EAE1223: Econometria III

Aula 5 - Metodologia de Box-Jenkins

Luis A. F. Alvarez

1 de abril de 2024

A metodologia

- A metodologia de Box-Jenkins consiste numa série de etapas para estimar um modelo univariado de previsão.
  - Ideia é estimar um modelo simples, embora flexível, aos dados.
- Trata-se da metodologia básica de previsão em séries de tempo.
  - Diversas metodologias modernas incorporam o "espírito" de Box-Jenkins.
  - *Benchmark* para avaliar outros modelos de previsão.

Etapas da metodologia

- A metodologia consiste de quatro etapas:
  1. Identificação: nessa etapa, avaliamos os dados e identificamos quais modelos são candidatos plausíveis para reproduzir os dados.
  2. Estimação: nessa etapa, estimamos os modelos candidatos.
  3. Diagnóstico: nessa etapa, avaliamos quais dos modelos se saíram melhor, de acordo com alguns critérios.
  4. Previsão: por fim, realizamos a previsão de acordo com nosso modelo.
- Nesta aula, discutiremos cada uma dessas etapas.
- Começaremos revisando a classe de modelos estudadas na metodologia de Box-Jenkins.

Modelos considerados

MA(q)

- Dizemos que uma série de tempo $\{Y_t\}$ segue um MA(q) se:

$$Y_t = \alpha + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

  onde $\{\epsilon_t\}$ é um ruído branco.

- Série hoje depende diretamente da realização atual e das últimas q realizações de um ruído branco.

- **Todo** processo de média móvel é fracamente estacionário.
  - De fato, $\mathbb{E}[Y_t] = \alpha$, $\mathbb{V}[Y_t] = \left(1 + \sum_{j=1}^{q} \theta_j^2\right) \sigma_\epsilon^2$, $\text{cov}(Y_t, Y_{t-s}) = (\theta_s + \sum_{j=s+1}^{q} \theta_j \theta_{j-s})\sigma_\epsilon^2$ se $s \leq q$ e 0 do contrário. Correlação morre após q períodos.

- Um processo MA(q) é dito **invertível** se pode ser escrito como:

$$Y_t = \omega + \sum_{j=1}^{\infty} \psi_j Y_{t-j} + \epsilon_t$$

  (MA(q) pode ser representado como AR(∞)).

AR(p) estacionário

- Dizemos que uma série de tempo $\{Y_t\}$ segue um AR(p) estacionário se ela se escreve como:

$$Y_t = \alpha + \sum_{j=1}^{p} \beta_j Y_{t-j} + \epsilon_t$$

  onde $\{\epsilon_t\}$ é um ruído branco e os coeficientes β_j são tais que o processo resultante é fracamente estacionário.

- Série hoje depende diretamente das realizações passadas nos últimos p períodos, mais um ruído branco.
  - Mas persistência não é tão grande, de modo que a série é estacionária (não há raiz unitária).

- Recorde-se que um AR(p) é estacionário se, e somente se, ele se escreve como um MA(∞):

$$Y_t = \kappa + \epsilon_t + \sum_{j=1}^{\infty} \tau_j \epsilon_{t-j}$$

ARMA(p,q) estacionário

- Dizemos que uma série de tempo $\{Y_t\}$ segue um ARMA(p,q) estacionário se:

$$Y_t = \alpha + \sum_{i=1}^{p} \beta_i Y_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

  onde $\{\epsilon_t\}$ é um ruído branco, e os β_i são tais que o processo resultante é estacionário.

- Combinação dos dois modelos anteriores.

- Metodologia de Box-Jenkins visará a estimar modelos na classe ARMA(p,q).

E se os dados forem não estacionários?

- Até agora, discutimos a modelagem supondo as séries estacionárias.
  - Como fazer a previsão em casos não estacionários?

- Se as séries apresentarem uma tendência estocástica, trabalhamos com os dados em primeira diferença $\{\Delta Y_t\}$.

- Conduzidas as etapas da metodologia Box-Jenkins com os dados diferenciados, e encontradas projeções para ΔY_t fora da amostra, recompomos as projeções em nível usando o fato de que $Y_{t+1} = Y_t + \Delta Y_{t+1}$.
  - Isto é, se temos T observações, projetamos $\widehat{Y_{T+1}} = Y_T + \widehat{\Delta Y_{T+1}}$, $Y_{T+2} = Y_T + \widehat{\Delta Y_{T+1}} + \widehat{\Delta Y_{T+2}}$, e assim por diante.

Modelos ARIMA(p,d,q)

- Em outras palavras, no caso de raiz unitária, a metodologia irá estimar um modelo ARIMA(p,1,q), da forma:

$$\Delta Y_t = \alpha + \sum_{i=1}^{p} \beta_i \Delta Y_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

- De modo geral, se a série é I(d), a modelagem considerará modelos ARIMA(p,d,q):

$$\Phi(L)(1-L)^d Y_t = \alpha + \Psi(L)\epsilon_t \,,$$

para polinômios $\Phi(L)$ e $\Psi(L)$ de grau p e q, respectivamente.

Previsão com tendência determinística

- Se séries apresentam tendência determinística, conduzimos a metodologia com dados *detrended*. Calculadas as projeções *detrended*, recompomos projeções em nível usando a tendência estimada.

- Por exemplo, se estimamos uma tendência linear:

$$Y_t = \tilde{a} + \tilde{b}t + \tilde{\xi}_t$$

  e ajustamos um modelo ARMA para $\tilde{\xi}_t$, a projeção para fora da amostra é:

$$\widehat{Y_{T+h}} = \tilde{a} + \tilde{b}(T+h) + \widehat{\tilde{\xi}_{T+h}},$$

  onde $\widehat{\tilde{\xi}_{T+h}}$ é a projeção do ARMA para $T+h$ (veremos como calculá-la na etapa de previsão).

- Estimação de ARMA para dado *detrended* é equivalente a estimar um modelo ARMA(p,q) com tendência determinística.

$$\Phi(L)Y_t = \alpha + \gamma t + \Psi(L)\epsilon_t\,.$$

"Filosofia" da metodologia de Box-Jenkins

- A restrição a modelos ARMA(p,q) pode ser entendida a partir do teorema de decomposição de Wold.

- Segundo esse teorema, **qualquer processo fracamente estacionário** pode ser representado pela soma de um MA(∞), acrescido de uma função determinística dos Y no passados:

$$Y_t = \epsilon_t + \sum_{l=0}^{\infty} \psi_l \epsilon_{t-l} + \kappa_t \,,$$

onde κ_t é função "aproximadamente" linear de Y_{t-1}, $Y_{t-2} \ldots$

- Ideia de Box-Jenkins é aproximar essa representação por um ARMA(p,q), com p e q pequenos.
  - Ideia é que aproximação parcimoniosa, por ser menos ruidosa, tende a funcionar melhor que modelos muito complexos.

Identificação

Identificação de um ARMA(p,q)

- A etapa de identificação da metodologia Box-Jenkins consiste em encontrar quais modelos da classe ARMA(p,q) melhor caracterizam a série de interesse.

- A identificação consiste em analisar a função de autocorrelação (FAC) e a função de autocorrelação parcial (FACP) da série estacionária.

Função de autocorrelação serial (FAC)

- A função de autocorrelação (FAC) de uma série $\{Y_t\}_t$ estacionária é o mapa que associa, a cada número $k > 0$, a autocorrelação de ordem k, i.e. $\gamma_k = \text{cor}(Y_t, Y_{t-k})$.

- Note que essa função está bem definida para processos estacionários, visto que $\text{cor}(Y_t, Y_{t-k})$ não depende de t.

$$\text{cor}(Y_t, Y_{t-k}) = \frac{\text{cov}(Y_t, Y_{t-k})}{\text{sd}(Y_t)\,\text{sd}(Y_{t-k})} = \frac{\text{cov}(Y_t, Y_{t-k})}{\mathbb{V}(Y_t)}$$

Função de autocorrelação serial parcial (FACP)

- A função de autocorrelação parcial (FACP) de uma série $\{Y_t\}_t$ estacionária é o mapa que associa, a cada número $k > 0$, a correlação θ_k entre Y_t e Y_{t-k}, controlando por $Y_{t-1}, Y_{t-2}, \ldots Y_{t-k+1}$.

- A autocorrelação parcial de ordem k (θ_k) é dada pelo coeficente τ_k associado a Y_{t-k} no modelo preditivo linear:

$$\begin{aligned} Y_t = \tau_0 + \tau_1 Y_{t-1} + \tau_2 Y_{t-2} \ldots + \tau_k Y_{t-k} + \nu_t \\ \mathbb{E}[\nu_t] = 0, \mathbb{E}[\nu_t Y_{t-j}] = 0, \quad j = 1, \ldots k \end{aligned} \tag{1}$$

- Note que função está bem definida para processos estacionários, visto que coeficientes do melhor preditor linear em (1) não dependem de t.

FAC e FACP estimadas

- Na prática, não observamos a FAC nem a FACP de um processo, mas podemos estimá-las usando as realizações da série de interesse.
  - Estimamos a FAC calculando as autocorrelações nos dados.

$$\hat{\gamma}_k = \frac{\sum_{t=1}^{T-k}(Y_t - \bar{y})(Y_{t+k} - \bar{y})}{\sum_{t=1}^{T}(Y_t - \bar{y})^2},$$

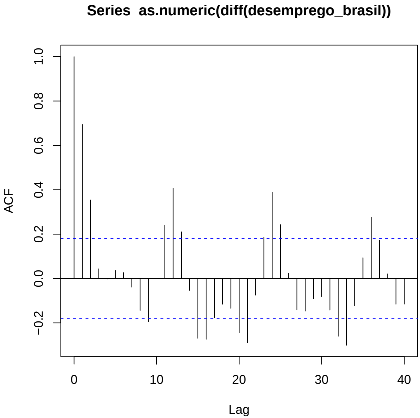
  onde $\bar{y} = T^{-1}\sum_{t=1}^{T} Y_t$.

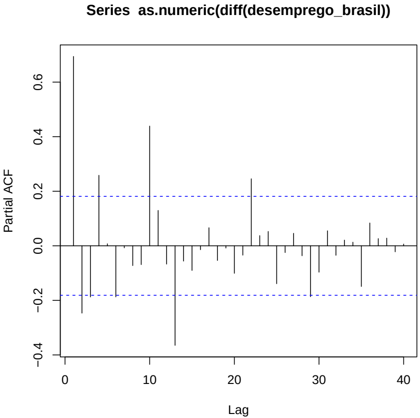
  - Estimamos a FACP ajustando o modelo (1) aos dados.

Inferência sobre FAC e FACP populacionais

- Como observamos apenas algumas realizações do processo, gostaríamos de testar hipóteses sobre a FAC e FACP populacionais.
- Sob algumas condições mais estringentes, o intervalo $[-2/\sqrt{T}, 2/\sqrt{T}]$ é uma região de aceitação aproximadamente válida para o teste da nula $\gamma_k = 0$ ($\theta_k = 0$ na FACP) contra a alternativa bilateral ao nível de significância de 5%, com base na estatística $\hat{\gamma}_k$ ($\hat{\theta}_k$).
    - São esses intervalos que são apresentados, no R, quando computamos a FAC e FACP.
    - Para autocorrelações (parciais) estimadas que excedem esses limites, rejeitamos a hipótese nula de não autocorrelação (parcial) a essa ordem.

FAC e FACP amostrais de ΔDesemprego





Inferência conjunta sobre a FAC

- Para testar a nula conjunta de que $\gamma_1 = \gamma_2 = \ldots = \gamma_s = 0$, onde s é pequeno relativamente a T, podemos usar a estatística de Ljung-Box:

$$\hat{Q} = T(T+2)\sum_{r=1}^{s} \frac{\hat{\gamma}_r^2}{T-r}\,.$$

- Com T grande, sob a nula, $\hat{Q}$ segue uma qui-quadrado com s graus de liberdade. Valores altos da estatística são evidência contra a nula, i.e. evidência de que ao menos uma autocorrelação entre as testadas é diferente de zero.

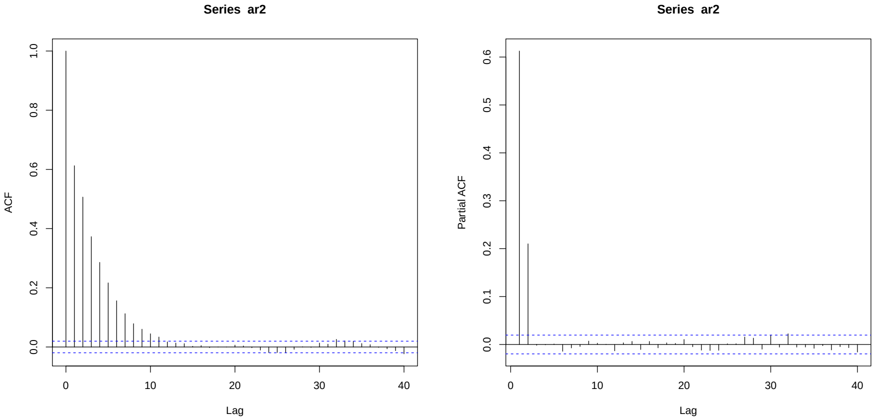
FAC e FACP de um AR(p) estacionário

- Como vimos em aula anterior, a FAC de um AR(1) estacionário é dada por:
$$\gamma_k = \text{cor}(Y_t, Y_{t-k}) = \beta_1^k, \quad |\beta_1| < 1$$
isto é, a FAC apresenta decaimento geométrico em direção a zero.

- De modo geral, a FAC de um AR(p) estacionário apresenta decaimento em direção a zero, visto que um AR(p) estacionário pode ser escrito como um MA(∞), i.e.
$$Y_t = \mu + \epsilon_t + \sum_{j=1}^{\infty} \omega_j \epsilon_{t-j}$$

- E a FACP de um AR(p) estacionário? Pela definição da FACP de ordem k como o coeficiente de Y_{t-k} na regressão populacional de Y_t em Y_{t-1}, Y_{t-2},... Y_{t-k}, esperamos que a FACP seja truncada em p, visto que o processo só depende diretamente das p primeiras defasagens.

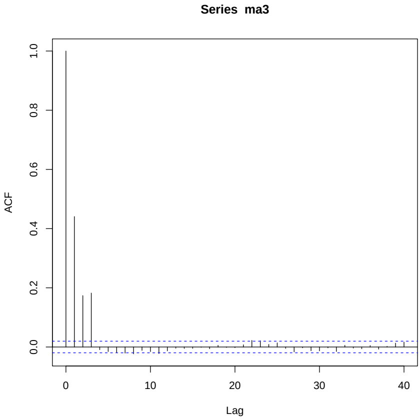
Exemplo: FAC e FACP estimadas de um AR(2) (T=10.000)

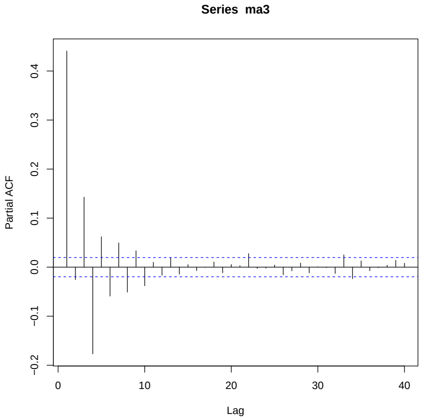


FAC e FACP de um MA(q) invertível

- Como vimos em aula anterior, a FAC de um MA(q) é truncada em q, visto que a correlação morre após q períodos.
- E a FACP? Se o processo MA for invertível, vimos que ele pode ser escrito como um $AR(\infty)$. Dessa representação, fica claro que a FACP de um MA(q) apresenta decaimento em direção a zero.

Exemplo: FAC e FACP estimadas de um MA(3) (T=10.000)

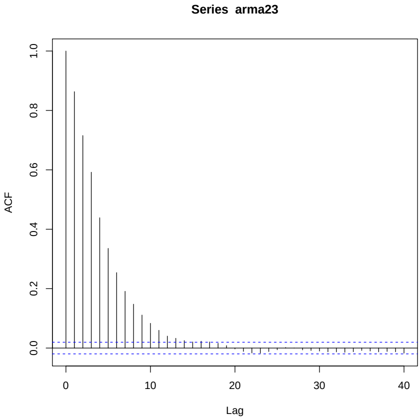


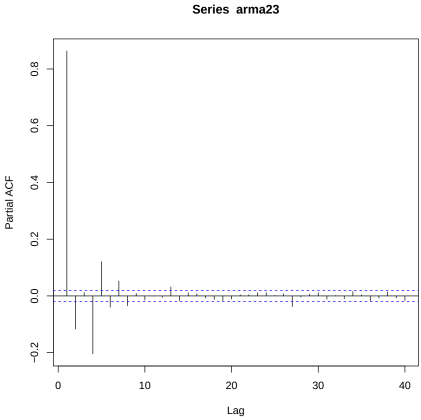


FAC e FACP de um ARMA(p,q) estacionário e invertível

- Generalizando a discussão anterior, um ARMA(p,q) estacionário cuja parte MA é invertível pode ser representado tanto como um $AR(\infty)$ como um $MA(\infty)$. Nesse caso, tanto a FAC como a FACP apresentam decaimento.
- Nesses casos, costuma-se considerar a ordem máxima q_{max} em que a FAC torna-se pouco significativa e a ordem p_{max} em que a FACP torna-se pouco signifcativa e considerar todos os ARMA(p,q), $0 \leq p \leq p_{\text{max}}$ e $0 \leq q \leq q_{\text{max}}$ como candidatos.

Exemplo: FAC e FACP estimadas de um ARMA(2,3) (T=10.000)





Resumo

Modelo	FAC	FACP
AR(p) estacionário	decai	truncada em p
MA(q) invertível	truncada em q	decai
ARMA(p,q) estacionário e invertível	decai (esp. após q)	decai (esp. após p)

Estimação

Estimação condicional vs. incondicional

- Para a estimação de modelos ARMA(p,q), há duas abordagens de estimação.
  - Na abordagem condicional, não utilizamos a informação acerca da distribuição das primeiras observações na estimação.
  - Na abordagem incondicional, fazemos hipóteses adicionais sobre a distribuição do ruído branco, que nos permitem incorporar a distribuição das primeiras observações na análise.
- Computacionalmente, a abordagem condicional é mais simples, embora menos eficiente que a segunda.
- Embora a abordagem incondicional aparente requerer mais hipóteses, visto que especificamos a distribuição do ruído branco, a estimação é robusta a violações dessa hipótese quando o número de observações é grande ("pseudo" máxima verossimilhança).

Estimação condicional do AR(p)

- Podemos estimar os parâmetros de um AR(p) através de mínimos quadrados ordinários:

$$(\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_p) \in \text{argmin}_{a,b_1,\ldots b_p} \frac{1}{T-p} \sum_{t=p+1}^{T} (Y_t - a - b_1 Y_{t-1} \ldots - b_p Y_{t-p})^2$$

- Note que não tentamos prever as p primeiras observações, para as quais não temos todas as defasagens.
- Estimadores $(\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ coincidem com o estimador de máxima verossimilhança que usa a distribuição de $(Y_{p+1}, Y_{p+2} \ldots Y_T)$ condicional a $(Y_1, Y_2, \ldots, Y_p)$, tomando o ruído branco como $\epsilon_{it} \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

Estimação condicional do MA(1)

- Para a estimação de um modelo MA(1) com base numa série $\{Y_t\}_{t=1}^T$, precisamos de um chute inicial para o ruído branco no período $t = 0$.

- Chamemos esse chute de $\tilde{\epsilon}_0$ (padrão é tomar $\tilde{\epsilon}_0 = 0$).

- Dado o chute inicial, e dado um valor candicato c para o parâmetro θ_1, e um valor candidato a para o intercepto α, podemos imputar o ruído branco em $t = 1$:

$$\tilde{\epsilon}_1(a, c) = Y_1 - a - c\tilde{\epsilon}_0 \,.$$

- Procedendo recursivamente, obtemos, para todo $t \geq 2$.

$$\tilde{\epsilon}_t(a, c) = Y_t - a - c\tilde{\epsilon}_{t-1}(a, c)$$

- MA(1) pode ser estimado como:

$$(\hat{\alpha}, \hat{\theta}_1) \in \text{argmin}_{a,c} \frac{1}{T-1} \sum_{t=2}^{T} (Y_t - a - \tilde{\epsilon}_t(a, c) - c\tilde{\epsilon}_{t-1}(a, c))^2$$

Estimação condicional do MA(1)

- Note que $(a, c) \mapsto a + \tilde{\epsilon}_t(a, c) + c\tilde{\epsilon}_{t-1}(a, c)$ varia não linearmente com (a, c).
    - Estimação se dá através de algoritmos para mínimos quadrados não lineares (não há expressão fechada para o mínimo).
- Observe também que a primeira observação não contribui à estimação.
    - De fato $(Y_1 - a - \tilde{\epsilon}_1(a, c) - c\tilde{\epsilon}_0) = 0$ para qualquer escolha de (a, c).
- Se o MA(1) é invertível, então, com T grande, efeito do chute inicial sobre a função objetivo desaparece.
    - Contribuição do chute inicial ao erro de previsão em t é da ordem de c^t, que desaparece quando $|c| < 1$.

Estimação condicional do ARMA(p,q)

- Estendendo a discussão anterior, a estimação condicional de um ARMA(p,q) é dada pela minimização de:

$$\sum_{t=p+q}^{T} (Y_t - a - b_1 Y_{t-1} \ldots - b_p Y_{t-p} - \tilde{\epsilon}_t(a, \boldsymbol{b}, \boldsymbol{c}) - c_1 \tilde{\epsilon}_{t-1}(a, \boldsymbol{b}, \boldsymbol{c}) \ldots - c_q \tilde{\epsilon}_{t-q}(a, \boldsymbol{b}, \boldsymbol{c}))^2$$

onde $\tilde{\epsilon}_t(a, \boldsymbol{b}, \boldsymbol{c})$ são definidos recursivamente, para cada valor candidato dos parâmetros $(a, \boldsymbol{b}, \boldsymbol{c})$ e chutes iniciais dos errors.

- Perdemos p observações pois não observamos os valores de Y anteriores a $t = 1$, e mais q observações na construção recursiva dos erros.

Estimação incondicional do ARMA(p,q)

- Na estimação condicional, perdemos a informação das $p+q$ primeiras informações.
- Se fizermos uma hipótese distributiva sobre os ruídos brancos, somos capazes de caracterizar a distribuição das $p+q$ primeiras observações.
  - Por exemplo, se $\epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$ e Y_t seguir um AR(1) estacionário, então $Y_1 \sim N(\alpha/(1-\alpha), \sigma^2/(1-\rho^2))$.
- A estimação incondicional de um ARMA(p,q) se dá pela máxima verossimilhança que usa a distribuição conjunta de $\{Y_1, \ldots, Y_T\}$, sob a hipótese auxiliar de que $\epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$.
  - Método padrão na função `arima` do `R`.
- Se ruído branco de fato é Gaussiano, estimador é eficiente: dentre todos os estimadores de um ARMA(p,q) Gaussiano, estimador é o de menor variância.
- Mesmo que o ruído branco não seja Gaussiano, estimador ainda é consistente para os parâmetros de um ARMA(p,q).

Diagnóstico

Diagnóstico

- Estimados os modelos candidatos, procedemos à etapa de diagnóstico. A ideia é avaliar os modelos conforme algumas métricas:
    1. Critérios de informação.
    2. Parcimônia.
    3. Não autocorrelação dos erros.
    4. Estabilidade e invertibilidade das partes AR e MA.
    5. Convergência numérica dos estimadores.
    6. Normalidade dos erros.

Critérios de informação

- A princípio, gostaríamos de uma métrica que indicasse quanto da variabilidade do processo é explicada pelo modelo.
  - Quantidade não observada; precisa ser estimada.
- Intuitivamente, um estimador dessa quantidade poderia ser dado pela soma dos quadrados dos resíduos (SSR) do modelo estimado.
- O problema dessa métrica é que modelos mais complexos **necessariamente** apresentam SSR menor.
  - Maior flexibilidade leva a melhor ajuste dentro da amostra.
  - Isso não significa que o modelo necessariamente explique bem a variação verdadeira do processo (em particular, **fora da amostra**).
- Se fôssemos escolher o modelo pelo menor SSR, sempre escolheríamos o modelo mais complexo, incorrendo num problema conhecido como sobreajuste (*overfitting*).
  - Modelo funcionará, em geral, muito mal fora da amostra, pois estimadores dos parâmetros apresentam alta variância.

Critérios de informação (cont.)

- A ideia de um critério de informação é penalizar a SSR pelo número de parâmetros estimados.
  - A penalização pode ser vista como uma forma de corrigir o viés do SSR em estimar a capacidade preditiva de um modelo.

- Os critérios de informação mais utilizados são o AIC e o BIC. Para um ARMA(p,q) com intercepto, eles são dados por:

$$AIC = T\ln(SSR) + 2(p+q+1)$$
$$BIC = T\ln(SSR) + ln(T)(p+q+1)$$

- Quanto menor o critério de informação, melhor.

- Para $T > 7$, BIC escolhe modelos com menos parâmetros.

- Se a estimação do modelo é condicional, importante ajustar amostra para que mesmo número de observações sejam usadas no cálculo dos critérios em todos os modelos comparados (isso já vale, por construção, na estimação incondicional).

Parcimônia

- Os critérios de informação induzem parcimônia no modelo escolhido, ajudando a evitar o problema de sobreajuste.
- Ainda sob essa lógica, é costumeiro verificar quais coeficientes são estatisticamente significantes na especificação: se houver muitos coeficientes insignificantes, talvez valha trabalhar com um modelo mais simples.

Não autocorrelação dos erros

- Os modelos ARMA discutidos supõem que os erros são ruído branco.
  - Dessa forma, esperaríamos que os *resíduos* de nosso modelo fossem aproximadamente não autocorrelacionados.
  - Se houver correlação nos resíduos, ainda há informação nos dados que a parte sistemática do modelo não está capturando.
- Teste da hipótese nula de que as autocorrelações dos erros de ordens 1 até s são zero, contra a alternativa de que ao menos uma é diferente de zero, podem ser conduzidos usando a estatística de Ljung-Box:

$$Q = T(T+2)\sum_{j=1}^{s} \hat{\gamma}_j^2/(T-j)\,,$$

onde $\hat{\gamma}_j$ é autocorrelação de ordem j estimada com base nos resíduos.
- Sob a nula, distribuição de Q é aproximadamente uma χ^2 com $s-(p+q)$ graus de liberdade.
- Quanto maior Q, maior a evidência contra a nula. Assim, região crítica do teste é da forma $Q > c$ onde c é o quantil apropriado da distribuição χ^2.

Estabilidade e invertibilidade

- Recorde-se que a análise de identificação dos modelos ARMA(p,q) pressupõe que os processos sejam estacionários e invertíveis. Assim, é costumeiro verificar se os coeficientes estimados de fato nos levam a processos estacionários e invertíveis.
  - Se isso não ocorrer, devemos suspeitar de nossas estimativas.

- Podemos checar a estacionariedade e invertibilidade do ARMA(p,q) resolvendo, respectivamente, as equações de grau p e q dos polinômios estimados, e avaliando se as raízes se encontram todas fora do cícrulo

Convergência numérica

- Os estimadores mais usados de modelos ARMA são não lineares e não possuem solução fechada.
  - Por esse motivo, pacotes estatísticos usam algoritmos de otimização para estimar o modelo.
- É importante checar se os algoritmos de otimização de fato convergiram para um mínimo.
  - Se esse não é o caso, devemos descartar as estimativas.

Normalidade dos erros

- Recorde-se que, se os ruídos brancos forem Gaussianos, o estimador de máxima verossimilhança do ARMA(p,q) é eficiente.
    - Ainda assim, mesmo que os ruídos brancos não sejam Gaussianos, o estimador é consistente.
- Nesse sentido, é costumeiro testar a hipótese de normalidade dos erros de um modelo ARMA.
- Isso é feito verificando a assimetria e curtose dos *resíduos* do modelo, e quanto elas distam do esperado em uma distribuição normal.
- Sob a nula de normalidade, a estatística de teste de Jarque-Bera possui distribuição χ^2 com **dois** graus de liberdade.
    - Valores grandes da estatística são evidência contra a nula.