

Metodologia Box & Jenkins para escolha de modelos ARIMA

- Dada uma série temporal, uma vez identificado o modelo $ARIMA(p,d,q)$ adequado, através da FAC/FACP e critérios de informação, não segue necessariamente que o modelo identificado deva ser mantido e utilizado para previsão.
- A razão principal é que a **FAC** e a **FACP amostrais** podem apresentar “falhas” na identificação da ordem “correta” do processo gerador da ST.
- Duas das principais razões destas falhas são:
 - i. **tamanho da amostra pequeno (T)**: as estimativas da FAC e FACP ficam pouco precisas, e assim sendo as FAC e FACP estimadas tornam-se estimativas não confiáveis das suas contrapartes teóricas.
 - ii. **influência desproporcional de um ou mais par de observações** (y_t, y_{t-k}) na estimativa da FAC/FACP, podendo resultar em valores espúrios, o que pode levar a identificação errônea da verdadeira ordem do modelo para a ST.

iii. o processo gerador dos dados pode ter componente de dependência não-linear, e o modelo ARIMA apenas captura a dependência linear.

- Como consequência, um modelo “mal” identificado pode apresentar algumas características estatísticas indesejáveis, tais como:

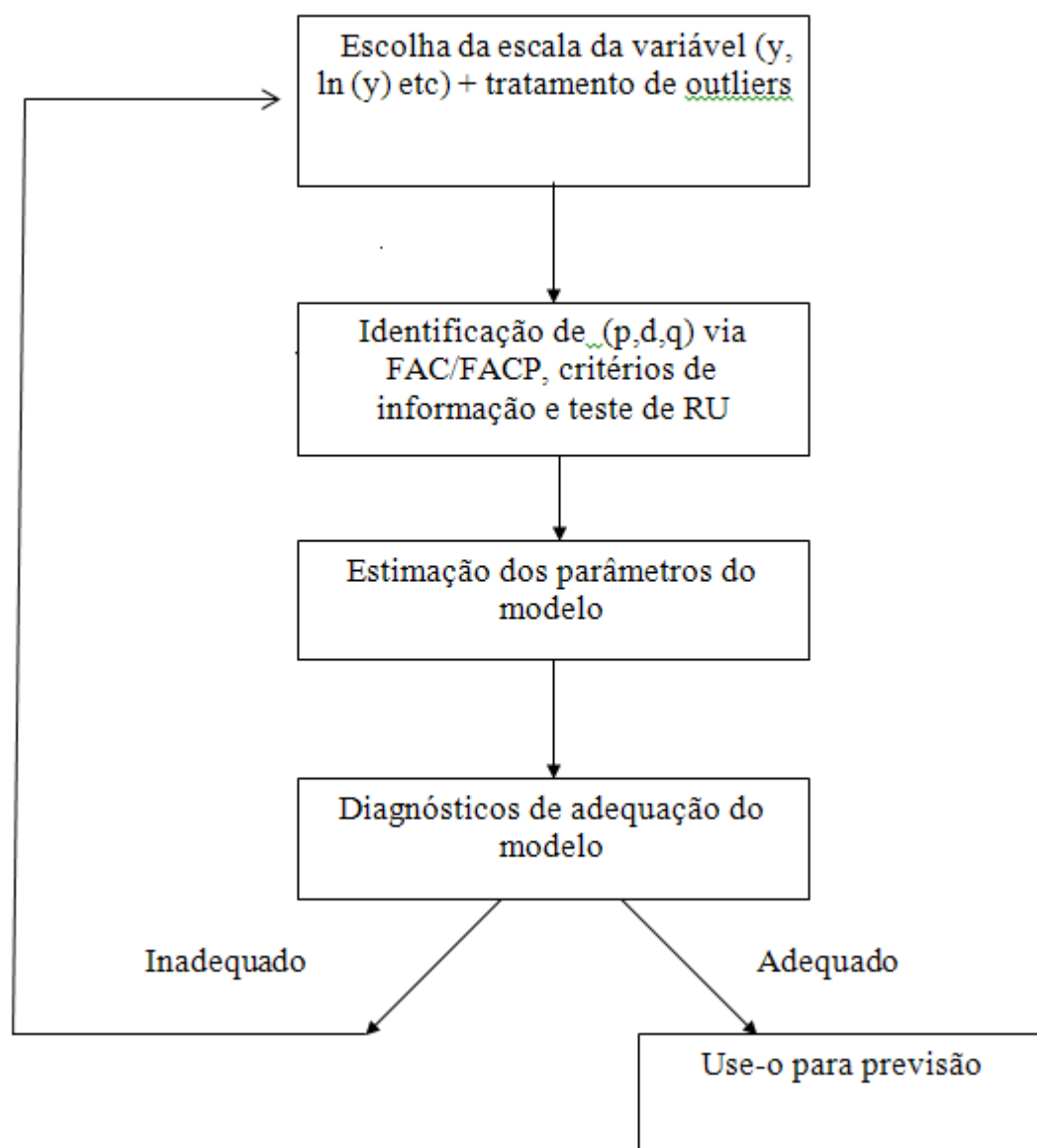
- alguns parâmetros podem ser estatisticamente insignificantes, implicando que o modelo identificado pode ser simplificado;

- alguns pressupostos referentes às propriedades probabilísticas do termo aleatório podem estar sendo violados. Por exemplo, erros não normais ou com variância não constante invalidam estatísticas de teste, intervalos de confiança e testes de hipótese.

- Assim, na prática, o processo de identificação não representa a etapa final na procura do melhor modelo para uma ST, tendo que ser complementado por outros procedimentos de **diagnóstico**, até que um modelo satisfatório seja encontrado.

- Este processo de natureza iterativa que compreende a identificação do modelo, estimação e diagnóstico dos resultados do melhor modelo $ARIMA(p,d,q)$ com repetição ou não do passo de identificação, define a **Metodologia Box & Jenkins para modelos ARIMA**, apresentada no fluxograma a seguir.
- As fases de identificação e estimação foram detalhadas em seções anteriores. As páginas seguintes serão dedicadas aos diagnósticos estatísticos da adequação do modelo identificado e estimado à ST.

Esquema da metodologia Box & Jenkins



- Os diagnósticos do modelo são em geral efetuados sob os “resíduos”, ou inovações estimadas do modelo, as quais são dadas por:

$$\hat{\varepsilon}_t = y_t - \hat{y}_{t|t-1}$$

em que $\hat{y}_{t|t-1} = E(y_t | Y_{t-1})$.

- Às vezes é mais adequado realizar os diagnósticos nos resíduos padronizados, dados por:

$$\hat{\varepsilon}_t^* = \hat{\varepsilon}_t / \hat{\sigma}$$

>> melhor para detectar outliers

- Na seqüência iremos apresentar os **testes para diagnósticos** mais utilizados na prática:
 - normalidade: Jarque-Bera
 - autocorrelação: Ljung Box e Breusch-Pagan
 - variância constante

Teste de Jarque-Bera para normalidade dos resíduos

A hipótese de normalidade do termo aleatório num modelo ARIMA é importante por diversas razões:

- é crucial no processo de estimação pontual e das variâncias dos estimadores, intervalos de confiança e testes de hipótese.
- é utilizada nos testes significância da FAC, FACP e CI's.

A normalidade será investigada através de dois coeficientes associados a distribuições de probabilidade:

- **assimetria** → S (*skewness*)

- **curtose** → K (*kurtosis*)

- **coeficiente de assimetria** = é uma medida de assimetria de uma densidade de probabilidade em relação à sua média.

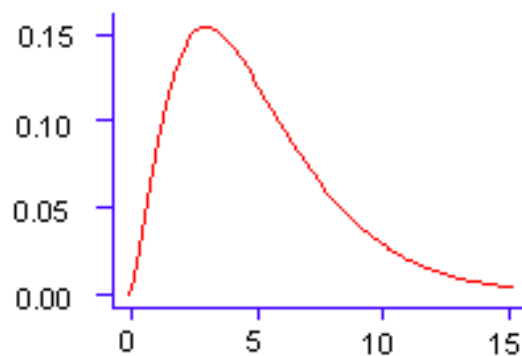
$$S = \frac{E(w_t - \mu)^3}{\sigma^3}$$

- Se a variável aleatória possuir densidade simétrica em torno da média, como é o caso da densidade

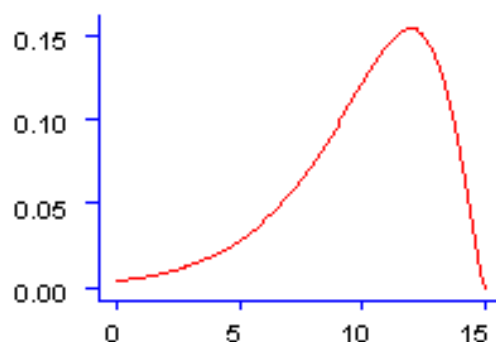
de probabilidade **normal**, prova-se que $E(w - \mu)^3 = 0$, pois a integral abaixo será nula.

$$E(w_t - \mu)^3 = \int_{-\infty}^{+\infty} (w_t - \mu)^3 f(w) dw$$

- Se $S > 0$, a variável aleatória tem assimetria positiva, com maior concentração de massa à esquerda da média, como a distribuição hipotética do exemplo abaixo.



- Se $S < 0$, a variável aleatória tem assimetria negativa, com maior concentração de massa à direita da média, como a distribuição hipotética do exemplo abaixo.



- Na prática, para um conjunto de observações, calculamos a estimativa amostral de S, dada por:

$$\hat{S} = (1/T) \sum_{t=1}^T (w_t - \bar{w})^3 / \left((1/T) \sum_{t=1}^T (w_t - \bar{w})^2 \right)^{3/2}$$

- **coeficiente de curtose**= mede simultaneamente o achatamento da densidade em torno do seu centro e a largura das caudas.

$$K = \frac{E(w_t - \mu)^4}{\sigma^4}$$

- Se a variável aleatória possuir densidade normal, então a integral

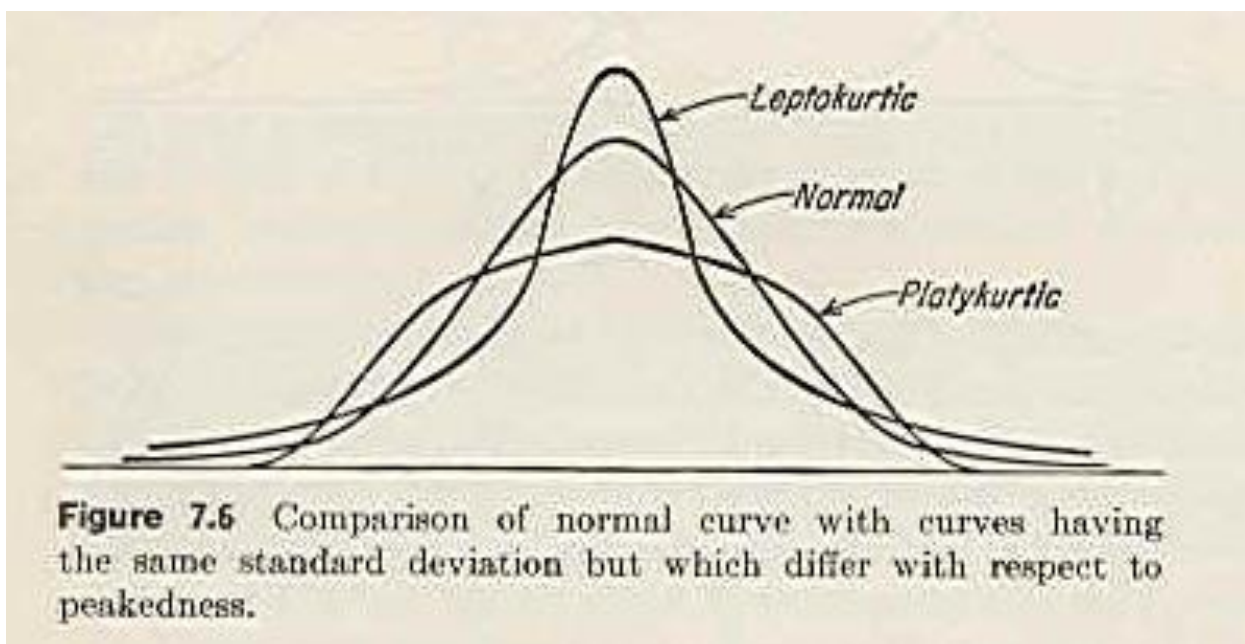
$$E(w_t - \mu)^4 = \int_{-\infty}^{+\infty} (w_t - \mu)^4 f(w) dw$$

é igual a $3\sigma^4$ e portanto a curtose da densidade normal padrão será $K=3$ (densidades **mesocúrticas**).

- Se $K > 3$, a variável aleatória é dita **leptocúrtica**, de caudas grossas, ou com excesso de curtose, possuindo assim mais concentração de massa de probabilidade nas caudas e no centro em relação à

densidade normal. Estas densidades produzem eventos extremos com maior probabilidade do que a normal e são muito importantes na avaliação de risco de mercado (Exs: densidade t, GED)

- Se $K < 3$, a variável aleatória é dita **platicúrtica**, com caudas que caem mais rapidamente do que a normal e mais achatada no centro do que a normal. (Exemplo: densidade uniforme)



- Na prática, para um conjunto de observações calculamos a estimativa amostral de K , dada por:

$$\hat{K} = (1/T) \sum_{t=1}^T (w_t - \bar{w})^4 / \left((1/T) \sum_{t=1}^T (w_t - \bar{w})^2 \right)^2$$

- O teste de **Jarque-Bera** investiga se séries temporais de observações ou de resíduos são

oriundas de uma variável aleatória com distribuição normal, utilizando que sob normalidade, $S=0$ e $K=3$.

- Formalmente:

- H_0 : $S=0$ e $K=3$ (série/resíduos seguem distribuição normal)

- H_a : cc \rightarrow série/resíduos não seguem distribuição normal.

- Demonstra-se, sob a hipótese nula, que as distribuição assintóticas dos estimadores de S e K possuem a seguinte forma:

$$\hat{S} \sim N(0, 6/n) \text{ e } \hat{K} \sim N(3, 24/n)$$

- Padronizando:

$$(\hat{S} - 0) / \sqrt{6/n} \sim N(0, 1)$$

$$\left((\hat{S} - 0) / \sqrt{6/n} \right)^2 = (n/6) (\hat{S} - 0)^2 \sim \chi^2(1)$$

$$(\hat{K} - 3) / \sqrt{24/n} \sim N(0, 1)$$

$$\left((\hat{K} - 3) / \sqrt{24/n} \right)^2 = (n/24) (\hat{K} - 3)^2 \sim \chi^2(1).$$

- Demonstra-se ainda que os dois momentos da distribuição são independentes. Por fim, sabe-se que a soma de **k** variáveis aleatórias Z elevadas ao quadrado segue uma distribuição Qui-quadrada com **k** graus de liberdade. Chega-se assim a estatística de teste para esse teste:

$$JB = (n/6) (\hat{S} - 0)^2 + (n/24) (\hat{K} - 3)^2 \sim \chi^2(2)$$

- Se **JB** > valor crítico a 100α % de $\chi^2(2)$, então se rejeita H_0 .

Problemas com o teste JB

- A aproximação normal para a distribuição da curtose estimada somente é válida para amostras muito grandes, de tamanho superior a 1000 observações. Portanto, a distribuição Qui-quadrada, usada nos testes, deve ser utilizada com cautela.
- Os autores, através de estudos de simulação, puderam obter os valores críticos para o teste sobre os resíduos para vários tamanhos de amostra:

T	5%	10%
30	3.71	2.49
50	4.26	2.90
75	4.27	3.09
100	4.29	3.14

300	4.60	3.68
500	4.82	3.91
∞	5.99	4.61

- Deb & Sefton (96) obtiveram os valores críticos do teste para estatísticas amostrais obtidas a partir dos dados originais

T	5%	10 %
20	3.77	2.33
50	5.00	3.19
75	5.30	3.49
100	5.44	3.67
200	5.71	4.05
500	5.89	4.35
∞	5.99	4.61

Teste para significância conjunta de autocorrelações (Ljung-Box):

- O objetivo desse teste é investigar se as primeiras **m** autocorrelações de uma série temporal são conjuntamente significantes estatisticamente (ou seja, se algum subconjunto das autocorrelações é não nulo).
- As hipóteses nula e alternativa para esse teste são:
 - $H_0: \rho(1) = \rho(2) = \dots = \rho(m) = 0$
 - H_a : pelo menos um dos ρ 's $\neq 0$
- A estatística de teste é:

$$LB(m) = T(T+2) \sum_{k=1}^m \hat{\rho}^2(k) / (T-k) \sim \chi^2(\nu)$$

- Os graus de liberdade ν , dependem se o teste é aplicado na ST original ou nos resíduos do modelo ARMA(p,q).
 - observações originais: $\nu = m$.
 - resíduos: $\nu = m - (p + q)$

- O teste é realizado comumente sobre os resíduos de um modelo ajustado.
- Rejeitar a hipótese nula deste teste implica na existência de autocorrelação nos resíduos do modelo ajustado.
- Portanto, o modelo é inadequado para descrever a série temporal e deverá ser modificado.

⇒ O teste de Ljung-Box possui dois inconvenientes:

- Para valores elevados de k , o teste pode apresentar baixa potência, pois haverá poucas observações para a estimação.
- O teste apenas indica se o modelo é inadequado, mas não sugere como o modelo deve ser modificado, na rejeição de H_0 .

Teste tipo multiplicador de Lagrange para um conjunto de autocorrelações (Breusch-Godfrey)

- Alternativa ao teste de Ljung-Box, possuindo algumas vantagens sobre esse:
 - estudos de simulação mostram que o teste Breusch-Godfrey é mais potente do que o teste Ljung-Box e tende a ter mais sensibilidade em detectar a presença de autocorrelação nos dados.
 - se utilizado de forma inteligente pode fornecer indicações sobre como o modelo deve ser corrigido.
- Esse teste é facilmente implementado, sendo automaticamente realizado pelo EViews.
- Suponha que um modelo AR(p) tenha sido estimado para a sua série. Para testar se essa especificação é satisfatória para a sua série deve-se estimar o seguinte modelo para a série de resíduos do seu modelo, e_t :

$$e_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \dots + \alpha_p y_{t-p} + \gamma_1 e_{t-1} + \gamma_2 e_{t-2} + \dots + \gamma_p e_{t-p} + \varepsilon_t$$

- Ou seja, na primeira parte da equação colocam-se os termos que foram incluídos no modelo, e depois

o resíduo com as defasagens que estão no seu modelo.

- O teste BG possui as seguintes hipóteses:
 - H_0 : o modelo estimado é satisfatório, ou seja, a estrutura de autocorrelação assumida é adequada (no caso do exemplo AR(p))

$$\gamma_1 = \gamma_2 = \dots = \gamma_p = 0;$$

- H_a : o modelo não é satisfatório.

onde e_t são os resíduos estimados do modelo original, nesse caso, um AR(p).

- Sob a hipótese nula de que o modelo ajustado é adequado (no caso um AR(p)):

$$nR^2 \sim \chi^2(s)$$

onde n é o número de observações da série e R^2 é o coeficiente de determinação da regressão dos resíduos.

- Obs: se o modelo j teste também pode ser utilizado para $s > p$. Nesse caso

Teste BDS (87) para independência dos resíduos (Brock, Dechert, Sheinkman)

- O teste de Ljung-Box apenas testa se as observações da série ou seus resíduos são descorrelatados.
- Aceitar a hipótese nula no teste LB não necessariamente implica que as observações da série ou seus resíduos estejam desprovidos de algum tipo de dependência, por exemplo, de ordem não linear.
- O teste de BDS é um teste mais geral do que o teste LB, onde a aceitação da sua hipótese nula implica ausência de qualquer estrutura de dependência nos dados.
- As hipóteses nula e alternativa do teste BDS são:
 - H_0 : os dados são i.i.d.
 - H_a : os dados apresentam dependência linear, não-linear (média e/ou variância) ou caos determinístico.

- A estatística de teste é dada por (detalhes mais adiante):

$$BDS(\varepsilon, m) = \sqrt{T} [C_m(\varepsilon) - (C_1(\varepsilon))^m] / V_m^{1/2} \sim N(0, 1)$$

em que V_m é a expressão da variância, cuja fórmula pode ser encontrada em Cromwell *et al.* (1994, pág. 34).

- Brock, Hsieh e Le Baron (1990) recomendam usar ε entre 0.5 e 2 vezes o desvio padrão da série, e o parâmetro dimensional m , entre 2 e 10.
- Esta distribuição assintótica é adequada desde que $T/m > 200$ ($T=n$, # de obs da série).
- Fora deste limite, ou quando a série é de resíduos de modelos do tipo GARCH (para ARMA não precisa) deve-se levantar os valores críticos da estatística BDS através da técnica de *bootstrap* (EViews 4.0).

Detalhes do teste BDS

- A base da construção do teste BDS é a chamada **integral de correlação**, definida pela seguinte expressão:

$$C_m(\varepsilon) = \sum_{t < s} I_\varepsilon(x_t^m, x_s^m) / \binom{T_m}{2} = \left[\frac{2}{(T_m(T_m - 1))} \right] \sum_{t < s} I_\varepsilon(x_t^m, x_s^m)$$

em que:

ε = é a distância entre os vetores;

m = é a dimensão de encaixamento;

$T_m = T - (m-1)$;

$x_t^m = [x_t, x_{t+1}, \dots, x_{t+m-1}]$;

$$I_\varepsilon(x_t^m, x_s^m) = \begin{cases} 1 & , se \ \|x_t^m - x_s^m\| < \varepsilon \\ 0 & , c.c \end{cases}$$

- A IC mede, para um dado m , a fração dos pares de "pontos" $[x_t^m, x_s^m]$, com distância máxima ε um do outro, isto é, $\|x_t^m - x_s^m\| < \varepsilon$.
- Se x_t^m e x_s^m estiverem muito próximos, então a integral de correlação assumirá um valor próximo ou igual a 1; caso contrário, essa assumirá um valor muito próximo de zero.

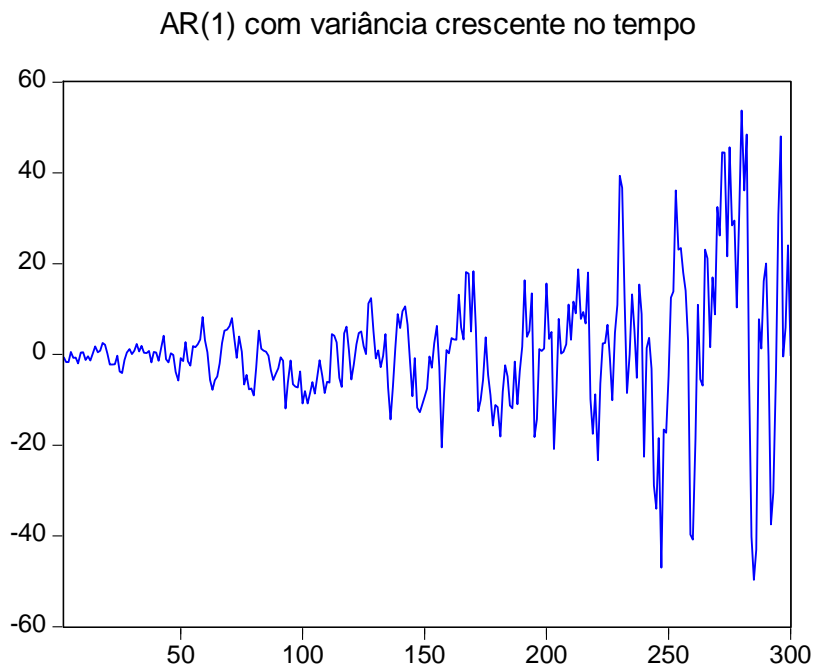
Teste de Homocedasticidade para o erro (variância constante)

- Uma das hipóteses dos modelos ARIMA(p,d,q) é que o termo aleatório tenha variância constante, i.e, que $E(\varepsilon_t^2) = \sigma^2, \forall t$.
- Esta hipótese é utilizada explicitamente na expressão das variâncias dos estimadores dos parâmetros do modelo, e assim na realização dos testes de significância dos parâmetros.
- Também na utilização dos intervalos de confiança para as previsões do modelo.
- Na prática existem duas maneiras principais desta hipótese ser violada:
 - i. se a variância do erro variar monotonicamente com o tempo, através de uma relação heterocedástica incondicional:

Ex: supor um processo AR(1) onde a variância é crescente no tempo

$$E(\varepsilon_t^2) = \sigma^2 \exp(bt) \Rightarrow \text{Var}(y_t) = \frac{\sigma^2 \exp(bt)}{(1 - \phi^2)}, b > 0$$

>> observar que neste caso o processo não mais será estacionário de 2ª ordem.



Para detectar este tipo de comportamento, além da inspeção visual da série, é possível realizar um teste de variância constante:

$$H(h) = \sigma_{final}^2 / \sigma_{início}^2$$

$H_o: H(h) = 1 \Leftrightarrow$ variância cte

$H_a: H(h) > 1 \Leftrightarrow$ variância crescente

$$\hat{\sigma}_{início}^2 = 1/h \sum_{t=1}^h \hat{\varepsilon}_t^2 \quad \hat{\sigma}_{final}^2 = 1/h \sum_{t=h+(T+1-2h)}^T \hat{\varepsilon}_t^2$$

Estatística de teste: $\hat{H}(h) = \hat{\sigma}_{final}^2 / \hat{\sigma}_{início}^2 \sim F(h, h)$,

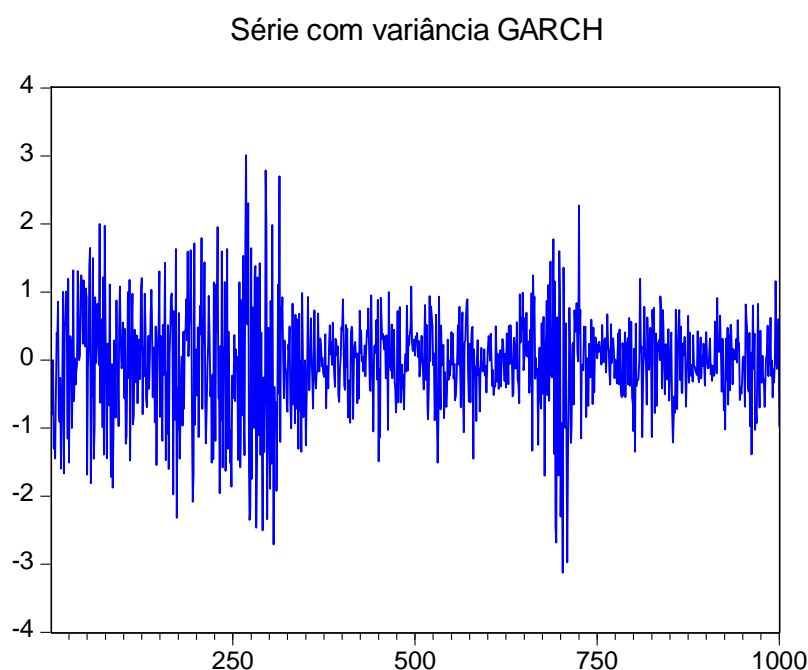
$$h \sim T / 3$$

Se $\hat{H}(h) > F^\alpha(h, h)$, rejeita-se a hipótese de variância constante.

ii. se a variância do erro seguir um processo tipo ARCH/GARCH (heterocedasticidade condicional)

$$E(\varepsilon_t^2 | Y_{t-1}) = \sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \quad (GARCH(1,1))$$

$$\alpha_0 > 0, \alpha_1 \geq 0, \beta_1 \geq 0, \alpha_1 + \beta_1 < 1$$



- Este teste investiga se há presença de volatilidade, ou seja, mudanças na variância condicional da série, testando-a nos resíduos.
- Se houver evidência de volatilidade, então se deve acoplar ao modelo ARIMA um modelo ARCH/GARCH para capturar as mudanças na variância condicional da série.
- A lógica do teste é do tipo multiplicador de Lagrange. As hipóteses nula e alternativa são:

H_0 : não existe efeito ARCH até ordem p nos resíduos

H_a : existe efeito ARCH

- Esse teste é facilmente implementado, sendo automaticamente realizado pelo EViews.
- Suponha que um modelo ARIMA(p, d, q) tenha sido estimado para a sua série temporal. Para testar que os resíduos possuem efeito ARCH(q) deve-se estimar:

$$\hat{\varepsilon}_t^2 = \alpha_0 + \sum_{i=1}^q \alpha_i \hat{\varepsilon}_{t-i}^2 + v_t$$

em que $\hat{\varepsilon}_t$ são os resíduos estimados do modelo ARIMA(p, d, q).

- Sob a hipótese nula de que não existe efeito ARCH até ordem q nos resíduos (para T grande):

$$TR^2 \sim \chi^2(q)$$

em que T é o número de observações e R^2 é o coeficiente de determinação da regressão dos resíduos.

Avaliação da capacidade preditiva de um modelo (*Forecast Evaluation*)

- Suponha que uma série temporal com T observações foi modelada por um determinado modelo dentro da classe ARIMA (p,d,q) .
- Idealmente devem ser utilizados dois períodos disjuntos para avaliar a capacidade preditiva de modelos:

-in-sample ou treinamento = a parte da amostra utilizada para estimar os parâmetros do modelo.

-out-of-sample, teste ou validação = a parte da amostra utilizada para validar o modelo estimado no período de treinamento.

- As medidas de capacidade preditiva, ou aderência do modelo aos dados estão baseadas nos resíduos ou erro de previsão um passo à frente $e_t = y_t - \hat{y}_{t|t-1}$ $t = 1, 2, \dots, m$.

- Algumas medidas de aderência (*goodness of fit*) usuais:

nome	expressão
i. Erro percentual absoluto médio (MAPE)	$MAPE = \frac{1}{m} \sum_{t=T-m+1}^T \left \frac{e_t}{y_t} \right \cdot 100$
ii. Coeficiente de determinação	$R^2 = [\text{corr}(y_t, \hat{y}_{t t-1})]^2, \quad 0 < R^2 < 1$
iii. Raiz do erro quadrático médio	$RMSE = \sqrt{\frac{1}{m} \sum_{t=T-m+1}^T e_t^2}$

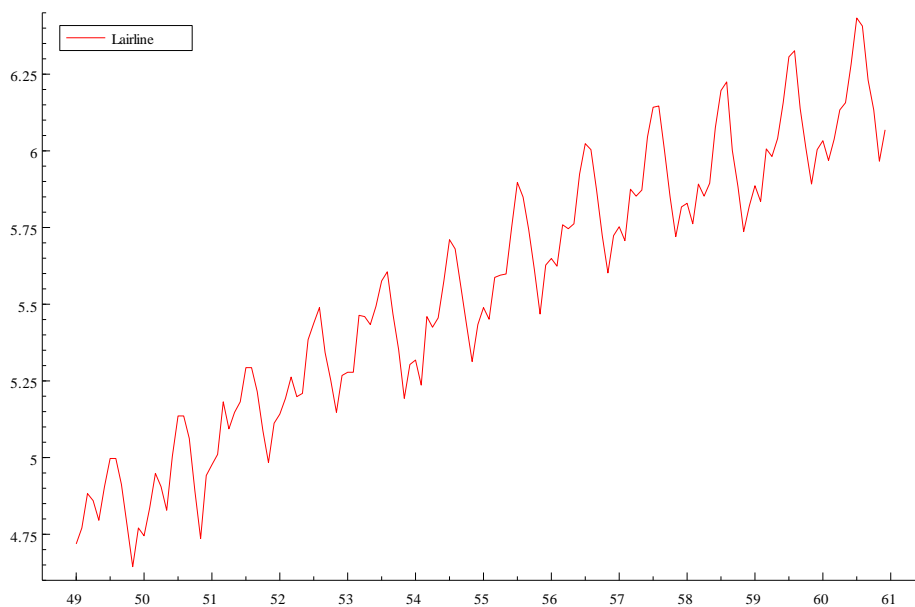
- A medida (i) é comparável entre modelos de escala diferente, enquanto que as medidas (ii) e (iii) não o são. Assim sendo, se queremos comparar um modelo p/ y_t com outro p/ $\log y_t$, só podemos utilizar a medida (i).
- Estas medidas são utilizadas tanto no período “dentro da amostra” como “fora da amostra”.

Sazonalidade nos processos ARIMA

- Muitas séries temporais econômicas, financeiras, físicas (meteorologia, hidrologia) e biológicas apresentam sazonalidade.
- Sazonalidade são flutuações periódicas que ocorrem num período máximo de um ano, e estão relacionadas a variações climáticas (estações do ano, ciclo dia-noite), convenções sociais (Carnaval, Páscoa, Natal, Dia das Mães, São João etc).
- **S** é o período da sazonalidade: o “tempo” que a série leva para se repetir dentro de um período máximo de 1 ano.
 - $s = 2$, séries semestrais;
 - $s = 4$, séries trimestrais;
 - $s = 12$, séries mensais ;
 - $s = 52$, séries semanais;
 - $s = 365$, séries diárias
- Em muitas circunstâncias, o termo sazonal pode ser visto como algo indesejável, pois pode obscurecer a visualização de outros fatores de interesse da série temporal, como por exemplo, a tendência.
- Denomina-se de ajuste sazonal ou de-sazonalização o processo de retirada/filtragem do

termo sazonal de uma série temporal de forma a se obter uma série livre das flutuações sazonais.

- Nos modelos lineares, a maneira padrão para filtragem da componente sazonal de uma série é considerar essa componente como **aditiva** às componentes de tendência e irregulares da série.



$$y_t = Tend + Saz + Irreg$$

$$y_t = \mu_t + \gamma_t + \varepsilon_t$$

- Para se obter uma série sazonalmente ajustada, é necessário que o modelo trate explicitamente a componente sazonal.
- A série sazonalmente ajustada teria portanto a seguinte forma:

$$y_t^{(a)} = y_t - \hat{\gamma}_t = \hat{\mu}_t + \hat{\varepsilon}_t$$

- O tratamento da componente sazonal pode ser efetuado por dois grupos de modelos, a saber:
 - *Sazonalidade determinística* (regressão)
 - variáveis *dummies*
 - funções trigonométricas
 - *Sazonalidade estocástica* (SARIMA)
 - variável endógena ou erros defasados: y_{t-s} , ε_{t-s} .

Sazonalidade com variáveis dummies

- Nesse modelo de simples implementação, o coeficiente de cada variável *dummy* representa o fator sazonal (mês, trimestre, etc.) de interesse.

Exemplo: série com frequência trimestral ($s=4$).

$$y_t = a_0 + \gamma_1 d_{1,t} + \gamma_2 d_{2,t} + \gamma_3 d_{3,t} + \gamma_4 d_{4,t} + \varepsilon_t$$

$$d_{i,t} = \begin{cases} 1 & , \text{se } t = i, s+i, 2s+i \\ 0 & , \text{caso contrário} \end{cases}$$

PROBLEMA: este modelo possui multicolinearidade perfeita e portanto, seus parâmetros não podem ser estimados simultaneamente.

- A solução é eliminar uma das variáveis dummies. O "mês" / "trimestre" sem *dummy* é o mês base, ou mês de referência. Assim sendo a flutuação sazonal (positiva ou negativa) será medida em relação a este mês (sempre que o modelo tiver intercepto).
- Por exemplo:

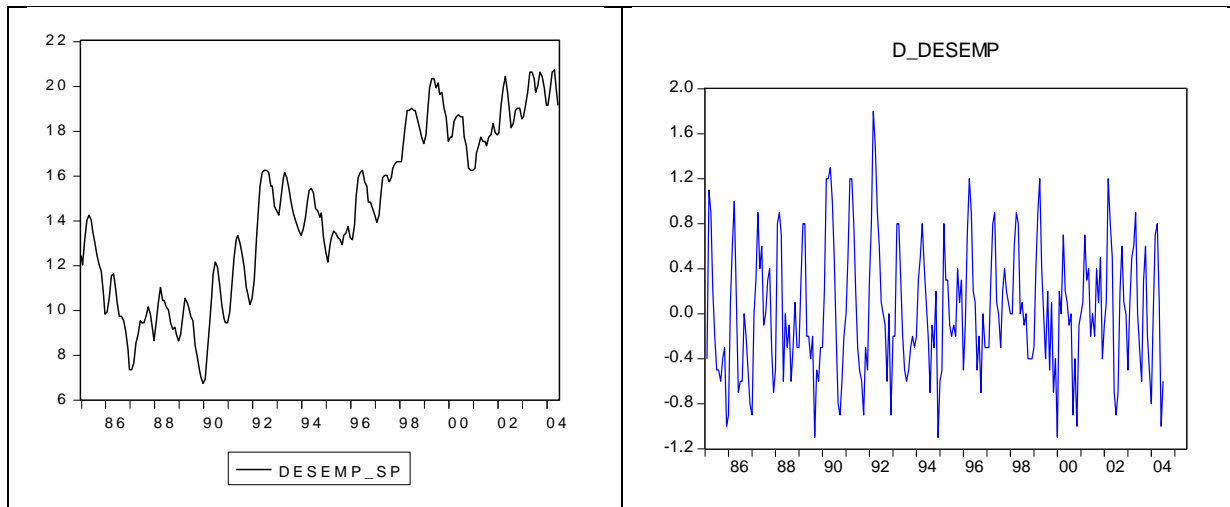
$$y_t = a_0 + \gamma_2 d_{2,t} + \gamma_3 d_{3,t} + \gamma_4 d_{4,t} + \varepsilon_t$$

$$E(y_t | 1^{\circ} \text{ trimestre}) = a_0$$

$$E(y_t | 2^{\circ} \text{ trimestre}) = a_0 + \gamma_2$$

$$\gamma_2 = E(y_t | 2^{\circ} \text{ trimestre}) - E(y_t | 1^{\circ} \text{ trimestre})$$

Exemplo: série da taxa de desemprego mensal e 1ª diferença (Região Metropolitana de São Paulo-jan/85 à jul/04 – DIEESE).



- Observem que no gráfico da série no tempo a sazonalidade se confunde com as variações da tendência da série, sendo as flutuações sazonais bem mais aparentes na série da 1ª diferença.
- A sazonalidade pode também ser inferida através da inspeção da FAC estimada da série da 1ª diferenças, mostrada a seguir:

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.576	0.576	78.715	0.000
		2 0.204	-0.191	88.654	0.000
		3 -0.235	-0.412	101.87	0.000
		4 -0.237	0.234	115.34	0.000
		5 -0.239	-0.153	129.13	0.000
		6 -0.215	-0.305	140.31	0.000
		7 -0.277	-0.077	158.99	0.000
		8 -0.290	-0.149	179.57	0.000
		9 -0.212	-0.118	190.58	0.000
		10 0.083	0.304	192.30	0.000
		11 0.411	0.283	234.10	0.000
		12 0.586	0.063	319.54	0.000
		13 0.419	-0.043	363.38	0.000
		14 0.079	-0.120	364.93	0.000
		15 -0.191	-0.073	374.16	0.000
		16 -0.265	-0.034	391.97	0.000
		17 -0.232	-0.061	405.68	0.000
		18 -0.232	-0.080	419.47	0.000
		19 -0.260	0.012	436.85	0.000
		20 -0.268	-0.023	455.34	0.000
		21 -0.142	-0.005	460.57	0.000
		22 0.128	0.135	464.83	0.000
		23 0.414	0.076	509.67	0.000
		24 0.520	-0.016	580.91	0.000
		25 0.370	0.020	616.99	0.000
		26 0.070	-0.046	618.30	0.000

- Vamos agora proceder à modelagem sazonal da série da 1ª diferença da taxa de desemprego utilizando variáveis dummies, e adotando o mês de janeiro como o mês base.
- O modelo será da forma:

$$\Delta y_t = a_0 + \gamma_2 d_{2,t} + \gamma_3 d_{3,t} + \gamma_4 d_{4,t} + \dots + \gamma_{12} d_{12,t} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, \sigma^2)$$

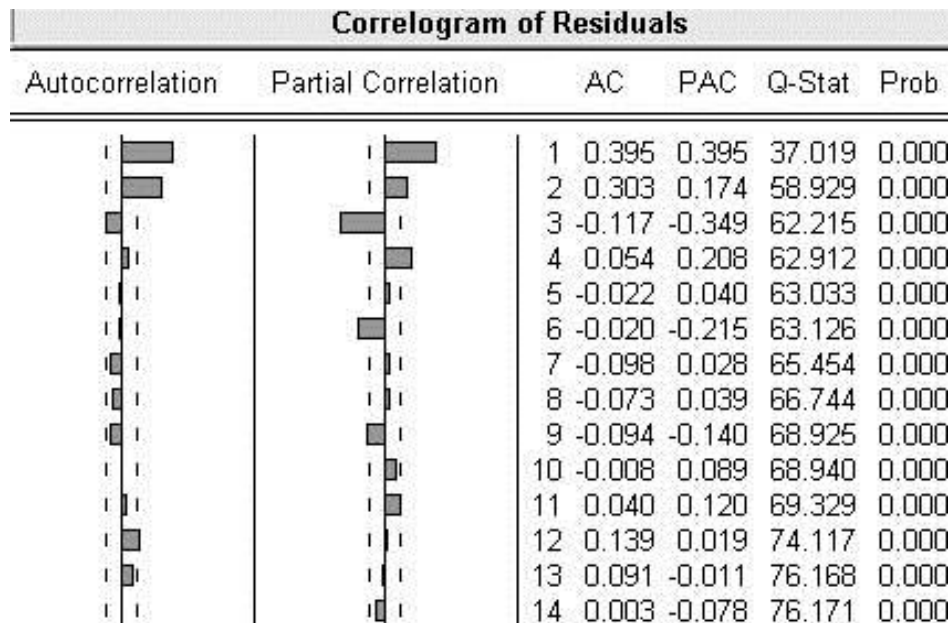
- Os comandos do EViews são:

```
d(desemp_sp)  c  @seas(2)  @seas(3)  @seas(4)
@seas(5)  @seas(6)  @seas(7)  @seas(8)  @seas(9)
@seas(10) @seas(11) @seas(12)
```

- A saída do EViews quando esse modelo é estimado é mostrada a seguir:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.378947	0.088469	-4.283400	0.0000
@SEAS(2)	0.498947	0.123540	4.038752	0.0001
@SEAS(3)	1.158947	0.123540	9.381152	0.0000
@SEAS(4)	1.248947	0.123540	10.10966	0.0000
@SEAS(5)	0.828947	0.123540	6.709952	0.0000
@SEAS(6)	0.358947	0.123540	2.905516	0.0040
@SEAS(7)	0.148947	0.123540	1.205661	0.2292
@SEAS(8)	0.078947	0.125114	0.631004	0.5287
@SEAS(9)	0.126316	0.125114	1.009607	0.3138
@SEAS(10)	0.205263	0.125114	1.640612	0.1023
@SEAS(11)	0.173684	0.125114	1.388210	0.1665
@SEAS(12)	-0.031579	0.125114	-0.252402	0.8010
R-squared	0.564230	Mean dependent var	0.028632	
Adjusted R-squared	0.542637	S.D. dependent var	0.570213	
S.E. of regression	0.385627	Akaike info criterion	0.982026	
Sum squared resid	33.01316	Schwarz criterion	1.159222	
Log likelihood	-102.8971	F-statistic	26.13115	
Durbin-Watson stat	1.197241	Prob(F-statistic)	0.000000	

- Esse modelo ainda não pode ser considerado definitivo pois os resíduos ainda apresentam autocorrelação:



- Devemos, então, reespecificar o modelo, adicionando um erro com estrutura ARMA adequada para capturar a autocorrelação presente nos resíduos.
- Portanto, o modelo adequado para a série da 1ª diferenças da taxa de desemprego mensal será dado por:

$$\Delta y_t = a_0 + \gamma_2 d_{2,t} + \gamma_3 d_{3,t} + \gamma_4 d_{4,t} + \dots + \gamma_{12} d_{12,t} + \varepsilon_t$$

$$\Phi_p(L) \varepsilon_t = \Theta_q(L) \eta_t, \quad \eta_t \sim \text{NID}(0, \sigma^2)$$

$$\Theta_q(L) = 1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q$$

$$\Phi_p(L) = 1 - \varphi_1 L - \varphi_2 L^2 - \dots - \varphi_p L^p$$

- A FAC e FACP dos resíduos sugerem um modelo MA (2) ou MA (3). O MA (3) foi escolhido pois minimiza o AIC.



























- A sintaxe do modelo a ser estimado no EViews será:

```
d(desemp_sp) c @seas(2) @seas(3) @seas(4)
@seas(5) @seas(6) @seas(7) @seas(8) @seas(9)
@seas(10) @seas(11) @seas(12) ma(1) ma(2)
ma(3)
```

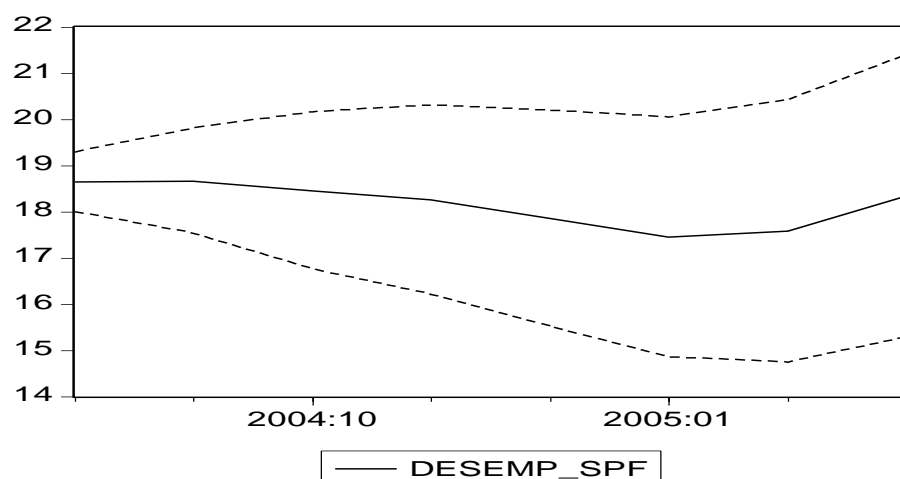
- O resultado do ajuste do modelo é dado a seguir:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.388791	0.088108	-4.412656	0.0000
@SEAS(2)	0.520096	0.093637	5.554369	0.0000
@SEAS(3)	1.167428	0.105748	11.03972	0.0000
@SEAS(4)	1.257698	0.131219	9.584741	0.0000
@SEAS(5)	0.838791	0.123332	6.801066	0.0000
@SEAS(6)	0.368791	0.123330	2.990275	0.0031
@SEAS(7)	0.158791	0.123330	1.287531	0.1993
@SEAS(8)	0.080740	0.124331	0.649390	0.5168
@SEAS(9)	0.150081	0.124593	1.204565	0.2297
@SEAS(10)	0.212618	0.132812	1.600892	0.1108
@SEAS(11)	0.188330	0.106601	1.766680	0.0787
@SEAS(12)	-0.035483	0.094226	-0.376570	0.7069
MA(1)	0.521503	0.065580	7.952187	0.0000
MA(2)	0.540334	0.065150	8.293733	0.0000
MA(3)	-0.209991	0.066665	-3.149960	0.0019
R-squared	0.732242	Mean dependent var	0.028632	
Adjusted R-squared	0.715125	S.D. dependent var	0.570213	
S.E. of regression	0.304343	Akaike info criterion	0.520636	
Sum squared resid	20.28486	Schwarz criterion	0.742131	
Log likelihood	-45.91438	F-statistic	42.77876	
Durbin-Watson stat	2.036700	Prob(F-statistic)	0.000000	

- Observe que o sinal e significância dos parâmetros sazonais permitem-nos afirmar que:
 - em relação à variação da taxa de janeiro, a variação mensal na taxa de desemprego é maior nos primeiros meses do ano, sendo abril o mês de maior variação da taxa de desemprego.
 - a variação da taxa de desemprego nos meses de julho a dezembro não são estatisticamente discerníveis da variação da taxa em janeiro.
- O correlograma dos resíduos, apresentado abaixo, indica que o modelo é satisfatório, pois os resíduos se comportam como ruído branco.

Correlogram of Residuals						
Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob	
		1	-0.021	-0.021	0.1039	
		2	0.039	0.039	0.4746	
		3	-0.018	-0.016	0.5524	
		4	0.047	0.045	1.0753	0.300
		5	-0.029	-0.026	1.2844	0.526
		6	0.015	0.010	1.3355	0.721
		7	-0.066	-0.062	2.3794	0.666
		8	-0.048	-0.055	2.9441	0.709
		9	-0.029	-0.024	3.1475	0.790
		10	-0.007	-0.008	3.1591	0.870
		11	-0.032	-0.026	3.4114	0.906
		12	0.117	0.119	6.8384	0.654
		13	0.082	0.092	8.5225	0.578
		14	-0.027	-0.036	8.7069	0.649

- Para obter previsões fora da amostra para a taxa de desemprego (a série original), temos os seguintes comandos do EViews:
 - inicialmente observar que a série de desemprego se inicia em jan 1985 e vai até julho de 2004.
 - aumentamos o *range* e o *sample* (nesta ordem), clicando a seta indicadora nos seus respectivos nomes que aparecem no canto superior esquerdo da janela *workfile*.
 - nas janelas que se abrem, colocamos uma data final que englobe o horizonte de previsão desejado. Por exemplo, se queremos obter previsões de agosto de 2004 até março de 2005, colocamos: 2005.03.
 - em seguida na janela que possui os resultados da estimação do modelo selecionamos a opção *Forecast* no canto superior direito.
 - na janela que se abre, em *Forecast Sample* colocamos 2004:08 2005:03.
 - é então criada a variável com as previsões (a qual possui o nome da variável original acrescida da letra *f* no final) e um gráfico das previsões com intervalos de confiança de 95%.



- A título de ilustração apresentamos uma tabela como os valores previstos e os atuais, fornecidos pelo DIEESE, na tabela subsequente.

Mês	previsão	atual
Ago.04	18.64	18.3
Set.04	18.66	17.9
Out.04	18.45	
Nov.04	18.26	
Dez.04	17.85	
Jan.05	17.44	
Fev.05	17.57	
Mar.05	18.35	

Resultado da Pesquisa

[[Altera pesquisa](#)]

[Região Metropolitana de São Paulo](#)

[Desemprego](#)

[Taxas de Desemprego Total \(em porcentagem\)](#)

Total

Anos	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
2004	19.1	19.8	20.6	20.7	19.7	19.1	18.5	18.3	17.9			

URL: <http://www.seade.gov.br>

- Observar que no EViews existem duas opções p/*forecast*:
 - *Static* - calcula a previsão 1 passo à frente, utilizando os valores reais da série. É adequado para *in-sample forecasts*.
 - *Dynamic* - calcula a previsão k passos à frente, iniciando a partir da 1ª data do *forecast sample*. É a escolha adequada para previsão *out-of-sample*.

Sazonalidade com funções trigonométricas

- A sazonalidade também pode ser capturada através de funções periódicas, as quais podem ser representadas através de uma série de Fourier.
- Teorema de Fourier (p/ seqüências)

- se $f(t)$ é uma seqüência periódica com período s , então $f(t)$ admite a seguinte expansão em séries trigonométricas (série de Fourier):

$$f(t) = a_0 + \sum_{j=1}^{[s/2]} (a_j \cos \omega_j t + b_j \sin \omega_j t), \quad t = 1, 2, \dots, T$$

$$\omega_j = \frac{2\pi j}{s} \quad [s/2] = \begin{cases} s/2, & s \text{ par} \\ (s+1)/2, & s \text{ ímpar} \end{cases}$$

- Portanto, sob a hipótese de que o resíduo deste modelo pode conter autocorrelação, uma ST estacionária com sazonalidade por termos trigonométricos pode ser representada pelo seguinte modelo geral:

$$\Delta^d y_t = a_0 + \sum_{j=1}^{[s/2]} (a_j \cos \omega_j t + b_j \sin \omega_j t) + \varepsilon_t$$

$$\Phi_p(L) \varepsilon_t = \Theta_q(L) \eta_t, \quad \eta_t \sim \text{NID}(0, \sigma^2)$$

$$\Theta_q(L) = 1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q$$

$$\Phi_p(L) = 1 - \varphi_1 L - \varphi_2 L^2 - \dots - \varphi_p L^p$$

- Observem que, ao contrário do tratamento por *dummies*, o fator sazonal associado ao “mês” t , não pode ser lido diretamente do modelo, tendo que ser avaliado.
- Assim sendo, para o “mês” t , o fator sazonal será dado por:

$$\gamma(t) = \sum_{j=1}^{[S/2]} (a_j \cos \omega_j t + b_j \sin \omega_j t)$$

- A implementação destes modelos no EViews será deixada como exercício.

Modelo SARIMA(p,d,q)(P,D,Q)_s

- Inicialmente, considere um modelo AR com *lag* no período sazonal, assumido como $s = 12$:

$$y_t = \varphi_{12} y_{t-12} + \eta_t, \quad |\varphi_{12}| < 1$$

$$(1 - \varphi_{12} L^{12}) y_t = \eta_t$$

- Este é um modelo AR (12) com parâmetros intermediários $\varphi_j = 0$, $j = 1, 2, \dots, 11$.
- A FAC do modelo é dada por:

$$\rho(k) = \varphi_{12}^{k/12}, \quad k = 12, 24, 36, \dots$$

que terá picos (decrecentes) apenas nos *lags* 12, 24, 36,...

- É fácil de ver que o modelo anterior pode ser generalizado através da seguinte expressão:

$$\Phi_P(L^s) y_t = \Theta_Q(L^s) \eta_t \quad (I)$$

onde:

$$- \Theta_Q(L^s) = 1 + \Theta_1 L^s + \Theta_2 L^{2s} + \dots + \Theta_Q L^{Qs}$$

$$- \Phi_P(L^s) = 1 - \Phi_1 L^s - \Phi_2 L^{2s} - \dots - \Phi_P L^{Ps}$$

- **Exemplo:** $s=12, P= Q = 2$.

$$(1 - \Phi_1 L^{12} - \Phi_2 L^{24}) y_t = (1 + \Theta_1 L^{12} + \Theta_2 L^{24}) \eta_t$$

$$y_t = \Phi_1 y_{t-12} + \Phi_2 y_{t-24} + \Theta_1 \varepsilon_{t-12} + \Theta_2 \varepsilon_{t-24} + \eta_t$$

- Na prática, o modelo acima será apenas adequado se apenas existir dependência sazonal na série, pois não incorpora a presença de dependência de curta duração na série.
- Isto poderia ser verificado formalmente obtendo a forma da FAC e FACP, e observando que nos lags diferentes de $s, 2s, 3s, \dots$ que os valores são nulos.
- Assim sendo, ao utilizarmos este modelo na prática os resíduos apresentarão estrutura de dependência, implicando que os erros η_t não são ruídos brancos, e assim sendo podem ser modelados por um modelo ARMA, i.e. :

$$\varphi_p(L) \eta_t = \theta_q(L) \varepsilon_t \therefore \eta_t = [\theta_q(L) / \varphi_p(L)] \varepsilon_t$$

- Finalmente, substituindo esta expressão na eq(I), obtemos o modelo SARIMA $(p,q) \times (P,Q)_s$, um modelo onde a parte sazonal e não-sazonal são combinadas multiplicativamente:

$$\Phi_P(L^S) \phi_p(L) y_t = \Theta_Q(L^S) \theta_q(L) \varepsilon_t \quad (II)$$

Deve-se notar que, em se tratando de séries com tendência estocástica e sazonalidade, a não-estacionariedade se manifesta através de duas formas:

- Através da tendência estocástica que pode ser removida por um número adequado de diferenças:

$$z_{1t} = \Delta^d y_t = (1 - L)^d y_t$$

- A sazonalidade implica que observações $y_t, y_{t-s}, y_{t-2s}, \dots$, apresentarão “picos”, induzindo não-estacionariedade, a qual pode ser removida pela operação de diferenças sazonais:

$$z_{2t} = (1 - L^s) y_t = y_t - y_{t-s}$$

ou de forma geral:

$$z_{2t} = \Delta_s^D y_t, \quad \Delta_s^D = (1 - L^s)^D \quad (D \text{ geralmente é } 1)$$

- Combinando as duas operações anteriores podemos definir o modelo **SARIMA (p,d,q)(P,D,Q)_s** para séries não estacionárias:

$$\Phi_p(L^s) \Phi_p(L) \Delta^d \Delta_s^D y_t = \Theta_q(L^s) \Theta_q(L) \varepsilon_t \quad (III)$$

$$- \Theta_q(L^s) = 1 + \Theta_1 L^s + \Theta_2 L^{2s} + \dots + \Theta_q L^{qs}$$

$$- \Phi_p(L^s) = 1 - \Phi_1 L^s - \Phi_2 L^{2s} - \dots - \Phi_p L^{ps}$$

$$- \Theta_q(L) = 1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q$$

$$- \Phi_p(L) = 1 - \varphi_1 L - \varphi_2 L^2 - \dots - \varphi_p L^p$$

Modelo Airline

- Uma estrutura particular dos modelos SARIMA é o modelo SARIMA (0,1,1) x (0,1,1)₁₂, conhecido como o modelo *Airline*:

$$\Delta \Delta_s y_t = (1 + \Theta_1 L^s)(1 + \theta_1 L) \varepsilon_t$$

$$z_t = \theta_1 \varepsilon_{t-1} + \Theta_1 \varepsilon_{t-12} + \Theta_1 \theta_1 \varepsilon_{t-13} + \varepsilon_t$$

- Pode-se mostrar que a função de autocovariância deste modelo é dada por:

$$\gamma(k) = E(z_t z_{t-k}), \quad k = 0, 1, 2, \dots$$

$$\gamma(0) = \sigma^2(1 + \theta_1^2 + \Theta_1^2 + \theta_1^2 \Theta_1^2)$$

$$\gamma(1) = \sigma^2(\theta_1 + \theta_1 \Theta_1^2)$$

$$\gamma(j) = 0, \quad j = 2, 3, \dots, 10$$

$$\gamma(11) = \sigma^2(\theta_1 \Theta_1)$$

$$\gamma(12) = \sigma^2(\theta_1 + \theta_1^2 \Theta_1)$$

$$\gamma(13) = \sigma^2(\theta_1 \Theta_1)$$

$$\gamma(j) = 0, \quad j > 13$$

- Portanto, a FAC, $\rho(k) = \gamma(k) / \gamma(0)$, tem lags diferentes de zero apenas em $k = 1, 11, 12$ e 13 .

- Comando para estimar o “airline” no EViews:

d (log(y), 1, 12) ma (1) sma (12)

- Observe que um modelo aditivo “equivalente” ao modelo *Airline* seria dado por:

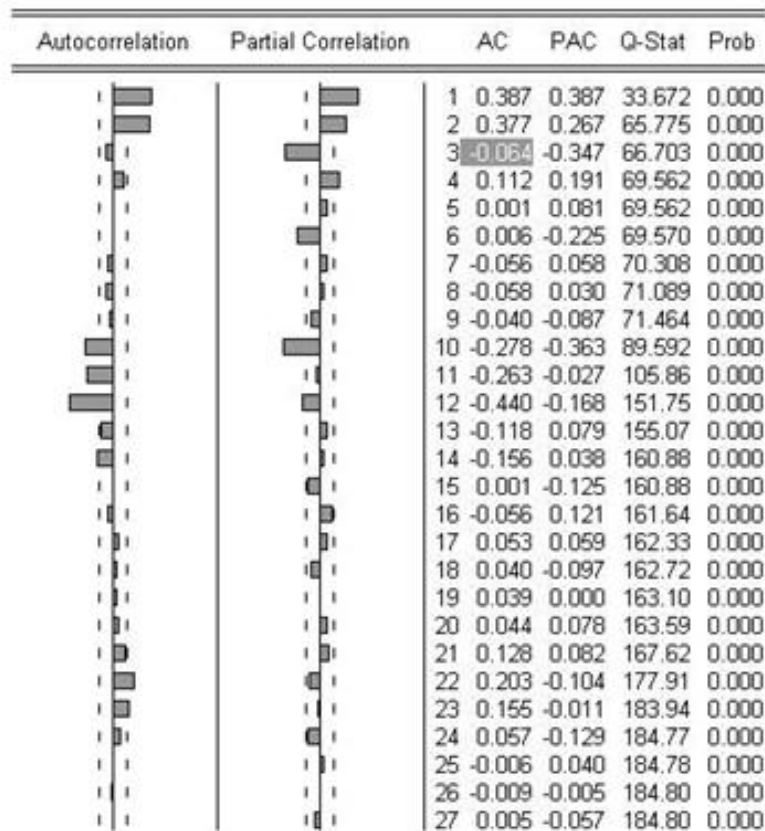
$$\Delta \Delta_s \log(y_t) = (1 + \theta_1 L + \Theta_1 L^{12} + \Theta_2 L^{13}) \varepsilon_t$$

- Deixamos como exercício o cálculo da FAC para este modelo, a qual possui valores diferentes para

os lags 11,12 e 13. Em algumas situações, pode ser mais adequado do que modelos multiplicativos.

- A identificação procede em linhas gerais como nos modelos não sazonais:
 - inicialmente tente utilizar a FAC e FACP para especificar um modelo inicial. O “airline” é sempre um bom ponto de partida;
 - caso necessário complemente o processo de identificação usando AIC e BIC;
 - observe a FAC dos resíduos, nas frequências baixas e nas sazonais.

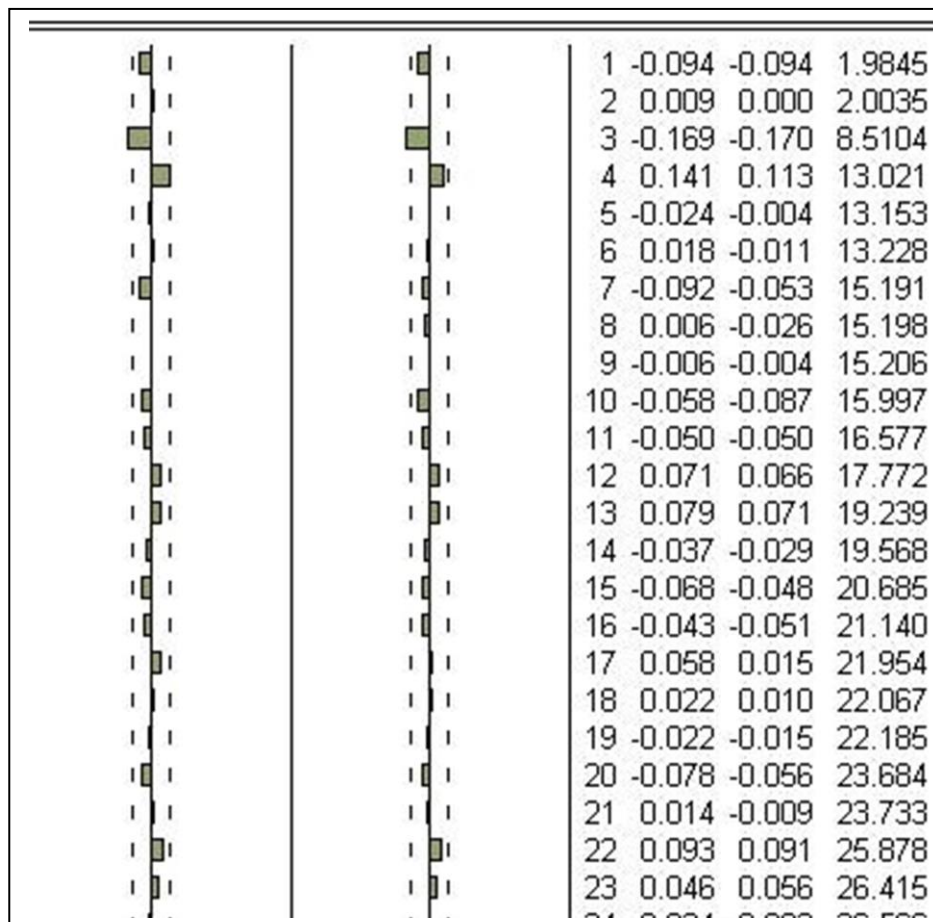
- Aplicação: 1ª diferença da série de desemprego de São Paulo



- Assim sendo, como primeira tentativa iremos estimar um modelo SARIMA(0,1,2) x (0,1,1) pelo EViews:

d(desemp_sp,1,12) ma(1) ma(2) sma(12)

- Observe que tem alguma correlação no lag 3 dos resíduos deste modelo. Vamos aumentar o lag da parte não sazonal para um $ma(3)$



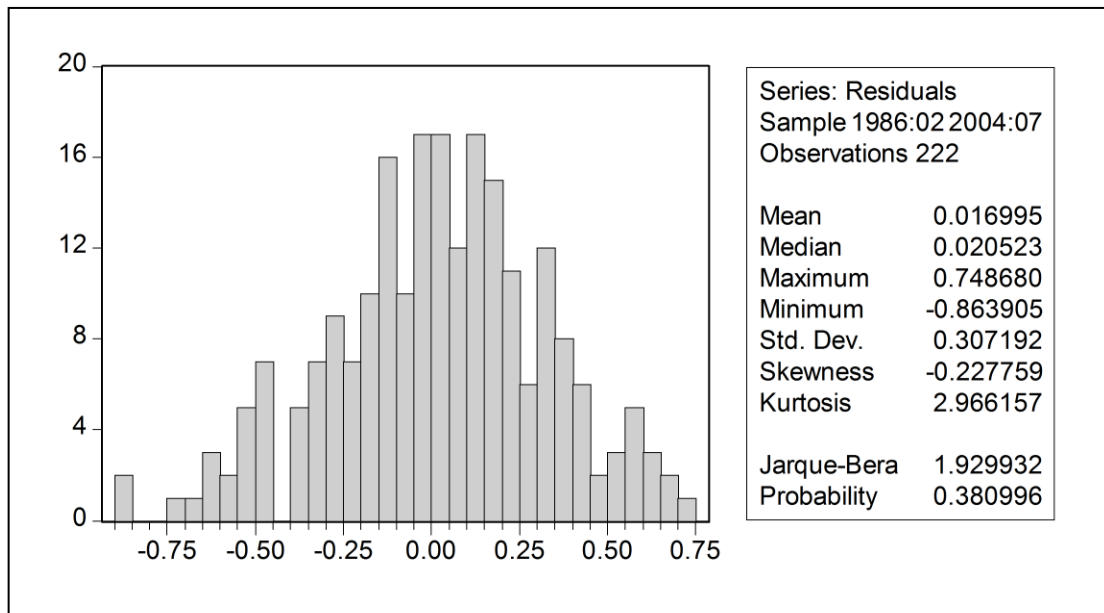
- O novo modelo resultante é:

Sample(adjusted): 1986:02 2004:07				
Included observations: 222 after adjusting endpoints				
Convergence achieved after 11 iterations				
Backcast: 1984:11 1986:01				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
MA(1)	0.526791	0.062858	8.380592	0.0000
MA(2)	0.544615	0.064722	8.414612	0.0000
MA(3)	-0.188332	0.059651	-3.157230	0.0018
SMA(12)	-0.911501	0.018472	-49.34427	0.0000
R-squared	0.605677	Mean dependent var		0.006306
Adjusted R-squared	0.600251	S.D. dependent var		0.489950
S.E. of regression	0.309774	Akaike info criterion		0.511907
Sum squared resid	20.91927	Schwarz criterion		0.573216
Log likelihood	-52.82164	Durbin-Watson stat		2.034011

- Observe que o termo ma(3) é estatisticamente significativo.
- A FAC dos resíduos indica que o modelo é adequado:

Q-statistic probabilities adjusted for 4 ARMA term(s)					
Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
1		-0.023	-0.023	0.1172	
2		0.040	0.040	0.4832	
3		-0.069	-0.067	1.5569	
4		0.048	0.044	2.0745	
5		-0.037	-0.030	2.3906	0.122
6		0.018	0.009	2.4624	0.292
7		-0.087	-0.079	4.2142	0.239
8		-0.054	-0.065	4.9023	0.297
9		-0.024	-0.016	5.0337	0.412
10		-0.044	-0.055	5.4880	0.483
11		-0.047	-0.049	6.0078	0.539
12		0.128	0.129	9.9072	0.272
13		0.061	0.065	10.799	0.290
14		-0.044	-0.061	11.255	0.338
15		-0.099	-0.103	13.622	0.255
16		-0.013	-0.027	13.665	0.323
17		0.032	0.027	13.910	0.380
18		-0.030	-0.051	14.129	0.440
19		0.001	0.012	14.129	0.516
20		-0.079	-0.053	15.654	0.477
21		-0.002	-0.013	15.655	0.548
22		0.115	0.123	18.932	0.396
23		0.042	0.041	19.366	0.434
24		0.013	-0.007	19.410	0.495
25		0.052	0.025	20.101	0.515

- A hipótese de normalidade dos resíduos não pode ser rejeitada:



- Previsão fora da amostra:

Mês	Saz c/ dummies	SARIMA	atual
Ago.04	18.64	18.66	18.3
Set.04	18.66	18.75	17.9
Out.04	18.45	18.64	
Nov.04	18.26	18.47	
Dez.04	17.85	18.09	
Jan.05	17.44	17.68	
Fev.05	17.57	17.79	