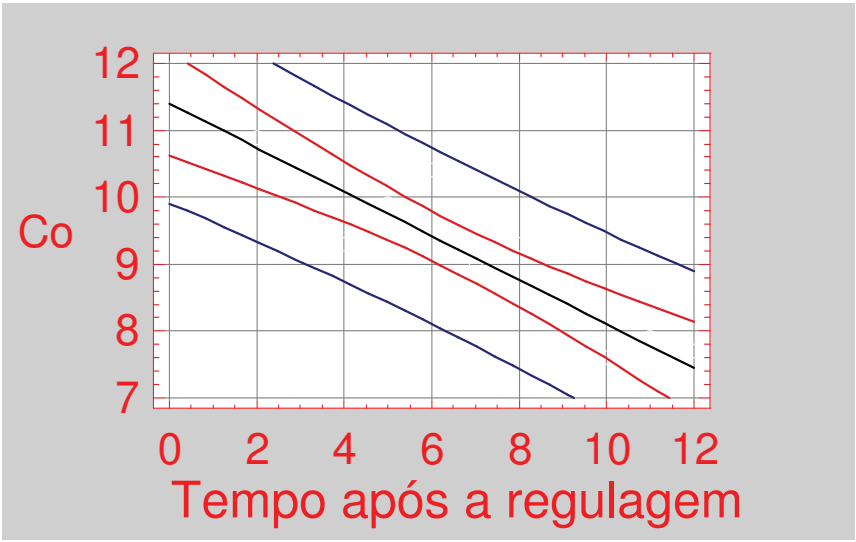


Figura 42 - Intervalo de Confiança de 95%

ANÁLISE DA VALIDADE

A adequação do ajuste e as suposições do modelo podem ser verificadas

DO MODELO

através de uma análise dos resíduos. Os resíduos padronizados são calculados como:

Eq 186:

$$R_i = \frac{y_i - (b_0 + b_1 x_i)}{S}$$

$$SQR = S_{YY} - b_1 S_{XY}$$

$$S^2 = SQR / n - 2$$

Adequação do ajuste

A adequação do ajuste é testada plotando os resíduos em função de X. Se o ajuste for bom, os resíduos seguirão um padrão aleatório. Caso contrário, alguma tendência curvilíneo será observada.

Na Figura 43, (a) representa uma situação onde o ajuste é adequado, enquanto que (b) representa uma situação onde o modelo linear não se ajusta bem aos dados.

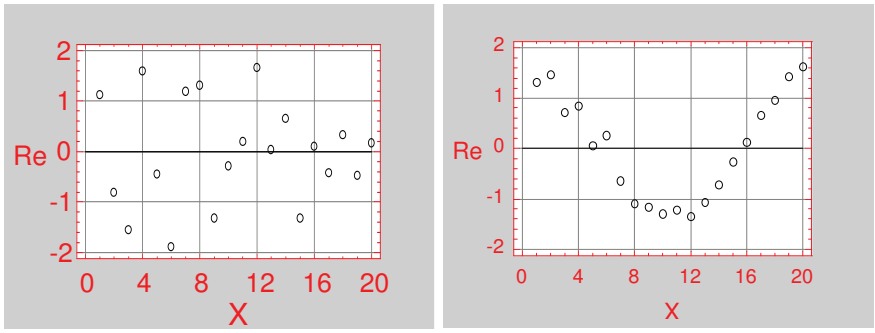


Figura 43 - Análise de resíduos.

(a) (b)

Se o modelo linear não fornece um bom ajuste, as vezes o problema pode ser contornado trabalhando-se com valores transformados de X *ou* Y, por exemplo,

Eq 187:

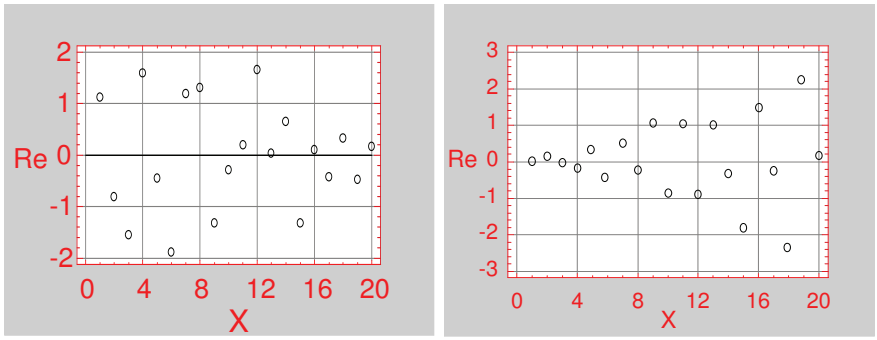
$$Y = b_0 + b_1 \sqrt{X}$$

$$Y = b_0 + b_1 X* \qquad \text{onde } X* = \sqrt{X}$$

Homogeneidade da variância

A suposição de homogeneidade da variância σ^2 ao longo de todo o intervalo de X também pode ser verificada analisando o gráfico de *Resíduos* $\times$ *X*.

A Figura 44 apresenta duas situações: (a) onde verifica-se a suposição de homogeneidade, e (b) onde essa suposição é violada.



(a) (b)

Figura 44 - Verificação da homogeneidade da variância.

Se a suposição de homogeneidade da variância é rejeitada, pode-se usar o método da regressão linear ponderada, onde se busca os valores de β_0 e β_1 que minimizam

Eq 188: $L = \Sigma\, w_i\, (y_i - (b_0 + b_1\, x_i))^2$

Nesse caso, os pesos w_i são inversamente proporcionais à variância.

Normalidade dos Resíduos

O teste da normalidade da distribuição dos resíduos pode ser feito plotando-se os resíduos em papel de probabilidade ou utilizando testes analíticos de normalidade, como o teste do Chi-quadrado ou o teste de Kolmorov-Smirnov.

Se a suposição de normalidade é rejeitada, muitas vezes uma transformação matemática nos valores de X e Y (logaritmo, inverso, raiz quadrada) irá gerar valores transformados com resíduos normalmente distribuídos.

Então o problema é analisado no espaço das variáveis transformadas e ao final retorna-se ao espaço original.

INTERVALO DE VARIAÇÃO PARA X

A variância da inclinação b_1 aumenta quando se reduz o intervalo de variação de X. Se o intervalo é pequeno, S_{b1} será grande e nesse caso será difícil rejeitar a hipótese $H_0 : b_1 = 0$. Em outras palavras, se a relação entre X e Y é medida em um intervalo reduzido de X, os parâmetros estimados não terão muito significado estatístico.

Se o objetivo é construir um modelo de regressão, deve-se coletar dados nos extremos do intervalo de X, ou seja, nos limites do interesse e viabilidade práticos ou nos limites em que se supõem válida a relação linear.

A ANÁLISE DE VARIÂNCIA E A REGRESSÃO

A análise de variância também é aplicável aos problemas de regressão. Na regressão simples, podemos decompor os resíduos da seguinte maneira:

Eq 189: $$(Y_i - \bar{Y}) = [y_i - (b_0 + b_1 X_i)] + [(b_0 + b_1 X_i) - \bar{Y}]$$

Elevando ao quadrado e somando, obtém-se:

Eq 190: $$\sum (Y_i - \bar{Y})^2 = \sum [y_i - (b_0 + b_1 X_i)]^2 + \sum [(b_0 + b_1 X_i) - \bar{Y}]^2$$

Uma vez que ;o produto cruzado resulta nulo. Essa equação também pode ser escrita como:

$S_{YY} = SQR + SQReg$

Cujos graus de liberdade valem respectivamente:

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

Assim, a média quadrada associada com o modelo de regressão e a média quadrada dos resíduos resultam:

$MQReg = SQReg / 1$

$MQR = SQR / (n - 2)$

E o teste F é feito comparando $MQReg$, com MQR, ou seja,

$F = MQReg / MQR$

A hipótese nula, $H_0 : \beta_1 = 0$, será rejeitada sempre que

$F > F_{\alpha, 1, n-2}$

A Tabela 12 apresenta a tabela ANOVA, contendo o formulário prático para o cálculo das Somas Quadradas e os demais desenvolvimentos até o teste F.

Tabela 12 - Tabela ANOVA para a análise de regressão.

Fonte de Variação	SQ	GDL	MQ	F
Regressão	SQReg = b_1 S_{XY}	1	MQReg	MQReg/MQR
Residual	SQR=S_{YY} - $b_1 S_{XY}$	n - 2	MQR	
Total	S_{YY}	n - 1		

Exemplo 9.5

Faça a análise de variância para o problema do rendimento de combustível e confirme a significância do modelo de regressão linear.

Solução:

Já tínhamos calculado as Somas Quadradas S_{YY} e SQR como:

$S_{YY} = 18,34; S_{XY} = - 46,45; b_1 = - 0,325$

$SQR = 3,24;$

Assim

$SQReg = b_1 S_{XY} = - 0,325 (- 46,45) = 15,10$

De modo que a ANOVA resulta conforme aparece na Tabela 13.

Fonte de Variação	SQ	GDL	MQ	F
Regressão	15,10	1	15,10	46,6
Residual	3,24	10	0,324	
Total	18,34	11		

Tabela 13 - Tabela ANOVA para o exemplo do combustível.

O valor de F calculado *(46,6)* é muito maior que o tabelado *(4,96)* e assim confirma-se a significância do modelo.

Nota: o coeficiente de determinação r^2 também pode ser calculado usando:

$$r^2 = \frac{SQ\,\mathrm{Re}\,g}{S_{YY}} = \frac{15{,}10}{18{,}34} = 0{,}82 \text{ ou } 82\%$$

DADOS ATÍPICOS

Algumas vezes, o conjunto de dados pode estar contaminado com alguns dados atípicos. Esses dados atípicos podem ser o resultado do efeito de algum fator externo ao estudo, ou podem ser simplesmente um erro de leitura e registro.

Existe um procedimento para testar a significância de um dado atípico. Este procedimento (ver Snedcor (1982)) está baseado na determinação de uma nova equação, com o dado atípico eliminado, seguido de um teste de hipótese comparando os valores preditos pela equação original com aqueles preditos pela nova equação.

Se o conjunto pode estar contaminado por vários dados atípicos, a solução será usar técnicas de regressão *robusta.* Neste tipo de análise, é dado um peso menor aqueles dados que se afastam do conjunto. Por exemplo, uma alternativa é minimizar

Eq 191: $$L = \Sigma\, w_i\, [y_i - (b_0 + b_1\, x_i)]^2$$

onde os pesos w_i são proporcionais ao inverso do resíduo R_i, e a solução é obtida após algumas iterações.

REGRESSÃO NÃO LINEAR SIMPLES

Se o ajuste linear é deficiente, muitas vezes é possível encontrar uma solução aproximada, e em geral satisfatória, utilizando uma transformação em X e/ou em Y.

Em forma genérica, teríamos:

Eq 192: $$f(y) = b_0 + b_1\, g(X) + \varepsilon$$

Eq 193: $$Y^* = b_0 + b_1\, X^* + \varepsilon$$

Os possíveis valores de $Y^* = f(y)$ seriam y, $1/y$, y^2, $\ln y$, etc. Igualmente, para $X^* = g(x)$ poderíamos usar x, $1/x$, x^2, $\ln x$, etc.

Uma vez definida a transformação, e confirmada em um gráfico de

dispersão a relação aproximadamente linear entre Y^* e X^*, poderia se usar o método apresentado anteriormente para obter-se as estimativas de β_0 e β_1.

Note-se que o método dos mínimos quadrados aplicado aos valores transformados, isto é, minimizando:

$$L = \Sigma [\, f(y_i) - (b_0 + b_i\, g(x_i))]^2$$

não vai fornecer os mesmos resultados que seriam obtidos minimizando:

$$L = \Sigma [\, y_i - h(x_i)]^2$$

onde $h(x)$ é uma função não linear de x. Contudo, as diferenças em geral são pequenas e não comprometem a análise.

Exercícios

Exercício 9.1

Em um processo químico a quantidade de sólidos depositada pode depender da concentração de um componente A que é adicionado à mistura. Ajuste um modelo de regressão linear aos dados que aparecem a seguir. Depois plote a reta de regressão e os valores observados

Conc.	0	0	0	2	2	2	4	4	4	6	6	6	8	8	8
Depos.	13,3	11,5	12,9	14,1	13,3	16,1	14,9	15,9	18,1	17,5	16,5	18,9	20,3	18,5	20,2

Exercício 9.2

Para os dados do exercício 9.1, calcule a variância residual e a variância dos parâmetros b_0 e b_1. Após construa um intervalo de confiança de 95% para a inclinação b_1 e verifique a hipótese $H_0: \beta_1 = 0$

Exercício 9.3

Calcule os resíduos padronizados $R_i = [Y_i - (b_0 + b_1 X_i)] / S$ para os dados do exercício 9.1. Em seguida, plote um gráfico de *Resíduos* $\times X$ e verifique se há evidências de falta de ajuste do modelo linear ou falta de homogeneidade da variância.

Exercício 9.4

Ainda em relação aos dados do exercício 9.1, calcule os intervalos de confiança para um valor médio e para um valor individual de Y usando $x_0 = 0$ e $x_0 = 8$.

Exercício 9.5

Um torno mecânico pode ser operado a diversas velocidades. Contudo, a qualidade do acabamento, ou seja, a rugosidade superficial, pode piorar com o aumento da velocidade de operação. Ajuste um modelo de regressão linear aos dados que aparecem a seguir e depois plote a reta de regressão e os valores observados.

Velocidade	3	3	3	6	6	6	9	9	9	12	12	12
Rugosidade	26,0	21,5	33,5	36,0	27,5	37,0	41,5	28,0	39,5	43,0	37,0	50,5

Exercício 9.6

Para os dados do exercício 9.5, calcule a variância residual e a variância dos parâmetros b_0 e b_1. Após construa um intervalo de confiança de 95% para a inclinação b_1 e verifique a hipótese da existência de uma relação entre velocidade e rugosidade superficial.

Exercício 9.7

Faça a análise de variância para os dados do exercício 9.5 e confirme a significância do modelo de regressão linear. Em seguida calcule o valor do coeficiente de determinação e indique qual o significado técnico desse coeficiente para o problema em questão.

Exercício 9.8

O gerente de uma indústria localizada em um país tropical suspeita que há uma correlação entre a temperatura do dia e produtividade. Dados coletados aleatoriamente ao longo de um período de seis meses revelaram o seguinte.

Temperatura	21,2	20,3	22,7	22,0	22,3	23,5	24,8	24,2	25,5	25,2	25,5	25,8
Produtividade	142	148	131	132	145	138	144	136	141	124	133	128

Temperatura	27,5	26,3	28,2	28,6	29,0	29,7	30,7	30,3	30,2	31,4	32,5	32,7
Produtividade	132	137	124	117	122	131	124	111	119	129	123	116

Calcule o valor do coeficiente de correlação entre a Temperatura e a produtividade e verifique a hipótese $H_0 : \rho = 0$. Depois plote um gráfico de dispersão e visualize a natureza da correlação entre Temperatura e Produtividade.

Exercício 9.9

A análise de 20 pares de valores indicou que a resistência à tração *(Y)* de uma fibra sintética usada na indústria têxtil guarda uma relação linear com a percentagem de algodão *(X)* presente na fibra. A equação obtida foi $Y = 35,7 + 0,85X$ (X fornecido em percentagem, equação válida para o intervalo de X entre 20 e 35%). Conhecidos os valores das Somas Quadradas S_{XY}=43,68 e S_{YY}=79,43 pede-se:

a) Faça a análise de Variância e conclua a respeito da significância do modelo.

b) Calcule o valor do coeficiente de determinação r^2 e indique qual o seu significado técnico.

Exercício 9.10

Um sofisticado simulador estocástico de tráfego fornece a velocidade média em avenidas de uma metrópole em função do volume de automóveis. O resultado de 14 simulações revelou o seguinte:

Vol. de Tráfego	3	3	5	5	10	10	15	15	20	20	25	25	30	30
Velocid. Média	95,6	93,8	74,4	74,8	50,5	51,5	44,6	42,4	35,8	38,7	32,0	3,2	30,1	29,1

Ajuste um modelo linear a esses dados e ache a equação de regressão $Y = b_0 + b_1 X$

Exercício 9.11

Calcule os resíduos padronizados para os dados do exercício 9.10. Após, plote um gráfico de *Resíduos* $\times X$ e verifique se há evidências de falta de ajuste do modelo linear.

Exercício 9.12

Utilize o seguinte modelo para ajustar os dados do exercício 9.10 $Y=b_0$ +

b_1 *(1 / √X).* Estime o valor dos coeficientes b_0 e b_1 para esse modelo não linear e depois repita a análise de resíduos pedida em 9.11 verificando se para o presente modelo há evidências de falta de ajuste.