2
Regressão Linear

Neste capítulo apresentamos um conjunto de técnicas estatísticas, denominadas análise de regressão linear, onde se procura estabelecer a relação entre uma *variável resposta y* e um conjunto de *variáveis de regressão*, ou *variáveis independentes*, $x_1, x_2, ..., x_k$. Cabe lembrar que, no nosso contexto, a variável *y* representa uma característica de qualidade de um processo produtivo e as variáveis $x_1, x_2, ..., x_k$ são os fatores que afetam o processo quando este está em operação.

Função de Resposta

Quando se afirma que a resposta *y* depende dos fatores, isto quer dizer que existe uma relação funcional entre *y* e $x_1, x_2, ..., x_k$ do tipo:

$$y = \Phi(\beta_0, \beta_1, \beta_2, \cdots, \beta_k, x_1, x_2, \cdots, x_k) + \varepsilon$$

onde $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ são os coeficientes desconhecidos e o termo ε (*erro*) representa outras fontes de variabilidade, que não estão contabilizadas em Φ. Assim, ε acumula efeitos tais como erros de medida e outras fontes de variabilidade inerentes ao processo, às vezes denominadas "ruído de fundo".

Geralmente não se conhece essa relação funcional, daí a utilização de modelos lineares de regressão

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \qquad (2.1)$$

nos quais os $p = k + 1$ parâmetros desconhecidos $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ são os coeficientes do modelo de regressão linear.

Cabe registrar que o modelo é dito linear porque é uma função linear dos coeficientes. Modelos que aparentemente são mais complexos podem ser representados pelo Modelo (2.1).

Por exemplo, considere um modelo de segunda ordem com duas variáveis:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon$$

Se fizermos $x_3 = x_1^2$, $x_4 = x_2^2$, $x_5 = x_1 x_2$, $\beta_3 = \beta_{11}$, $\beta_4 = \beta_{22}$ e $\beta_5 = \beta_{12}$, o modelo se torna

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

que é um modelo linear de regressão.

Os polinômios do primeiro grau nas variáveis de regressão são usados em experimentos fatoriais em dois níveis, completos (2^k) ou fracionados (2^{k-p}), e polinômios do segundo grau são usados em experimentos fatoriais em três níveis; completos (3^k) ou os denominados Experimentos Compostos Centrados (*Central Composite Designs*).

Neste capítulo apresentamos métodos de estimação dos coeficientes do modelo linear e para testar a significância dos coeficientes e, por conseguinte, ter indicações de quais fatores têm influência no processo produtivo. Isto geralmente é denominado **ajuste do modelo**. Finalmente, apresentamos métodos para verificar a adequação do modelo ajustado.

Para leituras adicionais a este capítulo e demonstrações, recomendamos: Atkinson, (1985), Atkinson e Riani (2000), Cook e Weisberg (1999), Myers e Montgomery (2002) e Myers, Montgomery e Vining (2002).

2.1. Estimação dos Parâmetros com Mínimos Quadrados

O método dos mínimos quadrados, tradicionalmente denominado de mínimos quadrados ordinário (MQ), é o método clássico de estimação dos parâmetros dos modelos lineares.

Suponha que foram realizadas *n* observações da variável resposta, $y_1, y_2, \ldots, y_n$. Conjuntamente com cada observação de *y* teremos uma observação, ou nível, de cada variável de regressão. Seja x_{ij} a i-ésima observação da variável x_j. Apresentamos os dados na Tabela 2.1.

Podemos escrever a equação do Modelo (2.1) em termos das observações da Tabela 2.1.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad i = 1,2,\ldots,n \qquad (2.2)$$

Tabela 2.1 - Dados para o Modelo de Regressão Linear.

x_1	x_2	...	x_k	y
x_{11}	x_{12}	...	x_{1k}	y_1
x_{21}	x_{22}	...	x_{2k}	y_2
.	.	...	.	.
.	.	...	.	.
.	.	...	.	.
x_{n1}	x_{n2}	...	x_{nk}	y_n

Assume-se que os diversos valores do termo do erro, ε_i, sejam variáveis aleatórias não correlacionadas, com média zero e variância constante σ^2, ou seja:

$$E(\varepsilon_i) = 0 \qquad \text{e} \qquad E(\varepsilon_i \varepsilon_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$$

A equação (2.2) pode ser escrita na forma matricial

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.3}$$

onde

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad e \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{bmatrix}$$

O vetor das observações **y** tem dimensão n, **X** é uma matriz $(n \times p)$ e o vetor dos níveis das variáveis independentes **β** tem dimensão $(p = k + 1)$.

O método dos MQ fornece o valor de **β** que minimiza a soma dos quadrados dos erros ε_i. A soma dos quadrados dos erros é

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \varepsilon_1^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

Os estimadores de mínimos quadrados devem, portanto, satisfazer a

$$\left.\frac{\partial S}{\partial \beta}\right|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0$$

ou

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{2.4}$$

A Equação (2.4) é denominada de equação normal dos mínimos quadrados na forma matricial. Desde que $\mathbf{X'X}$ seja positiva definida, podemos resolver a Equação (2.4) multiplicando ambos os seus membros por $(\mathbf{X'X})^{-1}$. Portanto, os estimadores de mínimos quadrados de $\boldsymbol{\beta}$ são

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y} \tag{2.5}$$

e o modelo de regressão ajustado é

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{2.6}$$

Na forma escalar, o modelo é $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$.

A diferença entre a observação y_i e o valor ajustado $\hat{y}$ é o resíduo $e_i = y - \hat{y}$. O vetor, de dimensão *n*, dos resíduos é $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.

Nos modelos lineares, o método dos MQ produz estimadores não enviesados dos parâmetros $\boldsymbol{\beta}$. Portanto, $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ (Myers *et al.* 2002, pág. 15).

A variância de $\hat{\boldsymbol{\beta}}$ pode ser obtida a partir da matriz de variância-covariância:

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = E\left\{\left[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right]'\right\}$$

que é uma matriz simétrica cujo *i*-ésimo elemento da diagonal principal é a variância do estimador do coeficiente de regressão $\hat{\beta}_i$ e o elemento (*ij*) é covariância entre $\hat{\beta}_i$ e $\hat{\beta}_j$.

Pode-se demonstrar (Myers *et al.* 2002, pág. 15) que:

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X'X})^{-1} \tag{2.7}$$

O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é um estimador linear não enviesado e de variância mínima, o que lhe confere o título de melhor estimador linear não enviesado.

Pode-se demonstrar (Myers *et al.* 2002, pág. 15) que a estimativa da variância σ^2 do erro ε é

$$\hat{\sigma}^2 = \frac{SS_E}{n-p} \tag{2.8}$$

onde SS_E é a soma dos quadrados dos resíduos: $SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

2.2. Testes de Hipótese em Regressão

São testes úteis para verificar quais os parâmetros significativos do modelo. Os procedimentos aqui descritos supõem que os erros ε_i têm distribuição normal e são independentes com média zero e variância constante. Por conseqüência, as observações y_i têm distribuição normal e são independentes com média igual a $\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$ e variância igual a σ^2.

Para a previsão de futuras observações de y, devemos usar modelos parcimoniosos, ou seja, modelos contendo apenas parâmetros significativos. Por conseguinte, devemos executar testes formais para determinar a significância de cada parâmetro.

Teste de Significância para a Regressão (Myers *et al.* 2002, pág. 17)

Este teste verifica se há uma relação linear entre y_i e as variáveis independentes $x_1, x_2, ..., x_k$. As hipóteses são

$$H_0: \beta_1 = ... = \beta_k = 0$$

$$H_1: \beta_j \neq 0 \text{ para ao menos um } j$$

A rejeição de H_0 implica que pelo menos uma das variáveis independentes contribui significativamente para o modelo. A hipótese nula pode ser testada por meio de uma análise de variância (ANOVA). O procedimento de teste começa com o parcelamento da soma total dos quadrados:

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (2.9)$$

onde $\bar{y}$ é a média aritmética da n observações da resposta y_i, e $\hat{y}_i$ é o valor ajustado pelo modelo.

A primeira parcela no membro direito de (2.9) mede o montante da variação de y_i devido à regressão; a segunda parcela é a soma dos quadrados dos resíduos, que mede o montante de variação não explicada pela regressão. A Equação (2.9) pode ser escrita da seguinte forma:

$$SS_T = SS_R + SS_E .$$

Se a hipótese nula H_0: $\beta_1 = ... = \beta_k = 0$ for verdadeira, pode-se demonstrar que SS_R/σ^2 tem distribuição qui-quadrado com k graus de liberdade χ^2_k e que SS_E/σ^2 tem distribuição χ^2_{n-k-1}.

Temos ainda que SS_R e SS_E são independentes e, sendo os respectivos quadrados médios dados por $MS_R = SS_R/k$ e $MS_E = SS_E/n-k-1$, o quociente MS_R/MS_E segue a distribuição $F_{k,\, n\text{-}k\text{-}1}$. A estatística de teste é então

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E}. \tag{2.10}$$

Rejeitamos H_0 se F_0 for maior do que $F_{\alpha,\, k,\, n\text{-}k\text{-}1}$. Podemos, alternativamente, calcular o P-valor, que é a probabilidade de $F_{k,\, n\text{-}k\text{-}1} > F_0$. Caso o P-valor seja menor do que α, rejeitamos H_0.

O coeficiente de determinação múltipla R^2 é definido como sendo o quociente de SS_R e SS_T:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}, \tag{2.11}$$

observando que: $0 \le R^2 \le 1$.

R^2 é a proporção da variabilidade "explicada" pelo modelo. Qualquer variável adicionada ao modelo, seja ela significativa ou não, provoca um aumento em R^2. Por conseguinte, é possível que haja um modelo com valor de R^2 elevado porém capacidade pobre de previsão. Devido a este fato, foi desenvolvido o R^2 ajustado:

$$R^2_{aju} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \tag{2.12}$$

Geralmente, o R^2 ajustado não é incrementado com a inclusão, no modelo, de variáveis desnecessárias. Na verdade, se acrescentamos variáveis desnecessárias ao modelo, o valor de R^2 ajustado deverá diminuir.

Quando R^2 e R^2_{aju} diferem muito, isso é uma indicação de que há parâmetros não significativos no modelo.

Testes para cada Coeficiente (Myers *et al.* (2002, pág. 21)

As hipóteses para testar a significância do coeficiente β_j são

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

Se H_0: $\beta_j = 0$ não é rejeitada, temos indicação de que x_j não deve ser incluída no modelo. A estatística de teste para esta hipótese é

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \tag{2.13}$$

onde C_{jj} é o elemento da matriz $(\mathbf{X}'\mathbf{X})^{-1}$que corresponde a $\hat{\beta}_j$.

A hipótese nula H_0: $\beta_j = 0$ é rejeitada se $|t_0| > t_{\alpha/2,n-k-1}$.

O denominador da Equação (2.13) é o **erro padrão** do coeficiente de regressão $\hat{\beta}_j$, ou seja,

$$ep(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}} \tag{2.14}$$

Exemplo 2.1.

Oliveira (1999) realizou um experimento para encontrar as condições de operação que maximizam a produção de polissacarídeos. Polissacarídeos são polímeros amplamente empregados nas indústrias alimentícia, petrolífera, farmacêutica, cosmética, têxtil, de produtos agrícolas, de tintas, entre várias outras. Os fatores considerados importantes foram: agitação (x_1), expressa em rotações por minuto (rpm), temperatura (x_2), expressa em graus centígrados (^{o}C), e aeração (x_3), expressa em litros de ar por minuto (L/min). A resposta medida foi o rendimento (y), expresso em gramas por litro (g/l), que é a medida da quantidade formada do produto.

Os níveis de cada fator são apresentados na Tabela 2.2. Os valores entre parênteses são os níveis codificados como (1, 0, -1).

Tabela 2.2 - Nível do Fator (Exemplo 2.1).

	Nível do Fator		
Fator	Alto	Médio	Baixo
Agitação (rpm)	800 (1)	650 (0)	500 (-1)
Temperatura (^{o}C)	36 (1)	28 (0)	20 (-1)
Aeração (L/min)	1,5 (1)	1,0 (0)	0,5 (-1)

Na Tabela 2.3 apresentamos os resultados correspondentes aos 16 experimentos realizados.

Tabela 2.3 - Dados Resultantes do Experimento (Exemplo 2.1)**.**

Agitação	Temperatura	Aeração	Resposta
x_1	x_2	x_3	y
-1	-1	-1	2,1
-1	1	-1	3,0
-1	-1	1	2,4
-1	1	1	3,3
1	-1	-1	2,3
1	1	-1	3,3
1	-1	1	2,5
1	1	1	3,7
0	-1	0	2,0
0	1	0	3,0
0	0	-1	5,6
0	0	1	6,0
-1	0	0	5,7
1	0	0	6,1
0	0	0	5,8
0	0	0	5,7

O modelo de segunda ordem a ser ajustado é constituído por $k = 9$ variáveis independentes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \varepsilon$$

O produto $x_i x_j$ é a variável que representa a interação do fator i com o fator j; e a variável x_j^2 é o termo quadrado do fator j.

Na Tabela 2.4, fornecida pela planilha Excel, apresentamos a ANOVA para o modelo de segunda ordem completo. Como o P-valor é inferior a 5%, não rejeitamos a hipótese de que a regressão é significativa, com pelo menos um coeficiente significativo.

Tabela 2.4 - ANOVA do Experimento (Exemplo 2.1).

Fonte de Variação	*gl*	*SS*	*MS*	*F*	*P-valor*
Regressão	9	38,2021	4,24467	932,542	9,75692E-09
Resíduo	6	0,0273	0,00455		
Total	15	38,2294			

Na Tabela 2.5, fornecida pela planilha *Excel*, apresentamos os testes de significância dos coeficientes para o modelo de segunda ordem completo.

Tabela 2.5 - Testes para os Coeficientes.

	Coeficientes	*Erro-padrão*	t_0	*P-valor*
Interseção	5,7155	0,0319	178,94	2,06E-12
X1	0,1400	0,0213	6,56	0,0006
X2	0,5000	0,0213	23,44	3,96E-07
X3	0,1600	0,0213	7,50	0,0003
X1X1	0,2017	0,0416	4,85	0,0028
X2X2	-3,1983	0,0416	-76,97	3,24E-10
X3X3	0,1017	0,0416	2,45	0,0499
X1X2	0,0500	0,0239	2,10	0,0809
X1X3	1,11E-16	0,0239	4,7E-15	1,0000
X2X3	0,0250	0,0239	1,05	0,3350

Na Tabela 2.6, fornecida pela planilha *Excel*, apresentamos os testes de significância dos coeficientes para o modelo de segunda ordem sem as interações x_1x_2, x_1x_3 e x_2x_3.

Tabela 2.6 - Testes para os Coeficientes.

	Coeficientes	*Erro-padrão*	t_0	*P-valor*
Interseção	5,7155	0,0361	158,35	8,12E-17
X1	0,1400	0,0241	5,81	0,0003
X2	0,5000	0,0241	20,74	6,59E-09
X3	0,1600	0,0241	6,64	9,52E-05
X1X1	0,2017	0,0470	4,30	0,0020
X2X2	-3,1983	0,0470	-68,12	1,6E-13
X3X3	0,1017	0,0470	2,17	0,0584

A significância do termo quadrático x_3^2 mudou para cerca de 5,8%. Optamos por excluí-la do modelo.

Na Tabela 2.7, fornecida pela planilha *Excel*, apresentamos os testes de significância dos coeficientes para o modelo de segunda ordem sem as variáveis x_3^2, x_1x_2, x_1x_3 e x_2x_3.

Tabela 2.7 - Testes para os Coeficientes

	Coeficientes	*Erro-padrão*	t_0	*P-valor*
Interseção	5,7386	0,0403	142,23	7,25E-18
X1	0,1400	0,0282	4,96	0,0006
X2	0,5000	0,0282	17,72	6,97E-09
X3	0,1600	0,0282	5,67	0,0002
X1X1	0,2341	0,0521	4,49	0,0012
X2X2	-3,1659	0,0521	-60,78	3,53E-14

Todos os termos permanecem significativos. O modelo é então

$$\hat{y} = 5{,}7386 + 0{,}14x_1 + 0{,}5x_2 + 0{,}16x_3 + 0{,}2241x_1^2 - 3{,}1659x_2^2 \qquad (2.15)$$

2.3. Verificação da Adequação do Modelo

Nesta seção trataremos de verificar se o modelo ajustado é adequado para descrever os dados. É necessário verificar se as suposições feitas não foram violadas, isto é, se os erros ε são normais, independentes e com variância constante.

As propriedades de melhor estimador linear não enviesado dos estimadores de mínimos quadrados não dependem da suposição de normalidade. Entretanto, se a variância não for constante, o estimador de mínimos quadrados, apesar de ser não enviesado, não terá mínima variância, e os erros-padrão dos estimadores dos coeficientes serão maiores que no caso de variância constante. Ademais, a violação das considerações de independência e variância constante pode tornar o modelo instável, no sentido que diferentes amostras podem resultar em modelos significativamente diferentes, levando a conclusões diferentes. Na Seção 4.2 mostramos que o modelo linear para o Exemplo 3.1 é instável.

Por conseguinte, não é prudente contar com o modelo até que a validade dessas suposições seja verificada. A violação de qualquer uma dessas suposições, assim como a adequação do modelo, pode ser investigada pela inspeção dos resíduos. Ademais, é necessário identificar se há observações atípicas (*outliers*) ou observações influentes. Mais adiante, nesta seção, discutiremos a importância desses tipos de observações.

2.3.1 Análise dos Resíduos

A investigação dos resíduos é uma etapa obrigatória de qualquer análise de regressão. Se o modelo é adequado, os resíduos devem se apresentar de forma aleatória, isto é, eles não devem conter nenhum padrão evidente. Desta forma, a verificação do modelo pode ser realizada pela análise de gráficos dos resíduos $e_i = y - \hat{y}_i$.

2.3.1.1. Verificação da Suposição de Normalidade

Um procedimento útil para verificação da consideração de normalidade é o gráfico de probabilidade normal dos resíduos. Sua construção começa com a ordenação dos resíduos $e_1, e_2, \ldots, e_n$ na ordem crescente $e_{(1)}, e_{(2)}, \ldots, e_{(n)}$, ou seja,

$e_{(1)}$ é o menor resíduo e $e_{(n)}$ é o maior resíduo. Os resíduos ordenados $e_{(j)}$ são então plotados *versus* a freqüência cumulativa $(j-0{,}5)/n$. A ordenada do gráfico é representada pelos valores da freqüência cumulativa em uma escala de probabilidade normal.

Vamos ilustrar este procedimento com o modelo (2.15), construído para o Exemplo 2.1.

Na Tabela 2.8 temos os resíduos ordenados e a freqüência cumulativa.

Tabela 2.8 - Resíduos Ordenados e Freqüência Cumulativa**.**

i	*Observação* y_i	*Valor Ajustado* $\hat{y}_i$	*Resíduo* $(e_i = y_i - \hat{y}_i)$	*Resíduos Ordenados*	*Freq. Cum.* $(j-0{,}5)/n$
1	2,1	2,0068	0,0932	-0,1327	0,0313
2	3,0	3,0068	-0,0068	-0,1068	0,0938
3	2,4	2,3268	0,0732	-0,0727	0,1563
4	3,3	3,3268	-0,0268	-0,0727	0,2188
5	2,3	2,2868	0,0132	-0,0386	0,2813
6	3,3	3,2868	0,0132	-0,0268	0,3438
7	2,5	2,6068	-0,1068	-0,0127	0,4063
8	3,7	3,6068	0,0932	-0,0068	0,4688
9	2,0	2,0727	-0,0727	0,0132	0,5313
10	3,0	3,0727	-0,0727	0,0132	0,5938
11	5,6	5,5786	0,0214	0,0214	0,6563
12	6,0	5,8986	0,1014	0,0614	0,7188
13	5,7	5,8327	-0,1327	0,0732	0,7813
14	6,1	6,1127	-0,0127	0,0932	0,8438
15	5,8	5,7386	0,0614	0,0932	0,9063
16	5,7	5,7386	-0,0386	0,1014	0,9688

Para construir o gráfico de probabilidade normal dos resíduos plotamos no eixo horizontal o resíduo ordenado e no eixo vertical plotamos a freqüência cumulativa, em uma escala de probabilidade normal. Na Figura 2.1 apresentamos este gráfico, fornecido pelo *Design Expert*.

Não observamos pontos muito fora do alinhamento. Por conseguinte, não há indicação de que a consideração de normalidade deva ser rejeitada.

O software *ARC* constrói este gráfico de outra maneira. No eixo vertical são plotados os resíduos ordenados e no eixo horizontal, em escala linear, são plotados os valores da inversa da normal padronizada (quartil normal) da freqüência cumulativa correspondente. Na Tabela 2.9 apresentamos estes valores.

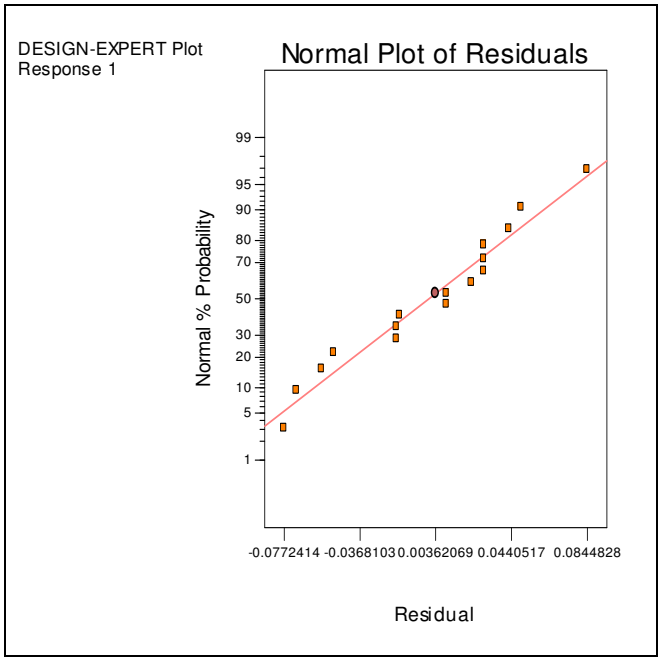


Figura 2.1 – Gráfico de Probabilidade Normal dos Resíduos

Tabela 2.9 - Resíduo Ordenado e Quartil Normal**.**

Resíduos Ordenados	*Freq. Cum.* $(j-0{,}5)/n$	*Quartil Normal*
-0,1327	0,0313	-1,8627
-0,1068	0,0938	-1,3180
-0,0727	0,1563	-1,0100
-0,0727	0,2188	-0,7764
-0,0386	0,2813	-0,5791
-0,0268	0,3438	-0,4023
-0,0127	0,4063	-0,2372
-0,0068	0,4688	-0,0784
0,0132	0,5313	0,0784
0,0132	0,5938	0,2372
0,0214	0,6563	0,4023
0,0614	0,7188	0,5791
0,0732	0,7813	0,7764
0,0932	0,8438	1,0100
0,0932	0,9063	1,3180
0,1014	0,9688	1,8627

Na Figura 2.2 apresentamos o gráfico de probabilidade normal dos resíduos com envelope, fornecido pelo software *ARC*. O procedimento para construção do envelope será descrito adiante.

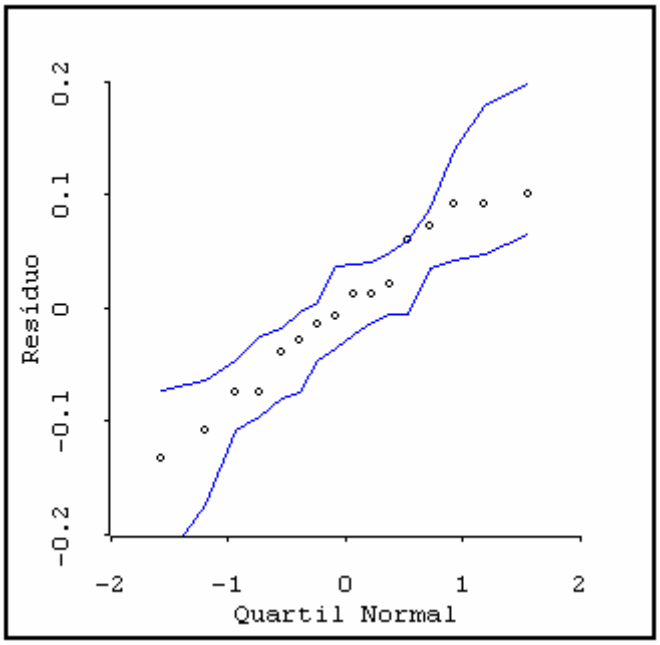


Figura 2.2 – Gráfico de Probabilidade Normal com Envelope

Este gráfico também pode ser feito na planilha *Excel* (mas aí sem envelope), construindo um gráfico de dispersão com (no caso em exemplo) a primeira e terceira coluna da Tabela 2.9.

Devido ao caráter subjetivo da análise desses gráficos, Atkinson (1985) desenvolveu um procedimento de simulação para a construção de linhas em volta dos pontos do gráfico. Tais linhas são denominadas de envelopes. Atkinson afirma que, além do caráter subjetivo da análise destes gráficos, há o problema da super-normalidade. Explicando: no caso de os erros não terem distribuição normal, ainda assim, os resíduos, devido ao fato de serem uma combinação linear de variáveis aleatórias, têm a tendência de terem uma distribuição mais próxima da normal do que os erros. Portanto, pontos aproximadamente alinhados ao longo de uma linha reta não significam necessariamente normalidade na distribuição do erro.

A construção de envelopes procura superar estes dois problemas. Weisberg (1999), fundamentado em Atkinson (1985), descreve como o software *ARC* procede para construir os envelopes. Eis o procedimento:

Construção de envelopes em *ARC*.

1. Traçar o gráfico de probabilidade normal dos resíduos versus os quartis da normal.

2. Supor que os valores dos parâmetros do modelo são os valores verdadeiros, e então gerar um vetor aleatório da resposta, baseado no modelo. Para modelos lineares normais, a *i*-ésima resposta é simplesmente igual ao *i*-ésimo valor ajustado mais um desvio aleatório com distribuição normal padronizada vezes a estimativa do desvio-padrão do erro $\hat{\sigma} = MS_E$.

3. Com as respostas aleatórias obtidas em (2) ajusta-se o mesmo modelo (i.e., reestimam-se os coeficientes do modelo) e calculam-se novos resíduos, que são salvos.

4. Repetir (2) e (3) 19 vezes. Para cada resposta, acrescentar ao gráfico de probabilidade, construído em (1), os valores máximo e mínimo dos resíduos gerados em (3).

Atkinson (1985) afirma que o propósito deste procedimento não é prover uma região de aceitação ou rejeição como em um teste formal, mas prover uma orientação sobre a forma ou linha que pode ser esperada deste gráfico. Mais do que o número de pontos fora do envelope, é importante o afastamento dos pontos em relação ao envelope, com especial atenção para os resíduos com valores mais elevados.

Resíduos Padronizados e Resíduos *Studentizados*

O resíduo padronizado é o quociente entre o resíduo e a estimativa do seu desvio padrão.

$$d_i = \frac{e_i}{\hat{\sigma}} \qquad i = 1,2,\ldots,n \tag{2.16}$$

onde $\hat{\sigma} = \sqrt{MS_E}$.

Esses resíduos têm média zero e variância aproximadamente igual a um. A maioria dos resíduos padronizados deve estar no intervalo $-3 \le d_i \le 3$. Qualquer observação cujo resíduo esteja fora deste intervalo é potencialmente uma observação atípica, e deve ser cuidadosamente examinada, uma vez que pode ser conseqüência de um erro de medição ou de registro. Entretanto, pode também corresponder a uma região especial no espaço da variável independente, onde o modelo ajustado representa pobremente o modelo real. Tal região pode ser de

grande interesse caso corresponda a um máximo (ou mínimo) da resposta, caso seja este o objetivo.

Ao dividirmos os resíduos pela estimativa do desvio-padrão, estamos na verdade dividindo-os pela média do desvio-padrão. De fato, o desvio-padrão dos resíduos não é constante. Ele é diferente para os diversos valores da variável de resposta. Ele é maior para respostas mais próximas da média desta variável. Na definição do resíduo *studentizado* isto é levado em conta. Vimos que os valores ajustados são calculados pela fórmula

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$
$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$
$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \qquad (2.17)$$

A matriz **H**, de dimensão $n \times n$, é conhecida como **matriz chapéu** (*hat*) porque ela "põe" um "chapéu" (acento circunflexo) em $\mathbf{y}$. A matriz chapéu é a matriz de projeção dos valores ajustados sobre os valores observados. Suas propriedades são importantes para a análise dos resíduos, como veremos daqui por diante. Os resíduos escritos sob forma matricial são

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} \qquad (2.18)$$

A matriz **H** é uma matriz simétrica $(\mathbf{H}' = \mathbf{H})$ e idempotente $(\mathbf{HH} = \mathbf{H})$ de dimensão $n \times n$. Da mesma forma, a matriz (**I** – **H**) é simétrica e idempotente.

A partir da Equação (2.18) temos que

$$\begin{aligned} \text{var}(\mathbf{e}) &= \text{var}[(\mathbf{I} - \mathbf{H})\mathbf{y}] \\ &= (\mathbf{I} - \mathbf{H})\text{var}(\mathbf{y})(\mathbf{I} - \mathbf{H})' \end{aligned}$$

Sabe-se que $\text{var}(\mathbf{y}) = \sigma^2\mathbf{I}$ e que a matriz (**I** – **H**) é simétrica e idempotente. Logo,

$$\text{var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H}) \qquad (2.19)$$

e então

$$\text{var}(e_i) = \sigma^2(1 - h_{ii})$$

onde h_{ii} é o i-ésimo elemento da diagonal da matriz **H**.

Os resíduos *studentizados* são então definidos como sendo

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} \tag{2.20}$$

onde $\hat{\sigma}^2 = MS_E$.

Temos que, quando o modelo é correto, $\text{var}(r_i) = 1$ qualquer que seja a localização de $\mathbf{x}_i$. Em muitos casos a diferença entre os resíduos padronizados e studentizados será pequena, contendo ambos informações equivalentes. Entretanto, no método dos mínimos quadrados, pontos com valores elevados de h_{ii} e e_i são potencialmente influentes no cálculo dos parâmetros do modelo. Por conseguinte, para diagnóstico do modelo recomenda-se o uso dos resíduos studentizados.

2.3.1.2. Verificação da Suposição de Independência

A suposição de independência e $E(e) = 0$ é verificada através do gráfico dos resíduos *studentizados* versus valores ajustados.

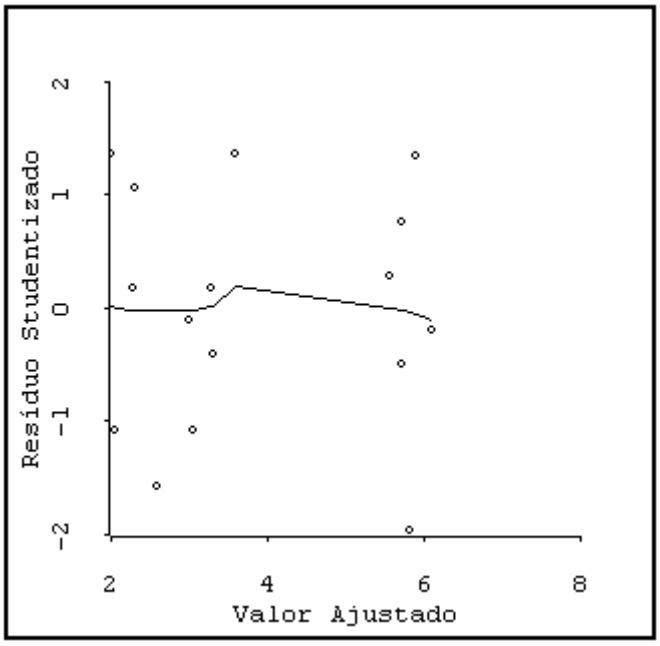


Figura 2.3 – Gráfico dos Resíduos *Studentizados*

No gráfico da Figura 2.3 os resíduos apresentam-se de forma desestruturada; isto é, eles não contêm nenhum padrão evidente, apresentando-se aleatoriamente distribuídos. A linha resultante do amortecimento (*lowess*) é aproximadamente

horizontal e próxima da reta horizontal de ordenada zero, indicando média zero para os resíduos.

A linha de amortecimento *lowess* (*locally weighted scaterplot smoother*), ou linha amortecida, no gráfico de dispersão, localmente ponderada, é uma técnica de estatística não paramétrica, indicada para visualizar tendências nos dados no gráfico. Cook e Weisberg (1999) descrevem na pág. 220 como o software *ARC* constrói esta linha.

Sejam duas variáveis x e y.

1. Selecionar um parâmetro de amortecimento f, no intervalo (0, 1). Geralmente são escolhidos valores entre 0,4 e 0,7. Escolhamos, por exemplo $f = 0{,}5$.
2. Selecionar um ponto x_l e escolhemos os $f \times n$ pontos mais próximos de x_l. No Exemplo 2.1 temos $n = 16$ e como escolhemos $f = 0{,}5$ temos $f \times n = 8$.
3. Com os $f \times n$ pontos mais próximos de x_l, fazer a regressão de y sobre x, com o método dos mínimos quadrados ponderados, com os pesos determinados de tal modo que os pontos mais próximos de x_l tenham maiores pesos, os quais decrescem até zero, à medida que os pontos se afastam de x_l. Em *ARC* é usada uma função triangular para os pesos, que decresce linearmente desde um valor máximo em x_l até zero para o ponto mais afastado.
4. Plotar o valor de $\hat{y}_l$, ajustado na regressão, que corresponde a x_l.
5. Repetir (1) a (4) para todos os valores de x_l e unir os pontos com os valores de $\hat{y}_l$ plotados.

2.3.1.3. Verificação da Suposição de Variância Constante

Cook e Weisberg (1999), pág. 346, propuseram um teste para verificar se a variância é constante. Para isso definem o seguinte modelo da variância da resposta

$$\mathrm{var}(y) = \sigma^2 \exp(\mathbf{x}'\boldsymbol{\gamma}) \tag{2.21}$$

onde $\boldsymbol{\gamma}$ é um vetor de parâmetros e a variância de y é constante quando $\boldsymbol{\gamma} = \mathbf{0}$.

Tomando o logaritmo de (2.21) temos que

$$\ln[\text{var}(y)] = \ln(\sigma^2) + \mathbf{x}'\boldsymbol{\gamma}$$

Freqüentemente a variância é função da média; e nesse caso, comumente, a variância aumenta quando a média aumenta.

Podemos então fazer $\mathbf{x}'\boldsymbol{\gamma} = \lambda\mathbf{x}'\boldsymbol{\beta} = \lambda E(y)$, resultando em

$$\begin{aligned}\ln[\text{var}(y)] &= \ln(\sigma^2) + \lambda\mathbf{x}'\boldsymbol{\beta} \\ &= \ln(\sigma^2) + \lambda E(y)\end{aligned}$$

onde $\ln[\text{var}(y)]$ é uma função linear da média de *y*. A função de variância é constante quando $\lambda = 0$.

Para testar se $\lambda = 0$, Cook e Weisberg usam um teste que requer a correta determinação de *E*(*y*). Para realizar o teste, ajusta-se o modelo linear $\hat{y} = \mathbf{x}'\boldsymbol{\beta}$ via MQ.

Os quadrados dos resíduos e^2 contêm informação sobre a função de variância. Procede-se então à regressão de e^2 sobre $\hat{y}$, por MQ. Calcula-se a soma dos quadrados devido a esta regressão

$$SSreg = \sum_{i=1}^{n}\left(e_i^2 - \sum_{i=1}^{n} e_i^2/n\right)^2 .$$

A estatística de teste é calculada dividindo-se *SSreg* pelo fator de escala $2\left(\sum e_i^2/n\right)$.

$$ET = \frac{SSreg}{2\left(\sum e_i^2/n\right)^2} .$$

Os autores asseguram que *ET* tem distribuição χ_1^2 com um grau de liberdade (número de termos da regressão de e^2 sobre $\hat{y}$), caso λ seja igual a zero.

Para o Exemplo 2.1, considerando a tabela 2.8, fazemos a regressão de e^2 sobre $\hat{y}$ na planilha *Excel*, obtendo *SSreg* = 0,0000016. Em seguida calculamos $2\left(\sum e_i^2/n\right) = 0{,}0000495$. Então,

$$ET = \frac{0{,}0000016}{0{,}0000495} = 0{,}0325$$

Na distribuição qui-quadrado com um grau de liberdade, *ET* = 0,0325 corresponde a um P-valor de 0,857. Então, aceitamos a hipótese de que a variância não aumenta quando a média aumenta.

O gráfico da Figura 2.4 vem confirmar a suposição de variância constante. Nesse gráfico temos o valor absoluto dos resíduos studentizados *versus* o valor ajustado. A linha resultante do amortecimento (lowess) não indica crescimento da variância com o aumento da média.

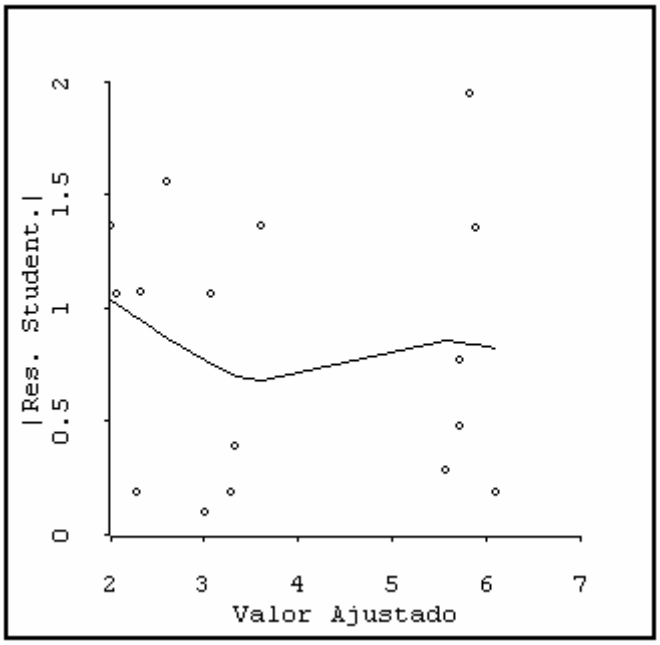


Figura 2.4 – Gráfico Valor Absoluto do Resíduo *Versus* Valor Ajustado

2.3.1.4. Verificação de Observações Atípicas (*Outliers*)

Uma observação atípica é aquela que não combina com o modelo obtido. Essas observações suspeitas podem dever-se a erros de medição da resposta, ou de transposição dos dados, ou de condução destes experimentos. Entretanto, as observações atípicas só devem ser descartadas caso se confirme erro de medida ou transcrição; a resposta obtida pode não ser fruto de um erro, mas um valor real e, caso trate-se de um extremo da resposta, pode mesmo — dependendo do objetivo — corresponder a um bom (senão ao melhor) ponto de operação do processo produtivo. Ademais, observações atípicas pode ser fruto de um modelo inadequado. Uma observação pode ser atípica em um modelo e não a ser em outro.

O resíduo studentizado (r_i) é freqüentemente considerado para um diagnóstico de observações atípicas. Cabe lembrar que, para calcular os resíduos studentizados de cada dado experimental, utiliza-se a média dos quadrados do erro (MS_E) como estimativa da variância ($\hat{\sigma}^2$). A MS_E foi gerada internamente e obtida a partir do ajuste do modelo às n observações. Portanto, o resíduo studentizado representa uma escala interna dos resíduos.

Um outro procedimento é considerar a exclusão da *i-ésima* observação e verificar qual é o efeito na estimativa da resposta *i*. Em particular, verificar se o valor observado y_i "concorda" com o valor ajustado $\hat{y}_{(i)}$, obtido quando a *i-ésima* observação é excluída da regressão, ou seja, faz-se a regressão com a *i-ésima* observação removida. Então, a estimativa de $\hat{\sigma}^2$ passa a ser $S^2_{(i)}$:

$$S^2_{(i)} = \frac{(n-p)MS_E - e_i^2/(1-h_{ii})}{n-p-1} \tag{2.22}$$

$S^2_{(i)}$ é usada no lugar da MS_E para gerar uma escala externa dos resíduos studentizados.

$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}} \quad i = 1,2,\ldots,n \tag{2.23}$$

Myers *et al*. (2002) afirmam que o resíduo t_i usualmente é denominado **R-*Student***, enquanto Atkinson (1985) denomina-o **resíduo de supressão (*deletion residual*)**, e Cook e Weisberg (1999): ***outlier-t***. Já que vamos usar os gráficos do software *ARC*, destes últimos, vamos adotar o nome *outlier-t*.

Em várias situações, o valor de *outlier-t* pouco diferirá em relação ao valor do resíduo *studentizado*. Entretanto, quando o valor $S^2_{(i)}$ da *i-ésima* observação diferir significativamente da MS_E, isso indicará que essa observação tem uma grande influência na determinação dos coeficientes de regressão do modelo, o que torna a estatística *outlier-t* mais sensível a observações atípicas do que o resíduo studentizado.

Quando a *i-ésima* observação se origina da mesma distribuição normal das outras observações, a estatística t_i tem uma distribuição t com n - p - 1 graus de liberdade, o que possibilita um procedimento mais formal para a detecção de observações atípicas, via teste de hipóteses.

Para obter conclusões com relação à existência ou não de observações atípicas, Myers *et al*. (2002) citam a abordagem de Bonferroni, que consiste em comparar todos os *n* valores de $|t_i|$ a $t_{(\alpha/2n),n-p-1}$. Entretanto, do ponto de vista desses autores, essa abordagem não é usualmente necessária, e apenas valores suspeitos devem ser testados. O software *Design-Expert* considera suspeita qualquer observação cujo valor da estatística *outlier-t* seja superior a 3,5 ou inferior a -3,5.

De fato, geralmente, a simples observação do gráfico de *outlier-t versus* os valores ajustados é suficiente para conclusões sobre observações atípicas. Na Figura 2.5 temos este gráfico. Já que todos os valores estão no intervalo (-3,5; 3,5), não há indicação de observações atípicas. Caso houvesse alguma observação fora deste intervalo, procederíamos ao teste com a distribuição *t*.

Atkinson (1985) sugere utilizar para os resíduos *outlier-t* em um gráfico de probabilidade normal com envelope.

O software *ARC* oferece esta opção. Na Figura 2.6 apresentamos o gráfico de probabilidade *t* para os resíduos *outlier-t*. Não observamos pontos muito fora do alinhamento. Por conseguinte, não temos indicação de observações atípicas.

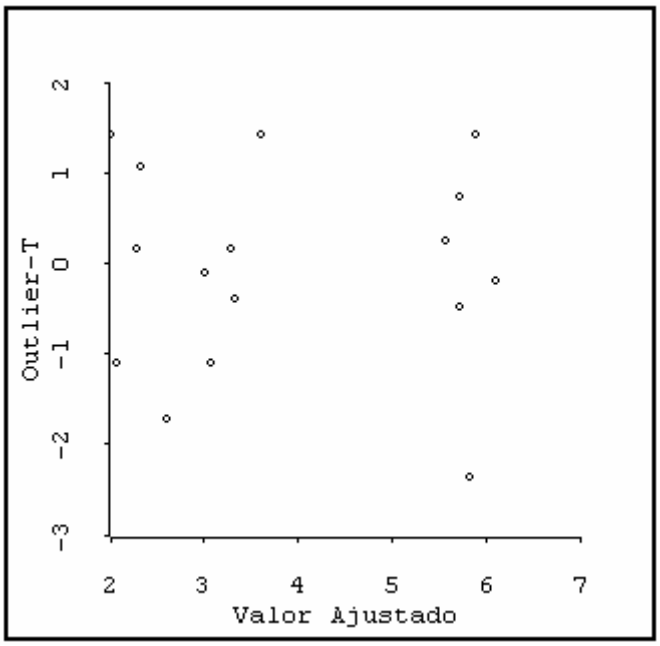


Figura 2.5 – Gráfico Resíduo *Outlier-t Versus* Valor Ajustado

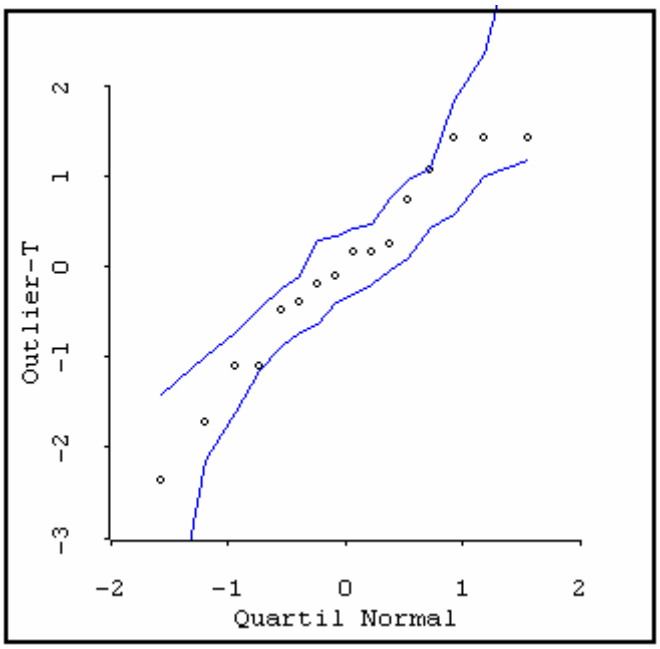


Figura 2.6 – Gráfico de Probabilidade Normal do Resíduo *Outlier-t*

2.3.1.5. Verificação de Observações Influentes

Ocasionalmente algumas observações exercem grande influência na determinação dos coeficientes de regressão do modelo. Tais observações são denominadas **observações influentes**. Pode haver uma certa confusão entre **observações influentes** e **observações atípicas**. Para ilustrar estes dois tipos de observações vamos considerar o conjunto de dados "*Ascomb*", apresentado na Tabela 2.11, fornecido junto com o software *ARC*.

Tabela 2.10 - Dados "*Ascomb*".

(a)	x_1	10	8	13	9	11	14	6	4	12	7	5
	y_1	7,46	6,77	12,74	7,11	7,81	8,84	6,08	5,39	8,15	6,42	5,73
(b)	x_2	8	8	8	8	8	8	8	19	8	8	8
	y_2	6,58	5,76	7,71	8,84	8,47	7,04	5,25	12,5	5,56	7,91	6,89

Na Figura 2.10 temos os gráficos de y_1 versus x_1 (esquerda) e de y_2 versus x_2 (direita).

As retas que aparecem nos gráficos são as retas de regressão por mínimos quadrados. O gráfico da esquerda apresenta uma observação atípica não muito influente na determinação dos coeficientes (marcada com um +). O gráfico da direita apresenta uma observação atípica de grande influência na determinação dos coeficientes (marcada com um ×). Por conseguinte, uma observação atípica

pode não ser muito influente (gráfico da esquerda) ou muito influente (gráfico da direita).

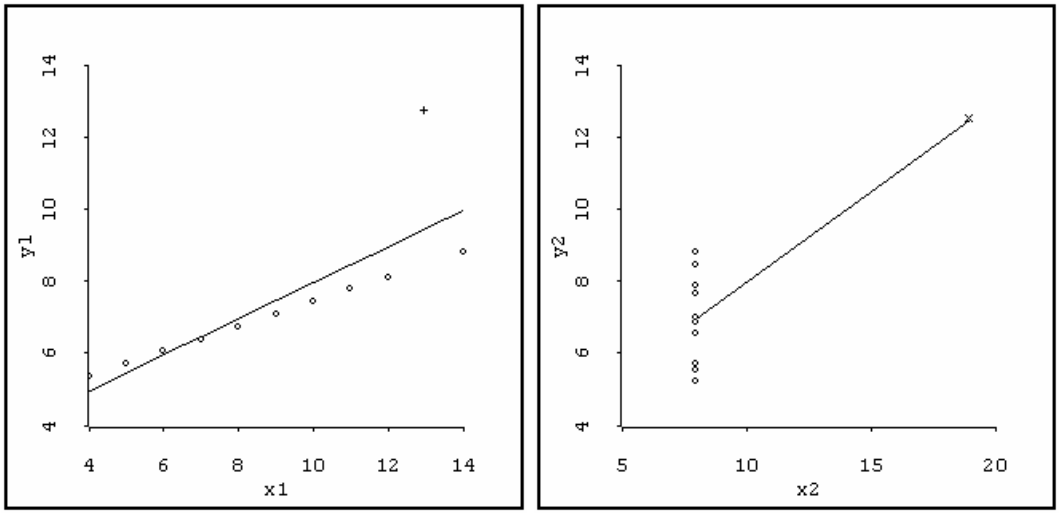


Figura 2.7 – Gráficos de y_1 *Versus* x_1 e de y_2 *Versus* x_2

Portanto, a detecção de observações atípicas deve ser considerada simultaneamente com a detecção das observações que exercem grande influência na determinação dos coeficientes de regressão do modelo.

Distância de Cook

Como foi visto, a disposição dos pontos no espaço das variáveis de regressão é importante na determinação das propriedades do modelo. Em particular, as observações remotas podem, potencialmente, exercer o efeito de uma "alavanca" nas estimativas dos parâmetros, nos valores previstos e nas estatísticas utilizadas.

A matriz chapéu $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ é útil na identificação dessas observações influentes. Os valores dos elementos h_{ij} da matriz **H** podem ser interpretados como a intensidade da "alavancagem" exercida pelos valores observados (y_i) sobre os valores ajustados ($\hat{y}_i$). Portanto, a inspeção dos elementos da matriz **H** pode revelar pontos potencialmente influentes devido à sua localização no espaço das variáveis independentes.

A atenção deve ser focada nos elementos h_{ii} da diagonal da matriz **H**. Pode-se demonstrar que $\sum_{i=1}^{n} h_{ii} = p$. Portanto, a magnitude média dos elementos h_{ii} da diagonal da matriz **H** é p/n.

Assim, uma forma aproximada, mas eficaz, de verificar se a *i-ésima* observação y_i é um ponto de grande "alavancagem", é verificar se o elemento h_{ii} da diagonal da matriz **H** é maior que $2p/n$, ou seja, se o valor de h_{ii} associado à *i-ésima* observação y_i é duas vezes maior que a média de todos os h_{ii} da diagonal da matriz **H**.

Caso seja constatado que um ou mais valores de h_{ii} são maiores que $2p/n$, então podemos concluir que as observações y_i são pontos de grande "alavancagem" (Myers *et al*. 2002).

No experimento em questão temos $n = 16$ e $p = 6$. Assim, são consideradas como possíveis pontos de alavancagem as observações cujos valores de h_{ii} sejam superiores a $2p/n = 0{,}75$.

Na tabela abaixo apresentamos os valores de h_{ii} no Exemplo 2.1.

i	1	2	3	4	5	6	7	8
h_{ii}	0,41	0,41	0,41	0,41	0,41	0,41	0,41	0,41
i	9	10	11	12	13	14	15	16
h_{ii}	0,42	0,42	0,30	0,30	0,42	0,42	0,20	0,20

Não há valores de h_{ii} maiores que 0,75. Portanto, não há indicação de observações que exerçam uma grande alavancagem sobre os parâmetros estimados, valores previstos e estatísticas empregadas.

Já vimos que, com a diagonal da matriz chapéu (**H**) identificamos pontos de potencial influência devido à sua localização no espaço das variáveis independentes. Entretanto, é desejável considerar, na medição da influência, não apenas a localização desses pontos mas, também, o valor da observação y_i.

Cook (1977) sugeriu a utilização de uma medida do quadrado da distância entre a estimativa dos mínimos quadrados $\hat{\boldsymbol{\beta}}$ e uma estimativa obtida excluindo-se o *i-ésimo* ponto, $\hat{\boldsymbol{\beta}}_{(i)}$.

Em geral, a medida desta distância, denominada distância de Cook, pode ser expressa como sendo:

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{pMS_E} \qquad i = 1,2,\ldots,n\,. \qquad (2.24)$$

Um ponto *i* correspondendo a um grande valor de D_i exerce uma influência considerável sobre a estimativa dos mínimos quadrados de **β**. Uma expressão alternativa para D_i:

$$D_i = \frac{(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})}{pMS_E} \qquad i = 1,2,\ldots,n\,. \tag{2.25}$$

Portanto, podemos interpretar D_i, também, como a medida da mudança no vetor dos valores ajustados quando não usamos a observação *i* para estimar **β**.

A versão escalar da expressão para a estatística D_i é (ver Atkinson, 1985)

$$D_i = \frac{e_i^2 h_{ii}}{p\hat{\sigma}^2(1-h_{ii})^2} = \frac{r_i^2 h_{ii}}{p(1-h_{ii})^2} \tag{2.26}$$

A distância de Cook provê uma ordenação das observações em termos da sua influência sobre o vetor das estimativas dos coeficientes. A intenção não é aplicar um teste formal, e sim fornecer uma ajuda para detectar as observações influentes. Cook e Weisberg (1999) afirmam que é **conveniente** analisar casos em que $D_i > 0{,}5$ e é **sempre importante** analisar casos em que $D_i > 1$. Esta análise consiste em verificar se a observação é realmente influente ou se é conseqüência de um modelo inadequado. Se o modelo for inadequado, deve-se construir outro modelo

Na Figura 2.10 temos o gráfico da distância de Cook. Não há indicação de observação influente. Todos os valores são inferiores a 0,5.

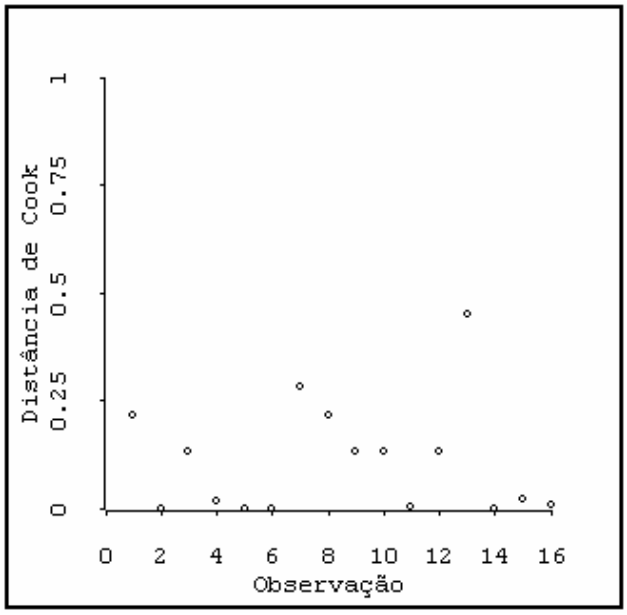


Figura 2.8. Gráfico da Distância de Cook

2.4 – Otimização do Processo

Para otimização do processo é conveniente escrever o modelo na forma seguinte: $\hat{y} = \hat{\beta}_0 + \mathbf{x}'\mathbf{b} + \mathbf{x}'\mathbf{B}\mathbf{x}$

Os vetores **x** e **b** e a matriz **B** da equação 5-39 são:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} \hat{\beta}_{11} & \hat{\beta}_{11}/2 & \cdots & \hat{\beta}_{1k}/2 \\ \hat{\beta}_{11}/2 & \hat{\beta}_{22} & \cdots & \hat{\beta}_{2k}/2 \\ \vdots & \vdots & & \vdots \\ \hat{\beta}_{1k}/2 & \hat{\beta}_{2k}/2 & \cdots & \hat{\beta}_{kk}/2 \end{bmatrix}$$

2.4.1 - Ponto Estacionário

Para achar as condições ótimas de operação é recomendável calcular a localização do ponto estacionário. Um ponto estacionário é aquele que possui uma localização tal que as derivadas parciais da equação para os *k* valores são nulas, isto é, $\partial\hat{y}/\partial x_1 = \partial\hat{y}/\partial x_2 = \ldots = \partial\hat{y}/\partial x_k = 0$.

A derivada de $\hat{y}$ com relação aos elementos do vetor **x** igualada a zero é

$$\frac{\partial\hat{y}}{\partial\mathbf{x}} = \hat{\mathbf{b}} + \mathbf{2}\hat{\mathbf{B}}\mathbf{x} = 0$$

Resolvendo a equação, temos o ponto estacionário:

$$\mathbf{x_s} = -\frac{1}{2}\hat{\mathbf{B}}^{-1}\hat{\mathbf{b}}$$

Para o Exemplo 2.1 temos as seguintes matrizes e vetores

$$\hat{\mathbf{B}} = \begin{bmatrix} 0{,}2017 & 0{,}0250 & 0{,}0000 \\ 0{,}0250 & -3{,}1983 & 0{,}0125 \\ 0{,}0000 & 0{,}0125 & 0{,}1017 \end{bmatrix} \qquad \hat{\mathbf{B}}^{-1} = \begin{bmatrix} 4{,}9525 & 0{,}0387 & -0{,}0048 \\ 0{,}0387 & -0{,}3122 & 0{,}0384 \\ -0{,}0048 & 0{,}0384 & 9{,}8258 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 0{,}1400 \\ 0{,}5000 \\ 0{,}1600 \end{bmatrix}$$

O ponto estacionário, em termos das variáveis codificadas, é obtido como se segue:

$$\mathrm{x}_s = -\frac{1}{2}\hat{\mathbf{B}}^{-1}\mathbf{b} = -½\begin{bmatrix} 4{,}9525 & 0{,}0387 & -0{,}0048 \\ 0{,}0387 & -0{,}3122 & 0{,}0384 \\ -0{,}0048 & 0{,}0384 & 9{,}8258 \end{bmatrix} \times \begin{bmatrix} 0{,}1400 \\ 0{,}5000 \\ 0{,}1600 \end{bmatrix} = \begin{bmatrix} -0{,}3560 \\ 0{,}0723 \\ -0{,}7953 \end{bmatrix}$$

A estimativa da resposta no ponto estacionário é obtida como se segue:

$$y_s = \hat{\beta}_0 + \frac{1}{2}\mathbf{x}_s'\mathbf{B}^{-1}\mathbf{b} = 5{,}7155 + \frac{1}{2}\begin{bmatrix}-0{,}3560 & 0{,}0723 & -0{,}7953\end{bmatrix}\times\begin{bmatrix}0{,}1400\\0{,}5000\\0{,}1600\end{bmatrix} = 5{,}645$$

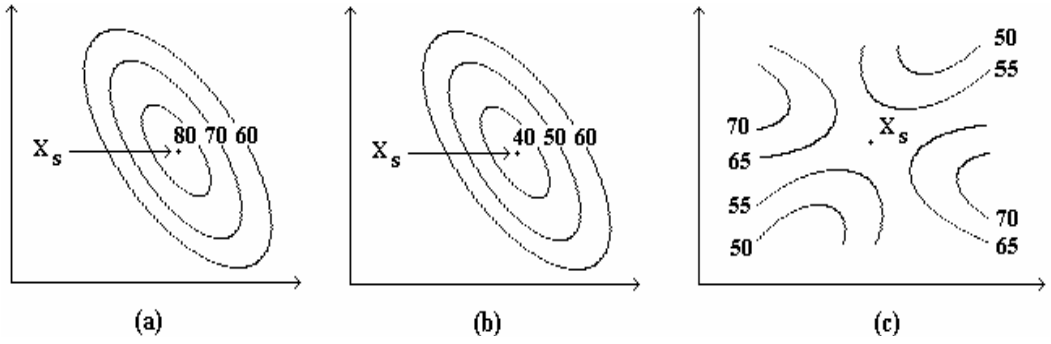
O ponto estacionário pode ser um máximo, mínimo ou ponto de sela. Pode-se caracterizar a superfície de resposta na vizinhança imediata deste ponto, procedendo-se à análise canônica, que consiste em se transformar o modelo para um novo sistema de coordenadas, denominado *forma canônica do modelo ajustado*, que utiliza o ponto estacionário $\mathbf{x}_s$ como origem, resultando na equação

$$\hat{y} = y_s + \lambda_1 w_1^2 + \lambda_2 w_2^2 + \cdots + \lambda_k w_k^2 .$$

Nesta equação, o termo y_s representa a estimativa da resposta no ponto estacionário, os termos w_i correspondem às variáveis independentes transformadas e as constantes λ_i são as raízes características, ou autovalores, da matriz **B**, dados por $\det|\mathbf{B} - \lambda\mathbf{I}| = 0$

1. Se os autovalores $\lambda_1, \lambda_2, \cdots \lambda_p$ são todos negativos, o ponto estacionário é um ponto de máximo.
2. Se os autovalores $\lambda_1, \lambda_2, \cdots \lambda_p$ são todos positivos, o ponto estacionário é um ponto de mínimo.
3. Se alguns dos autovalores $\lambda_1, \lambda_2, \cdots \lambda_p$ são positivos e outros, negativos, o ponto estacionário é um ponto de sela. [ver, p. ex., Myers & Montgomery, (2002)].

Na figura abaixo temos a representação dos três tipos de pontos estacionários: (a) Ponto de máximo, (b) Ponto de mínimo e (c) Ponto de sela.



Temos então

$$\mathbf{B}-\lambda\mathbf{I}=\begin{bmatrix} 0{,}2017-\lambda & 0{,}0250 & 0{,}0000 \\ 0{,}0250 & -3{,}1983-\lambda & 0{,}0125 \\ 0{,}0000 & 0{,}0125 & 0{,}1017-\lambda \end{bmatrix}$$

As raízes da equação $\det|\mathbf{B}-\lambda\mathbf{I}|=0$ são $\lambda_1=-3{,}1985$, $\lambda_2=0{,}1017$ e $\lambda_3=0{,}2019$.

Logo, a forma canônica do modelo ajustado é

$$\hat{y}=5{,}6450+0{,}2019w_1^2-3{,}1985w_2^2+0{,}1017w_3^2$$

Desta forma, como as três raízes possuem sinais variados entre positivo e negativo, conclui-se que o ponto estacionário é um ponto de sela.

Nesta equação, os termos w_1, w_2 e w_3 correspondem, respectivamente, aos fatores transformados agitação, temperatura e aeração. Quanto maior for o módulo da raiz característica λ_i que multiplica a variável independente transformada w_i, maior será a alteração do valor da variável resposta do sistema, em função da variação dos níveis do fator correspondente.

A sensibilidade do modelo, em sua forma canônica, com relação aos fatores, está diretamente associada às grandezas dos coeficientes λ_i. Isto significa, que quando há valores de λ_i muito pequenos ($\lambda_i \approx 0$), a variável resposta é praticamente insensível à variável w_i (já que ela é multiplicada por estes pequenos valores de λ_i). Por outro lado, quando há valores grandes de λ_i, a variável resposta $\hat{y}$ é extremamente sensível à variável w_i (já que ela é multiplicada por estes valores grandes de λ_i).

No exemplo em questão o rendimento do processo é moderadamente sensível aos fatores agitação (λ_1=0,20) e aeração (λ_3=0,10). Com relação ao fator temperatura (λ_2=-3,20), conclui-se que o sistema é muito sensível à sua variação.

2.4.2. Ponto de Máximo Rendimento

O objetivo do experimento é achar o ponto de máximo rendimento. No exemplo 2.1 temos as estimativas dos coeficientes do modelo completo:

$$\hat{y}=5{,}7155+0{,}1400x_1+0{,}5000x_2+0{,}1600x_3+0{,}0500x_1x_2+$$
$$+0{,}025x_2x_3+0{,}2017x_1^2-3{,}1983x_2^2+0{,}1017x_3^2$$

O objetivo é:

maximizar $\hat{y}$.

Sujeito a:

$$-1 \le x_1 \le 1, \quad -1 \le x_2 \le 1, \quad -1 \le x_3 \le 1$$

Resolvendo, com o módulo de otimização do software *Design Expert*, encontramos as seguintes coordenadas para o ponto ótimo:

$$x_1 = 1{,}00, x_2 = 0{,}09, x_3 = 1{,}00$$

Para os fatores em escala natural temos

$$x_1 = 800 \text{ rpm}, x_2 = 28{,}7^0\,\text{C}, x_3 = 1{,}5 \text{ l/min}$$

Para este ponto a previsão do rendimento é $\hat{y} = 6{,}34$ g/l

2.5. Intervalos para a Média da Resposta e para a Previsão da Resposta

Dado um ponto $x_{01}, x_{02}, \ldots, x_{0k}$, no espaço das variáveis regressoras, temos o vetor

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

Podemos obter um intervalo de confiança para a média da resposta e o intervalo de previsão para uma futura observação de um valor individual da resposta.

Intervalo de Confiança para a Média da Resposta

A média da resposta no ponto $\mathbf{x}_0$ é

$$\mu_{y|\mathbf{x_0}} = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \ldots + \beta_k x_{0k}.$$

O estimador da média da resposta neste ponto é

$$\hat{\mu}_{y|\mathbf{x}_0} = \hat{y}(\mathbf{x}_0) = \mathbf{x}_0'\hat{\boldsymbol{\beta}}. \tag{2.27}$$

O estimador não é enviesado, porque

$$E[\hat{\mu}_{y|\mathbf{x}_0}] = E(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \mathbf{x}_0'\boldsymbol{\beta} = \mu_{y|\mathbf{x}_0}.$$

A variância do estimador da média da resposta é

$$\text{var}\left[\hat{\mu}_{y|\mathbf{x}_0}\right] = \text{var}\left[\hat{y}(\mathbf{x}_0)\right] = \text{var}\left(\mathbf{x}_0'\hat{\boldsymbol{\beta}}\right). \tag{2.28}$$

Na Equação (2.7) tem-se que:

$$\text{var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

Então,

$$\text{var}\left[\hat{\mu}_{y|\mathbf{x}_0}\right] = \sigma^2\mathbf{x}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_0.$$

Como a distribuição de *y* é normal, o quociente

$$t = \frac{\hat{\mu}_{y|\mathbf{x}_0} - \mu_{y|\mathbf{x}_0}}{\text{var}\left(\hat{\mu}_{y|\mathbf{x}_0}\right)} \tag{2.29}$$

tem distribuição *t* com (*n*-*p*) graus de liberdade.

Portanto, para um intervalo de confiança de 100 (1-α)% tem-se que:

$$-t_{\alpha/2,n-p} \le \frac{\hat{\mu}_{y|\mathbf{x}_0} - \mu_{y|\mathbf{x}_0}}{\sqrt{\text{var}\left(\hat{\mu}_{y|\mathbf{x}_0}\right)}} \le t_{\alpha/2,n-p}. \tag{2.30}$$

Nas Equações (2.27) e (2.28) tem-se que $\hat{\mu}_{y|\mathbf{x}0} = \hat{y}\left(\mathbf{x}_0\right)$ e $\text{var}\left[\hat{\mu}_{y|\mathbf{x}_0}\right] = \sigma^2\mathbf{x}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_0$.

Substituindo em (2.30) tem-se que:

$$-t_{\alpha/2,n-p} \le \frac{\hat{y}\left(\mathbf{x_0}\right) - \mu_{y|\mathbf{x}_0}}{\sqrt{\sigma^2\mathbf{x}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_0}} \le t_{\alpha/2,n-p}.$$

O que é equivalente a

$$\hat{y}\left(\mathbf{x_0}\right) - t_{\alpha/2,n-p}\sqrt{\sigma^2\mathbf{x}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_0} \le \mu_{y|\mathbf{x}_0} \le \hat{y}\left(\mathbf{x_0}\right) + t_{\alpha/2,n-p}\sqrt{\sigma^2\mathbf{x}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_0} \tag{2.31},$$

que é o intervalo de confiança de 100 (1-α)% para a média da resposta no ponto $x_{01}, x_{02}, \ldots, x_{0k}$.

Exemplo 2.1 (cont.)

Obtivemos o modelo:

$$\hat{\mu} = 5{,}7386 + 0{,}14x_1 + 0{,}5x_2 + 0{,}16x_3 + 0{,}2341x_1^2 - 3{,}1659x_2^2 .$$

Ou seja, o vetor dos estimadores dos coeficientes é

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 5{,}7386 \\ 0{,}14 \\ 0{,}5 \\ 0{,}16 \\ 0{,}23411 \\ -3{,}1659 \end{bmatrix}.$$

O objetivo do experimento de Oliveira (1999) foi encontrar as condições de operação que maximizam a produção do polissacarídeo estudado.

O ponto de operação ótimo obtido dentro da região de experimentação foi:

	Nível do Fator	
Fator	Natural	Codificado
Agitação (rpm)	800	1
Temperatura (^{o}C)	28,64	0,08
Aeração (l/min)	1,5	1

No ponto de operação ótimo temos que

$$\mathbf{x}_0' = \begin{bmatrix} 1 & 1 & 0{,}08 & 1 & 1 & 0{,}0064 \end{bmatrix}.$$

Com a Equação 2.27 calculamos a estimativa da média da resposta neste ponto ótimo:

$$\hat{y}(\mathbf{x_0}) = \mathbf{x}_0'\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 1 & 0{,}08 & 1 & 1 & 0{,}0064 \end{bmatrix} \begin{bmatrix} 5{,}7386 \\ 0{,}14 \\ 0{,}5 \\ 0{,}16 \\ 0{,}23411 \\ -3{,}1659 \end{bmatrix} = 6{,}2924 .$$

A matriz **X** das variáveis regressoras é

$$\mathbf{X} = \begin{array}{c} \begin{matrix} 1 & x_1 & x_2 & x_3 & x_1^2 & x_2^2 \end{matrix} \\ \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & -1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

A matriz $(\mathbf{X}'\mathbf{X})^{-1}$ é

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0{,}205 & 0 & 0 & 0 & -0{,}114 & -0{,}114 \\ 0 & 0{,}1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0{,}1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}1 & 0 & 0 \\ -0{,}114 & 0 & 0 & 0 & 0{,}341 & -0{,}159 \\ -0{,}114 & 0 & 0 & 0 & -0{,}159 & 0{,}314 \end{bmatrix}$$

Com a Equação 2.28 calculamos a variância de $\hat{y}(\mathbf{x_0})$:

$$\text{var}[\hat{y}(\mathbf{x_0})] = \sigma^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = \sigma^2 0{,}4853$$

A estimativa de σ^2 é $\hat{\sigma}^2 = MS_E = 0{,}00796$.

Portanto, a variância de $\hat{y}(\mathbf{x_0})$ é

$$\text{var}[\hat{y}(\mathbf{x_0})] = \sigma^2 0{,}4853 = (0{,}00796)(0{,}4853) = 0{,}003863 .$$

Com a Equação (2.31) calculamos o intervalo de confiança de 95% para a média da resposta:

$$6{,}2924 - t_{0{,}25{,}9}\sqrt{0{,}003863} \le \mu_{y|\mathbf{x}_0} \le 6{,}2924 + t_{0{,}25{,}9}\sqrt{0{,}003863}$$

$$6{,}2924 - (2{,}2622)(0{,}062153) \leq \mu_{y|\mathbf{x}_0} \leq 6{,}2924 + (2{,}2622)(0{,}062153)$$

ou

$$6{,}1518 \leq \mu_{y|\mathbf{x}_0} \leq 6{,}4330$$

Intervalo de Previsão para uma Futura Resposta

O nosso modelo no ponto $\mathbf{x}_0$ é

$$y(\mathbf{x}_0) = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \ldots + \hat{\beta}_k x_{0k} + \varepsilon .$$

ou

$$y(\mathbf{x}_0) = \mathbf{x}_0' \hat{\boldsymbol{\beta}} + \varepsilon$$

A estimativa de uma nova resposta neste ponto é a mesma estimativa da média:

$$E[y(\mathbf{x}_0)] = \mathbf{x}_0' \hat{\boldsymbol{\beta}} .$$

A variância de uma nova resposta neste ponto é

$$\text{var}[y(\mathbf{x}_0)] = \text{var}(\mathbf{x}_0' \hat{\boldsymbol{\beta}} + \varepsilon) = \text{var}(\mathbf{x}_0' \hat{\boldsymbol{\beta}}) + \text{var}(e) = \sigma^2 \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 + \sigma^2$$

ou

$$\text{var}[y(\mathbf{x}_0)] = \sigma^2 [\mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 + 1] \tag{2.32}$$

e o intervalo de 100 (1-α)% de probabilidade para uma nova resposta no ponto $x_{01}, x_{02}, \ldots, x_{0k}$ é

$$\hat{y}(\mathbf{x_0}) - t_{\alpha/2,n-p} \sqrt{\text{var}[y(\mathbf{x}_0)]} \leq y(\mathbf{x}_0) \leq \hat{y}(\mathbf{x_0}) + t_{\alpha/2,n-p} \sqrt{\text{var}[y(\mathbf{x}_0)]}$$

$$\begin{aligned} &\hat{y}(\mathbf{x_0}) - t_{\alpha/2,n-p} \sqrt{\sigma^2 [\mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 + 1]} \leq y(\mathbf{x}_0) \\ &\leq \hat{y}(\mathbf{x_0}) + t_{\alpha/2,n-p} \sqrt{\sigma^2 [\mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 + 1]} \end{aligned} \tag{3.33}$$

Com a Equação 2.30 calculamos a variância de $\hat{y}(\mathbf{x_0})$:

$$\text{var}[y(\mathbf{x_0})] = \sigma^2 [\mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 + 1] = \sigma^2 (0{,}4853 + 1) = \sigma^2 (1{,}4853)$$

A estimativa de σ^2 é

$$\hat{\sigma}^2 = MS_E = 0{,}00796 .$$

Portanto, a variância de $y(\mathbf{x_0})$ é

$$\text{var}[y(\mathbf{x_0})] = \sigma^2 1{,}4853 = (0{,}00796)(1{,}4853) = 0{,}011823 .$$

Com a Equação 2.33 calculamos o intervalo de probabilidade de 95% para uma nova resposta:

$$6{,}2924 \pm t_{0{,}25{,}9}\sqrt{0{,}011823} \le y(\mathbf{x}_0) \le 6{,}2924 + t_{0{,}25{,}9}\sqrt{0{,}011823}$$

$$6{,}2924 \pm (2{,}2622)(0{,}10873) \le y(\mathbf{x}_0) \le 6{,}2924 + (2{,}2622)(0{,}10873)$$

ou

$$6{,}0464 \le y(\mathbf{x}_0) \le 6{,}5384$$

2.5. Mínimos Quadrados Ponderados

Na utilização do **método ordinário dos mínimos quadrados** pressupõe-se que os diversos valores da resposta, *y*, sejam variáveis aleatórias não correlacionadas, com média zero e variância constante σ^2. Ou seja: $\text{var}(y_i) = \sigma^2$.

No **método dos mínimos** quadrados **ponderados** as suposições são as mesmas, exceto que a variância pode não ser a mesma para todas as observações.

Neste caso supõe-se que:

$$\text{var}(y_i) = \sigma^2 v_i = \frac{\sigma^2}{w_i} \tag{2.34}$$

onde o termo w_i, associado à resposta *i*, é denominado **peso** correspondente a esta observação.

Podem-se citar três situações em que a variância do termo do erro não é constante:

1. O conhecimento prévio de características teóricas do processo produtivo pode ser usado para determinar os pesos.
2. A variabilidade da resposta pode estar relacionada com as variáveis de regressão $x_1, x_2, ..., x_k$, como foi visto no modelo 2.21.
3. A distribuição da variável resposta não é normal, como as distribuições gama, normal inversa e lognormal, como será visto no Capítulo 3.

Na forma matricial tem-se que a variância da resposta **y** é

$$\text{var}(\mathbf{y}) = \sigma^2\mathbf{V} = \sigma^2\begin{bmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_n \end{bmatrix}$$

onde **V** é uma matriz diagonal porque as repostas são não correlacionadas.

Seja $\mathbf{W} = \mathbf{V}^{-1}$. Como a matriz $\mathbf{V}$ é diagonal, a matriz $\mathbf{W}$ também é uma matriz diagonal cujos elementos da diagonal principal são os pesos $w_i, w_2, \ldots, w_n$.

Então, modelo considerado é

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \dot{\boldsymbol{\varepsilon}} \qquad \text{var}(\dot{\boldsymbol{\varepsilon}}) = \text{var}(\mathbf{y}) = \sigma^2 \mathbf{W}^{-1}, \tag{2.35}$$

onde $\dot{\boldsymbol{\varepsilon}} = [\dot{\varepsilon}_1, \dot{\varepsilon}_2, \ldots \dot{\varepsilon}_n]$.

Em mínimos quadrados ponderados é feita uma transformação de $\mathbf{y}$ para $\mathbf{W}^{1/2}\mathbf{y}$. Com isso o modelo passa a ser

$$\mathbf{W}^{1/2}\mathbf{y} = \mathbf{W}^{1/2}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{var}(\boldsymbol{\varepsilon}) = \sigma^2,$$

onde $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n]$ e $\varepsilon_i = \dot{\varepsilon}_i \sqrt{w_i}$.

As equações normais em mínimos quadrados ponderados são

$$(\mathbf{X}'\mathbf{W}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{W}\mathbf{y}$$

e o vetor

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} \tag{2.36}$$

é o vetor dos **estimadores de mínimos quadrados ponderados**.

2.6 *Forward Search*

A *Forward Search* (FS) é um procedimento gráfico, proposto por Atkinson e Riani (2000, Cap 2), que monitora os resíduos e outras estatísticas descritas neste capítulo. O objetivo da FS é não só detectar observações atípicas, mas, também, verificar se o modelo usado é apropriado. Há casos em que as observações atípicas são conseqüência de um modelo inadequado. Ao usarmos o modelo linear com mínimos quadrados (MQ), podemos detectar observações atípicas influentes, e o mesmo pode não acontecer quando usamos os modelos lineares generalizados (MLG). O Exemplo 4.2 é ilustrativo. Quando usamos MQ, na Seção 4.1.1, identificamos uma observação atípica, o que não ocorreu quando usamos MLG na Seção 4.1.3.

Se os parâmetros do modelo fossem conhecidos, não haveria dificuldade em detectar observações atípicas. A dificuldade ocorre devido à existência de observações atípicas nos dados usados para a estimativa dos parâmetros. Na FS,

dividem-se os dados em duas partes, uma parte "limpa" e outra com observações atípicas. A parte limpa é usada na estimação dos parâmetros.

Já vimos uma forma simples de dividir os dados. Nas definições de *outlier-t* e distância de Cook, excluímos uma observação e estimamos os parâmetros com as restantes. O problema com este método é que pode haver múltiplas observações atípicas.

Atkinson e Riani (2000) citam os métodos propostos por Cook e Weisberg (1982), Atkinson (1985) e Chatterjee e Hadi (1988), nos quais um pequeno número de observações, talvez duas ou três, são excluídas de cada vez. Entretanto, nestes métodos, o número de combinações pode crescer de maneira explosiva, o que não acontece na FS.

No procedimento da FS, são feitas sucessivas estimativas dos parâmetros com a regressão linear do modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, com $m \leq n$ observações, sendo $\hat{\beta}_j$ o estimador de β_j.

Com as n observações, os resíduos são $e_i(\hat{\boldsymbol{\beta}}) = y_i - \mathbf{x}'\hat{\boldsymbol{\beta}}$ (i = 1, 2, ...,n).

A FS começa usualmente com um sub-conjunto de p dados (p é o número de parâmetros do modelo), escolhidos a partir do conjunto total de n dados. Para a escolha do sub-conjunto inicial são avaliados todos os $\binom{n}{p}$ sub-conjuntos possíveis, ou, alternativamente, um número inferior, por exemplo 1000, caso o número de sub-conjuntos possíveis seja elevado. Para cada sub-conjunto ajustamos por MQ um modelo com as p observações e, então, selecionamos o sub-conjunto cuja mediana dos quadrados dos resíduos seja mínima, obtendo um sub-conjunto supostamente livre de observações atípicas. Temos então um modelo ajustado para as p observações escolhidas, ficando as $(n - p)$ observações restantes para a identificação de observações atípicas.

A vantagem de usar a mediana do quadrado dos resíduos em vez da sua soma é que a mediana é menos sensível a observações atípicas, o que torna o método de estimação mais "robusto".

Quando minimizamos a soma (ou a média) dos quadrados dos resíduos, uma só observação atípica pode modificar as estimativas dos coeficientes, o que geralmente não acontece com o método que minimiza a mediana do quadrado dos resíduos, usado na FS.

O sub-conjunto inicial vai sendo progressivamente incrementado, em $(n - p)$ etapas, até que todos os n dados sejam usados. Dado um sub-conjunto de $m \geq p$ observações, faz-se a regressão com MQ. A FS passa para um sub-conjunto de dimensão $(m + 1)$, selecionando as $(m + 1)$ observações (do conjunto total de n observações) que tenham os menores resíduos resultantes da regressão com MQ. Os autores afirmam que geralmente uma observação é adicionada ao sub-conjunto anterior em cada etapa, entretanto algumas vezes duas ou mais sejam adicionadas e uma ou mais subtraídas.

Na FS as observações atípicas não são identificadas com testes formais. O que interessa é a evolução das estatísticas nos diversos estágios. No conjunto selecionado, de tamanho m, que vai desde um tamanho p até n, o que se observa é a evolução dos resíduos, e de outras quantidades já vistas neste capítulo, tais como a distância de Cook e as estimativas da variância do termo do erro e do coeficiente de determinação múltipla. Essas quantidades são monitoradas e, quando ocorrem mudanças, podemos associar estas mudanças com os dados introduzidos neste estágio. A interpretação destas mudanças nas quantidades é complementada pela observação das mudanças ocorridas nos resíduos. Os autores destacam os seguintes aspectos da FS.

1. Devido à ausência de observações atípicas, as estimativas de $\boldsymbol{\beta}$, em cada estágio m, não são enviesadas. Por conseguinte, os resíduos e estas estimativas devem permanecer aproximadamente constantes durante a FS.

2. Se, por acaso, temos k observações atípicas e começamos com um sub-conjunto livre de observações atípicas, a FS incluirá estas observações atípicas até o fim do procedimento. Usualmente isto acontece nos k últimos estágios. Até então, os resíduos e a estimativa dos parâmetros devem permanecer aproximadamente constantes.

3. Observações atípicas detectadas em um modelo podem não ser consideradas atípicas para outros modelos. Particularmente, modelos lineares com a resposta transformada ou MLG podem não apresentar as observações atípicas detectadas em modelos lineares sem a resposta transformada. Se os dados são analisados com um modelo inadequado, as k eventuais observações atípicas podem aparecer antes do fim da FS.

Monitoramento das estatísticas:

Em cada etapa são monitoradas várias estatísticas de interesse. Para o caso de regressão linear os autores fornecem (entre outras) as seguintes estatísticas.

- **Estimativa da variância do termo do erro** – Equação (2.8).
- **Coeficiente de determinação múltipla** – Equação (2.11).
- **Resíduos padronizados –** Equação (2.16).
- **Estimativa dos coeficientes** – Equação (2.5).
- **Estatística para testar a significância das estimativas dos coeficientes** – Equação (2.13).
- ***i*-ésimo elemento da diagonal da matriz chapéu** $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ - Equação (2.17).
- **Distância de Cook** – Equação (2.26).

Os autores da FS fornecem um programa, compatível com o S-Plus. O software é composto de três módulos: Regressão Linear, Transformação em Regressão Linear e Modelos Lineares Generalizados. A diferença entre os módulos está nas estatísticas monitoradas.

No modulo de regressão linear são monitoradas as estatísticas acima mencionadas.

No módulo da transformação é monitorada uma estatística de teste, descrita na Seção 4.1.2, com o intuito de identificar a melhor transformação na resposta, entre os membros da família de transformações proposta por Box e Cox (1964).

No módulo de modelos lineares generalizados é monitorada uma estatística de teste, descrita por Atkinson e Riani (2000), pág. 200, com o intuito de identificar a melhor função de ligação, entre os membros da família de funções de ligação apresentada na Seção 3.2.

Exemplo 2.1 (cont.)

Na Figura 2.11 apresentamos os resultados da FS das estimativas da variância do termo do erro (direita) e do coeficiente de determinação múltipla (esquerda).

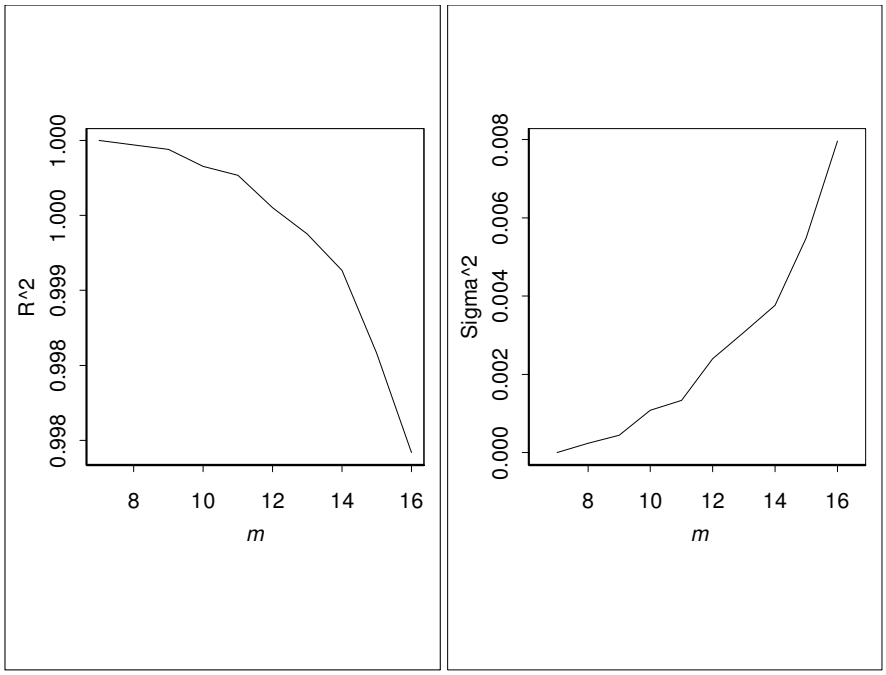


Figura 2.9 – *Forward Search* de R^2 e da Estimativa de σ^2

As duas estatísticas variam sem saltos, o que indica não haver observações atípicas mascaradas.

Na Figura 2.12 apresentamos o gráfico dos resíduos padronizados, durante os estágios da FS. As observações 8 e 13 são as últimas a serem incluídas na FS. Elas se destacam das demais; devem ser consideradas atípicas (ou influentes)? Os gráficos das figuras 2.11 e 2.13 indicam que não (nem atípicas nem influentes).

Na Figura 2.11, quando são incluídas nos dois estágios finais, não se observa alterações bruscas nas duas curvas.

Na Figura 2.13, quando são incluídas nos dois estágios finais, não se observa alterações bruscas nas estimativas dos coeficientes nem na estatística de teste *t*.

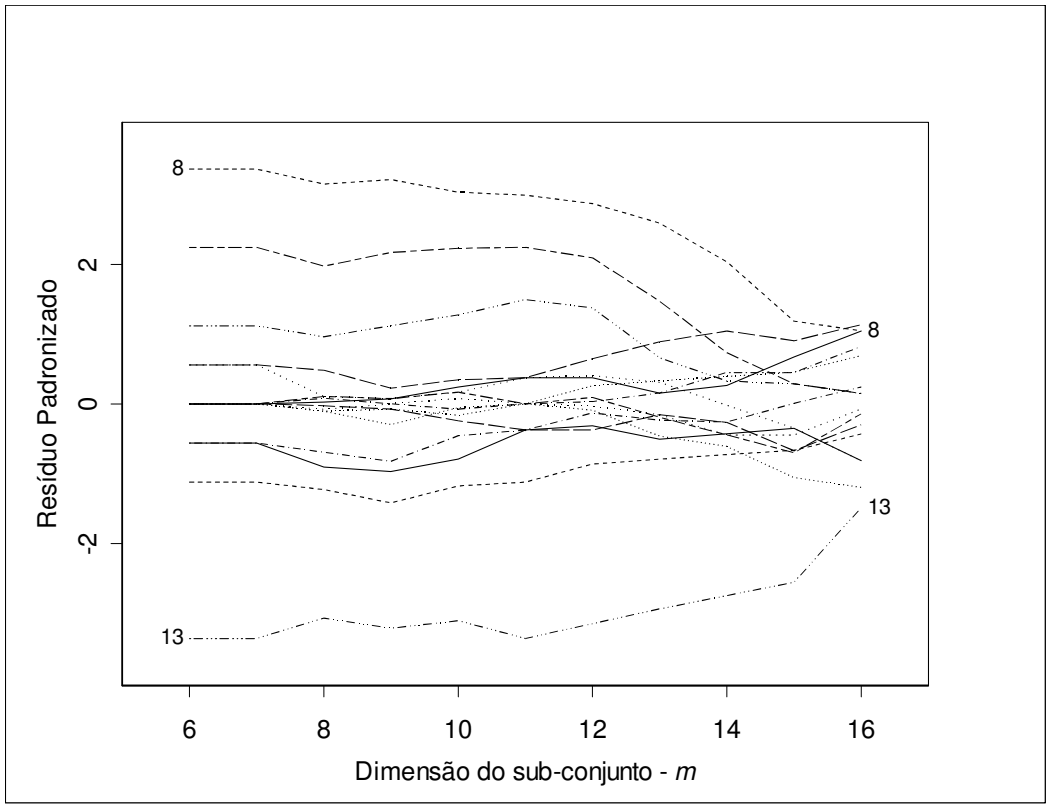


Figura 2.10 – *Forward Search* dos Resíduos Padronizados.

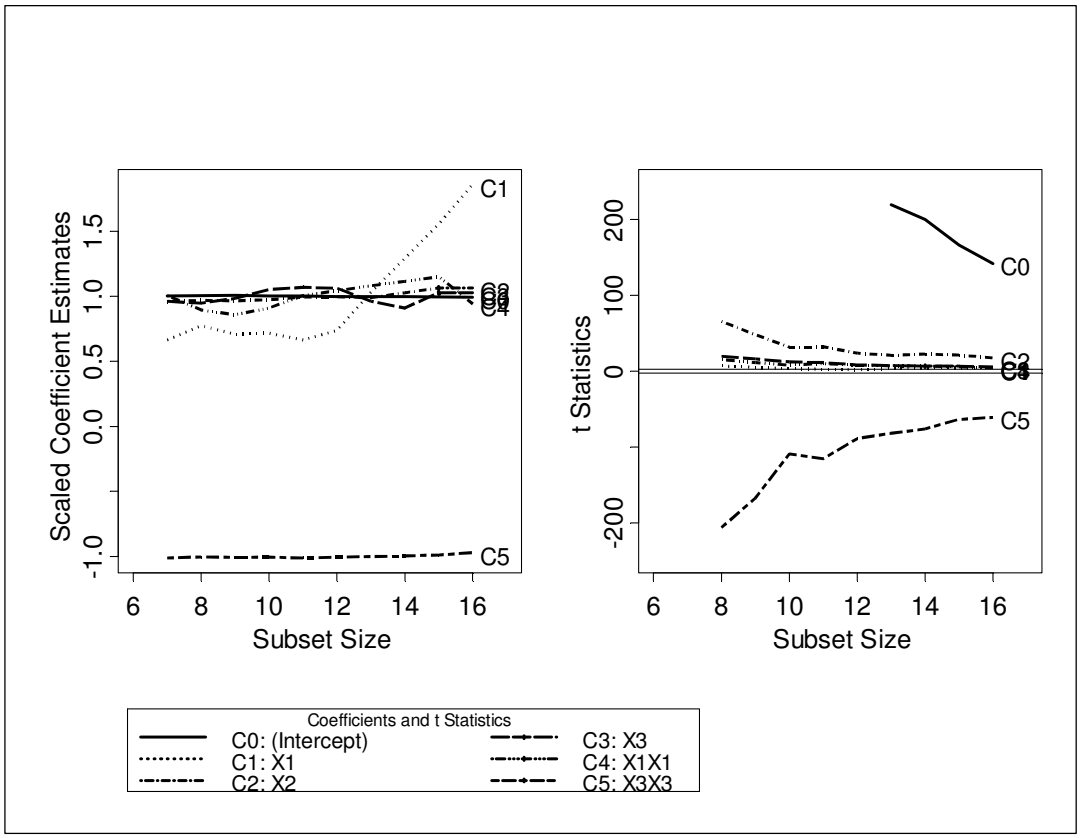


Figura 2.11 – *FS* das Estimativas dos Coeficientes e da Estatística de Teste *t*.