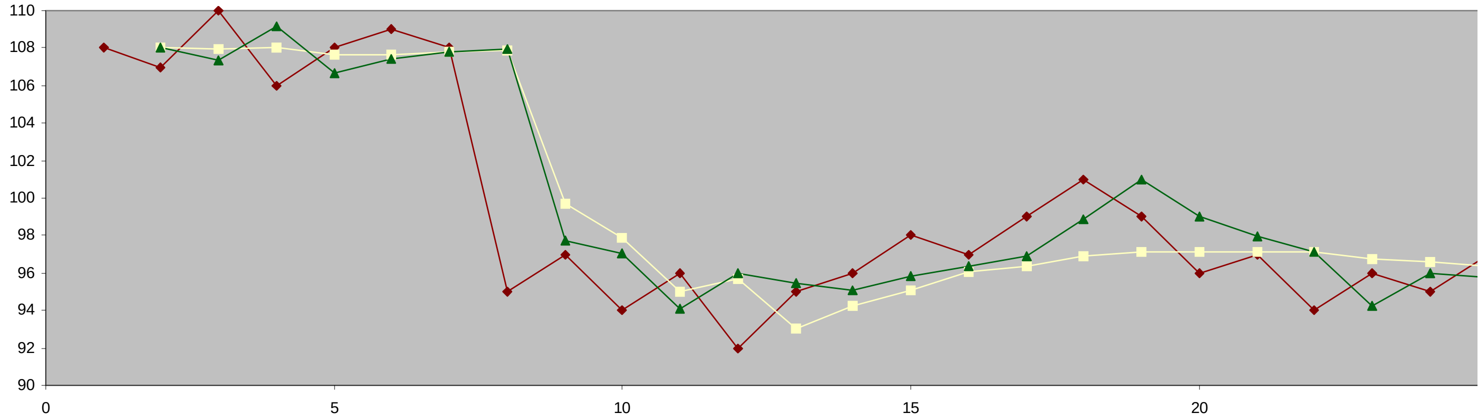
MÉTODOS ESTATÍSTICOS DE PREVISÃO



Análise de Erros

Bernardo Almada Lobo

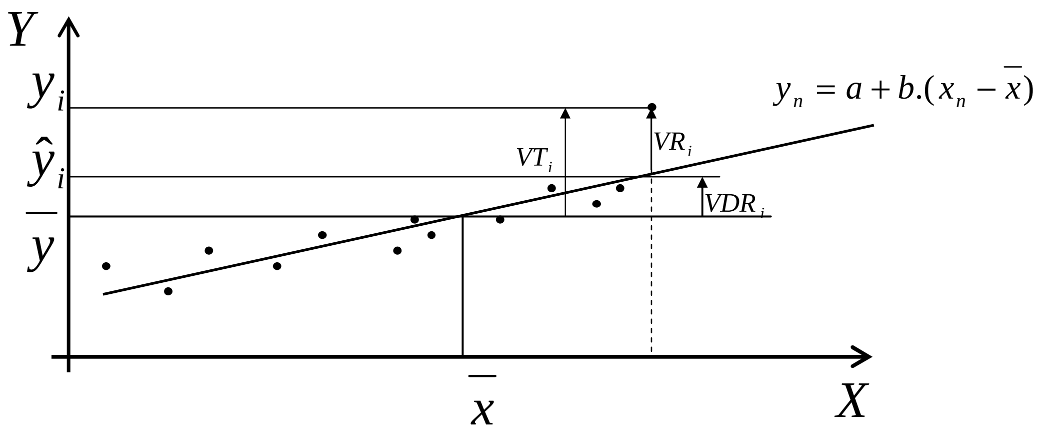
DEMEGI
Departamento de Engenharia Mecânica e Gestão Industrial

Universidade do Porto
FEUP Faculdade de Engenharia

Regressão Linear Múltipla

Coeficiente de Determinação (R^2_{XY})

$$R^2_{XY} = \frac{\text{VDR (variação de } Y \text{ explicada pela regressão)}}{\text{VT (variação total de } Y\text{)}} = \frac{B^2 . S_{XX}}{S_{YY}}$$



$$\left(Y_i - \bar{Y}\right) = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$

$$\sum_n \left(Y_n - \bar{Y}\right)^2 = \sum_n \left\{\left[A + B.\left(X_n - \bar{X}\right)\right] - \bar{Y}\right\}^2 + \sum_n \left\{Y_n - \left[A + B.\left(X_n - \bar{X}\right)\right]\right\}^2$$

VT = VDR + VR

$$\mathrm{B} = \frac{\mathrm{S_{XY}}}{\mathrm{S_{XX}}} \text{ (estimador de } \beta\text{)}$$

$$\mathrm{S_{XX}} = \sum_n \left(\mathrm{X_n} - \bar{\mathrm{X}}\right)^2$$

$$\mathrm{S_{YY}} = \sum_n \left(\mathrm{Y_n} - \bar{\mathrm{Y}}\right)^2$$

$$\mathrm{S_{XY}} = \sum_n \left[\left(\mathrm{X_n} - \bar{\mathrm{X}}\right).\left(\mathrm{Y_n} - \bar{\mathrm{Y}}\right)\right]$$

Regressão Linear Múltipla

Coeficiente de Determinação (Cont.)

- O coeficiente de determinação (R^2_{XY}), que traduz a proporção da variação total de Y explicada pela regressão ajustada, corresponde ao coeficiente de correlação r elevado ao quadrado;

- Este coeficiente apresenta uma limitação: o denominador da expressão que lhe está subjacente (ver página anterior) tem um valor fixo, enquanto que o numerador só pode aumentar. Assim, ao adicionar-se uma nova variável na equação da regressão, o numerador aumentará, no mínimo, ligeiramente, resultando num aumento do coeficiente de determinação, mesmo que a introdução da nova variável resulte numa equação menos eficiente;

- Em teoria, usando um número infinito de variáveis independentes para explicar a variação da variável dependente, resulta num R^2_{XY} igual a 1. Por outras palavras, o coeficiente de determinação pode ser manipulado, logo deve ser suspeitado;

Regressão Linear Múltipla

Coeficiente de Determinação Ajustado ($\overline{R}^2_{XY}$)

- Dado que a introdução de um regressor irrelevante aumentará ligeiramente o R^2_{XY}, é desejável tentar corrigi-lo, reduzindo-o de uma forma apropriada;
- O coeficiente de determinação ajustado, $\overline{R}^2_{XY}$, é uma tentativa de tentar corrigir o R^2_{XY}, ajustando o numerador e o denominador da expressão da página 2, através dos respectivos graus de liberdade;

$$\overline{R}^2_{XY} = 1-\left(1-R^2\right).\left(\frac{N-1}{N-K-1}\right)$$

N : n.º de observações
K : grau da regressão
N-1 : n.º total de graus de liberdade da VT
N-K-1 : n.º de graus de liberdade da VR

- Contrariamente ao coeficiente de determinação, o coeficiente de determinação ajustado pode diminuir em valor se a contribuição da variável adicional na explicação da VT, for inferior ao impacto que essa adição acarreta nos graus de liberdade.

Análise de Erros

"Measuring forecast accuracy"

Medidas Estatísticas Padrão

$$e_t = Z_t - \hat{Z}_t$$

$$EM = \frac{\sum_{t=1}^{n} e_t}{n}$$

$$EAM = \frac{\sum_{t=1}^{n} |e_t|}{n}$$

$$EQM = \frac{\sum_{t=1}^{n} e_t^2}{n}$$

EM : Erro Médio

EAM : Erro Absoluto Médio

EQM : Erro Quadrático Médio

Análise de Erros (cont.)

"Measuring forecast accuracy"

Medidas Relativas

$$EP_t = \left(\frac{Z_t - \hat{Z}_t}{Z_t} \right) x\,100$$

$$EPM = \frac{\sum_{t=1}^{n} EP_t}{n}$$

$$EPAM = \frac{\sum_{t=1}^{n} |EP_t|}{n}$$

EP_t : Erro Percentual

EPM : Erro Percentual Médio

$EPAM$: Erro Percentual Absoluto Médio

Análise de Erros (cont.)
Validade da Decomposição

Teste ao valor esperado dos erros

$$\overline{X} = \hat{\mu}_{E_t} = \frac{1}{N}\sum_{t=1}^{N} E_t$$

$$s_{\overline{X}} = \hat{\sigma}_{\hat{\mu}_{E_t}} = \frac{s_X}{\sqrt{N}} = \frac{1}{\sqrt{N}} \cdot \sqrt{\left[\frac{1}{N-1}\sum_{t=1}^{N}\left(E_t - \hat{\mu}_{E_t}\right)^2\right]}$$

- Amostra de pequena dimensão, população Normal (teste *t*):

$$H_0 : \mu = 0 \qquad H_1 : \mu \neq 0 \qquad \mathrm{ET} = \frac{\hat{\mu}_{E_t} - 0}{\hat{\sigma}_{\hat{\mu}_{E_t}}}$$

$$\text{Se } H_0 \text{ verd.} \Rightarrow \mathrm{ET} \rightarrow t_{N-1}(\alpha)$$

Análise de Erros (cont.)
Validade da Decomposição

Coeficiente de Autocorrelação dos Erros (*lag* 1)

$$r_1 = \frac{\sum_{t=2}^{N}\left(E_t - \hat{\mu}_{E_t}\right)\cdot\left(E_{t-1} - \hat{\mu}_{E_t}\right)}{\sum_{t=1}^{N}\left(E_t - \hat{\mu}_{E_t}\right)^2}$$

- NOTA 1: Para testar se r1=0, é preciso conhecer-se os parâmetros da distribuição dos coeficientes amostrais de autocorrelação;

- NOTA 2: A distribuição dos coeficientes de autocorrelação de uma série de números aleatórios, pode ser aproximada por uma distribuição normal de média zero e desvio padrão $1/\sqrt{n}$

Análise de Erros (cont.)

Validade da Decomposição (cont.)

Coeficiente de Autocorrelação dos Erros - *lag* 1 (cont.)

- LIMITES DO INTERVALO DE CONFIANÇA A 95 % PARA UMA SÉRIE ALEATÓRIA:

$$\pm 1.96 \times \frac{1}{\sqrt{n}} \rightarrow \text{valor crítico}$$

- Se r1 estiver dentro daquele intervalo, não há correlação significativa entre erros sucessivos.

Análise de Erros (cont.)

"Measuring forecast accuracy"

Estatística U (*Theil*)

$$U = \sqrt{\frac{\sum_{t_o}^{t_1} (VRP_t - VRR_t)^2}{\sum_{t_o}^{t_1} (VRR_t)^2}}$$

VRP_t : Variação relativa prevista

VRR_t : Variação relativa real

$$VRR_t = \frac{Z_t - Z_{t-1}}{Z_{t-1}} \quad ; \quad VRP_t = \frac{\hat{Z}_{t-1}(1) - Z_{t-1}}{Z_{t-1}}$$

U=1: O método "naive" é tão eficiente quanto o método em avaliação;

U<1: O método "naive" é menos eficiente que o método em avaliação;

U>1: O método "naive" é mais eficiente que o método em avaliação;

U=0: O método em avaliação é perfeito.

- Nota: no método "naive" as previsões a um passo correspondem ao último valor observado.

Análise de Erros (cont.)

"Measuring forecast accuracy"

Estatística D-W (*Durbin-Watson*)

A estatística D-W é usada para testar a presença de autocorrelação de primeira ordem (r_1) nos erros de previsão. O teste compara os erros do período t com os erros do período t-1 e desenvolve uma estatística que mede a significância da correlação entre estas duas séries.

$$D-W=\frac{\sum_{t=2}^{N}\left(e_t-e_{t-1}\right)^2}{\sum_{t=1}^{N}e_t^2}$$

- 0 < D-W < 4
- D-W ≈ 2 → erros aleatórios
- D-W >>2 → erros negativamente correlacionados
- D-W <<2 → erros positivamente correlacionados