

## 4

### Análise da Média

Neste Capítulo definiremos um roteiro para uma investigação que permita a escolha do modelo mais adequado para representar o comportamento da média de uma característica de qualidade (resposta) de um processo ou produto.

Um modelo adequado para a média é importante para ajustar os fatores que afetam a média e então conduzi-la a um valor desejado. Ademais, um modelo inadequado para a média em experimentos não replicados pode acarretar a identificação de fatores espúrios afetando a dispersão. Isto porque, não havendo replicações, o modelo para a dispersão é obtido através dos resíduos provenientes do modelo da média, o que exige uma identificação correta dos fatores que afetam a média, sob pena de identificação de efeitos espúrios na dispersão (ver Pan, 1999, e McGrath e Lin, 2001). No Capítulo 5 voltaremos ao assunto quando tratarmos da análise da dispersão.

O roteiro será apresentado em vários estágios da investigação. Cada parte destes estágios será explicada detalhadamente, definindo-se as atividades a serem realizadas e quais as decisões que devem ser tomadas. Exemplos ilustrarão os diversos estágios.

Na Seção 4.1 descreveremos as propriedades das variáveis de resposta que caracterizam a qualidade em produtos e processos produtivos. Na Seção 4.2 consideraremos, para modelagem destas variáveis, a metodologia apresentada no Capítulo 2, ou seja: usaremos os modelos lineares gaussianos juntamente com o método dos mínimos quadrados (MQ). Quando os pressupostos do modelo linear — normalidade, aditividade e variância constante — não puderem ser aceitos, a alternativa é a transformação da resposta ou os modelos lineares generalizados.

Na Seção 4.3 usaremos a alternativa de transformar a resposta  $y$  para outra escala, tal como  $\ln(y)$ , aplicando aos dados transformados ainda os mesmos procedimentos da Seção 4.2.

Na Seção 4.4 será descrita a utilização dos modelos lineares generalizados (MLG), incluindo aí a quase-verossimilhança (QV), como descritos no Capítulo 3.

#### **4.1. Variáveis de Resposta**

No começo do processo de modelagem é necessário caracterizar a variável de resposta como discreta ou contínua. A rigor, em aplicações práticas, toda variável aleatória é discreta, mesmo quando o espaço amostral da variável é contínuo. Isto porque o que se faz são medições na variável de resposta, que são limitadas pela precisão do instrumento de medida. Por conseguinte, existe um número finito de medidas possíveis, o que não ocorre com os modelos probabilísticos de variáveis contínuas. Entretanto, quando o espaço amostral é contínuo e o número de medidas possíveis é elevado em relação ao tamanho da amostra, a aproximação pelas distribuições de variáveis contínuas é satisfatória.

##### **Variáveis Contínuas**

As variáveis contínuas são utilizadas quando a característica de qualidade pode ser representada por uma grandeza definida num intervalo dos números reais, como é o caso das medidas de comprimento, superfície, volume, resistência, velocidade, voltagem, pureza, etc. Nestes casos o espaço amostral da variável de resposta é contínuo, sendo então elas serem representadas por variáveis contínuas. Portanto, nestes casos, podem ser consideradas as distribuições normal, gama e normal inversa.

A distribuição normal é simétrica em relação à média e tem variância constante. As distribuições gama e normal inversa são assimétricas e têm a variância aumentando junto com a média. Evidente que, se temos prévias informações sobre a simetria e o comportamento da variância da variável de resposta, poderemos usá-las na escolha da distribuição de probabilidade.

Raramente, em aplicações práticas, as medidas feitas na resposta tomam valores negativos. As grandezas acima mencionadas (comprimento, etc) são sempre positivas, sendo então mais comum haver assimetria com uma cauda mais longa do lado direito da curva. Por conseguinte, modelos com distribuição normal para a resposta devem ser utilizados com precaução. Eles devem ser considerados apenas nos casos em que a probabilidade de uma medida na resposta ser zero é praticamente nula. Nestes casos a normal pode proporcionar um ajuste

satisfatório, quando se observa que a resposta é aproximadamente simétrica em relação à média e que sua variância é aproximadamente constante. Se, adicionalmente, os efeitos a ela transmitidos pelas variáveis de regressão são aditivos, podem-se usar os métodos de análise apresentados na Seção 4.2. Entretanto, mesmo que não se tenha estas evidências, justifica-se a tentativa (de usar os métodos apresentados na Seção 4.2) porque em muitas aplicações industriais essa abordagem tem-se mostrado adequada [Box, *et al.* (1978), Box e Draper, (1987), Montgomery (2001) e Myers e Montgomery (2002)].

Caso um ou mais dos três pressupostos — normalidade, aditividade e variância constante — não possam ser considerados, uma tentativa válida é alterar a escala da variável de resposta, transformando-a e aplicando o método dos MQ (Seção 4.3). Cook e Weisberg (1999), pág. 317, afirmam que é surpreendente o grande número de casos em que os pressupostos são alcançados com uma única transformação. Vários destes casos podem ser vistos em Cook e Weisberg (1999), Atkinson e Riani (2000) e Atkinson (1985). Por conseguinte vale a pena tentar. Ademais, transformação e MLG frequentemente conduzem a modelos similares: ver, por exemplo, Lewis *et al.* (2001).

Entretanto, com uma única transformação, nem sempre se consegue atingir a normalidade, a equalização da variância e a aditividade dos efeitos. Ver o Exemplo 4.2. Portanto, para estes casos devemos considerar os MLG (Seção 4.4). Com os MLG podemos considerar outras distribuições que não a normal, não é necessário que a variância seja constante e conseguimos linearidade através da função de ligação. Temos três “tiros” em vez de um; um para cada alvo.

Ademais, Lewis, *et al.* (2001a) e (2001b), analisando vários exemplos, concluíram que os MLG devem ser preferidos à transformação da resposta, uma vez que em todos os casos os MLG apresentaram melhor desempenho na estimativa da resposta, pois propiciaram intervalos de confiança menores. Por conseguinte, mesmo que com uma única transformação venhamos a alcançar a normalidade, a equalização da variância e a aditividade dos efeitos, recomendamos utilizar os MLG quando for desejável uma estimativa da resposta mais confiável.

## Variáveis Discretas

Muitas das características de qualidade não podem ser representadas por medidas numéricas de grandezas físicas ou químicas. Em tais casos, usualmente classificam-se os produtos inspecionados como **conforme** ou **não conforme** (com) as especificações. A terminologia “**não defeituoso**” e “**defeituoso**” é freqüentemente empregada para estas duas classificações do produto. Um exemplo típico é o de um dispositivo que funciona ou não. Características de qualidade desse tipo são denominadas de **atributos**. O número de produtos defeituosos em uma amostra aleatória segue a distribuição binomial. A otimização é alcançada com a minimização do número esperado de produtos defeituosos por amostra. Para tanto é necessário se estabelecer uma relação (regressão) entre os fatores de produção (variáveis de regressão) e o número de produtos defeituosos por amostra (variável de resposta), de modo a identificar a combinação de valores para os fatores de produção que minimize a resposta.

Outro tipo de característica de qualidade discreta é o número de **defeitos** por unidade de inspeção. Em muitos casos a unidade de inspeção será uma unidade do produto, embora nem sempre. A unidade de inspeção é simplesmente tomada de acordo com a conveniência da coleta de dados. Geralmente a ocorrência de defeitos é um evento de probabilidade baixa, podendo serem inspecionadas várias unidades sem que ocorram defeitos. Um produto pode ter um ou mais defeitos e não ser considerado defeituoso (os defeitos podem ser reparados ou não). Um exemplo típico é o de uma máquina (ou um produto comercializado a metro) que pode ter um ou mais pequenos defeitos que não afetam o seu funcionamento, podendo então ser classificada como conforme (pode até ser comercializada por um preço reduzido se os defeitos não forem reparados). Entretanto, caso o número de defeitos seja elevado ou afete o seu funcionamento, e eles não possam ser reparados, a máquina será então classificada como não conforme. A contagem de defeitos (eventos) é feita em um intervalo contínuo (uni, bi, ou tridimensional). Frequentemente acontecem as três condições seguintes: i) independência dos eventos (defeitos), ii) os defeitos ocorrem aleatoriamente em qualquer ponto do intervalo e iii) não podem ocorrer dois ou mais defeitos em um mesmo ponto do intervalo. Neste caso, o número de defeitos em um determinado intervalo segue a distribuição de Poisson. A otimização é alcançada com a minimização do número

médio de defeitos por produto. Isto pode ser conseguido ao se estabelecer uma relação (regressão) entre os fatores de produção (variáveis de regressão) e a média de defeitos por unidade de inspeção do produtos (variável de resposta).

Como se sabe, nas distribuições de Poisson e binomial a variância é função da média. Ademais, nas situações em que essas distribuições são usadas, os efeitos das variáveis de regressão sobre a variável de resposta geralmente não são transmitidos aditivamente (geralmente são transmitidos multiplicativamente, daí a intensa utilização dos modelos log-lineares, [Myers *et al.* 2002, Cap. 3]). Por conseguinte, os métodos próprios para modelos lineares, propostos na Seção 4.2 não são recomendados para os casos em que se deve usar estas distribuições.

Tendo em vista alcançar os pressupostos de normalidade, variância constante e aditividade, vários autores já sugeriram transformar a escala da resposta. Entretanto, não há razão para se acreditar que, com uma única mudança de escala, esses pressupostos sejam alcançados. McCullagh e Nelder (1989), pág 22, afirmam que, nas aplicações em que a distribuição Poisson é adequada, a normalidade aproximada é alcançada com  $y^{2/3}$ , a equalização da variância com  $y^{1/2}$  e a aditividade com  $\ln(y)$ . Por conseguinte, os métodos propostos na Seção 4.3 não são recomendados para os casos em que se devem usar estas distribuições.

Como foi visto no Capítulo 3, as distribuições de Poisson e binomial são membros da família de distribuições exponencial. Portanto, nestas situações, recomenda-se o uso de modelos lineares generalizados (MLG), que serão vistos na Seção 4.4.

#### **4.2. Estágio 1: Modelo Linear e Mínimos Quadrados (MQ)**

No Estágio 1 é feita uma tentativa de utilizar o modelo linear com erro normal  $\mathbf{y}(\lambda) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , juntamente com o método dos mínimos quadrados (MQ) para estimativa do vetor  $\boldsymbol{\beta}$ . Para este estágio temos duas etapas:

- I) Estimativa e Teste de Significância dos Parâmetros do Modelo**
- II) Testes de Adequação do Modelo**

### I) Estimativa e Teste de Significância dos Parâmetros do Modelo

Para calcular os parâmetros do modelo são utilizados os MQ, descritos no Capítulo 2. Entretanto, eventualmente, temos problemas para identificar quais são os parâmetros significativos do modelo.

Quando o experimento tem repetições, pode-se estimar diretamente o desvio-padrão do “erro puro” (sem contaminação de componentes de erro de ajuste do modelo), que é utilizado para testar a significância dos efeitos dos fatores, o que não acontece quando o experimento não tem repetições. Uma abordagem para superar esta falta de medidas diretas do erro puro é assumir que os efeitos de interações de três ou mais fatores não são significativos, combinando-se então os respectivos quadrados médios para estimar o erro puro. Essa prática provém do princípio dos efeitos esparsos, que afirma que os efeitos principais e as interações de baixa ordem é que, geralmente, são importantes.

Acontece que em experimentos em dois níveis pode não haver um número de observações suficiente para estimar os quadrados médios das interações de três ou mais fatores. Isto é particularmente crítico em experimentos fracionados. Por exemplo, em um experimento  $2^{5-1}$  tem-se que estimar o termo constante, os 5 efeitos principais e os 10 efeitos de segunda ordem, num total de 16 efeitos, dispondo apenas de 16 observações. Ademais, ocasionalmente, em aplicações reais pode ocorrer que interações de três ou mais fatores sejam importantes. Para maiores detalhes ver Montgomery (2001), pg. 245.

Dadas essas considerações, para uma seleção preliminar dos parâmetros quando são usados experimentos fatoriais em dois níveis, recomenda-se o método proposto por Daniel (1959). Para identificar os efeitos fatoriais ativos em experimentos não replicados com dois níveis, ele propôs usar o gráfico de probabilidade normal das estimativas dos efeitos. A estimativa do efeito de um fator é definida como sendo igual à mudança média na resposta, quando se muda o nível do fator. Pode-se mostrar que, em experimentos fatoriais com dois níveis, a estimativa do efeito de um fator é igual ao dobro da estimativa do respectivo coeficiente, quando ajustamos um modelo linear por MQ (Montgomery 2001). Portanto, quando ajustamos um modelo de regressão, podemos calcular as estimativas dos efeitos de cada fator multiplicando por dois a estimativa do coeficiente de regressão correspondente.

Caso nenhum efeito seja importante, as suas estimativas terão distribuição normal. A construção do gráfico de probabilidade normal das estimativas dos efeitos é similar ao gráfico de probabilidade normal dos resíduos, vista na Seção 2.3.1. Os efeitos não importantes seriam distribuídos normalmente com média zero e variância constante e tenderiam a se alinhar numa reta, enquanto os efeitos importantes estariam fora desta reta. Um modelo preliminar é então especificado com os fatores associados aos efeitos que aparentemente são importantes. Os quadrados médios associados aos fatores não significativos são então combinados para estimar o erro puro; em seguida, pode-se testar a significância dos fatores.

Quando são usados experimentos fatoriais em três níveis, não podemos usar o método de Daniel; isso porque, nesse caso os efeitos não são medidos como nos experimentos de dois níveis. Para experimentos em três níveis usamos apenas os MQ. Combinamos então os respectivos quadrados médios de interações de três ou mais fatores para estimar o erro puro (ver o Exemplo 2.1).

## II) Testes de Adequação do Modelo

Para testar a adequação dos modelo são utilizados os gráficos dos resíduos, descritos no Capítulo 2. Vamos apresentar três exemplos ilustrativos.

### Exemplo 4.1.

Myers e Montgomery (2002) analisaram um experimento de moldagem cuja resposta é a contração. São estudados sete fatores (A-G). Trata-se de um experimento fatorial  $2^{7-3}$  com resolução IV. Os geradores são I=ABCE, I=BCDF e G=ACDG. Os dados são apresentados na Tabela 4.1.

Com base no gráfico de probabilidade Normal, mostrado na Figura 4.1, Montgomery considerou  $A$ ,  $B$  e  $AB$  como efeitos importantes. Usando-se o método MQ ajusta-se o seguinte modelo:

$$\text{Modelo M1: } \hat{y} = 27,31 + 6,94A + 17,81B + 5,94AB.$$

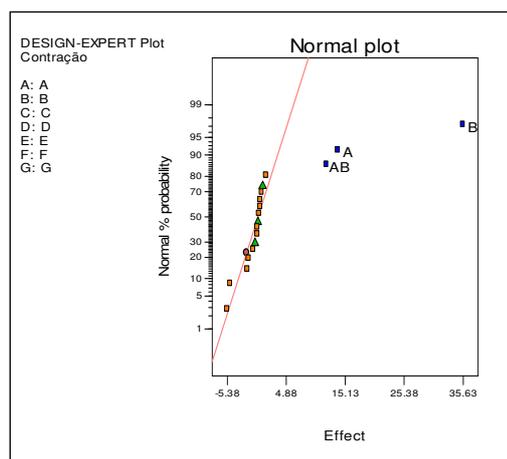
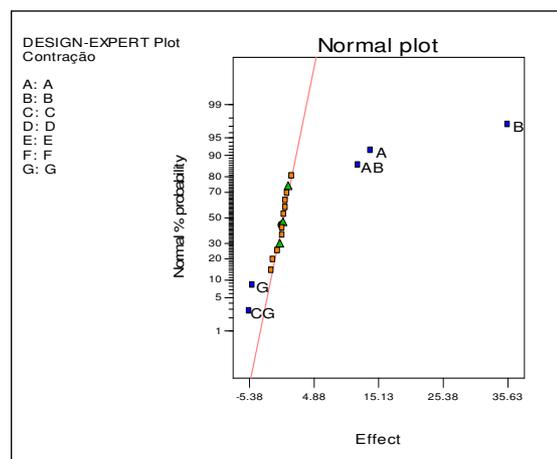
McGrath e Lin (2001) também analisaram este exemplo e, com base no gráfico de probabilidade normal dos efeitos (Figura 4.2), propuseram adicionar os efeitos  $G$  e  $CG$ . Usando-se o método MQ ajusta-se o seguinte modelo:

$$\text{Modelo M2: } \hat{y} = 27,31 + 6,94A + 17,81B + 5,94AB - 2,44G - 2,69CG$$

Para estimativa dos parâmetros dos modelos (M1) e (M2) e para construção dos gráficos das Figuras 4.1 e 4.2, foi utilizado o software *Design Expert*.

**Tabela 4.1** - Dados do Experimento de Moldagem.

A	B	C	D	E=ABC	F=BCD	G=ACD	Y
-1	-1	-1	-1	-1	-1	-1	6
1	-1	-1	-1	1	-1	1	10
-1	1	-1	-1	1	1	-1	32
1	1	-1	-1	-1	1	1	60
-1	-1	1	-1	1	1	1	4
1	-1	1	-1	-1	1	-1	15
-1	1	1	-1	-1	-1	1	26
1	1	1	-1	1	-1	-1	60
-1	-1	-1	1	-1	1	1	8
1	-1	-1	1	1	1	-1	12
-1	1	-1	1	1	-1	1	34
1	1	-1	1	-1	-1	-1	60
-1	-1	1	1	1	-1	-1	16
1	-1	1	1	-1	-1	1	5
-1	1	1	1	-1	1	-1	37
1	1	1	1	1	1	1	52

**Figura 4.1** - Gráfico dos Efeitos - Modelo (M1)**Figura 4.2** - Gráfico dos Efeitos Modelo (M2)

Observando os gráficos é difícil dizer quem está com a razão. O caráter subjetivo deste método eventualmente leva a dúvidas sobre quais fatores devem ser incluídos no modelo. Por via das dúvidas, devemos incluir os fatores  $G$  e  $CG$ , escolhendo, portanto, o modelo M2.

Entretanto, devemos aplicar o *princípio da hierarquia*. Montgomery (2001), pág. 203, descreve este princípio: se um modelo contém uma interação, deve conter também todos os fatores que a compõem. Ele afirma que isto promove uma consistência interna no modelo e que muitos estatísticos seguem este princípio. Nelder e Lee (1998), pág. 169, denominam este princípio de *regra marginal*. Com base nisto, incluímos o fator  $C$ , apesar dele não ter sido considerado significativo.

Para testar a significância dos parâmetros do modelo em questão, são usados os métodos descritos no Capítulo 2. Na Tabela 4.2, fornecida pela planilha *Excel*, apresentamos os testes para significância dos coeficientes para o modelo.

**Tabela 4.2** - Estimativas dos coeficientes e Teste  $t$  para o Modelo M2.

	<i>Coefficientes</i>	<i>Erro-padrão</i>	$t_0$	<i>P-valor</i>
Interseção	27,3125	0,4934	55,35	1,03E-12
A	6,9375	0,4934	14,10	1,98E-07
B	17,8125	0,4934	36,10	4,76E-11
C	-0,4375	0,4934	-0,89	0,398355
G	-2,4375	0,4934	-4,94	0,000803
AB	5,9375	0,4934	12,03	7,52E-07
CG	-2,6875	0,4934	-5,45	0,000408

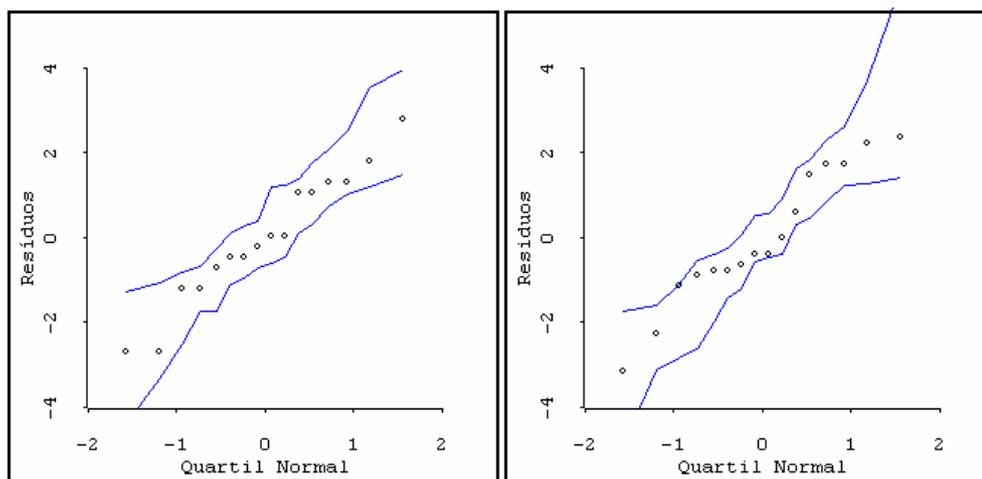
Então, o modelo considerado é:

$$\hat{y} = 27,31 + 6,94A + 17,81B + 5,94AB - 0,44C - 2,44G - 2,69CG$$

Para comparação dos dois modelos, com e sem o fator  $C$ , apresentamos na Figura 4.3 os gráficos de probabilidade normal dos resíduos, fornecidos pelo software *ARC*.

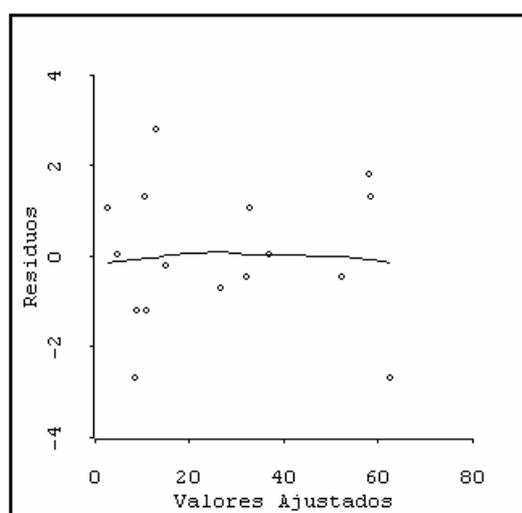
Ambos os gráficos da Figura 4.3 não indicam desvios na normalidade que justifiquem a invalidação dos modelos. Entretanto, notamos que no gráfico da esquerda os pontos estão mais alinhados. É plausível que seja consequência da exclusão de  $C$  no modelo.

Os gráficos apresentados nas figuras 4.4 a 4.7 foram fornecidos pelo software *ARC*.



**Figura 4.3** – Gráficos de Probabilidade Normal dos Resíduos. Modelos com o Fator C (esquerda) e sem o Fator C (direita).

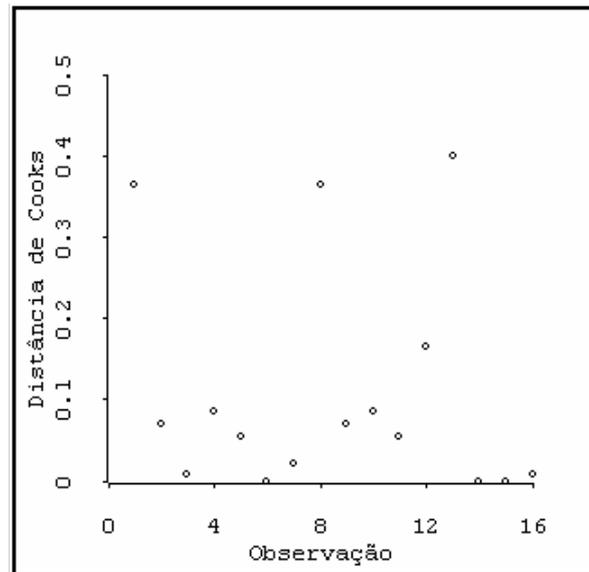
O gráfico da Figura 4.4 apresenta linha resultante do amortecimento (*lowess*) aproximadamente horizontal e próxima da reta horizontal de ordenada zero, indicando que a suposição de erros com média zero é satisfatória.



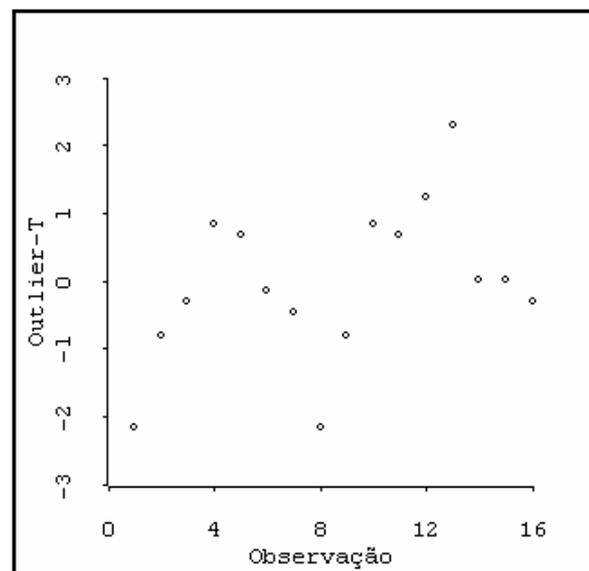
**Figura 4.4** – Gráfico dos Resíduos *versus* Valores Ajustados.

No gráfico da Figura 4.5 todos os valores da distância de Cook são inferiores a 0.5. Portanto, não há indicação de observações influentes.

No gráfico da Figura 4.6 temos os valores dos resíduos *outlier-t*. Cabe lembrar que este tipo de resíduo segue a distribuição *t* com  $n-p-1$  graus de liberdade, o que permite fazer um teste para verificar se existem observações atípicas. Entretanto, esse teste só é necessário quando há resíduo com valor absoluto superior a 3,5 (ver Capítulo 2). Portanto, não há indicação de observações atípicas.



**Figura 4.5** – Distância de Cook



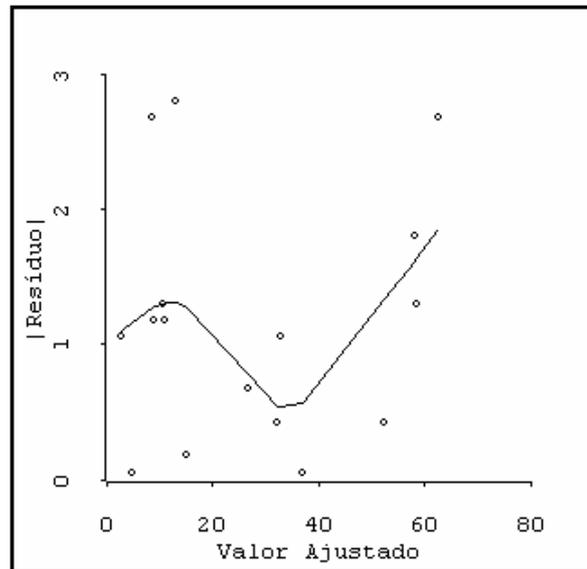
**Figura 4.6** – Gráfico dos Resíduos *Outlier-t*

Para verificar consideração de variância pode constante, temos o gráfico da Figura 4.7. Nesse gráfico, temos o valor absoluto dos resíduos *versus* o valor ajustado. A linha resultante do amortecimento (*lowess*) não indica crescimento sistemático da variância com o aumento da média.

A conclusão *proveniente* da análise do gráfico da Figura 4.7 é apenas uma indicação. Um teste formal para a variância, proposto por Cook e Weisberg (1999), foi mostrado no Capítulo 2. Neste exemplo a estatística de teste é

$$ET = \frac{SSreg(e^2 \text{ sobre } \hat{y})}{2(\sum e_i^2/n)^2} = 0,08.$$

A estatística segue uma distribuição qui-quadrado com um grau de liberdade; o valor obtido corresponde a um P-valor de 0,776. Então, não há indicação de que a variância não seja constante, o que vem corroborar a conclusão proveniente da análise do gráfico da Figura 4.7.



**Figura 4.7** – Gráfico Valor Absoluto do Resíduo *Versus* Valor Ajustado.

Em conclusão, os testes de adequação do modelo indicaram que o modelo M2 é adequado.

### Exemplo 3.1 (cont.)

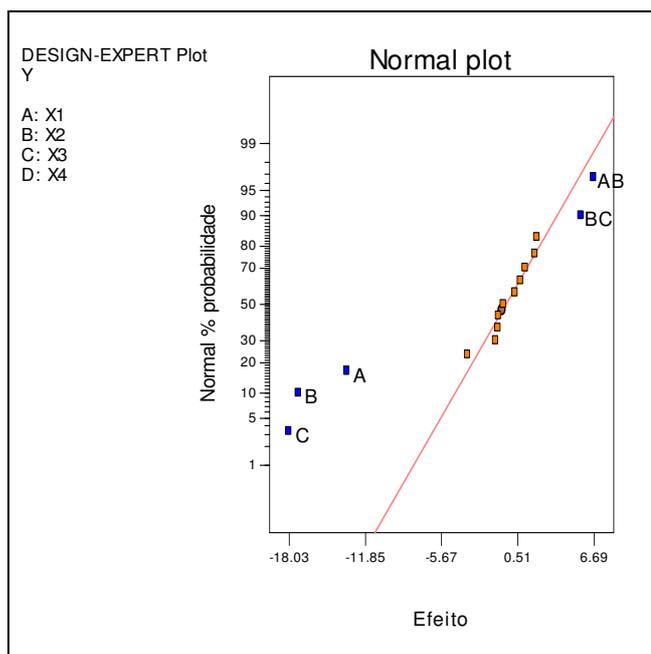
Na Tabela 3.3 apresentamos os dados de um experimento que foi gerado a partir do modelo

$$\mu = \frac{1}{0,040 + 0,008x_1 + 0,010x_2 + 0,012x_3 + 0,005x_1x_3 + 0,002x_2x_3}$$

e distribuição de probabilidade gama com parâmetro de dispersão  $\phi = 0,01$ .

É ilustrativo analisar este experimento como se o modelo fosse linear e os erros normais. Na Figura 4.8, temos o gráfico de probabilidade normal, fornecido pelo software *Design Expert*.

Os efeitos considerados importantes são os das variáveis  $x_1$ ,  $x_2$ ,  $x_3$ ,  $x_1x_2$  e  $x_2x_3$ . Para calcular os coeficientes destas variáveis e testar a sua significância, usamos a planilha *Excel*, que forneceu a Tabela 4.3.



**Figura 4.8** - Gráfico dos Efeitos – Modelo Linear

**Tabela 4.3** - Estimativas dos Coeficientes e Teste *t* para o Modelo Linear.

	<i>Coeficientes</i>	<i>Erro padrão</i>	<i>t</i> <sub>0</sub>	<i>P-valor</i>
Interseção	30,3381	0,8163	37,167	4,73E-12
X1	-6,6681	0,8163	-8,169	9,8E-06
X2	-8,6281	0,8163	-10,570	9,54E-07
X3	-9,0144	0,8163	-11,044	6,35E-07
X1X2	3,3456	0,8163	4,099	0,002149
X2X3	2,8244	0,8163	3,460	0,006121

Construímos então o modelo linear

$$\hat{y} = 30,3381 - 6,6681x_1 - 8,6281x_2 - 9,0144x_3 + 3,3456x_1x_2 + 2,8244x_2x_3.$$

No modelo construído observamos que a interação espúria  $x_1x_2$ , é considerada importante, tomando o lugar da interação  $x_1x_3$ , que é real. Se o objetivo fosse maximizar a média da variável de resposta, escolheríamos o nível -1 para as variáveis  $x_1$ ,  $x_2$ ,  $x_3$ , e estaríamos tomando a decisão correta, uma vez que o modelo real conduz à mesma decisão. O mesmo aconteceria se o objetivo fosse minimizar, quando o nível escolhido seria +1. Entretanto, isto só acontece porque o máximo e o mínimo correspondem a um dos pontos amostrais. Neste caso, não só a decisão seria correta, como as estimativas do máximo ou do mínimo seriam adequadas. Senão, vejamos.

As estimativas das respostas em pontos que não sejam os pontos amostrais pode resultar em diferenças significativas das respostas do modelo em relação à resposta real. Na Tabela 4.4 apresentamos as estimativas das respostas para o modelo linear e para o modelo linear generalizado ajustado na Seção 3.5, referentes a 12 combinações de níveis das variáveis de regressão.

**Tabela 4.4** - Estimativas das Respostas para o Modelo Linear e MLG-Gama.

Nível das Variáveis			Estimativa da Resposta			Diferença %	
$x_1$	$x_2$	$x_3$	Real	Linear	MLG(Gama)	Linear	MLG(Gama)
-1	-1	-1	58,82	60,82	60,61	0,03	0,03
1	-1	-1	43,48	40,79	41,15	-0,06	-0,05
-1	1	-1	30,30	31,22	30,96	0,03	0,02
1	1	-1	25,64	24,58	24,94	-0,04	-0,03
-1	-1	1	37,04	37,14	37,45	0,00	0,01
1	-1	1	18,87	17,11	16,86	-0,09	-0,11
-1	1	1	19,61	18,84	19,19	-0,04	-0,02
1	1	1	12,99	12,20	11,81	-0,06	-0,09
0	0	0	25,00	30,34	23,81	0,21	-0,05
-1	0	0	31,25	37,01	31,35	0,18	0,00
-0,5	0	0	27,78	33,67	27,06	0,21	-0,03
1	-0,5	-0,4	31,25	37,79	30,46	0,21	-0,03

As oito primeiras linhas correspondem às respostas dos pontos amostrais. Para estes pontos as estimativas da resposta, para os dois modelos, são adequadas; a diferença percentual em relação ao modelo real está em torno de 5%, sendo de 11% no pior caso.

As últimas quatro linhas correspondem às respostas de pontos dentro da região de experimentação, mas diferentes dos pontos amostrais. Para esses pontos, o modelo MLG-Gama continua fornecendo respostas cujas diferenças percentuais estão em torno de 5%, enquanto o modelo linear fornece respostas cujas diferenças percentuais estão em torno de 20%. Por conseguinte, a capacidade de predição do modelo linear cai muito quando tomamos pontos que não são pontos amostrais.

Vejamos um exemplo. Suponha que o valor desejado (“alvo”) da característica de qualidade seja 30. Caso usemos o modelo linear, podemos escolher o ponto (0 0 0), cuja resposta é 30,34. Nesse caso a estimativa da resposta estaria 21% acima do valor real (25,00). Caso usemos o modelo MLG-Gama, podemos escolher o ponto (1 -0,5 -0,4), cuja resposta é 30,46. Neste caso a estimativa da resposta estaria 3% abaixo do valor real (31,25).

Cabe lembrar que, no Exemplo 3.1, a variância da resposta tem distribuição gama, portanto cresce com o quadrado da média, e que o modelo linear torna-se instável quando a variância não é constante; o que significa que diferentes amostras podem resultar em modelos significativamente diferentes (ver Seção 2.3). Para ilustrar a instabilidade do modelo, tomamos outra amostra aleatória do experimento, apresentada na Tabela 4.5. O procedimento é o mesmo descrito na Seção 3.5, porém, com a semente para geração dos números aleatórios uniformes igual a 2487913.

**Tabela 4.5** - Dados do Experimento.

$x_1$	$x_2$	$x_3$	$x_4$	$\mu_i$	Uniforme [0, 1]	$\beta_i$	$y_i$
-1	-1	-1	-1	58,83	0,1862	0,5882	53,55
1	-1	-1	-1	43,48	0,1201	0,4348	38,44
-1	1	-1	-1	30,31	0,3556	0,3030	29,10
1	1	-1	-1	25,65	0,6543	0,2564	26,59
-1	-1	1	-1	37,04	0,5603	0,3704	37,48
1	-1	1	-1	18,87	0,9030	0,1887	21,36
-1	1	1	-1	19,61	0,6015	0,1961	20,06
1	1	1	-1	12,99	0,1108	0,1299	11,43
-1	-1	-1	1	58,83	0,0559	0,5882	49,79
1	-1	-1	1	43,48	0,8423	0,4348	47,84
-1	1	-1	1	30,31	0,2261	0,3030	27,99
1	1	-1	1	25,65	0,0926	0,2564	22,32
-1	-1	1	1	37,04	0,9191	0,3704	42,33
1	-1	1	1	18,87	0,3533	0,1887	18,11
-1	1	1	1	19,61	0,0997	0,1961	17,14
1	1	1	1	12,99	0,9263	0,1299	14,92

Com os dados da Tabela 4.4 e o software *Design Expert*, construímos o modelo linear

$$\mu = 20,9031 - 4,7769x_1 - 8,7094x_2 - 7,0494x_3 + 2,3981x_1x_2.$$

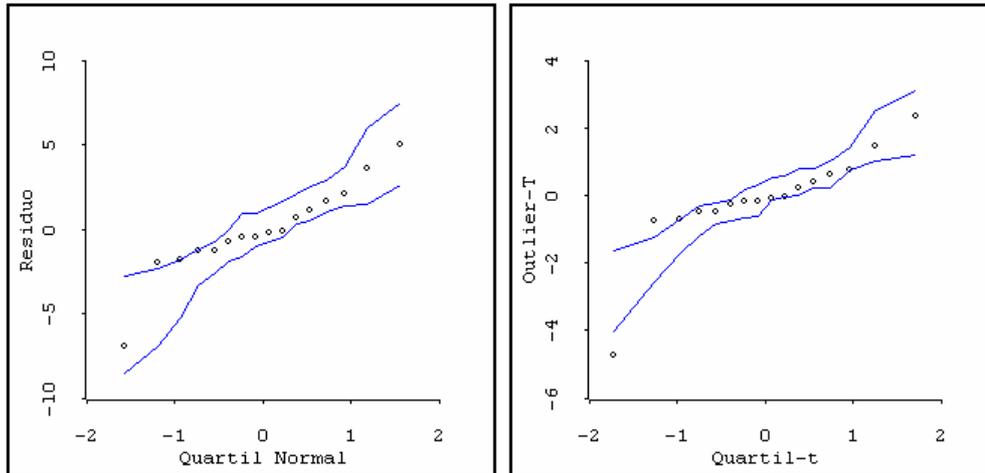
Para facilitar a comparação, repetimos aqui o modelo anterior:

$$\mu = 30,3381 - 6,6681x_1 - 8,6281x_2 - 9,0144x_3 + 3,3456x_1x_2 + 2,8244x_2x_3.$$

Observamos que: i) a interação espúria  $x_1x_2$  permanece no modelo; ii) a interação real  $x_1x_3$ , continua fora do modelo; iii) a interação real  $x_2x_3$  está fora do modelo, e iv) as estimativas de alguns coeficientes são bastante diferentes. Por conseguinte, para outras amostras, este novo modelo deverá levar a diferentes conclusões.

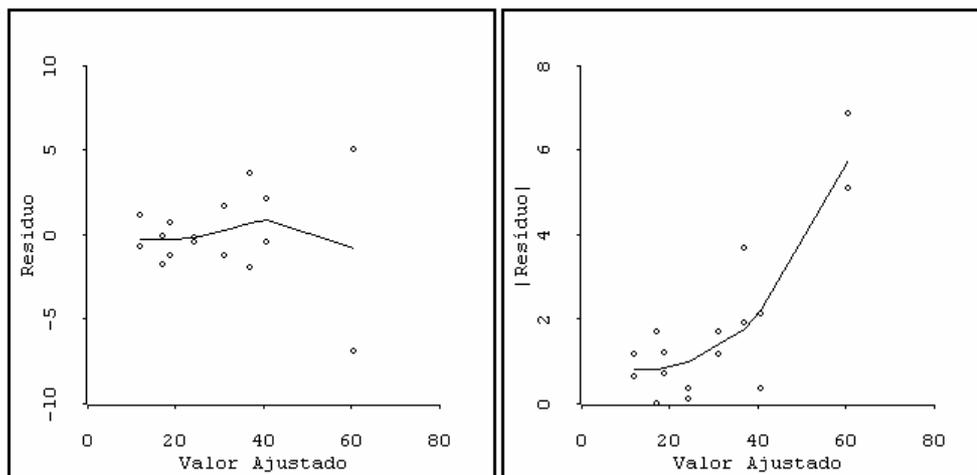
### Testes de Adequação do Modelo Linear para o Experimento do Exemplo 3.1

Na Figura 4.9 apresentamos, à esquerda, o gráfico de probabilidade normal dos resíduos e, à direita, o gráfico de probabilidade normal dos resíduos *outlier-t*, ambos fornecidos pelo software *ARC*. Ambos os gráficos indicam uma observação atípica. Trata-se da observação 9. O valor do *outlier-t* é 4,673, o que corresponde a um P-valor de 0,00088. Entretanto, sabemos que esta observação não é atípica, e sim que o modelo é que é inadequado.



**Figura 4.9** - Gráfico de Probabilidade dos Resíduos.

Na Figura 4.10 apresentamos o gráfico dos resíduos *versus* valores ajustados (esquerda) e o gráfico do valor absoluto dos resíduos *versus* valores ajustados (direita), fornecidos pelo software *ARC*.

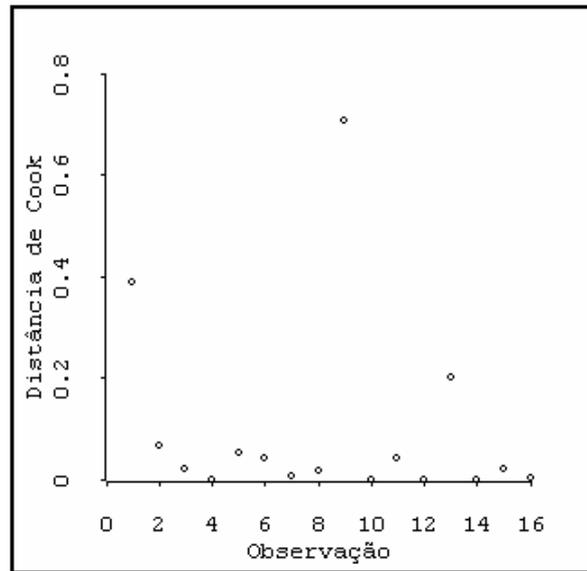


**Figura 4.10** - Gráficos dos Resíduos *versus* Valor Ajustado.

O gráfico da esquerda apresenta linha resultante do amortecimento (*lowess*) aproximadamente horizontal e próxima da reta horizontal de ordenada zero, indicando que a suposição de erros com média zero é satisfatória. No gráfico da direita há indicação de que a variância cresce com a média.

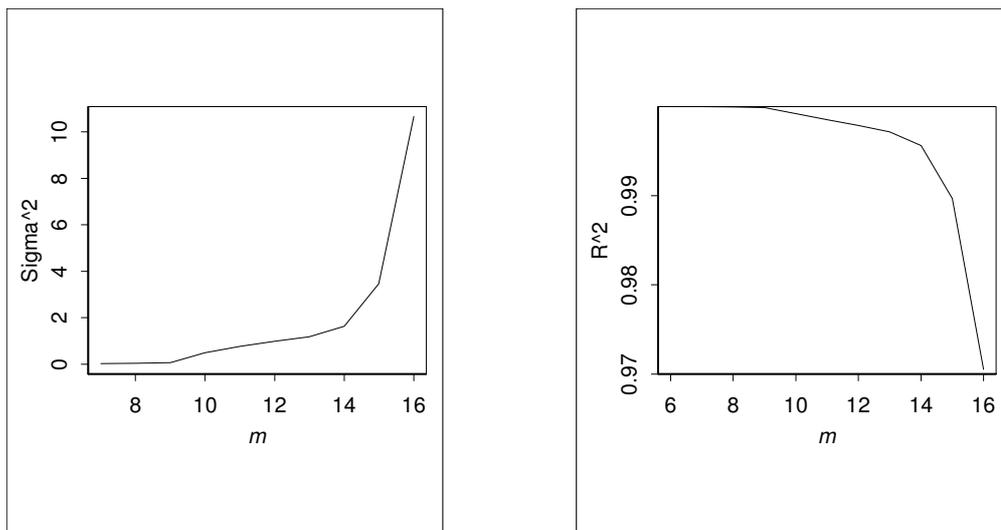
Um teste formal para a variância é o da estatística  $ET$  (ver Capítulo 2). Neste exemplo  $ET = 17.04$  com um P-valor de 0,000366. Então, o teste confirma que a variância não deve ser constante.

Na Figura 4.11 apresentamos o gráfico da distância de Cook, onde temos a indicação de que a observação 9 é influente, além de atípica. Sabemos que esta observação não é atípica, e sim que o modelo é inadequado.



**Figura 4.11** – Distância de Cook.

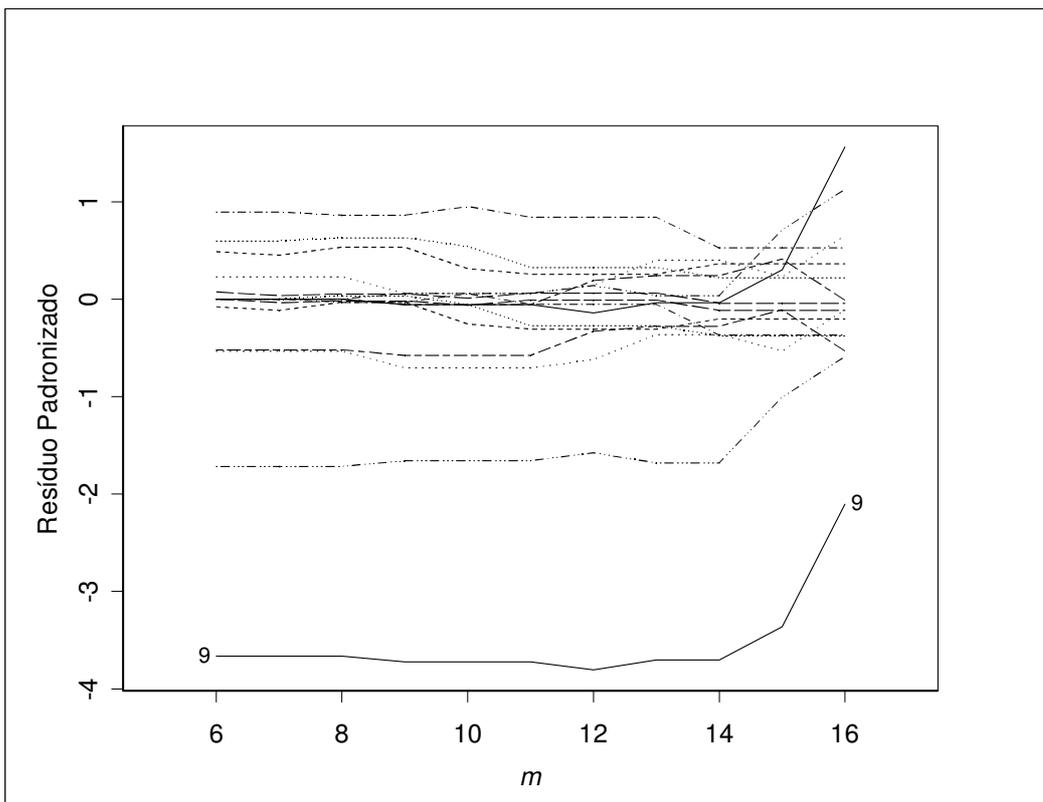
Na Figura 4.12 apresentamos os resultados da FS para a estimativa  $\hat{\sigma}^2$  da variância do termo do erro (esquerda) e do coeficiente de determinação múltipla  $R^2$  (direita).



**Figura 4.12** – *Forward Search* de  $\hat{\sigma}^2$  e de  $R^2$ .

A observação 9 é a última a ser incluída. Quando esta observação é incluída,  $\hat{\sigma}^2$  cresce drasticamente. Este fato confirma que a observação 9 é atípica e influente. A leitura do gráfico de  $R^2$  é similar à leitura do gráfico de  $\hat{\sigma}^2$ .

Na Figura 4.13 apresentamos o gráfico dos resultados da FS para os resíduos padronizados. Durante toda a FS, o resíduo da observação 9 destaca-se de todos os outros. Este fato confirma a indicação anterior de que é uma observação atípica.



**Figura 4.13** – *Forward Search* dos Resíduos Padronizados.

O modelo linear se mostrou inadequado para o experimento do Exemplo 3.1. A principal indicação da inadequação é a de que a variância não é constante. Mais adiante, na Seção 4.1.2, vamos tentar a alternativa de transformar a resposta para este exemplo.

#### **Exemplo 4.2**

Daniel (1976) analisou um experimento fatorial  $2^4$  que estuda a taxa de avanço y de uma perfuratriz de poços. Segundo Daniel, sabe-se que três fatores têm efeitos positivos: A, carga sobre a broca; B, fluxo da lama e C, velocidade

rotacional. O quarto fator, D, é o tipo de lama usado. Os dados estão na Tabela 4.5.

**Tabela 4.6** - Experimento da Perfuratriz.

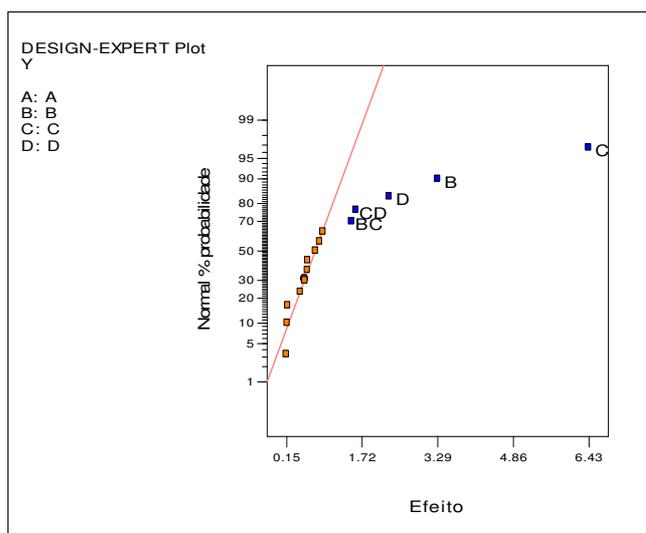
A	B	C	D	Y
-1	-1	-1	-1	1,68
1	-1	-1	-1	1,98
-1	1	-1	-1	3,28
1	1	-1	-1	3,44
-1	-1	1	-1	4,98
1	-1	1	-1	5,70
-1	1	1	-1	9,97
1	1	1	-1	9,07
-1	-1	-1	1	2,07
1	-1	-1	1	2,44
-1	1	-1	1	4,09
1	1	-1	1	4,53
-1	-1	1	1	7,77
1	-1	1	1	9,43
-1	1	1	1	11,75
1	1	1	1	16,30

Daniel (1976) usou as transformações  $\ln(y)$  e  $y^{-1/2}$ , concluindo que os efeitos dos fatores principais A, B, C e D são importantes, ficando em dúvida sobre a inclusão das interações BC e CD. Suas conclusões não foram definitivas, sugerindo a obtenção de mais dados.

Box *et al.* (1978) e Montgomery (2001), usaram a transformação  $\ln(y)$ , concluindo que apenas os efeitos dos fatores principais B, C e D são importantes. Myers *et al.* (2002) e Lewis *et al.* (2001a) usaram modelos lineares generalizados com distribuição gama e função de ligação logarítmica, concluindo igualmente que apenas os efeitos dos fatores principais B, C e D são importantes.

Com base no gráfico de probabilidade normal, construído com o software *Design Expert*, mostrado na Figura 4.14, podemos considerar B, C, D, BC e CD como efeitos importantes.

Na Tabela 4.6, fornecida pela planilha *Excel*, apresentamos os testes para significância dos coeficientes para o modelo. Confirmamos a indicação dos fatores significativos B, C, D, BC e CD, pois as estimativas de seus parâmetros apresentam P-valor inferior a 0,05.



**Figura 4.14** – Gráfico de Probabilidade Normal dos Efeitos

**Tabela 4.7** - Teste  $t$  para Experimento da Perfuratriz.

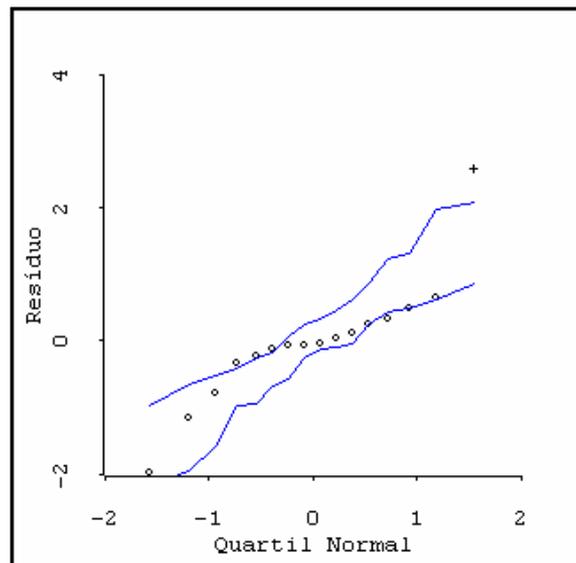
	<i>Coefficientes</i>	<i>Erro-padrão</i>	$t_0$	<i>P-valor</i>
Interseção	6,1550	0,29079	21,17	0,00000
B	1,6488	0,29079	5,67	0,00021
C	3,2163	0,29079	11,06	0,00000
D	1,1425	0,29079	3,93	0,00282
BC	0,7525	0,29079	2,59	0,02705
CD	0,7988	0,29079	2,75	0,02059

### Testes de Adequação do Modelo

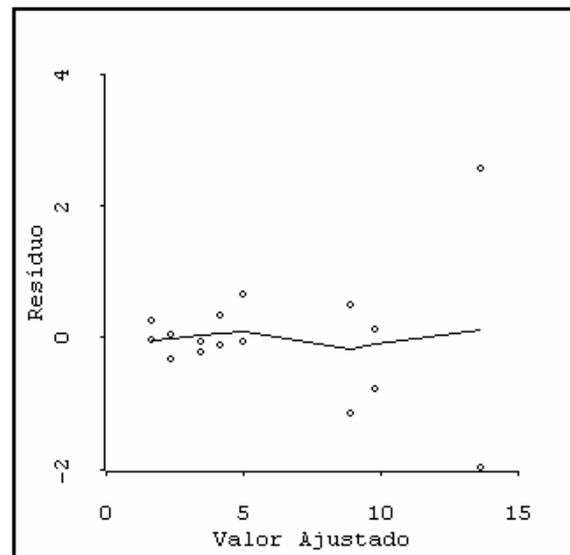
Para os testes de adequação, utilizamos os gráficos apresentados nas Figuras 4.15 a 4.19, fornecidos pelo software ARC.

O gráfico de probabilidade normal dos resíduos é apresentado na Figura 4.15. Este gráfico não apresenta um perfil condizente com a distribuição normal. Isto porque há cinco pontos fora do envelope, um dos quais (assinalado com um +) nitidamente não está alinhado com os demais, sugerindo uma observação atípica, o que será confirmado no gráfico da Figura 4.19.

O gráfico da Figura 4.16 apresenta linha resultante do amortecimento (*lowess*) aproximadamente horizontal e próxima da reta horizontal de ordenada zero, indicando que a suposição de erros com média zero é satisfatória.



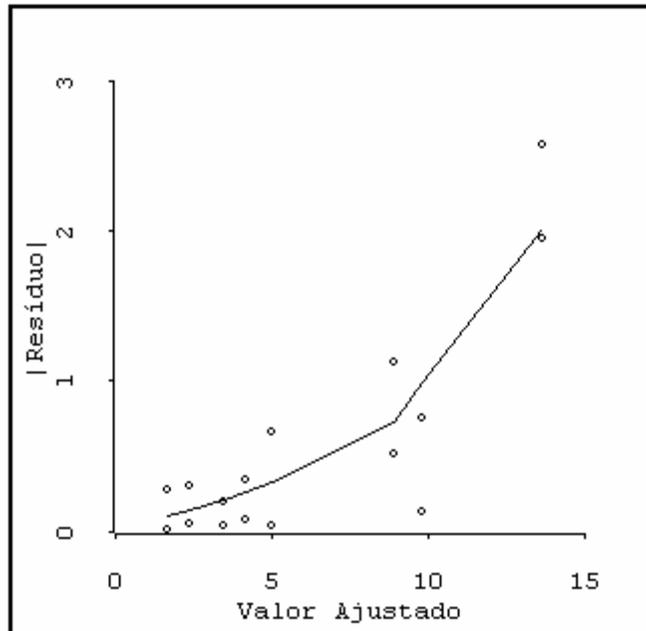
**Figura 4.15** – Gráfico de Probabilidade Normal dos Resíduos com Envelope



**Figura 4.16** – Gráfico dos Resíduos *versus* Valores Ajustados

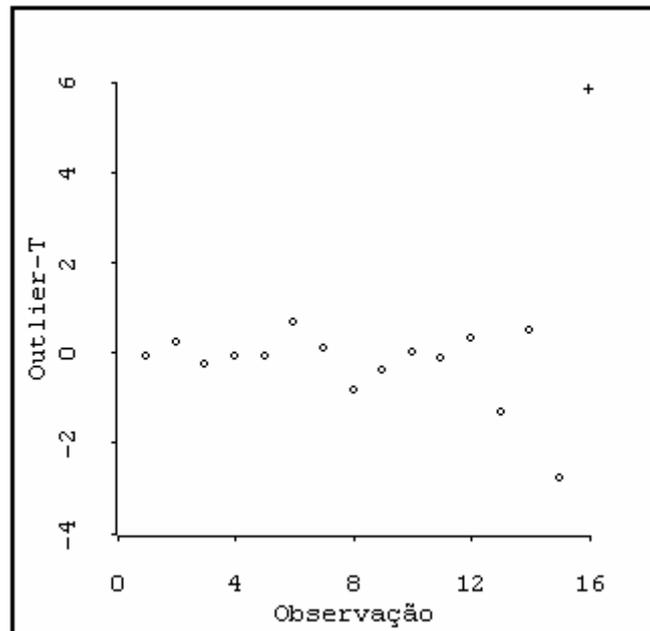
Para verificar se a variância pode ser considerada constante, temos, na Figura 4.17, o gráfico do valor absoluto dos resíduos *versus* o valor ajustado (ver Capítulo 2). A linha resultante do amortecimento (*lowess*) indica crescimento da variância com o aumento da média.

No Capítulo 2 vimos um teste formal para a variância é o da estatística *ET*. Neste exemplo  $ET = 20,09$  com um P-valor de 0,000007. Então, o teste confirma que a variância não deve ser constante.



**Figura 4.17** – Gráfico Valor Absoluto do Resíduo *versus* Valor Ajustado

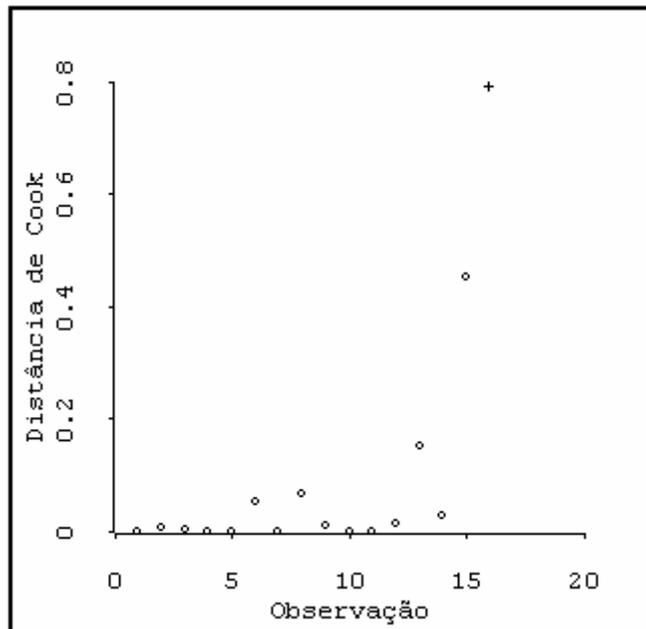
No gráfico da Figura 4.18, o *outlier-t* da resposta 16 (assinalado com um +) destaca-se dos demais. Seu valor é 5,84, o que corresponde, na distribuição *t* com 5 graus de liberdade, a um P-valor de 0,001044, indicando a resposta 16 como atípica.



**Figura 4.18** – Gráfico dos Resíduos *Outlier-t*

Cabe lembrar que a ocorrência de uma observação atípica pode ser causada por inadequação do modelo. O gráfico da Figura 4.17 e o teste de Cook e Weisberg indicaram que a variância não é constante. Por conseguinte, a resposta 16 pode não ser atípica.

No gráfico da Figura 4.19 a distância de Cook da resposta 16 é 0,7910, portanto, superior a 0,5. Neste caso temos indicação de que a resposta 16, além de atípica, é influente.



**Figura 4.19** – Gráfico da Distância de Cook

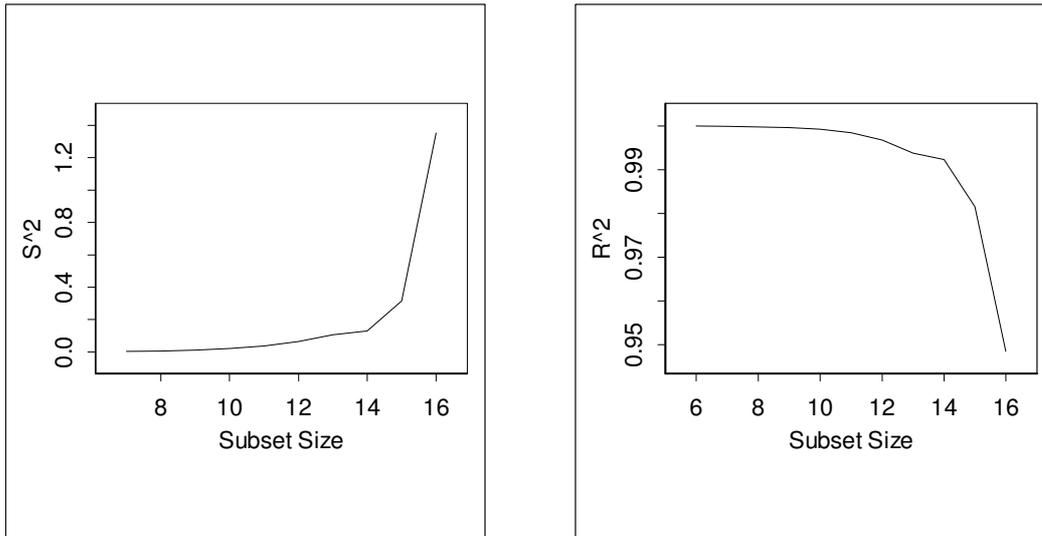
### ***Forward Search***

Os gráficos das figuras a seguir foram gerados com o módulo de regressão linear do programa *Forward Search* (FS), instalado no *S-Plus*. Este procedimento foi descrito no Capítulo 2.

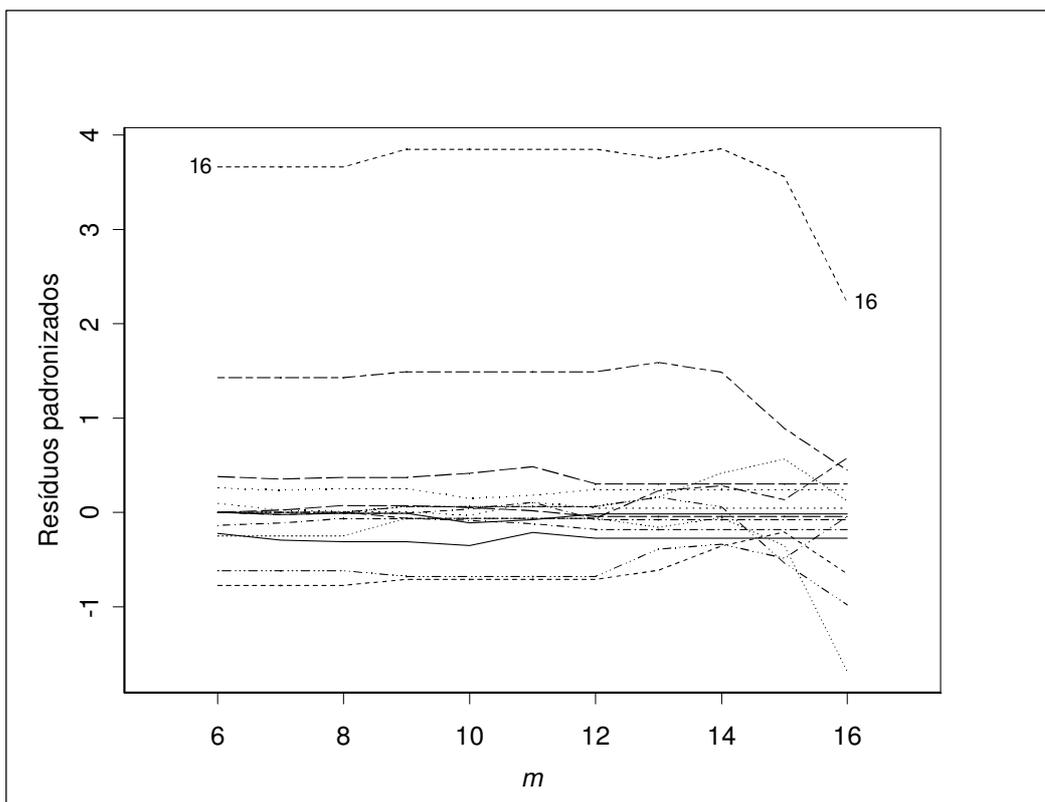
Na Figura 4.20 apresentamos os resultados da FS para a estimativa  $\hat{\sigma}^2$  da variância do termo do erro (esquerda) e do coeficiente de determinação múltipla  $R^2$  (direita). A observação 16 é a última a ser incluída. Quando esta observação é incluída,  $\hat{\sigma}^2$  cresce drasticamente. Este fato confirma que a observação 16 é atípica e influente.

Cabe registrar que a observação 16 foi incluída por último na FS porque é uma observação atípica, e não porque seja a última observação no experimento. A leitura do gráfico de  $R^2$  é similar à leitura do gráfico de  $\hat{\sigma}^2$ .

Na Figura 4.21 apresentamos o gráficos dos resultados da FS para os resíduos padronizados. Durante toda a FS, o resíduo da observação 16 destaca-se de todos os outros. Este fato confirma que ela é atípica.



**Figura 4.20** – Dados do Exemplo 4.2: *Forward Search* de  $\hat{\sigma}^2$  e de  $R^2$ .



**Figura 4.21** – *Forward Search* dos Resíduos Padronizados

O modelo linear mostrou-se inadequado para o experimento do Exemplo 4.2. A principal indicação é de que a variância não é constante. Para este exemplo, na Seção 4.3.1, vamos tentar a alternativa de transformar a resposta.

### 4.3. Estágio 2: Transformação da Resposta

Cook e Weisberg (1999, Cap. 13) e Atkinson e Riani (2000, Cap. 4) são excelentes referências para a análise da transformação da resposta.

Esse estágio torna-se necessário caso o modelo utilizado no Estágio 1 não tenha se mostrado adequado — especificamente, se um ou mais dos seguintes pressupostos não tiverem sido considerados satisfatórios: normalidade, variância constante e aditividade dos efeitos (linearidade). Nessas situações, há dois caminhos a seguir: transformação da resposta ou MLG.

A transformação da resposta muitas vezes pode ser satisfatória. Entretanto, com uma única transformação, nem sempre se consegue atingir a normalidade, a equalização da variância e a aditividade dos efeitos. Vejamos o Exemplo 4.2. Com o modelo linear e o método dos MQ a normalidade não é aceitável (Figura 4.15) e há indicação de que a variância cresce com a média estimada (Figura 4.17). Ademais, temos evidências de uma observação atípica (Figura 4.18) e de uma observação influente (Figura 4.19). Veremos que, com a transformação logarítmica, conseguimos normalidade satisfatória (Figura 4.23) e uma aparente aditividade nos efeitos (Figura 4.24), embora não tenhamos indicação de estabilização da variância (Figura 4.25) e ainda haja suspeita de uma observação influente (Figura 4.27). Para estes casos devemos considerar os MLG.

Lewis *et al.* (2001a) e (2001b), analisando vários exemplos, concluíram que os MLG devem ser preferidos à transformação da resposta, uma vez que em todos os casos os MLG apresentaram melhor desempenho na estimativa da resposta, pois propiciaram intervalos de confiança menores. Por conseguinte, mesmo que com uma única transformação venhamos a alcançar a normalidade, a equalização da variância e a aditividade dos efeitos, recomendamos utilizar os MLG quando for desejável uma estimativa da resposta mais confiável.

Por outro lado, a modelagem através da transformação da resposta pode oferecer indicações para a escolha das funções de ligação e de variância. Ainda no Exemplo 4.2, todos os métodos utilizados para a escolha da transformação da resposta, descritos na Seção 4.1.2, indicaram a transformação logarítmica, o que sugere o uso de uma função de ligação logarítmica para os MLG. Ademais, há indicação de que a variância cresce com o aumento da média (Figura 4.25), o que sugere as distribuições gama e inversa normal para os MLG. Por conseguinte,

mesmo que sejam utilizados os MLG, caso não haja informações sobre a forma dessas funções, a transformação da resposta é um estágio recomendado.

Quando o modelo gerado pela transformação da resposta é considerado adequado, em todos os casos por nós observados este modelo é praticamente o mesmo que o modelo gerado pelos MLG. Portanto, nestes casos, podemos ficar no estágio da transformação, desde que seja desejado apenas avaliar a importância dos fatores e haja dificuldade de utilizar os MLG.

Em síntese, sempre devemos preferir os MLG. Caso haja alguma dificuldade em utilizá-los, só recomendamos ficar no estágio da transformação da resposta se o modelo for considerado adequado e não for importante uma estimativa da resposta mais confiável.

Propomos então três etapas para o estágio de transformação da resposta.

**I) Definição da Transformação da Resposta**

**II) Estimativa e Teste de Significância dos Parâmetros do Modelo**

**III) Testes de Adequação do Modelo**

#### **4.3.1. Definição da Transformação da Resposta.**

Serão descritos quatro métodos voltados para definição da transformação: o de Box e Cox (1964), o de Cook e Weisberg (1999), o de Cook e Olive (2001) e o de Atkinson e Riani (2000). O primeiro é um método numérico que fornece um intervalo de confiança para  $\lambda$ , enquanto os outros três são métodos gráficos. O método de Atkinson e Riani usa um gráfico de uma estatística de teste.

Os métodos de Box e Cox e de Atkinson e Riani buscam obter normalidade na resposta transformada, enquanto os métodos de Cook e Weisberg e de Cook e Olive buscam a linearidade. Entretanto, os quatro métodos geralmente conduzem a uma só transformação. Exceção se faz quando há observações atípicas, quando então o método de Box e Cox pode apresentar resultados diferentes, uma vez que é sensível a este tipo de observações (Cook e Olive, 2001, e Atkinson e Riani, 2000).

Todos os quatros métodos são aplicados a modelos sem interações ou termos quadráticos. Embora não esteja explícito em nenhum dos textos, faz sentido que esses autores apliquem a transformação da resposta para simplificar

modelos que apresentam interações e termos quadráticos, para obter modelos sem esses termos.

Infelizmente, como veremos mais adiante, temos evidências de que *todos* os quatro métodos só fornecem a transformação adequada na ausência de interações importantes.

### 1. Método de Box e Cox

Box e Cox (1964) desenvolveram um procedimento numérico para escolher uma transformação da resposta tal que a distribuição da variável transformada esteja o mais próximo possível da distribuição normal. Para isso, é definida a seguinte família de transformações em potência:

$$y(\lambda) = \frac{y^\lambda - 1}{\lambda} \quad \text{para } \lambda \neq 0$$

$$y(\lambda) = \ln(y) \quad \text{para } \lambda = 0$$

Supõe-se que existe um valor de  $\lambda$  tal que o modelo de regressão linear na escala transformada,

$$\mathbf{y}(\lambda) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

tenha erros normalmente distribuídos com média zero e variância constante.

Uma possibilidade intuitiva é escolher o valor de  $\lambda$  que minimiza a soma dos quadrados dos resíduos,  $SQR(\lambda)$ , resultante da regressão com MQ. Entretanto, as unidades das  $SQR(\lambda)$  são diferentes para cada valor de  $\lambda$ . Por conseguinte, não se podem comparar as  $SQR(\lambda)$  para os diversos valores de  $\lambda$ . Para contornar este problema define-se a transformação de potência normalizada:

$$z(\lambda) = \frac{y^\lambda - 1}{\lambda \hat{y}^{\lambda-1}} \quad \text{para } \lambda \neq 0$$

$$z(\lambda) = \hat{y} \ln(y) \quad \text{para } \lambda = 0$$

onde  $\hat{y} = \exp(\sum \ln y_i / n)$  é a média geométrica das respostas.

Seja  $SQR_z(\lambda)$  a soma dos quadrados dos resíduos resultante da regressão com MQ com o modelo

$$\mathbf{z}(\lambda) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Atkinson (1985) demonstra que minimizar  $SQR_z(\lambda)$  conduz à maximização da função de log-verossimilhança da transformação de potência normalizada  $z(\lambda)$ .

Pode-se demonstrar, também, que as unidades das  $SQR_z(\lambda)$  são as mesmas para qualquer valor de  $\lambda$ . Portanto, pode-se comparar as  $SQR_z(\lambda)$  para os diversos valores de  $\lambda$ , escolhendo o valor de  $\lambda$  que minimiza  $SQR_z(\lambda)$ .

Resumindo:

- Escolhem-se valores para  $\lambda$ . Geralmente toma-se  $\lambda = -1, -0,5, 0, 0,5$  e  $1$ .
- Com os valores escolhidos de  $\lambda$ , usa-se a transformação de potência normalizada  $z(\lambda)$ .
- Para esses valores de  $\lambda$ , faz-se a regressão com MQ com o modelo  $\mathbf{z}(\lambda) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ .
- O valor escolhido de  $\lambda$  é aquele que minimiza a soma dos quadrados dos resíduos  $SQR_z(\lambda)$ .

Um intervalo para  $\lambda$  com  $100(1-\alpha)\%$  de confiança é calculado a partir dos valores para os quais

$$n \left[ \ln(SQR_z(\lambda)) - \ln(SQR_z(\hat{\lambda})) \right] \leq \chi_{1,\alpha}^2$$

onde  $\hat{\lambda}$  é o valor que minimiza  $SQR_z(\lambda)$ .

#### Exemplo 4.2 (cont.)

Para o exemplo da taxa de avanço da perfuratriz, o software *Design Expert* fornece o gráfico do  $\ln(SQR_z)$  versus  $\lambda$  apresentado na Figura 4.22.

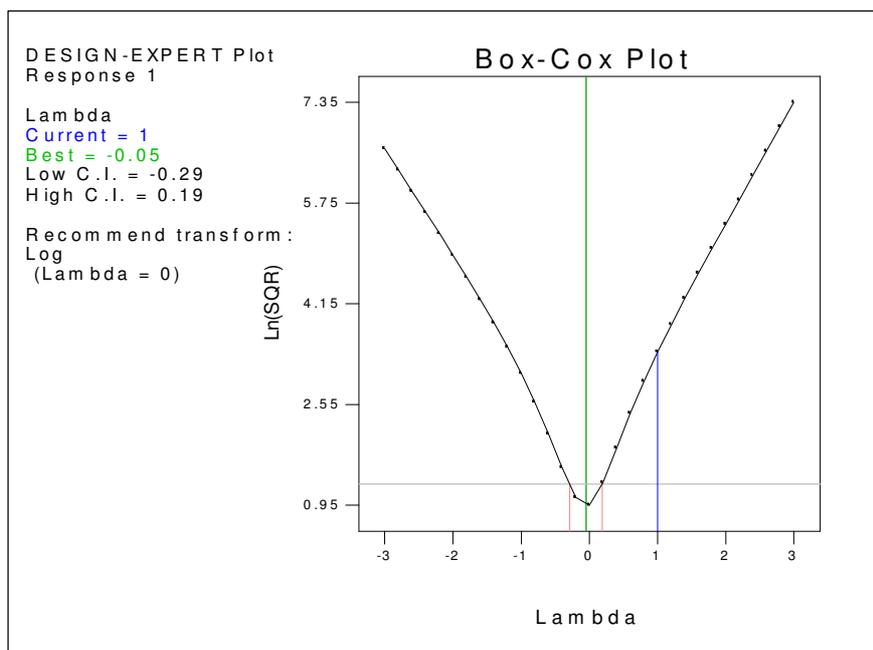


Figura 4.22 – Gráfico da Transformação de Box–Cox.

No gráfico também são fornecidos o valor de  $\hat{\lambda} = -0.05$  e o intervalo  $[-0,29, 0,19]$  para 95% de confiança. Portanto, a transformação recomendada por este método é  $\ln(y)$ , correspondendo a  $\lambda = 0$ .

## 2. Método de Cook e Weisberg

Cook e Weisberg (1999) propuseram um método gráfico muito simples. Eles supõem que para uma transformação  $T(y)$  há um modelo linear de tal modo que

$$T(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

É definida a transformação

$$\begin{aligned} T(y) &= y^\lambda && \text{para } \lambda \neq 0 \\ T(y) &= \ln(y) && \text{para } \lambda = 0 \end{aligned}$$

Para determinar o valor de  $\lambda$ , inicialmente é feita a regressão sem transformação ( $\lambda = 1$ ) resultando em  $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ , os valores ajustados.

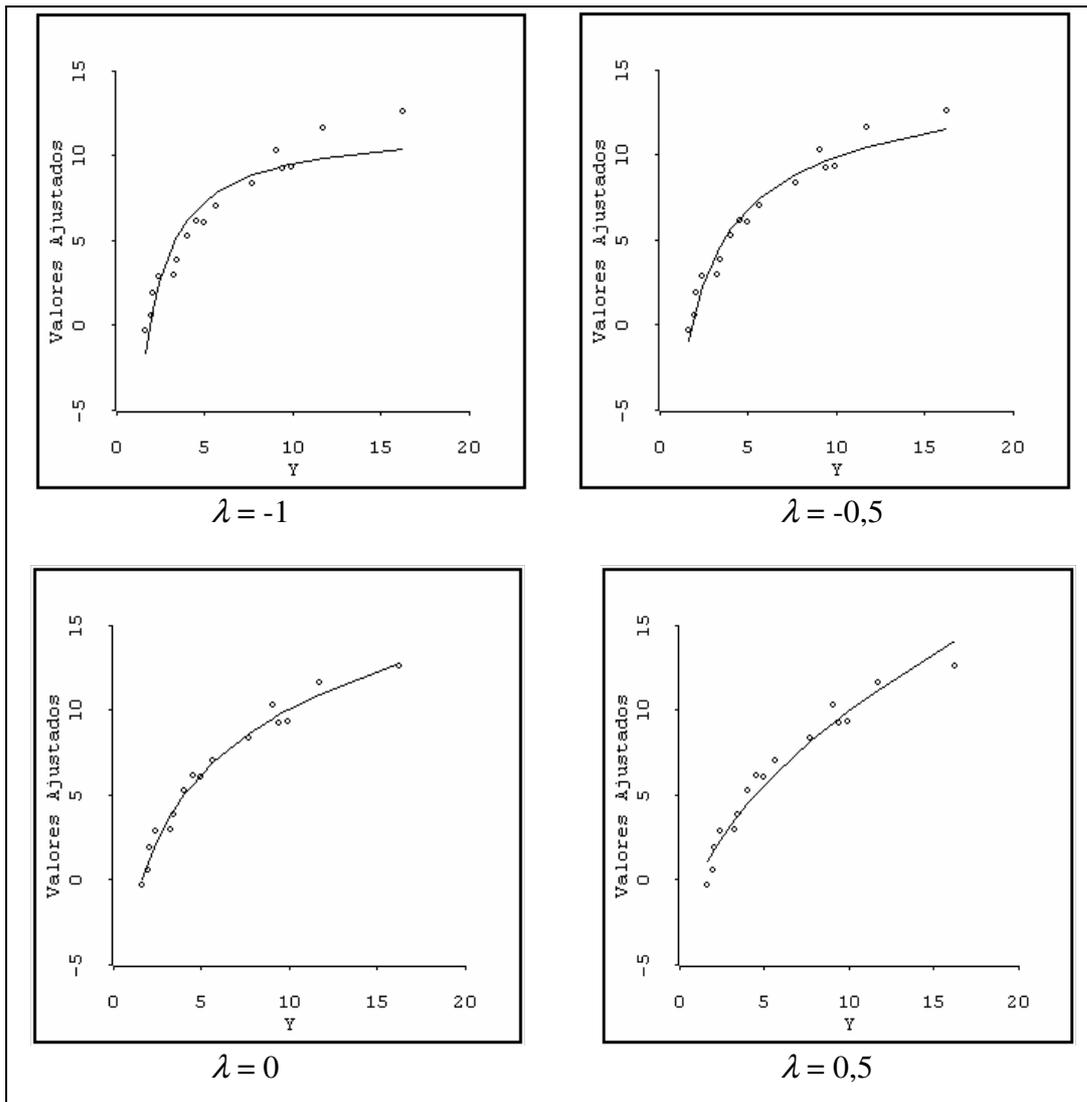
Em seguida é construído um gráfico com  $\mathbf{X}\hat{\boldsymbol{\beta}}$  no eixo vertical e  $\mathbf{y}$  no eixo horizontal, o qual eles denominam de **gráfico inverso**.

A esse gráfico acrescenta-se uma curva ajustada a partir da regressão de  $\mathbf{X}\hat{\boldsymbol{\beta}}$  sobre a transformação proposta  $T(\mathbf{y})$ . O valor de  $\lambda$  cuja curva fornecer o melhor ajuste deve ser o escolhido.

### Exemplo 4.2 (cont.)

O software ARC permite facilmente construir e comparar gráficos para diversos valores de  $\lambda$ , os quais são apresentados na Figura 4.23.

Observando os gráficos fica evidente que  $\lambda = 0$  propicia o melhor ajuste dentre os quatro valores experimentados.



**Figura 4.23** – Gráfico Inverso para Valores de  $\lambda$

### 3. Método de Cook e Olive

Cook e Olive (2001) consideram a família de transformações de potência:

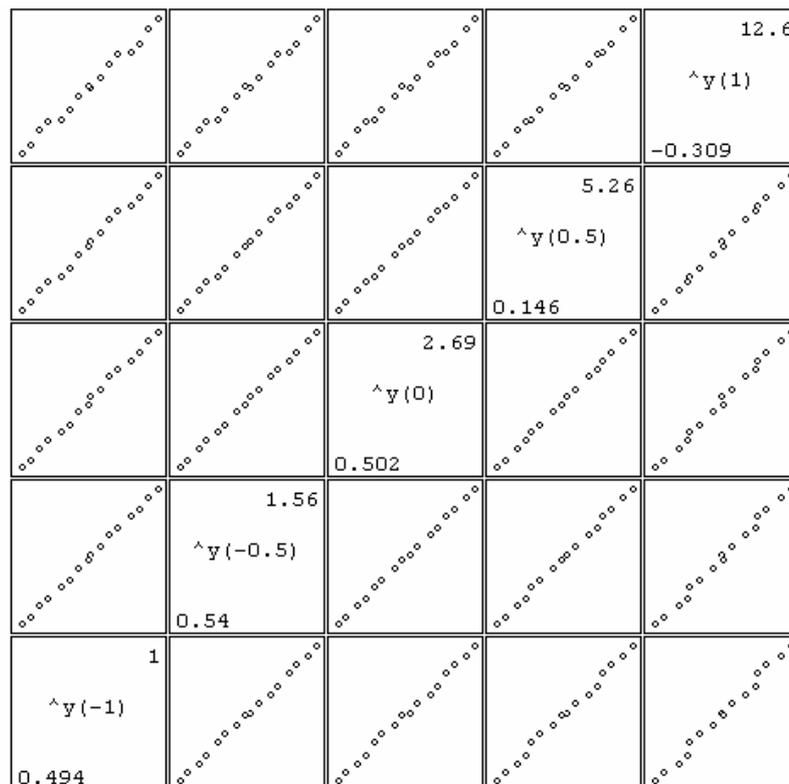
$$y(\lambda) = \frac{y^\lambda - 1}{\lambda} \quad \text{para } \lambda \neq 0$$

$$y(\lambda) = \ln y \quad \text{para } \lambda = 0$$

Os autores sugerem inicialmente considerar  $\lambda$  no intervalo  $\lambda \in [-1, 1]$ . Portanto, consideram-se as transformações  $y(-1)$ ,  $y(-0,5)$ ,  $y(0)$ ,  $y(0,5)$ ,  $y(1)$ . Para cada transformação e supondo o uso do método dos mínimos quadrados, procede-se ao ajuste dos modelos, calculando os valores ajustados:  $\hat{y}(-1)$ ,  $\hat{y}(-0,5)$ ,  $\hat{y}(0)$ ,  $\hat{y}(0,5)$ ,  $\hat{y}(1)$ . Em seguida faz-se uma série de gráficos de valor ajustado  $\hat{y}(\lambda_i)$  versus valor ajustado  $\hat{y}(\lambda_j)$ .

**Exemplo 4.2 (cont.)**

Na Figura 4.24 são apresentados esses gráficos para o Exemplo 4.2.

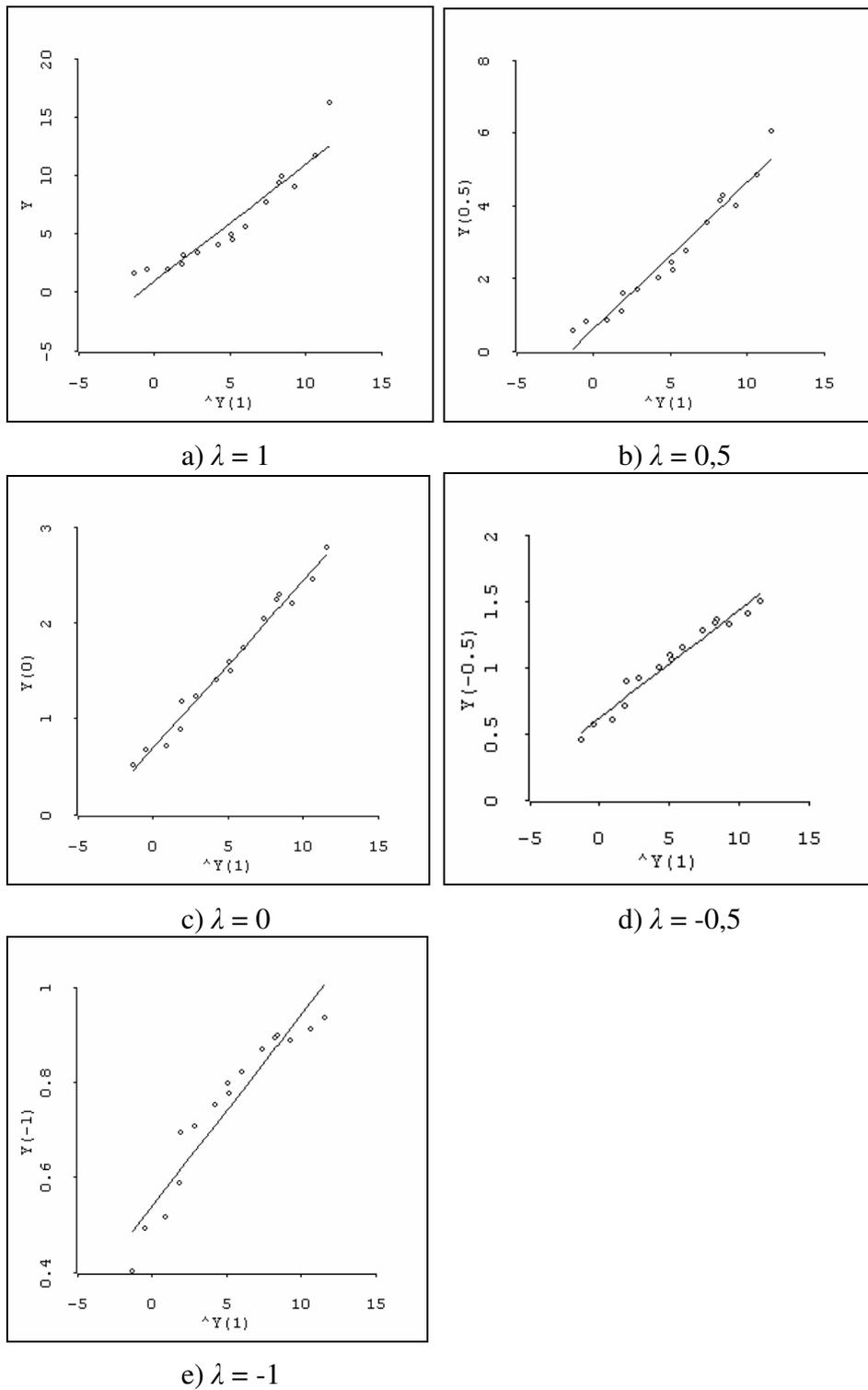


**Figura 4.24** - Gráficos de Valor Ajustado versus Valor Ajustado.

Os dois números que aparecem em cada quadrado do gráfico são os valores ajustados máximo e mínimo para cada transformação considerada.

Observa-se que há uma forte correlação dos diversos valores ajustados. O menor coeficiente de correlação é 0,9956. Cook e Olive asseguram que, caso todos os valores do coeficiente de correlação sejam superiores a 0,84, há indicação que a transformação adequada está no intervalo  $\lambda \in [-1, 1]$ .

Em seguida escolhe-se qualquer um dos conjuntos de valores ajustados, por exemplo  $\hat{y}(1)$ . Para cada um dos valores considerados para  $\lambda$  fazem-se gráficos dos diversos valores transformados  $y(\lambda)$  versus  $\hat{y}(1)$ . Na Figura 4.25 são apresentados esses gráficos.



**Figura 4.25** - Gráficos dos Valores Transformados  $y(\lambda)$  versus  $\hat{y}(1)$ .

Visualmente, o gráfico que fornece o melhor ajuste é o de  $\hat{y}(0)$ . O método de Box e Cox (1964) fornece um intervalo de confiança  $[-0,18 \ 0,28]$ . Portanto, a transformação indicada deve ser mesmo a logarítmica ( $\lambda = 0$ ).

#### 4. Método de Atkinson e Riani

Atkinson e Riani (2000), pág 86, também empregam a transformação de potência normalizada utilizada por Box e Cox:

$$z(\lambda) = \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} \quad \text{para } \lambda \neq 0$$

$$z(\lambda) = \dot{y} \ln(y) \quad \text{para } \lambda = 0$$

onde  $\dot{y} = \exp(\sum \ln y_i / n)$ .

Entretanto, para selecionar  $\lambda$ , eles propõem alternativamente um teste de uma estatística-escore proveniente da expansão de  $z(\lambda)$  em série de Taylor:

$$z(\lambda) \doteq z(\lambda_0) + (\lambda - \lambda_0)w(\lambda_0)$$

onde

$$w(\lambda_0) = \left. \frac{\partial z(\lambda)}{\partial \lambda} \right|_{\lambda=\lambda_0}$$

Combinando com o modelo de regressão  $y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$ , obtém-se

$$z(\lambda_0) = \mathbf{x}'\boldsymbol{\beta} - (\lambda - \lambda_0)w(\lambda_0) + \varepsilon$$

$$z(\lambda_0) = \mathbf{x}'\boldsymbol{\beta} + \gamma w(\lambda_0) + \varepsilon$$

onde:  $\gamma = -(\lambda - \lambda_0)$ .

A expressão acima é um modelo de regressão com uma variável a mais,  $w(\lambda_0)$ , proveniente da transformação. Esta nova variável é denominada de “variável construída”. A estimativa de  $\gamma$  é

$$\hat{\gamma}(\lambda) = \frac{\mathbf{w}'(\lambda)(\mathbf{I} - \mathbf{H})\mathbf{z}(\lambda)}{\mathbf{w}'(\lambda)(\mathbf{I} - \mathbf{H})\mathbf{w}(\lambda)}$$

e a variância da estimativa de  $\gamma$  é

$$\text{Var}(\hat{\gamma}) = \frac{s_w^2(\lambda)}{\mathbf{w}'(\lambda)(\mathbf{I} - \mathbf{H})\mathbf{w}(\lambda)}$$

onde  $s_w^2(\lambda)$  é a soma dos quadrados dos resíduos que estima  $\sigma^2$ , calculada pela expressão:

$$(n - p - 1)s_w^2(\lambda) = \mathbf{z}'(\lambda)(\mathbf{I} - \mathbf{H})\mathbf{z}(\lambda) - \frac{\mathbf{z}'(\lambda)(\mathbf{I} - \mathbf{H})\mathbf{w}(\lambda)^2}{\mathbf{w}'(\lambda)(\mathbf{I} - \mathbf{H})\mathbf{w}(\lambda)}$$

Então a estatística de teste é

$$T_p(\lambda) = -\frac{\hat{\gamma}(\lambda)}{\sqrt{s_w^2 / [\mathbf{w}'(\lambda)(\mathbf{I} - \mathbf{H})\mathbf{w}(\lambda)]}}$$

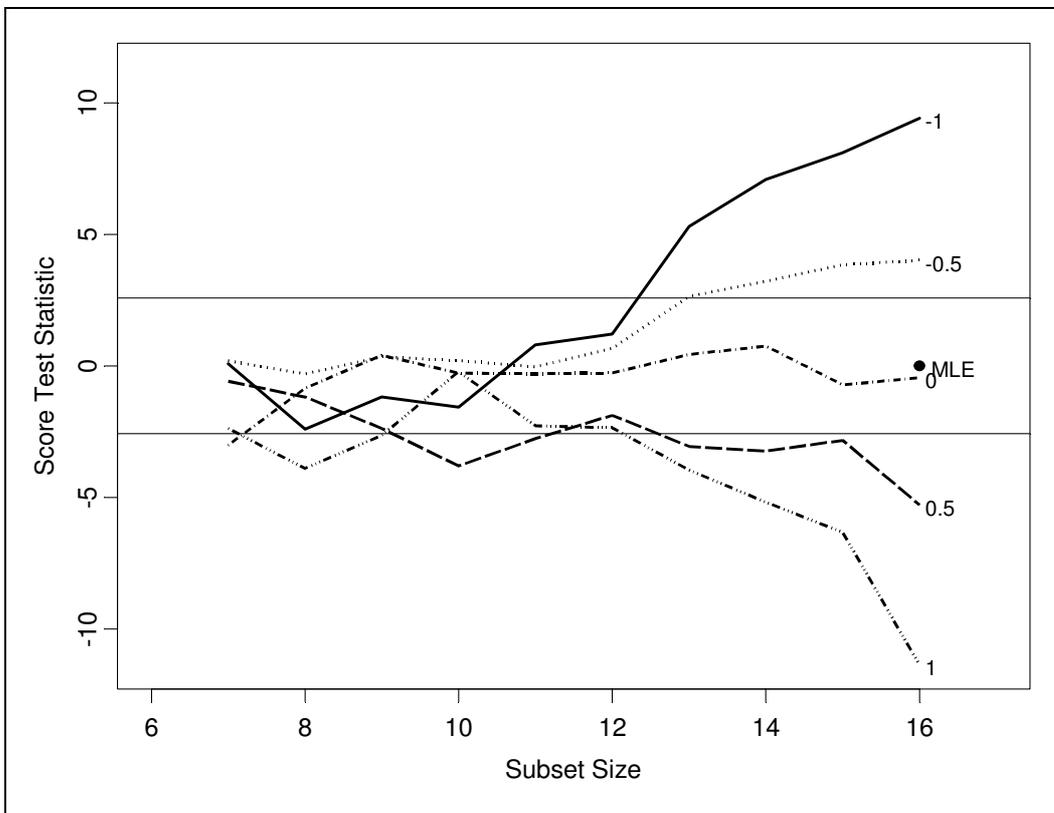
O teste para  $\gamma = 0$  é então o teste para  $\lambda = \lambda_0$ . Para calcular  $T_p(\lambda)$  é assumido um valor para  $\lambda$  e feita a regressão com  $\mathbf{X}$  e  $\mathbf{w}(\lambda)$ . Esta estatística pode ser monitorada com a *Forward Search*, descrita no Capítulo 2.

**Exemplo 4.2 (cont.)**

Na Figura 4.26 é apresentada a FS com os gráficos da estatística de teste  $T_p(\lambda)$ , fornecidos pelo módulo de transformação da resposta do programa *Forward Search* (FS), instalado no *S-Plus*. Foram considerados cinco valores de  $\lambda$ , perfazendo cinco modelos de regressão linear.

Os autores do método afirmam que, quando se inclui uma observação atípica no decorrer do procedimento, a linha do gráfico dá um salto, sendo então a observação identificada como atípica para o modelo correspondente à linha. Se há saltos em todas as linhas quando da inclusão de uma mesma observação, esta é verdadeiramente atípica.

No gráfico da Figura 4.26 há saltos para  $\lambda = -1, 0,5$  e  $1$ , mas não para  $0$  e  $0,5$ , o que indica que não há observações atípicas, e sim modelos inadequados para  $\lambda = -1, 0,5$  e  $1$ .



**Figura 4.26** – *Forward Search* - Gráfico da Estatística de Teste.

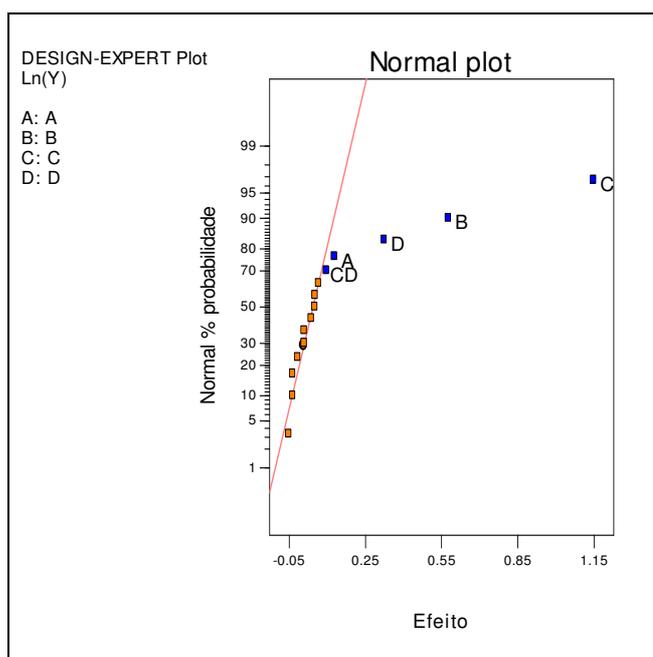
As linhas horizontais no gráfico correspondem a  $\pm 2,58$ , que são os limites de 99% de confiança para a distribuição normal padronizada. Observa-se que a transformação para  $\lambda = 0$  é a única que se mantém dentro dos limites durante todo o procedimento. O estimador de máxima verossimilhança é  $\hat{\lambda} = -0,05$ .

#### 4.3.2. Estimação e Teste de Significância dos Coeficientes

Uma vez escolhida a transformação, faz-se o ajuste do modelo da resposta transformada, com o método dos MQ.

#### Exemplo 4.2 (cont.)

Considerando a transformação logarítmica, indicada por todos os métodos, temos, na Figura 4.27, o gráfico de probabilidade normal dos efeitos, fornecido pelo *Design Expert*. Temos indicação que os fatores B, C e D são significativos, havendo dúvida sobre os fatores A e CD.



**Figura 4.27** – Gráfico de Probabilidade Normal dos Efeitos.

Na Tabela 4.7, fornecida pelo *Excel*, apresentamos os testes para significância dos coeficientes para o modelo. Confirmamos a indicação dos fatores significativos A, B, C, D e BC, pois as estimativas de seus parâmetros apresentam P-valor inferiores a 0,05.

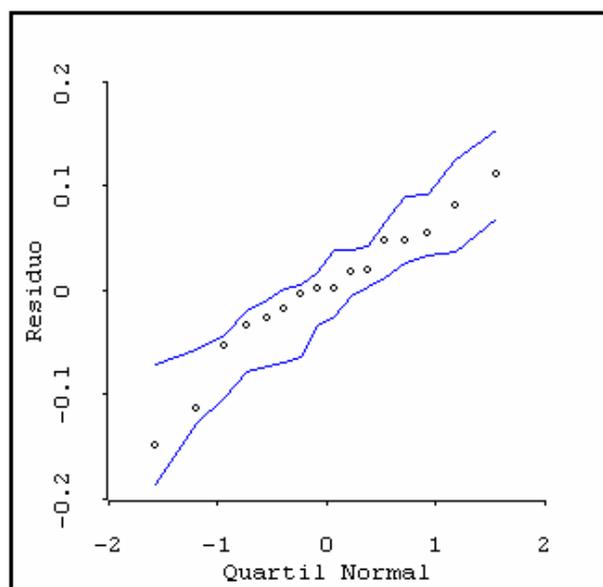
**Tabela 4.8** - Teste *t* para Experimento da Perfuratriz.

	<i>Coefficientes</i>	<i>Erro-padrão</i>	<i>t<sub>0</sub></i>	<i>P-valor</i>
Interseção	1,5977	0,02053	77,83	2,99E-15
A	0,0650	0,02053	3,17	0,010054
B	0,2900	0,02053	14,13	6,21E-08
C	0,5772	0,02053	28,12	7,52E-11
D	0,1633	0,02053	7,95	1,24E-05
CD	0,0491	0,02053	2,39	0,037936

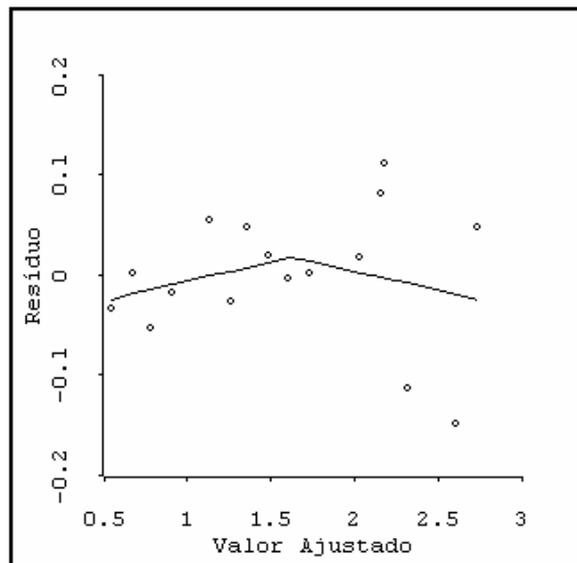
### 4.3.3. Testes de Adequação do Modelo

Para testar a adequação do modelo são utilizados os gráficos dos resíduos, descritos no Capítulo 2. Estes gráficos são apresentados nas Figuras 4.28 a 4.31. Para construção dos gráficos foi utilizado o software *ARC*.

O gráfico de probabilidade normal dos resíduos com envelope na Figura 4.28 não apresenta pontos fora do envelope e não observamos pontos muito fora do alinhamento. Por conseguinte, não há indicação de que deva ser rejeitada a consideração de normalidade.

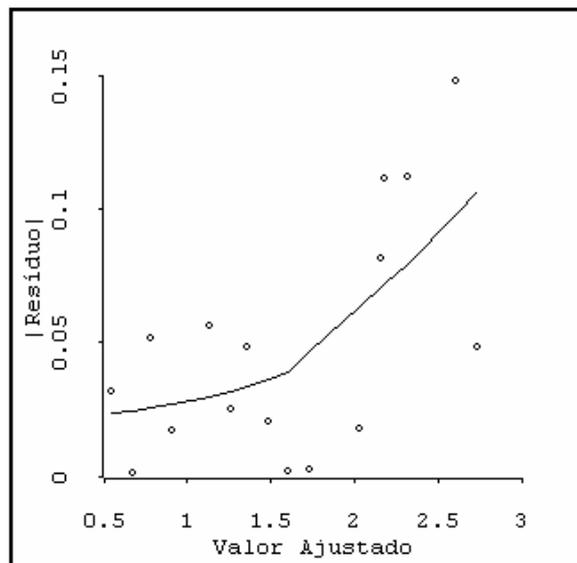
**Figura 4.28** – Gráfico de Probabilidade Normal com Envelope.

O gráfico da Figura 4.29 apresenta linha resultante do amortecimento (*lowess*) aproximadamente horizontal e próxima da reta horizontal de ordenada zero, indicando que a suposição de erros com média zero é satisfatória.



**Figura 4.29** – Gráfico dos Resíduos *versus* Valores Ajustados.

Para verificar se a variância deve ser considerada constante temos da Figura 4.30 o gráfico do valor absoluto dos resíduos *versus* o valor ajustado. A linha resultante do amortecimento (*lowess*) indica crescimento da variância com o aumento da média.

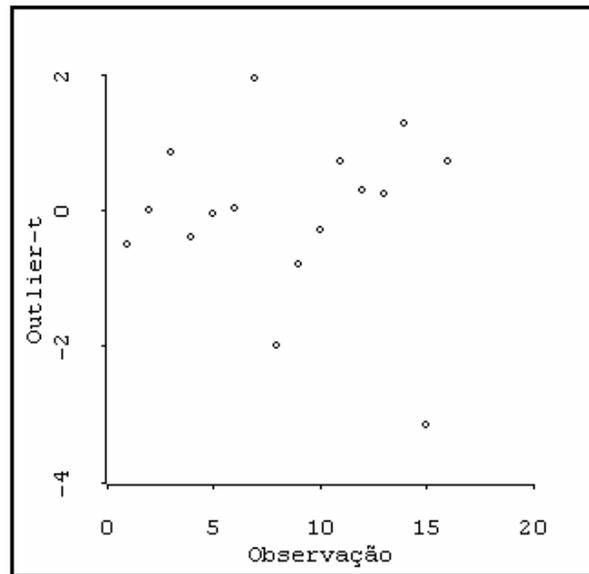


**Figura 4.30** - Gráfico dos Valores Absolutos dos Resíduos.

A indicação do gráfico da Figura 4.30 não é um teste formal. Portanto, efetuou-se o teste de Cook e Weisberg (1999), descrito no Capítulo 2. A estatística de teste  $ET = 6,41$  corresponde a um P-valor de 0,013. Então, o teste indica crescimento da variância com o aumento da média.

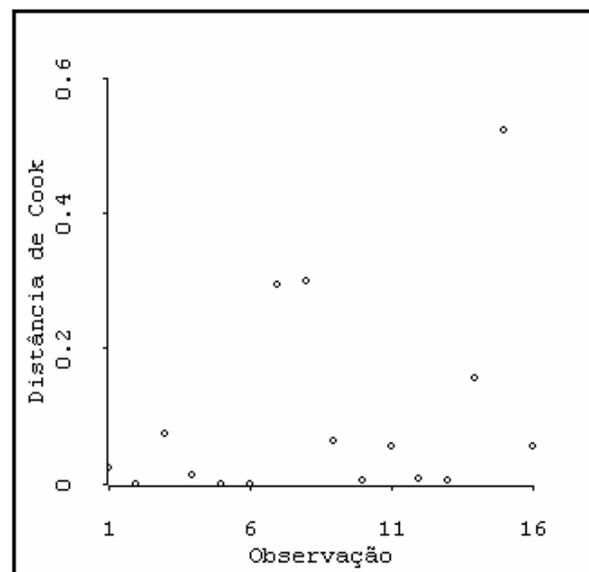
No gráfico da Figura 4.31 temos os valores dos resíduos *outlier-t*. Cabe lembrar que este tipo de resíduo segue a distribuição  $t$  com  $n-p-1$  graus de

liberdade, o que permite fazer um teste para verificar se existem observações atípicas. Entretanto, este teste só é necessário quando há resíduo com valor absoluto superior a 3,5. Portanto, não há indicação de observações atípicas.



**Figura 4.31** – Gráfico dos Resíduos *Outlier-t*.

No gráfico da Figura 4.32 o valor da distância de Cook da resposta 15 é 0,5232, portanto é maior do que 0,5. Logo, temos indicação de que é uma observação influente.



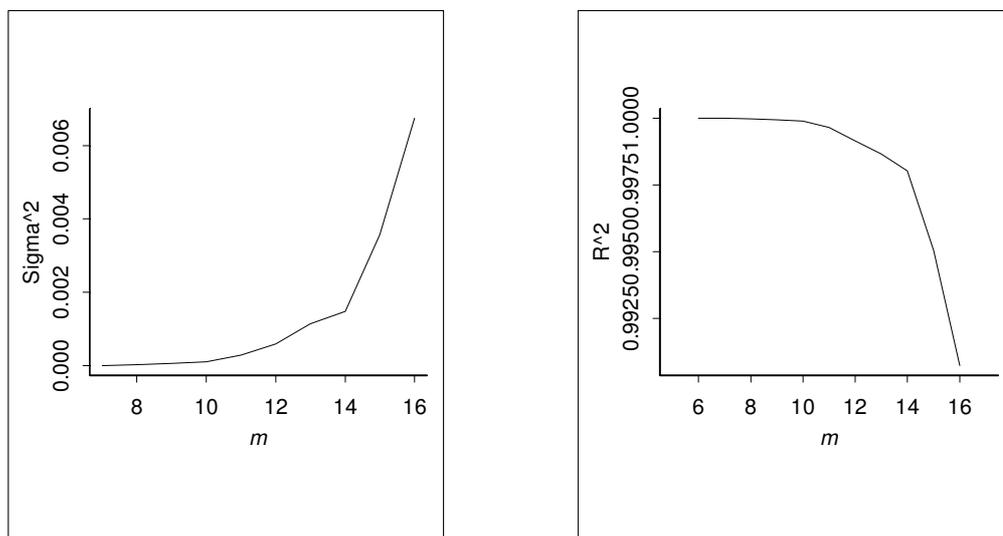
**Figura 4.32** – Gráfico da Distância de Cook.

### **Forward Search – Exemplo 4.2**

Os gráficos das figuras a seguir foram *gerados* com o módulo de regressão linear da FS, instalado no *S-Plus*. Este procedimento foi descrito no Capítulo 2.

Na Figura 4.33 apresentamos os resultados da FS para a estimativa  $\hat{\sigma}^2$  da variância do termo do erro (esquerda) e do coeficiente de determinação múltipla  $R^2$  (direita).

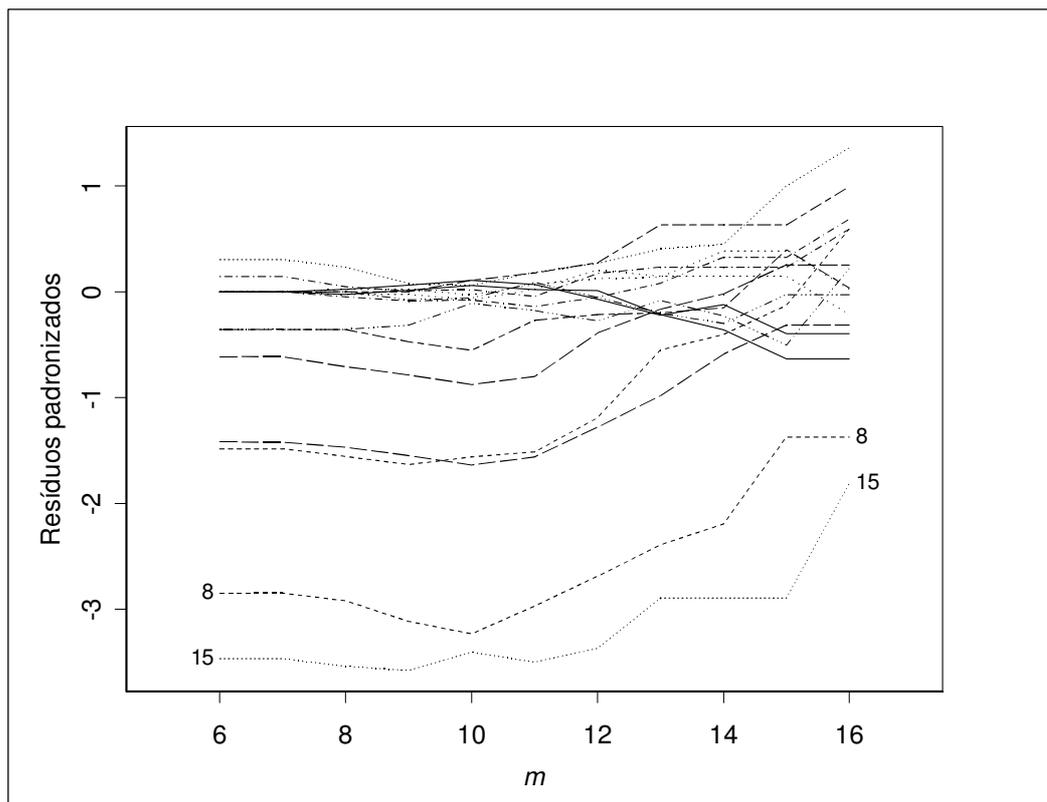
As respostas 8 e 15 são as mais influentes (ver Figura 4.32). Justamente, são as duas últimas a serem incluídas na FS. Até a inclusão da observação 8 os valores de  $\hat{\sigma}^2$  e de  $R^2$  mudam lentamente. Quando esta observação e a observação 15 são incluídas,  $\hat{\sigma}^2$  e  $R^2$  mudam drasticamente. Este fato confirma a influência das duas observações.



**Figura 4.33** – Dados do Exemplo 4.2: *Forward Search* de  $\hat{\sigma}^2$  e de  $R^2$ .

Na Figura 4.34 apresentamos o gráficos dos *resultados* da FS para os resíduos padronizados. Durante toda a FS, os resíduos das observações 8 e 15 destacam-se de todos os outros e permanecem elevados até serem incluídos. Isto indica observações atípicas mascaradas.

Cabe lembrar que a ocorrência de observações atípicas pode ser causada por inadequação do modelo. O gráfico da Figura 4.25 e o teste de Cook e Weisberg indicaram que a variância não deve ser considerada constante. Por conseguinte, as respostas 8 e 15 podem não ser atípicas.



**Figura 4.34** – *Forward Search* dos Resíduos Padronizados.

### Comentário sobre os métodos de identificação da transformação da resposta

Observamos que em *todos* os exemplos apresentados pelos autores desses métodos não há interações significativas. Os métodos são aplicados apenas com fatores principais. O mesmo acontece com os exemplos apresentados por Box e Cox, (1964), Box *et al* (1978), Montgomery (2001a) e Myers e Montgomery, (2002).

Embora não esteja explícito em nenhum texto, faz sentido que esses autores apliquem a transformação da resposta para simplificar modelos que apresentam interações e termos quadráticos, obtendo modelos sem esses termos. Entretanto, o que fazer se há interações significativas no experimento? A seguir vamos apresentar um exemplo em que há interações significativas e veremos que nenhum dos quatro métodos funciona.

### Exemplo 4.3 – Transformação da resposta

Na Tabela 3.3 apresentamos os dados de um experimento que foi gerado a partir do modelo

$$\mu = \frac{1}{0,040 + 0,008x_1 + 0,010x_2 + 0,012x_3 + 0,005x_1x_3 + 0,002x_2x_3}.$$

A distribuição de probabilidade escolhida foi a gama com parâmetro de dispersão  $\phi = 0,01$ . Entretanto, como os métodos de identificação assumem que a distribuição da resposta é normal, para aplicá-los, vamos gerar os dados a partir dessa distribuição. O modelo é  $y = \mu + \varepsilon$ . O erro  $\varepsilon$  tem média zero e  $\text{var}(\varepsilon) = 4$ .

Para gerar aleatoriamente os dados usamos a inversa da distribuição normal acumulada (INVNORMAL) do *Excel*. Para isto necessitamos de 16 números aleatórios uniformes entre zero e um, os quais foram gerados no software *S-Plus* a partir da semente 110945. Portanto, temos que

$$y_i = \text{INVNORMAL}\{\text{Uniforme}[0, 1]; \mu_i; \text{desvio padrão}\}$$

Na Tabela 4.8 apresentamos a média verdadeira, os números uniformes e as respostas geradas.

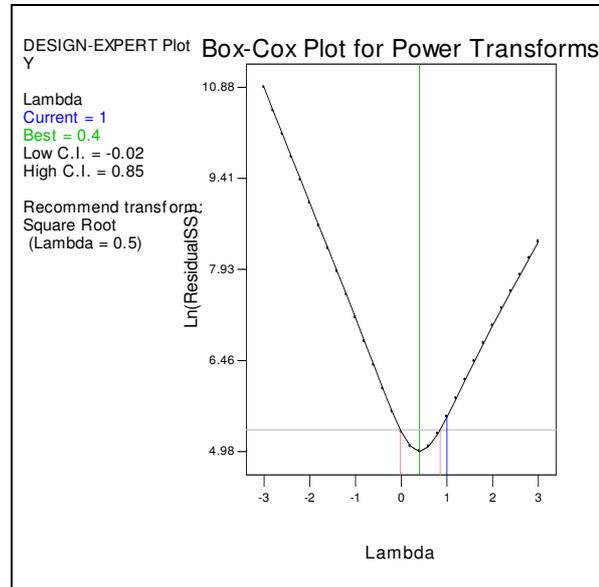
**Tabela 4.9 - Dados do Experimento.**

$x_1$	$x_2$	$x_3$	$x_4$	$\mu_i$	Uniforme [0, 1]	$y_i$
-1	-1	-1	-1	58,82	0,7566	60,21
1	-1	-1	-1	43,48	0,3935	42,94
-1	1	-1	-1	30,30	0,7853	31,88
1	1	-1	-1	25,64	0,3652	24,95
-1	-1	1	-1	37,04	0,3025	36,00
1	-1	1	-1	18,87	0,5514	19,13
-1	1	1	-1	19,61	0,1718	17,71
1	1	1	-1	12,99	0,0092	8,27
-1	-1	-1	1	58,82	0,6575	59,63
1	-1	-1	1	43,48	0,3657	42,79
-1	1	-1	1	30,30	0,4371	29,99
1	1	-1	1	25,64	0,5067	25,67
-1	-1	1	1	37,04	0,4269	36,67
1	-1	1	1	18,87	0,4393	18,56
-1	1	1	1	19,61	0,8106	21,37
1	1	1	1	12,99	0,7153	14,13

### Método de Box e Cox

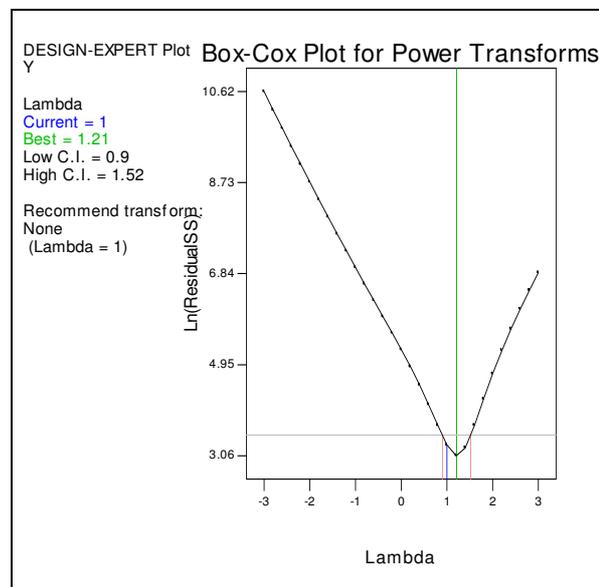
Na Figura 4.35 apresentamos o gráfico do  $\ln(SQR_z)$  versus  $\lambda$ , fornecido pelo *Design Expert*. No gráfico também são fornecidos o valor de  $\hat{\lambda} = 0,4$  e o intervalo

$[-0,02 \ 0,85]$  para 95% de confiança. O valor que sabemos ser o verdadeiro não está dentro do intervalo para 95% de confiança. Este método não indicou a transformação adequada.



**Figura 4.35** – Transformação de Box-Cox.

Vamos supor que soubéssemos que as interações reais  $x_1x_3$  e  $x_2x_3$  são significativas. Vamos incluí-las no modelo. Na Figura 4.36 apresentamos o gráfico do  $\ln(SQR_z)$  versus  $\lambda$ , fornecido pelo *Design Expert*.



**Figura 4.36** – Transformação de Box-Cox - com as Interações Reais.

No gráfico também são fornecidos o valor de  $\hat{\lambda} = 1,21$  e o intervalo  $[0,90 \ 1,52]$  para 95% de confiança. O valor que sabemos ser o verdadeiro ( $\lambda = -1$ ) não está dentro do intervalo para 95% de confiança.

Este método não indicou a transformação adequada. O fato é que, quando se inclui as interações, o modelo pode acompanhar as irregularidades dos dados, indicando o modelo linear sem transformação como adequado, mascarando as deficiências do modelo quanto à adequação da transformação. Mesmo quando usamos o modelo linear, constituído pelos fatores principais  $x_1$ ,  $x_2$ ,  $x_3$ ,  $x_4$ , e pelas interações reais  $x_1x_3$  e  $x_2x_3$ , este método não funcionou bem. Uma conclusão plausível é que este método não funciona para os casos em que as interações não são importantes, embora seja prematuro afirmar sem maior investigação. Entretanto, podemos afirmar que nem sempre este método funciona bem.

### Método de Cook e Weisberg

O software ARC permite facilmente construir e comparar gráficos para diversos valores de  $\lambda$ , os quais são apresentados na Figura 4.37.

O melhor ajuste é para  $\lambda = 0,5$ . Podemos tirar as mesmas conclusões que tiramos sobre o método de Box-Cox.

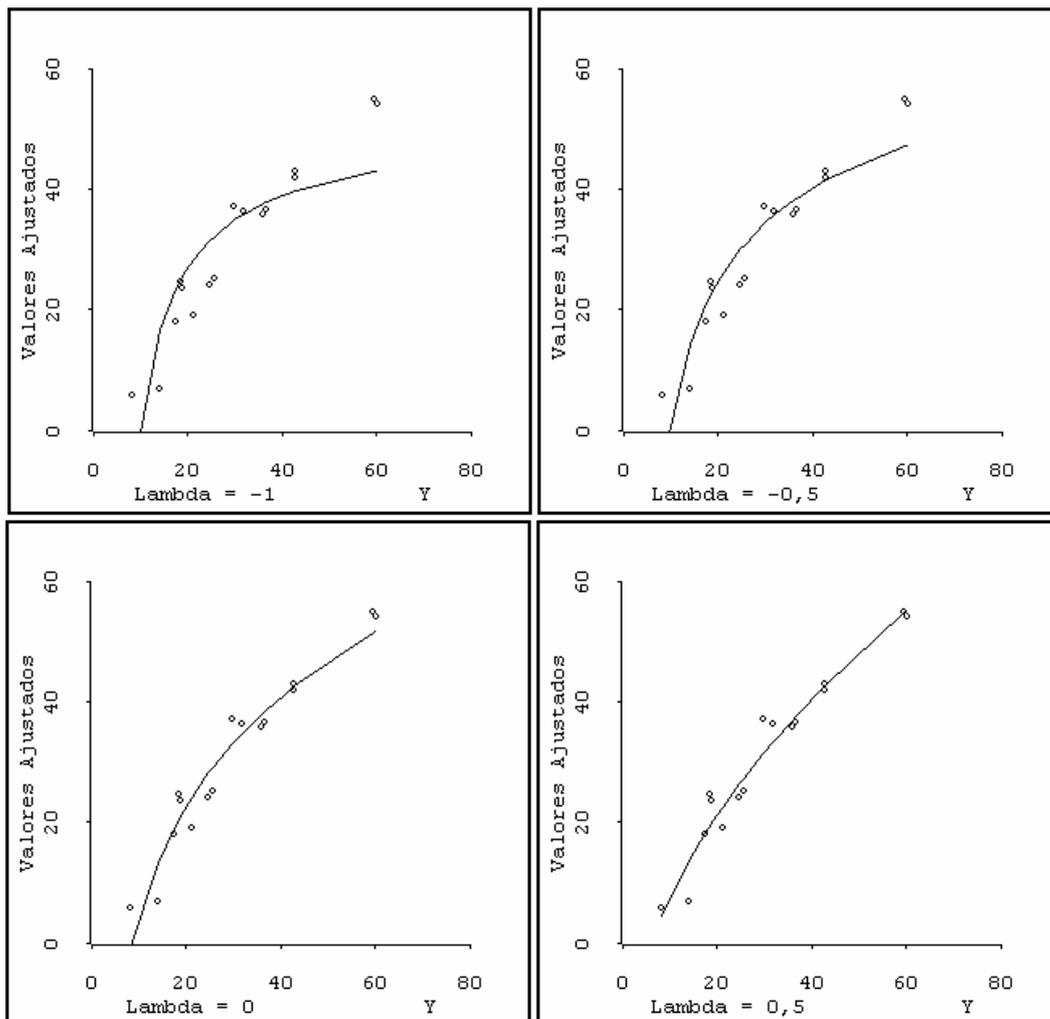
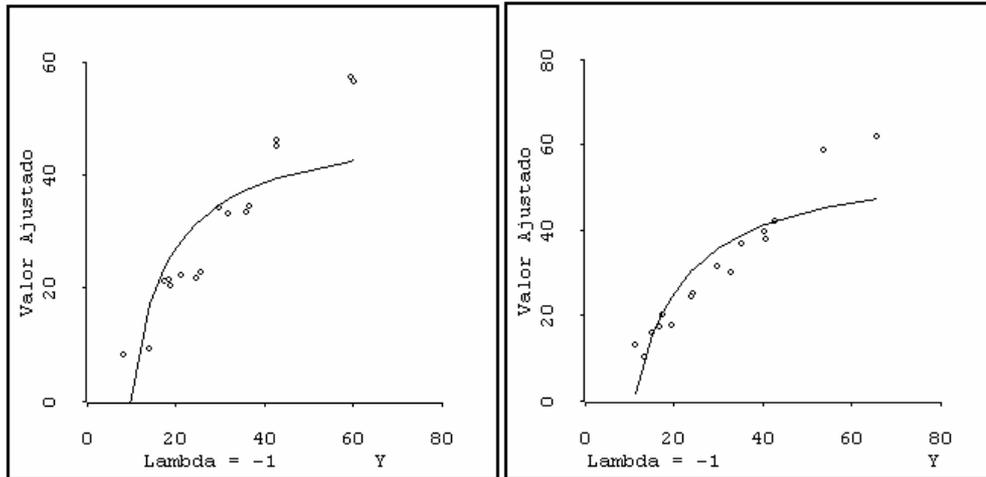


Figura 4.37 – Gráficos Inversos para Valores de  $\lambda$ .

Vamos supor que soubéssemos que as interações reais  $x_1x_3$  e  $x_2x_3$  são significativas. Vamos incluí-las no modelo. Para este modelo, experimentamos todos os valores de  $\lambda$  e o que ofereceu melhor ajuste foi  $\lambda = 1$ , ou seja, sem transformação, que não é a escolha adequada. Na Figura 4.38 apresentamos os gráficos inversos para  $\lambda = 1$  e  $\lambda = -1$ , fornecidos pelo software ARC.

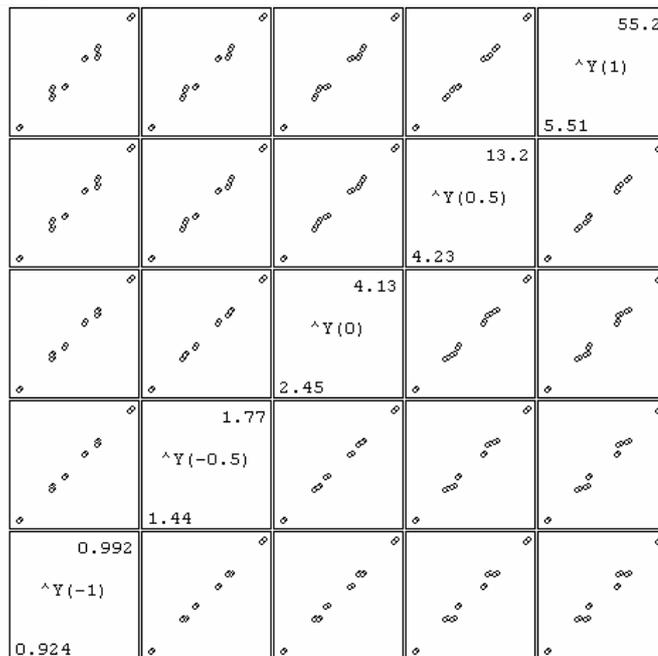


**Figura 4.38** – Gráficos Inversos para  $\lambda = 1$  e  $\lambda = -1$ .

Mesmo quando usamos o modelo linear, constituído pelos fatores principais  $x_1, x_2, x_3, x_4$ , e pelas interações reais  $x_1x_3$  e  $x_2x_3$ , este método também não funcionou bem.

**Método de Cook e Olive**

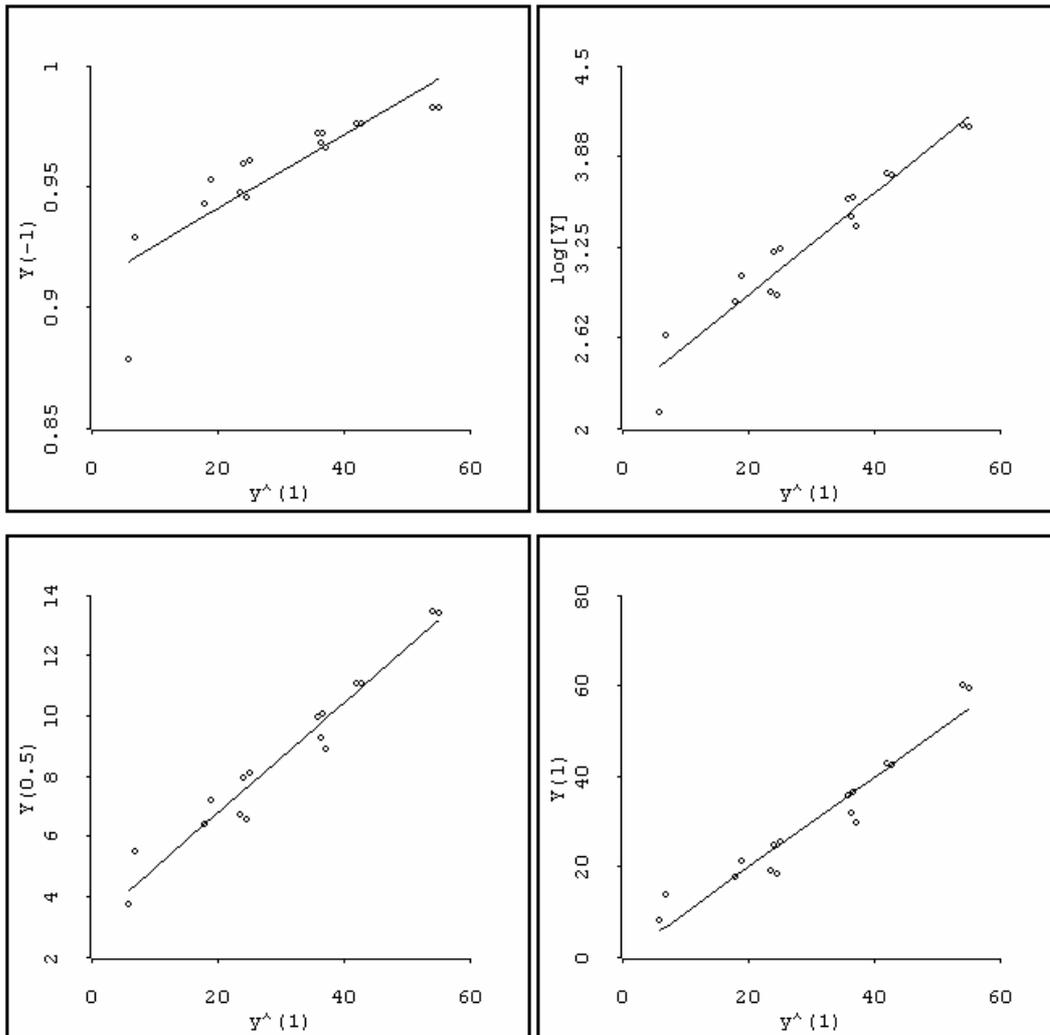
Na Figura 4.39, apresentamos os gráficos de valor ajustado  $\hat{y}(\lambda_i)$  versus valor ajustado  $\hat{y}(\lambda_j)$ , fornecidos pelo software ARC.



**Figura 4.39** - Gráficos de Valor Ajustado *Versus* Valor Ajustado.

Observa-se que há uma forte correlação dos diversos valores ajustados. O maior coeficiente de correlação é 0,9995 e o menor é 0,9919. Escolhemos um dos valores ajustados,  $\hat{y}(1)$ .

Para cada um dos valores considerados para  $\lambda$  fazem-se gráficos dos diversos valores transformados  $y(\lambda)$  versus  $\hat{y}(1)$ . Na Figura 4.39 são apresentados esses gráficos, fornecidos pelo software ARC.



**Figura 4.40** - Gráficos dos Valores Transformados  $y(j)$  Versus  $\hat{y}(1)$ .

Visualmente, os gráficos que oferecem o melhor ajuste é  $\hat{y}(0.5)$ . O gráfico de  $\hat{y}(-1)$ , que deveria ser o indicado, apresenta uma curvatura para baixo.

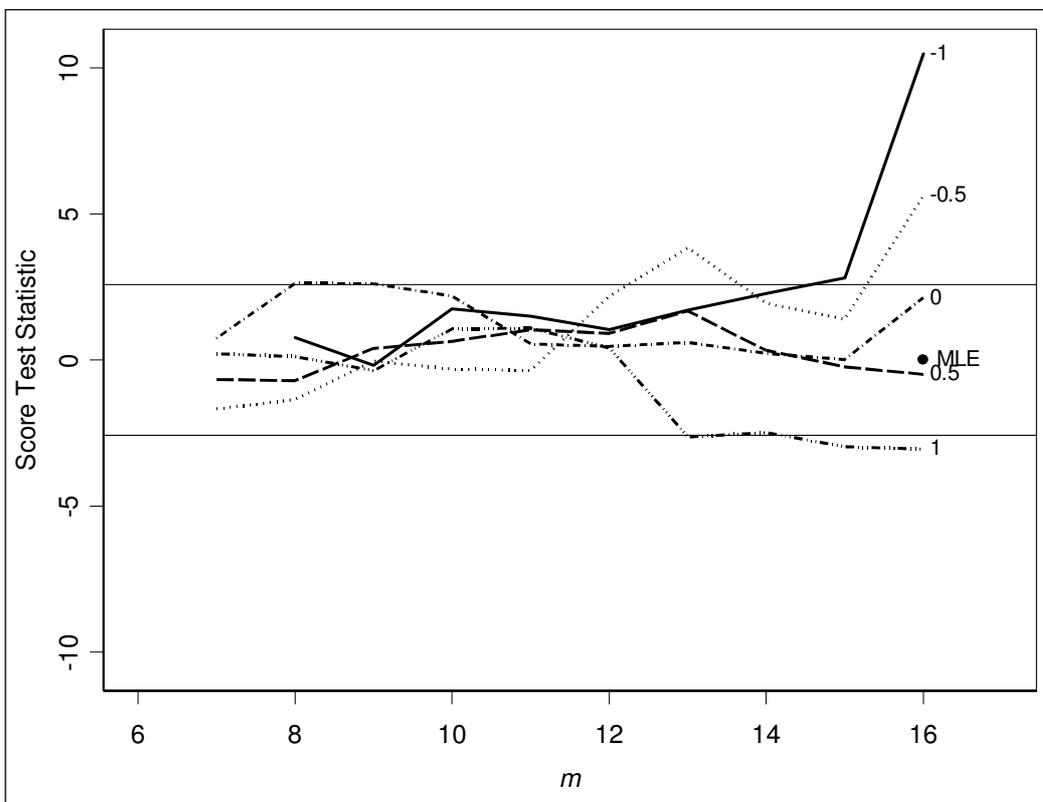
Consideramos o modelo linear, constituído pelos fatores principais reais  $x_1$ ,  $x_2$ ,  $x_3$  e pelas interações reais  $x_1x_3$  e  $x_2x_3$ . A melhor escolha indicada pelos gráficos foi a opção inadequada da transformação  $\lambda = 1$ . Mais uma vez as interações mascararam a transformação adequada.

Uma conclusão plausível é que, da mesma forma que os métodos de Box e Cox e de Cook e Weisberg, este método só funciona para os casos em que as interações não são importantes.

### Método de Atkinson e Riani

Este método foi descrito na Seção 4.3.1. Vamos considerar inicialmente o modelo constituído apenas pelos fatores principais  $x_1, x_2, x_3, x_4$ .

Na Figura 4.40 é apresentada a *Forward Search* (FS) com os gráficos da estatística de teste  $T_p(\lambda)$ , fornecidos pelo *S-Plus*. Foram considerados cinco valores de  $\lambda$ , perfazendo cinco modelos de regressão linear.

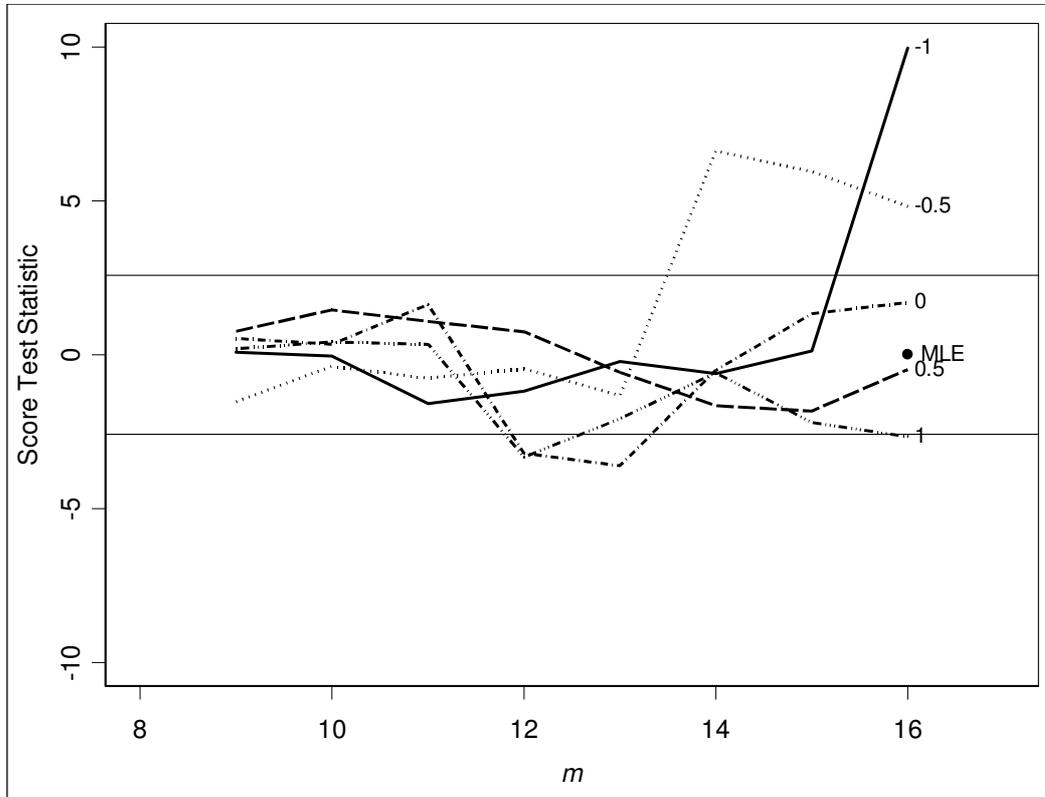


**Figura 4.41** – *Forward Search* da Estatística de Teste.

Observando o gráfico da estatística de teste constatamos que a transformação indicada é  $\lambda = 0,5$ , sendo, entretanto, aceitável  $\lambda = 0$ . A transformação que sabemos ser a adequada,  $\lambda = -1$ , não é considerada uma boa escolha.

Consideramos o modelo linear, constituído pelos fatores principais reais  $x_1, x_2, x_3$  e pelas interações reais  $x_1x_3$  e  $x_2x_3$ . Na Figura 4.41 é apresentada a *Forward Search* (FS) com os gráficos da estatística de teste  $T_p(\lambda)$ , fornecidos pelo *S-Plus*.

Observando o gráfico da estatística de teste constatamos que a transformação indicada é  $\lambda = 0,5$ , sendo, entretanto, aceitável  $\lambda = 0$ . A transformação que sabemos ser a adequada,  $\lambda = -1$ , não é considerada uma boa escolha.



**Figura 4.42** – *Forward Search* da Estatística de Teste com Interações Reais.

Mesmo quando usamos o modelo linear, constituído pelos fatores principais  $x_1, x_2, x_3, x_4$ , e pelas interações reais  $x_1x_3$  e  $x_2x_3$ , este método não funcionou bem.

Todos os quatro métodos não funcionaram bem para este exemplo. Uma conclusão plausível é que eles só funcionam para casos em que as interações não são importantes, embora seja prematuro afirmar sem maior investigação. Entretanto, podemos afirmar que nem sempre estes métodos funcionam bem. Para afirmações mais conclusivas sobre a eficácia dos quatro métodos é recomendável uma investigação mais ampla, que fica como indicação para pesquisa futura.

### Exemplo 3.1 Modificado

Na Tabela 3.3 apresentamos os dados de um experimento que foi gerado a partir do modelo

$$\mu_i = \frac{1}{0,040 + 0,008x_1 + 0,010x_2 + 0,012x_3 + 0,005x_1x_3 + 0,002x_2x_3}$$

Agora vamos usar o modelo sem interações

$$\mu_i = \frac{1}{0,040 + 0,008x_1 + 0,010x_2 + 0,012x_3}$$

A distribuição de probabilidade escolhida foi gama com parâmetro de dispersão  $\phi = 0,01$ .

Observe que estamos mantendo a distribuição da resposta como sendo gama. Os métodos de identificação assumem que a distribuição da resposta é normal. Apesar de usarmos uma distribuição diferente, veremos que os quatro métodos funcionam bem quando não há interação significativa.

Foi usado o mesmo procedimento descrito na Seção 3.4, inclusive os mesmos números aleatórios uniformes gerados a partir da semente 435187. Na tabela 4.10 apresentamos os dados de um experimento que foi gerado a partir do modelo sem interações.

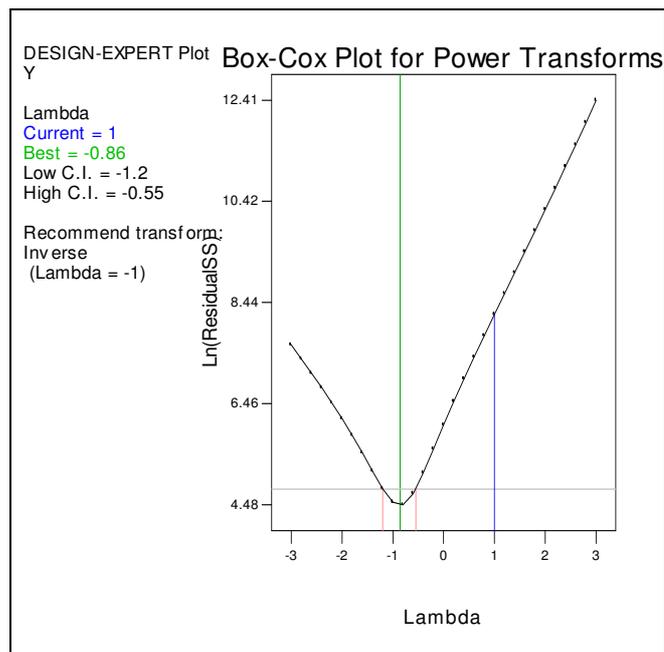
**Tabela 4.10** - Dados do Experimento.

$x_1$	$x_2$	$x_3$	$x_4$	$\mu_i$	Uniforme [0, 1]	$\beta_i$	$y_i$
-1	-1	-1	-1	100,00	0,8835	1,0000	112,05
1	-1	-1	-1	38,46	0,4619	0,3846	37,97
-1	1	-1	-1	33,33	0,4748	0,3333	33,02
1	1	-1	-1	21,74	0,3281	0,2174	20,72
-1	-1	1	-1	29,41	0,3201	0,2941	27,97
1	-1	1	-1	20,00	0,0260	0,2000	16,31
-1	1	1	-1	18,52	0,5019	0,1852	18,47
1	1	1	-1	14,29	0,6284	0,1429	14,72
-1	-1	-1	1	100,00	0,2063	1,0000	91,72
1	-1	-1	1	38,46	0,2466	0,3846	35,77
-1	1	-1	1	33,33	0,8101	0,3333	36,24
1	1	-1	1	21,74	0,2938	0,2174	20,52
-1	-1	1	1	29,41	0,8468	0,2941	32,42
1	-1	1	1	20,00	0,1717	0,2000	18,11
-1	1	1	1	18,52	0,1524	0,1852	16,63
1	1	1	1	14,29	0,1274	0,1429	12,68

### Método de Box e Cox - Exemplo 3.1 Modificado

Para o Exemplo 3.1 modificado, o software *Design Expert* fornece um gráfico do  $\ln(SQR_z)$  versus  $\lambda$  apresentado na Figura 4.42.

Na figura também é fornecido o valor de  $\hat{\lambda} = -1$  e um intervalo  $[-1,20 - 0,55]$  para 95% de confiança. Portanto, a transformação recomendada por este método é a inversa,  $\lambda = -1$ , que é a adequada.



**Figura 4.43** – Gráfico da Transformação de Box–Cox.

### Método de Cook e Weisberg - Exemplo 3.1 Modificado

O software *ARC* permite construir e comparar os gráficos inversos para diversos valores de  $\lambda$ : ver a Figura 4.43, gerada por ele. Observando os gráficos da Figura 4.43, fica evidente que  $\lambda = -0,67$  propicia o melhor ajuste dentre os seis valores experimentados e está dentro do intervalo de Box-Cox  $[-1,20 -0,55]$  para 95% de confiança. Entretanto a opção por  $\lambda = -1$  propicia um bom ajuste com exceção das duas maiores respostas, que o modelo subestima.

### Método de Cook e Olive - Exemplo 3.1 Modificado

Para este modelo temos, na Figura 4.44, os gráficos de valor ajustado  $\hat{y}(\lambda_i)$  versus valor ajustado  $\hat{y}(\lambda_j)$ , fornecidos pelo software *ARC*. Observa-se que há uma forte correlação dos diversos valores ajustados. O maior coeficiente de correlação é 0,9999 e o menor é 0,9973.

Para cada um dos valores considerados para  $\lambda$  fazem-se gráficos dos diversos valores transformados  $y(\lambda)$  versus  $\hat{y}(1)$ . Na Figura 4.45 são apresentados esses gráficos, fornecidos pelo software *ARC*. Visualmente, o gráfico que fornece o melhor ajuste é o de  $\hat{y}(-1)$ . O método de Box e Cox fornece um intervalo de confiança  $[-1,20 -0,55]$ . Portanto, a transformação indicada deve ser mesmo a inversa ( $\lambda = -1$ ).

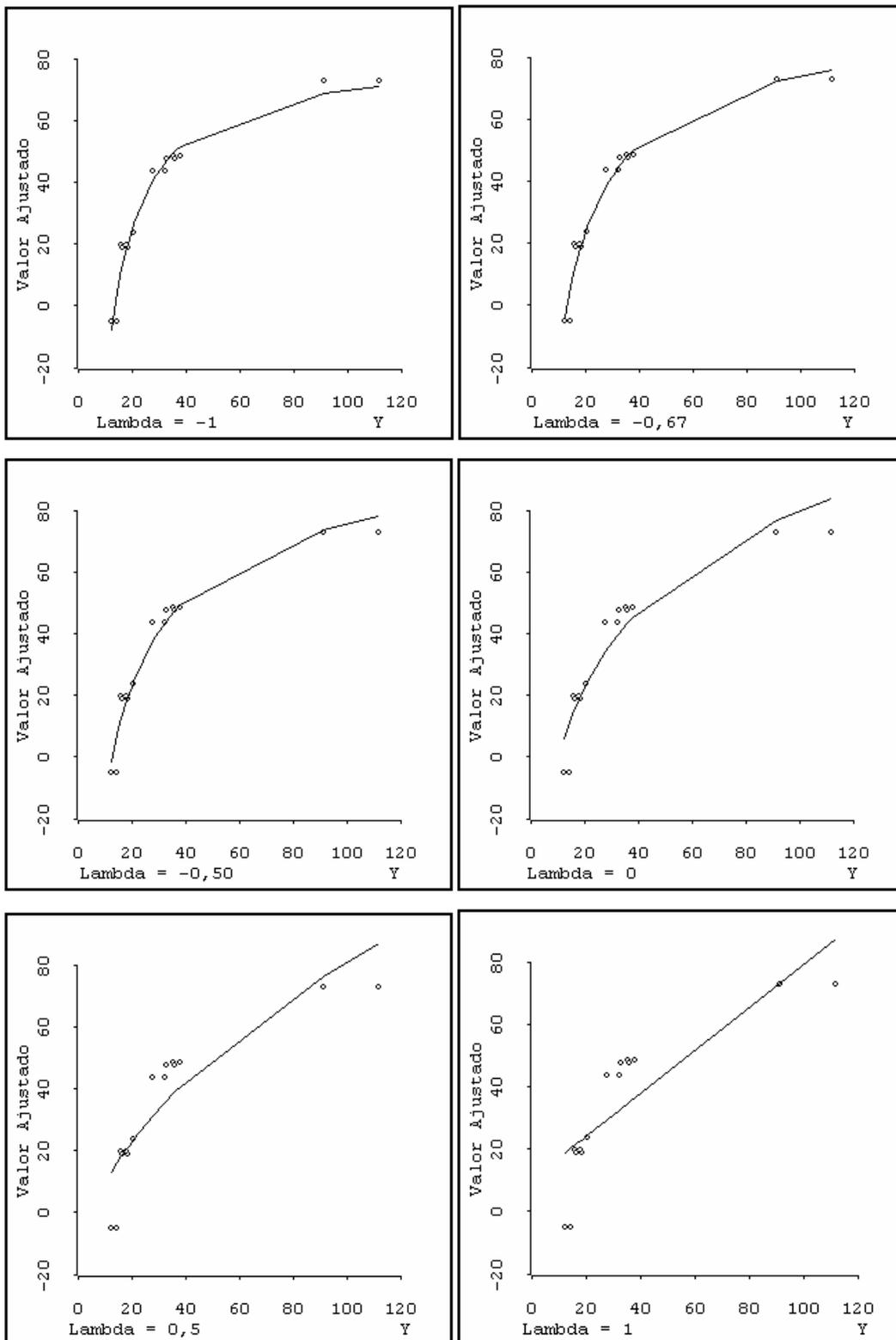
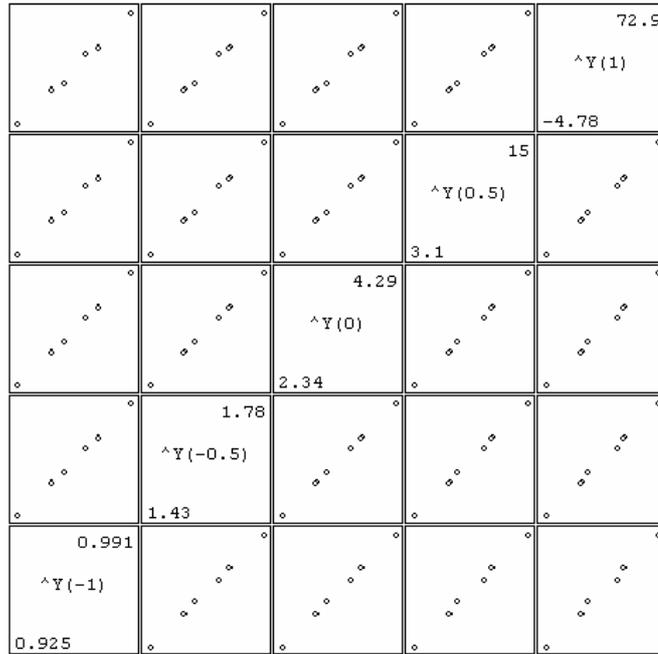
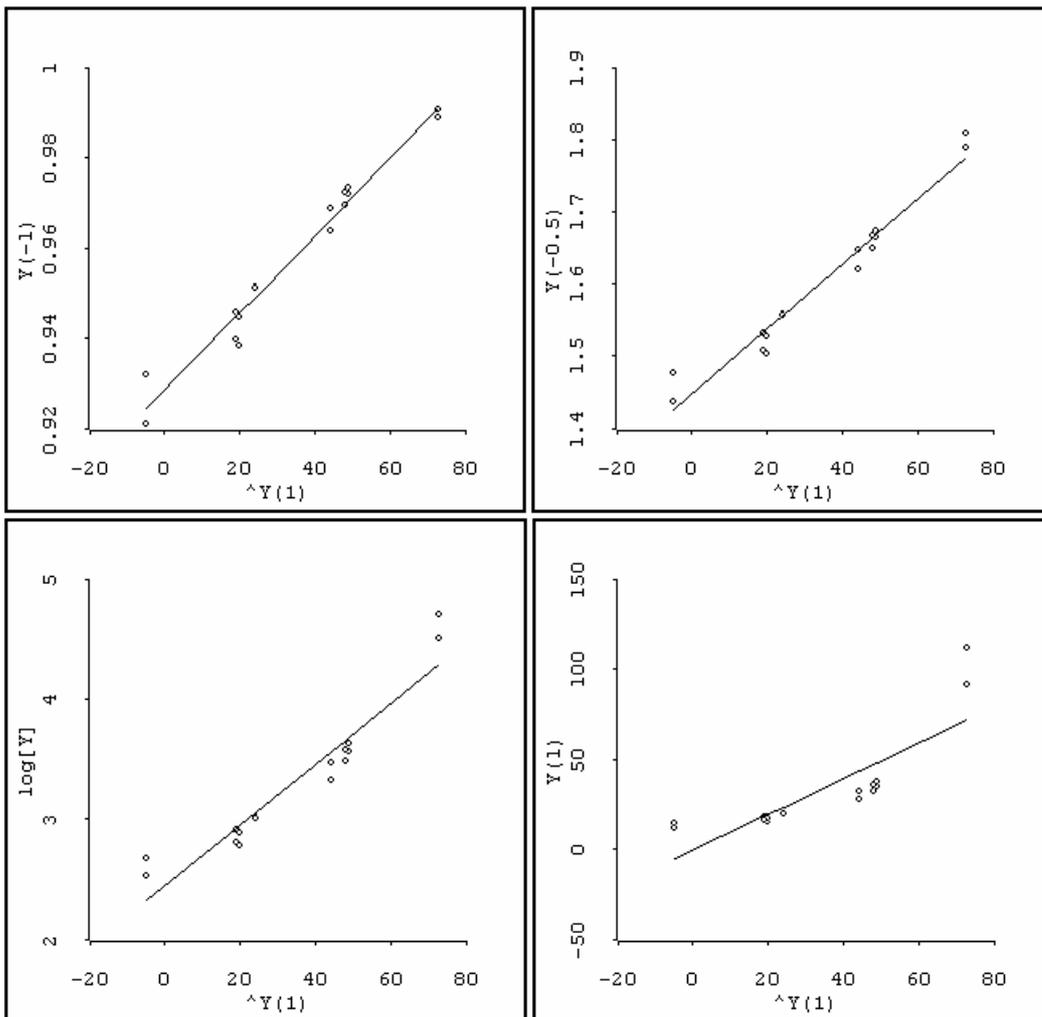


Figura 4.43– Gráficos Inversos para valores de  $\lambda$ .



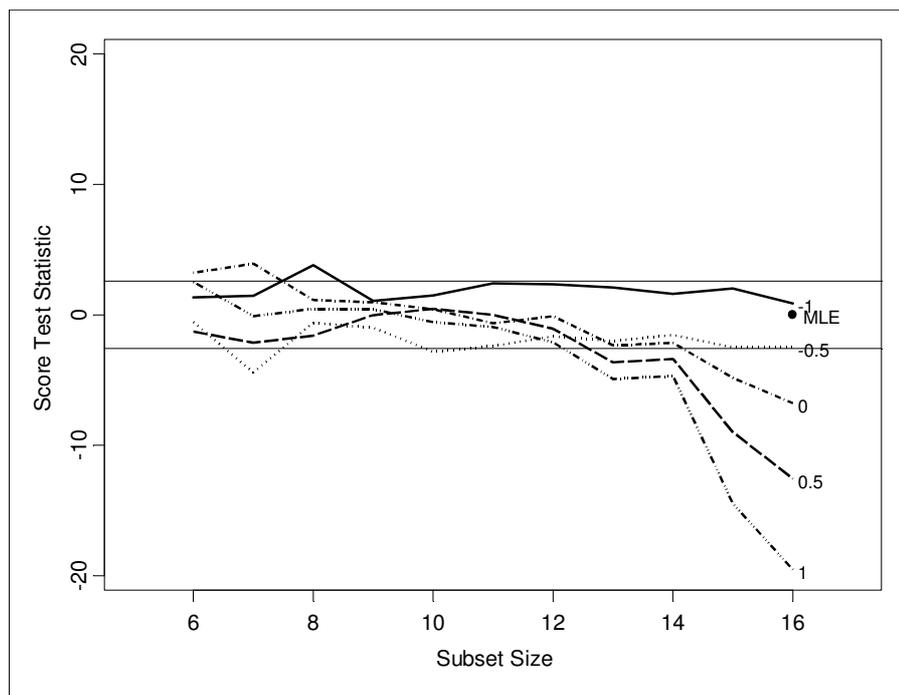
**Figura 4.44** - Gráficos de Valor Ajustado *Versus* Valor Ajustado.



**Figura 4.45** - Gráficos dos Valores Transformados  $Y(\lambda)$  *Versus*  $\hat{Y}(1)$ .

### Método de Atkinson e Riani - Exemplo 3.1 Modificado

Na Figura 4.46 é apresentada a *Forward Search* (FS) com os gráficos da estatística de teste  $T_p(\lambda)$ , fornecidos pelo S-Plus. Foram considerados cinco valores de  $\lambda$ , perfazendo cinco modelos de regressão linear. Observando o gráfico da estatística de teste, constatamos que a transformação indicada é  $\lambda = -1$ . Portanto, temos indicação da melhor escolha que sabemos ser a melhor.



**Figura 4.46** – *Forward Search* da Estatística de Teste.

A aplicação dos quatro métodos no Exemplo 4.2, no Exemplo 4.3 e no Exemplo 3.1 modificado foi elucidativa.

No Exemplo 4.2 os quatro métodos indicaram a transformação logarítmica, que parece ser a correta, isto porque, como veremos na Seção 4.4.3, no modelo escolhido a função de ligação é a logarítmica.

Na Exemplo 3.1 modificado, sem interação, todos os métodos funcionaram bem, indicando a transformação que sabíamos ser a correta.

Todos os quatro métodos não funcionaram bem quando havia interação significativa (como no Exemplo 4.3). Uma conclusão plausível é que eles só funcionam bem quando as interações não são significativas ou fracas, embora seja prematuro afirmar isso sem maior investigação. Entretanto, podemos afirmar que nem sempre estes métodos funcionam bem.

#### 4.4. Estágio 3 - Modelos Lineares Generalizados

Nos estágios 1 e 2 foram considerados apenas modelos com distribuição normal e função de ligação identidade; nesse caso a definição completa do modelo restringe-se à identificação dos efeitos significativos.

Com os modelos lineares generalizados (MLG), o conjunto de modelos possíveis é mais amplo. Além da normal, podemos considerar as distribuições lognormal, gama e normal inversa. Junte-se a isto a família  $g(\mu) = \mu^\lambda$  de cinco funções de ligação  $\mu^{-1}$ ,  $\mu^{-0,5}$ ,  $\ln(\mu)$ ,  $\mu^{0,5}$  e  $\mu$ , e teremos um conjunto de vinte alternativas para analisar.

É verdade que, em aplicações específicas, freqüentemente, sabemos de antemão que tais ou quais distribuições não são adequadas para representar a resposta em questão e, por conseguinte, podem ser excluídas de consideração.

Ademais, indicações da natureza da função de variância e do tipo de função de ligação emergem quando conduzimos os estágios anteriores, reduzindo o número de modelos candidatos. Por exemplo, no Exemplo 4.2 há duas indicações importantes: i) a variância da resposta cresce com o aumento da média e ii) a transformação logarítmica da resposta foi considerada por todos os métodos como a mais adequada para se alcançar a aditividade nos efeitos, o que sugere o uso de uma função de ligação logarítmica para os MLG. Por conseguinte, as distribuições gama e normal inversa (cuja variância cresce com o aumento da média) e a função de ligação logarítmica podem ser escolhidas para continuação da análise, o que será feito mais adiante.

Cabe acrescentar que não é surpresa a indicação de distribuições assimétricas no Exemplo 4.2. Isto porque a resposta é essencialmente positiva e há valores da resposta baixos, próximos de zero.

Caso não haja indicações para a função de variância ou para a função de ligação, a idéia é fazer uma seleção entre as vinte (ou menos) alternativas de modelagem, com base no valor da função de log-verossimilhança. Uma abordagem tem sido minimizar o critério de informação de Akaike (AIC) (Akaike, 1973):

$$AIC = -2 \sum_{i=1}^n \ln L(\hat{\mu}_i, y_i) + 2p$$

onde:

$y_i$  é o  $i$ -ésimo valor da resposta,  $i = 1, 2, \dots, n$ .

$\hat{\mu}_i = E[y_i] = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}'_i \hat{\boldsymbol{\beta}})$  é a estimativa de  $y_i$ , quando ajustamos um modelo de  $p$  parâmetros pela maximização da função de log-verossimilhança (FLV);

$\mathbf{x}'_i = [1, x_{i1}, x_{i2}, \dots, x_{ik}]$  é a  $i$ -ésima linha da matriz ( $\mathbf{X}$ ) das variáveis de regressão;

$\ln L(\hat{\mu}_i, y_i)$  é o valor da FLV maximizada, para a resposta  $y_i$ , quando ajustamos um modelo de  $p$  parâmetros.

Entre vários modelos candidatos, deve ser escolhido aquele que apresentar o menor valor de AIC. Isto significa escolher aquele modelo com maior valor do somatório da FLV  $\ln L(\hat{\mu}_i, y_i)$ , com uma penalidade  $p$  para modelos mais complexos.

O AIC origina-se da minimização da informação (ou distância) de Kullback-Leibler (K-L) como base para a seleção de modelos. A informação K-L é uma medida da distância entre o modelo verdadeiro e um modelo candidato. Na realidade este modelo verdadeiro quase sempre é uma abstração. De fato, estamos interessados em um bom modelo; um modelo que represente satisfatoriamente o mecanismo que gerou os dados em questão. Por conseguinte o que necessitamos é de uma medida da distância entre um bom modelo e vários modelos candidatos. Queremos evidências de um modelo que se destaque; as vezes conseguimos até dois modelos que se destacam.

Akaike (1973) desenvolveu uma estimativa da informação K-L, baseada na FLV em seu ponto máximo. Este desenvolvimento pode ser visto em Burnham e Anderson (2002). Estes autores só recomendam usar o AIC para selecionar modelos quando o número de respostas  $n$  é maior do que pelo menos 40 vezes o número de parâmetros  $p$ . Esta recomendação exclui todos os exemplos apresentados neste trabalho e todos os exemplos com experimentos fatoriais aplicados à qualidade que conhecemos na literatura. Davison (2001) afirma que o AIC não propicia uma seleção consistente de modelos, e que em aplicações práticas, freqüentemente indica modelos mais complexos do que deveriam ser. Nestes casos a origem da deficiência é que a complexidade é insuficientemente penalizada.

Hurvich e Tsai (1989) desenvolveram um critério de seleção de modelos, o  $AIC_c$ , indicado para pequenas amostras e respostas com distribuição normal.

$$AIC_c = -2 \sum_{i=1}^n \ln L(\hat{\mu}_i; y_i) + 2 \frac{n(n-p)}{n-p-2}$$

Davison (2001) afirma que o critério  $AIC_c$  pode aumentar apreciavelmente a probabilidade de se escolher o modelo correto, particularmente na seleção de modelos de regressão e de séries temporais. Burnham e Anderson (2002) afirmam que o  $AIC_c$  pode ser usado mesmo quando a distribuição não é normal, desde que as respostas sejam independentes, a distribuição seja unimodal e não seja fortemente assimétrica. Eles realizaram um estudo de simulação de Monte Carlo com a distribuição gama e concluíram que o  $AIC_c$  tem bom desempenho com amostras de tamanho superior a 10, desde que o parâmetro de dispersão  $\alpha = 1/\phi$  não seja inferior a um. Nos experimentos analisados até agora não encontramos valores de  $\alpha$  cuja estimativa fosse inferior a um. Contudo, consideramos arriscado usar para distribuições não normais um critério que foi desenvolvido para a distribuição normal.

Dos dois parágrafos anteriores podemos aduzir que AIC e  $AIC_c$  não são indicadores confiáveis para a seleção de modelos não normais e com reduzido número de observações (que é o caso dos experimentos usados na modelagem de uma característica de qualidade). Ambos os índices são calculados com base na FLV como estimativa da distância entre um bom modelo e vários modelos candidatos. O termo que se adiciona à FLV é um termo que tem a finalidade de corrigir um viés proveniente da comparação de modelos com diferentes números de parâmetros. Se dois ou mais modelos têm o mesmo número de parâmetros, porque não utilizar um critério de seleção baseado na FLV, pura e simples, em seu ponto máximo? Já não utilizamos a FLV para estimativa dos parâmetros? Mais adiante, quando descrevermos um procedimento para a Etapa 1 deste Estágio 3, a seleção de modelos será pelo valor da FLV resultante da estimativa dos parâmetros de cada modelo em questão.

Uma vez escolhidas a distribuição de probabilidade e a função de ligação, a segunda etapa é proceder à análise da significância dos parâmetros do modelo, ou seja, vamos definir quais os elementos significativos do vetor  $\hat{\beta}$  dos coeficientes

estimados. Para isto faremos uma análise de *deviance* (ANODE), que é uma série de testes de razão de log-verossimilhança. Essa análise será descrita mais adiante.

Com a conclusão da Etapa 2 temos um modelo completamente definido. Ou seja, temos definidas a distribuição de probabilidade e a função ligação, bem como os parâmetros considerados necessários para um bom ajuste dos dados. Finalmente utilizaremos os resíduos para verificar a adequação do modelo. Isto constitui a Etapa 3.

Em resumo, propomos as seguintes etapas de investigação:

**Etapa 1 Escolha da Distribuição de Probabilidade e da Função de Ligação.**

**Etapa 2 Definição dos Parâmetros Significativos: Análise de *Deviance*.**

**Etapa 3 Testes de Adequação do Modelo.**

#### **4.4.1. Etapa 1-Escolha da Distribuição de Probabilidade e Função de Ligação**

Nesta etapa vamos considerar as distribuições normal, lognormal, gama e normal inversa, juntamente com a família  $g(\mu) = \mu^\lambda$  de cinco funções de ligação:  $\mu^{-1}$ ,  $\mu^{-0.5}$ ,  $\ln(\mu)$ ,  $\mu^{0.5}$  e  $\mu$ , e teremos um conjunto de vinte combinações (distribuição x função de ligação) para analisar. Eventualmente algumas destas combinações poderão ser ignoradas.

Em nenhum software por nós utilizado conseguimos obter o valor da FLV maximizada quando ajustamos um modelo. Por conseguinte, vamos indicar como calcular este valor. Para isto, a seguir apresentamos a FLV de cada uma das distribuições mencionadas.

Para cada uma das FLV temos:

- $y_i$ , que é o  $i$ -ésimo valor da resposta  $i = 1, 2, \dots, n$ .
- $\hat{\mu}_i = E[y_i] = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i' \hat{\boldsymbol{\beta}})$ , que é a estimativa de  $\mu_i$  e  $y_i$ , quando ajustamos um modelo de  $p$  parâmetros pela maximização da função de log-verossimilhança (FLV);
- O parâmetro de dispersão  $\phi = \sigma^2$ , que deve ser estimado, também, pela maximização da FLV.

Observação: para a estimativa do vetor  $\hat{\beta}$  não há necessidade de se conhecer  $\phi$ , que só é estimado após o ajuste do modelo.

A estimativa da resposta,  $\hat{\mu}_i$ , pode ser obtida, com o ajuste do modelo, através de diversos softwares, como, por exemplo, *ARC*, *S-Plus* e *Genstat*. Estes foram os softwares por nós utilizados porque, em conjunto, admitem a escolha de todas as combinações de distribuição de probabilidade e função de ligação analisadas.

A seguir são apresentadas as FLV das distribuições consideradas no procedimento para cálculo de seus valores. Para isso, necessitamos apenas das estimativas da média e do parâmetro de dispersão.

$$\text{a) Normal: FLV} = \sum_{i=1}^n \ln L(\hat{\mu}_i, y_i) = \sum_{i=1}^n \left[ -\frac{(y_i - \hat{\mu}_i)^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2) \right].$$

É de conhecimento geral que o estimador de máxima verossimilhança de  $\phi = \sigma^2$  é dado por:

$$\hat{\sigma}^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{n}.$$

Substituindo  $\sigma^2$  na FLV da normal temos:

$$\sum_{i=1}^n \ln L(\hat{\mu}_i, y_i) = -\frac{n}{2} - \frac{n}{2} \ln(2\pi\hat{\sigma}^2).$$

Os softwares não fornecem a estimativa de máxima verossimilhança de  $\sigma^2$ . Entretanto, fornecem a soma dos quadrados dos resíduos,  $SQR = (y_i - \hat{\mu}_i)^2$ .

Então, a FLV da normal pode ser calculada pela fórmula:

$$\sum_{i=1}^n \ln L(\hat{\mu}_i, y_i) = -\frac{n}{2} - \frac{n}{2} \ln\left(2\pi \frac{SQR}{n}\right).$$

Logo, para calcular a FLV da normal basta saber a SQR, fornecida por todos os softwares por nós conhecidos.

$$\text{b) Gama: FLV} = \sum_{i=1}^n \ln L(\hat{\mu}_i, y_i) = \sum_{i=1}^n \left[ \alpha \ln \alpha - \alpha \ln \hat{\mu}_i + (\alpha - 1) \ln y_i - \frac{\alpha y_i}{\hat{\mu}_i} - \ln \Gamma(\alpha) \right]$$

onde  $\alpha = 1/\phi$ .

O método mais fácil de calcular o estimador de máxima verossimilhança  $\hat{\alpha}$  é maximizar diretamente a FLV da distribuição gama, através da ferramenta de otimização *Solver* da planilha *Excel*.

$$\max \sum_{i=1}^n \left[ \hat{\alpha} \ln \hat{\alpha} - \hat{\alpha} \ln \hat{\mu}_i + (\hat{\alpha} - 1) \ln y_i - \frac{\hat{\alpha} y_i}{\hat{\mu}_i} - \ln \Gamma(\hat{\alpha}) \right].$$

A solução é precisa e rápida. Entretanto, é recomendável fornecer um valor inicial de  $\alpha$  não muito diferente de  $\hat{\alpha}$ , pois o processo utilizado pelo *Solver* pode divergir. Uma sugestão é utilizar como valor inicial o desvio *deviance* médio.

$$\alpha_{inicial} = \bar{D} = \frac{D(y_i, \hat{\mu}_i)}{n}.$$

O desvio deviance  $D(y_i, \hat{\mu}_i)$  é fornecido por todos os softwares por nós conhecidos.

Por outro lado, McCullagh e Nelder (1989), pág 295, fornecem a seguinte aproximação para o estimador de máxima verossimilhança de  $\alpha$ .

$$\hat{\alpha} \cong \frac{6 + 2\bar{D}}{D(6 + \bar{D})}.$$

McCullagh e Nelder, pág. 295, afirmam que esta aproximação foi desenvolvida para valores grandes de  $\alpha$ , e, conseqüentemente, para pequenos valores de  $D(y_i, \mu_i)$ , sem indicar que valores seriam estes. Em alguns casos em que estimamos  $\alpha$  os dois métodos forneceram, praticamente, o mesmo resultado. A diferença máxima foi 0,3%. Entre estes casos o menor valor estimado de  $\alpha$  foi 5,18.

Uma vez feita a estimativa de  $\alpha$  podemos calcular  $L(\hat{\mu}_i, y_i)$ , para cada resposta, através da função *DISTGAMA*( $\alpha, \beta_i$ ) da planilha *Excel*, sendo que  $\hat{\beta}_i = \hat{\mu}_i / \hat{\alpha}$ , e então calcular

$$\sum_{i=1}^n \ln L(\hat{\mu}_i, y_i).$$

Em certos casos, quando  $\alpha$  é grande, o cálculo com a planilha *Excel* pode falhar, indicando divisão por zero. É possível que isto ocorra devido ao cálculo da função gama  $\Gamma(\alpha)$ .

Nesses casos, podemos usar esta planilha para calcular

$$\ln L(\hat{\mu}_i, y_i) = \hat{\alpha} \ln \hat{\alpha} - \hat{\alpha} \ln \hat{\mu}_i + (\hat{\alpha} - 1) \ln y_i - \frac{\hat{\alpha} y_i}{\hat{\mu}_i} - \ln \Gamma(\hat{\alpha}),$$

sendo que  $\ln \Gamma(\alpha)$  é calculado com a função LNGAMA, e então calcular

$$\sum_{i=1}^n \ln L(\hat{\mu}_i, y_i)$$

$$\text{c) Normal inversa: FLV} = \sum_{i=1}^n \ln L(\hat{\mu}_i, y_i) = \sum_{i=1}^n \left[ -\frac{1}{2} \frac{(y_i - \hat{\mu}_i)^2}{\sigma^2 \hat{\mu}_i^2 y_i} - \frac{1}{2} \ln(2\pi\sigma^2 y_i^3) \right]$$

Fazendo:

$$\frac{d}{d\sigma^2} \sum_{i=1}^n \ln L(\hat{\mu}_i, y_i) = 0,$$

o estimador de máxima verossimilhança de  $\phi = \sigma^2$  é dado por:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2 y_i}.$$

Sabemos que o desvio *deviance* da distribuição normal inversa é dado por

$$D(y_i, \hat{\mu}_i) = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2 y_i}.$$

Então temos que

$$\hat{\sigma}^2 = \bar{D} = \frac{D(y_i, \hat{\mu}_i)}{n}.$$

O desvio deviance  $D(y_i, \hat{\mu}_i)$  é fornecido por todos os softwares por nós conhecidos. Com a estimativa de  $\sigma^2$  e com os valores de  $\hat{\mu}_i$ , fornecidos pelo software, podemos utilizar a planilha *Excel* para calcular a FLV.

$$\sum_{i=1}^n \ln L(\hat{\mu}_i, y_i) = \sum_{i=1}^n \left[ -\frac{1}{2} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\sigma}^2 \hat{\mu}_i^2 y_i} - \frac{1}{2} \ln(2\pi\hat{\sigma}^2 y_i^3) \right].$$

$$d) \text{ Lognormal: FLV} = \sum_{i=1}^n \ln L(\hat{\mu}_i, y_i) = \sum_{i=1}^n \left[ -\frac{(\ln y_i - \hat{\mu}_i)^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2 y_i^2) \right].$$

Sendo  $y$  lognormal, sabemos que  $\ln(y_i)$  tem distribuição normal com média  $\hat{\mu}_i$  e variância  $\sigma^2$ . Então, tomamos o  $\ln(y_i)$  e procedemos como no caso da distribuição normal, calculando  $\hat{\mu}_i$  e estimando  $\sigma^2$  através da fórmula

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (\ln(y_i) - \hat{\mu}_i)^2}{n} = \frac{SQR}{n}.$$

Substituindo  $\hat{\sigma}^2$  na FLV da lognormal temos:

$$\sum_{i=1}^n \ln L(\hat{\mu}_i, y_i) = -\frac{n}{2} - \sum_{i=1}^n \left[ \frac{1}{2} \ln \left( 2\pi \frac{SQR}{n} y_i^2 \right) \right].$$

A  $SQR$  é fornecida pelo *software*. Podemos utilizar a planilha Excel para calcular a FLV.

$$\sum_{i=1}^n \ln L(\hat{\mu}_i, y_i) = -\frac{n}{2} - \frac{n}{2} \ln \left( 2\pi \frac{SQR}{n} \right) - \sum_{i=1}^n [\ln(y_i)].$$

A idéia básica nesta abordagem é utilizar a FLV como medida da distância de cada modelo em relação aos dados. Entretanto, se for incluído um número elevado de parâmetros, o modelo pode acompanhar as irregularidades dos dados e mascarar deficiências do modelo quanto à adequação da distribuição de probabilidade e, sobretudo, da função de ligação.

Por exemplo: suponhamos um caso em que o melhor modelo não inclui interações e a função de ligação é a logarítmica. Portanto, os efeitos se transmitem sobre a média multiplicativamente. Se usarmos um modelo com a função de ligação identidade, os testes indicarão erradamente interações significativas.

Por conseguinte, tomamos, para esta etapa de escolha da distribuição de probabilidade e da função de ligação, **um modelo com um número reduzido de parâmetros, sem as interações.**

Para cada uma das vinte combinações (ou menos) definimos um modelo que inclui apenas os termos correspondentes aos fatores principais (inclusive os termos quadráticos, se for o caso de experimentos em 3 níveis).

Entre as vinte combinações do modelo reduzido, a que apresentar menor valor da FLV terá uma grande chance de produzir um modelo satisfatório. Como será visto mais adiante, nos casos por nós analisados, tivemos indicação de que isto é verdadeiro quando as interações são não significativas ou fracas, em relação aos fatores principais.

#### **Exemplo 4.2 (continuação)**

Como foi dito, para esta etapa de escolha da distribuição de probabilidade e da função de ligação, consideramos um modelo que inclui apenas os termos correspondentes aos fatores principais, sem as interações.

Entretanto, apesar de ser a estratégia por nós adotada, a título de ilustração, vamos comparar os valores das FLV do modelo com as interações de segunda ordem (modelo maior). Este exemplo conta com quatro fatores principais, seis interações duplas e um termo constante, perfazendo 11 coeficientes a serem estimados. Portanto, o modelo maior é

$$\hat{\mu}_i^\lambda = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \\ + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \beta_{34} x_3 x_4$$

Na Tabela 4.11 são apresentados os valores da FLV deste modelo, em ordem decrescente, para vinte combinações de distribuição de probabilidade e função de ligação.

O maior valor da FLV corresponde à combinação de distribuição normal inversa e função de ligação inverso da raiz quadrada, aparentemente indicando esta combinação como a mais adequada. Contudo, como foi dito: quando um número elevado de parâmetros é incluído na análise, o modelo pode acompanhar as irregularidades dos dados e mascarar deficiências quanto à adequação da escolha da distribuição de probabilidade e da função de ligação. Por conseguinte, esta indicação da combinação normal inversa e inverso da raiz quadrada não é confiável. O que será confirmado na Etapa 2.

De qualquer forma, constatamos que as três combinações com maior valor da FLV são as distribuições normal inversa e gama, o que está de acordo com as indicações anteriores. Agora vamos considerar o modelo sem as interações (modelo menor):  $\mu_i^\lambda = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)$

**Tabela 4.11** - Função de Log-verossimilhança do Modelo Completo.

	<b>Distribuição</b>	<b>Função de Ligação (<math>\lambda</math>)</b>	<b>FLV</b>
1	Normal inversa	Inverso da Raiz Quadrada (-0,5)	5,91
2	Normal inversa	Logaritmo Neperiano (0)	5,22
3	Gama	Inverso (-1)	5,22
4	Normal	Inverso (-1)	5,19
5	Normal inversa	Inverso (-1)	4,58
6	Gama	Inverso da Raiz Quadrada (-0,5)	3,88
7	Normal inversa	Identidade (1)	3,79
8	Lognormal	Raiz Quadrada (0,5)	2,92
9	Lognormal	Inverso da Raiz Quadrada (-0,5)	1,85
10	Normal	Inverso da Raiz Quadrada (-0,5)	1,53
11	Normal inversa	Raiz Quadrada (0,5)	1,21
12	Gama	Logaritmo Neperiano (0)	1,11
13	Lognormal	Identidade (1)	1,11
14	Lognormal	Inverso (-1)	-0,15
15	Gama	Raiz Quadrada (0,5)	-0,69
16	Gama	Identidade (1)	-2,63
17	Normal	Logaritmo Neperiano (0)	-3,30
18	Normal inversa	Inverso Quadrado (Canônica) (-2)	-4,64
19	Normal	Raiz Quadrada (0,5)	-8,46
20	Normal	Identidade (1)	-13,50

Na Tabela 4.12 apresentamos o valor da função de log-verossimilhança e do  $AIC_c$  para o modelo maior e para o modelo menor. As combinações estão ordenadas pela FLV do modelo menor.

**Tabela 4.12** - Função de Log-verossimilhança dos Modelos Maior e Menor.

	<b>Distribuição</b>	<b>Função de Ligação (<math>\lambda</math>)</b>	<b>Modelo maior</b>		<b>Modelo menor</b>	
			<b>FLV</b>	<b><math>AIC_c</math></b>	<b>FLV</b>	<b><math>AIC_c</math></b>
1	Normal inversa	Logaritmo Neperiano (0)	5,22	117,56	-5,84	33,01
2	Gama	Logaritmo Neperiano (0)	1,11	125,78	-8,09	37,51
3	Lognormal	Identidade (1)	1,11	125,78	-8,13	37,59
4	Lognormal	Raiz Quadrada (0,5)	2,92	122,15	-11,82	44,97
5	Normal inversa	Raiz Quadrada (0,5)	1,21	125,57	-12,38	46,10
6	Gama	Inverso da Raiz Quadrada (-0,5)	3,88	120,25	-14,74	50,81
7	Gama	Raiz Quadrada (0,5)	-0,69	129,38	-15,29	51,92
8	Normal	Logaritmo Neperiano (0)	-3,30	134,60	-15,46	52,26
9	Normal	Inverso da Raiz Quadrada (-0,5)	1,53	124,95	-16,20	53,74
10	Normal inversa	Identidade (1)	3,79	120,41	-17,46	56,26
11	Normal	Inverso (-1)	5,19	117,62	-18,44	58,22
12	Normal	Raiz Quadrada (0,5)	-8,46	144,92	-20,03	61,39
13	Lognormal	Inverso da Raiz Quadrada (-0,5)	1,85	124,29	-20,25	61,84
14	Gama	Inverso (-1)	5,22	117,56	-20,72	62,77
15	Gama	Identidade (1)	-2,63	133,25	-21,19	63,71
16	Normal inversa	Inverso da Raiz Quadrada (-0,5)	5,91	116,17	-22,24	65,81
17	Lognormal	Inverso (-1)	-0,15	128,30	-22,47	66,28
18	Normal inversa	Inverso (-1)	4,58	118,84	-23,24	67,81
19	Normal	Identidade (1)	-13,50	155,01	-27,59	76,51
20	Normal inversa	Inverso Quadrado (Canônica) (-2)	-4,64	137,29	-28,84	79,02

Os dois primeiros lugares para as distribuições gama e normal inversa, ambas com a função de ligação logarítmica, vêm confirmar a indicação do Estágio 2. Ademais, as funções de variância das distribuições gama e normal inversa crescem com o aumento da média, o que corrobora a indicação do gráfico da Figura 4.30 e do resultado do teste de Cook e Weisberg (1999), cujo resultado apresentamos na Seção 4.3.3. O valor do  $AIC_c$  também indica o modelo menor com a combinação 1.

Vamos para a Etapa 2 com a indicação da combinação 1 da Tabela 4.12.

#### 4.4.2. Etapa 2 – Teste de Significância dos Coeficientes: Análise de Deviance.

Nesta etapa o objetivo é selecionar os coeficientes em busca de um modelo. Esta seleção é feita com a análise de *deviance* (ANODE) (Seção 3.4.1).

##### Exemplo 4.2 (continuação)

Primeiramente vamos ajustar o modelo maior, inclusive com as interações, com a distribuição normal inversa e função de ligação logarítmica. Na Tabela 4.13 temos a ANODE. O modelo mais simples é aquele que tem apenas o termo constante. Começando com este modelo vamos adicionando os termos que mais contribuem para o decréscimo no desvio *deviance* (coluna 1). Esta seqüência e os respectivos *deviances* (coluna 3) são fornecidos pelo software *ARC*. Calculamos a diferença no *deviance* ocasionada pelo acréscimo de cada termo (coluna 4), a estatística de teste  $F_0$  (coluna 5) e o P-valor desta estatística (coluna 6).

**Tabela 4.13** - Análise de *Deviance* (ANODE).

-1-	-2-	-3-	-4-	-5-	-6-
<b>Termo</b>	<b>Graus de</b>	<b>Diferença no Estatística</b>			
<b>Adicionado</b>	<b>Liberdade</b>	<b>Deviance</b>	<b>Deviance</b>	<b><math>F_0</math></b>	<b>P-Valor</b>
Constante	15	1,42253			
X3	14	0,44612	0,97641	30,64	0,0001
X2	13	0,10876	0,33736	40,32	0,0000
X4	12	0,03386	0,07490	26,54	0,0002
X1	11	0,01610	0,01776	12,13	0,0051
X3X4	10	0,00799	0,00811	10,15	0,0097
X1X2	9	0,00604	0,00195	2,90	0,1226
X2X3	8	0,00491	0,00113	1,84	0,2119
X1X4	7	0,00405	0,00086	1,49	0,2623
X2X4	6	0,00404	0,00001	0,01	0,9070
X1X3	5	0,00404	0,00000	0,00	1,0000

Os termos considerados significativos pelo teste  $F$  são  $x_1, x_2, x_3, x_4, x_3x_4$ , todos com P-valor inferior a 5%. O modelo com distribuição normal inversa e função de ligação logarítmica é

$$\hat{\mu} = \exp(1,6038 + 0,0652x_1 + 0,3059x_2 + 0,5857x_3 + 0,1642x_4 + 0,0532x_3x_4).$$

A alternativa com o segundo menor valor da função de log-verossimilhança é o modelo com a distribuição gama e a função de ligação logarítmica.

Procedemos à ANODE para esta alternativa, que resultou no modelo

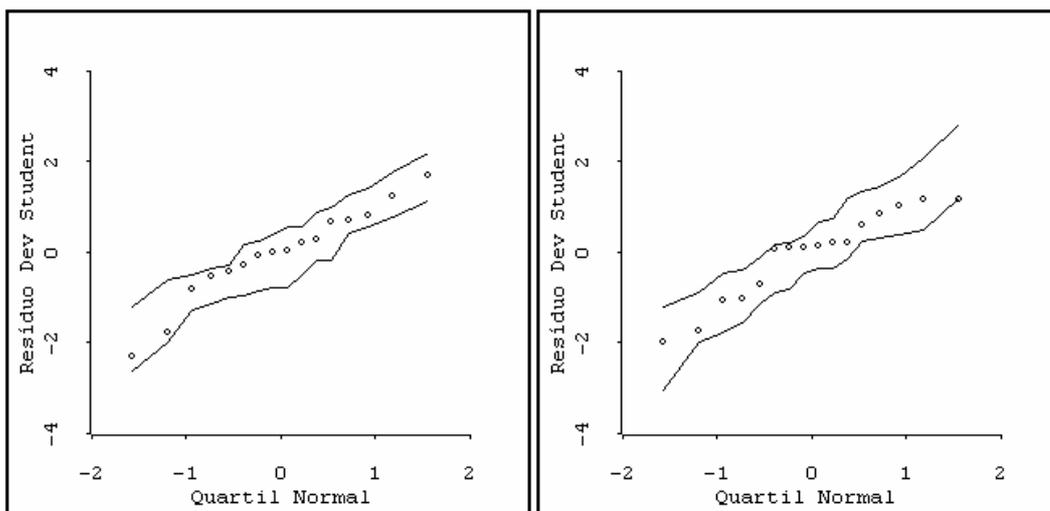
$$\hat{\mu} = \exp(1,5998 + 0,0643x_1 + 0,2914x_2 + 0,5787x_3 + 0,1635x_4 + 0,0492x_3x_4)$$

Comparando os dois modelos constatamos que são similares. Vamos proceder aos testes de adequação para cada um deles.

#### 4.4.3. Etapa 3 - Testes de Adequação dos Modelos

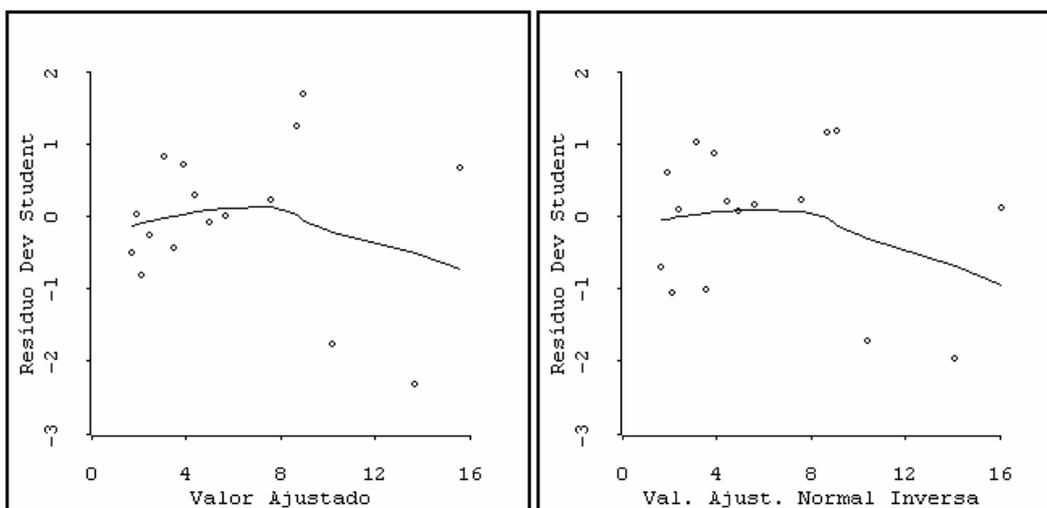
Para testar a adequação dos dois modelos são utilizados os gráficos dos resíduos *deviance*, descritos no Capítulo 3. Estes gráficos são apresentados nas Figuras 4.48 a 4.51. Para construção dos gráficos foi utilizado o software *ARC*.

Os gráficos de probabilidade normal com envelope na Figura 4.48 não apresentam muitos pontos fora do envelope e não observamos pontos muito fora do alinhamento. Por conseguinte, não há indicação de observações atípicas nem de que os modelos não sejam adequados.



**Figura 4.47** – Gráficos de Probabilidade Normal dos Resíduos *Deviance*. À Esquerda: Distribuição Gama. À Direita: Distribuição Normal Inversa.

Na Figura 4.49 temos os gráficos dos resíduos *deviance* studentizados *versus* valores ajustados.



**Figura 4.48** – Gráfico dos Resíduos *Deviance versus* Valores Ajustados. À Esquerda: Distribuição Gama. À Direita: Distribuição Normal Inversa.

Em ambos os gráficos a linha *lowess* apresenta uma curvatura para baixo; no entanto essa curvatura não constitui evidência forte para invalidar a função de ligação do modelo, se considerarmos que, em primeiro lugar, não é muita acentuada e, em segundo lugar, que um exame atento mostra que ela é ocasionada essencialmente pelo maior resíduo (em valor absoluto), numa região em que os pontos são mais esparsos e, portanto, a confiança na forma da linha *lowess* é menor.

Para verificar a função de variância dos dois modelos, temos na Figura 4.50 os gráficos do valor absoluto dos resíduos *deviance* studentizados *versus* valores ajustados (ver Capítulo 3).

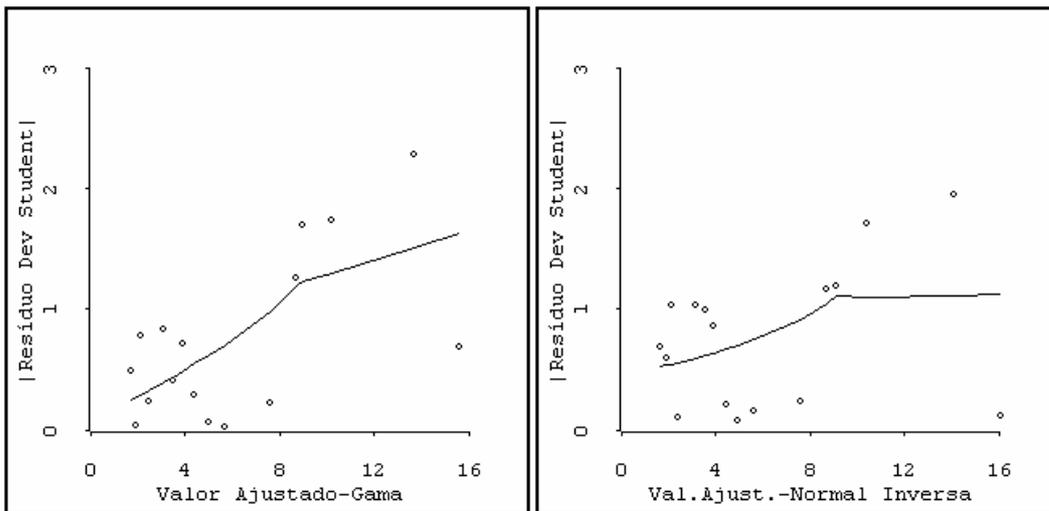
Como vimos na Seção 3.2, a função de variância geralmente é definida como uma função de potência da média:  $V(\mu) = \mu^\lambda$

Quando a linha *lowess* cresce sistematicamente com o aumento da média, indica que deve ser usado um maior valor para  $\lambda$  do que o valor correspondente à distribuição usada no modelo e quando decresce sistematicamente, indica que deve ser usado um menor valor para  $\lambda$ .

Para o gráfico do modelo com a *distribuição* gama (esquerda) a linha *lowess* apresenta um crescimento sistemático da esquerda para a direita. Isto indica que o expoente da média na função de variância deve ser maior do que o valor utilizado.

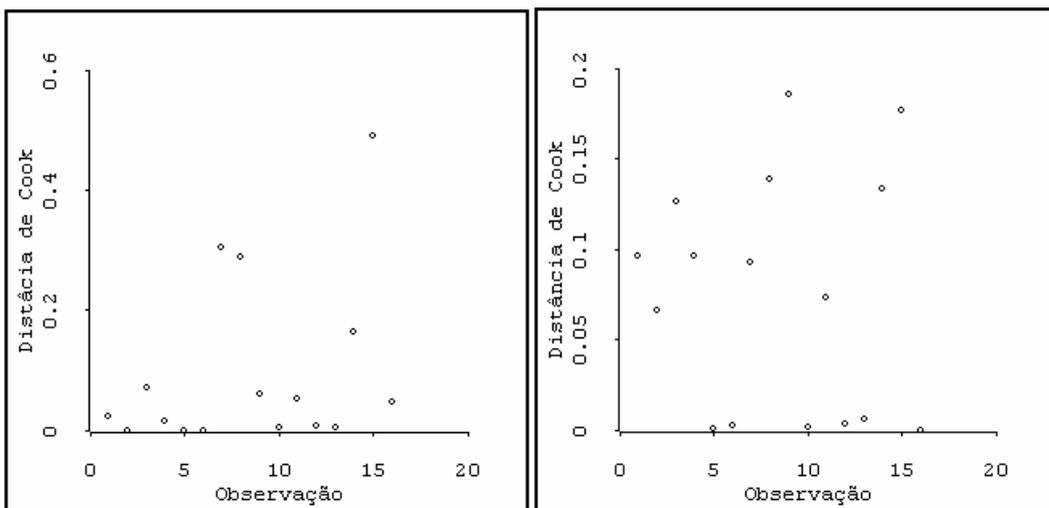
Para a distribuição gama a função de variância é  $V(\mu) = \mu^2$ . Portanto, o expoente deve ser maior do que 2.

Para o gráfico do modelo com a distribuição normal inversa (direita) a linha *lowess* não apresenta um crescimento sistemático da direita para a esquerda. Isto indica que a função de variância é correta. Como a função de variância da distribuição normal inversa é  $V(\mu) = \mu^3$ , temos a corroboração da indicação do gráfico da esquerda.



**Figura 4.49** - Gráfico dos Valores Absolutos dos Resíduos *Deviance Versus* Valores Ajustados. À Esquerda: Distribuição Gama. À Direita: Distribuição Normal Inversa.

Para verificar se temos indicação de observações influentes, apresentamos os gráficos da Figura 4.51, onde temos as distâncias de Cook.



**Figura 4.50** – Gráficos com as Distâncias de Cook. À Esquerda: Distribuição Gama. À Direita: Distribuição Normal Inversa.

No gráfico da esquerda o valor da observação 15 é 0,4928. Como este valor está muito próximo de 0,5, indica que esta pode ser uma observação influente. Entretanto, como a função de variância não se mostrou adequada, é plausível que esta seja a razão dessa indicação. No gráfico da direita não observamos valor superior a 0,5. Portanto, não temos indicação de observação influente para o modelo com a distribuição normal inversa.

Concluimos que o modelo com distribuição normal inversa e função de ligação logarítmica

$$\hat{\mu} = \exp(1,6038 + 0,0652x_1 + 0,3059x_2 + 0,5857x_3 + 0,1642x_4 + 0,0532x_3x_4)$$

é um modelo adequado para os dados apresentados no experimento da perfuratriz.

Cabe lembrar que, segundo Daniel (1976), os fatores A ( $x_1$ ), B ( $x_2$ ) e C ( $x_3$ ) são sabidamente importantes. Na sua análise ele usou várias transformações, concluindo que os efeitos dos fatores principais A, B, C e D ( $x_4$ ) são importantes, ficando em dúvida sobre a inclusão das interações BC ( $x_2 x_3$ ) e CD ( $x_3 x_4$ ). Suas conclusões não foram definitivas, sugerindo a obtenção de mais dados.

Myers *et al.* (2002) e Lewis *et al.* (2001a) usaram um modelo com distribuição gama e função de ligação logarítmica para ajustar esses dados, concluindo que apenas os efeitos dos fatores principais B, C e D são importantes.

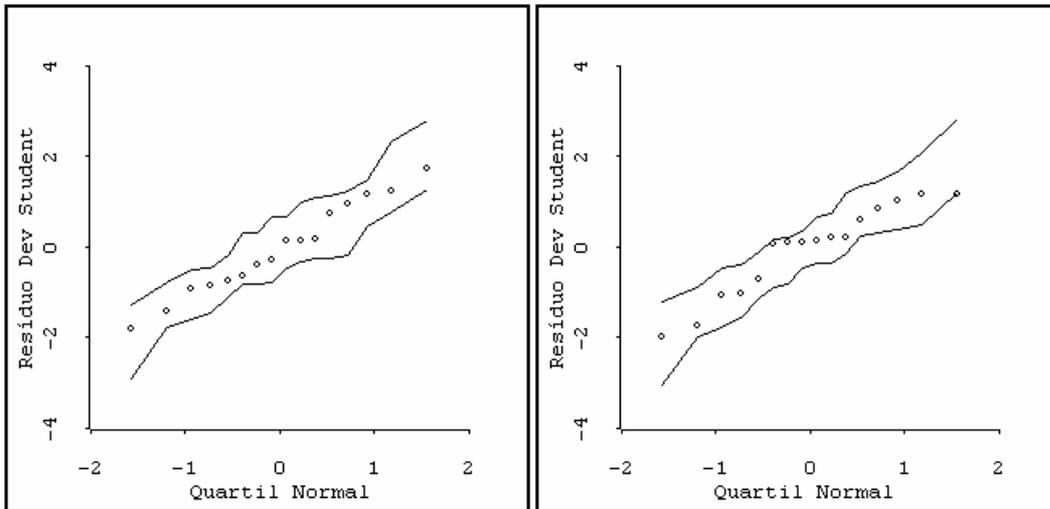
A nossa análise mostra que o modelo com distribuição normal inversa e função de ligação logarítmica é melhor, e que o fator A e a interação CD também são importantes.

Embora consideremos o modelo ajustado como sendo adequado para o experimento da perfuratriz, constatamos alguma imperfeição no gráfico da direita na Figura 4.49, mesmo que não haja indicação de que a função de ligação seja incorreta. Entretanto, pode ser que a interação  $x_3 x_4$  não seja realmente significativa, correspondendo a um erro do teste de hipótese na análise de *deviance*. Vamos ajustar um modelo sem esta interação. O modelo com distribuição normal inversa e função de ligação logarítmica e sem a interação  $x_3 x_4$  fornecido pelo ARC é

$$\hat{\mu} = \exp(1,6015 + 0,0667x_1 + 0,3047x_2 + 0,5799x_3 + 0,1365x_4).$$

A título de comparação, vamos repetir, ao lado dos gráficos que testam a adequação deste novo modelo (esquerda), os respectivos gráficos do modelo com

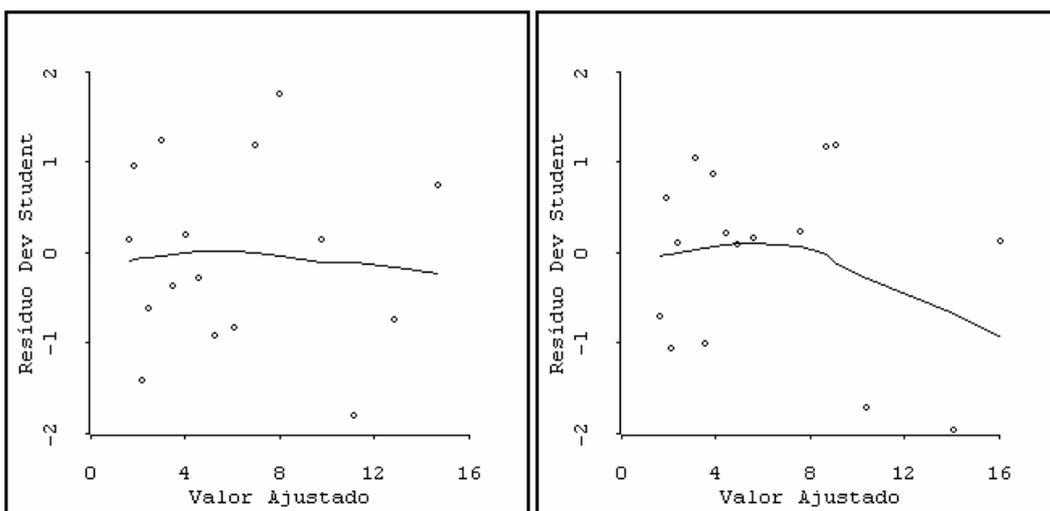
a interação (direita). Na Figura 4.52 apresentamos os gráficos de probabilidade dos resíduos *deviance studentizados*.



**Figura 4.51** – Gráficos de Probabilidade Normal dos Resíduos *Deviance*. À Esquerda: Sem Interação. À Direita: Com Interação.

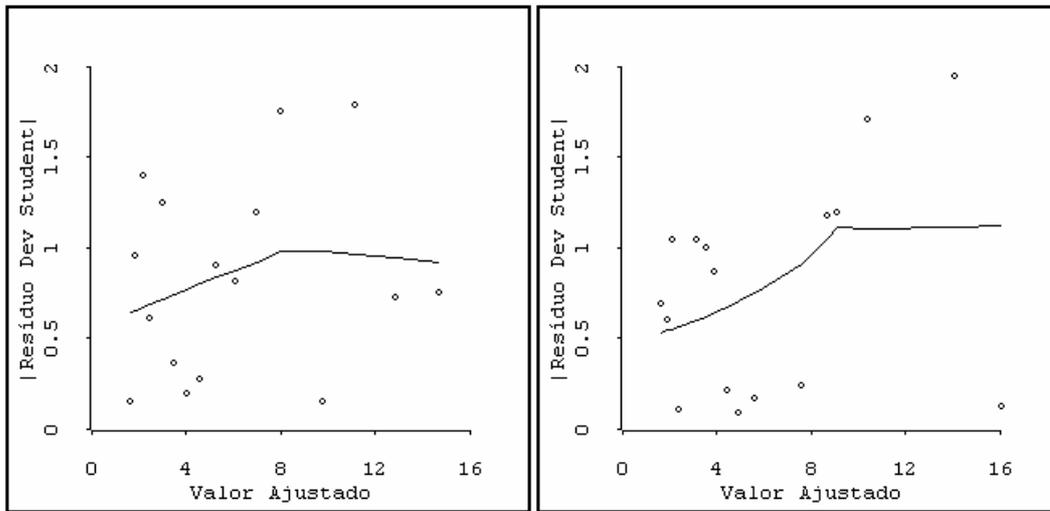
Estes gráficos não apresentam pontos fora do envelope. Entretanto, no gráfico da esquerda há um maior alinhamento e uma maior continuidade nos pontos, embora esses fatos não sejam decisivos para a escolha do modelo.

Na Figura 4.53 temos os gráficos dos resíduos *deviance* studentizados *versus* valores ajustados. Os gráficos não apresentam um perfil que invalide a função de ligação (ambas são logarítmicas). Entretanto, no gráfico da esquerda, a linha de amortecimento é aproximadamente horizontal e próxima da reta horizontal de ordenada zero, o que não acontece no gráfico da direita.



**Figura 4.52** – Gráficos dos Resíduos *Deviance Versus* Valores Ajustados À Esquerda: Sem Interação. À Direita: Com Interação.

Para verificar a função de variância dos dois modelos, temos, na Figura 4.54, os gráficos do valor absoluto dos resíduos *deviance* studentizados *versus* valores ajustados.



**Figura 4.53** - Gráfico do Valor Absoluto do Resíduo *Deviance Versus* Valores Ajustados. À Esquerda: Sem Interação. À Direita: Com Interação.

Em ambos os gráficos, a linha resultante do amortecimento não apresenta crescimento nem decrescimento sistemático da esquerda para a direita. Isso indica que a função de variância (a mesma para os dois modelos) é correta. Entretanto, no gráfico da esquerda, a linha de amortecimento é aproximadamente horizontal, o que não acontece no gráfico da direita.

Não chegamos a uma conclusão definitiva sobre o melhor modelo. Por um lado temos um teste formal (ANODE) que sugere a inclusão da interação no modelo. Por outro lado os resíduos do modelo sem interação têm um comportamento melhor do que o com interação.

Para os dois modelos, a taxa máxima de avanço ocorre quando  $x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1$ . Ou seja no ponto  $\mathbf{x}'_0 = [1 \ 1 \ 1 \ 1]$ .

Por conseguinte, independentemente de qual seja o modelo mais adequado, a otimização do desempenho da perfuratriz ocorre neste ponto. Entretanto, caso seja necessária uma estimativa da resposta neste ponto, dados adicionais são necessários.

A seguir apresentamos a estimativa da média da resposta para este ponto e o erro-padrão da média da resposta, fornecido pelo software *ARC*.

Modelo	Estimativa da Média	Erro-Padrão	Intervalo de Confiança (95%)
Com Interação	16,09	0,8323	[14,45 – 17,72]
Sem Interação	14,72	0,8200	[13,12 – 16,33]

O intervalo de confiança da média da resposta no ponto  $\mathbf{x}_0$  é calculado pela expressão abaixo.

$$\hat{\mu}(\mathbf{x}_0) - z_{\alpha/2}(\text{erro - padrão}) \leq \mu_{y|\mathbf{x}_0} \leq \hat{\mu}(\mathbf{x}_0) + z_{\alpha/2}(\text{erro - padrão}).$$

### Exemplo 3.1 (continuação)

Os dados do experimento encontram-se na Tabela 3.3. A distribuição de probabilidade que gerou os dados foi a gama com parâmetro de dispersão  $\phi = 0,01$ , e a função de ligação foi a função inverso.

O modelo sem as interações (modelo menor) é

$$\mu_i^\lambda = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4).$$

Na Tabela 4.14 apresentamos o valor da função de log-verossimilhança para o modelo menor. As combinações estão ordenadas pela FLV do modelo menor.

**Tabela 4.14** - Valor da Função de Log-verossimilhança do Modelo Menor.

Distribuição	Função de Ligação ( $\lambda$ )	Modelo Menor
		FLV
Gama	Logaritmo Neperiano (0)	-42,67
Lognormal	Identidade (1)	-42,80
Lognormal	Raiz Quadrada (0,5)	-42,91
Normal	Logaritmo Neperiano (0)	-43,48
Normal inversa	Logaritmo Neperiano (0)	-43,70
Lognormal	Inverso da Raiz Quadrada (-0,5)	-44,19
Normal	Raiz Quadrada (0,5)	-44,48
Gama	Inverso da Raiz Quadrada (-0,5)	-44,58
Normal	Inverso da Raiz Quadrada (-0,5)	-45,12
Lognormal	Inverso (-1)	-45,14
Normal	Inverso (-1)	-47,30
Gama	Inverso (-1)	-47,94
Normal inversa	Identidade (1)	-48,11
Gama	Identidade (1)	-48,20
Normal inversa	Inverso Quadrado (Canônica) (-2)	-48,21
Normal	Identidade (1)	-48,63
Normal inversa	Inverso da Raiz Quadrada (-0,5)	-53,68
Gama	Raiz Quadrada (0,5)	-56,95
Normal inversa	Raiz Quadrada (0,5)	-57,18
Normal inversa	Inverso (-1)	-60,10

A combinação que está em primeiro lugar é a formada pela distribuição gama com função de ligação logarítmica. O modelo que sabemos ser correto (gama-inverso) aparece em 12º lugar. A distribuição indicada é a que sabemos ser a correta, mas a função de ligação não o é.

Vamos prosseguir a análise com a indicação incorreta obtida.

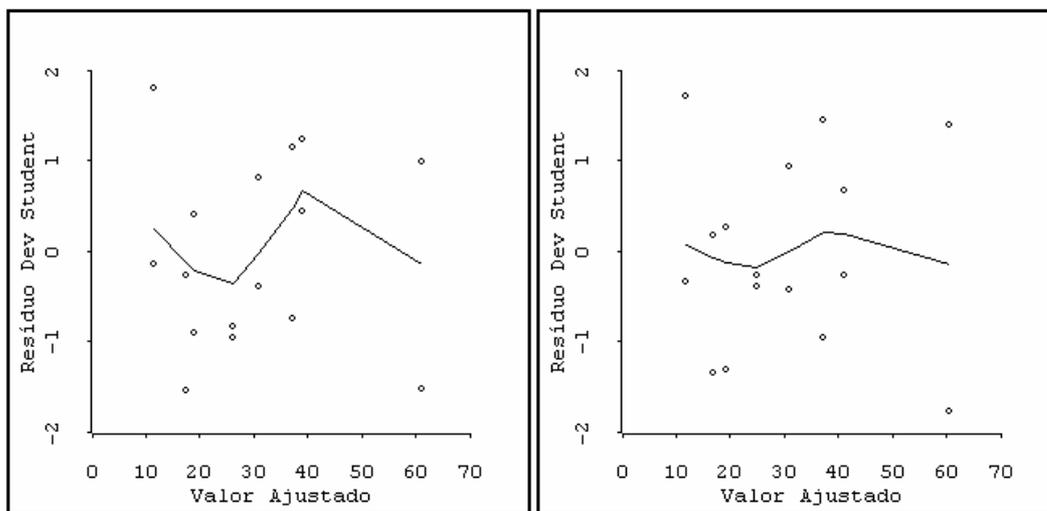
Procedemos à análise de *deviance* e o modelo escolhido foi:

$$\hat{\mu} = \exp(3,2929 - 0,23303x_1 - 0,2704x_2 - 0,3243x_3 + 0,00696x_1x_2 - 0,0788x_1x_3).$$

Observamos que foi selecionada interação  $x_1x_2$ , que sabemos não ser significativa, o que não aconteceu com  $x_2x_3$ , que sabemos ser significativa.

Os gráficos de verificação da adequação deste modelo nada indicaram que invalidasse o modelo. Entretanto, comparando os gráficos dos resíduos *deviance* studentizados *versus* valores ajustados, com o mesmo gráfico do modelo escolhido na Seção 3.4.1, temos indicação de que aquele deve ser preferido.

Na Figura 4.55 temos, para o modelo atual e para o modelo da Seção 3.4.1, os gráficos dos resíduos *deviance* studentizados *versus* valores ajustados.



**Figura 4.54** – Gráficos dos Resíduos *Deviance Versus* Valores Ajustados: À Esquerda: Função de Ligação Logarítmica. À Direita: Função de Ligação Inverso.

Observando os gráficos, damos preferência ao gráfico correspondente à função de ligação inverso. Isto porque, neste gráfico, a linha de amortecimento é mais próxima da reta horizontal de ordenada zero do que o gráfico correspondente à função de ligação logarítmica. Portanto, escolheríamos o modelo com a função de ligação inverso.

Na Seção 4.4.1 propusemos um método para a escolha da distribuição de probabilidade e da função de ligação. Tal escolha seria feita pela FLV do modelo sem as interações (modelo menor). Para identificação destas funções o método não funcionou bem.

É verdade que indicou a distribuição de probabilidade correta e que no final das contas tivemos indicação do modelo adequado, o que, todavia, pode ter sido fortuito. Entretanto, a FLV não indicou corretamente a função de ligação.

É plausível que esta indicação incorreta seja devido às interações significativas que são omitidas no modelo menor, e, por conseguinte, este método não funciona bem na presença de interações importantes.

Em certas situações, sabem-se de antemão os fatores que interagem e que não interagem. Nesses casos, a FLV do modelo com as interações importantes pode indicar corretamente a distribuição de probabilidade e da função de ligação.

### Exemplo 3.1 (continuação)

Suponha que soubéssemos que as interações  $x_{213}$  e  $x_2x_3$  são importantes. Teríamos o modelo:

$$\hat{\mu}_i^\lambda = (\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_{13}x_1x_3 + \beta_{23}x_2x_3)$$

Na Tabela 4.16 apresentamos o valor da função de log-verossimilhança para as quatro melhores combinações.

**Tabela 4.15** - Valor da Função de Log-verossimilhança.

Distribuição	Função de Ligação ( $\lambda$ )	FLV
Gama	Inverso (-1)	-32.20
Normal inversa	Inverso do Quadrado (Canônica) (-2)	-33.18
Normal inversa	Inverso (-1)	-33.29
Lognormal	Inverso (-1)	-33.65

A combinação que está em primeiro lugar é a formada pela distribuição gama com função de ligação inverso, que sabemos ser a indicação correta.

Em certas situações, sabem-se de antemão que os fatores não interagem. Nesses casos, a FLV do modelo sem interações importantes pode indicar corretamente a distribuição de probabilidade e a função de ligação.

**Exemplo 3.1 Modificado (continuação)**

Na Tabela 4.10 apresentamos os dados de um experimento que foi gerado a partir do modelo

$$\mu_i = \frac{1}{0,040 + 0,008x_1 + 0,010x_2 + 0,012x_3}$$

A distribuição de probabilidade escolhida foi gama com parâmetro de dispersão  $\phi = 0,01$ .

Na Tabela 4.16 apresentamos o valor da função de log-verossimilhança para as quatro melhores combinações.

**Tabela 4.15** - Valor da Função de Log-verossimilhança.

Distribuição	Função de Ligação ( $\lambda$ )	FLV
Normal	Inverso (-1)	-32.30
Gama	Inverso (-1)	-32.98
Normal inversa	Inverso (-1)	-34.21
Lognormal	Inverso (-1)	-36.50

Em primeiro lugar aparece a distribuição normal com função de ligação inversa. A opção gama com inverso, que sabemos ser a verdadeira, aparece em segundo. Observamos que as quatro primeiras tem função de ligação inverso. Vamos prosseguir com os modelos das duas primeiras indicações da tabela.

O software *ARC* forneceu os dois modelos:

Modelo Normal-Inverso:

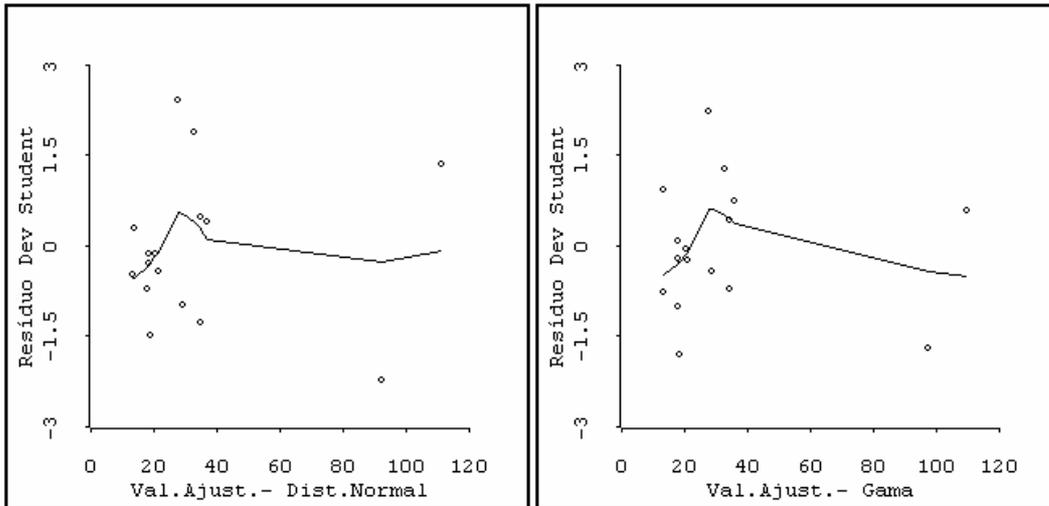
$$\mu_i = \frac{1}{0,0410 + 0,0089x_1 + 0,0098x_2 + 0,0124x_3 + 0,0009x_4}$$

Modelo Gama-Inverso:

$$\mu_i = \frac{1}{0,0417 + 0,0093x_1 + 0,0099x_2 + 0,0128x_3 + 0,0006x_4}$$

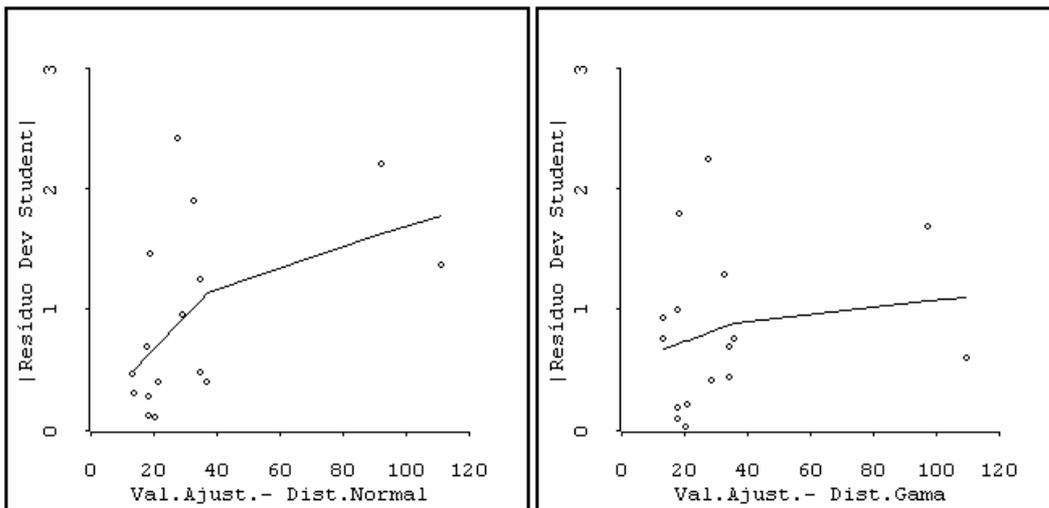
Os modelos são bastante parecidos. Praticamente fornecem as mesmas estimativas da média.

Na Figura 4.56 temos, para estes dois modelos, os gráficos dos resíduos *deviance* studentizados *versus* valores ajustados. Os gráficos não apresentam um perfil que indique problemas com a função de ligação (inverso em ambos os casos).



**Figura 4.55** – Gráficos dos Resíduos *Deviance Versus* Valores Ajustados: À Esquerda: Distribuição Normal. À Direita: Distribuição Gama.

Para verificar a função de variância dos dois modelos, temos, na Figura 4.57, os gráficos dos valores absolutos dos resíduos *deviance* studentizados *versus* valores ajustados.



**Figura 4.56** - Gráfico dos Valores Absolutos do Resíduos *Deviance Versus* Valores Ajustados. À Esquerda: Distribuição Normal. À Direita: Distribuição Gama.

Para o gráfico do modelo com a distribuição normal (esquerda) a linha *lowess* apresenta um crescimento sistemático da esquerda para a direita. Isto indica que o expoente da média na função de variância deve ser maior do que o valor utilizado. Para a distribuição normal a função de variância é constante. Portanto, temos indicação de que a variância cresce com o aumento da média.

Para o gráfico do modelo com a distribuição gama (direita) a linha *lowess* apresenta um discreto crescimento sistemático da direita para a esquerda, que não invalida a função de variância.

Na Seção 4.4.1 propusemos um método para a escolha da distribuição de probabilidade e da função de ligação. Tal escolha seria feita pela FLV do modelo sem as interações (modelo menor). Em alguns casos o método funcionou bem e em outros não.

Casos em que o método funcionou bem:

- ◆ Quando não houve interação significativa entre os fatores (Exemplo 3.1 modificado)
- ◆ Quando a interação entre os fatores, embora significativa, é fraca (Exemplo 4.2)
- ◆ Quando conhecemos os fatores que interagem e os que não interagem incluímos as interações conhecidas no modelo para o cálculo da FLV. (Exemplo 3.1).

Caso em que o método não funcionou bem:

- ◆ Quando interações entre os fatores são significativas e com efeitos grandes (Exemplo 3.1).

Nos casos em que o método não funciona bem, a distribuição de probabilidade (e a função de ligação) devem ser determinadas por tentativa, usando os testes de adequação do modelo para a decisão definitiva. O número de alternativas a testar não é necessariamente muito grande, pois pode-se dispor de informações que eliminem várias distribuições. Por exemplo, se houver indicação sobre o comportamento da função de variância, consideram-se somente as distribuições de probabilidade cuja função de variância seja compatível com esse comportamento.