IAG MASTER EM DESENVOLVIMENTO GERENCIAL 2006

Fundamentos de Economia de Energia
Mônica Barros, D.Sc.
Aula 2
24/08/2007

Uma visão gerencial da Estatística e Séries Temporais

- Séries Temporais
  - Conceitos
  - Parâmetros
  - Modelo Linear
  - Modelos AR(1), AR(2), AR(p) e PAR(p)
  - Modelos de Séries Temporais
  - Estatísticas de "erros"

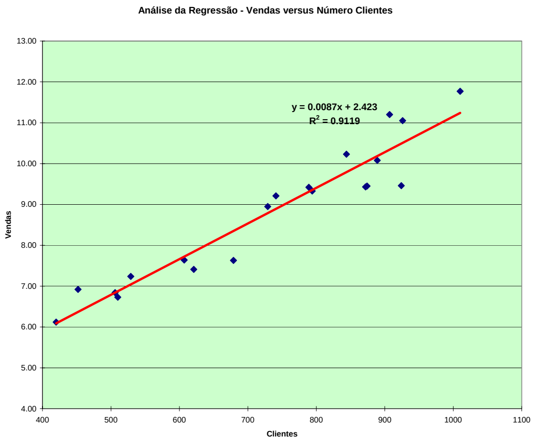
Série Temporal

- Conjunto de observações ordenadas no tempo.
- $y_1, y_2, \ldots, y_n$ – onde y_t é a carga no instante t.
- Existe dependência entre as cargas de diversos instantes.
- Isto nos permite usar cargas passadas para prever a carga futura.

Modelo

- É uma representação simplificada da realidade.
- Por exemplo, você observa duas variáveis X e Y e faz um gráfico dos n pontos (x_1,y_1), $x_2,y_2)$, ...(x_n,y_n). O gráfico tem a "cara":

Modelo



Modelo Linear

- Nesse exemplo a relação entre X e Y é, como o gráfico revela, aproximadamente linear.
- Nosso objetivo: achar reta que relaciona Y e X. Qual reta? Como encontrá-la?
- O método tradicional é chamado de **mínimos quadrados.**

Modelo Linear

- Método de Mínimos Quadrados
  - Como funciona? O objetivo é minimizar a soma dos quadrados dos “erros” (resíduos).
  - Que “erros” são esses?
  - Para cada ponto, o resíduo é o valor real menos o valor previsto pela reta.
  - Pode-se notar que o resíduo vai depender de qual reta foi ajustada.

Modelo Linear

- Por exemplo, se ajustamos a reta y = 3 + 2x, se o valor de x_i é 2, então y_i ajustado pela reat igual a 7. Se agora a reta é y = 4 + 2x, para x_i = 2 temos o y_i ajustado pela reta igual a 8.
- Imagine que o valor real de y_i seja 8.5 quando x_i é 2. Então, o resíduo gerado pela reta 3 + 2x é 8.5 – 7 = 1.5 e pela reta 4+2x é 8.5 – 8 = 8.

Modelo Linear

- Então, no que diz respeito a esse ponto (x_i =2), a reta 4 + 2x foi melhor que a reta 3 +2x pois gerou um resíduo menor.

- Em geral, dados n pares de pontos (x_1,y_1), x_2,y_2), ...(x_n,y_n), você pode ajustar uma reta y =b_0 + b_1.x a todos os pares.
- Note que os coeficientes b_0 e b_1 são os mesmos para todos os pares de pontos.

Modelo Linear

- O resíduo do i-ésimo par é:

$$\hat{e}_i = y_i - \hat{y}_i = real_i - previsto_i$$

- O método de mínimos quadrados procura obter a reta (ou seja, os coeficientes b_0 e b_1) tais que a soma dos quadrados dos resíduos seja minimizada.

Modelo Linear

- Método de Mínimos Quadrados
- Ache b_0 e b_1 tais que a expressão:

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

- seja minimizada.
- Esta soma acima é chamada de Soma dos Quadrados dos Resíduos (às vezes, Soma dos Quadrados dos Erros).

Modelo Linear

- Erro versus Resíduo
- Erro = variável aleatória não observável, com média zero e variância constante σ^2.

- Resíduo = estimador ("chute") para o erro. É observável, calculado como valor real – valor previsto pelo modelo.
- Isso causa alguma confusão, pois às vezes a gente chama de "erro" o que é, na verdade, o "resíduo".

Modelo Linear

- O modelo linear é:
- $y_i = b_0 + b_1 x_i + e_i$ para i =1, 2, ..., n

- Onde e_i é o erro, uma variável com média 0 e variância constante σ^2.

- Os parâmetros do modelo são b_0, b_1 e σ^2, são constantes desconhecidas que devem ser estimadas.

Modelo AR(1)

- Parâmetro = tudo que precisa ser estimado para especificar completamente o modelo.

- Modelo AR(1)
- AR(1) = autoregressivo de ordem 1

- $y_t = \varphi\, y_{t-1} + e_t$
  - y_t = carga no mês t
  - Y_{t-1} = carga no mês t-1
  - e_t = erro no mês t

Modelo AR(1)

- Parâmetros: φ, σ^2 (variância do erro)

- Caso Particular: Passeio Aleatório (Random Walk)
- AR(1) com $\varphi = 1 \Rightarrow y_t = y_{t-1} + e_t$
- Uso: mercado de ações
- Na média, preço da ação hoje = preço da ação ontem (não ajuda muito para ganhar dinheiro!!)

Random Walk

- O gráfico a seguir mostra um exemplo de um passeio aleatório em que $y_0 = 100$ e os erros são Normais com média zero e desvio padrão 2.

- Foram geradas 500 observações.

- Você poderia confundir o gráfico com a da evolução do preço de uma ação no tempo, não é mesmo?

Random Walk

Modelo AR(1)

Exemplo – efeito de σ^2

- $y_t = 1.05.y_{t-1} + e_t$
- Carga do mês t = 105% da carga do mês t-1 + erro.
- Suponha que $y_{t-1} = 100$.
  - a) $VAR(e_t) = 1$
  - b) $VAR(e_t) = 100$
  - Por exemplo, se fosse possível observar o erro poderíamos encontrar algo como:

Modelo AR(1)

a) $e_t = 1.05$ – valor pequeno pois foi “tirado” de uma distribuição de prob. com média 0 e variância 1.

b) $e_t = 10.32$ – valor grande porque a distribuição de prob. tem variância “grande” (100).

- Resultados:

a) $y_t = 105 + 1.05 = 106.05$

b) $y_t = 105 + 10.32 = 115.32$

Modelo AR(2)

- $y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + e_t$
- Onde:
  - e_t = erro no instante t com média 0 e variância σ^2
  - y_t = carga no instante t
  - y_{t-1} = carga no instante t-1
  - y_{t-2} = carga no instante t-2
- Parâmetros: φ_1, φ_2 e σ^2 .

Modelo AR(p)

- Modelo Autoregressivo de ordem p
- $y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots\ldots + \varphi_p y_{t-p} + e_t$

- A carga deste mês depende das cargas dos p últimos meses.
- Existem p + 1 parâmetros a serem estimados: os φ´s e a variância do erro.

Modelo PAR(p)

- PAR(p) = Modelo Autoregressivo Periódico de ordem p.

- Contexto: Newave
- Modelos para previsão de vazões.
- Um modelo AR diferente para cada mês. Restrição do Newave - p máximo = 6 meses. (p = ordem do modelo).

Modelo de Séries Temporais

- Resíduo no instante t:
- $\hat{e}_t$ = real – previsto no instante t

- Por exemplo, para dez/06, é o real de dezembro de 2006 menos o valor previsto para o mesmo mês. Só vai poder ser calculado quando o valor real de 12/2006 estiver disponível!

Modelo de Séries Temporais

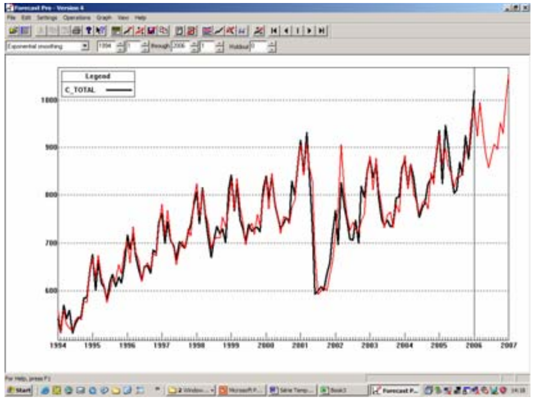
- Mas posso olhar para os resíduos “in sample” (dentro da amostra) no período em que o modelo foi ajustado.

- Por exemplo, suponha que o modelo foi ajustado no período jan/03 a abr/06. Neste período todo posso fazer a comparaçao entre o valor real e o ajustado pelo modelo (“fitted value”), calculando o resíduo a cada instante.

- Um exemplo está na próxima figura.

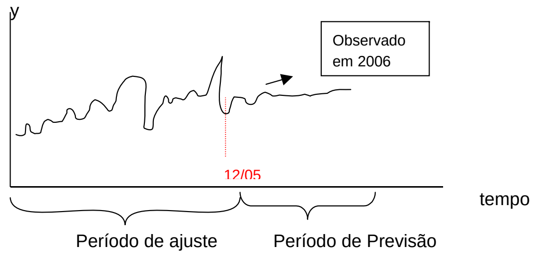
Modelo de Séries Temporais

- O gráfico a seguir mostra uma série de carga real (em preto) e ajustada por um modelo (em vermelho).



Modelo de Séries Temporais

- Período de ajuste (in sample)
- Período de previsão (out of sample)
- No gráfico anterior, o período de ajuste vai de janeiro de 1994 a dezembro de 2005, e o período de previsão é o ano de 2006.



Modelo de Séries Temporais

- Você ajustaria o seguinte modelo:
  - In sample = 12 meses,
  - Out of sample = 24 meses?
- Não! E o contrário? Sim, desde que a série não fosse muito mal comportada.
- Em resumo – previsão de séries temporais não é futurologia! Você precisa de dados passados para fazer previsão!

Estatísticas de “Erros”

- Estatística de “erros” (resíduos)

Mês	Real	Previsto	Resíduo
1	y1	^y1	ê1 = y1 - ^y1
2	y2	^y2	ê2 = y2 - ^y2
...	...	...	...
n	yn	^yn	ên = yn - ^yn

- MAPE = erro absoluto médio percentual
- “erro” percentual no instante i:
  - (yi - ^yi) / yi
- “erro” percentual absoluto no instante i:
  - \|yi - ^yi / yi \|

Estatísticas de "Erros"

- "Erro" absoluto percentual absoluto no instante i:
  - |(yi - ^yi)/ yi| = |(real – previsto)/real|
- MAPE = média destes erros absolutos percentuais

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{real_i - previsto_i}{real_i}\right|$$

- Vantagem: fácil de entender - a escala é %
- Desvantagem: Se o valor real é pequeno, qualquer discrepância na previsão faz o MAPE "explodir".

Estatísticas de "Erros"

- MAD = mean absolute deviation = desvio absoluto médio

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = \frac{1}{n}\sum_{i=1}^{n}|real_i - previsto_i|$$

- O MAD está nas mesmas unidades que a sua série.
- Por que o módulo?
- Para evitar soma = 0 quando um erro é grande e positivo e outro é grande e negativo!

Estatísticas de "Erros"

- EQM = erro quadrático médio

$$EQM = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(real_i - previsto_i)^2$$

- RMSE = root meam squared erro = raiz do EQM

$$RMSE = \sqrt{EQM} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(real_i - previsto_i)^2}$$

- Vantagem sobre EQM => mesma escala que os dados.

Estatísticas de "Erros"

- Erro "puro":

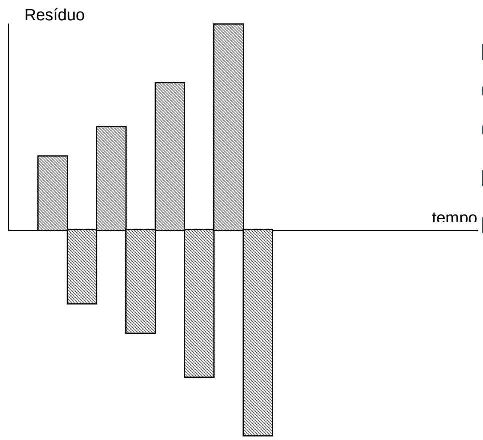
$$e_i = y_i - \hat{y}_i$$

- Vantagem: mesma escala que os dados.
- Desvantagem: não dá para sair somando! Se você fizer isso corre o risco de achar que seu modelo é bom quando não é, pois erros de sinais opostos poderão se cancelar.

Estatísticas de "Erros"

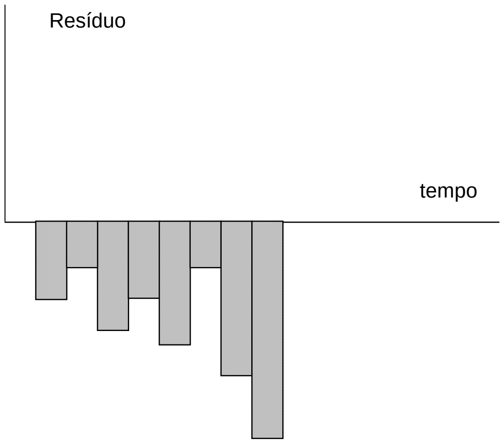
- O que fazer com tudo isso isso?
- O que é desejável? Se o modelo é bom, resíduo não tem padrão, não tem estrutura.
- Faça gráficos dos resíduos "puros" ou MAPE, ou MAD ou EQM, ou RMSE.
- Que tipo de gráficos?
  - Por exemplo, gráfico dos "erros" ao longo do tempo. Serve para responder como o seu modelo se comportou no "in sample" e verificar a existência de possíveis padrões.

Estatísticas de "Erros"



- Isto é tudo que eu não quero.
- Por que?
- Resíduo com um padrão claro.

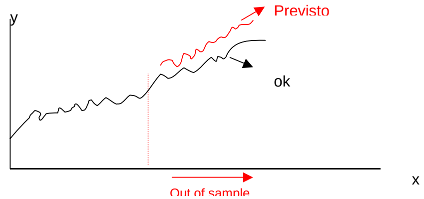
Estatísticas de "Erros"



Também não quero um padrão desses – todos os resíduos são negativos, indicando que a minha previsão esteve sempre ACIMA do real

Estatísticas de "Erros"

- Gráfico dos "erros" versus o horizonte de previsão.
  - Serve para responder se as previsões se deterioram rápido ou não.



Estatísticas de "Erros"

- Gráfico da Autocorrelação dos erros
- Correlação entre os erros em diversos instantes
- ACF = autocorrelation function
- Gráfico da ACF versus lag (defasagem) é também chamado de correlograma.

- O que você não quer?
- Que exista correlação entre resíduos em instantes diferentes.

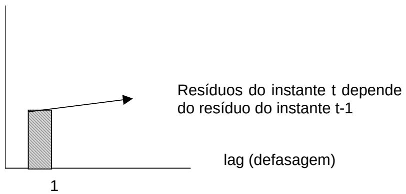
Estatísticas de "Erros"

- Por que você não quer que o gráfico da ACF dos resíduos mostre padrões?

- Porque se existe dependência temporal na série, toda esta dependência teria que ser, idealmente, capturada pelo modelo, não deveria ter sobrado nada nos resíduos!

Estatísticas de "Erros"

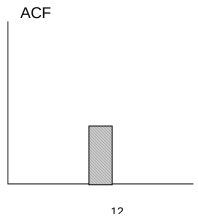
- Casos comuns...

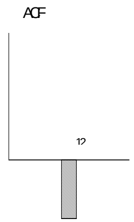


- Atenção: A autocorrelação é sempre um número entre –1 e +1
  - No gráfico anterior:
  - Resíduo grande e positivo no instante t leva a um resíduo grande e positivo no instante t+1

Estatísticas de "Erros"

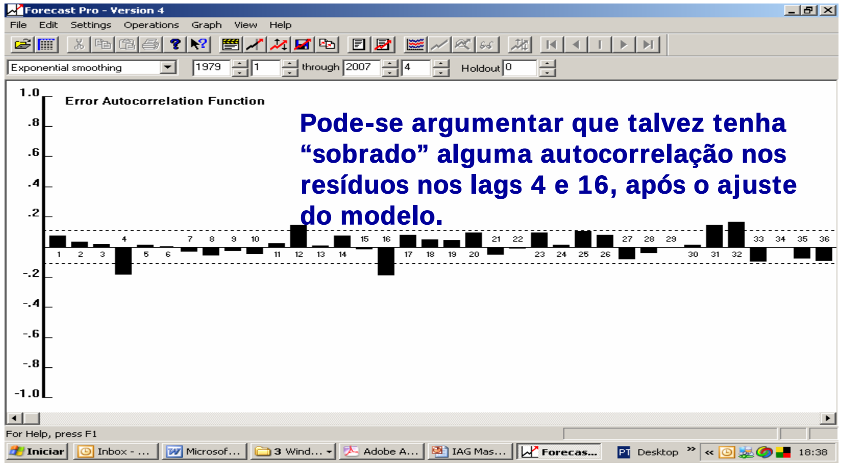
- Outros Padrões:





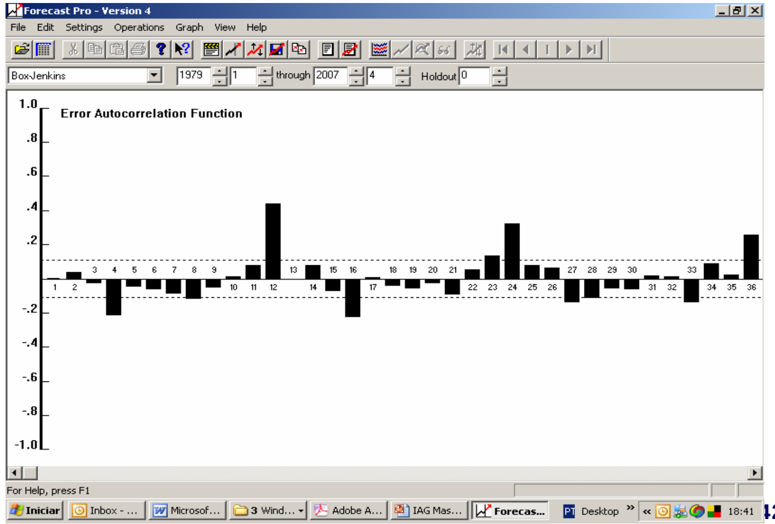
- Resíduos nos instantes t e t-12 têm dependência.

Estatísticas de "Erros"



Pode-se argumentar que talvez tenha "sobrado" alguma autocorrelação nos resíduos nos lags 4 e 16, após o ajuste do modelo.

Estatísticas de "Erros"

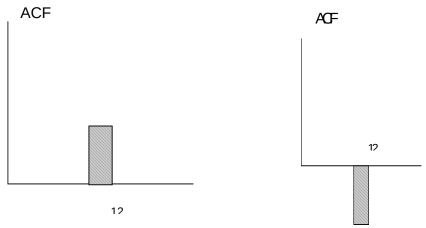


Estatísticas de "Erros"

- Neste caso, e sabendo que os dados são mensais, certamente esquecemos de "algo" na concepção do modelo.
- Note as fortíssimas autocorrelações dos erros.
- Certamente "sobrou" estrutura nos erros – isso não é bom.

Estatísticas de "Erros"

- Outros Padrões:



- Resíduos nos instantes t e t-12 têm dependência.

Estatísticas de "Erros"

- Em resumo...
- Se o modelo é bom o ruído é branco, o resíduo não tem estrutura.
- Se houver padrões no resíduo fique esperto!