

7500  
7000  
6500  
6000  
5500  
5000  
4500  
4000  
3500

# Econometria Aplicada no EViews®

Igor Morais  
Filipe Stona  
Gustavo Schuck

# Econometria Aplicada no EViews<sup>®</sup>

Igor Morais, Filipe Stona e Gustavo Schuck



GOVERNO DO ESTADO  
RIO GRANDE DO SUL

**SECRETARIA DO PLANEJAMENTO, MOBILIDADE E DESENVOLVIMENTO REGIONAL**

**FUNDAÇÃO DE ECONOMIA E ESTATÍSTICA Siegfried Emanuel Heuser**

CONSELHO DE PLANEJAMENTO: André F. Nunes de Nunes, Angelino Gomes Soares Neto, André Luis Vieira Campos, Leandro Valiati, Ricardo Franzói, Carlos Augusto Schlabit

CONSELHO CURADOR: Mayara Penna Dias, Olavo Cesar Dias Monteiro e Irma Carina Brum Macolmes

**DIRETORIA**

**DIRETOR TÉCNICO:** MARTINHO ROBERTO LAZZARI

**DIRETOR ADMINISTRATIVO:** NÓRA ANGELA GUNDLACH KRAEMER

**CENTROS**

**ESTUDOS ECONÔMICOS E SOCIAIS:** Vanclei Zanin

**PESQUISA DE EMPREGO E DESEMPREGO:** Rafael Bassegio Caumo

**INDICADORES ECONÔMICOS E SOCIAIS:** Juarez Meneghetti

**INFORMÁTICA:** Valter Helmuth Goldberg Junior

**INFORMAÇÃO E COMUNICAÇÃO:** Susana Kerschner

**RECURSOS:** Graziela Brandini de Castro

M827e Morais, Igor A. Clemente de.  
Econometria Aplicada no EViews® / Igor Morais, Filipe Stona  
e Gustavo Schuck. - Porto Alegre : FEE, 2016.  
182 p. : il.

ISBN 978-85-7173-141-7

1. Econometria. 2. Estatística. 3. EViews (programa de computador). I. Stona, Filipe. II. Schuck, Gustavo. III. Fundação de Economia e Estatística Siegfried Emanuel Heuser. IV. Título.

CDU 330.43

Bibliotecário responsável: João Vítor Ditter Wallauer — CRB 10/2016

© 2016 Igor Morais

Publicado pela Fundação de Economia e Estatística Siegfried Emanuel Heuser



É permitido reproduzir, compartilhar e derivar trabalhos desta obra, desde que citada a fonte, sendo proibido o uso para fins comerciais, a menos que haja permissão, por escrito, do detentor dos direitos autorais.

As opiniões emitidas neste livro são de exclusiva responsabilidade dos autores, não exprimindo, necessariamente, um posicionamento oficial da FEE ou da Secretaria do Planejamento, Mobilidade e Desenvolvimento Regional.

Capa: Laura Wottrich.

Como referenciar este trabalho:

MORAIS, I. A. C. de; STONA, F.; SCHUCK, G. **Econometria Aplicada no EViews®**. Porto Alegre: FEE, 2016.

FUNDAÇÃO DE ECONOMIA E ESTATÍSTICA Siegfried Emanuel Heuser (FEE)

Rua Duque de Caxias, 1691, Porto Alegre, RS — CEP 90010-283

Fone: (51) 3216-9132 Fax: (51) 3216-9134 E-mail: biblioteca@fee.tche.br Site: www.fee.rs.gov.br



# Sumário

<b>I</b>	<b>Parte Um</b>	
<b>1</b>	<b><i>EViews</i><sup>®</sup></b> .....	<b>9</b>
1.1	<b>Programando no Eviews</b>	<b>10</b>
1.1.1	Exemplo de Programação .....	10
1.2	<b>Como abrir dados no <i>EViews</i><sup>®</sup></b>	<b>12</b>
1.3	<b>Do Excel para o <i>EViews</i><sup>®</sup></b>	<b>13</b>
1.4	<b>Criando um Workfile</b>	<b>15</b>
1.5	<b>Abrindo os dados do FRED</b>	<b>16</b>
<b>2</b>	<b>Gráficos no <i>EViews</i><sup>®</sup></b> .....	<b>19</b>
2.1	<b>Dados Categóricos</b>	<b>28</b>
2.2	<b>Exemplos de programas.prg</b>	<b>29</b>
<b>3</b>	<b>Funções de Distribuição</b> .....	<b>31</b>
3.1	<b>A Curva Normal</b>	<b>33</b>
3.2	<b>A curva <i>t-student</i></b>	<b>40</b>
3.3	<b>A Curva Qui-Quadrado</b>	<b>42</b>
3.4	<b>Curva F</b>	<b>49</b>
3.5	<b>Distribuição de Poisson</b>	<b>51</b>
3.6	<b>Exercícios</b>	<b>52</b>
3.7	<b>Sites úteis</b>	<b>54</b>

<b>4</b>	<b>Estatísticas, testes de hipótese e ANOVA</b>	<b>55</b>
4.1	Histograma e Estatísticas	56
4.2	Estatísticas por classificação ( <i>Statistics by Classification</i> )	59
4.3	Testes de Hipótese	60
4.4	Teste de Igualdade por Classificação	61
4.5	Teste de Distribuição Empírica (Kolmogorov–Smirnov)	62
4.6	Teste de Igualdade ( <i>Test of Equality</i> )	64
4.7	Gráficos Analíticos – Fazendo a distribuição dos dados	64
4.8	Teste de Razão de Variância	65
4.9	Exercícios	72
<b>5</b>	<b>Séries de tempo</b>	<b>75</b>
<b>5.1</b>	<b>Ajuste Sazonal</b>	<b>75</b>
5.1.1	Método das Médias Móveis (Moving Average Methods)	77
5.1.2	TRAMO/SEATS	80
5.1.3	Método Census X-12	81
5.1.4	Método Census X-13	86
5.1.5	Alisamento Exponencial	88
<b>5.2</b>	<b>ETS-ERROR-trend-seasonal</b>	<b>93</b>
<b>5.3</b>	<b>Ciclo</b>	<b>98</b>
5.3.1	Filtro Hodrick-Prescott	98
5.3.2	Filtros de Frequência	100
5.3.3	O Filtro Corbae-Ouliaris	104
<b>5.4</b>	<b>Autocorrelação (Correlograma)</b>	<b>105</b>
<b>5.5</b>	<b>Análise Espectral</b>	<b>108</b>
<b>5.6</b>	<b>Exercícios</b>	<b>111</b>
<b>5.7</b>	<b>Bibliografia</b>	<b>112</b>
<b>6</b>	<b>Regressão Simples</b>	<b>115</b>
<b>6.1</b>	<b>Diagnóstico Dos Coeficientes</b>	<b>124</b>
6.1.1	Scaled Coefficients	125
6.1.2	Intervalo de Confiança	125
6.1.3	Teste de Wald	127
6.1.4	Confidence Ellipse	129
6.1.5	Variance Inflation Factors	130
6.1.6	Decomposição da Variância do Coeficiente	131
6.1.7	Variáveis Omitidas	131
6.1.8	Variáveis Redundantes	134
6.1.9	Teste Factor Breakpoint	135
<b>6.2</b>	<b>Diagnóstico Dos Resíduos</b>	<b>137</b>
6.2.1	Teste de Normalidade	137
6.2.2	O teste de Independência (BDS)	138
6.2.3	Correlograma – Q-stat	139
6.2.4	Correlograma dos Resíduos ao Quadrado	140
6.2.5	Teste de Autocorrelação – LM	140
6.2.6	Testes de Heteroscedasticidade	142

<b>6.3</b>	<b>Diagnóstico De Estabilidade</b>	<b>147</b>
6.3.1	Teste de Chow	147
6.3.2	Teste de Quandt-Andrews	150
6.3.3	Teste de Previsão de Chow	152
6.3.4	Teste de Ramsey	153
6.3.5	Estimativas Recursivas	153
6.3.6	Leverage Plots	157
6.3.7	Estatísticas de Influência	158
<b>6.4</b>	<b>Previsão - Forecast</b>	<b>158</b>
<b>6.5</b>	<b>ANEXO ESTATÍSTICO</b>	<b>164</b>
6.5.1	MÍNIMOS QUADRADOS ORDINÁRIOS	164
<b>6.6</b>	<b>Bibliografia</b>	<b>166</b>
<b>7</b>	<b>Regressão Múltipla</b>	<b>167</b>
<b>7.1</b>	<b>O modelo com duas variáveis independentes</b>	<b>168</b>
<b>7.2</b>	<b>Previsão - Forecast</b>	<b>175</b>
<b>7.3</b>	<b>Método STEPLS</b>	<b>176</b>
7.3.1	Os métodos de Seleção STEPLS	178
<b>7.4</b>	<b>Bibliografia</b>	<b>180</b>
	<b>Referências Bibliográficas</b>	<b>180</b>





# Parte Um

<b>1</b>	<b><i>EViews</i><sup>®</sup></b> .....	<b>9</b>
1.1	Programando no Eviews	
1.2	Como abrir dados no <i>EViews</i> <sup>®</sup>	
1.3	Do Excel para o <i>EViews</i> <sup>®</sup>	
1.4	Criando um Workfile	
1.5	Abrindo os dados do FRED	
<b>2</b>	<b>Gráficos no <i>EViews</i><sup>®</sup></b> .....	<b>19</b>
2.1	Dados Categóricos	
2.2	Exemplos de programas.prg	
<b>3</b>	<b>Funções de Distribuição</b> .....	<b>31</b>
3.1	A Curva Normal	
3.2	A curva <i>t-student</i>	
3.3	A Curva Qui-Quadrado	
3.4	Curva F	
3.5	Distribuição de Poisson	
3.6	Exercícios	
3.7	Sites úteis	
<b>4</b>	<b>Estatísticas, testes de hipótese e ANOVA</b>	<b>55</b>
4.1	Histograma e Estatísticas	
4.2	Estatísticas por classificação ( <i>Statistics by Classification</i> )	
4.3	Testes de Hipótese	
4.4	Teste de Igualdade por Classificação	
4.5	Teste de Distribuição Empírica (Kolmogorov-Smirnov)	
4.6	Teste de Igualdade ( <i>Test of Equality</i> )	
4.7	Gráficos Analíticos – Fazendo a distribuição dos dados	
4.8	Teste de Razão de Variância	
4.9	Exercícios	
<b>5</b>	<b>Séries de tempo</b> .....	<b>75</b>
5.1	Ajuste Sazonal	
5.2	ETS-ERROR-trend-seasonal	
5.3	Ciclo	
5.4	Autocorrelação (Correlograma)	
5.5	Análise Espectral	
5.6	Exercícios	
5.7	Bibliografia	
<b>6</b>	<b>Regressão Simples</b> .....	<b>115</b>
6.1	Diagnóstico Dos Coeficientes	
6.2	Diagnóstico Dos Resíduos	
6.3	Diagnóstico De Estabilidade	
6.4	Previsão - Forecast	
6.5	ANEXO ESTATÍSTICO	
6.6	Bibliografia	
<b>7</b>	<b>Regressão Múltipla</b> .....	<b>167</b>
7.1	O modelo com duas variáveis independentes	
7.2	Previsão - Forecast	
7.3	Método STEPLS	
7.4	Bibliografia	
	<b>Referências Bibliográficas</b> .....	<b>180</b>





Do ponto de vista operacional o *EViews*® é muito mais do que um simples pacote estatístico com uma boa interface. Esse software permite ao usuário manter seus modelos atualizados em tempo real, conectando o mesmo a dados na internet. Permite programar rotinas diversas com vários modelos e, **a despeito das falhas de testes estatísticos de fronteira**, o usuário pode ainda se conectar com outros softwares como o R e o Matlab. O *EViews*® pode ser utilizado para análises estatísticas e econométricas de três diferentes maneiras: interface gráfica, comandos individuais e arquivo de programa. A interface gráfica nos remete a tudo que o usuário visualiza e interage através do uso do mouse, barra de menus e as janelas, como *workfile*, *spreadsheet* e gráficos.



Figura 1.1: Acessando o *EViews*®

Outra forma de acessarmos as funções do *software* é por instruções de comando. O *EViews*® nos possibilita duas maneiras, a primeira é pela janela de comando em branco logo abaixo da barra de menus. Nesta podemos executar instruções de somente uma linha, como por exemplo, **wfcreate u 1000** e pressionar **enter**, pronto: criamos um *workfile* com 1000 observações. Torna-se útil e veloz quando se está trabalhando com a interface gráfica e quer executar comandos simples. A última maneira é por um arquivo de programa no formato “.prg”. Através dos programas podemos mandar instruções mais complexas, trabalhar com um conjunto superior de dados, salvar nossas linhas de programação para aplicações futuras e conectar o *EViews*® a diferentes bancos de dados

ou outros softwares. Nesse capítulo faremos uma breve introdução sobre essas três diferentes formas de usar o *EViews*<sup>®</sup>.

## 1.1 Programando no Eviews

Para criar um programa é necessário abrir uma porta específica que fica em **File/New/Program**. Como *workfiles* e demais objetos, o *EViews*<sup>®</sup> nos permite trabalhar com apenas um programa aberto sem nomear por vez, o *untitled*. Para dar nome ao seu programa e, conseqüentemente, salvar na extensão “.prg”, pressione *save* ou *save as* na barra de menu da janela do programa e escolha o local desejado.

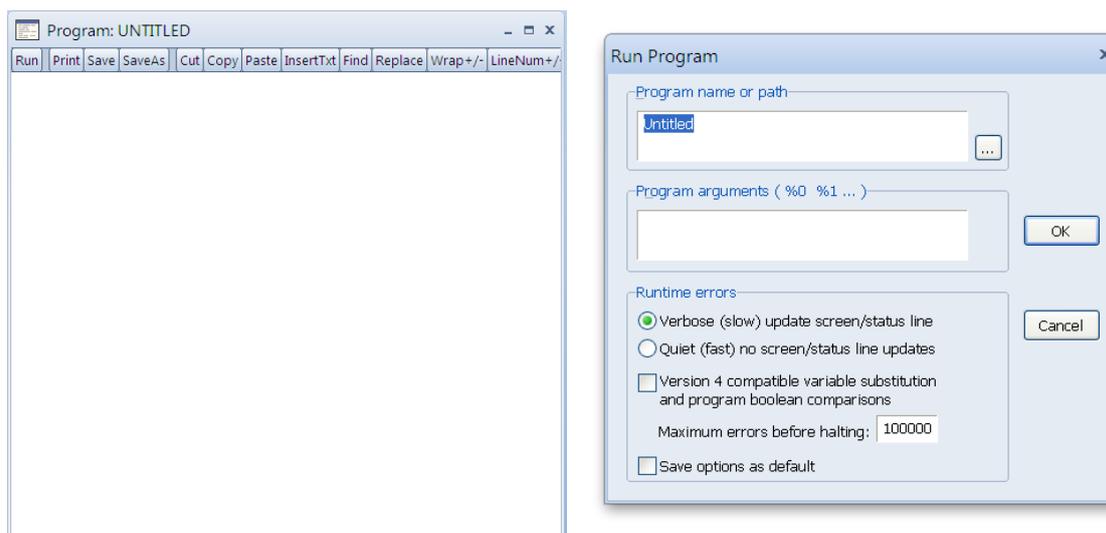


Figura 1.2: Programa sem Título

Uma vez salvo, os programas podem ser abertos através da barra de menus em **File, Open e, então, Programs**. Para executá-los basta pressionar **Run** na barra de menu da janela do programa aberto. Porém, muito cuidado ao fazer isso, pois se o caminho do programa não coincidir com o banco de dados ou se as variáveis que foram nomeadas não forem iguais, será retornado erro.

### 1.1.1 Exemplo de Programação

No *EViews*<sup>®</sup> os programas são executados linha por linha e cada linha é entendida como um comando. Comentários não executáveis podem ser adicionados depois do apóstrofo (') e tomam a cor verde na janela de programação.

**Programação 1.1.1** As linhas de comando a seguir criam um *workfile*, uma série aleatória, denominada dados e salva o valor da média no escalar *a*.

```
wfcreate u 100 'Comentário: workfile não estruturado com 100 observações
series dados = rnd
scalar a = @mean(dados)
show a
```

Primeiro é criado um *workfile* não estruturado utilizando o comando **wfcreate u**. Na segunda linha, **series** é o comando executado para criar ou alterar uma série de dados. Aqui nomeamos a série criada com o nome **dados** e pelo comando **rnd** geramos valores aleatórios. No caso de

alterarmos a linha 2 para **series dados = 2** a série dados irá tomar o valor 2 em cada observação.

Depois de criarmos dados é utilizado o comando **@mean(x)** para calcular sua média. Então, guardarmos esse valor dentro de um escalar denominado "a". O comando **show** apresenta qualquer objeto na tela, nesse caso "a".

Partindo do nosso programa inicial, podemos extrair mais informações da série dados. Por exemplo, para o número de observações, desvio padrão, valor máximo e mínimo utilizamos respectivamente os comandos **@obs(x)**, **@stdev(x)**, **@max(x)** e **@min(x)**.

```
wfcreate u 100
series dados = rnd
vector(5) a
a(1) = @obs(dados)
a(2) = @stdev(dados)
a(3) = @mean(dados)
a(4) = @max(dados)
a(5) = @min(dados)
show a
```

Note que, no lugar do escalar "a" utilizamos um vetor "a", isso nos possibilita guardarmos mais posições de informações. Esse vetor foi incluído para ter 5 linhas.

Na mesma linha de raciocínio, podemos desenvolver um programa que crie um *workfile* com, agora, cinco séries aleatórias e guarde o número de observações, desvio padrão, valor médio, máximo e mínimo.

```
wfcreate u 100
matrix(5,5) a
for !a = 1 to 5
series dados!a = rnd
a(1,!a) = @obs(dados!a)
a(2,!a) = @stdev(dados!a)
a(3,!a) = @mean(dados!a)
a(4,!a) = @max(dados!a)
a(5,!a) = @min(dados!a)
next
show a
```

Diferente do programa anterior, utilizamos uma matriz 5x5 "a" ao invés do vetor "a", para acomodar mais de uma coluna. Note que usamos o comando **!a**. Esse é para permitir que uma variável tenha um intervalo numérico. Também é aplicada a instrução **for**, que abre o *loop* encerrado pelo **next**. Este laço possibilita criarmos um circuito onde a variável "!a" tomará inicialmente o valor 1, procederá as linhas seguintes até o comando **next**, que aumenta "!a" em 1 e retorna a execução do programa para a linha do **for** até que "!a" guarde o valor 5 e quebre o circuito. Desta forma, sempre que houver um **for** existirá um **next** correspondente.

Para finalizar nosso programa, podemos adicionar um cabeçalho à nossa matriz "a". O *EViews*<sup>®</sup> não permite o uso de texto dentro de matrizes e por isso utilizamos **table(linha,coluna)** que cria um objeto tabela. Na terceira até a oitava linha adicionamos o cabeçalho na primeira coluna da tabela "a". Note que, textos são armazenados sendo colocados dentro de aspas.

```
wfcreate u 100
table(6,6) a
a(1,1) = "Estatísticas/Série"
a(2,1) = "Obs"
a(3,1) = "Desvio Padrão"
a(4,1) = "Média"
a(5,1) = "Máximo"
a(6,1) = "Mínimo"
for !a = 1 to 5
series dados!a = rnd
a(1,1+!a) = "dados"+@str(!a)
a(2,1+!a) = @obs(dados!a)
a(3,1+!a) = @stdev(dados!a)
a(4,1+!a) = @mean(dados!a)
a(5,1+!a) = @max(dados!a)
a(6,1+!a) = @min(dados!a)
next
show a
```

Outro detalhe a ser observado é na linha 11, onde é preenchido a primeira linha de " a" com o nome das séries. Para a tabela aceitar o "dados" seguido dos valores de "!a"(1,2,3,4 e 5) em forma de texto é utilizado o comando **@str(número)** que transforma os valores numéricos em texto.

Note que as variáveis de controle iniciadas com exclamação (!) são utilizadas para armazenar números. Já as variáveis iniciadas com o símbolo de porcentagem (%) guardam informações de texto.

**Programação 1.1.2** Uma maneira prática de manipular um conjunto de séries de tempo é agregando em um grupo. Abaixo agrupamos todas as séries do *workfile* em um grupo de nome "g". Na segunda linha, tiramos a série *resid* do nosso grupo e então, na terceira linha, instruímos o EViews® a buscar o nome da primeira série do grupo "g" e guardar esse informação em "%a".

```
group g *
g.drop resid
%a = g.@seriesname(1)
scalar b = @mean{%a}
show b
```

Na quarta linha, gravamos a média de "%a" dentro do escalar "b" e então exibimos "b". Note que adicionamos colchetes em "{%a}", isso faz com que o EViews® execute o texto dentro da variável. Não se esqueça de salvar.

Ao longo desse livro iremos exemplificar diversas ações que podem ser feitas criando seu próprio programa. A idéia é que, ao final do livro, você tenha desenvolvido as habilidades mínimas para criar um programa.

## 1.2 Como abrir dados no EViews®

Há várias formas de abrir dados no EViews® e cada uma delas irá depender do tipo de informação que será utilizado e dos objetivos de pesquisa. As opções para criar um banco de dados são muitas,

mas, para os propósitos desse livro, precisaremos apenas aprender como abrir ou criar os chamados *workfile*.

Para tanto, iremos dividir essa análise em duas partes. Primeiro abordando sobre a criação de um conjunto de dados no Excel que, posteriormente, são lidos no EViews®. A seguir, criando um *workfile* e copiando e colando dados. Qual das duas alternativas escolher fica a seu critério.

### 1.3 Do Excel para o EViews®

Vamos supor que se tenha um conjunto de séries de tempo de periodicidade trimestral, com início em 2006Q1 e término em 2014Q2. Essas podem ser vistas no arquivo em Excel de nome dados/exemplo1.

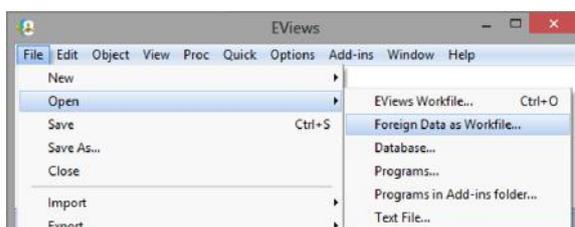


Figura 1.3: Importando dados do Excel

Como primeiro passo, abra o EViews®. Note que, por hora, não há nada disponível, nem dados, informação sobre a periodicidade e etc. A seguir, vá em **File/Open/Foreign Data as Workfile** (ver figura 1.3). E selecione o arquivo em Excel (vá até a pasta onde o mesmo foi salvo). Depois clique em **Ok**.

A janela de opções que se abre a seguir consiste de três passos. É muito comum que não se mudem as informações no primeiro e no segundo passos. Nesse caso, podemos clicar em **avanzar** nesses dois primeiros passos. Porém, no terceiro passo, caso não se modifique algumas opções, pode ser que o arquivo final não fique tal como desejado. Em especial se não especificarmos a periodicidade e as datas.

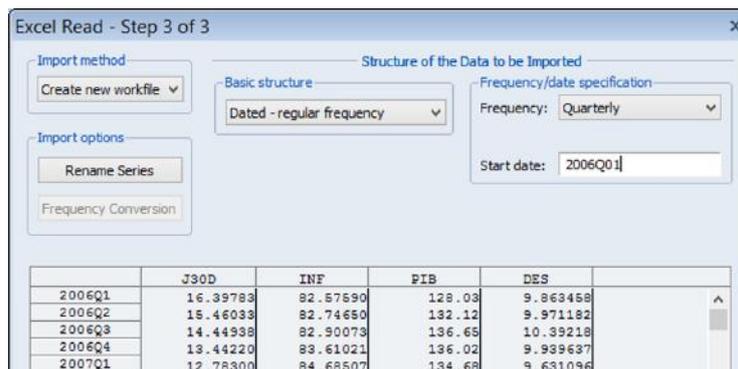
Sendo assim, no passo três, selecione **Dated - regular frequency**, que sempre será utilizado quando tivermos uma série de tempo e, depois, a periodicidade trimestral (**Quarterly**), conforme a figura 1.4a. Por vezes, o arquivo do Excel já tem uma série com os dados. Nesses casos, a opção **Dated - specified by date series** identifica automaticamente a frequência e o início da série, basta informar no campo **Date series** o nome da série que tem essa informação (ver figura 1.4b). Por fim, digite a data inicial como 2006Q1 e clique em **Finish**.

Pronto, agora temos um *workfile* de 30 trimestres contendo cinco séries de tempo com os respectivos nomes que estavam no Excel: J30D, INF, PIB e DES.

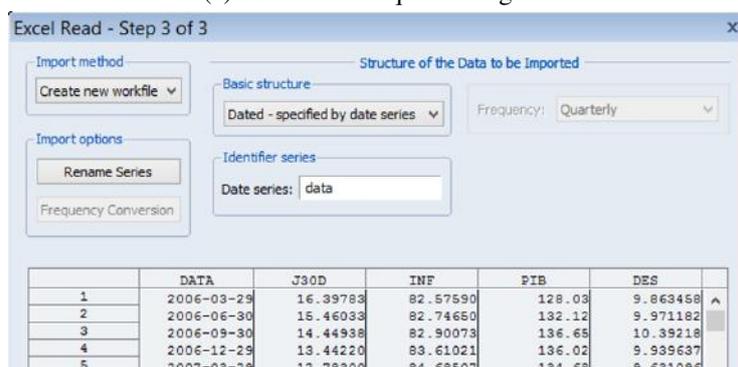
**Programação 1.3.1** Uma forma de fazer a abertura de um *workfile* é via programação, que dá muita agilidade ao trabalho. Basta aplicarmos o comando **wfopen** seguido do caminho onde se encontra o arquivo com as séries.

```
wfopen c:/exemplo1.xlsx
```

Um último ponto importante para salientar nesse momento é sobre a forma que as datas são inseridas no EViews®. Como pode ser visto no exemplo acima, especificamos primeiro o ano, seguido da letra que compõem a periodicidade, no caso de trimestre **Q** e, no caso de meses **M** para então, colocar o período. Como os dados começam no primeiro trimestre, colocamos "1". Se os dados tivessem como início março de 1996, especificaríamos 1996M3. Note que os dados estão no formato Inglês, onde os decimais são separados por ponto. Caso seu computador estiver no formato Português (Brasil), teremos problema na hora que o EViews® abrir esses dados do excel. Ele irá confundir os pontos com as vírgulas. A sugestão é reconfigurar o computador para o Inglês americano. A localização desta opção pode variar ligeiramente conforme a versão do Windows. Para o Windows 7, acesse **Painel de Controle/Relógio, Idioma e Região/Região e Idioma** na aba **Formatos** selecione o **Formato Inglês (Estados Unidos)**. Então clique em **Aplicar**



(a) Estrutura - frequência regular



(b) Estrutura - frequência definida por série

Figura 1.4: Importando Dados

e Ok.

**Programação 1.3.2** Sempre que for iniciar um programa pode digitar os comandos abaixo para que seu banco de dados seja aberto automaticamente.

```
%path = @runpath
cd %path
```

Ao rodar os comandos acima, o caminho utilizado para abrir os dados, mostrado na barra de status no canto inferior da tela, será alterado para o caminho que foi salvo o programa. Sendo assim, recomenda-se colocar o arquivo ".prg" na mesma pasta em que se encontra o ".wfl". Desta forma, se salvamos o exemplo1.xlsx dentro da mesma pasta do programa podemos importar os dados por programação.

```
%path = @runpath
cd %path
wfopen exemplo1.xlsx
```

Também é possível definirmos um caminho diferente do que o programa está salvo. Alterando a primeira linha de comando.

```
%path = "c:/nome da pasta/"
cd %path
```

```
wfopen exemplo1.xlsx
```

## 1.4 Criando um Workfile

Também podemos copiar os dados que estão no Excel e colar os mesmos no *EViews*<sup>®</sup>. Nesse caso precisamos criar, como primeiro passo, um *workfile*. Assim, abra um novo arquivo do *EViews*<sup>®</sup> que não contenha informações. A seguir, vá em **File/New/Workfile** (ou **Ctrl+N**). Dentre as diversas opções disponíveis, selecione **Dated - regular frequency**, a seguir **quarterly** e especifique o intervalo dos dados, escrevendo a data inicial e final (figura 1.5). Veja como é o formato de datas. Primeiro o ano, seguido da letra do período e depois o número do período.



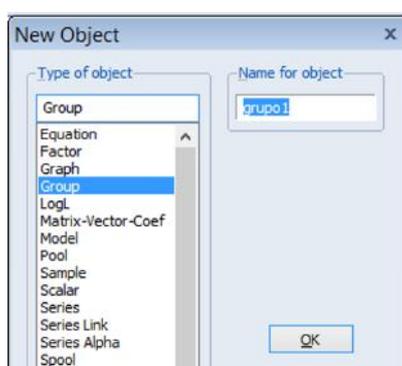
Figura 1.5: Criando Workfile

Primeiro o ano, seguido da letra do período e depois o número do período.

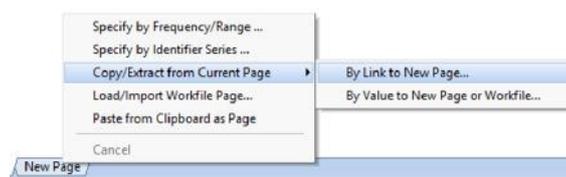
**Programação 1.4.1** Para criar um *workfile* não estruturado utilizamos o comando **wfcreate u** seguido do número de observações desejadas. Para criarmos um *workfile* mensal utilizamos a opção **m** seguida da data inicial e final. Da mesma forma, para frequência trimestral utilizamos **q** e anual **a**.

```
wfcreate u 100
wfcreate m 1990m1 2015m12
wfcreate q 1990q1 2015q4
wfcreate a 1990 2015
```

O arquivo que está sendo criado ainda não possui os dados, apenas criamos o chamado *workfile*. Para inserir os dados temos que, primeiro, criar um objeto (figura 1.6a). Vá em **Object/New Object** e selecione a opção **Group**. Do lado direito escolha um nome para o grupo (evite acentos, espaços e etc, seja bem simples nessas escolhas). A seguir, depois de clicar em **Ok**, o *EViews*<sup>®</sup> irá abrir uma janela que é bem semelhante com planilhas do Excel. Vá no Excel, selecione apenas os dados, não pegando as datas nem os nomes das séries, copie e cole no *EViews*<sup>®</sup>. A seguir feche o mesmo.



(a) Criando Workfile



(b) Criando página com vínculo

Figura 1.6: Novo Objeto

Note que agora temos um *workfile* e os dados, mas, as séries ficaram com nomes diferentes. Isso pode ser resolvido clicando com o botão direito na série e renomeando a mesma.

Após ter os dados no EViews®, há diversas outras formas de trabalhar com eles de forma a tornar a pesquisa mais fácil, em especial quando se trabalha com uma grande quantidade de informação e diversos testes e estimativas.

Uma opção interessante do EViews® é o uso de diversas planilhas ao mesmo tempo, sendo possível preservar o vínculo entre as variáveis. Tal recurso permite trabalhar com diversos modelos, separados por planilhas, sem poluir o *workfile* principal. Selecione as variáveis *des*, *inf*, *j30d* e *pi.b*. A seguir, clique com o botão direito do mouse na planilha de nome **New Page**, selecione **Copy/Extract from Current Page** e depois **By Link to New Page**.

Na janela que será aberta, ao escrever **@all**, o EViews® irá copiar todo o período amostral. Em **Objects to copy**, selecione **Listed Series**, como mostrado na figura 1.7, e deixe a opção **Include Links** selecionada. Caso queira dar um nome para a nova planilha, clique em **Page Destination** e, em **Page**: escreva o nome que quiser.

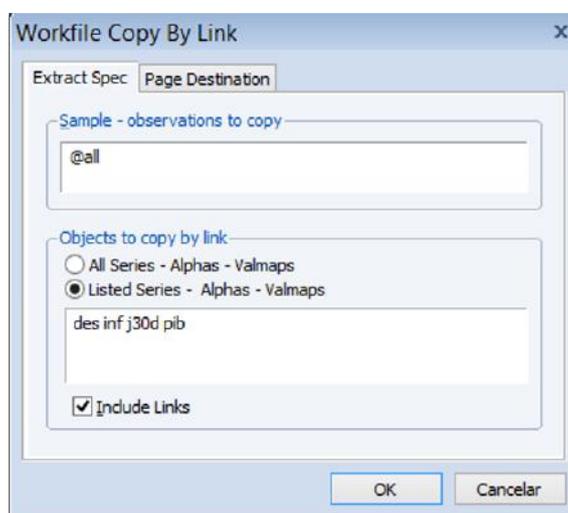


Figura 1.7: Objetos com vínculo

Note que será criada uma nova planilha com os dados selecionados com cores diferentes. Agora, sempre que os dados nas séries da planilha original forem modificados, o mesmo irá ocorrer com essas séries na nova planilha.

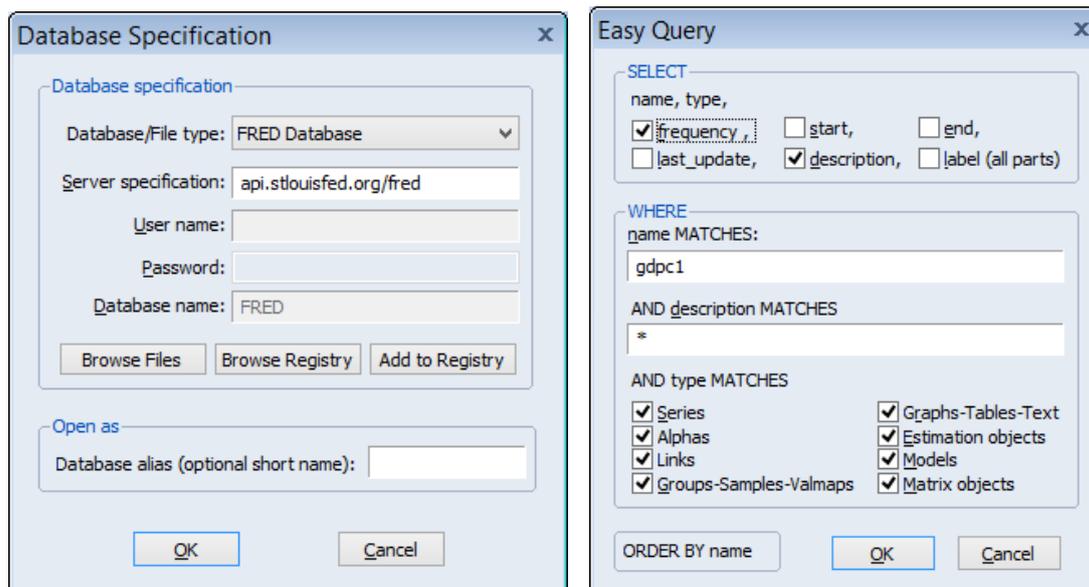
## 1.5 Abrindo os dados do FRED

Uma das funcionalidades interessantes do EViews® é poder abrir diversos formatos de dados, e um deles que é bastante útil para análise de conjuntura são os dados do FRED que é uma base de dados disponibilizada pelo *Federal Reserve of Saint Louis*<sup>1</sup>. Como esse é um banco de dados disponibilizado na internet, sua leitura só é possível se houver conexão à internet.

O primeiro passo é descobrir o nome da série que se quer abrir. Nesse caso, vá no site do *Federal Reserve Board of Saint Louis* e descubra o código da série que se busca. Como exemplo, vamos usar o PIB Real dos EUA em dólares de 2005, cujo código é GDPC1.

Agora, abra um arquivo do EViews®, vá em **File/ Open/ Database..** e selecione FRED database e clique e ok. A janela que irá ser aberta é a que permite fazer a conexão com o banco de dados, tal como a Figura 1.8a. A seguir, vá em EasyQuery, abrindo a caixa de diálogo da Figura

<sup>1</sup>Se você ainda não conhece esse recurso, vale a pena ver em: <http://research.stlouisfed.org/fred2/>



(a) Seleção da Base de Dados

(b) Easy Query

Figura 1.8: Abrindo dados do FRED

1.8b, e em **name MATCHES**, escreva o nome da série. No nosso caso, GDPC1 e clique em ok. A seguir, dê dois cliques na série e exporte a mesma para um banco de dados. Posteriormente iremos mostrar como é possível você mesmo criar um link entre o *EViews*<sup>®</sup> e um banco de dados que se queira para atualização automática. Também é possível criar um add-in que faz essa seleção automática.





## 2. Gráficos no *EViews*®

O recurso de gráficos em econometria é muito útil para uma detecção prévia das características de um conjunto de dados como, por exemplo, sua distribuição, a existência de tendência, movimentos cíclicos, sazonalidade, outliers, quebra estrutural, clusters dentre outras. No *EViews*® é possível personalizar a construção de gráficos, escolhendo cores, tamanho e estilo de letra, linhas de tendência, combinar diferentes tipos de gráficos, vincular os mesmos aos dados e demais aspectos. Há outras opções disponíveis em **Options/Graphics Default**. Deixamos para o leitor explorar esse ponto consultando o manual que acompanha o *software*.

Nesse capítulo iremos utilizar o arquivo do *EViews*® de nome *exemplo1.wfl*. Abra o mesmo. Ali irá ver cinco séries de dados de nome “qx, y, px, pm, qm”. Inicialmente, dê dois cliques na série de nome qx. O *EViews*® irá abrir uma janela que se parece com as planilhas do Excel. A sequência de dados que vemos é denominada de série de tempo. Note que, na primeira coluna, temos as respectivas datas que, para esse exemplo, é trimestral, com início no primeiro trimestre de 1997 e terminando no segundo trimestre de 2015. Porém, o intervalo vai até 2015Q4, o que resulta em uma sequência de células que estão vazias, com o termo “NA”. Isso irá facilitar quando quisermos prever o comportamento dos dados para alguns períodos a frente. Veremos isso no capítulo de regressão simples.

A seguir, a partir do menu **View/Graph...** Note que há várias opções de gráficos. O mais comum, e que será mais explorado aqui, é fazer um gráfico de linha. Selecione esse e o resultado é como aparece na figura 2.1. Alternativamente, podemos fazer um gráfico de barras para esse conjunto de dados. Clique com o botão direito do mouse sobre o gráfico e depois **Options** e selecione **Bar**. O mesmo pode ser aplicado a cada uma das outras opções. Outra alternativa é usar o menu opções, localizado logo acima do gráfico.

Note que ao fazer o gráfico aparece na parte inferior do mesmo uma barra de rolagem. A partir dessa podemos deslizar o gráfico para diferentes datas, basta que mova o cursor na barra para a esquerda ou para a direita.

O *EViews*® permite que se escolha entre diferentes maneiras de apresentar os gráficos, mudando o fundo para cor branca, tornando as linhas mais nítidas, mudando a cor das linhas e etc. Para verificar todas essas opções, com o gráfico aberto clique com o botão direito do mouse e selecione

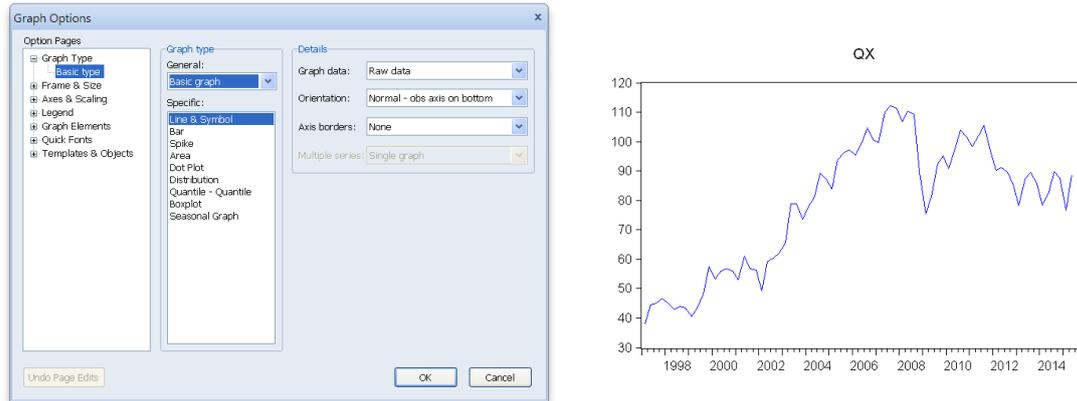


Figura 2.1: Opções de Gráficos

**Templates.** A seguir, escolha cada um dos modelos e, antes de clicar em **Ok**, clique em **Apply** para ver se te agrada.

**Programação 2.0.1** Para fazer um gráfico, usamos o comando “graph”. Para o exemplo de um gráfico de linha, aplicado à série *qx* colocamos o termo abaixo criando um gráfico de nome *gqx*. A seguir, especificamos que a linha tem cor vermelha, dado pelo  $RGB(255,0,0)^a$ .

```
graph gqx.line qx
gqx.setelem linecolor(255,0,0)
```

Dentre as várias opções a serem utilizadas em um gráfico de linha, uma das mais úteis para a econometria é a padronização dos dados. Nesse caso, o que fazemos é criar um gráfico onde cada informação é subtraída da média e depois dividida pelo desvio-padrão. Assim, o resultado final é uma nova sequência de dados onde a média é zero e o desvio-padrão é 1. Para essa opção use :

```
graph gqx.line(n) qx
```

<sup>a</sup>Se quiser outra cor, consulte os códigos de cores RGB

Após criar o gráfico, como mostrado no box de programação, o produto final é um gráfico no estilo congelado ou “frozen”. Esse é uma espécie de gráfico desvinculado dos dados. O inconveniente dessa opção é que toda vez que os dados originais forem atualizados isso não será feito no nosso gráfico, ou seja, ao aplicar o *freeze* no gráfico, o mesmo perde o vínculo com os dados.

Para contornar esse problema devemos voltar a vincular os dados ao gráfico. Dê dois cliques no gráfico *gqx*. A seguir selecione **Graph Updating** e, do lado direito as opções **Automatic** e, mais abaixo, **Update when data or the workfile sample changes**.

**Programação 2.0.2** Podemos montar um programa que faça automaticamente a atualização dos nossos gráficos. Primeiro criamos um gráfico de nome *gqx* e depois especificamos pelo comando **setupdate** e, entre parênteses “a”, que o mesmo seja atualizado sempre que o conjunto de dados mudarem. Ao fazer isso note que a cor da caixa que especifica o gráfico no workfile muda da cor verde para alaranjado.

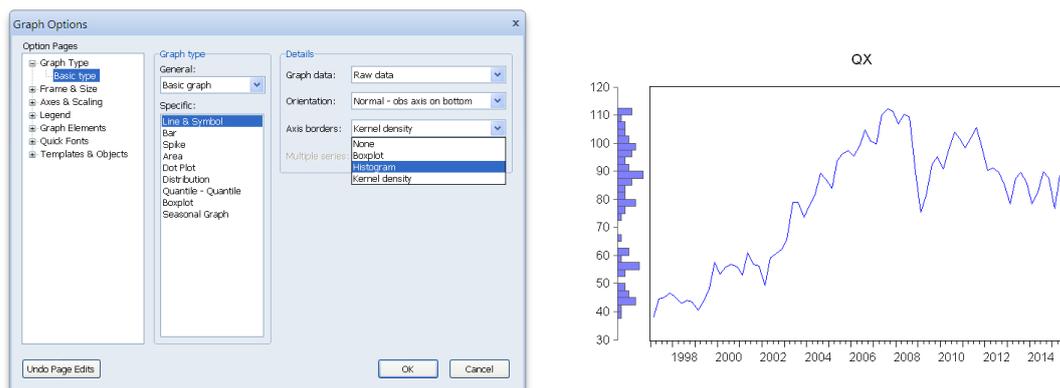


Figura 2.2: Gráfico de qx com a distribuição de frequência

```
graph gqx.line(n) qx
gqx.setupdate(a)
```

Vimos anteriormente que também temos a opção de criar um gráfico de barra. Porém, essa não é muito interessante quando há uma série de dados com muitas informações, isso porque as barras acabam ficando muito finas, fazendo com que o gráfico de barras se aproxime de um gráfico de área. Para o exemplo da série qx, selecione a opção de gráfico de barra e veja como fica. Caso a sua escolha seja para um gráfico de barra há várias opções interessantes. Clique duas vezes no gráfico e selecione **Graph Elements/Bar-Area-Pie**. Ali será possível escolher entre gráficos com efeito de cores, 3D, colocar os respectivos valores em cada barra e diversas outras opções.

Outra possibilidade de uso dos gráficos no *EViews*<sup>®</sup> é combinar diferentes informações. Por exemplo, vamos ver como fazer um gráfico que mostre simultaneamente a evolução dos dados no tempo e a distribuição dos mesmos.

Com a série qx aberta, vá em **View/Graph...**, selecione **Line&Symbol** e depois, na opção **Axis borders**, escolha **Histogram**. Também há a opção de usar a densidade de kernel. Note que a série é mostrada considerando as datas no eixo horizontal e as escalas no vertical. A distribuição de frequência dos dados é colocada nesse eixo.

**Programação 2.0.3** Esse gráfico também pode ser feito a partir da opção `ab=hist` no comando `line`, como mostrado a seguir:

```
graph gqx.line(ab=hist) qx
```

Alternativamente, se quisermos especificar ma distribuição de kernel ao invés da distribuição de frequência, podemos usar o comando:

```
graph gqx.line(ab=k) qx
```

Além disso, podemos adicionar um texto para identificar nosso gráfico. No exemplo abaixo colocamos um título série de dados qx, entre aspas, com uma fonte de tamanho 12, do tipo ubuntu light. Por fim, o comando `t` especifica que o texto é centralizado.

```
gqx.addtext(pt=12,face="ubuntu light",t) "Serie de dados qx"
```

Algumas opções para gráficos no *EViews*<sup>®</sup> somente se tornam disponíveis quando o gráfico é

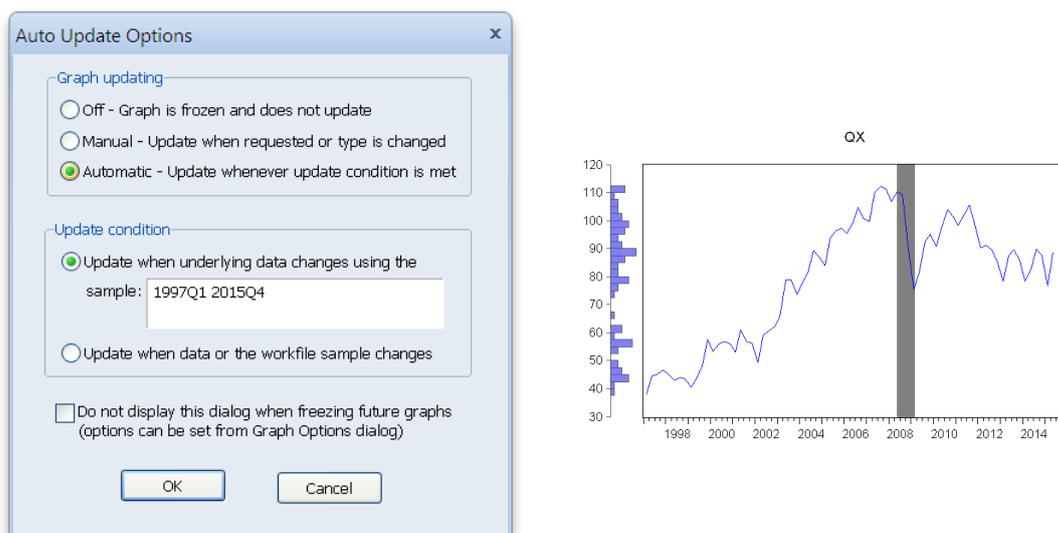


Figura 2.3: Gráfico de qx com area

um objeto(nomeado ou não). Pela linha de comando criamos automaticamente o objeto, a exemplo do gráfico `qx` criado acima. Para tanto, na interface gráfica utilizamos a função *Freeze*. Faça um gráfico da série `qx` e, no menu superior, poderá ver essa função. Uma das funções interessantes do *EViews*® é marcar períodos específicos de dados com uma área escura, muito útil quando estamos avaliando ciclo dos negócios e gostaríamos de sinalizar os períodos em que uma economia estava em recessão. Ou então queremos apenas sinalizar um intervalo de tempo para mostrar algum acontecimento.

Para usar essa função, clique com o botão direito do mouse no gráfico e, a seguir, selecione **Add lines & shading**. Note que esse recurso não está disponível para gráficos comuns. Como dito anteriormente, para habilitar essa função devemos selecionar antes o *Freeze*. A seguir, clique com o botão direito do mouse, selecione **Add lines & shading**, marque **Shaded Area**, deixe em **Vertical – Bottom axis** e mude o período para 2008Q2 até 2009Q1. Caso não esteja satisfeito com esse intervalo, clique duas vezes sobre a área cinza e modifique o intervalo.

Lembre que a opção *Freeze* tem a desvantagem de não ser atualizada sempre que os dados forem atualizados. Podemos contornar isso. Com o gráfico aberto dê dois cliques e depois selecione **Graph Updating**. A seguir, selecione a opção **Automatic** e **Update when data or the workfile sample changes**. Isso irá permitir que o gráfico seja atualizado sempre que os dados forem modificados no workfile.

**Programação 2.0.4** Uma opção interessante a ser utilizada em gráficos é especificar uma área em um determinado período. Isso pode ser feito a partir do comando `draw`. Dentre as opções, escolhemos que a área segue as datas na parte horizontal (bottom), a cor cinza (gray) e o período compreendido.

```
qx.draw(shade,bottom,color(gray)) 2008Q2 2009Q1
```

Outra opção que pode ser utilizada é mostrar duas séries de dados no mesmo gráfico, em especial quando as mesmas possuem escalas diferentes. Nesse caso, se fizermos esse gráfico com apenas um eixo vertical, visualmente podemos ter uma informação de baixa qualidade. O *EViews*® permite que se faça um gráfico com dois eixos, cada um com escala diferente.

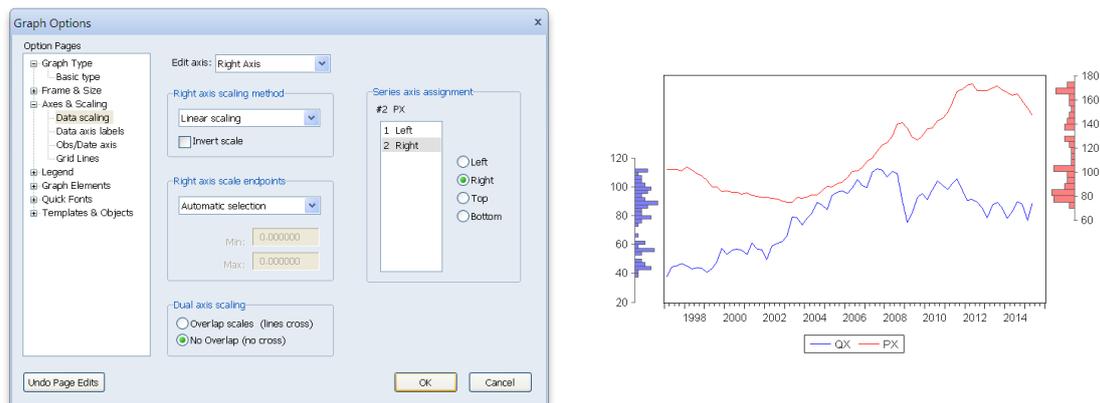


Figura 2.4: Gráfico de qx e px

Isso pode ser feito depois de se criar um grupo com as séries que se quer ilustrar. Selecione primeiro a série qx. Com o botão **Ctrl** do teclado pressionado, selecione a série px. A seguir, clique com o botão direito do mouse e **Open as Group**. O *EViews*<sup>®</sup> irá abrir as duas séries em conjunto, uma em cada coluna. A seguir, clique em **View/Graph...** e depois em **Ok**.

Note que temos uma única escala do lado esquerdo do gráfico. Agora, clique com o botão direito do mouse, vá em options e **Axes & Scaling** e, depois **Data scaling**. A seguir, do lado direito da tela, para cada série selecionada, escolha a escala que quer colocá-la, se esquerda ou direita. Nesse exemplo, escolhemos deixar a série qx no eixo esquerdo e a px no direito. Como exercício, veja se consegue também inserir a informação da distribuição de frequência para cada conjunto de dados como mostrado na figura 2.4.

**Programação 2.0.5** Um gráfico com duas linhas em duas colunas de escalas diferentes pode ser obtido a partir de uma instrução por linha de comando. Nesse caso, usamos “d”, que permite criar um gráfico com duas colunas. Não se esqueça de especificar qual é a segunda série de dados que se quer colocar junto. No exemplo abaixo usamos a série px. Note que também especificamos a opção de histograma.

```
graph gqx.line(ab=hist,d) qx px
```

Outra forma de usar os recursos gráficos é para identificar características estatísticas dos dados, uma possível relação entre diferentes variáveis dentre outras opções. Vamos iniciar essa discussão mostrando como são as funções de distribuição. Selecione a série qx. A seguir, vá em **View/Graph...** e, em **Graph Type**, selecione **Distribution**. Do lado direito, em **Details**, poderá ver que há diversas opções de gráfico. Selecionando **Histogram**, o *EViews*<sup>®</sup> irá retornar a distribuição dos dados de acordo com intervalos pré determinados.

Essa análise pode ser complementada com um gráfico que tem o mesmo formato, mas que, ao invés de ser uma distribuição de frequência, seja uma função de densidade ou então uma função de frequência relativa. Essas três opções podem ser selecionadas ao lado da opção **Histogram** na caixa **Options**. Vá em **Scaling** e selecione **Density**. O desenho não irá mudar, mas, note que a escala vertical sim. Isso porque, no caso da frequência temos, no eixo vertical, a informação do número de dados encontrados para cada intervalo. No caso da densidade estamos falando da área, o que também será diferente para o caso de se selecionar **Relative frequency**.

Vamos agora adicionar uma estimativa da função de distribuição utilizando uma função de Kernel. Com a série de dados qx aberta, faça o gráfico de distribuição e a seguir clique em **Options**.

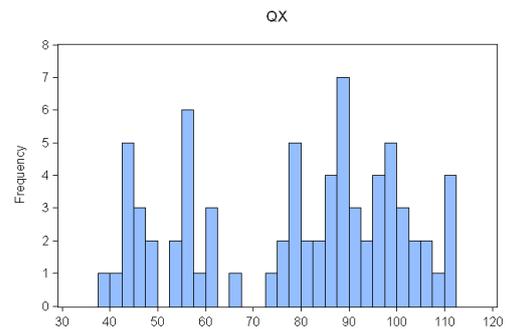
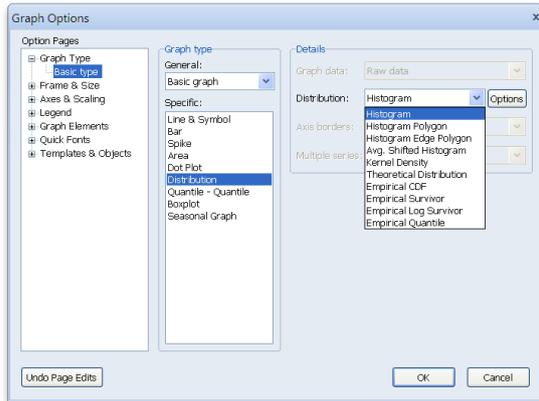


Figura 2.5: Gráfico de Distribuição de Frequência

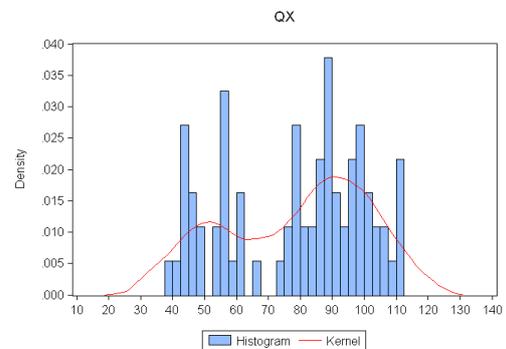
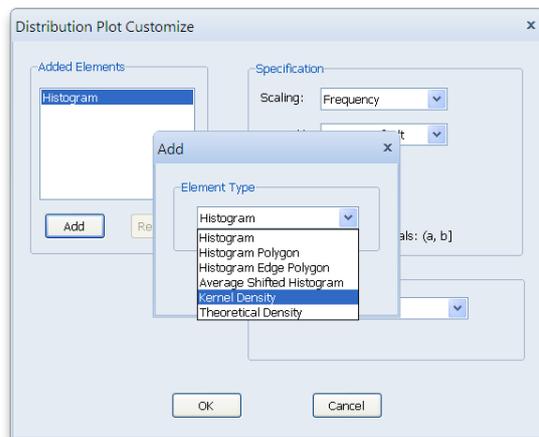


Figura 2.6: Adicionando uma densidade de Kernel

A seguir, na caixa **Details**, selecione **Options** e depois em **Add**. Escolha **Kernel density** e clique em ok. A figura 2.6 mostra o resultado<sup>1</sup>.

Note que há várias opções para a densidade de kernel. A figura 2.7 a seguir, mostra a comparação entre essas diversas funções utilizadas para estimar a função de densidade de kernel. Note que há pouca diferença entre os resultados.

**Programação 2.0.6** Para fazer um gráfico de distribuição conjugado com uma estimativa via densidade de Kernel, podemos usar o seguinte comando.

```
qx.distplot hist kernel
```

Ou então, se quisermos colocar em um único gráfico as diversas estimativas das funções de

<sup>1</sup> A ferramenta de determinar a densidade de kernel é uma forma não-paramétrica utilizada para determinar a densidade de uma função de distribuição de dados aleatórios, onde não conhecemos a função de distribuição verdadeira. Nesse caso, fazemos inferência sobre essa distribuição utilizando estatísticas da amostra que temos. Há várias funções de kernel disponíveis no *EViews*<sup>®</sup>: Epanechnikov, uniforme, triangular, normal, biweight, triweight e cosinus. Se a opção é utilizar a kernel normal, então, na sua estimativa é utilizada uma função de densidade normal padrão.

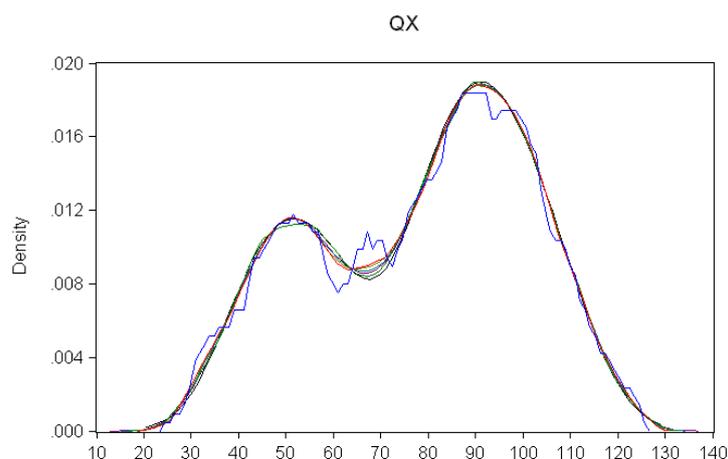


Figura 2.7: Comparação entre diversas funções de densidade de Kernel

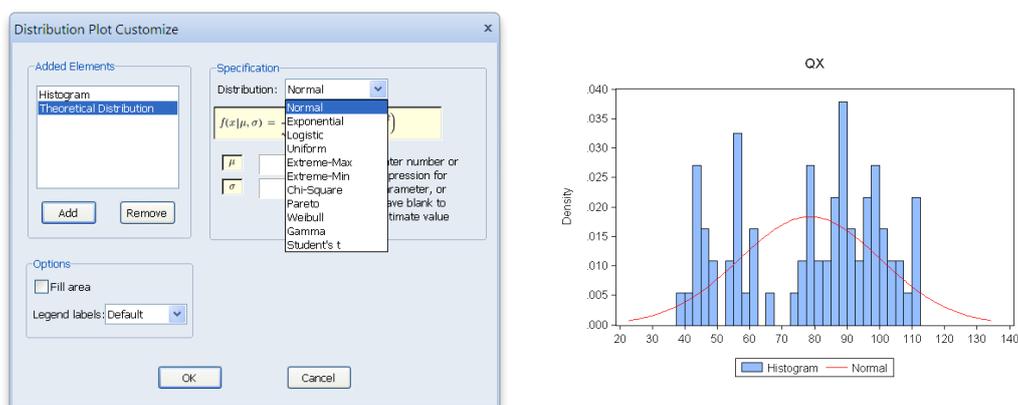


Figura 2.8: Gráfico de Distribuição de Frequência

kernel, usamos:

```
qx.distplot kernel(k=u,x) kernel(k=e) kernel(k=r) kernel(k=n) kernel(k=b)
kernel(k=t) kernel(k=c)
```

Alternativamente, com o gráfico aberto, clique em **Proc/Make Distribution Plot Data**. Como pode ser visto, há várias opções de distribuição que podemos investigar. Voltando ao nosso exemplo da distribuição de frequência, vá em **Options**, depois em **Add** e podemos ver que há diversas opções além do histograma. Já exemplificamos o uso da densidade de kernel. Selecione agora **Theoretical Density**, e clique em **Ok**. A seguir, clique novamente em **Theoretical distribution**, e veja que há diversas opções de funções de distribuição. Selecione a **Normal** e escolha os parâmetros. Se a escolha for  $\mu = 0$  e  $\sigma = 1$ , então simularemos uma curva normal padrão junto com nosso histograma dos dados, como pode ser visto na figura 2.8.

**Programação 2.0.7** Para inserir um gráfico com distribuição teórica junto com o histograma podemos usar o seguinte comando:

```
qx.distplot hist theory(dist=normal)
```

As opções de construção de gráficos também permite que sejam investigadas características dessa distribuição. Como se sabe, a função de distribuição cumulativa de dados que possuem uma distribuição normal tem o formato de um “S”. Mais a frente entraremos em detalhe sobre a função cumulativa e sua importância na determinação das probabilidades associadas a valores na construção de intervalos de probabilidade, teste de hipótese e uso em modelos como probit. Para investigar se os nossos dados possuem essa característica, com o gráfico aberto, clique com o botão direito do mouse e selecione **Options**. A seguir, do lado direito da tela, em **distribution**, selecione a opção **Empirical CDF**, que irá retornar os resultados para uma função de distribuição cumulativa. Como pode ser visto pela figura 2.9a, os nossos dados não parecem ter uma distribuição normal. Outra forma de verificar isso é via quantis. Abra a série **qx**, clique em **View/Graph...** e, na tela **Graph Type**, na parte **Specific**, clique em **Quantile-Quantile** e depois, em **Q-Q graph**, e selecione **Theoretical**. Note que, em ambos os resultados mostrados na figura 2.9, não há evidências de uma distribuição normal. Porém, para confirmar tal resultado é necessário que se faça um teste específico que será explicado no Capítulo 4.

A Figura 2.9 representa o gráfico da distribuição cumulativa associada a cada valor, no eixo horizontal, o percentual de vezes que o mesmo se encontra no conjunto de dados que são menores ou iguais a esse valor. Dessa forma, no eixo vertical fica descrita essa participação percentual, também denominada de frequência. Note que, como estamos falando de distribuição acumulada, ao final, teremos uma frequência de valor 1, ou seja, 100%. No gráfico da Figura 2.9a o valor 70 estaria associado a uma frequência de 0,33 no eixo vertical. Ou seja, a probabilidade de encontrarmos um valor, no nosso banco de dados, que é menor que 70,  $\mathbb{P}(x \leq 70) = 0,33$  é de 33%. Além de mostrar essa linha, o *EViews*® também coloca o intervalo de confiança, apresentado pela linha pontilhada.

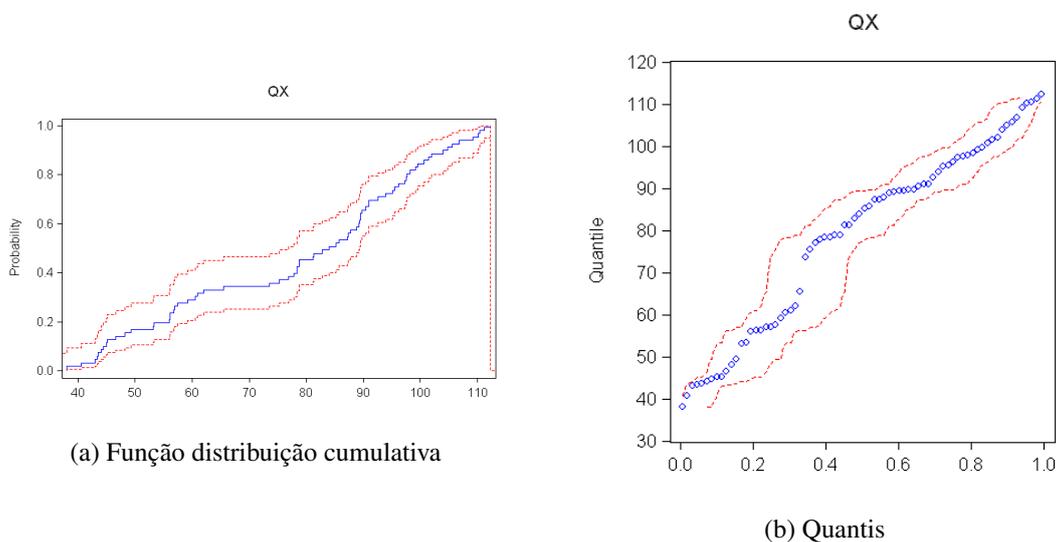


Figura 2.9: Gráfico da empirical CDF e quantile-quantile

**Programação 2.0.8** Para ver o gráfico da empirical CDF usamos o comando abaixo:

```
qx.distplot cdf
```

E, para fazer o gráfico do quantile-quantile usamos:

```
qx.distplot quantile
```

Além dos gráficos para uma série de dados apenas, podemos ver como se dá a relação entre dois conjuntos de dados, uma investigação prévia dos resultados da regressão simples. Nesse caso, vamos comparar o resultado da série  $y$  com a série  $qx$ , considerando que  $qx=f(y)$ .

Primeiro selecione a variável  $y$  e depois  $qx$  e clique com o botão direito do mouse abrindo ambas como grupo. A ordem das variáveis aqui importa na hora de verificar o resultado final. Selecione sempre a variável independente e depois a dependente para esse tipo de gráfico. A seguir, em **View/Graph...** selecione o gráfico tipo **Scatter** e em **Fit lines** escolha **Regression Line**. Isso irá adicionar uma linha de regressão entre as duas variáveis. Depois, para mostrar o resultado da linha de regressão clique em **Options** e, em **Legend Labels** selecione **Detailed**. Por fim, em **Axis borders** selecione **Kernel density** para termos a informação da distribuição de kernel para cada um dos dados. O gráfico resultante irá indicar a relação positiva entre os dois conjuntos de dados e, em cada eixo, a estimativa da distribuição de kernel para cada um desses conjuntos. Também será mostrado o resultado da equação de regressão simples.

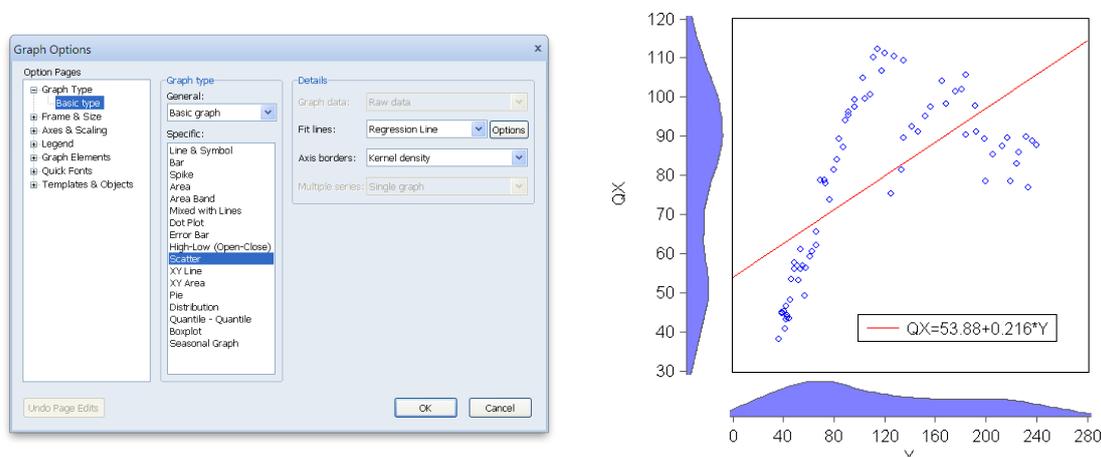


Figura 2.10: Scatter e linha de regressão entre  $qx$  e  $y$

**Programação 2.0.9** Para fazer um gráfico do tipo scatter plot entre duas variáveis, inserindo uma curva de regressão e mostrando o resultado da equação de regressão, devemos primeiro fazer o grupo com as variáveis de interesse e depois pedir o gráfico. Por fim, usamos a opção kernel para mostrar a distribuição de kernel nos eixos:

```
group g1 y qx
g1.scats(ab=kernel) linefit(leg=det)
```

Alternativamente, pode-se estar interessado em ver a relação de todas as variáveis em pares. Nesse caso, selecione todas as séries “ $qx$ ,  $y$ ,  $px$ ,  $pm$ ,  $qm$ ” e abra como grupo. A seguir, em **View/Graph...** escolha **Scatter**, em **Fit lines** selecione **Regression Line**, e em **Multiple series** selecione **Lower triangular matrix** (é uma matriz simétrica). O *EViews*<sup>®</sup> irá retornar a relação em par de todas as variáveis.

**Programação 2.0.10** Para fazer um gráfico do tipo scatter plot entre diversas variáveis, inse-

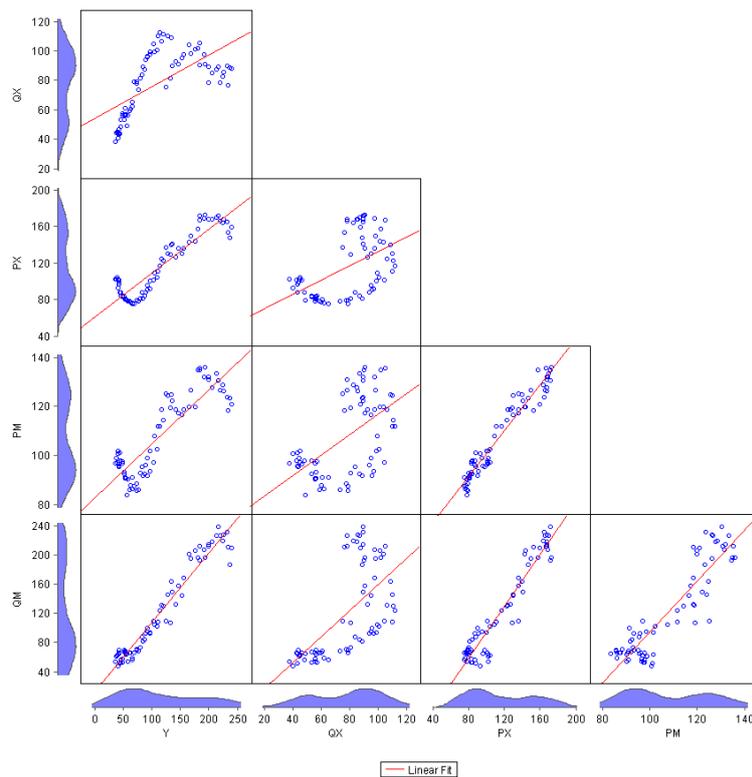


Figura 2.11: Scatter e linha de regressão entre todas as variáveis

rindo uma curva de regressão, devemos primeiro fazer o grupo com as variáveis de interesse e depois pedir o gráfico. Por fim, usamos a opção kernel para mostrar a distribuição de kernel nos eixos. O comando `m` especifica que são múltiplos gráficos. O comando `mult=l` especifica que é uma matriz de gráficos triangular inferior.

```
group g1 y qx px pm qm
g1.scat(m, mult=l, ab=kernel) linefit(leg=det)
```

## 2.1 Dados Categóricos

O formato de dados conhecido como categóricos é muito comum na investigação em economia. Podemos citar, por exemplo, o uso de microdados da PNAD/IBGE, onde temos informações de indivíduos com suas respectivas características como idade, cor, sexo, situação matrimonial, salário e etc. Os gráficos que são feitos considerando dados categóricos são diferentes daqueles utilizados em séries de tempo.

Para ilustrar o uso de gráficos com dados categóricos usamos os dados de exemplo do *EViews*® *gulfcoast.wfl*. Nesse estão informações sobre demografia de distritos localizados em uma região dos EUA. São 234 informações com 117 distritos, cada qual com duas informações em dois momentos do tempo. Os dados estão organizados no formato **Unstructured/Undated**. São quatro series: população em 1.000 para cada distrito, `pdiff`, `pop_previous` e `year`. Como são dois momentos no tempo, a organização dos dados segue uma lógica de primeiro mostrar os 117 resultados para o ano de 2005 e depois os 117 resultados para o ano de 2006. Note que são criados identificadores para os indivíduos. O `County_code` mostra o código de cada município, `County_name` o nome dos

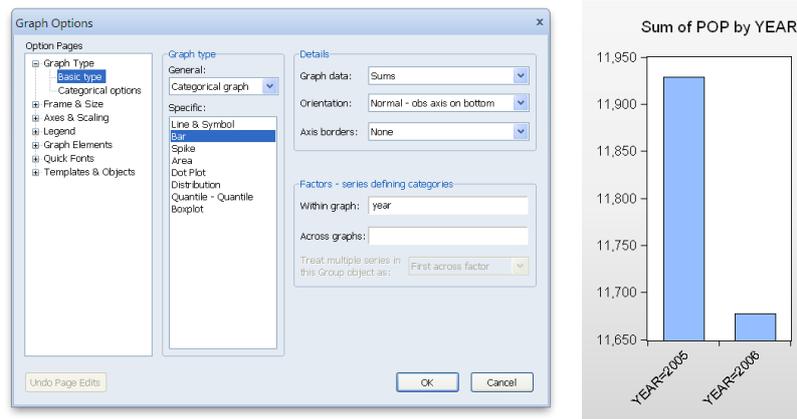


Figura 2.12: Dados categóricos - soma da população por ano

municípios, que se repetem a partir da observação de número 118. O `id` mostra o estado associado ao distrito; `state_code` o código do estado e `state_name` o nome do estado.

Vejamos como seria um gráfico que compara o total da população dos 117 distritos em cada um dos dois anos analisados. Abra a série `pop` e depois vá em **View/Graph...** e mude para a opção **Categorical graph**. A seguir selecione a opção **Bar**, para fazermos um gráfico de barras, e depois em **Details** use **Sums**, para termos a soma da população em cada um dos anos. Note no detalhe que especificamos na opção **Within graph** a série `year`. Isso irá fazer com que o programa entenda que há dois momentos no banco de dados.

## 2.2 Exemplos de programas.prg

Com os conhecimentos adquiridos neste capítulo somos capazes de criar programas para formatar nossos gráficos de uma mesma maneira, padronizando trabalhos de forma fácil. Para criar um programa clique em **File**, então **New** e **Program**.

**Programação 2.2.1** Com o `exemplo1.wfl` aberto, o programa abaixo irá gerar um gráfico de linha, na cor preta, para cada uma das séries padronizadas adicionando uma linha pontilhada, na média zero, e redimensionará o tamanho.

```
for %a y qx px pm qm
graph g{%a}.line(n) {%a}
g{%a}.setelem linecolor(0,0,0)
g{%a}.draw(dashline, left, rgb(172,172,172)) 0
g{%a}.options size(6,2)
show g{%a}
next
```

**Programação 2.2.2** Abaixo a sequência de comandos que utilizamos para abrir automaticamente o `exemplo1.wfl`, criar um gráfico com as séries `qx` e `px`, uma em cada eixo, com as respectivas funções de distribuição. Além de pintar na cor vermelha `qx` e `px` em azul, adicionar a barra cinza vertical, entre 2008Q2 e 2009Q1, e adicionar o título ao gráfico.

```
%path = @runpath
```

```

cd %path
load exemplo1.wf1
graph gqx.line(ab=hist,d) qx px
gqx.setelem(1) linecolor(255,0,0)
gqx.setelem(2) linecolor(132,112,255)
gqx.draw(shade,bottom,color(gray)) 2008Q2 2009Q1
gqx.addtext(pt=12,face="ubuntu light",t) "Series QX e PX"
show gqx

```

Com a utilização de sub-rotinas podemos sofisticar nossos programas. A criação destas é feita da mesma forma que um programa, **File/New/Program**. Para "chamar" uma sub-rotina dentro de um programa é necessário especificar o caminho exato da mesma. Caso o caminho inteiro não seja especificado o programa.prg deve estar salvo dentro do mesmo diretório da sub-rotina a ser executada.

**Programação 2.2.3** A sub-rotina `sub_recessoescodace.prg` descrita a seguir destaca as recessões do ciclo de negócios brasileiro datado pelo Comitê de Datação de Ciclos Econômicos (CODACE), em 30 de Julho de 2015.

```

subroutine recessoescodace(graph g1)
g1.draw(shade,bottom) 1981Q1 1983Q1
g1.draw(shade,bottom) 1987Q3 1988Q4
g1.draw(shade,bottom) 1989Q3 1992Q1
g1.draw(shade,bottom) 1995Q2 1995Q3
g1.draw(shade,bottom) 1998Q1 1999Q1
g1.draw(shade,bottom) 2001Q2 2001Q4
g1.draw(shade,bottom) 2003Q1 2003Q2
g1.draw(shade,bottom) 2008Q4 2009Q1
g1.draw(shade,bottom, color(255,100,100)) 2014Q2 2015Q2
endsub

```

Com o `exemplo2.wf1` aberto rode o programa `prog_recessoescodace.prg`, descrito abaixo. Esse utiliza da sub-rotina `sub_recessoescodace.prg` e por isso ambos devem ser salvos na mesma pasta antes da execução.

```

include sub_recessoescodace.prg 'Arquivo com a subrotina CODACE
graph gpx.line(d) pib x ' Cria o gráfico gpx com duas escalas
gpx.setelem(1) legend(PIB Brasil) ' Adiciona legenda da série 1
gpx.setelem(2) legend(Exportações Brasil) ' Adiciona legenda da série 2
gpx.setelem(2) linecolor(0,0,0) ' Altera cor da linha da série 2
' Chama subrotina para marcar as recessoes segundo CODACE
call recessoescodace(gpx)
show gpx 'Apresenta gráfico gpx na tela

```

Com base nos programas apresentados acima inclua a sub-rotina `sub_recessoescodace.prg` ao programa 2.2.2. Destacando as recessões do ciclo de negócios brasileiro datado pelo CODACE nos gráficos de todas as séries do `exemplo1.wf1`.



### 3. Funções de Distribuição

O *EViews*<sup>®</sup> permite a construção de diversas curvas de distribuição, que podem tanto serem discretas quanto contínuas. As mais utilizadas em testes de econometria são as funções normal, t-student, log-normal, F e qui-quadrado, que aqui iremos ilustrar<sup>1</sup>.

Ao trabalhar com funções de distribuição, devemos compreender dois pontos importantes. O primeiro é se a variável em questão é categórica ou numérica e, o segundo, as diferenças que existem entre uma função de probabilidade, ou densidade, uma distribuição cumulativa e uma distribuição inversa, que é a inversa da função cumulativa. As variáveis categóricas são fáceis de identificar. Ao aplicar um questionário com perguntas que contenham respostas como do tipo, sexo, nacionalidade e etc, obtemos como resposta características e não números. Essa classificação será importante para definir que tipo de teste irá usar para avaliar os resultados. Por exemplo, se perguntarmos o sexo dos entrevistados, temos respostas categóricas como homem ou mulher. Por outro lado, se perguntarmos a idade teremos respostas numéricas. Essas podem tanto serem discretas, ou seja, 25 anos, 35 anos, ou contínuas, expressando a idade inclusive em minutos, 13.140.325 minutos de vida.

A **função de densidade** representa a distribuição de probabilidade de uma variável aleatória. É como a probabilidade irá se comportar de acordo com os valores que essa variável aleatória irá assumir. É comum não conhecermos a função de densidade que irá representar o nosso conjunto de dados. Por isso que fazemos testes para ver se os nossos dados possuem uma distribuição que pode ser aproximada, por exemplo, por uma curva normal, uma curva t-student, uma curva F ou qualquer outra. Dada a nossa função de densidade, toda a área abaixo da curva deverá somar 1, que é a probabilidade da variável assumir qualquer valor. No *EViews*<sup>®</sup>, supondo uma curva normal, a função densidade é utilizada a partir do comando `@dnorm()`, onde dentro do parênteses podemos colocar os valores do banco de dados<sup>2</sup>. A função de densidade pode ser determinada fazendo a derivada da função de distribuição cumulativa. Em termos matemáticos uma função densidade de  $x$

<sup>1</sup>Há diversas outras distribuições contínuas em estatística como a Beta, de Cauchy, Exponencial, Gamma, Gumbel, Logística, Uniforme e de Weibull. Dentre as distribuições contínuas, destaque para a Binomial, Geométrica, Hipergeométrica, Multinomial e de Poisson.

<sup>2</sup>Os códigos das diferentes funções de densidade no *EViews* são precedidos da letra “d”. Por exemplo: `@dlogistic()`; `@dpareto()`; `@dpoisson()`; `@dtdist()`; `@dunif()`.

é representada a partir de  $F_x(x) = \mathbb{P}(a \leq x \leq b) = \int_a^b f(x)dx$ .

Descoberta a função de densidade, podemos usar a **distribuição cumulativa**. Esta irá determinar o quanto da curva, ou da probabilidade, existe até determinado valor que se queira avaliar. Para o exemplo de uma curva normal, podemos encontrar qual a probabilidade de se ter um valor menor que “x”, por exemplo. Esse é dado por toda a área abaixo da curva e que é inferior ao ponto “x”. O conceito de distribuição cumulativa é muito importante para os propósitos do entendimento da econometria e em testes de hipótese, pois usamos esse conceito para encontrar o p-valor nos testes. Para encontrar a resposta na distribuição cumulativa, especificamos o ponto da curva que se queira e encontramos a área (probabilidade) até esse ponto. A função do *EViews*<sup>®</sup> que iremos utilizar para a distribuição cumulativa, para o caso de termos uma distribuição normal, é a `@cnorm()`, e a área escura mostrada na figura 3.1 é a área resultante<sup>3</sup>. Mais a frente, ao estudarmos sobre os testes de hipótese, ficará claro que a área dada por  $1 - @cnorm$  representa o p-valor, ou como é comumente escrito, *probability*. Em termos matemáticos a representação da função de distribuição cumulativa é dada por:  $F(z) = \mathbb{P}(z \leq x)$ . No exemplo da Figura 3.1 a área dada por `@cnorm(x)` pode ser representada a partir de  $F(x) = \int_{-\infty}^x f(x)dx$ .

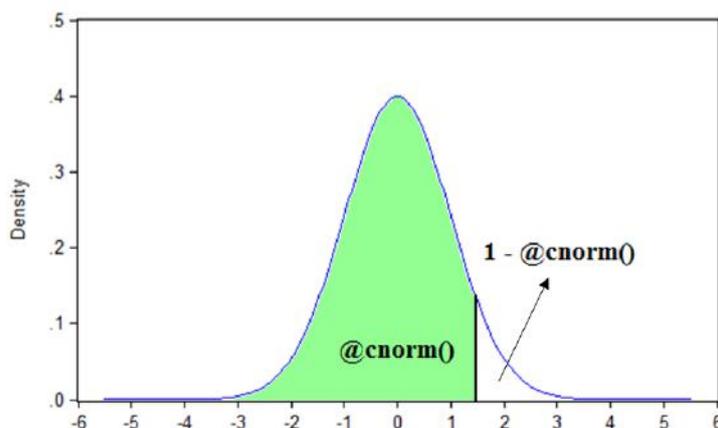


Figura 3.1: Distribuição Cumulativa

Por fim, a **distribuição inversa** irá representar a área da curva que é complementar à função de distribuição cumulativa. Agora fazemos o procedimento inverso da cumulativa, especificamos a área desejada e, com isso, obtemos o ponto na curva que representa essa área. Aqui, temos conhecimento da área da curva que estamos avaliando e queremos encontrar o ponto associado. No *EViews*<sup>®</sup> o comando utilizado para a distribuição inversa, para o exemplo de uma curva normal, é `@qnorm()`.

Todas essas três formas de avaliar uma função de distribuição estão disponíveis no *EViews*<sup>®</sup> e serão aplicadas a diferentes formas de distribuição a seguir. Nesse caso, para cada uma das opções de uma distribuição o *EViews*<sup>®</sup> fornece códigos diferentes. Por exemplo, para uma função de distribuição cumulativa, também denominada de CDF, usa-se o comando `@c`. Para uma função de probabilidade (densidade), usa-se `@d` e, por fim, para uma função inversa, `@q`. Também é possível criar funções de distribuição aleatórias a partir do comando `@r`, que gera números aleatórios. Veremos isso nas aplicações para as diferentes distribuições a serem analisadas nos tópicos a seguir.

<sup>3</sup>De maneira análoga ao visto, na função de densidade cumulativa é precedida da letra “c” nos comandos do *EViews*<sup>®</sup>. Por exemplo: `@clognorm()`; `@cpareto()`; `@cpoisson()`; `@ctdist()`.

### 3.1 A Curva Normal

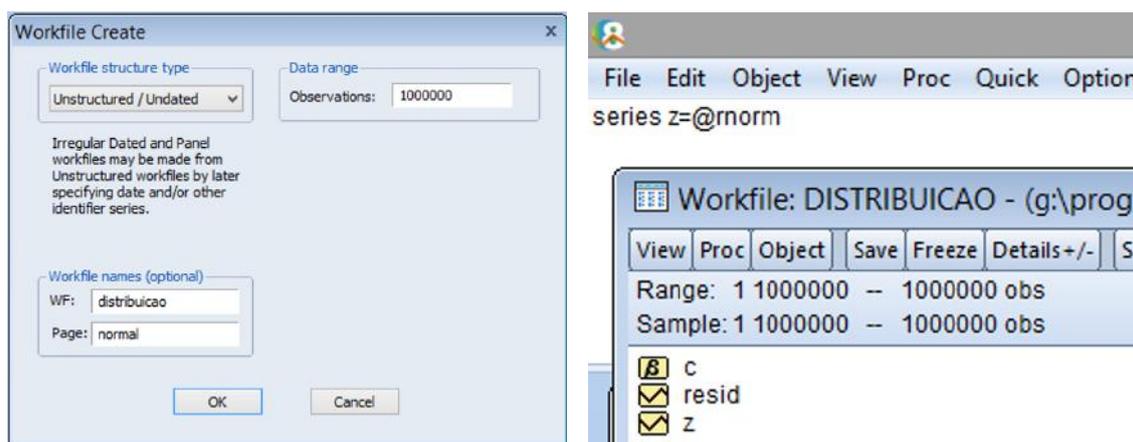
Essa é uma das mais importantes e também mais usadas funções de distribuição de probabilidade, também denominada de curva de Gauss. Suponha uma variável aleatória  $X$  com  $n$  dados. Se estamos assumindo que essa variável tem uma distribuição normal, podemos determinar cada ponto dessa curva a partir da equação:

$$z = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3.1)$$

Onde  $\mu$  é a média da variável aleatória  $X$ ,  $\sigma$  é o seu respectivo desvio-padrão e  $x$  o ponto a ser avaliado. Um caso particular e muito útil dessa curva é a “normal padrão”. Nesta, a média é zero e o desvio-padrão 1. Destaca-se que mesmo que nossa variável  $X$  não tenha média igual a zero e desvio-padrão 1 podemos converter os mesmos para essas medidas, no que se denomina de padronização.

Como forma de ilustrar o uso de funções de distribuição, vamos criar um arquivo com 1 milhão de dados aleatórios. Abra o *EViews*® e clique em **Create a New *EViews*® workfile**. A seguir, escolha uma estrutura tal como mostrado na figura 3.2a, digite 1000000 para especificar o número de observações que iremos usar e dê um nome para o WF (workfile) e a página. A partir de 3.1, podeos ver que uma curva norma padrão é representada por:

$$z = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (3.2)$$



(a) Distribuição

(b) Com densidade de Karna

Figura 3.2: Alterando o intervalo

Você pode modificar a qualquer momento o tamanho desse banco de dados, basta clicar duas vezes em “Range” e aumentar ou diminuir o intervalo. Note que, nesse momento, não há nenhuma informação, ou seja, nenhum dado associado. Como primeiro passo, vamos simular uma variável aleatória que tenha 1 milhão de dados definindo que a mesma tenha uma distribuição normal. Para fazer isso vamos usar o comando `@rnorm` como mostrado no box de programação.

**Programação 3.1.1** Podemos gerar números aleatórios no *EViews*® de várias formas. Para criar um arquivo do *EViews*® com dados inteiros no total de 1 milhão, ou seja, uma serie com 1000000 linhas, usamos o comando abaixo no arquivo:

```
rndseed 10
series z=@rnorm
```

Dica : Muitas vezes é melhor usar o conceito de series do que vector.

Ao iniciar os comandos descritos no box programação, determinamos a série aleatória utilizada com o comando `rndseed 10` e criamos uma série denominada `z` de 1 milhão de dados aleatórios com o comando `@rnorm`. Ao repetir esse procedimento sem aplicar `rndseed 10` ou utilizando qualquer outro gerador aleatório (`rndseed 1`, por exemplo), a sequência de dados irá diferir a cada momento. Porém, como especificamos que os dados seguem uma distribuição normal padrão a partir de `norm`, sempre que simular um novo conjunto de informações, ela terá a mesma distribuição.

Para confirmar, faça um gráfico de distribuição dos nossos dados. Abra a série `z`, vá em **View /Graph ...**, em tipo de gráfico selecione **distribution** e depois clique em **ok**. A seguir, adicione uma estimativa da curva a partir da densidade de kernel. Dica: com a opção gráfico aberta vá em **details** e crie um gráfico personalizado **custom**.

Outra contribuição interessante para visualizar é comparar nosso conjunto de dados com uma distribuição normal teórica, ou seja, uma curva normal que seja criada a partir da função. Com o gráfico aberto clique em **options**, a seguir, do lado direito, em **options** novamente. Depois em **add e theoretical density**. Vamos escolher primeiro uma curva normal e clique em **ok**. Note que a mesma fica praticamente imperceptível, uma vez que a curva teórica se mistura com a curva estimada pela densidade de kernel (Figura 3.3b).

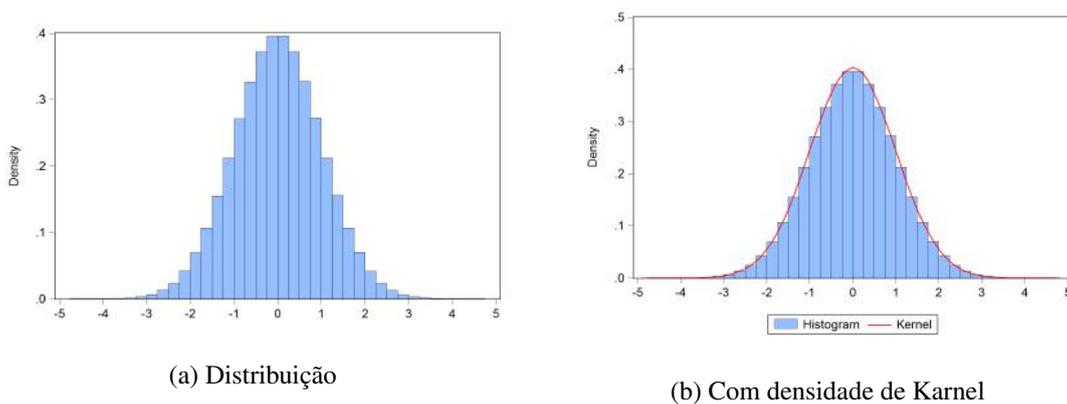


Figura 3.3: Distribuição Normal

**Programação 3.1.2** Para fazer um gráfico que tenha o histograma de uma série e mais duas curvas teóricas com diferentes valores para a média, podemos usar o comando abaixo. O termo `p1=1` representa a média=1:

```
z.distplot hist theory(dist=normal,p1=1) theory(dist=normal,p1=2)
```

Para fazer o mesmo gráfico, mas com diferentes valores para o desvio-padrão, especificando três diferentes curvas, que é o segundo parâmetro na curva normal, usamos:

```
z.distplot hist theory(dist=normal,p2=1)
theory(dist=normal,p2=2) theory(dist=normal,p2=3)
```

Podemos mudar os parâmetros dessa densidade teórica para que ela fique mais nítida. Repita os passos a seguir e, em **theoretical density** especifique média = 1 e desvio padrão = 1. Note que agora a curva de cor verde se desloca para a direita na Figura 3.4a.

Esse procedimento pode ser repetido para diferentes valores de média e desvio padrão e, dessa

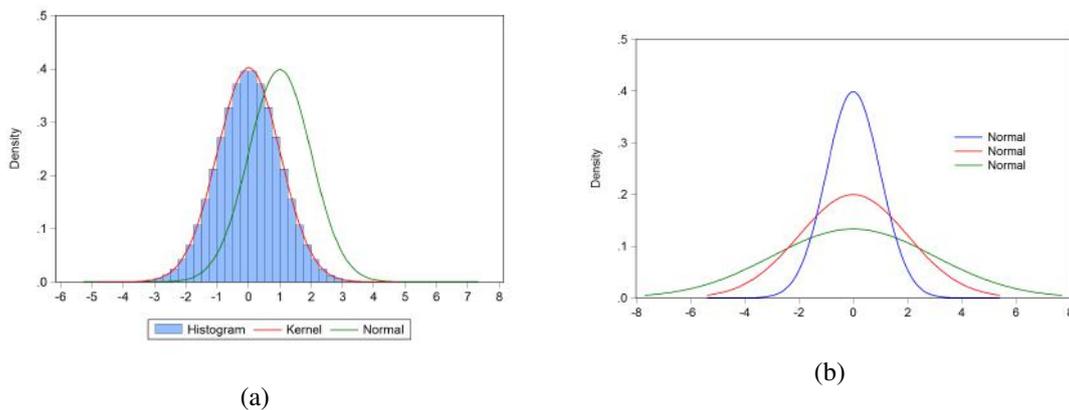


Figura 3.4: Alterando a média e o desvio-padrão

forma, podemos encontrar diferentes formatos para a curva normal. Para exemplificar isso, com o gráfico aberto clique em **options**. A seguir, em options novamente e, do lado esquerdo, apague os gráficos **histogram** e **kernel**. Acrescente mais duas curvas normais teóricas. No total, teremos três curvas (figura 3.4b). Agora, deixe todas com média igual a zero e faça para a primeira curva desvio padrão igual a 1, depois para a segunda um desvio padrão igual a 2 e, para a terceira curva, desvio padrão igual a 3. Clique em ok e você irá gerar o gráfico da Figura 3.4b.

Ao especificar diferentes valores para o desvio-padrão na curva, estamos determinando o que se conhece como curtose. Note que, para a curva azul no gráfico 3.4b, temos uma maior concentração de dados em torno da média e, na curva verde, mais achatada, os dados são mais espalhados. Iremos ver como obter o resultado estatístico da curtose a partir da média e do desvio padrão no próximo capítulo. Mas, o leitor já pode ir se familiarizando com o formato de uma distribuição de dados com diferentes desvios em torno da média.

**Programação 3.1.3** Para avaliar a função de densidade em um ponto qualquer da nossa função de distribuição podemos usar o comando “d” antes da distribuição que está sendo avaliada. Para o caso de uma distribuição normal, com média 0 e desvio-padrão 1, usamos:

```
scalar r=@dnorm(0)
```

Aqui, o comando scalar cria a caixa de nome “r” para receber o valor da distribuição. A seguir, especificamos d, para determinar que queremos a função de densidade, seguido de norm, que é a curva normal com média 0 e desvio-padrão 1 e, por fim, o valor “0” entre parênteses especifica que estamos avaliando a densidade naquele valor.

Agora que já sabemos como gerar uma curva normal aleatoriamente, vamos testar outras opções. Suponha que se queira um conjunto de dados que segue determinados parâmetros, por exemplo, média igual a 0 e desvio-padrão igual a 1. Nesse caso, podemos criar a série x usando apenas o comando nrnd.

Por outro lado, se queremos especificar uma média diferente, como por exemplo, 100 e variância igual a 22 o melhor é usar uma equação. Nesse caso, criamos a série y e o comando @sqrt representa a raiz de 22, que seria o desvio-padrão. A seguir, multiplicamos esse por uma série gerada aleatoriamente com distribuição normal.

**Programação 3.1.4** Também podemos gerar uma série de dados que segue uma distribuição

normal com média zero e desvio-padrão igual a 1 usando o comando `nrnd`:

```
Series x=nrnd
```

Alternativamente, para gerar uma série de dados que tem média igual a 100 e variância igual a 22, usa-se:

```
Series y=100+@sqrt(22)*nrnd
```

O comando que especifica uma distribuição inversa também pode ser utilizado para gerar uma sequência de números aleatórios porém, partindo de probabilidades. Vamos escolher a distribuição normal para exemplificar, criando uma série de nome `t`, e usando o comando “`q`”.

**Programação 3.1.5** Por fim, podemos gerar dados com distribuição, como, por exemplo, uma normal, com média zero e variância igual 1 usando uma função inversa. Para tanto, usamos o termo `q` que representa que estamos construindo uma função quantílica, ou seja, a inversa da função de distribuição cumulativa. O termo `rnd` é especificado para o parâmetro de probabilidade. Esse tem que ser entre 0 e 1. Nesse caso, ao colocar `rnd` construímos a curva normal a partir de diversos valores aleatórios para a probabilidade.

```
series t=@qnorm(rnd)
```

O comando `q` antes da especificação da curva também é útil para determinar o ponto da curva que é associado a uma determinada área. Para o exemplo de uma curva normal padrão sabemos que o ponto 0, que representa a média dos dados, divide a área em duas partes iguais, 50% antes e 50% depois. Se usarmos `scalar a=@qnorm(0.5)` encontraremos o valor 0, ou seja, o ponto  $a = 0$  representa 50% da curva acumulada. Teste `scalar a=@qnorm(0.025)`, que é uma área de 2,5%. O resultado será -1,959, ou seja, o ponto no qual a área a esquerda de  $x$  representa 2,5% do total.

O que está dizendo esse comando? Primeiro que a função utilizada `qnorm(·)` irá retornar um valor. Sendo assim, especificamos a como um escalar, exatamente porque irá receber um número. Em segundo lugar, o valor 0.5 representa uma probabilidade de 50% que será aplicada à função normal. Nesse caso, queremos saber qual é o valor na curva normal que irá resultar em uma área de 50%. Essa área é especificada como toda a área a esquerda do valor.

Agora, se estamos interessados em saber qual é o valor associado a uma curva normal padrão que irá determinar 95% da área, como podemos proceder? Usamos `scalar a = @qnorm(0.95)` o que irá retornar o valor 1,644854.

A informação sobre a função inversa é similar ao que obtemos ao usar a função cumulativa. Porém, enquanto que na função inversa usando o comando `@q` especificamos a área e obtemos o ponto, com a função cumulativa a partir de `@c` especificamos o ponto e obtemos a área.

**Exercício 3.1** Encontre a área entre dois pontos de curva normal padrão que preencha entre  $\pm 2,05$  desvios padrão. ■

**Exercício 3.2** Determine o formato de diferentes curvas normais variando apenas o desvio padrão. Para uma média igual a zero, use os seguintes valores para os desvios padrão: curva 1: 1,3; curva 2: 2,1; curva 3: 2,9. ■

Nesse momento podemos inserir os conceitos de “quantis”. Seja a curva normal padrão, imagine que se queira dividir sua área em 4 partes iguais. O que queremos obter aqui é o quantil de uma distribuição normal padrão. Nesse caso, quais seriam os respectivos pontos que permitem ter, em

cada quantil, 25% da área da curva normal? Isso pode facilmente ser obtido usando o comando `scalar quantil = @qnorm()` como valores 0.25; 0.5; 0.75 o que irá retornar os pontos -0,67; 0; 0,67 respectivamente. Assim, entre  $-\infty$  e -0,67 há 25% da área de uma curva normal padrão. Entre (-0,67; 0) há 25%, entre (0; 0,67) outros 25% e entre (0,67;  $\infty$ ) tem 25%.

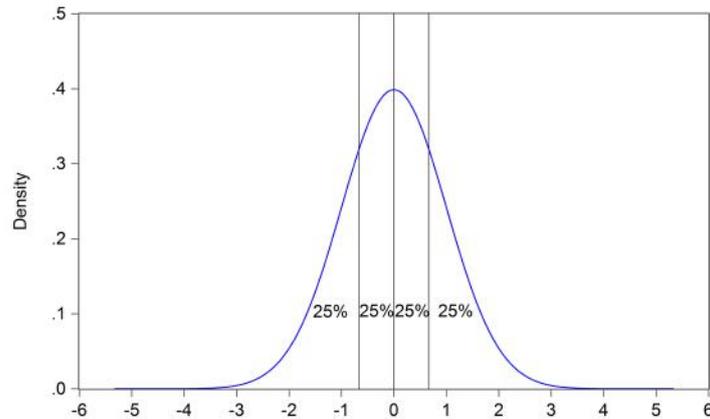
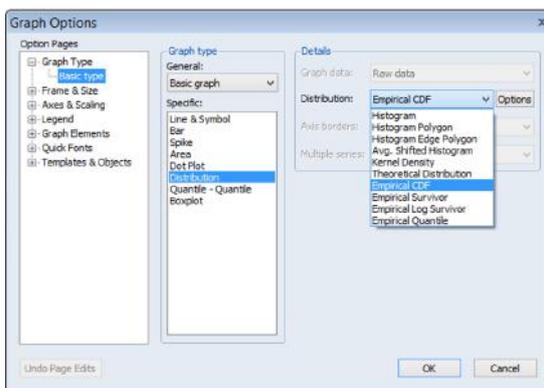


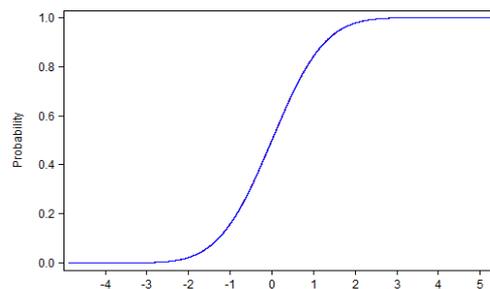
Figura 3.5: Divisão dos quantis da distribuição normal

Na estatística denominamos de tercís a divisão da área em 3 quantis; de quintis a divisão em 5 quantis; de decís a divisão em 10 quantis e de percentis a divisão em 100 quantis. Há diversas aplicações para os conceitos de quantis, sendo as mais comuns em análises de distribuição de renda e o uso da regressão quantílica.

Continuando com o nosso exemplo da distribuição normal, podemos especificar um gráfico que irá representar essa distribuição. Para tanto, abra a série z, a seguir em **view/graph** selecione **distribution** e depois **empirical CDF**.



(a)



(b)

Figura 3.6: Distribuição Cumulativa com dados normais (CDF)

Como apontado na introdução deste capítulo, em estatística, a distribuição cumulativa representa a probabilidade de se observar um valor de uma série de dados que não excede determinado valor específico. Esse cálculo pode ser representado a partir de:

$$F(z) = \mathbb{P}(z \leq r)$$

onde  $F(z)$  é a área da curva acumulada até o ponto  $r$ , ou seja, a estatística  $F(z)$  representa a função cumulativa. No exemplo da curva normal, temos que 50% dos dados se encontram abaixo da média e 50% acima. Como a média é zero para uma curva normal padrão então, a probabilidade acumulada até o valor “0” é 50% ou então, expresso de outra forma:  $F(z) = \mathbb{P}(z \leq 0) = 0,5$

**Programação 3.1.6** Usando como exemplo a nossa curva normal com média 0 e variância unitária, sabemos que o valor 0 divide ao meio a função de distribuição, colocando 50% da área para cada lado da distribuição. Nesse caso, isso pode ser verificado a partir de um comando do *EViews*<sup>®</sup> que usa o valor para encontrar a área a partir de:

```
Scalar r=@cnorm(0)
```

Aqui, primeiro criamos um scalar de nome  $r$  e que irá receber o valor da função. A seguir, o comando  $c$  usado antes da especificação da curva normal  $norm$  serve para determinar que estamos avaliando a função CDF – cumulativa. Por fim, o valor “0” entre parênteses significa que queremos avaliar a probabilidade de um valor não exceder o valor “0”. Isso irá retornar o valor 0,5. Ou seja, o total da distribuição acumulada até o valor 0 é de 50%.

Também podemos determinar a probabilidade associada a um valor mínimo especificado. Para tanto usamos a chamada *empirical survivor*. Com a série de dados  $z$  aberta, vá em **view/graph** e depois selecione **distribution** e, em **details, empirical survivor**. Note que o gráfico (figura 3.7) representa exatamente o inverso do gráfico da distribuição cumulativa. Sendo assim, a probabilidade de que um valor seja maior que 5, por exemplo, é quase 0%. Por outro lado, a probabilidade de que um valor seja maior que “0”, que é a média dos dados, é de 50%. Expresso de outra forma, como a área total da curva é 100% e a função cumulativa nos fornece a área até certo ponto, podemos usar o comando abaixo para especificar a área à direita de um ponto:

$$S(z) = 1 - F(z) = \mathbb{P}(z > 5) = 0$$

$$S(z) = 1 - F(z) = \mathbb{P}(z > 0) = 0,5$$

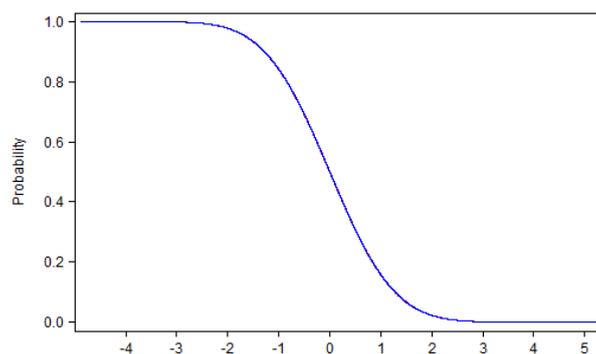


Figura 3.7: *Empirical Survivor*

Além de encontrar a área acumulada até um ponto ou acima de um determinado ponto, é muito comum querermos saber qual é a área definida entre dois pontos. Isso será útil para o entendimento de testes de hipóteses e construção de intervalos de confiança. Por exemplo, como podemos saber a área de uma curva normal entre  $z = -1$  e  $z = 1$ ? No box de programação 3.1.7 explicamos como encontrar essa área.

**Programação 3.1.7** Para conseguir determinar a área entre dois pontos da curva, podemos combinar duas funções cumulativas. Primeiro determinamos a área até o ponto maior e, depois, retiramos a área até o ponto menor.

Considerando uma curva normal padrão, vamos avaliar a área entre -1 e 1 desviopadrão usando o comando a seguir:

```
scalar area=@cnorm(1)-@cnorm(-1)
```

Esse irá retornar o valor de 0,682, que é o mesmo que dizer que 68,2% dos dados estão entre -1 e 1. Além desse, um intervalo muito utilizado é de  $z = \pm 2$  e também  $\pm 3$ . Esses podem ser encontrados apenas mudando o valor entre parênteses do comando acima.

No início desse tópico aprendemos a gerar uma série de números aleatórios usando o comando `vector` e dando o nome `z` para esse vetor. Porém, muitas vezes é útil que se tenha uma matriz de números aleatórios, ou seja, diversos vetores. Isso pode ser gerado no *EViews*<sup>®</sup> de forma simples usando o comando `matrix`, ao invés de criar um `scalar`, especificando `matrix`.

**Programação 3.1.8** A seguir, podemos criar uma matriz de números aleatórios que seguem uma distribuição normal, usando os comandos mostrados abaixo. Para uma matriz de 1.000.000 linhas e 30 colunas, usamos:

```
matrix b=@mnrnd(1000000,30)
```

Até esse ponto ilustramos o uso da curva normal considerando que a média é zero e o desvio-padrão 1, porém, o mais comum em investigações estatísticas é que os dados possuem média diferente de 0 e desvio padrão diferente de 1. Não se preocupe se seu banco de dados não possuir essa característica, isso é fácil de ser contornado a partir da padronização dos dados. Nesse caso, transformamos a distribuição de nossos dados que podem ter qualquer média e desvio padrão, em uma distribuição que tenha média=0 e desvio padrão=1. Isso é feito facilmente a partir de:

$$z = \frac{x - \bar{x}}{\sigma}$$

Onde  $z$  é o novo valor,  $x$  é o valor da série original,  $\bar{x}$  é a média dos dados e  $\sigma$  é o desvio padrão dos dados. Isso pode ser feito no *EViews*<sup>®</sup> especificando um comando.

**Programação 3.1.9** Suponha que tenhamos um conjunto de dados com média 35 e variância de 3,5. Podemos gerar esses dados utilizando:

```
Series n=35+@sqrt(3.5)*nrnd
```

Podemos transformar essa distribuição em média 0 e desvio padrão 1 usando o seguinte comando:

```
series y=(x-@mean(x))/@stdev(x)
```

Para o nosso exemplo, onde a série `n` tem média 35 e desvio-padrão de `@sqrt(3.5)`, fazemos:

```
Series n1=(n-35)/@sqrt(3.5)
```

Agora que aprendemos os comandos que especificam a densidade, a função cumulativa e a inversa de uma curva normal, podemos explorar um pouco o comportamento de outras funções que são muito úteis em econometria e testes estatísticos.

### 3.2 A curva *t-student*

A função de distribuição mais utilizada em testes de hipótese é a *t-student*, criada por William Sealy Gosset que acabou adotando o nome de *student* para representar a função. É uma distribuição simétrica, como a curva normal, mas possui caldas mais largas, o que a torna mais útil para representar distribuição de dados com valores extremos, como é comum não conhecermos a variância da população que estamos analisando, não podemos usar a curva normal. É aqui que a curva *t-student* se torna interessante e útil. Um parâmetro importante na curva *t-student* é o  $v = \text{graus de liberdade}$ . Quanto maior for seu valor, mais a curva *t-student* irá se aproximar da curva normal. Mas, o que significa os graus de liberdade? Suponha que temos um teste de laboratório a ser feito e coletamos uma amostra de 80 informações. Nesse caso, temos que  $v = n - 1$  ou seja  $v = 79$  graus de liberdade. Por isso que dizemos que quanto maior for o número de graus de liberdade da distribuição *t-student*, mais ela se aproxima da curva normal. Ou seja, quanto maior for a amostra “n”, maior será o valor de “v”.

Na Figura 3.8 estão simuladas uma curva normal e várias curvas *t-student* com diferentes graus de liberdade com  $v=2$ ,  $v=5$  e  $v=10$ . Note que, na medida em que esse parâmetro aumenta, a curva *t-student* vai se tornando mais próxima da curva normal, tornando a diferença entre elas quase imperceptível.

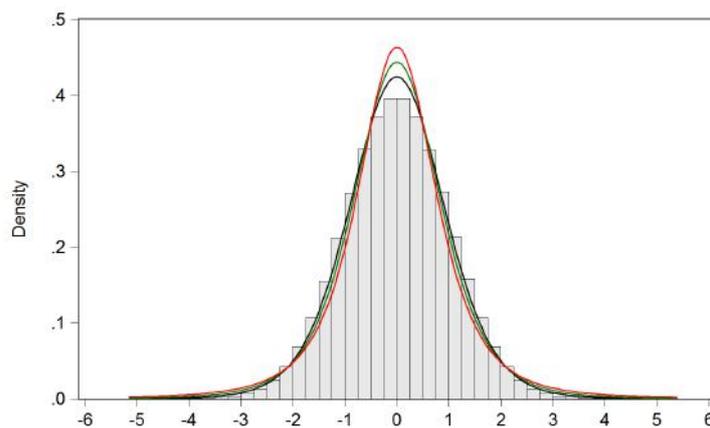


Figura 3.8: Curva *t-student* e curva normal padrão

Suponha que temos uma série de dados  $Z$  com distribuição normal padrão e um outro conjunto de dados  $Q(20)$  com distribuição  $q$ -quadrado com 20 graus de liberdade (veremos essa curva mais a frente). Além disso, suponha que  $Z$  e  $Q$  são séries de dados independentes. Se dividirmos uma série pela outra, teremos um conjunto de dados resultante com uma distribuição *t-student* com 20 graus de liberdade. Na fórmula abaixo o parâmetro  $v$  representa os graus de liberdade.

$$t_{(v)} = \frac{z}{\sqrt{\frac{Q_{(v)}}{v}}}$$

Para montar isso vamos especificar  $z = \text{rnorm}$  e  $q = \text{@qchisq}(\text{rnd}, 20)$  a seguir use o comando `series zq = z/@sqrt(q, 20)` e compare com uma curva encontrada a partir de `series t = @rtdist(20)`.

**Programação 3.2.1** Para criar uma variável aleatória que tenha distribuição *t-student*, usamos o comando abaixo. Note que há um parâmetro adicional a ser especificado,  $v$ , que representa os

graus de liberdade da curva *t-student*.

```
series z=@rtdist(v)
```

Tal qual na curva normal também podemos usar aqui o comando que especifica uma distribuição inversa para gerar uma sequência de números aleatórios. Além de ser útil para gerar uma curva qualquer, o comando `q` é útil para determinar o ponto da curva que é associado a uma determinada área. Para o exemplo de uma distribuição *t-student* a média dos dados divide a área em duas partes iguais, 50% antes e 50% depois. Se usarmos o termo `scalar a=@qtdist(0.5,50)` encontraremos o valor 0. Note que aqui não faz diferença os graus de liberdade, a média sempre irá dividir a área ao meio. Teste `scalar a=@qtdist(0.025,50)`, que é uma área de 2,5%. O resultado será -2,008, ou seja, o ponto no qual a área a esquerda representa 2,5% do total para uma curva *t-student* com 50 graus de liberdade. Esse resultado para uma curva normal seria -1,959 comprovando que a curva *t-student* é útil para representar dados com valores extremos.

**Programação 3.2.2** Para gerar uma distribuição de dados *t-student* também podemos recorrer a função inversa usando o termo `q`, a inversa da função de distribuição cumulativa. Como esse comando usa uma área para determinar os pontos, ao usar o comando `rnd`, teremos valores entre 0 e 1, exatamente o que precisamos para especificar as áreas da distribuição. Aqui usamos um exemplo com 50 graus de liberdade.

```
series t=@qtdist(rnd,50)
```

Lembre-se que, sempre que quiser encontrar um ponto que esteja associado a uma área da curva *t-student*, usamos o comando `q`. Análogo a esse comando, temos a distribuição cumulativa, que representa a probabilidade de se observar um valor de uma série de dados que não excede determinado valor específico. Tal como fizemos na curva normal, esse cálculo pode ser representado a partir de:

$$F(z) = \mathbb{P}(z \leq r)$$

onde  $F(z)$  é a área da curva acumulada até o ponto  $z$ . Na curva *t-student* temos que 50% dos dados se encontram abaixo da média e 50% acima. Com a média zero então, a probabilidade acumulada até o valor “0” é 50% ou então, expresso de outra forma:

$$F(z) = \mathbb{P}(z \leq 0) = 0,5$$

**Programação 3.2.3** Para encontrar a área acumulada até um determinado ponto na curva *t-student* podemos usar o comando `c`. Nesse caso, não se esqueça de também fornecer os graus de liberdade. Para uma curva *t*(50) usamos:

```
Scalar z=@ctdist(0,50)
```

O resultado aqui será 0,5, mostrando que toda a área da curva acumulada até o ponto 0 é de 50%. Note que isso independe de colocarmos o valor dos graus de liberdade em 100 ou 200. Isso porque estamos avaliando a curva em seu ponto médio. Agora, se avaliarmos a curva em outro ponto, os graus de liberdade produzirão resultados diferentes.

Note que a informação sobre a função inversa, dado por `q`, é similar ao que obtemos ao usar a função cumulativa. Porém, enquanto que na função inversa usando o comando `@q` e especificamos

a área para obtemos o ponto, no caso da função cumulativa usa-se @c e especificamos o ponto para obtermos a área.

**Programação 3.2.4** Para avaliar a função de densidade de uma curva t-student usamos:

```
scalar r=@dtdist(x,v)
```

Aqui, o comando scalar cria a caixa de nome “r” para receber o valor da distribuição. A seguir, especificamos d, para determinar que queremos a função de densidade, seguido do nome da distribuição tdist. Por fim, escolhemos o valor do ponto na distribuição x e os graus de liberdade em v.

O mais importante ao estudar a curva *t-student* é a construção de intervalos de confiança. Para tanto precisamos saber qual é a área definida entre dois pontos. Por exemplo, como podemos saber a área de uma curva *t-student* com 50 graus de liberdade entre -1 e 1? Veja no box de programação.

**Programação 3.2.5** Para encontrar a área entre dois pontos na curva t-student combinamos duas funções cumulativas. Suponha que se queira avaliar entre -1 e 1:

```
Scalar area=@ctdist(1,50)-@ctdist(-1,50)
```

O resultado será 67,78%, o que é menor que os 68,2% da curva normal. Agora vejamos no extremo da curva, quando consideramos entre 3 e -3. O resultado para a t-student será 99,57% enquanto que para a curva normal será de 99,73%.

**Exercício 3.3** Encontre a área entre dois pontos +2,50 e -2,50 para uma curva t-student com 50 graus de liberdade. ■

**Exercício 3.4** Encontre a área entre 3 e -3 para diferentes curvas t-student usando:

- Curva 1: 15 graus de liberdade;
- Curva 2: 30 graus de liberdade;
- Curva 3: 60 graus de liberdade.

### 3.3 A Curva Qui-Quadrado

A curva **qui-quadrado** ( $\chi^2_v$ ) possui um formato diferente da normal. Enquanto aquela tinha uma distribuição bi-caudal, essa é unicaudal. Isso é interessante, pois vários testes a serem feitos posteriormente irão considerar esse tipo de análise<sup>4</sup>, além de ser útil em diversas outras aplicações, principalmente em finanças<sup>5</sup>.

A sua função densidade é dada por:

$$f(z) = \frac{1}{2^{\frac{v}{2}} \Gamma\left(\frac{v}{2}\right)} x^{\left(\frac{v}{2}\right)-1} e^{-\frac{x}{2}} \quad (3.3)$$

Onde  $v \in \mathbb{N}^*$  são os graus de liberdade,  $x$  é uma variável aleatória no intervalo  $[0, \infty)$  e  $\Gamma(\cdot)$  é uma função Gamma<sup>6</sup>. Assim, podemos construir a curva a partir da definição do valor de  $v$  e, de

<sup>4</sup>Agradeça a Karl Pearson pelo desenvolvimento da distribuição qui-quadrado.

<sup>5</sup>Duas outras distribuições são próximas à qui-quadrado, a Poisson e a Weibull.

<sup>6</sup>A função Gamma é dada por  $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$

posse da variável aleatória  $x$ , encontrar seus diversos resultados.

Por exemplo, para uma função com 2 graus de liberdade,  $\nu = 2$ , teremos:

$$f(z) = \frac{e^{-\frac{z}{2}}}{2\Gamma(1)}$$

Um ponto a destacar aqui é que quanto maior forem os graus de liberdade da qui-quadrado mais sua distribuição vai se aproximando da normal.

No caso do uso do teste qui-quadrado também há uma particularidade a considerar. De uma forma geral esse teste é utilizado para identificar a existência ou não de diferenças em variáveis categóricas, como por exemplo, religião, sexo, raça, grupos de idade, ocorrência de evento e etc. Seu uso pode se dar para dois tipos de situações: (i) para comparar se o valor observado é diferente do valor esperado, ou então, se uma distribuição observada é diferente de uma esperada, fazendo comparação de frequências; (ii) identificar se duas variáveis aleatórias são independentes, usando tabelas de contingências.

Em ambas a aplicação poderá ver que o teste não usará as estatísticas de média e desvio padrão, ou seja, é um teste não paramétrico. Nesse caso, o que iremos fazer é comparar proporções. Como regra, ao definir as hipóteses a serem testadas seguimos que a hipótese nula é aquela onde as frequências observadas não são diferentes das frequências esperadas e, por consequência, a hipótese alternativa é onde as frequências são diferentes.

■ **Exemplo 3.1** Suponha que a razão de peso entre os estudantes homens e mulheres na universidade seja de 2:1, ou seja, os homens tem o dobro do peso das mulheres. Porém, essa relação tem sido de 1:1 em turmas de um curso específico por vários semestres. Essa relação seria estatisticamente diferente da esperada? O teste qui-quadrado é útil nesse caso.

Como forma de ilustrar como o teste qui-quadrado é utilizado, vamos usar um exemplo simples, que é descobrir se uma moeda é honesta. Esse teste também pode ser chamado de **Goodness of fit**. Nesse caso, o nosso resultado esperado é que, em 50% das vezes, se tenha cara e 50% coroa. Agora vamos ao experimento lançando uma moeda 200 vezes e anotando os resultados. Suponha que em 108 vezes se observe cara e 92 vezes coroa. Esse resultado estaria dentro do esperado?

O primeiro passo aqui é determinar a hipótese nula que, para nós, é ter uma distribuição igual entre cara e coroa, ou seja, em 200 tentativas, esperamos que 100 dessas seja cara. A seguir, podemos montar a seguinte tabela pra encontrar o valor da estatística qui-quadrado:

	Cara	Coroa	Total
Observado	108	92	200
Esperado	100	100	200
Diferença ( $O - E$ )	8	-8	0
$(O - E)^2$	64	64	128
$\chi^2 = (O - E)^2 / E$	0,64	0,64	1,28

Tabela 3.1: Testando se uma moeda é honesta



Figura 3.9

Como pode ser visto, temos duas categorias, cara e coroa. Nesse caso, a estatística qui-quadrado é dada pela soma da diferença das duas possibilidades em relação ao valor esperado, ou seja,  $\chi^2 = 1,28$ . O passo seguinte é determinar a probabilidade associada a esse valor. Mas, antes de fazer isso, vamos entender como é a distribuição qui-quadrado.

No *EViews*<sup>®</sup>, essa função de distribuição é encontrada a partir do comando `chisq`. Com o arquivo de antes aberto, vamos gerar uma distribuição aleatória com 1000000 de dados usando o comando `@qchisq()`. Um ponto importante a destacar é que o teste  $\chi^2$  só pode ser aplicado a números, não sendo aplicável a proporções, percentuais, médias e etc.

**Programação 3.3.1** A curva qui-quadrado tem um formato diferente. Usando o mesmo arquivo de antes, com 1000000 de dados vamos construir uma curva qui-quadrado com 1 grau de liberdade a partir do comando `q`, que fornece a inversa da curva:

```
rndseed 2
series q=@qchisq(rnd,1)
```

Aqui, o termo `rnd` é utilizado para gerar números aleatórios entre 0 e 1 e, nesse caso, representa diferentes valores para a probabilidade. Note que a probabilidade deve ficar entre 0 e 1.

Um exercício interessante é identificar o valor que representa determinado percentual de uma área. Por exemplo, determine o valor que representa 96% de uma amostra com distribuição qui-quadrado e 10 graus de liberdade ( $\chi_{10}^2$ ). Para encontrar esse valor, denomine o mesmo de `x` e podemos usar o comando `scalar x=@qchisq(0.96,10)` que irá retornar `x=19,02074`. Sendo assim, para os parâmetros especificados devemos esperar observar valores maiores que 19,02 em apenas 4% das vezes.

Após gerar os números aleatórios que irão seguir uma distribuição qui-quadrado, faça um gráfico combinando um histograma e uma densidade de kernel. Para tanto, abra a série `q`, vá em **view/graph**, selecione **distribution** e depois, do lado esquerdo, após escolher **histogram**, vá em **options** e escolha **kernel density**, conforme a Figura 3.9.

Note na Figura 3.10 que essa distribuição é unicaudal. Como forma de mostrar as mudanças na curva de acordo com os graus de liberdade, estimamos mais duas curvas qui-quadrado, uma com 2 graus de liberdade e outra com 5.

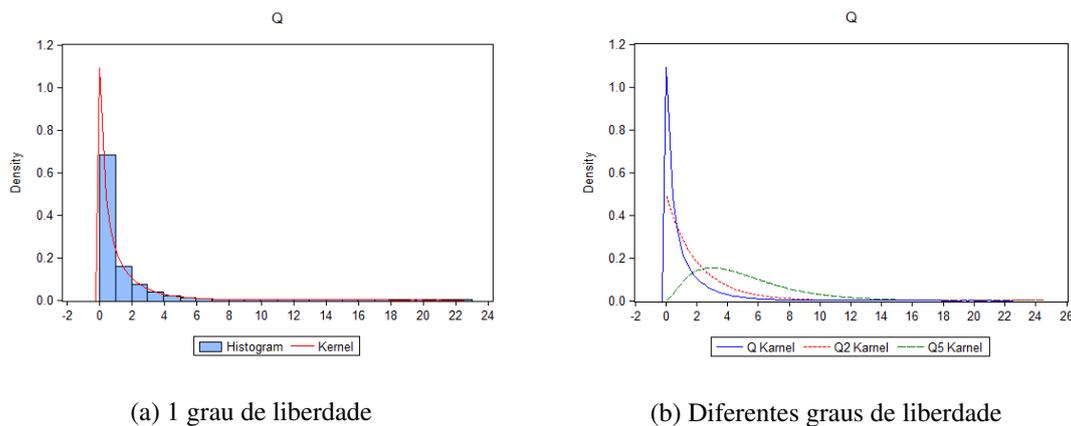


Figura 3.10: Curva Chi-Quadrado

Agora que conhecemos como é a distribuição qui-quadrado podemos retornar ao nosso exemplo das moedas e descobrir a probabilidade associada ao nosso teste. Pelos cálculos, obtemos  $\chi^2 = 1,28$ . Esse é o valor que tem que ser colocado na curva para avaliar a probabilidade associada. Assim, o total da curva entre 0 e 1,28 pode ser encontrado fazendo uso da opção de distribuição cumulativa CDF até o ponto 1,28.

**Programação 3.3.2** Para encontrar a área da curva entre o valor 0 e um ponto especificado, podemos usar o comando `@cchisq()`. Para o nosso exemplo, temos o valor de 1,28 com 1 grau de liberdade. Sendo assim, usamos:

```
scalar qq=@cchisq(1.28,1)
```

Fazendo isso, encontramos o valor de 0,7421, que representa 74,21% da curva entre 0 e 1,28. Ou seja, há uma probabilidade de 74% de nossa moeda ser viciada. O “famoso” **p-valor** associado a esse teste, que irá determinar se aceitamos ou rejeitamos a hipótese nula, é obtido a partir de  $1 - 0,7421 = 0,2579$ . Ou seja,  $p\text{-valor} = 0,25$  e, dependendo do nosso critério de significância podemos aceitar ou rejeitar a hipótese nula. Com um critério de 0,05 (ou 5%), então aceitamos a hipótese nula. Recorde-se que a nossa hipótese nula é de que o valor observado fosse igual ao esperado, ou seja, que a moeda era honesta. Portanto, podemos aceitar essa hipótese. ■

Aqui deve surgir a dúvida: porque 1 grau de liberdade? No nosso exemplo, estamos trabalhando com duas classes, cara e coroa. Nesse teste, sempre subtraímos o valor do total de classes de 1, portanto  $n - 1 = 2 - 1$  e, nesse caso, temos 1 grau de liberdade.

■ **Exemplo 3.2** Vejamos outra aplicação de um teste qui-quadrado do tipo **Goodness of fit** onde comparamos frequências. Nesse caso, vamos ver se um dado é honesto. Como se sabe, há a possibilidade de sair seis diferentes números e, nesse caso, a expectativa é que cada um tenha uma probabilidade igual. Ou seja, a probabilidade de sair o número 1 é de  $1/6$ , a mesma para sair o número 4 e assim por diante. Definimos as nossas hipóteses de teste como:

- $H_0$  : o dado é honesto (as proporções são iguais)
- $H_1$  : o dado não é honesto (as proporções são diferentes)

Agora, vamos lançar um dado 120 vezes e anotar os resultados observados junto com o esperado em uma tabela como mostrado abaixo.

Note que o resultado para alguns números supera em muito o valor que se esperava. Um indício de que o dado pode ser “viciado”. Para verificar isso, podemos usar o teste qui-quadrado comparando o valor observado com o esperado a partir da fórmula:

	Resultado	Esperado	$\frac{(O-E)^2}{E}$
1	30	20	5
2	12	20	3,2
3	27	20	2,45
4	18	20	0,20
5	17	20	0,45
6	16	20	0,80
Total	120	120	12,10

Tabela 3.2: Testando se um dado é honesto

$$\chi^2 = \frac{(O-E)^2}{E}$$

que é aplicada para cada um dos resultados. Ao final, somamos todos os seis. Essa é a estatística qui-quadrado. Para o nosso exemplo,  $\chi^2 = 12,1$ .

Para testar se esse valor corresponde ou não a aceitar ou rejeitar a hipótese nula, precisamos ter o número de graus de liberdade. Temos um procedimento com seis termos que foram utilizados para calcular a estatística, ou seja, nosso número de linhas. Sabemos que o número de graus de liberdade desse tipo de teste é dado por esse valor menos 1 ( $N^\circ$  de linhas  $- 1$ ). Sendo assim, nosso experimento tem 5 graus de liberdade,  $\chi^2_{(5)} = 12,10$ .

A seguir, devemos encontrar o p-valor. Esse pode ser dado no *EViews*<sup>®</sup> usando o comando `scalar qq=1-@cchisq(12.1, 5)` e que retorna como resultado 0,0334, ou então, 3,34%. Com esse resultado não é possível aceitar a hipótese nula, caso o nível de significância seja de 5%. O que nos leva a crer que existe uma chance pequena do dado ser honesto. Por outro lado, se o nosso nível de significância for de 1% para o teste, então pelo resultado do p-valor=0,034 aceitamos a hipótese nula de que o dado é honesto. ■

Vimos acima duas aplicações do teste qui-quadrado para o que se conhece como **Goodness of fit**. Esses testes são aplicados quando temos uma situação onde é possível determinar um valor esperado, ou seja, a nossa hipótese é baseada em uma teoria.

Outra possibilidade de aplicação desse teste é para exercícios do tipo teste de independência, ou então, como é conhecido, via tabela de contingência. Nesse caso queremos ver se duas variáveis são independentes e, para tanto, também fazemos uso do valor esperado. Mas, nesse tipo de teste, não conhecemos o valor esperado e, para tanto, devemos construir o mesmo utilizando os dados observados.

Como regra de formulação das hipóteses a serem testadas, definimos como hipótese nula o fato de que não há associação entre os grupos, ou distribuições, que estão sendo testadas, ou seja, as variáveis são independentes. Dessa forma, na hipótese alternativa teremos que as variáveis são dependentes, ou seja, há relação entre elas.

Vejamos um exemplo de teste de independência usando a função de distribuição qui-quadrado.

■ **Exemplo 3.3 Teste de Independência.** Considere que se tenha um experimento e que se queira verificar se há relação de dependência do resultado encontrado entre as diferentes categorias.

Nesse caso, suponha que, em determinado ano, tenha-se verificado a incidência de três diferentes tipos de pragas (onde praga é uma variável) em várias fazendas distribuídas em três estados (onde estado também é uma variável). Podemos afirmar que existe uma relação entre uma determinada praga e a localização da fazenda? Ou seja, é possível afirmar que quando há um problema em uma região podemos esperar que o mesmo irá ocorrer em outra região? Nesse caso queremos ver se

	Estado 1	Estado 2	Estado 3	Total
Praga 1	54	45	87	186
Praga 2	6	76	89	171
Praga 3	87	34	32	153
Total	147	155	208	510

Tabela 3.3: Incidência de praga em fazendas em três estados

	Tipo 1	Tipo 2	Tipo 3	Total
Categoria 1	a	b	c	a+b+c
Categoria 2	d	e	f	d+e+f
Categoria 3	g	h	i	g+h+i
Total	a+d+g	b+e+h	c+f+i	N

Tabela 3.4: Tabela de Contingência

existe uma relação entre duas variáveis, praga e estado.

Como primeiro passo, formulamos a hipótese nula e alternativa:

- $H_0$  : Não há relação entre região e diferentes tipos de praga (variáveis são independentes)
- $H_1$  : Há relação entre região e diferentes tipos de praga (variáveis são dependentes)

Como dito acima, a hipótese nula se refere ao caso de “independência” entre as duas variáveis. A seguir, fomos literalmente a campo e pesquisamos, nas três regiões, as fazendas que apresentaram cada uma dessas pragas. No total foram 510 fazendas que apresentaram problemas e que foram distribuídas de acordo com a tabela:

Note que temos os resultados observados, e não temos os valores esperados. Dessa forma, precisamos determinar qual é o valor esperado para esse tipo de teste. Como regra geral para um teste de independência, podemos determinar os valores esperados para cada uma das células usando uma fórmula específica. No caso de uma matriz 3x3, no geral temos:

Dessa forma, para encontrar o valor esperado da célula ‘i’, devemos usar:

$$\frac{(g + h + i)(c + f + i)}{N}$$

Onde N é dado por  $(a + b + c + d + e + f + g + h + i)$ . Usando esse procedimento, podemos produzir a matriz de valores esperados dos nossos resultados:

Depois de encontrar esses valores esperados o procedimento seguinte é encontrar a estatística qui-quadrado, que irá seguir exatamente os passos dados anteriormente quando do cálculo da moeda honesta. Primeiro encontra-se a diferença entre cada valor observado e o esperado. A seguir, eleva-se ao quadrado e divide pelo valor esperado da célula para, ao final, somar todos os resultados. Esse último valor é a estatística qui-quadrado. Esses resultados são mostrados na tabela a seguir,

	Estado 1	Estado 2	Estado 3
Praga 1	53,61	56,52	75,85
Praga 2	49,28	51,97	69,74
Praga 3	44,10	46,50	62,40

Tabela 3.5: Valores observados

	Estado 1	Estado 2	Estado 3	Total
Praga 1	0,0028	2,35	1,63	3,99
Praga 2	38,01	11,11	5,31	54,44
Praga 3	41,73	3,36	14,81	59,90
Total	16,82	16,82	21,76	118,34

Tabela 3.6: Estatística Qui-Quadrado

	Vitória	Não ganhou	total
Casa	103	76	179
Fora	42	137	179
Total	145	213	358

Tabela 3.7: Resultados de jogos do Grêmio

onde o resultado de cada célula é dado por  $\frac{(O-E)^2}{E}$ .

Observe que  $\chi^2 = 118,34$ . Agora falta determinar o número de graus de liberdade. A regra para testes do tipo tabela de contingência é usar:

$$(\text{N}^\circ \text{ de colunas} - 1)(\text{N}^\circ \text{ de linhas} - 1) = (3 - 1)(3 - 1) = 4$$

O que irá nos gerar um total de 4 graus de liberdade. Portanto, o nosso teste envolve uma estatística da forma  $\chi^2_{(4)} = 118,34$ . Usando a mesma função de antes para encontrar o p-valor no *EViews*<sup>®</sup>, ou seja, `scalar qq=1-@cchisq(118.34, 4)`, teremos `pvalor=0,000`. Para um critério de 5%, podemos concluir pela rejeição de  $H_0$ . Ou seja, não é possível aceitar  $H_0$  e, portanto, podemos afirmar que existe uma relação entre os três diferentes estados e as pragas que foram observadas em determinado ano. ■

■ **Exemplo 3.4** Muito se escuta falar que o fator “jogar em casa” costuma ser determinante para uma equipe de futebol no decorrer de um campeonato. Para comprovar esse fato, vamos testar essa hipótese para a equipe do Grêmio durante o campeonato brasileiro de 2003 a 2012. A tabela a seguir traz a divisão dos resultados, separados entre jogos em casa e fora e resultados de vitória ou não vitória, que pode tanto ser derrota quanto empate.

Tal como estruturado, as nossas hipóteses são assim dadas:

- $H_0$  : O fator “jogar em casa” não faz diferença (variáveis são independentes)
- $H_1$  : Jogar em casa faz diferença (variáveis são dependentes)

Como temos uma tabela 2x2, para encontrar o valor do teste qui-quadrado não é necessário encontrar a diferença entre cada valor observado e esperado, podemos usar, de forma direta, a fórmula:

$$\chi^2_{(1)} = \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)}$$

Como temos uma tabela 2x2, há 1 grau de liberdade. Dessa forma,  $\chi^2_{(1)} = 43,13$ . Usamos o comando `scalar qq=1-@cchisq(43.13, 1)` para encontrar o p-valor no *EViews*<sup>®</sup>, encontramos `qq=0,0000`. Ou seja, o p-valor é 0,00%. Nesse caso, podemos optar pela rejeição da hipótese nula se estivermos satisfeitos com um nível de significância de 5% ou até um nível de significância menor. Sendo assim, conclui-se que, pelo menos para o campeonato brasileiro, entre 2003 e 2012, para a equipe do Grêmio, jogar em casa ou não foi determinante. ■

Apesar de termos comentado sobre o uso de tabelas de contingência com o número de linhas igual ao número de colunas, é frequente termos tabelas de contingência que não são quadradas. Suponha um número de linhas “r” e de colunas “c”. De forma geral, a fórmula para calcular a frequência esperada para cada célula é dada por:

$$E = \frac{(\sum \text{da linha } r)(\sum \text{da linha } c)}{N}$$

onde N é o tamanho da amostra.

O último ponto de discussão sobre a aplicação do teste qui-quadrado é sobre amostras e valores esperados pequenos. Em algumas situações é comum nos depararmos com um experimento onde o número de resultados é menor do que 40. Nesse caso, claramente teremos um problema no teste. Além disso, também podemos ter uma situação onde o valor esperado de um evento, uma das células da tabela encontrada, tem um resultado menor do que 5.

Apesar de ser um problema, mesmo assim, podemos fazer o teste, basta que se faça uma correção que, na literatura de estatística, é denominada de **Correção de Yates**. E isso é simples. Quando for calcular o valor esperado de cada uma das células, ao invés de utilizar a fórmula:

$$\chi^2 = \frac{(O - E)^2}{E}$$

Usamos a seguinte expressão:

$$\chi^2 = \frac{(|O - E| - 0,5)^2}{E}$$

### 3.4 Curva F

Outra função de distribuição muito útil é a F, comumente conhecida como distribuição de Fisher, ou distribuição de Snedecor onde seu uso mais comum é na análise de variância, também conhecido como teste ANOVA. A distribuição F é uma distribuição encontrada a partir da razão da variância de duas populações independentes. Nesse caso, como estamos com duas populações, ou amostras, temos dois graus de liberdade. Por isso que a função F aparece sempre com  $F_{(v_1, v_2)}$  onde  $v_1$  são os graus de liberdade dados pelo número de amostras menos 1 e  $v_2$  é o número de tipos de medidas.

A função densidade de probabilidade de uma variável aleatória que tem distribuição F, com  $v_2$  e  $v_1$  graus de liberdade é dada por:

$$F(x) = \frac{\Gamma\left[\frac{v_1 + v_2}{2}\right] \left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} x^{\left(\frac{v_1}{2}\right) - 1}}{\Gamma\left[\frac{v_1}{2}\right] \Gamma\left[\frac{v_2}{2}\right] \left[\left(\frac{v_1}{v_2}\right)x + 1\right]^{\frac{(v_1 + v_2)}{2}}} \quad (3.4)$$

onde o valor de  $x$  é dado no intervalo  $x \in [0, \infty)$ , ou seja, assume valores positivos e  $\Gamma(\cdot)$  é uma função gamma. De forma geral, a curva  $F(\cdot)$  mede a razão entre duas distribuições qui-quadrado que sejam independentes.

Dentre as suas principais propriedades, temos que ela é assimétrica à direita, ou seja, seus valores sempre serão positivos. Dentre seus principais usos podemos destacar o teste para identificar se duas amostras independentes foram geradas por uma população com distribuição normal com a mesma variância e também se duas amostras independentes possuem mesma variância. Como hipótese principal tem o fato de que a distribuição da população no qual se está gerando a amostra é normal e que as duas populações são independentes.

Vejamos como podemos gerar 1.000.000 números aleatórios que descrevem uma distribuição F. Nesse caso, usamos, no *EViews*<sup>®</sup>, o comando `@qf.dist()`, onde o termo q representa a distribuição inversa, usada para gerar a curva procurada.

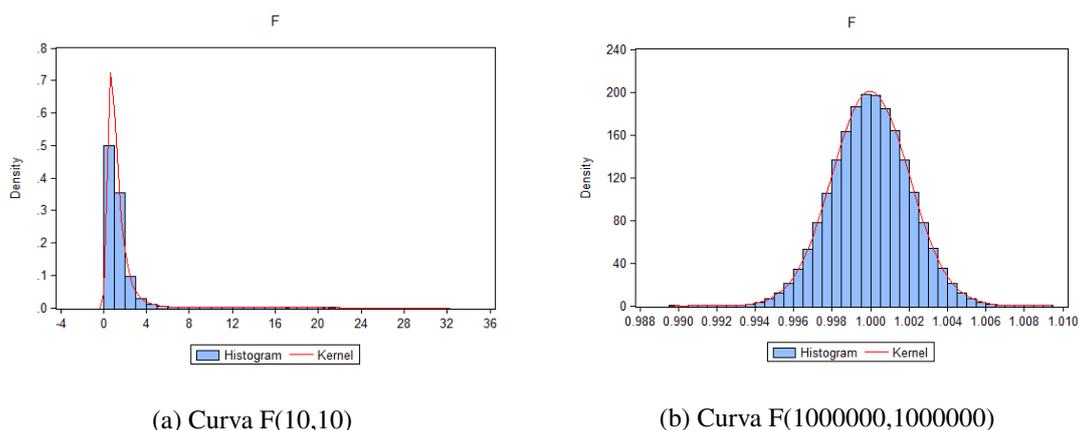
**Programação 3.4.1** A curva F também é muito útil para testes em estatística e econometria. Para simular essa curva no *EViews*<sup>®</sup> podemos usar os comandos a seguir:

```
rndseed 10
series f=@qfdist(rnd,10,10)
```

Para essa função temos 3 parâmetros a determinar dentro dos parênteses. O primeiro é a probabilidade associada. Como queremos 1.000.000 de números, usamos o termo *rnd*, que é utilizado para gerar números aleatórios entre 0 e 1 e, nesse caso, representa diferentes valores para a probabilidade. A seguir temos o número de graus de liberdade do numerador e o número de graus de liberdade do denominador. O mesmo gráfico pode ser gerado a partir de:

```
Series f=@rfdist(10,10)
```

Note que, ao especificar valores pequenos para os graus de liberdade, temos uma curva mais assimétrica (conforme a figura 3.11a). Na medida em que vamos aumentando os graus de liberdade, a curva F vai tendo outro formato, até que, ao ter um número grande de graus de liberdade, irá se aproximar da distribuição normal (conforme a figura 3.11b).



(a) Curva F(10,10)

(b) Curva F(1000000,1000000)

Figura 3.11: Curva Chi-Quadrado

Da mesma forma que para as demais curvas aqui avaliadas, para se encontrar a área abaixo da curva F podemos usar a função de distribuição cumulativa CDF. Por exemplo, para uma curva F(50,10) qual seria a área acumulada até o valor 2?

**Programação 3.4.2** Para encontrar a área da curva acumulada até determinado valor usamos a função abaixo:

```
scalar f4=@cdfist(x,v1,v2)
```

Onde *x* é o valor a determinar o ponto na curva, *v1* são os graus de liberdade do numerador e *v2* os graus de liberdade do denominador. Para o nosso exemplo, usamos:

```
series f4=@cdfist(2,50,10)
```

Que irá resultar em 0,8818, ou seja, 88,18% da área.

### 3.5 Distribuição de Poisson

Se estamos diante da possibilidade de ocorrência de um número muito grande de eventos e, que a probabilidade de ocorrência de um desses eventos seja bem pequena então, podemos usar a distribuição de Poisson. Seria como tentar medir a possibilidade de ocorrência de um evento raro, como um atropelamento em uma determinada rua de baixo movimento, o nascimento de quadrigêmeos dentre outros. A distribuição de Poisson é uma distribuição de probabilidade discreta.

Para medir essa chance de ocorrência de um evento, fazemos uso de três parâmetros. O primeiro, que se refere ao espaço de medida, pode tanto ser hora, minuto, segundo, dias, espaço, área, volume, peso ou qualquer outro campo contínuo. Na fórmula da distribuição é a variável  $t$ . Esse sempre vem acompanhado do parâmetro  $\lambda$ , que é utilizado para medir a frequência de ocorrência do evento. O último parâmetro,  $x$ , é utilizado para definir a possibilidade do número de ocorrências. A fórmula do teste de Poisson é dada por:

$$P(x) = \frac{e^{-\lambda t} (\lambda t)^x}{x!} \quad (3.5)$$

Imagine que se queira medir a probabilidade de que uma pessoa entre no restaurante a qualquer momento. Sabemos que o fluxo de clientes é medido por hora e que esse é de 3 por hora. Sendo assim,  $t = 1$  hora e  $\lambda = 3$ . Qual seria a probabilidade de não chegar nenhum cliente em 1 hora?

$$P(0) = \frac{e^{-3} (3)^0}{0!} = 0,049$$

Assim, a probabilidade de que em 1 hora não chegue nenhum cliente é de 4,9%. Outra pergunta interessante seria se, ao invés de querer saber o número exato, trabalharmos com um valor mínimo. Sendo assim, qual é a probabilidade de que chegue pelo menos um cliente? Nesse caso, podemos estimar via diferença de não chegar nenhum com o total da curva. O total é de 100% e então:

$$\mathbb{P}(\geq 1) = 1 - \mathbb{P}(0) = 1 - 0,049 = 0,9502$$

Ou seja, a probabilidade de que chegue pelo menos um cliente é de 95,02%.

**Programação 3.5.1** Para aplicar o teste de Poisson no *EViews*<sup>®</sup> podemos usar a fórmula da distribuição cumulativa (CDF). Nesse caso, é necessário especificar dois parâmetros,  $m$  e  $x$ . Com  $m = \lambda t$  e  $x$  tal como definido anteriormente.

```
scalar p=@cpoisson(x,m)
```

Para o nosso exemplo acima usamos, para medir a probabilidade de não chegar nenhum cliente:

```
scalar p=@cpoisson(0,3)
```

**Exercício 3.5** Suponha que em uma esquina ocorram, em média, 4 acidentes por semana. Encontre a probabilidade de que, em qualquer semana, ocorram 6 acidentes. Depois qual é a probabilidade de ocorrência de pelo menos 2 acidentes por semana?

Dica: na primeira pergunta  $x = 6$ ,  $\lambda = 4$ ,  $t = 1$ . Na segunda pergunta  $\mathbb{P}(\geq 2) = 1 - \mathbb{P}(1)$ ,  $x = 1$ ,  $\lambda = 4$ ,  $t = 1$  ■

Vimos nesse capítulo as curvas de distribuição e aplicação de testes, sejam esses paramétricos ou não paramétricos. Nesse ponto é importante entender a diferença entre esses dois tipos de

testes. Quando fazemos uso de estatísticas dos dados da amostra e da distribuição dos mesmos em algum teste como, por exemplo, o teste t, teste F, dentre outros, dizemos que o teste em questão é paramétrico. Ou então, denominados de testes clássicos. Nesse tipo de teste assumimos que a distribuição dos dados é conhecida.

Porém, há também os testes não paramétricos, onde não é feita nenhuma hipótese sobre o tipo de distribuição de probabilidade dos dados que estamos usando. Ou seja, nesse tipo de teste dizemos que estamos livres de especificar o tipo de distribuição. Portanto, usamos os testes não paramétricos quando desconhecemos essa distribuição ou os dados não satisfazem às suposições que são assumidas pelas técnicas tradicionais.

### 3.6 Exercícios

**Exercício 3.6** Sua namorada te liga, em média, 2 vezes por dia, considerando 24 horas. Qual é a probabilidade de ela não te ligar em 1 dia? Qual a probabilidade dela te ligar pelo menos 1 vez por dia? ■

**Exercício 3.7 Probabilidade** Considerando uma curva normal padronizada, encontre a probabilidade de se ter um valor tal como:

- (a)  $\mathbb{P}(z) = (0 < z < 1,18) = 30,10\%$
- (b)  $\mathbb{P}(z) = (0 < 2) = 97,72\%$
- (c)  $\mathbb{P}(z) = (-3,4 < z) = 99,96\%$
- (d)  $\mathbb{P}(z) = (2,45 < z) = 0,71\%$

**Exercício 3.8 Probabilidade.** Supondo que a renda da população do Brasil ( $r$ ) é de R\$ 6.200 por mês com um desvio padrão de R\$ 954. Imagine que a distribuição dessa renda seja normal. Responda aos itens a seguir. Dica: note que não temos uma distribuição normal padrão. Padronize os dados primeiro usando:

$$z = \frac{r - \bar{r}}{\sigma}$$

- (a)  $\mathbb{P}(r < 3.200) = \mathbb{P}(z < \frac{r - \bar{r}}{\sigma}) = 0,08\%$
- (b)  $\mathbb{P}(r < 9.000) = 0,16\%$
- (c)  $\mathbb{P}(3.560 < r < 6.340) = 55,55\%$

**Exercício 3.9** Considerando uma *t-student*, encontre a probabilidade de se ter um valor tal como:

- (a) use 20 graus de liberdade:  $\mathbb{P}(z) = (0 < z < 1,18) = 37,40\%$
- (b) use 30 graus de liberdade:  $\mathbb{P}(z) = (0 < z < 1,18) = 37,63\%$
- (c) use 300 graus de liberdade:  $\mathbb{P}(z) = (0 < z < 1,18) = 38,05\%$
- (d) use 20 graus de liberdade:  $\mathbb{P}(z) = (z < 2) = 97,03\%$
- (e) use 30 graus de liberdade:  $\mathbb{P}(z) = (z < 2) = 97,26\%$
- (f) use 20 graus de liberdade:  $\mathbb{P}(z) = (-3,4 < z) = 99,85\%$
- (g) use 30 graus de liberdade:  $\mathbb{P}(z) = (-3,4 < z) = 99,90\%$
- (h) use 20 graus de liberdade:  $\mathbb{P}(z) = (2,45 < z) = 0,11\%$
- (i) use 30 graus de liberdade:  $\mathbb{P}(z) = (2,45 < z) = 0,10\%$

**Exercício 3.10 Teste de independência.** Em uma pesquisa foram entrevistados 340 alunos de uma escola. Os entrevistados, separados por faixa de idade, deveriam apontar a preferência por uma cor. Sendo assim, estamos interessados em testar se existe uma relação entre idade e preferência por cor. Use como critério de significância 5%.

- Escolha a hipótese nula  $H_0$ ;
- Encontre a estatística qui-quadrado  $\chi^2$ ;
- Encontre o p-valor;
- Conclua.

Idade (anos)	Branco	Verde	Preto	Total
10-12	35	76	65	176
13-16	65	54	45	164
Total	100	130	110	340

**Exercício 3.11 Teste de independência.** Nas eleições para prefeito de 2012 tivemos vários votos nulos e brancos. Esses podem ser interpretados como uma forma de protesto. Com dados das eleições de 2012 no 1º turno para prefeito em todo o Brasil, separamos os mesmos entre capital e interior. A pergunta é: é possível afirmar que os eleitores das capitais estão mais “revoltados” do que os eleitores do interior?

	Votou	Branco + Nulo	Total
Capital	22.632.144	2.842.987	25.475.131
Interior	80.624.103	9.708.280	90.332.383
Total	103.256.247	12.551.267	115.807.514

**Exercício 3.12 Teste de independência.** Suponha que se queira testar se a faixa etária realmente faz diferença em relação a forma de dirigir. Nesse caso, com dados de jovens, adultos e idosos, separados entre números de acidentes e sem acidentes em um determinado ano, teste se há relação entre idade e condução ao volante.

	Acidente	Sem acidente	Total
Jovens	25	45	70
Adultos	15	25	40
Idosos	10	30	40
Total	50	100	150

**Exercício 3.13 Teste de independência.** Na tabela abaixo foram coletados dados sobre casamentos no Brasil no ano de 2011. Naquele ano ocorreram pouco mais de 1 milhão de casamentos divididos no estado civil do homem e da mulher na data do casamento. Por exemplo, 818.300 casamentos ocorreram entre homens e mulheres solteiros.

Homem↓/Mulher→	Solteira	Viúva	Divorciada	Total
Solteiro	818.3	5.876	50.696	874.872
Viúvo	8.557	2.925	5.297	16.779
Divorciado	88.805	4.806	38.221	131.832
Total	915662	13607	94214	1.023.483

### 3.7 Sites úteis

- [www.statistics.com](http://www.statistics.com)
- [www.portalaction.com.br](http://www.portalaction.com.br)
- <http://statlect.com/>
- <http://stat.unipg.it/iasc/>



## 4. Estatísticas, testes de hipótese e ANOVA

Fazer uma avaliação prévia de como um conjunto de dados se comporta é um dos procedimentos mais comuns em estatística e econometria, e deve ser feito antes de qualquer outra ação, pois irá permitir ter informações importantes sobre os passos a serem dados posteriormente.

Nesse caso, há diversas formas de se avaliar os dados, e que depende de como os mesmos são compostos, e que são classificados tanto em estatísticas descritivas como de inferência. No primeiro caso, há estatísticas que podem ser utilizadas para qualquer formato de conjunto de dados, como, por exemplo, a média, a moda e a mediana, referidas como medidas de tendência central. Por outro lado, quantis, variância e o desvio-padrão, por exemplo, são classificados como medidas de dispersão. Como o nome diz, no procedimento de estatística descritiva o que temos é apenas uma descrição do comportamento dos dados. No geral, os resultados gerados pela estatística descritiva aparecem no formato de gráficos ou de tabelas.

A inferência estatística envolve o conceito de amostragem. O mais comum em estatística e econometria é termos um conjunto de dados que representa uma amostra da população, uma vez que é muito difícil ter a informação da população. Nesse caso, estamos assumindo que a nossa amostra possa representar de maneira fiel o comportamento da população. Porém, nem sempre isso é verdade, o que acaba por resultar em erros de medida. Nesse caso, trabalhamos com diversos parâmetros como média, desvio padrão e etc, mas, os mesmos são estimados e são feitos testes de hipótese para confirmar a consistência dos mesmos. Em resumo, essa é a ideia da inferência estatística .

Portanto, enquanto que na **estatística descritiva** estamos apenas preocupados com a descrição dos dados, na **inferência estatística** estamos preocupados com a consistência dos mesmos.

Como exemplo, vamos usar a série  $z$  gerada na seção 3.1. Recorde-se que a mesma foi gerada para ter uma distribuição normal com média zero e variância unitária. A seguir, vá em **view/descriptive statistics & tests** e poderá ver que há diversas opções para se aplicar às séries de dados (conforme Figura 4.1). A seguir, mostraremos como interpretar cada uma dessas.

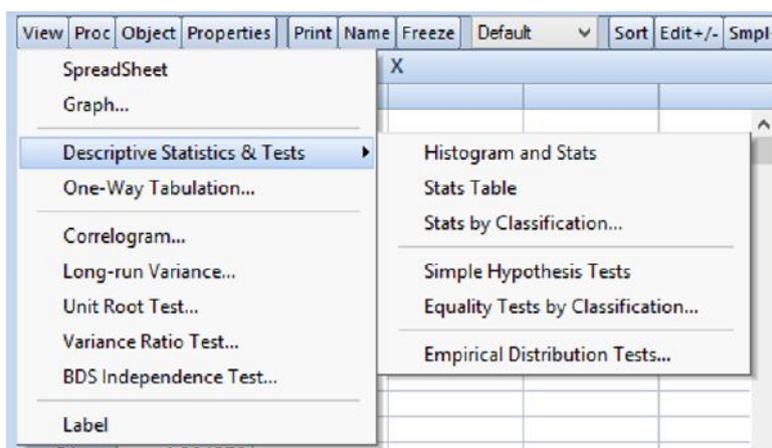


Figura 4.1: Testes e estatística descritiva

## 4.1 Histograma e Estatísticas

Selecionando a alternativa de **Histogram and Stats**, o *EViews*<sup>®</sup> irá retornar um resumo do que podemos entender como estatística descritiva, conforme Figura 4.3. Para o exemplo da série *z* podemos ver que os dados são bem distribuídos em torno da média, como mostra o gráfico à esquerda, que é conhecido como histograma.

A seguir, do lado direito, há diferentes estatísticas que são reportadas. As duas primeiras são medidas de tendência central, como a média que, tal como esperado, é próxima de zero. E, a seguir está a mediana, que representa o ponto onde a função de distribuição é dividida exatamente ao meio. Para o nosso exemplo ela também é próxima de zero. Essa é uma característica de um conjunto de dados que tem uma distribuição normal padrão, onde a média é zero.

Depois são reportados o valor máximo e o valor mínimo do nosso conjunto de dados. Note que ambos são muito próximos. Isso ocorre pois geramos uma função com distribuição normal e, nesse caso, os valores extremos, tanto para a esquerda quanto para a direita, conhecidos como caudas, devem ser próximos em módulo. Se, por exemplo, o valor máximo fosse bem diferente, em módulo, do valor mínimo, teríamos uma assimetria. A seguir está o desvio-padrão que, tal como especificado, esperava-se ter um valor unitário.

Por fim, duas outras estatísticas são importantes para avaliar os nossos dados, a assimetria e a curtose<sup>1</sup>. Ambas são estatísticas derivadas a partir da média e do desvio-padrão e úteis para caracterizar o tipo de distribuição dos dados.

**Programação 4.1.1** Podemos fazer todas essas estatísticas descritivas utilizando os comandos de programação do *EViews*<sup>®</sup>. Abaixo, vamos utilizar o `scalar` para apresentar a funções típicas para obter as estatística descritivas de uma série *x*:

```
scalar m = @mean(x)
scalar md = @median(x)
scalar mx = @max(x)
scalar min = @min(x)
scalar std = @stdev(x)
scalar assimetria = @skew(x)
scalar curt = @kurt(x)
```

Como vimos acima, o valor máximo e mínimo dos dados são muito próximos em módulo, o que acaba não gerando caudas para a nossa distribuição. Sendo assim, podemos esperar que os nossos

<sup>1</sup>Skewness e Kurtosis

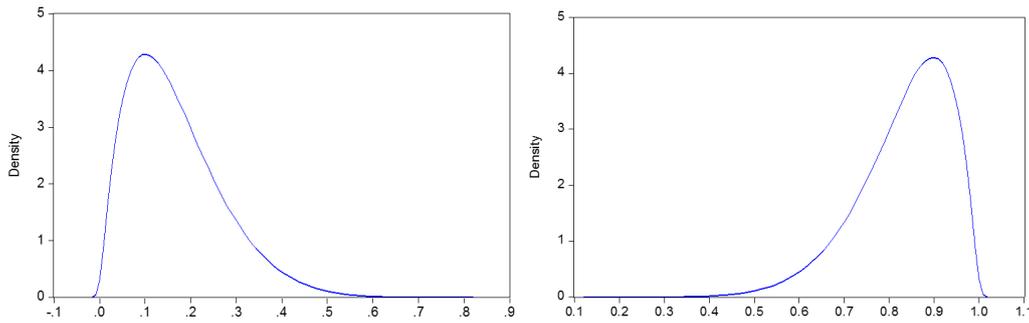


Figura 4.2: Assimetria à direita e assimetria à esquerda

dados tivessem uma distribuição simétrica, tal como sinalizado, por exemplo, pela igualdade entre a média e a mediana. Valores negativos para a assimetria indicam uma distribuição assimétrica para a esquerda, enquanto um valor positivo indica assimetria a direita. Os gráficos da Figura 4.2 mostra como se comporta a assimetria à direita e à esquerda. Para comprovar isso, calculamos a assimetria no *EViews*<sup>®</sup> com a seguinte fórmula:

$$S = \frac{1}{N} \sum_{i=1}^N \left( \frac{y_i - \bar{y}}{\hat{\sigma}} \right)^3$$

onde  $N$  é o número de observações que, no nosso caso é 1 milhão,  $y_i$  é cada uma das  $i$  observações,  $\bar{y}$  é a média dessas observações e  $\hat{\sigma}$  é o desvio-padrão amostral. Para o nosso exemplo, a assimetria é muito próxima do valor zero, o que é esperado para uma curva com distribuição normal.

Podemos facilmente mostrar como que apenas alguns valores extremos contribuem para gerar assimetria no banco de dados. Vá em **View** e depois **SpreadSheet**. Com a série aberta mude os cinco primeiros valores para números elevados, como 6, 7 e 8. Para tanto clique em **Edit +/-** na barra superior. Refaça o histograma e poderá ver como os dados apresentam assimetria à direita. Se repetir esse exemplo colocando elevados valores negativos, poderá ver que o histograma apresentará assimetria à esquerda.

A curtose, por outro lado, é uma medida relacionada à concentração dos dados, influenciando no desenho da curva verticalmente. Um conjunto de dados com um valor alto para a curtose concentra os dados na média, diminuindo bastante rapidamente quando se afasta da média. Por outro lado, dados com curtose baixa tendem a ser mais planos, com os dados mais distribuídos. Distribuições com curtose alta podem ser chamados de leptocúrticos, como os retornos das ações na bolsa de valores, enquanto distribuições com curtose mais baixa podem ser denominadas platicúrticas. Para o nosso exemplo, observamos na Figura 4.3 uma curtose com valor 3,0008, um valor muito próximo ao que se espera de uma curva normal, que é 3. O cálculo da curtose pode ser feito a partir de:

$$Z = \frac{1}{N} \sum_{i=1}^N \left( \frac{y_i - \bar{y}}{\hat{\sigma}} \right)^4$$

note que, também para esse cálculo, usamos apenas as estatísticas de média e desvio-padrão.

As duas últimas informações estão relacionadas a um teste de função de distribuição. Até então, fizemos uma avaliação na forma de estatística descritiva. Porém, somente a assimetria e curtose não são suficientes para confirmar que os dados possuem ou não uma distribuição normal. Há diversas formas para testar a possibilidade de um conjunto de dados terem uma distribuição normal ou não. Além disso, há testes que são aplicados para conjunto de dados multivariados, e também podemos testar outras distribuições. Nesse resumo de estatística descritiva, o *EViews*<sup>®</sup> retorna o resultado

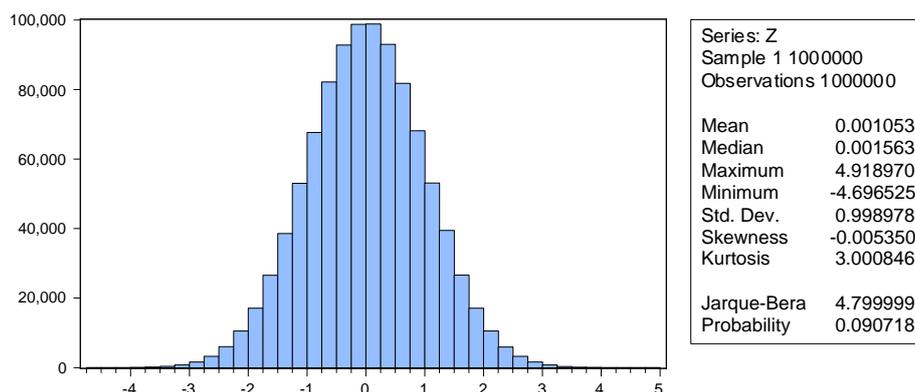


Figura 4.3: Histograma e Estatísticas de uma série Z

para o teste de normalidade de Jarque-Bera. Esse valor é encontrado usando a fórmula:

$$JB = \frac{N}{6} \left( S^2 + \frac{(k-3)^2}{4} \right)$$

onde  $N$  é o número de observações,  $S$  é o valor da assimetria e  $k$  a curtose. Substituindo os valores que vimos acima, encontraremos:

$$JB = \frac{1000000}{6} \left( -0,0053^2 + \frac{(3,0008 - 3)^2}{4} \right) = 4,799$$

Esse teste é aplicado sob a **hipótese nula de existência de distribuição normal** e, a hipótese alternativa seria que os dados não são distribuídos normalmente. Note que apenas estamos testando se a curva é normal, não estamos testando uma função de distribuição alternativa. Portanto, podemos apenas concluir se os dados são distribuídos normalmente ou não. Ou seja, o teste não permite inferir se a distribuição é qui-quadrado, F ou qualquer outra função.

No capítulo sobre funções de distribuição, aprendemos que a função qui-quadrado é utilizada em testes para verificar diferenças de distribuição entre duas amostras. No caso do teste de Jarque-Bera ocorre exatamente isso, temos um teste que tem uma estatística que usa a função qui-quadrado para testar a hipótese nula possuindo 2 graus de liberdade. Sendo assim, o mesmo é representado a partir de  $\chi^2_{(2)}$ .

Para o nosso exemplo, temos que  $\chi^2_{(2)} = 4,7999$  e usamos essa informação para encontrar o chamado p-valor, que no *EViews*® é o mesmo que *probability*. É essa estatística que irá dizer se aceitamos ou rejeitamos a hipótese nula. O número 4,7999 em uma distribuição  $\chi^2_{(2)}$  - qui-quadrado com 2 graus de liberdade produz p-valor=0,0907. Isso pode ser encontrado no *EViews*® a partir do comando `scalar qq=1-@cchisq(4.7999,2)`.

Sendo assim, não é possível rejeitar a hipótese nula de distribuição normal. As mesmas informações podem ser obtidas a partir da função **view/descriptive statistics & tests/stats table**, por isso não há necessidade de comentar seu uso. No box de programação mostramos como podemos montar um teste de Jarque-Bera usando os comandos que retornam o resultado para a assimetria e a curtose.

**Programação 4.1.2** Para fazer o histograma com a estatística dos dados podemos usar o comando “hist” para a série x e aplicar o comando freeze para salvar um gráfico com o nome “G1”:

```
x.hist
freeze(G1) = x.hist
```

Se estivermos interessados em ver apenas o resultado do teste de normalidade de Jarque-Bera, devemos construir o teste. Nesse caso, o primeiro passo é determinar um escalar e escolher um nome, suponha `jb` e depois aplicar seu resultado na curva qui-quadrado:

```
Scalar jb=((@obs(x))/6)*((@skew(x))^2+((@kurt(x)-3)\^2)/4)
Scalar testejb=@chisq(jb,2)
```

Na primeira parte construímos a estatística de Jarque-Bera usando os comandos `@obs()` para retornar o número de dados, `@skew()` para encontrar a estatística de assimetria e `@kurt()` para determinar a curtose. A seguir, encontramos o p-valor a partir da distribuição qui-quadrado, com 2 graus de liberdade.

## 4.2 Estatísticas por classificação (*Statistics by Classification*)

Quando estamos trabalhando com dados que podem ser separados por diferentes categorias ou mesmo se quisermos compreender melhor um determinado subconjunto de dados dentro do conjunto maior ou, então, comparar diferentes conjuntos de dados, podemos recorrer às estatísticas por classificação.

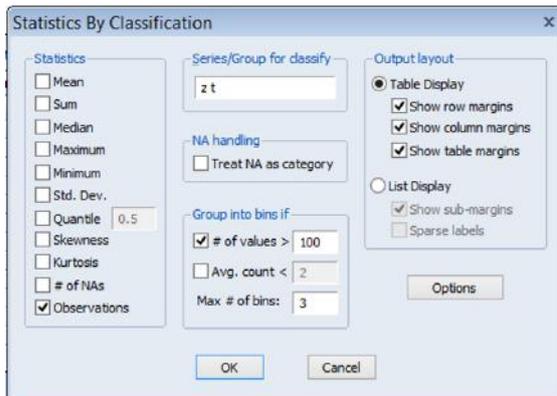
Com a série de dados aberta, clique em **view/ Descriptive Statistics/ Statistics by classification**. Do lado esquerdo da janela (ver Figura 4.4a), selecione apenas o número de observações. Depois, escreva o nome de duas séries, separadas por espaço. Vamos usar, para esse exemplo, a série aleatória `z`, com distribuição normal e a série `t`, que tem distribuição t-student com 50 graus de liberdade.

Na opção **Group into bins if**, deixe marcado apenas para valores  $>100$  e um número máximo de *bins* de 3 (isso representa o número de classes de distribuição dos dados), a seguir, clique em "ok". A Figura 4.4b apresenta os resultados. O *EViews*<sup>®</sup> mostra uma contagem dos dados dos dois grupos. Na linha estão aqueles referentes a `z`, com três intervalos e, na coluna, para a série `t` também com três intervalos. A última linha e coluna são dos totais.

Note que é feita a contagem de dados considerando a intersecção entre os dois conjuntos de dados. Por exemplo, no intervalo  $[-5,0)$  temos 249.688 dados. Porém, se avaliarmos apenas a linha do intervalo  $[-5,0)$  para `z` teremos um total de 499.392 dados onde há informações tanto de `z` quanto de `t` nesse intervalo. Por fim, o total de dados reportados tem que ser igual ao total de cada série. Do total de 1 milhão de dados, há 499.982 na série `x` que estão no intervalo  $[-5,0)$  e outros 500.011 que estão no intervalo  $[0,5)$ .

O mesmo tipo de análise pode ser feito para obter informações conjuntas sobre outras estatísticas, como mediana, desvio-padrão e etc. Vejamos como exemplo considerar o mesmo conjunto de dados e selecionar tanto a estatística de média (*Mean*) quanto a de assimetria (*skewness*). A tabela de resultado é como a tabela 4.5.

Mantemos o número máximo de classes em três, a última linha e a última coluna são os totais para cada subgrupo e o total de dados. Por exemplo, o valor -0,001053 na última célula da tabela refere-se à média do conjunto de dados `z` e, logo abaixo, o valor -0,005350 é a assimetria dos dados `z`. Isso acontece pois pedimos essa estatística a partir da abertura do conjunto de dados `z`. Se, ao invés disso tivéssemos aberto o conjunto de dados `t` e feito a estatística por classificação, essa última célula revelaria a média e assimetria para a série `t`. No intervalo  $[-5,0)$  de `z` com  $[-5,0)$  de `t` a média é -0,79 e significa que os 249.688 dados das duas amostras que caem neste intervalo possuem média -0,79 e uma assimetria de -1,004.



(a) Opções de classificação

Obs.		[-5, 0)	[0, 5)	T	All
	[-5, 0)	249688	249699	5	499392
Z	[0, 5)	250294	250312	2	500608
	All	499982	500011	7	1000000

(b) Classificação das observações

Figura 4.4: Statistics by Classification

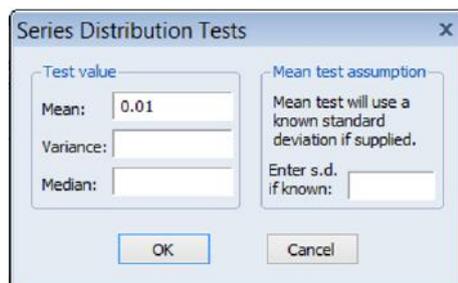
Mean		[-5, 0)	[0, 5)	T	All
Skew.	[-5, 0)	-0.796494	-0.797632	-0.522000	-0.797060
		-1.004476	-1.002469	-1.016303	-1.003471
Z	[0, 5)	0.796527	0.797930	0.520701	0.797228
		0.991711	0.987128	3.02E-16	0.989419
	All	0.000982	0.001127	-0.224085	0.001053
		-0.005304	-0.005399	-0.369841	-0.005350

Figura 4.5: Classificação por média e assimetria

### 4.3 Testes de Hipótese

Essa é uma importante ferramenta estatística para testar hipóteses em séries de dados individuais ou em conjunto. Vimos que a média da série de dados  $x$  é  $-0,001053$  e que seu desvio padrão é 1. Vamos testar a hipótese que a média é igual a 0,01. Vá em **view/descriptive statistics & tests/simple hypothesis tests** e, na caixa de diálogo que aparece (Figura 4.6a) especifique o valor da média a ser testado. No nosso exemplo 0.01. Podemos deixar em branco a informação do desvio padrão que é pedida à direita em “mean test assumption”.

Assim, na caixa que descreve *mean* digite o valor 0.01. E, na parte **Enter s.d. if known**, que corresponde ao desvio-padrão da nossa série de dados, não especifique nada. A seguir, clique em “ok”. Para esse exemplo, é possível ver como resultado apenas com a estatística  $t$ , o teste de média que segue uma distribuição *t-student*. Destaca-se que esse é um teste bi-caudal, pois estamos



(a)

Sample Mean = -0.000430		
Sample Std. Dev. = 1.000310		
Method	Value	Probability
t-statistic	-10.42665	0.0000

(b)

Figura 4.6: Teste de Hipótese

testando:

$$H_0 : \text{m\u00e9dia} = 0,01$$

$$H_0 : \text{m\u00e9dia} \neq 0,01$$

O resultado mostrado para o p-valor nos leva a rejeitar a hip\u00f3tese nula de igualdade inclusive a menos de 1% de signific\u00e2ncia. Ou seja, a m\u00e9dia de  $x$  \u00e9 estatisticamente diferente de 0,01. O teste \u00e9 realizado usando os valores amostrais para a m\u00e9dia e o desvio padr\u00e3o, e a f\u00f3rmula:

$$\text{t-statistic} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Substituindo os valores da Figura 4.6b teremos

$$\text{t-statistic} = \frac{-0.00043 - 0.01}{(1.00031)^2/\sqrt{1000000}} = -10.4265.$$

Por fim, o *probability* \u00e9 dado usando `prob = @ctdist(-10.4265, 999999)`. Lembre que os graus de liberdade s\u00e3o dados por  $N - 1$  e que esse \u00e9 um teste bicaudal.

Segue-se o mesmo procedimento para testar a igualdade da vari\u00e2ncia ou da mediana. Podemos refazer o teste especificando o desvio-padr\u00e3o. Nesse caso s\u00e3o reportados dois resultados, um para a estat\u00edstica Z, que segue uma distribui\u00e7\u00e3o normal, e outro para uma estat\u00edstica t, com desvio padr\u00e3o desconhecido. Se esse teste for aplicado para identificar se a vari\u00e2ncia \u00e9 igual a determinado valor, a hip\u00f3tese nula \u00e9 de igualdade, e usa-se a estat\u00edstica  $\chi^2_{(N-1)}$  para o teste. Sendo assim, \u00e9 aplicada a f\u00f3rmula

$$\chi^2 = \frac{(N-1)s^2}{\sigma^2} \quad (4.1)$$

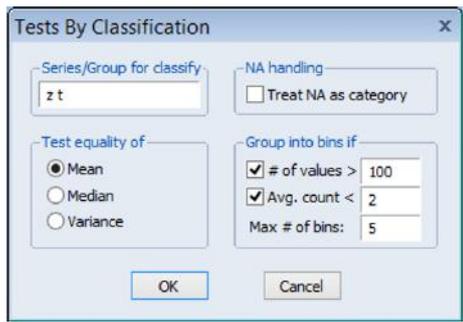
onde  $s^2$  \u00e9 a vari\u00e2ncia amostral.

#### 4.4 Teste de Igualdade por Classifica\u00e7\u00e3o

Esse teste \u00e9 muito utilizado no caso de dados categ\u00f3ricos e para verificar a rela\u00e7\u00e3o entre subconjuntos de dados. Por exemplo, \u00e9 poss\u00edvel testar se a renda m\u00e9dia \u00e9 a mesma para homens e mulheres. Os testes assumem que as subamostras s\u00e3o independentes. Indo em **view/descriptive statistics & tests/equality tests by classification...** ser\u00e1 apresentada a caixa de di\u00e1logo 4.7a. Existem as op\u00e7\u00f5es de realizar testes de igualdade entra a m\u00e9dia, a mediana e a vari\u00e2ncia das s\u00e9ries. Em **Series/Group for classify** informa-se as categoriza\u00e7\u00f5es de an\u00e1lise. As op\u00e7\u00f5es dispon\u00edveis em “Group into bins if” s\u00e3o as mesmas descritas na se\u00e7\u00e3o 4.2.

O teste de igualdade de m\u00e9dia \u00e9 um teste ANOVA<sup>2</sup>. A hip\u00f3tese nula \u00e9 que os subgrupos tem a mesma m\u00e9dia e que, dessa forma, a vari\u00e2ncia entre as m\u00e9dias da amostra devem ser as mesmas que as vari\u00e2ncias entre quaisquer subgrupos. Comparando a s\u00e9rie z categorizada pela s\u00e9rie t, observamos pela Figura 4.7b que h\u00e1 uma alta probabilidade que z n\u00e3o difira entre os grupo definido por t, pois tanto o teste ANOVA padr\u00e3o quanto o teste de Welch apresentam probabilidade acima de 70%. Ou seja, n\u00e3o \u00e9 poss\u00edvel rejeitar a hip\u00f3tese nula de igualdade. Caso o teste fosse categorizado por dois grupos, digamos t e q, seria apresentado apenas o teste ANOVA padr\u00e3o. Em ambos os casos, o *EViews*<sup>\u2122</sup> retorna uma tabela com a fonte da vari\u00e2ncia, comparando resultados entre os grupos (*between groups*) e dentro dos grupos (*within groups*). O resultado do teste \u00e9 via

<sup>2</sup>O teste ANOVA, tamb\u00e9m conhecido como an\u00e1lise de vari\u00e2ncia, \u00e9 uma t\u00e9cnica de teste de hip\u00f3tese usada para testar a igualdade de duas ou mais m\u00e9dias amostrais de uma popula\u00e7\u00e3o, tamb\u00e9m denominadas de tratamento. Na se\u00e7\u00e3o 4.8 ser\u00e1 abordado esse tema com mais detalhamento.



Method	df	Value	Probability
Anova F-test	(2, 999997)	0.180415	0.8349
Welch F-test*	(2, 16.0003)	0.304775	0.7415

\*Test allows for unequal cell variances

(b)

(a)

Figura 4.7: Teste de Igualdade

distribuição  $F_{(G-1, N-G)}$ , onde  $G$  é o número de grupos, no exemplo  $G = 2$ , e  $N$  é o número de observações.

Para o teste de igualdade de mediana, o *EViews*<sup>®</sup> calcula vários testes com a hipótese nula de que os subgrupos têm a mesma distribuição geral, contra a hipótese alternativa de que pelo menos um subgrupo tem uma distribuição diferente. Caso sejam definidos dois subgrupos, a hipótese nula é de que os dois subgrupos são amostras independentes da mesma distribuição.

Os testes de igualdade da variância avaliam a hipótese nula de que a variância em todos os subgrupos é igual, enquanto a hipótese alternativa é de que pelo menos um dos subgrupos tem variância diferente. Os principais testes oferecidos pelo *EViews*<sup>®</sup> para testar a igualdade da variância são: teste F, teste de Levene e o teste de Brown-Forsythe. Ao utilizar o teste F para atestar diferença de variância entendemos que os grupos tem distribuição normal, tornando os outros dois mais robustos.

#### 4.5 Teste de Distribuição Empírica (Kolmogorov–Smirnov)

De posse de um conjunto de dados, é muito comum não conhecermos como os mesmos são distribuídos. Para tanto, podemos aplicar um teste de distribuição para comprovar se possuem uma distribuição normal, por exemplo, como vimos no teste de Jarque-Bera, ou então, podemos estar interessados em saber se a distribuição de nossos dados é igual a alguma outra distribuição teórica. Nesse caso, há várias outras opções que podem ser verificadas no *EViews*<sup>®</sup>, conhecidas como *EDF test*.

Por exemplo, usando os dados do Capítulo 3, pode-se investigar se a distribuição da série de dados  $z$  pode ser aproximada por uma normal. Nesse caso, com a série  $z$  aberta, clique em **View / Descriptive statistics & tests / Empirical distribution tests ...**. A seguir, dentre as opções que existem vamos testar se a série de dados  $z$  tem uma distribuição normal. Deixe a opção para escolha dos parâmetros vazia, isso fará com que o *EViews*<sup>®</sup> estime os mesmos.

Note na figura 4.8b que há vários resultados de testes, e que são mostrados em duas partes. Na primeira, estão diversos testes estatísticos para verificar a hipótese nula de igualdade entre a distribuição empírica e a teórica que, nesse caso, é a curva normal. Assim, temos o teste de Lilliefors, Cramer-von Mises, Watson e Anderson-Darling. Na primeira coluna temos o valor do teste e, na última, o p-valor. Pelo resultado do p-valor, aceitamos a hipótese nula de distribuição normal em todos os quatro testes propostos. Ou seja, os dados em  $z$  possuem distribuição normal.

A segunda parte mostra os parâmetros estimados da nossa distribuição teórica. A média<sup>3</sup> (“MU”) é  $-0,001053$  e o desvio-padrão<sup>4</sup> (“SIGMA”) de  $0,998978$ . Note que esses dois resultados

<sup>3</sup>“MU” representa a letra grega  $\mu$

<sup>4</sup>“SIGMA” representa a letra grega  $\sigma$

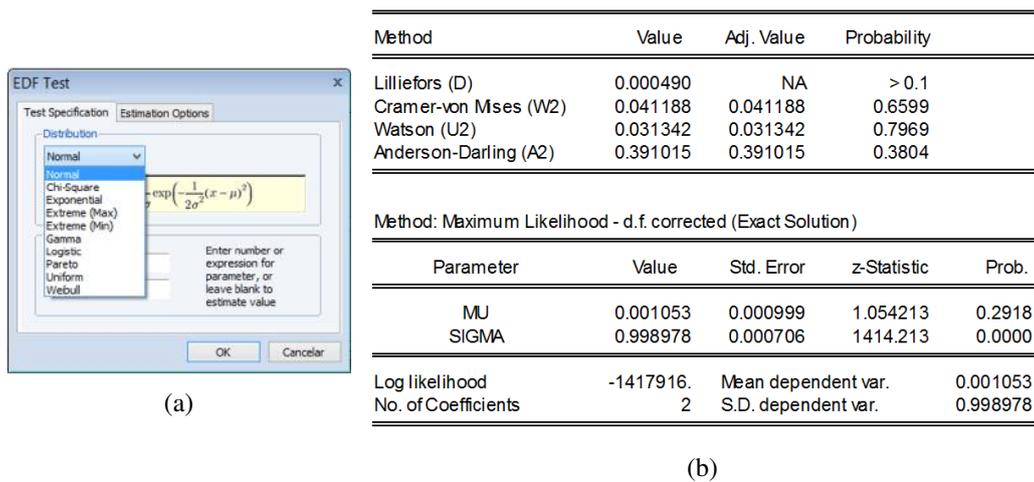


Figura 4.8: Teste de Distribuição Empírica

para a média e desvio-padrão, são iguais aos obtidos quando pedimos o Histogram & Statistics (Figura 4.3). A seguir, em Prob. temos o teste para identificar se esses valores são estatisticamente iguais a zero. No primeiro caso, o p-valor=0,2918 sinaliza que o valor da média é estatisticamente igual a zero, esse foi encontrado usando  $z = \frac{0.001053-0}{0.000999} = 1.054$ . Logo abaixo temos p-valor=0,0000 que significa que o valor de SIGMA, ou seja, o desvio-padrão, é estatisticamente diferente de zero, o que era esperado. Lembre-se que simulamos um conjunto de dados com desvio-padrão igual a 1. Se tentar testar outras distribuições teóricas, irá perceber que muitas não são possíveis, pois temos valores negativos.

Agora, faça o mesmo teste para identificar se a série de dados  $q$ , simulada para ter uma distribuição qui-quadrado, possui uma distribuição normal. O resultado é como mostrado na Figura 4.9a. Note que, agora, rejeitamos fortemente a hipótese nula de igualdade da distribuição empírica e a teórica. Nesse caso, pela segunda tabela de resultados, podemos ver que os parâmetros de média estimados para a distribuição teórica, nesse caso a normal, são média igual a 1,000459 e desvio padrão igual a 1,416870. Pelos resultados dos testes na primeira parte da tabela na Figura 4.9a rejeitamos a hipótese nula de distribuição normal dos dados.

De fato, como a série  $q$  foi gerada de acordo com uma distribuição qui-quadrado, podemos testar essa distribuição. Refazendo o teste EDF só que, agora, especificando como função teórica a curva qui-quadrado (deixe o *EViews*<sup>®</sup> estimar o número de graus de liberdade), teremos um resultado diferente. Nesse caso, pelo p-valor, todos <1, aceitamos a hipótese nula de igualdade das distribuições. Mais abaixo, na segunda tabela, podemos ver a estimativa dos graus de liberdade<sup>5</sup> (“NU”) =0,999453, praticamente o mesmo utilizado para formar a série, onde consideramos  $\nu = 1$ ).

**Programação 4.5.1** Para fazer o teste de distribuição empírica no *eviews* via programação podemos usar o comando abaixo. Nesse caso, o default é testar se a série de dados em questão possui uma distribuição normal onde os parâmetros de média e desvio padrão são estimados.

```
x.edftest
```

Alternativamente, podemos testar se a série  $q$  possui uma distribuição qui-quadrado usando:

<sup>5</sup>“NU” representa a letra grega  $\nu$

```
q.edftest(dist=chisq)
```

Method	Value	Adj. Value	Probability
Lilliefors (D)	0.240062	NA	0.0000
Cramer-von Mises (W2)	16243.46	16243.47	0.0000
Watson (U2)	14203.82	14203.83	0.0000
Anderson-Darling (A2)	88211.87	88211.93	0.0000

Method	Value	Adj. Value	Probability
Cramer-von Mises (W2)	0.036160	0.036160	<1
Watson (U2)	0.035331	0.035331	<1
Anderson-Darling (A2)	0.295167	0.295167	<1

Parameter	Value	Std. Error	z-Statistic	Prob.
MU	1.000459	0.001417	706.1047	0.0000
SIGMA	1.416870	0.001002	1414.213	0.0000

Parameter	Value	Std. Error	z-Statistic	Prob.
NU	0.999453	0.000900	1110.632	0.0000

(a)

(b)

Figura 4.9: Teste de Distribuição Empírica

#### 4.6 Teste de Igualdade (*Test of Equality*)

É comum querer testar se dois grupos de dados, sejam eles categóricos ou então séries de tempo, possuem média ou variância iguais. Para fazer isso no *EViews*<sup>®</sup> devemos primeiro criar um grupo. Esse procedimento é conhecido como ANOVA, e pode ser melhor entendido na Seção 4.8.

#### 4.7 Gráficos Analíticos – Fazendo a distribuição dos dados

Anteriormente, no capítulo sobre gráficos, aprendemos a fazer alguns tipos diferentes de gráficos misturando curvas teóricas com estimativas de kernel e histograma. Porém, naquele momento, o resultado conhecido era apenas de um gráfico, o que inviabilizava usar os dados gerados para outra análise.

Felizmente o *EViews*<sup>®</sup> permite salvar os resultados desses gráficos em uma matriz. Assim, o objetivo dessa função é poder salvar os resultados que são úteis para avaliar a distribuição dos dados criando os intervalos. Vejamos um exemplo. Abra a série de dados *z* e, a seguir em **Proc /Make Distribution Plot Data ...**. Note que irá abrir a janela representada na Figura 4.10a. Nesta, há várias opções que podem ser testadas e customizadas, sendo que as especificações do lado direito da tela mudam conforme a seleção com o tipo de dado selecionado no lado esquerdo da janela.

Para iniciar, imagine que se queira salvar os dados que podem ser utilizados para construir o histograma da série *z*. Nesse caso, selecione a opção **Histogram**. Mais abaixo escolha um nome (para poder diferenciar das demais estimativas, escolhemos como nome para essa matriz *histograma\_z*) e, do lado direito, vamos pedir que sejam salvos os dados de frequência. A seguir, clique em **ok**.

A matriz *histograma\_z* que é salva contém três colunas, conforme a Figura 4.10b. As duas primeiras, C1 e C2, são os diversos intervalos do histograma. A última coluna, a C3, é a quantidade de dados, ou seja, a frequência dos mesmos, que aparece naquele intervalo. Por exemplo, entre -4 e -3,75 temos 54 dados. As outras duas opções para dados de histograma (*Scaling* na Figura 4.10a) são densidade e frequência relativa. Ainda na parte de **Specification**, é possível ver a opção **Bin Width**. Esse se refere ao tamanho do intervalo que será utilizado para gerar o histograma. Nesse caso, podemos escolher entre um default do *EViews* ou diversas outras opções.

Uma alternativa interessante para ver como é o formato da distribuição dos dados é via **Density of Kernel**. Para a série de dados *z*, vá em **Proc /Make Distribution Plot Data ...** e depois selecione **Kernel Density**. Nas demais opções, deixe em **bandwidth** selecionado *EViews* e 100 **grid points**. Para esse exemplo o *EViews*<sup>®</sup> retorna duas colunas. Na primeira é o intervalo

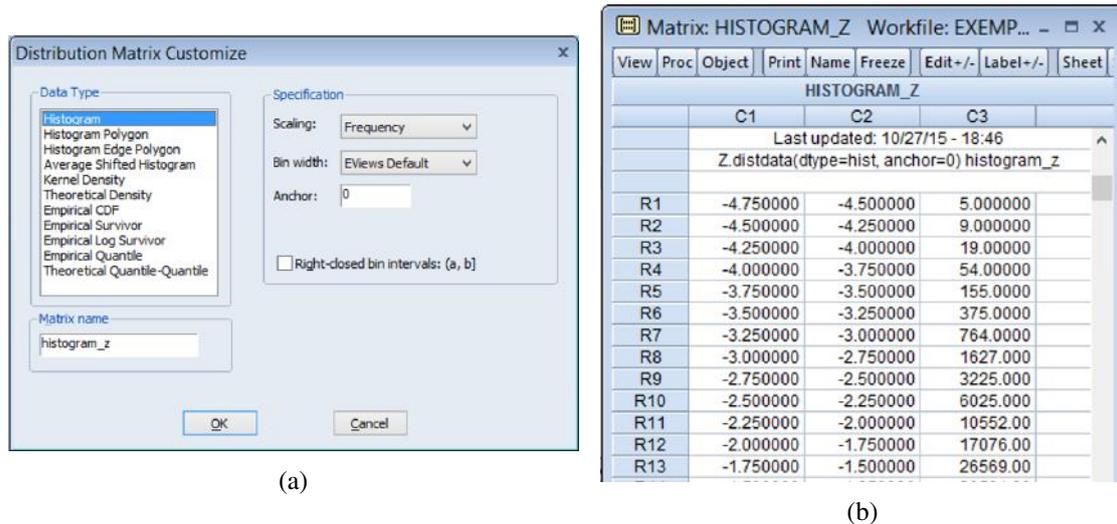


Figura 4.10: Matriz de Distribuição

superior da classe e, na segunda coluna, sua respectiva densidade. Faça o gráfico da coluna 2 (C2) e verá que temos uma distribuição próxima da curva normal.

A última opção interessante é usar em Data Type a função de densidade teórica, selecionando Theoretical Density. Do lado esquerdo há diversas funções que podem ser selecionadas e que irão retornar os resultados para a estimativa de uma função.

**Programação 4.7.1** Para obter os resultados de um histograma ou de uma função de densidade qualquer, podemos usar alguns comandos específicos. Para fazer um histograma da serie x e depois salvando o resultado com o nome de histograma\_x, usamos:

```
x.distdata(dtype=hist) histograma_x
```

Para fazer uma estimativa usando a densidade de kernel, usamos:

```
x.distdata(dtype=kernel) kernel_x
```

A opção **Unit Root Test ...** será vista quando estudarmos séries de tempo, bem como o teste de razão de variância. O **BDS Independence Test ...** será visto em regressão simples, bem como o correlograma

## 4.8 Teste de Razão de Variância

A análise de variância, conhecida como ANOVA, é uma técnica de teste de hipótese usada para testar a igualdade de duas ou mais médias amostrais de uma população, também denominadas de tratamento. Para tanto, a análise é feita via variância amostral. Com essa técnica é possível determinar se a diferença entre duas amostras é causada por um erro aleatório ou então é uma diferença estrutural.

Para o uso da análise de variância, temos que assumir três hipóteses: (i) todas as populações que estão sendo usadas devem seguir uma distribuição normal, o que acaba por caracterizar o teste como sendo paramétrico<sup>6</sup>; (ii) todas as populações devem ter a mesma variância; (iii) as amostras

<sup>6</sup>Isso não quer dizer que não possa ser feito uma análise de variância de forma não-paramétrica.

devem ser selecionadas de forma aleatória, ou seja, devem ser independentes.

Ao fazer o teste temos que ter em mente que a hipótese nula assumida sempre será de que a média das amostras selecionadas é igual. Além disso, como estamos trabalhando com a razão de variância nos dados, usamos a distribuição F para o teste.

Há basicamente quatro tipos de teste ANOVA. O primeiro é o teste **one-way between groups**. Esse é o teste ANOVA mais simples, e o objetivo é testar se existe diferença entre os grupos. O segundo é o **one-way repeated**, usado para ver, por exemplo, diferenças em um experimento repetido ou, então, para ver mudanças ao longo do tempo. Os dois testes seguintes são mais complexos: o two-way between group e two-way repeated. Nesses é feita uma investigação iterativa entre os diferentes grupos.

Vamos ver um exemplo simples para fixar o conceito, e que se encontra no arquivo de nome distribuição na planilha ANOVA. Suponha que uma empresa aplicou três diferentes métodos para a produção de um produto e, para cada um desses métodos, coletou os resultados encontrados de forma aleatória durante um mês. Ou seja, pro método 1, temos 10 informações de produtividade, para o método 2 e 3 de forma similar, completando um universo de 30 resultados. Esses métodos são descritos como c1, c2 e c3.

View	Proc	Object	Print	Name	Freeze	Default	Sort	Transpose	Edit+
obs				C1		C2		C3	
1				6.270000		3.070000		4.040000	
2				5.360000		3.290000		3.790000	
3				6.390000		4.040000		4.560000	
4				4.850000		4.190000		4.550000	
5				5.990000		3.410000		4.530000	
6				7.140000		3.750000		3.530000	
7				5.080000		4.870000		3.710000	
8				4.070000		3.940000		7.000000	
9				4.350000		6.280000		4.610000	
10				4.950000		3.150000		4.550000	

Figura 4.11: Dados da Planilha ANOVA

O natural nessa avaliação é responder se a média de produção difere entre os três métodos. Em uma avaliação prévia, podemos ver que o método 1 tem uma média de produtividade de 5,44, ao passo que para o segundo método é 3,99 e o terceiro método 4,48. Para ver as estatísticas dos dados, selecione as três séries, clique com o botão direito, abra como grupo. A seguir, vá em **Stats**, na barra de ferramentas.

Mas, será que essa média é estatisticamente diferente entre c1, c2 e c3? Qual é o melhor método e qual é o pior? Ou, reformulando a pergunta, será que o método de produção utilizado influencia na produção? Para responder a esses pontos vamos usar o método ANOVA.

Para tanto, iremos fazer uso de três estatísticas que representam a variabilidade dos dados, seja dentro do grupo ou entre grupos: (i) SQT – Soma ao quadrado total; (ii) SQE – Soma ao quadrado do erro; (iii) SQG – Soma ao quadrado dos grupos.

De uma forma geral, uma tabela de teste ANOVA é apresentada da seguinte forma, onde n representa o número total de dados, m é o número de grupos.

Origem da variabilidade	Soma dos quadrados	Graus de liberdade	Variância do quadrado médio	Razão F
Entre médias	10,82	2	5,41	5,70
Dados dos grupos ( <i>within groups</i> )	25,62	27	0,95	
Total	36,44	29		

Tabela 4.2: Resultados das estatísticas para análise da variância dos dados

Origem da variabilidade	Soma dos quadrados	Graus de liberdade	Variância do quadrado médio	Razão F
Entre médias	$SQG = n \sum_{j=1}^m (x_j - \bar{x})^2$	$m - 1$	$MSG = \frac{SQG}{m-1}$	$F_{ratio} = \frac{MSG}{MSE}$
Dados dos grupos ( <i>within groups</i> )	$SQE = \sum_{i=1}^n \sum_{j=1}^m (x_{ij} - \bar{x}_j)^2$	$n - m$	$MSE = \frac{SQE}{n-m}$	
Total	$SQT = SQE + SQG$	$n - 1$		

Tabela 4.1: Estatísticas para análise da variância dos dados

Para encontrar a primeira estatística, SQT, devemos calcular a média de todos os 30 dados, denominada média total ( $\bar{x}$ ). Para o nosso exemplo, essa é 4,643. A seguir, encontrar o desvio de cada dado  $x_j$  em relação a essa média, elevar ao quadrado e somar. É a medida de variabilidade total de todo o conjunto de dados. Assim, SQT=36,44.

A segunda estatística, SQE, é uma medida de variabilidade que deve ser encontrada para cada grupo (*within group*). Nesse caso, para o primeiro método, temos a média dos 10 dados que o integram e, encontramos o desvio de cada dado em relação a essa média, elevamos ao quadrado e, depois, somamos. Sendo assim, para o nosso exemplo, teremos três valores de SQE, um para cada um dos métodos que estamos usando. Para o método 1 temos um SQE de 8,57, para o método 2 um SQE de 8,49 e, para o método 3 um SQE de 8,55. A seguir, ao somar os três resultados, encontramos que SQE=25,62.

Por fim, a terceira estatística, SQG, é uma medida de variabilidade entre os diferentes grupos (*between group*), e que também é referida como entre médias. Nesse caso, ela representa a soma do quadrado dos desvios da média de cada grupo em relação a média total. Ou seja, encontramos a variabilidade da média do grupo que representa o método 1 em relação a média total, elevado ao quadrado. Isso é feito para cada uma das informações. Assim, no nosso exemplo teremos um resultado que se repete por 10 vezes no grupo 1. Depois fazemos o mesmo para o método 2 e para o método 3. Sendo assim, teremos 30 resultados para SQG. Ao fim, somamos todos e obtemos SQG=10,82.

De forma geral, essas três estatísticas são encontradas sempre que se vai fazer o teste ANOVA, independente de quantos grupos se está trabalhando. Outro ponto interessante é a relação que existe entre elas, dada a partir de:

$$SQT = SQE + SQG$$

$$SQT = 25,62 + 10,82 = 36,44$$

Note que a variabilidade total pode ser dividida em duas partes, uma (SQE) que representa as características de cada grupo, ou seja, representa a diferença dos grupos, cada qual com seu “tratamento” e, a segunda (SQG), as diferenças entre os grupos, a partir de um tratamento comum, que seria considerando a média global. Portanto, a origem da variabilidade total pode estar ligada a cada uma dessas duas causas.

No nosso exemplo, cada grupo tem 10 dados. Dessa forma, não há problema em usar a medida de variabilidade. Porém, pode ocorrer de compararmos grupos que possuem uma quantidade diferente de dados. Nesse caso, o grupo com maior número de dados irá ter, naturalmente, um maior valor para a variabilidade. Aqui é que entra um ponto importante no uso da ANOVA, devemos computar os graus de liberdade.

Para o conjunto total de dados, usamos  $n-1$ , onde  $n$  é o número de dados. Sendo assim, com 30 dados, os graus de liberdade de SQT é 29. No caso do SQE usamos  $n-m$ , onde  $n$  é o número de dados e  $m$  o número de grupos. No nosso exemplo,  $n-3=30-3=27$ . Sendo assim, SQE (*within group*) tem 27 graus de liberdade. Por fim, para SQG temos a diferença entre os graus de liberdade de SQT e SQE, ou seja, SQG tem 2 graus de liberdade.

De posse dos valores referentes aos graus de liberdade, podemos agora fazer a respectiva “ponderação” nas variabilidades, chegando a uma medida mais próxima da variância. Isso é feito simplesmente dividindo os valores pelos seus graus de liberdade. Em livros de estatística essa medida é denominada de MS – *Mean Square*. Assim, temos MST, para representar a estatística SQT ponderada pelos graus de liberdade,  $MSE=0,949$  relativa a SQE e  $MSG=5,411$  que se relaciona com SQG.

Por fim, encontramos a estatística F, que é dada por:

$$F_{ratio} = \frac{MSG}{MSE} = \frac{5,411}{0,949} = 5,70$$

Se essa razão for igual a 1, então, a parcela de variação explicada entre os grupos e a explicada pelo respectivo grupo é igual, ou seja, as médias são iguais. Porém, podemos chegar a essa mesma conclusão para valores diferentes de 1. Lembre-se, isso é estatística e, nesse caso, podemos ter um resultado que seja estatisticamente significativo.

Porque estamos usando a estatística F para esse teste? Na discussão sobre funções de distribuições, ilustramos que a distribuição F é dada a partir da razão de variâncias sob a hipótese nula. Portanto, a curva F irá ter todos os resultados possíveis para as razões de variância. A seguir, calculamos o  $F_{ratio}$  e identificamos se seu valor pode ser considerado estatisticamente significativo comparando o mesmo com a distribuição F.

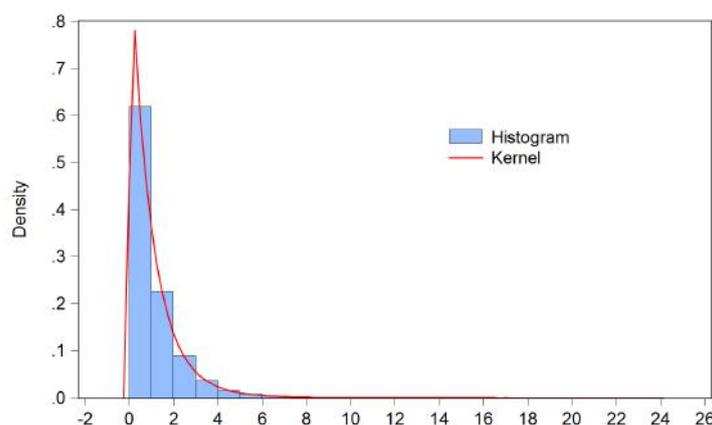
No nosso exemplo, temos uma distribuição  $F_{(2,27)}$ , ou seja, com 2 graus de liberdade no numerador e 27 no denominador. Podemos encontrar seu desenho a partir de um conjunto de 1000000 de dados aleatórios para ver como seria essa distribuição.

**Programação 4.8.1** Para encontrar a forma como os dados de uma distribuição F se comportam, com 1000000 de dados aleatórios, 2 graus de liberdade no numerador e 27 no denominador, podemos usar:

```
series f5 = @qfdist(rnd, 2,27)
```

Para fazer essa estimativa não se esqueça de usar uma planilha que tenha uma dimensão de 1000000 de dados, como a usada no exemplo distribuição.

Note na Figura 4.12 que, como temos 27 graus de liberdade no denominador, a curva tem sua área um pouco menos concentrada perto do valor zero. O próximo passo seria determinar qual o

Figura 4.12: Curva  $F_{2,27}$ 

p-valor associado a estatística  $F_{ratio} = 5,7$  que foi encontrada no nosso teste. Para tanto, podemos fazer uso da função cumulativa `@cdfist()` - veja o box de programação 4.8.2.

Esse irá produzir como resultado p-valor=0,008, que é a área da curva à direita do valor  $F=5,7$ . Sendo assim, podemos concluir que os três métodos apresentam diferença no resultado final, ou seja, rejeitamos a hipótese nula a 0,8%.

**Programação 4.8.2** Para encontrar o p-valor associado ao valor do teste F, devemos ter em mente que a função cumulativa fornece a área até determinado valor. Sendo assim, devemos subtrair de 1, a partir de:

```
scalar f=1-@cdfist(5.70, 2,27)
```

Esse procedimento pode ser facilmente feito no *EViews*<sup>®</sup>, sem a necessidade de todos esses cálculos. Na planilha de nome ANOVA, temos as nossas três séries de dados referentes aos nossos três métodos. Como primeiro passo, crie um grupo com essas três séries. A seguir, vá em **View /Tests of equality ...**, selecione mean e clique em ok. Os resultados são apresentados em três partes. Na primeira está o resultado final (Figura 4.13a), que aponta o teste F e também o teste de Welch.

A seguir, está o bloco com o resultado da análise de variância (Figura 4.13b), com suas respectivas estatísticas SQG, SQE e SQT, além das MSG, MSE e MST, que são ponderadas pelos graus de liberdade.

Por fim, no terceiro bloco (Figura 4.13c), são mostradas as estatísticas referentes às séries de dados que foram avaliadas, suas respectivas médias, desvio padrão e erro padrão, tanto por grupo quanto no conjunto.

Vale destacar que apenas concluir que as médias são diferentes, como identificado pelo teste acima, não é o suficiente. Muitas vezes estamos interessados em saber a origem dessa diferença, e isso pode ser verificado a partir do intervalo de confiança. O primeiro passo é determinar o tamanho do intervalo. Vamos supor 95% para uma estatística t. Nesse caso, com 27 graus de liberdade, o valor de  $t_{95\%} = 2,05$  e, o intervalo para cada grupo é construído a partir de:

$$\text{média} \pm t_{95\%} \sigma$$

Como obtemos esse resultado para t? Usando a função do EViews que descreve o ponto a partir da área. Lembre-se que a curva t é bicaudal. Como queremos 95% de intervalo de confiança, sobra

Test for Equality of Means Between Series

Date: 02/02/13 Time: 02:55

Sample: 1 10

Included observations: 10

Method	df	Value	Probability
Anova F-test	(2, 27)	5.702374	0.0086
Welch F-test*	(2, 17.9999)	5.496724	0.0137

\*Test allows for unequal cell variances

Analysis of Variance

Source of Variation	df	Sum of Sq.	Mean Sq.
Between	2	10.82275	5.411373
Within	27	25.62215	0.948969
Total	29	36.44490	1.256721

(a) Testes F e de Welch

Category Statistics

Variable	Count	Mean	Std. Dev.	Std. Err. of Mean
C1	10	5.445000	0.975981	0.308632
C2	10	3.999000	0.971750	0.307294
C3	10	4.487000	0.974714	0.308232
All	30	4.643667	1.121035	0.204672

(c) Estatísticas do Grupo

(b) Análise da Variância

Figura 4.13: Teste de Igualdade das Médias entre as Séries - ANOVA

5% para ser dividido nas duas áreas, uma à esquerda com 2,5% e outra à direita com 2,5%. Assim, usamos a função `scalar intervalo = @qtdist(0.025, 27)`. Aplicando isso para os nossos valores da tabela anterior, podemos encontrar os resultados apresentados na Tabela 4.3.

	Mínimo	Média	Máximo
C1	4,81	5,44	6,07
C2	3,36	3,99	4,62
C3	3,85	4,48	5,11

Tabela 4.3: Intervalo de Confiança para a Média 95%

■ **Exemplo 4.1** Também há outra forma de fazer o teste ANOVA conhecendo apenas o número de observações, a média e a variância dos dados em questão. Suponha, por exemplo, que se queira verificar se o nível de qualificação de um trabalhador em determinada empresa influencia na sua produtividade. Nesse caso, selecionamos três tipos de trabalhadores: estagiários, formado, pós-graduado para serem avaliados. Os resultados são mostrados na tabela.

	Nº	Média	Variância
Estagiário	23	29,1	18,3
Graduado	21	28,1	16,9
Pós-graduado	16	21,3	15,2

Como primeiro passo, definimos as hipóteses:

- $H_0$ : não há diferença entre os níveis de qualificação e produtividade
- $H_a$ : Existe diferença de produtividade entre os níveis de qualificação

No total foram 60 dados distribuídos em 23 estagiários, 21 trabalhadores graduados e 16 com pós-graduação. A seguir temos as respectivas médias de tempo gasto para executar uma tarefa e a variância. Note que aqui não temos os dados da pesquisa, apenas os resultados de média e variância. Mas, podemos fazer o teste ANOVA mesmo assim.

O primeiro passo é determinar a média total entre os três grupos. No nosso exemplo essa é dada por 21,16. A seguir, fazemos a soma do quadrado total, que consiste em fazer a diferença

entre a média de cada grupo e a média total:

$$\begin{aligned} SQG &= n_1(x_1 - \bar{x}) + n_2(x_2 - \bar{x}) + n_3(x_3 - \bar{x}) \\ SQG &= 23(29,1 - 26,16) + 21(28,1 - 26,16) + 16(21,3 - 26,16) \\ SQG &= 655,34 \end{aligned}$$

A seguir encontramos a estatística SQE, que é uma medida de variabilidade de cada grupo (within group) usando a formula do SQE, onde  $(s_i^2)$  é a variância do grupo i, temos:

$$\begin{aligned} SQE &= (n_1 - 1)(s_1^2) + (n_2 - 1)(s_2^2) + (n_3 - 1)(s_3^2) \\ SQE &= (22)(18,3) + (20)(16,9) + (15)(15,2) \\ SQE &= 968,60 \end{aligned}$$

Agora, devemos fazer o ajuste para cada uma das estatísticas pelos graus de liberdade. No caso da SQG, os graus de liberdade são dados pela diferença entre o número de argumentos menos um. Como temos três diferentes argumentos, estagiário, graduado e pós-graduado então, há 2 graus de liberdade para SQG. No caso de SQE, os graus de liberdade são dados pela diferença entre o total de dados utilizados e o número de argumentos. Como temos um total de 60 dados então, os graus de liberdade de SQE serão 57.

Podemos, assim, encontrar a estatística F:

$$F = \frac{SQG/m-1}{SQE/(n-m)} = \frac{655,34/2}{968,60/57} = 19,2828$$

Com esse resultado rejeitamos fortemente a hipótese nula, basta ver em `scalar f= 1-@cfdist(19.2828, 2, 57)` no *EViews*<sup>®</sup>, que produz um p-valor=0,000. Sendo assim, o nível de qualificação é importante para determinar diferenças na produtividade.

Descobrimos que existe diferença, mas, não de onde vem essa diferença. Para responder a esse ponto, aplicamos um teste de diferença de média que usa a curva t. Como temos três argumentos, para descobrir a origem da diferença temos que testar aos pares. Nesse tipo de teste temos que determinar apenas qual é o nível de significância procurado para que se construa o intervalo de confiança.

Como regra geral, ao avaliar a diferença entre a média do grupo 1 com a média do grupo 2, usamos:

$$\mu_1 - \mu_2 \pm t_{\alpha/2c} \sqrt{\frac{SQE}{(m-n)} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

onde  $\mu_1$  é a média do grupo 1,  $t_{\alpha/2c}$  é a estatística t avaliada em um ponto,  $\alpha$  é o nível de significância,  $(m-n)$  é o número de graus de liberdade  $n_1$  é o total de dados do grupo 1 e  $c$  é dado por:

$$c = \frac{m(m-1)}{2}$$

Como regra de decisão, se o intervalo de confiança não contiver o valor 0 então, rejeitamos a hipótese nula. Primeiro vamos descobrir se tem diferença entre o resultado para estagiário e graduado:

- H0:  $\mu_{\text{estagiário}} = \mu_{\text{graduado}}$ ;
- Ha: As médias são diferentes

Nesse caso temos:

$$c = \frac{3(3-1)}{2} = 3$$

e, para  $\alpha = 0,05$  temos que encontrar o valor de  $t_{\alpha/2c} = t_{0,05/6} = 0,0083$ . Isso pode ser feito no *EViews*® utilizando `scalar t=@qtdist(0.0083, 57)`. Com isso, temos um valor de 2,46. Agora é só determinar o intervalo:

$$29,1 - 28,1 \pm 2,46 \sqrt{16,992 \left( \frac{1}{23} + \frac{1}{21} \right)}$$

Dessa forma, não rejeitamos a hipótese nula. Ou seja, a média entre estagiários e graduados é estatisticamente igual. Faça a mesma conta para verificar a diferença de média entre estagiário e pós-graduado. O resultado será:  $4,49 < \mu_1 - \mu_3 < 11,10$ , fazendo com que se rejeita a hipótese nula, ou seja, há diferença de média entre estagiários e pós-graduados. Por fim, podemos fazer para verificar a diferença entre graduado e pós-graduado, o que irá resultar em:  $3,43 < \mu_1 - \mu_3 < 10,16$ . Também apontando para a rejeição da hipótese nula, ou seja, temos diferença entre as médias.

**Programação 4.8.3** Diante de dados como o apontado no exemplo da qualificação, podemos usar de programação para produzir os resultados do teste ANOVA de maneira direta.

```

'programa para calcular o intervalo de confiança em um teste ANOVA
'os parametros abaixo podem ser modificados
'n representa o total de dados por grupos
'm representa a media do grupo
scalar n1=23
scalar n2=21
scalar n3=16
scalar total=n1+n2+n3
scalar m1=29.1
scalar m2=28.1
scalar m3=21.3
scalar sqe1=968.60/(total-3)

'parâmetros de escolha para o intervalo
scalar alfa=0.05

'resultado para a estatística t
scalar t=-@qtdist(alfa/6, total-3)

'testando a diferença entre m1 e m2
scalar minimo=m1-m2-t*@sqrt(sqe1*((1/n1)+(1/n2)))
scalar maximo=m1-m2+t*@sqrt(sqe1*((1/n1)+(1/n2)))

```

■

## 4.9 Exercícios

**Exercício 4.1** Três tipos de baterias estão sendo testadas sob condições de alta pressão. Na tabela abaixo está o tempo, em horas, que 10 baterias de cada marca funcionou antes de ficar sem energia.

Marca da bateria		
1	2	3
5,60	5,38	6,40
5,43	6,63	5,91
4,83	4,60	6,56
4,22	2,31	6,64
5,78	4,55	5,59
5,22	2,93	4,93
4,35	3,90	6,30
3,63	3,47	6,77
5,02	4,25	5,29
5,17	7,35	5,18

(a) Use a análise de variância para determinar se as baterias de cada marca levaram tempos significativamente diferentes para descarregar por completo. Se o tempo de descarregamento for significativamente diferente (ao nível de confiança de 0,05) determine qual marca de bateria diferem uma das outras. Especifique e verifique os pressupostos do modelo.

(b) Podemos dizer que resultados da marca 1 tem distribuição normal a 5% de significância?

A tabela ANOVA do Exercício 4.1 é:

	Soma dos quadrados	Graus de liberdade	Var. do quadrado médio	Razão F
Entre médias	10,77	2	5,39	4,79
<i>Within groups</i>	30,33	27	1,12	
Total	41,11			

Testando 5% de significância, a região crítica inclui os valores superiores a  $F_{2,27}(0,95) = 3,354$ . O resultado da Razão F 4,79 fica na região crítica, portanto, rejeitamos a hipótese das médias serem iguais. O teste indica que não há diferença entre as marcas 1 e 2, mas a marca 3 difere-se da marca 2.

O resultado do teste de Jarque-Bera foi 5,0603 e, aplicando à uma distribuição qui-quadrado com 2 graus de liberdade temos que  $\chi^2_{(2)} = 0,0796$ . Portanto, não podemos rejeitar a hipótese nula de existência de distribuição normal.

**Exercício 4.2** Uma siderúrgica está testando a eficiência de seus alto-fornos. Para a produção de uma peça específica, o forno precisa alcançar rapidamente a temperatura de 900 °C. Quatro fornos foram testados várias vezes para determinar o tempo (em minutos) que levavam para atingir essa temperatura e foram obtidos os seguintes resultados:

Forno	$n_i$	$\bar{x}_i$	$s_i$
1	15	14,21	0,52
2	15	13,11	0,47
3	10	15,17	0,60
4	10	12,42	0,43

O tempo médio de aquecimento dos fornos são diferentes? Caso sejam, qual forno é o mais rápido? E qual é o mais lento?

A tabela ANOVA para o Exercício 4.2 é:

	Soma dos quadrados	Graus de liberdade	Var. do quadrado médio	Razão F
Entre médias	47,106	3	15,702	61,303
<i>Within groups</i>	11,782	46	0,2561	
Total	58,888	49		

Testando um nível de 5% de significância,  $F_{3,46}(0,95) = 2,802$ . Considerando que  $61,303 > 2,806$  rejeitamos a hipótese nula. Assim, consideramos que o tempo médio de aquecimento dos fornos diferem-se. Realizando múltiplas comparações, concluímos que o forno número 4 é o mais rápido e o número 3 o mais lento.



## 5. Características dos dados de séries de tempo

Um banco de dados pode ser organizado de várias formas e os testes e modelos aplicados seguem esse desenho. Para dados com periodicidade definida, como mês, trimestre ou ano, usamos os conceitos de série de tempo. Por outro lado podemos ter dados que descrevem as características, em um dado momento, de vários indivíduos, denominados de cross section. Também há a opção de dados em painel que agrega informações de indivíduos com o tempo. Nesse capítulo serão apresentadas as principais características de uma série de tempo, assim como os ajustes e filtros possíveis de serem aplicados com o *EViews*<sup>®</sup>. Com conjunto de dados de série de tempo é possível extrair várias informações que ajudam a compreender o comportamento desses ao longo do período.

### 5.1 Ajuste Sazonal

A sazonalidade é entendida como um processo que pode ter diferentes periodicidades dentro de um determinado período. Podemos identificar a presença de sazonalidade em dados trimestrais ou mensais, os mais comuns, mas também é possível que se tenha um comportamento sazonal em dias dentro de uma semana, horas e etc. As primeiras investigações<sup>1</sup> sobre essa característica dos dados remontam a 1884 e, até mais recentemente, a forma de identificar essa era decompondo a série de dados  $y_t$  a partir de seus componentes, como tendência ( $T_t$ ), ciclo ( $C_t$ ), sazonalidade ( $S_t$ ) e componentes irregulares ( $I_t$ ).

Os modelos construídos a partir de então são denominados de “modelos de componentes não-observáveis”, podendo ter a forma de aditivo:

$$y_t = T_t + C_t + S_t + I_t$$

Ou então, multiplicativo:

$$y_t = T_t * C_t * S_t * I_t$$

De início, os modelos que procuravam determinar o comportamento sazonal de uma série de tempo assumiam que esse era constante ao longo do tempo. Porém, há diversos fatores,

<sup>1</sup>Uma boa referência para essa discussão histórica está em Hylleberg(1986).

como mudanças na temperatura média, diferentes expectativas, mudança de comportamento do consumidor, efeito feriado e outros, que podem produzir um padrão sazonal diferente hoje do que se identificava no passado. Um ponto importante a lembrar é que a não correção da característica sazonal dos dados, antes de se fazer uma análise de regressão, bem como, a aplicação de um filtro errado para corrigir a sazonalidade, podem distorcer os resultados finais e prejudicar a interpretação. Nesse caso, podemos escolher resolver o problema sazonal de maneira integrada com o modelo final ou então, de maneira individual antes da modelagem final. Esse caso é o mais comum, onde são usadas variáveis *dummy* para corrigir o problema da sazonalidade. Outra alternativa é o uso do **Band Pass Filter** onde a análise é feita a partir do domínio da frequência e é utilizada uma transformação de Fourier na série de dados.

Os modelos de série de tempo para correção da sazonalidade, como apontado por Hylleberg(2006) podem ser de vários tipos. No caso univariado: (i) modelos de Box-Jenkins; (ii) modelos de componentes não-observáveis; (iii) modelos de parâmetros variáveis no tempo. Para o caso multivariado: (i) cointegração sazonal; (ii) cointegração periódica; (iii) características sazonais comuns.

Como primeiro passo de investigação de uma característica sazonal vamos ver sua representação gráfica. Para essa seção vamos usar a série que descreve o PIB mensal do Brasil e calculada pelo Banco Central, o IBC-BR, número 17439, sem ajuste sazonal. Você pode fazer o download da mesma no site do BC ou abrir o arquivo de nome IBCbr.wf1. Selecione a série *ibcbr* e clique em **View/Graph.../Seasonal Graph**, tal como mostrado na figura 5.1.

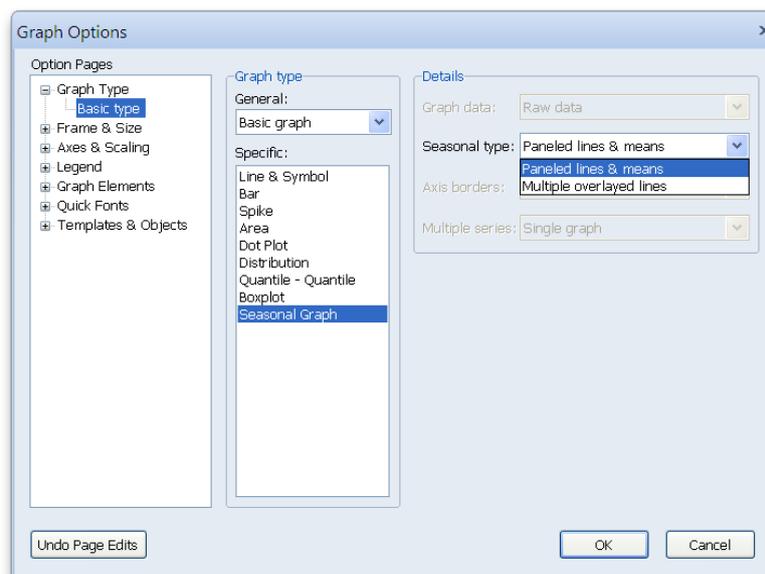


Figura 5.1: Opções de Gráfico Sazonal

Note que em seasonal type, temos duas opções, a primeira, quando é selecionado **Paneled lines & means**, irá mostrar como a série de dados se comporta para cada mês ou trimestre e, na segunda opção, em **Multiple overlaid lines**, os dados são divididos em diferentes linhas, cada qual representando o período específico, para todo o período amostral. Vai da opção de cada um ver qual dos dois gráficos melhor descreve o comportamento dos dados, não havendo regra. Ambos os gráficos são mostrados na figura 5.2. O primeiro, 5.2a, indica doze linhas de cor azul – lembre-se que estamos usando dados mensais – e que sinalizam como evoluíram os dados em cada mês durante todo o período de análise. Esse é complementado com a informação da média para cada mês, indicada pela linha vermelha. Por exemplo, a primeira informação relacionada ao mês de

fevereiro ocorre em 2002M02, e tem valor 99, ao passo que a última, em 2015M2, tem valor 138. A média dos valores do mês de fevereiro é 121, e é exatamente a linha vermelha horizontal. Para identificar esse valor o leitor deve deixar o mouse posicionado sobre a linha vermelha no gráfico no *EViews*<sup>®</sup>. Note que a média de valores do mês está bem longe dos extremos, sinalizando que, durante o período de análise, a sazonalidade do mês teve forte modificação, tendo atingido um mínimo de 99 e um máximo de 144. Certamente seria um erro considerar a sazonalidade média como representativa de tudo.

A segunda forma de ver o comportamento dos dados separados para cada um dos meses é selecionando a opção de múltiplos gráficos: **Multiple overlaid lines**, cujo resultado está mostrado na figura 5.2b. Note que há uma tendência de crescimento dos valores para cada mês ao longo do tempo. Isso tem uma implicação importante como comentado acima, em especial pelo fato de que usar a média de cada mês para identificar e corrigir padrão sazonal estaria incorreto, ou seja, a média de cada trimestre não é constante ao longo do tempo. Uma parte da literatura em econometria usa a média como fator de dessazonalização. Isso é conhecido como “**sazonalidade determinística**” e sua correção é feita com o uso de variáveis *dummy* (valores zero e um).

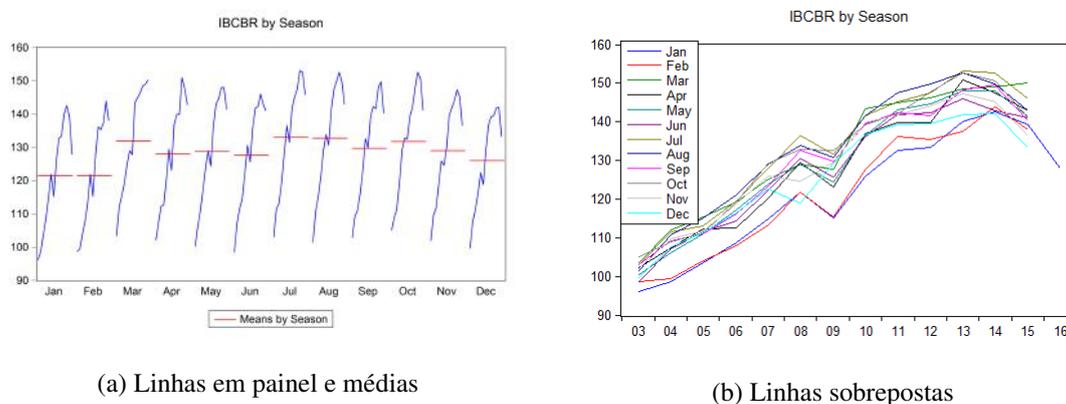


Figura 5.2: Gráfico da Sazonalidade

Mas há outros métodos mais sofisticados e específicos que podem ser utilizados, e o *EViews*<sup>®</sup> permite seu uso. Com a série *ibcbr* aberta, vá em **Proc/Seasonal Adjustment**. Note que são fornecidas cinco diferentes opções para se dessazonalizar os dados. Vamos discutir os aspectos gerais do método X-12 comparativamente ao método das médias móveis e TRAMO/SEATS, sem entrar no detalhe técnico, que pode ser visto em outros livros de econometria.

### 5.1.1 Método das Médias Móveis (Moving Average Methods)

Esse método é simples a ponto de resultar em uma importante perda de informação do comportamento sazonal dos dados. Nesse caso, a modelagem é feita a partir de:

$$y_t = \sum_{s=1}^S \delta_{st} m_s + \varepsilon_t$$

Onde  $S$  é o número de períodos, se dados mensais  $S=12$  e se forem trimestrais  $S=4$ ;  $\delta_{st}$  assume valores 1 para o respectivo período sazonal em questão e zero caso contrário;  $m_s$  é o valor da média desses períodos e, por fim  $\varepsilon_t$  é estacionário com média zero. Sendo assim, a equação geral que irá medir a sazonalidade por médias para dados trimestrais, é dada por:

$$qx_t = \delta_{1t} m_1 + \delta_{2t} m_2 + \delta_{3t} m_3 + \delta_{4t} m_4 + \varepsilon_t$$

No caso de uma série de dados com periodicidade mensal teremos 12 variáveis  $\delta_{st}$ . Para encontrar os respectivos valores devemos criar séries de dados usando variáveis *dummy* de valor 1 e 0, tal como mostrado abaixo no caso trimestral:

	qx	Primeiro trimestre	Segundo trimestre	Terceiro trimestre	Quarto trimestre
1997Q1	38.027	1	0	0	0
1997Q2	44.520	0	1	0	0
1997Q3	45.070	0	0	1	0
1997Q4	46.547	0	0	0	1
1998Q1	45.003	1	0	0	0
1998Q2	42.943	0	1	0	0
1998Q3	44.047	0	0	1	0

A seguir, rodamos a regressão para encontrar os respectivos valores de  $m_s$ . Note que essa regressão é feita sem o uso da constante. Caso contrário seria encontrado cinco valores para a média em dados trimestrais e treze em dados mensais, e a matriz não seria simétrica. Vejamos como o *EViews*<sup>®</sup> faz essa estimativa. Com a série *ibcbr* aberta, vá em **Proc/Seasonal Adjustment/Moving Average Methods**. Escolha o método multiplicativo e um nome para a série resultante (aqui no exemplo colocamos o número 1 na frente para diferenciar esse método de dessazonalização do X-12 a ser visto a seguir).

**Programação 5.1.1** Também pode ser usado um comando para se fazer a dessazonalização. Nesse caso, para o método multiplicativo, podemos escrever:

```
seas(m) ibcbr ibcbr_sa ibcbr_sf
```

A letra 'm' representa o método multiplicativo. Caso queira o método aditivo, use 'a'. O comando é seguido pelo nome da série, o nome da série ajustada sazonalmente e o fator sazonal.

É comum trabalhar com modelos com várias séries de tempo, o que demandaria tempo para aplicar o método de dessazonalização para cada uma. Como forma de operacionalizar isso de maneira rápida, podemos usar um *loop* para dessazonalizar todas as séries do banco de dados ao mesmo tempo com apenas um comando simples. Porém, nesse caso, é necessário abrir um programa antes. Vá em **File/New/Program**. A seguir, escreva o programa abaixo e salve em qualquer lugar do computador e feche o mesmo.

```
for %a qx y px pm qm
seas(m) {%a} {%a}_sa {%a}_sf
next
```

Aqui, o termo *%a*, denominado no *EViews*<sup>®</sup> como "string variable" indica para o programa que ele irá aplicar a fórmula a todas as séries descritas na sequência, seguindo uma de cada vez (qx, y, px, pm, qm). A seguir, estão os comandos para salvar as respectivas séries ajustadas sazonalmente e o fator sazonal.

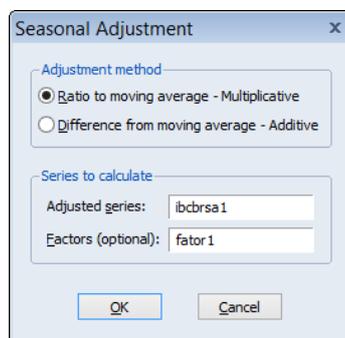
Para rodar o programa, abra o arquivo onde estão os dados. Depois vá em *window/command*. Note que foi aberta uma janela onde pode ser escrito qualquer fórmula ou programa. Assim, para rodar o nosso programa, escreva o comando *run*, seguindo a localização do programa no

computador, como por exemplo, c:\....:

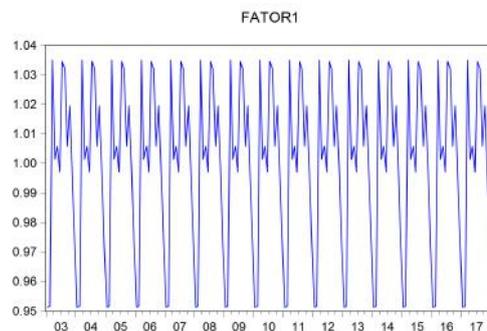
run “localização do programa”

A seguir, aperte o “enter” e o programa será executado. Esse procedimento é válido para todas as demais aplicações a seguir que envolvem a construção de um programa.

Como selecionamos a opção de aparecer o fator sazonal, o *EViews*<sup>®</sup> retorna 12 fatores, um para cada mês, em uma tabela. Para não perder essa informação clique em Freeze, escolha um nome e depois salve a mesma. Esses representam exatamente os fatores sazonais para cada trimestre. Se estivéssemos com dados mensais, seriam 12 fatores. Como nesse método é assumido que os fatores ficam contínuos durante todo o período amostral, o gráfico do padrão sazonal apresenta um fator constante, como pode ser visto na figura 5.3.



(a) Ajuste sazonal - médias móveis



(b) Fator sazonal - médias móveis

Figura 5.3: Gráfico da Sazonalidade - Método das Médias Móveis

Uma forma interessante de representar esse comportamento sazonal é via uma função trigonométrica:

$$qx_t = \alpha_0 + \sum_{k=1}^{S/2} \left\{ \alpha_k \cos\left(\frac{2\pi kt}{S}\right) + \beta_k \sin\left(\frac{2\pi kt}{S}\right) \right\} + \varepsilon_t$$

Onde o  $\alpha_0$  é uma constante que representa a média e  $S$  é o número de componentes sazonais. Suponha o exemplo de uma série trimestral. Nesse caso  $S=4$  e teremos:

$$qx_t = \alpha_0 + \sum_{k=1}^2 \left\{ \alpha_k \cos\left(\frac{2\pi kt}{4}\right) + \beta_k \sin\left(\frac{2\pi kt}{4}\right) \right\} + \varepsilon_t$$

$$qx_t = \alpha_0 + \alpha_1 \cos\left(\frac{2\pi t}{4}\right) + \beta_1 \sin\left(\frac{2\pi t}{4}\right) + \alpha_2 \cos\left(\frac{2\pi 2t}{4}\right) + \beta_2 \sin\left(\frac{2\pi 2t}{4}\right) + \varepsilon_t$$

$$qx_t = \alpha_0 + \alpha_1 \cos\left(\frac{\pi t}{2}\right) + \beta_1 \sin\left(\frac{\pi t}{2}\right) + \alpha_2 \cos(\pi t) + \beta_2 \sin(\pi t) + \varepsilon_t$$

Mas,  $\sin(\pi t) = 0$ , sendo assim, teremos:

$$qx_t = \alpha_0 + \alpha_1 \cos\left(\frac{\pi t}{2}\right) + \beta_1 \sin\left(\frac{\pi t}{2}\right) + \alpha_2 \cos(\pi t) + \varepsilon_t$$

Onde  $t=1,2,3,\dots$ , de acordo com o período amostral, e o comportamento cíclico para as trajetórias anuais e semi-anual é dado por:

$$\cos\left(\frac{\pi t}{2}\right) = 0, -1, 0, 1, 0, -1, \dots$$

$$\text{sen}\left(\frac{\pi t}{2}\right) = 1, 0, -1, 0, 1, 0, -1, \dots$$

$$\cos(\pi t) = -1, 1, -1, 1, \dots$$

Os componentes  $\alpha_1$  e  $\beta_1$  representam a oscilação anual nos dados, ao passo que  $\alpha_2$  representa o componente semi-anual. Para encontrar os valores desses componentes, podemos usar:

$$\alpha_1 = \frac{1}{2}(-m_2 + m_4)$$

$$\beta_1 = \frac{1}{2}(m_1 - m_3)$$

$$\alpha_2 = \frac{1}{2}(-m_1 + m_2 - m_3 + m_4)$$

Vejamos para o nosso um onde  $m_1 = 0.9479$ ,  $m_2 = 1.0105$ ,  $m_3 = 1.035$ ,  $m_4 = 1.0078$ . Sendo assim, teremos:

$$\alpha_1 = \frac{1}{2}(-m_2 + m_4) = \frac{1}{2}(-1.0105 + 1.0078) = -0.00133$$

$$\beta_1 = \frac{1}{2}(m_1 - m_3) = \frac{1}{2}(0.9479 - 1.035) = -0.0438$$

$$\alpha_2 = \frac{1}{2}(-m_1 + m_2 - m_3 + m_4) = \frac{1}{2}(-0.9479 + 1.0105 - 1.035 + 1.0078) = 0.0086$$

E o ciclo que domina todo o processo é o anual ( $\alpha_1$  e  $\beta_1$  são maiores que  $\alpha_2$ ). Ou seja, a frequência é mais forte no ciclo anual. Os valores são pequenos pois o conjunto de dados tem uma pequena sazonalidade, como mostrado pelas médias de cada trimestre. Para construir a série do fator sazonal podemos usar o fato de que:  $\alpha_0 = 1$ ,  $\alpha_1 = -0.00133$ ,  $\beta_1 = -0.0438$  e  $\alpha_2 = 0.0086$  em:

$$qx_t = 1 - 0.00133\cos\left(\frac{\pi t}{2}\right) - 0.0438\text{sen}\left(\frac{\pi t}{2}\right) + 0.0086\cos(\pi t) + \varepsilon_t$$

### 5.1.2 TRAMO/SEATS

Na técnica TRAMO/SEATS<sup>2</sup> de dessazonalização combina dois métodos TRAMO e SEATS para decompor a série em seus componentes não observados. A primeira é similar a uma regressão ARIMA, e é utilizada antes como uma espécie de ajuste dos dados<sup>3</sup>. Ao passo que o SEATS é usado para extrair os sinais da série de tempo, ou seja, os componentes não observados a partir de um modelo aditivo:

$$y_t = TC_t + S_t + I_t$$

Sendo que  $TC_t$  é o componente tendência-ciclo e os demais como dito anteriormente, o componente sazonal e o irregular. Para selecionar esse método, com a série de dados *ibcbr* aberta, vá em **Proc/Seasonal Adjustment/TRAMO/SEATS....** O *EViews*<sup>®</sup> irá abrir uma caixa que contém três diferentes opções. A primeira delas refere-se às especificações básicas. No

<sup>2</sup>TRAMO – Time Series Regression with ARIMA Noise, Missing Observation and Outliers. SEATS – Signal extraction in ARIMA time series.

<sup>3</sup>Mais a frente iremos aprender como são os modelos ARIMA.

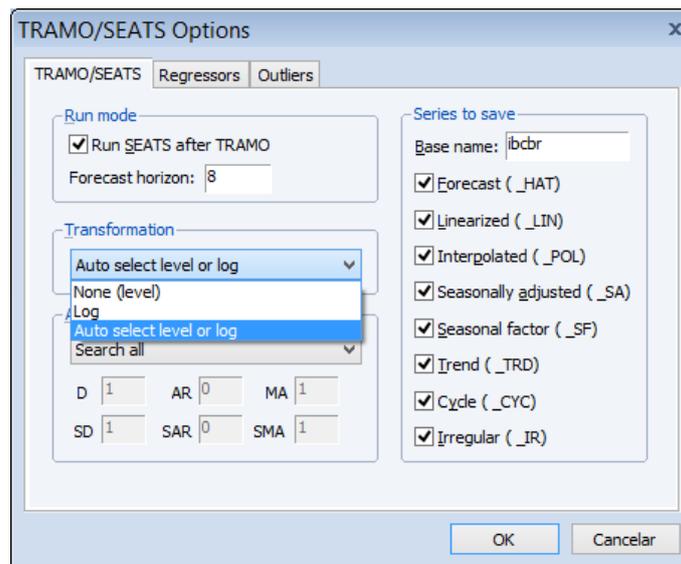


Figura 5.4: Opções TRAMO/SEATS

primeiro bloco, podemos escolher rodar apenas o filtro TRAMO, somente o SEATS ou então esse após o TRAMO, método mais recomendado. Normalmente deixamos o horizonte de previsão tal como o default do *EViews*<sup>®</sup>, sem prejuízo dos resultados finais. Na escolha do modelo ARIMA, podemos determinar que a mesma é feita via seleção de dados em nível, com transformação log ou uma seleção automática. No último bloco podemos deixar o filtro TRAMO escolher a melhor especificação ARIMA ou, então, fazermos a escolha do modelo especificando os parâmetros. Esse ponto é interessante pois muitas séries de dados que são dessazonalizadas por institutos de pesquisas já contém o modelo ARIMA e são fornecidos para uso por parte de terceiros. Nesse caso, se quisermos reproduzir o mesmo modelo, basta imputar os dados, tal que D é o número de diferenciações, AR é o número do componente autorregressivo e MA o de médias móveis. Do lado direito estão as opções para salvar as séries de dados. Podemos escolher todas as opções, encontrar as séries de tendência e do fator sazonal e depois fazer o gráfico. Para extrair o componente sazonal, temos que encontrar o fator sazonal pelo método aditivo. Há duas outras abas com opções que podem ser úteis. A **Regressors** é para especificar se no processo de identificação queremos colocar alguma variável exógena. Na aba **Outliers** podemos escolher se tem algum, especificando o período, ou então deixar que o programa faça a identificação.

Ao clicar em OK o *EViews*<sup>®</sup> irá mostrar um relatório que contém todos os procedimentos, testes e ajustes necessários no processo de estimativa. Podemos salvar esse relatório clicando em Freeze. Atualizando o mesmo poderá ver que o modelo final é da forma (2,1,1)(0,1,1) sem média, sem correções para dias da semana ou páscoa. A ordem dos números mostrados acima é (AR, D, MA)(SAR, SD, SMA) ou seja, temos um modelo ARIMA (2,1,1) com sazonalidade SARIMA (0,1,1). Veremos isso mais a frente. A seguir o *EViews*<sup>®</sup> salva todos os resultados em um grupo de séries. Salve esse como grupo 1 para consulta futura. Agora selecione a série *ibcbr\_trd* e *ibcbr\_sf* e faça um gráfico com dois eixos como mostrado em 5.5.

### 5.1.3 Método Census X-12

Esse é, sem dúvida, um dos métodos de identificação dos componentes de uma série de dados mais utilizado na literatura até o momento. Quando esse é selecionado, é possível identificar várias opções. A primeira delas é a **X-11 Method**. Há pequenas diferenças no uso de cada uma, mas recomenda-se ao leitor que utilize o método *Additive* caso tenha valores negativos ou zero. Na

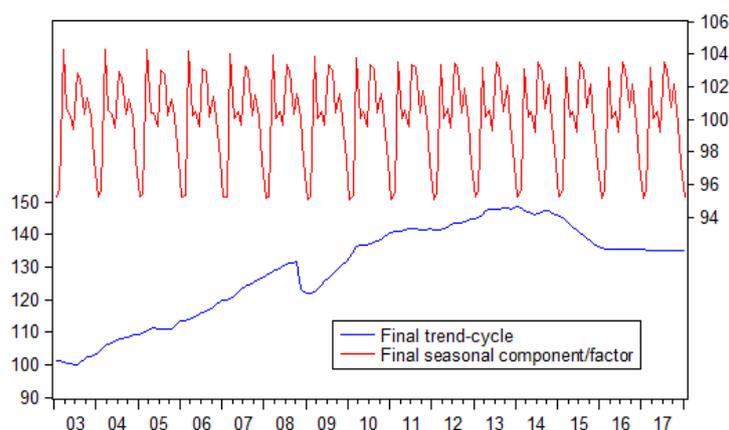


Figura 5.5: Tendência e fator sazonal

opção **Seasonal Filter**, que irá gerar os fatores sazonais, está selecionado como *default* o método **X-12**. Sugere-se fortemente seu uso<sup>4</sup>.

Na opção **Trend Filter**, o leitor poderá notar que o *default* é deixar o *EViews*<sup>®</sup> determinar quantos termos serão utilizados nas médias móveis para extrair a tendência. Em **Component Series to Save**, pode-se especificar o nome da série que será salva além de outros seis componentes. A primeira opção retorna a série ajustada sazonalmente. Além disso, a aplicação do filtro **X-12** permite que se tenha a informação de três importantes características dos dados: os fatores sazonais; a tendência cíclica; o componente irregular. Nos dois últimos estão as opções de ajuste dos fatores ao efeito calendário de dias de negociação no mercado ou para feriados. Esses estão definidos no *EViews*<sup>®</sup> para os feriados dos EUA e páscoa no Canadá. **Quando aplicados, duas opções são selecionadas, abre outras opções na aba Trading Day/Holiday.**

Há três outras abas com diferentes opções. Em **Outliers** podemos especificar se em determinada data haverá um outlier. Na opção **ARIMA Options** escolhemos se há ou não transformação dos dados se teremos repressores exógenos ou se queremos usar alguma amostra de dados na estimativa. Por fim, na aba **Diagnostics** podemos pedir para que seja feita uma análise da sazonalidade e que seja mostrado o diagnóstico dos resíduos, detecção de outliers ou gráficos.

Na aba **Seasonal Adjustment** vamos selecionar o **método multiplicativo**, selecionando as quatro opções de componentes, como mostra a figura 5.6a, e então clique em **Ok**. Não esqueça de mudar o nome base para *ibcbr2* para que a nova estimativa não apague a anterior. Ao fazer a dessazonalização, o *EViews*<sup>®</sup> retorna uma página com diversas descrições do processo implementado. Essa pode ser fechada sem prejuízo da análise futura ou então salva com clicando em **Freeze**. Ao voltar para a página do *workfile*, poderá ver que foram criadas quatro novas séries de dados, todas com o nome da série original mais os termos que representam cada uma dos componentes. No nosso caso: *ibcbr2\_sa*, *ibcbr2\_sf*, *ibcbr2\_tc*, *ibcbr2\_ir*.

A figura 5.6b mostra a evolução do fator sazonal para a série *qx*. Note que o mesmo não é constante ao longo do tempo, sugerindo que a correção pela sazonalidade deve preservar essa diferença.

Esse fator sazonal pode então ser utilizado para dessazonalizar os dados originais. Para tanto, no *workfile*, clique em **Genr/Generate Series by Equation**. Essa opção abre uma janela e permite que

<sup>4</sup>O método X-12 ARIMA é melhor do que o X-11, e incorpora diversos pontos interessantes, como por exemplo, a possibilidade de detectar outliers, mudanças no padrão sazonal, mudanças de nível na série, melhor para tratar com séries de dados com falhas de informação, efeito calendário e testes de diagnóstico.

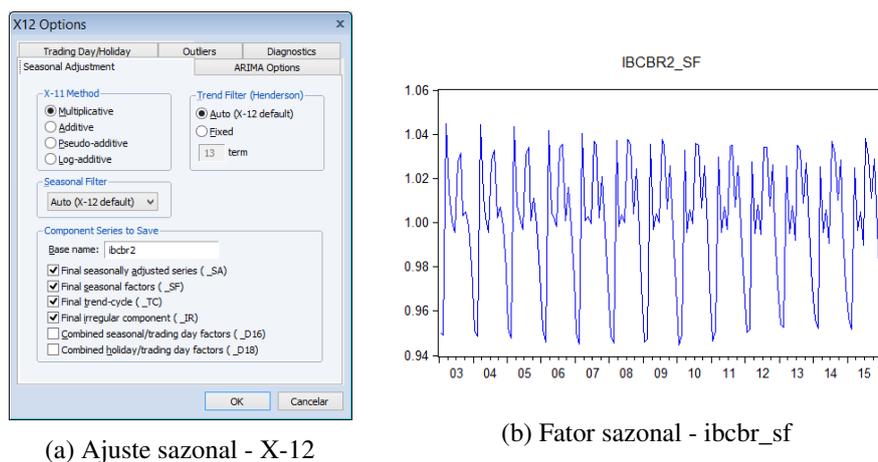


Figura 5.6: Gráfico da Sazonalidade - Método X-12 multiplicativo

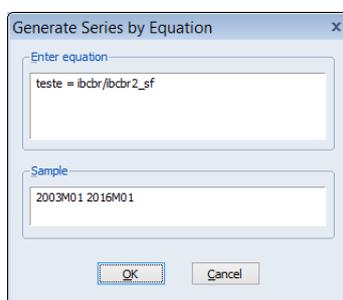


Figura 5.7: Gerar série por equação

se crie uma série nova a partir de outras existentes. Vamos usar um nome “*teste*” para representar essa dessazonalização, tal como mostrado na Figura 5.7, onde vamos dividir a série original pelo fator sazonal:  $teste = \frac{ibcbr}{ibcbr_{sf}}$ . A seguir, clique em **Ok** e confira os dados com os obtidos em *ibcbr\_sa*. O leitor poderá ver que são idênticos.

Os dois outros componentes são a tendência cíclica descrito como *ibcbr\_tc* e o componente irregular *ibcbr\_ir*, cujos gráficos estão dispostos na figura 5.8. Note que, juntamente com os mesmos, foi escolhida a opção **Kernel density** em **Axis borders**. Isso ajuda a compreender como os dados estão distribuídos, possibilitando observar que os resíduos do modelo X-12 ARIMA possuem distribuição normal, tal como esperado.

De forma geral, o que obtemos aqui é uma decomposição da nossa série original em 3 importantes fatores: (i) fator sazonal; (ii) tendência cíclica; (iii) componente irregular. Assim, também podemos obter a série original a partir desses 3 fatores, basta fazer:

$$ibcbr = ibcbr_{sf} * ibcbr_{tc} * ibcbr_{ir}.$$

Nesse caso, escolhemos um nome para essa nova série “teste”, e construímos uma fórmula para ela a partir da multiplicação dos três componentes anteriores. A seguir, o leitor poderá ver que foi criada uma série de nome “teste” no *workfile*.

Além de determinar os componentes de uma série de tempo, o *EViews*<sup>®</sup> também faz a correção sazonal dos dados, como o leitor pode ver no *workfile* a partir da série *ibcbr2\_sa*. Essa é obtida dividindo-se a série original pelo seu fator sazonal:

$$ibcbr2_{sa_t} = \frac{ibcbr2_t}{ibcbr2_{sf_t}}$$

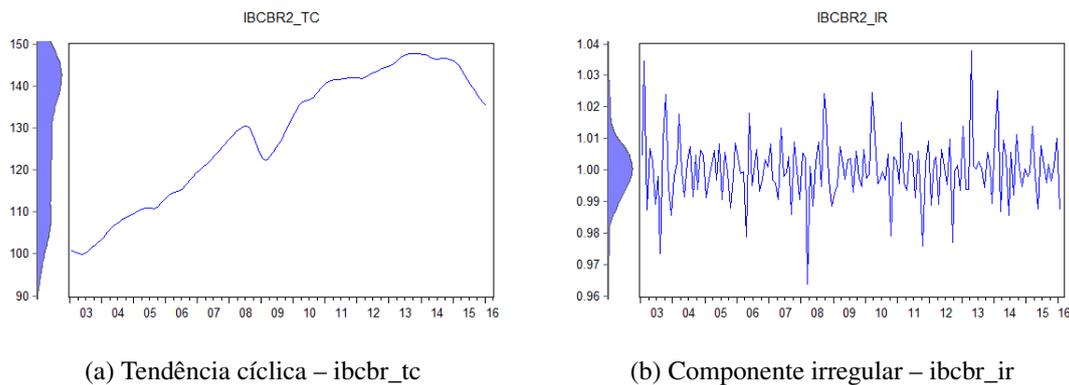


Figura 5.8: Gráfico dos componentes da série `ibcbr`

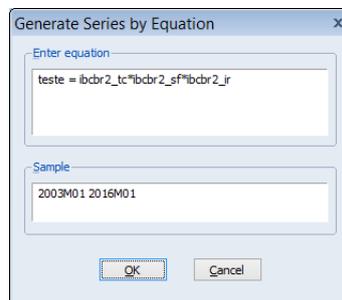


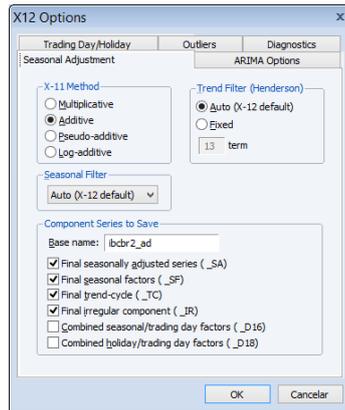
Figura 5.9: Gerar série por equação

**Programação 5.1.2** O método de dessazonalização X-12 possui muitas opções. A forma mais básica pode ser aplicada como a seguir, seguindo o exemplo utilizado para as médias móveis. Assim, usamos um procedimento para aplicar tanto o método das médias móveis quanto o X12 a várias séries ao mesmo tempo. Abra o mesmo programa de antes e agora acrescente o termo para a dessazonalização pelo X-12. Depois, vá ao arquivo original e rode o mesmo.

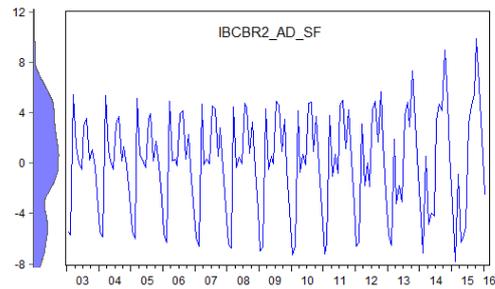
```
for %a qx y px pm qm
seas(m) {%a} {%a}_sa {%a}_sf
{%a}.x12(mode=m, filter=msr, save= "d10 d11 d12") {%a}_x12
next
```

Nesse caso, usamos o método multiplicativo (se quiser usar o método aditivo mude de 'm' para 'a'), como filtro selecione o X-12 e salvamos, respectivamente, os fatores sazonais (d10), as séries ajustadas sazonalmente (d11) e a tendência cíclica (d12). Ao final, o termo `{%a}` serve para usar o nome da série como nome base. Por exemplo, quando o programa estiver aplicando a dessazonalização para a série `qx`, irá salvar a série de fatores sazonais como `qx_sf`.

Uma alternativa de dessazonalização é via **método aditivo**. Nesse caso, também podemos selecionar a opção de obter os três fatores: (i) fator sazonal; (ii) tendência; (iii) componente irregular, além da série ajustada sazonalmente. Algumas diferenças de resultado aparecerão entre o método multiplicativo e aditivo, como por exemplo, o fator sazonal e o irregular. Mas, a série ajustada sazonalmente irá produzir resultados semelhantes. Porém, ao invés de dividir a série original pelo seu fator sazonal, como feito no método multiplicativo, agora diminuimos a mesma de seu fator sazonal.



(a) Ajuste sazonal - X-12



(b) Fator sazonal - qx\_ad\_sf

Figura 5.10: Gráfico da Sazonalidade - Método X-12 aditivo

**Programação 5.1.3** Como forma de complementar a análise das séries quando se tem mais de uma, pode ser mais útil agrupar as mesmas. Nesse caso, recorreremos ao comando “**group**”, como mostrado abaixo, onde agrupamos todas as séries ajustadas sazonalmente no seu banco de dados feitas anteriormente em um grupo de nome “ajustada”.

```
group ajustada qx_x12_sa px_x12_sa pw_x12_sa pr_x12_sa y_x12_sa
```

Para ver como isso ocorre repita os passos anteriores mas escolha o método aditivo. Para que as séries salvas sejam diferentes das anteriores, use um subíndice “ad”, tal como mostrado na figura 5.10a:

Note que o resultado do fator sazonal oscila em torno do valor zero, ao invés de oscilar em torno de 1, como no método multiplicativo mas, o resultado é o mesmo. A seguir, use:

$$ibcbr2\_ad\_sa_t = ibcbr2_t - ibcbr2\_ad\_sf_t$$

Para encontrar a série ajustada sazonalmente. Por fim, para obter a série original não multiplicamos os três fatores tal como no caso do método multiplicativo e, sim, somamos os mesmos:

$$ibcbr2_t = ibcbr2\_ad\_sf_t + ibcbr2\_ad\_tc_t + ibcbr2\_ad\_ir_t$$

Há diversas opções que o *EViews*<sup>®</sup> permite aplicar no ajuste sazonal. Em especial, e muito comum para o Brasil, seria um ajuste que considerasse os feriados. Apesar de disponibilizar essa opção em **Trading day/Holiday**, a mesma está formatada para feriados nos EUA.

**Programação 5.1.4** O gráfico do fator sazonal para cada uma das séries do seu banco de dados pode ser solicitado. Nesse caso, usamos o objeto “graph”. Como queremos um gráfico de linha, usamos o comando “line”. Por fim, é especificada a série que será feito o gráfico. Nesse caso, `{%a}_x12_sf`.

```
for %a qx y px pm qm
seas(m) {%a} {%a}_sa {%a}_sf
{%a}.x12(mode=m, filter=msr, save= "d10 d11 d12") {%a}_x12
graph gra {%a}x12.line {%a}_x12_sf
next
```

**Programação 5.1.5** Alternativamente, podemos estar interessados em avaliar como fica cada uma das séries ajustadas sazonalmente a partir de dois diferentes métodos. Nesse caso, aplicamos o método das médias móveis e depois o X-12. Em ambos, fazemos tanto a sazonalidade aditiva quanto multiplicativa. A seguir, é calculada a correlação entre as séries ajustadas sazonalmente e o resultado é armazenada em uma tabela de nome `correl`.

```

scalar sum=1
table(3,4) correl
correl(2,1)="aditivo"
correl(3,1)="multiplicativo"
correl(1,2)="ctotal"
correl(1,3)="preco"
correl(1,4)="renda"
for %a qx y px pm qm
seas(a) {%a} {%a}asa {%a}asf
seas(m) {%a} {%a}msa {%a}msf
{%a}.x12(mode=a, filter=msr, save= "d10 d11 d12") {%a}a
{%a}.x12(mode=m, filter=msr, save= "d10 d11 d12") {%a}m
correl(2,sum+1)=@cor({%a}asa,{%a}a_sa)
correl(3,sum+1)=@cor({%a}msa,{%a}m_sa)
sum=sum+1
next

```

#### 5.1.4 Método Census X-13

Esse é um dos mais novos métodos de dessazonalização disponível e que foi desenvolvido pelo U.S. Census. Sua aplicação deve ser feita apenas para dados mensais ou trimestrais sendo necessário ter, ao menos, três anos completos de dados. Com a série de dados `ibcbr` aberta clique em **view/seasonal adjustment/censos x-13...** A caixa de diálogo que aparece, como mostrado na Figura 5.11, permite especificar aspectos da variável, como alguma transformação que tenha sido feita do tipo log ou logit, determinar o modelo ARIMA, escolher o método de ajuste sazonal e os resultados a serem mostrados.

A opção **X-13 built in regressors** permite inserir uma constante no modelo, sazonalidade via dummy ou trigonometricamente, especificar os dias de negociação, determinar os feriados ou escolher o tipo de outlier que pode ser usado no processo de estimativa. A seguir podemos especificar, em **User-defined regressors**, se queremos usar alguma variável exógena para melhorar o modelo proposto. Na opção ARIMA podemos escolher o tipo de modelo, caso se tenha um conhecimento prévio, selecionando a opção “manual”. Nesse caso, os parâmetros são (p, d, q)(P, D, Q) com as letras minúsculas representando o componente ARIMA e as letras maiúsculas os componentes sazonais. Por exemplo, a série do PIB trimestral do IBGE para “serviço de informação” tem uma decomposição dos componentes do modelo ARIMA aditivo e dado da forma (0,1,1)(0,1,1). Já a série da indústria de transformação tem um método aditivo do tipo (2,1,0)(0,1,1) mas com três intervenções dummy: AO 1996.3 – representa uma dummy aditiva no mês de março de 1996; LS 2008.4 – é dada por uma dummy do tipo “level-shift” mudança no nível no mês de abril de 2008; TC 2009.1 – é uma dummy definida como “constant-level-change”, ou seja, uma mudança no nível em janeiro de 2009. Todas essas intervenções podem ser facilmente inseridas via **X-13 built in regressors**, juntamente com ARIMA model e escolhendo “manual” e colocando (2,1,0)(0,1,1).

Logo abaixo da opção “manual”, está a opção “X-11 Auto”. Nessa o *EViews*<sup>®</sup> irá estimar todos os modelos que estão especificados na lista (você pode inserir mais opções) e modificar as opções de escolha. Uma opção interessante é fazer a especificação “with limits” que irá estimar

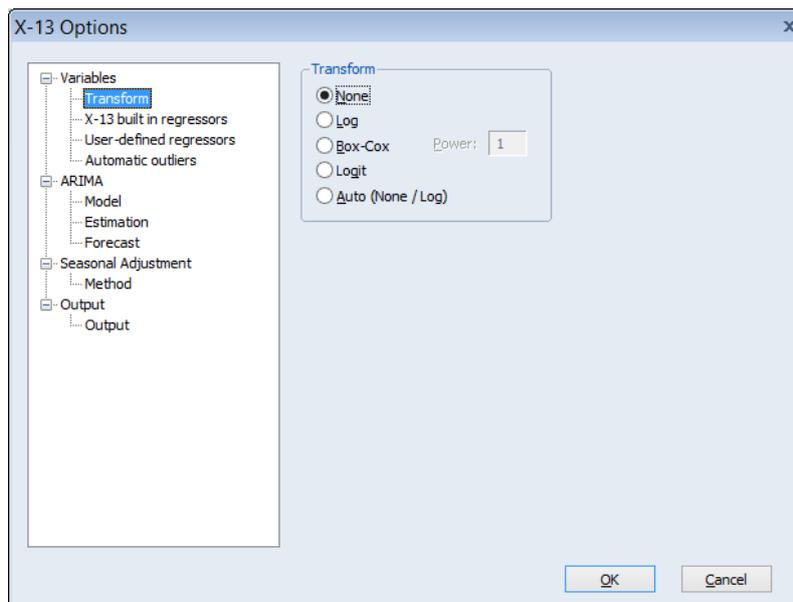


Figura 5.11: Opções do Método X-13

todas as possíveis combinações de modelos com AR, MA e D especificadas. Veremos um exemplo mais a frente. Por fim temos a opção TRAMO Auto que permite a escolha o modelo ARIMA e diferenciação máximos. A seguir, temos as opções de estimação do modelo ARIMA, onde escolhemos os critérios e o período a ser utilizado. Na opção **ARIMA forecast**, podemos usar o melhor modelo encontrado para prever dados futuros com base em suas características encontradas.

Na opção **Seasonal Adjustment** podemos escolher o método de ajuste sazonal entre x-11 ou SEATS. Se escolher “none”, não será feito nenhum ajuste sazonal na série, sendo apenas escolhido o melhor modelo ARIMA. A última escolha é para **Output**, onde selecionamos os resultados a serem mostrados. Note os códigos das séries resultantes: D11 – dados ajustados sazonalmente; D12 – tendência; D10 – fator sazonal; D13 – componente irregular. Vejamos como usar essas opções do x-13 na série do *ibcbr*. Abra a mesma e clique em **Proc /Seasonal Adjustment /Census X-13 ...**. Nas opções **Variables** vamos deixar como default, ou seja, os dados não possuem transformação (transform option); não é feita intervenção no **X-13 built in regressors**; não usamos variável exógena e, por fim, não usamos **Automatic outliers**. Como primeiro passo vamos investigar qual seria o melhor modelo ARIMA para descrever a sazonalidade de *ibcbr*. Na opção **ARIMA /Model ...** selecione TRAMO Auto tal como mostrado na Figura 5.12.

Após clicar em OK será produzido um relatório de resultados. Sugiro fortemente olhar o relatório, pois ali irá constar as características do melhor modelo final selecionado, que é da forma  $(3,1,1)(0,1,1)$ . Esse processo convergiu após 64 iterações, tendo sido investigadas 415 funções. Os coeficientes estimados e os erros padrão também são fornecidos, bem como alguns critérios que são utilizados para comparar modelos, como AIC, BIC e Hannan–Quinn (veremos isso mais a frente). Como não fizemos nenhuma seleção adicional, o *EViews*<sup>®</sup> irá retornar a série *ibcbr\_d11* que descreve os dados ajustados sazonalmente. Para ver os demais resultados das séries vá em output e escolha D\_12, D\_10 e D\_13.

Uma opção interessante é tentar identificar se existe outlier ou não no modelo ARIMA. Com a série *ibcbr* aberta selecione X-13 e na opção **Automatic outliers** clique em “Temporary change” (TC). Mantenha todo o período amostral e o processo de seleção “One at a time”, como mostrado na Figura 5.13a. A seguir, em **ARIMA /Model ...**, clique em manual e especifique o modelo  $(3,1,1)(0,1,1)$ , tal como na Figura 5.13b. Por fim, em **Output** selecione todas as opções e clique

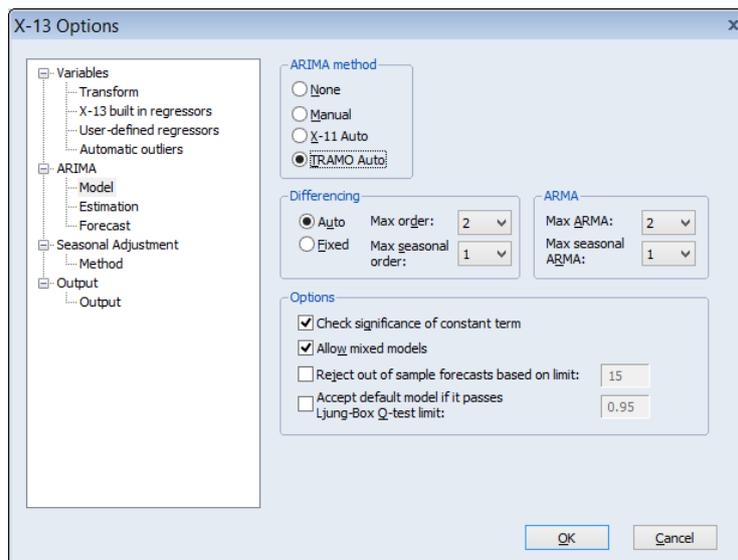
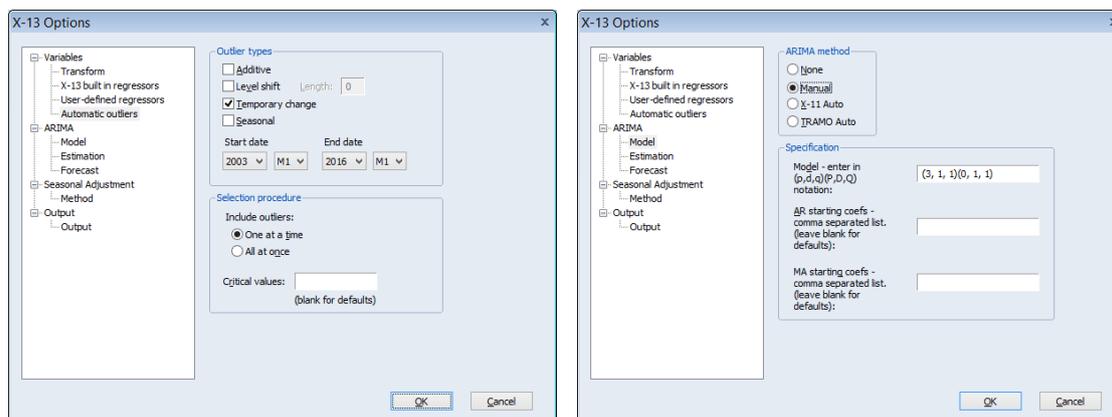


Figura 5.12: Métodos ARIMA em X-13



(a) Automatic outliers

(b) ARIMA Model

Figura 5.13: Identificação de outliers

em OK. Na página de resultados são mostradas as estimativas para o outlier do tipo TC. Primeiro veja o valor crítico,  $|t| \leq 3,91$ . A data com o resultado do t-valor mais alto é dezembro/2008 com  $t = -3,60$ . Note que esse resultado fica dentro do intervalo de confiança especificado  $-3,91 \leq t \leq 3,91$  e, com isso, aceitamos a  $H_0$  de não existência do outlier do tipo TC. Dado que ele não é significativo, as estimativas seguintes são testes sem a presença do outlier.

### 5.1.5 Alisamento Exponencial

Esse recurso é muito útil para fazer previsões, especialmente para séries de dados de curta periodicidade, e com a vantagem de que seus coeficientes são atualizados a cada momento, não permanecendo fixos ao longo do processo<sup>5</sup>. Há dois tipos de alisamento disponíveis no *EViews*<sup>®</sup>, o SES - *Simple Exponential Smoothing* e o ETS - *Exponential Smoothing*.

Vamos exemplificar seu uso com a série de dados qx, que tem periodicidade trimestral. Com a série qx aberta, selecione **Proc/Exponential Smoothing/ Simple Exponential Smoothing...** Como mostra a figura 5.14, o *EViews*<sup>®</sup> permite que se escolha dentre 5 diferentes opções de

<sup>5</sup>Porém, no processo de previsão, os mesmos tornam-se fixos.

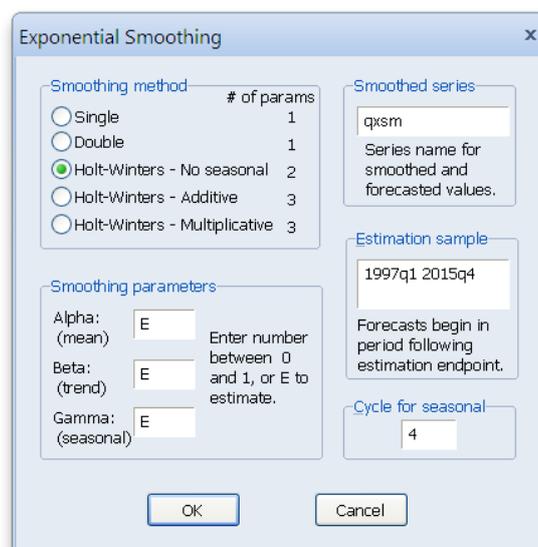


Figura 5.14: Alisamento exponencial da série qx

método para se fazer o alisamento exponencial

Além de selecionar o método, o leitor também tem a opção de determinar os parâmetros a serem utilizados ou então deixar a letra **E** para que o *EViews*<sup>®</sup> estime os mesmos. Valores próximos a zero significam que informações passadas são importantes para determinar o futuro. Ao passo que, valores mais próximos de 1 representam um comportamento tipo *random walk*, onde apenas a última informação é útil para prever o futuro. Recomenda-se deixar o *EViews*<sup>®</sup> estimar o valor dos parâmetros. Na tabela 5.1 está uma descrição das equações e aplicações desses diferentes métodos.

Tabela 5.1: Diferentes métodos de alisamento exponencial

Método	Equação	Aplicação
Simple	$\hat{y}_t = \alpha y_t + (1 - \alpha)\hat{y}_{t-1}$	Em séries sem constante, tendência ou sazonalidade.
Duplo	$\hat{y}_t = \alpha y_t + (1 - \alpha)\hat{y}_{t-1}$ $\hat{D}_t = \alpha \hat{y}_t + (1 - \alpha)\hat{D}_{t-1}$	Melhor para séries com tendência linear. Aplica o método simples duas vezes.
Holt-winters – sem sazonalidade	$\hat{y}_{t+k} = \alpha + tk$	Útil para séries com tendência linear e sem sazonalidade.
Holt-winters –mult.	$\hat{y}_{t+k} = (\alpha + tk)c_{t+k}$	Útil para séries com tendência linear e sazonalidade multiplicativa.
Holt-winters – adit.	$\hat{y}_{t+k} = \alpha + tk + c_{t+k}$	Útil para séries com tendência linear e sazonalidade aditiva.

Nota: o termo  $\alpha$  é o parâmetro de alisamento,  $t$  é a tendência e  $c$  a sazonalidade.

Do lado direito da janela que será aberta, o *EViews*<sup>®</sup> sugere um nome para a série alisada “qxsm”. Logo abaixo tem o período de especificação da amostra. Se deixarmos como data final 2015Q4, o *EViews*<sup>®</sup> irá fazer a previsão a partir desse ponto. O problema com essa escolha é que, após feita a previsão, não há informação verdadeira para comparar com essa previsão. Portanto, se

o objetivo é apenas prever, tudo bem, podemos usar como data a última observação. Por outro lado, se o objetivo é testar essa previsão, o melhor seria determinar uma data anterior ao final, reservando dados para comparação.

Por fim, tem a opção do **Cycle for seasonal**. Note que, para esse exemplo, temos o número 4, que representa a quantidade de trimestres no ano. Se os dados forem mensais, o *EViews*<sup>®</sup> irá retornar o número 12. Caso o leitor tenha dados sem periodicidade, ou então dados diários do mercado financeiro, pode escolher um número diferente. Como exemplo, vamos estimar cada um dos cinco métodos para a série *qx*. Para o primeiro método, selecionamos a série de resultado como *qxsm1*. Para o segundo método, *qxsm2* e assim sucessivamente. Além disso, vamos deixar quatro trimestres de dados para comparar com as previsões, digitando como data final em **estimation sample**, 2014q4. Os resultados são mostrados na tabela 5.2.

Tabela 5.2: Resultados do alisamento exponencial para *qx*

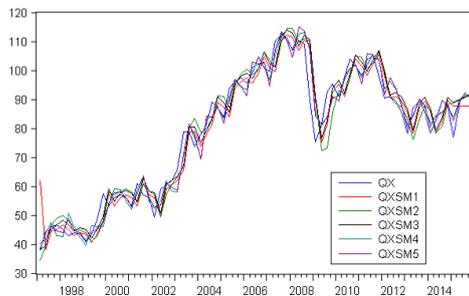
		Simple	Duplo	Holt-winters sem sazonalidade	Holt-winters adit.	Holt-winters –mult.
Parameters:	Alpha	0.9990	0.4680	1.0000	1.0000	1.0000
	Beta			0.0100	0.0000	0.0000
	Gamma				0.0000	0.0000
Sum of Squared Residuals		2963.3370	3060.6500	2413.8800	1353.1410	1339.7440
Root Mean Squared Error		6.4154	6.5199	5.7902	4.3352	4.3136
End of Period Levels:	Mean	87.7719	87.8750	87.7700	87.4709	86.8356
	Trend		0.9136	1.0562	0.6047	0.6047
	Seasonals:					
	2014Q1				-4.2175	0.9429
	2014Q2				0.8709	1.0105
	2014Q3				3.0475	1.0358
	2014Q4				0.2991	1.0108

Note que o coeficiente  $\alpha$  varia de um valor mínimo de 0,48 a um máximo de 1 sinalizando que as informações passadas são úteis para prever o comportamento futuro. O valor zero para o parâmetro beta e gamma indicam que os mesmos foram constantes. Por exemplo, para o modelo simples, temos:  $\hat{y}_t = 0.999y_t + (1 - 0.999)\hat{y}_{t-1}$ . Também são fornecidas duas estatísticas de previsão que podem ser utilizadas para comparar os diferentes modelos: SSR – soma do quadrado dos resíduos<sup>6</sup>; RMSE – raiz do erro quadrado médio<sup>7</sup>. Comparando essas para os cinco modelos, podemos ver que praticamente não há diferenças entre o Holt-Winters aditivo e o multiplicativo e que, ambos, são os melhores modelos de previsão. Os valores de média e de tendência fornecidos

$${}^6 SSR = \sum_{t=1}^n (y - \hat{y})^2$$

$${}^7 RMSE = \sqrt{\frac{\sum_{t=1}^n (y - \hat{y})^2}{n}}$$

para o fim do período são usados para fazer a previsão, bem como a estimativa da sazonalidade. A figura 5.15 mostra o comportamento da estimativa (entre 1997Q1 e 2014Q4) e também da previsão para os quatro trimestres à frente 2015Q1 a 2015Q4, usando esses cinco métodos, juntamente com o resultado de verdadeiro de qx.



(a) Previsão dos dados

	2015Q1	2015Q2	2015Q3	2015Q4
Simples	87.7719	87.7719	87.7719	87.7719
Duplo	88.7886	89.7022	90.6159	91.5295
Holt-winters sem sazonalidade	88.8262	89.8825	90.9387	91.9949
Holt-winters adit.	83.8581	89.5513	92.3326	90.1889
Holt-winters -mult.	82.4517	88.9725	91.8198	90.2149
qx	76.8533	88.7567	-	-

(b) Valores previstos

Figura 5.15: Alisamento exponencial da série qx

Note que, pelo método mais simples, que não considera a presença de tendência e sazonalidade, as previsões são constantes, e refletem exatamente o valor da média (87,7719) obtido para o último período da estimativa (2014Q4). Já o método Duplo tem a influência de uma tendência de valor 1,0167. Nesse caso, o resultado para a primeira previsão é dado por:

$$duplo_{2015q1} = \text{média} + \text{tendência}$$

$$duplo_{2015q1} = 87,8750 + 0,9136 = 88,7886$$

No segundo momento, a previsão passa a diferir apenas na magnitude da tendência. Ou seja, usa-se a estimativa da média do momento anterior e, com base nela, é somada a tendência. Assim, a previsão do próximo trimestre é dada por:

$$duplo_{2015q2} = 88,7886 + 0,9136 = 89,7022$$

Ou então:

$$duplo_{2015q2} = 87,8750 + 2 * 0,9136 = 89,7022$$

E assim sucessivamente para mais períodos a frente:

$$duplo_{2015q3} = 87,8750 + 3 * 0,9136 = 90,6159$$

No modelo Holt-Winters sem sazonalidade, também há dois resultados para se fazer a previsão, a média e a tendência. E basta fazer a previsão para os trimestres a frente como fizemos no método duplo. Porém, os dois últimos métodos contemplam a presença da sazonalidade. Nesse caso, as previsões devem considerar essa influência em seus respectivos trimestres. Por exemplo, no Holt-Winters aditivo, a previsão para 2014Q1 é dada pela soma da média, da tendência e também da sazonalidade do primeiro trimestre:

$$HWaditivo_{2015q1} = \text{média} + \text{tendência} + \text{sazonalidade}_{2014q1}$$

$$HWaditivo_{2015q1} = 87,4709 + 0,6047 - 4,2175 = 83,8581$$

Na previsão do segundo trimestre, multiplicamos a tendência por 2 e aplicamos a sazonalidade de 2015q2;

$$HWaditivo_{2015q2} = \text{média} + 2 * \text{tendência} + \text{sazonalidade}_{2014q2}$$

$$HW_{aditivo}_{2015q2} = 87,4709 + 2 * 0,6047 + 0,8785 = 89,5513$$

E assim sucessivamente, sempre aplicando um multiplicador para a tendência e considerando o fator sazonal do respectivo trimestre que está sendo feita a previsão. Por exemplo, se quisermos fazer essa previsão para 6 trimestres à frente, usamos:

$$HW_{aditivo}_{2016q2} = 87,4709 + 6 * 0,6047 + 0,8785 = 91,9701$$

Por fim, temos o método Holt-Winters multiplicativo. Nesse caso, a sazonalidade é multiplicativa, e fazemos a previsão para 2015Q1 da seguinte forma;

$$HW_{multiplicativo}_{2015q1} = (\text{média} + \text{tendência}) * \text{sazonalidade}_{2014q1}$$

$$HW_{multiplicativo}_{2015q1} = (86.8356 + 0.6047) * 0.9429 = 82.4517$$

Para prever o segundo trimestre, multiplicamos a tendência por 2 e consideramos a sazonalidade de 2015q2:

$$HW_{multiplicativo}_{2015q1} = (\text{média} + 2 * \text{tendência}) * \text{sazonalidade}_{2014q1}$$

$$HW_{multiplicativo}_{2015q1} = (86.8356 + 2 * 0.6047) * 1.0105 = 88.9725$$

**Programação 5.1.6** O método de alisamento exponencial permite que sejam escolhidas cinco diferentes alternativas (**s,d,n,a,m**), e que seguem respectivamente as opções de escolha entre os modelos **simple**, **duplo**, **Holt-winters no seasonal**, **Holt-winters seasonal aditivo** e, por último o **multiplicativo**.

Para usar o método multiplicativo em uma única série de dados, como por exemplo, *qx*, deixando que os parâmetros sejam estimados, usamos:

```
qx.smooth(m, e, e, e) qxsm1
```

Alternativamente, dando sequencia ao programa anterior, podemos determinar que o alisamento exponencial seja feito para uma sequencia de séries de dados. Nesse caso, apenas acrescentamos ao nosso programa a opção abaixo:

```
for %a qx y px pm qm
seas(m) {%a} {%a}_sa {%a}_sf
{%a}.x12(mode=m, filter=msr, save= "d10 d11 d12") {%a}_x12
graph gra {%a}x12.line {%a}_x12_sf
{%a}.smooth(m, e, e, e) {%a}sm1
next
```

Alternativamente, para uma única série de dados, podemos ver como se comportam as previsões a partir dos cinco diferentes métodos de alisamento exponencial. Nesse caso, usando a série *qx* de exemplo, o “*loop*” pode se modificar para:

```
for %a s d n a m
smooth({%a}, e, e, e) qx qx{%a}
next
```

Ou então, podemos pedir que os cinco métodos sejam aplicados para cada uma das séries de dados que temos. Nesse caso, podemos usar um comando “for” dentro de outro comando “for”:

```

for %b qx y px pm qm
  for %a s d n a m
    smooth({%a},e,e,e) {%b} {%b}{%a}
  next
next

```

**Programação 5.1.7** Alternativamente, podemos fazer os cinco diferentes métodos de alisamento exponencial, para cada uma das séries de dados e, a seguir, armazena o resultado do RMSE em uma tabela de nome “alisa”.

```

table(6,4) alisa
alisa(2,1)="single"
alisa(3,1)="doble"
alisa(4,1)="no seas"
alisa(5,1)="HW-no seas"
alisa(6,1)="HW-seas"
alisa(1,2)="ctotal"
alisa(1,3)="preco"
alisa(1,4)="renda"
scalar sum=1
scalar numero=1
for %b ctota preco renda
  for %a s d n a m
    smooth({%a},e,e,e) {%b} {%b}alisa{%a}
    alisa(sum+1,numero+1)=@rmse({%b},{%b}alisa{%a})
    sum=sum+1
  next
  numero=numero+1
  scalar sum=1
next

```

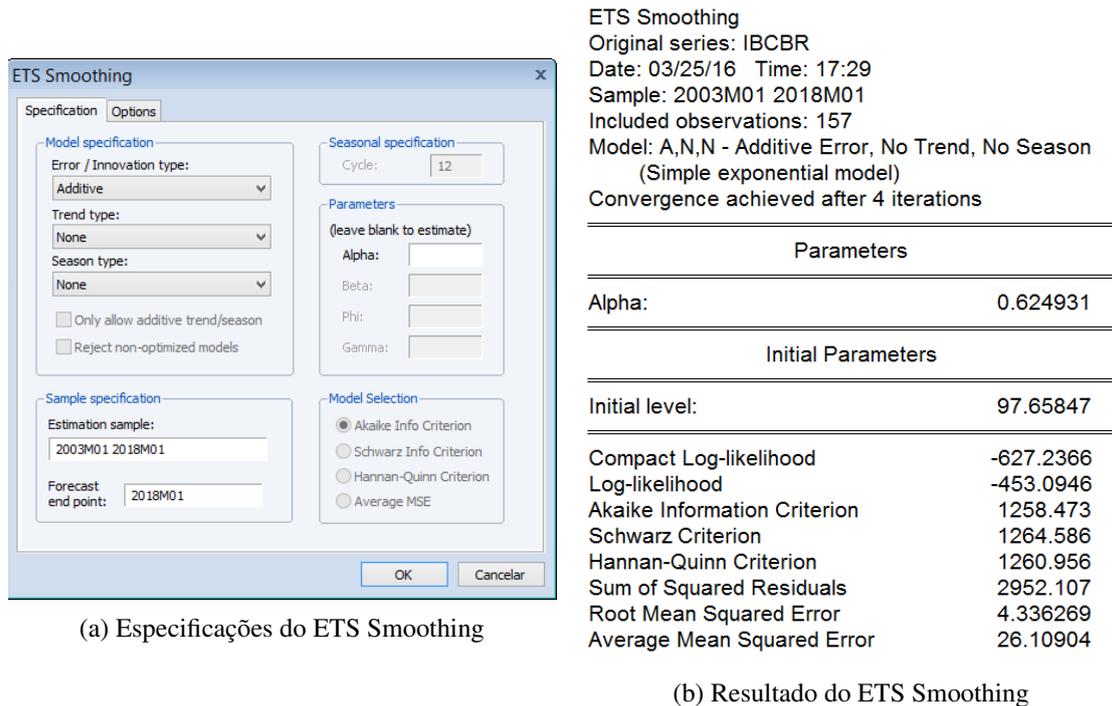
## 5.2 ETS-ERROR-trend-seasonal

Os modelos ETS são bem mais complexos e eficientes que a proposta anterior (ES) e se diferenciam por incorporar o erro de previsão do passo anterior para melhorar a estimativa no momento presente. Aqui a ideia é decompor a série de dados em três componentes T - tendência; S - sazonalidade e I - componente irregular, ou resíduo. Tal como visto anteriormente nos métodos de dessazonalização, aqui podemos ter modelos aditivos e multiplicativos, ou então combinados totalizando 30 diferentes tipos de modelos. Em resumo temos:

- Modelo aditivo puro:  $y = T + S + I$
- Modelo multiplicativo puro:  $y = T \times S \times I$
- Modelo misto:  $y = (T \times S) + I$

### Modelo ANN

Vejam como é o modelo mais simples de todos, dado por A, N, N (erro aditivo, sem tendência, sem sazonalidade) aplicado ao IBC-Br. Para estimá-lo, abrimos a série `ibcbr` e vamos em **Proc /Exponential Smoothing /ETS Exponential Smoothing ...**, abrindo a caixa de diálogo apresentada Figura 5.16a, a qual também já apresenta as configuração utilizadas para o modelo mais simples. O resultado é tal como mostrado na figura 5.16b. De forma geral, teremos que



(a) Especificações do ETS Smoothing

(b) Resultado do ETS Smoothing

Figura 5.16: ETS Smoothing Simples

$\hat{y}_t = \hat{y}_{t-1} + \alpha \varepsilon_{t-1}$ . Dito de outra forma, nossa previsão é corrigida pelo erro de previsão do passo anterior.

Ali temos a estimativa do parâmetro  $\alpha=0,624931$  e o valor inicial de 97.65847. Logo abaixo temos diversas estatísticas de comparação de modelos. Você deve estar se perguntando: “de onde vem esse valor inicial?”. E o alfa? Aqui começamos a ter o primeiro contato com o processo de maximização em série de tempo e iteração. Para começar o modelo precisamos de um valor inicial, a semente, e um valor de  $\alpha$ . O valor inicial é para representar a previsão do primeiro mês que, no nosso exemplo, é de  $y_1=96,15$ . O valor de  $\alpha$  é para encontrar a evolução da nossa estimativa. Suponha um valor inicial de  $\hat{y}_1 = 97,6584$ . Com esse encontramos um erro de previsão  $\varepsilon$  de:

$$\begin{aligned} y_1 - \hat{y}_1 &= \varepsilon_1 \\ 96,15 - 97,65 &= -1,508 \end{aligned}$$

Considerando  $\alpha = 0,6249$ , podemos fazer:

$$\begin{aligned} \hat{y}_2 &= \alpha y_1 + (1 - \alpha) \hat{y}_1 \\ \hat{y}_2 &= (0,62) \cdot 96,15 + (0,38) \cdot 97,65 = 96,71 \end{aligned}$$

Como podemos prever o momento  $\hat{y}_2$ ? Precisaremos do valor de  $\alpha$ . Veja que, para prever o passo atual, usamos a informação verdadeira em  $t - 1$  e a estimada ou então, a previsão em  $t + 1$  pode ser encontrada aplicando  $\hat{y}_2 = \hat{y}_1 + \alpha \varepsilon_1$  ou  $\hat{y}_2 = 97,65 + 0,62(-1,508)$ . Agora podemos encontrar o erro de previsão no passo 2 ( $\varepsilon_2$ ) da mesma forma que antes,

$$\begin{aligned} \varepsilon_2 &= y_2 - \hat{y}_2 \\ &= 98,67 - 96,71 \\ &= 1,954, \end{aligned}$$

ou então, usando o erro de previsão anterior:  $\hat{y}_3 = \hat{y}_2 + \alpha \varepsilon_2 = 96,71 + 0,62(1,95) = 97,93$ .

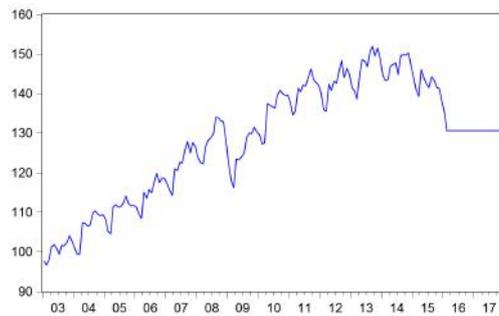


Figura 5.17: Previsão simples no modelo ETS

Fazemos isso sucessivamente e encontraremos diversos erros de previsão. Esses são utilizados para encontrar as estatísticas SSR, RMSE e AMSE bem como o valor do log verossimilhança (veremos isso mais a frente). Porém, o que garante que o valor inicial que usamos (97,65) e o  $\alpha = 0,624931$  produzem o melhor modelo? Automaticamente, no processo de iteração são testadas combinações de diferentes valores iniciais com diferentes valores de  $\alpha$ , até que se encontre aquele que gera o menor erro. Esse é o produto final mostrado nas estimativas.

Agora vamos ver como fica a previsão “n” passos a frente:

$$\begin{aligned}\hat{y}_{2016M2} &= \alpha y_{2016M1} + (1 - \alpha)\hat{y}_{2016M1} \\ &= (0,62)127,92 + (0,38)135,20 \\ &= 130,6514.\end{aligned}$$

Daí em diante, como não há mais valor conhecido, a previsão será dada por:

$$\hat{y}_{2016M3} = y_{2016M2} = 130,6514.$$

E nosso gráfico de previsão é tal como mostrado na Figura 5.17.

### Modelo MAN

Esse modelo também é conhecido como método de holt com erros multiplicativos e uma tendência aditiva. Na caixa de diálogo do ETS Smoothing selecionamos Multiplicative em Erros / Innovation type, Additive em Trand Type e deixamos None em Seasonal Type, conforme a Figura 5.18a. A Figura 5.18b apresenta os resultados desse modelo.

Note que agora temos um parâmetro adicional,  $\beta$  e valor inicial para a tendência em 0,312577. Todas as demais estatísticas de comparação são como antes. Com a incorporação da tendência a previsão no momento t fica da forma:

$$\hat{y}_t = (\hat{y}_{t1} + T_{t-1}) + \alpha \varepsilon_{t1}$$

Quando a taxa de crescimento do componente tendência for zero, ou seja,  $\beta = 0$ . O valor inicial estimado para a tendência é  $T = 0,312577$  e o valor do nível inicial é 97,44160. Sendo assim, nosso valor inicial é dado por:

$$\begin{aligned}\hat{y}_1 &= N_1 + T_1 \\ &= 97,4416 + 0,312577 = 97,7541\end{aligned}$$

Como temos uma tendência, essa deve ser incorporada na previsão dos passos seguintes e, também devemos usar o erro de previsão do passo anterior para melhorar o modelo no passo

ETS Smoothing

Specification Options

Model specification

Error / Innovation type: Multiplicative

Trend type: Additive

Season type: None

Only allow additive trend/season

Reject non-optimized models

Seasonal specification

Cycle: 12

Parameters (leave blank to estimate)

Alpha:

Beta:

Phi:

Gamma:

Sample specification

Estimation sample: 2003M01 2018M01

Forecast end point: 2018M01

Model Selection

Akaike Info Criterion

Schwarz Info Criterion

Hannan-Quinn Criterion

Average MSE

OK Cancelar

(a) Especificações do ETS Smoothing MAN

ETS Smoothing  
 Original series: IBCBR  
 Date: 03/25/16 Time: 18:00  
 Sample: 2003M01 2018M01  
 Included observations: 157  
 Model: M,A,N - Multiplicative Error, Additive Trend,  
 No Season  
 Convergence achieved on boundaries.

Parameters	
Alpha:	0.540088
Beta:	0.000000
Initial Parameters	
Initial level:	97.44160
Initial trend:	0.312577
Compact Log-likelihood	-628.5626
Log-likelihood	-454.4207
Akaike Information Criterion	1265.125
Schwarz Criterion	1277.350
Hannan-Quinn Criterion	1270.090
Sum of Squared Residuals	0.184462
Root Mean Squared Error	0.034277
Average Mean Squared Error	25.46566

(b) Resultado do ETS Smoothing MAN

Figura 5.18: ETS Smoothing MAN

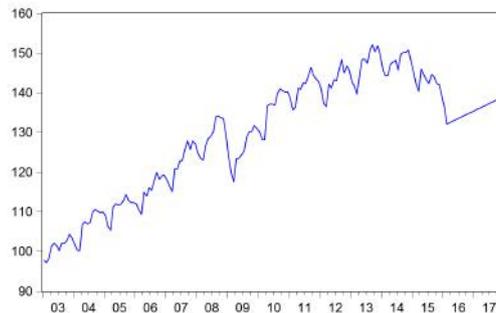


Figura 5.19: Gráfico da previsão conforme a especificação MAN

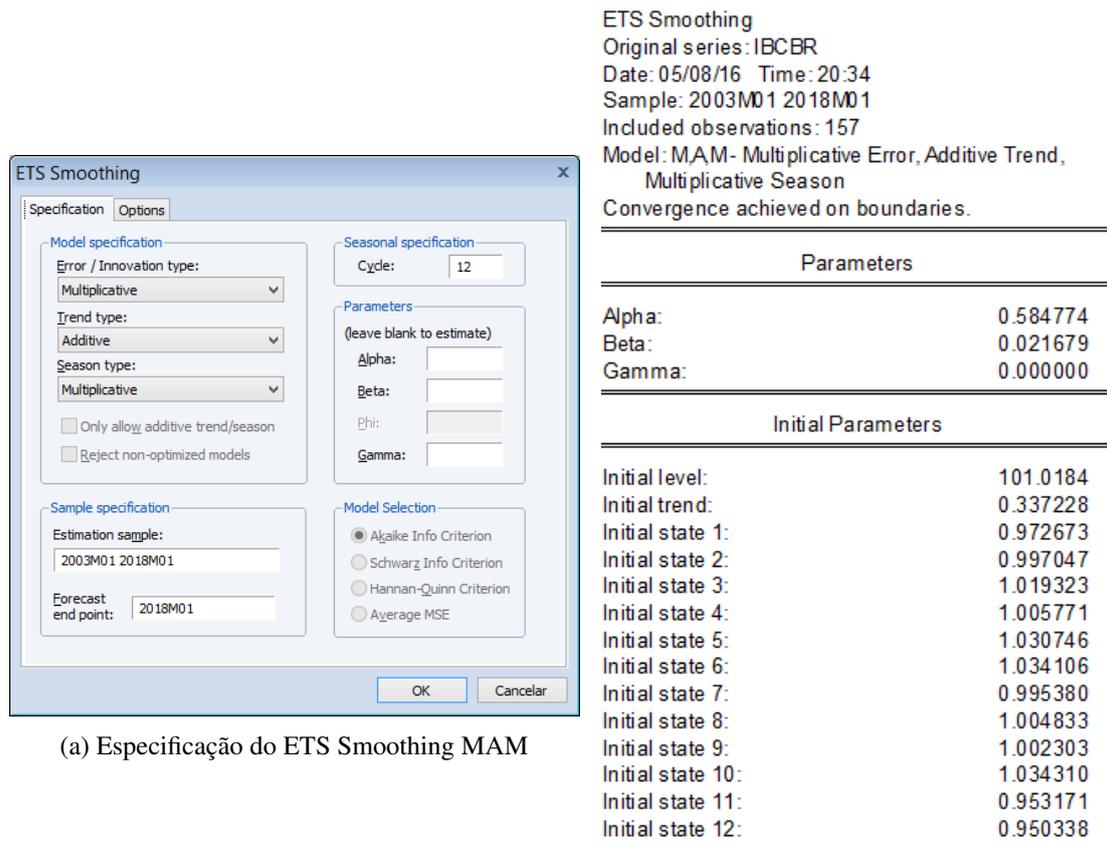
seguinte. Sendo assim, nossa previsão para o passo dois é dada por:

$$\begin{aligned}
 \hat{y}_2 &= (\hat{y}_1 + T_1) + \alpha \varepsilon_1 \\
 &= (97,75 + 0,3125) + 0,54(-1,604) \\
 &= 97,2003
 \end{aligned}$$

Com essa estimativa, encontramos o erro de previsão no passo dois,

$$\begin{aligned}
 \varepsilon_2 &= y_2 - \hat{y}_2 \\
 &= 98,67 - 97,20 \\
 &= 1,47,
 \end{aligned}$$

que será útil para corrigir a previsão no passo três. Fazemos isso até a última observação. A partir de então, a previsão passa a evoluir de acordo com a tendência, como mostra a Figura 5.19.



(a) Especificação do ETS Smoothing MAM

(b) Resultado do ETS Smoothing MAM

Figura 5.20: ETS Smoothing MAM

### Modelo MAM

Esse é o descrito pela presença de erro multiplicativo, tendência aditiva e sazonalidade multiplicativa. Com a série *ibcbr* aberta selecione ETS e depois as opções como mostrado na Figura 5.20a. Note que agora abre a opção de especificação cíclica. Como estamos com dados mensais, temos um valor  $cycle=12$ .

Os resultados agora possuem estimativa de 3 parâmetros ( $\alpha, \beta, \gamma$ ). O primeiro para atualização do erro de previsão, o  $\beta$  para a tendência e o  $\gamma$  para a sazonalidade. Logo abaixo estão os valores iniciais para o nível, a tendência e os 12 estados, cada qual representando um mês; ver Figura 5.20b.

O valor inicial estimado corresponderá à soma de valor do nível, da tendência e, como temos uma sazonalidade multiplicativa, essa soma é multiplicada pelo respectivo estado que corresponde ao mês anterior:

$$\begin{aligned}\hat{y}_1 &= (N_1 + T_1) \cdot S_{t-1} \\ &= (101,0184 + 0,3372) \cdot 0,9503 \\ &= 96,322.\end{aligned}$$

Como o primeiro mês é janeiro, usamos estado dezembro = 0,950338. Lembre-se que esses valores de estado correspondem aos fatores sazonais vistos anteriormente. Com base em  $\hat{y}_1$  podemos determinar o erro de previsão no primeiro passo usando:

$$\begin{aligned}\varepsilon_1 &= y_1 - \hat{y}_1 \\ &= 96,15 - 96,32 = -0,172\end{aligned}$$

A seguir, para prever o passo seguinte usamos o erro de previsão do passo anterior, o valor de  $\alpha$  e, como  $\beta = 0,021679$ , temos que considerar a taxa de crescimento da tendência. O gráfico de previsão pode ser visto na Figura 5.21.

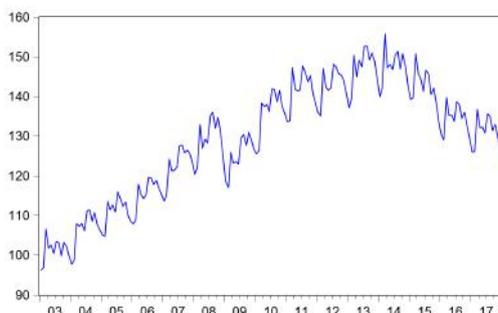


Figura 5.21: Gráfico da previsão conforme a especificação MAM

### 5.3 Ciclo

Outra característica observada nas séries de tempo é o componente cíclico, ( $C_t$ ). Entender como é o comportamento cíclico de uma série de tempo tem sido objeto de estudo com aplicações principalmente na macroeconomia. Há diversas técnicas paramétricas e não paramétricas que foram desenvolvidas para esse fim. A seguir são apresentados os diferentes filtros disponíveis no *EViews*<sup>®</sup> para a estimar a tendência de longo prazo e ciclo.

#### 5.3.1 Filtro Hodrick-Prescott

Esse é mais conhecido como filtro HP, em referência a seus autores, ver Hodrick e Prescott (1997) e é usado para estimar o componente de tendência de longo prazo de uma série de tempo. Sua estimativa considera a escolha de um parâmetro que irá determinar o grau de “aderência” dessa tendência à série de tempo. Quanto maior for, mais linear torna-se a tendência de longo prazo estimada.

Essa técnica de extração do componente cíclico é do grupo das que são aplicadas no domínio do tempo. Com a série *qx* aberta, selecione **Proc/Hodrick-Prescott Filter**. O filtro produz duas estimativas, uma para a série filtrada, ou seja, a estimativa de tendência de longo prazo e uma outra para o ciclo, que é a diferença entre a série original e filtrada. Escolha um nome para cada uma das opções. No nosso exemplo escolhemos *qxhp* e *qxciclo*. A seguir estão as opções de escolha para o parâmetro de alisamento. Como *default*, é feita a sugestão com base em Hodrick e Prescott (1997), que leva em conta a periodicidade dos dados. Como temos dados trimestrais, sugere-se usar 1600. Outra alternativa é determinar o valor de com base em Ravn e Uhlig (2002) escolhendo a potência. Ao escolher a opção de 1600, duas séries de dados serão salvas no workfile: *qxhp* e *qxciclo*. É simples o leitor confirmar como que se obtém a série de ciclo, basta fazer:

$$qxciclo = qx - qxhp$$

Ou seja, o ciclo representa a diferença da série original em relação a sua tendência de longo prazo e o resultado de *qxciclo* é muitas vezes visto como “gap”. Valores acima de zero significam que estamos acima da tendência de longo prazo. No caso de usar o PIB, esse seria um exemplo de produção acima do potencial, uma informação útil para avaliação de conjuntura e que o leitor interessado pode ver em relatórios de bancos, corretoras e também do Banco Central. Por outro lado, valores abaixo de zero são indicações de que estamos abaixo da tendência de longo prazo.

Na figura 5.22 estimamos três tendências de longo prazo para diferentes valores de  $\lambda$ : (i)  $\lambda = 0$  nome qxhp1; (ii)  $\lambda = 1600$  nome qxhp; (iii)  $\lambda = 100000$  nome qxhp2. Como pode ser visto, para um valor de  $\lambda = 0$ , a tendência de longo prazo é igual à série em questão (linha azul). No valor sugerido de  $\lambda = 1600$ , a tendência de longo prazo oscila um pouco (linha verde). Por fim, para um valor muito alto,  $\lambda = 100000$ , a tendência de longo prazo se aproxima de uma reta.

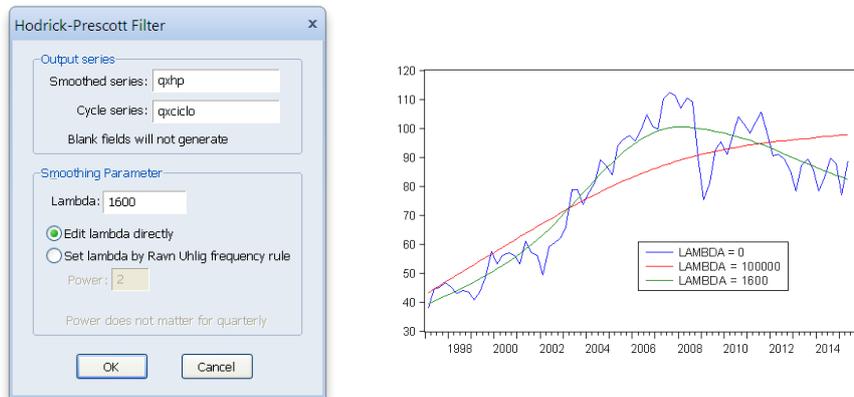


Figura 5.22: Filtro HP aplicado a qx

Um ponto interessante que o leitor poderá notar é que, ao se escolher como parâmetro de alisamento a alternativa de Ravn e Uhlig(2002) mas, deixando o valor 2 em “**power**”, os resultados serão idênticos ao aplicado o parâmetro  $\lambda = 1600$  sugerido por Hodrick e Prescott(1997).

**Programação 5.3.1** O método de Hodrick-Prescott também pode ser aplicado a partir de uma programação. Nesse caso, usamos:

```
qx.hpfilter(lambda=1600) qx_hp @qx_ciclo
```

Entre parênteses é colocado o lambda e o valor que se pretende para o parâmetro. Nesse exemplo, escolhemos 1600. A seguir estão os nomes das duas séries que serão geradas, a primeira é o componente de tendência de longo prazo e, a segunda, o componente cíclico. Note que, nessa função, é exigido que se tenha @ antes do nome da segunda série de dados. Seguindo a mesma linha de raciocínio, podemos juntar, em uma única função, a dessazonalização dos dados, o alisamento exponencial e a aplicação do filtro HP a partir de:

```
for %a qx y px pm qm
    seas(m) {%a} {%a}_sa {%a}_sf
    {%a}.x12(mode=m, filter=msr, save= "d10 d11 d12") {%a}_x12
    {%a}.smooth(m,e,e,e) {%a}_sm1
    {%a}.hpfilter(lambda=1600) {%a}_hp @{%a}_ciclo
next
```

**Programação 5.3.2** A OECD(2008), em suas análises de ciclo e construção de indicadores antecedentes e coincidentes sugere a utilização de dupla filtragem pelo método de Hodrick-Prescott para extração do ciclo de crescimento. Primeiramente é feita uma filtragem ajustando um  $\lambda$  de alto valor para extrair a tendência de longo prazo. Para então, com um  $\lambda$  menor extrair os movimentos de alta frequência e alisar o ciclo. Desta forma a OECD extrai o componente cíclico dentro de uma banda de frequência de 12 a 120 meses que corresponde respectivamente ao

$\lambda_2 = 13.93$  e  $\lambda_1 = 133107.94$ .

```

scalar lambda1= 133107.94
scalar lambda2= 13.93
for %a qx
  {%a}.hpf(lambda={lambda1}) {%a}_hptrend1 @{%a}_hpciclo1
  {%a}_hpciclo1.hpf(lambda={lambda2}) {%a}_hptrend2 @{%a}_hpciclo2
  genr {%a}_chp = ({%a}_hptrend2-@mean({%a}_hptrend2))/
  @stdev({%a}_hptrend2) + 100
next

```

Note que, em primeiro lugar, definimos o valor dos  $\lambda_1$  e  $\lambda_2$  dentro das variáveis de nome **lambda1** e **lambda2** pelo comando **scalar**. Isso facilita visualmente na hora de reescrever a programação para testar diferentes lambdas. Em seguida, declaramos um **loop** onde indicamos que **%a** tomará os valores de **qx**. Então, utilizando duas vezes o comando **.hpf** aplicamos o filtro HP, com os lambdas definidos anteriormente. Além de rodarmos a dupla filtragem, padronizamos o ciclo e adicionamos média 100, conforme sugerido OECD(2008). A programação pode incluir diversas séries e testes, abaixo trazemos um exemplo da dessazonalização pelo método X-12 multiplicativo seguido da extração do ciclo de crescimento sugerido em OECD(2008).

```

scalar lambda1= 133107.94
scalar lambda2= 13.93
for %a qx y px pm qm
  {%a}.x12(mode=m, filter=msr, save= "d10 d11 d12") {%a}_sa
  {%a}_sa.hpf(lambda={lambda1}) {%a}_hptrend1 @{%a}_hpciclo1
  {%a}_hpciclo1.hpf(lambda={lambda2}) {%a}_hptrend2 @{%a}_hpciclo2
  genr {%a}_chp = ({%a}_hptrend2-@mean({%a}_hptrend2))/
  @stdev({%a}_hptrend2) + 100
next

```

A dupla filtragem aproxima o filtro HP aos *Band-Pass filters* mostrados a seguir.

### 5.3.2 Filtros de Frequência

Também conhecidos como *Band-Pass filter*, é um filtro linear que extrai o componente cíclico de uma série de tempo a partir de um intervalo de duração do mesmo. Aqui, a análise é feita no domínio da frequência, e a série de tempo é representada a partir de uma soma ponderada de oscilações seno e cosseno. Sendo assim, a questão é como encontrar essa matriz de pesos que será aplicada à série de dados.

Há vários métodos de aplicação do filtro. O que irá diferenciá-los é a forma de cálculo das médias móveis. São três alternativas. As duas primeiras consideram um filtro simétrico e são diferentes apenas na forma como a função objetivo estima os pesos das médias móveis. Ao selecionar um desses dois métodos, e escolher os *Lead/lags* – refere-se ao comprimento da frequência do ciclo, é importante ter em mente que são perdidos os dados do início e fim da série para que seja feita a estimativa. Destaca-se que o comprimento da frequência do ciclo fica constante durante toda a série de dados, por isso que esse é um filtro de comprimento fixo.

O terceiro filtro, de nome Christiano-Fitzgerald, é assimétrico com as ponderações sendo diferentes no tempo e se comportando de acordo com os dados. O fato de ser um filtro que é variante no tempo, o torna mais completo para se determinar os ciclos de uma série. Nesse caso,

não é necessário especificar o comprimento da frequência do ciclo.

Com a série *qx* aberta, vá em **Proc/Frequency Filter...**, e aparecerá uma tela para selecionar as opções do filtro. Escolha o primeiro deles (Baxter-king). A seguir, do lado direito, a opção *Lead/lags* refere-se ao comprimento da frequência com que ocorre o ciclo. Vamos deixar o valor 12. Isso irá resultar na perda de informação do ciclo, 12 trimestres antes e 12 trimestres depois, reduzindo a estimativa para apenas 36 trimestres.

A parte do Cycle periods, se refere à duração do ciclo. Como *default* o *EViews*® retorna o valor **Low=6** e **High=32**. Ou seja, o ciclo de menor duração tem 6 trimestres e, o de maior duração, 32 trimestres. Depois, escolha os nomes para os resultados como mostrado na Figura 5.23. Do lado esquerdo está a escala para a série *qx* e *qxbp*, esse sendo o componente de longo prazo.

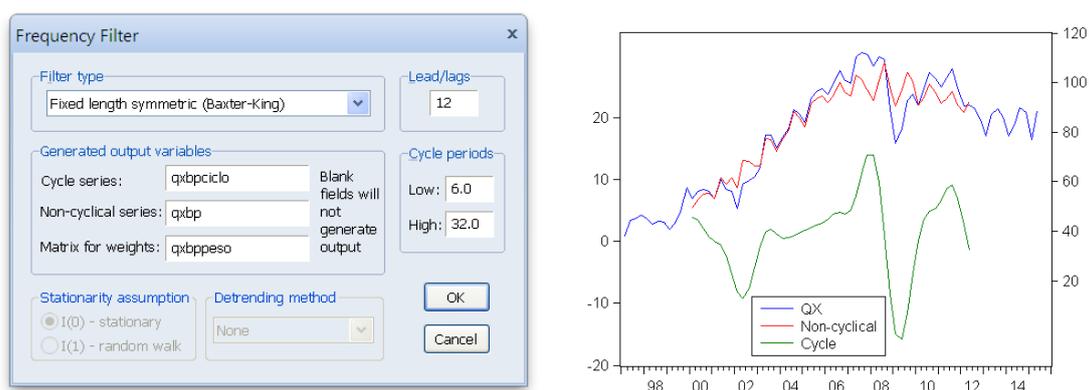


Figura 5.23: Filtro Baxter-King aplicado a *qx*

Como procedimento na estimativa vemos que, primeiro, é encontrada a matriz de pesos *bpeso*. Como escolhemos **Lead/lags** igual a 12, a matriz terá 13 colunas (será sempre uma a mais que o número de **Lead/lags**). Destaca-se que essa matriz é posteriormente utilizada para gerar a série *qxbpciclo* a partir de:

$$qxbpciclo_t = \sum_{c=1}^{q+1} w(1,c)y_{t+1-c} + \sum_{c=2}^{q+1} w(1,c)y_{t+c-1}$$

Com  $t=q+1, q+2, \dots, n-q$  onde  $w(1,c)$  é a matriz linha de pesos, aqui denominada de *bpeso*,  $c$  é cada uma das colunas dessa matriz,  $q$  é o número de **Lead/lags** (no nosso exemplo é 12), e  $n$  é o número de dados, no nosso exemplo, 74 observações. Assim, o intervalo de *qxbpciclo* será dado por  $t = 13, 14, \dots, 62$ . Portanto, o primeiro resultado, com  $t = 13$ , é encontrado usando:

$$qxbpciclo_{13} = \sum_{c=1}^{13} w(1,c)y_{14-c} + \sum_{c=2}^{13} w(1,c)y_{12+c}$$

Note que, tal como anteriormente:

$$qxbpciclo_t = qx_t - qxbp_t$$

Uma última informação fornecida diz respeito à resposta que a série filtrada *qxbp*, responde à série *qx*, em uma dada frequência, ver figura 5.24. A linha vermelha mostra a resposta ideal que deve estar no intervalo  $\left(\frac{1}{P_U}, \frac{1}{P_L}\right)$ , onde  $P_U$  é o maior período e  $P_L$  o menor. No nosso exemplo,  $P_U = 32$  e  $P_L = 6$ , e o intervalo ótimo é entre (0,031;0,167).

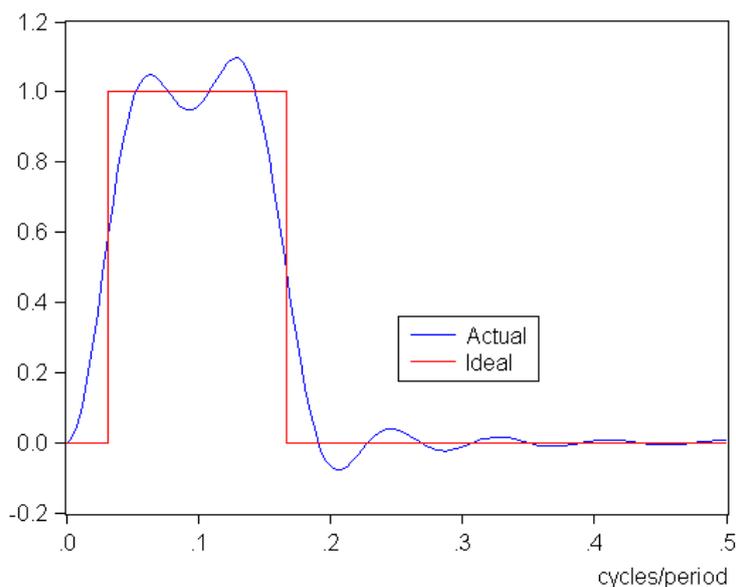


Figura 5.24: Função resposta de frequência – Baxter-King

**Programação 5.3.3** Podemos fazer a estimativa do filtro Baxter-King via programação. Nesse caso, a função utilizada é dada por:

```
qx.bpf(type=bk, low=6, high=32, lag=12, noncyc=qxbpfciclo, w=wqxbp) qxbpf
```

Dentre as várias opções que podem ser especificadas, o termo **type=bk** refere-se ao filtro Baxter-King. Se quiser escolher o filtro Christiano-Fitzgerald fixo, coloque **type=cffix** e, para o filtro assimétrico use **type=cfasym**. A seguir está o período mínimo do ciclo e o máximo. Depois, a série de dados ajustada pelo ciclo. Por fim, podemos selecionar os resultados a serem mostrados. Ainda no conjunto de opções, podemos escolher o nome da série ajustada pelo ciclo a partir de **noncyc=qxbpfciclo**. Podemos especificar a matriz de pesos do ciclo usando **weight=wqxbp**. A seguir, colocamos o nome da série do ciclo, qxbpf. Se o nome da série do ciclo (qxbpf) for omitido, o *EViews*<sup>®</sup> irá criar uma série de nome BPFILTER01. Assim, é possível agregar essa estimativa às anteriores, a partir de:

```
for %a qx y px pm qm
  seas(m) {%a} {%a}_sa {%a}_sf
  {%a}.x12(mode=m, filter=msr, save= "d10 d11 d12") {%a}_x12
  {%a}.smooth(m,e,e,e) {%a}_sm1
  {%a}.hpf(lambda=1600) {%a}_hp @{%a}_ciclo
  {%a}.bpf(type=bk,low=6,high=32,noncyc={%a}bpf_ciclo,w=w{%a}bp) {%a}_bpf
next
```

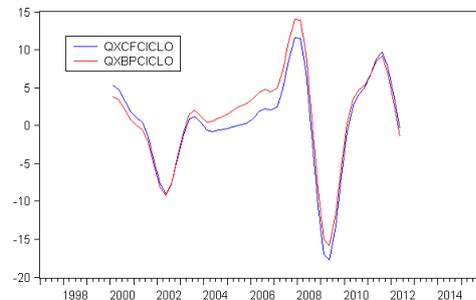
O segundo filtro simétrico que pode ser utilizado é o CF. Ao fazer essa escolha, será habilitada a opção de estacionariedade dos dados, além do método de diferenciação da série<sup>8</sup>. Ao selecionar a opção  $I(0)$ , há três alternativas para se proceder à diferenciação dos dados. Por outro lado, ao

<sup>8</sup>Para maiores esclarecimentos sobre o grau de integração de uma série de dados, o leitor deve consultar a seção sobre Raiz Unitária.

escolher que o processo é um *random walk*, há uma opção adicional. Assuma por hora que a série  $qx$  é um processo  $I(0)$  e que vamos usar o método **Remove linear trend**. Mantenha todas as demais opções como anteriormente, ou seja, **Lead/lags** igual a 12, a mesma periodicidade para o ciclo e dê nomes para as variáveis, como mostra a figura 5.25a. A forma de cálculo de  $qxcfciclo$  é a mesma de antes, usando a matriz de pesos.



(a) Opções filtro CF simétrico



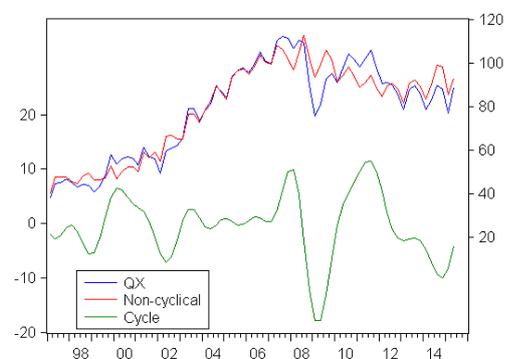
(b) Ciclo comparado pelos dois métodos

Figura 5.25: Filtro CF simétrico aplicado a  $qx$ 

No geral, os resultados são muito parecidos. A matriz de pesos em pouco difere (não foi aqui mostrada, mas o leitor pode comparar  $bp peso$  com  $cf peso$ ) e, como mostrado na figura 5.25b, o componente cíclico, resultante da série filtrada, também é similar.

O ganho maior no **Band-Pass filter** está em usar o método assimétrico. Nesse caso, não perdemos informações com o uso de *lead/lags*. Aliás, como pode ser visto, a opção do terceiro filtro não habilita a escolha do número de *lead/lags*. Escolha a opção **Full sample asymmetric**, mantenha a periodicidade cíclica de 6 e 32,  $I(0)$  e use o método **remove linear trend**. Escolha nomes diferentes para os resultados para não coincidir com as estimativas anteriores. Tal procedimento é mostrado na figura 5.26.

A determinação da periodicidade cíclica pode variar de acordo com a percepção sobre a duração do ciclo. O menor valor a ser especificado em **Low** é 2, o que irá produzir uma estimativa de ciclo bem errática. Obviamente, a duração máxima em **High** tem que ter um valor maior que o especificado em **Low**. Outra opção que precisa ser avaliada em **Stationarity Assumption** é se a série em questão que estamos extraíndo o ciclo é estacionária  $I(0)$  ou então possui raiz unitária  $I(1)$  e, por fim, tem-se que especificar o método para eliminar essa não estacionariedade.

Figura 5.26: Filtro CF assimétrico aplicado a  $qx$

Note que o ciclo agora é estimado para todo o conjunto de dados. Além disso, o leitor poderá ver que, ao analisar a matriz de pesos, a mesma é de dimensão  $74 \times 74$ , refletindo o fato de que os pesos variam no tempo. Para encontrar o resultado do primeiro trimestre, usa-se o primeiro vetor linha, multiplicado pelo vetor coluna de `qx`. Isso irá produzir como resultado, a primeira informação do ciclo, no nosso exemplo, a série `qxcfciclo2`. Na linha 1, as primeiras 13 informações de pesos são idênticas às encontradas pelo método CF simétrico. Na figura 5.27, comparamos os resultados da estimativa pelo filtro HP com a obtida pelo método CF assimétrico. Note que esse tem uma estimativa de ciclo mais suavizada.

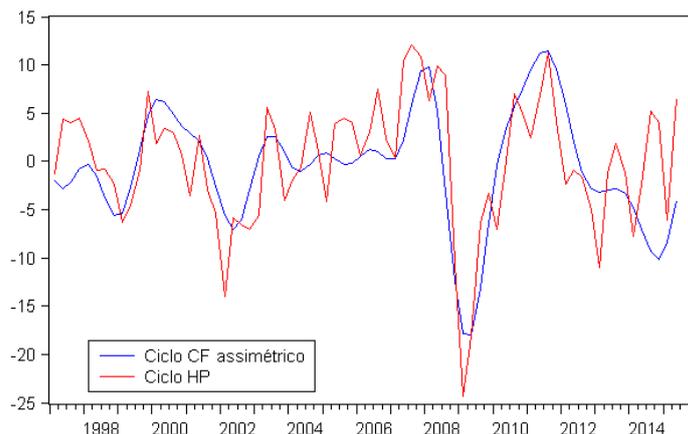


Figura 5.27: Ciclos de `qx` comparados

**Programação 5.3.4** Podemos fazer a estimativa do filtro Christiano-Fitzgerald assimétrico usando diversas combinações entre ciclo mínimo e máximo. Nesse caso especificamos primeiro um escalar de valor 4, a duração mínima do ciclo. A seguir usamos `type=cfasym` e, em `low` denominamos esse escalar fixando o máximo em 60. Depois, especificamos que a série seja diferenciada para eliminar a tendência especificando uma ordem de integração `iorder=1`.

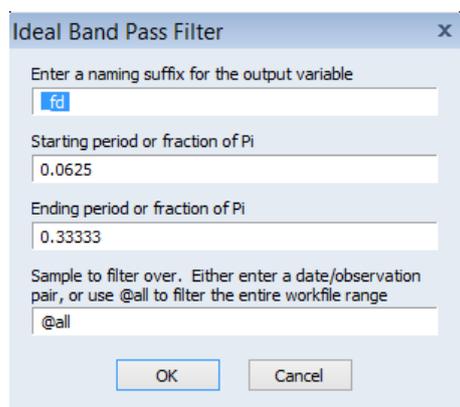
A seguir especificamos a série ajustada pelo ciclo a partir de `noncyc=qxbpfciclo`. Podemos especificar a matriz de pesos do ciclo usando `weight=wqxbp`. A seguir, colocamos o nome da série filtrada, `qxbpf`. Se o nome da série do ciclo (`qxbpf`) for omitido, o *EViews*® irá criar uma série de nome `BPFILTER01`.

```
scalar num=4
for %a qx y px pm qm
  {%a}.bpf(type=cfasym,low=num,high=60,detrend=t,iorder=1,nogain,
noncycle={%a}cf) {%a}bpf
  num=num+1
next
```

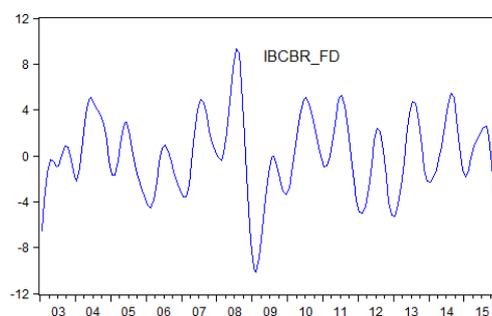
### 5.3.3 O Filtro Corbae-Ouliaris

As técnicas de extração do componente cíclico são divididas, de forma geral, em dois grupos, ou no domínio do tempo ou da frequência. Esse é particularmente importante na análise de séries econômicas devido a não-estacionariedade, ver Corbae e Ouliaris (2006). Após feita a instalação do add-in `fdfilter` no *EViews* podemos usar esse filtro (veja o capítulo que fala sobre add-in). Abra a série mensal do `ibcbr` em `proc/add-ins` selecione **corbae\_ouliaris FDFilter**. A Figura 5.28a mostra

a caixa de diálogo onde devemos preencher com os valores.



(a) Caixa de diálogo do add-in



(b) Estimativa do ciclo com o filtro Corbae-Ouliaris

Figura 5.28: Filtro Corbae-Ouliaris

**Programação 5.3.5** Uma vez que o add-in `FDfilter` esteja instalado no *EViews*<sup>®</sup> é possível aplicar o filtro a partir do menu ou então a partir de um comando da sub-rotina. Para o exemplo do IBCBR teremos:

```
call ideal bandpass (ibcbr, 0.0625, 0.033, "ibcbr_cicle", "data inicial, data final")
```

Caso tenha várias séries de dados podemos usar um loop para aplicar o filtro a todas elas:

```
For %a a1 a2 a3
%name = "ciclo" + %a
call ideal bandpass (%a, 0.062, 0.333, %name, "2003M01, 2016M1")
next
```

Como primeira opção colocamos a extensão do nome da série do ciclo que será calculada. Nesse exemplo usamos `_FD`. As duas opções seguintes se referem aos valores dos períodos iniciais e finais ou, fração de  $\Pi$ , que será usado na determinação da frequência do ciclo. Como default usamos 0.0625 e 0.3333. Por fim especifique o período de análise dos dados. Como queremos uma estimativa para todo o período escrevemos `@all`. A Figura 5.28b mostra a estimativa do ciclo do IBCBr.

## 5.4 Autocorrelação (Correlograma)

O conceito de autocorrelação será bem útil quando analisarmos os modelos ARIMA, mas já podemos começar a compreender algumas características e implicações da autocorrelação. Como o próprio nome diz, a autocorrelação descreve a relação de correlação que uma variável aleatória, o PIB por exemplo, tem com ela mesma no passado. Em séries de tempo de economia é muito comum vermos a presença de autocorrelação, bem como em séries financeiras. Imagine a taxa de câmbio hoje. Seu resultado será altamente correlacionado com o valor da taxa de câmbio ontem. Quanto maior for essa relação, maior será a medida de autocorrelação. Há formas de ver a presença ou não de autocorrelação em uma série de dados, sendo a mais comum fazer o correlograma.

Com uma série de tempo aberta, `qx` por exemplo, selecione **View/Correlogram...**. A janela de opções, conforme figura 5.29a, possibilita analisarmos a série em nível e primeira ou segunda diferença como adicionar o número de defasagens. Um correlograma em nível avalia a série original. Quando selecionamos *1st difference*, aplicamos o conceito de autocorrelação para a série de dados

$\Delta qx$ , ou seja, na primeira diferença da variável em questão. A opção dos lags a incluir é apenas para o teste e a visualização gráfica. O programa nos retorna dois gráficos de barras (correlogramas) e quatro estatísticas vinculadas: autocorrelação (AC), autocorrelação parcial (PAC), estatística Q e a probabilidade, conforme figura 5.29b.

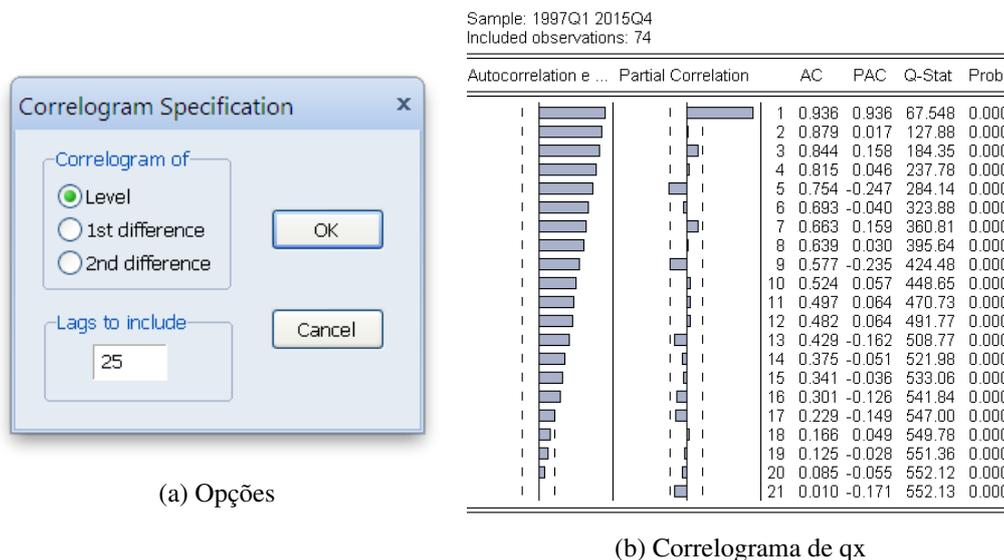


Figura 5.29: Correlograma

A função de autocorrelação (AC) mensura a correlação de uma variável e suas defasagens. Seu cálculo é obtido através da divisão da covariância com a defasagem  $k$  pela variância da amostra. Ligeiramente diferente da definição teórica o *EViews*<sup>®</sup> estima autocorrelação pela seguinte fórmula:

$$t_k = \frac{\sum_{t=k+1}^n (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2},$$

tal que, quando  $k = 1$ , estamos medindo a autocorrelação de ordem um e assim sucessivamente.

Já a autocorrelação parcial (PAC) calcula a autocorrelação da defasagem  $k$  descontando o poder preditivo das defasagens anteriores,  $t_1, t_2, \dots, t_{k-1}$ . Os resultados de AC e PAC são apresentados de forma gráfica nos dois correlogramas, onde a linha vertical continua indica o zero e as duas linhas pontilhadas aproximadamente dois desvios padrão, computados por  $\pm 2/(\sqrt{n})$ . Ou seja, para que o coeficiente,  $t_k$ , seja significativamente diferente de zero, ao nível de significância de aproximadamente 5%, este não pode estar entre as linhas pontilhadas.

No exemplo apresentado na figura 5.29b,  $qx$  tem 74 observações então  $\pm 2/(\sqrt{74}) = \pm 0.2325$ . Logo, para que o coeficiente seja significativo e estatisticamente diferente de zero, não pode pertencer ao intervalo de confiança de 95%:

$$Prob(\hat{t}_k - 0.2325 \leq t_k \leq \hat{t}_k + 0.2325) = 0.95$$

Além de calcularmos a significância estatística para determinada defasagem individualmente, podemos utilizar estatística Q de Ljung-Box (Q-Stat) para uma hipótese conjunta. Esse teste estatístico avalia a autocorrelação na defasagem  $k$  sob a hipótese nula de que todos coeficientes,  $t_1, t_2, \dots, t_k$ , são simultaneamente iguais a zero. A fórmula da estatística Q é dada por:

$$Q_{LB} = T(T+2) \sum_{j=1}^k \frac{\hat{t}_j^2}{T-j}.$$

Assim, supondo a avaliação da autocorrelação até  $k=1$ , teremos:

$$Q_{LB} = 74(76) \sum_{j=1}^1 \frac{0,936^2}{74-1} = 67,54.$$

Além disso, a estatística Q e seu p-valor, apresentados nas últimas duas colunas do correlograma, são comumente utilizados para testar se a série é ruído branco. Cabe destacar nesse caso que, considerando uma série qualquer  $y_t = \varepsilon_t$ , tal que o choque  $\varepsilon_t$  não é serialmente correlacionado, esse processo, com média zero e variância constante, será denominado ruído branco. Adicionalmente, se  $\varepsilon_t$  e, conseqüentemente,  $y_t$ , forem serialmente independentes, podemos dizer que  $y$  é ruído branco independente escrevendo  $y_t \sim iid(0, \sigma^2)$ , ou seja,  $y$  é independentemente e identicamente distribuído com média zero e variância constante.

O correlograma também nos permite algumas considerações sobre modelagem das séries de tempo. Se a autocorrelação apresentar coeficientes significativos que diminuam lentamente de forma geométrica e a autocorrelação parcial for para zero depois da defasagem  $p$ , podemos evidenciar que a série obedece um processo autorregressivo puro de ordem  $p$ ,  $AR(p)$ . Como o correlograma na figura 5.29b, que nos sugeriu que a série  $qx$  segue um processo autorregressivo de primeira ordem  $AR(1)$ .

Enquanto processos puros de médias móveis (MA) apresentam autocorrelação próxima a zero depois de algumas defasagens, junto de autocorrelação parcial persistente caindo gradualmente para zero, conforme figura 5.30a. Da mesma forma, um correlograma que apresente um padrão sazonal de movimentos recorrentes como ondas sugeriu a presença de sazonalidade, figura 5.30b. Vale ressaltar, o padrão de séries não-estacionárias mostram coeficientes de autocorrelação altos e persistentes em diversas defasagens, como  $qx$  na figura 5.29b.

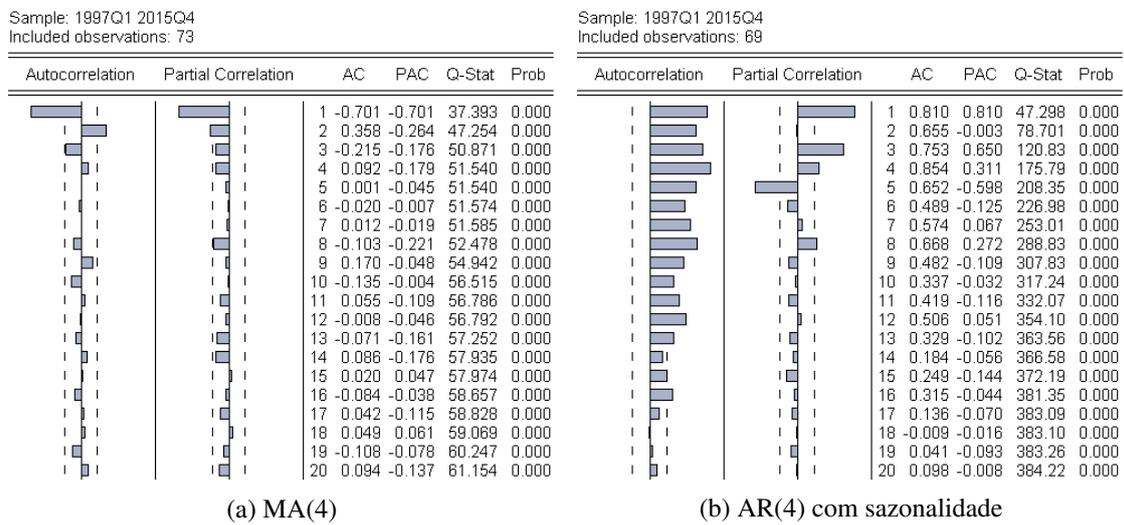


Figura 5.30: Correlograma

**Programação 5.4.1** A programação para acessarmos o correlograma é dada pelo comando `.correl(k)`, onde  $k$  é a quantidade de defasagens a serem testadas. Abaixo executamos o correlograma na variável `qx` com 21 defasagens.

```
qx.correl(21)
```

Também podemos criar para diversas séries de tempo um loop que salve as informações estatísti-

cas do correlograma em uma tabela de resumo.

```

scalar k = 25
table corres
!j=0
for %a qx y px
  for !i = 1 to k
    freeze(mode = overwrite, temp) {%a}.correl(k)
    corres(1,1+!j) = %a
    corres(2,1+!j) = "k"
    corres(2,2+!j) =temp(5,4)
    corres(2,3+!j) =temp(5,5)
    corres(2,4+!j) =temp(5,6)
    corres(2,5+!j) =temp(5,7)
    corres(2+!i,1+!j) =temp(6+!i,3)
    corres(2+!i,2+!j) =temp(6+!i,4)
    corres(2+!i,3+!j) =temp(6+!i,5)
    corres(2+!i,4+!j) =temp(6+!i,6)
    corres(2+!i,5+!j) =temp(6+!i,7)
  next
  !j = !j+5
next

```

Note que inicialmente criamos o escalar **k**, que recebe o número de defasagens, a tabela resumo **corres**, que receberá as estatísticas calculadas, e a variável de contagem **!j**, que organizará as colunas em **corres** quando houver mais de uma série de tempo. Então, é aplicado o comando **.correl** em **qx,y** e **epx** e guardamos as informação dentro da tabela temporária, **temp**, usando o comando **freeze**. Para preenchermos **corres** com os dados contidos em **temp**.

## 5.5 Análise Espectral

A análise espectral tem muita aplicação na física, química e demais ciências. Na economia, sua importância está, principalmente, na explicação das informações de frequência que podemos extrair e que acaba por revelar características cíclicas. Toda série de tempo pode ser expressa a partir da soma de senos e cossenos que oscilam de acordo com uma determinada frequência. O desafio é poder identificar essas frequências, e isso pode ser feito via estimativa do periodograma. Esse é conhecido como densidade espectral e relaciona as variabilidades do conjunto de dados com as frequências, ao passo que, na análise de série de tempo, as variabilidades são relacionadas com o domínio do tempo. Um dos pontos importantes é utilizar séries de dados que sejam estacionárias.

Sendo assim, podemos afirmar que a densidade espectral é uma representação das características da série de tempo, mas no domínio da frequência. O canal para se fazer essa relação, entre uma série de tempo expressa no domínio do tempo com uma que é expressa no domínio da frequência é a transformada de Fourier. Na literatura da área são disponíveis diversos métodos paramétricos e não-paramétricos para estimar a densidade espectral de um conjunto de dados.

Diversos pontos emergem a partir dessa relação e estão relacionados, principalmente, a variância dos dados. Primeiro, podemos citar que a integral da densidade espectral é igual a variância da série de dados. Na verdade, o espectro de uma série de tempo pode ser visto como a distribuição de variância dessa série como uma função da frequência. Em segundo lugar, que há uma relação entre o espectro, que contém informações do conjunto de dados no domínio da frequência, com a função de autocovariância, que contém informações no domínio do tempo.

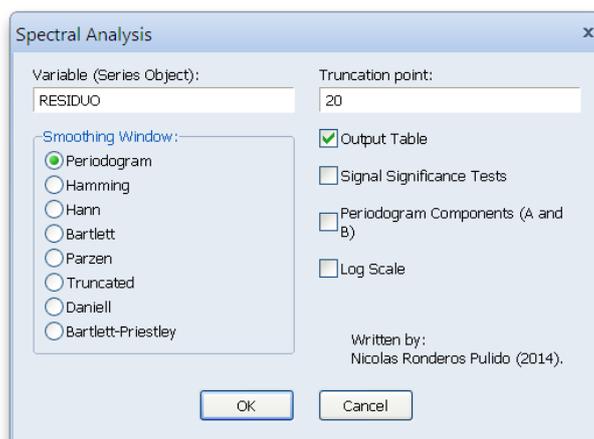


Figura 5.31: Opções da Análise Espectral

Uma vez identificada a densidade espectral podemos avaliar os picos de frequência e os períodos associados a ele. Suponha uma série de dados mensal e que na análise espectral tenha-se encontrado, por exemplo, um pico na frequência "a". Nesse caso, o período associado a esse ciclo, medido em meses, é dado por  $1/a$ . É normal termos mais de um pico na densidade espectral e veja que quanto maior for o valor de a, menor será o período, em tempo, associado a essa frequência.

O Eviews faz análise espectral, mas é necessário instalar o add-in **Spectral Analysis**. Vejamos como é a aplicação desse método a série mensal IBCBR do Banco Central do Brasil sem ajuste sazonal. Lembre-se que essa análise deve ser feita com a série estacionária. Como a nossa série possui tendência, primeiro temos que eliminar essa tendência, o que é feito a partir de uma regressão simples tendo como variável independente o tempo e uma constante. A seguir, analisamos os resíduos dessa equação.

Com a série residuo aberta vá em **Proc/Add-ins/Spectral Analysis**. A janela que será aberta é como mostrado na figura 5.31. Note que há várias opções de escolha para o processo de alisamento do periodograma. Vamos usar como *default* o ponto 20 como de truncagem e, por enquanto, não vamos selecionar as demais opções, apenas deixe *output table*.

Após clicar em **ok** é perguntado se queremos gerar o ciclo ótimo. Clique, novamente, em **ok**. A seguir é aberta uma janela que pergunta o p-valor e o número de ciclos. Digite 0.05. Deixe selecionada a opção **weighted cycle** e selecione **individual-cycles**. Clique em **ok**. Diversos resultados são reportados, mas vamos olhar primeiro para o gráfico do periodograma, como mostrado na figura 5.32. Note que o mesmo não foi alisado e apresenta diversos picos. Cada um desses picos, na respectiva frequência, possui um ciclo no tempo.

Mas, tal como colocado no gráfico não seria possível identificar essas frequências. Felizmente esses resultados são salvos em uma tabela no *workfile* de nome "data". Abra e poderá ver que o mesmo possui quatro colunas, como mostrado na figura 5.33. A segunda coluna corresponde ao eixo horizontal do gráfico do periodograma e traz a relação ciclo/tempo. A última coluna, de nome *periodogram*, corresponde ao eixo vertical do gráfico, e permite identificar os picos da nossa densidade espectral. Veja por exemplo que a primeira frequência, de valor 0,006369 tem o maior pico encontrado, de valor 4,65. A terceira coluna nos mostra a relação tempo/ciclo, ou seja, o período de ocorrência do ciclo, dado por  $1/\text{frequência}$ . Sendo assim, para a frequência 0,006369 temos um período cíclico de 157 meses. Note que há um pico no periodograma de valor 2,89 associado com a frequência 0,025478 e que gera um período de 39 meses. Há outro pico na frequência 0,082803 e que gera um período cíclico de 12 meses, revelando a existência de sazonalidade no nosso banco de dados.

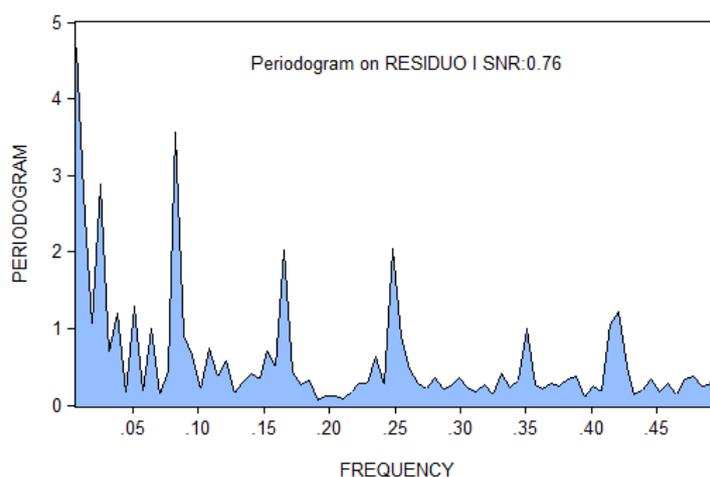


Figura 5.32: Periodograma da série residuo

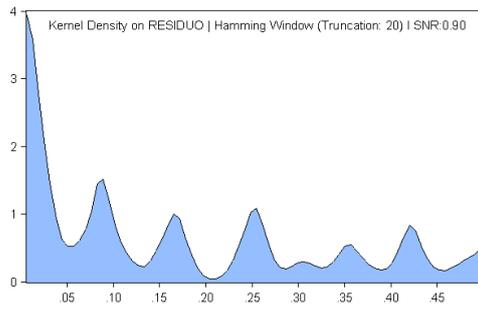
	Angular Frequency (Omega)	Frequency (Cycle/Time Unit)	Cycle Frequency (Time Unit/Cycle)	Periodogram
1				
2	0.040020	0.006369	157.0000	4.657059
3	0.080041	0.012739	78.50000	2.749385
4	0.120061	0.019108	52.33333	1.048113
5	0.160081	0.025478	39.25000	2.890035
6	0.200101	0.031847	31.40000	0.681870
7	0.240122	0.038217	26.16667	1.193706
8	0.280142	0.044586	22.42857	0.154285
9	0.320162	0.050955	19.62500	1.282258
10	0.360183	0.057325	17.44444	0.178488
11	0.400203	0.063694	15.70000	0.989477
12	0.440223	0.070064	14.27273	0.130936
13	0.480243	0.076433	13.08333	0.446892
14	0.520264	0.082803	12.07692	3.553907
15	0.560284	0.089172	11.21429	0.914875

Figura 5.33: Data

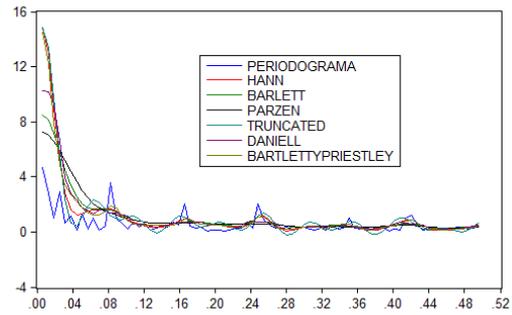
Alternativamente ao uso do periodograma para encontrar a densidade espectral, podemos usar os demais filtros. O gráfico 5.34a mostra a estimativa da densidade espectral usando o método de Hamming. Note que o resultado é mais suave que o apontado pelo periodograma e sinaliza para a presença dos mesmos picos identificados anteriormente. No gráfico 5.34b estão todas as estimativas de densidade. Para fazer esse gráfico primeiro faça a estimativa considerando cada um dos métodos disponíveis. A seguir, monte um grupo com todas as séries denominadas de "spectral density" e que estão na última coluna da tabela que é salva. Por fim, selecione **View/Graph./XY line** e, do lado direito, em **details**, onde está **multiple graphs**, escolha **single graph - First vs. All**.

Vejamos agora como pode ser obtido o ciclo. Para esse exercício vamos primeiro extrair o ciclo pelo filtro HP. Isso irá produzir uma série estacionária. Abra a série do ciclo resultante da aplicação do filtro HP e vamos usar o add-in de **Spectral Analysis**, selecionando o filtro de Bartlett e selecione as opções como mostrado na figura 5.35a. Na opção do filtro spectral vamos selecionar um teste a 0,05% e ciclos individuais tal como mostrado no gráfico 5.35b. Note que há a opção de **Cycle Sum**. Essa é a soma dos ciclos individuais pedidos acima. O número de ciclos individuais que são gerados são quatro: `sfw_13`, `sfw_26`, `sfw_39` e `sfw_4`. Se somarmos os quatro teremos como resultante o ciclo estimado para a nossa série.

No conjunto de gráficos 5.36a estão os ciclos individuais estimados, e no gráfico 5.36b está a soma dos quatro ciclos individuais.

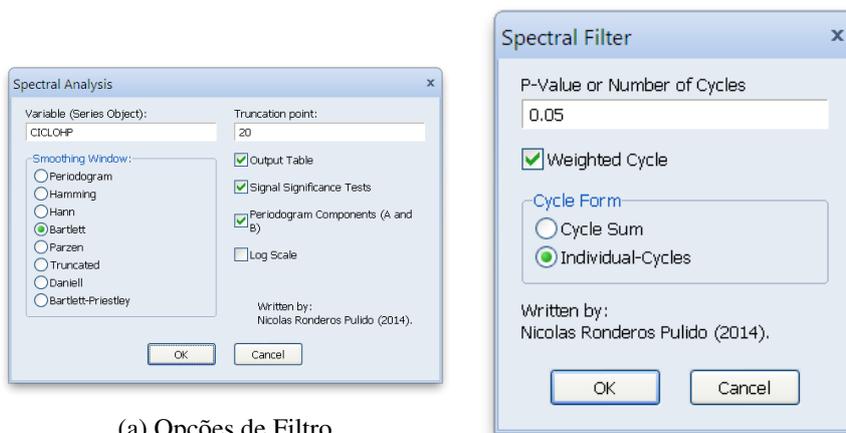


(a) Hamming



(b) Comparativo de Densidade Espectral

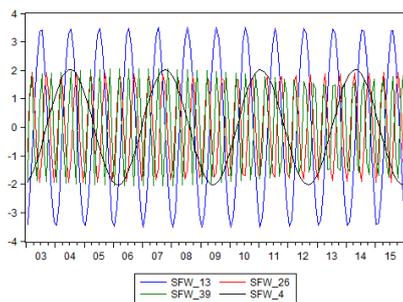
Figura 5.34: Estimativa Espectral



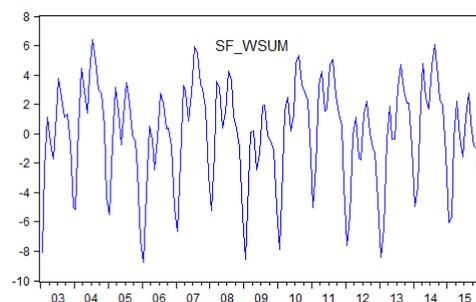
(a) Opções de Filtro

(b) Opções do Ciclo

Figura 5.35: Análise Espectral



(a) Ciclos Individuais



(b) Ciclos da série IBCR

Figura 5.36: Ciclos estimados

## 5.6 Exercícios

**Exercício 5.1** Aplique os quatro diferentes métodos de dessazonalização na série  $qx$ , agrupe em um gráfico e discuta as diferenças.

**Exercício 5.2** Aplique os cinco diferentes métodos de alisamento exponencial na série  $y$  e discuta as diferenças de resultado entre eles. ■

**Exercício 5.3** Utilize os diferentes métodos de alisamento exponencial para prever 6 trimestres de  $px$ . ■

**Exercício 5.4** Monte um gráfico de  $y$  que compare os três diferentes métodos da opção **detrending method** para o filtro CF simétrico em  $I(0)$ . ■

**Exercício 5.5** Monte um gráfico de  $y$  que compare os três diferentes métodos da opção **detrending method** para o filtro CF assimétrico  $I(0)$ . ■

**Exercício 5.6** Monte um gráfico de  $y$  que compare os quatro diferentes métodos da opção **detrending method** para o filtro CF assimétrico  $I(1)$ . ■

**Exercício 5.7** Extraia o ciclo de  $y$  pelo método de dupla filtragem HP, utilizado pela OECD(2008), e compare aos resultados encontrados com uma única filtragem. ■

**Exercício 5.8** Compare as melhores estimativas para  $y$  encontradas nos exercícios 5.5, 5.6 e 5.7. ■

**Exercício 5.9** Quais são as características de uma série de ruído branco? E por que a estatística  $Q$  é útil para identificá-la? ■

**Exercício 5.10** Crie uma série de ruído branco e prove as afirmações feitas no exercício 5.9 utilizando o correlograma e a estatística  $Q$ . ■

Quais são as características de uma série de não-estacionária? E como podemos utilizar a autocorrelação para inicialmente identificá-la?

**Exercício 5.11** Por que consideramos o cálculo de autocorrelação feito pelo *EViews*<sup>®</sup> diferente da definição teórica? ■

**Exercício 5.12** Calcule o correlograma de  $y$  para 30 defasagens e indique quais autocorrelações são estatisticamente diferentes de zero ao nível de significância de 5%. ■

**Exercício 5.13** Calcule o correlograma de  $y$  para 30 defasagens e indique quais autocorrelações são estatisticamente diferentes de zero ao nível de significância de 10%. ■

## 5.7 Bibliografia

- Christiano, L. J. e Fitzgerald, T. J. (2003), The Band Pass Filter. *International Economic Review*, 44: 435–465.
- Corbae, Dean e Ouliaris, Sam (2006). Extracting Cycles from Nonstationary Data. In: Dean

- Corbae et al. (eds.) *Econometric Theory and Practice*. Cambridge: Cambridge University Press, pp. 167-177.
- Gyomai, G., e Guidetti, E. (2008). OECD system of composite leading indicators. Organisation for Economic Co-Operation and Development (OECD). Disponível em: <http://www.oecd.org/std/leading-indicators/41629509.pdf>.
  - Hodrick, R. J., e Prescott, E. C. (1997). Postwar US business cycles: an empirical investigation. *Journal of Money, credit, and Banking*, 1-16.
  - Hylleberg, Svend (1986). *Seasonality in Regression*.
  - Morais, I.A.C., Bertoldi, A., Anjos, A.T.M. (2010), Um modelo não-linear para as exportações de borracha. *Revista Sober*.
  - Nilsson, R., e Gyomai, G. (2011). Cycle extraction: A comparison of the Phase-Average Trend method, the Hodrick-Prescott and Christiano-Fitzgerald filters.





## 6. Regressão Simples

O primeiro contato com modelos de econometria começa agora. Entretanto esse livro não tem a intenção de esgotar o assunto do ponto de vista metodológico, e sim com aplicações. Nesse sentido, se o leitor precisar de fundamentos e discussões técnicas sobre o tema, diversos livros técnicos podem ser consultados. O procedimento aqui é simples. Começamos com a estimação de um modelo com apenas uma variável independente e explicamos todas as opções de testes e identificação de problemas que por ventura possam aparecer e que estão disponíveis no *EViews*<sup>®</sup>. Entendido esse ponto, o capítulo seguinte passa a explicar um modelo de regressão múltipla.

O primeiro passo na estimativa de um modelo de regressão é definir as variáveis dependentes e independentes. No nosso exemplo a ideia é trabalhar com uma curva de demanda aplicada a exportação de móveis ( $qx$ ) e que pode ser explicada pela variável renda ( $yw$ ), que representa o número índice do PIB mundial. Vejamos como estimar uma regressão simples. Abra o arquivo do *EViews*<sup>®</sup> *regressão simples.wfl*. Há um conjunto de variáveis, mas usaremos apenas duas nesse momento. Nesse caso, vamos rodar a seguinte equação de regressão <sup>1</sup>:

$$qx_t = \alpha_1 + \beta_1 yw_t + \varepsilon_t$$

Há três caminhos possíveis no *EViews*<sup>®</sup> para se estimar uma equação. O mais simples deles é selecionar cada uma das variáveis a constar nessa equação, sempre selecionando em primeiro lugar a variável dependente e, a seguir, clicar com o botão direito e clicar em **Open/as Equation....** A segunda maneira é ir em **Quick/Estimate Equation...** e escrever o formato da equação. Esses dois métodos são mostrados na figura 6.1. Note a diferença sutil que existe, podemos escrever nossa equação de duas maneiras. Na primeira opção, aparece apenas o nome das variáveis, sempre seguindo a ordem da dependente como a inicial. No segundo método, é necessário escrever a equação, onde o termo  $c(1)$  e  $c(2)$  representam os coeficientes a serem estimados.

<sup>1</sup>Note que há dados com e sem ajuste sazonal, onde esses são representados por *\_sa*. Aqui foi usado o método X-12. Desse ponto em diante usaremos apenas os dados com ajuste sazonal.

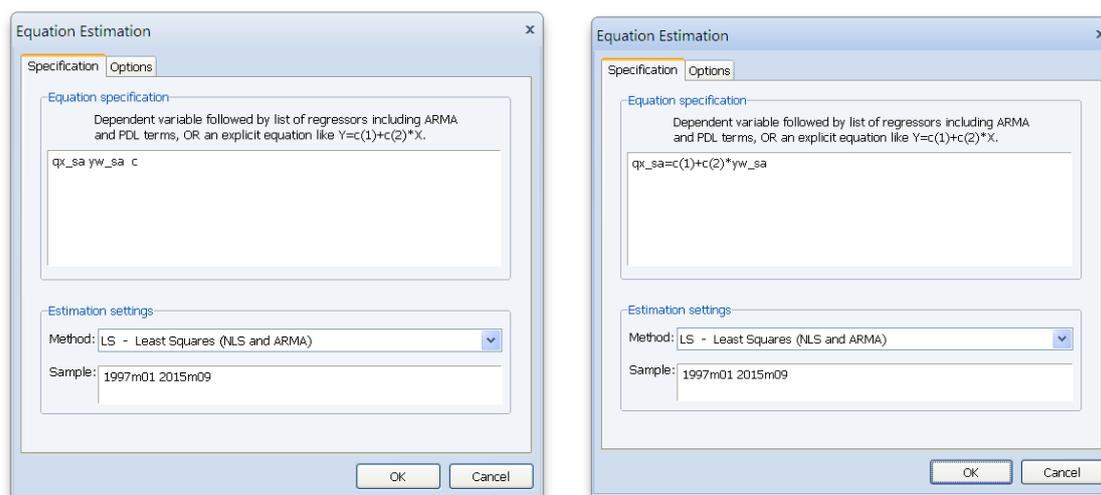


Figura 6.1: Como Estimar Uma Equação no *EViews*®

Logo abaixo do campo onde se especifica as equações, na figura 6.1, está o método de estimação, que no nosso caso é o **LS – Least Squares**, também conhecido como mínimos quadrados, e o **Sample** que é o período amostral onde serão feitas as estimativas. Clique em **OK**. Qualquer que seja a forma utilizada para rodar essa regressão, o resultado será o mesmo, como mostrado na figura 6.2. Diversas estatísticas podem ser visualizadas. Na primeira linha está descrita a variável dependente, seguido do método de estimação, a data em que foi feita essa estimativa (útil para ver se os alunos fizeram o exercício na data certa), o período utilizado para gerar os resultados e o total de dados. Note que são usados 187 dados que vão de janeiro de 2000 a julho de 2015. Logo abaixo, em uma tabela, são mostrados os resultados da nossa equação e que, normalmente, são assim representados em livros e artigos de econometria:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)            (0.063)

onde, entre parênteses, ficam descritos os valores dos respectivos desvio-padrão. A terceira forma de estimar uma equação no *eviews* é via programação e pode ser visualizado no box 6.0.1.

**Programação 6.0.1** No caso da programação, há duas formas de se rodar uma regressão. Na primeira, escrevemos o método que, no presente caso, é dado pelo comando “*ls*”, que significa “*Least Square*” (Mínimos Quadrados) seguindo pela ordem das variáveis onde primeiro é colocada a dependente. Há uma lista de opções que podem ser colocadas depois do termo *ls*, consulte o manual. Antes de qualquer coisa, o melhor a fazer é especificar o intervalo de dados que estamos trabalhando que, no presente exemplo, é de 2000M1 a 2015M7.

```
smp1 2000M1 2015M7
ls qx_sa yw_sa c
```

A segunda maneira seria escrever o comando “*equation*” seguido do nome a ser dado para a equação e da lista das variáveis. Há algumas vantagens nesse segundo método que vão ficar mais claras mais a frente. Uma delas é o fato de já especificarmos o nome da nossa regressão como “*eq1*”:

Dependent Variable: QX\_SA  
Method: Least Squares  
Date: 11/07/15 Time: 21:31  
Sample (adjusted): 2000M01 2015M07  
Included observations: 187 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
YW_SA	0.647967	0.063427	10.21588	0.0000
C	18.38936	6.616833	2.779178	0.0060
R-squared	0.360667	Mean dependent var		85.16810
Adjusted R-squared	0.357212	S.D. dependent var		17.50486
S.E. of regression	14.03436	Akaike info criterion		8.131532
Sum squared resid	36438.21	Schwarz criterion		8.166089
Log likelihood	-758.2983	Hannan-Quinn criter.		8.145535
F-statistic	104.3642	Durbin-Watson stat		0.338638
Prob(F-statistic)	0.000000			

Figura 6.2: Resultado da Regressão Simples

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
```

**Programação 6.0.2** O arquivo *regressão simples.wf1*, também, contém as séries originais sem ajuste sazonal. Podemos adicionar os comandos aprendidos no capítulo anterior, para dessazonalizar as séries pelo método X-12 multiplicativo e, então, rodar a regressão simples pelo método dos mínimos quadrados.

```
qx.x12(mode=m) qx
yw.x12(mode=m) yw
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
```

Os resultados para a nossa primeira estimativa de regressão simples podem ser visualizados na figura 6.2.

Após os valores dos coeficientes (parâmetros), estão os desvios-padrão (*StdError*) e, na coluna seguinte da tabela está a estatística t (*t-statistic*) e o p-valor (*Prob*). A primeira é utilizada para testar se o seu respectivo coeficiente é estatisticamente diferente de zero, a partir da fórmula:

$$t = \frac{x - \mu}{\sigma}$$

Por exemplo, podemos testar se  $\alpha_1 = 0$  que é a nossa constante. Nesse caso, a estatística t é dada por:

$$t = \frac{\alpha_1 - 0}{\sigma_\alpha} = \frac{18.389 - 0}{6.616} = 2.779$$

O mesmo podendo ser feito para testar se  $\beta_1 = 0$ , onde:

$$t = \frac{\beta_1 - 0}{\sigma_\beta} = \frac{0.647 - 0}{0.063} = 10.216$$

Por fim, o resultado do *Prob* irá indicar se aceitamos ou rejeitamos a hipótese nula de que o coeficiente em questão é estatisticamente igual a zero. O *Prob* aqui é o mesmo que o *P-valor*. Destaca-se que, para esse teste, estamos assumindo uma distribuição *t-student* e que é bicaudal. No nosso exemplo, tanto para o coeficiente da constante, quanto para o da renda, rejeitamos a hipótese nula de que são estatisticamente iguais a zero.

O valor *Prob* também pode ser encontrado a partir da função **tdist**. Nesse caso, como o resultado é um número, criamos primeiro um escalar e especificamos os valores para a função **tdist** a partir de `scalar pvalor=@tdist(10.216,187)`. O valor 10.216 é o valor da estatística *t* e 187 representa o número de graus de liberdade do teste, equivalente ao número de observações utilizadas após o ajuste (veja no início dos resultados na fig. 6.2).

**Programação 6.0.3** Dando sequência à nossa regressão simples, os comandos abaixo podem ser usados para testar se o parâmetro da elasticidade-renda é igual a zero  $\beta_1 = 0$ . Nesse caso, primeiro especificamos a estatística *t* e armazenamos a mesma em um escalar de nome `valort`, salvamos o número de observações no escalar `obs` e, a seguir, aplicamos o teste para encontrar seu respectivo *p*-valor e armazenar o resultado em um escalar de nome `pvalor`:

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
scalar valort=eq1.@tstats(1)
scalar obs=eq1.@regobs
scalar pvalor=@tdist(valort,obs)
```

Além desses resultados básicos, há diversos outros que são mostrados logo abaixo e que servem para avaliar o modelo em questão. Por exemplo, no caso do *R-squared*, conhecido como  $R^2$  ou  $R^2$ , o valor de 0.360 deve ser interpretado como: “cerca de 36% das variações em *qx* são explicadas por variações em *yw*”. Alguns costumam afirmar que esse resultado, na verdade, estaria se referindo ao grau de explicação do modelo, o que não deixa de ser verdade. A fórmula é dada por:

$$R^2 = 1 - \frac{\sum_{t=1}^T \hat{\varepsilon}_t^2}{\sum_{t=1}^T (Y_t - \bar{Y})^2}$$

O termo  $\sum_{t=1}^T \hat{\varepsilon}_t^2$  é denominado de “soma do quadrado dos resíduos”, e que está mostrado na tabela como “*Sum squared resid*”.  $\sum_{t=1}^T \hat{\varepsilon}_t^2 = 36438.21$ .

Esse resultado pode ser feito manualmente. Primeiro pegue todos os resíduos da regressão e eleve cada um deles ao quadrado e depois some todos.

Esse termo também poderia ser expresso da seguinte forma:

$$\sum_{t=1}^T (\hat{\varepsilon}_t - \bar{\varepsilon})^2$$

Onde  $\bar{\varepsilon}$  é a média dos resíduos. Porém, por definição, a média dos resíduos é igual a zero, uma vez que a reta de regressão foi estimada de forma a passar exatamente na média de todos os pontos. Sendo assim, tudo o que se erra na estimativa para cima, também se erra para baixo. Ou seja, teremos valores positivos e negativos que se anulam e, sua média daria zero. Sendo assim, acabamos por fazer:

$$\sum_{t=1}^T (\hat{\varepsilon}_t - \bar{\varepsilon})^2 = \sum_{t=1}^T (\hat{\varepsilon}_t - 0)^2 = \sum_{t=1}^T \hat{\varepsilon}_t^2$$

Caso queira verificar a série de resíduos, com a equação aberta, vá em **View/Actual,Fitted,Residual**. Ou então, se quiser gerar a série dos resíduos, vá em **Proc/Make Residual Series...**, e escolha um nome para essa série.

**Programação 6.0.4** Uma alternativa interessante é rodar várias regressões com uma janela fixa de, por exemplo, 60 dados, ou seja, 5 anos. Nesse caso, iniciamos uma regressão em 2000M1 que vai até 2004M12. A seguir, a segunda regressão vai de 2000M2 até 2005M1 e assim sucessivamente. Isso irá representar 99 regressões no total, com a última indo de 2008M4 a 2013M3. Para tanto, podemos declarar um loop usando o comando “for”:

```
for !i=1 to 99
smp1 2000M1+!i 2004M12+!i
equation eq2.ls qx_sa yw_sa c
next
```

Porém, isso irá gerar apenas um resultado para as nossas estimativas, qual seja, a última regressão. Nesse caso, não iríamos saber como evoluiu, por exemplo, ao longo dessas 99 regressões, o valor do coeficiente da elasticidade renda-demanda. O ideal seria comparar essa estimativa com a que envolve todos os dados, como feito anteriormente em eq1. Para tanto, podemos usar o comando `matrix`, para criar uma matriz de 100 linhas de nome *coef* e, depois, pedir para salvar esse coeficiente nessa matriz.

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
matrix(100) coef
coef(100)=eq1.@coefs(1)
for !i=1 to 99
smp1 2000M1+!i 2004M12+!i
equation eq2.ls qx_sa yw_sa c
coef(!i)=eq2.@coefs(1)
next
```

Como último complemento, note que, após fazer isso, seu conjunto de dados amostral se reduziu para 60 dados, mostrado na parte superior do workfile em `sample`. Para fazer o banco de dados contemplarem todos os dados escreva no final do programa:

```
smp1 @all
```

A figura 6.3a mostra como são os resíduos e a 6.3b a distribuição dos mesmos. Veja que a média é zero, satisfazendo a premissa do modelo de regressão  $E(\varepsilon) = 0$ . Mas não possuem uma distribuição normal, sinalizando que podemos melhorar essa estimativa no futuro.

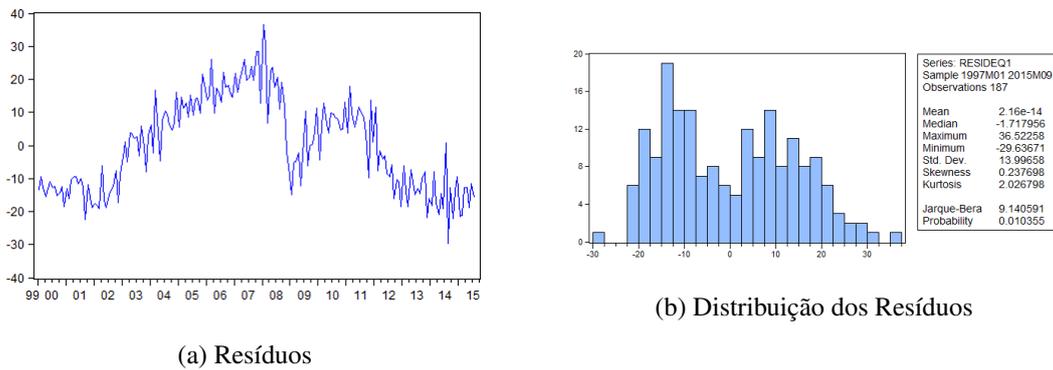


Figura 6.3: Resíduos da Regressão Simples

A seguir, o termo  $\sum_{t=1}^T (Y_t - \bar{Y})^2$  representa o quanto a variável dependente desvia em relação à sua média. O termo é elevado ao quadrado exatamente para evitar que desvios positivos sejam anulados por desvios negativos. A média da variável dependente é mostrada na tabela como “*Mean dependent var*” e, para o nosso exemplo, tem valor  $qx_t = 85.168$ . Para encontrar esse valor podemos usar o comando `scalar media = eq1.@meandep`. Isso também pode ser feito manualmente, onde teremos  $\sum_{t=1}^T (Y_t - \bar{Y})^2 = 85.168$ . Por fim, no nosso exemplo, basta encontrar:

$$R^2 = 1 - \frac{36438.21}{56994.139} = 0.360$$

Veja que, independente do modelo que for utilizado, o denominador da equação acima nunca se modifica. Porém, o numerador, ou seja, o desvio dos erros em relação a sua média, que é igual a zero, será diferente para cada modelo. Ou seja, tem modelos que erram mais que outros. Dessa forma, quanto maior for o numerador, relativamente ao resultado do denominador, mais o modelo estará errando e, com isso, menor será o valor de  $R^2$ . Um modelo que tem erro próximo a zero irá produzir um  $R^2$  próximo ao valor 1.

Logo abaixo dessa estatística há outra que deve ser considerada mais útil, é a “*Adjusted R-squared*”. Nessa, o valor do  $R^2$  é corrigido pelo número de coeficientes que estão sendo utilizadas no modelo. Sua fórmula geral é dada por:

$$\bar{R}^2 = 1 - (1 - R^2) \frac{T - 1}{T - k}$$

Onde T é o número de observações utilizadas e k é o número de coeficientes. No nosso exemplo, temos 187 dados e dois coeficientes, um para a constante e outro para a variável independente. Sendo assim:

$$\bar{R}^2 = 1 - (1 - 0.360) \frac{187 - 1}{187 - 2} = 0.357$$

Porque utilizar o  $\bar{R}^2$  e não o  $R^2$ ? Em regressão simples os dois valores são bem parecidos, pois temos no máximo dois coeficientes a utilizar, a constante e o  $\beta$ . Mas em modelos de regressão múltipla onde k é maior as estimativas podem diferir de forma significativa.

**Programação 6.0.5** Seguindo no exemplo das nossas 100 regressões, podemos pedir agora para que seja criada uma série com todos os valores dos  $R^2$ . Isso será útil para identificar em qual sequência de regressões obtemos a melhor estimativa. Assim, criamos mais uma matriz, só que agora de nome “*explicado*” e pedimos para salvar os valores nela. Note que os valores da regressão com o conjunto total dos dados ficam na última linha dessa matriz:

```

smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
matrix(100) coef
coef(100)=eq1.@coefs(1)
matrix(100) explicado
explicado(100)=eq1.@r2
for !i=1 to 99
smp1 2000M1+!i 2004M12+!i
equation eq2.ls qx_sa yw_sa c
coef(!i)=eq2.@coefs(1)
explicado(!i)=eq2.@r2
next
smp1 @all

```

Veja que o valor do  $R^2$  é obtido usando o comando `.@r2` logo depois do nome da equação (nesse exemplo `eq1` e `eq2`). Caso queira determinar a série de valores do  $R^2$  ajustado use `.@rbar2`.

A seguir, na tabela com o resultado da regressão, há a informação do desvio padrão da regressão, ou então, “*S.E. of regression*”. Seu resultado é dado de forma direta a partir do conhecimento da variância dos resíduos, visto anteriormente:

$$\sum_{t=1}^T (\hat{\varepsilon}_t - \bar{\varepsilon})^2$$

Ou seja, como a média dos resíduos é igual a zero,  $\bar{\varepsilon} = 0$ , a variância pode ser encontrada a partir de:

$$s = \sqrt{\frac{\sum_{t=1}^T \varepsilon_t^2}{(T - k)}}$$

No nosso exemplo,  $s = 14.034$ . O comando para obter esse valor é dado por `scalar desvio = eq1.@se`.

A estatística seguinte mostrada na tabela de resultados é a “*log likelihood*”, ou então, o log da verossimilhança, onde os erros são avaliados supondo-se uma distribuição normal a partir de:

$$l = -\frac{T}{2} \left( 1 + \ln(2\pi) + \ln \left( \frac{\sum_{t=1}^T (\varepsilon_t - \bar{\varepsilon})^2}{T} \right) \right)$$

Para os dados do nosso exemplo, temos que  $T = 187$  e, sendo assim:

$$l = -\frac{187}{2} \left( 1 + \ln(2\pi) + \ln \left( \frac{36438.213}{187} \right) \right) = -758.298$$

O comando no *EViews*<sup>®</sup> para determinar o valor do log da verossimilhança é dado por `scalar logver = eq1.@logl`.

A seguir, está a estatística F. Essa tem como objetivo testar se todos coeficientes das variáveis independentes no modelo, em conjunto, são estatisticamente iguais a zero. Esse teste não inclui a constante. É usada a seguinte fórmula geral para o teste:

$$F = \frac{R^2}{(k - 1)} \frac{T - k}{(1 - R^2)}$$

Para o nosso exemplo, teremos:

$$F = \frac{0.360}{(2-1)} \frac{187-2}{(1-0.360)} = 104.364$$

E, com base no p-valor, podemos rejeitar a hipótese nula de que  $\beta_1 = 0$ . O *Prob* pode ser encontrado usando `scalar probf = 1 - @fdist(104.364, 1, 185)`. E usando `scalar f = eq1.@f` encontramos o teste F.

**Programação 6.0.6** O teste F pode ser feito via programação. Primeiro criamos o scalar de nome *f* que calcula o valor da estatística. A seguir, criamos o scalar de nome *testef* para especificar o p-valor dessa estatística que tem  $k-1$  graus de liberdade no numerador e  $T-k$  graus de liberdade no denominador:

```
scalar f = (eq1.@r2)*(eq1.@npers-eq1.@ncoef)/(eq1.@ncoef-1)*(1-eq1.@r2)
scalar testef=(1-@cdfist(f,eq1.@ncoef-1,eq1.@npers-eq1.@ncoef))
```

Além da estatística  $R^2$ , muito utilizada para comparar modelos, o *EViews*<sup>®</sup> fornece outras três que são bem mais eficientes e que são conhecidas como critérios de comparação. Em ambas, quanto menor o valor, em módulo, melhor. A primeira delas é o critério de Akaike. De forma geral, sua fórmula é dada por:

$$AIC = \frac{2}{T}(k-l)$$

Onde  $l$  é o log da verossimilhança. Usando os dados do nosso exemplo, vemos que:

$$AIC = \frac{2}{187}(2 - (-758.298)) = 8.131$$

Esse valor também pode ser encontrado usando `scalar aic = eq1.@aic`. A segunda estatística é o critério de informação de Schwarz. A vantagem desse método em relação ao de AIC é que agora é aplicada uma espécie de penalidade para o uso de coeficientes adicionais:

$$SC = \frac{1}{T}(k \ln(T) - 2l)$$

O comando no *EViews*<sup>®</sup> que retorna essa estatística é dado por `scalar sc = eq1.@schwarz`. Para os dados do nosso exemplo, teremos:

$$SC = \frac{1}{187}(2 \ln(187) - 2(-758.298)) = 8.166$$

Por fim, também pode ser usado o critério de comparação de Hannan-Quinn, que adiciona mais uma penalidade:

$$HQ = \frac{2}{T}(k \ln(\ln(T)) - l)$$

Usando os dados do nosso exemplo, encontramos:

$$HQ = \frac{2}{187}(2 \ln(\ln(187)) - (-758.298)) = 8.145$$

Para encontrar essa estatística podemos usar o comando `scalar hq = eq1.@hq`. Um ponto importante a destacar é que essas três estatísticas não são comparáveis entre si. Ou seja, de posse de diferentes modelos, comparamos o AIC do modelo 1 com o AIC dos demais modelos. Não usamos a comparação entre AIC e HQ, por exemplo.

**Programação 6.0.7** Ao rodar as 100 regressões, podemos estar interessados em criar uma série de dados que mostre a evolução dos critérios de comparação. Como iremos usar os três critérios, a nova matriz que usaremos, de nome “*critério*”, tem que ter 3 colunas. Criamos a mesma e salvamos os valores desses critérios para a *eq1*. A seguir, ao rodar o loop, fazemos o mesmo para cada uma das outras 99 regressões:

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
matrix(100) coef
coef(100)=eq1.@coefs(1)
matrix(100) explicado
explicado(100)=eq1.@r2
matrix(100,3) critério
critério(100,1)=eq1.@aic
critério(100,2)=eq1.@hq
critério(100,3)=eq1.@schwarz
for !i=1 to 99
smp1 2000M1+!i 2004M12+!i
equation eq2.ls qx_sa yw_sa c
coef(!i)=eq2.@coefs(1)
explicado(!i)=eq2.@r2
critério(!i,1)=eq2.@aic
critério(!i,2)=eq2.@hq
critério(!i,3)=eq2.@schwarz
next
smp1 @all
```

Até esse momento vimos como avaliar os resultados das estatísticas do modelo de regressão e como as mesmas são calculadas. A figura 6.4 traz um resumo das funções utilizadas até o presente momento. Esses comandos devem ser aplicadas em uma equação. Por exemplo, para determinar a número de observações do modelo de de nome *eq1*, é utilizado o comando *eq1.@regobs*.

Dependent Variable: QX_SA				
Method: Least Squares				
Included observations: .@regobs				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
YW_SA	.@coef(1)	.@stderrs(1)	.@tstats(1)	.@pval(1)
C	.@coef(2)	.@stderrs(2)	.@tstats(2)	.@pval(2)
R-squared	.@r2	Mean dependent var		.@meandep
Adjusted R-squared	.@rbar2	S.D. dependent var		.@sddep
S.E. of regression	.@se	Akaike info criterion		.@aic
Sum squared resid	.@ssr	Schwarz criterion		.@schwarz
Log likelihood	.@logl	Hannan-Quinn criter.		.@hq
F-statistic	.@f	Durbin-Watson stat		.@dw
Prob(F-statistic)	.@fprob			

Figura 6.4: Comandos para Resultados do Modelo de Regressão

Após avaliar esses resultados, podemos ver, graficamente, como o nosso modelo, para o conjunto de dados, se comportou. Para tal, com a janela de resultados da nossa regressão aberta,

clique em **Resids**. O mesmo irá mostrar o gráfico conforme figura 6.5.

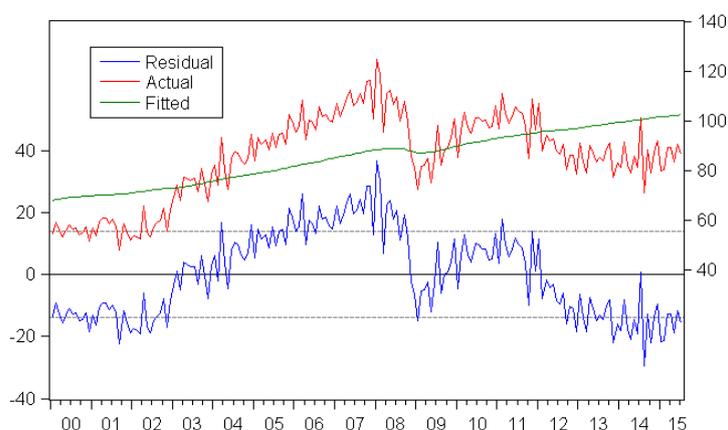


Figura 6.5: Resultados do Modelo de Regressão

Tal como citado na legenda do gráfico, a linha vermelha mostra os verdadeiros valores da variável dependente, no nosso caso,  $qx$ . A linha verde são as estimativas obtidas a partir do modelo de regressão. E, por fim, a linha azul é a série de resíduos que nada mais é que a diferença entre o verdadeiro valor e o estimado. Note que o nosso modelo não é tão bom para reproduzir o comportamento de  $qx$  em determinados momentos, errando muito.

Nesse momento, o leitor pode estar se perguntando como é feita a estimativa dos valores para cada período. Vamos recordar a equação encontrada:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)      (0.063)

Com base nela podemos determinar qualquer valor de  $qx$  ao longo do tempo. Por exemplo, a estimativa para janeiro de 2000 pode ser dada a partir da substituição do respectivo valor da variável independente naquela data. Ou seja, olhando a série de  $yw_t$ , ajustada sazonalmente, vemos que, em janeiro de 2000 tem-se  $yw_{jan/2000} = 76.333$ . Sendo assim, podemos encontrar o valor de  $qx_{jan/2000}$  fazendo:

$$qx_{jan/2000} = 18.389 + 0.647(76.333) = 67.850$$

Esse procedimento pode ser repetido para qualquer mês que se queira avaliar, modificando apenas o respectivo valor de  $yw_t$  e mantendo fixo o coeficiente da constante, 18,389, e da inclinação, 0,647.

Após fazer a regressão é necessário proceder a uma investigação detalhada sobre os resultados. Há no *EViews*® 3 blocos de testes que são explorados nas seções a seguir. Primeiro é feita a investigação sobre os coeficientes. A seguir sobre os resíduos e, por fim, sobre a estabilidade do modelo.

## 6.1 Diagnóstico Dos Coeficientes

Alguns estatísticas podem ser avaliadas para testar a robustez dos coeficientes. Com uma equação aberta, o diagnóstico dos coeficientes pode ser acessado em **View/Coefficient Diagnostics**. Note que são nove diferentes tipos de testes que avaliaremos na sequência.

### 6.1.1 Scaled Coefficients

Essa opção só funciona se a equação for estimada a partir de um comando de lista. Lembre-se disso, pois vários outros testes exigem esse formato. Como é o modelo estimado em lista? Você terá que, ao abrir a janela de estimativa da equação, escrever as variáveis em ordem. No nosso exemplo colocamos `qx_sa yw_sa c`.

Essa opção permite que se tenha uma visão da estimativa dos coeficientes, os coeficientes padronizados e as elasticidades médias. Para o nosso exemplo, temos os resultados mostrados na figura 6.6.

Variable	Coefficient	Standardized Coefficient	Elasticity at Means
YW_SA	0.647967	0.600556	0.784082
C	18.38936	2.04E-24	0.215918

Figura 6.6: Scaled Coefficients

Na primeira coluna estão as variáveis, na segunda coluna o valor dos coeficientes estimados. Na terceira coluna os coeficientes padronizados e, por fim, a estimativa das respectivas elasticidades no ponto médio. Essa tabela também pode ser encontrada usando o comando `eq1.coefscale`.

### 6.1.2 Intervalo de Confiança

Uma informação útil na interpretação dos resultados de uma regressão é usar o intervalo de confiança dos coeficientes. Ao clicar em **View/Coefficient Diagnostics**, selecione **Confidence Intervals...** Na janela que irá abrir, podemos selecionar qualquer tamanho para o intervalo de confiança. Por exemplo, na figura 6.7 mostramos como seriam os resultados para estimativas com 95% de significância.



Coefficient Confidence Intervals  
Date: 11/09/15 Time: 18:40  
Sample: 2000M01 2015M07  
Included observations: 187

Variable	95% CI	Coefficient	95% CI
YW_SA	0.522833	0.647967	0.773101
C	5.335207	18.38936	31.44351

Figura 6.7: Intervalo de Confiança

Note que, ao não marcar a opção **Arrange in pairs**, os resultados mostrados são mais fáceis de interpretar, com o intervalo mínimo à esquerda, no meio a média do coeficiente e, depois, o intervalo máximo. A tabela com os intervalos de confiança pode ser obtida usando o comando `eq1.cinterval(nopair) .95`. Para encontrar esses valores a um nível de significância de 95% e uma distribuição *t-student*, o resultado para o coeficiente de  $yw_t$  será dado por:

$$\begin{aligned}
 yw - 1,972\sigma_{y,w} < yw < yw + 1,972\sigma_{y,w} \\
 0.647 - 1,972(0.063) < yw < 0.647 + 1,972(0.063) \\
 0,522 < yw < 0,773
 \end{aligned}$$

Com 99% de significância, usamos:

$$\begin{aligned}
 yw - 2,346\sigma_{y,w} < yw < yw + 2,346\sigma_{y,w} \\
 0.482 < yw < 0.813
 \end{aligned}$$

O mesmo também pode ser feito para todos os demais coeficientes encontrados, inclusive a constante. A forma de interpretar esse resultado é: “Acredita-se que o valor de  $yw_t$  tem 95% de probabilidade de ficar entre 0,522 e 0,773”.

**Programação 6.1.1** Para o nosso exemplo de 100 regressões, podemos pedir para que seja criado, a cada passo, o intervalo de confiança para o primeiro coeficiente. Nesse caso, mudamos a matriz “coef” para 3 colunas onde, na primeira, temos o intervalo inferior, a 95%; na segunda coluna temos a estimativa do coeficiente; na terceira coluna o intervalo superior a 95%. Note que também é modificada a parte do loop:

```

smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
matrix(100,3) coef
coef(100,2)=eq1.@coefs(1)
coef(100,1)=eq1.@coefs(1)-1.975*eq1.@stderrs(1)
coef(100,3)=eq1.@coefs(1)+1.975*eq1.@stderrs(1)
for !i=1 to 99
smp1 2000M1+!i 2004M12+!i
equation eq2.ls qx_sa yw_sa c
coef(!i,2)=eq2.@coefs(1)
coef(!i,1)=eq2.@coefs(1)-1.972*eq2.@stderrs(1)
coef(!i,3)=eq2.@coefs(1)+1.972*eq2.@stderrs(1)
next
smp1 @all

```

Um ponto importante na construção do intervalo de confiança é definir o valor na curva de distribuição. Nesse caso, como usamos a curva *t-student*, devemos ter em mente que é necessário especificar também os graus de liberdade. Sendo assim, o valor de 1,972 para 95% só é válido para 185 graus de liberdade do nosso modelo (N-k), onde N é o número de dados e k o número de coeficientes. Se o número de dados ou o número de coeficientes variarem, o valor para 95% não será mais 1,972. Felizmente existe uma função no *EViews*<sup>®</sup> que permite encontrar esse ponto na curva de distribuição: **@qtdist(área, graus de liberdade)**. No nosso exemplo, queremos saber o ponto para 95%. Note que, como temos uma curva bi-caudal, devemos especificar uma área dividida em dois (5%/2=0,025), sendo assim, o valor de área=0,975 e os graus de liberdade=185. Com isso podemos encontrar 1,972.

**Programação 6.1.2** Encontrando o ponto na curva *t-student* que especifica o intervalo de confiança de acordo com uma área e um valor dos graus de liberdade. Para encontrar o ponto no qual a área interna é 90% e temos 185 graus de liberdade:

Coefficient Confidence Intervals  
 Date: 03/28/16 Time: 22:24  
 Sample: 1997M01 2015M09  
 Included observations: 187

Variable	99% CI	95% CI	90% CI	Coefficient	90% CI	95% CI	99% CI
YW_SA	0.482887	0.522833	0.543113	0.647967	0.752821	0.773101	0.813047
C	1.167956	5.335207	7.450861	18.38936	29.32786	31.44351	35.61076

Figura 6.8: Intervalo de Confiança 90% 95% 99%

```
scalar ponto=@qtdist(0.95,185)
```

Alternativamente podemos pedir uma estimativa de intervalo de confiança com vários níveis de significância. Para tanto podemos usar o comando `eq1.cinterval(nopair) .90 .95 .99` que irá produzir os resultados da figura 6.8.

### 6.1.3 Teste de Wald

Ao criar intervalos de confiança, podemos ter uma idéia de inferência sobre valores mínimos e máximos. Porém, podemos querer testar algumas restrições nos coeficientes. Isso pode ser feito a partir do teste de Wald. Vá em **View/Coefficient Diagnostics/Wald Test...** A seguir, vamos testar se o coeficiente de  $yw_t$  é estatisticamente igual a 2. Nesse caso, temos:

$$H_0 : c(1) = 2 \text{ ou então: } H_0 : c(1) - 2 = 0$$

$$H_a : c(1) \neq 2 \text{ ou então: } H_a : c(1) - 2 \neq 0$$

Como temos que  $c(1) = 0.647$ , então  $c(1) - 2 = -1.353$ . Esse é o valor reportado no sumário da hipótese nula e que deverá ser testado. O *Std. Error*, ou seja, o desvio-padrão a ser usado nesse teste é o mesmo da estimativa de regressão para o coeficiente em questão. Nesse caso,  $\sigma_{c(1)} = 0.063$ . Sendo assim, podemos encontrar a estatística *t-student* a partir de:

$$t = \frac{x - \mu}{\sigma}$$

$$t = \frac{-1.353}{0.063} = -21.316$$

No caso da estatística *t* o *probability* é dado a partir de  $(1 - @ctdist(-21.316, 185))$ . Lembre-se que esse é um teste bicaudal. Note que também é mostrado o resultado para um teste F. No geral, o teste F que compara dois modelos é dado por:

$$F = \frac{\left[ \frac{SSE_2 - SSE_1}{k_1 - k_2} \right]}{\left[ \frac{SSE_1}{n - k_1} \right]}$$

Onde  $n$  é o número de observações de um modelo não restrito, que no nosso caso é o resultado com  $c$  e  $yw\_sa$  e dado por 187;  $k_1$  é o número de parâmetros do modelo não restrito,  $k_1 = 2$  no nosso exemplo, dado pelo parâmetro da constante e do coeficiente de  $yw\_sa$ ;  $SSE_1$  é a soma ao quadrado dos resíduos de um modelo não restrito, que para o nosso exemplo é dado por 36438,2. Esse modelo não restrito é combinado ao modelo restrito onde teríamos que testar a hipótese de  $C(1) = 2$ . Para tanto vamos estimar uma equação onde  $qx_t = c(1) + 2yw_t$  o resultado será  $qx_t = 120,9 + \varepsilon_t$ ,

uma equação com apenas um parâmetro, ou seja,  $k_2 = 1$ . Tendo a  $SSE_2 = 125934,7$ . Substituindo esses valores no teste F encontramos:

$$F = \frac{\left[ \frac{125934,7 - 36438,2}{2-1} \right]}{\left[ \frac{36438,2}{187-2} \right]} = 454,3817$$

Para encontrar o p-valor desse teste é só fazer `scalar pvalorf = (1-@cfdist(454.3817, 1, 185))`. O teste F é válido nesse caso apenas se assumirmos que os erros são independentes e com distribuição normal. Assim, pelo resultado do p-valor (*probability*), podemos dizer que o coeficiente de  $yw_t$  é estatisticamente diferente de 2. Como pode ser visto na figura 6.9.

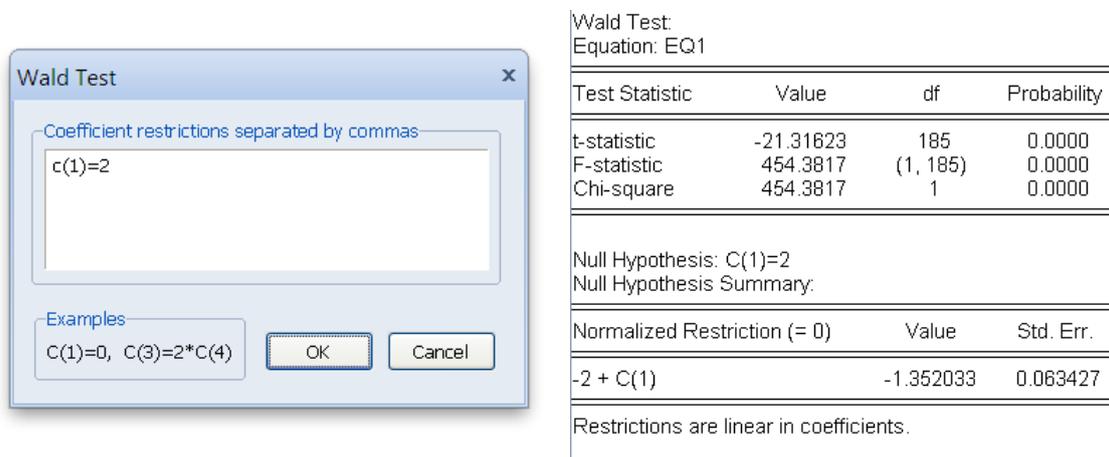


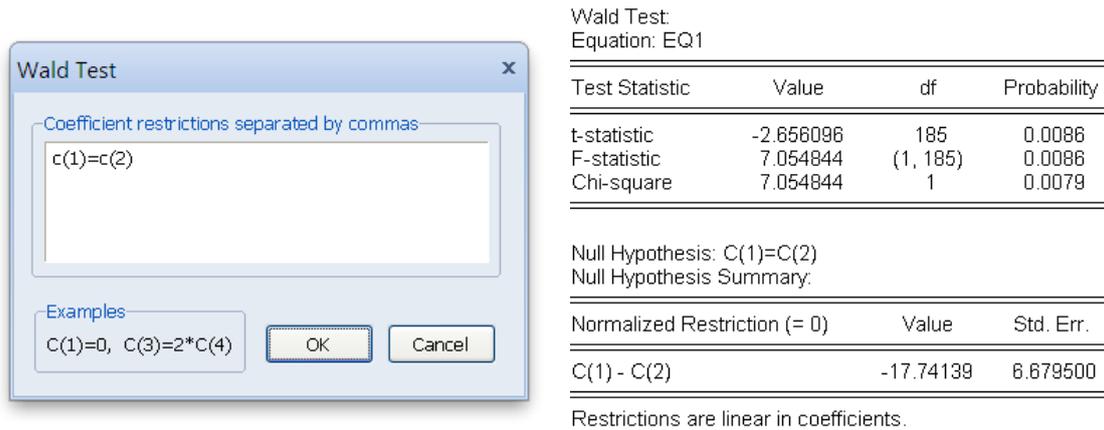
Figura 6.9: Teste de Wald  $c(1)=2$

Também podemos estar interessados em testar se os coeficientes do nosso modelo de regressão são estatisticamente iguais. Nesse caso, devemos mudar a hipótese nula para:

$$H_0 : c(1) = c(2) \text{ ou então: } H_0 : c(1) - c(2) = 0$$

$$H_a : c(1) \neq c(2) \text{ ou então: } H_a : c(1) - c(2) \neq 0$$

Para fazer isso no *EViews*<sup>®</sup>, vá em **View/Coefficient Diagnostics/Wald Test...**, e especifique tal como mostrado na figura 6.10. Pelo resultado do teste, não é possível aceitar a hipótese nula. Sendo assim, os dois coeficientes são estatisticamente diferentes.

Figura 6.10: Teste de Wald  $c(1)=c(2)$ 

### 6.1.4 Confidence Ellipse

Apesar do teste de Wald ser muito útil, é normal que se queira testar mais de uma restrição, como por exemplo, se  $c(1)=0$  e ao mesmo tempo, se  $c(2)=0$ . Nesse caso, o teste de Wald não é o mais apropriado, e devemos recorrer a **View/Coefficient Diagnostics/Confidence Ellipse**. Isso pode ser feito apenas digitando os coeficientes, omitindo o valor “0”. Assim, da forma como digitado,  $c(1)$  é o mesmo que testar se  $c(1)=0$ . Como também deixamos  $c(2)$ , estamos, na verdade, testando se  $c(1)=c(2)=0$ . Em confidence levels, selecione 0.95 (95%). Em individual intervals, selecione **Shade**, que é uma opção melhor de visualizar os resultados. A seguir, clique em **Ok**.

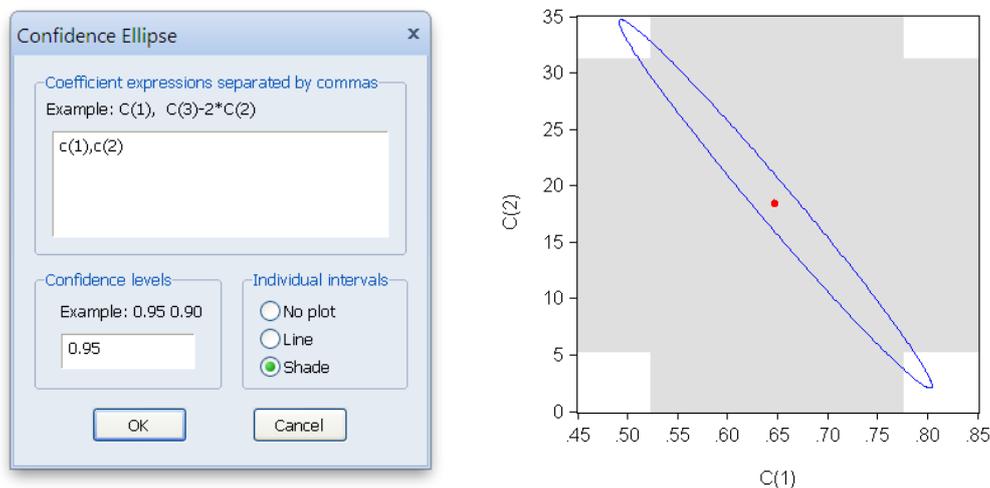


Figura 6.11: Confidence Ellipse

Como mostra a figura 6.11, há um ponto central na cor vermelha. Esse representa a estimativa dos dois coeficientes na equação de regressão, com  $c(1) = 0,647$  e  $c(2) = 18,389$ . Para conferir isso coloque o mouse sobre o ponto vermelho que esses valores irão aparecer. A área que está na cor cinza representa o intervalo de confiança individual para um teste a 95% de significância, ou seja, para cada um dos coeficientes. Note que, para o coeficiente

$c(1)$  esse é dado por  $0,522 < c(1) < 0,773$  no eixo horizontal. Lembre-se que encontramos esse valor do intervalo de confiança anteriormente. Para a constante, que é o segundo coeficiente,  $5,335 < c(2) < 31,443$  e está no eixo vertical. Dentro do círculo está o resultado do teste conjunto. No nosso caso, testando se  $c(1) = c(2) = 0$ . Esse gráfico pode ser gerado usando o seguinte comando no *EViews*<sup>®</sup>: `eq1.cellipse(ind=shade) C(1)=0, C(2)=0`.

A análise pode ser feita tanto para um teste individual quanto para um teste conjunto. Por exemplo, se quisermos testar a 95% se  $c(1) = -1,2$ , vemos que esse valor está fora da área cinza do gráfico na linha horizontal. Sendo assim, rejeitamos a hipótese nula. Para comprovar esse resultado faça o teste de Wald para  $c(1)$ . Da mesma forma, podemos testar se  $c(2) = 2,5$ . Olhando no gráfico vemos que esse valor está fora da área cinza (não se esqueça de agora ver a linha vertical). Dessa forma, rejeitamos a hipótese nula.

Mas, se queremos um teste conjunto entre dois coeficientes, como no nosso caso, entre  $c(1)$  e  $c(2)$ , devemos olhar para a elipse. Sempre que a combinação entre os dois pontos ficar dentro da elipse, não é possível rejeitar a hipótese nula. Uma outra opção interessante é colocar mais de um intervalo. Na caixa de opção **Confidence levels** digite **0.99 0.90** e em **Individual intervals** a opção **Line**. Tal como no comando `eq1.cellipse(ind=line, size = 0.99 0.90) C(1)=0, C(2)=0`. O resultado é como na mostrado na figura 6.12.

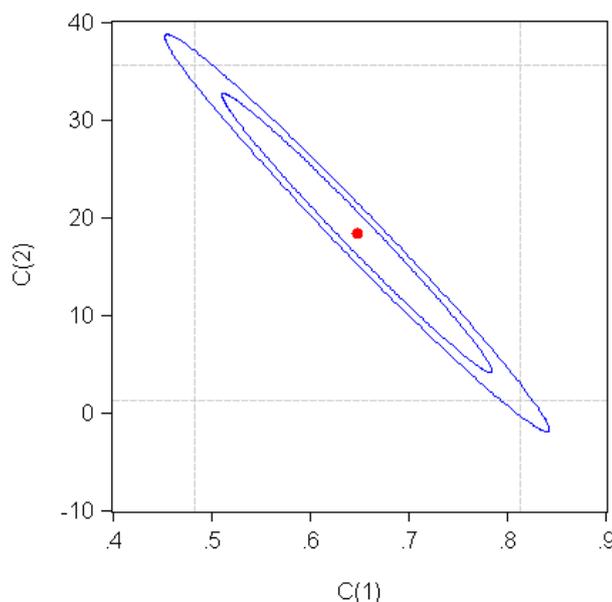


Figura 6.12: Confidence Levels: 0.99 0.90

### 6.1.5 Variance Inflation Factors

Sample: 1997M01 2015M09  
Included observations: 187

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
YW_SA	0.004023	41.56776	1.000000
C	43.78248	41.56776	NA

Figura 6.13: Variance Inflation Factors

Essa medida tem como objetivo apontar o nível de colinearidade que existe entre as variáveis independentes do modelo. Duas variáveis são ditas serem colineares se todos os pontos estiverem sob uma linha reta. Nesse sentido, se duas variáveis possuem determinado grau de colinearidade é natural esperar que uma esteja influenciando a estimativa do coeficiente da outra. O VIF permite identificar a presença de colinearidade na nossa equação dividindo a variância dos parâmetros em questão. O resultado é apresentado de duas formas. O VIF centrado é encontrado a partir da divisão da variância do coeficiente obtida no modelo completo, no nosso caso 0,004023, pela variância do mesmo coeficiente mas estimado a partir de um modelo que contenha apenas a constante e o coeficiente em questão. Como temos um

modelo de regressão simples, esses dois valores são iguais, resultando em um VIF centrado em  $y$  de 1. Veja na figura 6.13.

A segunda medida é o VIF não-centrado. Esse é dado pela razão da variância do coeficiente obtida a partir de um modelo completo (0,004023) e um modelo sem constante (faça uma regressão  $qx\_sa \ c(2)yw\_sa$  e encontrará  $\beta = 0,82211$  com variância de 0,00010028). Esse resultado pode ser acessado a partir de `eq1.varinf` na janela de comando.

### 6.1.6 Decomposição da Variância do Coeficiente

#### Coefficient Variance Decomposition

Date: 03/29/16 Time: 15:09

Sample: 1997M01 2015M09

Included observations: 187

Eigenvalues	43.78641	9.68E-05
Condition	2.21E-06	1.000000

#### Variance Decomposition Proportions

Variable	Associated Eigenvalue	
	1	2
YW_SA	0.975947	0.024053
C	1.000000	1.98E-10

#### Eigenvectors

Variable	Associated Eigenvalue	
	1	2
YW_SA	-0.009469	-0.999955
C	0.999955	-0.009469

Figura 6.14: Decomposição da Variância do Coeficiente

por 0,00000221, associado ao autovalor 43,78. Isso sinalizaria que temos colinearidade. Porém, estamos trabalhando apenas com uma variável independente. Esse tipo de investigação faz sentido em um modelo com mais de uma variável independente. A tabela com os resultados apontados acima pode ser facilmente encontrada usando `eq1.cvardec`.

Essa é uma ferramenta útil para determinar a existência de uma possível colinearidade entre as variáveis independentes. O método se dá pela construção da matriz de covariância dos coeficientes, a seguir, são encontrados os autovetores e, por fim, a proporção da decomposição da variância. Vejamos como interpretar esses resultados para a regressão que estamos usando. A figura ?? mostra esses cálculos.

A última parte da tabela mostra a estimativa dos autovetores para os dois parâmetros do modelo (para entender isso consulte o capítulo sobre análise de componente principal). A partir desses autovetores obtém-se a proporção da decomposição da variância, mostrado no meio da tabela. Por fim, é feito o cálculo do *condition number*. Como regra se esse valor é menor que  $1/900 = 0,001$  então há colinearidade. Se for verificado na linha *condition* mais de um resultado menor que 0,001 então é necessário avaliar a proporção da decomposição da variância. Veja que no nosso exemplo o *condition* apresenta um resultado menor que 0,001 dado

### 6.1.7 Variáveis Omitidas

Frequentemente nos deparamos com a possibilidade de inserir uma nova variável no modelo de regressão como forma de melhorar o poder de explicação do mesmo. Porém, pode ocorrer de, ao se fazer isso, a contribuição não seja tão boa. Nesse caso, o ideal seria fazer um teste de variáveis omitidas. Já fizemos o modelo de regressão mais básico, onde:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)                      (0.063)

Agora, vamos investigar se a omissão, por exemplo, da variável  $px_t$  foi ruim para o modelo. Nesse caso, com a janela da equação acima aberta, vá em **View/Coefficient Diagnostics/Omitted Variables Test...** A seguir, digite o nome da variável em questão (ela tem que existir no *workfile*). Os resultados são apresentados na figura 6.15. Ao final será possível identificar a regressão na forma:

$$qx_t = -21.625 + 1.531yw_t - 0.428 px_t + \varepsilon_t$$

(10.096)                      (0.185)                      (0.085)

Esse é o primeiro contato com a ideia de regressão múltipla. O resultado dessa regressão aparece na parte final da tabela. A estatística  $t$  e o *Prob* são avaliados como anteriormente.

Ou seja, para poder fazer o teste, primeiro é rodada uma regressão com a presença da variável que está supondo ter sido omitida. Posteriormente, são feitos os testes e apresentados no início da tabela. O primeiro resultado para o teste *t-statistic*, refere-se apenas à hipótese de o coeficiente da nova variável, no nosso caso,  $px_t$ , ser estatisticamente igual a zero. Pelo p-valor, rejeitamos a hipótese nula e, individualmente, o coeficiente é diferente de zero. Ou seja, desse ponto de vista, ele seria importante para o modelo. Veja a primeira parte da tabela na figura 6.15. Aqui é desnecessário mostrar como chegamos no *Probability*, pois já comentamos isso anteriormente.

Logo abaixo está o teste *F-statistic*, que representa o teste conjunto para ver se todas as variáveis são estatisticamente iguais a zero, ou seja, se  $c(1) = c(2) = c(3) = 0$ . Porém, esse teste é feito com base em um modelo restrito (sem a variável  $px_t$ ) relativamente a um modelo não-restrito, com a presença da variável  $px_t$ . O conjunto de informações em “*F-test summary*” mostra os resultados para a soma do quadrado dos resíduos para os dois modelos, o restrito(sem a variável  $px$ ) e o não restrito (com a variável  $px$ ).

$$F_{stat} = \frac{\frac{(SSR_R - SSR_{UR})}{q}}{\frac{SSR_{UR}}{(T-k)}}$$

Onde  $SSR_R$  é a soma dos resíduos ao quadrado do modelo restrito,  $SSR_{UR}$  é a soma ao quadrado do modelo não-restrito, com todas as variáveis,  $q$  é o número de restrições impostas,  $T$  é o número de observações e  $k$  é o número de parâmetros presentes no modelo não restrito. A hipótese nula é que a variável que foi omitida não é significativa para o modelo. Substituindo esses valores encontramos:

$$F_{stat} = \frac{\frac{(36438,21 - 32031,26)}{1}}{\frac{32031,26}{(187-3)}} = 25,315$$

Assim, o valor de  $F = 25,315[0,000]$  sinaliza que rejeitamos a hipótese nula e, os coeficientes não são iguais e, dessa forma, adicionar a variável  $px_t$  no modelo representa ganhos. Note que o teste  $F$  para variáveis omitidas tem distribuição  $X_q^2$  onde  $q$  é o número de restrições impostas. Nesse caso, podemos encontrar o p-valor diretamente no *EViews*<sup>®</sup>.

**Programação 6.1.3** Podemos encontrar o p-valor do teste escrevendo um comando no *EViews*<sup>®</sup>. Na barra de ferramentas, clique em **Window** e depois selecione “**Command**”. Essa ação irá abrir uma parte em branco na parte superior do *EViews*<sup>®</sup>. Ali podemos escrever o comando abaixo e verificar que ele cria uma variável escalar de nome *testef* com o resultado do p-valor.

```
scalar testef
testef=@chisq(25.315,1)
```

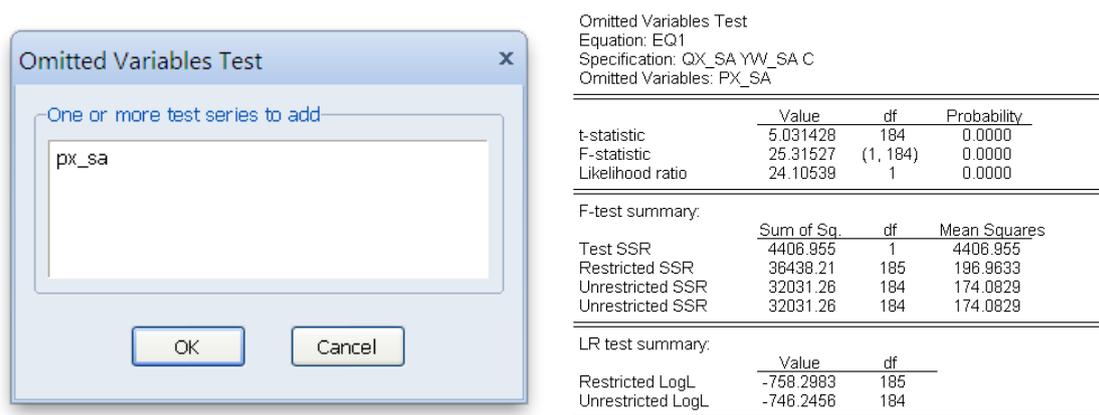


Figura 6.15: Variáveis Omitidas - px\_sa

Note que devido ao fato de executarmos as linhas de programação pela janela de comandos, devemos executar cada linha de uma vez.

Por fim, temos o teste de razão de verossimilhança, conhecido como LR. Esse também tem como objetivo comparar o modelo restrito e o não-restrito e tem, como hipótese nula, que adicionar uma nova variável não seria significativa para o modelo. De forma geral, o teste é dado por:

$$LR = -2(l_{restrito} - l_{nao-restrito})$$

Onde  $l_{restrito}$  é o log da verossimilhança para o modelo restrito. No nosso exemplo, olhando os resultados das estimativas, temos que:

$$LR = -2(-758.298 - (-746.245)) = 24.105$$

E, pelo resultado do p-valor, mostrado no início da tabela, rejeitamos a hipótese nula de que inserir a variável não é estatisticamente significativo para o modelo. Portanto, concluímos pela importância de inserir a variável  $px_t$ . Um lembrete importante: esse teste não se aplica quando usamos variáveis dependentes defasadas. Isso ficará mais claro após ter estudado os modelos autoregressivos. A tabela com os resultados para o teste de variáveis omitidas pode ser facilmente encontrada usando: `eq1.testadd px_sa` para o nosso exemplo.

Também podemos testar a omissão de mais de uma variável. Seja por exemplo, o modelo básico, restrito, dado por:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)            (0.063)

E queremos testar se a omissão da variável  $px_{sa}$  e  $pm_{sa}$  são estatisticamente significativas para o modelo ou não. Nesse caso, o modelo completo seria dado por:

$$qx_t = -126.641 + 1.815yw_t - 1.343 px_t + 1.686 pm_t + \varepsilon_t$$

(15.682)            (0.163)            (0.135)            (0.209)

Para fazer esse teste, com a janela da equação acima aberta, vá em **View/Coefficient Diagnostics/Omitted Variables Test...** A seguir, digite o nome das variáveis que estão sendo omitidas, tal como mostrado na figura 6.16.

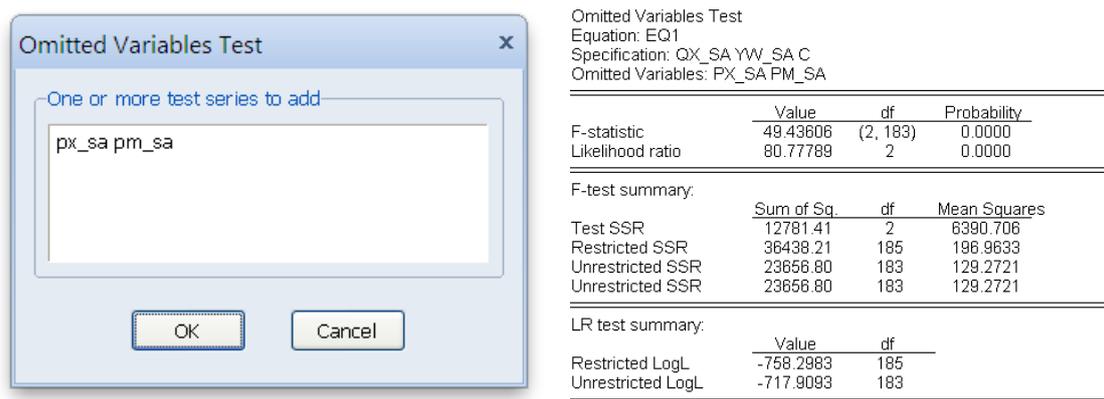


Figura 6.16: Variáveis Omitidas - px\_sa e pm\_sa

Note que não é mostrado o resultado para o teste t. Isso ocorre pois estamos testando mais de uma variável. Tanto pelo teste F quanto pelo LR rejeitamos a hipótese nula de que inserir as variáveis não é estatisticamente significativo para o modelo. Ou seja, a inclusão dessas variáveis no nosso modelo deve resultar em melhora nas estimativas. Nesse caso, o teste F é dado a partir de:

$$F_{stat} = \frac{\frac{(36438.213 - 23656.802)}{2}}{\frac{23656.802}{(187-4)}} = 49.436$$

E o teste LR é dado por:

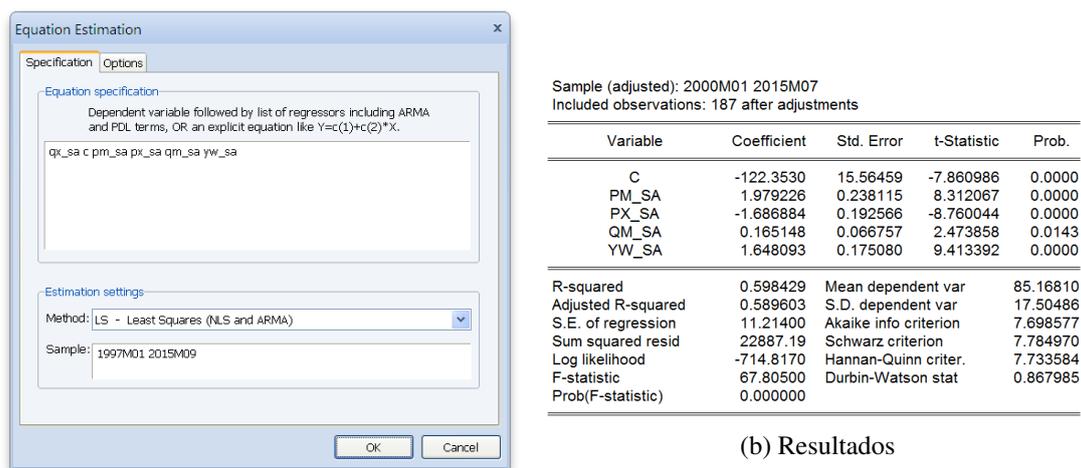
$$LR = -2(-758.298 - (-717.909)) = 80.777$$

**Programação 6.1.4** Podemos fazer o teste LR para variáveis omitidas a partir da programação. Nesse caso, de acordo com o nosso exemplo, especifique a equação restrita, que tem apenas uma variável independente e a não-restrita, com duas variáveis independentes. Após estimar, calcule o teste usando o comando do log da verossimilhança.

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c 'equação restrita'
equation eq4.ls qx_sa yw_sa px_sa c 'equação não-restrita'
matrix(1,2) testeomitida 'cria uma matriz com 1 linha e 2 colunas'
testeomitida(1,1)=-2*(eq1.@logl)+2*(eq4.@logl)
testeomitida(1,2)=@chisq(testeomitida(1,1),1)
' o número de graus de liberdade no teste quiquadrado é igual ao número de restrições, variáveis omitidas.
```

### 6.1.8 Variáveis Redundantes

Um teste complementar ao teste de variáveis omitidas seria verificar se um conjunto de variáveis do modelo poderia ser excluído sem prejuízo. Esse é o tipo de investigação que só faz sentido em modelos de regressão múltipla, onde o método de estimação foi mínimos quadrados, TSLS, binário do tipo logit e demais que possuem variável dependente do tipo ordenada. Outro ponto importante para fazer esse teste é que ele só funciona se quando for estimar a equação utilizar variáveis em lista. Vamos exemplificar esse teste usando um modelo completo, especificado em lista, como na figura 6.17a.

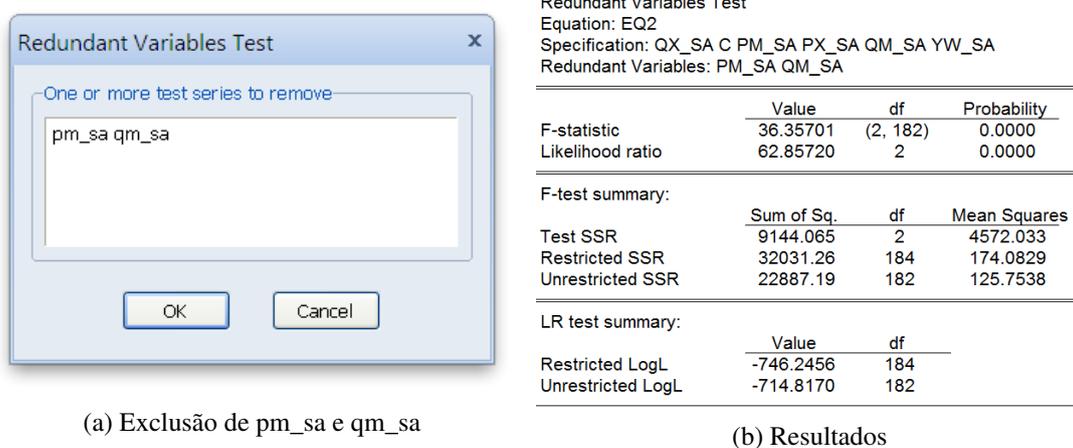


(a) Especificação em Lista

(b) Resultados

Figura 6.17: Variáveis Redundantes - Regressão eq2

Note que primeiro escreve-se a variável dependente, seguida das demais independentes e, se for o caso, a constante. A figura 6.17b mostra os resultados dessa regressão. Nomeie-a como eq2. A seguir vá em **View/Coefficient Diagnostics/Redundant Variables Test...** e especifique as variáveis que quer investigar se podem ser excluídas do modelo. No nosso exemplo vamos escolher pm\_sa e qm\_sa, como mostrado na figura 6.18a. Os resultados dos testes aparecem como mostrado na figura 6.18b.



(a) Exclusão de pm\_sa e qm\_sa

(b) Resultados

Figura 6.18: Variáveis Redundantes

Na parte inferior do resultado aparece a regressão sem os dois parâmetros de restrição que estamos testando. Veja que, da mesma forma do teste de variáveis omitidas, saí fornecido os resultados par o teste LR e F. Seus valores são calculados como anteriormente, dispensando apresentação. Esse resultado também pode ser obtido usando o comando `eq2 .testdrop pm_sa qm_sa`.

### 6.1.9 Teste Factor Breakpoint

Esse teste é uma forma de encontrar uma possível mudança estrutural na equação. A maneira de fazer isso é estimar a equação em diferentes subperíodos da amostra e depois comparar os respectivos

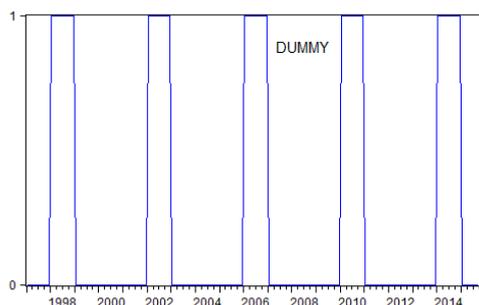
modelos via soma dos resíduos ao quadrado em um típico teste F. Fizemos isso anteriormente ao comparar o SSE de um modelo restrito com outro não restrito.

Além do teste F também é reportado o teste LR. Onde a hipótese nula é de ausência de quebra estrutural e tem distribuição  $X_2^2$  com  $(m - 1)k$  graus de liberdade. Aqui  $k$  é o número de parâmetros na equação e  $m$  o número de subamostras. Por fim tem o teste de Wald, onde a hipótese nula é de ausência de mudança estrutural. Esse é um teste que, para ser feito, tem que ter uma variável *dummy* especificando as datas em que possivelmente tenha ocorrido uma quebra estrutural. Até esse momento o leitor não foi apresentado ao conceito de quebra estrutural nem variáveis *dummy*. Uma variável *dummy* é uma variável indicador, que assume valores 0(zero) e 1(um). É uma típica variável categórica, e que veremos sua aplicação em várias áreas da econometria, como modelos probit, logit, em quebra estrutural e etc. Vamos criar uma variável *dummy* no nosso banco de dado que separe dois intervalos de tempo. Nesse caso, o objetivo é testar se os anos eleitorais no Brasil resultaram em problemas para o nosso modelo. O box de programação 6.1.5 mostra como criar essa *dummy*.

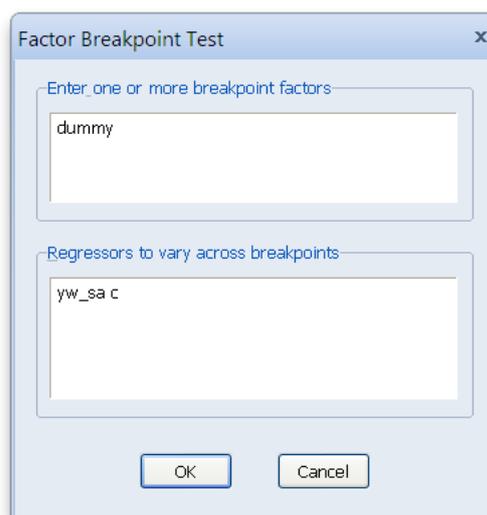
**Programação 6.1.5** Para criar uma variável *dummy* podemos usar vários recursos; digitar os valores diretamente no *EViews*<sup>®</sup>; criar a série no Excel e copiar para o *EViews*<sup>®</sup>; usar programação. Abaixo mostramos como criar uma *dummy* de valor 1 para os anos eleitorais e 0(zero) para os demais anos.

```
series dummy=0
smpl 1998m1 1998m12 2002m1 2002m12 2006m1 2006m12 2010m1 2010m12 2014m1
2014m12
dummy=1
smpl @all
```

A figura 6.19a mostra como ficou nosso gráfico da variável *dummy*. Veja que nos anos eleitorais esta assume o valor 1. Agora vamos ver se esses períodos têm impacto no modelo. Vamos usar a equação eq1 como base. Abra ela e depois vá em **View/Coefficient Diagnostics/Factor Breakpoint Test...** e insira a variável *dummy* no quadro, como mostrado na 6.19b.



(a) Variável Dummy



(b) Inserindo Dummy

Figura 6.19: Teste Factor Breakpoint

O resultado é tal como mostrado na figura 6.20. Nas primeiras linhas estão descritas a variável considerada como fator no teste, a hipótese nula e o fato de que estamos testando um impacto em todos os parâmetros do modelo. São fornecidas três estatísticas, em todas não é possível rejeitar a hipótese nula de ausência de quebra estrutural. Sendo assim, não podemos afirmar que os ciclos eleitorais estejam afetando o nosso modelo.

Factor Breakpoint Test: DUMMY			
Null Hypothesis: No breaks at specified breakpoints			
Varying regressors: All equation variables			
Equation Sample: 2000M01 2015M07			
F-statistic	0.443602	Prob. F(2,183)	0.6424
Log likelihood ratio	0.904406	Prob. Chi-Square(2)	0.6362
Wald Statistic	0.887204	Prob. Chi-Square(2)	0.6417
Factor values:	DUMMY = 0		
	DUMMY = 1		

Figura 6.20: Resultados do Teste Factor Breakpoint

## 6.2 Diagnóstico Dos Resíduos

Além do diagnóstico dos coeficientes há uma série de opções para diagnóstico dos resíduos e que serão vistas nessa seção. O processo de avaliar os resíduos é muito importante, pois é ali que ficam caracterizados todos os problemas que possam existir na especificação do modelo. Após estimar uma equação de regressão, há uma série de pressupostos que devem ser investigados como forma de validar o modelo. Por exemplo, de uma forma geral, ao especificar nosso modelo de regressão colocamos:

$$qx_t = \alpha_1 + \beta_1 yw_t + \varepsilon_t$$

E, na verdade, apesar de não ter sido afirmado, estamos supondo que:

$$\varepsilon_t \sim NIID(0, \sigma^2)$$

Ou seja, estamos supondo que os resíduos têm distribuição normal, são independentes e identicamente distribuídos, tem média zero e variância finita. E esses pressupostos são importantes para garantir que o modelo tem boa especificação. Assim, esse passo tem como objetivo investigar cada uma dessas afirmações. Vamos começar pela mais simples e que menos influência pode ter nos resultados, que é a distribuição normal dos resíduos.

### 6.2.1 Teste de Normalidade

Já vimos anteriormente como podemos testar se uma série de dados possui distribuição normal. Agora, queremos saber se os resíduos da regressão (eq1) são distribuídos normalmente:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)      (0.063)

Com a janela da regressão aberta, vá em **View/Residual Diagnostics/Histogram – Normality test**. Ao fazer isso, será retornado o gráfico da distribuição dos resíduos bem como diversas estatísticas descritivas, como média, mediana, desvio-padrão, assimetria, curtose e, a mais importante, o teste de Jarque-Bera. Já vimos isso em capítulo anterior. Portanto, não há necessidade de explorar

os resultados. Pelo resultado reportado,  $JB=9,140[0,010]$ , rejeitamos a hipótese nula, ou seja, não podemos afirmar que os resíduos possuem distribuição normal <sup>2</sup>.

### 6.2.2 O teste de Independência (BDS)

Já para o teste de independência, opção que não está disponível no diagnóstico dos resíduos, temos que primeiro salvar a série dos resíduos. Esse teste pode ser feito para qualquer série de tempo e o objetivo é saber se os dados podem ser considerados independentes. Nesse caso, há dois importantes parâmetros para escolher.

O primeiro é a distância entre um par de pontos, denominado de  $\varepsilon$  (epsilon). Para uma série ser verdadeiramente *iid*, considerando qualquer par de pontos, a probabilidade de que a distância entre esses pontos seja menor ou igual a  $\varepsilon$ , ou seja,  $c_1(\varepsilon)$ , deve ser constante. O segundo parâmetro é a dimensão do teste, ou seja, em quantos pares de pontos o mesmo é aplicado. Por exemplo, a partir de uma série de dados qualquer  $y_t$  com  $t=1,2,\dots,T$  podemos criar vários pares de mesma distância:

$$\{y_t, y_s\}, \{y_{t+1}, y_{s+1}\}, \dots, \{y_{t+m-1}, y_{s+m-1}\}$$

Note que foram criados  $m$  pares que possuem  $c_m(\varepsilon)$  probabilidades associadas. Assim, como temos  $m$  pares, então,

$$H_0 : c_m(\varepsilon) = c_1^m(\varepsilon) = \text{independência}$$

Ou seja, a probabilidade associada a todos os pares  $c_m(\varepsilon)$  é igual ao produto de todas as probabilidades individuais  $c_1^m(\varepsilon)$ . Se isso se verificar, então os dados são independentes. Vamos aplicar esse teste nos resíduos da equação 1:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)                      (0.063)

Com a eq1 aberta vá em **Proc/Make Residual Series...** e escolha um nome para a série dos resíduos da equação 1. Abra a série de resíduos e, a seguir, vá em **View/BDS Independence test**. A seguir, selecione como mostrado na figura 6.21 e clique em **OK**.

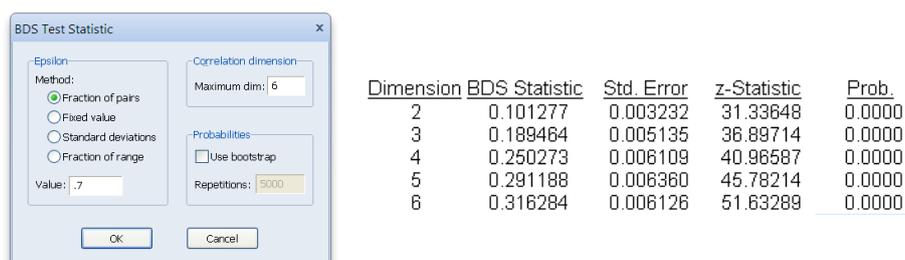


Figura 6.21: Teste BDS

Dentre as opções de escolha de  $\varepsilon$ , que irá determinar a distância para os pontos, recomenda-se **Fraction of pairs**, que tem menos influência da distribuição dos dados. As demais opções são variações para a definição do valor de  $\varepsilon$ . Ao especificar a dimensão máxima em 6, o teste é aplicado para cada valor de  $m=2,\dots,m=6$ . O terceiro conjunto de opção é para o cálculo das probabilidades do teste. Essa pode ser utilizada em séries de dados pequenos que não possuem uma distribuição muito bem definida. Nesse caso, a distribuição do teste BDS seria diferente da curva normal. A parte de resultados que interessa analisar é a mostrada na figura 6.21. Note que pelo p-valor, rejeitamos

<sup>2</sup>Lembre-se que a hipótese nula nesse caso é de distribuição normal.

a hipótese nula de independência, ou seja, os resíduos não são independentes. O teste pode ser apresentado da seguinte forma:

$$BDS_{m=2} = 0,101[0,000]$$

e assim sucessivamente até o valor de  $m = 6$ .

### 6.2.3 Correlograma – Q-stat

Para entender o teste de Ljung-Box, é necessário compreender o que o cálculo da autocorrelação representa para uma série de tempo. Conhecemos a correlação que existe entre duas variáveis. A ideia é a mesma para o caso da autocorrelação. Nesse caso, queremos justamente medir o grau de relação que existe entre a informação no tempo  $t$ , para uma variável  $y$  e a informação no tempo  $k$ , para a mesma variável. Isso é feito no *EViews*® a partir de:

$$t_k = \frac{\sum_{t=k+1}^T (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2}$$

Onde  $k$  é justamente o lag entre as duas informações, e  $\bar{y}$  é a média da série. Vejamos o exemplo dos resíduos da equação de regressão:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)                      (0.063)

Esses possuem média zero e, considerando  $k=1$  temos:

$$t_1 = \frac{\sum_{t=2}^{187} (y_t y_{t-1})}{\sum_{t=1}^{187} (y_t)^2} = \frac{y_2 y_1 + y_3 y_2 + \dots + y_{187} y_{186}}{y_1^2 + y_2^2 + \dots + y_{187}^2} = 0.825$$

O mesmo pode ser feito para a autocorrelação de ordem  $k$  que se desejar. Com isso, construímos a função de autocorrelação, que irá mostrar como essa se comporta ao longo do tempo. O passo seguinte seria testar se essa autocorrelação é estatisticamente significativa. Nesse caso, recorreremos ao teste de Ljung-Box, que tem a seguinte forma:

$$Q = T(T+2) \sum_{j=1}^k \frac{t_j^2}{(T-j)}$$

onde  $T$  é o número de observações,  $k$  é o lag máximo para o teste e  $t_j$  é a autocorrelação de ordem  $j$ . A hipótese nula para o teste é ausência de autocorrelação até o lag  $k$  e o mesmo possui distribuição qui-quadrado com os graus de liberdade dados pelo número de autocorrelações que se está medindo.

Para operacionalizar esse teste, após rodar uma regressão vá em **View/Residual Diagnostic/Correlogram - Q-statistics...** A seguir, escolha o número de lags e clique em **OK**, conforme figura 6.22. Um ponto importante para lembrar é que o teste pode ser sensível ao número de lags que é escolhido.

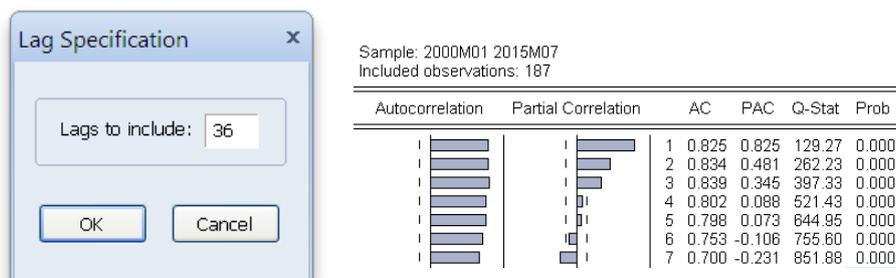


Figura 6.22: Teste de Ljung-Box

Para o nosso exemplo, o teste de Ljung-Box para 36 lags sinaliza que, pelo menos até o lag 7 não é possível aceitar a hipótese nula de ausência de autocorrelação nos resíduos. Ou seja, há evidência de autocorrelação. Podemos usar a fórmula acima para encontrar o valor do teste Q onde  $T=187$  e a autocorrelação de ordem 1 é  $t_1 = 0.825$ :

$$Q = 187(187 + 2) \sum_{j=1}^1 \frac{0.825^2}{(187 - 1)} = 129.270$$

A forma de apresentar os resultados é tal como:

$$Q(1) = 129,270[0,000]$$

### 6.2.4 Correlograma dos Resíduos ao Quadrado

O correlograma pode ser usado para identificar a presença ou não de heteroscedasticidade nos dados. Nesse caso, ao invés de calcular a função de autocorrelação considerando os resíduos, como no teste Q anterior, a mesma é feita com base nos resíduos ao quadrado. Sendo assim, primeiro é calculada a função de autocorrelação para cada lag e, a seguir, é aplicado o teste Q. Sua forma de avaliação é tal como anteriormente.

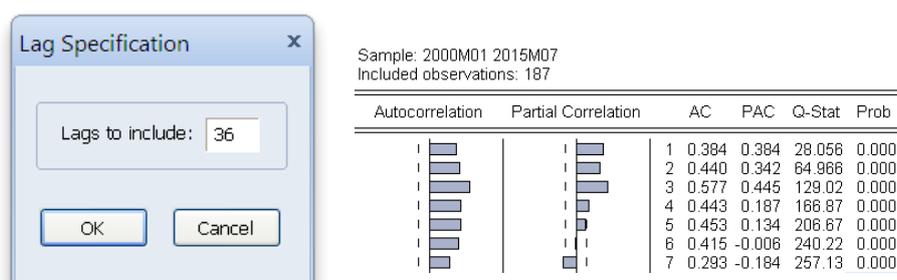


Figura 6.23: Correlograma dos Resíduos ao Quadrado

Fazendo esse teste para os resíduos ao quadrado da equação 1, figura 6.23, podemos ver que não é possível aceitar a hipótese nula de ausência de autocorrelação nos resíduos ao quadrado, sinalizando que os mesmos podem ter heteroscedasticidade.

### 6.2.5 Teste de Autocorrelação – LM

Esse teste é feito com base na hipótese nula de ausência de autocorrelação até o lag especificado. Após ter estimado a equação de regressão, como fizemos anteriormente no teste Q, vá em **View/Residual Diagnostics/Serial Correlation LM Test...** A seguir escolha o lag máximo que gostaria de testar, no nosso exemplo colocamos 2, e clique em **OK**. O que o *EViews*<sup>®</sup> faz é pegar a série de resíduos da primeira regressão e fazer uma nova regressão entre esses resíduos, seus valores passados e também a variável independente. A figura 6.24 reporta o resultado do teste.

**Programação 6.2.1** O teste LM tem distribuição qui-quadrado e com graus de liberdade de acordo com o número de lags avaliados sob a hipótese nula. Para encontrar o respectivo p-valor do teste no *EViews*<sup>®</sup>, clique em **Window** e depois selecione **Command...** A seguir, escreva o comando abaixo para encontrar o p-valor.

```
scalar testef
testef=@chisq(142.223, 2)
```

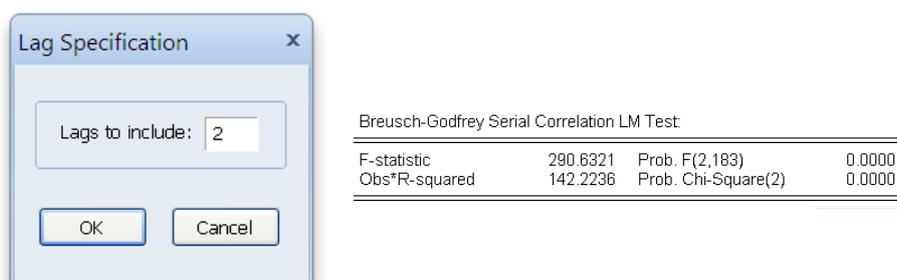


Figura 6.24: Teste LM para Autocorrelação

Dica: lembre de executar uma linha por vez, quando utilizar a janela de comandos.

Note que há duas estatísticas mostradas. O teste F não tem uma distribuição amostral finita conhecida sob a hipótese nula, mas, mesmo assim, é mostrado seu resultado. A seguir, tem a segunda estatística, que possui uma distribuição qui-quadrado, sendo mais recomendada para avaliação do teste de autocorrelação. Para entender como foi feito o teste basta olhar no final dos resultados a estimativa de uma equação para os resíduos, figura 6.25.

Como pode ser visto pelos resultados acima, para ambas as estatísticas rejeitam-se a hipótese nula de ausência de autocorrelação nos resíduos. Esse teste pode ser apresentado da seguinte forma:

Test Equation:  
 Dependent Variable: RESID  
 Method: Least Squares  
 Date: 11/11/15 Time: 10:30  
 Sample: 2000M01 2015M07  
 Included observations: 187  
 Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
YW_SA	-0.013795	0.031215	-0.441926	0.6591
C	1.313906	3.256267	0.403501	0.6871
RESID(-1)	0.425279	0.064559	6.587485	0.0000
RESID(-2)	0.491200	0.064692	7.592863	0.0000

R-squared	0.760554	Mean dependent var	1.58E-14
Adjusted R-squared	0.756629	S.D. dependent var	13.99658
S.E. of regression	6.904890	Akaike info criterion	6.723495
Sum squared resid	8724.984	Schwarz criterion	6.792610
Log likelihood	-624.6468	Hannan-Quinn criter.	6.751500
F-statistic	193.7547	Durbin-Watson stat	2.341606
Prob(F-statistic)	0.000000		

Figura 6.25: Teste LM - Regressão dos Resíduos

$$LM_{(2)} = 142,223[0,000]$$

**Programação 6.2.2** Para fazer o teste LM de autocorrelação, rodamos a regressão e salvamos a série dos resíduos. A seguir, fazemos uma regressão desses resíduos com a variável independente e o resíduo com 1 defasagem. Por fim, é usado um teste quiquadrado com 1 grau de liberdade no valor de  $T * R^2$  da regressão dos resíduos.

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
eq1.makesresid resid1
equation autocor.ls resid1 yw_sa c resid1(-1)
matrix(1,2) testelm
testelm(1,1)=autocor.@regobs*@r2
testelm(1,2)=@chisq(testelm(1,1),1)
```

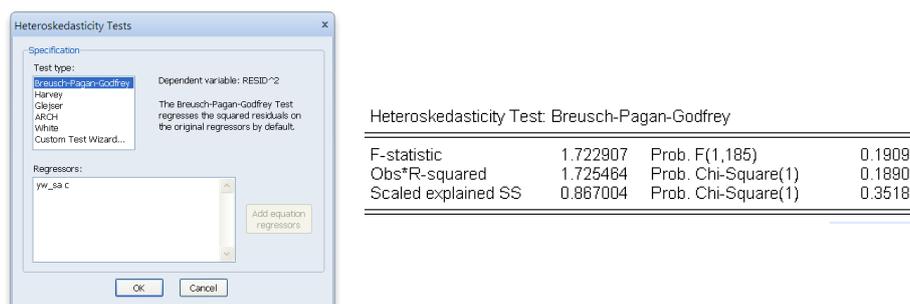


Figura 6.26: Teste de Heteroscedasticidade – Breusch-Pagan-Godfrey

Alternativamente, podemos fazer um loop para que sejam testados vários lags no teste LM e armazenar os resultados em uma tabela:

```
smp1 2000M1 2015M7
table(11,2) teste_lm
teste_lm(1,1)="valor do teste"
teste_lm(1,2)="p-valor"
equation eq1.ls qx_sa yw_sa c
eq1.makesresid resid1
for !i=1 to 10
equation eq10.ls resid1 yw_sa c resid1(-1 to -!i)
teste_lm(!i+1,1)=eq10.@regobs*@r2
teste_lm(!i+1,2)=@chisq(eq10.@regobs*@r2, !i)
next
```

### 6.2.6 Testes de Heteroscedasticidade

Na literatura da área há vários testes de heteroscedasticidade que podem ser aplicado a uma série de tempo. O *EViews*<sup>®</sup> apresenta algumas opções que discutiremos a seguir e que são aplicados à série dos resíduos da nossa equação.

#### Breusch-Pagan-Godfrey

Esse teste é feito a partir de uma regressão auxiliar dos resíduos ao quadrado relativamente a todas as variáveis independentes. Nesse caso, suponha que se tenha feito a seguinte regressão:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)      (0.063)

E que se pretende investigar a existência de heteroscedasticidade nos resíduos. Nesse caso, podemos fazer a regressão:

$$\varepsilon_t^2 = c(1) + c(2)yw_t$$

E testar a hipótese nula de ausência de heteroscedasticidade a partir de 3 diferentes estatísticas propostas pelo *EViews*<sup>®</sup>. Portanto, após feita a regressão, podemos ir em **View/Residual Diagnostics/Heteroskedastic Tests...** e selecionar o teste de **Breusch-Pagan-Godfrey**, tal como mostrado na figura 6.26. O primeiro teste mostrado é um teste F. Esse irá testar se todos os coeficientes da equação são estatisticamente iguais a zero. Note que seu valor é igual ao teste F mostrado ao fim da regressão. Pelo p-valor de 0,190 podemos dizer que não é possível rejeitar a hipótese nula de ausência de heteroscedasticidade a pelo menos 18% de significância.

O segundo teste é dado pela multiplicação do número de observações e o  $R^2$  da regressão. No nosso exemplo:

$$Obs * R^2 = 187 * 0,0092 = 1,725$$

O mesmo tem distribuição  $X^2$  e, pelo resultado, não é possível rejeitar a hipótese nula a, por exemplo, 18% de significância.

**Programação 6.2.3** Os testes de heteroscedasticidade são aplicados após ter rodado uma regressão. Dessa forma, só é solicitado após a equação de regressão ter sido especificada. Para aplicar um teste à *eq1* usamos a função abaixo.

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
eq1.hettest(type=BPG) @regs
```

Utilizando essa função, é aberta uma janela com o resultado. Alternativamente, podemos aplicar o teste BPG por uma equação de regressão. Note que construímos o teste a partir da regressão original, fazendo a série dos resíduos e aplicando uma nova regressão de nome “*bpg*”. A seguir, armazenamos a estimativa do teste no escalar “*bpgtest*” e, depois, o p-valor no escalar “*quiteste*”.

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
eq1.makesresid res1
equation bpg.ls res1^2 yw_sa c
scalar bpgtest=bpg.@r2*@regobs
scalar quiteste=@chisq(bpgtest,bpg.@ncoef-1)
```

Agora que sabemos como fazer o teste para uma única equação, podemos inserir o mesmo no loop de 100 regressões que usamos anteriormente. Note que, nesse caso, não usamos mais o termo “*scalar*” e, sim, criamos a matriz que irá armazenar os resultados dos vários testes de heteroscedasticidade “*heterosc*”. Nesse caso, na primeira coluna estão os vários resultados para o valor do teste e, na segunda coluna, o p-valor do mesmo, seguindo um teste qui-quadrado.

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
matrix(100,3) coef
coef(100,2)=eq1.@coefs(1)
coef(100,1)=eq1.@coefs(1)-1.975*eq1.@stderrs(1)
coef(100,3)=eq1.@coefs(1)+1.975*eq1.@stderrs(1)
matrix(100,2) heterosc
eq1.makesresid res1
equation bpg.ls res1^2 yw_sa c
heterosc(100,1)=bpg.@r2*@regobs
heterosc(100,2)=@chisq(bpg.@r2*@regobs,bpg.@ncoef-1)
for !i=1 to 99
smp1 2000M1+!i 2004M12+!i
equation eq2.ls qx_sa yw_sa c
```

Heteroskedasticity Test: Harvey			
F-statistic	0.205893	Prob. F(1,185)	0.6505
Obs*R-squared	0.207887	Prob. Chi-Square(1)	0.6484
Scaled explained SS	0.108039	Prob. Chi-Square(1)	0.7424

Figura 6.27: Teste de Heteroscedasticidade - Harvey

```
coef(!i,2)=eq2.@coefs(1)
coef(!i,1)=eq2.@coefs(1)-1.975*eq2.@stderrs(1)
coef(!i,3)=eq2.@coefs(1)+1.975*eq2.@stderrs(1)
eq2.makesresid res2
equation bpg.ls res2^2 yw_sa c
heterosc(!i,1)=bpg.@r2*@regobs
heterosc(!i,2)=@chisq(bpg.@r2*@regobs,bpg.@ncoef-1)
next
smpl @all
```

### Harvey

Para fazer esse teste primeiro rodamos a regressão normal, tal como mostrado abaixo:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)      (0.063)

A seguir, salvamos os resíduos e fazemos uma regressão no qual, diferentemente do teste de Breusch-Pagan-Godfrey, no teste de Harvey usamos o logaritmo, tal como mostrado a seguir:

$$\log \varepsilon_t^2 = c(1) + c(2)yw_t$$

Tal regressão irá produzir um  $R^2$  e, com isso, podemos construir a estatística do teste a partir de  $T * R^2$ . Outra estatística fornecida é a F-statistics, que irá testar se todos os coeficientes da regressão dos resíduos são estatisticamente iguais a zero, como apresentado na figura 6.27.

**Programação 6.2.4** Para fazer o teste de Harvey usamos o mesmo comando de antes, “*hettest*” mas, modificamos o tipo para “*harvey*”.

```
smpl 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
eq1.hettest(type=harvey) @regs
```

Como a diferença entre o método de Harvey e o de BPG está apenas no fato de que aquele usa  $\log \varepsilon_t^2$ , podemos usar a mesma sequência de comandos de antes e modificar apenas a estimativa de regressão do teste, como mostrado abaixo.

```
smpl 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
eq1.makesresid res1
equation harvey.ls @log(res1^2) yw_sa c
scalar harveytest=harvey.@r2*@regobs
scalar quiteste=@chisq(harveytest,harvey.@ncoef-1)
smpl @all
```

**Glejser**

Para fazer esse teste primeiro rodamos a regressão normal, tal como mostrado abaixo:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)      (0.063)

A seguir, salvamos os resíduos e fazemos uma regressão no qual, diferentemente do teste de Harvey, utilizamos os resíduos em módulo, tal como mostrado a seguir:

$$|\varepsilon_t| = c(1) + c(2)yw_t$$

A figura 6.28 mostra os resultados para esse teste. Na primeira linha está o teste F, que testa se todos os coeficientes da regressão dos resíduos são iguais a zero. A seguir está o teste que considera  $T * R^2$ , tal como feito anteriormente nos dois outros testes.

Heteroskedasticity Test: Glejser			
F-statistic	1.086321	Prob. F(1,185)	0.3031
Obs*R-squared	1.071672	Prob. Chi-Square(1)	0.3006
Scaled explained SS	0.676555	Prob. Chi-Square(1)	0.4108

Figura 6.28: Teste de Heteroscedasticidade - Glejser

**Programação 6.2.5** O teste de Glejser pode ser feito modificando no comando “*hettest*” o tipo para “*glejser*”.

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
eq1.hettest(type=glejser) @regs
```

Nesse teste, usamos o valor absoluto dos resíduos, e não os resíduos ao quadrado. E isso pode facilmente ser modificado no nosso comando usando “*@abs*”, tal como mostrado a seguir.

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
eq1.makesresid res1
equation glejser.ls @abs(res1) yw_sa c
scalar glejsertest=glejser.@r2*@regobs
scalar quiteste=@chisq(glejsertest,glejser.@ncoef-1)
smp1 @all
```

**ARCH**

Sem dúvida esse é um dos testes mais recomendados para identificar a presença de heteroscedasticidade nos resíduos de uma regressão. Partindo do nosso modelo de regressão:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)      (0.063)

Salvamos os resíduos e fazemos uma nova regressão no qual, diferentemente dos testes anteriores, também usamos os resíduos ao quadrado em defasagens como variável explicativa, tal como mostrado a seguir:

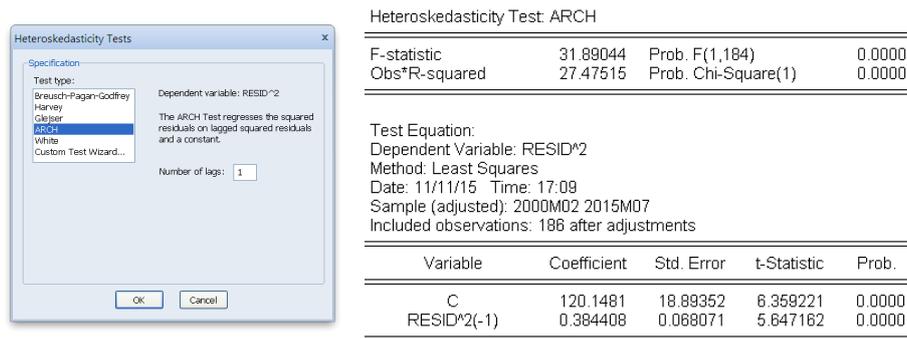


Figura 6.29: Teste de Heteroscedasticidade - ARCH

$$\varepsilon_t^2 = c(1) + c(2)\varepsilon_{t-1}^2$$

São mostrados dois testes, o  $F$ -statistic e o  $T * R^2$ . Em ambos podemos ver que não é possível aceitar a hipótese nula de homoscedasticidade. Um ponto interessante desse modelo é que ele difere do encontrado anteriormente pelos outros testes de heteroscedasticidade. Porém, devido ao poder do teste recomendamos que o leitor considere fortemente o teste ARCH como o mais importante.

**Programação 6.2.6** Para fazer o teste ARCH via programação, modificamos no comando “*hetttest*” o tipo de teste para “*arch*”. Porém, nesse caso, devemos especificar quantos *lags* serão utilizados para o teste. Como exemplo, usamos uma defasagem para os resíduos ao quadrado, como mostrado a seguir:

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
eq1.hetttest(type=arch, lags=1) @regs
```

Esse teste também pode ser construído no *EViews*® a partir dos comandos mostrados a seguir. Primeiro é feita a estimativa do modelo inicial e os resíduos são salvos. A seguir, como queremos apenas 1 lag, fazemos uma regressão dos resíduos ao quadrado tendo como variável independente a dependente com 1 defasagem. Por fim, aplicamos a estatística qui-quadrado.

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
eq1.makesresid res1
equation arch.ls res1^2 c res1^2(-1)
scalar archtest=arch.@r2*@regobs
scalar quiteste=@chisq(archtest,1)
smp1 @all
```

**Programação 6.2.7** O loop a seguir faz 100 regressões acrescentando, a cada passo, um novo mês na amostra. A seguir, faz os quatro testes de heteroscedasticidade apresentados, BPG, Glejser, Harvey e ARCH, e salva o p-valor em uma matriz com 100 linhas e quatro colunas. Esse exercício irá permitir avaliar, para qual sequência de dados, é possível aceitar ou rejeitar a hipótese nula de ausência de heteroscedasticidade.

```

smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
matrix(100,4) testeh
for !i=1 to 100
smp1 2000M1+!i 2004M12+!i
equation eq3.ls qx_sa yw_sa c
eq3.makesresid res3
equation bpg.ls res3^2 yw_sa c
testeh(!i,1)=@chisq(bpg.@r2*@regobs,bpg.@ncoef-1)
equation harvey.ls @log(res3^2) yw_sa c
testeh(!i,2)=@chisq(harvey.@r2*@regobs,harvey.@ncoef-1)
equation glejser.ls @abs(res3) yw_sa c
testeh(!i,3)=@chisq(glejser.@r2*@regobs,glejser.@ncoef-1)
equation harch.ls res3^2 c ar(1)
testeh(!i,4)=@chisq(harch.@r2*@regobs,1)
next
smp1 @all

```

### 6.3 Diagnóstico De Estabilidade

Avaliamos anteriormente diversas características dos resíduos que são importantes para sinalizar a eficácia do modelo formulado. Essas são investigações consideradas padrão, como a normalidade nos resíduos, a autocorrelação, a independência e a heteroscedasticidade. Porém, alguns desses resultados podem estar sendo influenciados pela presença de quebra estrutural e que pode se manifestar de várias formas, na média, nos parâmetros ou na tendência. Em síntese, os coeficientes podem não ser estáveis ao longo do tempo, e isso resulta em problemas de formulação. A seguir apresentamos alguns testes disponíveis no *EViews*<sup>®</sup> para lidar com a estabilidade nos parâmetros.

#### 6.3.1 Teste de Chow

Esse é um dos testes mais antigos e simples para identificar a existência ou não de quebra estrutural. Aqui a proposta é, a partir da especificação de uma data na amostra de dados, dividir o mesmo em 2 partes, rodar três regressões e comparar os resultados. A primeira regressão, denominada de modelo não-restrito, é feita para todo o conjunto de dados. A segunda, considerada modelo restrito, estima uma regressão entre a data inicial e a data especificada como de mudança estrutural. E, por fim, a terceira regressão é feita entre essa data especificada e o fim do período amostral. Assim, suponha que se tenha feito a seguinte regressão:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)      (0.063)

E que se quer verificar se ocorreu uma mudança estrutural em 2008M1, tanto na constante quanto no coeficiente de inclinação. Nesse caso, especificamos essa data e o *EViews*<sup>®</sup> irá rodar duas regressões da forma:

$$qx_t = c(1) + c(2)yw_t + \varepsilon_t \text{ (entre 2000M1 e 2007M12)}$$

$$qx_t = c(3) + c(4)yw_t + \varepsilon_t \text{ (entre 2008M1 e 2015M7)}$$

Onde a primeira usa os dados entre a data inicial e 2007M12 e, a segunda entre 2008M1 e a data final. Note que a data escolhida é utilizada na segunda regressão. A seguir, é feito um teste F para

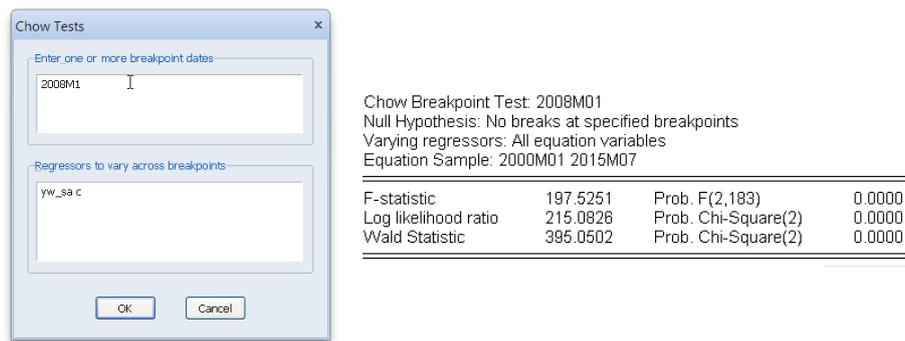


Figura 6.30: Teste de Chow

comparar os dois modelos com a estimativa para todo o período amostral. Esse teste utiliza a soma ao quadrado dos resíduos (SSR) de cada uma das três regressões.

$$F_{stat} = \frac{\frac{(SSR - (SSR_1 + SSR_2))}{k}}{\frac{(SSR_1 + SSR_2)}{(T - 2k)}}$$

Onde SSR é a soma ao quadrado dos resíduos da regressão que usa todo o conjunto de dados.  $SSR_1$  é a soma dos resíduos ao quadrado para a regressão do período pre mudança estrutural e  $SSR_2$  é a soma dos resíduos ao quadrado da segunda regressão, que usa o período pos quebra estrutural. Por fim, T é o número de dados e k o número de parâmetros da regressão. A hipótese nula é de que as duas subamostras são iguais, ou seja, não tem mudança estrutural. Um ponto importante nesse teste é que, caso não seja especificado, o *EViews*® irá testar a instabilidade em todos os parâmetros.

Vejam como isso pode ser feito no *EViews*®. Após estimar a equação de regressão para todo o período amostral, vá em **View/Stability Diagnostics/Chow Breakpoint Test...** Na janela que vai abrir, escreva a data de mudança estrutural que, para o nosso exemplo, é 2008M1 (janeiro de 2008). Abaixo estão os parâmetros que se quer testar a mudança estrutural, tanto para a constante quanto para a elasticidade-renda. Nesse nosso exemplo vamos testar uma mudança estrutural na constante e elasticidade-renda. Na janela de resultados, figura 6.30, primeiro é especificada a data de quebra estrutural, a seguir a hipótese nula e os parâmetros onde ocorreram a mudança. Por fim, a amostra de dados utilizada.

Para encontrar a estatística F, vamos especificar cada uma das 3 regressões. Para a que contempla todo o conjunto amostral, temos que  $SSR = 36438.213$ . Já na regressão que vai de 2000M1 até 2007M12, encontra-se  $SSR_1 = 2816.406$  e, para a da segunda parte de dados,  $SSR_2 = 8719.257$ . O conjunto amostral é  $T=187$  e o número de parâmetros é  $k=2$ . Portanto, a estatística F é:

$$F_{stat} = \frac{\frac{(36438.213 - (2816.406 + 8719.257))}{2}}{\frac{(2816.406 + 8719.257)}{(187 - 4)}} = 197.525$$

A estatística F, os testes de razão de verossimilhança e Wald trabalham sob a hipótese nula de não existência de mudança estrutural para toda a amostra. Esses dois últimos possuem distribuição  $X^2$  com  $mk_v$  graus de liberdade, onde  $m$  a quantidade de quebras e  $k_v$  os número de parâmetros testados na mudança estrutural. Para o nosso exemplo, figura 6.30, baseado nos três testes rejeitamos a hipótese nula, ao nível de confiança de 99% e, portanto, a data escolhida 2008M1, pode ser considerada como de quebra estrutural do modelo especificado.

**Programação 6.3.1** O teste de Chow tem distribuição qui-quadrado para o teste LR e de Wald, considerando como graus de liberdade  $q = mk_v$ . Por isso utilizamos o comando `@chisq(valor do teste, graus de liberdade)`. Para o teste F, a distribuição é a  $F$ . Logo, seu p-valor é dado por `1-@cdfist(F-stat, q, T-(q+k))`, onde  $F\text{-stat}$  é o valor do teste F,  $q = mk_v$  é o número de restrições sob a hipótese nula e  $T - (q + k)$  é o número de observações menos as restrições e os parâmetros da regressão original.

Assim, para encontrar o respectivo p-valor dos testes no *EViews*<sup>®</sup>, clique em **Window** e depois selecione **Command...** A seguir, escreva os comandos abaixo e execute um de cada vez.

```
scalar testeF=1-@cdfist(197.525,2,183)
scalar testeLR=@chisq(215.082,2)
scalar testeWald=@chisq(395.050,2)
```

**Programação 6.3.2** Para aplicar o teste de Chow, primeiro especificamos a regressão e, a seguir, o teste colocando a data que queremos testar para ver se ocorreu uma mudança estrutural.

```
smpl 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
eq1.chow 2008M1
```

Um aspecto ruim do teste de Chow é que devemos especificar a data da quebra estrutural, o que dificulta encontrar o ponto ideal da quebra. Uma forma de contornar isso é usar um loop. No exemplo abaixo, começamos com a primeira data de quebra em 2008M1 e, a seguir, são rodadas 50 regressões. Note que, a cada momento, após escolher as datas de quebra, são feitas duas regressões restritas, uma para a primeira parte e outra para a parte final. A seguir, feito os testes F e de Wald, calculados os p-valores e armazenados os resultados em uma tabela chamada `chowresult`.

```
smpl 2000M1 2015M7
equation eq5.ls qx_sa yw_sa c
scalar chowfteste
scalar chowfpvalor
scalar chowwaldteste
scalar chowwaldpvalor
table chowresult
chowresult(1,1)="Data de Mudança Estrutural"
chowresult(1,2)="Estatística F"
chowresult(1,3)="Prob"
chowresult(1,4)="Teste de Wald"
chowresult(1,5)="Prob"
for !i=1 to 50
smpl 2000M1 2007M11+!i
equation eq6.ls qx_sa yw_sa c
smpl 2007M12+!i 2015M7
equation eq7.ls qx_sa yw_sa c
chowfteste=((eq5.@ssr-eq6.@ssr-eq7.@ssr)/(eq5.@ncoef))/((eq6.@ssr
+eq7.@ssr)/(eq5.@regobs-2*eq5.@ncoef))
chowfpvalor=1-@cdfist(chowfteste,2,eq5.@regobs-eq5.@ncoef-eq6.@ncoef)
```

```

chowwaldteste=@transpose(@identity(eq5.@ncoef)*(eq6.@coefs- eq7.@coefs
))*@inverse(eq6.@cov/(eq6.@se)^2+eq7.@cov/(eq7.@se)^2)*@identity
(eq5.@ncoef)*(eq6.@coefs-eq7.@coefs))*((eq5.@regobs-2*eq5.@ncoef)
/(eq6.@ssr+eq7.@ssr))
chowwaldpvalor=@chisq(chowwaldteste,eq5.@ncoef)
chowresult(!i+1,1)=@otods(1)
chowresult(!i+1,2)= chowfteste
chowresult(!i+1,3)=chowfpvalor
chowresult(!i+1,4)=chowwaldteste
chowresult(!i+1,5)=chowwaldpvalor
next
smp1 @all

```

O teste de Chow também pode ser especificado de forma a identificar a presença de quebra estrutural apenas em um dos parâmetros ou em parte. Com a equação aberta, vá em **View/Stability Diagnostics/Chow Breakpoint Test...** e a seguir, para testar mudanças na constante especifica-se a data e depois deixa escrito apenas “c” na parte de baixo da janela.

**Programação 6.3.3** Para especificar em quais parâmetros queremos aplicar o teste de Chow, adicionamos ao comando **chow**, depois declaração da data de mudança estrutural, a instrução **@** seguido do nome das variáveis.

```

eq1.chow 2008M1 @ yw_sa
O EViews® também permite testarmos mais de uma quebra estrutural com o teste de Chow.

eq1.chow 2008M1 2010M7 @ yw_sa c

```

### 6.3.2 Teste de Quandt-Andrews

O teste de Chow é muito simples e de difícil solução prática, uma vez que devemos testar várias datas e formatos para ter certeza de onde veio a instabilidade e em que parâmetro. Uma evolução natural seria permitir que fossem feitos diversos testes ao mesmo tempo em uma sequência e, ao final, escolher a data apropriada. Essa é justamente a proposta do teste de Quandt-Andrews.

Esse teste pode ser aplicado para identificar mais de uma data de quebra estrutural, usando como base a ideia do teste de Chow. Nesse caso, o mesmo é aplicado a cada informação entre a data de início e final do conjunto de dados. A hipótese nula é de ausência de quebra estrutural, e pode ser feito para toda a equação, considerando todos os parâmetros ao mesmo tempo ou, então, para o caso de uma equação linear, para cada um dos parâmetros de forma isolada.

Uma diferença importante entre esse teste e do de Chow é que aqui especificamos o “*trimming*” ou seja, o percentual de dados que são isolados do teste e não são utilizados. No *eviews*, como *default*, é fornecido o valor 15. Se escolher esse então, na verdade, estamos retirando 15% dos dados, 7,5% do início da amostra e 7,5% do final, e o teste é feito com o restante dos dados entre dois pontos  $\tau_1$  e  $\tau_2$ .

Como pode ser visto na figura 6.31, também escolhemos qual parâmetro será utilizado para o teste. No exemplo abaixo, aplicamos o mesmo para os dois parâmetros da nossa regressão simples, o da elasticidade-renda e a constante. Por fim, como opcional, especificamos um nome para a série dos testes, tanto para o LR quanto para o teste de Wald. O que acontece com esse teste é que, como o mesmo é aplicado a cada uma das datas entre  $\tau_1$  e  $\tau_2$ , então, iremos produzir uma estatística LR e Wald para cada uma dessas datas. Com isso, estaremos formando uma série com o resultado do teste. A estatística LR, com distribuição F, compara um modelo restrito com um não restrito e, após

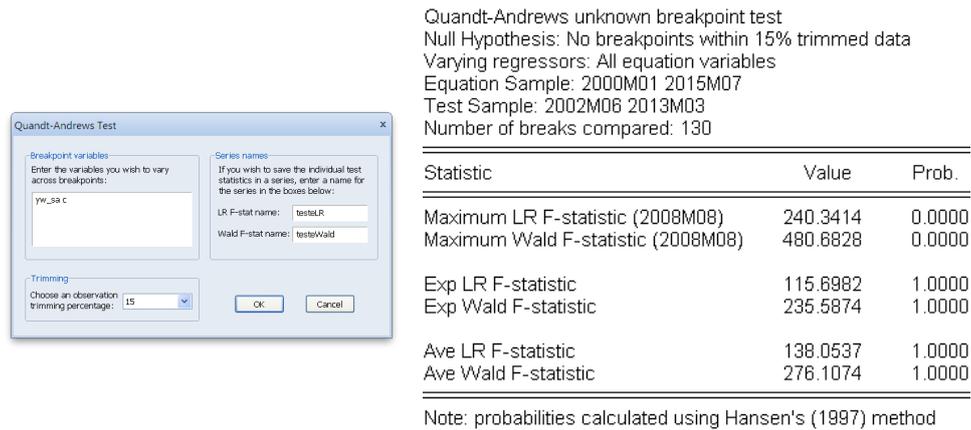


Figura 6.31: Teste de Quebra Estrutural de Quandt-Andrews

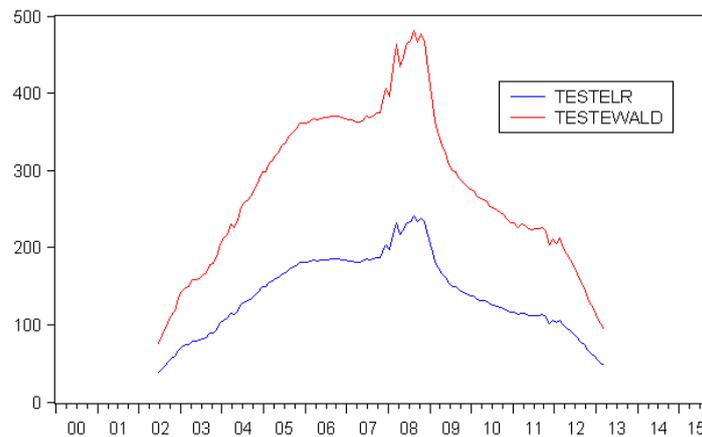


Figura 6.32: Resultados dos Testes LR e Wald

ter todos os resultados, a data da quebra é escolhida a partir do valor máximo do teste, como em:

$$MaxF = \max_{\tau_1 \leq \tau \leq \tau_2} (F(\tau))$$

Pelos resultados do teste de Quandt-Andrews, aplicado ao modelo de regressão simples:

$$qx_t = 18.389 + 0.647yw_t + \varepsilon_t$$

(6.616)            (0.063)

podemos ver que rejeitamos a hipótese nula de ausência de quebra estrutural. Nesse caso, há uma mudança estrutural e essa é especificamente em agosto de 2008.

As duas estatísticas são mostradas no gráfico 6.32. Note que ambas revelam que o valor máximo para o teste, tanto o LR quanto Wald, é em agosto de 2008. Com a escolha de um trimming de 15% foram eliminados 29 dados do início e outros 28 do final da série, restando 130 datas para serem testadas.

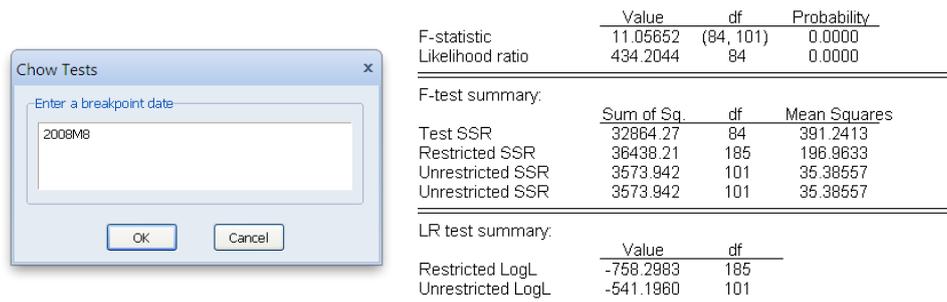


Figura 6.33: Teste de Previsão de Chow – Quebra Estrutural

**Programação 6.3.4** Para fazer o teste de Quandt-Andrews, usamos a função *ubreak* e que pode ser aplicada a qualquer equação. No exemplo abaixo, aplicamos a mesma à regressão simples. Uma opção interessante é pedir a série dos testes de wald e de LR, usados para selecionar o ponto de quebra estrutural. O número após a função se refere ao tamanho da parte da amostra que é retirada do teste. No caso de 15, estamos escolhendo 15% dos dados.

```
smp1 2000M1 2015M7
equation eq1.ls qx_sa yw_sa c
eq1.ubreak(wfname=testewald,lfname=testelr) 15
```

### 6.3.3 Teste de Previsão de Chow

Aqui é feita duas regressões, uma para todo o conjunto de dados e outra apenas com os dados que vão até a data anterior da quebra estrutural. São reportadas duas estatísticas, a primeira é o teste F, que é dado por:

$$F_{stat} = \frac{\frac{(SSR - SSR_1)}{T_2}}{\frac{SSR_1}{(T_1 - k)}}$$

Onde  $SSR$  é a soma dos resíduos ao quadrado da regressão completa,  $SSR_1$  é a soma dos resíduos ao quadrado da regressão com dados que vão até  $T_1$ , que é o número de dados utilizados nesse período.  $T_1$  é o número de dados da segunda parte da regressão e  $k$  é o número de parâmetros da regressão completa. A hipótese nula é de ausência de quebra estrutural.

$$LR = -2(l_{restrito} - l_{nao-restrito})$$

Para fazer esse teste, vá em **View/Stability Diagnostics/Chow Forecast Test...** Os resultados do exemplo de um teste de quebra estrutural para a data 2008M8 são mostrados na figura 6.33.

Note que, pelos resultados do teste F, rejeitamos a hipótese nula de ausência de quebra estrutural, confirmando o resultado encontrado pelo teste de Quandt-Andrews anterior. A estatística F pode ser calculada a partir de:

$$F_{stat} = \frac{(36438,213 - 3573,942)}{\frac{84}{\frac{3573,942}{(103 - 2)}}} = 11,056$$

E, para a estatística LR usamos:

$$LR = -2(-758,298 - (-541,196)) = 434,204$$

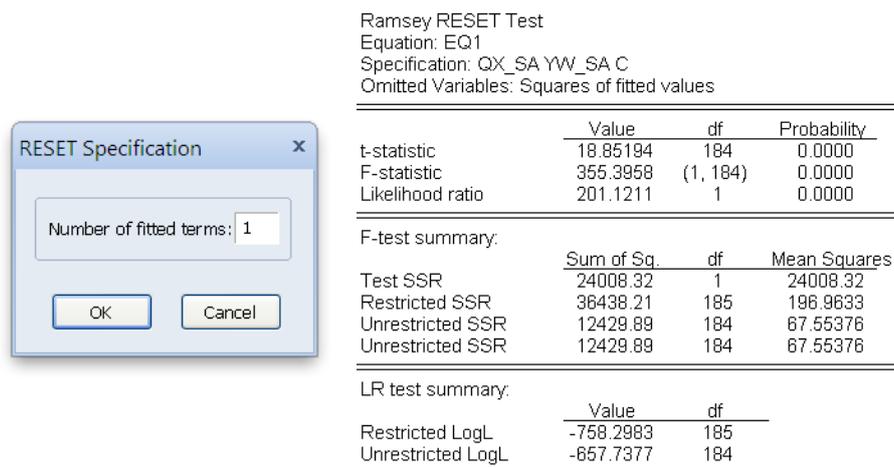


Figura 6.34: Teste de Ramsey – Quebra Estrutural

### 6.3.4 Teste de Ramsey

Aprendemos anteriormente a testar se os resíduos possuem distribuição normal, se há presença de heteroscedasticidade ou então autocorrelação. Porém, há outros problemas que podem aparecer na nossa regressão, como por exemplo, de variável omitida, de má especificação da forma funcional, ou a correlação entre a variável independente e os resíduos. Esses aspectos irão resultar que o estimador de mínimos quadrados é viesado e não consistente e, dessa forma, o vetor dos resíduos não terá média zero, ver Ramsey(1969).

Assim, o teste é feito considerando como hipótese nula que os resíduos da equação são distribuídos normalmente, com média zero e variância constante, contra a hipótese alternativa de que a média dos resíduos não é zero. Na figura 6.34, especificamos que o número de termos a serem considerados no teste é 1, ou seja, usamos como variável adicional o quadrado da variável dependente:

$$qx_t = \alpha_1 + \beta_1 y w_t + qx_t^2 + \varepsilon_t$$

Assim, temos que o teste considera as seguintes hipóteses:

$$H_0 : qx_t = \alpha_1 + \beta_1 y w_t + \varepsilon_t$$

$$H_a : qx_t = \alpha_1 + \beta_1 y w_t + qx_t^2 + \varepsilon_t$$

E pode ser visto como um teste de variável omitida. A figura 6.34 apresenta os resultados para esse teste. Note que são fornecidas três estatísticas para o teste e todas apontam para a não aceitação da hipótese nula. Portanto, a nossa equação possui problema de especificação.

### 6.3.5 Estimativas Recursivas

As Estimativas Recursivas podem ser acessadas em **View/Stability Diagnostics/Recursive Estimates....** Essa seção é aplicada de seis formas diferentes, cada uma fornecendo uma informação específica.

#### Recursive Residual

Para esse teste são feitas várias regressões, a partir do método MQO, mudando apenas o período amostral. A primeira regressão é feita com uma quantidade de informações igual ao número de coeficientes. Considerando a nossa equação básica, com a quantidade como função da renda, temos 2 coeficientes, a constante e a elasticidade renda, tal como:

$$qx_t = \alpha_1 + \beta_1 y w_t + \varepsilon_t$$

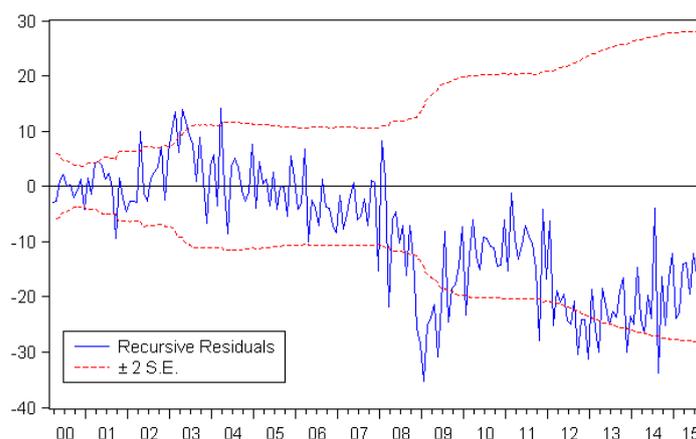


Figura 6.35: Resultado dos Resíduos Recursivos

Note que temos  $k=2$ , onde  $k$  é o número de coeficientes. Assim, a primeira regressão é feita considerando os 2 primeiros dados da amostra. O resultado para esses coeficientes é utilizado para prever o valor da variável dependente no período seguinte. Depois é calculada a diferença entre o valor previsto e o verdadeiro valor e dividida pela variância da previsão. Esse resultado é denominado de “*recursive residual*” e seu valor é armazenado em um vetor.

A seguir, acrescentamos o terceiro dado da amostra e fazemos novamente a regressão, encontrando os coeficientes, fazendo a previsão para um passo à frente, dividindo pela variância da previsão e encontrando o resíduo. Esse resíduo é armazenado no vetor de resíduos. Esse procedimento continua até que se utilize todo o conjunto amostral, ou seja, as  $T$  observações. Sendo assim, fazemos um total de  $T - k + 1$  regressões e obtemos um total de  $T - k + 1$  estimativas para os resíduos. Vejamos como fica esse processo a partir dos dados da equação acima. A primeira regressão, com apenas os dois primeiros dados, ou seja, usando 2000M1 a 2000M2, produz o seguinte resultado:

$$qx_t = -640,656 + 9,106yw_t + \varepsilon_t$$

Se usarmos esses coeficientes para prever o valor de  $qx$  para 2000M3, encontramos:

$$qx_{2000M3} = -640,656 + 9,106 * (77,169) + \varepsilon_t = 62,082$$

O verdadeiro valor é  $qx_{2000M3} = 55,747$ , gerando um resíduo de valor 6,335. A seguir, temos que calcular a variância da previsão e depois:

$$recursive - residual_{2000M3} = \frac{6,335}{-2,169} = -2.920$$

Esse procedimento é repetido até o fim da amostra, gerando uma sequência de valores para os resíduos recursivos. O *EViews*<sup>®</sup> retorna o gráfico dessa sequência com o respectivo intervalo de confiança, conforme apresentado na figura 6.35. Valores que estão fora do intervalo sinalizam instabilidade nos parâmetros da equação. Note que a data entre 2008M8 e 2009M6 está fora do intervalo de confiança, sinalizando possível quebra estrutural nesse período.

### Teste CUSUM

Os resíduos recursivos obtidos do teste anterior “*recursive residual*”, são usados para produzir o teste CUSUM, ou seja, esse teste nada mais é que a soma cumulativa dos resíduos encontrados no teste anterior. Nesse caso, esse é dividido pelo seu respectivo desvio-padrão e depois é feita

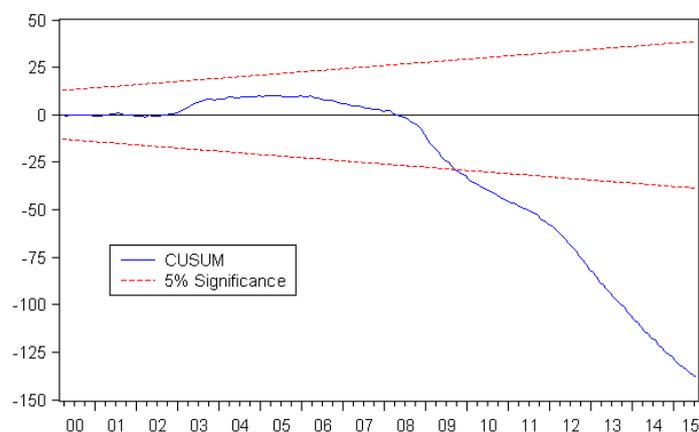


Figura 6.36: Teste CUSUM

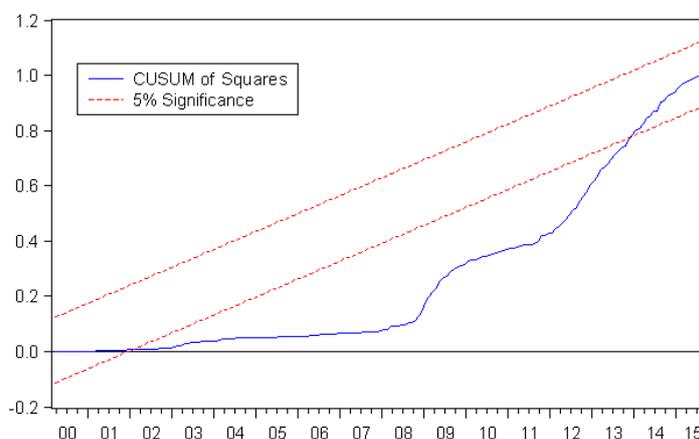


Figura 6.37: Teste do CUSUM ao Quadrado

a soma desses resíduos. Se o valor ficar fora do intervalo crítico de 5% do teste, então, há uma sinalização de instabilidade nos coeficientes da equação. Ao aplicar esse teste à nossa equação acima, encontramos o seguinte resultado, mostrado na figura 6.36. Note que, em 2009M9, o teste ultrapassa o valor crítico a 5%, sinalizando uma instabilidade no modelo.

#### Teste do CUSUM ao Quadrado

Da forma como o teste é calculado, seria como obter a variância dos resíduos recursivos. Na verdade deriva do teste CUSUM e do teste recursivo, só que aqui elevamos os resíduos ao quadrado e depois somamos os mesmos. A expectativa do resultado desse teste, sob a hipótese de estabilidade dos parâmetros, é que inicie em zero e termine em 1 e que seu resultado fique dentro do intervalo de 5% de significância. Aplicando o teste ao nosso modelo, encontramos que há uma instabilidade entre 2001M12 e 2013M12, figura 6.37.

#### Teste de Previsão One-Step

Esse teste também utiliza os resultados dos resíduos recursivos, complementando a análise do mesmo com o desvio-padrão da amostra total. Seu resultado, para o nosso modelo, é mostrado na figura 6.38. Note que há duas informações. Primeiro, a série do resíduo recursivo é mostrada

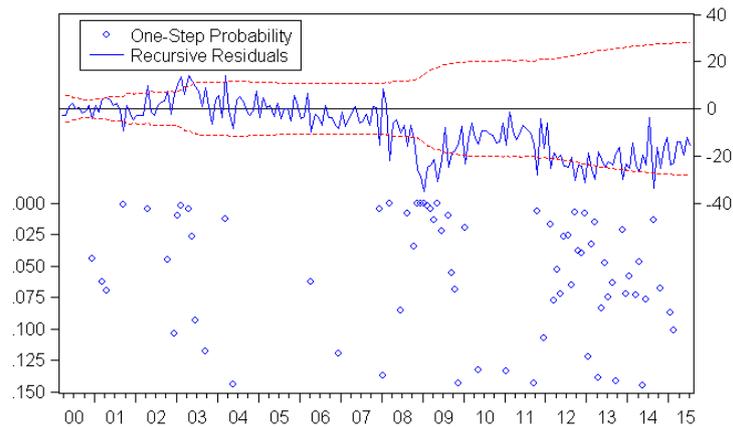


Figura 6.38: Teste de Previsão One-Step

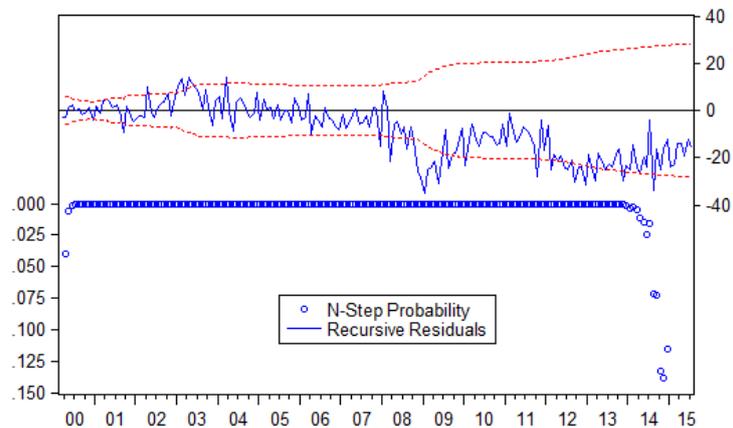


Figura 6.39: Teste de Previsão N-Step

novamente, com seus valores no eixo direito. Além desse, também é mostrado o p-valor do teste, ou seja, a probabilidade associada à rejeição da hipótese de estabilidade dos parâmetros. Nos pontos com valores menores, significa a não aceitação da hipótese nula de estabilidade. Como pode ser visto, há vários desses pontos, em especial entre 2007 e 2015

### Teste de Previsão N-step

Esse teste também usa os resultados dos resíduos recursivos e é equivalente ao teste de Chow, mas, sem a necessidade de informar cada uma das datas que se queira testar. Ou seja, o teste é feito para várias datas, e retorna o valor dentro de um intervalo de confiança e com o p-valor. Os resultados para o nosso modelo são apresentados na figura 6.39.

### Coeficientes Recursivos

Esse teste pode ser utilizado para identificar como é o comportamento de cada um dos coeficientes ao longo do tempo. Para tanto, o método segue a estimativa feita anteriormente, quando foram obtidos os resíduos recursivos para encontrar o valor a cada momento do tempo, adicionando, a cada passo, uma nova observação.

Para o nosso modelo, o teste foi feito para os dois coeficientes e seus resultados sinalizam

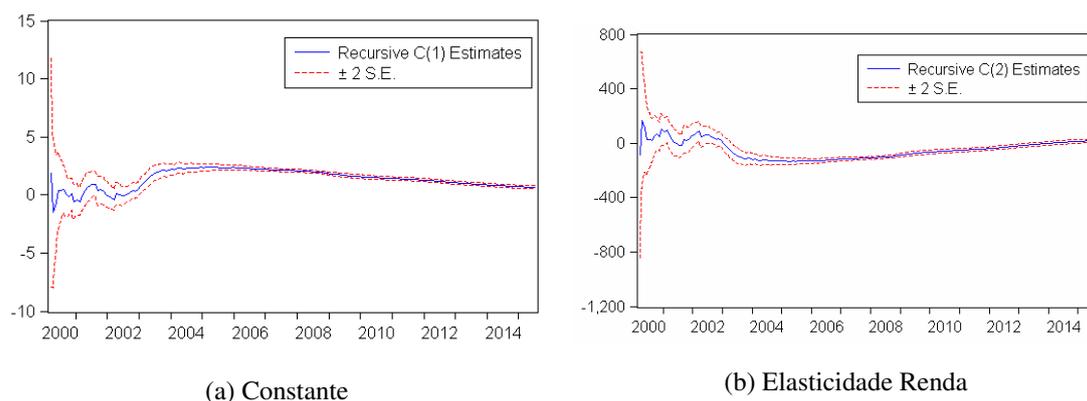


Figura 6.40: Coeficientes Recursivos

instabilidade presente nas informações adicionadas até meados de 2004, conforme mostrados na figura 6.40.

### 6.3.6 Leverage Plots

Esse método gráfico fornece a informação sobre a presença de possíveis outliers a partir de várias regressões e a comparação dos resíduos. Clique em **View/Stability Diagnostics/Leverage Plots...** que serão mostradas as opções em uma janela. Na primeira parte, devem ser definidas as variáveis que serão utilizadas na análise. No nosso exemplo, sabemos que a variável `qx_sa` é dependente e, especificamos `yw_sa` e `c` (constante) como regressores. A seguir, selecione a opção para adicionar uma linha de tendência e para a informação parcial, que é a mais ilustrativa. Por fim, especifique um nome para que, ao salvar as séries de resíduos resultantes, se tenha um nome como complemento. Nesse exemplo, serão geradas quatro séries adicionais no *workfile*. Duas séries são geradas usando a variável dependente. Na primeira, temos a série de nome `qx_sa_p_yw_sa_lv` que representa os resíduos da regressão:

$$qx_t = c + \varepsilon_t$$

A segunda é dada pela série de nome `qx_sa_p_c_lv`, e representa os resíduos da regressão:

$$qx_t = \beta yw_t + \varepsilon_t$$

A seguir, temos duas outras séries que são geradas a partir do uso das variáveis independentes. Nesse caso, como temos apenas uma independente, teremos duas séries derivadas. A primeira é a



Figura 6.41: Opções Leverage Plots

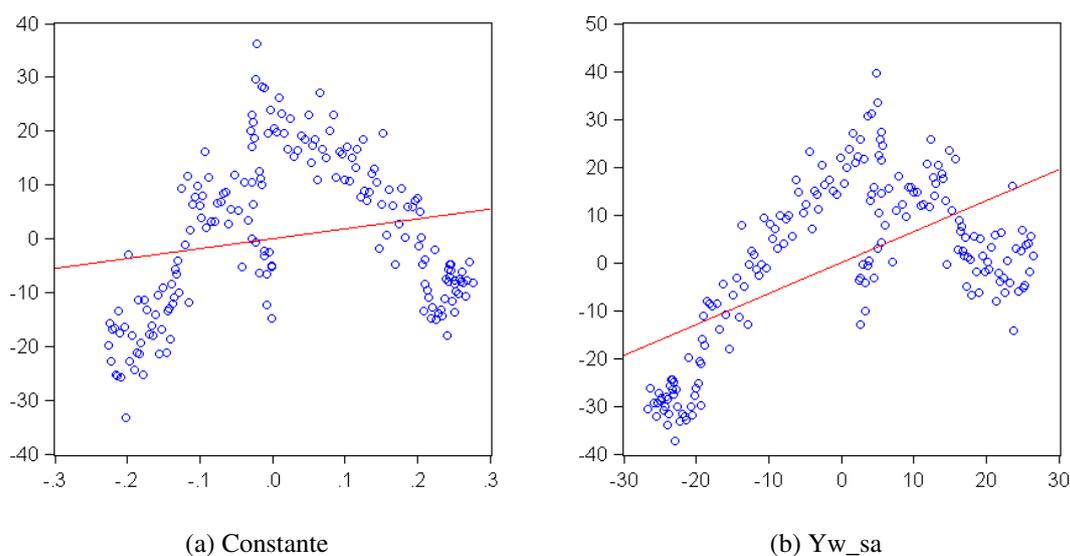


Figura 6.42: Leverage Plots

série de nome `c_lv`, que representa a série de resíduos da regressão:

$$yw_t = c + \varepsilon_t$$

De acordo com o nosso exemplo, serão gerados dois gráficos de dispersão. O primeiro, mostrado na figura 6.42a, representa a relação entre a série `c_lv` e `qx_sa_p_c_lv`. A seguir, o segundo gráfico, figura 6.42b, representa a relação entre a série `yw_sa_lv` e `qx_sa_p_yw_sa_lv`.

### 6.3.7 Estatísticas de Influência

Uma forma de identificar a presença de *outliers* é através das estatísticas de influência. Uma informação é considerada como um *outlier* se ela produz um impacto significativo na regressão. Assim, partindo dessa definição, essa estatística é aplicada ao conjunto de dados para identificar o quanto que uma única observação pode modificar o modelo de regressão. São seis diferentes estatísticas que podem ser utilizadas. Vá em **View/Stability Diagnostics/Influence Statistics...** A seguir, selecione três estatísticas, como mostrado na figura 6.43.

Os resultados serão salvos nas respectivas séries IS1, IS2 e IS3 e mostrados em um conjunto de gráficos. Note na figura 6.44 que, para cada uma das estatísticas, há um intervalo de confiança. Os testes RStudent e COVRATIO apontam 2008M1 como um *outlier* e, também, RStudent junto do teste DFFITS sinalizam para a existência de um *outlier* em 2014M8.

## 6.4 Previsão - Forecast

A partir do momento que temos a estimativa dos parâmetros do modelo, podemos fazer previsões para o futuro e, mais do que isso, encontrar um intervalo de confiança para essa previsão. De um modo geral, nosso modelo simples pode ser representado pela equação linear:

$$qx_t = \alpha_1 + \beta_1 yw_t + \varepsilon_t$$

onde os valores de `yw` são conhecidos. Lembre-se que os dados são ajustados sazonalmente. Para cada valor de `ywt+n` utilizado, podemos encontrar um respectivo valor de `qxt+n`, o que nos permitirá obter, futuramente, o erro de previsão. Sendo assim, podemos modificar essa equação linear como forma de obter o erro de previsão:

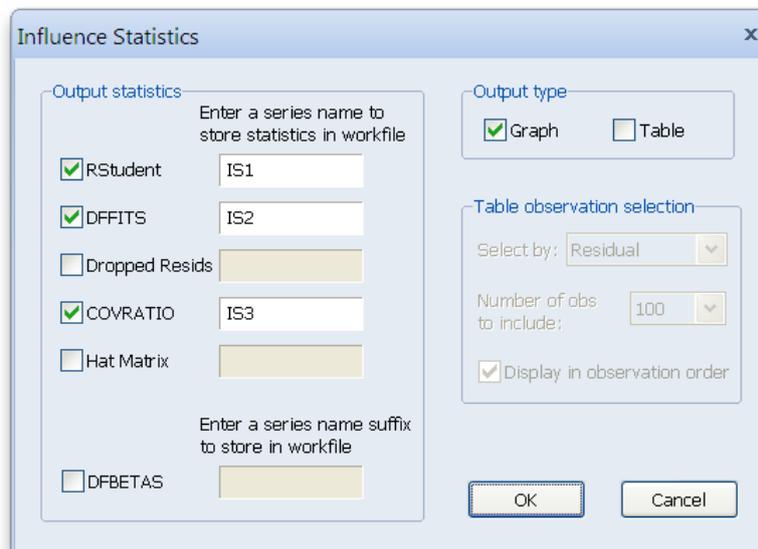


Figura 6.43: Opções Estatísticas de Influência

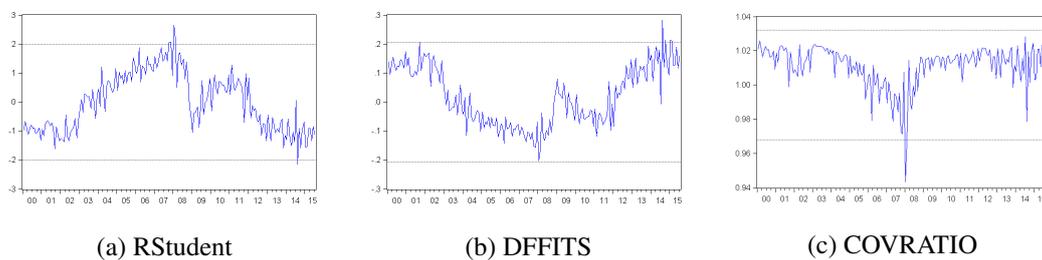


Figura 6.44: Teste de Quebra Estrutural – Estatística de Influência

$$(q\hat{x}_t - qx_t) = (\hat{\alpha} + \hat{\beta}yw_t + \varepsilon_t) - qx_t$$

Além disso, usando o fato <sup>3</sup> de que  $qx_t = \alpha_1 + \beta_1yw_t + \varepsilon_t$  podemos substituir o mesmo na equação acima e obter:

$$(q\hat{x}_t - qx_t) = (\hat{\alpha} + \hat{\beta}yw_t + \varepsilon_t) - (\alpha + \betayw_t)$$

$$(q\hat{x}_t - qx_t) = (\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)yw_t + \varepsilon_t$$

Esse valor que encontramos corresponde ao valor médio de erro de previsão. Porém, para fazer inferência estatística sobre a previsão, devemos conhecer outros resultados, em especial a variância do erro de previsão. A partir da equação acima, aplicamos o operador  $V()$ , que corresponde à variância, encontramos a variância do erro de previsão:

$$V(q\hat{x}_t - qx_t) = V[(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)yw_t + \varepsilon_t]$$

$$V(q\hat{x}_t - qx_t) = V(\hat{\alpha} - \alpha) + yw_t^2 V(\hat{\beta} - \beta) + 2yw_t \text{cov}(\hat{\alpha} - \alpha, \hat{\beta} - \beta) + V(\varepsilon_t)$$

Usando o fato de que:

$$V(\hat{\alpha} - \alpha) = \hat{\sigma}^2 \left( \frac{1}{T} + \frac{y\bar{w}^2}{\sum_{i=1}^T yw^2 - Ty\bar{w}^2} \right)$$

$$V(\hat{\beta} - \beta) = \frac{\hat{\sigma}^2}{\sum_{i=1}^T yw^2 - Ty\bar{w}^2}$$

$$\text{cov}(\hat{\alpha} - \alpha, \hat{\beta} - \beta) = \hat{\sigma}^2 \frac{y\bar{w}}{\sum_{i=1}^T yw^2 - Ty\bar{w}^2}$$

Então, temos que a variância do erro de previsão pode ser calculada a partir de:

$$V(q\hat{x}_t - qx_t) = \hat{\sigma}^2 \left( \frac{1}{T} + \frac{y\bar{w}^2}{\sum_{i=1}^T yw^2 - Ty\bar{w}^2} \right) + yw_t^2 \frac{\hat{\sigma}^2}{\sum_{i=1}^T yw^2 - Ty\bar{w}^2} + 2yw_t \frac{\hat{\sigma}^2 y\bar{w}}{\sum_{i=1}^T yw^2 - Ty\bar{w}^2} + \hat{\sigma}^2$$

onde  $\hat{\sigma}^2$  é a variância da regressão. Colocando  $\hat{\sigma}^2$  em evidência, chegamos a uma formulação mais reduzida da variância do erro de previsão:

$$V(q\hat{x}_t - qx_t) = \hat{\sigma}^2 \left[ 1 + \frac{1}{T} + \frac{(yw - y\bar{w}^2)}{\sum_{i=1}^T yw^2 - Ty\bar{w}^2} \right]$$

Essa equação mostra para cada informação de  $yw_{t+n}$  prevista, a variância dessa previsão. Sendo assim, para qualquer valor de  $yw_{t+n}$  que utilizarmos, o correspondente valor de  $qx_{t+n}$  irá se encontrar exatamente na reta de regressão que estimamos. Seria como se estivéssemos prolongando a nossa reta de regressão para poder fazer uma previsão dos valores futuros<sup>4</sup>.

Mas essa é uma estimativa por ponto e, uma vez que estamos diante de incerteza, o que acaba por incorporar a presença de probabilidade de ocorrência de um evento e devemos ter cuidado ao trabalhar com essa informação. Sendo assim, recorreremos à estimativa de um intervalo para a nossa previsão. E, como vimos anteriormente, para construir esse intervalo, precisamos do cálculo da variância.

<sup>3</sup>O leitor deve prestar bastante atenção à diferença que existe entre a equação conhecida  $y = \alpha + \beta x$  e a estimada  $\hat{y} = \hat{\alpha} + \hat{\beta}x + \varepsilon$ .

<sup>4</sup>Considere que a escolha de um modelo econométrico para fazer previsão resulta em um casamento com a relação entre as variáveis independentes e a dependente. Com a vantagem de ser menos burocrático trocá-lo.

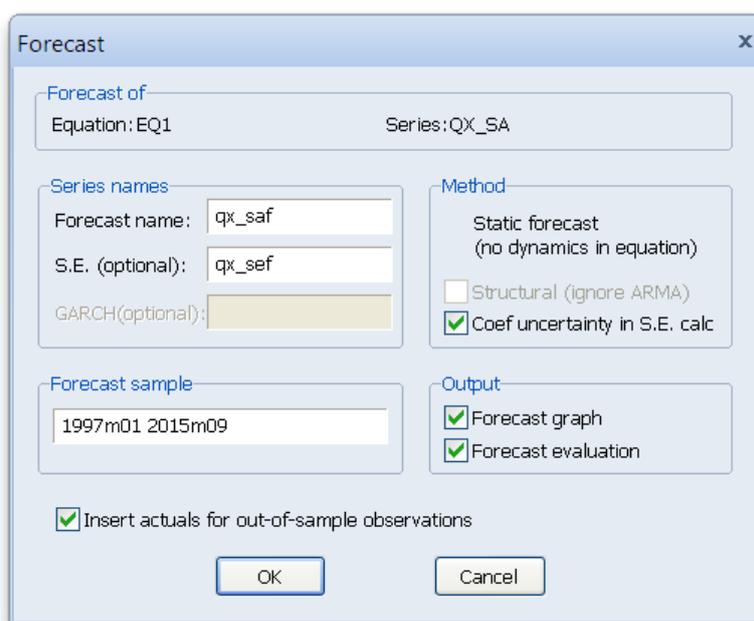


Figura 6.45: Fazendo a Previsão

A previsão da variável dependente da nossa equação no *EViews*<sup>®</sup> pode ser feita a partir de três diferentes formas. Na primeira, com os resultados da equação aberto, clique em **Forecast**, que irá aparecer a janela conforme figura 6.45. O que temos que fazer é especificar o nome da série prevista na parte **Forecast name** e, caso queira, o desvio-padrão da previsão, especificado como **S.E.**. Esse irá construir um intervalo de previsão para os dados. O mais interessante aqui é construir o intervalo com 2 desvios padrão em relação à média. Nesse caso, a informação de S.E deve ser multiplicada por 2 e depois acrescida e subtraída da série prevista para se ter o intervalo. No *box* de programação a seguir construímos esse intervalo mínimo e máximo.

Os resultados podem ser mostrados tanto em um gráfico quanto em uma tabela com estatísticas de informação que são úteis para comparar diferentes modelos. A segunda forma de fazer essa previsão é clicar em **Proc/Forecast**, que irá abrir a mesma janela de previsão

Na primeira linha é mostrado que a previsão é feita para a equação de nome *eq1* e a série que está sendo prevista é *qx\_sa*. Uma vez que temos uma equação de regressão simples, apenas algumas opções em **Forecast** são abertas. No caso do **Method**, como não temos variável dependente defasada, fazemos uma previsão que não é dinâmica. Podemos mudar o intervalo da previsão no campo denominado de **Forecast Sample**. Além disso, podemos mudar, na parte de **Output**, o que queremos que seja mostrado, os gráficos e os resultados de avaliação dessa previsão. Por economia de espaço foi pedido apenas os resultados da previsão e não mostramos o gráfico na figura 6.45 .

Vamos discutir as estatísticas mostradas nesse cálculo e que se referem aos erros de previsão. Em todas as três primeiras estatísticas quando for comparar diferentes modelos podemos escolher aquele que tem o menor valor para essas estatísticas.

A primeira delas é o RMSE, e que é dado a partir de:

$$RMSE = \sqrt{\frac{(\sum_{t=T+1}^{T+h} (\hat{q}x_t - qx_t)^2)}{h}}$$

Ou seja, calculamos o erro de previsão para cada ponto do intervalo  $t$ , elevamos ao quadrado, somamos todos eles, dividimos pelo número de dados utilizados  $h$  e, por fim, extraímos a raiz.

A segunda estatística é o erro médio absoluto, também referido como MAE. Esse é dado a partir de:

$$MAE = \frac{(\sum_{t=T+1}^{T+h} |q\hat{x}_t - qx_t|)}{h}$$

Nesse caso, para cada erro de previsão é aplicado o operador módulo, que transforma valores negativos em positivos e, a seguir, cada um desses erros absolutos são divididos pelo total de dados. Por fim, somam-se todos esses erros. A terceira estatística é o erro percentual médio absoluto, também referido como MAPE, e que é dado por:

$$MAPE = 100 \frac{(\sum_{t=T+1}^{T+h} |\frac{q\hat{x}_t - qx_t}{qx_t}|)}{h}$$

Nesse caso, cada erro de previsão é dividido pelo valor observado, extraído o valor absoluto e dividido pelo número de dados. Por fim, esses resultados são somados e multiplicados por 100.

**Programação 6.4.1** A terceira forma de fazer previsão é via programa. A primeira linha determina a equação de regressão de nome *eq1*. A seguir, é especificado o período para a previsão. Por fim, é feita a previsão para a equação e o resultado colocado na série *qx\_saf*.

Também é pedido que seja fornecido o S.E., e damos o nome ao mesmo de *qx\_sase*. Por fim, construímos outras duas séries de dados, uma para especificar o intervalo de previsão mínimo, com 2 desvios-padrão e outra série para o intervalo máximo, também com 2 desvios-padrão.

```
equation eq1.ls qx_sa yw_sa c
smp1 2000M1 2015M7
eq1.fit qx_saf qx_sase
series minimo=qx_saf-2*qx_sase
series maximo=qx_saf+2*qx_sase
```

Por fim, há um bloco de resultados que se refere ao coeficiente de desigualdade de Theil onde os resultados oscilam entre 0 e 1, sendo que um modelo com resultado 0 é considerado como um que faz a perfeita previsão dos dados. A primeira estatística é uma relação do RMSE total com suas partes, a prevista junto com a observada. Note que no numerador temos o resultado para a estatística RMSE que calcula o erro para cada ponto de previsão. Por outro lado, no denominador, essa estatística é quebrada em duas partes. Na primeira, cada valor previsto da variável dependente é elevado ao quadrado e dividido pelo número total de dados. Ao final, esses valores são somados e extraídos a raiz. Na segunda parte do denominador estão os valores observados, onde os mesmos são elevados ao quadrado, divididos pelo total de dados e, por fim, somados e extraída a raiz.

$$Theil = \frac{\sqrt{\frac{(\sum_{t=T+1}^{T+h} (q\hat{x}_t - qx_t)^2)}{h}}}{\sqrt{\frac{(\sum_{t=T+1}^{T+h} (q\hat{x}_t)^2)}{h}} + \sqrt{\frac{(\sum_{t=T+1}^{T+h} (qx_t)^2)}{h}}}$$

As três estatísticas seguintes de previsão são proporções. A primeira delas, denominada de *Bias Proportion*, relaciona duas medidas. No numerador temos a diferença entre o valor previsto médio (valor previsto dividido pelo total de dados) e a média do valor observado. A seguir, esse valor é elevado ao quadrado. Note que  $\sum_{t=T+1}^{T+h} q\hat{x}_t/h$  é a média do valor previsto. No denominador temos a estatística de RMSE sem a extração da raiz. Essa estatística mostra o quanto a média da previsão se distancia da média da série atual. Ela irá oscilar entre 0 e 1. Se for 0, significa que a média dos valores previstos é igual à média dos valores observados. Por outro lado, se for 1, significa que a média dos valores previstos são bem diferentes dos valores observados. Portanto, quanto mais próximo de 0 for o valor de *bias*, melhor é o modelo estimado.

$$bias = \frac{\left( \frac{\sum_{t=T+1}^{T+h} \hat{q}x_t}{h} - q\bar{x}_t \right)^2}{\frac{\sum_{t=T+1}^{T+h} (q\hat{x}_t - qx_t)^2}{h}}$$

A segunda estatística é a *Variance Proportion*, justamente porque relaciona a variância. Nesse caso, no numerador temos a diferença entre a variância da previsão com a variância do valor observado. A seguir, essa diferença é elevada ao quadrado. Essa estatística mostra o quanto a variância do erro de previsão se distancia da variância do erro do valor observado. Quando esse valor for próximo de 0, menor é a diferença das variâncias entre o valor previsto e observado, ou seja, o modelo é melhor do que aquele que apresenta uma estatística de *variance* mais próxima de 1.

$$variance = \frac{(\sigma_{\hat{y}} - \sigma_y)^2}{\frac{\sum_{t=T+1}^{T+h} (q\hat{x}_t - qx_t)^2}{h}}$$

A terceira estatística é a *Covariance Proportion*, que considera a estimativa da covariância entre os valores previstos e observados. No numerador temos que  $r$  é a correlação entre o valor previsto e observado. Essa estatística mede os erros de previsão restante. Quanto melhor for o modelo, menor deve ser a estatística *bias* e *variance* o que, por sua vez, faz com que a maioria do viés do modelo esteja concentrado na estatística de covariância. Essa estatística também vai de 0 a 1.

$$covariance = \frac{2(1-r)\sigma_{\hat{y}}\sigma_y}{\frac{\sum_{t=T+1}^{T+h} (q\hat{x}_t - qx_t)^2}{h}}$$

Vamos agora juntar os conhecimentos adquiridos com a regressão simples e a estimativa por alisamento exponencial para produzir uma previsão da variável  $qx$  alguns meses à frente. No arquivo *regressão simples.wfl* as séries já estão ajustadas sazonalmente e nomeadas com o sufixo "\_sa". A ideia é fazer uma regressão simples com todos os dados disponíveis. Como não sabemos a trajetória futura das variáveis independentes, usamos o método do alisamento exponencial para prever vários passos a frente. A seguir, fazemos uma previsão da variável dependente considerando essas trajetórias.

**Programação 6.4.2** Podemos usar a técnica de alisamento exponencial para definir uma trajetória para as variáveis independentes e, de posse desses valores, usar o recurso de previsão do *EViews*<sup>®</sup> para prever o comportamento da variável dependente.

```

smpl @first @last
for %a px_sa yw_sa
%a.smooth(m,e,e,e) %asm
next
smpl @first 2013M7
series pxsa=px_sa
series ywsa=yw_sa
smpl 2013M7 2015M12
pxsa=px_sasm
pxwsa=yw_sasm
smpl @first @last
equation eq1.ls qx_sa ywsa pxsa c
smpl 2013m7 2015m12

```

```
eq1.fit qx_saf
smp1 @first @last
```

## 6.5 ANEXO ESTATÍSTICO

### 6.5.1 MÍNIMOS QUADRADOS ORDINÁRIOS

Vimos anteriormente que, em um modelo de regressão simples, partindo dos dados de  $y$  e  $x$  queremos encontrar a equação que melhor irá descrever o comportamento dos mesmos. Nesse caso, considerando a relação  $\text{lineary} = \alpha + \beta x + \varepsilon$ , procuramos os valores de  $\hat{\alpha}$  e  $\hat{\beta}$ . Um dos métodos que podem ser empregados para estimar esses valores é o MQO (Mínimos Quadrados Ordinários), que consiste na minimização da soma ao quadrado dos resíduos:

$$\underset{(\alpha, \beta)}{\text{Min}} \sum_{i=1}^n (\varepsilon_i)^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

Inicialmente, vamos resolver esse problema para  $\alpha$ .

$$\frac{\partial (\sum_{i=1}^n \varepsilon_i^2)}{\partial \alpha} = - \sum_{i=1}^n 2(y_i - \alpha - \beta x_i) = 0$$

$$- \sum y_i + n\alpha + \beta \sum x_i = 0$$

$$n\alpha = \sum y_i - \beta \sum x_i$$

$$\alpha = \frac{\sum y_i}{n} - \beta \frac{\sum x_i}{n}$$

$$\hat{\alpha} = \bar{y} - \beta \bar{x}$$

Para facilitar o cálculo de  $\hat{\beta}$ , podemos substituir o valor de  $\hat{\alpha}$  encontrado na equação dos resíduos:

$$\varepsilon_i = y_i - \hat{\alpha} - \hat{\beta} x_i$$

$$\varepsilon_i = y_i - (\bar{y} - \beta \bar{x}) - \hat{\beta} x_i$$

$$\varepsilon_i = (y_i - \bar{y}) - \hat{\beta} (x_i - \bar{x})$$

O termo  $(y_i - \bar{y})$  representa o desvio de cada  $y_i$  em relação à média amostral  $\bar{y}$ . Dessa forma, teremos  $i$  desvios, que podem ser representados por  $y_i^*$ . O mesmo se aplica para os desvios de  $x_i$ , no qual temos,  $x_i^*$ . Assim, elevando esse termo ao quadrado e somando para todos os valores  $i$ :

$$\sum_{i=1}^n (\varepsilon_i)^2 = \sum_{i=1}^n (y_i^* - \hat{\beta} x_i^*)^2$$

Minimizando esse termo em relação a  $\beta$  encontraremos:

$$\frac{\partial (\sum \varepsilon^2)}{\partial \beta} = - \sum 2(y_i^* - \hat{\beta} x_i^*) x_i^* = 0$$

$$- \sum y_i^* x_i^* + \hat{\beta} \sum (x_i^*)^2 = 0$$

$$\hat{\beta} \sum (x_i^*)^2 = \sum y_i^* x_i^*$$

$$\hat{\beta} = \frac{\sum y_i^* x_i^*}{\sum (x_i^*)^2}$$

$$\hat{\beta} = \frac{\text{cov}(y, x)}{\text{var}(x)}$$

**Exercício 6.1** Utilizando o arquivo *regressão simples.wfl*, repita os testes e confirme os resultados apresentados nesse capítulo. ■

**Exercício 6.2** Imagine duas regressões distintas, *eq1* e *eq2*, que possuem a mesma variável dependente,  $y$ . Onde, *eq1* é uma regressão simples, *eq2* possui três variáveis independentes, estatisticamente diferentes de zero, e a soma dos resíduos ao quadrado de *eq2* é maior que a de *eq1*. Podemos dizer que o  $R^2$  de *eq2* é maior que o  $R^2$  de *eq1*, pois *eq2* explica melhor os movimentos de  $y$ ? Por quê? ■

**Exercício 6.3** Em posse do arquivo *regressão simples.wfl*, adote *qx\_sa* como variável dependente e rode cinco regressões simples, utilizando as variáveis *px\_sa*, *pm\_sa*, *qm\_sa*, *y\_sa* e *yw\_sa* como explicativa. Então, escolha o melhor modelo pelo  $R^2$ . ■

**Exercício 6.4** Em posse do arquivo *regressão simples.wfl*, adote *qx\_sa* como variável dependente e rode cinco regressões simples, utilizando as variáveis *px\_sa*, *pm\_sa*, *qm\_sa*, *y\_sa* e *yw\_sa* como explicativa. Então, escolha o melhor modelo pelo critério de Schwartz. ■

**Exercício 6.5** Em posse do arquivo *regressão simples.wfl*, adote *qm\_sa* como variável dependente e rode cinco regressões simples, utilizando as variáveis *px\_sa*, *pm\_sa*, *qx\_sa*, *y\_sa* e *yw\_sa* como explicativa. Então, escolha o melhor modelo pelo critério de Hannan-Quinn. ■

**Exercício 6.6** Considerando a equação  $qm_t = \alpha_1 + \beta_1 y_t + \varepsilon_t$ , onde  $qm_t$  representa as importações de produtos de borracha e material plástico e  $y_t$  o PIB do Brasil. Utilize o método dos mínimos quadrados para encontrar a elasticidade da renda, apresente seu intervalo de confiança de 95% e explique se o resultado está de acordo com o esperado. ■

**Exercício 6.7** A partir da equação da quantidade importada como função da renda, teste se a inclusão de *px\_sa*, *pm\_sa*, *qm\_sa*, *y\_sa* e *yw\_sa* são significativas para o modelo. ■

**Exercício 6.8** Analise os resíduos na regressão da equação  $qm_t = \alpha_1 + \beta_1 y_t + \varepsilon_t$  e responda:  
 A) Os resíduos apresentam distribuição normal?  
 B) Os resíduos são independentes?  
 C) Existe autocorrelação nos resíduos?  
 D) Os resíduos apresentam comportamento homocedástico ou heteroscedástico? ■

**Exercício 6.9** Preencha a tabela a seguir com os resultados dos testes de heteroscedasticidade, apresentados nesse capítulo. Então, conclua sobre o padrão de comportamento dos resíduos da

regressão na equação  $qm_t = \alpha_1 + \beta_1 y_t + \varepsilon_t$ .

Heteroscedasticidade				
Teste	Estatística F	Prob. F	Obs*R2	Prob. Qui Quadrado
Breusch-Pagan-Godfrey				
Harvey				
Glejser				
ARCH ( 1 lag)				

**Exercício 6.10** Utilize o teste Quandt-Andrews para verificar a possível existência de quebra estrutural na regressão  $qm_t = \alpha_1 + \beta_1 y_t + \varepsilon_t$ .

**Exercício 6.11** Com o teste de Chow, comprove o resultado sobre a existência ou não de quebra estrutural encontrado no exercício anterior.

**Exercício 6.12** Utilize a estatística F e a razão de verossimilhança do teste RESET de Ramsey para concluir se o modelo está mal especificado na regressão  $qm_t = \alpha_1 + \beta_1 y_t + \varepsilon_t$ .

**Exercício 6.13** Teste a existência de *outliers* na regressão  $qm_t = \alpha_1 + \beta_1 y_t + \varepsilon_t$ , utilizando os testes RStudent, DFFITS e COVRATIO.

**Exercício 6.14** Baseado no coeficiente de desigualdade de Theil escolha qual dos modelos a seguir apresenta o menor erro de previsão.

A)  $qm_t = \alpha_1 + \beta_1 y w_t + \varepsilon_t$

B)  $qm_t = \alpha_1 + \beta_1 y_t + \varepsilon_t$

C)  $qm_t = \alpha_1 + \beta_1 y_t + \beta_2 p m_t + \varepsilon_t$

D)  $qm_t = \beta_1 y_t + \varepsilon_t$

## 6.6 Bibliografia

- Hodrick, R. J., e Prescott, E. C. (1997). Postwar US business cycles: an empirical investigation. *Journal of Money, credit, and Banking*, 1-16.
- Morais, I.A.C., Bertoldi, A., Anjos, A.T.M. (2010), Um modelo não-linear para as exportações de borracha. *Revista Sober*.
- Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 350-371.

## 7. Regressão Múltipla

A passagem da análise de regressão simples para múltipla nada mais é do que acrescentar mais variáveis independentes ( $x$ ), resultando em um modelo da forma:

$$y_t = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon_t$$

Aqui assumimos as mesmas hipóteses relativas aos resíduos que foram feitas anteriormente, ou seja, que possuem uma média zero  $E(\varepsilon_t) = 0$ , variância constante  $E(\varepsilon_t^2) = \sigma^2$ , são independentes entre eles  $E(\varepsilon_t \varepsilon_{t-i}) = 0$  e também entre as diversas variáveis independentes  $E(\varepsilon_t x_i) = 0$  e são distribuídos normalmente  $\varepsilon_t \sim N(0, \sigma^2)$ .

Uma hipótese adicional importante a ser feita aqui é que as variáveis independentes não possuem uma relação linear determinística. Ou seja, que as mesmas não possam ser combinadas de maneira a se produzir uma outra série. Para exemplificar essa questão, suponha um modelo com duas variáveis do tipo:

$$y_t = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_t$$

se existir colinearidade entre as duas variáveis independentes ( $x_1, x_2$ ), como por exemplo  $ax_1 + bx_2 = c$ , então podemos dizer que  $x_2 = \frac{c}{b} - \frac{a}{b}x_1$ , e o modelo final seria diferente do original:

$$y_t = \alpha + \beta_1 x_1 + \beta_2 \left( \frac{c}{b} - \frac{a}{b} x_1 \right) + \varepsilon_t$$

$$y_t = \alpha + \beta_1 x_1 + \frac{c\beta_2}{b} - \frac{a}{b} \beta_2 x_1 + \varepsilon_t$$

$$y_t = \left( \alpha + \frac{c\beta_2}{b} \right) \left( \beta_1 - \frac{a}{b} \beta_2 \right) x_1 + \varepsilon_t$$

ou seja, ao invés de estimar  $\alpha$ , podemos então estimar  $\left( \alpha + \frac{c\beta_2}{b} \right)$ . Além disso, ao invés de estimar  $\beta$ , seria encontrado  $\left( \beta_1 - \frac{a}{b} \beta_2 \right)$ . Portanto, se as variáveis independentes forem correlacionadas, o modelo irá produzir parâmetros bem diferentes dos originais.

## 7.1 O modelo com duas variáveis independentes

Vamos exemplificar o uso da regressão múltipla acrescentando apenas uma variável independente. Considere a estimativa de um modelo linear:

$$y_t = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\varepsilon}_t,$$

onde os resíduos são obtidos a partir de

$$\hat{\varepsilon}_t = y_t - \hat{\alpha} - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2$$

e queremos encontrar os parâmetros  $\hat{\alpha}$ ,  $\hat{\beta}_1$  e  $\hat{\beta}_2$ .

Para tanto, podemos fazer uso do método dos mínimos quadrados ordinários, da mesma forma que foi aplicado para o modelo de regressão simples. Ou seja, vamos minimizar a soma ao quadrado dos resíduos:

$$Q = \sum (\varepsilon_t^2)$$

$$\min Q = \min \sum (y_t - \hat{\alpha} - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2)^2$$

que irá produzir os valores de  $\hat{\alpha}$ ,  $\hat{\beta}_1$  e  $\hat{\beta}_2$ , tal como a seguir<sup>1</sup>:

$$\hat{\alpha} = \bar{y} - \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2,$$

$$\hat{\beta}_1 = \frac{S_{22}S_{1y} - S_{12}S_{2y}}{S_{11}S_{22} - S_{12}^2},$$

$$\hat{\beta}_2 = \frac{S_{11}S_{2y} - S_{12}S_{1y}}{S_{11}S_{22} - S_{12}^2},$$

onde defini-se  $S_{11} = \sum x_1^2 - n\bar{x}_1^2$ ,  $S_{22} = \sum x_2^2 - n\bar{x}_2^2$ ,  $S_{1y} = \sum x_1 y - n\bar{x}_1 \bar{y}$ ,  $S_{2y} = \sum x_2 y - n\bar{x}_2 \bar{y}$  e  $S_{yy} = \sum y^2 - n\bar{y}^2$ .

Da mesma forma que para a regressão simples, além dos coeficientes estimados, na regressão múltipla também é possível encontrar as seguintes estatísticas:

- Soma ao quadrado dos resíduos (RSS) =  $S_{yy} - \hat{\beta}_1 S_{1y} - \hat{\beta}_2 S_{2y}$
- Soma ao quadrado da regressão (ESS) =  $\hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y}$
- Soma ao quadrado total (TSS) =  $ESS + RSS = S_{yy}$
- Coeficiente de determinação.  $\frac{ESS}{TSS} = R_{12}^2 = \frac{\hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y}}{S_{yy}}$

Destaca-se que o valor de RSS é a parte da regressão que não é explicada pelo modelo com duas variáveis, ou seja, está relacionada ao resíduo<sup>2</sup>. Já ESS define a parte explicada. Dessa forma, a soma da parte explicada com a não explicada, nos fornece o total, ou seja, TSS. Por fim, relacionando a parte explicada com o total, temos a parcela da variável dependente que é explicada pelo modelo, ou seja, o  $R_{12}^2$ .

Assim como no modelo de regressão simples, aqui podemos encontrar as estatísticas associadas a cada parâmetro. Porém, devido o fato de se ter mais de uma variável independente, é necessário considerar a relação que existe entre elas. Para tanto, usamos o coeficiente de correlação ao

<sup>1</sup>Os passos para se encontrar essas relações podem ser vistos em qualquer livro texto de econometria

<sup>2</sup>Como pode ser visto, a diferença entre esse resultado e o encontrado para o modelo de regressão simples, com uma única variável dependente, deve-se a  $\hat{\beta}_2 S_{2y}$ .

quadrado<sup>3</sup> que, no caso de duas variáveis, é dado por  $r_{12}^2 = \rho^2$ . O conjunto de equações que irá determinar as estatísticas dos coeficientes do modelo de regressão múltipla são dadas por:

$$\begin{aligned}\hat{\sigma}_{\hat{\beta}_1}^2 &= \frac{\hat{\sigma}^2}{S_{11}(1-r_{12}^2)}, \hat{\sigma}_{\hat{\beta}_2}^2 = \frac{\hat{\sigma}^2}{S_{22}(1-r_{12}^2)}, \\ cov(\hat{\beta}_1, \hat{\beta}_2) &= \frac{-\hat{\sigma}^2 r_{12}}{S_{12}(1-r_{12}^2)} \text{ e} \\ \hat{\sigma}_{\hat{\alpha}}^2 &= \frac{\hat{\sigma}^2}{n} + \bar{x}_1^2 \sigma_{\hat{\beta}_1}^2 + 2\bar{x}_1 \bar{x}_2 cov(\hat{\beta}_1, \hat{\beta}_2) + \bar{x}_2^2 \sigma_{\hat{\beta}_2}^2.\end{aligned}$$

tal que o coeficiente de correlação entre  $x_1$  e  $x_2$  é dado por  $r_{12}$ .

Um resultado interessante aqui é que, quanto maior for a correlação entre as duas variáveis,  $x_1$  e  $x_2$ , mantendo tudo o mais constante, maior será o  $r_{12}^2$ . Como  $r_{12}^2$  também está presente no cálculo da variância de  $\hat{\beta}_1$  e  $\hat{\beta}_2$  então, quanto maior for a correlação entre as duas variáveis, maior será a variância desses parâmetros. Da mesma forma, como a variância de  $\hat{\beta}_1$  e  $\hat{\beta}_2$  fazem parte do cálculo da variância de  $\hat{\alpha}$ , podemos inferir que, uma maior correlação entre as variáveis independentes irá resultar em maior variância do intercepto. Portanto, uma elevada correlação entre as variáveis independentes torna insignificante a estimativa de seus coeficientes. Por fim, enquanto no modelo de regressão simples os graus de liberdade utilizados para se fazer os testes estatísticos eram dados por  $n - 2$ , no modelo de regressão múltipla, com 2 variáveis independentes, tem-se  $n - 3$ . No limite, para  $k$  variáveis independentes teremos que os graus de liberdade são dados por  $n - k - 1$ .

Vejam como seria o exemplo da estimativa de um modelo de regressão múltipla acrescentando apenas uma variável ao modelo de regressão simples feito anteriormente. Nesse caso, escolhemos adicionar os preços praticados pelo exportador, dado por  $px_t$  mas, ajustado sazonalmente, e a nossa equação ficaria:

$$qx_t = \alpha_1 + \beta_1 yw_t + \beta_2 px_t + \varepsilon_t \quad (7.1)$$

Tal como antes, temos duas formas distintas de estimar essa equação, como mostrado na Figura 7.1. A primeira seria selecionando a variável dependente e a seguir, todas as outras independentes. Depois, clique em **open /as equation ...**, abrindo a janela da Figura 7.1a. A segunda forma seria selecionar **quick /estimate equation ...** e escrever a equação, conforme a Figura 7.1b.

As duas formas de estimativa irão conduzir ao mesmo resultado e o *EViews*<sup>®</sup> irá mostrar um conjunto de informações, como mostrado na Figura 7.2. Como sugestão, prefira estimar conforme a Figura 7.1a, pois tal procedimento é condição necessária para realizar alguns testes no futuro.

Note que aparece um coeficiente a mais na nossa equação. Nesse caso, o  $c(3)$ , que é o parâmetro relacionado ao preço de exportação  $px_t$ . Todas as demais estatísticas informadas são iguais ao modelo de regressão simples mas, com algumas diferenças na interpretação. Para mostrar esse resultado em formato de equação, usamos:

$$qx_t = -21,6254 + 1,5312yw_t - 0,4280px_t + \varepsilon_t$$

(10,0968)      (0,1854)      (0,0851)

onde, entre parênteses, ficam descritos os valores dos respectivos desvio-padrão.

Tal qual nos resultados apresentados para o modelo de regressão simples, na coluna especificada como **“StdError”**, estão os desvios-padrão de cada parâmetro. Depois, a estatística  $t$  (*t-statistic*) e o  $p$ -valor (*Prob*). A primeira é utilizada para testar se o seu respectivo parâmetro é estatisticamente diferente de zero, a partir da fórmula:

$$t = \frac{x - \mu}{\sigma}.$$

<sup>3</sup>Repare a diferença que existe entre  $R_{12}^2$  e  $r_{12}^2$ . O primeiro representa a relação entre as duas variáveis independentes e a dependente. Por outro lado,  $r_{12}^2$  está relacionado apenas à relação que existe entre as variáveis independentes.

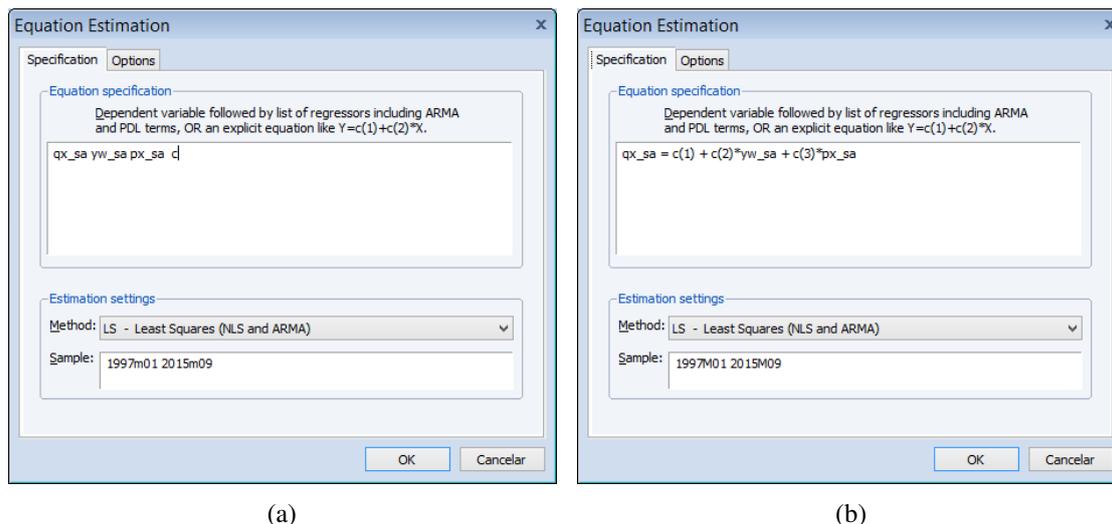


Figura 7.1: Como estimar uma regressão múltipla

Dependent Variable: QX\_SA  
 Method: Least Squares  
 Date: 12/22/15 Time: 09:35  
 Sample (adjusted): 2000M01 2015M07  
 Included observations: 187 after adjustments  
 QX\_SA = C(1) + C(2)\*YW\_SA + C(3)\*PX\_SA

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-21.62543	10.09684	-2.141802	0.0335
C(2)	1.531192	0.185393	8.259169	0.0000
C(3)	-0.428029	0.085071	-5.031428	0.0000
R-squared	0.437990	Mean dependent var		85.16810
Adjusted R-squared	0.431882	S.D. dependent var		17.50486
S.E. of regression	13.19405	Akaike info criterion		8.013322
Sum squared resid	32031.26	Schwarz criterion		8.065158
Log likelihood	-746.2456	Hannan-Quinn criter.		8.034325
F-statistic	71.69825	Durbin-Watson stat		0.398232
Prob(F-statistic)	0.000000			

Figura 7.2: Resultado da Regressão Múltipla

Por exemplo, podemos testar se o parâmetro da elasticidade-preço da demanda é estatisticamente igual a zero ( $\beta_2 = 0$ ) a partir de:

$$t = \frac{\beta_2 - 0}{\sigma_{\beta_1}} = \frac{-0,4280}{0.085071} = -5,0314.$$

Por fim, o resultado do Prob irá indicar se aceitamos ou rejeitamos a hipótese nula de que o coeficiente em questão é estatisticamente igual a zero. Destaca-se que, para esse teste, estamos assumindo uma distribuição t-student. No nosso exemplo, tanto para o coeficiente da constante, quanto para o da renda, rejeitamos a hipótese nula de que são estatisticamente iguais a zero.

**Programação 7.1.1** Tal qual na regressão simples, a outra forma de estimar um modelo de regressão múltipla é via programação, apenas acrescentando o nome das novas variáveis a serem utilizadas. Vejamos o exemplo de se ter uma regressão com duas variáveis independentes, adicionando apenas os preços internacionais “ $px_i$ ”

```
Smpl 2000m01 2015m07
equation eq1.ls qx_sa yw_sa px_sa c
```

A estatística  $t$  e seu respectivo teste podem ser aplicados a partir de uma programação, tal qual mostrado na regressão simples. Nesse caso, queremos testar se  $\beta_2 = 0$ . Primeiro especificamos a estatística  $t$  e armazenamos a mesma em um escalar de nome **estatisticat**. A seguir, criamos uma tabela com três linhas e uma coluna de nome **testet**, e armazenamos na primeira linha o valor de **estatisticat**, na segunda linha o p-valor e, na terceira linha uma variável “string” que irá nos dizer se aceitamos ou rejeitamos a hipótese nula. Para fazer isso, usamos o comando “if” e também como nível de significância 5%.

```
scalar estatisticat=eq1.@tstats(2)
table(3,1) testet
testet(1,1)=estatisticat
teste(2,1)=@tdist(estatisticat,157)
if testet(2,1)>0.05 then
    estet(3,1)="aceitamos h0"
else
    estet(3,1)="rejeitamos h0"
endif
```

Além desses resultados básicos, tal qual no modelo de regressão simples, há diversos outros que são mostrados logo abaixo e que servem para avaliar o modelo em questão. O R-squared, conhecido como  $R^2$ , tem um valor de 0,437990 mas, deve ser interpretado de maneira diferente ao valor encontrado para o  $R^2$  do modelo de regressão simples. Aqui, dizemos que: “cerca de 43,80% das variações em  $qx$  são explicadas por variações em  $yw$  e  $px$ ”. A fórmula é tal como antes, e dada por:

$$R^2 = 1 - \frac{\sum_{t=1}^T \hat{\varepsilon}_t^2}{\sum_{t=1}^T (y_t - \bar{y})^2},$$

onde  $\sum_{t=1}^T \hat{\varepsilon}_t^2$  é denominado de “soma do quadrado dos resíduos” (*sum squared resid*) e, no nosso exemplo tem valor de 32031,26. Tal qual na regressão simples, o termo  $\sum_{t=1}^T (y_t - \bar{y})^2$  representa o quanto a variável dependente desvia em relação à sua média, ou então, mantendo a notação anterior, é o mesmo que  $S_{yy} = \sum y^2 - n\bar{y}^2$ . Antes de prosseguir no cálculo, cabe destacar que para realizar a regressão o *EViews*<sup>®</sup> precisou ajustar a amostra, como visto na Figura 7.2 em **Sample (adjusted): 2000M01 2015M07**. Essa alteração é feita pois em algumas das séries utilizadas faltam as observações anteriores a janeiro de 2000, como a série  $yw\_sa$ . Assim, para encontrar o valor de  $S_{yy}$ , utiliza-se a média da variável dependente nesse período, representada na Figura 7.2 por **Mean dependent var**. No presente caso, a média da variável dependente é 85,16810. E, se fizermos o quadrado da diferença de cada observação da variável dependente em relação a sua média e somarmos, encontraremos 56.994,14. Assim:

$$R^2 = 1 - \frac{32031,26}{56994,14} = 0,437990284.$$

O valor de  $R^2$  ajustado, “Adjusted R-squared” corrigido pelo número de coeficientes ( $k$ ) que estão sendo utilizadas no modelo. Sua fórmula geral é dada por:

$$\bar{R}^2 = 1 - (1 - R^2) \frac{T - 1}{T - k},$$

e, aplicando para os dados do modelo:

$$\bar{R}^2 = 1 - (1 - 0,437990) \frac{187 - 1}{187 - 3} = 0,431881483.$$

No caso da regressão simples, o  $R^2$  tem uma interpretação direta. Porém, na regressão múltipla, podemos estar interessados não em identificar quanto o modelo é explicado pelas variáveis independentes, mas, sim, quanto que cada variável explica o modelo. Nesse caso, usamos a estatística de correlação parcial, dada por:

$$R_{y,x}^2 = \frac{t_x^2}{t_x^2 + (T - k)},$$

onde  $t_x$  é a estatística t do coeficiente  $x$ ,  $T$  é o número de observações e  $k$  o número de parâmetros do modelo completo.

Suponha, por exemplo, que na nossa regressão acima, se queira determinar o efeito da elasticidade-renda da demanda ( $yw\_sa$ ), mantendo todas as demais variáveis independentes constantes, eliminando o impacto que  $\beta_2$  tem sobre  $\beta_1$ . Assim, usamos:

$$R_{qx,yw}^2 = \frac{t_{yw}^2}{t_{yw}^2 + (T - k)},$$

$$R_{qx,yw}^2 = \frac{8,2592^2}{8,2592^2 + (187 - 3)} = 0,270460.$$

Para o caso de se querer saber o impacto apenas da elasticidade-preço usamos:

$$R_{qx,px}^2 = \frac{-5,0314^2}{-5,0314^2 + (187 - 3)} = 0,120943.$$

Portanto, note que a elasticidade tem uma capacidade explicativa mais do que o dobro da variável preço. A soma de ambas dá 0,39 de um total de 0,43 do valor de  $R^2$ .

O desvio-padrão da regressão (**S.E. of regression** na Figura 7.2) é dado por:

$$\sigma = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{RSS}{T - k}}$$

onde, além do número de observações ( $T$ ) e do número de parâmetros ( $k$ ), temos  $RSS$  como a soma ao quadrado dos resíduos. Na regressão representada pela Equação 7.1, que temos como exemplo, o desvio-padrão da regressão será dado por:

$$\sigma = \sqrt{\frac{32031,26}{187 - 3}} = 13,19405$$

A estatística do log da verossimilhança (log likelihood) e os critérios de comparação de Akaike, Schwarz e Hannan-Quinn são feitos da mesma forma que para a regressão simples. Por isso não entramos no detalhe desses resultados. E estatística F também é calculada da mesma forma, mas, sua interpretação é feita de maneira diferente. Nesse caso, como a nossa regressão tem três parâmetros, o teste-F está testando, na hipótese nula, se:

$$\alpha_1 = \beta_1 = \beta_2 = 0$$

Pelos resultados apresentados no teste F, onde seu p-valor=0, não podemos aceitar a hipótese nula. Nesse caso, os parâmetros em conjunto são estatisticamente diferentes de zero.

**Programação 7.1.2** Como forma de ilustrar cada uma das funções que são utilizadas para gerar as estatísticas apresentadas acima, essa rotina monta uma tabela com quatro colunas e 8 linhas e, a seguir, especifica cada estatística e coloca a mesma de tal forma que seja possível comparar com os resultados apresentados pelo *EViews*<sup>®</sup>. Note que, na última linha (oitava linha) foi colocada a correlação parcial, primeiro da elasticidade-renda e depois da elasticidade-preço.

```
Smpl 1997m1 2015m09
equation eq1.ls qx_sa c yw_sa px_sa
table(8,4) result
result(1,1)"R2"
result(1,2)=eq2.@r2
result(2,1)="R2 ajustado"
result(2,2)=eq2.@rbar2
result(3,1)="erro padrão da regressão"
result(3,2)=eq2.@se
result(4,1)="Soma dos resíduos ao quadrado"
result(4,2)=eq2.@ssr
result(5,1)="Log da verossimilhança"
result(5,2)=eq2.@logl
result(6,1)="estatística F"
result(6,2)=eq2.@f
result(7,1)="p-valor da estatística f"
result(7,2)=eq2.@fprob
result(1,3)="média da variável dependente"
result(1,4)=eq2.@meandep
result(2,3)="desvio-padrão da variável dependente"
result(2,4)=eq2.@sddep
result(3,3)="Akaike"
result(3,4)=eq2.@aic
result(4,3)="Schwarz"
result(4,4)=eq2.@schwarz
result(5,3)="Hannan-Quinn"
result(5,4)=eq2.@hq
result(6,3)="Durbin-watson"
result(6,4)=eq2.@dw
result(8,1)="parcela explicada por yw"
scalar ryw=(eq2.@tstats(2)^2/(eq2.@tstats(2)^2+(eq2.@regobs-eq2.@ncoef))
result(8,2)= ryw
result(8,3)="parcela explicada por px"
scalar rpx=(eq2.@tstats(3)^2/(eq2.@tstats(3)^2+(eq2.@regobs-eq2.@ncoef))
result(8,4)= rpx
```

Com os resultados da equação abertos, podemos ver o gráfico clicando em **Resids** ou **View /Actual, Fitted, Residual /Actual, Fitted, Residual Graph ...**. Observando a Figura 7.3, note que, agora, nosso modelo erra menos do que no modelo de regressão simples, demonstrado na Figura 6.5, e que também pode ser comprovado pelo resultado do R2.

A estimativa dos valores para cada período é feita tal como no modelo de regressão simples. Só que, agora, temos uma variável a mais para especificar, como demonstrado pela Equação 7.1 do modelo de regressão múltipla. Nesse caso, vejamos como é a estimativa do valor de  $qx$  para janeiro

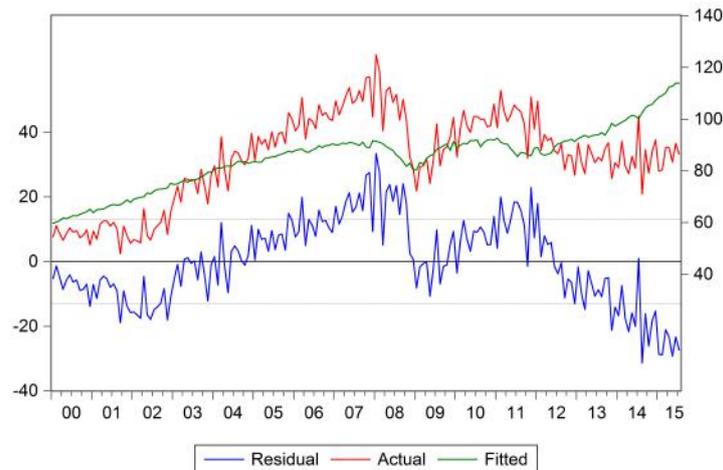


Figura 7.3: Resultado do Modelo de Regressão Múltipla

de 2003:

$$qx_t = -21,6254 + 1,5312yw_t - 0,4280px_t + \varepsilon_t.$$

(10,0968)
(0,1854)
(0,0851)

Naquela data, temos os seguintes valores para as variáveis independentes  $yw_{jan/2003} = 1,917077$  e  $px_{jan/2003} = 1,904287$ . Substituindo esses valores na equação acima encontramos:

$$qx_{jan/2003} = -21,6254 + 1,5312(1,917077) - 0,4280(1,904287) = -19,5050.$$

(10,0968)
(0,1854)
(0,0851)

A seguir, todos os procedimentos de diagnósticos dos coeficientes ou então dos resíduos pode ser feito tal como no modelo de regressão simples. Por isso, não iremos apresentá-los aqui.

**Programação 7.1.3** O loop a seguir pode ser usado para rodar várias regressões e colocar os resultados em uma tabela, permitindo que se faça a escolha do melhor modelo através do R2 e dos critérios de comparação.

```
smp1 1997m01 2015m09
table(5,5) modelos
modelos(1,2)='eq1'
modelos(1,3)='eq2'
modelos(1,4)='eq3'
modelos(1,5)='eq4'
modelos(2,1)='R2'
modelos(3,1)='akaike'
modelos(4,1)='Schwarz'
modelos(5,1)='Hannan-Quinn'
equation eq1.ls qx_sa c yw_sa
equation eq2.ls qx_sa c yw_sa px_sa
equation eq3.ls qx_sa c yw_sa px_sa pxw_sa
equation eq4.ls qx_sa c yw_sa px_sa pxw_sa e_sa
```

```

for !i=1 to 4
  modelos(2,!i+1)=eq!i.@r2
  modelos(3,!i+1)=eq!i.@aic
  modelos(4,!i+1)=eq!i.@schwarz
  modelos(5,!i+1)=eq!i.@hq
next

```

Os testes de diagnóstico de estabilidade, tal como o Teste de Chow, Teste de Quandt-Andrews, Teste de Previsão de Chow, Teste de Ramsey, estimativas recursivas e estatísticas de influência, podem ser aplicados da mesma forma apresentada no capítulo de Regressão Simples. A sua interpretação também é feita da mesma forma.

## 7.2 Previsão - Forecast

A previsão em modelos de regressão múltipla pode ser feita tal como nos modelos de regressão simples. Com a equação aberta seleccione **Forecast**, escolha um nome para a série de resultados da previsão, aqui usamos `qx_saf`, e um nome para a série do desvio-padrão, usamos `qx_sef`. Por fim seleccione o intervalo de previsão e clique em ok.

No resultado, tal como mostrado na Figura 7.4, podemos ver a série de previsão com seu respectivo intervalo de confiança com 2 desvios. Para comparação dos resultados do modelo de regressão simples com o modelo de múltiplas variáveis, observa-se as estatísticas de erro de previsão (RMSE, MAE e MAPE). Enquanto os resultados de RMSE, MAE e MAPE da regressão simples foram, respectivamente, 18,9004, 16,1485 e 23,8140, a Figura 7.4 apresenta os resultados dessas estatísticas de previsão para o modelo de regressão múltipla com as variáveis independentes `px_sa` e `yw_sa`. Assim, as estatísticas de previsão mostram que, em comparação com o modelo de regressão simples, o modelo com duas variáveis adere melhor aos dados, apesar dos resultados ainda estarem longe do ideal.

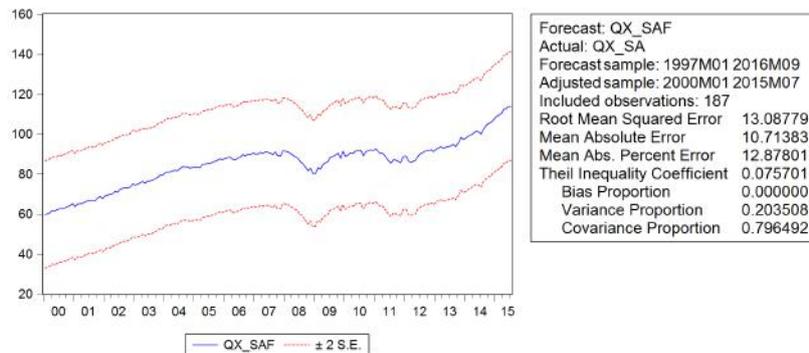


Figura 7.4: Previsão do modelo de regressão múltipla

**Programação 7.2.1** Também podemos fazer uma previsão de um modelo de regressão múltipla via programa. A primeira linha determina a equação de regressão, onde colocamos primeiro a variável dependente e, a seguir, as independentes. A seguir, é especificado o período para a previsão. No exemplo abaixo colocamos para todo o período amostral. Por fim, é feita a previsão para a equação e o resultado colocado na série `qx_saf`, juntamente com o desvio padrão com nome `qx_sef`. Esse último irá permitir que seja construído o intervalo de confiança da previsão, referente aos comandos `series min` e `series max`.

```
equation eq1.ls qx_sa c yw_sa px_sa
smp1 1997m01 2015m09
eq1.forecast qx_saf qx_sef
series min=qx_saf-2*qx_sef
series max=qx_saf+2*qx_sef
```

### 7.3 Método STEPLS

A programação não é a única maneira de se fazer várias regressões, testes e a aplicação de diversas outras ferramentas estatísticas no *EViews*<sup>®</sup>. Também podemos aplicar o método STEPLS. A partir deste, várias equações são estimadas, considerando as variáveis em questão e, fornecidos os resultados para que seja selecionada a melhor.

Para no nosso exemplo, temos cinco variáveis que podem ser combinadas de várias formas. Além do quantum de exportações do setor de produtos de borracha e materiais plásticos (*qx\_sa*), do PIB mundial (*yw\_sa*) e do índice de preços das exportações desse setor (*px\_sa*), temos o PIB brasileiro (*y\_sa*), o índice de preço das importações (*pm\_sa*) e o quantum das exportações do setor de produtos de borracha e material plástico (*qm\_sa*). As mesmas estão no arquivo *07rm.wf1*. Os mesmos já estão ajustados sazonalmente. Nosso objetivo é encontrar a melhor equação linear com, no máximo cinco variáveis independentes.

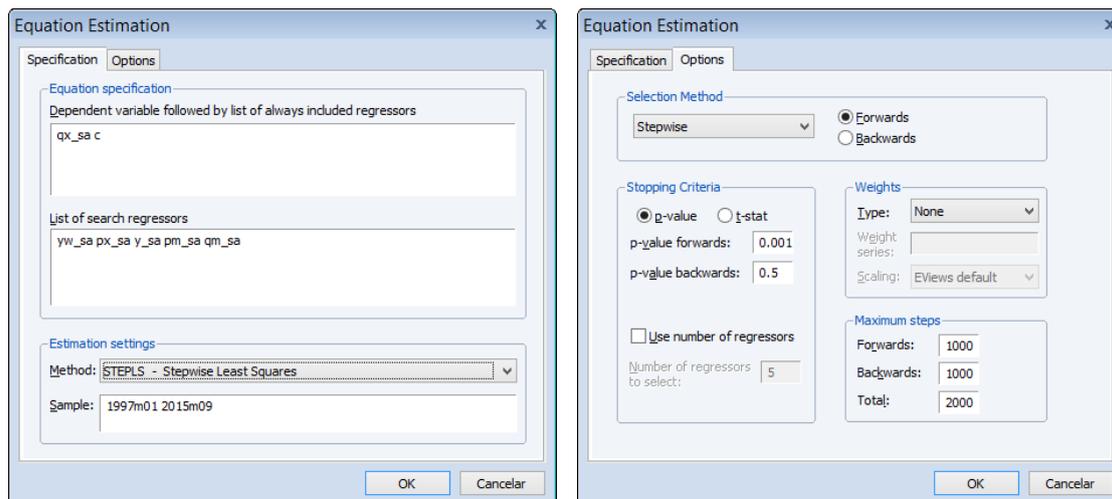
Como já foi demonstrado, para estimar uma equação podemos selecionar as variáveis e abrir como equação ou via **quick /estimate equation ...** e escrever a equação que queremos estimar. A seguir, na opção **method**, selecione **STEPLS - Stepwise Least Squares**. Na parte onde é possível especificar a variável dependente, coloque *qx\_sa* e a constante, depois, na parte da lista dos repressores, especifique todas as demais independentes, conforme a Figura 7.5a.

Na aba options podemos escolher diversas formas de estimação, conforme Figura 7.5b. Vamos começar selecionando stepwise forwards. A diferença entre as opções de forwards e backwards está na adição ou remoção de variáveis independentes com o menor p-valor ou maior estatística t na equação, conforme critério definido. Além disso, também é possível selecionar o método *uni-directional*, *swapwise* e *combinational*. O método unidirecional adiciona (forward) ou remove (backward) variáveis até que o menor p-valor das variáveis não incluídas ser maior que o critério de parada definido. O método stepwise nada mais é que uma combinação da rotina uni-direcional forward e backward. No método swapwise é necessário optar por máximo ou mínimo incremento ao R<sup>2</sup>. Esse método busca maximizar o R<sup>2</sup>, sendo que o *Max R-Squared Increment* adiciona as variáveis que mais contribuem para o aumento do R<sup>2</sup> e o *Min R-Squared Increment* adiciona as variáveis que menos contribuem para o incremento do R<sup>2</sup>. O método *combinational* testa todas as combinações de variáveis e seleciona o modelo com maior R<sup>2</sup>. Esses métodos serão explorados na próxima subseção, onde haverá uma explanação mais detalhada para cada um deles.

Na opção de critério de parada (*Stopping Criteria*), definimos se o método irá ser coordenado pelo menor p-valor ou maior resultado da estatística t. Essa opção só aparece para o método unidirecional e stepwise. No nosso exemplo, utilizando o stepwise forward, coloque 0.001 para o p-valor forwards e deixe o resto tal como sugerido. Um ponto importante nesse passo é a opção “use number of regressors”, utilizada em todos métodos. Se colocarmos o valor 1, o melhor modelo terá apenas uma variável independente. Se selecionarmos o valor 2, o melhor modelo terá duas variáveis independentes. Se não selecionarmos essa opção, o procedimento irá determinar o número de variáveis independentes a serem consideradas.

O resultado será a seleção de uma equação com todos os coeficientes. O método stepwise inicia, no nosso exemplo, com uma regressão da forma:

$$qx_t = c + \varepsilon_t.$$



(a) Estimando regressão pelo método STEPLS

(b) Opções do método STEPLS

Figura 7.5: Método STEPLS

A seguir, é inserida uma variável independente gerando mais quatro regressões simples, cada uma com uma constante. Por exemplo, será feita uma regressão com o seguinte formato abaixo, onde a variável independente é  $yw$ :

$$qx_t = c + \beta_1 yw\_sa + \varepsilon_t.$$

As demais regressões simples irão ter uma constante e uma variável independente diferente. Dessa forma, teremos uma regressão apenas com  $px\_sa$  como variável independente e assim sucessivamente. De acordo com os nossos dados o modelo final sugerido é dado por:

$$qx_t = c + \beta_1 yw\_sa + \beta_2 y\_sa + \beta_3 qm\_sa + \beta_4 px\_sa + \varepsilon_t.$$

No resultado (Figura 7.6) da estimativa podemos ver que o método manteve apenas uma variável em todas as regreções (Number of always included regressors), a constante, e que o total de variáveis independentes foi 4. O método de seleção é o Stepwise forwards e o critério de inclusão é o p-valor ao nível de 0,001. Note que todos os coeficientes são estatisticamente diferentes de zero e as demais estatísticas podem ser interpretadas de forma igual ao que vimos em modelos de regressão por mínimos quadrados. Além disso, percebemos que a variável  $pm\_sa$  não foi adicionada à regressão, pois o p-valor ficava acima do critério determinado. Em comparação com as outras regressões apresentadas, destacamos o resultado do R2 de 0,8659, maior que a regressão simples e a regressão múltipla apenas com as variáveis  $yw\_sa$  e  $px\_sa$ .

**Programação 7.3.1** O método STEPLS pode ser feito via programação. O default é o método stepwise, para utilizar as outras opções utilizamos `method = uni` (para o uni-directional), `text` (swapwise) ou `comb` (combinatorial). De qualquer forma, vamos utilizar o padrão stepwise. O procedimento forward também é default, não sendo necessário especificar o mesmo. Caso contrário, podemos especificar `back`. No critério de seleção o p-valor é default e, de outra forma, podemos escolher `tstat`. Para definir os critério, utilizamos `ftol=0.001` para o critério forward, sem termos que utilizar `btol = 0.5` para o critério backward, pois este é o valor padrão. O procedimento acima pode ser feito via:

```
eq1.steps(method=stepwise, ftol=0.001) qx_sa c @ yw_sa y_sa qm_sa px_sa
```

Dependent Variable: QX\_SA  
Method: Stepwise Regression  
Date: 12/29/15 Time: 21:01  
Sample (adjusted): 2000M01 2015M07  
Included observations: 187 after adjustments  
Number of always included regressors: 1  
Number of search regressors: 5  
Selection method: Stepwise forwards  
Stopping criterion: p-value forwards/backwards = 0.001/0.5

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
C	-216.9735	10.12210	-21.43562	0.0000
YW_SA	4.366129	0.151420	28.83447	0.0000
Y_SA	-1.132276	0.047430	-23.87232	0.0000
QM_SA	0.358269	0.038817	9.229643	0.0000
PX_SA	-0.414353	0.056301	-7.359660	0.0000

R-squared	0.865897	Mean dependent var	85.16810
Adjusted R-squared	0.862949	S.D. dependent var	17.50486
S.E. of regression	6.480361	Akaike info criterion	6.601804
Sum squared resid	7643.104	Schwarz criterion	6.688197
Log likelihood	-612.2686	Hannan-Quinn criter.	6.636810
F-statistic	293.7906	Durbin-Watson stat	1.411983
Prob(F-statistic)	0.000000		

Selection Summary

---

Added YW\_SA  
Added Y\_SA  
Added QM\_SA  
Added PX\_SA

---

\*Note: p-values and subsequent tests do not account for stepwise selection.

Figura 7.6: Resultado do método STEPLS

### 7.3.1 Os métodos de Seleção STEPLS

Dentro do procedimento de escolha do melhor modelo de regressão é possível selecionar dentre vários tipos de métodos, sendo que, os mesmos, podem ser divididos entre *forward* e *backward* e também tendo como opção de seleção das variáveis o p-valor ou a estatística t. A seguir fazemos uma breve explanação sobre esses métodos.

#### Uni-direcional

Esse processo pode ser utilizado tanto para adicionar variáveis ao modelo (*forward*) quanto para retirar (*backward*) e, nesse caso, a decisão é feita com base ou no p-valor ou na estatística t, sendo necessário escolher qual o critério de decisão para essas duas estatísticas.

Suponha que seja o p-valor. Com essa opção o modelo começa como uma regressão simples, rodando várias regressões com diferentes combinações, sempre tendo apenas uma variável. A variável com menor p-valor fica. A seguir, são feitas regressões múltiplas adicionando mais uma variável dentre todas as que foram especificadas. Aquela que atender o critério estabelecido e tiver o menor p-valor, é acrescentada ao modelo. O procedimento continua até que não seja mais possível adicionar variável que atenda aos critérios especificados, p-valor e número de passos *Maximum steps*.

Já no método unidirecional *backwards* o procedimento se inicia com todas as variáveis e vai retirando aquelas com maior p-valor até que restem apenas aquelas que atendam aos critérios especificados. Nesse caso, como o método é *backward*, é utilizado o critério *Maximum steps backwards*.

Esse procedimento é complementado com a escolha da opção *User Number of Regressors*, que

determina quantas variáveis devem constar no modelo final. Se não selecionar essa opção, o modelo irá conter o máximo de variáveis que atendem as especificações anteriores. Caso contrário, se o mesmo for selecionado, podemos especificar quantas variáveis queremos que o modelo final tenha.

### Stepwise

Da mesma forma que no método unidirecional o método *Stepwise* pode ser escolhido com a opção de *forwards* e *backwards*. Independente da escolha da opção, o fato é que o método *Stepwise* é uma combinação do método unidirecional *forward* com o unidirecional *backward*. O que muda é a ordem de execução da seleção e escolha das variáveis.

Por exemplo, suponha que se tenha escolhido o método *Stepwise forward* com opção de p-valor. Aqui, o processo começa sem variável, são feitas diversas regressões simples, ou seja, adicionando apenas uma variável. A seguir, aquela que apresentar o menor p-valor é mantida no modelo. O procedimento se repete, testando todas as demais variáveis e escolhendo aquela que também irá ter o menor p-valor. Nesse momento teremos um modelo com duas variáveis independentes. Antes de testar a terceira variável, é feito o procedimento *backward* no modelo com duas variáveis independentes. Se alguma delas não atender ao critério do p-valor ou estatística t, é removida.

No passo seguinte é escolhida a terceira variável a ser adicionada no modelo e que deve atender aos critérios especificados (p-valor ou estatística t). Escolhida essa terceira variável, é feito novamente o procedimento *backward* com o modelo tendo três variáveis para confirmar as mesmas. A seguir, para toda e qualquer variável que se queira acrescentar ao modelo é feito o mesmo procedimento, primeiro testando *forward* e, a seguir, *backward*.

Podemos comparar o resultado do método *Stepwise forward* com o método *backward* para ver se encontramos o modelo com o mesmo número de variáveis. Para tanto selecione *backward* e o critério de seleção, que pode tanto ser o p-valor quanto a estatística t. Destaca-se que o método *Stepwise backward* é exatamente o contrário do *Stepwise forward*. Primeiro todas as variáveis são inseridas no modelo e a que tiver o maior p-valor é excluída. A seguir, dentro daquelas que ficaram no modelo é feita a investigação *forward* para confirmar a presença das mesmas.

O procedimento se repete e, as variáveis que foram excluídas são verificadas pelo método *forward*. Se alguma delas tiver um p-valor mais baixo ou uma estatística t maior, é inserida novamente no modelo. O procedimento se repete até que todos os critérios sejam atendidos.

### Swapwise

Esse método utiliza dois importantes critérios de escolha, a estimativa do R quadrado para fazer a seleção do melhor modelo dividindo a escolha entre um incremento máximo ou mínimo e o número de variáveis independentes a considerar.

Vejamos como é o exemplo do método via R quadrado máximo. O procedimento se inicia sem variável independente e, após feitas várias regressões simples, é escolhida aquela que maximiza o R quadrado. A seguir, são testadas as demais variáveis adicionando uma a uma no modelo. A que gerar o maior incremento no R quadrado permanece. Para confirmar a presença dessas duas variáveis, as mesmas são comparadas com cada uma das que estão fora do modelo. Ou seja, imagine que temos uma regressão do tipo:

$$y_t = c + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t.$$

Para confirmar a presença dessas variáveis, são feitas regressões do tipo:

$$y_t = c + \beta_1 x_{1t} + \beta_3 x_{3t} + \varepsilon_t$$

e também

$$y_t = c + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t,$$

e assim sucessivamente para todas as variáveis que não estavam no modelo básico. Isso é feito para ver se as diferentes combinações não gera um R quadrado incremental maior. Uma vez descoberta a melhor combinação com duas variáveis, o procedimento continua para a terceira variável que gera o melhor incremento no R quadrado.

A seguir, partindo de um modelo de três variáveis independentes, são feitas várias combinações para descobrir qual gera o melhor incremento no R quadrado. De outra forma, se escolhermos o método *Swapwise* com R quadrado mínimo, o procedimento é parecido com o que considera o R quadrado máximo. A diferença é que, na hora de testar as diferentes combinações o procedimento é feito escolhendo aquela que gera o menor incremento no R quadrado.

### **Combinatorial**

Nesse método devemos especificar quantas variáveis independentes queremos testar no modelo e as mesmas são testadas em várias combinações e é selecionada aquela combinação que produz o maior R quadrado. Esse método é o que requer o maior número de estimativas e, dependendo do número de variáveis a serem especificadas, o resultado pode demorar em ser fornecido.

## **7.4 Bibliografia**

- Hamilton, J. (1994). Linear Regression Model. In: \_\_\_\_\_. Time Series Analysis. Princeton University Press, pp. 200 - 232.
- Wansbeek, T., e Meijer, E. (2008). Measurement error and latent variables. In: Baltagi, B. H. (Ed.). A companion to theoretical econometrics. John Wiley & Sons, pp. 162 - 179.

The background of the slide is a complex financial data visualization. It features a grid with various charts and data points. On the left, there is a line graph showing market activity from July 2007 to February 2008, with values ranging from 3000 to 8000. A prominent blue line with diamond markers shows a peak in late 2007 followed by a decline. Other lines with square markers show different trends. In the center, there are large, semi-transparent numbers like '42', '58', '47', and '22'. On the right, there are more line graphs with various markers and labels like 'HRT', 'RTG', 'TRG', 'GHT', 'IBNIO', and 'FRT division'. The overall color scheme is blue and white, with a grid pattern.

## Referências Bibliográficas

- [1] Lawrence J. Christiano and Terry J. Fitzgerald. The band pass filter\*. *International Economic Review*, 44(2):435–465, 2003.

