IV - TESTES DE AJUSTE

ou

TENTATIVAS DE INVALIDAÇÃO

IV.1 INTRODUÇÃO

Weisberg (1983): "O comportamento de um procedimento de diagnóstico tem que ser conhecido, ao menos aproxidamente, tanto para o modelo correto quanto para o modelo incorreto com alguma hipótese adicional modificada. Os métodos de diagnóstico não devem ser computacionalmente intensivos e de preferencia ter sempre um equivalente gráfico e sugerir alguma ação alternativa".

Box: "Nenhum modelo é bom, o importante é que ele seja útil".

Karl Popper, em interpretação livre: "não existem meios de se provar a validade de um modelo, existem apenas meios para provar a sua não validade. Quando da proposição de um modelo científico o único mecanismo para prová-lo é tentar de todo modo invalidá-lo". Neste sentido, basta uma evidência física, e apenas uma, para pôr em cheque toda uma teoria.

Stéphane Mallarmé: "Un coup de dés, quand même bien lancé dans des conditions idéales, jamais n´abolira le hasard".

Não existe um conjunto de testes padrão, adotados como dogmas para a "validação" ou invalidação de um modelo. Diferentes autores propõem, ou adotam, diferentes conjuntos de diagnósticos. Não existe um número mágico que diga, de maneira biunívoca e incontestável, se um modelo é adequado para representar um determinado fenômeno, ou conjunto de fenômenos. É importante desde o início abandonar esta esperança e entender que os procedimentos de diagnóstico têm que ser os mais variados, e isto sim, têm que ser corretamente interpretados para que não se diagnostique de maneira errônea um determinado resultado.

Os diagnósticos de modelos são baseados no resultado do ajuste do modelo, que pode ser verificado através:

- da comparação entre as predições obtidas através do modelo ajustado, com os dados experimentais;
- da análise de inferência sobre os parâmetros (com a qual, eu pessoalmente não concordo).

A finalidade dos diagnósticos é a de realimentar o procedimento de identificação fornecendo pistas sobre eventuais modificações que abrangem:

1) a estrutura do modelo determinístico;
2) a estrutura do modelo probabilístico;
3) os dados experimentais.

Na figura 1 apresentamos uma representação livre do procedimento de estimação de parâmetros.

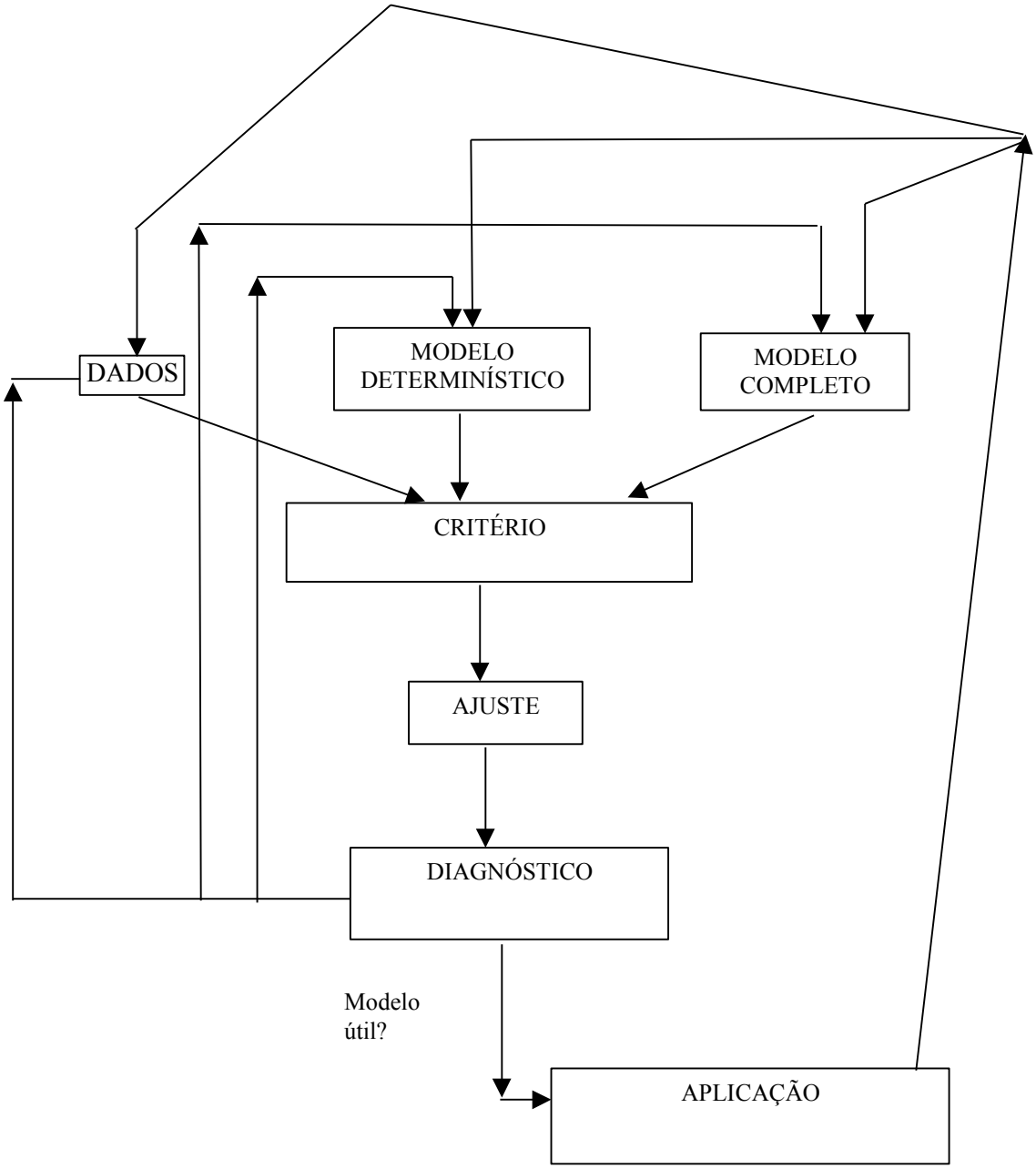

Figura 1 - Representação livre do procedimento de identificação

IV.2 PATOLOGOGIAS A DETECTAR

A seguir enumeramos as principais patologias a serem evitadas quando do diagnóstico do ajuste. O objetivo de um ajuste "ideal" é propor um ajuste que permita distinguir "ao ruído o que é do ruído e ao determinístico o que é determinístico", e que além disto, descreva o ruído corretamente.

IV.2.1) Falta de ajuste

É importante detectar se toda a informação determinística foi extraída durante a estimação. Para tal no ruído, ou no que é considerado como uma das suas manifestações, o resíduo, não deve sobrar nada de determinístico. Se a informação que corresponde à parte determinística do processo não for totalmente extraída, as predições obtidas com o modelo apresentarão um erro determinístico, além do erro aleatório intrínseco ao processo. Por exemplo, caso um sistema apresente exatamente um comportamento polinomial de ordem dois, se um modelo de ordem abaixo de dois for ajustado aos dados será obtida uma falta de ajuste.

IV.2.2) Sobreajuste

O sobreajuste tem um efeito contrário à falta de ajuste. No sobreajuste os componentes determinísticos são usados para ajustar uma realização particular do ruído. Seguindo o mesmo exemplo anteriormente utilizado, o sobreajuste seria obtido quando um modelo de ordem superior a três fosse usado para modelar um sistema que apresente exatamente um comportamento polinomial de ordem dois. Neste caso, o sobreajuste causará problemas de predição, ou mais exatamente, de extrapolação.

Na verdade como na maioria dos casos não se conhece exatamente a forma do modelo *a priori*, como no exemplo que foi proposto, e portanto o limite entre a falta de ajuste e o sobreajuste é difícil de ser detectado.

IV.2.3) Hipóteses do ruído

A inferência depende da proposição adequada de uma distribuição de amostragem. A forma da distribuição de amostragem depende da verificação das hipóteses sobre o ruído. Por exemplo, se dizemos que uma estatística se distribui segundo

uma distribuição de Student, isto depende de por exemplo, o ruído ser normal. Se o ruído não é normal a distribuição de amostragem de Student não corresponde.

Da mesma forma, um ajuste mais adequado pode ser obtido com a proposição de um critério mais adequado a descrever o ruído.

IV.2.4) Problemas numéricos de ajuste

Os problemas numéricos para se obter um determinado ajuste de um modelo não fazem parte das patologias de um modelos, apesar de muitos autores confundirem as origens destes problemas com o de diagnóstico de problemas de identificação, estes são de outra ordem e portanto têm que ser tratados de outra forma, conforme será visto em outro tópico.

IV.3 - INFERÊNCIA

IV.3.1 Teste de Hipóteses

Os testes de hipóteses são processos de decisão utilizados para, baseado no resultado de um sorteio aleatório, por exemplo uma estimativa particular ($\hat{\theta}$) de uma determinada variável, θ decidir sobre a variável θ. Como todo processo decisório este procedimento envolve um compromisso. A seguir apresentamos simplificadamente a maneira como este processo é apresentado em De Groot (1975). Um teste de hipóteses é formalizado ao se dividir o universo em dois conjuntos disjuntos Ω_o e Ω_1 de tal forma a que:

$\theta \in \Omega_o \Rightarrow \theta \notin \Omega_1$

$\theta \in \Omega_1 \Rightarrow \theta \notin \Omega_o$

$\Omega_o \cup \Omega_1 = \Omega$

As hipóteses são formalizadas em termos destes conjuntos:

$H_o = \theta \in \Omega_o$ (em geral denominada hipótese fundamental)

$H_1 = \theta \in \Omega_1$ (denominada hipótese alternativa)

O procedimento para tomada de decisões é baseado na definição de uma região crítica, c, do espaço amostral. A região crítica é portanto uma região do espaço das variáveis X_1, X_2, ..., X_n. Pensemos por exemplo no caso de uma média aritmética, ao definirmos uma região para a média, isto é:

$$b \leq \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \leq a$$

estamos de fato definindo uma região para as variáveis X_1, X_2, ..., X_n. O procedimento decisório proposto é simplesmente traduzido em: "se o resultado de um sorteio, ou bem, se a realização de um conjunto de processos aleatórios, estiver na região c, então rejeita-se a hipótese H_o, caso contrário aceita-se a hipótese".

Existe uma uma probabilidade para que o resultado de um sorteio, ou bem, para que a realização de um conjunto de processos aleatórios esteja na região c, para cada diferente valor de θ. A partir desta probabilidade (e do conjunto crítico c) é definida a "função de potência" do teste, que é uma função que associa a todo valor de θ uma probabilidade (que não é a probabilidade de θ, é claro), mas sim a probabilidade que X_1, X_2, ..., X_n pertençam a c dado que q é q:

$\Pi(\theta) = P(X_1, X_2, ..., X_n \in c \mid \theta)$

Antes que as coisas comecem a degringolar vamos a um exemplo prático. Vamos supor um evento composto por 10 dez sorteios independentes de uma variável aleatória normal com média zero e variância 1. A média destes dez sorteios é uma variável aleatória com média zero e variância 1/10 (ver qualquer livro de estatística sobre a distribuição de amostragem, "sampling distribution", da média de uma população normal). Definamos a região crítica como sendo a região em que:

$\bar{x} \le -1$ ou $1 \le \bar{x}$

Vamos calcular alguns valores da função de potência de μ. Se μ vale zero, a probabilidade de obter um resultado menor que -1 ou maior que 1 é dada por (para uma variável normal com média zero e variância 1/10):

$P(\bar{x} \le -1 \text{ ou } 1 \le \bar{x}) = 0.0016$

Se μ valer 0.5, a probabilidade de obter um resultado menor que -1 ou maior que 1 é dada por (para uma variável normal com média 0.5 e variância 1/10):

$P(\bar{x} \le -1 \text{ ou } 1 \le \bar{x}) = 0.0569$

Podemos construir o gráfico de $\Pi(\mu)$:

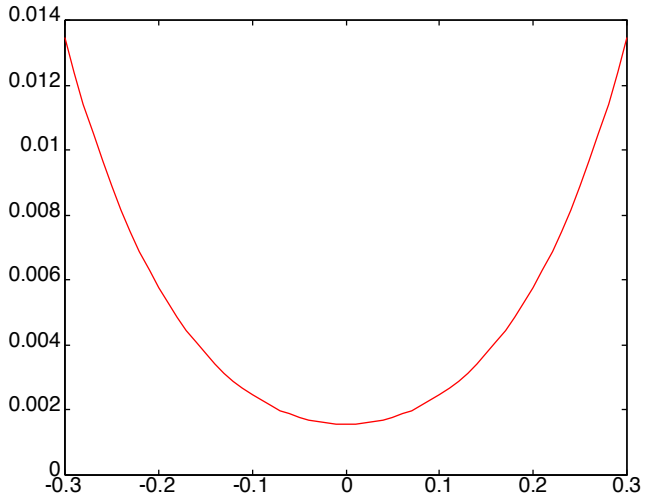

Figura 1 - $\Pi(\mu)$

Voltando portanto à pergunta: "o que isto tem a ver com Ω_o e Ω_1 ?" Define-se o tamanho de um teste por:

$$\alpha = \max_{\theta \in \Omega_0} \Pi(\theta)$$

Um teste portanto é baseado em uma região crítica (c) e um tamanho. O conjunto Ω_0 define a hipótese que se quer testar. Por exemplo:

- se $\Omega_o = \{0\}$ para a região crítica definida acima o tamanho é $\Pi(0) = 0.0016$
- se $\Omega_o = [-0.1, 0.1]$ para a região crítica definida acima o tamanho é $\Pi(0.1) = 0.0025$, pois é o maior valor de θ dentro do intervalo Ω_0.

O tamanho é portanto a probabilidade máxima de termos um resultado dentro da região crítica sendo que a hipótese é verdadeira. Ou seja, que a hipótese seja verdadeira mas que o resultado indica que ela deve ser rejeitada. A isto se dá o nome de erro tipo I (rejeitar a hipótese pois o resultado se encontra na região crítica apesar de $\theta \in \Omega_o$). Através da aplicação do procedimento de decisão estabelece-se uma probabilidade máxima de cometer este tipo de erro.

Um procedimento de decisão é portanto baseado na escolha de um região crítica e de um tamanho do teste. Suponhamos que $\Omega_o = \{0\}$ (hipótese simples) a região crítica que faria com que a probabilidade de cometer um erro tipo I fosse zero teria que ser $]-\infty,+\infty[$. Neste caso jamais seria cometido um erro tipo I, ou seja rejeitar a hipótese, sendo ela verdadeira. No entanto, a probabilidade de cometer um erro tipo II, aceitar a hipótese sendo ela errada, seria muito elevada.

A esta probabilidade dá-se o nome de β. Ela é muito difícil de avaliar. Um artigo muito interessante é o de Buzzi Ferraris (em anexo).

IV.4 RESÍDUOS

IV.4.1 Definição

Os resíduos ordinários são definidos como:

$$\boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} \qquad (1)$$

onde $\hat{\mathbf{y}}$ são os valores da variável dependente obtida a partir do modelo ajustado. A análise dos resíduos é interessante pois os mesmos podem ser vistos como uma realização particular do suposto ruído do sistema. No caso de regressão linear, pode-se escrever:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\,\mathbf{y} \qquad (2)$$

H é a famosa matriz-chapéu (hat matrix), pois ela põe o chapéu em **y**. Portanto, para modelos lineares em relação aos parâmetros:

$$\boldsymbol{\varepsilon} = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{y} \qquad (3)$$

Estes serão utilizados para estimar propriedades

Os resíduos ordinários tem características um pouco estranhas pois sobretudo com relação à variância, já que:

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \qquad (4)$$

$$\mathrm{V}(\boldsymbol{\varepsilon}) = (\mathbf{I} - \mathbf{H})\,\sigma^2 \qquad (5)$$

onde σ^2 é a variância do ruído do modelo. Em consequência, os resíduos ordinários não são independentes. Definem-se diferentes tipos de estatísticas calculadas a partir dos resíduos, que são acessórios ao diagnóstico.

Definido-se a média dos resíduos:

$$\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \qquad (6)$$

e a variância calculada a partir dos resíduos (que sob algumas condições é uma estimativa não tendenciosa de σ^2):

$$s^2 = \frac{1}{(n-p)}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 \qquad (7)$$

onde n é o número total de pontos e p o número de parâmetros. Definem-se os resíduos normalizados:

$$\boldsymbol{\varepsilon}^{*} = \frac{\left(\boldsymbol{\varepsilon} - \mathbf{I}\bar{\varepsilon}\right)}{\sqrt{s^2}} \qquad (8)$$

Para modelos não lineares nos parâmetros eles devem ser obtidos calculando a média dos resíduos, subtraindo-a de cada resíduo e dividindo o produto pela variância estimada a partir da fórmula não tendenciosa. Infelizmente, apesar do nome, os resíduos normalizados não são nem independentes (mesmo que **e** o fosse) nem tem variância unitária.

Definem-se os resíduos Studentizados (na verdade normalizados), que têm variância unitária e média nula, através de:

$$t_i = \frac{\varepsilon_i}{\sqrt{s^2}\sqrt{(1 - h_{ii})}} \qquad (9)$$

h_{ii} são os elementos da diagonal de **H** definido na equação (2). Estes resíduos só podem ser calculados para regressões lineares.

IV.4.2 Análise Gráfica Comparativa

As diferentes hipóteses sobre o ruído são em geral testadas utilizando-se os resíduos. A independência entre os resíduos e as variáveis independentes pode ser rapidamente intuída através de gráficos de $\boldsymbol{\varepsilon}$ em função de cada uma das variáveis independentes $\mathbf{x}_i$. Isto pode levar à conclusões de falta de ajuste (insuficiência de detalhes no modelo determinístico). Os gráficos podem ser feitos com $\boldsymbol{\varepsilon}^{*}$ ou **t**.

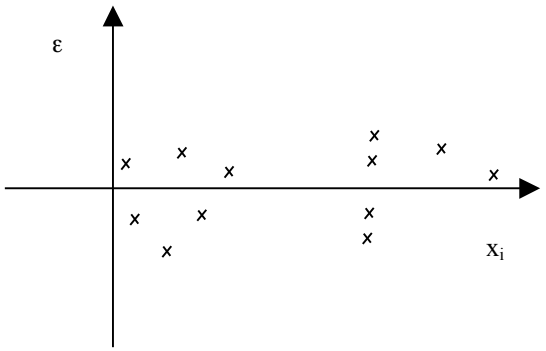

figura 2 - aspecto não patológico de um gráfico de resíduo em função de uma variável independente.

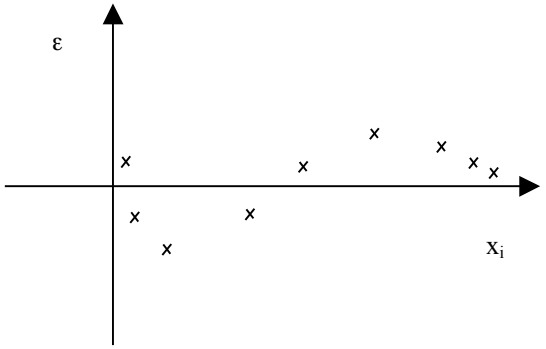

Figura 3 - possível falta de ajuste

Outro gráfico bastante informativo é o do resíduo em função das variáveis dependentes (simuladas ou medidas) que podem também revelar problemas como os ilustrados nas figuras 2 e 3. Outro problema que pode aparecer nestes gráficos é da heterocedacidade, tal como apresentado na figura 4. Em geral faz-se a hipótese que a variância é constante (homocedacidade), isto é, não depende de y, mas às vezes isto não é verdade e podem aparecer dependências tais como:

$$\sigma^2 = a \mid y \mid^b \qquad (10)$$

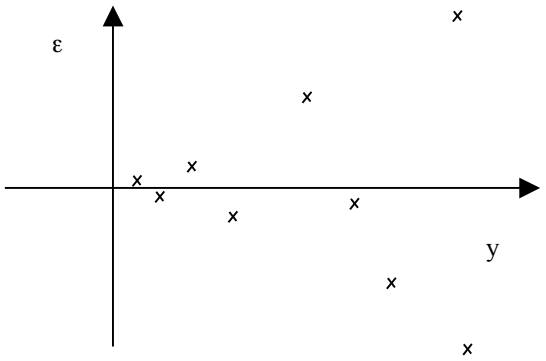

Figura 4 - possível problema de heterocedacidade

Para problemas no tempo, em que é formulada a hipótese de ruído branco, pode-se fazer um gráfico do resíduo em instantes diferentes, tal como aparece nas figuras 5 e 6.

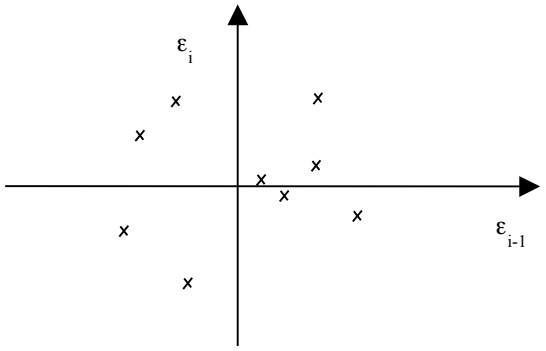

Figura 5 - resíduos sem correlação aparente no tempo

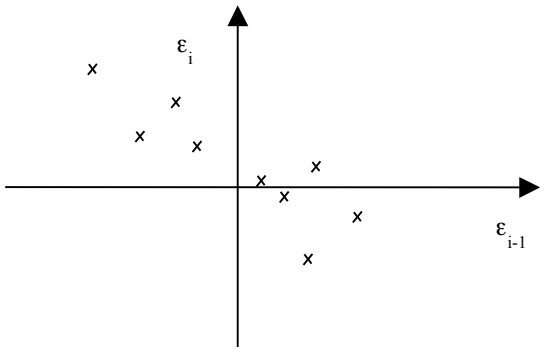

Figura 6 - resíduos com correlação aparente no tempo

IV.4.3 Análise da distribuição do resíduo

Em geral o critério utilizado para o ajuste obedece a determinadas hipóteses sobre a distribuição do ruído. Neste caso é importante verificar se estas são verificadas ou contraditas no resíduo. Um forma simples de verificar a distribuição do resíduo é fazendo-se um histograma do mesmo. Note-se que se o número de pontos experimentais disponíveis for pequeno raramente se obtém um histograma significativo.

Uma maneira simples de se fazer a comparação da distribuição do resíduo é fazer um pp-plot (Walter e Pronzato, 1997). Para tal calcula-se para cada ponto resíduo a distribuição experimental. Isto é feito ordenando-se os resíduos em ordem crescente e atribuindo a seguinte freqüência a cada resíduo:

$$\mathrm{fac}(\varepsilon_i) = \frac{i}{n} \qquad (11)$$

Há na literatura a proposição de fórmulas um pouco mais acuradas para esta freqüência experimental. A análise gráfica pode ser feita plotando-se esta frequência experimental em função da frequência teórica que seria obtida para um distribuição normal. No entanto, um meio mais prático de fazê-la é através do gráfico de probabilidade normal (Mandansky, 1988) que pode ser obtido através da função normplot do Matlab. Este tipo de gráfico possui nas ordenadas uma escala particular que faz com que uma variável cuja frequência fosse de fato normal apresente um comportamento linear. Na figura 7 é apresentado um gráfico obtido no Matlab para um conjunto de 15 variáveis obtidas através da dunção randn do Matlab.

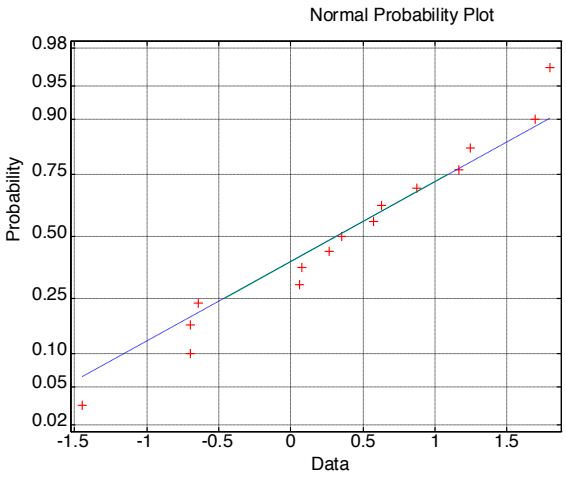

Figura 7 - Gráfico de probabilidade Normal para uma amostra de 15 variáveis obtidas através da função randn do Matlab

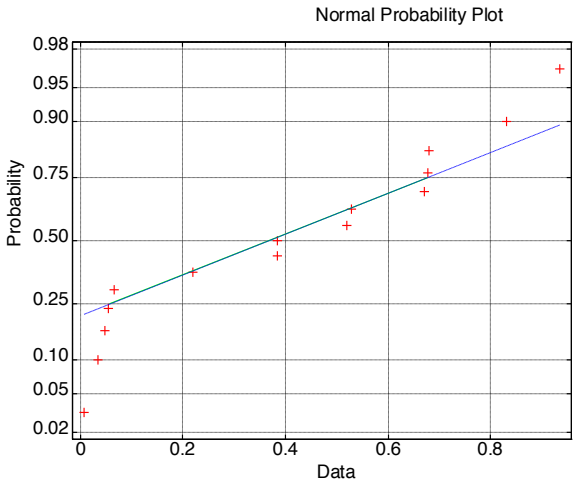

Figura 8 - Gráfico de probabilidade Normal para uma amostra de 15 variáveis obtidas através da função rand do Matlab (distribuição uniforme)

Na figura 8, é apresentado um gráfico de probabilidade normal para uma amostra de 15 variáveis obtidas a partir de uma distribuição uniforme. Como afirma Mandansky (1988) muitas vezes romanceia-se o resultado desta análise, no sentido em que ela não é

muito evidente e portanto quando deve-se fazer outros tipos de teste quanto à distribuição do resíduo. Na figura 9 apresentamos o gráfico de probabilidade Normal para uma amostra de 150 variáveis com distribuição uniforme. Pode-se apreciar que para grandes amostras a visualização pode ser mais significativa.

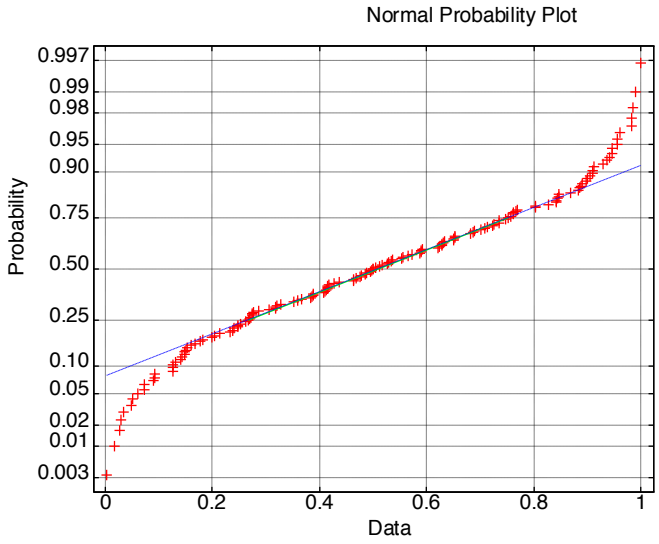

Figura 9 - Gráfico de probabilidade Normal para uma amostra de 150 variáveis obtidas através da função rand do Matlab (distribuição uniforme)

Para se testar a normalidade os testes mais comuns são o de Shapiro-Wilk (p.21, Mandansky, 1988), de Filliben (p.30, Mandansky, 1988), o D´Agostino (p.31, Mandansky, 1988) e o de Studentized range. Outras distribuições podem ser supostas e testadas através de testes de ajustes de distribuições tais como o de χ^2 (baseado na diferença entre a distribuição teórica e experimental em diferentes intervalos), o teste de Kolmogorov (baseado na comparação da diferença máxima entre a distribuição empírica e teórica) e o teste de Durbin (similar ao de Kolmogorov, só que baseado na discrepância entre diferenças entre distribuição empírica e teórica sucessivas).

IV.5 AJUSTE

É impossível quantificar o ajuste de um modelo de maneira absoluta pois existe o compromisso entre a falta de ajuste e o sobre-ajuste. A dificuldade da definição de um limite claro para este compromisso decorre principalmente do fato que à medida que aumenta a complexidade de um modelo, aumenta a sua capacidade de ajustar um conjunto de dados. Sejam por exemplo dois modelos M_1 e M_2, de forma que M_1 esteja contido em M_2. M_1 estar contido em M_2 significa que se forem assumidos fixados alguns valores de parâmetros do modelo M_2 o modelo obtido é o M_1. O ajuste obtido com o modelo M_1 tem necessariamente que ser pior que o do modelo M_2. É claro que existem dificuldades numéricas, ou limites ligados à informatividade dos dados. Isto é, se dispõe-se de um conjunto de 5 pares variável dependente-variável independente, as soluções obtida para modelos com 6, 7 ou mais parâmetros proporão possivelmente um mesmo ajuste, apesar de fornecerem predições, fora dos pontos experimentais que eles ajustam exatamente, totalmente diferentes.

Para modelos caixa-preta, isto é modelos sem correlação com os princípios fenomenológicos relativos ao sistema, a falta de ajuste pode ser compensada pelo aumento da complexidade do modelo. Um modelo caixa-preta típico é o modelo polinomial correlacionando uma variável independente a uma variável dependente. O aumento da complexidade do modelo pode ser entendido como o aumento do grau do polinômio de correlação. Um modelo polinomial de grau 4 ajustado a um conjunto de 5 pares variável dependente / variável independente fornece não mais uma correlação senão uma interpolação, pois ele é capaz de interpolar exatamente todos os pontos.

Para modelos fenomenológicos, a falta de ajuste é um elemento enriquecedor, pois em geral denota alguma limitação das hipóteses assumidas para a geração do modelo e portanto, a existência de algum fenômeno que não foi contemplado nestas hipóteses.

A maneira mais trivial, tradicional e comum de se acompanhar o ajuste de um modelos é ode construir um gráfico de variáveis simuladas em função das variáveis medidas (ou vice-versa, tanto faz), também chamado de gráfico de paridade. Nas figuras 10, 11, 12, 13 e 14 são apresentados diferentes gráficos de paridade.

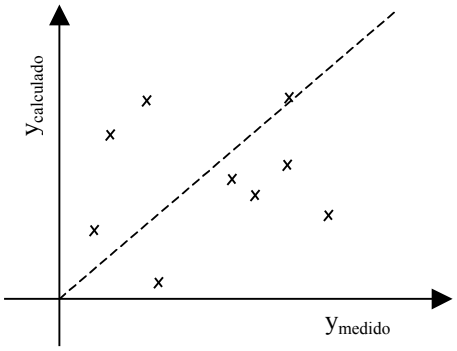

Figura 10 - Exemplo de gráfico de paridade

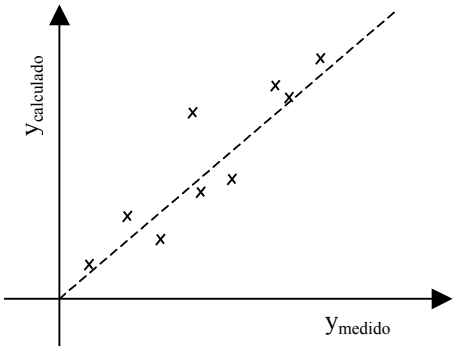

Figura 11 - Exemplo de gráfico de paridade

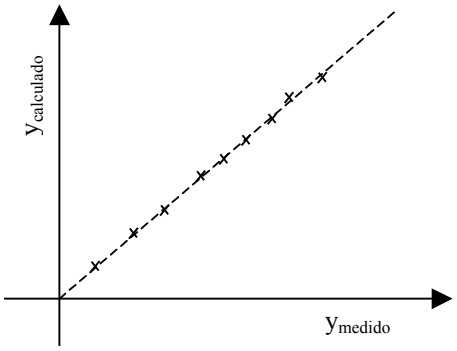

Figura 12 - Exemplo de gráfico de paridade

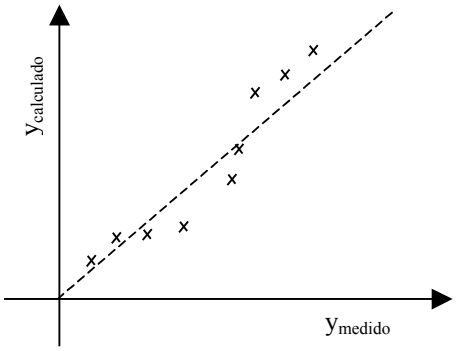

Figura 13 - Exemplo de gráfico de paridade

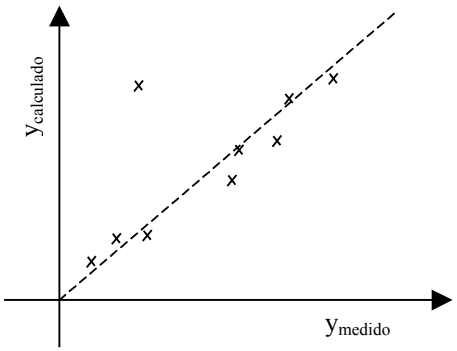

Figura 14 - Exemplo de gráfico de paridade

Na figura 10 fica praticamente evidente que o ajuste é péssimo. Nas figura 11 e 12 são apresentados dois ajustes diferentes, no entanto é difícil dizer se um é bom e o outro é ruim. O ajuste da figura 12 pode representar um sistema sobreajustado (mas a quantificação disto depende de outros parâmetros). O ajuste da figura 13 pode representar

um problema de falta de ajuste. O gráfico de paridade da figura 14 apresenta um ponto esquisito, no entanto o problema pode ser tanto de falta de ajuste quanto de ponto fora de padrão (outlier) que será discutido mais tarde. Os gráficos de paridade são ferramentas interessantes mas que por si só não permitem classificar univocamente o ajuste.

Um outro elemento gráfico elegante de comparação do ajuste são gráficos em que se comparam as predições do modelo com os dados experimentais em função de uma determinadas variável independente. Na figura 15 apresentamos um destes gráficos.

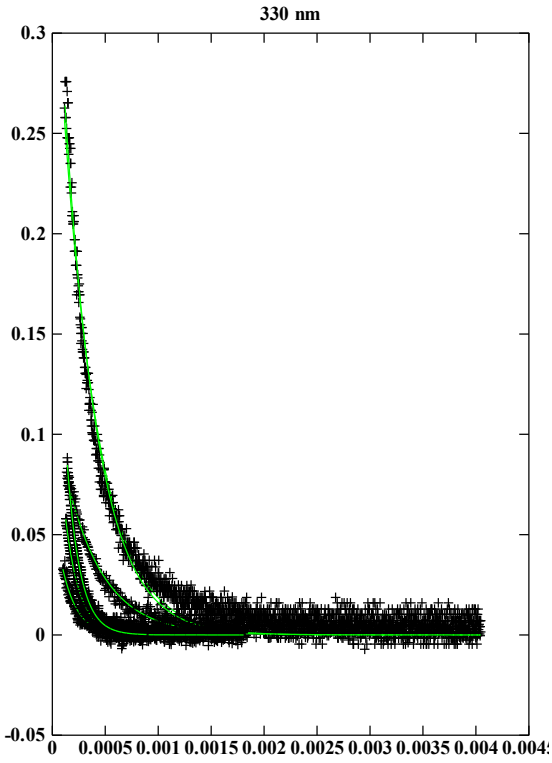

Figura 15 - Comparação entre predições e dados experimentais

IV.6 Estatísticas de ajuste

Existe uma diversidade enorme na literatura de estatísticas para quantificar o ajuste. O grande problema, como já foi dito anteriormente, é que em geral o critério de

ajuste diminui monoticamente com a complexidade do modelo, e portanto se se considerar apenas esta estatística para classificar o ajuste, haverá uma tendência a favorecer modelos sobre-ajustados.

Um estatística que sofre exatamente deste mesmo problema, e que é muito usada, é a chamada R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad (12)$$

onde $\hat{y}_i$ são os valores preditos e $\bar{y}$ a média dos valores medidos. Quando os dados experimentais são interpolados exatamente esta estatística tem o valor de 1 (valor máximo).

IV.6.1 Testes de ajuste

Testes de ajuste são baseados na comparação da variância de **e** (ruído, σ^2) com a variância estimada a partir do ajuste. Existem duas hipóteses possíveis para este teste. Se σ^2 for conhecida a priori (do que eu duvido), calcula-se a estimativa da variância, digamos, após ajuste:

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (13)$$

neste caso a variável $\hat{\sigma}^2 / \sigma^2$ tem distribuição χ^2_{n-p} e em consequência pode ser testada a hipótese de igualdade entre $\hat{\sigma}^2$ e σ^2 . É importante frisar que esta distribuição de amostragem depende de que as hipóteses de ruído distribuído normalmente e independente se confirmem.

Caso não se conheça o valor de σ^2 a priori, uma estimativa independente do modelo tem que ser obtida. Esta estimativa pode ser obtida através de repetições de ensaios. Ao teste que é proposto é dado o nome de teste de falta de ajuste (lack of fit) cujo procedimento está resumido na tabela 1.

Tipo de repetição	Número de pontos repetidos por tipo de repetição	média de cada repetição	Soma dos Quadrados	Graus de Liberdade
1	n_1	$\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$	$\sum_{i=1}^{n_1}(y_i - \bar{y}_1)^2$	n_1-1
2	n_2	$\bar{y}_2$	$\sum_{i=1}^{n_2}(y_i - \bar{y}_2)^2$	n_2-1
...	...	...	...	
K	n_k	$\bar{y}_k$	$\sum_{i=1}^{n_k}(y_i - \bar{y}_k)^2$	n_k-1
		Total	S_E (soma dos erros puros)	n_E

Tabela 1 - Procedimento de implementação das estatísticas envolvidas no teste de falta de ajuste

Calcula-se a soma de falta de ajuste, que é dada pela diferença entre a soma dos erros ao quadrado (norma do resíduo) e soma dos erros puros:

$$S_L = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 - S_E \qquad (14)$$

que tem n-p-n_E graus de liberdade. A razão:

$$\frac{S_L / n_L}{S_E / n_E} \qquad (15)$$

tem distribuição $F_{nL,nE}$ (se todas as outras hipóteses adicionais forem válidas) e um teste pode ser realizado para decisão sobre a igualdade dos termos.

IV.6.2 Estatísticas envolvendo penalização por complexidade

Existem algumas estatísticas em que se penaliza a complexidade do modelo. Algumas delas são baseadas em nada, ou seja são puramente empíricas, outras são baseadas em tentativas de se levar em conta o error de predição, algumas são derivadas de hipóteses simplistas e outras são baseadas em critérios Bayesianos.

Um dos critérios empíricos é o critério de R^2 ajustado (a notação na literatura não é nada homogênea, portanto não adianta lhe dar um nome particular):

$$\overline{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ (n-p)}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \Big/ (n-1)} \qquad (16)$$

A grande vantagem deste critério é que ele contém uma certa penalização pela complexidade, ou seja, à medida que o número de graus de liberdade (n-p) diminui o critério piora. Outros critérios serão vistos quando for tratado o problema da complexidade do modelo.

IV.7 Inferências sobre os parâmetros

Um estimador para a variância dos parâmetros, no caso de modelos monovariáveis seria:

$$\hat{\mathbf{V}}_\theta = \hat{\sigma}_e^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \qquad (22)$$

Ou, no caso mais geral:

$$\hat{\mathbf{V}}_{\theta} = \left[\sum_{i=1}^{n} \left(\mathbf{R}_i^{T} \hat{\mathbf{V}}_{\varepsilon}^{-1} \mathbf{R}_i \right) \right]^{-1} \qquad (23)$$

Algumas estatísticas podem ser calculadas a partir destas matrizes:

IV.7.1 Coeficientes de correlação:

A matriz de coeficientes de correlação é dada como uma matriz de coeficientes dados por:

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{\theta ij}^{2}}{\sqrt{\hat{\sigma}_{\theta ii}^{2} \hat{\sigma}_{\theta jj}^{2}}} \qquad (24)$$

Esta matriz tem diagonal unitária. Diz-se na literatura que "se algum coeficiente for próximo a um ou –1, os parâmetros são altamente correlacionados", mas esta frase naõ faz sentido.

IV.7.2 Decomposição em componentes principais

A decomposição em componentes principais da matriz de covariância leva a:

$$\mathbf{S}^{T} \mathbf{DS} = \hat{\mathbf{V}}_{\theta} \qquad (25)$$

como a matriz **V** é simétrica semi-positiva **S** é uma matriz ortogonal, e **D** uma matriz diagonal. Às coluna de **S** dá-se o nome de auto-vetores e aos elementos da diagonal de **D** dá-se o nome de auto-valores. A ortogonalidade da matriz **S** se traduz por:

$$\mathbf{S}^{T} \mathbf{S} = \mathbf{S}\mathbf{S}^{T} = \mathbf{I} \qquad (26)$$

Dá-se o nome de componente principal a cada um dos auto-vetores. Os componentes principais de um estimador, que podemos definir como (válido somente para modelos lineare em relação aos parâmetros):

$$\theta_i^{CP} = \mathbf{s}_i^T \hat{\boldsymbol{\theta}} \qquad (27)$$

têm a propridade de ter a matriz de covariância diagonal.

IV.7.3 Região de Confiança

Algumas estimativas de regiões de confiança para os parãmetros (para modelos lineares em relação aos parâmetros) são definidas através de:

- se $\mathbf{V}_\theta$ for conhecida, a região é definida pelos $\boldsymbol{\theta}$ tais que:

$$\left(\theta - \hat{\theta}\right)^T \mathbf{V}_\theta^{-1} \left(\theta - \hat{\theta}\right) \le \chi^2\left(p, 1-\alpha\right) \qquad (28)$$

- se $\mathbf{V}_\theta$ não for conhecida, a região é definida pelos $\boldsymbol{\theta}$ tais que:

$$\left(\theta - \hat{\theta}\right)^T \hat{\mathbf{V}}_\theta^{-1} \left(\theta - \hat{\theta}\right) \le p\, F\left(p, n-p; 1-\alpha\right) \qquad (29)$$

Em ambos os casos a região de confiança dos parâmetros é uma hiper-elipse. Os eixos da hiperelipse têm justamente a direção dada pelos auto-vetores da matriz $\mathbf{V}_\theta$. Os comprimentos relativos são inversamente proporcionais à raiz quadrada dos autovalores. A figura 16 ilustra a região de confiança conjunta para o caso de um modelo linear com dois parâmteros.

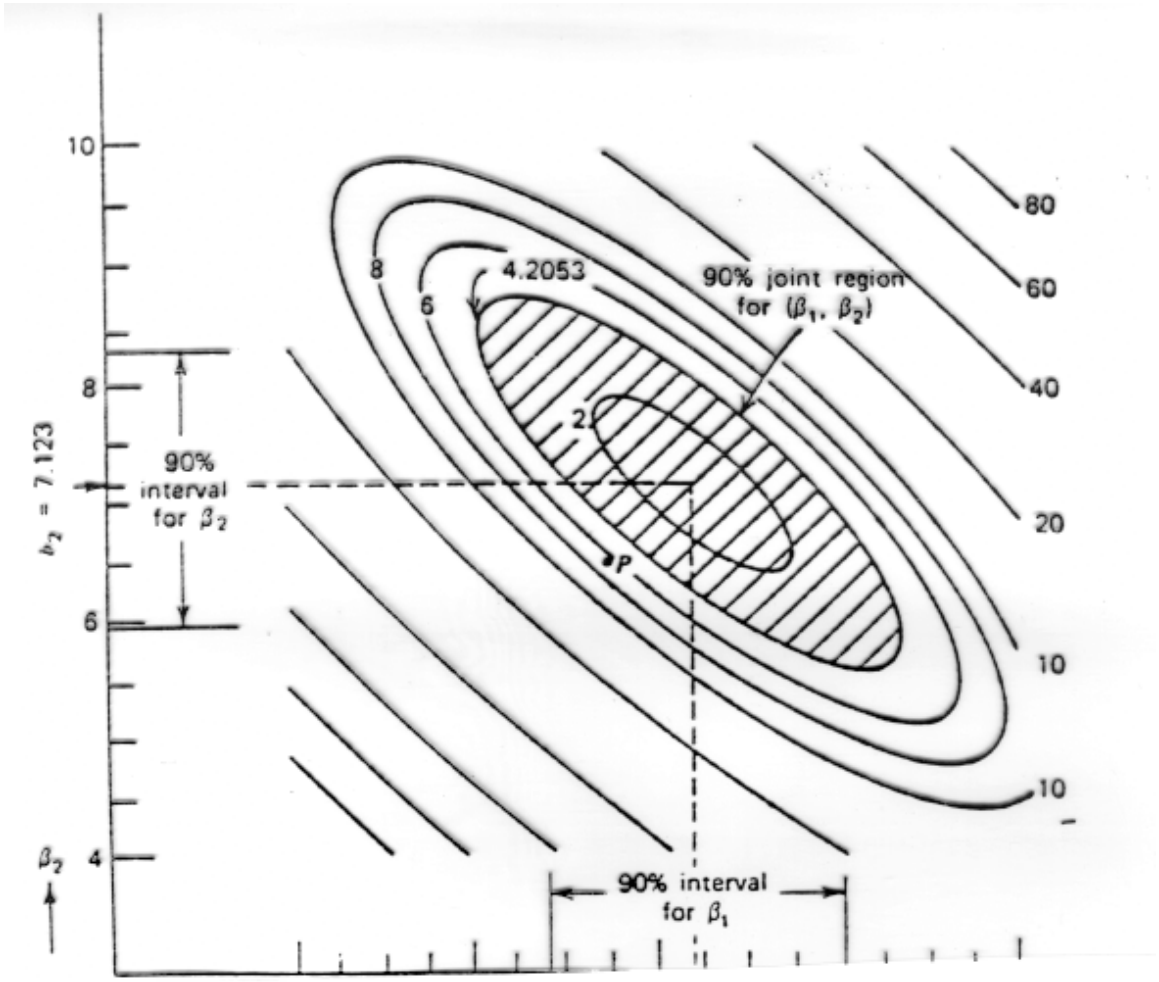

Figura 16 - Variância Marginal e Variância Conjunta, ilustração de Box H.&H.

IV.7.4 Intervalo de confiança marginal

Define-se o intervalo de confiança marginal (quando a matriz de coovariância não é conhecida a priori) por:

$$\theta_i = \hat{\theta}_i \pm \sqrt{\hat{\sigma}_{\theta_{ii}}^2}\; t(n-p, \alpha/2) \qquad (30)$$

Dizer que um parâmetro pode ser considerado nulo devido ao fato que o valor zero se encontra dentro do intervalo de confiança é um tanto absurdo e simplista. Além do quê, o significado da região de confiança do intervalo marginal de confiança são bastante diferentes, como pode ser observado na figura 16.

Se se quer fazer a "seleção de regressores" a metodologia a ser usada está longe de ser aquela em que se retiram os parâmetros "não significativos" do modelo. Mas isto será visto com mais detalhes quando falarmos de complexidade.

IV.7.5 Variância de Predição

A variância de predição permite quantificar o quanto a incerteza sobre os parâmetros influencia as predições em diferentes pontos. No caso de um modelo LP as predições para um conjunto de variáveis indetependentes **x** é calculada por:

$$\hat{y} = \mathbf{r}^{\mathbf{T}}(\mathbf{x})\,\hat{\theta}$$

A variância de predição é dada por:

$$\hat{\sigma}_p^2 = \hat{\sigma}_\varepsilon^2 + \mathbf{r}^{\mathbf{T}}(\mathbf{x})\ \mathbf{V}_\theta^{-1}\mathbf{r}\ (\mathbf{x})$$

IV.8 Detecção de Dados fora de padrão

O grande dilema na detecção de outiers é que para classificar dados fora de padrão é ncessário ter uma idéia do que seja um padrão. Na figura 17 ilustra-se o que pode ser classficado como um dado fora de padrão por um procedimetno puramente estatístico:

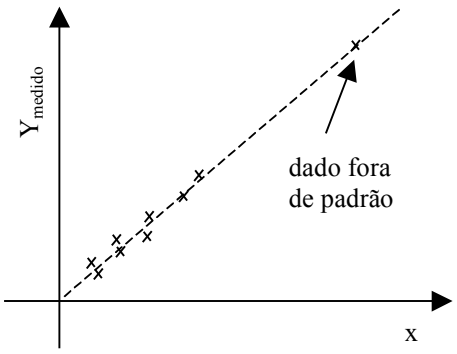

Figura 17 – análise de dados for a de padrão de um ponto de vista puramente estatístico

Ao contrário se a abordagem é partir de um modelo para decidir sobre outliers, é necessário ter resíduos para saber o que não é parecido com os outros. No entanto, limpando-se as observações fora de padrão perturba-se a distribuição do que resta, pois o procedimento produz erros de tipo I e II.

Por outro lado, os melhores procedimentos de rejeição não levam ao mesmo desempenho que os procedimentos de estimação robusta, pois estes últimos permitem que dados fora de padrão possam ser meio que levados em conta na análise.

No caso de regressões uma idéia seria calcular a regressão com todos os pontos e posteriormente calcular a regressão eliminado os ponto um a um. Por exemplo, seja $\mathbf{X}$ a matriz de regressores considerando todos os ponto. Chamemos $\mathbf{X}_{(i)}$ a matriz obtida retirando-se a i-ésima linha de $\mathbf{X}$. É válida a seguinte propriedade:

$$\left(\mathbf{X}_{(i)}{}^{\mathbf{T}}\mathbf{X}_{(i)}\right)^{-1} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1} + \frac{\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{r}^{\mathrm{T}}\left(\mathbf{x_i}\right)\mathbf{r}\left(\mathbf{x_i}\right)\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}}{1 - h_{ii}}$$

onde h_{ii} é o elemento da diagonal da matriz chapéu correspondente a "i". A diferença entre as estimativas considerando ou não o ponto I é dada por:

$$\hat{\boldsymbol{\theta}}_{(i)} = \hat{\boldsymbol{\theta}} - \frac{\hat{y}_i\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)\mathbf{r}^{\mathrm{T}}\left(\mathbf{x_i}\right)}{1 - h_{ii}}$$

Uma comparação entre esses dois resultados permite medir o quanto um único ponto tem influência na estimativa. Cook definiu:

$$d_i = \frac{\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^T \left(\mathbf{X}^T\mathbf{X}\right)\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^T}{p\sigma^2}$$

como a distância entre $\hat{\theta}_{(i)}$ e $\hat{\theta}$. Esta distância pode ser calculada através de:

$$d_i = \frac{t_i^{\,2}}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

onde os t_I são os resíduos studentizados definidos em (9). Para n grandes recomenda-se que d seja menor que 1. Outros critérios como o DFITS podem ser encontrados em Madansky (pp. 142 a 144).

Bibliografia

de Groot, M.H., *Probability and Statistics*, Addison-Wesley, Philippines, 1975

Hocking, R.R., Developments in Linear Regression Methodology: 1959-1982, *Technometrics*, 25 (3), 219-230, 1983

Mandansky, A., Prescriptions for Working Statisticians, Springer Texts in Statistics, 1988

Sanford Weisberg, Some Principles for Regression Diagnostic and Influence Analysis, *Technometrics*, 25 (3), 240-244, 1983

Walter, E., Pronzato, L., Identification of Parametric Models from Experimental Data, Springer, 1997

Welsch, R.E., Discussion on Developments in Linear Regression Methodology: 1959-1982, *Technometrics*, 25 (3), 219-230, 1983