UNIVERSIDADE FEDERAL DA BAHIA

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DEPARTAMENTO DE ESTATÍSTICA

NOTAS DE AULA

MAT236 – MÉTODOS ESTATÍSTICOS

3ª UNIDADE

2017.2

Esta apostila foi elaborada em 2004.1 pelas professoras
Giovana Silva, Lia Moraes, Rosana Castro e Rosemeire Fiaccone

Revisada em 2010.2
Monitora: Tatiana Felix da Matta

Revisada em 2013.1 pelas professoras:
Gecynalda Gomes e Silvia Regina

Revisada em 2014.1 pela professora:
Silvia Regina

Revisada em 2017.2 pelas professoras:
Giovana Silva e Verônica Lima

Revisada em 2017.2 pelos monitores:
Ícaro Augusto e Matheus Borges

Conteúdo

14 Análise de Regressão

Frequentemente, estamos interessados em estudar como duas ou mais variáveis estão associadas. Algumas vezes o interesse é apenas medir o grau de associação e outras vezes desejam-se obter um modelo matemático-estatístico que seja capaz de descrever a relação funcional entre as variáveis. Para investigar e modelar a relação entre elas, usa-se a Análise de Regressão.

Quando estamos estudando o comportamento de apenas duas variáveis x e y que supostamente se relacionam através de uma função linear, devemos considerar a seguinte equação:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad ,$$

em que β_0 é o intercepto e β_1 a inclinação. O erro aleatório ε pode ser pensado como uma "falha" da equação linear em se ajustar aos dados exatamente. Este modelo é chamado de Modelo de Regressão Linear Simples. Para estimar os parâmetros β_0 e β_1, uma amostra de pares (x; y) deve ser coletada e analisada. A variável x é conhecida como variável preditora ou independente e y é conhecida como variável resposta ou dependente.

Obtemos um modelo mais geral quando a variável resposta pode ser relacionada a k variáveis preditoras, $x_1, x_2, ..., x_k$ e, neste caso, o modelo adequado seria:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon.$$

Este modelo é chamado Modelo de Regressão Linear Múltipla. Nem sempre um modelo de regressão linear é o mais adequado para uma determinada situação. Algumas vezes, devemos modelar a relação entre variáveis utilizando funções não lineares ou mesmo fazendo alguma transformação funcional na(s) variável(s) de modo a obter linearidade.

Em todos os casos é importante destacar que um modelo de regressão não implica numa relação de causa-e-efeito. Para estabelecer causalidade, a relação entre as variáveis preditoras e a resposta deve ter uma base além do conjunto de dados. Por exemplo, o relacionamento entre variáveis pode ser sugerido por considerações teóricas. A Análise de Regressão pode apenas ajudar a confirmar esta relação.

14.1 Diagrama de Dispersão e Coeficiente de Correlação

Como dissemos anteriormente, para estudar a relação entre duas variáveis devemos partir da coleta de uma amostra de pares de observações. Para isto, é necessário realizar um experimento em que se faz simultaneamente medidas de duas variáveis x e y para uma amplitude de diferentes condições experimentais. Sejam $(x_1; y_1), (x_2; y_2), ..., (x_n; y_n)$ os n pares de observações.

Um procedimento para visualizarmos a forma da relação entre as variáveis x e y é o **diagrama de dispersão**, que nada mais é do que a representação dos pares de valores num sistema cartesiano.

Exemplo 14.1: (Werkema, 1996): Uma indústria fabricante de eletrodomésticos da chamada "linha branca", tem como objetivo resolver o problema apresentado pelo elevado índice de refugo da gaveta de legumes de um modelo de refrigerador produzido pela empresa. A observação do problema indicou que a maior parte das gavetas refugadas era considerada defeituosa por apresentarem corte fora de esquadro. Os técnicos da empresa suspeitaram que a ocorrência do corte de gavetas fora de esquadro pudesse estar relacionada à variação de tensão na rede elétrica, que poderia prejudicar o desempenho do equipamento de corte. Para a verificação da validade desta hipótese, foram coletados dados sobre a tensão na rede elétrica (x) e a variação no corte (y), os quais estão apresentados na tabela abaixo.

Tabela 14.1: Medidas da Tensão na Rede Elétrica (Volts) e Variação no Corte das Gavetas (mm).

Número da Medida	**Tensão na Rede Elétrica (Volts)**	**Variação no Corte (mm)**	**Número da Medida**	**Tensão na Rede Elétrica (Volts)**	**Variação no Corte (mm)**
1	222,7	15,7	19	219,9	16,2
2	217,7	17,0	20	222,2	15,9
3	219,4	16,3	21	213,9	19,1
4	220,9	16,1	22	216,0	18,0
5	214,4	18,6	23	218,1	17,0
6	216,5	17,8	24	222,0	16,0
7	213,0	19,5	25	224,1	15,4
8	221,7	16,0	26	214,9	18,6
9	224,7	15,3	27	214,2	18,7
10	215,5	18,3	28	223,3	15,6
11	220,0	16,3	29	216,7	17,6
12	218,6	16,7	30	215,3	18,5
13	223,5	15,7	31	223,8	15,5
14	217,0	17,4	32	220,6	16,1
15	221,5	16,1	33	215,8	18,2
16	218,4	16,8	34	217,3	17,3
17	213,6	19,3	35	219,2	16,5
18	221,2	16,2			

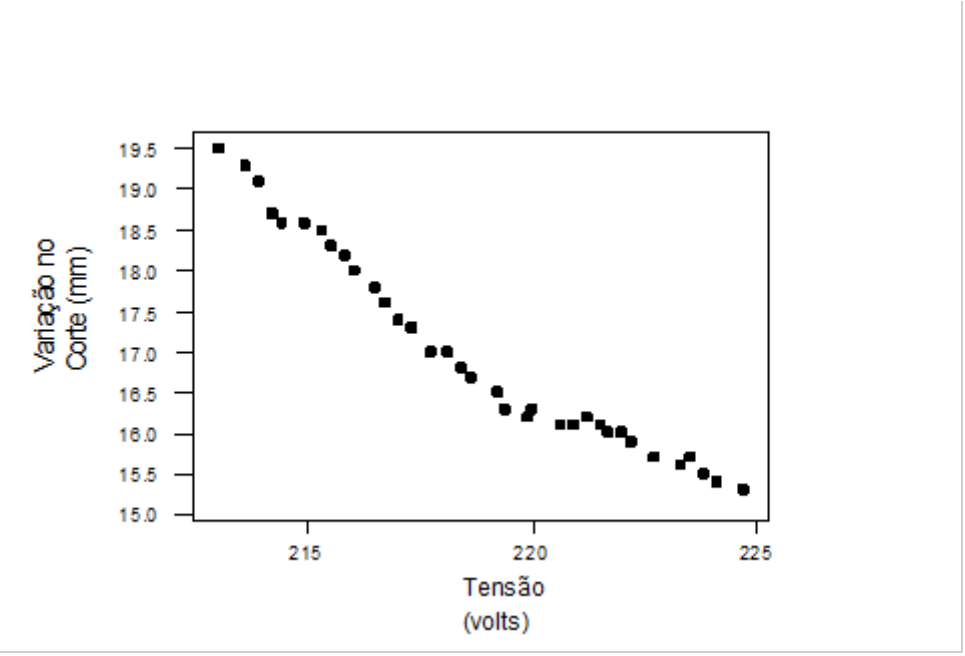


Figura 14.1: Diagrama de dispersão da Tensão da Rede Elétrica e da Variação no Corte.

Pela Figura 14.1, podemos constatar que existe uma tendência decrescente, já que maiores valores para a tensão na rede elétrica correspondem a menores valores para a variação no corte. Porém, observada esta associação, é útil quantificá-la.

Podemos utilizar o **coeficiente de correlação** para quantificar esta associação. Em geral, a letra r é usada para representar este coeficiente. Valores de r variam de –1,0 a + 1,0. Um r próximo a + 1,0 corresponde a um diagrama de dispersão em que os pontos caem em torno de linha reta com inclinação positiva, e um r próximo a –1,0 corresponde a um diagrama em que os pontos caem em torno de uma linha reta com inclinação negativa. Um r próximo a 0 corresponde a um conjunto de pontos que não mostram nenhuma tendência, nem crescente, nem decrescente. A Figura 14.2, a seguir, mostra cinco diagramas de dispersão de Y e X.

Os diagramas das Figuras 14.2(a) e 14.2(b) mostram duas situações em que os pontos estão em torno de uma reta imaginária ascendente. Valores pequenos de X estão associados a valores pequenos de Y, o mesmo acontecendo para valores grandes. Estes dois casos indicam o que chamamos de correlação linear positiva de Y e X. Porém, os dados em 14.2(b) apresentam uma correlação linear positiva mais forte que em 14.2(a).

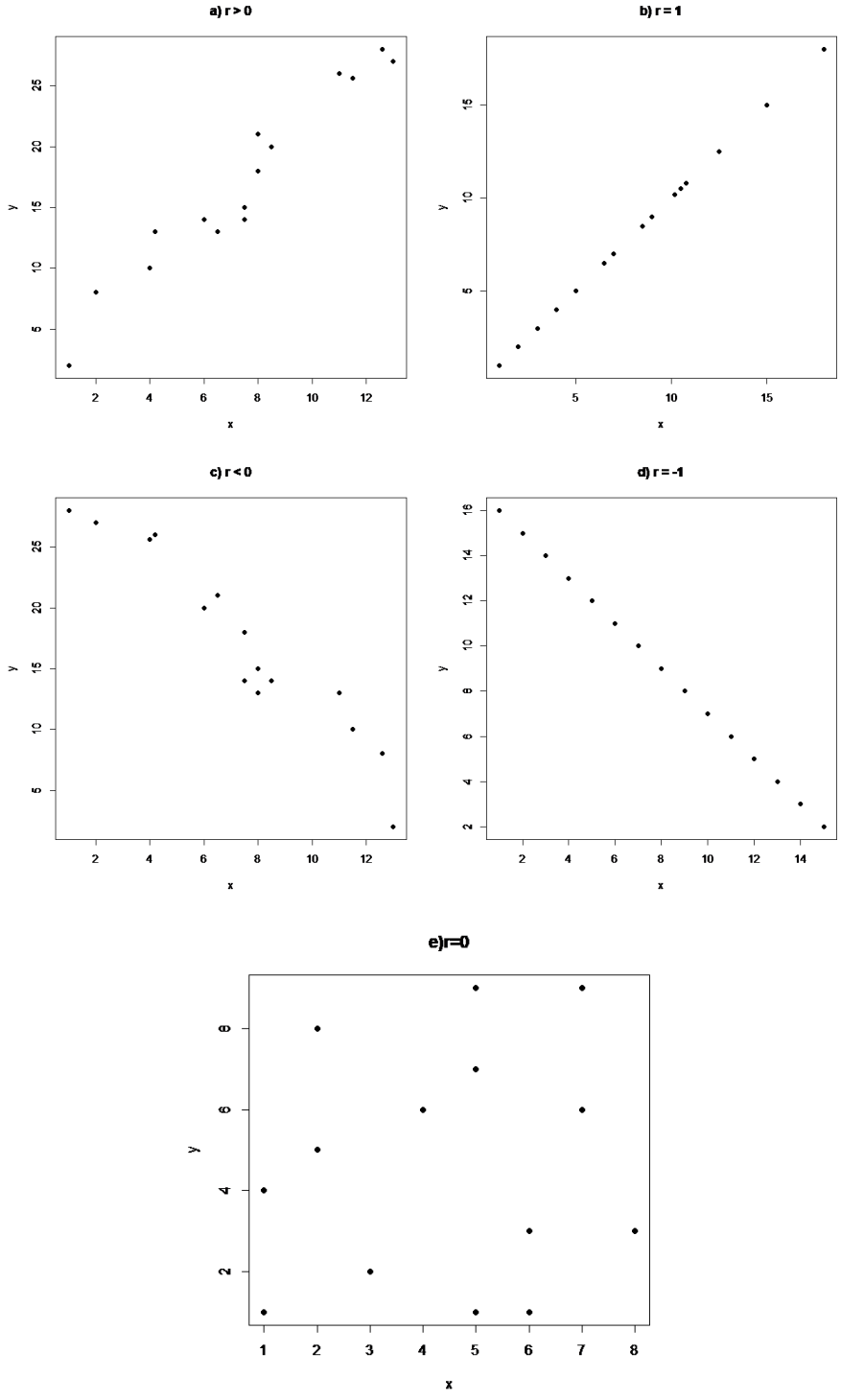


Figura 14.2 Gráficos de Dispersão

As Figuras 14.2(c) e 14.2(d) mostram que os pontos em X e Y estão em torno de uma reta imaginária descendente, indicando o que chamamos de correlação linear negativa, ou seja, valor de r menor que zero. Observe que em 14.2(d) a correlação é igual a -1.

Os valores de X e Y na Figura 14.2(e) não sugerem uma associação entre duas variáveis,

pois valores pequenos ou grandes de X estão associados tanto a valores pequenos quanto a valores grandes de Y. Os pontos do diagrama não se posicionam em torno de uma linha imaginária ascendente ou descendente.

O coeficiente de correlação, também chamado de **Coeficiente de Correlação de Pearson**, é calculado por:

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

ou,

$$r = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{[(\sum_{i=1}^{n} x_i^2) - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2][(\sum_{i=1}^{n} y_i^2) - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2]}}$$

em que x_i e y_i são os valores observados de X e Y, respectivamente; i=1,2,...,n e n é o número de observações para cada variável $\bar{x}$ e $\bar{y}$ são as médias de X e Y, respectivamente.

Exemplo 14.2: Calculando o coeficiente de correlação linear para os dados do exemplo 14.1, r = -0,976 , um valor muito próximo de –1, podemos concluir que existe uma forte correlação negativa entre a tensão na rede elétrica e a variação no corte das gavetas de legumes do refrigerador produzido pela indústria.

Tabela 14.2: Dados para o Cálculo do Coeficiente de Correlação para o Exemplo 14.1. (Continua)

i	x	y	x_2	y_2	xy
1	222,70	15,70	49595,29	246,49	3496,39
2	217,70	17,00	47393,29	289,00	3700,90
3	219,40	16,30	48136,36	265,69	3576,22
4	220,90	16,10	48796,81	259,21	3556,49
5	214,40	18,60	45967,36	345,96	3987,84
6	216,50	17,80	46872,25	316,84	3853,70
7	213,00	19,50	45369,00	380,25	4153,50
8	221,70	16,0	49150,89	256,00	3547,20
9	224,70	15,3	50490,09	234,09	3437,91
10	215,50	18,3	46440,25	334,89	3943,65
11	220,00	16,3	48400,00	265,69	3586,00

Tabela 14.2: Dados para o Cálculo do Coeficiente de Correlação para o Exemplo 14.1. (Conclusão)

i	x	y	x_2	y_2	xy
12	218,60	16,7	47785,96	278,89	3650,62
13	223,50	15,7	49952,25	246,49	3508,95
14	217,00	17,4	47089,00	302,76	3775,80
15	221,50	16,1	49062,25	259,21	3566,15
16	218,40	16,8	47698,56	282,24	3669,12
17	213,60	19,3	45624,96	372,49	4122,48
18	221,20	16,2	48929,44	262,44	3583,44
19	219,90	16,2	48356,01	262,44	3562,38
20	222,20	15,9	49372,84	252,81	3532,98
21	213,90	19,1	45753,21	364,81	4085,49
22	216,00	18,0	46656,00	324,00	3888,00
23	218,10	17,0	47567,61	289,00	3707,70
24	222,00	16,0	49284,00	256,00	3552,00
25	224,10	15,4	50220,81	237,16	3451,14
26	214,90	18,6	46182,01	345,96	3997,14
27	214,20	18,7	45881,64	349,69	4005,54
28	223,30	15,6	49862,89	243,36	3483,48
29	216,70	17,6	46958,89	309,76	3813,92
30	215,30	18,5	46354,09	342,25	3983,05
31	223,80	15,5	50086,44	240,25	3468,90
32	220,60	16,1	48664,36	259,21	3551,66
33	215,80	18,2	46569,64	331,24	3927,56
34	217,30	17,3	47219,29	299,29	3759,29
35	219,20	16,5	48048,64	272,25	3616,80
Total	7657,60	595,30	1675792,38	10178,11	130103,39

$$r = \frac{130103,39 - \frac{1}{35}(7657,60x595,30)}{\sqrt{[1675792,38 - \frac{1}{35}(7657,6)^2][10178,11 - \frac{1}{35}(595,30)^2]}} = -0,976$$

Cuidados com correlações

Um dos cuidados que devemos ter quando a correlação é interpretada é saber que correlação não é o mesmo que causalidade (relação de causa e efeito). Isto é, quando duas

variáveis são altamente correlacionadas, não significa, necessariamente, que uma causa a outra. Em alguns casos, podem existir relações causais, mas não se saberá isso pelo coeficiente de correlação. Provar uma relação de causa e efeito é muito mais difícil do que somente mostrar um coeficiente de correlação alto.

Um outro cuidado que deve ser tomado ao se interpretar correlação é associar um diagrama de dispersão ao conjunto de dados. Veja o exemplo abaixo.

Exemplo 14.3: Vamos calcular para cada um dos quatro conjuntos de dados abaixo o coeficiente de correlação.

Conjunto 1		*Conjunto* 2		*Conjunto* 3		*Conjunto* 4	
X	Y	X	Y	X	Y	X	Y
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,10	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,10	4	5,39	19	12,50
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

Para cada um deles, temos: r = 0,82 (Verifique!). Porém, estes conjuntos de dados apresentam disposições completamente diferentes no diagrama.

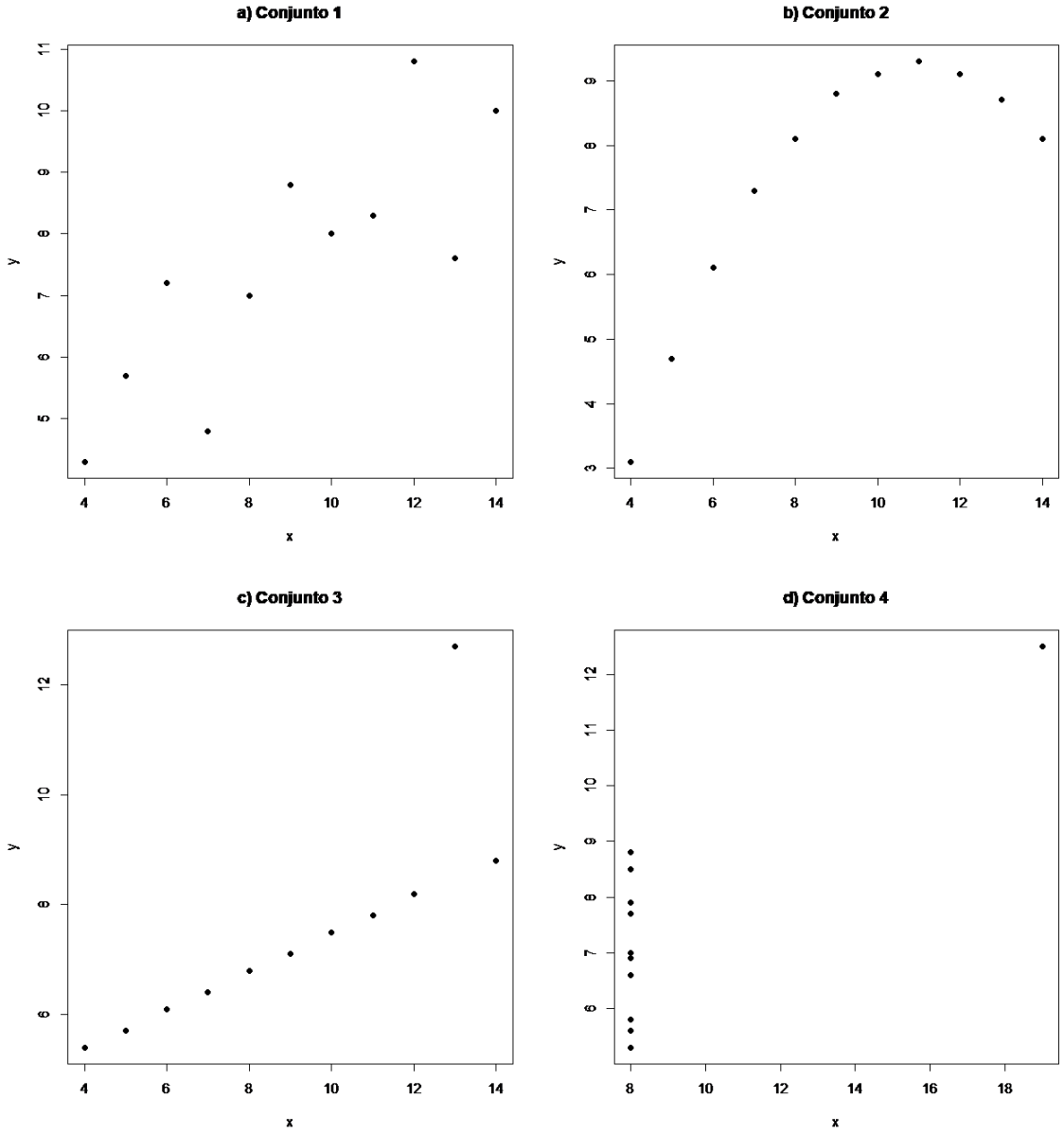


Figura 14.3 Diagramas de Dispersão.

Avaliando a Figura 14.3 (a) mostra que existe uma relação linear entre as variáveis x e y. Os dados em 14.3 (b) sugerem uma relação quadrática entre as variáveis. O diagrama da Figura 14.3 (c) apresentaria um coeficiente de correlação aproximadamente igual a 1, mas devido a um valor atípico apresentou um valor 0,82. Na Figura 14.3 (d) indica que não existe uma relação linear entre as variáveis, mas o valor 0,82 do coeficiente de correlação é devido à observação atípica presente no conjunto de dados.

Questões Não Respondidas pela Correlação

Ao estudarmos a relação entre variação no corte (mm) e tensão (volts) surgem algumas questões importantes tais como:

1. Qual é a previsão de variação no corte (mm) usando uma tensão especificada em volts?
2. Qual é a média estimada de variação no corte (mm) para uma especificada tensão em volts?
3. Quais são os limites de confiança para variação no corte (mm) predita?

Questões deste tipo podem ser respondidas com uma análise de regressão dos dados, que é o assunto das próximas seções.

Principais Objetivos da Análise de Regressão

De maneira geral, os modelos de regressão podem ser usados para vários propósitos, dentre os quais é possível destacar:

a) Descrição dos dados

b) Estimação dos parâmetros

c) Predição

d) Controle

Descrição dos dados

É muito comum a utilização da análise de regressão para descrever um conjunto de dados. Isto é, a construção de um modelo que relacione, por exemplo, o efeito do ar condicionado no consumo de energia elétrica é uma maneira muito mais efetiva de conhecer o relacionamento entre estas variáveis em comparação a uma tabela ou mesmo um gráfico.

Estimação dos parâmetros

No exemplo sobre o consumo de energia elétrica, podemos utilizar a análise de regressão para conhecermos qual o número médio de kilowatt/hora consumido usando o ar condicionado por uma hora.

Predição

É possível também, utilizar regressão para predizer valores para a variável resposta. Voltando ao Exemplo 14.1, o fabricante pode estar interessado em conhecer quanto será a variação do corte da gaveta (em mm) para uma determinada tensão na rede elétrica (em volts).

Controle

A Análise de Regressão pode ser usada com o objetivo de controlar a variável resposta. Considere, como exemplo, um engenheiro químico que está interessado em controlar o rendimento de um processo químico através das variáveis temperatura e tempo de reação. Esta equação poderá ser utilizada para determinar a natureza dos ajustes a serem realizados nas variáveis temperatura e tempo de reação, para que o rendimento possa ser mantido num intervalo pré-estabelecido.

É importante destacar que, quando o modelo de regressão for empregado com o objetivo de controle, a relação existente entre a variável de interesse e as variáveis utilizadas para seu controle sejam do tipo causa-e-efeito.

14.2 Regressão Linear Simples por Mínimos Quadrados

Um coeficiente de correlação descreve a associação linear entre variáveis porém, para investigar e modelar a relação entre elas, usa-se a Análise de Regressão.

Para se ajustar um modelo de regressão por mínimos quadrados a variável resposta deve ser quantitativa.

O que se deseja, freqüentemente, com base em dados amostrais, é estimar o valor da variável y, correspondente ao conhecimento de uma variável x. Isto pode ser feito mediante a estimativa da função linear $f(x) = y = \beta_0 + \beta_1 x$.

Observe, porém, que as linhas que várias pessoas podem traçar para este conjunto de pontos seriam, provavelmente, similares, desde que o gráfico tenha um padrão bem definido. Porém, elas não seriam idênticas, de forma que os valores preditos para variável resposta poderiam diferir também.

Para um conjunto de dados sem um padrão óbvio no gráfico; diferentes pessoas poderiam traçar diferentes linhas sobre os dados, permitindo grandes diferenças entre os valores preditos. Usando a Análise de Regressão, qualquer um obterá exatamente a mesma linha reta. Este processo é chamado de ajuste de uma reta de regressão. O método usado mais freqüentemente para ajustar uma reta usa um princípio chamado de Mínimos Quadrados. Este método será descrito posteriormente.

Observe a Figura 14. 4 a seguir. O princípio de mínimos quadrados envolve ajustar uma reta passando por pontos de forma que as diferenças verticais entre todos os pontos e a reta

são calculadas. Então, estas diferenças são elevadas ao quadrado para dar aos pontos acima e abaixo da reta a mesma importância (as diferenças ao quadrado são todas positivas). As diferenças são então somadas. A "melhor" reta é aquela que minimiza esta soma das diferenças ao quadrado, sendo chamada, de **mínimos quadrados.**

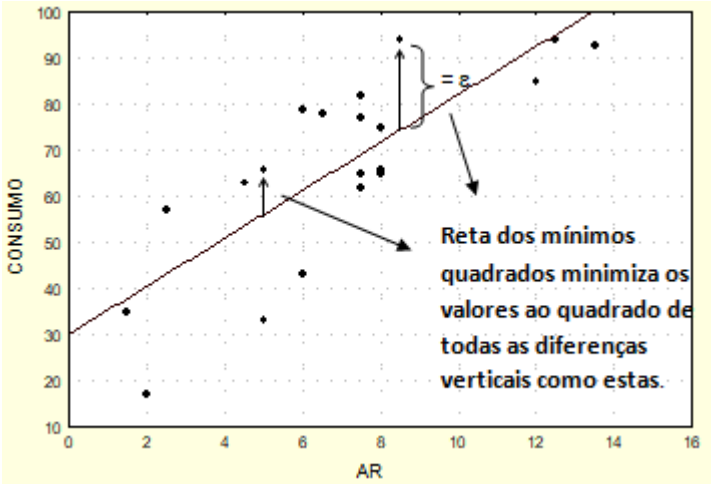


Figura 14.4 – Processo de Mínimos Quadrados

Já vimos que uma relação linear entre duas variáveis pode ser expressa através da equação:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad ,$$

em que, β_0é o intercepto, β_1 é a inclinação e ε é o erro. Esta equação é a que se obteria medindo-se a população inteira de valores de x e y. Na realidade, apenas uma amostra é medida e usa-se esta amostra para estimar a reta. A reta estimada por meio da amostra pela regressão de mínimos quadrados será denotada por:

$$\hat{y} = b_0 + b_1 x,$$

em que b_0 e b_1 são estimativas de β_0 e β_1 , respectivamente. O valor b_0 é o valor predito de $\hat{y}$ quando x é zero e é chamado de **intercepto** da reta desde que ele é o local em que a reta intercepta o eixo vertical. O valor b_1 é o incremento em $\hat{y}$ resultante do incremento de uma unidade em x e é chamado de **inclinação** da reta.

O método de Mínimos Quadrados é baseado na soma dos quadrados dos resíduos, ε, ou seja:

$$l(\beta_0, \beta_1) = \sum_{i=0}^{n} \varepsilon_i^2 = \sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 x_1)^2,$$

fazendo-se tal soma a menor possível. Esta minimização implica em derivar $l(\beta_0, \beta_1)$ com respeito a β_0 e β_1 e igualar estas derivadas a zero. Assim, a solução deste sistema de

equações fornece as seguintes expressões como estimadores de mínimos quadrados para estes parâmetros:

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad e \quad \hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x}$$

ou

$$b_0 = \frac{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_iy_i}{n\sum_{i=1}^{n}x_i^2 - (\sum_{i=1}^{n}x_i)^2}$$

e

$$b_1 = \frac{n\sum_{i=1}^{n}y_ix_i - \sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i}{n\sum_{i=1}^{n}x_i^2 - (\sum_{i=1}^{n}x_i)^2}$$

em que y_i e x_i são os valores observados das variáveis Y e X e $\bar{x}$ e $\bar{y}$ são as respectivas médias amostrais destas variáveis.

15 Relação com máxima verossimilhança

Observe que nenhuma suposição sobre a distribuição dos erros foi feita. Entretanto, quando o tipo da distribuição dos erros é conhecida, o método de estimação por máxima verossimilhança pode ser aplicado. Assumindo que os erros são independentes e têm distribuição $N(0, \sigma^2)$, isto é, $\epsilon_i \sim N(0, \sigma^2)$, então $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ e são independentes. Neste caso, a função de densidade de y_i é dada por:

$$f(y_i) = \frac{1}{\sqrt{(2\pi\sigma^2)}e^{-(y_i-\beta_0-\beta_1x_i)/(2\sigma^2)}} \quad \text{para} \quad -\infty < y_i < \infty.$$

Logo como visto na Seção 12.1, a função de verossimilhança é dada por:

$$\ln L(\beta_0, \beta_1, \sigma) = \ln(f(y_1)) + \cdots + \ln(f(y_n) = -n\ln(\sigma) - n\ln(\sqrt{(2\pi)}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1x_i).$$

Para $\sigma > 0$ fixo, o logaritmo da função de verossimilhança atinge seu máximo quando $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1x_i)$ é mínimo, isto é, o estimador de máxima verossimilhança de β_0 e β_1

coincide com os estimadores produzidos pelo método de mínimos quadrados.

O modelo de regressão adotado para o Exemplo 14.1 é dado por:

$$y = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, 35.$$

Vamos ajustar agora um modelo de regressão linear simples com dados do Exemplo 14.1.

$$b_0 = \frac{1675792,38x595,3 - 7657,60x130103,39}{35x1675792,38 - (7657,6)^2} = 94,96$$

e

$$b_1 = \frac{35x130103,39 - 595,3x7657,6}{35x1675792,38 - (7657,6)^2} = -0,3563 \approx -0.36$$

portanto, o modelo de regressão ajustado é expresso por:

$$\hat{y}_i = 94,96 - 0,36x_i, \quad i = 1, ..., 35$$

Esta equação de regressão mostra que para cada aumento de um volt na tensão na rede elétrica a variação no corte das gavetas diminui, em média, 0,36 mm. Como o intervalo dos valores observados de x não contempla o valor zero, o valor 94,96 não tem um significado particular como termo separado do modelo de regressão.

Após o ajuste do modelo de regressão é importante verificar se existe, de fato, uma relação linear entre as variáveis x e y. A seguir, apresentaremos como verificar essa hipótese de linearidade.

Análise de Variância no Modelo de Regressão

Um dos objetivos da análise de regressão é testar a significância do coeficiente do modelo, isto é, queremos testar:

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

Neste teste, rejeitar H_0 significa que existe relação linear entre x e y, ou que x é importante para explicar a variabilidade em y. Para testar estas hipóteses de interesse, vamos assumir que os erros são independentes e possuem distribuição normal com média 0 e variância constante σ^2, isto é $\varepsilon_i \sim N(0, \sigma^2)$ para i=1,...,n.

Um primeiro procedimento para testar H_0 é utilizar a estatística de teste dada por

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{QM_{Residual}}{S_{XX}}}},$$

em que $QM_{Residual} = SQ_{Residual}/(n-2)$. Sob H_0, a estatística t_0 possui distribuição t com n-2 graus de liberdade. Rejeita-se H_0, ao nível de significância α, se $|t_0| > t_{\alpha/2,n-2}$.

Um segundo procedimento, é baseado na partição da variação total da variável dependente Y que pode ser decomposta em duas partes: uma explicada pelo modelo de regressão ajustado e outra não explicada, conforme mostra a equação abaixo.

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{1}$$

O termo do lado esquerdo de (1) é a soma dos quadrados das observações em relação ao seu valor médio e representa uma medida da variabilidade total dos dados de Y. Esta soma é denotada por $SQ_{Total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$. O primeiro termo do lado direito de (1) é a soma dos quadrados explicada pelo modelo de regressão, sendo denotada por $SQ_{Regressão} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ enquanto o segundo termo é a soma de quadrados residual $SQ_{Residual} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ que não é explicada pelo modelo de regressão. O modelo será melhor ajustado quanto maior for a variação explicada $SQ_{Regressão}$ em relação á variação total SQ_{Total}. A equação (1) pode ser representada como:

$$SQ_{Total} = SQ_{Regressão} + SQ_{Residual}.$$

Assim, para testar estas hipóteses será usada a estatística $F = \frac{MQ_{regressão}}{MQ_{residual}}$, a qual tem distribuição F (também conhecida como distribuição de Fisher-Snedecor) com 1 e n-2 graus de liberdade que correspondem ao numerador e ao denominador, respectivamente.

O critério do teste é o seguinte: rejeita-se H_0, ao nível α de significância, se F > F(α, 1 ; n - 2), em que F(α, 1; n - 2) é o α percentil da distribuição F com 1 e n - 2 graus de liberdade, respectivamente. Caso contrário, a hipótese H_0 não deve ser rejeitada. Este teste pode ser resumido através da Tabela de Análise de Variância como mostrado a seguir (ver Tabela 14.3).

Tabela 14.3: Tabela de Análise de Variância para o Modelo de Regressão Linear Simples.

Fonte de variação	**Graus de liberdade (gl)**	**Soma de quadrados (SQ)**	**Quadrados médios (MQ)**	**F**
Regressão	1	$\mathbf{SQ}_{Regressão}$	$MQ_{Regressão} = \frac{SQ_{Regressão}}{1}$	$F = \frac{MQ_{Regressão}}{MQ_{Residual}}$
Residual	n-2	$\mathbf{SQ}_{Residual}$	$MQ_{Residual} = \frac{SQ_{Residual}}{n-2}$	
Total	n-1	$\mathbf{SQ}_{Total}$		

Para uma amostra n pares (x, y), a soma de quadrados total associada a variabilidade total de Y tem n-1 graus de liberdade e a soma de quadrados de resíduo tem n-2 graus de liberdade. Os quadrados médios são obtidos dividindo as somas de quadrados pelos correspondentes graus de liberdade.

Quando as somas de quadrados forem calculadas manualmente, elas podem ser obtidas através das seguintes expressões dadas adiante.

$$SQ_{Total} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 \tag{2}$$

$$SQ_{Regressão} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = b_1[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})] = b_1[\sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}] \tag{3}$$

e

$$Q_{Residual} = SQ_{Total} - SQ_{Regressão} \tag{4}$$

Para o Exemplo 14.1 será realizado o seguinte teste de hipóteses:

$H_0 : \beta_1 = 0$ (Não existe relação linear entre a tensão da rede elétrica e o corte da gaveta)
$H_1 : \beta_1 \neq 0$ (Existe relação linear entre a tensão da rede elétrica e o corte da gaveta)

A estatística t_0 é dada por:

$$t_0 = \frac{-0,36}{\sqrt{\frac{0,0762}{1675792,38-35x218,79^2}}} = -25,96$$

Com auxílio da Tabela da distribuição t, obtemos $t_{0,025;33} = 2,03$. Como $|t0| > t_{0,025;33}$ rejeitamos a hipótese nula e concluímos que os dados estão indicando a existência de uma relação linear entre a tensão na rede elétrica (volts) e a variabilidade no corte das gavetas (mm) produzidas pela fabrica, ao nível de significância de 5%. A soma de quadrados para compor a Tabela da Análise de Variância é calculada conforme as equações (2) a (4),

$$SQ_{Total} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 =$$

$$10178,11 - \frac{1}{35}(595,3)^2 \approx 52,907$$

$$SQ_{Regressão} = b_1[\sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}] =$$

$$-0,3563[130103,39 - 35(\frac{7657,60}{35})(\frac{595,30}{35})] \approx 50,397$$

Uma vez calculadas as duas somas de quadrados, obtemos a terceira soma de quadrados conforme apresenta adiante:

$$SQ_{Residual} = SQ_{Total} - SQ_{Regressão} = 52,91 - 50,397 = 2,513.$$

Os quadrados médios (MQ) sãos calculados através das expressões adiante.

$$MQ_{Regressão} = \frac{SQ_{Regressão}}{1} = 50,397 e MQ_{Residual} = \frac{SQ_{Residual}}{n-2} = \frac{2,513}{33} = 0,0762.$$

O valor da estatística de teste

$$F_{calculado} = \frac{MQ_{Regressão}}{MQ_{Residual}} = \frac{50,367}{0,0762} = 661,377$$

Os resultados estão resumidos na Tabela da Análise de Variância a seguir.

Tabela 14.4: A Tabela Anova para o modelo de Regressão Linear Simples

Fonte de variação	Graus de liberdade (gl)	Soma de quadrados (SQ)	Quadrados médios (MQ)	$F_{calculado}$	*p-valor
Regressão	1	50,397	50,397	661,377	$2,071x10^{-23}$
Residual	33	2, 513	0, 0762		
Total	34	52,907	...	...	

$$*\text{p-valor} = P(F_{1;33} \geq F_{Calculado}) = P(F_{1;33} \geq 661,377) \approx 2,071x10^{-23}$$

Com auxílio da Tabela da distribuição F, obtemos $F_{0,05;1;33} = 4,139$ (ver Tabela 2 em Anexo). Como $F_{0,05;1;33} = 4,139 < F_{Calculado} = 661,38$ rejeitamos a hipótese nula e concluímos que os dados estão indicando a existência de uma relação linear entre a tensão na rede elétrica (volts) e a variabilidade no corte das gavetas (mm) produzidas pela fabrica, ao nível de significância de 5%. Chegamos a mesma conclusão ao observarmos o p-valor $= 2,071x10^{-23} < \alpha = 0,05$.

No caso de regressão linear uma forma de medir a proporção da redução na variação total em Y associada com o uso da variável explicativa X é o coeficiente de determinação expresso por:

$$R^2 = \frac{SQ_{Regressão}}{SQ_{Total}} = 1 - \frac{SQ_{Residual}}{SQ_{Total}}$$

O valor de R_2 varia no intervalo [0, 1]. Desta forma, quanto maior for o coeficiente de determinação, maior será a redução na variação total de Y pela introdução da variável independente X. Entretanto, o coeficiente de determinação dever ser empregado com muita cautela. Por exemplo, quando temos dados envolvendo séries temporais que tendem a se mover na mesma direção, refletindo uma forte tendência, qualquer modelo que detecte essa tendência terá um de R_2 alto, o que pode ser espúrio (não refletir a verdadeira relação linear entre as variáveis envolvidas) (Souza, 1998).

O coeficiente de determinação $R^2 = \frac{50,397}{52,907}$ revela que aproximadamente 95,3% da variabilidade no corte das gavetas produzidas pela fabricadas é explicada pela tensão na rede elétrica (através do modelo proposto) e que 4,7% são atribuídas a outras causas.

Saída do software Excel

Estatística de regressão	
R múltiplo	0,9760
R-Quadrado	**0,9525**
R-quadrado ajustado	0,9511
Erro padrão	0,2760
Observações	35

Coeficiente de Determinação para Regressão Linear Simples

A Tabela Anova

P- valor

ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	1,00	50,39	50,39	661,60	$2{,}071x10^{-23}$
Resíduo	33,00	2,51	0,08		
Total	34,00	52,91			

	Coeficientes	Erro padrão	Estatística t	P- valor	95% inferiores	95% superiores
Interseção	**94,957**	3,031	31,330	0,000	**88,791**	**101,124**
Tensão	-0,356	0,014	-25,722	0,000	-0,384	-0,328

Estimativa para β_1

Estimativa para β_0

Intervalos de confiança para β_0 e β_1

Outra maneira de verificar a adequação do modelo de regressão linear simples é apresentada adiante. Para tal, é necessário supor que o erro ε **tem distribuição normal com média 0 e variância** σ^2. Na Seção adiante será estudada a análise de resíduo para a verificação desta suposição.

O intervalo de confiança para β_1 com (1-α)100% de confiança é dado por:

$$b_1 \mp t_{(\alpha/2;n-2)} \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}} \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Exemplo 14.5: Vamos calcular o intervalo de confiança para β_1 com 95% para o modelo de regressão linear simples com os dados do Exemplo 14.1.

$$-0,36 \mp 2,0345x\sqrt{\frac{2,5136}{33}}\sqrt{\frac{1}{397,0154}} = [-0,388; -0,332]$$

Baseado neste intervalo pode-se concluir que existe evidência que $\beta_1 \neq 0$, com 95% de con-

fiança e, portanto, há evidência de que o modelo de regressão linear é adequado.

Predição de Novas Observações

Suponha que se queira predizer uma nova observação y correspondendo a um nível especificado da variável preditora x. Denotando $x = x^*$ como sendo este o valor de interesse, então,

$$y^* = b_0 + b_1 x^*$$

é uma estimativa pontual para o novo valor da resposta. Considerando que o erro ε tem distribuição normal com média 0 e variância σ^2, o intervalo de predição para y^* com (1-α)% de confiança é dado por:

$$b_0 + b_1 x^* \mp t_{\alpha/2;n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

em que

$$s^2 = \frac{[(\sum_{i=1}^{n} y_i^2) - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2] - b_1[\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)]}{n-2}$$

ou

$$s^2 = \frac{SQ_{Residual}}{n-2}$$

s é o desvio padrão do erro e $t_{(\alpha/2;n-2)}$ é o ponto que define uma área de $(\alpha/2)$ na cauda superior da distribuição t com n-2 graus de liberdade.

Exemplo 14.5: Suponha que se queira predizer a variação no corte (mm) quando a tensão é 200 volts. Neste caso, $x^* = 200$ volts, e, portanto, variação predita = 95,03 – (0,36 x 200) = 23,03.

O intervalo de 95% confiança é: $(\alpha = 0,05 \rightarrow t_{0,025;33} = 2,035;\ n = 35;\ s = \sqrt{0,0762} = 0,276)$

$$[23,03 \mp (2,035)(0,276)\sqrt{1 + \tfrac{1}{35} + \tfrac{(200-218,79)^2}{397,015}}] = [22,3; 23,8]$$

Isto significa que você pode estar confiante com 95% que a variação do corte (mm) quando a tensão é de 200 volts varia entre 22,3 e 23,7.

Observação: Deve-se tomar cuidado quando estender uma reta de regressão ajustada para se fazer predições fora do intervalo de variação dos valores de x, usados para ajustar a reta de regressão. Não somente o intervalo de predição começa a se tornar mais largo,

tornando as previsões de pouca confiança, como o padrão da relação entre as variáveis pode mudar drasticamente para valores distantes de x. Os dados coletados não dão nenhuma indicação sobre a natureza desta mudança.

Diagnósticos Básicos em Regressão

Como determinar se um modelo representa adequadamente os dados? Como saber se mais termos devem ser adicionados ao modelo? Como identificar *outliers*, isto é, observações que não são típicas do restante da massa de dados? Estas são questões que podem ser respondidas examinando-se os resíduos do modelo ajustado, isto é, as diferenças entre os valores observados e preditos pelo modelo.

Para que um modelo de regressão possa ser empregado como base para outros estudos, é necessário que as suposições feitas durante sua construção sejam válidas. Se algumas destas suposições não se confirmarem, o modelo poderá ser inadequado para fazer as inferências de interesse. Neste caso, deve ser procurado outro modelo mais adequado ou ser empregada outra abordagem para a análise do problema.

As suposições que devem ter sua validade verificada são:

- O relacionamento entre y e x é linear;
- O erro ε tem média zero;
- O erro ε tem variância constante;
- Os erros são não correlacionados;
- O erro ε tem distribuição normal.

Diagnósticos básicos em regressão e ajuste de modelos são interdependentes. Primeiro um modelo é ajustado, e então se examina o modelo usando diagnósticos. Isso pode levar ao ajuste de um segundo modelo, o qual deve ser examinado por meio da análise dos resíduos. O processo continua até que se encontre um modelo que se ajuste bem aos dados. Note que é possível não se encontrar um modelo que represente adequadamente os dados.

Nesta seção serão discutidos métodos úteis para o estudo da adequação do modelo de regressão.

Análise de Resíduos

Um resíduo é definido por:

$$e_i = y_i + \hat{y}_i, \quad i = 1, 2, 3, ..., n \qquad ,$$

em que y_i é o valor observado e $\hat{y}_i$ é o correspondente valor estimado por meio do modelo de regressão.

È conveniente visualizar os resíduos como valores observados para o erro ε que aparecem no modelo. Portanto, é razoável esperar que quaisquer desvios das suposições feitas sobre o erro poderão ser detectados se for realizada uma análise de resíduos.

Gráficos dos Resíduos (e_i) contra os Valores Preditos ($\hat{y}_i$)

Se o modelo tem todos os termos que precisa, então o gráfico dos resíduos contra os valores preditos ou contra as variáveis independentes deveria parecer como uma distribuição aleatória de pontos sem tendência (numa faixa horizontal). Se o modelo precisa de outros termos, então o gráfico dos resíduos tem um padrão que sugere que tipo de termo deveria ser adicionado ao modelo. Alguns padrões são mostrados na Figura 14.5(a) seguir.

O padrão da Figura 14.5(a) representa a situação satisfatória. Nela os resíduos estão situados, aproximadamente, em uma faixa horizontal centrada em $e_i = 0$. Já os padrões b, c e d da Figura 14.5, indicam a presença de inadequações no modelo.

O padrão apresentado na Figura 14.5(b), o qual é semelhante à forma de um funil, indica que a variância do erro não é constante. Nesta figura a variância do erro é uma função crescente de $\hat{y}_i$. No entanto também existem situações em que a variância do erro aumenta com o decréscimo de $\hat{y}_i$.

O padrão apresentado na Figura 14.5(c) ocorre quando a variância dos erros é maior para valores intermediários de y e, portanto, também indica que erros não têm variância constante.

A Figura 14.5 (d) indica não linearidade. Este padrão pode indicar a necessidade da inclusão no modelo de um termo quadrático em x.

Quando é detectada que a variância do erro não é constante uma solução para este problema consiste em realizar transformações na variância resposta para estabilizar a variância.

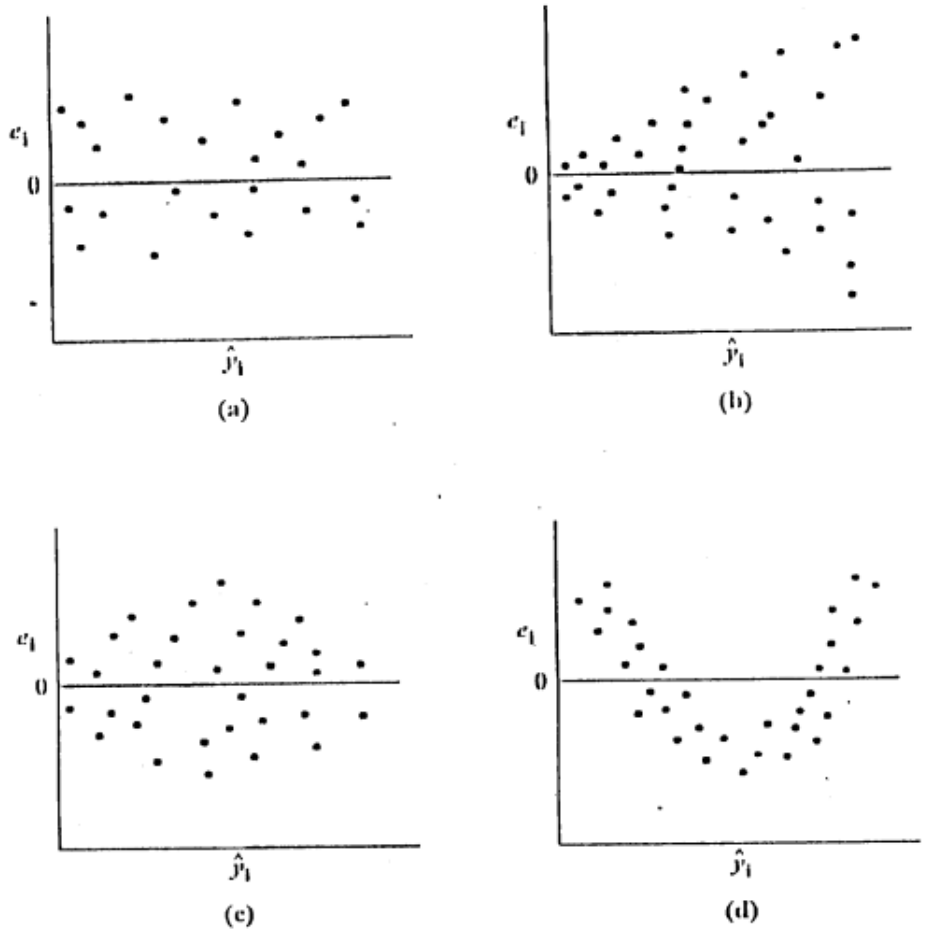


Figura 14.5: Gráficos de Resíduos contra Valores Ajustados.

Gráficos de Resíduos (e_i) Contra Valores da Variável Preditora (x)

No caso do modelo de regressão linear simples, um gráfico dos resíduos contra os valores da variável preditora fornece o mesmo tipo de informação gerada pelo gráfico de resíduos contra os valores ajustados. A configuração dos gráficos e_i versus x_i poderá corresponder a um dos quatro padrões gerais já apresentados na Figura 14.5, bastando para isso que, nesta Figura, $\hat{y}_i$ seja substituído por x_i. A interpretação dos padrões representados na Figura 14.5, após a substituição de $\hat{y}_i$ por x_i , é semelhante à já apresentada na seção anterior.

Gráfico de Resíduos Contra o Tempo

A validade da suposição de que os erros não são correlacionados pode ser verificada por meio de um gráfico de resíduos contra o tempo ou ordem de coleta das observações. A presença de configurações especiais neste gráfico pode indicar que os erros são correlacionados.

As duas configurações apresentadas na Figura 17.6 a seguir indicam a presença de correlação entre os erros, que representam uma séria violação das suposições associadas ao modelo de regressão.

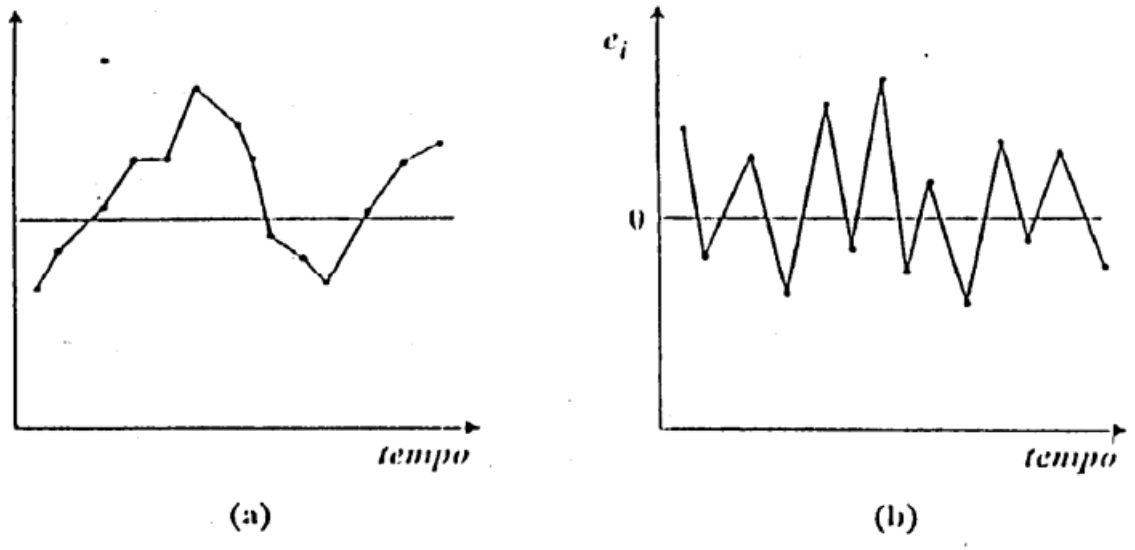


Figura 14.6: Gráficos de Resíduos contra o Tempo Indicando a Presença de Autocorrelação.

Gráfico de Probabilidade Normal para os Resíduos

A validade da suposição de normalidade pode ser verificada por meio do gráfico de probabilidade normal para os resíduos. A suposição de normalidade será considerada válida se os pontos do gráfico estiverem localizados, aproximadamente, ao longo de uma linha reta. Como esta avaliação é subjetiva, um teste estatístico pode ser utilizado para complementar esta avaliação.

Tabela 14.5: Valores previstos e os resíduos do modelo linear simples ajustado para a variação no corte.

Previsto	**Resíduo**	**Previsto**	**Resíduo**	**Previsto**	**Resíduo**	**Previsto**	**Resíduo**
15,62	0,08	16,58	-0,28	18,75	0,35	15,22	0,28
17,40	-0,40	17,08	-0,38	18,00	0,00	16,36	-0,26
16,79	-0,49	15,33	0,37	17,25	-0,25	18,07	0,13
16,26	-0,16	17,65	-0,25	15,86	0,14	17,54	-0,24
18,57	0,03	16,04	0,06	15,12	0,28	16,86	-0,36
17,82	-0,02	17,15	-0,35	18,39	0,21		
19,07	0,43	18,86	0,44	18,64	0,06		
15,97	0,03	16,15	0,05	15,40	0,20		
14,90	0,40	16,61	-0,41	17,75	-0,15		

Exemplo 14.6: Vamos agora examinar os resíduos para o modelo linear simples ajustado para a variação no corte.

Análise de Resíduos

Figura 14.7: Gráfico de Probabilidade Normal

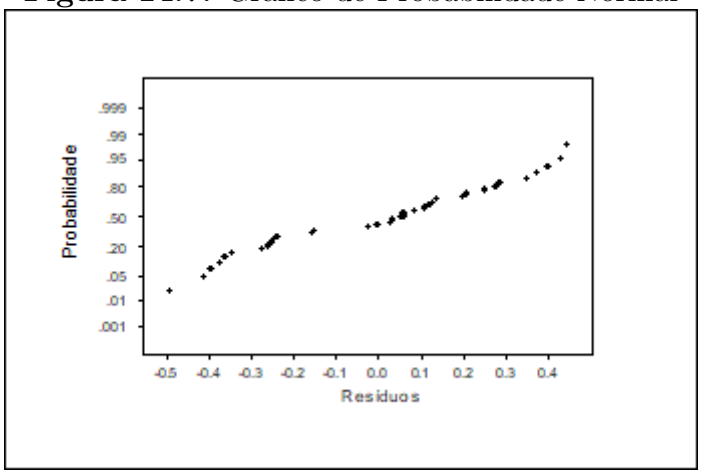


Figura 14.8: Histograma dos Resíduos

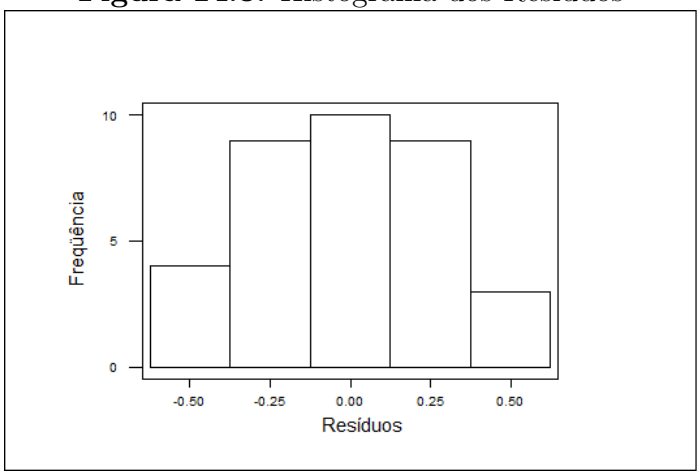


Figura 14.9: Resíduos vs valores ajustados

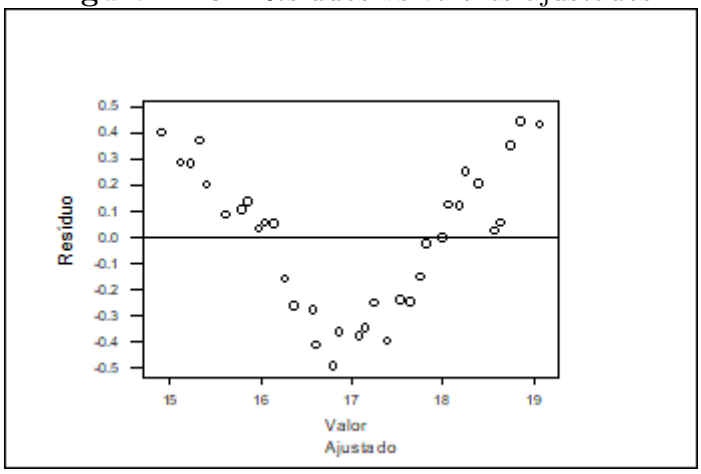
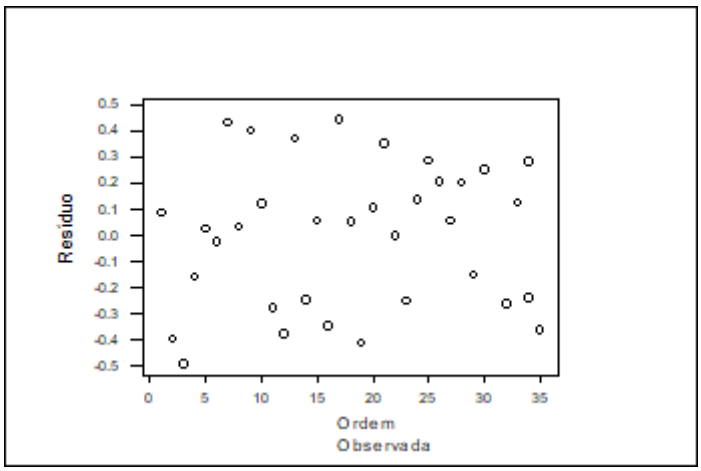


Figura 14.10: Resíduos vs Ordem dos dados



O primeiro gráfico serve para verificar a suposição de normalidade dos resíduos. Este gráfico parece indicar que os resíduos seguem uma distribuição normal. O segundo gráfico é o histograma dos resíduos também serve para verificar normalidade.

O terceiro gráfico apresenta os valores preditos *versus* resíduos. O padrão deste gráfico é semelhante ao apresentado na Figura 14.5 (d), o que indica a necessidade da inclusão no modelo de um termo quadrático em X.

O quarto gráfico apresenta a ordem em que os valores foram observados versus resíduos (foi considerado que as observações estão listadas no Exemplo 14.1 na ordem em que foram observadas). Pode-se notar que a relação entre os valores preditos e a ordem de observação é aleatória.

Exercícios de fixação:

1. Uma indústria produz grandes quantidades de alumina (Al_2O_3 de elevado teor de pureza) para a fabricação de alumínio metálico. A matéria prima para a fabricação da alumina é a bauxita, um mineral com cerca de 55% de óxido de alumínio (Al_2O_3). No processo de produção da alumina, o teor da Na_2O (óxido de sódio) ocluído no produto é um fator importante do ponto de vista da qualidade da alumina fabricada. O Na_2O é uma impureza, e, portanto é desejável que o seu teor na alumina seja o mais baixo possível. Com o objetivo de minimizar o teor da Na_2O ocluído no produto durante a etapa de precipitação, um dos estágios do processo de produção da alumina, a indústria iniciou trabalhos para melhoria. Os técnicos da empresa sabiam que a razão Al_2O_3 / NaOH era um dos fatores responsáveis pelas variações no teor de Na_2O da alumina. Nesta razão, o símbolo Al_2O_3 está representando a massa de óxido de alumínio proveniente da bauxita que entra no processo de produção, e o símbolo NaOH se refere à massa de hidróxido de sódio, um dos reagentes do processo, que é empregada na fabricação de alumina. Durante a etapa de observação do problema,

para se conhecer melhor a relação entre estas duas variáveis (variável resposta: Na_2O e variável preditora: Al_2O_3 / NaOH), os técnicos da indústria coletaram os dados apresentados na tabela a seguir.

Tabela: Teor de Na_2O ocluído na Alumina em Função da Razão Al_2O_3 / NaOH

Índice	**Razão Al_2O_3 / NaOH – (x)**	**Teor Na_2O (%)–(y)**
1	0,645	0,46
2	0,643	0,46
3	0,648	0,45
4	0,639	0,44
5	0,641	0,45
6	0,648	0,47
7	0,635	0,42
8	0,646	0,47
9	0,646	0,45
10	0,643	0,44
11	0,641	0,40
12	0,643	0,42
13	0,637	0,42
14	0,635	0,42
15	0,64	0,41
16	0,646	0,43
17	0,636	0,41
18	0,639	0,40
19	0,634	0,39
20	0,636	0,38
21	0,643	0,40
22	0,647	0,43
23	0,637	0,42
24	0,631	0,37
25	0,633	0,41

a) Com base no diagrama de dispersão, o que você pode dizer sobre a relação entre essas duas variáveis.

b) Calcule o coeficiente de correlação. O que você pode concluir?

c) Ajuste o modelo de regressão linear simples aos dados e interprete o coeficiente de regressão.

d) Calcule os resíduos do modelo.

e) Com base na análise dos resíduos do modelo ajustado no item b, decida sobre a adequabilidade do modelo.

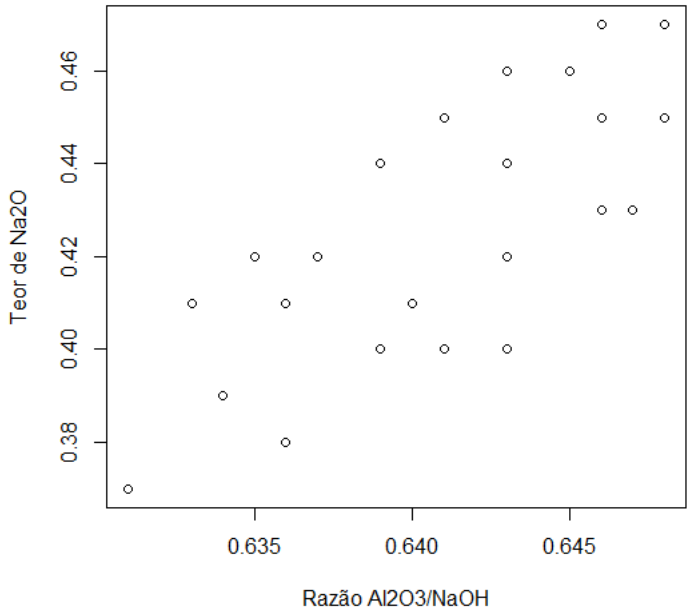
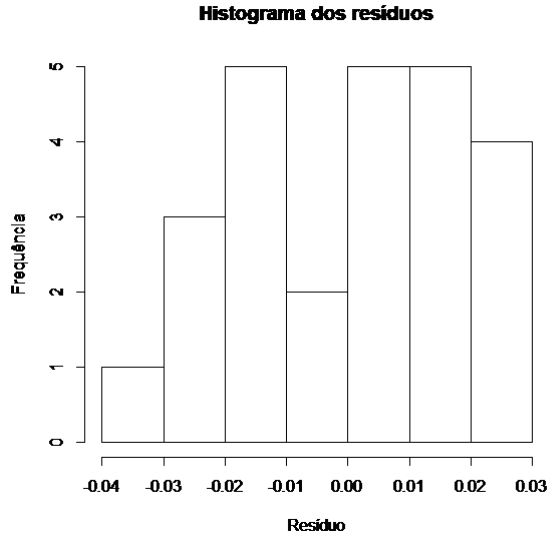
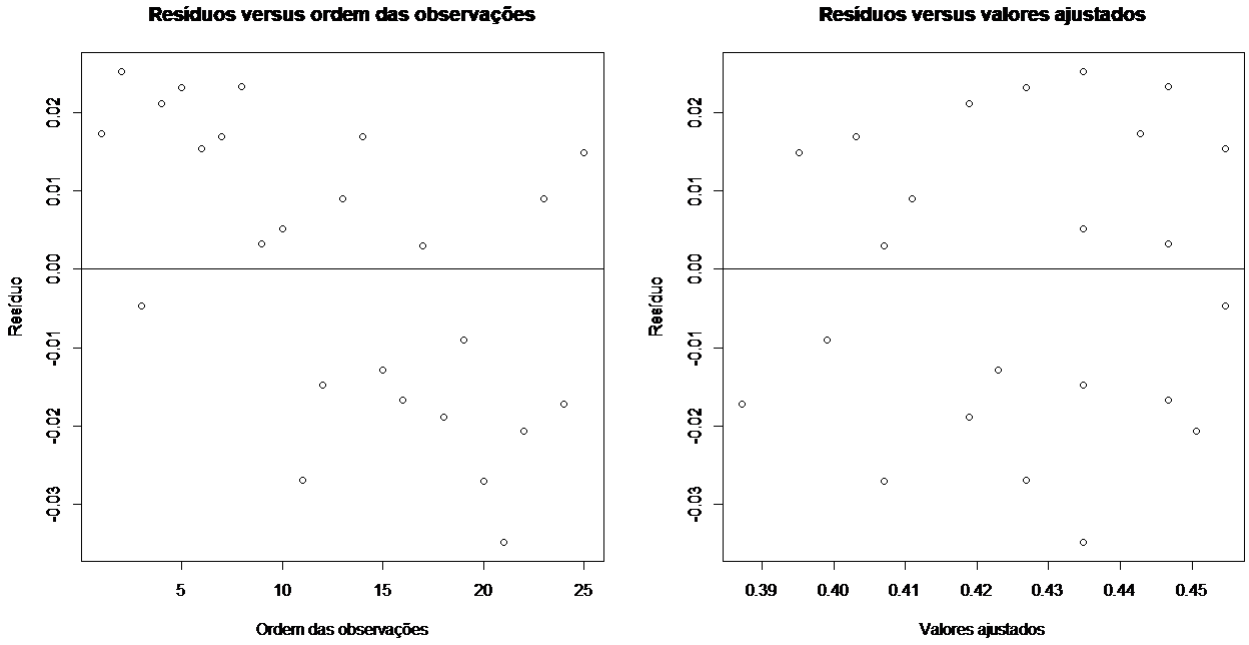


Figura: Diagrama de Dispersão: Teor de Na_2O ocluído na Alumina em Função da Razão Al_2O_3 / NaOH





R= Resp. r = 0,7321; Teor Na_2O (%) = - 2.12 + 3.97 Razão Al_2O_3 / NaOH

16 Testes de Aderência (ou Testes de Bondade de Ajustamento)

No conteúdo apresentado na apostila da segunda unidade foi admitido que a variável aleatória de interesse tivesse uma determinada distribuição de probabilidade. O problema era relacionado a ter um ou mais parâmetros desconhecidos, associado a uma distribuição de probabilidade conhecida (ou aproximada). Entretanto, pode acontecer de termos observações de uma variável aleatória e não se ter a menor idéia de sua distribuição de probabilidade. Neste caso, uma das formas iniciais de análise é construir um gráfico (colunas, histograma ou *boxplot*, etc.) com os valores da variável cuja distribuição na população é desconhecida para tentar entender o comportamento desta variável. E, em seguida, sugerir um modelo adequado para os dados. O modelo probabilístico proposto pode ser testado através do Teste de Aderência. A idéia básica é que, dada uma amostra aleatória de tamanho n, observada de uma variável aleatória X, nosso objetivo é testar:

$H_0 : X$ tem distribuição f

$H_1 : X$ não tem distribuição f.

A distribuição f nas hipóteses pode ser, por exemplo, Normal, Exponencial, Poisson ou qualquer outra distribuição.

Na literatura, existem várias maneiras de realizar os Testes de Aderência, porém neste texto será apresentado apenas o Teste Qui-Quadrado (χ^2).

16.1 Teste de Qui-Quadrado (χ^2) de Aderência

O teste de Qui-quadrado de Aderência é utilizado para comparar se as freqüências observadas da variável de interesse obtida na amostra aleatória diferem muito das freqüências esperadas. Estas, geralmente, sendo especificadas por uma distribuição de probabilidade.

Considere n observações independentes de uma variável aleatória X com função de distribuição não especificada. Cada observação é classificada em uma das k categorias, de forma que a seguinte tabela de contingência pode ser construída.

Variável	Categorias				
	1	2	3	...	k
Freqüência Observada	O_1	O_2	O_3	...	O_k

Na tabela acima, Oi representa a freqüência observada na célula i, para i= 1, 2, 3,..., k.
As hipóteses estatísticas a serem testadas são:
H_0: A variável X segue o modelo proposto;
H_1: A variável X não segue o modelo proposto.

A estatística de teste é dada por:

$$\chi^2_{cal} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \approx \chi^2_v$$

em que χ^2 tem distribuição aproximadamente Qui-Quadrado com υ graus de liberdade, supondo que a hipótese nula seja verdadeira, e:
k: é o número de categorias;
O_i: frequência observada na i-ésima categoria;
E_i: frequência esperada na i-ésima categoria, dada por $E_i = np_i$;
p_i = probabilidade da categoria i, supondo que H_0 é verdadeira.
υ = k-1 se as frequências esperadas puderem ser calculadas sem precisar estimar os parâmetros da distribuição.

Para um dado nível de significância α, rejeitar a hipótese nula se $\chi^2_{cal} > \chi^2_{\alpha;v}$, em que $\chi^2_{\alpha;v}$ é uma constante tal que $P(\chi^2_v > \chi^2_{\alpha;v}) = \alpha$. Ou pelo p-valor, rejeitar a hipótese nula

se p-valor $< \alpha$. Vale mencionar que a estatística de teste apresentada acima tem distribuição aproximadamente Qui-Quadrado e esta aproximação torna-se satisfatória quando as frequências esperadas são grandes. Para assegurar que esta aproximação seja boa é necessário levar em consideração as seguintes observações adiante:

1. Quando o número de categorias for igual a dois (k=2) as frequências esperadas dentro de cada categoria devem ser iguais ou superiores a 5.

2. Quando k $>$ 2, não deve ter mais de 20% das categorias com frequências esperadas menores que 5 e nenhuma frequência esperada igual a zero.

3. Quando as categorias apresentarem pequenas frequências esperadas elas podem ser combinadas com outras categorias, de tal forma que o sentido do trabalho seja conservado.

Quando desejamos testar se uma variável segue um determinado modelo, mas são desconhecidos um ou mais parâmetros da distribuição, devemos primeiro estimá-los de forma apropriada. Nestes casos, υ = k-m-1, em que m é o número de parâmetros que precisam ser estimados.

Exemplo 15.1: (Adaptado de Magalhães & Lima, 2006) Deseja-se verificar a afirmação de que a porcentagem de cinzas contidas em carvão, produzido por uma empresa, segue distribuição Normal. Os dados, apresentados a seguir, representam a quantidade percentual de cinzas encontradas em 250 amostras de carvão analisadas em laboratório. Qual decisão deve-se tomar ao nível de significância de 2,5%?

i	**Cinzas (em %)**	**Número de observações**
1	09,5|− 10,5	2
2	10,5|− 11,5	5
3	11,5|− 12,5	16
4	12,5|− 13,5	42
5	13,5|− 14,5	69
6	14,5|− 15,5	51
7	15,5|− 16,5	32
8	16,5|− 17,5	23
9	17,5|− 18,5	9
10	18,5|− 19,5	1

Solução: A média e a variância, da distribuição Normal que será testada, são desconhecidas, precisamos obter suas estimativas a partir da amostra. Os melhores estimadores para os parâmetros μ *e* σ^2 são a média amostral ($\bar{X}$) e a variância amostral (S^2), respectivamente. Calculando esses valores temos que

$$\bar{X} = \frac{\sum_{i=1}^{10} x_i f_i}{\sum_{i=1}^{10} f_i} = \frac{10x2 + 11x5 + 12x16 + ... + 19x1}{250} \cong 14,5$$

e

$$s^2 = \frac{\sum_{i=1}^{10}(x_i - \bar{x})^2 f_i}{(\sum_{i=1}^{10} f_i) - 1} = 2,7$$

Seja a variável aleatória X: porcentagem de cinzas contidas no carvão produzido pela empresa. As hipóteses a serem testadas são:

H_0 : A porcentagem de cinzas contidas no carvão segue distribuição Normal.

H_1 : A porcentagem de cinzas contidas no carvão não segue distribuição Normal.

As diversas faixas que constituem as categorias de valores da variável X serão enumeradas de 1 a 10. De modo a varrer os valores do intervalo $(-\infty, \infty)$, correspondentes ao modelo Normal, acrescentando às categorias 1 e 10 os valores, respectivamente, menores que 9,5 e maiores que 19,5. Dessa forma, para calcular as frequências esperadas, procedemos da seguinte forma, por exemplo, para categoria 1,

$$E_1 = 250P(X < 10,5) = 250P(Z < \frac{10,5 - 14,5}{\sqrt{2,7}}) =$$

$$250P(Z < -2,43) = 1,875$$

Para categoria 2

$$E_2 = 250P(10,5 < X < 11,5) =$$

$$250P(\frac{10,5 - 14,5}{\sqrt{2,7}} \leq Z < \frac{11,5 - 14,5}{\sqrt{2,7}}) =$$

$$250P(-2,43 < Z < -1,83) = 6,525$$

Para as categorias de 3 a 9, são calculados de forma análoga. A última categoria,

$$E_10 = 250P(X > 18,5) = 250P(Z > \frac{18,5 - 14,5}{\sqrt{2,7}}) =$$

$$250P(Z > 2,43) = 1,875$$

As probabilidades calculadas anteriormente supõem que H_0 é verdadeira, assim foi usada

a tabela da Normal Padrão.

As frequências esperadas são apresentadas na tabela, a seguir, e devem somar 250, o que não foi possível devido aos arredondamentos efetuados.

Categorias	**Frequência observada**	**Frequência esperada**
1	2	1,875
2	5	6,525
3	16	19,400
4	42	39,925
5	69	57,275
6	51	57,275
7	32	39,925
8	23	19,400
9	9	6,525
10	1	1,875

Observamos que exatamente 20% das categorias apresentaram frequências inferiores a 5, as categorias 1 e 10. Efetuando o cálculo da estatística de teste, temos

$$\chi^2_{cal} = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} + \frac{(2-1,875)^2}{1,875} + \frac{(5-6,525)^2}{6,525} + ... + \frac{(1-1,875)^2}{1,875} = 7,74$$

Para determinar a região crítica, utilizamos a distribuição Qui-Quadrado com 10-1-2=7 graus de liberdade, pois perdemos dois graus de liberdade devido à estimação dos parâmetros μ *e* σ^2 . Com auxílio da tabela da Qui-Quadrado, obtemos $\chi^2_{7;2,5\%} = 16,01$.

Conclusão: Como $\chi^2_{cal} = 7,74 < \chi^2_{7;2,5\%} = 16,01$ (valor tabelado, ver Tabela 1 em Anexo), logo não rejeitamos a hipótese nula, ou seja, não existem evidências para rejeitar a hipótese de que a amostra é proveniente de uma variável aleatória com distribuição normal, ao nível de significância de 2,5%.

Exercícios de fixação

1. O quadro abaixo se refere ao número de acidentes sofridos por um grupo de mineiros durante um trabalho numa mina de carvão. O interesse é investigar se a distribuição do número de acidentes segue o modelo Poisson ($\lambda = 1,45$). (Use $\alpha = 5\%$).

Número de acidentes	0	1	2	3	4	5
Número de mineiros	35	47	39	20	5	2

R: Modelo não é rejeitado, Região crítica [11,07; ∞) e $\chi^2_{cal} = 1{,}0833$.

2. Uma indústria registra, em cada semana, o número de dias em que ocorrem acidentes de trabalho. Para uma amostra de 200 semanas, verifique se os dados apresentados a seguir, aderem ao modelo Binomial com parâmetros n = 5 e p = 0,2. (Use $\alpha = 10\%$).

Número de dias com acidentes	0	1	2	3	4	5
Freqüência	64	56	40	24	8	8

R: Modelo é rejeitado, Região crítica [7,78; ∞) e $\chi^2_{cal} = 189{,}2$.

16.2 Outros testes de normalidade

Na literatura existem outros procedimentos para realizar o Teste de Aderência, a saber: o Teste Kolmogorov- Smirnov, o Teste de Shapiro-Wilk para Normalidade e o Teste de Lilliefors para Normalidade. Para maiores detalhes sobre estes testes consultar as seguintes referências: Siegel & Castellan (2006), Campos (1979), Conover (1999) e Hollander & Wolf (1999).

16 Comparação de Médias Populacionais

Na apostila da segunda unidade foi apresentado teste de hipóteses para apenas uma única média. No entanto, não é raro encontrar situações em que se deseja verificar se há diferenças significativas entre as médias de k populações distintas.

A análise usada para comparação de k médias populacionais ou de tratamentos é comumente realizada por uma Análise de Variância (ANOVA). Grande parte da teoria de Análise de Variância foi desenvolvida por um grupo de pesquisadores estatísticos que trabalhou na Estação Experimental de Agricultura de Rothamstead, na Inglaterra. As análises destes experimentos agronômicos desenvolvidos por estes pesquisadores, atualmente, se aplicam na maioria das áreas de conhecimento, a saber: engenharia, medicina, educação, psicologia, economia, odontologia, dentre outras. De qualquer forma, é a origem agrícola das ciências experimentais que explica o uso de alguns termos técnicos que serão apresentados adiante.

Alguns termos técnicos utilizados em Planejamento de Experimentos e Análise de Variância.

- **Fator e Nível**

  Fator é uma variável independente obtida quando é realizado um estudo de investigação e o nível é a forma particular deste fator. Por exemplo, em um estudo sobre os efeitos da presença de três tipos de diferentes soluções de açúcar (glicose, sacarose e frutose)

no crescimento de bactérias, o fator é o açúcar e cada tipo de solução é um nível em estudo. Neste caso, o fator açúcar tem três níveis (glicose, sacarose e frutose). Considere outro exemplo, um fabricante de papel, usado para a confecção de sacolas de mercearia, realiza um experimento para investigar se a concentração de madeira de lei em polpa (5%, 10%, 15% e 20%) tem efeito sobre a resistência à tração das sacolas fabricadas da polpa. A concentração de madeira de lei é o fator sob estudo e os níveis são as diferentes aplicações, diz-se que o fator concentração de madeira de lei tem quatro níveis (5%, 10%, 15% e 20%). No primeiro exemplo, o fator é de natureza qualitativa, ou seja, é um fator em que os níveis não podem ser arranjados em ordem crescente de magnitude. No segundo exemplo, o fator é de natureza quantitativa, ou seja, é um fator em que os níveis podem ser associados a pontos na escala aritmética.

- **Tratamento**

  Um tratamento é uma condição imposta ou objeto que se deseja medir ou avaliar em um experimento. Em outras palavras, denomina-se de tratamento, o nível de um fator sob análise ou uma combinação de fatores e níveis em estudo com dois ou mais fatores. Por exemplo, se o interesse é estudar os efeitos de cinco diferentes marcas de gasolina na eficiência operacional (milhas/galão) de motores de automóvel, o fator é a marca e cada marca constitui um tratamento. Em um estudo para comparar duas diferentes marcas de canetas (A e B) e dois diferentes tipos de lavagem (1 e 2) em relação à capacidade de remover manchas em um determinado tipo de tecido, existem 4 combinações possíveis, a saber: marca A e lavagem 1, marca A e lavagem 2, marca B e lavagem 1 e, marca B e lavagem 2. Cada uma destas combinações é chamada de tratamento, de modo que há 4 tratamentos diferentes envolvidos

- **Unidade experimental**

  A aplicação do tratamento é feita na unidade experimental que fornece os dados para serem avaliados. Dependendo do experimento, a unidade experimental pode ser um motor, uma peça do motor, uma porção de algum alimento, um vaso, um animal, um indivíduo, etc.

  As unidades experimentais podem ser formadas por grupos ou indivíduos, cujo uso depende do fenômeno que se está estudando, da forma como o experimento é conduzido e dos recursos disponíveis. De modo geral, a escolha da unidade experimental deve ser feita de forma a minimizar o erro experimental e representar satisfatoriamente o processo de estudo.

- **Repetição**

  Repetição é o número de vezes que um tratamento aparece no experimento. O número de repetições, em um experimento, vai depender também dos recursos disponíveis, do delineamento do experimento e, também, da variabilidade do experimento ou da variável resposta. Existem várias metodologias para estimar o número satisfatório de repetições em um experimento. Mas, em função das possíveis limitações citadas acima, a definição do número de repetições, na maioria vezes, depende da experiência do pesquisador sobre o fenômeno em estudo. Além disso, as metodologias empregadas, para esse cálculo, pressupõem que uma estimativa do erro experimental seja conhecida. Para calcular o número de repetições (ou tamanho da amostra) que deve ser usado no experimento consultar a referência Dean & Voss (1999).

16.1 Análise de Variância

Suponha um procedimento experimental com k tratamentos (populações) ou diferentes níveis de um único fator. A variável resposta para cada k tratamento é uma variável aleatória. Na tabela de dados (Tabela 16.1), y_{ij} é a observação da j-ésima unidade experimental no i-ésimo tratamento ou fator. Existem n observações no i-ésimo tratamento. Inicialmente, a análise de variância será apresentada para o caso em que as amostras em cada tratamento (ou população) têm o mesmo tamanho, neste caso é conhecido como dados balanceados.

Tabela 16.1: Dados para experimento com um único fator

Tratamento(Nível)	Observações				Total	Média
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\overline{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\overline{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kn}	$y_{k.}$	$\overline{y}_{k.}$
					$y_{..}$	$\overline{y}_{..}$

Em que $y_{i.}$ representa a soma total das observações do i-ésimo tratamento, $\overline{y}_{i.}$ representa a média das observações do i-ésimo tratamento, $y_{..}$ a soma de todas as observações e $\overline{y}_{..}$ representa a média de todas as observações, denominada média global amostral. Simbolicamente expressos por:

$$y_{i.} = \sum_{j=1}^{n} y_{ij} \text{ e } \overline{y}_{i.} = \frac{y_{i.}}{n},\ i = 1, 2, \ldots, k.$$

$$y_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij} \text{ e } \overline{y}_{..} = \frac{y_{..}}{n}$$

em que N=n×k é o número total de observações. Observe que o "ponto" subscrito na notação matemática representa a soma.

Assim, suponha k tratamentos (ou populações) cada um com n repetições e os valores numéricos das observações representados por y_{ij}. Um modelo para descrever os dados é

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \ldots, k \quad \text{e} \quad j = 1, \ldots, n, \tag{5}$$

em que: y_{ij} é a observação do i-ésimo tratamento na j-ésima unidade experimental; μ_i é a média do i-ésimo nível do fator ou tratamento, sendo um valor fixo e desconhecido, ϵ_{ij} é o erro aleatório associado ao i-ésimo tratamento na j-ésima unidade experimental assumido como: $\epsilon_{ij} \sim N(0; \sigma^2)$, independentes e identicamente distribuído. A variância σ^2 é assumida como constante para todos nos níveis de fator. Isto implica que $y_{ij} \sim N(\mu_i; \sigma^2)$. Assim, μ_i é a parte sistemática que representa a média da população i, que é fixa, e ϵ_{ij} é a parte aleatória, a informação referente a outros fatores que podem influenciar as observações, mas não são incorporadas em μ_i.

A equação (5) é denominada modelo μ, porque ele usa as médias $\mu_1, \mu_2, \ldots, \mu_k$ como parâmetros básicos na expressão matemática do modelo. Uma forma alternativa para escrever o modelo (5) para os dados é

$$\mu_i = \mu + \tau_i, \quad i = 1, \ldots, k. \tag{6}$$

E a equação (6) acima torna-se

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \ldots, k \quad \text{e} \quad j = 1, \ldots, n. \tag{7}$$

Nesta forma de modelo, μ é o parâmetro média comum a todos os tratamentos, chamado de média global, e τ_i é o parâmetro do i-ésimo tratamento, denominado efeito do tratamento. Os modelos (5) e (7) são também denominados de Análise de Variância de fator único (ANOVA) porque apenas um único fator é investigado. Além disso, será necessário que a alocação do material experimental às diversas condições experimentais seja aleatória e que o meio em que os tratamentos sejam aplicados (chamado de unidades experimentais) seja tão uniforme quanto possível. Assim, o planejamento experimental é denominado de completamente aleatorizado. O objetivo será o de testar hipóteses apropriadas sobre as médias dos tratamentos.

A análise dos efeitos dos tratamentos pode ser feita de duas maneiras. Na primeira, os

tratamentos podem ser escolhidos de acordo com o interesse do pesquisador. Nesta situação, as inferências extraídas serão aplicáveis e restritas somente aos níveis de fator considerados na análise, não podendo ser estendidos a outros níveis não investigados. Sob estas condições, o modelo (7) é denominado de modelo de efeitos fixos. Já quando os tratamentos analisados representam uma amostra aleatória de uma população de níveis de fator ou fatores, podem-se estender as conclusões da análise feitas para essa amostra, para todos os outros tratamentos da população, nesse caso tem-se análise de um modelo de efeitos aleatórios. Considere, por exemplo, que foram selecionadas três máquina de uma população de 75 máquinas distribuídas numa fábrica e suas produções foram medidas por um período de 10 dias. As três máquinas constituem três níveis do fator em estudo, porém, o interesse nas conclusões não se restringe apenas àquelas três nas quais os dados foram mensurados, mas a todas as máquinas da fábrica.

A análise de um modelo de efeitos aleatórios não será abordado nesta apostila e o leitor interessado poderá consultar as seguintes referências: Montgomery (2005), Neter (1974) e Peter & Wasserman (1970).

Análise de um modelo com efeitos fixos

Considere um experimento completamente aleatorizado e que a análise de variância será para um único fator com efeitos fixo. O interesse é testar a igualdade média dos tratamentos. Assim, as hipóteses apropriadas são:

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_{\mathrm{k}}$

$H_1 : \mu_{\mathrm{i}} \neq \mu_{\mathrm{j}}$ para algum i e algum j tais que $i \neq j$.

A hipótese nula supõe que as observações amostrais dentro de cada tratamento podem ser vistas como provenientes de populações com médias iguais. Reescrevendo $\mu_{\mathrm{i}} = \mu + \tau_{\mathrm{i}}, i = 1, \ldots, k$. A média μ é a média geral calculada da seguinte forma:

$$\mu = \frac{\sum\limits_{\mathrm{i}\,=\,1}^{\mathrm{k}} \mu_{\mathrm{i}}}{k}$$

Implicando que $\sum\limits_{\mathrm{i}\,=\,1}^{\mathrm{k}} \tau_{\mathrm{i}} = 0$. Consequentemente, é possível reescrever as hipóteses acima em termos dos efeitos dos tratamentos, ou seja,

$H_0 : \tau_1 = \tau_2 = \ldots = \tau_{\mathrm{k}} = 0$

$H_1 : \tau_{\mathrm{i}} \neq 0$ para algum i.

Então, podemos testar a igualdade de médias de tratamentos ou testar se os efeitos dos tratamentos τ_{i} são iguais a zero.

Ao realizar a análise de variância, a idéia básica é de que existe uma distribuição de probabilidade para a variável resposta (dependente (y_{ij})) em cada nível do fator. Para efeito

de inferências sobre o modelo (16.2) é necessário assumir que:

i. y_{ij} são variáveis aleatórias independentes

ii. y_{ij} tem distribuição normal com média μ_1, $i = 1, \ldots, k$ e $j = 1, \ldots, n$

iii. $\mathrm{Var}(y_{\mathrm{ij}}) = \sigma^2, i = 1, \ldots, k$ e $j = 1, \ldots, n$, ou seja, todas as k populações devem ter variâncias homogêneas ($\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_{\mathrm{k}}^2 = \sigma^2$), propriedade conhecida como Homocedasticidade. Em outras palavras, a variância σ^2 deve ser constante para todos nos níveis de fator.

Decomposição da soma total de quadrado

O termo análise de variância pode induzir a um equívoco, uma vez que a finalidade é investigar diferenças entre médias dos tratamentos, e não diferenças significativas entre as variâncias dos grupos. O nome análise de variância é atribuído devido a uma decomposição da variabilidade total das suas componentes.

A soma total de quadrado é dada por:

$$\sum_{\mathrm{i}\,=\,1}^{\mathrm{k}} \sum_{\mathrm{j}\,=\,1}^{\mathrm{n}} (y_{\mathrm{ij}} - \overline{y}_{\mathrm{..}})^2$$

é uma medida de variabilidade total dos dados. Esta soma pode ser subdividida em duas partes da seguinte forma:

$$\sum_{\mathrm{i}\,=\,1}^{\mathrm{k}} \sum_{\mathrm{j}\,=\,1}^{\mathrm{n}} (y_{\mathrm{ij}} - \overline{y}_{\mathrm{..}})^2 = n \sum_{\mathrm{i}\,=\,1}^{\mathrm{n}} (\overline{y}_{\mathrm{i.}} - \overline{y}_{\mathrm{..}})^2 + \sum_{\mathrm{i}\,=\,1}^{\mathrm{k}} \sum_{\mathrm{j}\,=\,1}^{\mathrm{n}} (y_{\mathrm{ij}} - \overline{y}_{\mathrm{i.}})^2 \tag{8}$$

O termo do lado esquerdo de (8) é a soma dos quadrados das observações em relação à média global e representa uma medida da variabilidade total dos dados, denotada por SS_{T}. O primeiro termo do lado direito de (8) é a soma dos quadrados das diferenças entre as médias de cada tratamento e a média global (ou seja, aquela decorrente das diferenças entre os grupos de tratamentos), sendo denotada por $SS_{\mathrm{Tratamento}}$. O segundo termo do lado direito de (8) é a soma de quadrados das diferenças de cada observação dentro dos tratamentos em relação à média do tratamento (ou seja, aquela decorrente da variação dentro do grupo), sendo denotado por SS_{E}. Em outras palavras, $SS_{\mathrm{Tratamento}}$ é a soma de quadrados devido ao tratamento (ou seja, entre tratamentos), e SS_{E} é a soma de quadrados residual (ou seja, dentro dos tratamentos). Simbolicamente, podemos representar a equação (8) por:

$$\mathrm{SS_T} = SS_{\mathrm{Tratamento}} + SS_{\mathrm{E}} \tag{9}$$

Considere o segundo termo do lado direito da expressão (9)

$$SS_{\mathrm{E}} = \sum_{\mathrm{i}\,=\,1}^{\mathrm{k}} \left[\sum_{\mathrm{j}\,=\,1}^{\mathrm{n}} (y_{\mathrm{ij}} - \overline{y}_{\mathrm{i.}})^2 \right]$$

Observe que a soma dentro do colchete dividido por (n-1) é a variância amostral do i-ésimo tratamento, ou seja,

$$S_{\mathrm{i}}^2 = \frac{\sum_{\mathrm{j}\,=\,1}^{\mathrm{n}} (y_{\mathrm{ij}} - \overline{y}_{\mathrm{i.}})^2}{n-1}, i = 1, 2, \ldots, k$$

As variâncias amostrais podem ser combinadas para encontrar um estimador da variância populacional, σ^2, como se segue

$$\frac{(n-1)S_1^2 + (n-1)S_2^2 + \ldots + (n-1)S_{\mathrm{k}}^2}{(n-1) + (n-1) + \ldots + (n-1)} = \frac{\sum_{\mathrm{i}\,=\,1}^{\mathrm{k}} \left[\sum_{\mathrm{j}\,=\,1}^{\mathrm{n}} (y_{\mathrm{ij}} - \overline{y}_{\mathrm{i.}})^2 \right]}{\sum_{\mathrm{i}\,=\,1}^{\mathrm{k}} (n-1)} = \frac{SS_{\mathrm{E}}}{N-K}$$

em N=n×k. Assim, $\frac{SS_{\mathrm{E}}}{N-k}$ é uma média ponderada das k variâncias individuais dentro de cada um dos tratamentos. De forma análoga, a expressão

$$\frac{SS_{\mathrm{Tratamento}}}{k-1} = \frac{n \sum_{\mathrm{j}\,=\,1}^{\mathrm{n}} (\overline{y}_{\mathrm{i.}} - \overline{y}_{\mathrm{..}})^2}{k-1}$$

é um estimador de σ^2, se não existe diferença entre as k médias dos tratamentos. Portanto, a quantidade $\frac{SS_{\mathrm{E}}}{N-K}$ é uma estimador de σ^2 e se não existe diferença entre as k médias dos tratamentos $\frac{SS_{\mathrm{Tratamento}}}{k-1}$também é um estimador de σ^2.

Análise de Variância pode ser resumida através da Tabela 16.2 adiante. Esta tabela pode ser utilizada para testar as seguintes hipóteses:

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_{\mathrm{k}}$

$H_1 : \mu_{\mathrm{i}} \neq \mu_{\mathrm{scriptsize\ j}}$ para algum i e algum j tais que $i \neq j$.

Para testar estas hipóteses de interesse, será usando a estatística

$$F = \frac{\frac{SS_{\mathrm{Tratamento}}}{k-1}}{\frac{SS_{\mathrm{E}}}{\mathrm{N\ -\ K}}} = \frac{MS_{\mathrm{Tratamento}}}{MS_{\mathrm{E}}}.$$

Supondo que a hipótese nula é verdadeira e que o erro $\epsilon_{\mathrm{ij}} \sim \mathrm{N}(0; \sigma^2)$, é possível mostrar que F tem distribuição de Fisher-Snedecor com (k-1) e (N-k) graus de liberdade que correspondem ao numerador e ao denominador, respectivamente.

Tabela 16.2: Análise de variância

Fonte de variação	**Graus de liberdade**	**Soma de quadrados**	**Quadrados médios**	**F calculado**
Entre tratamentos	k-1	$SS_{\text{Tratamento}}$	$MS_{\text{Tratamento}}$	$\frac{MS_{\text{Tratamento}}}{MS_{\text{E}}}$
Resíduo (dentro do tratamento)	N-k	SS_{E}	MS_{E}	
Total	N-1	SS_{T}		

$MS_{\text{Tratamento}} = \frac{SS_{\text{Tratamento}}}{k-1}$ e $MS_{\text{E}} = \frac{SS_{\text{E}}}{\text{N - K}}$.

Supondo que a hipótese nula é verdadeira, tanto $MS_{\text{Tratamento}}$ quanto MS_{E} estimam a variância comum σ^2 e espera-se que $F_{\text{calculado}}$ seja aproximadamente 1. Se há diferença entre os tratamentos, a variância entre os tratamentos excede a de dentro dos tratamentos e espera-se que $F_{\text{calculado}}$ seja maior que 1. Consequentemente, quando utiliza o procedimento de ANOVA, rejeita-se a hipótese de nula H_0 em favor de H_1, a um nível de significância α, se $F_{\text{calculado}} > F_{[\alpha;\ \text{(k-1)};\ \text{(N-k)}]}$, ou seja, existem evidências de diferença significativa entre pelo menos um par de médias de tratamentos. Caso contrário, não rejeitamos a hipótese H_0, ou seja, não há evidências de diferença significativa entre tratamentos, ao nível α de significância escolhido.

Outra maneira de avaliar a significância da estatística F é utilizando o p-valor. Se o p-valor$< \alpha$, rejeitamos a hipótese H_0. Caso contrário, não rejeitamos a hipótese de nulidade H_0, ou seja, não há evidências de diferenças significativas entre os tratamentos, ao nível α de significância escolhido.

Quando as somas de quadrados forem calculadas manualmente, elas podem ser obtidas através das seguintes expressões dadas adiante.

$$SS_{\text{T}} = \sum_{\text{i}\,=\,1}^{\text{k}} \sum_{\text{j}\,=\,1}^{\text{n}} y_{\text{ij}}^2 - \frac{y_{\cdot\cdot}^2}{N}, \tag{10}$$

$$SS_{\text{Tratamento}} = \frac{\sum_{\text{i}\,=\,1}^{\text{k}} y_{\text{i}\cdot}^2}{n} = \frac{y_{\cdot\cdot}^2}{N}, \tag{11}$$

e

$$SS_{\text{E}} = SS_{\text{T}} - SS_{\text{Tratamento}} \tag{12}$$

Os quadrados médios dos resíduos e dos tratamentos são obtidos dividindo as somas de

quadrados pelos correspondentes graus de liberdade, ou seja,

$$MS_{\mathrm{E}} = \frac{SS_{\mathrm{E}}}{N-K}$$

e

$$MS_{\mathrm{Tratamento}} = \frac{SS_{\mathrm{Tratamento}}}{k-1}$$

Dados desbalanceados

Em alguns experimentos de um único fator o número de observações obtidas dentro de cada tratamento pode ser diferente. Neste caso, é mencionado que os dados são desbalanceados. A análise de variância, descrita acima, ainda pode ser usada, mas pequenas modificações devem ser realizadas nas fórmulas das somas de quadrados. Considere que n_{i} observações são realizadas no tratamento i (i = 1, ..., k) e o número total de observações nos k grupos é igual a $\mathrm{N} = \sum_{\mathrm{i}=1}^{\mathrm{k}} n_{\mathrm{i}}$. As fórmulas (10) a (12) tornam-se:

$$SS_{\mathrm{T}} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \sum_{\mathrm{j}=1}^{n_{\mathrm{i}}} y_{\mathrm{ij}}^2 - \frac{y_{..}^2}{N},$$

$$SS_{\mathrm{Tratamento}} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \frac{y_{\mathrm{i.}}^2}{n_{\mathrm{i}}} - \frac{y_{..}^2}{N},$$

e

$$SS_{\mathrm{E}} = SS_{\mathrm{T}} - SS_{\mathrm{Tratamento}}$$

As médias geral e dos grupos são dados por:

$$\overline{y}_{..} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{k}} \sum_{\mathrm{j}=1}^{n_{\mathrm{i}}} y_{\mathrm{ij}}}{K}$$

$$\overline{y}_{\mathrm{i.}} = \frac{\sum_{\mathrm{j}=1}^{n_{\mathrm{i}}} y_{\mathrm{ij}}}{n_{\mathrm{i}}},\ \mathrm{i}=1,2,\ldots,\mathrm{k}.$$

Nenhuma outra alteração é necessária para a análise de variância. Segundo Montgomery (2005) há duas desvantagens na escolha de um experimento com dados desbalanceado. Primeira desvantagem, a estatística de teste é relativamente sensível a pequenos desvios da suposição de homogeneidade variância quando os tratamentos têm números de observações diferentes em cada amostra. Caso contrário ocorre quando os tamanhos das amostras são iguais nos tratamentos. Segunda desvantagem, a potência do teste diminui se as amostras são de tamanhos desiguais.

16.2 Teste de Tuckey

O procedimento seguinte quando se rejeita a hipótese nula na análise de variância é o de comparar as médias de tratamentos utilizando algum teste de comparação de médias ou contrastes para identificar qual(is) tratamento(s) é (são) diferente(s). Existem vários procedimentos para realizar comparações múltiplas de médias, e alguns deles podem ser vistos em Montgomery (2005). Aqui será apresentado apenas um deles.

O teste de Tukey permite testar qualquer contraste, sempre, entre duas médias de tratamentos. Nesse caso, as hipóteses estatísticas são:

$$H_0 : \mu_{\mathrm{i}} = \mu_{\mathrm{j}},$$
$$H_1 : \mu_1 \neq \mu_{\mathrm{j}}$$

para todo $i \neq j$. O teste proposto por Tukey baseia-se na diferença significante HSD = Δ, denominada de Honestly Significant Difference. Esta diferença, para dados balanceados, é dada da seguinte forma:

$$\triangle_\alpha = q_\alpha(k; f)\sqrt{\frac{MS_{\mathrm{E}}}{n}}, \tag{13}$$

em que, f é o número de graus de liberdades associado a MS_{E}, q é a amplitude total studentizada (valor tabelado, ver Tabela 4 em Anexo) e MS_{E} é o quadrado médio dos resíduos. O valor de q depende do número de tratamentos e do número de graus de liberdade associada com a soma de quadrados dos resíduos. Também, em um teste de comparações de médias, deve-se determinar um nível de significância α para o teste. Normalmente, utiliza-se o nível de 5% ou 1% de significância.

Como o teste de Tukey é, de certa forma, independente do teste F , é possível que, mesmo sendo significativo o valor de $F_{\mathrm{calculado}}$, não se encontrem diferenças significativas entre as médias.

As duas médias, $\mu_{\mathrm{i}} e \mu_{\mathrm{j}}$, (i$\neq$ j), são consideradas significantemente diferentes se

$$|\overline{y}_{\mathrm{i.}} - \overline{y}_{\mathrm{j.}}| > \Delta_\alpha$$

Quando os dados são desbalanceados, o teste de Tukey descrito acima apresenta a seguinte modificação na equação (13)

$$\Delta_\alpha = \frac{q_\alpha(k; f)}{\sqrt{2}}\sqrt{MS_{\mathrm{E}}\left(\frac{1}{n_{\mathrm{i}}} + \frac{1}{n_{\mathrm{j}}}\right)}, \quad i \neq j \tag{14}$$

Exemplo 16.1: (Montgomery, Goldman e Borror, 2006) Um fabricante de papel usado para a confecção de sacolas de mercearia está interessado em melhorar a força de resistência

do produto. A engenharia de produção acha que a força de resistência é uma função da concentração de madeira de lei na polpa, e que a amplitude das concentrações da madeira de lei de interesse prático está entre 5% e 20%. Uma das engenheiras responsáveis pelo estudo decide investigar quatro níveis de concentração de madeira de lei: 5%, 10%, 15% e 20%. Ela decide, também, fazer seis repetições de teste de cada nível de concentração usando uma usina-piloto. Todos os 24 espécimes são testados em um testador de tração de laboratório, em ordem aleatória. Os dados desse experimento constam na Tabela 15.3. Esse é um exemplo de um experimento de fator único completamente aleatorizado, com quatro níveis do fator, ou seja, quatro tratamentos. E cada tratamento tem seis observações ou repetições.

Tabela 16.3: Força de resistência do papel (ψ)

Concentração de madeira de lei	**Repetição (ou observação)**						**Totais**	**Médias**
	1	**2**	**3**	**4**	**5**	**6**		
5%	7	8	15	11	9	10	60	10,00
10%	12	17	13	18	19	15	94	15,67
15%	14	18	19	17	16	18	102	17,00
20%	19	25	22	23	18	20	127	21,17
Total							**383**	**15,96**

É importante que se realize uma análise descritiva nos dados obtidos no experimento realizado. Na Tabela 16.3 nota-se que para a concentração de 5%, a resistência do papel foi, em média, menor. Conforme mostra a Tabela 16.4, observa-se que a menor e maior dispersão relativa ocorreu nas concentrações de 15% e 5% de madeiras, respectivamente. Na Figura 16.1, é possível visualizar que a força de resistência da sacola aumenta à medida que a concentrações de madeira de lei aumenta, ou seja, suspeita-se de que a mudança na concentração de madeira de lei tem um efeito na força de resistência da sacola. Também percebe-se um forte indicativo que a concentração de 5% difere da concentração de 20% no que diz respeito à resistência do papel, pois não há sobreposição dos gráficos de ambos os tratamentos. Além disso, há indicativo que os valores das resistências dos papeis nos quatros tratamentos são provenientes de distribuiçõs assimétricas.

O modelo de análise de variância adotado é dado por:

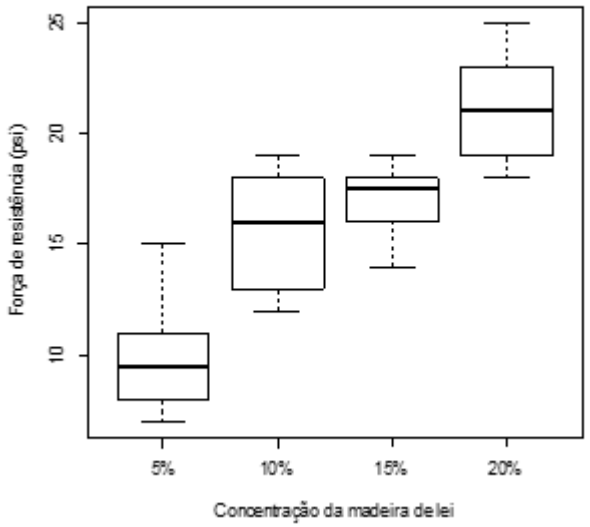
$$\mu_{\mathrm{i}} = \mu + \tau_{\mathrm{i}} + \epsilon_{\mathrm{ij}}, i = 1, \ldots, 4 \text{ e j} = 1, \ldots, 6$$

em que y_{ij} é a força de resistência do papel observada na j-ésima sacola para a i-ésima concentração de madeira de lei, μ é a média geral, τ_{i} é o efeito da i-ésima concentração de madeira de lei e ϵ_{ij} é o efeito do erro experimental suposto normal e independentemente distribuído com média 0 e variância comum σ^2.

Tabela 16.4: Medidas descritivas dos dados da força de resistência dos papeis para cada concentração de madeira de lei.

Medidas Descritivas	**Concentração de madeira de lei**			
	5%	**10%**	**15%**	**20%**
Mediana	9,5	16,0	17,5	21,0
Desvio-padrão	2,83	2,80	1,79	2,64
Coeficiente de variação	0,28	0,18	0,11	0,12
Mínimo	7,0	12,0	14,0	18,0
Máximo	15,00	19,00	19,00	25,00

Figura 16.1: Boxplot dos dados da força de resistência dos papeis para cada concentração de madeira de lei



Para comparar se as médias das forças de resistências do papel, para fabricação de sacolas, são diferentes quando é usado diferentes tipos de concentrações de madeira de lei, será usado a análise de variância. As hipóteses estatísticas a serem testadas são:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ (A força de resistência média do papel são as mesmas nas quatro concentrações de madeira de lei analisadas).

H_1: Pelos menos uma das médias,da força de resistência do papel,é diferente das demais.

A soma de quadrados para compor a Tabela da Análise de Variância é calculada conforme as equações (10) a (12),

$$SS_{\mathrm{T}} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \sum_{\mathrm{j}=1}^{n_{\mathrm{i}}} y_{\mathrm{ij}}^2 - \frac{y_{\cdot\cdot}^2}{N} = (7)^2 + (8)^2 + \ldots + (20)^2 - \frac{(383)^2}{24} = 625 - \frac{(383)^2}{24} = 512,96$$

$$SS_{\mathrm{Tratamento}} = \sum_{\mathrm{i}=1}^{\mathrm{k}} y_{\mathrm{i}\cdot}^2 - \frac{y_{\cdot\cdot}^2}{N} = \frac{1}{6}[(60)^2 + (94)^2 + \ldots + (127)^2] - \frac{(383)^2}{24} = 382,79$$

Uma vez calculadas as duas somas de quadrados, obtemos sem dificuldades a terceira

soma de quadrados conforme apresenta adiante:

$SS_{\text{E}} = 512,96 - 382,79 = 130,17$

Os resultados estão resumidos na Tabela da Análise de Variância a seguir.

Tabela 16.5: Análise de Variância para a força de resistência da sacolas

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	$F_{\text{calculado}}$	p-valor
Entre tratamentos	3	382,79	127,60	19,61	$3,59 \times 10^{-6}$
Resíduo (dentro do tratamento)	20	130,17	6,51		
Total	23	512,96			

Sendo $MS_{\text{Tratamento}} = \frac{SS_{\text{Tratamento}}}{k-1} = \frac{382,79}{3} = 127,60; MS_{\text{E}} = \frac{SS_{\text{E}}}{N-a} = \frac{130,17}{20} = 6,51$ e $F_{\text{calculado}} = \frac{MS_{\text{Tratamento}}}{MS_{\text{E}}} = \frac{127,60}{6,51} = 19,61$

Conclusão: Como $F_{\text{calculado}} = 19,61 > F_{0,05;\ 3;\ 20} = 3,03$ (valor tabelado, ver Tabela 2 em Anexo), rejeitamos H_0 e concluímos que a concentração da madeira de lei afeta a resistência do papel, ao nível de significância de 5%. Chegamos à mesma conclusão ao observarmos o p-valor = 0,00000359 < 0,05 = α. Portanto, pelo menos uma das médias de tratamento difere das demais.

Como o teste acima rejeitou a hipótese nula será aplicado o teste de Tukey para realizar as comparações múltiplas de médias nos quatro tratamentos. Lembrando que k = 4, n = 6, $MS_{\text{E}} = 6,51$ e $f = 20$. As médias amostrais dos tratamentos são:

$$\overline{y}_{1.} = 10,00\psi,\ \overline{y}_{2.} = 15,67\psi,\ \overline{y}_{3.} = 17,00\psi \text{ e } \overline{y}_{4.} = 21,17\psi.$$

Através da Tabela da Distribuição de Amplitude Total Studentizada, com $\alpha = 0,05$ encontramos o valor $q_{0,05}(4;20) = 3,96$ (ver Tabela 4 em Anexo). Calculando Δ_α (equação (16.9)), temos

$$\Delta_\alpha = q_{0,05}(4;20)\sqrt{\frac{MS_{\text{E}}}{n}} = 3,96\sqrt{\frac{6,51}{6}} = 4,12$$

Portanto, concluímos que as duas médias são significantemente diferentes se

$$|\overline{y}_{\text{i.}} - \overline{y}_{\text{j.}}| > 4,12$$

As diferenças nas médias dos tratamentos são:

$$|\overline{y}_{1.} - \overline{y}_{2.}| = |10,00 - 15,67| = 5,67^*,$$
$$|\overline{y}_{1.} - \overline{y}_{3.}| = |10,00 - 17,00| = 7,00^*,$$
$$|\overline{y}_{1.} - \overline{y}_{4.}| = |10,00 - 21,17| = 11,17^*,$$
$$|\overline{y}_{2.} - \overline{y}_{3.}| = |15,67 - 17,00| = 1,33,$$
$$|\overline{y}_{3.} - \overline{y}_{4.}| = |17,00 - 21,17| = 4,17^*,$$

Os asteriscos nos valores indicam que os pares de médias μ_{i} e μ_{j}, i $\neq$ j, são significantes. Portanto, com base no conjunto de dados analisados, há evidência de diferenças significativas entre todos os pares de médias, exceto entre os tratamentos 2 e 3, ao nível de significância mínimo de 5%.

16.3 Análise de diagnóstico básico em ANOVA

Ao realizar o teste de hipóteses para comparação de médias é necessários que sejam satisfeitas certas suposições. Especificadamente, que as observações sejam adequadamente descrita pelo modelo proposto

$$y_{\mathrm{ij}} = \mu + \tau_{\mathrm{i}} + \epsilon_{\mathrm{ij}}, i = 1, \ldots, k \text{ e j} = 1, \ldots, \mathrm{n}$$

em que $\epsilon_{\mathrm{ij}} \sim N(0; \sigma^2)$, independentes e identicamente distribuído. Se estas suposições são violadas, as inferências realizadas a partir da ANOVA são seriamente afetadas, ou seja, o teste F usado para testar as diferenças nas médias de tratamento pode não ser válido.

As estimativas dos erros recebem o nome de resíduos. Define-se o resíduo como:

$$\epsilon_{\mathrm{ij}} = y_{\mathrm{ij}} \text{ - } \hat{y}_{\mathrm{ij}}, \mathrm{i} = 1 \text{ ,}\ldots, \mathrm{k} \text{ e j} = 1, \ldots, \mathrm{n}$$

em que $\hat{y}_{\mathrm{ij}}$ é o valor ajustado pelo modelo proposto correspondente ao valor observado y_{ij}, obtido como segue

$$\hat{y}_{\mathrm{ij}} = \hat{\mu} + \hat{\tau} + \overline{y}_{\mathrm{i.}}$$

De forma geral, violação das suposições básicas da ANOVA pode ser detectada através das seguintes análises gráficas dos resíduos apresentadas adiante.

Gráfico de Probabilidade Normal

A validade da suposição de normalidade pode ser avaliada por meio do gráfico de probabilidade normal para os resíduos. Os resíduos são colocados no eixo das abscissas e os escores de uma distribuição normal no eixo das ordenadas. A suposição de normalidade será considerada válida se os pontos do gráfico estiverem localizados, aproximadamente, ao longo de uma linha reta que passa pela origem e tem coeficiente angular 1 (uma reta de 45^{o}). Outros gráficos como histograma e *boxplot* podem ser usados para verificar a suposição de normalidade. Como a avaliação gráfica é subjetiva, um teste estatístico pode ser utilizado para complementar esta verificação. Aplicam-se os chamados Testes de Aderência, nesta apostila é apresentado o Teste de Qui-Quadrado de Aderência.

Gráfico de Resíduos Contra Ordem das Observações Coletadas

A validade da suposição de que os erros não são correlacionados pode ser verificada por meio de um gráfico de resíduos contra a ordem das observações coletadas. Se os resíduos estiverem aleatoriamente situados, aproximadamente, em torno de uma faixa horizontal centrada em $\epsilon_{ij} = 0$, sem nenhum padrão definido, é uma indicação da validade da suposição de independência. Por outro lado, configurações especiais, tais como a presença de sequências de resíduos positivos e negativos, ou padrões de alternância de sinais, podem indicar que as observações não são independentes.

Gráficos dos Resíduos (ϵ_{ij}) contra os Valores Preditos (y_{ij})

A validade da suposição homogeneidade das variâncias dos erros em todos os níveis do fator. A suposição de homogeneidade não é viola se a dispersão dos resíduos não depende dos valores preditos $\hat{y}_{ij}$ (para o modelo de um fator $y_{ij} = \overline{y}_{i.}$). Por exemplo, se as variâncias dos resíduos crescem quando os valores preditos crescem ou se a variâncias dos resíduos decrescem á medida que valores preditos decrescem é indicativo de violação de homogeneidade variâncias. Adicionalmente, quando o gráfico apresenta um padrão parecido com um "funil"ou "megafone também é um indicativo de variância não constante.

Exemplo 15.1: Examinar os resíduos do modelo ajustado para a força de resistência das sacolas.

Os cálculos dos resíduos para os dados da Tabela 16.3 estão a seguir:

Tabela 16.6: Resíduos dos dados apresentados na Tabela 16.3

Concentração de madeira de lei			
5%	10%	15%	20%
7-10=-3,0	12-15,67=-3,7	14-17=-3,0	19-21,17=-2,2
8-10=-2,0	17-15,67=1,3	18-17=1,0	25-21,17=3,8
15-10=5,0	13-15,67=-2,7	19-17=2,0	22-21,17=0,8
11-10=1,0	18-15,67=2,3	17-17=0,0	23-21,17=1,8
9-10=-1,0	19-15,67=3,3	16-17=-1,0	18-21,17=-3,2
10-10=0,0	15-15,67=-0,7	18-17=1,0	20-21,17=-1,2

O Gráfico 16.2 serve para verificar a suposição de normalidade dos resíduos. Neste gráfico, a hipótese de normalidade para os resíduos pode ser aceita, pois o gráfico revela-se aproximadamente linear. Para confirmar esta suposição foi realizado o Teste Qui-Quadrado de Aderência. As hipóteses testadas foram:

Figura 16.2: Gráfico Q-Q Normal dos Resíduos do Exemplo 15.1

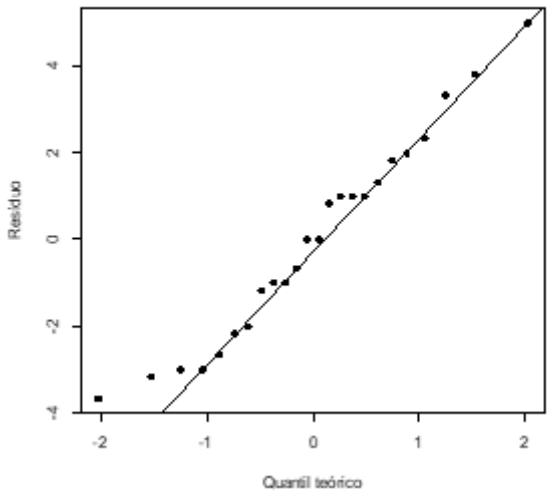
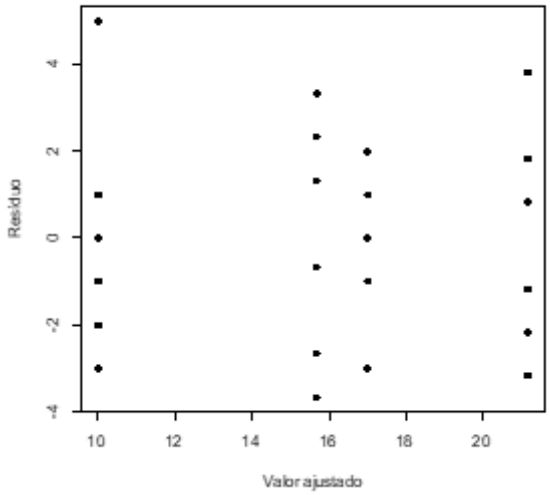


Figura 16.3: Resíduos versus valores ajustados do Exemplo 15.1



H_0: Os dados dos resíduos se ajustam a uma distribuição normal.

H_1: Os dados dos resíduos não se ajustam a uma distribuição normal.

O resultado do teste está na caixa adiante, observe que o p-valor = 0,6487 $> \alpha = 5\%$. Portanto, não existem evidências para rejeitar a suposição de que os resíduos se ajustam a uma distribuição normal, ao nível de significância de 5%.

No software R o teste para normalidade pode ser realizado usando a seguinte sintaxe:
> library(nortest)
> pearson.test(Resíduo)
Pearson chi-square normality test
data: Resíduo
P = 3.3333, **p-value = 0.6487**

O segundo gráfico (Gráfico 16.3) apresenta os valores ajustados, através do modelo, versus resíduos. O padrão deste gráfico não indica evidência de violação da suposição de que as variâncias dos erros são constantes em todos os níveis do fator, pois a variabilidade dos resíduos não parece crescer quando os valores preditos crescem. Para confirmar esta suposição será realizado o Teste de Homogeneidade de Variâncias na Seção 17.

17 Homogeneidade das Variâncias

Uma importante pressuposição para aplicação da técnica de Análise de Variância é a homogeneidade das variâncias da variável de interesse das populações envolvidas, ou seja, que a variância seja homogênea em todos os níveis de fator. Para testar a homogeneidade das variâncias, utilizam-se das seguintes hipóteses estatísticas:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_{\mathrm{k}}^2 = \sigma^2$$
$$H_1 : \text{pelo menos uma variância difere das demais.}$$

em que k é o número de níveis do fator de interesse e σ_{i}^2 é a variância do i-ésimo nível, i=1,...,k. O procedimento usado para testar as hipóteses acima será o Teste de *Bartlett* (ver Montegomery, 2005). Considere que $S_1^2, S_2^2, \ldots, S_{\mathrm{k}}^2$ são as variâncias amostrais de tamanho $n_1, n_{\mathrm{k}}, \ldots, n_{\mathrm{k}}$, respectivamente sendo $\mathrm{N} = \sum_{\mathrm{i}\,=\,1}^{\mathrm{k}} n_{\mathrm{i}}$. O estimador da variância combinada das k tratamentos (ou populações) é dado por:

$$S_{\mathrm{P}}^2 = \frac{\sum_{\mathrm{i}\,=\,1}^{\mathrm{k}} (n_{\mathrm{i}}-1)S_{\mathrm{i}}^2}{N-k}$$

A estatística de teste é dada por:

$$\chi^2_{\text{calc.}} = 2,3026\frac{q}{c'}$$

em que:

$$\mathrm{q} = (N-k)\log_{10} S_{\mathrm{P}}^2 - \sum_{\mathrm{i}\,=\,1}^{\mathrm{k}} (n_{\mathrm{i}}-1)\log_{10} S_{\mathrm{i}}^2,$$

e

$$\mathrm{c} = 1 + \frac{1}{3(k-1)}\left(\sum_{\mathrm{i}\,=\,1}^{\mathrm{k}}\left(\frac{1}{(n_{\mathrm{i}}-1)}\right) - \left(\frac{1}{(N-k)}\right)\right).$$

Supondo que a hipótese nula é verdadeira, a estatística de teste possui distribuição aproximadamente Qui-Quadrado com (k - 1) graus de liberdade. Para um dado nível de significância α, rejeitar a hipótese nula se $\chi^2_{\text{calc.}} > \chi^2_{\alpha;\ (\mathrm{k}\ -\ 1)}$, em que $\chi^2_{\alpha;\ (\mathrm{k}\ -\ 1)}$ é uma constante tal que $\mathrm{P}(\chi^2_v > \chi_{\alpha;\ (\mathrm{k}\ -\ 1)}) = \alpha$. Ou pelo p-valor, rejeitar a hipótese nula se p-valor $< \alpha$.

Exemplo 17.1: Use o Exemplo 16.1 para realizar o teste de hipóteses, ao nível de significância de 1%, de que as variâncias populacionais para o conjunto de dados sobre a força de resistência de papel usado para a confecção de sacolas, com diferentes concentrações madeira de lei, são iguais.

Solução: As hipóteses estatísticas a serem testadas são:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$$

(As variâncias das forças de resistência dos papeis são homogêneas nas quatro concentrações de madeira de lei.)

H_1: Pelo menos uma das variâncias, da força de resistência do papel, é diferente das demais.

Tem-se que $n_1^2 = n_2^2 = n_3^2 = n_4^2 = 6, k = 4$. As variâncias amostrais são:

$$s_1^2 = 8,00$$
$$s_2^2 = 7,87$$
$$s_3^2 = 3,20$$
$$s_4^2 = 6,97$$

A variância combinada

$$S_{\mathrm{P}}^2 = \frac{5(8,00)+5(7,87)+5(3,20)+4(6,97)}{20} = \frac{130,17}{20} = 6,5083$$

Agora, calcule

$q = (20)\log_{10}(6,5083) - 5[\log_{10}(8) + \log_{10}(7,87) + \log_{10}(3,20) + \log_{10}(6,97)] = 0,5321.$

$1 + \frac{1}{3(3)}\left(\frac{4}{5} - \frac{1}{20}\right) = 1,0833$

O valor da estatística de teste $\chi^2_{\text{calc.}} = 2,3026\frac{0,5321}{1,0833} = 1,1310$

Conclusão: Como $\chi^2_{\text{calc.}} = 1,1310 < \chi^2_{1\%;\ (3)} = 11,34$ (valor tabelado, ver Tabela 1 em anexo), não rejeitamos a hipótese nula e concluímos que não existem evidência de que as variâncias populacionais para dos conjunto de dados sobre a força de resistência de papel

usado para a confecção de sacolas, com diferentes concentrações madeira de lei, são diferentes, ao nível de significância de 1%.

Na literatura existem outros procedimentos para realizar o teste de homogeneidade das variâncias, a saber: teste de *Cochran* e *Levene*. O teste de *Bartlett* é mais eficiente para testar a homogeneidade de variâncias quando as variáveis envolvidas no estudo possuem distribuição normal (ou aproximadamente normal). Se a hipótese de normalidade for violada, é melhor utilizar o teste proposto por *Levene*.

18 1ª Lista de Exercícios

1. Abaixo você encontra uma lista de situações de pesquisa. Para cada uma indique se o apropriado é uma análise de correlação ou uma de regressão.

   a) A quantidade procurada da carne gado depende do preço da carne de porco? R:Correlação

   b) O objetivo é estimar o tempo necessário para a realização de certa tarefa usando para tanto o tempo de treinamento do executor. R:Regressão

   c) O preço de uma reforma depende dos valores dos artigos usados no acabamento? R:Correlação

   d) Estime o número de milhas que um pneu radial possa rodar antes de ser substituído. R:Regressão

   e) Deseja-se prever quanto tempo será necessário para uma pessoa completar determinada tarefa, com base no número de semanas de treinamento. R:Regressão

   f) Decida se o número de semanas de treinamento é uma variável importante para avaliar o tempo necessário para realizar uma tarefa. R:Regressão

   g) Verificar se existe dependência entre os salários mensais (em milhares de reais) recebidos por executivos homens e mulheres. R:Correlação

2. Um modelo genérico especifica que os animais de certa população devam ficar classificados em quatro categorias, com probabilidades $p_1 = 0{,}656$, $p_2 = p_3 = 0{,}093$ e $p_4 = 0{,}158$. Dentre 197 animais, obtivemos as seguintes freqüências observadas: $O_1 = 125$, $O_2 = 18$, $O_3 = 20$ e $O_4 = 34$. Teste se esses dados estão de acordo com o modelo genérico postulado. (Use $\alpha = 1\%$). R: Os dados estão de acordo com o modelo postulado, Região Crítica $[11{,}34;\ \infty)$ e $\chi^2_{\text{calc.}} = 0{,}5635$

3. Uma empresa localizada na cidade de São Paulo, produtora de pneumáticos, possui uma rede distribuidora por todo o interior do Estado. Realizou um estudo para determinar qual a função que ligava o preço do produto e a distância do mercado consumidor da cidade de São Paulo. Os dados são os seguintes:

Preço	36	48	50	70	42	58	91	69
Distância (Km)	50	240	150	350	100	175	485	335

   a) Calcule o coeficiente de correlação e interprete o resultado. R.: r = 0,959

   b) Estimar a reta de regressão; $P_i = 30{,}19 + 0{,}12D_i$, i=1,..., 8.

   c) Calcule um intervalo com confiança de 5% para o preço quando a distância é 250Km. R.: [45,91; 74,47]

   d) A empresa tem uma filial no Rio de Janeiro e o preço de venda do pneumático lá produzido, na cidade B, é de R$160,00. Sabendo-se que a distância entre São Paulo e a cidade B é de 250 km, pergunta-se qual produto deve ser vendido: o produzido no Rio de Janeiro ou o fabricado em São Paulo.R.: São Paulo.

4. Após ser derrotado por um amigo num jogo de dado, você suspeita que o dado que ele deu a você seja desonesto. Para verificar, você lança o dado 60 vezes, registrado o número de vezes que cada face aparece. Os resultados estão adiante.

Face	1	2	3	4	5	6
Frequência	11	7	9	15	12	6

   a) Se o dado for honesto, quantas vezes você esperaria que cada face aparecesse? R:10.

   b) Para verificar se o dado é honesto, qual teste você usaria? R:Teste de Aderência

   c) Teste a hipótese de que o dado é honesto. (Use $\alpha = 5\%$).R: Não há evidências de que o dado seja desonesto. Região Crítica [11,07; ∞) e $\chi^2_{\text{calc.}} = 5{,}635$.

5. Suponhamos que uma cadeia de supermercados tenha financiado um estudo dos gastos com mercadoria para famílias de 4 pessoas. A investigação se limitou a famílias com renda líquida entre R$8.000 e R$20.000. Obteve-se a seguinte equação:

$$\hat{y}_i = -200 + 0,10x_i$$

em que: y = despesa anual estimada com mercadorias e x = renda líquida anual. Suponha que a equação proporcione um ajustamento razoavelmente bom.

a) Estime a despesa de uma família de quatro com renda de R$15.000. R.: 1.300,00

b) Um dos vice-presidentes da firma ficou intrigado com o fato de a equação aparentemente sugerir que uma família com R$2.000 de renda não gaste nada em mercadorias. Qual a explicação? R. Observe que o intervalo de x não contempla o valor R$ 2.000, então não é recomendado estender a reta de regressão ajustada para fazer predições fora do intervalo de x.

c) Explique por que a equação acima não poderia ser usada nos seguintes casos:
i. estimar despesas com mercadorias para famílias de cinco pessoas.R: O modelo foi ajustado com despesas de 4 pessoas.

ii. estimar despesas com mercadorias para famílias com renda de 20 a 35 s.m.R: O modelo foi ajustado com renda líquida entre R$8.000 e R$20.000.

6. Três diferentes bancos possuem agências de mesmo porte em uma avenida movimentada de Salvador, BA. Para testar se essas agências têm movimento médio equivalente, foi escolhida uma semana típica de trabalho e o desempenho, nesses dias, foi registrado. Os dados obtidos, em milhares de reais, estão apresentados nas tabelas a seguir.

Banco		
146,4	194,3	173,7
199,2	227,2	246,5
179,5	203,4	289,8
98,4	111,8	127,4
263,7	275	265,6

a) É razoável afirmar que as variâncias das três distribuições são homogêneas? (Use $\alpha = 1\%$). R. Como $\chi^2_{\text{calc.}} = 0{,}0679 < \chi^2_{1\%;\,(2)} = 9{,}21$ (Concluímos que não existem evidências de que as variâncias dos movimentos financeiros nos três bancos sejam diferentes.)

b) A partir da análise de variância (ANOVA) adiante verifique se as agências têm movimentos médios equivalentes. Use $\alpha = 5\%$. R. Não rejeita H_0, ao nível de 5% de significância.

Fonte de variação	Soma de quadrados	Graus de liberdade	Quadrados médios	$F_{\text{calculado}}$
Entre grupos	4693,705	2	2346,853	0,590894
Dentro dos grupos	47660,38	12	3971,699	
Total	52354,09	14		

7. A fim de testar se o tempo médio necessário para misturar um lote de materiais é o mesmo para máquinas produzidas por três diferentes fabricantes, a Jacobs Chemical Company obteve os seguintes dados sobre o tempo (em minutos) necessário para misturar os materiais.

   a) É razoável afirmar que as variâncias das três distribuições do tempo médio necessário para misturar um lote de materiais são homogêneas? (Use α =0,01). R.: Sim. Não rejeita H_0, pois $\chi^2_{\text{calc.}} = 7{,}47 < \chi^2_{1\%;\,(2)} = 9{,}21$.

   b) Realize um teste para verificar se o tempo médio para misturar um lote de materiais difere em relação aos três fabricantes, use $\alpha = 0{,}05$.R: Não existe evidências de que o tempo médio necessário para misturar um lote de materiais em cada fábrica seja diferente, ao nível de significância de 5%.

Fábrica		
1	2	3
21	34	21
14	28	17
25	38	23
32	25	22
31	26	28
35	27	24
8	25	24
21	27	20

8. Os dados a seguir dão um custo líquido por real de prêmio (Y) e o tempo de apólice em meses (X).

X	8	29 47	24	57	45	39	14	70	40	66	55
Y	1,26	1,15	0,81	1,14	0,61	0,88					
	0,99	1,11	0,58	0,74	0,67	0,70					

   a) Estimar a reta de regressão. R: $Y_{\text{i}} = 1{,}35 - 0{,}01\ X_{\text{i}}$, i = 1, 2, 3,. . ., 12.

   b) Calcule um intervalo de confiança de 95% de confiança para a inclinação β_1. Baseado no intervalo, qual a conclusão sobre a relação linear entre x e y.
   R: $\left[-0{,}01 - 2{,}228\sqrt{\frac{0{,}07452}{10}}\sqrt{\frac{1}{4225{,}67}};\ -0{,}01 + 2{,}228\sqrt{\frac{0{,}07452}{10}}\sqrt{\frac{1}{4225{,}67}}\right]$

   c) Construir um IC para o valor de um prêmio cuja apólice tem 3 anos; $\alpha = 5\%$.
   R: $0{,}99 \mp 2{,}228\sqrt{\frac{1}{4225{,}67}}\sqrt{1 + \frac{1}{2} + \frac{(36-41{,}17)^2}{4225{,}67}}$

9. Os valores do módulo de elasticidade (MOE, a razão da força, isto é, força por área unitária, para o escoamento, ou seja, deformação por comprimento unitário, em GPa)

e a resistência à reflexão (uma medida da capacidade de resistência a falhas decorrentes de desdobramento, em MPa) foram determinados para um tipo de amostra de vigas de concreto, gerando os dados a seguir (reproduzidos de um gráfico do artigo *"Effects of Aggregate and Microfilleres on the Flexural Proprties of Concrete", Magazine of Concrete Research, 1997, p.81-98*):

MOE	Resistência
29,8	5,9
33,2	7,2
33,7	7,3
35,3	6,3
35,5	8,1
36,1	6,8
36,2	7,0
36,3	7,6
37,5	6,8
37,7	6,5
38,8	6,3
39,6	7,9
41,0	9,0
42,8	8,2
42,8	8,7
43,5	7,8
45,6	9,7
46,0	7,4
46,9	7,7
48,0	9,7
49,3	7,8
51,7	7,7
62,6	11,6
69,8	11,3
79,5	11,8
80,0	10,7

a) O valor da resistência é determinado exclusivamente pelo valor do MOE? R: Não, porque há observação com os valores idênticos de MOE com diferentes valores de y.

b) Use os resultados da saída do software Excel a seguir e apresente a equação ajustada do modelo de regressão. R: $\hat{y}_i = 3,34 + 0,107x_i, i = 1, 2, 3, \ldots, 26$

c) Calcule o coeficiente de determinação. R: 0,736

Estatística de regressão	
R múltiplo	0,858
R-quadrado	0,7364
R-quadrado ajustado	0,7253
Erro padrão	0,8785
Observações	26

Fonte de Variação	Graus de Liberdade	Soma de Quadrado	Quadrado Médio	Estatística F
Regressão	1	51,7325	51,7325	67,035
Resíduo	24	18,5214	0,77172	
Total	25	70,2539		

	Coeficientes	Erro padrão	Estatística t	P-valor
Interseção	3,3400	0,6163	5,4200	0,00001442
MOE	0,1068	0,0130	8,1875	0,00000002

10. Nova York, Boston e o Vale do Silício na Califórnia estão entre as regiões que apresentam os maiores salários no setor de tecnologia nos Estados Unidos (*USA Today*, 28 de fevereiro de 2002). Os dados amostrais seguintes apresentam os salários anuais individuais expressos em milhares de dólares

    a) Verifique se existe diferença entre a média populacional de salários do setor de tecnologia correspondente nas três localidades. Use $\alpha = 5\%$.

Nova York	Boston	Vale do Silício
82	85	82
79	80	91
72	74	94
89	78	88
79	75	85
85	80	81
86	79	90

R: Existe diferença entre as médias de salários nas três localidades, observe a tabela da anova apresentada adiante

Grupo	Contagem	Soma	Média	Variância
Nova York	7	572	81,71	31,90
Boston	7	551	78,71	13,24
Vale do Silício	7	611	87,29	23,24

11. Realiza-se um estudo para se determinar o efeito da velocidade de corte sobre a duração (em horas) de uma máquina particular. Quatro níveis de velocidade de corte são selecionados para o estudo, com os seguintes resultados:

ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrado	Quadrado Médio	Estatística F	valor-P	F crítico
Tratamento	264,86	2,00	132,43	5,81	0,01	3,55
Erro	410,29	18,00	22,79			
Total	675,14	20,00				

Durabilidade da ferramenta

Velocidade de corte	Repetição (ou observação)					
	1	2	3	4	5	6
1	41	43	33	39	36	40
2	42	36	34	45	40	39
3	34	38	34	34	36	33
4	36	37	36	38	35	35

Fonte: Hines, Montgomery, Goldman e Borror (2006). Probabilidade e Estatística na Engenharia. 4ª ed.

a) A velocidade de corte afeta a durabilidade da máquina?. Use $\alpha = 0{,}01$. R.:A velocidade média de corte não afeta a durabilidade da máquina, ao nível de significância de 1%.

b) Você usaria o Teste de Tukey para fazer comparações entre os pares de médias dos níveis de velocidade de corte? R: Não usaria o teste, pois não há evidências de que existam diferenças significativas entre os pares de médias.

ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrado	Quadrado Médio	Estatística F	valor-P
Tratamentos (Velocidade)	3	80,17	26,722	3,175	0,0465
Resíduo	20	168,33	8,417		
Total	23	248,5			

12. O conjunto de dados a seguir consiste de 26 observações sobre a resistência à fratura do prato de base do aço temperado com níquel a 18% (de "Fracture Testing of Weldments", ASTM Special Publ. Nº 381, 1965, p. 328-356). Um indivíduo suspeita que estes dados possam ser ajustados através de uma Distribuição Normal.

a) Construa um histograma para os dados acima e verifique se é razoável supor que os dados segue uma distribuição normal. Justifique sua resposta.

Resistência	Número de observações
65⊢ 70	3
70⊢ 75	5
75⊢ 80	10
80⊢ 85	6
85⊢ 100	2
Total	26

b) Verifique através do teste se os dados seguem a distribuição normal. Use $\alpha = 10\%$. R: Modelo normal não é rejeitado, Região Crítica [6,251; ∞) e $\chi^2_{\text{calc.}} = 1{,}1$.

13. Teste se os dados abaixo são observações de uma distribuição normal com média $\mu = 10$ e variância $\sigma^2 = 25$. Os dados estão apresentados na tabela adiante. (Use $\alpha = 5\%$)

Variável	**Número de observações**
01,0 ⊣ 6,6	4
06,6 ⊣ 10,0	11
10,0 ⊣ 13,4	9
13,4 ⊣ 22,0	6
Total	30

R: Modelo é não é rejeitado, Região Crítica [7,81; ∞) e $\chi^2_{\text{calc.}} = 3{,}7346$

14. Uma regressão de y=volume de cálcio (g/l) em x = material dissolvido (mg/cm^2) foi descrita em um artigo "Use of Fly Ash or Silica Feed Acids"(Magazine of Concrete Research, 1997, p. 337-344). A questão da reta de regressão estimada foi: $\hat{y} = 3{,}678 + 0{,}144\ x_i$i = 1,2,3,...,23. e R^2 =0,860.

a) Interprete o coeficiente estimado 0,144. R.: Estima-se que 0,144 é a mudança esperada no conteúdo de cálcio associado com 1mg/cm2 de aumento na quantidade dissolvida de material.

b) Interprete o coeficiente de determinação 0,860. R.: Aproximadamente, 86% da variabilidade observada no volume de cálcio pode ser atribuída a quantidade dissolvida de material (através do modelo proposto), 14% é devido a outros fatores.

c) Calcule uma estimativa pontual do volume médio real de cálcio quando a quantidade de material dissolvido for igual a 50 mg/cm2. R.: 10,88.

15. Oito programas foram monitorados para estudar a demanda por recursos. Neste trabalho, a variável resposta (dependente) é o tempo de CPU, e a variável independente é o número de acessos ao disco (disk I/O)

Tempo de CPU (Y)	Número de acessos ao disco (X)
2,0	14
4,6	15
5,7	23
7,3	31
9,8	38
10,9	40
12,6	53
13,2	51

a) Faça o diagrama de dispersão. Conclua sobre a correlação entre as duas variáveis. R: as variáveis são correlacionadas.

b) Calcule o coeficiente de correlação de Pearson. Conclua sobre a correlação entre as duas variáveis. R: 0,979793, existe uma forte correlação positiva entre o tempo de CPU e o número de acessos ao disco.

16. Os 12 pares de valores são relativos às variáveis tamanho da memória (mbytes) e bytes transferidos (mbytes). Observe que para cada tamanho de memória foram realizados três experimentos (repetições).

Tamanho de memória em mbytes (X)	Bytes transferidos em mbytes (Y)
0,238	39,058
0,238	42,967
0,238	35,118
0,286	37,938
0,286	41,257
0,286	32,921
0,334	36,531
0,334	40,368
0,334	30,563
0,381	35,484
0,381	39,203
0,381	30,823

a) Faça o diagrama de dispersão. Conclua se existe uma relação linear entre bytes transferidos e o tamanho da memória. R: Parece existir uma relação linear entre as variáveis.

b) Calcule os coeficientes de regressão do modelo linear simples. Apresente a reta de regressão ajustada aos dados. $\hat{y}$ = 45,42 - 27,67 x_i ,i=1, . . ., 12.

c) Dê a estimativa pontual do número de bytes transferidos para tamanhos de memória iguais a 0,255 e 0,355. R:38,36 e 35,60.

d) Estime a variância dos erros do modelo de regressão utilizado. R: 15,09

e) Predição de uma nova observação. Obter o intervalo de predição no número de bytes transferidos para X = 0,255 e X = 0,355, use $\alpha = 5\%$. R: [29,67; 47,06] e [29,68; 47,05].

f) Fazer a análise de variância do modelo de regressão (apresentar a tabela da análise de variância).

ANOVA

	Graus de Liberdade	**Soma de Quadrado**	**Quadrado Médio**	**F**	**F de significação**
Regressão	1	26,12	26,12	1,73	0,22
Resíduo	10	150,94	15,09		
Total	11	177,07			

g) Teste de significância do modelo. Fazer o teste F para verificar se existe relação linear entre o número de bytes transferidos e o tamanho de memória, use $\alpha = 0{,}01$. Escreva a conclusão do teste. R: Ao nível de significância de 1%, ao existe relação linear entre o número de bytes transferidos e o tamanho de memória.

h) Coeficiente de determinação. Calcular o coeficiente de determinação para os dados de bytes transferidos e tamanho de memória. R: 0,148

i) Análise de resíduos. Calcular os resíduos e fazer os gráficos dos resíduos versus os valores preditos pelo modelo de regressão e resíduos versus a variável independente (X). As suposições do modelo parecem satisfeitas? R: Não

j) É válido construir o intervalo de confiança do número médio de bytes transferidos para x= 35 mbytes de tamanho de memória, com um grau de confiança de 95%? R: Não, pois os resíduos não se ajustam a uma distribuição normal.

17. A quantidade de chuva é um fator importante na produtividade agrícola. Para medir esse efeito, foram anotadas, para 8 regiões diferentes produtora de soja, o índice pluviométrico e a produção do último ano.

    Para analisar os dados descritos acima, considere o modelo de regressão linear simples dado por:

$$Y_{\mathrm{i}} = \beta_0 + \beta_1 X_{\mathrm{i}} + \varepsilon_{\mathrm{i}}, i = 1, 2, \ldots, n.$$

Chuva (mm)	**Produção (tonelada)**
120	40
140	46
122	45
150	37
115	25
190	54
130	33
118	30
120	37
155	40
90	20
100	24
116	27
130	34
127	43
160	48
110	28
154	39
180	50

com $\varepsilon_{\mathrm{i}} \sim N(0;\sigma^2)$ e ε_{i} não correlacionado com ε_{i} para i,j = 1,2,..., n e i $\neq$ j, σ^2 desconhecido. Siga os passos a seguir:

a) Esboçar o gráfico de dispersão da chuva versus produção. Verifique se existe uma relação linear entre as variáveis. R: Sim, parece existir uma relação linear entre a chuva e a produção agrícola.

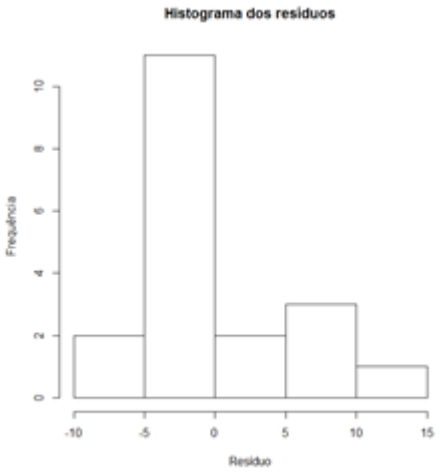
b) Obtenha o ajuste do modelo de regressão linear simples e apresente as estimativas de β_0, β_0 e σ^2. Descreva o modelo ajustado matematicamente. R.: 1,55; 0,27 e 298,04. y_{i}=1,55 + 0,27x_{i},i=1,2,3,...,8.

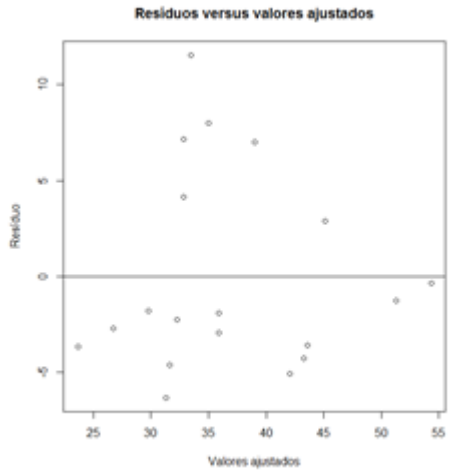
c) Faça o teste da significância da regressão via Tabela ANOVA e verifique possível falta de ajuste. R.: A análise da Tabela ANOVA indica evidências da existência de uma relação linear entre a quantidade de chuva e a produção agrícola.

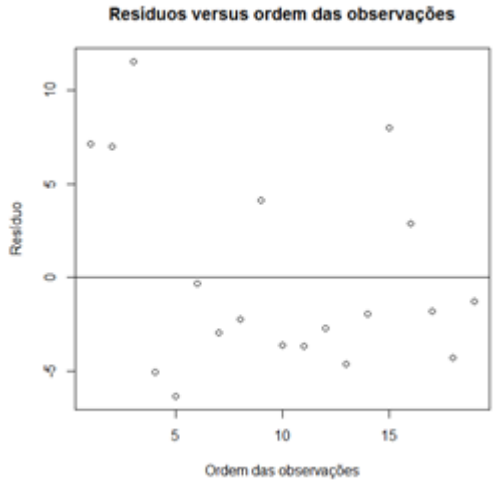
d) Analise os gráficos adiante e faça uma análise de resíduos verificando os possíveis padrões indesejáveis ou presença de observações com valores extremos. R.: Converse com o seu professor

e) Calcule os coeficientes de determinação e correlação. R. 0,52; 0,7246.

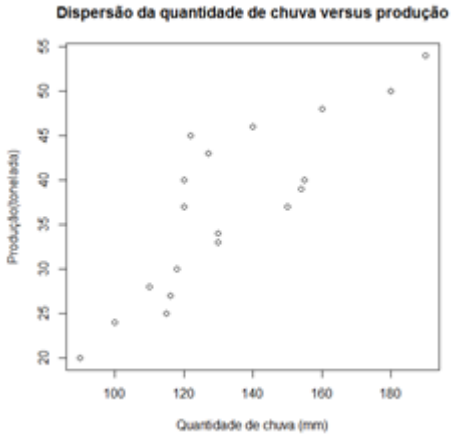
f) Finalize sua análise concluindo se o modelo é ou não adequado para os dados em







questão. R:Converse com o seu professor.



18. Quatro tipos de fertilizantes estão sendo comparados para ver qual deles apresenta

Estatística de regressão	
R múltiplo	0,839529615
R-Quadrado	0,704809974
R-quadrado ajustado	0,687445855
Erro padrão	5,334000957
Observações	19

ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrado	Quadrado Médio	Estatística F	P-valor
Regressão	1	1154,85	1154,85	40,590	$6,9710^{-}6$
Resíduo	17	483,68	28,45		
Total	18	1638,53			

ANOVA

	Coeficientes	Erro padrão	Estatística t	P-valor	95% inferiores	95% superiores
Interseção	-3,96	6,52	-0,61	0,55	-17,72	9,79
Chuva (mm)	0,31	0,05	6,37	0,00	0,21	0,41

maior produção de sementes de milho. Quarenta áreas de terra similares foram disponibilizadas para realizar o experimento. As 40 áreas de terra foram divididas aleatoriamente em quatro grupos, dez áreas em cada grupo. Fertilizante 1 foi aplicado em cada uma das dez áreas no grupo 1. Similarmente, os fertilizantes 2, 3 e 4 foram aplicados nas áreas do grupo 2, 3 e 4, respectivamente. Os resultados de produção de milho (y) das 40 áreas foram:

Fertilizante 1	Fertilizante 2	Fertilizante 3	Fertilizante 4
31	27	36	33
34	27	37	27
34	25	37	35
34	34	34	25
43	21	37	29
35	36	28	20
38	34	33	25
36	30	29	40
36	32	36	35
45	33	42	29

a) Verifique se, em média, os três tipos de fertilizantes tem um efeito sobre a produção de sementes de milho, ao nível de significância de 5%. (Fazer a tabela da análise de variância e o teste F). R: Tem efeito, pois o p-valor=0,005 $<$ 0,05.

RESUMO

Grupo	Contagem	Soma	Média	Variância
Fertilizante 1	10	366	36,6	18,71
Fertilizante 2	10	299	29,9	22,77
Fertilizante 3	10	349	34,9	16,99
Fertilizante 4	10	298	29,8	35,51

ANOVA

Fonte da variação	SQ	gl	MQ	F	valor-P	F crítico
Entre grupos	362,6	3	120,867	5,144	0,005	2,866
Dentro dos grupos	845,8	36	23,494			
Total	1208,4	39				

19. A Butler Trucking Company, uma companhia de transporte do sul da Califórnia tem seus maiores negócios envolvendo entregas na região. Para desenvolver um trabalho melhor, os gerentes supõem que o modelo de regressão linear simples poderia ser usado para descrever a relação entre o tempo total de viagem (Y) e a quilometragem percorrida (X1). Foi selecionada uma amostra aleatória simples de 10 tarefas de entrega, que forneceu os dados da tabela abaixo.

Tarefa	X1: Quilometragem	X2: Número de entregas	Y: Tempo de entrega
1	100	4	9,3
2	50	3	4,8
3	100	4	8,9
4	100	2	6,5
5	50	2	4,2
6	80	2	6,2
7	75	3	7,4
8	65	4	6,0
9	90	3	7,6
10	90	2	6,1

a) Determine o grau de correlação linear entre Y e X_1. R: 0,815

b) Teste se o tempo de viagem está relacionado linearmente com a quilometragem percorrida, com um nível de significância de 5%. R: Sim, pois p-valor (F de significação) = 0,004 < 0,05.

c) Qual o percentual da variabilidade do tempo de viagem que pode ser explicado pelo efeito linear da quilometragem percorrida? R: 66,4%

d) Utilize o modelo de regressão linear simples para descrever a relação entre Y e X_1.

i. Determine a equação de regressão linear. $Y_{\text{i}} = 1{,}27 + 0{,}07x_1$, i=1,2,3,...,10.

ii. Interprete o coeficiente β_1 da reta de regressão estimada. R: Estima-se que 0,07 é a quantidade acrescida ao tempo de entrega para cada aumento de 1 quilômetro percorrido.

e) Os gerentes resolveram acrescentar outra variável independente para explicar alguma variabilidade remanescente na variável dependente. Acharam que o número de entregas também poderia contribuir para o tempo de viagem. Considerando o número de entregas como X2, determine a equação de regressão linear. $y_{\text{i}} =$ -0,869+0,061$x_{1\text{i}}$ + 0,923$x_{2\text{i}}$ i = 1,2,3,...,10.

f) Interprete os coeficientes da equação obtida no item anterior. R: O valor 0,061 é a quantidade acrescida ao tempo de entrega para cada aumento de 1 quilômetro percorrido permanecendo constante o número de entrega. O valor 0,923 é a quantidade acrescida ao tempo de entrega para cada aumento de 1 quantidade de entrega permanecendo constante a quilometragem.

Correlação

	X1: Quilometragem	X2:Número de entregas	Y:Tempo de entrega
X1: Quilometragem	1,000		
X2: Número de entregas	0,162	1,000	
Y: Tempo de entrega	0,815	0,615	1,000

Estatística de regressão	
R múltiplo	0,815
R-Quadrado	0,664
R-quadrado ajustado	0,622
Erro padrão	1,002
Observações	10

ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	1	15,871	15,871	15,815	0,004
Resíduo	8	8,029	1,004		
Total	9	23,902			

	Coeficientes	Erro padrão	Estatística t	P-valor	95% inferiores	95% superiores
Interseção	1,27	1,40	0,91	0,39	-1,96	4,50
X1: Quilometragem	0,07	0,02	3,98	0,00	0,03	0,11

Estatística de regressão	
R múltiplo	0,951
R-Quadrado	0,904
R-quadrado ajustado	0,876
Erro padrão	0,573
Observações	10

ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	2	21,601	10,800	32,878	0,000
Resíduo	7	2,299	0,328		
Total	9	23,900			

	Coeficientes	Erro padrão	Estatística t	P-valor	95% inferiores	95% superiores
Interseção	-0,869	0,952	-0,913	0,392	-3,119	1,381
X1: Quilometragem	0,061	0,010	6,182	0,000	0,038	0,085
X2: Número de entregas	0,923	0,221	4,176	0,004	0,401	1,446

Referências Bibliográficas

CONOVER, W. J. (1999). **Practical Nonparametric Statistics**. 3rd. ed. New York: Chichester: John Wiley & Sons (Asia).

DEAN, A. & VOSS, D. (1999). **Desing and Analysis of Experiments**. New York: Springer.

FERNANDES, Gilênio Borges, (2002). **Notas de Aula MAT 229- Análise de Regressão**.

HINES, W. William, MONTGOMERY, C. Douglas, GOLDSMAN, M. David e BORROR, M. Cannie (2006). **Probabilidade e Estatística na Engenharia**. 4ª ed., Rio de Janeiro: LTC.

HOLLANDER, Myles; WOLFE, Douglas A (1999). **Nonparametric Statistical Methods**. 2nd. ed. New York: John Wiley & Sons.

CAMPOS, Humberto de (1979). **Estatística Experimental Nao-Paramétrica**. 3. ed. Piracicaba: Departamento de Matemática e Estatística da Escola Superior de Agricultura 'Luiz de Queiroz.

MORAES, Lia Terezinha L. P. (2006). **Notas de Aula MAT 187- Métodos Não Paramétricos**

MAGALHÃES, Marcos Nascimento e LIMA, Antônio Carlos P. (2007). **Noções de Probabilidade e Estatística**. 6a edição rev. 1a reimpressão, São Paulo, Edusp.

MONTGOMERY, Douglas C.; RUNGER, George C.; HUBELE e Norma Faris (2004). **Estatística Aplicada à Engenharia**. Rio de Janeiro: LTC.

MONTGOMERY, Douglas C. (2005). **Design and Analysis of Experiments**. 3ed. New

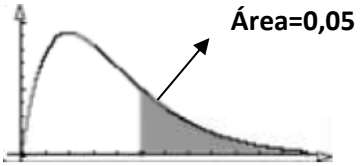
York, John Wiley.
MORETTIN, Pedro Alberto e BUSSAB, Wilton de Oliveira (2006). **Estatística Básica**. 5. ed. São Paulo: Saraiva.
NETER, J. e Wasserman, W. (1974). **Applied linear statistical models**. Richard D. Irwin Inc. Homewood, Illinois.
Peter W. M. John. (1970). **Statistical Design and Analysis of Experiments**. Macmillan Co., New York.
SIEGEL, Sidney; CASTELLAN, N. John (2006). **Estatística Não-paramétrica para Ciências do Comportamento**. 2. ed. Porto Alegre, RS.
SOUZA, G. S. (1998). **Introdução aos Modelos de Regressão Linear e Não-Linear**. Brasília: Embrapa-SPI / Embrapa-SEA.
WERKEMA, Maria Cristina Catarino; AGUIAR, Silvio (1996). **Analise de Regressão: Como Entender o Relacionamento Entre as Variáveis de um Processo**. Belo Horizonte, MG: UFMG. Escola de Engenharia.

Tabela 1: **Distribuição de Qui-Quadradro χ^2** com os valores críticos de Qui-Quadradro tais que a probabilidade de a variável aleatória χ^2 ser maior do que χ^2_c vale α, ou seja, Prob($\chi^2 \geq \chi^2_c$) = α.

Graus de Liberdade	Valores de α									
	0,995	0,99	0,975	0,95	0,50	0,10	0,05	0,025	0,01	0,005
1	0,00	0,00	0,00	0,00	0,45	2,71	3,84	5,02	6,63	7,88
2	0,01	0,02	0,05	0,10	1,39	4,61	5,99	7,38	9,21	10,60
3	0,07	0,11	0,22	0,35	2,37	6,25	7,81	9,35	11,34	12,84
4	0,21	0,30	0,48	0,71	3,36	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	4,35	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	5,35	10,64	12,59	14,45	16,81	18,55
7	0,99	1,24	1,69	2,17	6,35	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	7,34	13,36	15,51	17,53	20,09	21,95
9	1,73	2,09	2,70	3,33	8,34	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	9,34	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	10,34	17,28	19,68	21,92	24,72	26,76
12	3,07	3,57	4,40	5,23	11,34	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	12,34	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	13,34	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	14,34	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	15,34	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	16,34	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	17,34	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	18,34	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	19,34	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	20,34	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	21,34	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	22,34	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	23,34	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	24,34	34,38	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	25,34	35,56	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	26,34	36,74	40,11	43,19	46,96	49,64
28	12,46	13,56	15,31	16,93	27,34	37,92	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	28,34	39,09	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	29,34	40,26	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	39,34	51,81	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	49,33	63,17	67,50	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	59,33	74,40	79,08	83,30	88,38	91,95
70	43,28	45,44	48,76	51,74	69,33	85,53	90,53	95,02	100,43	104,21
80	51,17	53,54	57,15	60,39	79,33	96,58	101,88	106,63	112,33	116,32
90	59,20	61,75	65,65	69,13	89,33	107,57	113,15	118,14	124,12	128,30
100	67,33	70,06	74,22	77,93	99,33	118,50	124,34	129,56	135,81	140,17

Tabela Gerada no Excel

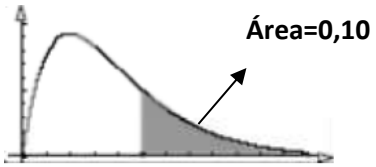
Tabela 2: Distribuição Fisher-Snedecor F com os valores críticos da F tais que a probabilidade de a variável F ser maior que F_c vale 0,05, ou seja, Prob(F ≥F_c)=0,05.



gl Denominador	gl Numerador									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
45	4,06	3,20	2,81	2,58	2,42	2,31	2,22	2,15	2,10	2,05
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93

Tabela Gerada no Excel

Tabela 3: Distribuição Fisher-Snedecor F com os valores críticos da F tais que a probabilidade de a variável F ser maior que F_c vale 0,10, ou seja, Prob(F ≥F_c)=0,10.



gl do Denominador	gl do Numerador									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82
35	2,85	2,46	2,25	2,11	2,02	1,95	1,90	1,85	1,82	1,79
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76
45	2,82	2,42	2,21	2,07	1,98	1,91	1,85	1,81	1,77	1,74
50	2,81	2,41	2,20	2,06	1,97	1,90	1,84	1,80	1,76	1,73
100	2,76	2,36	2,14	2,00	1,91	1,83	1,78	1,73	1,69	1,66

Tabela Gerada no Excel

Tabela 4: Amplitude q para os procedimentos de Tukey.

gl ***(N - k)***	***α***	***k níveis***								
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
5	0,05	3,64	4,6	5,22	5,67	6,03	6,33	6,58	6,8	6,99
	0,01	5,7	6,98	7,8	8,42	8,91	9,32	9,67	9,97	10,24
6	0,05	3,46	4,34	4,9	5,3	5,63	5,9	6,12	6,32	6,49
	0,01	5,24	6,33	7,03	7,56	7,97	8,32	8,61	8,87	9,1
7	0,05	3,34	4,16	4,68	5,06	5,36	5,61	5,82	6	6,16
	0,01	4,95	5,92	6,54	7,01	7,37	7,68	7,94	8,17	8,37
8	0,05	3,26	4,04	4,53	4,89	5,17	5,4	5,6	5,77	5,92
	0,01	4,75	5,64	6,2	6,62	6,96	7,24	7,47	7,68	7,86
9	0,05	3,2	3,95	4,41	4,76	5,02	5,24	5,43	5,59	5,74
	0,01	4,6	5,43	5,96	6,35	6,66	6,91	7,13	7,33	7,49
10	0,05	3,15	3,88	4,33	4,65	4,91	5,12	5,3	5,46	5,6
	0,01	4,48	5,27	5,77	6,14	6,43	6,67	6,87	7,05	7,21
11	0,05	3,11	3,82	4,26	4,57	4,82	5,03	5,2	5,35	5,49
	0,01	4,39	5,15	5,62	5,97	6,25	6,48	6,67	6,84	6,99
12	0,05	3,08	3,77	4,2	4,51	4,75	4,95	5,12	5,27	5,39
	0,01	4,32	5,05	5,5	5,84	6,1	6,32	6,51	6,67	6,81
13	0,05	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32
	0,01	4,26	4,96	5,4	5,73	5,98	6,19	6,37	6,53	6,67
14	0,05	3,03	3,7	4,11	4,41	4,64	4,83	4,99	5,13	5,25
	0,01	4,21	4,89	5,32	5,63	5,88	6,08	6,26	6,41	6,54
15	0,05	3,01	3,67	4,08	4,37	4,59	4,78	4,94	5,08	5,2
	0,01	4,17	4,84	5,25	5,56	5,8	5,99	6,16	6,31	6,44
16	0,05	3	3,65	4,05	4,33	4,56	4,74	4,9	5,03	5,15
	0,01	4,13	4,79	5,19	5,49	5,72	5,92	6,08	6,22	6,35
17	0,05	2,98	3,63	4,02	4,3	4,52	4,7	4,86	4,99	5,11
	0,01	4,1	4,74	5,14	5,43	5,66	5,85	6,01	6,15	6,27
18	0,05	2,97	3,61	4	4,28	4,49	4,67	4,82	4,96	5,07
	0,01	4,07	4,7	5,09	5,38	5,6	5,79	5,94	6,08	6,2
19	0,05	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04
	0,01	4,05	4,67	5,05	5,33	5,55	5,73	5,89	6,02	6,14
20	0,05	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,9	5,01
	0,01	4,02	4,64	5,02	5,29	5,51	5,69	5,84	5,97	6,09
24	0,05	2,92	3,53	3,9	4,17	4,37	4,54	4,68	4,81	4,92
	0,01	3,96	4,55	4,91	5,17	5,37	5,54	5,69	5,81	5,92
30	0,05	2,89	3,49	3,85	4,1	4,3	4,46	4,6	4,72	4,82
	0,01	3,89	4,45	4,8	5,05	5,24	5,4	5,54	5,65	5,76
40	0,05	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,73
	0,01	3,82	4,37	4,7	4,93	5,11	5,26	5,39	5,5	5,6
60	0,05	2,83	3,4	3,74	3,98	4,16	4,31	4,44	4,55	4,65
	0,01	3,76	4,28	4,59	4,82	4,99	5,13	5,25	5,36	5,45
120	0,05	2,8	3,36	3,68	3,92	4,1	4,24	4,36	4,47	4,56
	0,01	3,7	4,2	4,5	4,71	4,87	5,01	5,12	5,21	5,3
∞	0,05	2,77	3,31	3,63	3,86	4,03	4,17	4,29	4,39	4,47
	0,01	3,64	4,12	4,4	4,6	4,76	4,88	4,99	5,08	5,16

f* graus de liberdade associado ao quadrado médio dos resíduos (MS_E).