# Técnicas de Previsão de Box-Jenkins – ARIMA[1]

## Introdução

**Metodologia Box-Jenkins Ou Método de Previsão ARIMA:** Os modelos de previsão Box-Jenkins são baseados em conceitos e princípios estatísticos e são capazes de modelarem um amplo espectro do comportamento de séries temporais. O objetivo fundamental deste método de auto-projeção para previsão de series temporais é encontrar uma fórmula apropriada para que os erros/resíduos sejam tão pequenos quanto possível e não apresentem padrões. O processo de construção do modelo envolve quatro passos. Repetindo sempre que for necessário, até acabar com uma fórmula específica que replique os padrões da série o mais próximo possível e produza também previsões acuradas. (O termo ARIMA e Box-Jenkin são usados indistintamente)

A metodologia tem uma grande classe de modelos à escolha e uma abordagem sistemática para identificar a correta forma de modelar. Existem testes estatísticos para verificar a validade do modelo e medidas estatísticas de incerteza das previsões. Em contraste, os modelos de previsão tradicionais oferecem um número limitado de modelos em relação ao comportamento complexo de muitas séries temporais, com pouca coisa na forma de orientações e testes estatísticos, para verificar a validade do modelo selecionado. (Isso você aprendeu no texto Métodos Básicos de Previsão de Séries Temporais no Excel).

**Modelo Básico:** Com uma série estacionária no lugar, um modelo básico pode agora ser identificado. Existem três modelos básicos, **AR** (*autoregressivos*), **MA** (*moving average*) e um combinado **ARMA** em adição ao RD (diferenciação regular) especificado anteriormente, os quais se combinam para fornecer as ferramentas disponíveis. Quando a diferenciação regular (RD) for aplicada junto com ao **AR** e **MA**, eles são referidos como **ARIMA**, com o **I** indicando "*integrado*" e referindo-se ao procedimento de diferenciação.

Os modelos **ARIMA** são largamente usados nas situações:

1. previsão de preços de estoques,
2. vendas da companhia,
3. números de manchas solares,
4. lançamento de casas e muitos outros campos.

Os modelos **ARIMA** são também *univariados*, isto é, eles são baseados numa única variável de série temporal. (Existem modelos multivariados que estão além do escopo deste texto e não serão discutidos)

Os processos **ARIMA** parecem, à primeira vista, envolverem apenas uma variável e a sua própria história. A nossa intuição nos diz que qualquer variável econômica é dependente de muitas outras variáveis. Como podemos então considerar o sucesso relativo da metodologia Box Jenkins? O uso de previsões *univariadas* deve ser importante por várias razões:

- Em alguns casos temos uma escolha de modelagem, digamos, a saída de um grande número de processos ou de saídas agregadas, deixando o modelo *univariado* como a única abordagem possível por causa da magnitude completa do problema.
- Deve ser difícil encontrar variáveis que estejam relacionadas à variável que está sendo projetada, deixando o modelo *univariado* como o único meio de previsão.
- Onde os métodos *multivariados* estiverem disponíveis o método *univariado* fornece um parâmetro contra o qual os métodos mais sofisticados podem ser avaliados.

[1] Auto-Regressive Integrated Moving Average

- A presença de grandes resíduos num modelo *univariado* deve corresponder aos eventos anormais – greves, etc.
- O estudo dos modelos *univariados* pode dar informação útil sobre ciclos de tendências de longo prazo, efeitos sazonais, etc., nos dados.
- Alguma forma de análise *univariada* deve ser um pré-requisito necessário à análise multivariada se regressões espúrias e problemas relacionados devam ser evitados.

Embora os modelos *univariados* funcionem bem no curto prazo, provavelmente os métodos multivariados fazem uma apresentação de melhor qualidade ao levar mais termos, se as variáveis relacionadas à variável que está sendo projetada flutuarem de várias maneiras, e de formas diferentes aos seus comportamentos no passado.

Box e Jenkins desenvolveram procedimentos para esta modelagem multivariada. Entretanto, na prática, mesmo sua abordagem *univariada*, algumas vezes, não é tão bem entendida quanto o método de regressão clássico. O objetivo deste texto é descrever o básico dos modelos univariados de Box- Jenkins em termos simples e não especializados.

## O Modelo Matemático

Os modelos **ARMA** podem ser descritos por uma série de equações. As equações são de certa forma mais simples se as series temporais primeiro forem reduzidas à média zero, subtraindo delas a média amostral. Portanto, trabalharemos com a série *ajustada à média*

$$\mathbf{y}_{\text{ajustada}}(t) = y(t) - \bar{Y} \qquad (1)$$

Onde *y(t)* é a série temporal original, $\bar{Y}$ é sua media amostral, e $\mathbf{y}_{\text{ajustada}}(t)$ é a série ajustada à média[2]. Um subconjunto dos modelos **ARMA** são aqueles chamados de *autoregressivos*, ou modelos **AR**. Um modelo **AR** expressa uma série temporal como uma função linear dos seus valores passados. A *ordem* do modelo **AR** diz quantos valores atrasados (lags) no passado são incluídos. O modelo **AR** mais simples é o auto-regressivo de *primeira ordem*, ou modelo *AR(1)*,

$$\mathbf{y}(t) = \mathbf{a}(1)*\mathbf{y}(t-1) + \mathbf{e}(t) \qquad (2)$$

onde **y**(t) é a série ajustada à media no período t, **y**(t-1) é o valor do período anterior na série , **a**(t) é o coeficiente auto-regressivo de *lag-1*, e **e**(t) é o *ruído*. O ruído também é conhecido por vários outros nomes: *erro*, *choque aleatório* e *resíduo*. Os resíduos **e**(t) são assumidos serem aleatórios no tempo (não auto-correlacionados), e normalmente distribuídos. Podemos ver que o modelo *AR(1)* tem a forma de um modelo de regressão em que **y**(t) é regredido ao seu valor anterior. Desta forma, **a**(t) é análogo ao coeficiente de regressão, e **e**(t) ao resíduo de regressão. O nome *auto-regressivo* se refere à regressão em si mesmo (auto).

Os modelos regressivos de ordem superior incluem mais termos de defasagens em **y**(t) como preditores. Por exemplo, o modelo auto-regressivo de segunda ordem, *AR(2)*, é dado por

$$\mathbf{y}(t) = \mathbf{a}(1)*\mathbf{y}(t-1) + \mathbf{a}(2)*\mathbf{y}(t-2) \qquad (3)$$

onde: **a**(1) , **a**(2), são os coeficientes auto-regressivos sobre as defasagens 1 e 2. O modelo auto-regressivo de ordem ***p-ésima***, *AR(p)* inclui os termos de defasagens dos períodos $t-1$ até $t-p$ .

O modelo *média móvel* (*moving average*) (**MA**) é uma forma do modelo **ARMA** em que a série temporal é tomada como uma média móvel (pesos desiguais) de uma série de choques aleatórios **e**(t). A *média móvel de primeira ordem*, ou modelo *MA(1)*, é dada por

$$\mathbf{y}(t) = \mathbf{e}(t) + \mathbf{c}(1)*\mathbf{e}(t-1) \qquad (4)$$

[2] Representaremos esta série ajustada à média simplesmente por **y**(t)

onde **e**(t), **e**(t-1), são os *resíduos* no período *t* e *t-1*, e **c**(1) é o coeficiente de média móvel de primeira ordem. Como com os modelos **AR**, modelos **MA** de ordem superiores incluem termos de defasagens mais altos. Por exemplo, o modelo de media móvel de segunda ordem, *MA(2)*, é

$$\mathbf{y(t) = e(t) + c(1)*e(t-1) + c(2)*e(t-2)} \quad (5)$$

A letra **q** é usada para a **ordem** do modelo de *média móvel*. O modelo de média móvel de segunda ordem é MA(**q**), com **q** = 2.

Temos visto que o modelo auto-regressivo inclui termos de defasagens na sua própria série, e que o modelo de média móvel inclui termos de defasagens nos ruídos ou resíduos.

Por incluir ambos os tipos de termos de defasagens, chegamos ao que é chamado de *média móvel auto-regressiva*, ou modelos **ARMA**.

A **ordem** do modelo **ARMA** está incluída nos parênteses como: **ARMA(p,q)**, onde ***p*** é a ordem auto-regressiva e **q** a ordem de média móvel. O mais simples, e mais frequentemente usado modelo **ARMA** é o modelo *ARMA(1,1):*

$$\mathbf{y(t) = d + a(1)*y(t-1) + e(t) – c(1)*e(t-1)} \quad (6)$$

O *processo de média móvel autoregressivo* geral com **AR** de ordem **p e MA** de ordem **q** pode ser escrito como

$$\mathbf{y(t) = d+ a(1)*y(t-1) + a(2)*y(t-2) + … + a(p)*y(t-p) – e(t) – c(1)*e(t-1) – c(2)*e(t-2) -…- c(p)*e(t-p)} \quad (7)$$

O parâmetro **d** será explicado mais tarde.

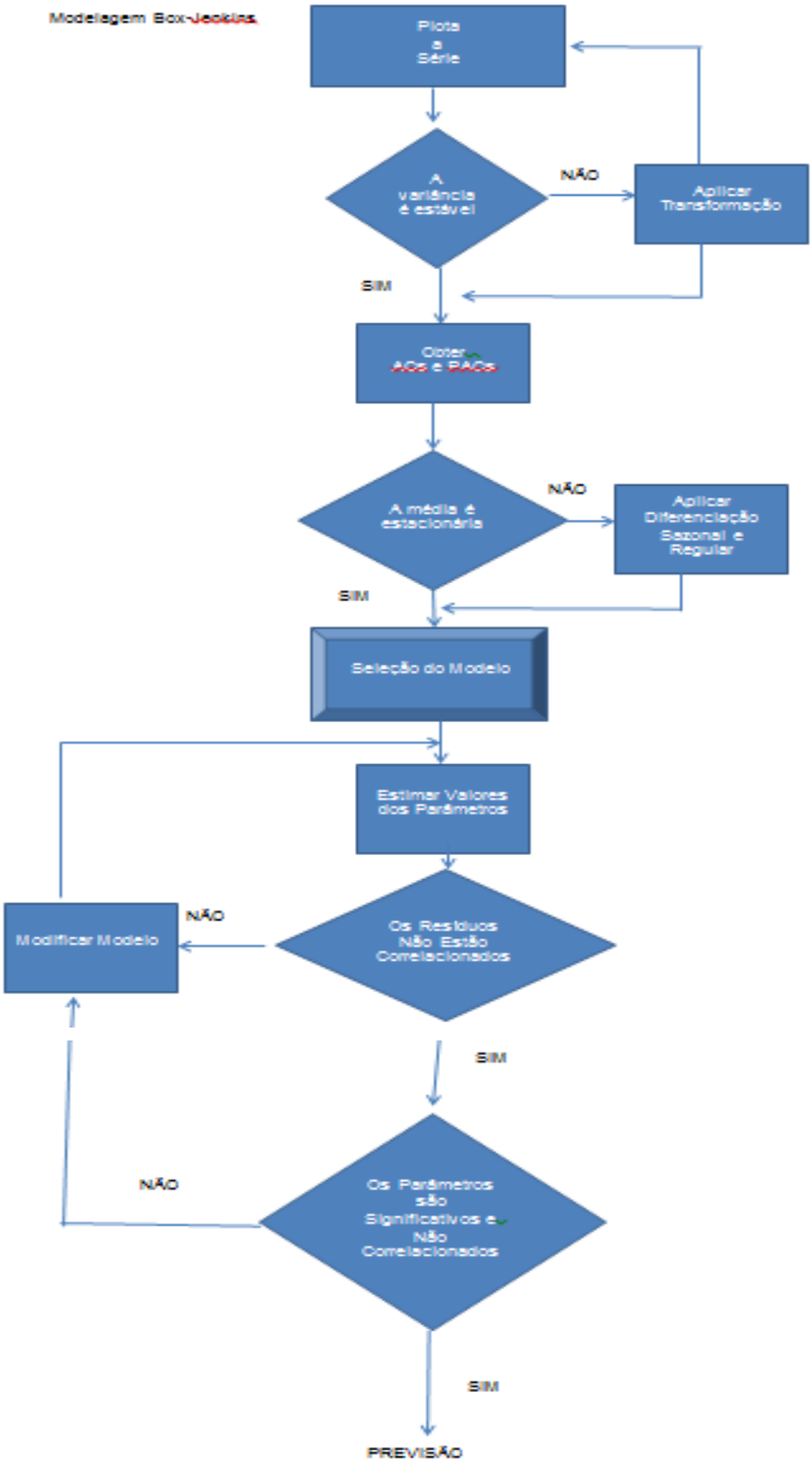
## Modelagem ARIMA

O propósito da modelagem **ARIMA** é estabelecer uma relação entre o valor presente de uma série temporal e seus valores passados de modo que as previsões possam ser feitas somente com base nos valores passados.

***Séries Temporais Estacionárias:*** A primeira exigência para a modelagem **ARIMA** é que a série temporal de dados a ser modelada tenha **estacionariedade** ou possa ser transformada nela. Podemos definir que uma série temporal é estacionária se tiver uma média constante e não tiver tendência no decorrer do tempo. Um gráfico dos dados é geralmente o bastante para ver se os mesmos são estacionários. Na prática, poucas séries temporais podem ser encontradas nesta condição, mas sempre que os dados puderem ser transformados numa série estacionária, um modelo **ARIMA** pode ser desenvolvido. (Explicarei adiante este conceito com mais detalhes).

Enfatizamos novamente que, para projetar uma série temporal usando esta abordagem de previsão, precisamos saber se a série temporal é estacionária. Se não for, para produzir previsões aceitáveis e acuradas, precisamos determinar a classe e a ordem do modelo, i.é, se ele é um modelo **AR**, **MA** ou **ARMA** e quantos coeficientes **AR** e **MA** (**p** e **q**) são apropriados. A análise das funções de **auto-correlação** (ACF) e **auto-correlação parcial** (PACF) fornece pista para todas estas questões. Ambas as exigências acima serão calculadas e implementadas em duas planilhas-exemplos no Excel posteriormente.

Os passos gerais para a modelagem **ARIMA** estão mostrados no diagrama abaixo:


Modelagem Box-Jenkins
Plota a Série
A variância é estável
NÃO
Aplicar Transformação
SIM
Obter ACs e PACs
A média é estacionária
NÃO
Aplicar Diferenciação Sazonal e Regular
SIM
Seleção do Modelo
Estimar Valores dos Parâmetros
Modificar Modelo
NÃO
Os Resíduos Não Estão Correlacionados
SIM
NÃO
Os Parâmetros são Significativos e Não Correlacionados
SIM
PREVISÃO

## O Processo de Modelagem

A modelagem Box-Jenkins ou ARIMA de uma série temporal estacionária envolve os quarto principais passos seguintes:

A) Identificação do modelo

B) Estimativa do modelo

C) Diagnóstico de Verificação

D) Previsão

Os quatro passos são semelhantes àqueles exigidos para a regressão linear, exceto o *Passo A* ser um pouco mais envolvido. Box-Jenkins usa um *procedimento estatístico* para identificar um modelo, que pode ser complicado. Os outros três passos são muito simples. Vamos primeiro discutir a mecânica do *Passo A*, identificação do modelo, a qual será feita em grande detalhe. Depois então usaremos um exemplo para ilustrar o processo de modelagem por completo.

### A) IDENTIFICAÇÃO DO MODELO

**ARIMA** significa Autoregressive – Integrated - Moving Average. A letra "**I**" (Integrado) indica que a modelagem da série temporal a transformará numa série estacionária. **ARIMA** representa três tipos diferentes de modelos: Ele pode ser um modelo **AR** (*autoregressivo*), ou um modelo **MA** (*moving average*), ou um modelo **ARMA** que inclua ambos os termos AR e MA. Note que tivemos de tirar o "**I**" do ARIMA por simplicidade.

Vamos brevemente definir estas três formas de modelos novamente.

***Modelo AR:***

Um modelo **AR** se parece com uma modelo de regressão linear, exceto que num modelo de regressão a variável dependente e suas variáveis independentes são diferentes, enquanto no modelo AR as variáveis independentes são simplesmente os valores defasados no tempo da variável dependente, por isso é *auto-regressivo*. Um modelo **AR** pode incluir diferentes números de termos auto-regressivos.

Se um modelo AR incluir somente um termo auto-regressivo, ele é um modelo *AR (1)*; podemos também ter *AR (2)*, *AR (3)*, etc. Um modelo **AR** pode ser linear ou não linear. O que se segue, são uns poucos exemplos:

***AR(1)***

$$y(t) = d + a(1)* y(t-1) + e(t) \quad (8)$$

***AR(3)***

$$y(t) = d + a(1)*y(t-1) + a(2)*y(t-2) + a(3)*y(t-3) + e(t) \quad (9)$$

Eu explicarei posteriormente mais sobre o **d**.

***Modelo MA:***

Um modelo MA é uma media móvel ponderada, de número fixo, de erros de previsões, produzidas no passado, por isso é chamado média móvel. Diferentemente da media móvel tradicional, os pesos numa **MA** não são iguais e não somam 1. Numa média móvel tradicional, o peso atribuído a cada um dos *n* valores a ser feita a média, iguala-se a 1/*n*; os *n* pesos são iguais e somam 1. Numa **MA**, o número de termos para o modelo e o peso de cada termo são estatisticamente determinados pelo padrão dos dados; os pesos não são iguais e não somam 1. Geralmente, numa **MA** o valor mais recente carrega um peso maior que os valores atrasados mais distantes. Para uma série temporal estacionária, pode-se usar sua média ou valor passado imediato como uma previsão para o próximo período futuro. Cada previsão produzirá um

*erro de previsão*. Se os erros assim produzidos no passado exibirem qualquer padrão, podemos desenvolver um modelo **MA**. Note que estes *erros de previsão* não são valores observados; eles são valores gerados. Todos os modelos MA, tal como *MA(1)*, *MA(2)*, *MA(3)*, são **não** lineares. O que segue são uns poucos exemplos:

***MA(1)***

$$y(t) = e(t) + c(1)*e(t-1) \quad (10)$$

***MA(2)***

$$y(t) = e(t) + c(1)*e(t-1) + c(2)*e(t-2) \quad (11)$$

***Modelo ARMA:***

Um modelo **ARMA** requer ambos os temos: **AR** e **MA**. Dada uma série temporal estacionária, devemos primeiro identificar uma forma apropriada de modelo. É um **AR,** ou um **MA** ou um **ARMA**? Quantos termos nós precisamos no modelo identificado? Para responder estas questões podemos usar dois métodos:

1) Podemos usar um modo subjetivo calculando a **função autocorrelação** (ACF) e a **função autocorrelação parcial** (PACF) da série.

2) Ou usar métodos objetivos de identificação do melhor modelo ARMA para os dados em mãos. (ARIMA Automatizado)

## Método 1- para identificação do melhor modelo ARMA

### *i) O que são: Função de Autocorrelação (ACF) e Função de Autocorrelação Parcial (PACF)?*

Entender a ACF e a PACF é muito importante para se usar o método (1) para identificar qual modelo usar.

Sem entrar na matemática, os valores ACF caem entre -1 e +1, calculados da série temporal nas diferentes defasagens para medir a significância das correlações entre a *observação presente* e as *observações passadas*, e determinar o quanto a se voltar no tempo (i.é, de quantas defasagens no tempo eles estão correlacionados).

Os valores PACF são os *coeficientes* de uma regressão linear da série temporal usando seus valores defasados como variáveis independentes. Quando a regressão incluir somente uma variável independente de um período defasado, o coeficiente da variável independente é chamado **função autocorrelação parcial de primeira ordem**; quando um segundo termo de dois períodos de defasagem for adicionado à regressão, o coeficiente do segundo termo é chamado **função de autocorrelação parcial de segunda ordem** , etc. Os valores de PACF também caem entre -1 e +1, se a série temporal for estacionária.

Deixe-me mostrar-lhe como calcular o **ACF** e o **PACF** com um exemplo:

Abra o arquivo Arima.xls da pasta. Selecione a planilha *(acf).* Esta planilha contém os valores de fechamento diário da *Dow Jones Industrial Composite Index* (DJI)[3] entre 20 de Julho de 2009 e 29 de Setembro de 2009. No total, isto inclui *51* valores diários na série.

**ACF**

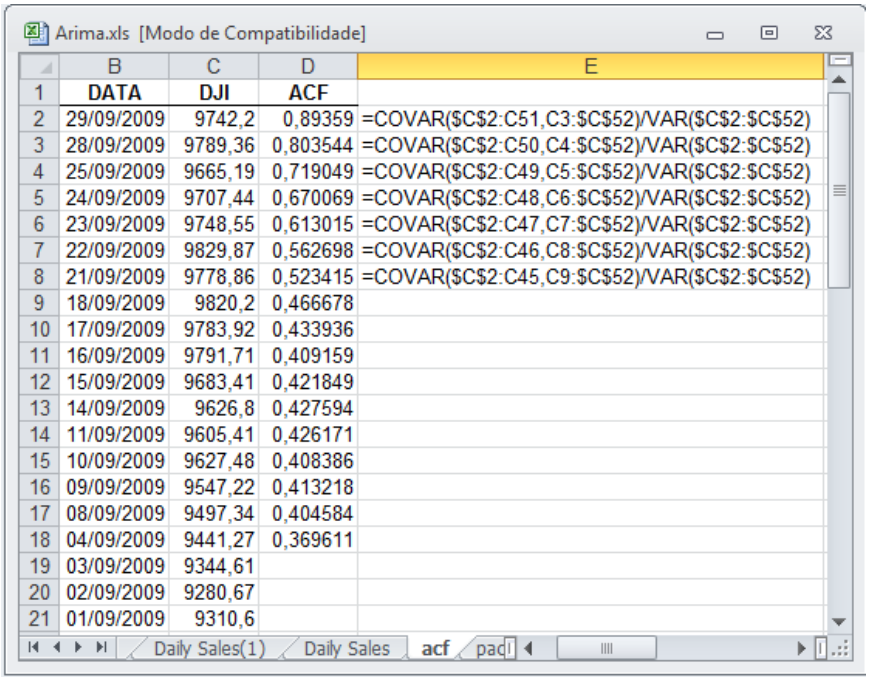
Abaixo está a formula geral para a **função Autocorrelação** (ACF):

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\widehat{cov}(R_{it};R_{i,t-k})}{\widehat{var}(R_{it})} \quad (12)$$

[3] Índice Dow Jones.

Não se intimide com esta fórmula. É facilmente implementada numa planilha usando uma função Excel. Podemos simplificar este procedimento usando alguma das muitas fórmulas embutidas do Excel. A fórmula acima essencialmente nos diz que o coeficiente de correlação para alguma k-defasagem é calculada como a covariância entre a série original e a série removida k defasagens, dividido pela variância da série original.

O Excel contém ambas as funções covariância e variância, e elas são: =VAR(intervalo), e, =COVAR(intervalo, intervalo). A planilha (*acf*) contém os detalhes de como estas duas funções podem ser usadas para calcular os coeficientes de autocorrelação:




| | B | C | D | E |
|---|---|---|---|---|
| 1 | DATA | DJI | ACF | |
| 2 | 29/09/2009 | 9742,2 | 0,89359 | =COVAR($C$2:C51,C3:$C$52)/VAR($C$2:$C$52) |
| 3 | 28/09/2009 | 9789,36 | 0,803544 | =COVAR($C$2:C50,C4:$C$52)/VAR($C$2:$C$52) |
| 4 | 25/09/2009 | 9665,19 | 0,719049 | =COVAR($C$2:C49,C5:$C$52)/VAR($C$2:$C$52) |
| 5 | 24/09/2009 | 9707,44 | 0,670069 | =COVAR($C$2:C48,C6:$C$52)/VAR($C$2:$C$52) |
| 6 | 23/09/2009 | 9748,55 | 0,613015 | =COVAR($C$2:C47,C7:$C$52)/VAR($C$2:$C$52) |
| 7 | 22/09/2009 | 9829,87 | 0,562698 | =COVAR($C$2:C46,C8:$C$52)/VAR($C$2:$C$52) |
| 8 | 21/09/2009 | 9778,86 | 0,523415 | =COVAR($C$2:C45,C9:$C$52)/VAR($C$2:$C$52) |
| 9 | 18/09/2009 | 9820,2 | 0,466678 | |
| 10 | 17/09/2009 | 9783,92 | 0,433936 | |
| 11 | 16/09/2009 | 9791,71 | 0,409159 | |
| 12 | 15/09/2009 | 9683,41 | 0,421849 | |
| 13 | 14/09/2009 | 9626,8 | 0,427594 | |
| 14 | 11/09/2009 | 9605,41 | 0,426171 | |
| 15 | 10/09/2009 | 9627,48 | 0,408386 | |
| 16 | 09/09/2009 | 9547,22 | 0,413218 | |
| 17 | 08/09/2009 | 9497,34 | 0,404584 | |
| 18 | 04/09/2009 | 9441,27 | 0,369611 | |
| 19 | 03/09/2009 | 9344,61 | | |
| 20 | 02/09/2009 | 9280,67 | | |
| 21 | 01/09/2009 | 9310,6 | | |

Da fórmula (mostramos somente os primeiros sete valores e cálculos) fica claro que a parte da *variância* é fácil, isto é, apenas o intervalo $C$2:$C$52 no nosso caso. A *covariância* é apenas um pouco mais difícil para calcular. Os intervalos são:

$C$2:C51;C3:$C$52

$C$2:C50;C4:$C$52

$C$2:C49;C5:$C$52

$C$2:C48;C6:$C$52, etc.

Isto significa que se copiarmos para as células abaixo, C51 tornar-se-á C52, depois C53, etc. Para evitar este problema, podemos copiar a fórmula para baixo na coluna, mas precisamos manualmente mudar C51 progressivamente numa sequência descendente. Vamos lá, com você. Os valores **ACF** são calculados na coluna D.

## PACF

O gráfico PACF é um gráfico dos coeficientes de correlação parciais entre a série e as defasagens dela própria. Uma *autocorrelação parcial* é quantia de correlação entre uma variável e uma defasagem dela própria que não é explicado pelas correlações em todas as *defasagens de ordem inferior*. A autocorrelação de uma série temporal Y na defasagem 1 é o coeficiente de correlação entre Y(t) e Y(t-1), o qual é presumivelmente também a correlação entre Y(t-1) e Y(t-2). Mas se Y(t) está correlacionado com Y(t-1), e Y(t-1) está igualmente correlacionado com Y(t-2), então devemos também esperar encontrar correlação entre Y(t) e Y(t-2). (De fato, a quantia de correlação que devemos esperar na defasagem 2 é precisamente o *quadrado* da correlação na defasagem 1). Assim, a correlação na defasagem 1 "propaga-se" para a

defasagem 2 e presumivelmente para defasagens de ordem superior. A autocorrelação *parcial* na defasagem 2 é portanto a diferença entre a correlação atual na defasagem 2 e a correlação esperada devido à propagação da correlação na defasagem 1.

Selecione agora a planilha (*pacf*). Esta mostra como a PACF é implementada e calculada no Excel. Os valores da PACF são especificados na coluna C. Os coeficientes de autocorrelação são definidos como o último coeficiente de uma equação de autoregressão parcial de ordem k. Esta é a fórmula geral:

$$\pi_\tau = \frac{\rho_\tau - \sum_{j=1}^{\tau-1} \pi_{\tau-1,j} \cdot \rho_{\tau-j}}{1 - \sum_{j=1}^{\tau-1} \pi_{\tau-1,j} \cdot \rho_{\tau-j}} \quad (13)$$

Onde $\tau > 1$, $\rho$ é a autocorrelação, $\pi$ é a PACF.

A fórmula acima é implementada nas células E4, F5, G6, H7, I8, e assim por diante. (ver Figura 2.1 abaixo).

Arima.xls [Modo de Compatibilidade]

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | **Lag** | **AC** | **PAC** | | | | |
| 2 | k | $r_k$ | $r_{kj}$ | k,1 | k,2 | k,3 | k,4 |
| 3 | 1 | 0,8936 | 0,89 | **0,89** | | | |
| 4 | 2 | 0,8035 | 0,03 | 0,87 | **0,03** | | |
| 5 | 3 | 0,719 | -0,02 | 0,87 | 0,04 | **-0,02** | |
| 6 | 4 | 0,6701 | 0,13 | 0,87 | 0,03 | -0,13 | **0,13** |
| 7 | 5 | 0,613 | -0,05 | 0,88 | 0,03 | -0,13 | 0,17 |
| 8 | 6 | 0,5627 | 0,00 | 0,88 | 0,03 | -0,13 | 0,17 |
| 9 | 7 | 0,5234 | 0,05 | 0,88 | 0,03 | -0,14 | 0,18 |
| 10 | 8 | 0,4667 | -0,11 | 0,89 | 0,03 | -0,14 | 0,20 |
| 11 | 9 | 0,4339 | 0,09 | 0,90 | 0,01 | -0,14 | 0,21 |
| 12 | 10 | 0,4092 | 0,05 | 0,89 | 0,02 | -0,15 | 0,21 |
| 13 | 11 | 0,4218 | 0,15 | 0,89 | 0,01 | -0,12 | 0,18 |
| 14 | 12 | 0,4276 | 0,04 | 0,88 | 0,02 | -0,12 | 0,19 |
| 15 | 13 | 0,4262 | -0,02 | 0,88 | 0,02 | -0,12 | 0,19 |
| 16 | 14 | 0,4084 | -0,03 | 0,88 | 0,02 | -0,12 | 0,19 |
| 17 | 15 | 0,4132 | 0,12 | 0,88 | 0,02 | -0,12 | 0,17 |
| 18 | 16 | 0,4046 | -0,06 | 0,89 | 0,01 | -0,12 | 0,18 |
| 19 | 17 | 0,3696 | -0,12 | 0,88 | 0,03 | -0,14 | 0,18 |
| 20 | | | | | | | |

Daily Sales(1) Daily Sales

Figura 2.1

Podemos ver que o cálculo da PACF é um pouco mais difícil e complexo. Felizmente, escrevi uma macro para simplificar os seus cálculos. Para usar esta macro, você precisa carregar o **nn_Solver** no seu Excel. Eu mostrarei a você os passos de como fazer isto com um exemplo mais tarde. (ver Apêndice A sobre como carregar **nn_Solver**).

***ii) Como usar o par de funções ACF e PACF para identificar um modelo apropriado?***

Um gráfico de pares nos fornecerá uma boa indicação de qual tipo de modelo queremos tomar em consideração. O gráfico de um par de ACF e PACF é chamado de **correlograma**. A Figura 2.2 mostra três pares de correlogramas ACF e PACF.

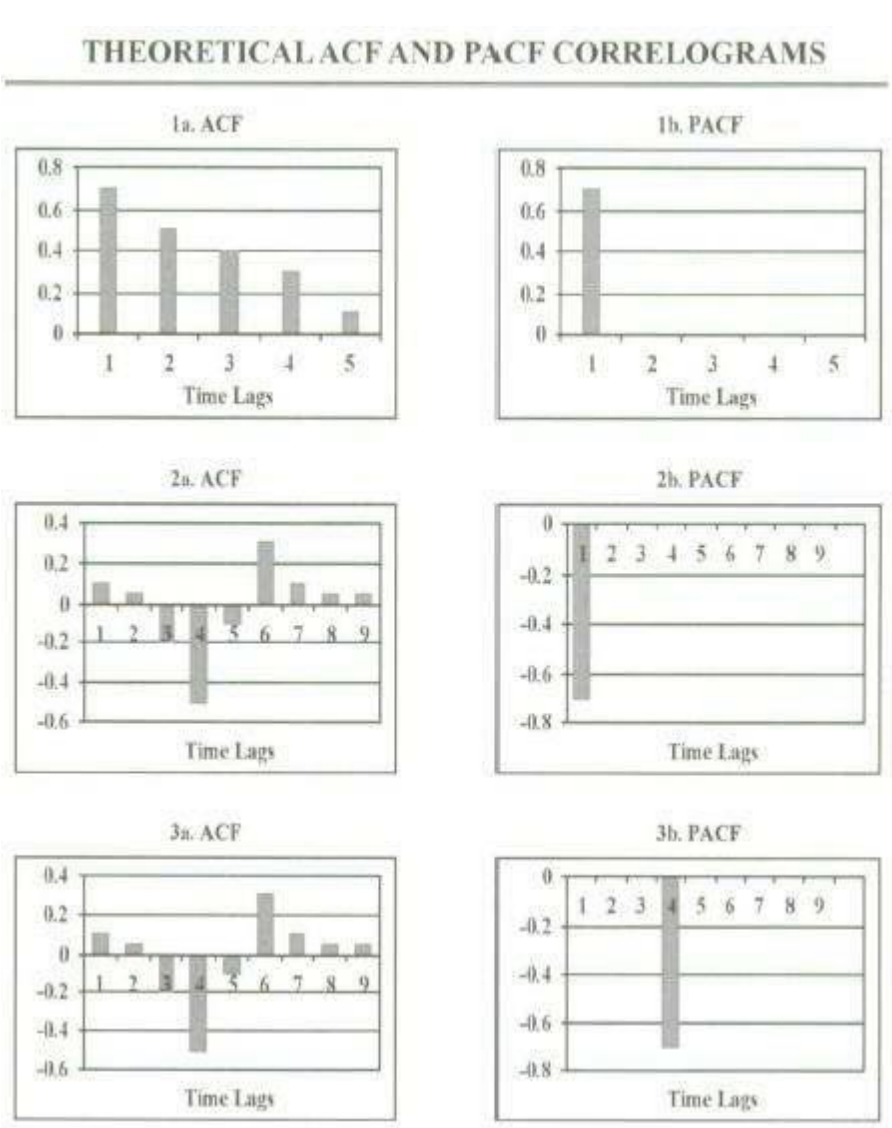



Figura 2.2

Na modelagem, se o correlograma atual se parecer com um destes três correlogramas teóricos, em que o ACF diminui rapidamente e o PACF tem somente um grande pico, escolheremos um modelo *AR(1)* para os dados. O "*1*" nos parênteses indica que o modelo **AR** precisa somente um termo autoregressivo, e o modelo é um AR de ordem 1. Note que os padrões ACF em 2a e 3a são os mesmos, mas o pico PACF maior em 2b ocorre na defasagem 1, enquanto que em 3b, ele ocorre na defasagem 4. Embora ambos correlogramas sugiram um modelo *AR(1)* para os dados, os padrões 2a e 2b indicam que um termo autoregressivo no modelo é de defasagem 1; mas o 3a e o 3b indicam que um termo autoregressivo no modelo é de defasagem 4.

Suponha que na Figura 2.2, ACF e PACF troquem seus padrões, isto é, os padrões do PACF se parecerão com aqueles da ACF e os padrões do ACF se parecerão com aqueles da PACF tendo somente uma estaca larga, então escolheremos um modelo *MA(1)*. Suponha que o PACF em cada par pareça o mesmo que o ACF, e então tentaremos um ARMA(1,1).

Até agora descrevemos os modelos **AR**, **MA** e **ARMA** mais simples. Os modelos de ordem superior podem ser assim identificados, é claro, com diferentes padrões de correlogramas.

Embora o catálogo acima não seja exaustivo, ele nos dá uma ideia razoável do que esperar quando se decidir sobre os modelos mais básicos. Infelizmente, o catálogo comportamental acima, das funções de autocorrelação e autocorrelação parcial, é somente teórico. Na prática, as autocorrelações e

autocorrelações parciais somente seguem vagamente estes padrões, que é o que torna esta abordagem subjetiva de previsão muito difícil. Em adição a isso, as séries temporais da vida real podem ser tratadas exatamente como uma amostra dos processos subjacentes. Portanto, as autocorrelações e autocorrelações parciais que são calculadas são apenas estimativas dos valores reais, sujeitos aos erros de amostragem.

As autocorrelações e autocorrelações parciais também fazem um importante papel na decisão se uma série temporal é estacionária, para que classe de modelos ela pertence e quantos coeficientes são caracterizados por ela. A questão que ainda está aberta é como calcular os coeficientes **a** e **c** que constituem um modelo particular.

Antes de continuarmos com como estimar **a** e **c**, retornaremos à questão de diferenciação e estacionariedade como prometido antes. Em geral devemos ser cautelosos onde a diferenciação está envolvida, a qual influenciará a classe do modelo. Será errado assumir que quando se garante que se a série é não estacionária, ela deverá simplesmente ser diferenciada. Muita diferenciação pode levar-nos a acreditar que a série temporal pertença a uma classe completamente diferente, que é apenas um dos problemas.

### Regras para diferenciação

Como, então, sabemos se temos exagerado e diferenciado demais a série? Uma das **regras básicas** é: se a primeira autocorrelação da série diferenciada for negativa e mais que -0,5, a série provavelmente foi diferenciada demais. Outra regra básica: se a variância para o nível superior de diferenciação crescer, devemos retornar ao nível anterior de diferenciação. Um dos princípios básicos é que o nível de diferenciação corresponda ao grau uma tendência polinomial que pode ser usada para ajustar a série temporal real.

A noção completa de diferenciação está relacionada ao conceito da assim chamada *raiz unitária*. **Raiz unitária** significa que **um** coeficiente *AR(1)* ou um *MA(1)* seja igual a um (unidade). Para modelos de ordem superior, isto significa que a soma de todos os coeficientes seja igual a um. Se isto acontecer temos um problema. Se um modelo *AR(1)* tiver uma raiz unitária, então este coeficiente **AR** deverá ser eliminado e o nível de diferenciação deverá ser aumentado. Para modelos *AR(p)* superiores, o número de coeficientes AR tem que ser reduzido e o nível de diferenciação aumentado. Para modelos MA mostrando raiz unitária, um coeficiente MA deverá também ser removido, mas o nível de diferenciação tem que ser diminuído. Algumas vezes não "pegamos" raízes unitárias anteriores suficientes, e produzimos previsões, que concentram muitos erros. Isto também é uma consequência das raízes unitárias, que significa que a redução nos coeficientes AR ou MA seja necessária.

Outra questão que precisamos responder é: *qual é o significado de* ***d****, como calculá-lo e quando o incluímos num modelo*?

Essencialmente, **d** nos modelos ARMA faz o mesmo papel que o *intercepto* na regressão linear. Nosso modelo aqui é chamado um modelo ARMA com um **nível**, onde **d** representa este nível inicial do modelo (um intercepto). Algumas vezes ele também é referido como parâmetro de tendência, ou uma constante.

Se quisermos calcular este parâmetro tendência, precisamos começar com a fórmula para o valor esperado de um processo **AR**, isto é, o *valor médio*. A **média** de qualquer processo *AR(p)* é calculada como:

$$Z = \frac{d}{(1-a(1)-\cdots-a(p))} \tag{17}$$

A qual, para AR(2), conduz:

$$Z = \frac{d}{(1-a(1)-a(2))} \tag{18}$$

Desta fórmula, o nível **d** (ou componente de tendência) para o processo *AR(2)* é calculado como:

$$d = Z^* [1 - a(1) - a(2)] \tag{19}$$

Em geral, o nível para qualquer processo *AR(p)* é calculado como:

$$d = Z * \left[1 - \sum_{i=1}^{p} a(p)\right] \quad (20)$$

Agora sabemos o que é e como calculá-lo, a parte aberta da questão ainda é: *quando o incluímos no nosso modelo*?

O conjunto de regras pode ser resumido como segue:

- Se uma série temporal é não estacionária na sua forma original e tivemos que diferenciá-la para torna-la estacionária, então a constante **não** é geralmente necessária.
- Séries temporais diferenciadas mais do que duas vezes **não** precisam de uma constante
- Se a série temporal original for estacionária com média zero, **não** é necessária uma constante.
- Se a série original for estacionária, mas com uma média significativamente grande (que efetivamente significa $\bar{x} \pm \sigma_x > 1$), a constante **é** necessária
    - Se o modelo não tiver uma componente AR (i.é, ele for um modelo MA ou IMA), então **a constante é igual ao valor médio** da série.
    - Se o modelo tiver um componente AR, a constante é calculada como em (2.20)

(Mostrarei outro exemplo onde calcularemos a constante **d**, mais tarde).

### Testando a *média zero* para indicar estacionariedade

O que acontece se o nível de diferenciação não for suficiente e o segundo nível é muito? Isto algumas vezes acontece na prática e uma série temporal parece ser estacionária, mesmo que o seu valor médio não seja zero, a despeito da exigência de estacionariedade que deverá existir. Se isto acontece, temos de assegurar que a média é no mínimo próxima de zero. A maneira mais fácil de fazer isto é calcular a média $\bar{w}$ da série diferenciada $w_t$, e subtraí-la de cada observação. (Vá à planilha (*Vendas Diárias*)). Os dados de A2:A101 são as vendas diárias de uma loja esportiva em milhares.

$z_t = w_t - \bar{w}$ , implementada na coluna B.

Uma vez tendo transformada a série temporal diferenciada de tal maneira, podemos calcular este valor médio da série transformada, $\bar{z}$ na célula E3 e verificar se ele é zero. Como verificamos se a média é zero ou próxima de zero? Primeiro precisamos estimar este erro padrão da série transformada. Você recordará que SE[4] é a razão entre o desvio padrão e a raiz quadrada do número de observações:

$$SE(z) = \frac{\sigma_z}{\sqrt{n}} \quad (14)$$

A média da *série temporal transformada*, $\bar{z}$, é considerada **não** zero se:

$$|\bar{z}| < 1{,}96 * SE(z) \quad (15)$$

Entrar com isso na célula E5. Não se preocupe com os símbolos matemáticos acima. Eles podem ser facilmente implementados numa planilha Excel. (ver Figura 2.3 abaixo). O resultado é **não zero** (ver célula E5) e do gráfico podemos ver que a série temporal parece não estacionária, i.é, ela está com tendência para cima. Então precisamos antes processar a série temporal que está sendo diferenciada. A diferenciação de **uma defasagem**, i.é, y(t) = y(t) – y(t-1), é aplicada. Os valores em C2:C100 são os valores de diferenciação de uma defasagem[5].

[4] Erro padrão
[5] Foram feitas as diferenças sobre os valores originais y(t).

| | D | E | F | G |
|---|---|---|---|---|
| 1 | | | | |
| 2 | **Média Original** | 7,146824125 | =MÉDIA(A2:A101) | |
| 3 | **Média Transformada** | -4,53859E-15 | =MÉDIA(B2:B101) | |
| 4 | **1.96*SE** | 0,0088899266453850 | =1.96*(RAIZ(B2:B101)/CONT.NÚM(B2:B101)) | |
| 5 | **Teste de Média Zero** | Não-zero | =SE(E3<E4,"Não-zero","Zero") | |
| 6 | | | | |

Vendas Diárias(1) | **Vendas Diárias** | acf | pacf

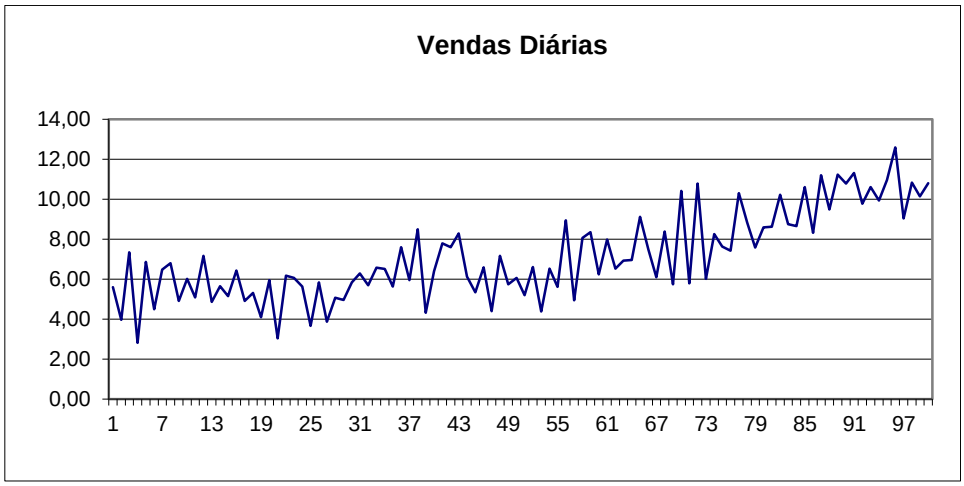



Figura 2.3

Há outra abordagem comum para transformações, que evita diferenciação. Em finanças, por exemplo, frequentemente estamos mais interessados nos retornos, i.é, se vendermos as ações hoje ($y_t$), quanto ganharemos quando comparado com quando as comparamos ($y_{t-1}$). Matematicamente isto é simplesmente: $\frac{y_t - y_{t-1}}{y_{t-1}}$. Mesmo se os valores das ações estiverem pulando descontroladamente, a série de tais retornos calculados geralmente será estacionária. A expressão matemática acima é conhecida ser aproximadamente igual a log ($y_t$)-log($y_{t-1}$), que frequentemente é usada para se calcular retornos. Esta expressão pode também ser usada para transformar uma série temporal para uma forma estacionária. Algumas séries estacionárias não são estritamente estacionárias e embora tenham uma média constante, suas variâncias não são constantes (lembre-se da ideia da homocedasticidade?). A transformação log sugerida aqui é sabida reduzir a heterocedasticidade.

Após uma série estacionária for colocada no lugar, um modelo básico pode agora ser identificado. Existem três modelos básicos, **AR** (autoregressivo), **MA** (média móvel) e, um combinado, **ARMA**, em adição aos RD (regular diferenciação) especificados anteriormente se combinam para fornecer as ferramentas disponíveis. Quando a diferenciação regular for aplicada junto com **AR** e **MA**, eles são referidos como **ARIMA**, com o **I** indicando "integrado" e se referindo ao procedimento de diferenciação.

Tenha em mente que estamos usando o método (1) para identificar o modelo. Até agora tenho 3 componentes que são importantes para nós entendermos para identificar o modelo:

- A ACF e PACF
- Dados estacionários
- Diferenciação

Vamos usar um exemplo de planilha para mostrar como calcular a ACF e PACF primeiro e depois então demonstrar o que acabamos de discutir, i.é, usar ACF e PACF para determinar os parâmetros **p** e **q** como no **ARMA(p,q)**.

Copiamos os valores diferenciados com uma defasagem em C2:C100 e os colocamos no intervalo C2:C100 na planilha (*Vendas Diárias (1)*). O gráfico abaixo[6] agora se parece estacionário e aleatório. E o teste indica também que a série temporal tem média zero na célula L4.

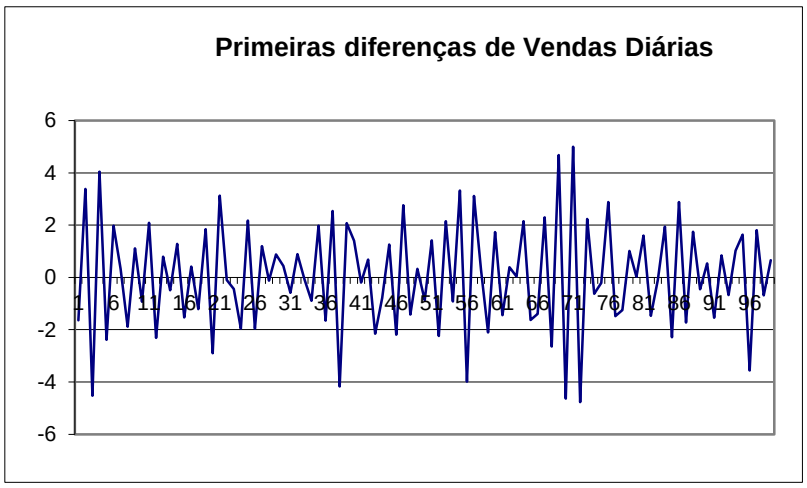



Figura 2.4

Agora precisamos calcular o ACF e o PACF. Embora, mostrei a você como os calcular manualmente (ver planilha *(acf)* e planilha *(pacf)*), é, ainda, muito tedioso, especialmente quando você calcula o PACF. Felizmente, você pode usar o suplemento **Resolve_Previsão** escrito por mim para calcular a *ACF* e a *PACF* automaticamente, como também os seus respectivos *correlogramas*. Carregue o **Resolve_Previsão** no seu Excel. (ver Apêndice sobre como carregar o **Resolve_Previsão**).

1. Selecione *ACF-PACF* no menu **Resolve_Previsão** (ver Figura 2.4a)

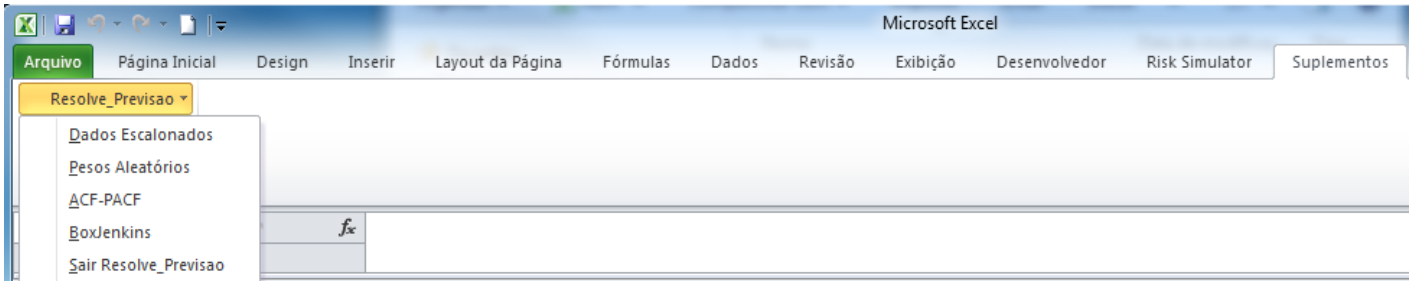



Figura 2.4a

Entre com a referência que você quer calcular no **Intervalo de Dados**. No nosso caso, entramos com C2:C100. (ver Figura 2.4b abaixo). **O intervalo de dados não pode começar com linha 1 como C1, A1, B1 e assim por diante. O Resolve_Previsao dará um erro. Sempre entre com os dados que você quer calcular iniciando na linha 2 como C2, A2, B2 e assim por diante numa planilha**.

[6] Da série original diferenciada de uma defasagem (lag-1)

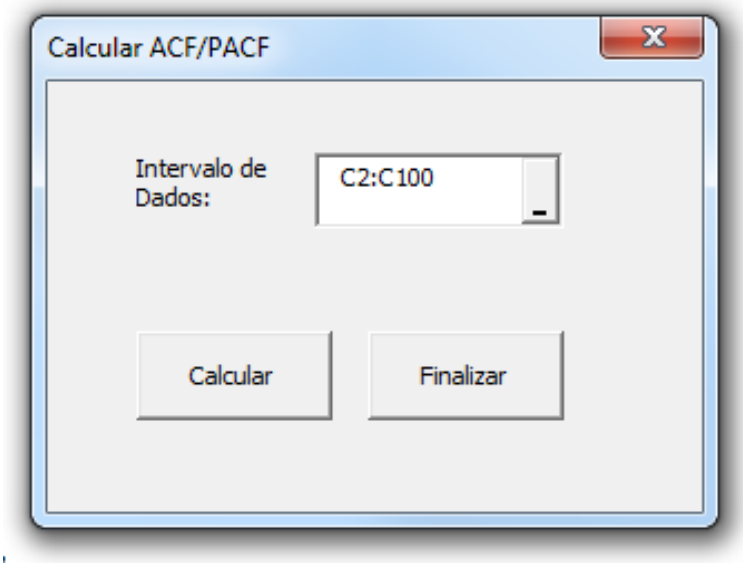



Figura 2.4b

2. Depois então clique no botão Calcular. O **ACF**, **PACF** e o **Erro Padrão** serão calculados. (ver Figura 2.4c)

Arima.xls [Modo de Compatibilidade]

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | AC | SE+ | SE- | PAC | SE+ | SE- | |
| 2 | | | -1,63509 | -0,77171 | 0,148769 | -0,14877 | -0,77171 | 0,1005 | -0,1005 | |
| 3 | | | 3,386727 | 0,480725 | 0,163709 | -0,16371 | -0,28388 | 0,100504 | -0,1005 | |
| 4 | | | -4,52892 | -0,28475 | 0,168638 | -0,16864 | -0,07402 | 0,100504 | -0,1005 | |
| 5 | | | 4,051568 | 0,086718 | 0,169088 | -0,16909 | -0,23466 | 0,100504 | -0,1005 | |
| 6 | | | -2,3803 | 0,029377 | 0,169139 | -0,16914 | -0,10969 | 0,100504 | -0,1005 | |
| 7 | | | 1,982004 | -0,07349 | 0,169462 | -0,16946 | -0,06177 | 0,100504 | -0,1005 | |
| 8 | | | 0,334183 | 0,107416 | 0,170148 | -0,17015 | 0,017955 | 0,100504 | -0,1005 | |
| 9 | | | -1,89037 | -0,03725 | 0,17023 | -0,17023 | 0,205491 | 0,100504 | -0,1005 | |
| 10 | | | 1,106627 | -0,02557 | 0,170269 | -0,17027 | 0,121509 | 0,100504 | -0,1005 | |
| 11 | | | -0,93435 | 0,042956 | 0,170379 | -0,17038 | 0,068459 | 0,100504 | -0,1005 | |
| 12 | | | 2,084561 | -0,06642 | 0,17064 | -0,17064 | 0,042051 | 0,100504 | -0,1005 | |
| 13 | | | -2,31443 | 0,042944 | 0,170749 | -0,17075 | -0,05157 | 0,100504 | -0,1005 | |
| 14 | | | 0,790386 | -0,02524 | 0,170787 | -0,17079 | -0,10051 | 0,100504 | -0,1005 | |
| 15 | | | -0,49097 | 0,047749 | 0,170922 | -0,17092 | -0,00648 | 0,100504 | -0,1005 | |
| 16 | | | 1,280339 | -0,0709 | 0,171218 | -0,17122 | -0,12331 | 0,100504 | -0,1005 | |
| 17 | | | -1,52955 | 0,07532 | 0,171553 | -0,17155 | -0,09188 | 0,100504 | -0,1005 | |
| 18 | | | 0,410206 | -0,09645 | 0,1721 | -0,1721 | -0,10647 | 0,100504 | -0,1005 | |
| 19 | | | -1,21685 | 0,084461 | 0,172518 | -0,17252 | -0,08978 | 0,100504 | -0,1005 | |
| 20 | | | 1,84515 | -0,02287 | 0,172548 | -0,17255 | 0,102158 | 0,100504 | -0,1005 | |
| 21 | | | -2,90236 | -0,02769 | 0,172593 | -0,17259 | 0,069004 | 0,100504 | -0,1005 | |
| 22 | | | 3,127629 | 0,020603 | 0,172618 | -0,17262 | -0,06735 | 0,100504 | -0,1005 | |
| 23 | | | -0,10112 | 0,011322 | 0,172626 | -0,17263 | 0,082795 | 0,100504 | -0,1005 | |
| 24 | | | -0,44532 | -0,04783 | 0,172759 | -0,17276 | 0,021475 | 0,100504 | -0,1005 | |
| 25 | | | -1,96425 | 0,109838 | 0,173463 | -0,17346 | 0,132232 | 0,100504 | -0,1005 | |
| 26 | | | 2,177237 | 0,199 | 0,176764 | -0,17676 | 0,10597 | 0,100504 | -0,1005 | |

Vendas Diárias(1) | Vendas Diárias | acf | pacf | Plan1

Construa os gráficos abaixo usando os dados calculados. (ver Fig. 2.5 e Fig. 2.6). A função autocorrelação e a função autocorrelação parcial para os dados das receitas de vendas diferenciadas são dados na Fig. 2.5 e Fig. 2.6.

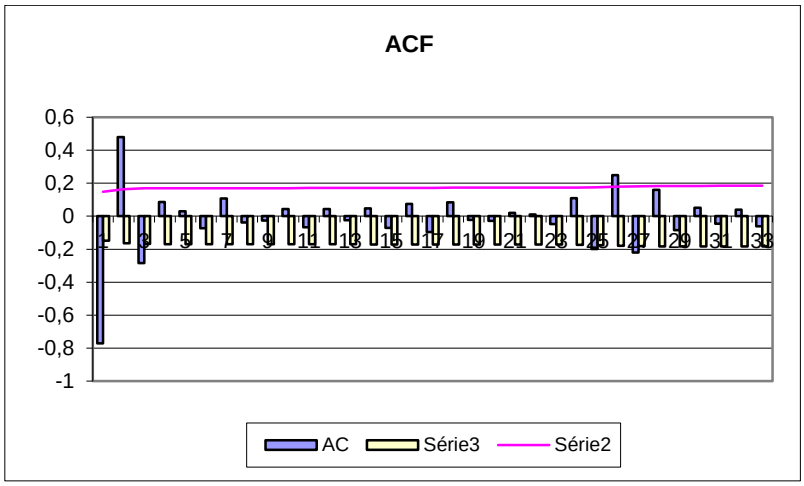



Figura 2.5

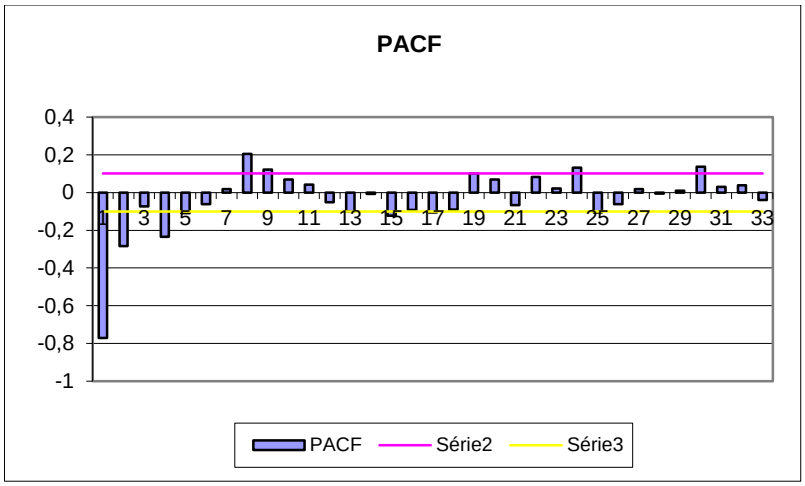
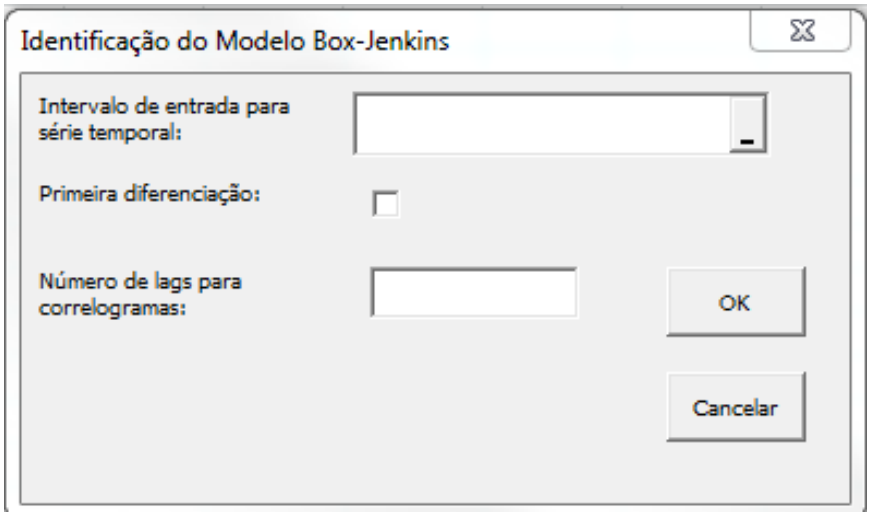



Figura 2.6

Isto pode ser feito usando também o suplemento **Resolve_Previsao** apenas clicando no item de menu BoxJenkins. Irá aparecer a janela:




Entre com o intervalo de células com os dados na *combobox*, marque a caixa de verificação para primeira diferenciação e adote o número de lags como 20 na caixa de texto. Clique OK e aparecerá uma nova pasta com os resultados e gráficos. Interessante, não!

A função auto-correlação parcial na Fig. 2.6 mostra dois coeficientes como significativamente não zero, implicando que isto é um modelo ARMA(p,q). A função autocorrelação confirma esta suposição como mostra o padrão usualmente associado com um modelo ARMA(p,q). Dado que temos que diferenciar a série temporal original, o modelo que usaremos, portanto, é ARIMA(2,1,1) ou ARMA(2,1)

## B) ESTIMAÇÃO DO MODELO

A equação para este modelo é:

$$\mathbf{y(t) = d + a(1)*y(t-1) + a(2)*y(t-2) – e(t) – c(1)*e(t-1)} \quad (16)$$

Vamos implementar esta fórmula numa planilha para otimizar os coeficientes, ajustar o modelo e produzir previsões. Abra a planilha (*Vendas Diárias(2)*). Os valores das vendas diferenciadas com 1 defasagem ($y_t - t_{t-1}$) são entrados na coluna A. Na coluna B estão os resíduos. Na coluna C está a fórmula completa. Pressione CTRL + ~, para ver a fórmula na sua planilha Excel. (ver Figura 2.7 abaixo):

| | A | B | C |
|---|---|---|---|
| 1 | **y(t)** | **e(t)** | **y(t) = d + a(1)y(t-1) + a(2)y(t-2) - c(1)e(t-1)** |
| 2 | -1,635086447 | 0,00 | =E8 |
| 3 | 3,386726733 | 0,00 | =E8 |
| 4 | -4,528916922 | =A4-($E$8+($E$2*A3+$E$3*A2)-$E$4*B3) | =$E$8+($E$2*A3+$E$3*A2)-$E$4*B3 |
| 5 | 4,051567782 | =A5-($E$8+($E$2*A4+$E$3*A3)-$E$4*B4) | =$E$8+($E$2*A4+$E$3*A3)-$E$4*B4 |
| 6 | -2,380298944 | =A6-($E$8+($E$2*A5+$E$3*A4)-$E$4*B5) | =$E$8+($E$2*A5+$E$3*A4)-$E$4*B5 |
| 7 | 1,982004137 | =A7-($E$8+($E$2*A6+$E$3*A5)-$E$4*B6) | =$E$8+($E$2*A6+$E$3*A5)-$E$4*B6 |
| 8 | 0,334182644 | =A8-($E$8+($E$2*A7+$E$3*A6)-$E$4*B7) | =$E$8+($E$2*A7+$E$3*A6)-$E$4*B7 |
| 9 | | | |

Podemos atribuir valores iniciais a **a**(1) = **a**(2) = **c**(1) = 0,1 e verificar se precisamos calcular **d**. A Figura 2.8 abaixo fornece um gabarito (template) para isso, e alguns outros cálculos, que explicaremos mais abaixo.

| | D | F |
|---|---|---|
| 1 | | **Inicial** |
| 2 | a(1) | 0,1 |
| 3 | a(2) | 0,1 |
| 4 | c(1) | 0,1 |
| 5 | Média | 0,05 |
| 6 | Desv. Pad. | 2,08 |
| 7 | Medida | 2,13 |
| 8 | d | 0 |
| 9 | μ implicado | 0,0000 |
| 10 | SSE | 377,07 |
| 11 | | |
| 12 | -1<a(2)<1 | 0,1 |
| 13 | a(1)+a(2)<1 | 0,2 |
| 14 | a(2)-a(1)<1 | 0 |
| 15 | –1<c(1)<1 | 0,1 |
| 16 | | |
| 17 | $\bar{e}$ | 0,23750 |
| 18 | $SE_e$ | 0,199149636 |
| 19 | Valor | 0,390333287 |
| 20 | Verdito: | **Média zero** |
| 21 | | |
| 22 | | |
| 23 | | **Teste de Durbin - Watson** |
| 24 | | 1292,5951 |
| 25 | | 377,0698 |
| 26 | | 3,42800 |

Figura 2.8

Os valores iniciais de **a**(1), **a**(2) e **c**(1) estão definidos nas células F2, F3 e F4. As células E5 e E6 contêm a média e o desvio padrão da série temporal[7]. Desde que temos aplicado a diferenciação à série temporal, o valor **d** não é necessário e entra-se com 0 na célula E8. Mostrarei para você outro exemplo onde calcularemos o **d** mais tarde quando usarmos o método (2).

Nosso conjunto de dados de receitas de venda era originalmente não estacionário e teve que ser diferenciado antes que a modelagem pudesse ser aplicada. Esta é a razão pela omissão da constante **d** em primeiro lugar. Assim definimos **d** como 0 neste exemplo. (Ver Fig. 2.8 acima).

Da fórmula 2.16 podemos facilmente extrair **e(t)**, que é:

$$\mathbf{e(t) = y(t) - [d + a(1)*y(t-1) + a(2)*y(t-2) - c(1)*e(t-1)]}$$

A fórmula acima mostra como calcular **e(1)**. Mas precisamos conhecer **e(0)**, que não conhecemos. A convenção é atribuir zeros para todos os valores desconhecidos de **e(0)**. Na Figura 2.7 acima, podemos ver zero na célula B2 e B3, que são as primeiras células necessárias para realizar este cálculo. Como o modelo é um ARMA(2,1), atribuímos 0 também para B3.

Agora temos todos os erros e(t), dado apenas os valores iniciais de a(1), a(2) e c(1), podemos calcular a assim chamada *soma condicional dos quadrados dos resíduos* (SSE), que é condicional nos valores de **a**(1) e **c**(1). A fórmula para **SSE** é:

$$SSE(a,c) = \sum_{t=1}^{n} e^2(t)$$

A célula F10 nos dá o valor de **SSE** = 377,07 inicialmente, que foi obtido usando a função Excel =SOMAQUAD(B2:B100). Esta célula é instrumental para estimar o valor ótimo de **a**(1), **a**(2) e **c**(1), que esperamos conduzir à melhor previsão possível. Para chegar a isto, usaremos o Solver do Excel. Nosso objetivo é *minimizar* o **SSE** (i.é, o valor da célula F10), mudando os valores de F2:F4, i.é, os valores de **a**(1), **a**(2) e **c**(1). Como antes, precisamos definir a **região admissível** que garantirá que o nosso modelo seja estacionário e invertível. Para processos, ARIMA(2,1,1), isto é: -1 < a(1) < 1 e -1 < c(1) < 1, ou, |a(1)| < 1 e |c(1)| < 1. As células F12 até F15 definem estas condições.

Antes de mostrarmos como usar o Solver, precisamos entender mais um ponto sobre os coeficientes de **AR(p)**, **a**(1), **a**(2), etc. Um processo que é gerado usando estes coeficientes tem que ser estacionário. Em outras palavras, certos valores de **a**(1), **a**(2), etc., não necessariamente gerarão um processo estacionário. Para satisfazer esta condição estrita de estacionariedade, precisamos definir a **região admissível** para estes coeficientes.

No caso de **AR(1)**, esta **região admissível** é *definida* como:

-1 < **a**(1) < 1 (ou, |**a**(1)| < 1).

No caso de **AR(2)**, esta **região admissível** é *definida* por três condições:

**a**(2) + **a**(1) < 1, **a**(2) – a(1) < 1 e -1 < **a**(2) < 1 (ou, |**a**(2)| < 1) e -1 < **c**(1) < 1

Podemos ver que nossas estimativas iniciais de **a**(1), **a**(2), **c**(1) na Figura 2.8 satisfazem todas estas condições de estacionariedade. Estes parâmetros são entrados na célula F12 até a F15. Uma última coisa antes de mostrar como usar o Solver para calcular os coeficientes.

Agora que entendemos modelagem (no mínimo para esta classe de modelos), devemos estabelecer se os valores estimados dos coeficientes do modelo são verdadeiramente aqueles melhores disponíveis. Tradicionalmente esta questão envolve cálculos complicados e muito complexos, que garantem que o

[7] Valores das vendas diferenciadas com um defasagem.

máximo dos estimadores mais prováveis seja selecionado. Felizmente com ajuda do Solver do Excel, muitas destas operações não são necessárias. Vamos fazê-las...

Nosso objetivo é minimizar o valor **SSE** na célula F10.

A célula F10 nos dá o valor de **SSE** = 377,07 inicialmente, que foi obtido usando a função = SOMAQUAD(B2:B100) do Excel.

Esta célula, juntamente com as células F12 até F15 é o instrumental para se estimar o valor ótimo de **a**(1), **a**(2) e **c**(1), que esperamos conduzir à melhor previsão possível. Para chegar a isto, usaremos o Solver do Excel. Nosso objetivo é minimizar **SSE** (i.é, o valor da célula F10), mudando os valores de F2:F4, i.é, os valores de **a**(1), **a**(2) e **c**(1). Como antes, precisamos definir a **região admissível** que garantirá que o nosso modelo seja estacionário e invertível. Para processos ARIMA(2,1,1), isto é: -1 < **a**(1) < 1 e -1 < **c**(1) <1, ou, |**a**(1)| < 1 e |**c**(1)| < 1. As células F12 até F15 definem estas condições.

Depois de invocar Solver no grupo de ferramentas Análise na guia Dados, uma caixa de diálogo aparece como mostrado na Figura 2.9 abaixo, onde entraremos com todos os parâmetros nesta caixa de diálogo.

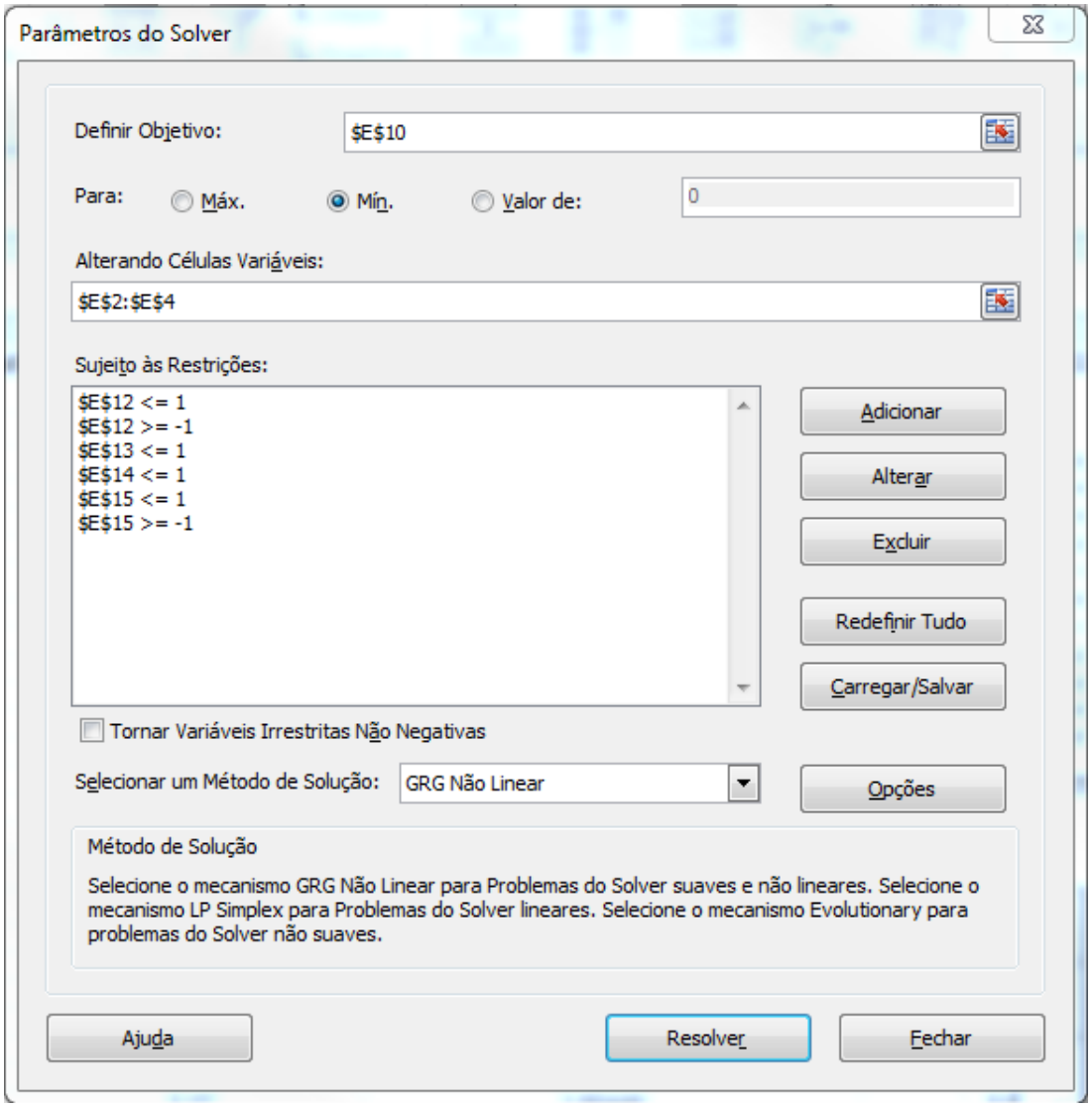


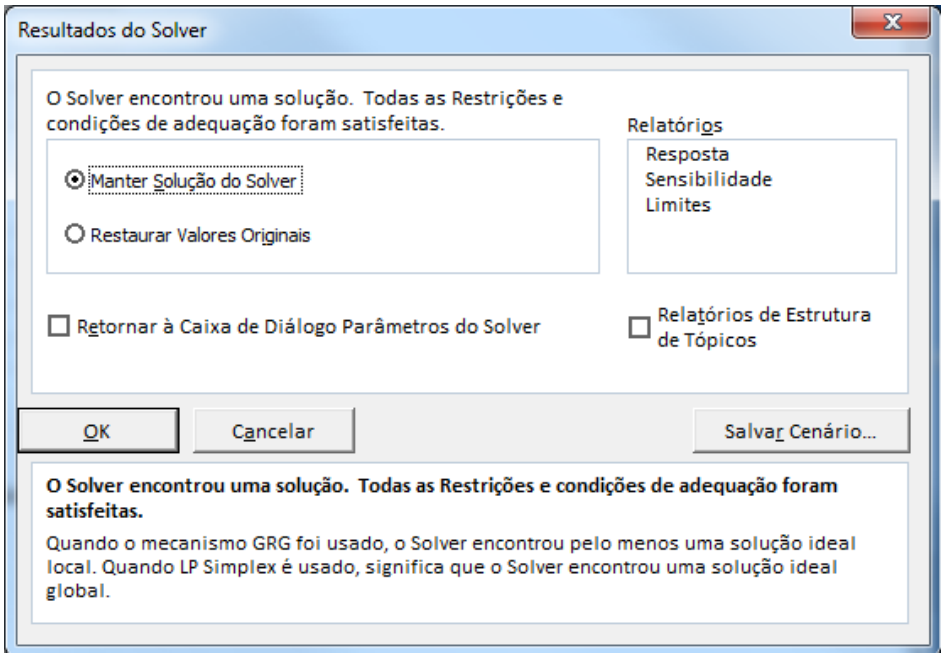

Figura 2.10

a. Defina a Célula Alvo: F10 (o **SSE**)
b. Mudando as Células: F2:F4 (**a**(1), **a**(2), **c**(1))
c. Os vínculos como mostrados na célula F12:F15

| | D | F | G |
|---|---|---|---|
| 1 | | Inicial | |
| 2 | a(1) | 0,1 | |
| 3 | a(2) | 0,1 | |
| 4 | c(1) | 0,1 | |
| 5 | Média | | |
| 6 | Desv. Pad. | | |
| 7 | Medida | | |
| 8 | d | | |
| 9 | μ implicado | | |
| 10 | SSE | 377,07 | |

Figura 2.11- **Antes da otimização**

d. Clique no botão Resolver. O Solver iniciará a otimização.




e. Manter Solução do Solver

| | D | E | F | G |
|---|---|---|---|---|
| 1 | | **Final** | **Inicial** | |
| 2 | a(1) | -0,537871274 | 0,1 | |
| 3 | a(2) | 0,058098633 | 0,1 | |
| 4 | c(1) | 0,614100745 | 0,1 | |
| 5 | Média | 0,05 | | |
| 6 | Desv. Pad. | 2,08 | | |
| 7 | Medida | 2,13 | | |
| 8 | d | 0 | | |
| 9 | μ implicado | 0,0000 | | |
| 10 | SSE | 137,09 | 377,07 | |
| 11 | | | | |
| 12 | -1<a(2)<1 | 0,058098633 | 0,1 | |
| 13 | a(1)+a(2)<1 | -0,479772641 | 0,2 | |
| 14 | a(2)-a(1)<1 | 0,595969908 | 0 | |
| 15 | −1<c(1)<1 | 0,614100745 | 0,1 | |
| 16 | | | | |
| 17 | $\bar{e}$ | 0,16058 | | |
| 18 | $SE_{\bar{e}}$ | 0,118952502 | | |
| 19 | Valor | 0,233146903 | | |
| 20 | Verdito: | **Média zero** | | |
| 21 | | | | |
| 22 | | | | |
| 23 | | **Teste de Durbin - Watson** | | |
| 24 | | 261,0780 | =SOMAXMY2(B3:B100;B2:B99) | |
| 25 | | 137,0857 | =SOMAQUAD(B2:B100) | |
| 26 | | 1,90449 | =E24/E25 | |
| 27 | | | | |

Figura 2.13 - **Após a otimização**

A solução é instantaneamente encontrada e os valores aparecem como você pode ver na Figura 2.13 acima.

Como podemos ver, **a**(1) agora é -0,537871274, **a**(2) é 0,058098633 e **c**(1) torna-se 0,614100745, que dá um valor muito inferior de **SSE** = 137,09, comparado com o valor anterior de 377,07.

Como calculamos o **e**(t), implicitamente na coluna B, se realmente o quisermos, podemos explicitamente calcular os valores da série temporal ajustada, i.é, as previsões *ex-post* y(t) de acordo com este modelo. Nós usamos a fórmula:

$$y(t) = d + (-0{,}537871274*y(t-1)) + 0{,}058098633*y(t-2) - 0{,}614100745*e(t-1) \qquad (18)$$

em vez de

$$\mathbf{y(t) = d + a(1)*y(t-1) + a(2)*y(t-2) - e(t) - c(1)*e(t-1)}$$

Você pode ver que eu abandonei o **e**(t) da nossa fórmula quando os calculamos na coluna B para derivar os valores na Coluna C. A coluna C, na Figura 2.15, mostra os valores para y(t) e a Figura 2.14 mostra a fórmula usada para produzir a Figura 2.15.

| | C | D | E | F | G |
|---|---|---|---|---|---|
| 1 | y(t) = d + a(1)y(t-1) + a(2)y(t-2) - c(1)e(t-1) | | Final | Inicial | |
| 2 | =E8 | a(1) | -0,537871274 | 0,1 | |
| 3 | =E8 | a(2) | 0,058098633 | 0,1 | |
| 4 | =$E$8+($E$2*A3+$E$3*A2)-$E$4*B3 | c(1) | 0,614100745 | 0,1 | |
| 5 | =$E$8+($E$2*A4+$E$3*A3)-$E$4*B4 | Média | =MÉDIA(A2:A100) | 0,05 | |
| 6 | =$E$8+($E$2*A5+$E$3*A4)-$E$4*B5 | Desv. Pad. | =DESV.PAD(A2:A100) | 2,08 | |
| 7 | =$E$8+($E$2*A6+$E$3*A5)-$E$4*B6 | Medida | =E5+E6 | 2,13 | |
| 8 | =$E$8+($E$2*A7+$E$3*A6)-$E$4*B7 | d | 0 | 0 | |
| 9 | | μ implicado | =E8/(1-E2) | 0,0000 | |
| 10 | | SSE | =SOMAQUAD(B2:B100) | 377,07 | |
| 11 | | | | | |
| 12 | | -1<a(2)<1 | =E3 | 0,1 | |
| 13 | | a(1)+a(2)<1 | =E2+E3 | 0,2 | |
| 14 | | a(2)-a(1)<1 | =E3-E2 | 0 | |
| 15 | | −1<c(1)<1 | =E4 | 0,1 | |
| 16 | | | | | |
| 17 | | $\bar{e}$ | =MÉDIA(B2:B100) | 0,23750 | |
| 18 | | $SE_{\bar{e}}$ | =DESV.PAD(B3:B100)/RAIZ(CONT.NÚM(B3:B100)) | 0,199149636 | |
| 19 | | Valor | =1,96*E18 | 0,390333287 | |
| 20 | | Verdito: | =SE(E17>E19;"Média não zero";"Média zero") | Média zero | |
| 21 | | | | | |
| 22 | | | | | |
| 23 | | | | Teste de Durbin - Watson | |
| 24 | | | | =SOMAXMY2(B3:B100;B2:B99) | |
| 25 | | | | =SOMAQUAD(B2:B100) | |
| 26 | | | | =E24/E25 | |

Figura 2.14

| | A | B | C |
|---|---|---|---|
| 1 | y(t) | e(t) | y(t) = d + a(1)y(t−1) + a(2)y(t−2) − c(1)e(t−1) |
| 2 | -1,63509 | 0,00 | 0 |
| 3 | 3,38673 | 0,00 | 0,00 |
| 4 | -4,52892 | -2,61 | -1,92 |
| 5 | 4,05157 | -0,19 | 4,24 |
| 6 | -2,3803 | -0,05 | -2,33 |
| 7 | 1,982 | 0,43 | 1,55 |
| 8 | 0,33418 | 1,81 | -1,47 |
| 9 | -1,89037 | -0,72 | -1,17 |
| 10 | 1,10663 | -0,37 | 1,48 |

Figura 2.15

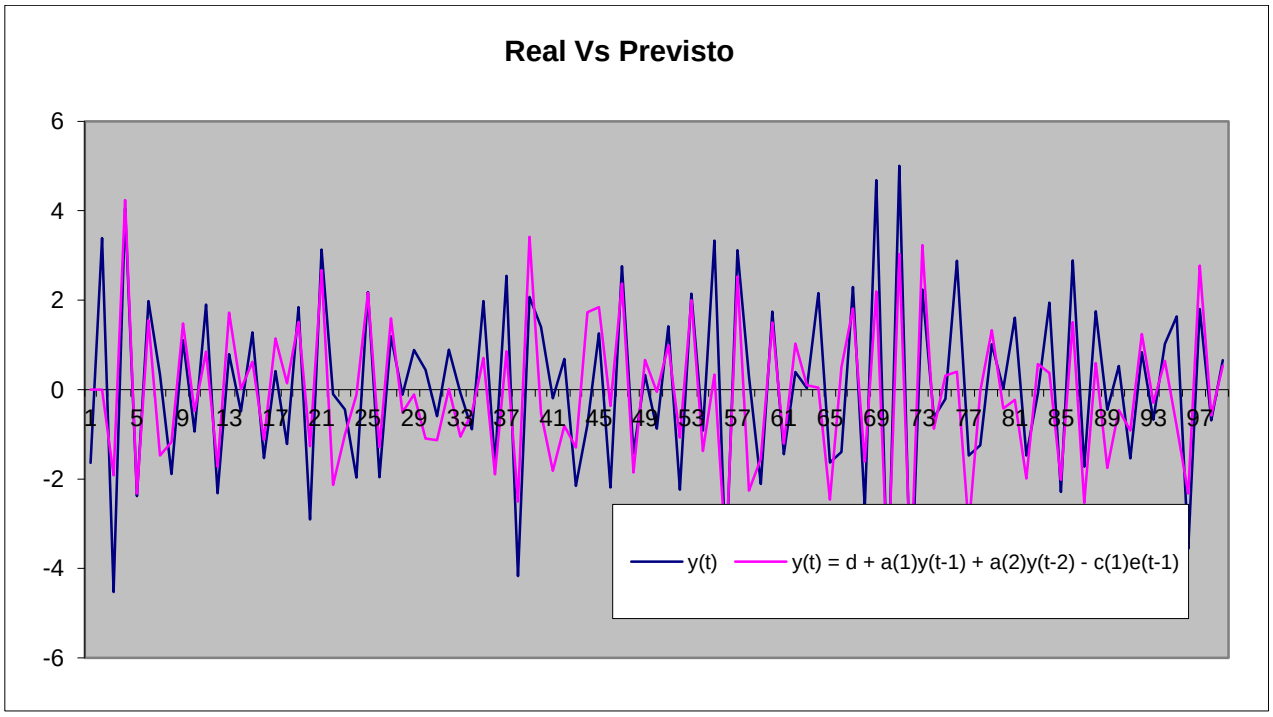



Figura 2.16

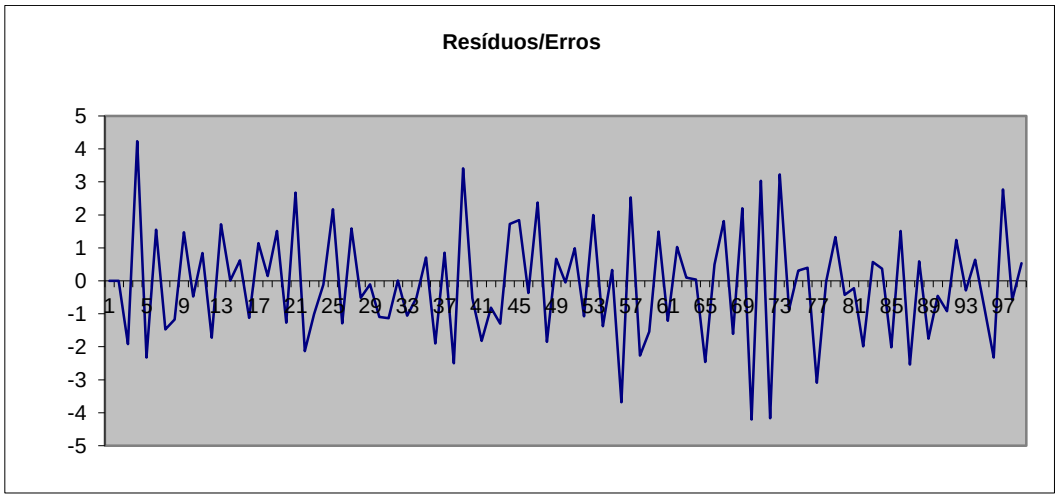



Figura 2.17

Quão aproximadamente os valores ajustados estão se casando com os da série temporal original pode ser visto na Figura 2.16 acima. Erros de previsão da coluna B são mostrados na Figura 2.17 acima e eles parecem distribuídos aleatoriamente, como esperado. Antes de aprontarmos para a previsão, precisamos fazer um diagnóstico verificando primeiro.

## C) DIAGNÓSTICO DE VERIFICAÇÃO

Como saber que produzimos um modelo razoável e que nosso modelo realmente reflete a série temporal real? Isto é uma parte do processo que Box e Jenkins se referem como diagnóstico de verificação. Usarei dois métodos para conduzir o diagnóstico.

Como esperamos, os erros de previsão são completamente aleatórios, o primeiro passo é plotá-los, como fizemos na Figura 2.17 acima por exemplo. Um dos requisitos é que a média residual deverá ser zero, ou próxima à zero. Para estabelecer que este é ocaso, precisamos estimar o erro padrão do erro médio. Isto é calculado como:

$$\sigma_e = \sqrt{\frac{\sum_{t=1}^{n}(e_t - \bar{e})^2}{n}} \quad (19)$$

$$SE_{\bar{e}} = \frac{\sigma_e}{\sqrt{n}} \quad \text{... na célula E18} \quad (20)$$

Onde $\sigma_e$ é o *desvio padrão residual*, $\bar{e}$ é o *erro médio*, $n$ é o número de erros e $SE_{\bar{e}}$ é o *erro padrão do erro médio*. Se a média residual $\bar{e}$ for maior que 1,96 erros padrões, então podemos dizer que ela é significativamente **não** zero:

$$\bar{e} > 1{,}96\, SE_{\bar{e}} \quad \text{... na célula E20} \quad (21)$$

Podemos tomar um exemplo da Coluna B para a qual os erros **e**(t) são calculados e mostrados nela. Como estimar o *erro residual padrão* $SE_e$ (erro padrão), está mostrado abaixo na Figura 2.18 e a fórmula está dada na Figura 2.19 abaixo:

| | D | E | F | G |
|---|---|---|---|---|
| 1 | | **Final** | **Inicial** | |
| 2 | a(1) | -0,537871274 | 0,1 | |
| 3 | a(2) | 0,058098633 | 0,1 | |
| 4 | c(1) | 0,614100745 | 0,1 | |
| 5 | Média | 0,05 | | |
| 6 | Desv. Pad. | 2,08 | | |
| 7 | Medida | 2,13 | | |
| 8 | d | 0 | | |
| 9 | μ implicado | 0,0000 | | |
| 10 | SSE | 137,09 | 377,07 | |
| 11 | | | | |
| 12 | -1<a(2)<1 | 0,058098633 | 0,1 | |
| 13 | a(1)+a(2)<1 | -0,479772641 | 0,2 | |
| 14 | a(2)-a(1)<1 | 0,595969908 | 0 | |
| 15 | −1<c(1)<1 | 0,614100745 | 0,1 | |
| 16 | | | | |
| 17 | $\bar{e}$ | 0,16058 | | |
| 18 | $SE_e$ | 0,118952502 | | |
| 19 | Valor | 0,233146903 | | |
| 20 | Verdito: | **Média zero** | | |
| 21 | | | | |
| 22 | | | | |
| 23 | | **Teste de Durbin - Watson** | | |
| 24 | | 261,0780 | =SOMAXMY2(B3:B100;B2:B99) | |
| 25 | | 137,0857 | =SOMAQUAD(B2:B100) | |
| 26 | | 1,90449 | =E24/E25 | |

Figura 2.18

| | C | D | E | F | G |
|---|---|---|---|---|---|
| 1 | y(t) = d + a(1)y(t-1) + a(2)y(t-2) - c(1)e(t-1) | | Final | Inicial | |
| 2 | =E8 | a(1) | -0,537871274 | 0,1 | |
| 3 | =E8 | a(2) | 0,058098633 | 0,1 | |
| 4 | =$E$8+($E$2*A3+$E$3*A2)-$E$4*B3 | c(1) | 0,614100745 | 0,1 | |
| 5 | =$E$8+($E$2*A4+$E$3*A3)-$E$4*B4 | Média | =MÉDIA(A2:A100) | 0,05 | |
| 6 | =$E$8+($E$2*A5+$E$3*A4)-$E$4*B5 | Desv. Pad. | =DESV.PAD(A2:A100) | 2,08 | |
| 7 | =$E$8+($E$2*A6+$E$3*A5)-$E$4*B6 | Medida | =E5+E6 | 2,13 | |
| 8 | =$E$8+($E$2*A7+$E$3*A6)-$E$4*B7 | d | 0 | 0 | |
| 9 | | μ implicado | =E8/(1-E2) | 0,0000 | |
| 10 | | SSE | =SOMAQUAD(B2:B100) | 377,07 | |
| 11 | | | | | |
| 12 | | -1<a(2)<1 | =E3 | 0,1 | |
| 13 | | a(1)+a(2)<1 | =E2+E3 | 0,2 | |
| 14 | | a(2)-a(1)<1 | =E3-E2 | 0 | |
| 15 | | −1<c(1)<1 | =E4 | 0,1 | |
| 16 | | | | | |
| 17 | | $\bar{e}$ | =MÉDIA(B2:B100) | 0,23750 | |
| 18 | | $SE_{\bar{e}}$ | =DESV.PAD(B3:B100)/RAIZ(CONT.NÚM(B3:B100)) | 0,199149636 | |
| 19 | | Valor | =1,96*E18 | 0,390333287 | |
| 20 | | Verdito: | =SE(E17>E19;"Média não zero";"Média zero") | Média zero | |
| 21 | | | | | |
| 22 | | | | | |
| 23 | | | | Teste de Durbin - Watson | |
| 24 | | | | =SOMAXMY2(B3:B100;B2:B99) | |
| 25 | | | | =SOMAQUAD(B2:B100) | |
| 26 | | | | =E24/E25 | |

Figura 2.19

A célula E20 contém uma breve declaração SE avaliando se a média $\bar{e}$, calculada em E17, é maior que o erro padrão vezes 1,96. No nosso modelo, isso não acontece e, então, temos **média zero,** a qual passa no teste.

Outro teste que é muito popular é o teste de **Durbin-Watson**, o qual é usado no contexto de verificação da validade dos modelos ARIMA.

A **estatística Durbin-Watson** é um teste estatístico usado para detectar a presença de autocorrelação nos *resíduos* de uma análise de regressão. É assim chamado depois de James Durbin e Geoffrey Watson. Se $\mathbf{e}_t$ é o *resíduo* associado com a observação no tempo t, então o teste estatístico é:

$$w = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \qquad (22)$$

Como *w* na célula E26 é aproximadamente igual a 2(1 – r), onde *r* é a autocorrelação amostral dos resíduos, *w* = 2 indica nenhuma autocorrelação. O valor de ***w*** sempre cai entre 0 e 4. Se a estatística de Durbin-Watson for substancialmente menor que 2, há evidência de correlação serial positiva. Como um princípio básico grosseiro, se Durbin-Watson for menor que 1,0, deverá ser causa para alarme. Valores pequenos de *w* indicam que os termos de erros sucessivos são, na média, próximos em valor um do outro, ou positivamente correlacionado. Se *w* > 2 os termos de erros sucessivos são, na média, muito diferentes em valor um do outro, i.é, negativamente correlacionados. Em regressões, isto pode implicar uma sub estimação do nível de significância estatística.

| | E | F | G |
|---|---|---|---|
| 22 | | | |
| 23 | Teste de Durbin - Watson | | |
| 24 | 261,0780 | =SOMAXMY2(B3:B100;B2:B99) | |
| 25 | 137,0857 | =SOMAQUAD(B2:B100) | |
| 26 | 1,90449 | =E24/E25 | |
| 27 | | | |

Figura 2.20

No nosso modelo temos 1,90449 na célula E26 o qual está muito próximo de 2, o que indica: nenhuma autocorrelação. Ver Figura 2.20 acima. Podemos agora seguir com a previsão.

## D) PREVISÃO

Agora estamos prontos para produzir previsões reais, i.é, aquelas que vão adiante no futuro. A equação pode ser aplicada "um passo adiante" para estimar $\hat{y}$(t) do observado y(t-1). Uma previsão "*k*-passos adiante" pode também ser feita pela aplicação recorrente da equação. Numa aplicação recorrente, o y observado no tempo 1 é usado para gerar o $\hat{y}$ estimado no tempo 2. Esta estimativa é então substituída com y(t-1) para obter o $\hat{y}$ estimado no tempo 3, e assim por diante. A previsão *k*-passos adiante eventualmente converge a zero quando o **horizonte de previsão**, *k*, aumentar. Vá célula A101:A105.

Faremos a previsão de acordo com a fórmula abaixo: ARIMA(2,1,1) ou ARMA(2,1). A fórmula é

$$y(t) = -0{,}537871274*y(t-1) + 0{,}058098633*y(t-2) - 0{,}614100745*e(t-1)$$

| | A | B | C |
|---|---|---|---|
| 1 | y(t) | e(t) | y(t) = d + a(1)y(t−1) + a(2)y(t−2) − c(1)e(t−1) |
| 2 | -1,63509 | 0,00 | 0 |
| 3 | 3,38673 | 0,00 | 0,00 |
| 4 | -4,52892 | -2,61 | -1,92 |
| 5 | 4,05157 | -0,19 | 4,24 |
| 6 | -2,3803 | -0,05 | -2,33 |
| 98 | 1,80283 | -0,97 | 2,77 |
| 99 | -0,68614 | -0,10 | -0,58 |
| 100 | 0,65656 | 0,12 | 0,54 |
| 101 | -0,58 | | -0,46 |
| 102 | 0,32 | | 0,29 |
| 103 | -0,22 | | -0,18 |
| 104 | 0,38868 | | 0,11 |
| 105 | 0,03799 | | -0,07 |
| 106 | | | |
| 107 | | | |

Figura 2.21

| | A | B | C |
|---|---|---|---|
| 1 | y(t) | e(t) | y(t) = d + a(1)y(t-1) + a(2)y(t-2) - c(1)e(t-1) |
| 2 | -1,635086447 | 0,00 | =E8 |
| 3 | 3,386726733 | 0,00 | =E8 |
| 4 | -4,528916922 | =A4-($E$8+($E$2*A3+$E$3*A2)-$E$4*B3) | =$E$8+($E$2*A3+$E$3*A2)-$E$4*B3 |
| 5 | 4,051567782 | =A5-($E$8+($E$2*A4+$E$3*A3)-$E$4*B4) | =$E$8+($E$2*A4+$E$3*A3)-$E$4*B4 |
| 6 | -2,380298944 | =A6-($E$8+($E$2*A5+$E$3*A4)-$E$4*B5) | =$E$8+($E$2*A5+$E$3*A4)-$E$4*B5 |
| 7 | 1,982004137 | =A7-($E$8+($E$2*A6+$E$3*A5)-$E$4*B6) | =$E$8+($E$2*A6+$E$3*A5)-$E$4*B6 |
| 8 | 0,334182644 | =A8-($E$8+($E$2*A7+$E$3*A6)-$E$4*B7) | =$E$8+($E$2*A7+$E$3*A6)-$E$4*B7 |
| 98 | 1,802826736 | =A98-($E$8+($E$2*A97+$E$3*A96)-$E$4*B | =$E$8+($E$2*A97+$E$3*A96)-$E$4*B97 |
| 99 | -0,686138281 | =A99-($E$8+($E$2*A98+$E$3*A97)-$E$4*B | =$E$8+($E$2*A98+$E$3*A97)-$E$4*B98 |
| 100 | 0,656560484 | =A100-($E$8+($E$2*A99+$E$3*A98)-$E$4*I | =$E$8+($E$2*A99+$E$3*A98)-$E$4*B99 |
| 101 | -0,58 | | =$E$8+($E$2*A100+$E$3*A101)-$E$4*B100 |
| 102 | 0,32 | | =$E$8+($E$2*C101+$E$3*A100) |
| 103 | -0,22 | | =$E$8+($E$2*C102+$E$3*C101) |
| 104 | 0,388682142 | | =$E$8+($E$2*C103+$E$3*C102) |
| 105 | 0,037989519 | | =$E$8+($E$2*C104+$E$3*C103) |
| 106 | | | |

Figura 2.22

A Figura 2.21 mostra a planilha contendo os números básicos e a Figura 2.22 mostra todos os cálculos.

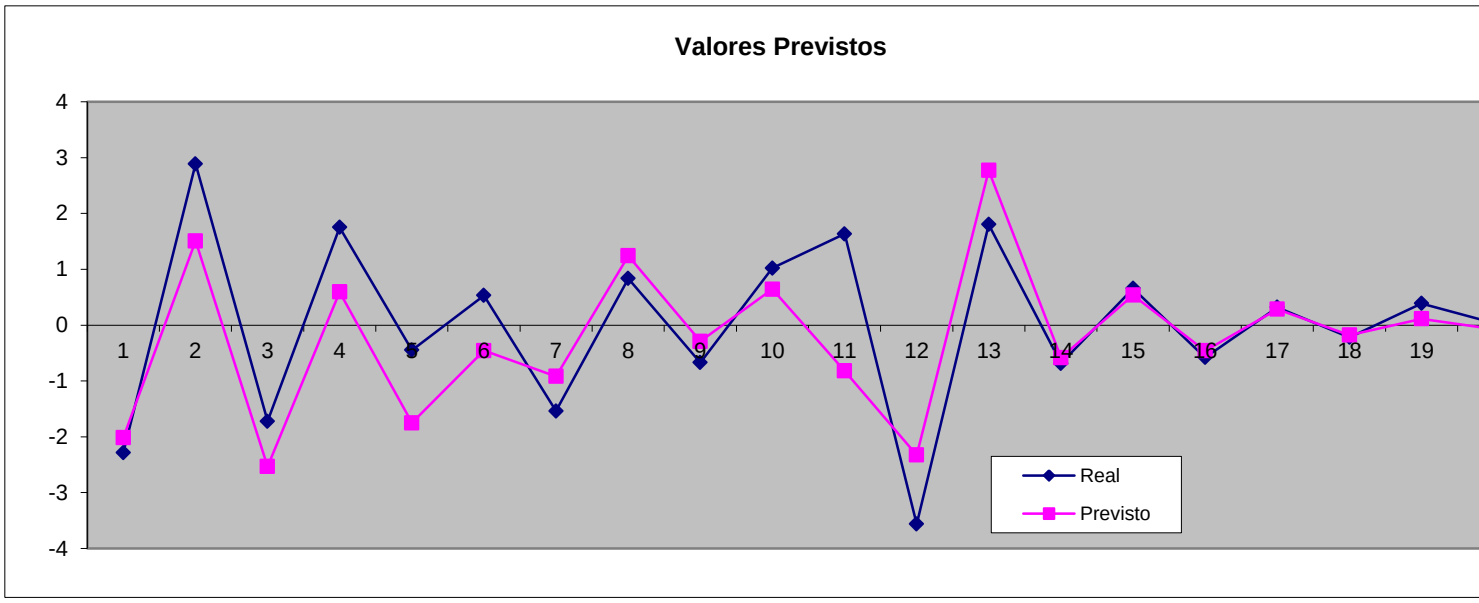



Figura 2.23

Como já explicamos, uma vez tendo executado os valores reais, os valores reais de y(t) são trocados pelos seus valores ajustados (iniciando em C102). Isto inevitavelmente degrada as previsões, e explicamos como modelos diferentes se comportam. Como podemos ver, nossa previsão para a célula C102 e C103 na Figura 2.21 é muito boa (como sabemos os valores reais, os colocamos nas células A101:A105). Infelizmente nossa previsão para a célula C104 começa a ser significativamente diferente do valor real conhecido na célula A104. Isto implica que para muitas séries temporais, o método de Box-Jenkins é um bom ajuste, mas somente para previsões de curto prazo.

Para resumir, nesta seção não somente foi mostrado o processo completo de identificação do modelo, ajustando-os e fazendo previsão, mas também apresentamos uma maneira muito ágil de fazê-la. Vinculamos os valores dos coeficientes ARMA diretamente com a soma dos quadrados dos resíduos, a

qual se tornou um valor alvo no Solver, e que num único passo produziu valores ótimos para estes coeficientes.

## Box Jenkins Automático/ARIMA: Método 2

O procedimento subjetivo delineado acima requer intervenção considerável dos economistas e estatísticos complementando a previsão. Várias tentativas foram feitas para automatizar as previsões. A mais simples delas ajusta uma seleção de modelos aos dados, decide qual é o "melhor" e depois então se o "melhor" for suficientemente bom usa este. Por outro lado a previsão é referida para voltar pela análise "padrão" pelos estatísticos e economistas. A seleção será baseada num critério tal como o **AIC** (*Akaike´s Information Criterion*) e critério Bayesiano (Schwarz) (**BIC** ou **SIC**).

Antes de seguir com a implementação do modelo automatizado numa planilha, há 2 importantes funções do Excel que quero explicar:

**i) SOMARPRODUTO()**

**ii) DESLOC()**

### i) SOMARPRODUTO()

No Excel, a função SOMARPRODUTO() multiplica os itens correspondentes nas matrizes e retorna a soma dos resultados.

A sintaxe para a função **SOMARPRODUTO()** é:

SOMARPRODUTO(matriz1; matriz2; ...; matrizN)

*matriz 1*, *matriz2*, ..., *matrizN* são intervalos de células ou matrizes que você quer multiplicar. Todas as matrizes devem ter o mesmo número de linhas e colunas. Você deve entrar com no mínimo 2 matrizes e você pode ter até 30 matrizes.

Nota: Se todas as matrizes fornecidas como parâmetros não tiverem o mesmo número de linhas e colunas, a função SOMARPRODUTO retornará o erro #VALUE!.

Se existirem valores não numéricos nas matrizes, estes valores são tratados como 0´s pela função SOMARPRODUTO().

Vamos dar uma olhada num exemplo:

=SOMARPRODUTO({1,2;3,4};{5,6;7,8})

O exemplo acima retornará 70. A SOMARPRODUTO calcula estas matrizes como segue:

=(1*5) + (2*6) + (3*7) + (4*8)

Você poderá também ter intervalos de referência no Excel.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | 1 | 2 | | 5 | 6 | |
| 2 | 3 | 4 | | 7 | 8 | |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |
| 6 | | | | | | |

Baseado na planilha Excel acima, você poderá entrar com a seguinte fórmula:

=SOMARPRODUTO(A1:B2;D1:E2)

Isto retornará também o valor 70. Outro exemplo:

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | 2 | | 1 | | | |
| 2 | 3 | | 2 | | | |
| 3 | 4 | | 3 | | | |
| 4 | 5 | | 4 | | | |
| 5 | | | | | | |
| 6 | 40 | =SOMARPRODUTO(A1:A4;C1:C4) | | | | |
| 7 | | | | | | |

Isto será (2*1) + (3*2) + (4*3) + (5*4) = 40

**ii) DESLOC()**

A função DESLOC() retorna uma célula ou intervalo de células que é um número especificado de linhas e/ou colunas de uma célula de referência. Neste tutorial explicaremos as aplicações mais comuns de DESLOC() e os erros que são cometidos quando se usa esta função no MS-EXCEL.

A sintaxe para **DESLOC()** é:

DESLOC(célula de referência; linhas; colunas; [altura]; [largura])

Os componentes entre colchetes podem ser omitidos na fórmula.

**Como funciona a função DESLOC do Excel?**

A função DESLOC() retorna uma célula ou intervalo de células que for especificado no número de linhas e/ou colunas da célula de referência. Para descrições específicas de cada componente, por gentileza ver o arquivo Ajuda do Excel.

Se algum componente, "linhas", "colunas", "altura" ou "largura", for deixado em branco, o Excel assumirá seu valor como zero. Por exemplo, se a fórmula for escrita como DESLOC(C38;;1;; ), o Excel interpretará isto como DESLOC(C38;0;1;0;0). Isto pode também ser escrito como DESLOC(C38;;1), desde que "altura" e "largura" podem ser omitidos.

Note que se "altura" e "largura" forem incluídos na fórmula, eles não podem ser iguais à zero ou resultará um erro #REF!. Os exemplos abaixo ilustram a função.

**Exemplo 1 de DESLOC()**

DESLOC(D10;1;2) dará o valor em F11 ou 7, i.é, o Excel retorna o valor da célula 1 linha abaixo e 2 colunas à direita de D10.

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 9 | | | | | | | | |
| 10 | | | | 1 | 2 | 3 | 4 | |
| 11 | | | | 5 | 6 | 7 | 8 | |
| 12 | | | | 9 | 10 | 11 | 12 | |
| 13 | | | | | | | | |

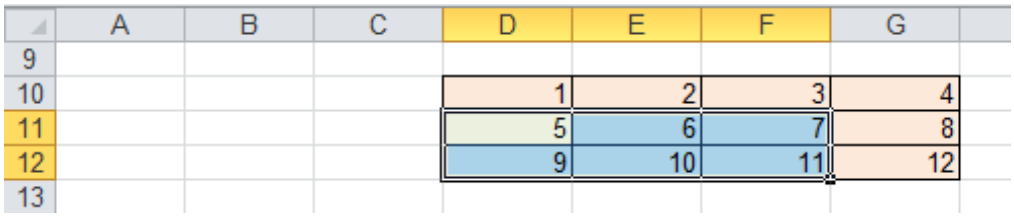
**Exemplo 2 de DESLOC()**

DESLOC(G12;-2;-2) dará o valor em E10 ou 2, i.é, o Excel retorna o valor da célula 2 linhas acima e duas colunas para a esquerda de G12.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 9 | | | | | | | |
| 10 | | | | 1 | 2 | 3 | 4 |
| 11 | | | | 5 | 6 | 7 | 8 |
| 12 | | | | 9 | 10 | 11 | 12 |
| 13 | | | | | | | |

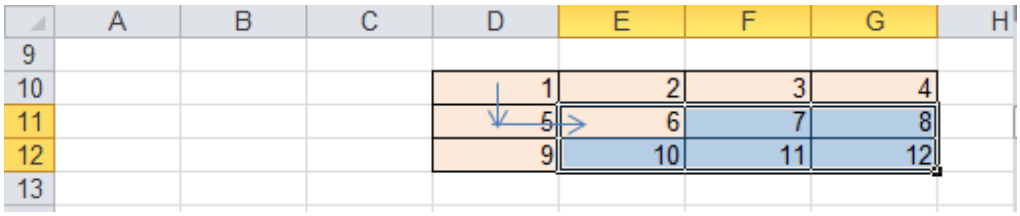
**Exemplo 3 de DESLOC()**

DESLOC(F12;;;-2;-3) retornará o intervalo de 2 linhas por três colunas, D11:F12. Note que a célula de referência F12 está incluída neste intervalo.




**Exemplo 4 de DESLOC()**

DESLOC(D10;1;1;2;3) retornará o intervalo de 2 linhas por três colunas, E11:G12, i.é, o Excel primeiro calcula DESLOC(D10;1;1) que é E11 (1 linha abaixo e 1 coluna à direita da célula de referência D10), depois então aplica a fórmula DESLOC(E11;;;2;3).




**Problemas e erros comuns com a função DESLOC()**

Quando reconstituir os passos das funções DESLOC(), somente a célula de referência é retornada. Por exemplo, quando seguir o precedente de DESLOC(D10;1;1;2;3) a célula retornada é D10 e não E11:G12.

O Excel exclui a célula de referência quando calcula os componentes "linha" e "colunas", mas inclui a célula de referência quando calcula os componentes "altura" e "largura".

Isto pode ser confuso, e requer extremo cuidado. DESLOC() é um conceito complexo para entender que reduz a confiança do usuário no modelo pois ele não é entendido facilmente.

**Combinando DESLOC() com outras Funções**

Como DESLOC() retorna uma célula ou um intervalo de células, ele pode ser facilmente combinado com outras funções tais como SOMA(), SOMARPRODUTO(), MIN(), MAX(), etc.

Por exemplo, SOMA(DESLOC()) calcula a soma da célula ou intervalo de células retornado pela função DESLOC(). Estendendo do Exemplo 4 acima, SOMA(DESLOC(D10;1;1;2;3)) é equivalente a escrever SOMA(E11:G12) (pois DESLOC(D10;1;1;2;3) retorna o intervalo E11:G12) que é igual a 54 (6+7+8+10+11+12). Similarmente, MÉDIA(DESLOC(D10;1;1;2;3)) é equivalente à MÉDIA(E11:G12).

Como explicado no Método 1, a modelagem ARIMA envolve 4 passos principais:

## A) IDENTIFICAÇÃO DO MODELO

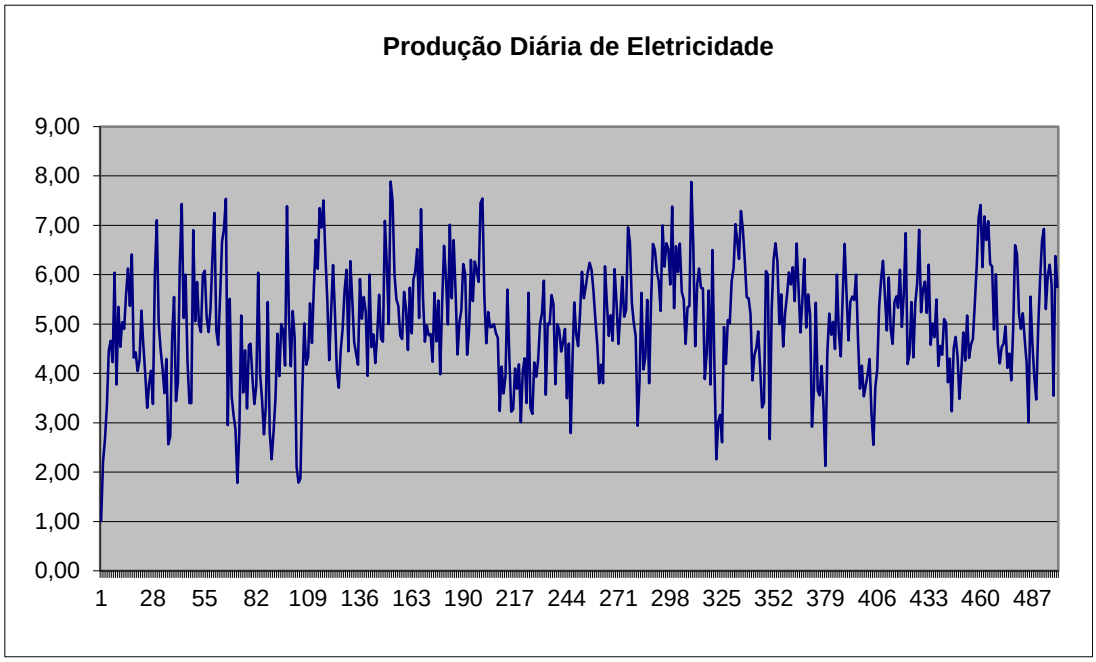
Como mencionado anteriormente, há uma clara necessidade de métodos objetivos, automáticos, de identificação do melhor modelo ARIMA para os dados em mãos. Métodos objetivos tornam-se particularmente cruciais quando especialistas treinados na construção de modelos não estiverem disponíveis. Além disso, mesmo para especialistas, métodos objetivos fornecem uma ferramenta adicional muito útil, pois o correlograma e correlograma parcial nem sempre apontam claramente para o único melhor modelo. Os dois critérios mais largamente usados são o **critério de informação de *Akaike*** (AIC), e o **critério Bayesiano (Schwarz)** (BIC ou SIC):

$$AIC(p,q) = \ln(\hat{\sigma}^2) + \frac{2(p+q)}{T}$$

$$BIC(p,q) = \ln(\hat{\sigma}^2) + \frac{\ln(T)(p+q)}{T}$$

$$\hat{e}^2 = estimativa\ de\ \sigma^2\ de\ ARMA(p,q)$$

Não se preocupe com as fórmulas acima. Elas são facilmente implementadas numa planilha Excel. Deixe-me mostrar-lhe como construir esta identificação de modelo automatizada com um exemplo de planilha. Abra a planilha (*Trabalho(2)*). Os dados são da produção diária de eletricidade num país em desenvolvimento: milhões de quilowatts por dia são entrados no intervalo de células A2:A501.




Do gráfico podemos ver que os dados são estacionários. Entretanto, podemos confirmar isto com o teste de média zero. Na célula I28 confirme que os dados têm uma média zero.

A equação geral para o ARMA(p,q) é:

**y(t) = d + a(1)*y(t-1) + a(2)*y(t-2)+ ... + a(p)*y(t-p) + e(t) – c(1)*e(t-1) – c(2)*e(t-2) - ... –c(p)*e(t-p)**

Os parâmetros para **p** e **q** são entrados nas células L1 e M1, respectivamente.

Os coeficientes para **p** são entrados em L2:L11 e os coeficientes para **q** são entrados em M2:M11.

Note que coloquei o máximo ARMA(10,10). O modelo pode ter qualquer **p** e qualquer **q**. Para usar o **AIC**, **BIC** para identificar um modelo **ARMA(p,q)** automaticamente precisamos definir os limites superiores, **p** e **q** para a ordem **AR** e **MA**, respectivamente. No nosso caso os limites superiores, **p** e **q** é 10. Os valores nas células L2:M11 são os coeficientes correspondentes.

Por exemplo, se o modelo é um ARMA(3,2), a célula em L1 mostrará um 3 e M1 será um 2. Os coeficientes correspondentes são as células L9, L10, L11 para o **AR** e células M10, M11 para o **MA**. (ver os números em azul na Figura 2.24 abaixo).

| | K | L | M | N |
|---|---|---|---|---|
| 1 | **p** | **3** | **2** | **q** |
| 2 | a(10) | 0,1 | 0,1 | c(10) |
| 3 | a(9) | 0,1 | 0,2 | c(9) |
| 4 | a(8) | 0,1 | 0,3 | c(8) |
| 5 | a(7) | 0,1 | 0,1 | c(7) |
| 6 | a(6) | 0,1 | 0,1 | c(6) |
| 7 | a(5) | 0,1 | 0,1 | c(5) |
| 8 | a(4) | 0,1 | 0,1 | c(4) |
| 9 | a(3) | 0,1 | 0,1 | c(3) |
| 10 | a(2) | 0,1 | 0,1 | c(2) |
| 11 | a(1) | 0,1 | 0,1 | c(1) |
| 12 | | 3 | 2 | |

Figura 2.24

Os coeficientes são como segue:

**a**(1) = L11, **a**(2) = L10, **a**(3) = L9

**c**(1) = M11, **c**(2) = M10

Ou se for um ARMA(2,1), então a célula em L1 mostrará um 2 e M1 será um 1. Os coeficientes correspondentes são as células L10:L11 para o **AR** e célula M11 para o **MA**. (ver os números em azul na Figura 2.25 abaixo):

| | K | L | M | N |
|---|---|---|---|---|
| 1 | **p** | **2** | **1** | **q** |
| 2 | a(10) | 0,1 | 0,1 | c(10) |
| 3 | a(9) | 0,1 | 0,2 | c(9) |
| 4 | a(8) | 0,1 | 0,3 | c(8) |
| 5 | a(7) | 0,1 | 0,1 | c(7) |
| 6 | a(6) | 0,1 | 0,1 | c(6) |
| 7 | a(5) | 0,1 | 0,1 | c(5) |
| 8 | a(4) | 0,1 | 0,1 | c(4) |
| 9 | a(3) | 0,1 | 0,1 | c(3) |
| 10 | a(2) | 0,1 | 0,1 | c(2) |
| 11 | a(1) | 0,1 | 0,1 | c(1) |
| 12 | | 2 | 1 | |

Figura 2.25

A função INT do Excel foi usada para remover todas as casas decimais deixando somente o número inteiro. Remover casas decimais, ou a parte fracionária de um número é necessário para usar o Solver do Excel para nossa modelagem.

A célula L12 está relacionada à L1 quando L1 for um número inteiro ou inteiro de L12. Precisamos entrar com a função INT() na célula L1 pois o Solver do Excel retornará um erro se L1 não for um inteiro. O mesmo vale para M1 e M12. Eles estão relacionados pela mesma razão.

Como prometido anteriormente, incluirei o cálculo de **d** neste exemplo. A fórmula para **d** é entrar na célula I5. (Você pode se referir em como **d** é derivado olhando a página 35 acima).

Para o entendimento mais fácil, a equação geral acima é desdobrada em 3 partes:

i. Fórmula **d** é entrada na célula I5

ii. **a(1)*y(t-1) + a(2)*y(t-2)+ ... + a(p)*y(t-p)** entrado na coluna B, e

iii. **e(t) – c(1)*e(t-1) – c(2)*e(t-2) - ... –c(p)*e(t-p)** entrado na coluna C

**i)** Ler a página xxxxx acima para entender como **d** é calculado

A fórmula para **d** em I5 é:

=I2*(1-(SOMA(DESLOC(L12;-1;0):DESLOC(L12;-L1;0))))

Para um ARMA(2,1), isto significa I2*(1-(SOMA(L10:L11)).

Para um ARMA(3,2), então é I2*(1-(SOMA(L9:L11)).

**ii)** A fórmula é como esta para a segunda linha na célula B3:

=SE($L$1 <= 1;SOMARPRODUTO(DESLOC($L$12;-1;0):DESLOC($L$12;-$L$1;0);DESLOC(A3;-$L$1;0):DESLOC(A3;-1;0));0)

| | A | B |
|---|---|---|
| 1 | $y_t$ | a(p)*y(t) |
| 2 | 1,01159 | 0,00000 |
| 3 | 2,20618 | =SE($L$1 <= 1;SOMARPRODUTO(DESLOC($L$12;-1;0):DESLOC($L$12;-$L$1;0);DESLOC(A3;-$L$1;0):DESLOC(A3;-1;0));0) |
| 4 | 2,66032 | =SE($L$1 <= 2;SOMARPRODUTO(DESLOC($L$12;-1;0):DESLOC($L$12;-$L$1;0);DESLOC(A4;-$L$1;0):DESLOC(A4;-1;0));0) |
| 5 | 3,34366 | =SE($L$1 <= 3;SOMARPRODUTO(DESLOC($L$12;-1;0):DESLOC($L$12;-$L$1;0);DESLOC(A5;-$L$1;0):DESLOC(A5;-1;0));0) |
| 6 | 4,46771 | =SE($L$1 <= 4;SOMARPRODUTO(DESLOC($L$12;-1;0):DESLOC($L$12;-$L$1;0);DESLOC(A6;-$L$1;0):DESLOC(A6;-1;0));0) |
| 7 | 4,66031 | =SE($L$1 <= 5;SOMARPRODUTO(DESLOC($L$12;-1;0):DESLOC($L$12;-$L$1;0);DESLOC(A7;-$L$1;0):DESLOC(A7;-1;0));0) |
| 8 | 4,22850 | =SE($L$1 <= 6;SOMARPRODUTO(DESLOC($L$12;-1;0):DESLOC($L$12;-$L$1;0);DESLOC(A8;-$L$1;0):DESLOC(A8;-1;0));0) |
| 9 | 6,04132 | =SE($L$1 <= 7;SOMARPRODUTO(DESLOC($L$12;-1;0):DESLOC($L$12;-$L$1;0);DESLOC(A9;-$L$1;0):DESLOC(A9;-1;0));0) |
| 10 | 3,77283 | =SE($L$1 <= 8;SOMARPRODUTO(DESLOC($L$12;-1;0):DESLOC($L$12;-$L$1;0);DESLOC(A10;-$L$1;0):DESLOC(A10;-1;0));0) |
| 11 | 5,34599 | =SE($L$1 <= 9;SOMARPRODUTO(DESLOC($L$12;-1;0):DESLOC($L$12;-$L$1;0);DESLOC(A11;-$L$1;0):DESLOC(A11;-1;0));0) |
| 12 | | |

Figura 2.26

Como você pode ver, a fórmula inicia com uma função SE. A razão é que esta célula será calculada se o **p** na célula L1 for maior que ou igual a 1. De outra forma o valor será zero.

Uma declaração SE é usada no Excel para fazer certas ações somente se alguma coisa for verdadeira. Por exemplo, você poderia querer imprimir a mensagem "*Estamos perdendo dinheiro*" se as vendas totais para este mês ficarem abaixo de certa quantia. Por outro lado, você apenas irá querer imprimir "*Estamos fazendo dinheiro!*" Assim, a célula B3 significa

=SE(**p** <= 1;então calcule; caso contrário o valor da célula = 0) em termos não profissionais.

Por exemplo $L$1 = 3 (i.é, **p** = 3) então os primeiros 3 dados da série não são calculados, i.é., os valores em y(t-1) na célula A4, y(t-2) na célula A3 e y(t-3) na célula A2 são usados para calcular y(t) na célula B5 (ver Figura 2.27 abaixo).

| | A | B | C | D | E | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|
| 1 | $y_t$ | a(p)*y(t) | c(q)*e(t) | e(t) | y(t) | p | 3 | 2 | q |
| 2 | 1,01 | 0 | 0 | 0,00 | 0,00 | a(10) | 0,1 | 0,1 | c(10) |
| 3 | 2,21 | 0 | 0 | -1,30 | 3,50 | a(9) | 0,1 | 0,1 | c(9) |
| 4 | 2,66 | 0 | -0,129630379 | -0,97 | 3,63 | a(8) | 0,1 | 0,1 | c(8) |
| 5 | 3,34 | 0,5878097 | -0,22681047 | -0,97 | 4,32 | a(7) | 0,1 | 0,1 | c(7) |
| 6 | 4,47 | 0,821016308 | -0,194524832 | -0,05 | 4,52 | a(6) | 0,1 | 0,1 | c(6) |
| 7 | 4,66 | 1,047168393 | -0,102377148 | 0,01 | 4,65 | a(5) | 0,1 | 0,1 | c(5) |
| 8 | 4,23 | 1,247168091 | -0,004204317 | -0,53 | 4,75 | a(4) | 0,1 | 0,1 | c(4) |
| 9 | 6,04 | 1,335651792 | -0,051708173 | 1,15 | 4,89 | a(3) | 0,1 | 0,1 | c(3) |
| 10 | 3,77 | 1,49301367 | 0,062611324 | -1,16 | 4,93 | a(2) | 0,1 | 0,1 | c(2) |
| 11 | 5,35 | 1,404265406 | -0,000858266 | 0,44 | 4,91 | a(1) | 0,1 | 0,1 | c(1) |
| 12 | 4,54 | 1,516014477 | -0,072168172 | -0,55 | 5,09 | | 3 | 2 | |

Figura 2.27

Usando **p** = 3 isto

SOMARPRODUTO(DESLOC($L$12;-1;0):DESLOC($L$12;-$L$1;0);DESLOC(A3;-$L$1;0):DESLOC(A3;-1;0)), será:

**a(1)*y(t-1) + a(2)*y(t-2)+ a(3)*y(t-3) ⇒** L11*A4 + L10*A3 + L9*A2

Se **p** = 2, então ela será:

**a(1)*y(t-1) + a(2)*y(t-2) ⇒** L11*A3 + L10*A2

etc...

Agora podemos ver como usei as funções **SOMARPRODUTO()** e **DESLOC()** para configurar a fórmula geral **AR(p)**.

**iii.** Agora explicarei a parte **MA(q)**

A fórmula acima implica que para calcular **e**(t-1), por exemplo, precisamos conhecer **e**(t), o qual não conhecemos. A convenção é atribuir zeros a todos os valores desconhecidos de **e**(t). Assim, entramos com zero na célula C2, que é a primeira célula necessária para realizar este cálculo.

A fórmula é como esta para a segunda linha na célula C3 (ver Figura 2.28 abaixo)

=SE($M$1 <= 1;SOMARPRODUTO(DESLOC($M$12;-1;0):DESLOC($M$12;-$M$1;0);DESLOC(D3;-$M$1;0):DESLOC(D3;-1;0));0)

| | A | C |
|---|---|---|
| 1 | $y_t$ | c(q)*e(t) |
| 2 | 1,01159 | 0 |
| 3 | 2,20618 | =SE($M$1 <= 1;SOMARPRODUTO(DESLOC($M$12;-1;0):DESLOC($M$12;-$M$1;0);DESLOC(D3;-$M$1;0):DESLOC(D3;-1;0));0) |
| 4 | 2,66032 | =SE($M$1 <= 2;SOMARPRODUTO(DESLOC($M$12;-1;0):DESLOC($M$12;-$M$1;0);DESLOC(D4;-$M$1;0):DESLOC(D4;-1;0));0) |
| 5 | 3,34366 | =SE($M$1 <= 3;SOMARPRODUTO(DESLOC($M$12;-1;0):DESLOC($M$12;-$M$1;0);DESLOC(D5;-$M$1;0):DESLOC(D5;-1;0));0) |
| 6 | 4,46771 | =SE($M$1 <= 4;SOMARPRODUTO(DESLOC($M$12;-1;0):DESLOC($M$12;-$M$1;0);DESLOC(D6;-$M$1;0):DESLOC(D6;-1;0));0) |
| 7 | 4,66031 | =SE($M$1 <= 5;SOMARPRODUTO(DESLOC($M$12;-1;0):DESLOC($M$12;-$M$1;0);DESLOC(D7;-$M$1;0):DESLOC(D7;-1;0));0) |
| 8 | 4,22850 | =SE($M$1 <= 6;SOMARPRODUTO(DESLOC($M$12;-1;0):DESLOC($M$12;-$M$1;0);DESLOC(D8;-$M$1;0):DESLOC(D8;-1;0));0) |
| 9 | 6,04132 | =SE($M$1 <= 7;SOMARPRODUTO(DESLOC($M$12;-1;0):DESLOC($M$12;-$M$1;0);DESLOC(D9;-$M$1;0):DESLOC(D9;-1;0));0) |
| 10 | 3,77283 | =SE($M$1 <= 8;SOMARPRODUTO(DESLOC($M$12;-1;0):DESLOC($M$12;-$M$1;0);DESLOC(D10;-$M$1;0):DESLOC(D10;-1;0));0) |
| 11 | 5,34599 | =SE($M$1 <= 9;SOMARPRODUTO(DESLOC($M$12;-1;0):DESLOC($M$12;-$M$1;0);DESLOC(D11;-$M$1;0):DESLOC(D11;-1;0));0) |

Figura 2.28

Novamente uma função **SE** está em uso. Temos **q** = 2 na célula M1, assim as células C2 e C3 = 0.

Usando **q** = 2, a fórmula (cálculo parte de C4):

SOMARPRODUTO(DESLOC($M$12;-1;0):DESLOC($M$12;-$M$1;0);DESLOC(D3;-$M$1;0):DESLOC(D3;-1;0)), será:

**c(1)*e(t-1) – c(2)*e(t-2) ⇒** M11*D3 – M10*D2

Se **q** = 3, então a fórmula é (o cálculo parte de C5):

**c(1)*e(t-1) – c(2)*e(t-2) - c(3)*e(t-3) ⇒ M11*D4 – M10*D3 – M9*D2**

etc...

Como para os resíduos ou erros e(t), estes são entrados na coluna D. (ver Figura 2.29 abaixo)

| | D | E |
|---|---|---|
| 1 | e(t) | y(t) |
| 2 | 0,00 | 0,00 |
| 3 | =A3-($I$5+(B3-C3)) | =$I$5 +B3-C3 |
| 4 | =A4-($I$5+(B4-C4)) | =$I$5 +B4-C4 |
| 5 | =A5-($I$5+(B5-C5)) | =$I$5 +B5-C5 |
| 6 | =A6-($I$5+(B6-C6)) | =$I$5 +B6-C6 |
| 7 | =A7-($I$5+(B7-C7)) | =$I$5 +B7-C7 |
| 8 | =A8-($I$5+(B8-C8)) | =$I$5 +B8-C8 |
| 9 | =A9-($I$5+(B9-C9)) | =$I$5 +B9-C9 |
| 10 | =A10-($I$5+(B10-C10)) | =$I$5 +B10-C10 |

Figura 2.29

E por último a fórmula completa é entrada na coluna E:

- D = 15
- Coluna B = **a(1)*y(t-1) + a(2)*y(t-2)+ ... + a(p)*y(t-p)**
- Coluna C = **e(t) – c(1)*e(t-1) – c(2)*e(t-2) - ... –c(p)*e(t-p)**
- Coluna D = **e(t)**

Então a fórmula completa na coluna E (ver Figura 2.29 acima)

Antes de invocar o Solver do Excel para resolver os parâmetros p, q e seus coeficientes, deixe-me explicar a você o **AIC** e **BIC** como usamos num método objetivo para identificação do modelo na próxima seção.

## B) ESTIMAÇÂO DO MODELO

Devido à natureza altamente subjetiva da metodologia de Box-Jenkins, os analistas de séries temporais têm perseguido métodos objetivos alternativos para identificaçãode modelos ARMA. Funçõesestatísticas de penalização, tais como Akaike Information Criterion (AIC) ou Critério Final Predictor Error (FPE) (Akaike, 1974), Critério Schwarz (SC) ou Bayesian Information Criterion (BIC) (Schwarz, 1978) foram usados para auxiliar os analistas de séries temporais na reconciliação da necessidade de minimizar erros com o desejo conflitante para economia do modelo. Estas estatísticas todas tomam a forma de minimização da soma da soma dos quadrados residuais mais um termo "penalidade" que incorpora o número de coeficientes paramétricos estimados pelo fator na economia do modelo..

**Critério de Informação Akaike** (AIC):

$$AIC = log\left(\frac{rss}{n}\right) + \left(2 * \frac{k}{n}\right)$$

**Critério de Informação Bayesiano** (BIC)

$$BIC = log\left(\frac{rss}{n}\right) + \left(\log(n) * \frac{k}{n}\right)$$

onde,

*k* = número de coeficientes estimados (1 + p + q + P + Q)

*rss* = soma dos quadrados residuais

*n* = Número de observações

Assumindo que haja um modelo ARMA verdadeiro para a série temporal, o BIC e HQC têm as melhores propriedades teóricas. O BIC é fortemente consistente enquanto o AIC usualmente resultará num modelo sobre parametrizado; isto é fácil de verificar que para *n* maior que sete o BIC impõem uma penalidade maior para parâmetros adicionais do que faz o AIC.

Assim, na prática, usar o critério de seleção objetivo de modelo envolve estimar um intervalo de modelos e aquele um com o critério de informação mais baixo é selecionado. Estas duas fórmulas são entradas nas células I11 para o AIC e I12 para o BIC (ver Figura 2.30 abaixo):

| | H | I |
|---|---|---|
| 1 | | |
| 2 | **Média** | =MÉDIA(A2:A501) |
| 3 | **Desv. Pad.** | =DESVPAD.N(A2:A501) |
| 4 | **Medida** | =I2+I3 |
| 5 | **d** | =I2*(1-(SOMA(DESLOC(L1;-1;0;):DESLOC(L1;L1;0)))) |
| 6 | $\bar{e}$ | =MÉDIA(D3:D501) |
| 7 | $SE_e$ | =DESVPAD.N(D3:D501)/RAIZ(CONT.NÚM(D3:D501)) |
| 8 | **Valor** | =1,96*I7 |
| 9 | **Verdito:** | **=SE(I6>I8;"Média não zero";"Média zero")** |
| 10 | | |
| 11 | **AIC** | =LN(DESVPAD.N(D3:D501)/CONT.NÚM(D3:D501))+(2*2/CONT.NÚM(D3:D501)) |
| 12 | **BIC** | =LN(DESVPAD.N(D3:D501)/CONT.NÚM(D3:D501))+(LN(CONT.NÚM(D3:D501)*2/CONT.NÚM(D3:D501))) |
| 13 | **SSE** | =SOMAQUAD(D2:D501) |

Figura 2.30

Precisamos também definir a região admissível que garantirá que nosso modelo seja estacionário e invertível. Os coeficientes do modelo **AR** devem estar dentro de uma região permitida para garantir a estacionariedade e há também uma região permitida para os coeficientes do modelo **MA** que garanta a invertibilidade. Cada modelo **MA** é estacionário por definição, mas é invertível somente se certas condições forem satisfeitas. A propósito, modelos **AR** são invertíveis para todos os valores dos coeficientes, mas somente estacionários se os coeficientes estiverem numa região admissível particular. Na Figura 2.30 acima a região admissível que garante estacionariedade é dada na célula I20 e a região admissível garantindo a invertibilidade é dada na célula I21. Quando tivermos um modelo generalizado para ARIMA automático, as fórmulas para garantir a estacionariedade e a invertibilidade são:

$$-1 \leq \sum p \leq 1$$

e

$$-1 \leq \sum q \leq 1$$

***Agora é a vez de usar o Solver do Excel para o nosso modelo ARIMA(p,q) Automatizado***

Abra a planilha (*Trabalho(3)*). A planilha (*Trabalho(3)*) é apenas uma cópia da planilha (*Trabalho(2)*). Para usar o Solver, clique no botão em Arquivo > Opções, para aparecer a seguinte janela:

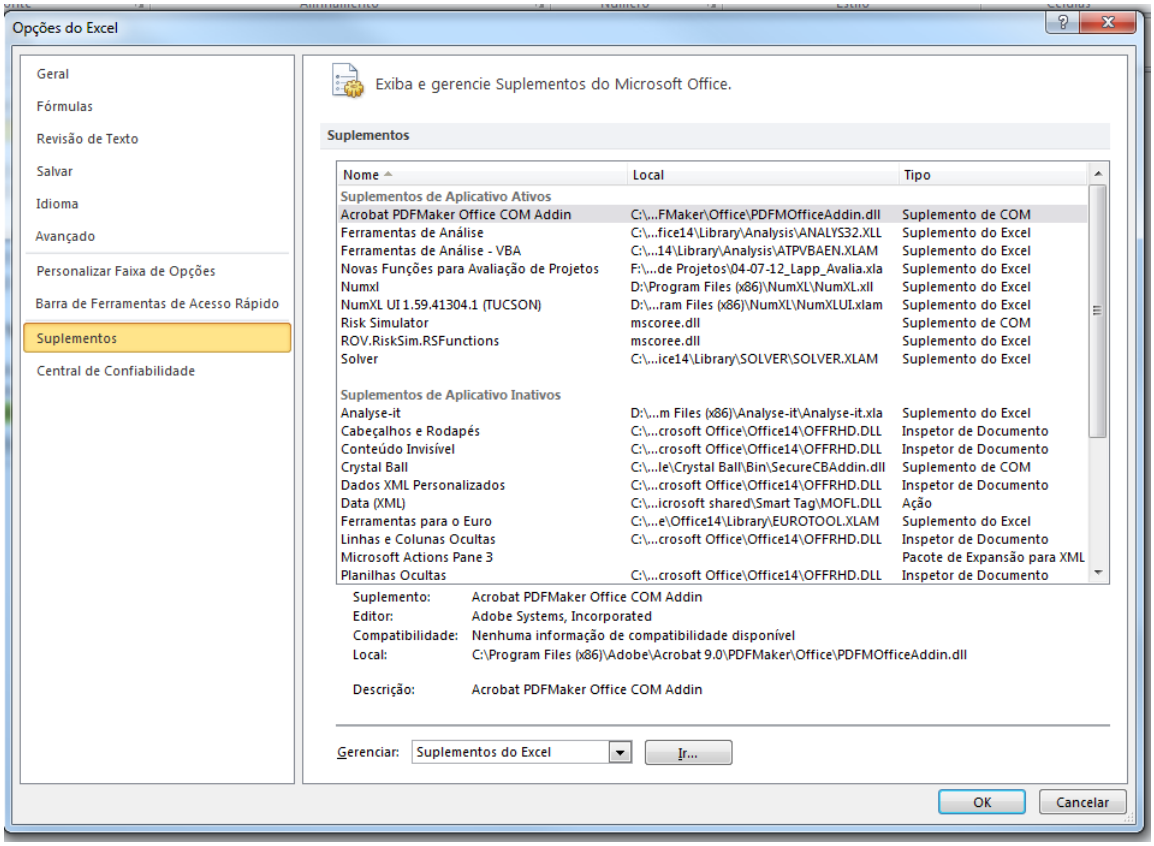



Figura 2.31

Verifique se na caixa Gerenciar aparece *Suplementos do Excel*, depois então pressione o botão Ir... Selecione a caixa SOLVER se esta estiver desmarcada nos Suplementos disponíveis.

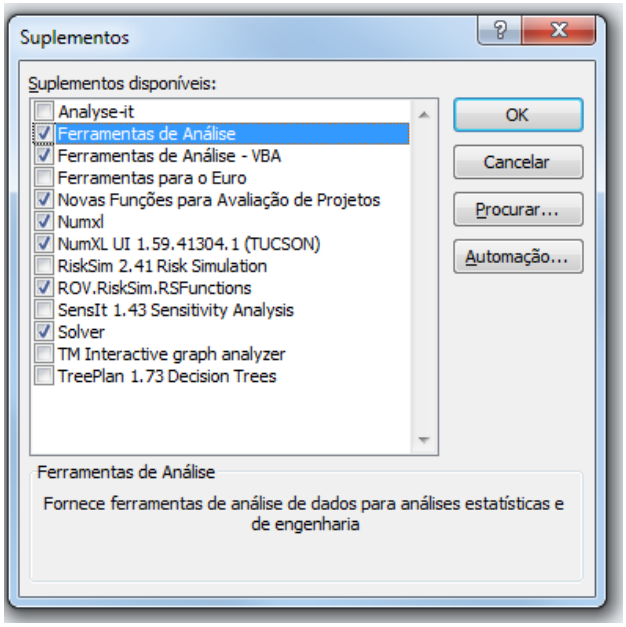



Figura 2.32

Depois de selecionar o suplemento Solver e clicar no botão OK, o Excel leva um momento para chamá-lo e o adiciona no grupo de ferramentas de Análise da guia Dados.

Depois de executar o Solver, você será apresentado aos parâmetros do Solver na caixa de diálogo abaixo:

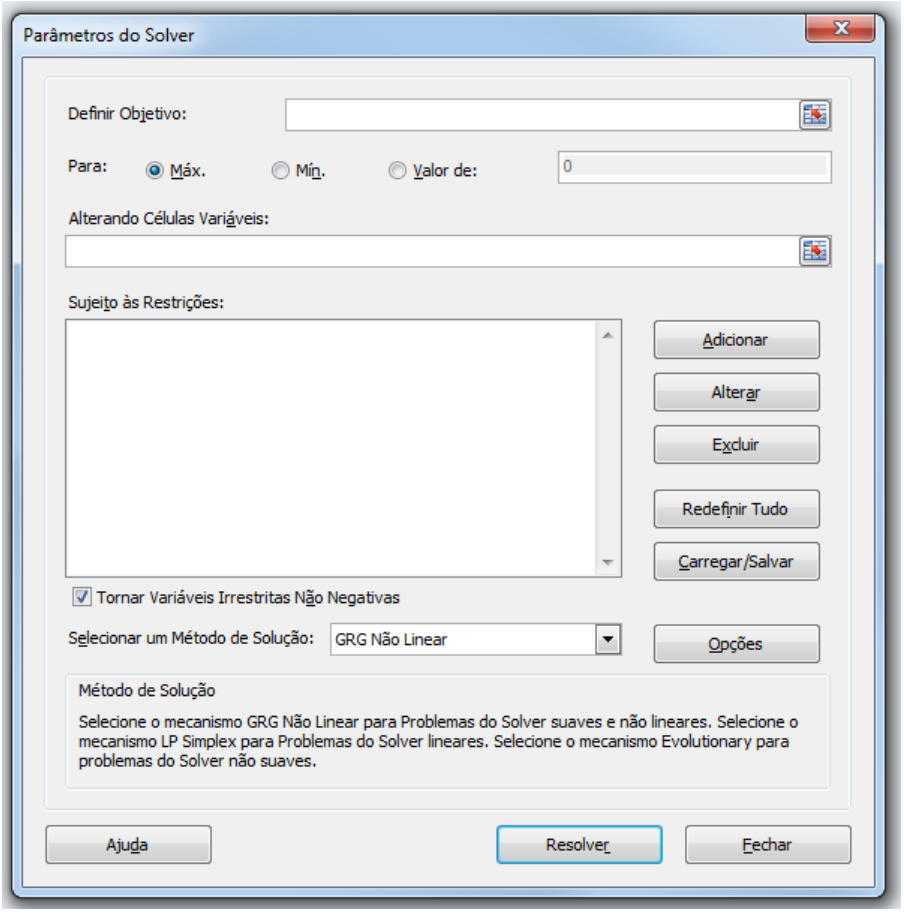



Figura 2.33

Vamos revisar cada parte desta caixa de diálogo, uma de cada vez.

**Definir Objetivo:** é onde você indica a função objetivo (ou meta) a ser otimizada. Esta célula deve conter uma fórmula que depende de uma ou de outras células (incluindo aquela última "célula variando"). Você pode ou digitar no endereço de células ou clicar na célula desejada. Aqui entramos com a célula **I11**.

No nosso modelo ARIMA, a função objetivo é minimizar o **AIC** na célula I11. Ver Figura 2.34 abaixo

**Para:** lhe dará a opção de tratamento da Célula Alvo em três modos alternativos. **Max** (o *default*) diz ao Excel para maximizar a Célula Alvo e **Min**, minimizá-la, enquanto **Valor de**: é usada se você quiser atingir certo valor particular na Célula Alvo escolhendo um valor particular da variável endógena.

Aqui, selecionamos **Min** pois queremos minimizar o **AIC**. (Você pode também tentar I12 o **BIC**).

Para valor inicial, usei **p** e **q** = 5. Os coeficientes =0,1 (ver Fig. 2.34 abaixo).

| | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | | p | 5 | 5 | q |
| 2 | | Média | 5,00 | | a(10) | 0,1 | 0,1 | c(10) |
| 3 | | Desv. Pad. | 1,17 | | a(9) | 0,1 | 0,1 | c(9) |
| 4 | | Medida | 6,17 | | a(8) | 0,1 | 0,1 | c(8) |
| 5 | | d | 2,501777294 | | a(7) | 0,1 | 0,1 | c(7) |
| 6 | | $\bar{e}$ | 0,04369 | | a(6) | 0,1 | 0,1 | c(6) |
| 7 | | $SE_e$ | 0,051352445 | | a(5) | 0,1 | 0,1 | c(5) |
| 8 | | Valor | 0,100650793 | | a(4) | 0,1 | 0,1 | c(4) |
| 9 | | Veredito: | Zero mean | | a(3) | 0,1 | 0,1 | c(3) |
| 10 | | | | | a(2) | 0,1 | 0,1 | c(2) |
| 11 | | AIC | -6,067329742 | | a(1) | 0,1 | 0,1 | c(1) |
| 12 | | BIC | -5,382198594 | | | 5 | 5 | |
| 13 | | SSE | 656,27066726 | | | | | |
| 14 | | | | | | | | |
| 15 | | Teste Durbin - Watson | | | | | | |
| 16 | | 635,1584 | | | | | | |
| 17 | | 656,2707 | | | | | | |
| 18 | | 0,96783 | | | | | | |
| 19 | | | | | | | | |
| 20 | | Regiões permissíveis para p | 0,5 | | | | | |
| 21 | | Regiões permissíveis para q | 0,5 | | | | | |
| 22 | | | | | | | | |
| 23 | | | | | | | | |
| 24 | | | | | | | | |
| 25 | | | | | | | | |
| 26 | | Média | 5,00 | | | | | |
| 27 | | 1.96*SE | 0,00763 | | | | | |
| 28 | | Teste Média Zero | Zero | | | | | |

Figura 2.34

**Alterando Células Variáveis:** permite-lhe indicar quais células são as células ajustáveis (i.é., variáveis endógenas). Como na caixa Definir Objetivo:, você deve digitar um endereço de célula ou clicar numa célula da planilha. O Excel manipula problemas de otimização multi-variável permitindo-lhe incluir células adicionais na caixa Alterando Células Variáveis. Cada variável escolhida não contígua é separada por ponto e vírgula. Se você usar a técnica do mouse (clicando nas células), a separação de ponto e vírgula é automática.

Aqui, as células que precisam ser mudadas são aquelas dos parâmetros **p** e **q** e seus coeficientes. No modelo, os parâmetros **p** e **q** e seus coeficientes estão contidos no Intervalo L2:M12 e L2:M11 respectivamente. Então entramos com, L12:M12;L2:M11. Ver Figura 2.35 abaixo:

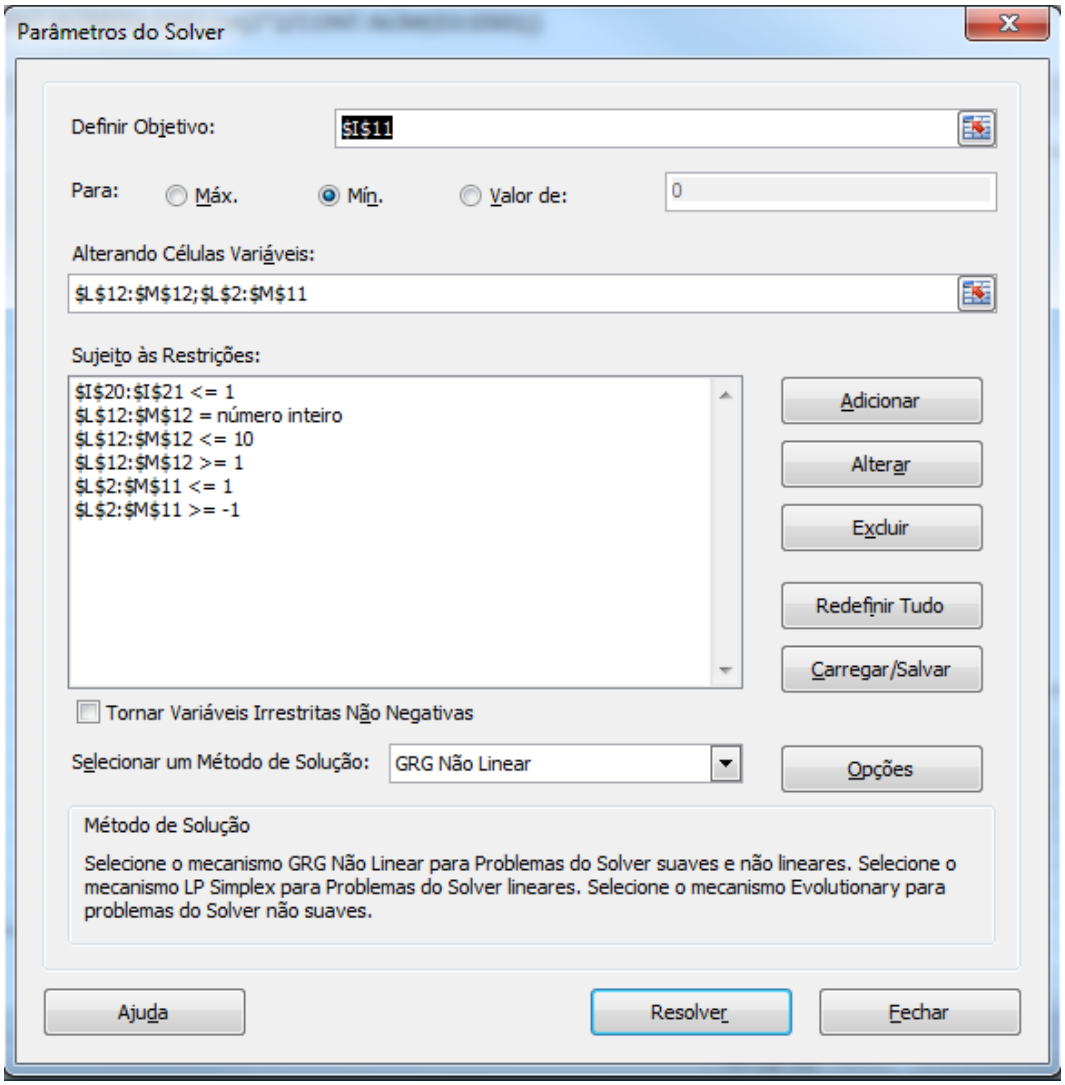



Figura 2.35

**Sujeito às Restrições:** é usado para impor vínculos nas variáveis endógenas. Recorreremos a esta importante parte do Solver quando fizermos os problemas de Otimização de Vínculos. Teremos uns poucos vínculos que precisam ser entrados como mostrado na Figura 2.35 acima.

Clique no botão Adicionar para adicionar estas restrições.

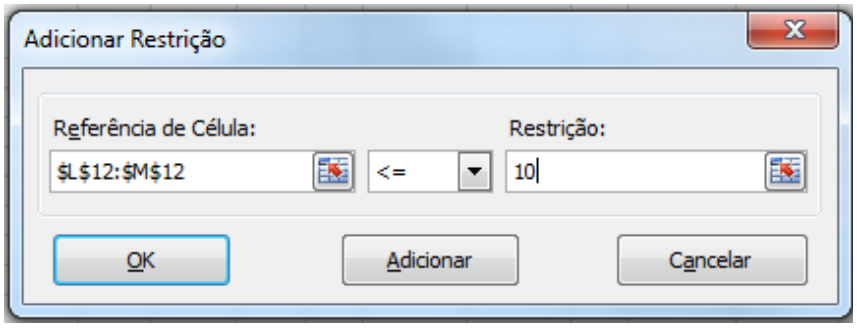



Figura 2.36

Estas restrições são:

a. I20:I21 ≤ 1 : As Regiões Permissíveis

b. I20:I21 ≥ -1

c. L12:M12 ≤ 10 : o **p** e o **q**

d. L12:M12 ≥ 1

e. L12:M12 = número inteiro

f. L2:M11 ≤ 1 : os coeficientes

g. L2:M11 ≥ -1

**Após isto selecione as Opções**. Isto lhe permitirá ajustar as maneiras nas quais o Solver abordará a solução (ver Figura 2.37)

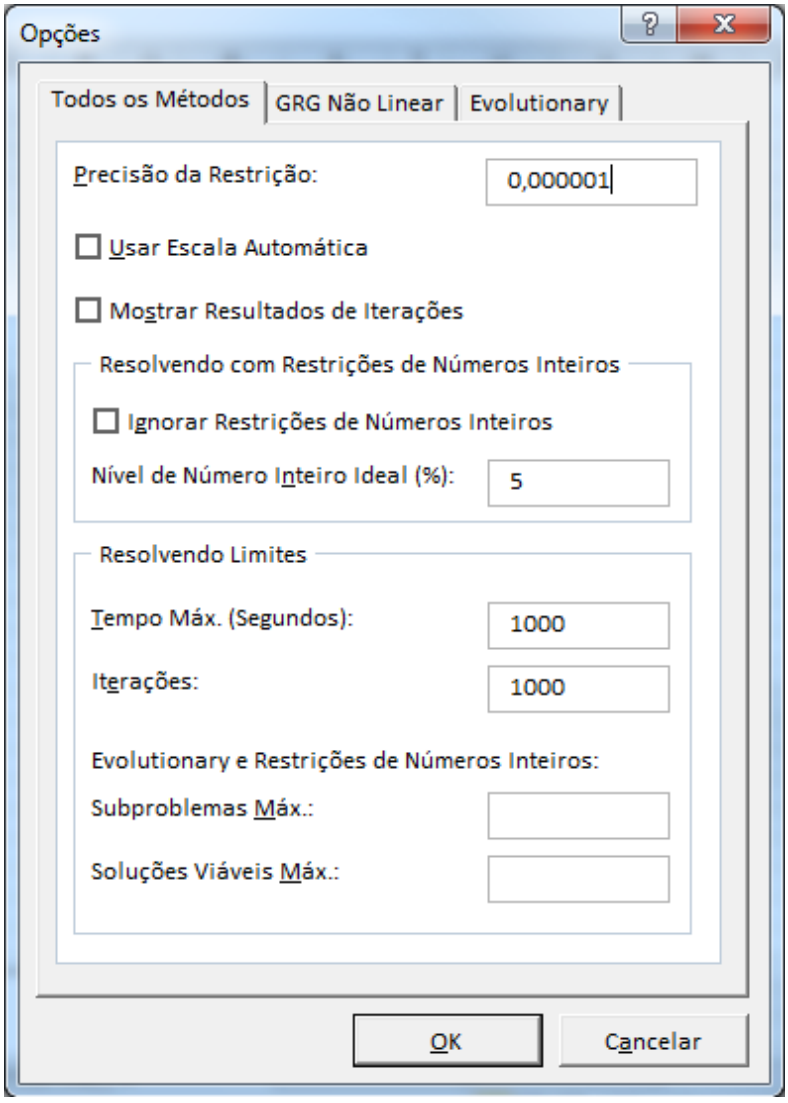



Figura 2.37

Como você pode ver, uma série de escolhas está incluída na caixa de diálogo Opções do Solver que direcionam a busca do Solver pela solução ótima e a duração da busca. Estas opções podem ser mudadas se o Solver estiver tendo dificuldade de encontrar a solução ótima. Abaixando a Precisão, o Nível de Número Inteiro Ideal (%), etc., diminui a velocidade do algoritmo mas deve capacitar o Solver a encontrar uma solução.

Para um modelo ARIMA, você pode definir:

i. Tempo Máx. (Segundos): 1000

ii. Iterações: 1000

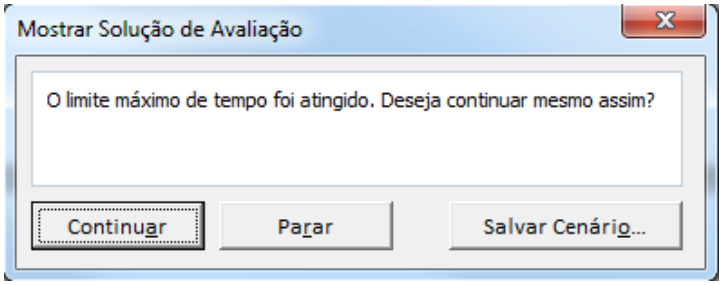
iii. Precisão da Restrição: 0,000001

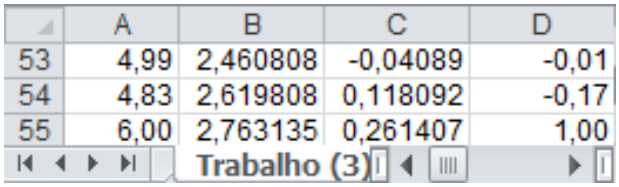
iv. Nível de Número Inteiro Ideal (%): 5%

Selecione **Todos os Métodos** como o método de procura. Isto prova ser muito efetivo na minimização do AIC.

Clicando OK retorne à caixa de diálogo Parâmetros do Solver.

**Resolver:** é obviamente o botão que você tem de clicar para fazer o Solver, do Excel, encontrar uma solução. Esta é a última coisa que você deve fazer na caixa de diálogo de Parâmetros do Solver. Então, clique Resolver para iniciar o treinamento.



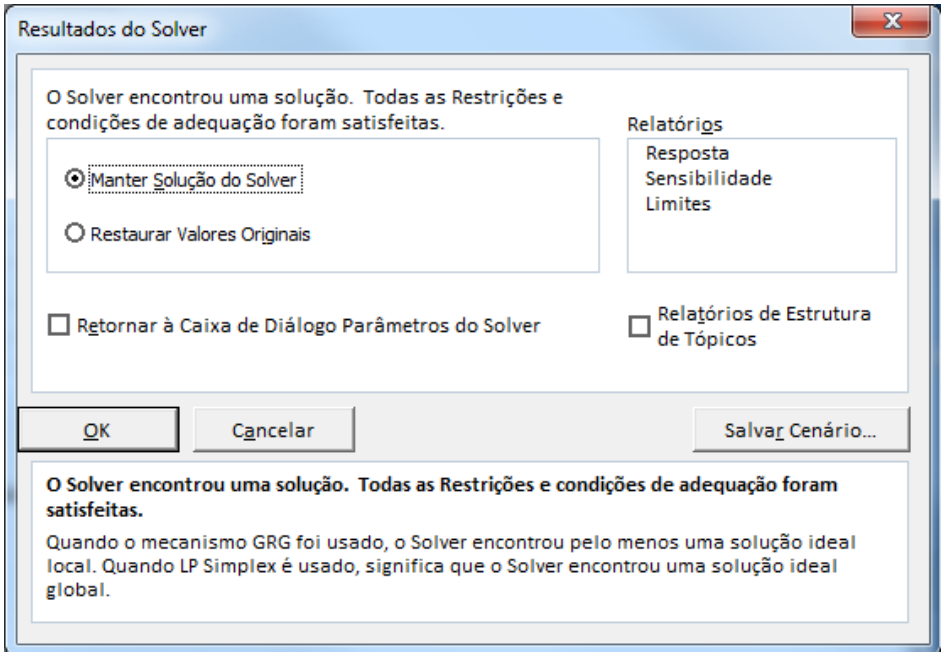



| | A | B | C | D |
|---|---|---|---|---|
| 53 | 4,99 | 2,460808 | -0,04089 | -0,01 |
| 54 | 4,83 | 2,619808 | 0,118092 | -0,17 |
| 55 | 6,00 | 2,763135 | 0,261407 | 1,00 |

Trabalho (3)

Figura 2.39

Quando o Solver iniciar a otimização, você verá a Solução Tentativa no canto esquerdo inferior de sua planilha. Ver Figura 2.39 acima.




Uma mensagem aparecerá depois que o Solver tiver convergido (ver Figura 2.40). Nesse caso, o Excel relata que: "*Solver has converged to the current solution. All constraints are satisfied*". Esta é uma boa notícia.

Algumas vezes, a solução não é satisfatória e o Solver se torna incapaz de encontrar a solução de uma só vez. Por exemplo deve ter falhado o teste estacionário como indicado na célula I9, i.é, nenhuma média zero. Se este for o caso então você, deve mudar os parâmetros iniciais de **p** e **q** e os coeficientes e executar o Solver novamente. Siga os passos discutidos acima. Da minha experiência, geralmente você irá precisar executar o Solver umas poucas vezes antes de chegar à solução satisfatória.

A má notícia é uma mensagem como, "*Solver could not find a solution*." Se isto acontecer, você deve diagnosticar, debugar, e por outro lado pensar sobre o que esteve errado e como poderia ser fixado. As duas fixações mais rápidas são tentar diferentes parâmetros iniciais **p** e **q** e seus coeficientes.

Na caixa de diálogo Resultados do Solver, você elegeu se o Excel escreverá a solução que ele encontrou nas Células Objetivo (i.é, *Manter a Solução do Solver*) ou se deixará a planilha somente e NÃO escreverá o valor da solução nas Células Objetivo (i.é, *Restore Original Values*). Quando o Excel relatar uma execução bem sucedida, você deverá geralmente querer Manter a Solução do Solver. No lado direito da caixa de diálogo Resultados do Solver, o Excel apresenta uma série de relatórios. Os relatórios Resposta, Sensibilidade e Limites são planilhas adicionais inseridas na pasta corrente. Elas contêm os diagnósticos e outras informações e deverá ser selecionada se o Solver estiver problemas ao encontrar uma solução.

| | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|
| 1 | | | | p | 1 | 1 | q |
| 2 | Média | 5,00 | | a(10) | 0,10000 | 0,10000 | c(10) |
| 3 | Desv. Pad. | 1,17 | | a(9) | 0,10000 | 0,10000 | c(9) |
| 4 | Medida | 6,17 | | a(8) | 0,10000 | 0,10000 | c(8) |
| 5 | d | 1,302352095 | | a(7) | 0,10000 | 0,10000 | c(7) |
| 6 | $\bar{e}$ | 0,01318 | | a(6) | 0,10000 | 0,10000 | c(6) |
| 7 | $SE_e$ | 0,04337707 | | a(5) | 0,10000 | 0,10000 | c(5) |
| 8 | Valor | 0,085019058 | | a(4) | 0,10000 | 0,10000 | c(4) |
| 9 | Veredito: | Média zero | | a(3) | 0,10000 | 0,10000 | c(3) |
| 10 | | | | a(2) | 0,10000 | 0,10000 | c(2) |
| 11 | AIC | -6,236111323 | | a(1) | 0,73971 | 0,32466 | c(1) |
| 12 | BIC | -5,550980175 | | | 1 | 1 | |
| 13 | SSE | 467,66068620 | | | | | |
| 14 | | | | | | | |
| 15 | Teste Durbin - Watson | | | | | | |
| 16 | 920,1784 | | | | | | |
| 17 | 467,6607 | | | | | | |
| 18 | 1,96762 | | | | | | |
| 19 | | | | | | | |
| 20 | Regiões permissíveis para p | 0,100000000000 | | | | | |
| 21 | Regiões permissíveis para q | 0,100000000000 | | | | | |
| 22 | | | | | | | |
| 23 | | | | | | | |
| 24 | | | | | | | |
| 25 | | | | | | | |
| 26 | Média | 5,00 | | | | | |
| 27 | 1.96*SE | 0,0076314255549013 | | | | | |
| 28 | Teste Média Zero | Zero | | | | | |
| 29 | | | | | | | |

Figura 2.41

Minha primeira execução do Solver do Excel veio a ter a solução acima. AIC = -6,236111282 na célula I11. Como indicado na Figura 2.41 acima, temos um modelo ARMA(1,1). Os coeficientes estão nas células L11 e M11. Ele passou todos os testes como você pode ver nas células I9 e H18 na Figura 2.41. (Note: Dependendo dos dados que você tiver, algumas vezes você irá precisar executar o Solver umas poucas vezes antes de você chegar a uma solução satisfatória).

## D) DIAGNÓSTICO DE VERIFICAÇÃO

Como sabemos que produzimos um modelo razoável e que nosso modelo realmente reflete a série temporal real? Esta é uma parte do processo que Box e Jenkins se referem como diagnóstico de verificação. Usarei dois métodos para conduzir o diagnóstico. Como esperamos que os erros de previsão sejam completamente aleatórios, o primeiro passo é plotá-los, como fora feito na Figura 2.42 abaixo, por exemplo. Este diagrama dos resíduos indica aleatoriedade. Mas queremos garantir isto e precisamos fazer os cálculos

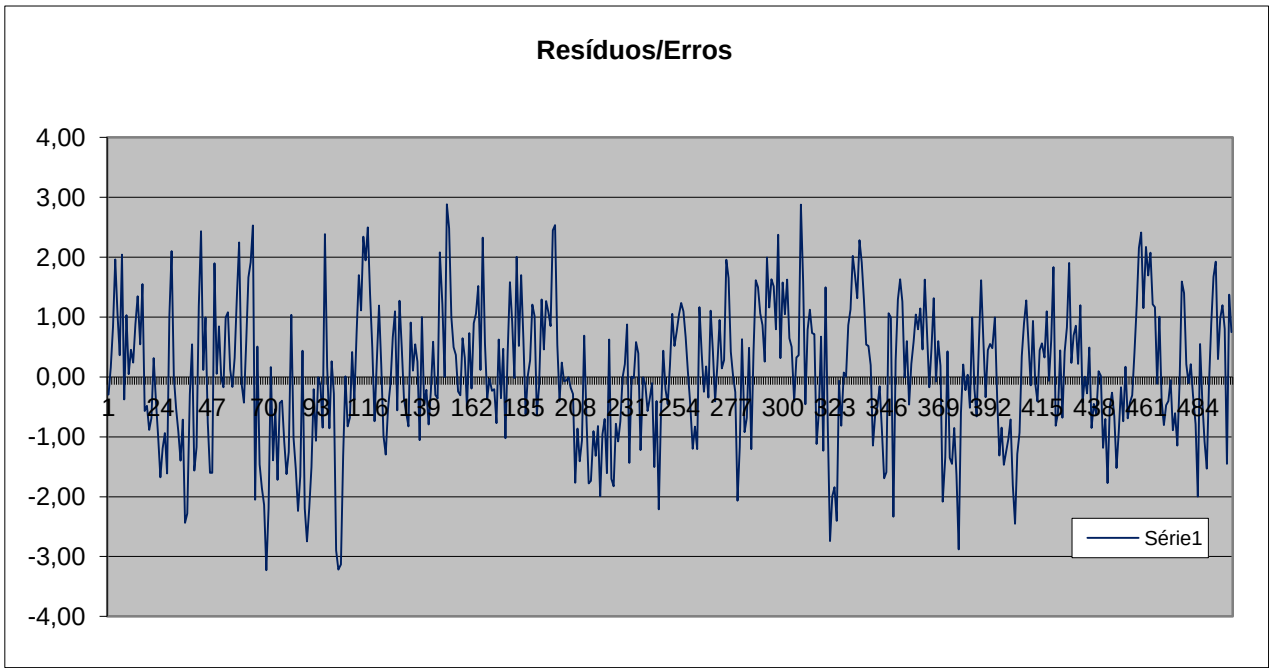



Figura 2.42

Uma das exigências é que a média residual deverá ser zero, ou próxima de zero. Para estabelecer que este é o caso, precisamos estimar o erro padrão do erro médio. Isto é calculado como:

$$\sigma_e = \sqrt{\frac{\sum_{t=1}^{n}(e_t - \bar{e})^2}{n}}$$

$$SE_{\bar{e}} = \frac{\sigma_e}{\sqrt{n}}$$

na célula I7, onde $\sigma_e$ é o desvio padrão residual, $\bar{e}$ é o erro médio, $n$ o número de erros e $SE_{\bar{e}}$ é o erro padrão do erro médio. Se a média residual $\bar{e}$ for maior que 1,96 erros padrões ($E_{\bar{e}}$), então podemos dizer que ela é significativamente não zero:

$$\bar{e} > 1{,}96\, E_{\bar{e}}$$

na célula I9.

Como estimar o erro residual padrão $SE_e$ (erro padrão) está mostrado abaixo na Figura 2.43 e as fórmulas estão dadas na Figura 2.44 abaixo.

| | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|
| 1 | | | | p | 5 | 5 | q |
| 2 | Média | 5,00 | | a(10) | 0,1 | 0,1 | c(10) |
| 3 | Desv. Pad. | 1,17 | | a(9) | 0,1 | 0,1 | c(9) |
| 4 | Medida | 6,17 | | a(8) | 0,1 | 0,1 | c(8) |
| 5 | d | 2,501777294 | | a(7) | 0,1 | 0,1 | c(7) |
| 6 | $\bar{e}$ | 0,04369 | | a(6) | 0,1 | 0,1 | c(6) |
| 7 | $SE_e$ | 0,051352445 | | a(5) | 0,1 | 0,1 | c(5) |
| 8 | Valor | 0,100650793 | | a(4) | 0,1 | 0,1 | c(4) |
| 9 | Veredito: | Zero mean | | a(3) | 0,1 | 0,1 | c(3) |
| 10 | | | | a(2) | 0,1 | 0,1 | c(2) |
| 11 | AIC | -6,067329742 | | a(1) | 0,1 | 0,1 | c(1) |
| 12 | BIC | -5,382198594 | | | 5 | 5 | |
| 13 | SSE | 656,27066726 | | | | | |
| 14 | | | | | | | |
| 15 | Teste Durbin - Watson | | | | | | |
| 16 | 635,1584 | | | | | | |
| 17 | 656,2707 | | | | | | |
| 18 | 0,96783 | | | | | | |
| 19 | | | | | | | |
| 20 | Regiões permissíveis para p | 0,5 | | | | | |
| 21 | Regiões permissíveis para q | 0,5 | | | | | |
| 22 | | | | | | | |
| 23 | | | | | | | |
| 24 | | | | | | | |
| 25 | | | | | | | |
| 26 | Média | 5,00 | | | | | |
| 27 | 1.96*SE | 0,00763 | | | | | |
| 28 | Teste Média Zero | Zero | | | | | |
| 29 | | | | | | | |

Figura 2.43

| | H | I |
|---|---|---|
| 4 | Medida | =I2+I3 |
| 5 | d | =I2*(1-(SOMA(DESLOC(L1;-1;0;):DESLOC(L1;L1;0)))) |
| 6 | $\bar{e}$ | =MÉDIA(D3:D501) |
| 7 | $SE_e$ | =DESVPAD.N(D3:D501)/RAIZ(CONT.NÚM(D3:D501)) |
| 8 | Valor | =1,96*I7 |
| 9 | Verdito: | =SE(I6>I8;"Média não zero";"Média zero") |
| 10 | | |
| 11 | AIC | =LN(DESVPAD.N(D3:D501)/CONT.NÚM(D3:D501))+(2*2/CONT.NÚM(D3:D501)) |
| 12 | BIC | =LN(DESVPAD.N(D3:D501)/CONT.NÚM(D3:D501))+(LN(CONT.NÚM(D3:D501)*2/CONT.NÚM(D3:D501))) |
| 13 | SSE | =SOMAQUAD(D2:D501) |
| 14 | | |
| 15 | Teste Durbin - Watson | |
| 16 | =SOMAXMY2(D3:D501;D2:D500) | |
| 17 | =SOMAQUAD(D2:D501) | |
| 18 | =H16/H17 | |
| 19 | | |
| 20 | Regiões permissíveis para p | =SOMA(DESLOC($L$1;1;0):DESLOC($L$1;$L$1;0)) |
| 21 | Regiões permissíveis para q | =SOMA(DESLOC($M$1;1;0):DESLOC($M$1;$M$1;0)) |
| 22 | | |
| 23 | | |
| 24 | | |
| 25 | | |
| 26 | Média | =MÉDIA(A2:A501) |
| 27 | 1.96*SE | =1,96*(RAIZ(A2:A501)/CONT.NÚM(A2:A501)) |
| 28 | Teste Média Zero | =SE(I26<I27;"Não-zero";"Zero") |
| 29 | | |

A célula I9 contém uma breve declaração SE avaliando se a média calculada em I6 é maior que o erro padrão vezes 1,96. No nosso modelo tivemos média zero a qual passou no teste.

Outro teste que é bem popular é o teste de Durbin-Watson, que é usado no contexto de verificação de validade do modelo ARIMA. A **estatística Durbin-Watson** é um teste estatístico usado para detectar a presença de autocorrelação nos resíduos de uma análise de regressão. Foi assim chamado após James Durbin e Geofrey Watson.

Se $e_t$ é o resíduo associado com a observação no tempo t, então a estatística do teste é

$$w = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

A célula H16 contém a parte superior da fórmula acima e H17 contém a parte inferior. Como w na célula H18 é aproximadamente igual a 2(1 - r), onde ***r*** é a autocorrelação amostral dos resíduos, $w = 2$ indica nenhuma correlação. O valor de $w$ sempre fica entre 0 e 4. Se a estatística de Durbin-Watson for essencialmente menor que 2, há evidência de correlação serial positiva. Como um princípio básico grosseiro, se Durbin-Watson for menor que 1,0, haverá motivo de alarme. Pequenos valores de $w$ indicam que os sucessivos termos de erros são, em média, próximos em valor um do outro, ou positivamente correlacionado. Se $w > 2$, os sucessivos termos de erro são, em média, muito diferentes em valor uns dos outros, i.é., negativamente correlacionados. Em regressões, isto pode implicar uma sub-estimação do nível de significância estatística.

No nosso modelo temos 1,96812 na célula H18 o qual está muito próximo de 2 e que indica nenhuma correlação. Ver Figura 2.43 acima. Podemos agora seguir com a previsão.

## E) PREVISÃO

Agora estamos prontos para produzir previsões reais, i.é, aquela que entram no futuro. A equação pode ser aplicada "um passo à frente" para obter a estimativa de y(t) do y(t-1) observado. Uma previsão "*k* passos à frente" pode também ser feita pela aplicação recorrente da equação. Na aplicação recorrente, o y observado no tempo 1 é usado para gerar o y estimado no tempo 2. Esta estimativa é então substituída em y(t-1) para obter o y estimado no tempo 3, e assim por diante. As previsões *k* passos adiante eventualmente convergem a zero quando o horizonte de previsão, *k*, cresce. Na planilha *Trabalho(3)* vá à célula A502:A507.

Projetaremos conforme a fórmula seguinte: ARIMA(1,0,1) ou ARMA(1,1)

Usamos a fórmula:

$$y(t) = 1{,}30335 + 0{,}73951*y(t-1) - 0{,}32419*e(t-1)$$

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 497 | 6,20 | 4,425613 | 0,167304 | 0,64 | 5,56 | |
| 498 | 5,80 | 4,584962 | 0,206935 | 0,12 | 5,68 | |
| 499 | 3,55 | 4,292052 | 0,039717 | -2,01 | 5,56 | |
| 500 | 6,38 | 2,625261 | -0,65023 | 1,80 | 4,58 | |
| 501 | 5,75 | 4,717826 | 0,5838 | 0,31 | 5,44 | |
| 502 | 5,38 | | | | 5,454217182 | |
| 503 | 5,44 | | | | 5,336824083 | |
| 504 | 5,28 | | | | 5,250010712 | |
| 505 | 5,00 | | | | 5,185811357 | |
| 506 | 4,82 | | | | 5,138335291 | |
| 507 | 4,72 | | | | 5,103226266 | |

Figura 2.45

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 500 | 6,3797 | =SE($L$1 | =SE($M$1 | =A500-($I$ | =$I$5 +B500-C500 |
| 501 | 5,7500 | =SE($L$1 | =SE($M$1 | =A501-($I$ | =$I$5 +B501-C501 |
| 502 | 5,3760 | | | | =$I$5+($L$11*A501)-$M$11*D501 |
| 503 | 5,4356 | | | | =$I$5+($L$11*E502) |
| 504 | 5,2790 | | | | =$I$5+($L$11*E503) |
| 505 | 5,0000 | | | | =$I$5+($L$11*E504) |
| 506 | 4,8195 | | | | =$I$5+($L$11*E505) |
| 507 | 4,7217 | | | | =$I$5+($L$11*E506) |

Figura 2.46

A Figura 2.45 mostra a planilha contendo os valores de previsão e a Figura 2.46 mostra todos os cálculos e fórmulas.

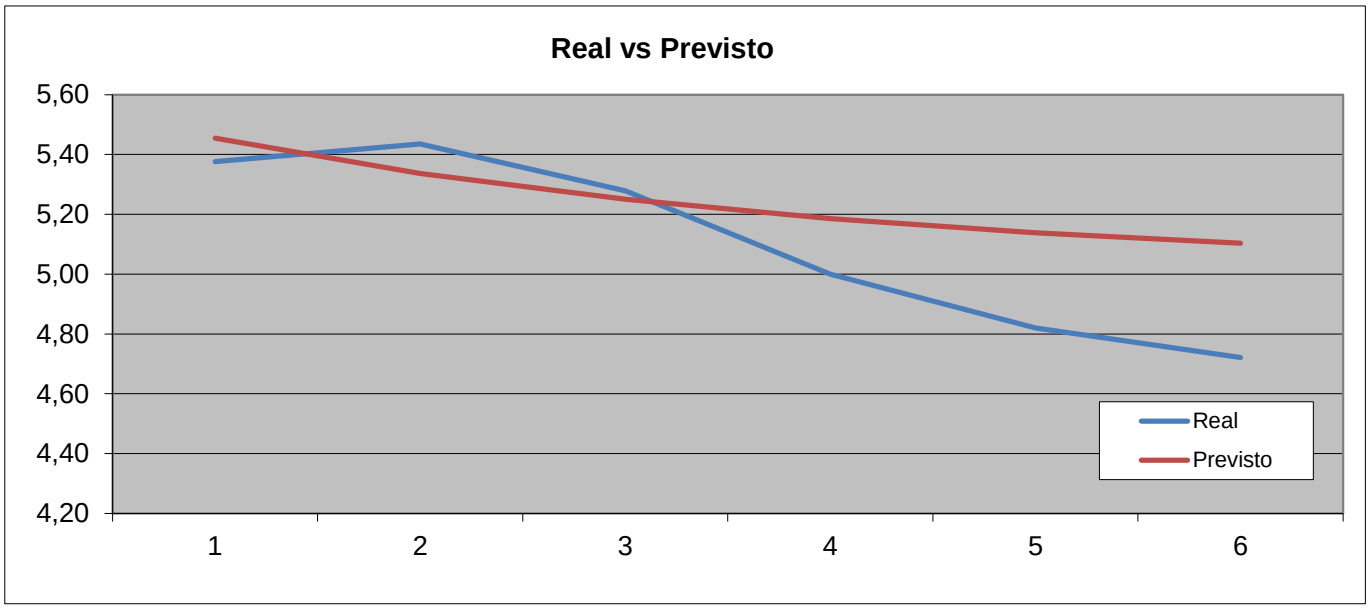



Figura 2.47

Como já explicamos, uma vez tendo executado os valores reais, os valores reais de y(t) trocados pelos seus valores ajustados (começando de E503). Isto inevitavelmente degrada a previsão, e explicamos como diferentes modelos se comportam. Como podemos ver, nossa previsão para a célula E502 e E505 na Figura 2.45 é muito boa (como conhecemos os valores reais, os colocamos nas células A502:A507 para comparar).Infelizmente nossa previsão para a célula E506 começa a ser significativamente diferente dos valores reais conhecidos na célula A506. Isto implica que para muitas séries o método ARIMA é um bom ajuste, mas somente para previsões de curto prazo.

Você pode não ter um modelo tão ideal. Ele levou-me a cerca de 10 execuções do Solver Excel sobre o modelo antes de apresentar estes resultados. Mudando o **p**, **q** e seus valores iniciais dos coeficientes e depois então executando o Solver até você chegar a uma solução satisfatória. Ele leva um pouco mais de teste e execuções.

Outra maneira que você pode fazer uso do modelo é usar o Solver do Excel para otimizar os coeficientes somente. Você entra com o **p** e o **q** manualmente e usa o Solver para otimizar o coeficiente. Deixe-me dar-lhe um exemplo. Abra a planilha *Trabalho(4)*. Entre com 2 em ambas as células L12 e M12. (ver Figura 2.48 abaixo). Então estamos usando um modelo ARMA(2,2). Invoque o Solver do Excel e entre com os parâmetros como mostrados na Figura 2.49 abaixo:

| | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | p | 2 | 2 | q | |
| 2 | Média | 5,00 | | a(10) | 0,1 | 0,1 | c(10) | |
| 3 | Desv. Pad. | 1,17 | | a(9) | 0,1 | 0,2 | c(9) | |
| 4 | Medida | 6,17 | | a(8) | 0,1 | 0,3 | c(8) | |
| 5 | d | 4,002843671 | | a(7) | 0,1 | 0,1 | c(7) | |
| 6 | $\bar{e}$ | 0,01151 | | a(6) | 0,1 | 0,1 | c(6) | |
| 7 | $SE_e$ | 0,051401222 | | a(5) | 0,1 | 0,1 | c(5) | |
| 8 | Valor | 0,100746395 | | a(4) | 0,1 | 0,1 | c(4) | |
| 9 | Veredito: | Média zero | | a(3) | 0,1 | 0,1 | c(3) | |
| 10 | | | | a(2) | 0,1 | 0,1 | c(2) | |
| 11 | AIC | -6,066380354 | | a(1) | 0,1 | 0,1 | c(1) | |
| 12 | BIC | -5,381249205 | | | 2 | 2 | | |
| 13 | SSE | 656,62962414 | | | | | | |
| 14 | | | | | | | | |
| 15 | Teste Durbin - Watson | | | | | | | |
| 16 | 636,4766 | | | | | | | |
| 17 | 656,6296 | | | | | | | |
| 18 | 0,96931 | | | | | | | |
| 19 | | | | | | | | |
| 20 | Regiões permissíveis para p | 0,2 | | | | | | |
| 21 | Regiões permissíveis para q | 0,3 | | | | | | |
| 22 | | | | | | | | |
| 23 | | | | | | | | |
| 24 | | | | | | | | |
| 25 | | | | | | | | |
| 26 | Média | 5,00 | | | | | | |
| 27 | 1.96*SE | 0,00763 | | | | | | |
| 28 | Teste Média Zero | Zero | | | | | | |
| 29 | | | | | | | | |
| 30 | | | | | | | | |
| 31 | | | | | | | | |

Figura 2.48

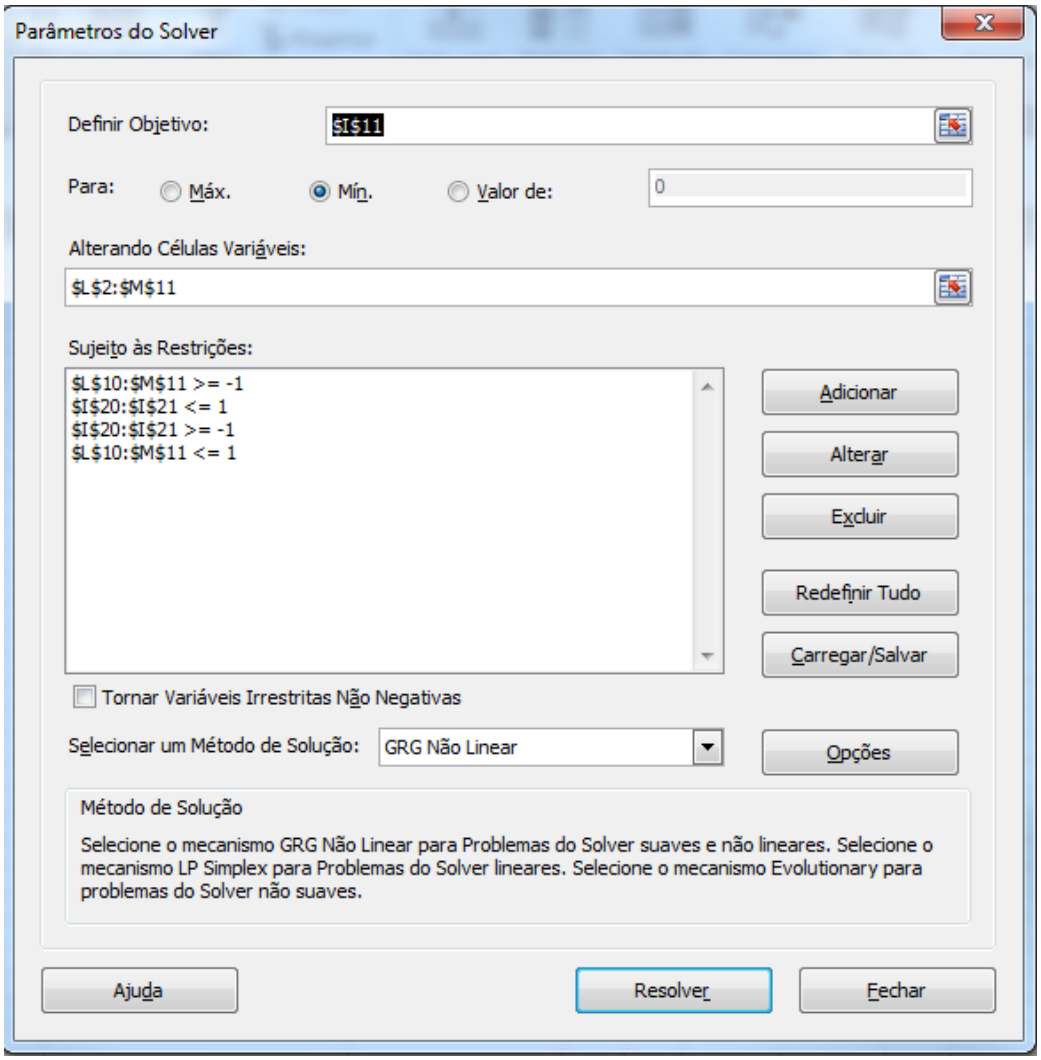



Figura 2.49

Assim as Células variáveis agora são somente L10:M11. A L12:M12 que são o **p** e o **q** não está lá pois temos fixadas estas células com 2 como num ARMA(2,2). (Você pode também experimentar diferentes valores de **p** e **q**).

Após isto apenas otimize os coeficientes na célula L10:M11 com o Solver até você obter uma solução satisfatória. Faça o teste e a previsão como a que foi mostrada acima. Chamei isto de uma modelagem ARIMA semi-automatizada. O resultado que tivemos, foi: **a**(1) = 0,160689, **a**(2) = 0,455254, **c**(1) = -0,27953, **c**(2) = 0,27179. Ver o resultado na Figura 2.50 abaixo.

A fórmula é:

y(t) = 1,9216479 + 0,16069*y(t-1) + 0,455255*y(t-2) – (-0,27953*e(t-1)) – 0,271794*e(t-2)

| | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|
| 1 | | | | p | 2 | 2 | q |
| 2 | Média | 5,00 | | a(10) | 0,1 | 0,1 | c(10) |
| 3 | Desv. Pad. | 1,17 | | a(9) | 0,1 | 0,2 | c(9) |
| 4 | Medida | 6,17 | | a(8) | 0,1 | 0,3 | c(8) |
| 5 | d | 1,921647982 | | a(7) | 0,1 | 0,1 | c(7) |
| 6 | $\bar{e}$ | 0,01395 | | a(6) | 0,1 | 0,1 | c(6) |
| 7 | $SE_e$ | 0,043229965 | | a(5) | 0,1 | 0,1 | c(5) |
| 8 | Valor | 0,084730731 | | a(4) | 0,1 | 0,1 | c(4) |
| 9 | Verdito: | Zero mean | | a(3) | 0,1 | 0,1 | c(3) |
| 10 | | | | a(2) | 0,45525 | 0,27179 | c(2) |
| 11 | AIC | -6,239508409 | | a(1) | 0,16069 | -0,27953 | c(1) |
| 12 | BIC | -5,554377261 | | | 2 | 2 | |
| 13 | SSE | 464,50509663 | | | | | |
| 14 | | | | | | | |
| 15 | Teste de Durbin - Watson | | | | | | |
| 16 | 932,6662 | | | | | | |
| 17 | 464,5051 | | | | | | |
| 18 | 2,00787 | | | | | | |
| 19 | | | | | | | |
| 20 | Regiões permissíveis para p | 0,2 | | | | | |
| 21 | Regiões permissíveis para q | 0,3 | | | | | |
| 22 | | | | | | | |
| 23 | | | | | | | |
| 24 | | | | | | | |
| 25 | | | | | | | |
| 26 | Média | 5,00 | | | | | |
| 27 | 1.96*SE | 0,00763 | | | | | |
| 28 | Teste Média Zero | Zero | | | | | |
| 29 | | | | | | | |

Figura 2.50

Para resumir, nesta seção nós não somente mostramos o processo todo de identificação de modelos automaticamente, ajustando-os e fazendo previsões, mas também apresentamos uma maneira muito mais rápida de fazer isto. Vinculamos os valores dos coeficientes ARMA diretamente com o AIC, que se tornou o valor alvo no Solver, e que em poucos passos simples produziu valores ótimos para estes **p**, **q** e seus coeficientes.

## Modelagem ARIMA Sazonal (SARIMA)

Podemos usar modelos ARIMA para previsão de séries temporais sazonais. Os princípios subjacentes são idênticos àqueles para séries temporais não sazonais, descritas acima. Estas séries temporais sazonais mostram tendências sazonais com periodicidade s.

As séries sazonais se repetem após certo número de meses, geralmente após doze meses, ou a cada quatro meses (sazonalidade quadrimestral). Elas podem ser estacionárias ou não. As séries temporais sazonais não estacionárias, precisam ser diferenciadas. Infelizmente, a diferenciação ordinária não é boa suficiente para tais casos. A diferenciação sazonal é que é necessária.

Por exemplo,

- Dados mensais têm 12 observações por ano
- Dados quadrimestrais têm 4 observações por ano
- Dados diários têm 5 ou 7 (ou algum outro número) de observações por semana.

Um processo SARIMA tem quatro componentes:

- Auto-regressivo (AR)
- Média móvel (MA)
- Diferenciação de um passo
- Diferenciação sazonal

Para uma série temporal com um padrão de 112 meses, a diferenciação sazonal é executada como segue:

A fórmula de diferenciação exige que numa série temporal sazonal, precisamos encontrar diferenças entre dois meses comparáveis, melhor do que entre dois meses sucessivos como faz mais sentido. Neste exemplo, 12 é o número de meses. Se atribuirmos a letra s para a sazonalidade, então a diferenciação sazonal é em geral descrita como:

$$w(t) = y(t) - y(t-s)$$

Como na diferenciação ordinária, algumas vezes um segundo nível de diferenciação é necessário. Isto é feito como:

$$\nabla w(t) = w(t) - w(t-s)$$

Se substituirmos $\nabla w(t) = w(t) - w(t-s)$, mas $w(t) = y(t) - y(t-s)$, obtemos:

$$w(t) = [y(t) - y(t-s)] - [y(t-1) - y(t-s-1)] = \mathbf{y(t) - y(t-1) - y(t-s) + y(t-s-1)}$$

Por exemplo, para s = 12, dá:

$$\nabla w(t) = y(1) - y(t-1) - y(t-12) + y(t-13)$$

A fórmula acima mostra que: y(t) = y(t-1) + y(t-12) – y(t-13), i.é, neste caso a observação corrente é igual à observação anterior, mais aquela de doze períodos atrás, menos aquela que o precede! Parece raro mas se a reescrevermos diferentemente ela fará um pouco de sentido:

$$y(t) - y(t-12) = y(t-1) - y(t-13)$$

Assim estamos dizendo que estas diferenças sazonais periódicas são as mesmas que as diferenças sazonais observadas no período anterior, que são mais lógicas.

Podemos fazer uma interessante digressão aqui e perguntar-nos como serão as diferenças sazonais do próximo período. É razoável assumir que elas serrão alguma coisa como: $y_{t+1} - y_{t-11} = y_t - y_{t-12,}$ o que é muito interessante porque podemos ver acima, que $y_t - y_{t-12} = y_{t-1} - y_{t-13}$. Essencialmente estamos dizendo que $y_{t+1} - y_{t-11} = y_{t-1} - y_{t-13}$. Isto significa que $y_{t+1} - y_{t-11} = y_t - y_{t-12}$? Sim, isto significa que a origem da previsão determinará todas as diferenças sazonais futuras.

Vamos retornar à modelagem de séries temporais sazonais. As explicações acima implicaram que numa ordem para ajustar uma série temporal com um modelo ARIMA, não é suficiente apenas ter um modelo de ordem (p,d,q). Precisamos também uma ordem sazonal (P,D,Q), que será combinado com estes coeficientes não sazonais (p,d,q). A fórmula geral é

**ARIMA(p,d,q)(P,D,Q)**

Como combinamos os dois? Podemos usar por exemplo, um **SARIMA(1,1,1)(1,1,1)$_4$**, i.é, um modelo com s = 4. Este modelo é descrito como:

$$(1 - \phi_1 B)(1 - \Phi_1 B^4)(1 - B)(1 - B^4)y_t = (1 - \theta_1 B)(1 - \Theta_1 B^4)\, e_t$$

Onde $\phi$ e $\theta$ são os coeficientes ARMA ordinários, $\Phi$ e $\Theta$ são os coeficientes ARMA sazonais e B é o operador de retardo. Se desatarmos a equação acima, obtemos:

$$y_t = (1 + \phi_1)y_{t-1} - \phi_1 y_{t-2} + (1 + \phi_1)y_{t-4} - (1 + \phi_1 + \Phi_1 + \phi_1\Phi_1)y_{t-5} + (\phi_1 + \phi_1\Phi_1)\, y_{t-6} - \Phi_1 y_{t-8} +$$
$$+ (\Phi_1 + \phi_1\Phi_1)\, y_{t-9} - \phi_1\Phi_1 y_{t-10} + e_t - \theta_1 e_{t-1} - \Theta_1 e_{t-4} + \theta_1\Theta_1 e_{t-5}$$

Como podemos ver, é uma equação bastante longa e confusa. Poderíamos usar notação abreviada no lugar para fazer um pouco mais de sentido. Assim, um modelo sazonal ARIMA(p,d,q)(P,D,Q) pode ser escrito numa forma geral curta como:

$$(1 - \phi_1 B)(1 - \Phi_1 B^S)\, y_t = (1 - \theta_1 B)(1 - \Theta_1 B^S)\, e_t$$

Que é muito mais elegante. Um modelo ARIMA $(2,1,0)(0,1,0)_{12}$, por exemplo, é portanto escrito como:

$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12})(1 - B)\, y_t = e_t$$

Onde, $(1- \phi_1 B - \phi_2 B^2)$ representa uma parte AR(2) não sazonal do modelo, $(1 - \Theta_1 B^{12})$ representa a parte AR(1) sazonal e $(1 – B)$ são as diferenças não sazonais.

Os parâmetros e os coeficientes sazonais são Φ, Φ, P, D, Q e $\phi$, $\theta$, p, d, q são para as séries temporais não sazonais. O s denota a sazonalidade.

Seguindo os três passos da metodologia Box e Jenkins (1976), a *identificação*, a *estimação* e o *diagnóstico de verificação* dos modelos SARIMA são ajustados aos dados da série temporal estacionária ou fracamente estacionária. A estimativa dos parâmetros AR(p, P) e MA(q, Q) para ajustamento de um modelo SARIMA é aproximadamente a mesma quando você modela um modelo ARIMA.

Por exemplo, um modelo **SARIMA(1,0,0)(0,1,1)$_{12}$**, numa forma abreviada é:

$$(1 - \phi_1 B)(1 - B^{12})\, y_t = (1 - \Theta_1 B^{12})\, e_t$$

Que conduz a

$$y_t = \phi_1 y_{t-1} + y_{t-12} - \phi_1 y_{t-12} + e_t - \Theta_1\, e_{t-12}$$

Por exemplo um modelo **SARIMA(0,1,1)(0,1,1)$_4$**, numa forma curta é

$$(1 - B)(1 - B^4)\, y_t = (1 - \theta_1 B)(1 - \Theta_1 B^4)\, e_t$$

Que conduz a

$$y_t = y_{t-1} + y_{t-4} - y_{t-5} + \theta_1\, e_{t-1} - \Theta_1 e_{t-4} + \theta_1\Theta_1\, e_{t-5}$$

O que deveríamos esperar das funções autocorrelação e autocorrelação parcial para estes modelos? De muitas maneiras elas são idênticas, em termos de inferência, que os modelos não sazonais. Um modelo $ARIMA(0,0,0)(1,0,0)_{12}$, por exemplo, terá uma utocorrelação parcial significativa na defasagem 12 e as autocorrelações decairão exponencialmente para todas as defasagens sazonais, i.é., 12, 24, 36, etc. Um modelo $ARIMA(0,0,0)(0,0,1)_{12}$, por outro lado, terá uma autocorrelação significativa na defasagem 12 e autocorrelações parciais decairão exponencialmente para todas as defasagens sazonais.

Os princípios de estimação de parâmetros para modelos sazonais são os mesmos que para os modelos não sazonais, embora as equações para o SARIMA possam ser mais confusas.

Por favor seja consciente de que é impraticável e desnecessário fazer diferenciação sazonal da série duas vezes. É uma boa prática não diferenciar a série temporal mais do que duas vezes, a despeito da espécie de diferenciação é usada, i.é., use um máximo de uma sazonal e uma ordinária ou, no máximo, faça diferenciação ordinária duas vezes. Um dos mais populares e frequentemente usado modelo sazonal na prática é um **ARIMA(0,1,1)(0,1,1)$_S$**. A maioria das séries temporais podem ser ajustadas com um **ARIMA(0,1,1)(0,1,1)$_S$**, então não exagere.

# CONCLUSÕES

O modelo ARIMA oferece um boa técnica para prever a magnitude de qualquer variável. Sua força está no fato de que o método é adequado para quaisquer séries temporais com qualquer padrão de variação e não requer que o planejador escolha a priori o valor de qualquer parâmetro.

Os modelos ARIMA fornecem também ferramentas úteis para as partes interessadas para serem usadas como ponto de referência ao desempenho de outros modelos de previsão como redes neurais, regressão de kernel e assim por diante. Entretanto, por favor tenha em mente que a imprecisão de previsão aumenta quanto mais longe a previsão estiver dos dados usados, o que é consistente com a expectativa dos modelos ARIMA. É preciso muita prática e experiência. Felizmente com todos os exemplos apresentados neste Capítulo podemos acelerar e encurtar a sua curva de aprendizagem