Universidade Federal de Santa Catarina
Centro Sócio–Econômico
Departamento de Ciências Econômicas

Curso de graduação em Ciências Econômicas
a distância

Estatística Econômica e Introdução à Econometria

Milton Biage

B576e Biage, Milton

Estatística Econômica e Introdução à Econometria. / Milton Biage . 3.impri. - Florianópolis : Departamento de Ciências Econômicas/UFSC, 2012.

179p. : il

Curso de Graduação em Ciências Econômicas

Inclui bibliografia

ISBN 978-85-7426-087-7

1. Estatística Econômica. 2. Econometria. 3. Análise de regressão. 4. Educação a distância I.Universidade Federal de Santa Catarina.Departamento de Ciências Econômicas. II. Título

CDU: 330

GOVERNO FEDERAL

Presidente da República	Dilma Vana Rousseff
Ministro da Educação	Fernando Haddad
Secretário de Educação a Distância	Carlos Eduardo Bielschowsky
Coodenador Nacional da Universidade Aberta do Brasil	João Carlos Teatini de Souza Clímaco

UNIVERSIDADE FEDERAL DE SANTA CATARINA

Reitora	Roselane Neckel
Vice–Reitora	Lúcia Helena Martins Pacheco
Pró-Reitor de Desenvolvimento Urbano e Social	Luiz Henrique Vieira Silva
Pró-Reitora de Assuntos Estudantis	Cláudio José Amante
Pró-Reitora de Pesquisa e Extensão	Débora Peres Menezes
Pró-Reitora de Pós-Graduação	Maria Lucia de Barros Camargo
Pró-reitor de Ensino de Graduação	Yara Maria Rauh Müller
Secretário de Planejamento e Finanças	Luiz Alberton
Secretário de Cultura e Arte	Maria de Lourdes Alves Borges
Coordenadora UAB - UFSC	Eleonora Milano Falcão Vieira
Coordenadora Adjunta UAB - UFSC	Dulce Márcia Cruz

CENTRO SÓCIO-ECONÔMICO

Diretor	Ricardo José Araújo Oliveira
Vice-Diretor	Alexandre Marino Costa

DEPARTAMENTO DE CIÊNCIAS ECONÔMICAS

Chefe do Departamento	Armando de Melo Lisboa
Subchefe do Departamento	Brena Paula M. Fernandez
Coordenadora Geral na modalidade a distância	Marialice de Moraes

EQUIPE DE PRODUÇÃO DE MATERIAL - PRIMEIRA EDIÇÃO

Coordenação de Design Instrucional	Fernanda Pires Teixeira
Design Instrucional	Patrícia Cella Azzolini
Revisão Textual	Júlio César Ramos
Coordenação de Design Gráfico	Giovana Schuelter
Design Gráfico	Ana Flávia Maestri
	Felipe Augusto Franke
	Natalia de Gouvêa Silva
Ilustrações	Laís Barbosa
Design de Capa	Guilherme Dias Simões
	Felipe Augusto Franke
	Steven Nicolás Franz Peña
Projeto Editorial	André Rodrigues da Silva
	Felipe Augusto Franke
	Max Vartuli
	Steven Nicolás Franz Pena

EQUIPE DE PRODUÇÃO DE MATERIAL - TERCEIRA EDIÇÃO

Coordenação de Design Instrucional	Andreia Mara Fiala
Coordenação de Design Gráfico	Giovana Schuelter
Design Gráfico	Fabrício Sawczen
	Felipe Augusto Franke
	Patrícia Cella Azzolini
Revisão de Material	Patrícia Cella Azzolini
Ilustrações	Laís Barbosa
	Aurino Manoel dos Santos Neto
Design de Capa	Guilherme Dias Simões
	Felipe Augusto Franke
	Steven Nicolás Franz Peña
Projeto Editorial	André Rodrigues da Silva
	Felipe Augusto Franke
	Max Vartuli
	Steven Nicolás Franz Pena

Sumário

UNIDADE 1

NÚMEROS ÍNDICES

1.1 Números índices, aplicabilidade e construção 14
1.2 Critérios de avaliação da Fórmula de um índice (propriedades dos relativos) 16
1.3 Números índices simples 17
1.4 Critério de decomposição das causas (ou inversão dos fatores) 19
1.5 Números índices de base móvel (ou relativos de ligação) e números índices de base fixa 19
1.6 Mudança de base de um número índice de base fixa 21
1.7 Números índices compostos (ou agregados) 23
Índice agregado simples (de preços e de quantidades) 23
Índices agregados ponderados 27
Índices especiais de preço e quantidade: de Fischer, de Drobish e de Marshal-Edgeworth 37
1.8 Números índices agregados ponderados de base móvel 41
Número Índice de Theil (média geométrica ponderada) 41
Número Índice de Laspeyres com base móvel 41
Índice de Bureau (ou índice de Laspeyres modificado, com base móvel) 42
Considerações sobre o Exemplo 12 do material complementar 44
1.9 Deflação de uma série temporal 46
1.10 Poder aquisitivo 48
1.11 Taxa real ou taxa deflacionada 48
1.12 Deflator implícito de preço e índice quantum 49
Deflator Implícito 49
Índice de Quantum 50
1.13 Índices brasileiros 52
Índice Nacional de Preços ao Consumidor (INPC) e o Índice Nacional de Preços ao Consumidor Amplo (IPCA) 52
Índice de Preços ao Consumidor da FIPE, IPC/FIPE 60
Índice de Custo de Vida do DIEESE (ICV-DIEESE) 61
Índice Geral de Preços de Mercado (IGP-M) da Fundação Getúlio Vargas (FGV) 65
Índice Geral de Preços - Disponibilidade Interna (IGP-DI) da Fundação Getúlio Vargas (FGV) 66

UNIDADE 2

ECONOMETRIA E ANÁLISE DE REGRESSÃO

2.1 Introdução: método científico 76
Pesquisa Indutiva 77
Pesquisa Dedutiva 78
2.2 O que é econometria? 79
2.3 Metodologia da econometria 79
Formulação da teoria ou da hipótese 80
Observação do problema levantado para a pesquisa 80
Especificação do modelo matemático do consumo 82
Especificação do modelo econométrico de consumo 84
Estimativa dos parâmetros do modelo econométrico 85
Teste de hipótese 87
Previsão ou predição 90
2.4 Tipos de econometria 90
Pré-requisitos Matemáticos e Estatísticos 90
O papel do computador 91
2.5 Natureza da análise de regressão 91
Exemplos de dependência de uma variável em relação à outra 92
Relações estatísticas *versus* deterministas 94
Regressão *versus* Causação 94
Regressão *versus* Correlação 95
Diferenças fundamentais entre regressão e correlação 95
Terminologia e Notação 96
Estrutura dos dados econômicos 96
2.6 Análise de regressão de duas variáveis: alguns conceitos básicos 98
O conceito de função de regressão da população (FRP) 103
O significado do termo linear nas variáveis e nos parâmetros 104
Especificação estocástica da FRP 105
O significado do termo perturbação estocástica 106
Função de regressão amostral (FRA) 108

UNIDADE 3

MODELO DE REGRESSÃO DE DUAS VARIÁVEIS: O PROBLEMA DE ESTIMATIVA

3.1 Construção de um modelo de regressão 120
3.2 Dados experimentais e observacionais 123
3.3 Análise de correlação linear 125
3.4 Revisão da concepção de modelos de regressão 129
Uma FRP de duas variáveis 129
Uma FRA de duas variáveis 131
Hipóteses Adjacentes 132
Funções Amostrais e mecanismo dos Mínimos Quadrados Ordinários (MQO) 138
3.5 Considerações sobre o MQO 151
3.6 Regressão de duas variáveis: estimativa de intervalo e teste de hipótese 158
Estimativa de intervalo: alguns conceitos básicos 158
Intervalos de confiança para os coeficientes de regressão β_1 e β_2 159
Intervalos de confiança para Σ^2 163
3.7 Teste de hipótese: comentários gerais 165
Teste de hipótese: a abordagem do intervalo de confiança 165
Teste de hipótese: a abordagem do teste de significância 167

Referências 178

Palavra do professor

Estimado aluno, seja bem-vindo!

O objetivo deste livro de *Estatística Econômica e Introdução à Econometria*, é apresentar, por meio de uma linguagem simples e clara, os conceitos da Teoria de Números Índices e Modelos de Regressão.

Você deve levar em consideração que a disciplina exige alguns conhecimentos básicos de estatística, estudados na disciplina Introdução à Estatística, como: conceitos de projetos experimentais de amostragem, medidas de tendências e de dispersão (conceito de normalidade de distribuição), intervalos de confiança e testes de hipóteses. Além disso, esta disciplina também exige alguns conhecimentos de matemática, que envolvem cálculo diferencial e integral, funções de uma e várias variáveis, além de conceitos específicos de taxa de crescimento, taxa marginal e elasticidade.

O livro foi estruturado de maneira a abordar o conteúdo programático de forma simples, que facilitará a compreensão dos conceitos.

A Unidade 1 aborda, de forma detalhada, a teoria básica de Números Índices, tanto daqueles de base móvel, como daqueles de base fixa, além das mudanças de base fixa para móvel, de móvel para fixa e de fixa para fixa, como frequentemente é exigido em estudos econômicos. Foram abordados, ainda, os conceitos de deflacionamento, deflator implícito, número de Quantum e as definições dos principais números índices da economia brasileira.

Na Unidade 2 você encontrará os princípios metodológicos que norteiam os modelos econométricos, os principais conceitos de modelos de correlação, os conceitos básicos de modelos de regressão e as diferenças básicas entre os modelos de regressão populacional e amostral.

Já a Unidade 3 encontra-se focada na teoria do modelo clássico de regressão, que utiliza o Método dos Mínimos Quadrados Ordinários (MQO). Apresentaremos os seus princípios de estimação, tomando como base um modelo de regressão linear com duas variáveis (o mais simples de todos) e estabeleceremos as hipóteses inerentes ao MQO, o que o condiciona como o modelo clássico de regressão. Além disso, esta unidade aborda os conceitos de distribuição dos resíduos e dos coeficientes, o que permite estabelecer inferências sobre os parâmetros (coeficientes) e estatísticas do modelo de regressão, possibilitando estabelecer a qualidade desses

modelos, por meio de testes de intervalo e de confiança, e testes de hipótese com nível de significância. Na Unidade 3 você verá, também, a importância de utilizar softwares específicos para econometria, conforme exemplos de aplicações econômicas que serão apresentados no material complementar, resolvidos por meio do código computacional GRETL (*software* econométrico livre).

Finalmente, pode-se dizer que os conceitos básicos abordados neste livro lhe servirão como ferramentas de aplicabilidade para um grande número de problemas econômicos. Contudo, deve ficar claro que a teoria econométrica é muito mais ampla, e que outras disciplinas, que serão cursadas futuramente, lhe permitirão aprofundar ainda mais os conhecimentos sobre a teoria quantitativa aplicada no campo econômico, com a finalidade de constituir uma base profunda para entender o mundo econômico atual. Mas não se esqueça de que você é o principal autor do seu conhecimento. A sua curiosidade e o seu esforço como aluno-pesquisador são aspectos fundamentais para o seu desempenho.

Bons estudos!

Prof. Milton Biage

1

NÚMEROS ÍNDICES

Nesta unidade você verá:

- qual a função dos números índices;
- como deve proceder para construir um número índice e quais as propriedades que eles deverão satisfazer;
- o que são números índices simples de base fixa (relativos de ligação de base fixa) e de base móvel (relativos de ligação de base móvel);
- como efetuar mudanças de base fixa para móvel, mudanças de base móvel para fixa e mudanças de base fixa para base fixa;
- como aplicar os conceitos de números índices agregados simples de preço, de quantidade e de valor (Badstreet-Dutot, Sauerbeck, Média Geométrica e Média Harmônica) na construção de números índices, assim como entender as suas limitações;
- como aplicar os conceitos de números índices agregados ponderados de preço, de quantidade e de valor (Laspeyres, Paache, Divisia, Marchal-Edgeworth, Fish e Drobish) na construção de números índices, assim como entender as suas limitações e as suas vantagens;
- como aplicar os conceitos de números índices agregados ponderados de base móvel, de preço e de quantidade (Laspeyres, Paache, Divisia, Marchal-Edgeworth, Fish e Drobish) na construção de números índices de base móvel, assim como entender as suas limitações e as suas vantagens;
- como estabelecer processos de deflacionamento e determinar taxas reais de variação;
- como estimar, a partir de receitas correntes, as receitas reais, assim como os deflatores implícitos e as taxas de crescimento reais, em uma empresa ou em uma economia agregada;
- como são estimados e quais as finalidades de deflacionamento dos principais indicadores de preço macroeconômicos da economia brasileira.

Olá caro aluno! Antes de iniciarmos os nossos estudos, é necessário chamar a sua atenção para os materiais que estão disponíveis no AVEA. Trata-se de um material complementar, que você deverá imprimir e ter sempre em mãos, pois senão será muito difícil acompanhar o desenvolvimento dos conteúdos do livro! Agora, mãos à obra!

1.1 NÚMEROS ÍNDICES, APLICABILIDADE E CONSTRUÇÃO

Os números índices são medidas estatísticas frequentemente usadas por economistas, administradores e engenheiros, com o objetivo de comparar grupos de variações entre si e obter um quadro simples e resumido das mudanças significativas em áreas relacionadas com preços de matéria prima, preços de produtos acabados, volume físico de produção (FONSECA et al., 1982).

O emprego de números índices permite estabelecer:

- comparações entre variações ocorridas ao longo do tempo, de um produto ou de uma cesta de produtos;
- diferenças entre comportamentos de variáveis em lugares diferentes;
- diferenças entre comportamento de produtos ou categorias de organizações semelhantes, etc.

Os **números índices** são usados para indicar variações relativas em quantidades, preços ou valores de um artigo (ou artigos) durante certo período de tempo. Eles sintetizam as modificações nas condições econômicas ocorridas em um espaço de tempo, através de uma **razão**. Se apenas **um item** (produto) é computado, trata-se de um **número índice simples** (sem agregação). Porém, se **vários itens** (produtos) têm suas variações computadas conjuntamente, tem-se um **número índice composto**.

Os números índices constituem indicativos de mudanças, como quando, por exemplo, a moeda sofre uma desvalorização, ou quando o processo de desenvolvimento econômico acarreta mudanças contínuas nos hábitos dos consumidores, provocando com isso modificações qualitativas e quantitativas na composição da produção nacional e, em consequência, na produção

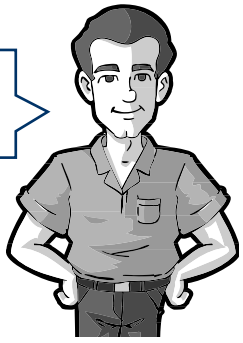
de cada empresa individualmente. A utilização de números índices se torna indispensável quando o fator monetário se encontra presente, em qualquer análise, quer no âmbito interno de uma empresa ou mesmo fora dela, sob pena de o analista ser conduzido a conclusões totalmente falsas e prejudiciais.

Cada número índice de uma série costuma vir expresso em termos percentuais. Os índices mais empregados medem, em geral, variações ao longo do tempo, e é exatamente neste sentido que iremos tratá-los.

Ao construir um número índice, deve-se considerar alguns fatores importantes, como veremos a seguir.

- **Seleção dos dados.** Quando se pretende medir a variação nos custos com educação, deve-se tomar em consideração somente as variáveis que afetam diretamente o custo da educação. Por exemplo, não se deve considerar itens como vinagre ou canela, ou seja, itens extremamente à parte da composição desses custos.
- **A eleição do período base.** Ao eleger o período base, deve-se considerar que naquele período houve uma estabilidade relativa. Por exemplo: quando da determinação de séries mensais de preço, não se deve eleger os meses de Dezembro ou Janeiro como meses base, pois nesses meses, normalmente, ocorrem significativas alterações de preços ou de quantidades, especialmente em serviços públicos. Deve-se ainda considerar que esse período tomado como base deve ser recente, pois, se estiver distante, isso resultará em uma não uniformidade na composição dos dados, tanto em quantidade como em preço.
- **A importância relativa das variáveis (ou itens) no conjunto de elementos que compõem a variável em análise.** Deve-se atribuir a cada variável sua importância relativa real dentro do conjunto, já que nenhum item apresenta o mesmo efeito sobre o preço total do conjunto. Por exemplo: uma alta ou baixa no preço do vinagre, ou uma alta ou baixa no preço do leite, deve influenciar diferentemente a composição nos custos de alimentação.

Caro aluno, chegou o momento de consultar o Exemplo 1 do material complementar que você imprimiu! Leia o exemplo, analise e depois continue a sua leitura da Seção 1.2.



1.2 CRITÉRIOS DE AVALIAÇÃO DA FÓRMULA DE UM ÍNDICE (PROPRIEDADES DOS RELATIVOS)

Não existe um número índice considerado como perfeito, ou uma fórmula definitiva para quantificar, de modo inequívoco e exato, as variações de preço e quantidades, especialmente quando os índices se referem não a um, mas a um conjunto de bens.

Existe uma variedade de métodos de cálculo de número índices. A escolha do número índice será facilitada se houver algum critério que possibilite salientar as vantagens e as limitações de cada um deles. Irving Fisher (1922) desenvolveu alguns testes ou critérios matemáticos muito úteis para comparar as várias fórmulas propostas de números índices.

Palavra do Professor

Deve ficar claro que os testes (ou propriedades desejáveis) propostos podem ser aplicados a qualquer número índice.

Portanto, os números índices definidos a partir de uma forma geral devem cumprir algumas propriedades gerais, principalmente os números índices simples, pois, alguns números índices compostos não cumprem, de forma simultânea, todas as propriedades gerais de Fisher, apresentadas no Quadro 1.1 abaixo.

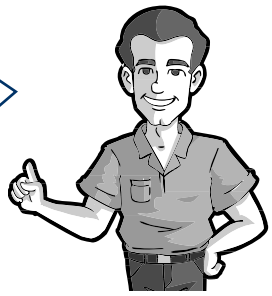
PROPRIEDADE	DESCRIÇÃO
De existência	O índice deve constituir-se num valor real e finito: $I_{0,t} \neq 0$ *.
De identidade	Se coincidir o período base e o período atual do número índice, então, $I_{0,0} = 100\%$.
De inversão	O produto de dois índices invertidos é igual a 1. Ou seja, $I_{0,t} \times I_{t,0} = 1 \Rightarrow I_{t,0} = \frac{1}{I_{0,t}}$.
Circular	Representa a generalização para vários períodos. Ou seja, $I_{0,t'} \times I_{t',t} \times I_{t,0} = 1 \rightarrow I_{0,t'} \times I_{t',t} = \frac{1}{I_{t,0}} \rightarrow I_{0,t'} \times I_{t',t} = I_{0,t}$

De proporcionalidade	Se a magnitude da variável, X_t, aumenta numa proporção K (ou seja, $X_t = X_{t'} + K \times X_{t'}$), então, o número índice, $I_{0,t}$, aumenta proporcionalmente à variável, tal como: $I_{0,t} = I_{0,t'} + K \times I_{0,t'}$.

Quadro 1.1 – Propriedades dos números índices de Fisher (1967).

* O primeiro subíndice representa o período base; e o segundo, o período atual do número índice.

Caro aluno, agora você deve consultar o Exemplo 2 do seu material complementar! Leia o exemplo, analise e depois continue a sua leitura da Seção 1.3.



1.3 NÚMEROS ÍNDICES SIMPLES

Os números índices simples podem ser chamados (assim como os compostos) de **relativos de base fixa** ou **relativos de ligação** (ou relativos de base móvel). Esses números índices são empregados de forma intensa no mundo empresarial, a fim de estudar o comportamento das produções e de vendas dos produtos fabricados.

São esses índices simples que são utilizados para determinar as **variações de preço**, **valor** ou **quantidade** de sacas de café exportadas por ano e de barris de petróleo produzidos e exportados pelo mundo árabe, por exemplo.

Os números índices simples podem ser:

- **número índice de preço:** quando se calcula a razão entre o preço observado de um artigo em um período qualquer e o preço do mesmo artigo no período base;
- **número índice de quantidade:** quando se calcula a razão entre a quantidade observada de um artigo em um período qualquer e a quantidade no período base; e
- **número índice de valor:** quando se calcula a razão entre o produto do preço pela quantidade de um artigo em um período qualquer e o produto do preço pela quantidade do mesmo artigo no período base.

Vejamos as definições acima expressas na forma de equações, no Quadro 1.2 a seguir.

PREÇO	QUANTIDADE	VALOR
$p_{0,t} = \frac{p_t}{p_0} \times 100$	$q_{0,t} = \frac{q_t}{q_0} \times 100$	$v_{0,t} = \frac{p_t \times q_t}{p_0 \times q_0} \times 100$
(1)	**(2)**	**(3)**

Quadro 1.2 – Fórmulas dos índices simples de preço, de quantidade e de valor.

Onde:

p_0 = o preço do artigo no período base

p_t = o preço do artigo em um período qualquer

q_0 = a quantidade do artigo no período base

q_t = a quantidade do artigo em um período qualquer

Outros fatores a serem observados nos números índices simples são com relação à base do número índice. Uma maneira de calcular um número índice é mantendo a base fixa. Nesse caso, a base de referência mantém-se a mesma para todos os cálculos relativos, e a comparação é feita com relação ao período base. Esse procedimento é chamado de **número índice de base fixa** (ou **relativos de base fixa**).

Outra forma de calcular os índices consiste em variar a base de um período para outro. Quando esse for o caso, diz-se que o número índice é de base móvel (variável) ou chamamos **Relativos de Base Móvel**. Nesse caso, a comparação é feita com relação ao período anterior ao do número índice.

Palavra do Professor

Caro aluno, chegou o momento de consultar o Exemplo 3 do seu material complementar! Leia e analise o exemplo. Em seguida, recomendo que você resolva a Questão 1 das Atividades de aprendizagem. Isso facilitará muito o seu trabalho. Só depois, continue a sua leitura da Seção 1.4.

1.4 CRITÉRIO DE DECOMPOSIÇÃO DAS CAUSAS (OU INVERSÃO DOS FATORES)

O critério de decomposição das causas sustenta que o produto de um número índice de preço pelo correspondente número índice de quantidade deve ser igual ao valor total relativo, ou ao índice de valor. Portanto, supondo que:

$p_{0,t}$ = índice de preços

$q_{0,t}$ = índice de quantidade

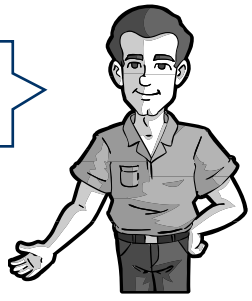
$v_{0,t}$ = índice de valor

Então

$$v_{0,t} = p_{0,t} \times q_{0,t} \qquad (4)$$

Cabe aqui uma observação: todos os números simples satisfazem o critério de decomposição das causas; contudo, poucos números índices agregados satisfazem esta propriedade.

Caro aluno, veja agora o Exemplo 4 do seu material complementar! Leia e analise o exemplo. Em seguida, continue a sua leitura da Seção 1.5.



1.5 NÚMEROS ÍNDICES DE BASE MÓVEL (OU RELATIVOS DE LIGAÇÃO) E NÚMEROS ÍNDICES DE BASE FIXA

Provavelmente, devido à cultura inflacionária existente no Brasil, não costumamos encontrar índices em valores absolutos (de base fixa), tais como os calculados na questão **(a)** do Exemplo 3 (material complementar). Por outro lado, é bastante comum nos depararmos com os Números Índices de Base Móvel (ou Relativos de Ligação), que sintetizam as variações econômicas entre dois períodos consecutivos (variação percentual em relação ao mês imediatamente anterior).

Portanto, para determinar um índice de base fixa (ou relativo de base fixa) a partir de um número índice de base móvel, basta aplicar as propriedades **circular** e de **inversão**. Ou seja:

$$I_{0,t} \times I_{t,t'} \times I_{t',0} = 1 \rightarrow I_{0,t} \times I_{t,t'} = \frac{1}{I_{t',0}} \qquad \textbf{(5)}$$

Portanto, aplicando em (5) a **propriedade de inversão**, obtém-se:

$$I_{0,t'} = I_{0,t} \times I_{t,t'} \qquad \textbf{(6)}$$

Assim, os subíndices em (6) são definidos, tal que:

t' = período atual

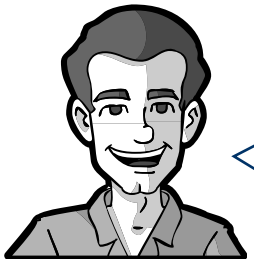
t = período anterior ao atual

0 = período base a ser fixado como base fixa

Portanto, movendo t' e t em cada estimativa, obtém-se a **transformação dos números índices de base móvel para base fixa.**

Em resumo, para obter os números índices de base fixa (ou relativos de base fixa) de um período a partir de um número índice de base móvel, basta multiplicar o índice de base fixa do período anterior pelo índice de base móvel atual, nas formas fracionárias. Para obter resultados na forma percentual, basta multiplicar o resultado por 100. Ou seja:

$$I_{0,t'} = I_{0,t} \times I_{t,t'} \times 100$$



Caro aluno, chegou o momento de consultar o Exemplo 5 do material complementar! Leia o exemplo e analise-o. Depois, continue a leitura desta seção.

Por outro lado, para determinar um índice de base móvel (ou relativo de ligação) a partir de um número índice de base fixa, também basta aplicar as **propriedades circular** e **de inversão**. Ou seja:

$$I_{0,t} \times I_{t,t'} \times I_{t',0} = 1 \rightarrow I_{0,t} \times I_{t,t'} = \frac{1}{I_{t',0}} \rightarrow I_{0,t} \times I_{t,t'} = I_{0,t'} \rightarrow I_{t,t'} = \frac{I_{0,t'}}{I_{0,t}} \qquad \textbf{(7)}$$

Na relação (7) os resultados são dados na forma fracionária, entretanto, ela pode ser representada na forma percentual, como segue:

$$I_{0,t} \times I_{t,t'} = I_{0,t'} \rightarrow \ I_{t,t'} = \frac{I_{0,t'}}{I_{0,t}} \times 100 \qquad (8)$$

Assim, os subíndices em (7) ou em (8) são definidos, tal que:

t' = período atual

t = período anterior ao atual

0 = período base no número-índice de base fixa

Portanto, movendo t' e t em cada estimativa, obtém-se a **transformação dos números índices de base fixa para base móvel**.

Em resumo, para obter os números índices de base móvel (ou relativos de ligação) de um período, a partir de um número índice de base fixa, basta dividir o índice do período de interesse pelo índice do período imediatamente anterior (os resultados serão na forma fracionária, conforme vimos em (7). Mas, se desejar os resultados na forma percentual, conforme aparece em (8), o resultado deve ser multiplicado por 100.

Caro aluno, chegou o momento de vermos mais alguns exemplos! Primeiro, leia e analise o Exemplo 6 do seu material complementar. Em seguida, partiremos para a análise do Exemplo 7; afinal, geralmente, conhecemos apenas as variações de um índice e não o próprio índice. Neste caso, podemos facilmente criar o índice e trabalhar com ele normalmente, da forma que será mostrado neste exemplo. O terceiro passo será responder à Questão 2 das Atividades de aprendizagem. Bom trabalho!

1.6 MUDANÇA DE BASE DE UM NÚMERO ÍNDICE DE BASE FIXA

A escolha da base de um número índice é muitas vezes uma tarefa difícil. É preciso escolher um **período relativamente estável**, o mais "típico" possível, quando a atividade econômica não estiver sendo afetada por variações estruturais ocasionais.

No Brasil, apesar da estabilidade atual, a economia parece estar sendo sempre sacudida, em maior ou menor grau, por flutuações e crises. Assim, a escolha da base torna-se ainda mais controvertida e, talvez por isso, haja tanta predileção pelos índices relativos de ligação.

De qualquer forma, independentemente do índice, pode ser interessante ou necessário mudar a base de um número índice por duas razões:

1. **Atualizar a base, tornando-a mais próxima da realidade atual** (por este motivo, periodicamente, o IBGE realiza Pesquisas de Orçamento Familiar (POF) com a finalidade de incluir as mudanças nos hábitos de consumo nas ponderações dos seus índices);

Links

Você pode consultar as Pesquisas de Orçamento Familiar (POF) no site do Instituto Brasileiro de Geografia e Estatística (IBGE), através do *link*: http://www.ibge.gov.br/home/mapa_site/mapa_site.php#populacao

2. **Para permitir a comparação de duas séries de índices que tenham bases diferentes**.

O procedimento é extremamente simples, pois basta dividir toda a série de números índices originais pelo número índice do período escolhido como nova base. Isso preservará as diferenças relativas entre eles. Matematicamente, para obter o valor de um novo índice numa nova base, basta aplicar as propriedades circular e de inversão de índices, como segue:

$$I_{0,t'} \times I_{t',t} \times I_{t,0} = 1 \Rightarrow I_{0,t'} \times I_{t',t} = \frac{1}{I_{t,0}} = I_{0,t} \Rightarrow I_{t',t} = \frac{I_{0,t}}{I_{0,t'}} \times 100 \quad \textbf{(9)}$$

É importante notar que em **(9),** o primeiro subíndice indica o período base e o segundo subíndice indica o período atual do número índice.

Caro aluno, chegou o momento de consultar o Exemplo 8 do seu material complementar! Leia e analise-o. Em seguida, recomendo que você resolva as Questões 3, 4 e 5 das Atividades de aprendizagem. Isso facilitará muito o seu trabalho. Só depois, continue a sua leitura da Seção 1.7.

1.7 NÚMEROS ÍNDICES COMPOSTOS (OU AGREGADOS)

Os números índices compostos expressam variações no preço, quantidade ou valor de um grupo de itens. São chamados de **agregados simples** quando atribuem a mesma ponderação para todos os itens, desconsiderando a importância relativa de cada um. Já os índices **agregados ponderados** atribuem ponderações diferentes para os itens, o que pode permitir dar maior ênfase às variações em determinado item, característica que faz desta a forma mais utilizada.

1.7.1 Índice agregado simples (de preços e de quantidades)

> Os índices agregados simples circunscrevem as comparações entre preços, quantidades ou valores de um único item.

Esses relativos, os índices agregados simples, são úteis para a compreensão de conceitos básicos e para a averiguação das propriedades de avaliação de um número índice. Entretanto, a quase totalidade dos números índices envolve avaliações simultâneas de variações de preço e quantidade para vários itens, o que denominamos uma cesta de produtos. Este é o caso dos **números índices agregados simples** ou **ponderados**.

Portanto, para avaliar a variação de preço ou quantidade entre dois períodos de uma cesta de produtos, sem levar em consideração a importância relativa de cada item que compõe a cesta, utilizamos proposições como: *média aritmética simples*, *média geométrica, média harmônica* e *índices agregados simples.*

Os índices agregados simples mais utilizados são:

- o **índice da média agregada simples** (ou **índice de Bradstreet-Dutot**), que é a razão entre as médias aritmética simples dos *n* preços (ou quantidades) no período *t* e os *n* preços (ou quantidades) no período 0, tomado como base;
- o **índice de Sauerbeck**, que é constituído pela média aritmética dos índices de preços simples (ou de quantidade) para cada item (ou seja, é simplesmente o índice de média aritmética simples);
- o **índice média geométrica**, que é constituído pela média geométrica dos índices de preços simples (ou de quantidade) para cada item; e

- o **índice média harmônica**, que é constituído pela média harmônica dos índices de preços simples (ou de quantidade) para cada item.

Índice de preço de Badstreet-Dutot

Esse índice é expresso pela relação, em porcentagem, entre a somatória dos preços dos *n* artigos num período *t* e a somatória dos preços dos mesmos *n* artigos em um período 0, tomado como base. Assim:

$$BD_{0,t}^{p} = \frac{\frac{1}{n}\sum_{i=1}^{n}(p_t)^i}{\frac{1}{n}\sum_{i=1}^{k}(p_0)^i} = \left(\frac{\sum_{i=1}^{n}(p_t)^i}{\sum_{i=1}^{k}(p_0)^i}\right) \times 100 \qquad \textbf{(10)}$$

Onde:

$BD_{0,t}^{p}$= índice de preço agregado simples de Badstreet-Dutot para o período *t*

$(p_t)^i$ = preço do item *i* no período *t*

$(p_0)^i$ = preço do item *i* para o período base

Índice de quantidade de Badstreet-Dutot

Esse índice é expresso pela relação, em porcentagem, entre a somatória das quantidades dos *n* artigos num período *t* e a somatória das quantidades dos mesmos *n* artigos em um período 0, tomado como base. Assim:

$$BD_{0,t}^{q} = \frac{\frac{1}{n}\sum_{i=1}^{n}(q_t)^i}{\frac{1}{n}\sum_{i=1}^{k}(q_0)^i} = \left(\frac{\sum_{i=1}^{n}(q_t)^i}{\sum_{i=1}^{k}(q_0)^i}\right) \times 100 \qquad \textbf{(11)}$$

Onde:

$BD_{0,t}^{q}$ = índice de quantidade agregado simples de Badstreet-Dutot para o período *t*

$(q_t)^i$ = quantidade do item *i* no período *t*

$(q_0)^i$ = quantidade do item *i* para o período base

Índice de preço de Sauerbeck

Esse índice é expresso pela média aritmética dos *n* índices de preço simples correspondentes a cada item, ou seja:

$$S_{0,t}^{p} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{p_t}{p_0}\right)^{i} \times 100 \qquad \textbf{(12)}$$

Onde:

$S_{0,t}^{p}$ = índice de preço média aritmética simples para o período *t*

$(p_t)^{i}$ = preço do item *i* no período *t*

$(p_0)^{i}$ = preço para o item *i* no período base

Índice de quantidade de Sauerbeck

Esse índice é expresso pela média aritmética dos *n* índices de quantidade simples correspondentes a cada item, ou seja:

$$S_{0,t}^{q} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{q_t}{q_0}\right)^{i} \times 100 \qquad \textbf{(13)}$$

Onde:

$S_{0,t}^{q}$ = índice de quantidade média aritmética simples para o período *t*

$(q_t)^{i}$ = quantidade do item *i* no período *t*

$(q_0)^{i}$ = quantidade do item *i* para o período base

Índice de preço média geométrica

Esse índice é expresso pela média geométrica dos *n* índices de preços correspondentes a cada item, ou seja:

$$MG_{0,t}^{p} = \sqrt[n]{\prod_{i=1}^{n}\left(\frac{p_t}{p_0}\right)^{i}} \qquad \textbf{(14)}$$

Onde:

$MA_{0,t}^{p}$ = índice de preço média aritmética simples para o período t

$p_{t,i}$ = preço do item i no período t

$p_{0,i}$ = preço do item i para o período base

Índice de quantidade média geométrica

Esse índice é expresso pela média geométrica dos n índices de quantidade correspondentes a cada item, ou seja:

$$MG_{0,t}^{q} = \sqrt[n]{\prod_{i=1}^{n}\left(\frac{q_t}{q_0}\right)^{i}} \qquad \textbf{(15)}$$

Onde:

$MA_{0,t}^{q}$ = índice de quantidade média geométrica para o período t

$(q_t)^{i}$ = quantidade do item i no período t

$(q_0)^{i}$ = quantidade do item i no período base

Índice de preço da média harmônica

Esse índice é expresso pela média harmônica dos n índices de preço simples correspondentes a cada item, ou seja:

$$MH_{0,t}^{p} = \frac{n}{\sum_{i=1}^{n}\left(\frac{p_0}{p_t}\right)^{i}} \times 100 \qquad \textbf{(16)}$$

Onde:

$MH_{0,t}^{p}$ = índice de preço média harmônica para o período t

$(p_t)^{i}$ = preço do item i no período t

$(p_0)^{i}$ = preço para o item i no período base

ÍNDICE DE QUANTIDADE DA MÉDIA HARMÔNICA

Esse índice é expresso pela média harmônica dos n índices de quantidades simples correspondentes a cada item, ou seja:

$$MH_{0,t}^{q} = \frac{n}{\sum_{i=1}^{n}\left(\frac{q_0}{q_t}\right)^{i}} \times 100 \qquad \textbf{(17)}$$

Onde:

$MH_{0,t}^{q}$ = índice de quantidade média aritmética harmônica para o período t

$(q_t)^i$ = quantidade do item i no período t

$(q_0)^i$ = quantidade do item i no período base

Agora chegou a hora de você consultar o Exemplo 9 do material complementar! Leia e analise o exemplo. Em seguida, resolva a Questão 6 das Atividades de aprendizagem. Depois, continue a sua leitura da Subseção 1.7.2.

1.7.2 ÍNDICES AGREGADOS PONDERADOS

Os índices agregados ponderados são determinados quando se pretende estabelecer a importância relativa de cada item (produto) que compõe o conjunto (ou cesta) de base de cálculo.

Esta importância relativa é expressa pela quantidade monetária gasta durante um período com cada item do conjunto. Obtemos a quantidade monetária gasta com cada produto multiplicando o preço do item pela sua quantidade consumida no período *t*.

Os índices agregados ponderados mais utilizados são:

- **Índice de Valor** (ou Índice de preço ponderado por quantidade consumida): a ponderação é feita em função do produto do preço de cada item pela sua respectiva quantidade, em um determinado período 0 (zero), tomado como base. Pode ser calculado somente o índice de preço.

- **Índice de Laspeyres** (época básica): a ponderação é feita em função dos preços ou quantidades do período base. Podem ser calculados índices de preço e quantidade. Em particular, o índice de Laspeyres não cumpre a propriedade de inversão e, tampouco, a propriedade circular. Contudo, assume-se que esse índice as cumpre.
- **Índice de Paasche** (época atual): a ponderação é feita em função dos preços ou quantidades do período atual. Esse índice também não cumpre as propriedades de inversão e circular. Podem ser calculados índices de preço e quantidade.
- **Índice de Divisia:** é uma média geométrica ponderada dos relativos, com sistema de pesos fixos na época da base. A principal vantagem desse número índice reside no fato de ele ser o único a satisfazer as propriedades circular e de inversão.
- **Índice de Marshall-Edgeworth:** a ponderação é feita pela soma das bases de ponderação do índice de Laspeyres e do índice de Paasche. Como esse índice envolve os índices citados, ele também não cumpre as propriedades de inversão e circular. Mas também se assume que esse índice as cumpre.
- **Índice de Fisher:** esse índice é definido como a média geométrica entre os índices de Laspeyres e de Paasche. Como esse índice envolve os índices citados, ele também não cumpre as propriedades de inversão e circular. Também se assume que esse índice as cumpre.
- **Índice de Drobish:** esse índice é definido como a média aritmética entre os índices de Laspeyres e de Paasche. Como esse índice envolve os índices citados, ele também não cumpre as propriedades de inversão e circular. Mas também se assume que esse índice as cumpre.

Índice de Laspeyres

Os índices agregados ponderados de preço e quantidade de Laspeyres são definidos, de forma similar a uma média aritmética ponderada, como vemos a seguir:

$$L_{0,t}^{p} = \frac{\sum_{i=1}^{k}\left(\frac{p_t}{p_0}\right)^{i} \times w_i}{\sum_{i=1}^{k} w_i} \times 100 \rightarrow L_{0,t}^{p} = \left(\sum_{i=1}^{k}\left(\frac{p_t}{p_0}\right)^{i} \times w_i\right) \times 100 \qquad \textbf{(18)}$$

Onde:

w_i = base de ponderação

$$\sum_{i=1}^{k} w_i = 1$$

E, para o índice agregado ponderado de quantidade:

$$L_{0,t}^{q} = \frac{\sum_{i=1}^{k}\left(\frac{q_t}{q_0}\right)^{i} \times w_i}{\sum_{i=1}^{k} w_i} \times 100 \rightarrow \; L_{0,t}^{q} = \left(\sum_{i=1}^{k}\left(\frac{q_t}{q_0}\right)^{i} \times w_i\right) \times 100 \qquad \textbf{(19)}$$

Onde:

w_i = base de ponderação

$$\sum_{i=1}^{k} w_i = 1$$

k = número de itens

$p_{t,i}$ = preço do item *i* no período atual

$p_{0,i}$ = preço do item *i* no período base

$q_{t,i}$ = quantidade do item *i* no período atual

$q_{0,i}$ = quantidade do item *i* no período base

No índice de Laspeyres, a ponderação é feita em função dos preços e quantidades do período base. Assim, ele tem a **vantagem** de que as ponderações para todos os períodos se mantêm fixas, mas tem a **desvantagem** de que a representatividade do efeito de ponderação diminui quando o período de cálculo do índice se distancia do período base. Por causa disso, ele tende a exagerar a alta, pois considera as quantidades (ou preços) como sendo sempre os mesmos do período base.

O índice de Laspeyres, tanto de preço quanto de quantidade, é mais utilizado nos indicadores gerais de preços e produção. O seu projeto e elaboração exigem uma rigorosa seleção de seus componentes e das ponderações de cada componente no conjunto. Afinal, na medida em que o cálculo do índice se distancia do período base, torna-se necessário fixar novo período base e estabelecer uma nova estrutura de ponderações de cada item que compõe o conjunto, em função do fato de que o índice de Laspeyres diminui a sua significância uma vez que a base se distancia do período atual, que determinará o índice.

Para o número Índice de Laspeyres utiliza-se uma base de ponderação definida em função do período base, como segue:

$$w_i = \frac{(p_0 \times q_0)^i}{\sum_{i=1}^{k}(p_0 \times q_0)^i} \qquad \textbf{(20)}$$

Então, aplicando (20) em (18), obtém-se uma outra forma para estimar o Índice de Preço de Laspeyres:

$$L_{0,t}^{p} = \left[\sum_{i=1}^{k}\left(\left(\frac{p_t}{p_0}\right)^i \times \frac{(p_0 \times q_0)^i}{\sum_{i=1}^{k}(p_0 \times q_0)^i}\right)\right] \times 100 \rightarrow L_{0,t}^{p} = \frac{\sum_{i=1}^{k}(p_t \times q_0)^i}{\sum_{i=1}^{k}(p_0 \times q_0)^i} \times 100 \qquad \textbf{(21)}$$

Da mesma forma, substituindo (18) em (19), obtém-se o Índice de Quantidade de Laspeyres:

$$L_{0,t}^{q} = \left[\sum_{i=1}^{k}\left(\left(\frac{q_t}{q_0}\right)^i \times \frac{(p_0 \times q_0)^i}{\sum_{i=1}^{k}(p_0 \times q_0)^i}\right)\right] \times 100 \rightarrow L_{0,t}^{q} = \frac{\sum_{i=1}^{k}(q_t \times p_0)^i}{\sum_{i=1}^{k}(p_0 \times q_0)^i} \times 100 \qquad \textbf{(22)}$$

Para se estimar os números índices de Laspeyres de preço, poderão ser utilizadas conjuntamente as fórmulas (18) e (20) ou somente a fórmula (21). Da mesma forma, para se estimar os números índices de Laspeyres de quantidade poderão ser utilizadas conjuntamente as fórmulas (19) e (20) ou somente a fórmula (22).

Palavra do Professor

Caro aluno, observe a seguir as vantagens e desvantagens da formulação de Laspeyres.

Vantagens da formulação de Laspeyres:

- o número índice de Laspeyres é mais fácil de construir do que qualquer outro número índice ponderado por quantidades;
- os índices de Laspeyres de preço ou quantidade apresentam evoluções uniformes, sem grandes instabilidades, pois **as quantidades permanecem constantes** de um período para outro (quando do cálculo do índice de preço) e os preços permanecem constantes quando do cálculo do índice de quantidade, o que permite observar somente o efeito das mudanças de preço (ou de quantidade).

Desvantagem da formulação de Laspeyres:

- O índice de Laspeyres, ao ponderar os preços dos artigos *i* no período *t*, por quantidades consumidas no período base, quando da determinação do índice de preços, tende a dar maior importância relativa dentro do conjunto aos itens que tiveram os seus preços alterados mais significativamente, já que as quantidades consumidas estão sujeitas à lei da oferta e da demanda, induzindo ao fato de que quando o preço sobe, **as quantidades consumidas tendem a diminuir.** O mesmo raciocínio de análise é aplicado quando da determinação do índice de Laspeyres de quantidade.

Índice de Paasche

No índice de Paasche a ponderação é feita com o valor das transações, determinadas em função dos preços e quantidades do período atual. Esse índice tem a vantagem de que os pesos relativos dos distintos itens atualizam-se para cada período. Contudo, o seu agravante é que ele apresenta maior complexidade e maiores custos nas suas determinações (a mudança constante da base atual pode encarecer a pesquisa necessária para identificar os pesos).

Além disso, ele tende a exagerar as quedas, por considerar como base as quantidades (ou preços) iguais aos do período atual. Os índices agregados ponderados de preço e quantidade de Paasche são definidos por meio de uma média harmônica ponderada, como segue:

$$PA_{0,t}^{p} = \frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{k}\left(\frac{p_0}{p_t}\right)^{i} \times w_i} \times 100 \rightarrow PA_{0,t}^{p} = \frac{1}{\left(\sum_{i=1}^{k}\left(\frac{p_0}{p_t}\right)^{i} \times w_i\right)} \times 100 \quad \textbf{(23)}$$

Onde:

w_i = base de ponderação

$\sum_{i=1}^{k} w_i = 1$

E, para o índice agregado ponderado de quantidade tem-se:

$$PA_{0,t}^{q} = \frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{k}\left(\frac{q_0}{q_t}\right)^{i} \times w_i} \times 100 \rightarrow PA_{0,t}^{q} = \frac{1}{\left(\sum_{i=1}^{k}\left(\frac{q_0}{q_t}\right)^{i} \times w_i\right)} \times 100 \quad \textbf{(24)}$$

Onde:

w_i = base de ponderação

$\sum_{i=1}^{k} w_i = 1$

k = número de itens

$(p_t)^i$ = preço do item *i* no período atual

$(p_0)^i$ = preço do item *i* no período base

$(q_t)^i$ = quantidade do item *i* no período atual

$(q_0)^i$ = quantidade do item *i* no período base

A função de ponderação para o índice agregado de preço e para o índice agregado de quantidade é a seguinte:

$$w_i = \frac{(p_t \times q_t)^i}{\sum_{i=1}^{k}(p_t \times q_t)^i} \qquad \textbf{(25)}$$

Portanto, substituindo a equação (25) nas equações (23) e (24) temos, respectivamente, as seguintes fórmulas:

$$PA_{0,t}^{p} = \frac{1}{\left(\sum_{i=1}^{k}\left(\frac{p_0}{p_t}\right)^i \times w_i\right)} \times 100$$

$$PA_{0,t}^{p} = \frac{1}{\left[\sum_{i=1}^{k}\left(\left(\frac{p_0}{p_t}\right)^i \times \frac{(p_t \times q_t)^i}{\sum_{i=1}^{k}(p_t \times q_t)^i}\right)\right]} \times 100 \rightarrow$$

$$PA_{0,t}^{p} = \frac{1}{\frac{\sum_{i=1}^{k}(p_0 \times q_t)^i}{\sum_{i=1}^{k}(p_t \times q_t)^i}} \times 100 \rightarrow PA_{0,t}^{p} = \frac{\sum_{i=1}^{k}(p_t \times q_t)^i}{\sum_{i=1}^{k}(p_0 \times q_t)^i} \times 100 \qquad \textbf{(26)}$$

e

$$PA_{0,t}^{q} = \frac{1}{\left(\sum_{i=1}^{k}\left(\frac{q_0}{q_t}\right)^i \times w_i\right)} \times 100$$

$$PA_{0,t}^{q} = \frac{1}{\left[\sum_{i=1}^{k}\left(\left(\frac{q_0}{q_t}\right)^{i} \times \frac{(p_t \times q_t)^{i}}{\sum_{i=1}^{k}(p_t \times q_t)^{i}}\right)\right]} \times 100 \rightarrow$$

$$PA_{0,t}^{q} = \frac{1}{\frac{\sum_{i=1}^{k}(p_t \times q_0)^{i}}{\sum_{i=1}^{k}(p_t \times q_t)^{i}}} \times 100 \rightarrow \quad PA_{0,t}^{q} = \frac{\sum_{i=1}^{k}(p_t \times q_t)^{i}}{\sum_{i=1}^{k}(p_t \times q_0)^{i}} \times 100 \qquad \textbf{(27)}$$

Para se estimarem os números índices de preço de Paasche poderão ser utilizadas conjuntamente as fórmulas (23) e (25) ou somente a fórmula (26). Da mesma maneira, para se estimarem os números índices de Paasche de quantidade poderão ser utilizadas conjuntamente as fórmulas (24) e (25) ou somente a fórmula (27).

Palavra do Professor

Caro aluno, observe a seguir as vantagens e desvantagens da formulação de Paasche.

Vantagem da formulação de Paasche:

- Mede de forma combinada as mudanças nos preços e nos padrões de consumo.

Desvantagens da formulação de Paasche:

- Não apresenta muita uniformidade na evolução de preços, porque as quantidades de base são diferentes de um período para outro, o que torna impossível atribuir diferenças entre dois períodos somente em função das mudanças de preços.
- O índice de Paasche tende a diminuir as importâncias relativas dos itens que subiram de preço mais intensamente, pois, como se observou anteriormente, as quantidades consumidas estão sujeitas às leis da oferta e da demanda.

ÍNDICE DE PREÇO E DE QUANTIDADE DE DIVISIA

O método proposto por Divisia é uma média geométrica ponderada dos relativos, com sistema de pesos fixos na época da base. A principal **vantagem** desse número índice reside no fato de ele ser o único a satisfazer as propriedades circular e de inversão. Neste caso, bases de cálculos móveis podem ser construídas com precisão, a partir de estimativas desse número com base de ponderação fixa.

Como **desvantagens** desse número índice, podemos apontar o fato de ele não satisfazer a propriedade de decomposição das causas, e o fato de que o caráter essencialmente variável dos pesos não pode ser captado pela formulação proposta.

Portanto, os Índices de Divisia de preço e quantidade utilizam médias geométricas ponderadas, definidas, respectivamente, como podemos ver a seguir:

$$D_{0,t}^{p} = \prod_{i=1}^{n}\left(\frac{P_t^i}{P_0^i}\right)^{w_0^i} \quad \textbf{e} \quad D_{0,t}^{q} = \prod_{i=1}^{n}\left(\frac{q_t^i}{q_0^i}\right)^{w_0^i} \qquad \textbf{(28) e (29)}$$

com

$$w_0^i = \frac{p_0^i \times q_0^i}{\sum_{i=1}^{n} p_0^i \times q_0^i} \quad \text{tal que} \quad \sum_{i=1}^{n} w_0^i = 1 \qquad \textbf{(30)}$$

ÍNDICE DE VALOR (OU ÍNDICE DE PREÇO PONDERADO POR QUANTIDADE CONSUMIDA)

Supondo que existam informações sobre preços e quantidades dos k produtos e serviços que integram a base de dados, é evidente que podemos obter os números índices de valor. O Índice de Valor é determinado pela relação expressa em porcentagem, dada abaixo:

$$IV_{0,t} = \frac{\sum_{i=1}^{k}(p_t \times q_t)^i}{\sum_{i=1}^{k}(p_0 \times q_0)^i} \times 100 \qquad \textbf{(31)}$$

Onde:

$IV_{0,t}$ = Índice de Valor

k = número de itens

$(p_t)^i$ = preço do item *i* no período atual

$(p_0)^i$ = preço do item *i* no período base

$(q_t)^i$ = quantidade do item *i* no período atual

$(q_0)^i$ = quantidade do item *i* no período base

A interpretação desse índice é a de que ele mede a variação percentual dos valores de um conjunto de artigo, de um período para outro.

Vantagem da formulação do Índice de Valor:

- Mede de forma combinada as mudanças nos preços e nos padrões de consumo.

Desvantagens da formulação do Índice de Valor:

- Falta de uniformidade na evolução dos preços, devido ao fato de ele caracterizar mudanças ocorridas tanto em quantidade como em preços, de período para período.
- Dificuldades para a construção do índice, já que, em cada período, necessita-se de informações diferentes, tanto com relação às quantidades como com relação aos preços.

Considerações Gerais

- Deve ficar evidente que só se pode determinar o **Índice de Valor** quando se tem informações, ao longo de toda a série, sobre preços e quantidades (hábitos de consumo) dos itens que compõem a cesta de produtos, obtendo, neste caso, o valor agregado da cesta.
- Caso não se tenha informações atuais sobre preços e quantidades dos itens que compõem a cesta de produtos, deve-se estimar o **Índice de Laspeyres de preço (ou de Divisia)**, utilizando hábitos de consumo num período base e considerando que estes não se alteram. Nesse caso, o número de informações atuais necessárias para construir o índice é reduzido, tornando-o menos oneroso em termos de custo e tempo para a sua construção.

- A mesma consideração acima deve ser feita quando se pretende construir um número **Índice de Laspeyres de quantidade (ou de Divisia)**. Neste caso, as informações sobre quantidades (hábitos de consumo) devem ser atuais e os preços somente na base.
- As informações necessárias para construir o **número índice de preço ou de quantidade de Paasche** são as mesmas necessárias para construir o índice de valor; portanto não há economia nos seus custos operacionais ou redução do tempo empregado.
- O uso de um índice agregado ponderado deve ser feito com base na distribuição de probabilidade subjacente aos dados, como no caso de índices agregados simples:

a) Se os relativos de preços (ou de quantidade) seguirem uma distribuição normal, utilize o **índice média aritmética (Índice de Laspeyres)**;

b) Se o inverso dos relativos de preços (ou de quantidade) seguir uma distribuição normal, utilize o **índice média harmônica (de Paasche)**; e

c) Se o logaritmo dos relativos de preços (ou de quantidade) seguir uma distribuição normal, utilize o **índice de média geométrica** (no caso, **de Divisia**).

Caro aluno, chegou o momento de consultar o Exemplo 10 do material complementar! Leia e analise o exemplo. Em seguida, continue a sua leitura da Subseção 1.7.3.

1.7.3 Índices especiais de preço e quantidade: de Fischer, de Drobish e de Marshal-Edgeworth

São denominados aqui como números índices especiais aqueles que têm como finalidade corrigir a tendência do número índice de preço de Laspeyres de **superestimar as suas estimativas**, assim como a tendência do número índice de preço de Paasche de **subestimar os seus valores estimados**. Veremos a seguir os números índices de Fisher, de Drobrish e de Marshal Edgeworth.

Índice de preços e de quantidade de Fischer

Como observado anteriormente, o índice de Laspeyres tende a dar maior importância relativa aos itens cujos preços subiram de forma mais intensa. Por outro lado, de maneira diferente, o índice de Paasche tende a diminuir a importância relativa desses itens cujos preços subiram de forma mais intensa.

Dentro desse contexto, podemos ser induzidos a pensar que o índice de preço correto seja caracterizado por um valor médio entre os índices de Laspeyres e de Paasche. Esse raciocínio foi a lógica utilizada por Fisher para idealizar o seu índice.

O **Índice de preço e de quantidade de Fisher** pode ser obtido calculando-se a média geométrica entre os índices de Laspeyres e de Paasche, respectivamente, como vemos a seguir:

$$F_{0,t}^{p} = \sqrt{L_{0,t}^{p} \times PA_{0,t}^{p}} \quad \textbf{e} \quad F_{0,t} = \sqrt{L_{0,t}^{q} \times PA_{0,t}^{q}} \qquad \textbf{(32)} \text{ e } \textbf{(33)}$$

Os índices de Fisher são melhores estimadores dos índices de preço e de quantidade do que os correspondentes de Laspeyres e de Paasche. Mas, na prática, as fórmulas dos números índices de Fisher são de pouco uso, tendo em vista que elas são funções do índice de Paasche e, como já se observou, esse índice utiliza, nas suas estimativas, quantidades e preços atuais, o que resulta em certas dificuldades para a sua implantação, tanto em termos de custos como em termos de tempo.

Adicionalmente, o índice de Fisher não apresenta uniformidade em suas evoluções que permitam estabelecer comparações de preços e quantidades em séries que envolvam mais de dois períodos.

Índice de preços e quantidade de Drobish

Seguindo a mesma metodologia do número índice de Fisher, a fórmula do número índice de Drobish sugere que se calcule a média aritmética dos índices de Laspeyres e de Paasche. Os Índices de Preço e Quantidade de Drobish são definidos, respectivamente, como segue:

$$Dr_{0,t_p} = \frac{L_{0,t_p} + PA_{0,t_p}}{2} \quad \textbf{e} \quad Dr_{0,t_q} = \frac{L_{0,t_q} + PA_{0,t_q}}{2} \qquad \textbf{(34)} \text{ e } \textbf{(35)}$$

Também, como o índice de Fisher, esse índice de preço e de quantidade tem como meta contrabalançar os efeitos de subestimação de quedas provocados no índice de Paasche, que considera como base as quantidades (ou preços) iguais aos do período atual, e os efeitos de superestimação das altas, provocados no índice de Laspeyres, que considera as quantidades (ou preços) como sendo sempre os mesmos do período base.

Índice de preços e quantidade de Marshal-Edgeworth

Para contornar os efeitos de viés antagônico presentes nos índices de Laspeyres e de Paasche, o índice de Marshal-Edgeworth foi proposto, procedendo a uma média aritmética entre os componentes do numerador e do denominador que compõem, respectivamente, os índices de Laspeyres e de Paasche.

Portanto, o **Índice de Marshal-Edgeworth de preço** é definido a partir das relações (23) e (26), como vemos a seguir:

$$ME_{0,t}^{p} = \frac{\dfrac{\sum_{i=1}^{n}(p_t \times q_0 + p_t \times q_t)^i}{2}}{\dfrac{\sum_{i=1}^{n}(p_0 \times q_0 + p_0 \times q_t)^i}{2}} \times 100 \rightarrow$$

$$ME_{0,t}^{i} = \frac{\sum_{i=1}^{n}(p_t \times q_0 + p_t \times q_t)^i}{\sum_{i=1}^{n}(p_0 \times q_0 + p_0 \times q_t)^i} \times 100$$

ou

$$ME_{0,t}^{p} = \frac{\sum_{i=1}^{n} p_t (q_0 + q_t)^i}{\sum_{i=1}^{n} p_0 (q_0 + q_t)^i} \times 100 \qquad \textbf{(36)}$$

Da mesma forma, o **Índice de Marshal-Edgeworth de quantidade** é definido a partir das relações (24) e (27), como podemos observar abaixo:

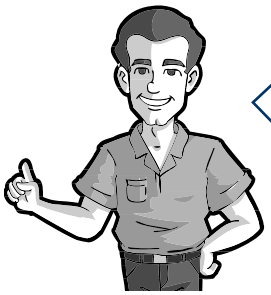
$$ME_{0,t}^{q} = \frac{\dfrac{\sum_{i=1}^{n}(p_0 \times q_t + p_t \times q_t)^i}{2}}{\dfrac{\sum_{i=1}^{n}(p_0 \times q_0 + p_t \times q_0)^i}{2}} \times 100 \rightarrow$$

$$ME_{0,t}^{q} = \frac{\sum_{i=1}^{n}(p_0 \times q_t + p_t \times q_t)^i}{\sum_{i=1}^{n}(p_0 \times q_0 + p_t \times q_0)^i} \times 100$$

ou

$$ME_{0,t}^{q} = \frac{\sum_{i=1}^{n} q_t(p_0 + p_t)^i}{\sum_{i=1}^{n} q_0(p_0 + p_t)^i} \times 100 \qquad \textbf{(37)}$$

Portanto, as equações (36) e (37) permitem estimar os índices de preço e de quantidade de Marshal-Edgeworth.



Caro aluno, agora você deve consultar o Exemplo 11 do seu material complementar! Leia e analise o exemplo. Em seguida, recomendo que você resolva a Questão 7 das Atividades de aprendizagem. Isso facilitará muito o seu trabalho. Depois, continue a sua leitura da Seção 1.8.

1.8 NÚMEROS ÍNDICES AGREGADOS PONDERADOS DE BASE MÓVEL

No Brasil, devido aos fatores históricos, como evidenciado anteriormente, existe uma predileção pelos índices relativos de ligação, sendo necessário saber como estimá-los. Portanto, vamos apresentar a seguir as principais formulações de números índices de base móvel.

1.8.1 NÚMERO ÍNDICE DE THEIL (MÉDIA GEOMÉTRICA PONDERADA)

Henri Theil propôs um índice que se constitui num sistema de ponderação, no qual os pesos compõem uma média ponderada entre os pesos das épocas consideradas na estimativa do número índice. No caso particular de estabelecer a comparação entre as épocas *t* e *(t-1),* para os índices de preço e quantidade, tem-se, respectivamente:

$$T^{p}_{(t-1,t)} = \prod_{i=1}^{n}\left(\frac{p^{i}_{t}}{p^{i}_{t-1}}\right)^{\frac{w_{t-1}+w_{t}}{2}} \quad \textbf{e} \quad T^{q}_{(t-1,t)} = \prod_{i=1}^{n}\left(\frac{q^{i}_{t}}{q^{i}_{t-1}}\right)^{\frac{w_{t-1}+w_{t}}{2}} \qquad \textbf{(38) e (39)}$$

Embora não haja restrições teóricas ao emprego da fórmula de Theil, subsiste o problema das restrições operacionais, devido à necessidade de estimar as bases de ponderação a cada período atual. As bases de ponderação são as seguintes:

$$w^{i}_{t} = \frac{p^{i}_{t} \times q^{i}_{t}}{\sum_{i=1}^{n} p^{i}_{t} \times q^{i}_{t}} \quad \textbf{e} \quad w^{i}_{t-1} = \frac{p^{i}_{t-1} \times q^{i}_{t-1}}{\sum_{i=1}^{n} p^{i}_{t-1} \times q^{i}_{t-1}} \qquad \textbf{(40)} \text{ e } \textbf{(41)}$$

1.8.2 NÚMERO ÍNDICE DE LASPEYRES COM BASE MÓVEL

O Índice de Laspeyres modificado com base móvel (que satisfaz as propriedades de inversão e circular) é um método alternativo, que consiste em calcular o índice em cadeia a partir de índices intermediários, que são obtidos mediante o emprego de médias aritméticas em vez de geométricas, e definido como vemos a seguir:

$$LM^{p}_{t-1,t} = \sum_{i=1}^{n}\left[\left(\frac{p_t}{p_{t-1}}\right)^{i} \times w_0^i\right] \times 100 \qquad (42)$$

e

$$LM^{q}_{t-1,t} = \sum_{i=1}^{n}\left[\left(\frac{q_t}{q_{t-1}}\right)^{i} \times w_0^i\right] \times 100 \qquad (43)$$

Trata-se de um índice com sistema de ponderação fixo em uma época básica fixa, definida como segue:

$$w_0^i = \frac{p_0^i \times q_0^i}{\sum_{i=1}^{n} p_0^i \times q_0^i} \quad \text{com} \quad \sum_{i=1}^{n} w_{0,t-1}^i = 1 \qquad (44)$$

Portanto, em resumo, o Índice de Laspeyres de Base Móvel apresenta um sistema de ponderação fixa em uma época básica fixa e com base de comparação móvel.

1.8.3 Índice de Bureau (ou índice de Laspeyres modificado, com base móvel)

O Índice de Bureau é um índice de Leysperes modificado, mas que se utiliza de base móvel de comparação e ponderação, com quantidades fixas em determinada época 0, sendo definido para a época *t,* em relação à época imediatamente anterior *(t-1)*, ou seja:

$$w_{0,t-1}^i = \frac{p_{t-1}^i \times q_0^i}{\sum_{i=1}^{n} p_{t-1}^i \times q_0^i} \quad \text{com} \quad \sum_{i=1}^{n} w_{0,t-1}^i = 1 \qquad (45)$$

Portanto, considerando que o Índice de Bureau é definido por uma média aritmética ponderada, usando a base de ponderação (42), temos:

$$B^{p}_{t-1,t} = \sum_{i=1}^{k}\left[\left(\frac{p_t}{p_{t-1}}\right)^{i} \times w^{i}_{0,t-1}\right] \times 100 \qquad \textbf{(46)}$$

Portanto, considerando as relações (45) e (46), obtém-se:

$$B^{p}_{t-1,t} = \left[\sum_{i=1}^{k}\left(\left(\frac{p_t}{p_{t-1}}\right)^{i} \times \frac{(p_{t-1}\times q_0)^{i}}{\sum_{i=1}^{n}(p_{t-1}\times q_0)^{i}}\right)\right] \times 100 \rightarrow B^{p}_{t-1,t} = \frac{\sum_{i=1}^{k}(p_t \times q_0)^{i}}{\sum_{i=1}^{n}(p_{t-1}\times q_0)^{i}} \times 100 \qquad \textbf{(47)}$$

E, para o Índice de Bureau de quantidade, considera-se que:

$$B^{q}_{t-1,t} = \sum_{i=1}^{k}\left[\left(\frac{q_t}{q_{t-1}}\right)^{i} \times w^{i}_{0,t-1}\right] \times 100 \qquad \textbf{(48)}$$

Utilizando a base de ponderação dada pela relação (45) na equação (48), obtém-se:

$$B^{q}_{t-1,t} = \left[\sum_{i=1}^{k}\left(\left(\frac{q_t}{q_{t-1}}\right)^{i} \times \frac{(p_0\times q_{t-1})^{i}}{\sum_{i=1}^{n}(p_0\times q_{t-1})^{i}}\right)\right] \times 100 \rightarrow B^{q}_{t-1,t} = \frac{\sum_{i=1}^{k}(q_t \times p_0)^{i}}{\sum_{i=1}^{n}(q_{t-1}\times p_0)^{i}} \times 100 \qquad \textbf{(49)}$$

Para se estimar os números índices de preço e de quantidade de Theil, utilizam-se, respectivamente, as fórmulas (46) e (48), conjuntamente com a fórmula (45). Também se podem estimar os números índices de preço e de quantidade de Theil, utilizando as relações (47) e (49), respectivamente.

Palavra do Professor

Caro aluno, consulte agora o Exemplo 12 do seu material complementar! Leia e analise-o. Em seguida, continue a sua leitura da Subseção 1.8.4.

1.8.4 Considerações sobre o Exemplo 12 do material complementar

Os Índices de Bureau, de Laspeyres modificado e o Índice de Theil, estimados no Exemplo 12, são índices de base móvel; portanto, são comparados no Quadro 1.3 abaixo, com a finalidade de estabelecer uma análise comparativa.

	LASPEYRES		BUREAU		THEIL	
	$LM^{p}_{t-1,t}$	$LM^{q}_{t-1,t}$	$B^{p}_{t-1,t}$	$B^{q}_{t-1,t}$	$T^{p}_{t-1,t}$	$T^{q}_{t-1,t}$
06-07	118,50%	151,81%	118,5%	151,81%	117,64%	133,64%
07-08	143,76%	129,95%	143,73%	131,25%	139,76%	125,18%

Quadro 1.3 – Comparação entre as estimativas dos Índices de Preço e de Quantidades de base móvel, de Theil, de Laspeyres e de Bureau.

O que se observa neste quadro é que os índices de Laspeyres e de Bureau são maiores que os correspondentes índices de Theil. Logicamente, tal comportamento é esperado, tendo em vista que os números índices de Laspeyres e de Bureau são médias aritméticas, e o índice de Theil é definido por uma média geométrica ponderada. Lembre-se de que a média geométrica ponderada conduz a um resultado sempre menor do que o da média aritmética ponderada.

Então, a questão a ser respondida é: que fórmula de número índice de base móvel deverá ser utilizada?

Como para os números índices agregados simples, a escolha deve ser feita com base na distribuição de probabilidade subjacente aos dados observados:

a) se os relativos de preços (ou de quantidade) seguirem uma distribuição normal, utilizar-se-á um índice formulado por meio de média aritmética (Índice de Laspeyres de base móvel ou o Índice de Bureau);

b) se o logaritmo dos relativos de preços (ou de quantidade) seguir uma distribuição normal, será utilizado um índice com formulação baseada na média geométrica, no caso, o índice de Theil; e

c) se o inverso dos relativos de preços (ou de quantidade) seguir uma distribuição normal, será utilizado um índice com formulação baseada numa média harmônica, no caso, o índice de Paasche (equações (26) e (27)), considerando o período base 0 como o período *(t-1)*.

Podemos observar, ainda, que todos os números índices de base móvel satisfazem às propriedades circular e de inversão, pois, ao converter esses números índices para base fixa, aplicam-se, simultaneamente, essas próprias propriedades, por meio da fórmula (6), não causando, portanto, viés nas estimativas de conversão de base móvel para base fixa.

Assim, vamos considerar como exemplo os resultados do Índice de Laspeyres de Base Móvel e converter os Índices de Preço e de quantidade de Laspeyres de Base Móvel para Base Fixa. Aplicando as propriedades de Inversão e Circular, obtém-se:

$$LM_{06,07_p} \times LM_{07,08_p} \times LM_{08,06_p} = 1 \rightarrow$$

$$LM_{06,07_p} \times LM_{07,08_p} = \frac{1}{LM_{08,06_p}} \rightarrow$$

$$LM_{06,08_p} = LM_{06,07_p} \times LM_{07,08_p} \qquad \textbf{(50)}$$

Assim, conforme a equação (50), para os índices de preço e de quantidade (conforme o Quadro 28 do Exemplo 12 do seu material complementar), tem-se, respectivamente:

$$LM^{p}_{06,08} = 1{,}185 \times 1{,}4376 \times 100 = 170{,}35\% \qquad \textbf{(51)}$$

e

$$LM_{96,98_q} = 1{,}5185 \times 1{,}2995 \times 100 = 197{,}32\% \qquad \textbf{(52)}$$

Portanto, comparando os resultados obtidos para o índice de Laspeyres de base fixa (apresentados no Quadro 12 do Exemplo 10 do material complementar) com os transformados (utilizando as equações (50) e (51)), a partir das estimativas dos índices de base móvel (Quadro 28 do Exemplo 12 do material complementar), podemos observar que os números índices de base fixa e móvel para dois períodos consecutivos são iguais, no caso, $LM^{p}_{06,07}$ e $LM^{q}_{06,07}$. Este comportamento é caracterizado no resumo apresentado no Quadro 1.4 abaixo:

NÚMEROS ÍNDICES DE PREÇO E QUANTIDADE	DE LASPEYRES DE PREÇO DE BASE FIXA (%)	DE LASPEYRES DE PREÇO TRANSFORMADO PARA BASE FIXA (%)	DE LASPEYRES DE QUANTIDADE DE BASE FIXA (%)	DE LASPEYRES DE QUANTIDADE TRANSFORMADO PARA BASE FIXA (%)
	$L_{0,t\,p}$ (%)	$LM_{0,t\,p}$(%)	$L_{0,t\,q}$ (%)	$LM_{0,t\,p}$(%)
06-06	100,0	100,0	100,0	100
06-07	118,5	118,5	151,8	151,8
06-08	170,3	170,3	188,9	197,32

Quadro 1.4 – Comparação entre os Números índices de Preço e de Quantidade de Laspeyres de base fixa e os respectivos de base móvel transformados.

Contudo, também se observa no Quadro 1.4 que a mudança do Índice de Laspeyres de base móvel para o índice de base fixa nem sempre conduz aos mesmos valores daqueles índices de Laspeyres de Base Fixa estimados (por exemplo, no caso do Índice de Quantidade 06 – 08), porque as estimativas dos índices de base fixa estão sempre sujeitas à introdução de vieses.

Mas por que isso acontece?

Porque o índice de Laspeyres de Base fixa não satisfaz a propriedade circular, introduzindo viés em suas estimativas. No caso do número índice de quantidade isso se deve, possivelmente, ao fato de que a base de ponderação fixa não funciona adequadamente para os Índices de Lapeyres de quantidade e, em consequência, não satisfaz à propriedade circular.

Palavra do Professor

Caro aluno, agora que você está com o conteúdo fresco em sua memória, aproveite e resolva a Questão 8 das Atividades de aprendizagem. Bom trabalho!

1.9 DEFLAÇÃO DE UMA SÉRIE TEMPORAL

As variações de preço causadas por inflação ou deflação podem obscurecer as variações de quantidade. Isso significa que, às vezes, o que parece ser um crescimento de vendas ou aumento na participação no mercado (por apresentar maior faturamento) pode ser mais um efeito de flutuações de preços ou de desvalorizações cambiais, do que realmente um acréscimo nas quantidades vendidas.

Este problema torna-se mais grave quando se examinam longas séries temporais, incluindo vários anos. Em especial, o problema é bastante sério em economias como a do Brasil, que sofreu grandes mudanças estruturais em seu processo econômico ao longo de décadas e, ainda hoje, apresenta momentos de instabilidades. Portanto, é absolutamente necessário fazer deflações das séries temporais. Em outras palavras, há que se remover o efeito da inflação nos valores das séries temporais. Para tanto, devemos procurar um número índice apropriado para isso:

- se tratarmos de alguma atividade de uma empresa que vende diretamente ao consumidor final, no varejo, por exemplo, devemos utilizar como deflator um índice de preços ao consumidor, como o IPC-A do IBGE ou o IPC da FIPE, etc.;
- se tratarmos de uma empresa de vendas de bens de capital ou de vendas no atacado, por exemplo, devemos utilizar um índice que retrate as flutuações de tal mercado, como o IGP-M ou o IGP-DI da Fundação Getúlio Vargas, para os quais 60% desses índices são compostos pelo Índice de Preços por Atacado, calculado pela mesma instituição;
- contudo, se tratarmos de uma atividade de exportação, por exemplo, então seria interessante incluir também a flutuação da taxa de câmbio do país ou dos países de destino.

É importante ressaltar que no processo de deflacionamento é preciso ter os números índices de base fixa. Se apenas os relativos de ligação estiverem disponíveis, será necessário obter os números índices de base fixa por meio de transformação. Independentemente do deflator (índice) escolhido, o procedimento é similar:

$$\text{Valor Real} = \left(\frac{\text{Valor Nominal ou corrente}}{\text{Índice d Inflação}} \right) \times 100 \qquad (53)$$

Outras observações sobre o conceito de deflator:

Sempre se tem como meta analisar uma série de valores monetários, em termos de suas variações, e eliminar uma das causas (variação de preço ou variação de quantidade). Contudo, quando se pretende eliminar o impacto da variação de preço em uma série de valores monetários, utiliza-se um deflator, o que permite analisar a evolução do faturamento (ou das despesas) apenas em função da variação da quantidade, eliminando da série os efeitos decorrentes de variação de preço.

Palavra do Professor

Vamos analisar agora mais alguns exemplos! Consulte os Exemplos 13 e 14 do seu material complementar. Leia e analise-os. Em seguida, continue a sua leitura da Seção 1.10.

1.10 PODER AQUISITIVO

O **Poder Aquisitivo (PA)** de um determinado volume de unidades monetárias, com relação a uma determinada época, é definido da seguinte maneira:

$$PA = \frac{1}{P_{0,t}} \times 100 \qquad \textbf{(54)}$$

Portanto, para calcular o poder aquisitivo de uma unidade monetária, basta calcular o inverso do índice de preço.

Caro aluno, chegou o momento de consultar o Exemplo 15 do material complementar! Leia e analise o exemplo. Em seguida, recomendo que você resolva a Questão 9 das Atividades de aprendizagem.

1.11 TAXA REAL OU TAXA DEFLACIONADA

É necessário converter a taxa a ser deflacionada em índice para, em seguida, aplicar o deflator. Portanto, suponha que *i* seja a taxa nominal (ou a taxa a ser deflacionada) e *j* seja a taxa de inflação. Assim, para deflacionar a taxa deve-se fazer conforme mostra o Quadro 1.5 abaixo.

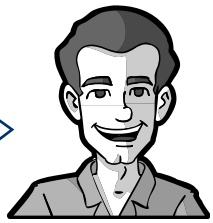
TAXA		ÍNDICE		DEFLACIONAMENTO
i	$\rightarrow$	(1+i)	$\Rightarrow$	$r = \left(\frac{(1+i)}{1+j} - 1\right) \times 100$
j	$\rightarrow$	(1+j)		

Quadro 1.5 – Relações entre número índice e taxa.

Ou seja, a taxa real é obtida, em termos percentuais, por meio da seguinte razão:

$$r = \left(\frac{(1+i)}{(1+j)} - 1\right) \times 100 \qquad \textbf{(55)}$$

Agora, caro aluno, vamos observar os Exemplos 16, 17 e 18 do material complementar! Leia e analise-os com atenção. Em seguida, resolva a Questão 10 das Atividades de aprendizagem. Só depois inicie a sua leitura da Seção 1.12.



1.12 DEFLATOR IMPLÍCITO DE PREÇO E ÍNDICE QUANTUM

Os termos deflator implícito de preço e o índice de quantum são aplicados para referenciar impactos inflacionários e impactos de variações reais relacionados a medidas de crescimento do PIB de uma economia agregada, no caso, nacional, estadual ou municipal. Entretanto, esses indicadores não são nada mais do que, respectivamente, um índice de preço e um índice de quantidade, estudados nas formas clássicas apresentadas nas seções anteriores.

1.12.1 Deflator Implícito

O Deflator Implícito de Preços é um índice de preço calculado a partir de dados da Renda Nacional ou do Produto Nacional. Mas, também, eles podem referir-se a índices obtidos internamente em uma empresa.

Supõe-se, por exemplo, que uma empresa atacadista deseja comparar as vendas entre dois anos quaisquer. As vendas totais durante cada ano contém, implicitamente, cada transação individual efetuada e registrada no caixa. Nos resultados das vendas estão os impactos de preço e de quantidade; portanto, será necessário separar esses efeitos inseridos na variação monetária de valores.

Para separar os impactos dos componentes de preço e de quantidade do resultado do valor da Renda Nacional ou do Produto Nacional, existem dois caminhos:

a) fazer uma estimativa direta do componente quantidade, deduzindo o componente preço e, em seguida, dividindo o índice de valor pelo índice de quantidade estimado; e

b) estimar o componente preço, deduzindo o componente quantidade e, em seguida, dividindo o índice de valor pelo índice de preço.

Primeiramente, definimos o **Índice de Quantum** ($IQ_{0,t}$), que descreve o comportamento da produção física de bens finais de uma economia, entre dois ou mais períodos de tempo. O Índice de Quantum é definido pelo Índice de Laspeyres de Quantidade, como você vê a seguir:

$$IQ_{0,t} = L^{q}_{0,t} \rightarrow \text{Índice de Laspeyres de Quantidade} \qquad (56)$$

O **Índice de Valor** ($V_{0,t}$) é obtido através da comparação da renda ou do produto entre dois períodos, conforme definido pela equação (31). Finalmente, utilizando o procedimento **(a)** acima, o Deflator Implícito da Renda Nacional ou do Produto Interno é o Índice de Preço, denominado $DI_{0,t}$, e estimado como vemos abaixo:

$$DI_{0,t} = \frac{V_{0,t}}{IQ_{0,t}} \qquad (57)$$

O Deflator Implícito (ou Índice de Preço) determinado pela relação (57) acima é um índice de preço de Paasche.

É importante lembrar que o produto cruzado de um índice de quantidade de Laspeyres por um índice de preço de Paasche satisfaz ao critério de decomposição das causas. Ou seja, $L^{q}_{0,t} \times PA^{p}_{0,t} = V_{0,t}$.

Palavra do Professor

Caro aluno, chegou o momento de consultar o Exemplo 19 do seu material complementar! Leia e analise o exemplo. Em seguida, inicie a leitura da Subseção 1.12.2.

1.12.2 Índice de Quantum

A princípio, o Centro de Contas Nacionais do Instituto Brasileiro de Economia da Fundação Getúlio Vargas (CCN/IBRE/FGV) era o responsável pelas estimativas das contas nacionais. Na atualidade, o IBGE é o órgão que estima, a preços constantes, o produto interno, segundo os setores de atividade da economia brasileira, bem como os principais componentes agregados das contas nacionais, permitindo a obtenção do total de despesa nacional, a preço de um determinado ano.

Saiba Mais

!

Em 1973, o IBGE passou a ser o responsável pela coordenação do sistema estatístico nacional, mas delegou à FGV a continuidade dos trabalhos respectivos às contas nacionais. Somente em meados de 1985, o IBGE iniciou o programa para a elaboração de um sistema de produção de séries anuais de Contas Nacionais completas para a economia brasileira, em colaboração com o *Institut National de La Statistique et de Etudes Economiques (INSEE)*. A Fundação Getúlio Vargas deixou de ser responsável pelas contas nacionais em dezembro de 1986, e parte dos profissionais do Centro de Contas Nacionais do Instituto Brasileiro de Economia (CCN/IBRE/FGV) foi incorporada ao quadro do IBGE. (IBGE, 1990, p.87).

As estimativas do PIB, a preços constantes, são calculadas a partir de índices do produto real (Índice de Quantum $IQ_{t-1,t}$). Conforme já anunciado, a fórmula utilizada para a obtenção do $IQ_{t-1,t}$ é a de Laspeyres com base móvel. A vantagem desse procedimento reside na possibilidade de aplicação do critério circular e de inclusão de novos produtos no painel pesquisado. Assim:

$$IQ_{t-1,t} = L^{q}_{t-1,t} = \frac{\sum_{i=1}^{n} q_t^i \times p_{t-1}^i}{\sum_{i=1}^{n} p_{t-1}^i \times q_{t-1}^i} \times 100 \qquad \textbf{(58)}$$

A fórmula (58) é aplicada na obtenção do Índice de Quantum, agregando os principais setores do aparelho de produção da economia nacional, conforme vemos a seguir:

a) **Agricultura** – lavoura; produto animal e de derivados;

b) **Indústria** – Produção Extrativa Mineral; Indústria de Transformação; Construção Civil, Industrial e Serviços de Utilidade Pública; e

c) **Serviços** – Comércio, transportes, comunicação, instituições financeiras, outros serviços e administração pública.

A agregação dos impactos dos principais setores da economia sobre o Índice de Quantum é realizada por meio de uma média aritmética ponderada sobre os índices setoriais especificados de (a) a (c). Atualmente, as participações dos setores produtivos na estrutura do índice de Quantum correspondem, aproximadamente, às seguintes proporções: **(a)** Agricultura, 9,65%; **(b)** Indústria, 37,18%; e **(c)** Serviços, 53,15%.

Caro aluno, chegou o momento de consultar o Exemplo 20 do seu material complementar! Leia e analise-o. Em seguida, resolva a Questão 11 das Atividades de aprendizagem.

1.13 ÍNDICES BRASILEIROS

Abordaremos nesta seção os principais índices brasileiros, dentre os quais:

- o Índice Nacional de Preços ao Consumidor (INPC) e o Índice Nacional de Preços ao Consumidor Amplo (IPCA), ambos estimados pelo Instituto Brasileiro de Geografia e Estatística (IBGE);
- o Índice de Preços ao consumidor da Fundação Instituto de Pesquisas Econômicas (IPC/FIPE);
- o Índice de Custo de Vida do Departamento Intersindical de Estatística e Estudos Socioeconômicos (ICV-DIEESE); e
- o Índice Geral de Preços de Mercado (IGP-M) e o Índice Geral de Preços – Disponibilidade Interna (IGP-DI), ambos estimados pela Fundação Getúlio Vargas (FGV).

Foi utilizada como fonte de informações aqui apresentadas, a série de relatórios metodológicos (volume 14) do Sistema Nacional de Índices de Preços ao Consumidor – Métodos de Cálculos (5a Edição), publicado pelo IBGE (Instituto Brasileiro de Geografia e Estatística), em 2007.

1.13.1 Índice Nacional de Preços ao Consumidor (INPC) e o Índice Nacional de Preços ao Consumidor Amplo (IPCA)

Os índices mais tradicionais estimados pelo governo são o Índice Nacional de Preços ao Consumidor (INPC) e o Índice Nacional de Preços ao Consumidor Amplo (IPCA). As diferenças metodológicas entre esses indicadores decorrem dos objetivos definidos para cada um, o que, em geral, implica em distinguir a população-objetivo e/ou o período de coleta.

A partir de julho de 1978, o IBGE assumiu integralmente esta responsabilidade, por determinação legal. As características básicas do INPC e do IPCA estão evidenciadas no Quadro 1.6 a seguir.

Entre 1948 e 1978, esteve a cargo do Ministério do Trabalho a produção do Índice de Preços ao Consumidor para 13 capitais/cidades brasileiras (Belém, Fortaleza, Natal, Recife, Salvador, Belo Horizonte, Niterói, Rio de Janeiro, São Paulo, Curitiba, Florianópolis, Porto Alegre e Cuiabá), além de um indicador nacional.

	INPC	IPCA
DEFINIÇÃO	Ambos são medidas sínteses de movimentos de preços de um conjunto de mercadorias consumidas (Cesta de Mercadorias), representativo de um determinado grupo populacional, em certo período de tempo.	
INSTITUIÇÃO RESPONSÁVEL	Fundação Instituto Brasileiro de Geografia e Estatística (IBGE)	
ABRANGÊNCIA GEOGRÁFICA	Envolvem as regiões metropolitanas de Rio de Janeiro, Porto Alegre, Belo Horizonte, Recife, São Paulo, Belém, Fortaleza, Salvador e Curitiba, além do Distrito Federal e do município de Goiânia. Os índices são calculados para cada região.	
MOTIVAÇÃO E OBJETIVO	Medir as variações de preços da cesta de consumo das populações assalariadas e com baixo rendimento.	Medir as variações de preços referentes ao consumo pessoal.
PRINCIPAL FINALIDADE	Fornecer subsídios para as decisões de reajustes de remunerações, não apenas aos agentes diretamente afetados pelos dissídios, mas a qualquer categoria de trabalhadores, sindicalizados ou não. Tem sido usado, também, como indexador de outros preços da economia, especialmente, daqueles com maior influência sobre a capacidade de consumir das famílias de mais baixos rendimentos.	Utilizado pelo Banco Central do Brasil para o acompanhamento dos objetivos estabelecidos no sistema de metas de inflação, adotado a partir de julho de 1999, para o balizamento da política monetária.
DEFLATOR	Deflator salarial.	Deflator de taxas de juros de curto e médio prazo, assim como indexador de letras do tesouro nacional de curto e médio prazo.
DADOS PRODUZIDOS	Dados disponíveis (índice nacional) desde 1979. Sendo o período da coleta o mês calendário.	Dados disponíveis (índice nacional) desde 1980. Sendo o período da coleta o mês calendário.

Quadro 1.6 – Características básicas do INPC e do IPCA.

Os procedimentos metodológicos para a implantação dos índices INPC e IPCA utilizaram-se das seguintes pesquisas:

- O **Sistema Nacional de Índices de Preços ao Consumidor (SNIPC)**, quando de sua criação, forneceu os dados necessários para a definição das populações objetivo, para a montagem da cesta de produtos e serviços, bem como para a sua estrutura de pesos, que foram extraídos da pesquisa *Estudo Nacional da Despesa Familiar (ENDEF) 1974-1975*, de objetivo mais amplo que o da Pesquisa de Orçamentos Familiares (POF), porém com características semelhantes.
- A partir da década de 1980, a **POF** tem sido a base de definição das populações objetivo para o INPC e o IPCA. Atualmente, essa pesquisa, realizada no período 2002-2003, fornece as estruturas de ponderação das populações objetivo, tanto para o INPC como para o IPCA.

- A **Pesquisa de Especificação de Produtos e Serviços (PEPS)**, realizada na época de implantação da pesquisa nas 13 cidades integradas ao sistema de estimativa desses índices, para todos os produtos e serviços constantes da estrutura de ponderações, forneceu o cadastro de produtos e serviços pesquisados, que é permanentemente atualizado com o objetivo de acompanhar a dinâmica de mercado.

- A **Pesquisa de Locais de Compra (PLC)**, realizada no período de maio a junho de 1988, nas 11 áreas de abrangência, forneceu o cadastro de informantes da pesquisa feita através de visitas aos domicílios de uma amostra previamente selecionada, cuja manutenção é contínua.

Atualmente, as 13 regiões, citadas anteriormente, foram reduzidas para 11 (aquelas citadas no Quadro 1.6).

- A base de ponderação para as estimativas do INPC e do IPCA é resultado de uma recente atualização, que utiliza as estruturas de ponderação do **Sistema Nacional de Índices de Preços ao Consumidor** (SNIPC), realizada a partir das informações sobre as despesas realizadas pelas famílias, que foram obtidas através da POF 2002-2003, implantada a partir de julho de 2006.

Utilizando os procedimentos metodológicos acima especificados, foram definidas as populações objetivo para o INPC, assim como para o IPCA, seguindo os critérios resumidos nos Quadros 1.7 e 1.8 abaixo.

POPULAÇÃO OBJETIVO DO INPC
A população objetivo tem sido focalizada no atendimento ao seu objetivo original: medir a variação agregada dos preços dos bens e serviços consumidos pelas famílias com baixos rendimentos e cujos chefes são assalariados. Os critérios de cobertura populacional e de estabilidade da estrutura de consumo têm sido aplicados, segundo os parâmetros que seguem: • **Cobertura populacional:** foi arbitrado, desde a implantação do INPC, que o índice assegurasse a cobertura populacional de cerca de 50% das famílias com chefes assalariados; e • **Estabilidade da estrutura de consumo:** foram excluídas as famílias com rendimentos menores que um salário mínimo, com base no argumento de que esse segmento tem renda e estrutura de consumo instável ou atípica. Além disso, a exclusão dessa faixa de rendimentos justifica-se, tendo em vista que o INPC visa à correção monetária de salários, não sendo procedente incluir famílias com renda inferior ao menor salário legal do País.

Neste processo de implementação dos pesos da Pesquisa de Orçamento Familiar (POF) 2002-2003 (a mais recente), decidiu-se, dado o objetivo original do INPC:

- manter a exclusão das famílias com chefes assalariados, cujos rendimentos são inferiores a um salário mínimo; e
- manter o parâmetro histórico para o critério da cobertura. Ou seja, que aproximadamente 50% das famílias com chefes assalariados sejam cobertas, considerando-se as famílias com os rendimentos mais baixos, desde que iguais ou superiores a um salário mínimo.
- Na implantação do INPC, em 1979, o IBGE definiu como população objetivo as famílias, cujos chefes eram assalariados e tinham rendimentos monetários disponíveis situados entre 1 e 5 salários mínimos. Atualmente, segundo a POF mais recente, o intervalo é de 1 a 6 salários mínimos, ficando, assim, mantidas as proporções inicialmente estabelecidas a partir dos dados do ENDEF.
- Hoje, o critério da estabilidade aplicado aos dados apresentados pela POF indica a exclusão de 4,07% das famílias, ou seja, daquelas com rendimentos menores que R$ 200,00 (duzentos reais), valor equivalente a 1 salário mínimo (15 de janeiro de 2003);
- O critério da robustez, pelo qual se busca assegurar a cobertura de cerca de 50% das famílias com mais baixos rendimentos, leva a considerar as famílias com rendimentos de até R$ 1.200,00 (um mil e duzentos reais), ou 6 salários mínimos (15 de janeiro de 2003).

Na definição dos limites de renda, foram considerados alguns fatores, tal que contemplasse mais de 50% das famílias com chefes assalariados:

- Fixou-se o limite inferior em 1 salário mínimo, a fim de não acarretar distorções à cesta, isto porque se acredita que as famílias com rendimento mensal inferior a este valor tenham sua subsistência complementada, pelo menos em parte, através de auto-consumo, doações, trocas, etc., não sendo possível caracterizar suas cestas de compras;
- A fixação do limite superior privilegiou os dois objetivos já mencionados:
  - a necessidade de pesquisar uma cesta que fosse, de fato, representativa de um maior número possível de famílias com chefes assalariados; e
  - o grupo contemplado ser aquele que tem a menor capacidade de defesa contra a inflação, ou seja, as famílias de baixa renda.

Quadro 1.7 – Critérios de definição da População objetivo do INPC.

POPULAÇÃO OBJETIVO DO IPCA
A motivação para sua criação foi oferecer, para todos os fins práticos, a medida do movimento geral dos preços no mercado varejista. Trata-se, portanto, do indicador da inflação segundo o consumo pessoal, bem como encontrar, nas Contas Nacionais, um campo de importante utilização. A definição da população objetivo do IPCA tem levado em conta o objetivo de medida da inflação sob a ótica do consumo pessoal. Além disso, foram sempre considerados os seguintes critérios e parâmetros: • **Cobertura:** acima de 90% das famílias residentes nas áreas urbanas de abrangência do SNIPC, qualquer que seja a fonte de rendimentos, de modo a assegurar cobertura próxima da totalidade, tendo em vista o objetivo do IPCA; e • **Estabilidade da estrutura de consumo:** são excluídos os extremos da distribuição, ou seja, aquelas famílias cujos rendimentos estão abaixo de 1 salário mínimo e aquelas com rendimentos considerados muito altos. Os argumentos são a instabilidade e atipicidade dos hábitos de consumo das famílias componentes desses segmentos.
A distribuição do número de famílias obtida na Pesquisa de Orçamento Familiar (POF) 2002-2003, segundo o rendimento familiar monetário disponível para o total das 11 áreas do SNIPC, permitiu aplicar os seguintes critérios: • O **critério da estabilidade** que indicou a exclusão de 8,2% das famílias, ou seja, aquelas com rendimentos menores que 1 salário mínimo de 15 de janeiro de 2003, correspondendo a 5,8%, e as famílias com rendimentos superiores a R$ 8.000,00 (oito mil reais), ou seja, 40 salários mínimos, perfazendo 2,4%. • O **critério da robustez,** que assegura a cobertura de mais que 90% das famílias, levou a considerar aquelas famílias com rendimentos de R$ 200,00 (duzentos reais) até R$ 8.000,00. Assim, a população objetivo do IPCA adotada desde julho de 2006 é a que segue: famílias residentes nas áreas urbanas das regiões de abrangência do SNIPC com rendimentos de 1 a 40 salários mínimos, qualquer que seja a fonte dos rendimentos.

Quadro 1.8 – Critérios de definição da População objetivo do IPCA.

A estrutura das famílias que integram a faixa de renda de 1 a 6 salários mínimos (no caso do INPC) é diferente daquela cuja faixa de renda compreende 1 a 40 salários mínimos (no caso do IPCA). Estas diferenças podem ser tanto para as espécies de produtos e serviços, quanto para as despesas relativas efetuadas. Atualmente, as estruturas de ponderações utilizadas para o cálculo dos índices resultaram da consolidação dos orçamentos familiares levantados pela POF 2002-2003.

As estruturas são montadas de forma que as categorias de consumo de mesma natureza fiquem juntas, resultando nos seguintes níveis de agregação, assim hierarquizados: Grupo (por exemplo, alimentação e bebidas), Subgrupo (por exemplo, alimentação), Item (por exemplo, frutas) e Subitem (por exemplo, laranja).

Os métodos de determinação dos pesos e os principais critérios adotados na montagem das estruturas de ponderações dos índices regionais são as seguintes:

a) expandir, ao longo do ano, os valores das despesas de consumo familiar provenientes da POF, coletados em diferentes períodos de referência;

b) deflacionar as despesas anuais para 15 de janeiro de 2003, ponto referencial para a transformação dos valores monetários a preços constantes;

c) somar, para cada subitem, as despesas realizadas pelas famílias pertencentes à população objetivo; e

d) calcular a razão entre a soma obtida em (c) e a despesa total (relativa a todos os subitens) de todas as famílias da região em questão (*ou seja, define-se uma base de ponderação ou pesos*).

Calculados os pesos, são montadas as estruturas de consumo, podendo-se constatar, neste momento, que há subitens com participações inexpressivas. Dessa forma, estruturas originais poderão ser muito extensas, dificultando o acompanhamento eficaz dos preços mês a mês. Quando isto ocorre, são realizados alguns estudos que resultam na simplificação das estruturas, sem, contudo, comprometer sua representatividade. No caso das estruturas do INPC e do IPCA, foi necessário proceder a simplificações. Assim, a montagem das estruturas definitivas destes índices obedeceu aos seguintes critérios, conforme explicado a seguir:

a) subitens com participação igual ou superior a 0,07% fazem parte das estruturas;

b) subitens com participação inferior a 0,01% em hipótese alguma fazem parte das estruturas. Os valores dessas despesas são distribuídos, proporcionalmente, entre outras despesas do mesmo gênero, ou seja, no item; e

c) subitens com ponderação igual ou superior a 0,01% e inferior a 0,07% podem fazer parte da estrutura para assegurar que o item do qual fazem parte tenha cobertura de 70% dos gastos realizados com os componentes do item. Esta cobertura é estabelecida em relação à estrutura completa definida de início.

Constituídas as estruturas, observa-se que, no nível de subitem, são evidenciadas peculiaridades estruturais relativas a cada área e população objetivo, podendo existir certo subitem numa área e não existir em outra, ou existir em determinada área para a estrutura de pesos do INPC e não para a do IPCA.

Entretanto, no nível de item, o processo de agregação e hierarquização das despesas é realizado de modo que garanta a existência da categoria em todas as estruturas de pesos. Assim, os itens são agregados em caráter nacional e, por serem

comuns às diversas áreas, todos os resultados produzidos a partir deste nível de agregação das despesas são passíveis de comparação. Esse procedimento possibilita estimar, em média, que as estruturas dos gastos dos grupos distribuam-se, aproximadamente, conforme o Quadro 1.9, para o INPC e para o IPCA.

INPC		IPCA	
TIPO DE GASTOS	PESO DO GASTO (EM %)	TIPO DE GASTOS	PESO DO GASTO (EM %)
Alimentação e bebidas	33,10	Alimentação e bebidas	25,21
Despesas Pessoais	13,36	Transporte e comunicação	18,77
Vestuário	13,16	Despesas Pessoais	15,68
Habitação	12,53	Vestuário	12,49
Transporte e comunicação	11,44	Habitação	10,91
Artigos de residência	8,85	Saúde e cuidados pessoais	8,85
Saúde e cuidados pessoais	7,56	Artigos de residência	8,09
TOTAL	100,00	TOTAL	100,00

Quadro 1.9 – Ponderações dos gastos dos subitens agregados, para a determinação do INPC e do IPCA.

O ponto de partida para o cálculo mensal dos índices INPC e IPCA é a série histórica de dois meses que contém, para cada produto, o preço obtido no mês corrente (mês em que se está calculando o índice) e no mês anterior. Ressalta-se que para os produtos cujos métodos de coleta possibilitam a obtenção de mais de um preço por local, o preço registrado na série histórica corresponde à média aritmética dos preços obtidos no respectivo estabelecimento. Tal procedimento constitui-se, a rigor, na primeira etapa de agregação para o cálculo dos indicadores dos produtos, subitens, itens, subgrupos e grupos com esta característica.

Assim, a partir das informações da série histórica de dois meses, a estimativa das variações mensais dos preços dos produtos, referenciadas por *j*, ou de seus relativos de preços, permite levantar os índices de preços regionais e nacionais, cujo procedimento está apresentado no Quadro 1.10 abaixo. Deve ser lembrado que os procedimentos para determinar o INPC ou o IPCA são os mesmos, o que difere entre eles é a população foco e os pesos dos itens que integram a cesta de produtos.

NÍVEL DE AGREGAÇAO	PROCEDIMENTO
Coleta que obtenha mais de um preço por local de compra, ao longo do mês atual.	Procede-se a média aritmética dos preços obtidos no respectivo estabelecimento.
Variação de preços do produto *j* (ou índice do produto *j*), entre o mês *t-1*, mês anterior e *t*, o mês corrente ($R^{j}_{t\ -1,t}$). Fórmula adotada, a partir de Junho de 1980.	$R^{j}_{t-1,t} = \frac{P^{j}_{t}}{P^{j}_{t\ -1}} = \frac{\frac{1}{n_t}\sum_{L=1}^{n_t} P^{j.L}_{t}}{\frac{1}{n_{t-1}}\sum_{L=1}^{n_{t-1}} P^{j.L}_{t-1}}$ onde $P^{j.L}_{t}$ é o preço do produto *j*, no local *L*, no mês *t* e $P^{j.L}_{t-1}$ o preço do produto *j*, no local *L*, no mês *t-1*, n_t o número de locais de compra no mês *t* e n_{t-1} o número de locais de compra, no mês *t-1*.
Variação de preço do subitem *k*[1], também chamado de índice do subitem *k* (estimado por meio de uma média geométrica), $R^{k}_{t-1,t}$.	$R^{k}_{t\ -1,t} = \sqrt[n]{\prod_{j=1}^{m} R^{j}_{t\ -1,t}}$ onde $R^{k}_{t-1,t}$ é variação de preços entre os meses *t-1* e *t*, dos produtos que compõem o subitem k, $R^{j}_{t-1,t}$ é a variação do produto *j*, entre os meses *t-1* e *t* e *m* o número de produtos que compões o subitem *k*.
Índice do Item *m* (expresso, a partir da fórmula de Laspeyres de preço de base móvel), $I^{m}_{t-1,t}$.	$L^{p}_{t-1,t} = \sum_{k=1}^{n} w^{k}_{0} \cdot \left(\frac{p^{k}_{t}}{p^{k}_{t-1}}\right) = \sum_{k=1}^{n} w^{k}_{0} \cdot R^{k}_{t-1,t}$ onde $w^{k}_{0} = \frac{(p_0 \cdot p_0)^k}{\sum_{k=1}^{n}(p_0 \cdot p_0)^k}$, ou seja, o peso do subitem *k*, obtido pela pesquisa POF. Contudo, a estimativa, $L^{p}_{t-1,t}$ serve para corrigir a base de ponderação para o período anterior ao mês atual, conforme o seguinte procedimento: $I^{m}_{t-1,t} = \frac{\sum_{k=1}^{n} w^{k}_{t-1} \cdot R^{k}_{t-1,t}}{\sum^{n} w^{k}_{t-1}}$, tal que: $w^{k}_{t-1} = w^{k}_{0} \cdot \prod_{j=0}^{t-2} \frac{R^{k}_{j,j+1}}{L^{p}_{j,j+1}}$
Índice de Preço ao Consumidor Regional ($IPC^{A,F}_{t-1,t}$) da área denominada *A*, para a população especifica *F*, distinta para o INPC e IPCA.	$IPC^{A,F}_{t-1,t} = \sum_{m}^{M} w^{m}_{t-1} \cdot I^{m}_{t-1,t}$, onde $w^{m}_{t-1} = w^{m}_{0} \prod_{j=0}^{t-2} \frac{R^{m}_{j,j+1}}{L^{p}_{j,j+1}}$, $I^{m}_{t-1,t} = \frac{\sum_{k=1}^{n} w^{k}_{t-1} \cdot R^{k}_{t-1,t}}{\sum_{k=1}^{n} w^{k}_{t-1}}$ e $L^{p}_{t-1,t} = \sum_{m=1}^{m} w^{m}_{0} \cdot R^{m}_{t-1,t}$
Índice Nacional, $IPCA_{t-1,t}$ ou $INPC_{t-1,t}$ obtido a partir dos índices regionais.*	$INPC_{t-1,t} = \sum_{A=1}^{11} w^{A,F}_{t-1} \cdot InPC^{A,F}_{t-1,t}$ ou $IPCA_{t-1,t}$, o que muda é a população foco. onde $W^{A,F}$ é o peso da área *A*, definido como: $WA,F = \frac{\text{população da grande região x coeficiente de proporcionalidade}}{\text{população urbana brasileira}}$ **

Quadro 1.10 – Procedimento metodológico para a determinação dos índices INPC e IPCA.
(Como podemos observar, todos os produtos participam do subitem k com a mesma ponderação).

Saiba Mais

* De junho de 1989 a dezembro de 1993, com a redefinição da estrutura de ponderações, o índice restrito de cada área passou a ser ponderado pela População Urbana de seu estado e parte das populações urbanas não cobertas pelo SNIPC, pertencentes à mesma Grande Região, através da utilização dos dados da projeção de população residente urbana de 1985, realizada pelo então Departamento de População e Indicadores Sociais. A partir de janeiro de 1994, a fonte passou a ser o Censo Demográfico, realizado em 1991 (substituindo a projeção populacional para 1985). Em 1999, os pesos implantados foram gerados a partir dos dados da Contagem da População de 1996. Na presente atualização, os novos pesos das regiões baseiam-se nas mais recentes estimativas da População Residente Urbana, obtida por meio da POF 2002-2003. Para o IPCA, de junho de 1989 a dezembro de 1993, utilizou-se a variável Rendimento Total Urbano como ponderadora regional, com base nos dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 1987 e, a partir de janeiro de 1994, da PNAD de 1990. As ponderações regionais para o IPCA foram novamente atualizadas, agora com base nas estimativas de Rendimento Familiar Monetário, disponíveis mensalmente, obtidas da POF 2002-2003. Ressalta-se que as fontes das variáveis ponderadoras foram substituídas por pesquisas mais recentes, visando a maior precisão no cálculo dos estimadores nacionais. Acrescenta-se, ainda, que a fórmula de agregação dos índices regionais para obtenção dos índices nacionais continuará sendo a média aritmética ponderada.

** Na mais recente atualização, tendo como fonte a POF 2002-2003, os pesos das regiões foram obtidos com base nas estimativas da população urbana para os estados, para as Grandes Regiões e para o Brasil. Primeiramente, calculam-se os coeficientes de proporcionalidade referentes às Unidades da Federação da Grande Região com cobertura do SNIPC, que retratam individualmente a participação da população urbana dos estados no total da população da Grande Região, excluindo deste cálculo as Unidades da Federação não pesquisadas pelo SNIPC.

1.13.2 Índice de Preços ao Consumidor da FIPE, IPC/FIPE

O Índice de Preços ao Consumidor do Município de São Paulo, que deu origem ao IPC-FIPE, é o mais tradicional indicador da evolução do custo de vida das famílias paulistanas e o mais antigo do Brasil. Foi criado pela Prefeitura de São Paulo em 1939, com o objetivo de calcular reajustes de salários dos servidores municipais. No início, sua apuração era de responsabilidade da Divisão de Estatística e Documentação da Prefeitura. Em 1968, passou a ser calculado pelo Instituto de Pesquisas Econômicas da USP e, em 1973, pela FIPE (ano em que esta foi fundada).

O Índice IPC-FIPE tem uma história de credibilidade, conquistada ao longo do período de inflação elevada, que chegou aos 80% em junho de 1994, pois era a

única instituição a fazer apurações e divulgações semanais de preços para balizar as expectativas de inflação diariamente projetadas pelo mercado. Fez isso antes das demais instituições que se dedicavam a acompanhar a taxa de inflação.

Em épocas de inflação alta, índices como o IPC-FIPE recebem grande importância ao serem utilizados, livremente ou oficialmente, como indexadores dos mais variados tipos de contratos, com a finalidade de corrigir as perdas de poder de compra dos rendimentos de salários, pensões, aposentadorias, aluguéis, juros ou contratos de prestação de serviços, de entregas futuras, da cesta básica, etc.

Podemos definir o IPC-FIPE como o índice que mede a variação do custo de vida das famílias com renda de 1 a 20 salários mínimos, do município de São Paulo. Quanto à metodologia, o cálculo do IPC-Fipe manteve-se sem grandes alterações desde o início dos anos 1970. Ele é feito da seguinte forma: o período de coleta é diário, e semanalmente ocorrem divulgações prévias, chamadas variações quadrissemanais, que comparam os preços médios das últimas quatro semanas apurados com os das quatro semanas imediatamente anteriores.

Para o cálculo das variações quadrissemanais, esse índice leva em consideração a amostra total do IPC mensal de aproximadamente 110.000 tomadas de preços, subdividida em quatro subamostras, cada uma delas pesquisadas em um período de no mínimo 7 e no máximo 8 dias, que constituem a semana de coleta. O sistema de cálculo sempre abrange um período total de 8 semanas e as variações são obtidas fazendo-se a divisão dos preços médios das 4 semanas de referência pelos preços médios das 4 semanas anteriores (base).

Dessa forma, para se obter uma série sequencial de índices quadrissemanais, consideram-se sempre 8 semanas, incluindo-se no cálculo as informações sobre os preços coletados na última semana, eliminando automaticamente da operação os dados referentes à semana mais antiga. São apresentadas, portanto, 3 prévias durante o mês, sendo a 4ª quadrissemana o resultado definitivo do mês.

1.13.3 Índice de Custo de Vida do DIEESE (ICV-DIEESE)

O ICV-DIEESE é um número índice que tem como objetivo medir o movimento dos preços de um conjunto de bens e serviços que formam uma cesta de consumo fixa, com itens e quantidades apurados através de uma pesquisa de orçamento familiar (POF), nos seus segmentos finais de comercialização, entre um mês civil e o seu anterior.

Sua principal utilidade é medir e apurar o poder de compra desses bens e serviços pelos trabalhadores (levando-se em consideração as diferentes faixas salariais) e servir de base para negociações de melhores salários, ou ainda para o cálculo da inflação. A população objetivo é composta por famílias com renda entre 1 e 30 salários mínimos. O ICV-DIEESE atualmente é construído sobre a base da Pesquisa de Orçamento Familiar 1994/95 para a cidade de São Paulo, elaborada pelo Departamento Intersindical de Estatística e Estudos Socioeconômicos (DIEESE).

Desde 1983, sucederam-se planos de estabilização econômica, viveu-se sob a hiperinflação, sob recessão e, a partir de 1994, vive-se um plano de estabilização que vem tendo êxito na redução da inflação e que tem servido de instrumento para o aprofundamento da integração da economia brasileira no processo de globalização financeira, ao mesmo tempo em que mantém a economia interna condicionada a níveis insuficientes de crescimento econômico.

Nesse quadro, é evidente que a população, particularmente a que trabalha, foi levada a alterar seu padrão de vida, procurando se adaptar, do ponto de vista social e econômico, às novas condições da sociedade, da economia e do mercado de trabalho.

A Pesquisa de Orçamentos Familiares 1994/95 é a quarta das pesquisas voltadas para esse mesmo tema, realizadas pelo DIEESE desde 1958, com o objetivo mais imediato de definir a estrutura do ICV, mas, ao mesmo tempo, de levantar dados que permitam analisar as condições de sobrevivência da população em geral e, em particular, dos trabalhadores. A evolução da estrutura de consumo das famílias de assalariados da cidade de São Paulo é demonstrada no Quadro 1.11 a seguir.

	1958	1969/70	1982/83	1994/95
ALIMENTAÇÃO	45,00	39,00	28,13	27,44
NO DOMICÍLIO	-	37,13	23,22	21,40
HORTIFRUTAS	-	4,37	3,03	3,23
CARNES, PEIXES E OVOS	-	9,28	6,57	5,55
LEITE E DERIVADOS	-	4,21	3,40	3,25
CEREAIS, MASSAS, PÃES ETC.	-	11,15	4,62	3,76
OUTROS NO DOMICÍLIO	-	8,12	5,60	5,61
FORA DO DOMICÍLIO	-	1,87	4,91	6,04
HABITAÇÃO	33,00	25,20	24,87	23,52
LOCAÇÃO, IMPOSTOS E TAXAS	-	14,71	9,74	10,32

	1958	1969/70	1982/83	1994/95
MANUTENÇÃO	-	2,40	6,39	3,25
SERVIÇOS PÚBLICOS	-	6,39	4,72	6,19
OUTROS	-	1,70	4,02	3,76
TRANSPORTE	2,00	8,80	19,30	13,62
COLETIVO	-	4,74	2,95	3,41
INDIVIDUAL	-	4,06	16,35	10,21
SAÚDE	4,00	3,60	4,95	8,18
ASSISTÊNCIA MÉDICA	-	2,17	3,71	5,91
MEDICAMENTOS E PRODUTOS FARMACÊUTICOS	-	1,43	1,24	2,20
APARELHOS	-	-	-	0,07
VESTUÁRIO	10,00	7,48	6,54	7,87
EDUCAÇÃO E LEITURA	1,00	3,50	4,80	6,91
EDUCAÇÃO	-	3,10	4,45	6,33
LEITURA	-	0,40	0,35	0,58
EQUIPAMENTOS DOMÉSTICOS	3,00	7,12	4,89	6,13
ELETRODOMÉSTICOS	-	3,21	2,42	3,47
MÓVEIS	-	2,55	1,32	1,42
OUTROS	-	1,36	1,15	1,24
DESPESAS PESSOAIS	1,50	5,18	4,72	3,96
RECREAÇÃO	0,50	0,12	1,63	2,08
DESPESAS DIVERSAS	-	-	0,17	0,29
TOTAL	100,00	100,00	100,00	100,00

Quadro 1.11 – Estrutura de orçamento doméstico – Município de São Paulo, 1958, 1969/70, 1982/83 e 1994/95, em termos percentuais (%).

O Quadro 1.11 tem como fonte DIEESE-POFs 1958, 1969/70, 1982/83 e 1994/95. As estruturas das pesquisas de 1969/70 e 1982/83 foram ajustadas à de1994/95.

No levantamento da POF 1994/95 ampliou-se o conceito para a família assalariada, discriminando as famílias por estratos, em função dos rendimentos dos assalariados, de forma a captar o impacto do custo de vida para toda a população residente em São Paulo. Nesse contexto, três estratos de estrutura de gastos familiares foram definidos, tal que:

a) o **estrato 1** encontra-se associado a famílias com chefes assalariados com salários médios de aproximadamente até três salários mínimos;

b) o **estrato 2** refere-se a famílias com chefes assalariados com salários médios de aproximadamente dez salários mínimos; e

c) o **estrato 3** refere-se a famílias com chefes assalariados com salários médios de aproximadamente vinte salários mínimos.

A estrutura de gastos para as famílias que integram estes estratos encontra-se demonstrada no Quadro 1.12, abaixo.

ITENS DE CONSUMO	TOTAL	ESTRATO 1	ESTRATO 2	ESTRATO 3
TOTAL GERAL	100,00	100,00	100,00	100,00
ALIMENTAÇÃO	27,44	35,71	31,19	23,80
HABITAÇÃO	23,52	25,50	23,75	22,95
TRANSPORTE	13,62	7,74	12,29	15,62
SAÚDE	8,18	6,55	6,73	9,22
VESTUÁRIO	7,87	8,78	8,39	7,43
EDUCAÇÃO E LEITURA	6,91	3,25	4,14	9,02
EQUIPAMENTOS DOMÉSTICOS	6,13	5,56	7,18	5,80
DESPESAS PESSOAIS	3,96	5,38	4,37	3,44
RECREAÇÃO	2,08	1,23	1,74	2,44
DESPESAS DIVERSAS	0,28	0,30	0,23	0,29

O Quadro 1.12 tem como fonte DIEESE - POF 1994/95. Observação: a preços de junho/96; deflator INPC/ SP – IBGE.

Quadro 1.12 – Gasto mensal médio por domicílio – Município de São Paulo dezembro de 1994 a novembro de 1995 - (em %).

O ICV-DIEESE é integrado por cerca de 1.000 participantes, incluindo seis grandes centrais sindicais: Força Sindical, Confederação Geral dos Trabalhadores do Brasil (CGTB), Coordenação Nacional de Lutas (Conlutas), Nova Central Sindical dos Trabalhadores (NCST), Central Única dos Trabalhadores (CUT) e União Geral dos Trabalhadores (UGT). A Metodologia utiliza a fórmula de Laspeyres, supondo que não há substituição de bens. As quantidades apuradas, quando da realização da última POF, 1994/95, são mantidas constantes. A cesta de consumo fixa obtida na POF mantém-se, portanto, inalterada, até que uma nova pesquisa domiciliar seja realizada.

Supõe-se rigidez nos hábitos de consumo. A atual composição dos grupos de despesas para o cálculo do índice, de forma global, está dada na coluna **Total** do Quadro 1.12 acima.

Links

Para obter maiores informações a respeito desse índice, consulte:

www.dieese.org.br (clique no link *metodologias* e, em seguida, em *ICV*).

1.13.4 Índice Geral de Preços de Mercado (IGP-M) da Fundação Getúlio Vargas (FGV)

O IGP-M é o Índice Geral de Preços do Mercado, calculado pela Fundação Getúlio Vargas. A coleta de preços é feita entre os dias 21 do mês anterior e 20 do mês corrente, com divulgação no dia 30. É composto por três índices: **Índice de Preços no Atacado (IPA-M)**, **Índice de Preços ao Consumidor (IPC-M)** e **Índice Nacional do Custo da Construção (INCC)**, que representam 60%, 30% e 10%, respectivamente, do IGP-M. Portanto, definiremos estes índices, como segue:

- O **IPA** é calculado pela FGV com base na variação dos preços no mercado atacadista na economia interna do país, envolvendo produtos produzidos internamente e importados, com abrangência nacional;
- O Índice de Preços ao Consumidor, calculado pela FGV, mede, a partir de 1989, a inflação para famílias, com abrangência nacional. Os preços pesquisados pertencem a uma cesta de consumo de famílias com renda de até trinta e três salários mínimos, sendo a cesta composta por 432 produtos, com denominações de **IPC-BR** e **IPC-M**. O primeiro compõe o Índice Geral de Preço – Disponibilidade Interna (IGP-DI), e o segundo compõe o IGP-M, com participação de 30%. Em 1994 houve uma nova reestruturação na estrutura do índice, sendo então, a cesta composta por 381 produtos.
- O **INCC**, calculado pela FGV, mede a variação de preços de uma cesta de produtos e serviços atualizados pelo setor de construção civil. Este índice é calculado para três tipos de construções diferentes, com abrangência nacional.

Portanto, conforme acima descrito, podemos definir o IGP-M como o índice que mede a variação de preços no mercado de atacado, de consumo e construção civil, pois é formado pela soma ponderada de outros três índices.

O índice é elaborado pela Fundação Getúlio Vargas, com a mesma abrangência geográfica do IPA-M e INCC-M, que são pesquisados nas principais capitais do país, e do IPC-M, que abrange os municípios do Rio de Janeiro e São Paulo. A apuração do índice é efetuada em três etapas: 1º decêndio, 2º decêndio e 3º decêndio. O 1º decêndio compara os preços dos primeiros 10 dias do período e os preços dos 30 dias do período anterior. O 2º decêndio

compara os preços dos primeiros 20 dias do período e os 30 dias do período anterior. Já o 3º decêndio compara os preços dos 30 dias do período e os 30 dias do período anterior. Portanto, os dois primeiros decênios são considerados resultados parciais, e o 3º é o resultado definitivo do índice do mês.

O IGP-M considera todos os produtos disponíveis no mercado, inclusive o que é importado, diferentemente do Índice Geral de Preço – Disponibilidade Interna (IGP-DI), que considera somente os produtos produzidos e comercializados no mercado. Ele também difere do IGP-DI pelo período de coleta.

Esse índice foi criado com o objetivo de ser um indicador confiável para as operações financeiras, especialmente para as de longo prazo, sendo utilizado para correções de Notas do Tesouro Nacional (NTN) dos tipos B e C e para os Certificados de Depósito Bancário (CDB) pós-fixados com prazos acima de um ano. Posteriormente, passou a ser o índice utilizado para a correção de contratos de aluguel e como indexador de algumas tarifas como energia elétrica.

1.13.5 Índice Geral de Preços - Disponibilidade Interna (IGP-DI) da Fundação Getúlio Vargas (FGV)

O IGP-DI, Índice Geral de Preços – Disponibilidade Interna, é realizado pela Fundação Getúlio Vargas, com abrangência geográfica do IPA-M e INCC-M e IPC-BR que são pesquisados nas principais capitais do país, todos com abrangência nacional.

O IPA-DI e o IPC-BR não consideram em suas estimativas produtos importados e comercializados no país.

Esse índice foi instituído em 1944, com a finalidade de medir o comportamento de preços, em geral, da economia brasileira. Trata-se da média ponderada de seus três índices: IPA-DI, IPC-BR e INCC, com pesos de 60%, 30% e 10%, respectivamente. Difere do IGP-M especialmente pela periodicidade de coleta que, nesse caso, coincide com o mês calendário. Além disso, os índices que o compõem são estimados somente para produtos produzidos e comercializados no país.

A metodologia DI (ou Disponibilidade Interna) é a consideração das variações de preços que afetam diretamente as atividades econômicas localizadas no território brasileiro. Não se considera a variação de preços dos produtos exportados, que é considerada somente no caso de variação no aspecto de Oferta Global.

O IGP-DI mede a variação dos preços conforme acima descrito no período do primeiro ao último dia de cada mês de referência. Portanto, este índice mede a variação de preços de um determinado mês por completo. O IGP-DI/FGV é calculado mensalmente pela FGV.

Palavra do Professor

Agora que você terminou a Unidade, não esqueça de assistir à Videoaula 1. Ah! Lembre-se também de resolver as Atividades complementares que estão no AVEA; afinal, elas farão parte da sua avaliação. Não se esqueça de que você é o principal responsável pela construção do seu conhecimento; então, pesquise sempre que for necessário, busque informações complementares nos livros e *sites* citados nas referências que estão no final do livro, interaja com os seus colegas e procure o tutor sempre que estiver com dúvidas. Bom trabalho!

Atividade de Aprendizagem – 1

1) Considere as estimativas dos índices de quantidades fixo e móvel do Exemplo 3 do material complementar e verifique se a definição de número índice simples de quantidade satisfaz as propriedades de Fisher de número índice.

2) Os índices abaixo foram determinados tomando-se como base o ano imediatamente anterior:

Anos	2005	2006	2007	2008
Índice	104	103	103,8	107,2

Com base em 2004, qual será o índice de 2008?

3) Considere o quadro abaixo e complete os valores nas células vazias.

Quadro – Mudança de base de 1998 a 2008

ANO	ÍNDICE PREÇO (1998=100)	ÍNDICE NOVO **(2003=100)**
1998	100,0	
1999	104,0	
2000	108,0	
2001	113.2	
2002	117,4	
2003	119,4	100,0
2004		106,0
2005		109,6
2006		111,4
2007		116,2
2008		122,3

4) O índice constante do quadro abaixo foi calculado com base móvel.

ANOS	2003	2004	2005	2006
Índice	105	103	106	104

A partir dos dados acima, calcule:

a) a) O índice de 2006 com base em 2004.

b) b) O índice de 2005 com base em 2006.

c) c) O índice de 2006 com base em 2002.

d) d) O índice de 2002 com base em 2004.

5) Selecione no *site* www.ipea.gov.br (clique, consecutivamente, nos links: ipeadata, macroeconômico, temas, preços) três séries de dados de índices de preços mensais (IGP-DI, INPC e IPC-FIPE), de Jan/2006 a Dez/2008. Utilize as planilhas do EXCELL para realizar as seguintes operações:

a) determine os correspondentes números índices de base fixa e justifique a escolha do período da base;

b) estabeleça uma análise comparativa de evolução de preços descrita pelos números índices estimados e comente o porquê da diferença entre eles;

c) a partir dos números índices de base fixa, determine os números índices de base móvel.

6) Os dados apresentados no quadro abaixo são referentes aos preços médios dos meses de abril e maio de 2006, de 10 produtos de alimentação. Portanto:

a) calcule os números índices simples agregado de preços de Badstreet-Dutot, média harmônica, média geométrica e média aritmética (Sauerbeck);

b) em seguida, comente os resultados especificando as vantagens e desvantagens de cada um desses índices e especifique quais são os critérios que permitem estabelecer a escolha de cada um.

PRODUTOS	ABRIL /2006 (R$)	MAIO /2006 (R$)
Coxão mole (kg)	8,80	8,70
Arroz (kg)	1,08	1,35
Feijão (kg)	2,78	3,10
Óleo de soja (900ml)	2,43	2,85
Leite em pó (454 g)	5,38	5,95
Margarina (250g)	1,80	1,92
Bolacha água e sal (200g)	1,57	1,62

PRODUTOS	ABRIL /2006 (R$)	MAIO /2006 (R$)
Macarrão com ovos (500g)	2,10	2,38
Farinha de trigo (kg)	1,60	1,63
Farinha de mandioca (500g)	1,97	2,00

7) Vamos supor que dispomos de informações sobre os preços e as quantidades de cinco produtos de alimentação, conforme apresentado no quadro abaixo.

T	ARROZ		FEIJÃO		PÃO		LEITE		OVOS	
	Q(KG)	P (R$)	Q(KG)	P (R$)	Q(KG)	P (GRAMAS)	Q(L)	P (R$)	Q(DZ)	P (R$)
2000	9	2,36	5	2,75	14,0	1,75	30	1,05	8	2,48
2001	8,5	2,45	5	3,05	13,5	1,90	30	1,05	8	2,58
2002	7,5	2,70	4,5	3,20	13,5	2,00	29	1,20	7,5	2,68
2003	7,5	2,80	4,5	3,45	12,5	2,30	29	1,25	7,5	2,75
2004	7,0	2,85	4,0	3,53	12,0	2,50	28	1,40	7	2,80
2005	7,0	3,00	3,5	3,60	12,0	2,80	27	1,55	7	3,10

Obtenha a série de índices de preços utilizando as seguintes fórmulas:

a) de Laspeyres;

b) de Paasche;

c) de Divisia;

d) de Fisher;

e) de Drobish;

f) de Marshal-Edworth.

g) Verifique se estes números satisfazem as propriedades de inversão e circular e decomposição das causas;

h) Comente as vantagens e desvantagens de cada um desses números índices.

8) Considere os dados apresentados no quadro abaixo, que representam evoluções de preço e quantidade para uma cesta de quatro produtos, no período de 2004 a 2007, e determine os seguintes números índices de base móvel e base fixa:

a) o índice de preço de Laspeyres de base móvel;

b) o índice de preço Theil;

c) o índice de preço de Bureau;

d) os índices de preço e quantidade de Paasche de Base Móvel;

e) oÍndice de Preço e quantidade de Paasche (2004 = 100).

f) A partir dos resultados de (d), obtenha o índice com base fixa, comparando-o com o resultado do item (e) e comente: por que não são iguais?

g) Compare os resultados dos itens (a) a (d) e comente: por que não são iguais? Quando se deve utilizar uma ou outra entre essas formulações? Os índices satisfazem as propriedades circulares e de inversão? Eles satisfazem as propriedades de decomposição dos fatores?

ANOS	2004		2005		2006		2007	
Produtos	P	Q	P	Q	P	Q	P	Q
A	3	5	4	6	5	8	10	8
B	2	3	2	5	3	6	4	10
C	4	6	5	8	6	12	8	15
D	1	5	1	5	2	6	2	12

9) O quadro abaixo fornece dados relativos à evolução do PIB de uma economia hipotética e o deflator implícito da renda, no período de 1997 a 2006. Portanto:

a) calcule o PIB real no período de 1997 a 2006, a preços de 2006;

b) calcule a taxa real de variação anual do PIB (%), no período de 1997 a 2006; e

c) calcule a taxa média real anual de variação do PIB no período 1997 a 2006.

ANO	PIB R$ MILHÕES CORRENTES	DEFLATOR IMPLÍCITO (VARIAÇÃO ANUAL %)
1997	208.301,00	5,8
1998	276.807,00	7,3
1999	363.167,00	6,4
2000	498.307,00	]5,5
2001	719.519,00	5,5
2002	1.009.380,00	4,7
2003	1.560.271,00	6,9
2004	2.321.925,00	4,2
2005	3.410.019,00	3,5
2006	5.511.654,00	4,2

10) Renata aplicou, ao final de dezembro de 2002, a quantia de R$ 200.000,00 em três modalidades de papéis, recebendo, ao final de 2005, as importâncias discriminadas no quadro a seguir.

TÍTULO	VALOR DE AQUISIÇÃO (EM R$)	VALOR DE RESGATE (EM R$)
1	40.000,00	45.120,00
2	50.000,00	57.800,00
3	110.000,00	122.570,00

Os índices normalmente empregados para medir a perda de poder aquisitivo da moeda encontram-se no quadro seguinte.

ÍNDICE / ANO	ÍNDICE GERAL DE **PREÇOS DE MERCADO (IGP-M)**
2000	2,48
2001	2,97
2002	3,43
2003	3,97
2004	534
2005	6,90

Portanto:

a) calcule a taxa real de juros (*r*) de cada aplicação; e

b) calcule a taxa real média de juros para aplicação nos três títulos.

Anotações

d m a

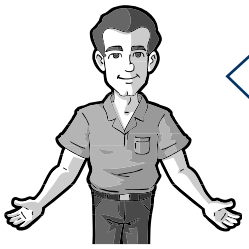
importante

2

ECONOMETRIA E ANÁLISE DE REGRESSÃO

Nesta unidade você deverá:

- entender qual a concepção metodológica que estabelece a estrutura de um modelo econométrico;
- saber especificar um modelo e a necessidade de estabelecer testes de hipótese, a fim de levantar a confiabilidade do modelo e da teoria testada;
- saber quais são os pré-requisitos matemáticos e estatísticos para o aprendizado de modelos econométricos;
- identificar a importância do computador e de *softwares* aplicados a estes fins para executar modelos econométricos aplicados a situações reais;
- entender que a principal ideia subjacente à análise de regressão é a dependência estatística de uma variável, denominada de dependente, de uma ou mais variáveis explanatórias;
- entender como estimar ou prever um valor médio da variável dependente com base nos valores conhecidos ou fixados das variáveis explanatórias;
- entender que o sucesso da análise de regressão depende da disponibilidade adequada de dados;
- entender que o conceito-chave subjacente à análise de regressão é o da função de valor esperado condicional ou função de regressão populacional;
- entender que um modelo de regressão estima parâmetros médios e, consequentemente, estima a variável independente, em seu valor médio;
- entender as diferenças entre funções de regressões populacionais e amostrais;
- entender o que são funções de regressão lineares. Ou seja, regressões lineares nos parâmetros e regressões não lineares nas variáveis;
- entender por que utilizamos, na prática, funções de regressões amostrais e não populacionais e quais são as dificuldades surgidas neste contexto.



Olá caro aluno! Pronto para iniciar o estudo da Unidade 2? Antes de iniciarmos, é necessário que você imprima o material complementar que está no AVEA. Tenha-o sempre em mãos, pois senão será muito difícil acompanhar o desenvolvimento dos conteúdos do livro! Agora, mãos à obra!

2.1 INTRODUÇÃO: MÉTODO CIENTÍFICO

Dados numéricos são, de fato, uma parte da Estatística, mas são apenas a matéria-prima, que precisa ser transformada pelos "métodos estatísticos" para posterior análise. A Estatística, como método científico, refere-se ao projeto de experimentos e à descrição e interpretação de observações que são feitas.

A pesquisa científica é um processo de aprendizagem dirigida. O objetivo dos métodos estatísticos é tornar este processo o mais eficiente possível. As características mais importantes da estatística são o uso de métodos científicos e a construção de modelos de decisão.

O método científico consiste em observar, medir, registrar e refinar os dados. Ou seja, trata-se de construir um modelo que descreva, explique e prediga o comportamento do sistema sob estudo, o que permite testar e melhorar modelos como meta de aumentar a eficiência multigerencial.

Dois importantes aspectos da ciência ilustram o método científico:

- Ele compõe um corpo de conhecimentos sistemáticos, formado de conceitos, leis, princípios e teorias usados na explicação de um conjunto de fenômenos. O corpo de conhecimento matemático é acumulado por meio de pesquisas **dedutivas** e **indutivas**, com o objetivo de explicar fenômenos.
- Ele envolve um processo de pesquisa ou um procedimento para responder a questões, resolver problemas e esquematizar melhor o procedimento.

Assim, pode-se dizer que um problema de decisão pode ser formulado e resolvido de diferentes modos, mas sempre através da aplicação da lógica simples para construir modelos matemáticos elaborados. Normalmente, o ponto de vista quantitativo é discutido através de duas aproximações:

- a **Pesquisa indutiva**, que tem como base a observação, a definição do problema, a formulação de hipóteses, o teste de hipóteses, a implementação e o controle; e
- a **Pesquisa dedutiva**, na qual um modelo geral, previamente definido por relações empíricas ou por fórmulas matemáticas, é usado para resolver um problema específico.

2.1.1 Pesquisa Indutiva

A Pesquisa Indutiva é baseada, fundamentalmente, em evidências empíricas. Ela envolve a coleta de dados em situações específicas, que devem ser testados em situações diferentes, mas com certo grau de similaridade, com o intuito de generalizar as hipóteses formuladas. A direção de inferência (análise) deve ser do **específico** para o **geral,** conforme ilustrado na Figura 2.1.

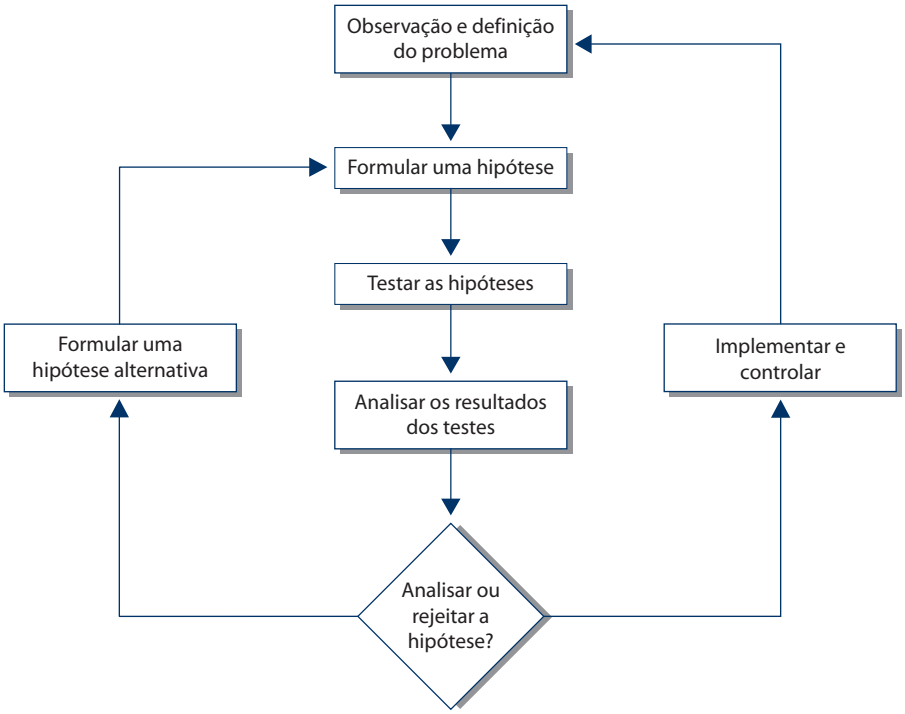


Figura 2.1 – Aproximação indutiva.

Como já foi enfatizado, a pesquisa indutiva requer que a análise seja baseada, de forma mais intensa possível, na utilização do esforço sistemático de observação, na definição concisa do problema, na formulação de hipóteses, nos testes das hipóteses e na implementação dos resultados obtidos.

2.1.2 Pesquisa Dedutiva

A pesquisa dedutiva, por outro lado, tem como meta obter conclusões que dependem de técnicas matemáticas, ao invés de evidências empíricas. Seus dados são fatos ou fenômenos para os quais existe uma teoria previamente comprovada e demonstrada.

Por exemplo, o cálculo do ponto máximo ou do ponto mínimo de uma curva (como o da função $y = 2x^2 + 3$) é obtido determinando-se a primeira derivada da função e igualando-a a zero. Assim, se aceita a verdade do procedimento geral para determinar o máximo ou o mínimo de uma curva (máximo retorno ou mínimo risco). Nenhuma medida de parâmetro é necessária; a validade da conclusão depende da validade do procedimento matemático empregado.

Na pesquisa dedutiva, parte-se de uma teoria geral para resolver um problema especifico. São utilizados axiomas, postulados ou leis determinadas cientificamente e teoremas, a fim de definir alguma proposição ou teorias gerais.

As **aproximações indutivas** e **dedutivas** relacionam-se uma com a outra e podem ser usadas separadamente, sequencialmente ou através de uma combinação básica entre elas, conforme podemos ver na Figura 2.2 abaixo.

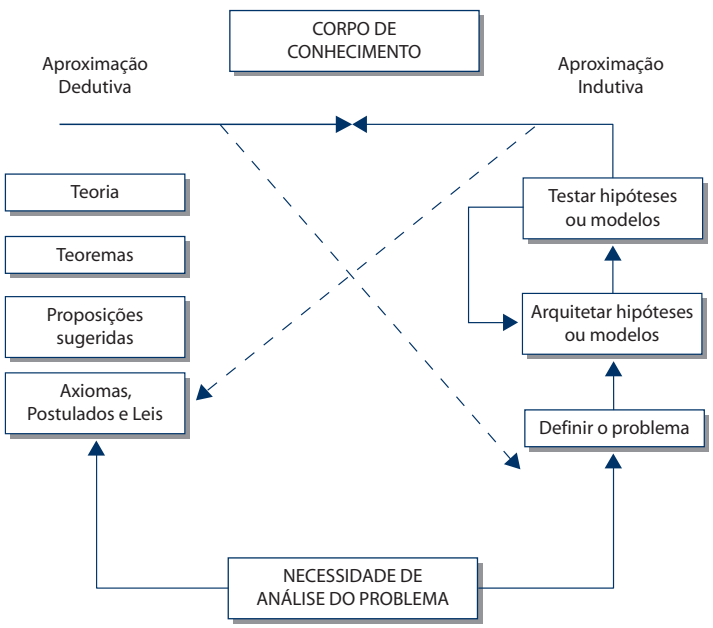


Figura 2.2 – Processo de acumulação de conhecimento sistemático.

2.2 O QUE É ECONOMETRIA?

Econometria significa medida econômica. Embora a medida econômica, por si só, seja também uma parte importante da econometria, a finalidade desta é mais abrangente. Assim, ela trata:

- da aplicação da estatística matemática aos dados, para dar suporte empírico aos modelos econômicos;
- da análise quantitativa de fenômenos econômicos concretos, baseada no desenvolvimento simultâneo da teoria e da observação, e do uso de métodos de inferência adequados;
- da determinação empírica das leis econômicas; e
- do levantamento de hipóteses econômicas apropriadas ao problema, e estabelecimento das suas comprovações.

Observação: O método de pesquisa econométrica visa, essencialmente, analisar a teoria econômica com medidas concretas, por meio da metodologia científica.

2.3 METODOLOGIA DA ECONOMETRIA

Em linhas gerais, a metodologia econométrica tradicional se utiliza da metodologia científica e é estruturada por uma combinação de procedimentos de análise indutivos e dedutivos, por meio dos seguintes passos:

1. formulação de teoria ou hipótese;
2. observação do problema, que caracteriza o levantamento dos dados, por meio de pesquisa de campo ou através de fontes secundárias de informação;
3. especificação do modelo matemático da teoria, utilizando uma relação funcional conhecida e, supostamente, considerada apropriada para o estudo do problema;
4. especificação do modelo econométrico da teoria, que se dá transformando um modelo funcional determinístico numa relação, contendo uma componente de tendência que satisfaça a relação funcional, sobreposta por uma perturbação estocástica;

5. estimativa dos parâmetros do modelo econométrico;
6. teste de hipótese;
7. previsão ou predição; e
8. utilização do modelo para fins de controle ou política.

2.3.1 Formulação da teoria ou da hipótese

Com a finalidade de evidenciar os passos metodológicos de (1) a (8) que acabamos de ver, para a estruturação de um problema econométrico clássico, importaremos um exemplo do texto de Gujarati (2006), que parte da seguinte declaração de Keynes:

> "A lei psicológica fundamental [...] é que os homens (mulheres) como regra e na média, se dispõem a aumentar seu consumo quando sua renda aumenta, mas não tanto quanto o aumento em sua renda."

Logicamente, se a nossa intenção é estudar um problema relacionando renda e consumo, então a formulação de teoria ou hipótese é definida à luz da concepção Keynesiana que, em suma, estabelece que a propensão marginal a consumir *(PMqC)*, ou a taxa de variação do consumo, é maior que zero, mas menor que 1.

2.3.2 Observação do problema levantado para a pesquisa

Normalmente, levanta-se um conjunto de observações dos valores que uma variável assume, através de pesquisas diretas, por um processo de amostragem na população foco, ou, de forma secundária, por meio de coleta em fontes especificas de informação. Esses dados podem caracterizar evoluções de comportamento em diferentes momentos, em diferentes situações ou em diferentes locais. Tais dados podem ser coletados, por exemplo:

a) em intervalos de tempo regulares para séries temporais, como diariamente (preços de ações), semanalmente (suprimento monetário fornecido pelo Banco Central), mensalmente (taxa de inflação, taxa de desemprego, etc.) ou anualmente (PIB);

b) considerando espécies diferentes (como segmentos femininos ou pessoas de cor negra) com o intuito de caracterizar um comportamento atípico para estes grupos; ou

c) em diversos locais, com o objetivo de caracterizar aspectos econômicos ou sociais diferentes.

De posse da base de dados a ser utilizada na análise econométrica, pode-se estruturar o modelo de regressão apropriado, que atenda aos requisitos pretendidos.

Utilizaremos, como exemplo dos passos de elaboração de um modelo econométrico, os dados de consumo agregado das famílias brasileiras e o PIB como *proxy* de renda agregada da economia brasileira.

Estes dados estão ilustrados no Quadro 2.1, para o período de 1990 a 2008. Estes resultados também estão ilustrados graficamente, por meio de espalhamento de pontos, na Figura 2.3, o que permite estabelecer uma relação funcional para a tendência sistemática dos dados.

Às vezes, informações sobre nossa variável explicativa não estão disponíveis por falta de estatísticas. Para solucionar problemas como este, pode ser utilizada uma variável "proxy", que é uma variável que substitui aproximadamente a que estamos procurando. Por exemplo, podemos medir a renda *per capita de* uma dada cidade (informação não disponível) pela arrecadação de impostos (imposto de renda ou imposto sobre produtos industrializados) ou ainda pelo consumo de energia elétrica. Informação disponível em: http://www.capitao.pro.br/apostilas/estatistica/E4%20-%20AN%C1LISE%20DE%20REGRESS%C3O%20SIMPLES.pdf .

Os dados foram obtidos no site www.ipea.gov.br, ipeadata.

ANO	CONSUMO (EM MILHÕES DE R$)	PIB (EM MILHÕES DE R$)
1990	9,08	11,55
1991	47,91	60,29
1992	503,68	640,96
1993	10960,00	14097,11
1994	270644,26	349204,68
1995	589145,34	705640,89
1996	715338,77	843965,63
1997	796147,11	939146,62
1998	832102,19	979275,75
1999	905549,87	1064999,71
2000	985026,00	1179482,00
2001	1084511,00	1302136,00
2002	1216102,00	1477822,00
2003	1382355,00	1699948,00
2004	1533895,00	1941498,00
2005	1721783,00	2147239,00
2006	1903679,00	2369797,00
2007	2096902,88	2597611,42
2008	2337822,61	2889718,58

Quadro 2.1 – Consumo agregado e PIB, em milhões de reais, para a economia brasileira, para o período de 1990-2008.
Fonte: www.ipea.gov.br | ipeadata.

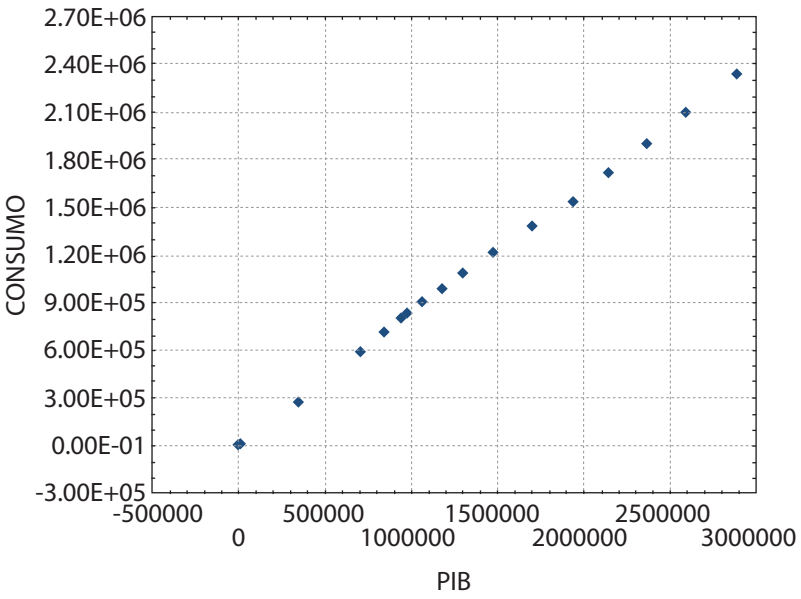


Figura 2.3 – Evolução da relação entre consumo agregado das famílias brasileiras e PIB, em milhões de R$, para o período de 1990-2008.

2.3.3 Especificação do modelo matemático do consumo

Palavra do Professor

Embora Keynes tenha postulado uma relação positiva entre consumo e renda, ele não especificou a forma precisa da relação funcional entre os dois. Como fazemos, então?

Neste caso, a especificação de um modelo matemático entre consumo e renda deve ser estabelecida de acordo com o bom senso. Assim, no sentido lógico dos princípios econômicos, a direção de causalidade é tal que **a renda causa consumo e não o inverso**, como sugere a relação funcional abaixo (ou seja, Y está relacionado com X, por meio da função f):

$$Y=f(X) \quad \textbf{(59)}$$

onde: **Y** = despesa de consumo
X = renda

A forma da função *f* (equação (59)) dependerá das características da relação entre consumo e renda, o que constitui um caso particular para cada conjunto de dados, dependendo dos hábitos de consumo da população foco, influenciada pelos seus fatores socioeconômicos.

Portanto, observando a Figura 2.3, que estabelece a evolução do consumo agregado em função do PIB, constata-se a presença de uma componente sistemática, praticamente linear, com pequenos desvios. Nesse caso, o modelo apropriado é uma função linear, como a equação (60) abaixo.

$$Y = \beta_1 + \beta_2 X, \quad como \quad 0 < \beta_2 < 1 \quad \textbf{(60)}$$

onde β1 e β2 são conhecidos como parâmetros do modelo e são, respectivamente, o **intercepto** e a **declividade**.

A equação (59) estabelece uma relação linear entre renda e consumo, que representa o modelo matemático denominado de **Função Consumo.** Como o modelo apresenta uma única equação, ele é denominado de modelo de **Equação Única**, ao passo que, se tivesse mais do que uma equação, seria um modelo de **Equações Múltiplas**.

Na equação (59), a variável que aparece à esquerda do sinal de igualdade chama-se **variável dependente** e a variável (ou variáveis) à direita chama(m)-se **variável(eis) independente(s)** ou **explicativa(s).**

No modelo da equação (60), o coeficiente de declividade β_2 mede a PMgC (propensão marginal a consumir) e o coeficiente β_1 mede o consumo autônomo. Econometricamente, a equação (60) é mostrada na Figura 2.4.

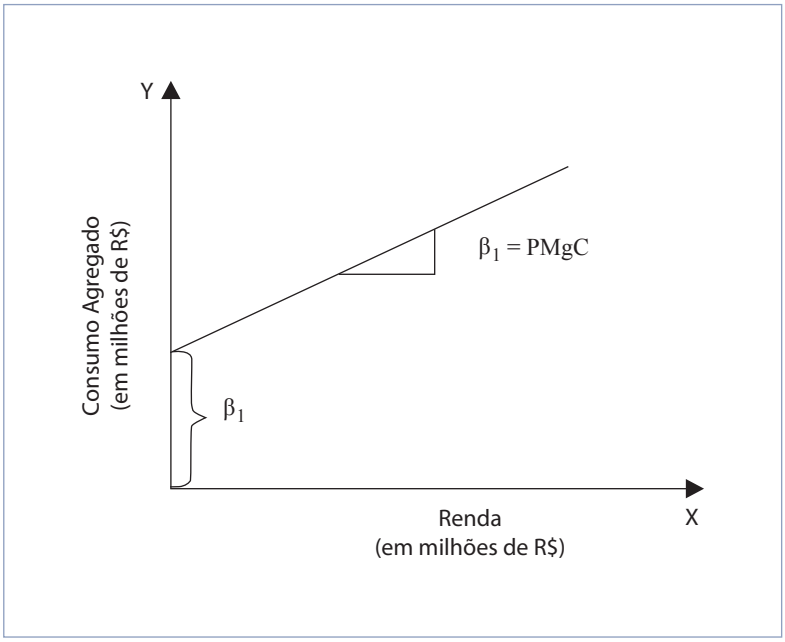


Figura 2.4 – Função Consumo Keynesiana, tendo uma componente sistemática linear.

2.3.4 Especificação do modelo econométrico de consumo

Um modelo, como aquele dado pela equação (60), tem interesse restrito para o econometricista, uma vez que supõe uma relação exata ou determinística entre consumo e renda. Mas, em geral, as relações entre as variáveis econômicas são insensatas ou aleatórias, apresentando desvios da componente sistemática, caracterizados por erros de estimativas ou fatores desconhecidos que interferem no consumo das famílias.

Assim, se considerarmos uma amostra de 500 famílias, e quisermos obter uma relação entre despesa de consumo e renda, não poderemos esperar que as 500 observações se situem exatamente sobre a reta de ajuste da equação (60). Isto porque, além da renda, outras variáveis (tamanho da família, idade de seus membros, religião, etc.) afetam o consumo. Assim, para admitir relações insensatas entre as variáveis econômicas, a função consumo determinística, dada pela equação (60), é modificada para a seguinte forma:

$$y_t = \beta_1 + \beta_2 X_t + \varepsilon_t \qquad \textbf{(61)}$$

onde ε_t é termo de perturbação ou erro, sendo uma variável aleatória (estocástica), que possui distribuição probabilística bem definida. O termo de perturbação ε_τ pode representar bem todos os fatores que afetam o consumo, mas que não são considerados explicitamente no modelo.

O modelo econométrico da função consumo, dado pela equação (61), pode ser representado pela Figura 2.5 que você verá a seguir. Esta equação é um exemplo de modelo econométrico, mas, tecnicamente, é um exemplo de modelo de regressão linear. A função consumo econométrica dada por esta equação assume que a variável dependente *Y* (consumo) relaciona-se linearmente com a variável explicativa *X* (renda), mas que a relação entre as duas não é exata, estando sujeita a variações individuais, conforme podemos ver na Figura 2.5.

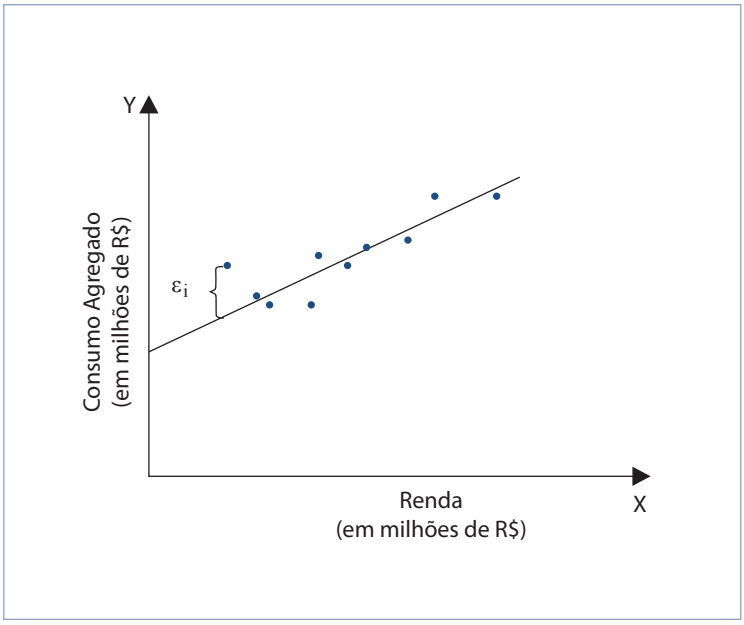


Figura 2.5 – Modelo econométrico da função consumo Keynesiana, com tendência sistemática linear e com erros estocásticos caracterizando os níveis de dispersões assistemáticos.

2.3.5 Estimativa dos parâmetros do modelo econométrico

Para a estimativa do modelo econométrico (modelo de regressão), para os dados de consumo agregado e PIB (*proxy* da renda agregada das famílias brasileiras) apresentados no Quadro 2.1, utilizaremos a função de regressão dada pela equação (61). Logicamente, esta relação se adapta perfeitamente para descrever a evolução sistemática dos dados do Quadro 2.1, conforme foi possível perceber na Figura 2.3. Devemos lembrar que, caso os dados (consumo e renda) se relacionassem descrevendo o comportamento de uma função exponencial ou logarítmica, o modelo econométrico a ser proposto deveria seguir estas relações funcionais.

A estimativa é realizada a partir da técnica estatística de análise de regressão, que é a principal ferramenta para obter modelos de previsão. Portanto, considerando que na equação (61) **Y** é o **consumo** agregado das famílias brasileiras e **X** é o **PIB** (uma medida da renda agregada), e utilizando o software Gretl 1.8, obteremos os resultados apresentados no Quadro 2.2 abaixo.

Trata-se de um pacote de softwares, com plataforma para análise econométrica, escrita na linguagem de programação de C. É livre; um software de fonte aberta. Você pode redistribuir este pacote e/ou modificá-lo sob a condição do GNU, General Public Licence (GPL) como publicado pela Fundação de Software Grátis.

	COEFICIENTE	ERRO PADRÃO	RAZÃO-T	P-VALOR	
CONST	16010	7485,68	2,1387	0,04727	**
X	0,8034	0,00441877	181,8151	<0,00001	***

Média var. dependente	967501,3	D.P. var. dependente	731564,1
Soma resíd. quadrados	8,42e+09	E.P. da regressão	22258,39
R-quadrado	0,999126	R-quadrado ajustado	0,999074
F(1, 17)	33056,71	P-valor(F)	2,11e-29
Log da verossimilhança	-216,1022	Critério de Akaike	436,2044
Critério de Schwarz	438,0933	Critério Hannan-Quinn	436,5241

Quadro 2.2 – Estimativas pelos Mínimos Quadrados (MQO) usando as 19 observações 1-19, Variável dependente: Y e Variável independente X
(Heteroscedasticidade-robusta erros padrão, variante HC1)

Mínimos Quadrados Ordinários (MQO), em inglês Ordinary Least Squares (OLS).

Observação: Teste da normalidade dos resíduos – Hipótese nula: o erro tem distribuição Normal; Estatística de teste: Qui-quadrado(2) = 0,665278, com p-valor = 0,717029

Logicamente, o modelo de regressão apresentado no Quadro 2.2 apresenta os resultados das estimativas para os coeficientes: parâmetros (coeficiente **const**, conforme especificado neste quadro) e (coeficiente de *X*). O modelo traz ainda um conjunto amplo de informações que possibilitam analisar a confiabilidade do modelo. Entretanto, essa análise não será feita no momento, servindo os resultados apresentados no Quadro 2.2 somente para efeito ilustrativo, pois, não detemos ainda os conhecimentos para tal análise.

A função consumo estimada no modelo é:

$$\hat{Y}_t = 16010,00 + 0,8034 X_t + e_t \quad \textbf{(62)}$$

onde $\hat{Y}$ é o valor do consumo agregado estimado.

O coeficiente de declividade β_2 = 0,8034 caracteriza a propensão marginal a consumir, *PMgC* (isto é, β_2 = *PMgC* ≈ 0,8034, sugere que um aumento de R$ 1,00 na renda nacional Brasileira provocará, em média, um aumento de R$ 0,80 na despesa real de consumo). O coeficiente β1, por sua vez, mede o consumo autônomo. Consequentemente, observamos que o consumo autônomo agregado da população brasileira é da ordem de R$ 16.010,00 milhões, conforme o valor estimado de β_1 = 16010,00, a preço de 1990.

A Figura 2.6 apresenta a evolução da curva de regressão, dada pela equação (62), comparativamente com os valores efetivos da relação consumo e renda, fornecidos no Quadro 2.1. Podemos ver, na figura abaixo, que se tem um modelo de regressão praticamente perfeito, com baixo nível de dispersão aleatória, dado pela distância entre a reta de regressão e os pontos efetivos. Esta distância é representada pelo erro estocástico, ε_t.

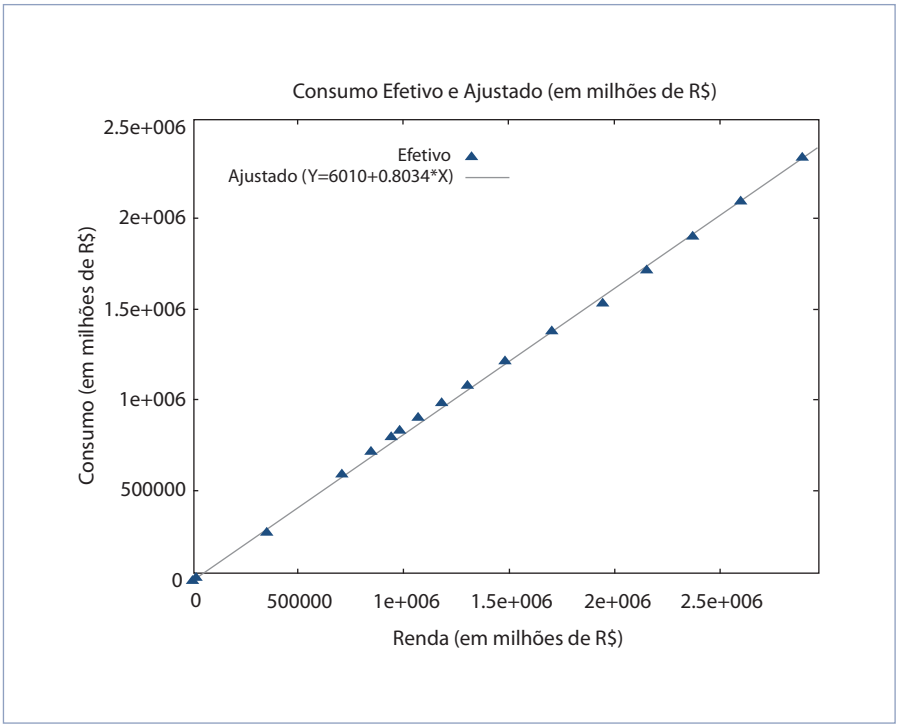


Figura 2.6 – Consumo agregado em relação ao PIB (1990-2008) e curva de regressão linear entre os dados do Quadro 2.1

2.3.6 Teste de hipótese

No momento da estimativa de um modelo econométrico, é necessário desenvolver critérios adequados para descobrir se a estimativa obtida (diga-se, aquela representada pela equação (62)) satisfaz as expectativas da teoria que está sendo testada. Uma teoria ou hipótese que não seja verificável por meio de evidência empírica, não pode ser admitida como parte de uma investigação científica. Assim, primeiramente, **(a)** devemos testar se realmente existe **uma relação de causalidade**, tal que a renda agregada cause o consumo agregado, e que haja um consumo autônomo. Estes testes são feitos através da verificação

de que os coeficientes e β_2 = 0,8034 são estatisticamente significantes (quer dizer, diferentes de zero). Em seguida, **(b)** devemos testar a teoria estabelecida como hipótese, no caso a teoria de Keynes, que supôs que 0< *PMgC* < 1. Nesse exemplo, obteve-se: *PMgC* ≈ 0,8034.

Portanto, antes de aceitarmos este resultado como uma confirmação da teoria de consumo Keynesiana, devemos averiguar se esta estimativa está suficientemente abaixo de 1 para nos convencermos de que não se trata de uma ocorrência casual ou uma peculiaridade dos dados específicos que foram utilizados.

Em outras palavras, 0,8034 é estaticamente menor que 1? Se for, pode-se sustentar a teoria de Keynes.

A confirmação ou rejeição de teorias econômicas com base na evidência de amostras é baseada num ramo da teoria estatística conhecido como **inferência estatística** (teste de hipótese), que será abordado em detalhes na Unidade 3.

Considerando os resultados do modelo apresentado no Quadro 2.2, observa-se que o *software* GRETL 1.8 já apresenta os testes de significância para parâmetros β_1 = 16010,00 e β_2 = 0,8034, por meio das estatísticas *t* e os seus correspondentes *p*_valor (probabilidades de aceitar as hipótese H_o, especificadas abaixo), o que significa verificar o item **(a)** citado acima. Esses testes de hipótese testam, respectivamente:

$$\begin{cases} H_0: \beta_1 \geq 1 \\ H_1: \beta_1 < 1 \end{cases} \quad \text{e} \quad \begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases} \qquad \textbf{(63)} \text{ e } \textbf{(64)}$$

Aceitar as hipóteses H_1 para β_1 e β_2 significa aceitar as significâncias estatísticas para estes coeficientes, com um nível de erro da ordem dos respectivos *p*_valor, estimados a partir das correspondentes estatísticas *t*, com *(n - k)* graus de liberdades (onde *n* é o tamanho da amostra (*n*=19) e *k* é o número de parâmetros estimados pelo modelo (*k*=2)).

No caso dos resultados apresentados no Quadro 2.2, observamos que o *p*_valor para β_1 foi de 0,04227, o que corresponde a um nível de erro de 4,727% ou a uma confiabilidade na estimativa de 95,273%, o que é satisfatório estatisticamente. Também observamos, no Quadro 2.2, que o *p*_valor para β2 foi de 0,000001, o que corresponde a um nível de erro menor que 0,0001% ou a

uma confiabilidade na estimativa de aproximadamente 100%, o que é bastante satisfatório estatisticamente. Portanto, concluímos que o modelo, do ponto de vista estatístico, é aceitável e realizável. Logicamente, devemos analisar o comportamento de outras estatísticas estimadas pelo *software* incluídas no Quadro 2.2; contudo, no momento, estamos apresentando estes cálculos somente para efeito ilustrativo e, portanto, não entraremos em detalhes.

A verificação do item **(b)** citado acima passa pelo fato de construir o seguinte teste de hipótese:

$$\begin{cases} H_0 : \beta_2 \geq 1 \\ H_1 : \beta_2 < 1 \end{cases} \qquad \textbf{(65)}$$

Para verificar a hipótese acima, deve-se estimar a estatística *t*, pela seguinte relação:

$$t = \frac{\hat{\beta}_2 - \beta_2}{ep\,(\hat{\beta}_2)} \qquad \textbf{(66)}$$

onde

$\hat{\beta}_2$ = valor estimado para β_2

β_2 = valor sustentado por H_0 para este parâmetro

$ep\,(\hat{\beta}_2)$ = erro padrão para β_2

Portanto, utilizando as estimativas para e para apresentadas no Quadro 2.2, que são, respectivamente, da ordem de 0,8034 e de 0,00441877, podemos estimar a relação (66) da seguinte maneira:

$$t = \frac{0{,}8034 - 1}{0{,}00441877} = -44{,}49$$

Tomando t=-44,49 e com GL=n-k=19-2=17 graus de liberdade, obtém-se um p_valor da ordem de 0,0000, o que nos induz a aceitar a hipótese H_1 de (65), com uma confiabilidade maior que 99% (praticamente 100%). Neste caso, o modelo em estudo confirma a teoria de consumo keynesiana. Portanto, de acordo com os testes de hipótese realizados, o modelo é aceitável para realizar previsões econômicas.

2.3.7 Previsão ou predição

Se o modelo escolhido confirmar a hipótese ou a teoria, pode-se usá-lo para prever os valores futuros da variável dependente (ou variável de previsão) *Y*, com base nos valores futuros conhecidos ou esperados da variável explicativa (ou preditiva), conforme foi desenvolvido de forma detalhada no subitem anterior.

Caro aluno, agora você deve consultar o Exemplo 1 do material complementar da Unidade 2, disponível no AVEA! Imprima, leia e analise o exemplo. Depois, continue a sua leitura da Seção 2.4.

2.4 TIPOS DE ECONOMETRIA

A inferência clássica assume, a priori, que os dados se distribuem normalmente, conforme a Teoria de Gauss.

A inferência bayesiana (Teorema de Bayes (Thomas Bayes (1702? - 1761)) é um tipo de inferência estatística que descreve as incertezas sobre quantidades invisíveis de forma probabilística, também conhecida como inferência não paramétrica.

A Econometria pode ser subdividida em duas grandes categorias: a Econometria Teórica e a Econometria Aplicada. Em cada categoria, pode-se abordar o assunto segundo a tradição clássica ou de acordo com a inferência bayesiana (GUJARATI, 2006).

A **Econometria Teórica** se ocupa do desenvolvimento de métodos apropriados para medir relações econômicas específicas pelos modelos econométricos. Nesse caso, a econometria deve explicitar as hipóteses deste método, suas propriedades e o que ocorre com essas propriedades quando uma ou mais hipóteses do método não são satisfeitas (exemplo, Método dos Mínimos Quadrados).

Na **Econometria Aplicada**, as ferramentas da Economia Teórica são utilizadas para estudar algumas partes específicas da economia e de negócios, tais como: função de produção, função de investimento, função de demanda e oferta e teoria de carteiras. Você verá diversos exemplos no decorrer da disciplina.

2.4.1 Pré-requisitos Matemáticos e Estatísticos

Os pré-requisitos para a econometria são conceitos básicos de **estimativa estatística** (procedimentos de amostragem, estimadores, intervalos de confiança e testes de hipótese, estatística não paramétrica). Por outro lado, os **pré-requisitos matemáticos** são conceitos básicos de matemática, tais como: diferenciação, integração, conceitos de álgebra matricial, etc.

2.4.2 O papel do computador

A análise de regressão é uma ferramenta comum na econometria, o que torna necessário proceder a um amplo conjunto de testes, tais como: intervalos de confiança, testes de hipótese, teste de causalidade de Granger, testes de verificação de autocorrelação, de homoscedasticidades, de multicolinearidade, etc, conforme será discorrido ao longo desta disciplina e da disciplina *Econometria.* Estes testes são de extrema importância na econometria. Portanto, este estudo torna-se impensável sem o uso de *softwares* aplicativos.

Na atualidade, existem vários pacotes computacionais como ferramentas de análise econométrica. Dentre eles, podemos citar os *softwares* ESTATISTICA-8, SPSS-14, E-VIEWS-6, STATA-10 e o *software* livre GRETL 1.8. Estudos econométricos complexos só se tornaram possíveis na atualidade, em função do desenvolvimento desses vários *softwares* aplicados.

Links

Você pode fazer o download do software livre GRETL 1.8.1 (versão em português) no link: < http://gretl.sourceforge.net/gretl_portugues.html >. Leia atentamente e siga os procedimentos do site. Qualquer dúvida, consulte o seu tutor.

2.5 NATUREZA DA ANÁLISE DE REGRESSÃO

A regressão é a principal ferramenta da econometria. O termo regressão foi introduzido por Francis Galton (e depois confirmada por seu amigo Karl Pearson), ao verificar que, embora houvesse uma tendência de pais altos terem filhos altos e de pais baixos terem filhos baixos, a altura média de filhos de pais de uma dada altura tendia a se deslocar ou "regredir" até certa altura média da população (GUJARATI, 2006).

Francis Galton - (1822 — 1911) foi um antropólogo, meteorologista, matemático e estatístico inglês. Curiosidade: Era primo de Charles Darwin!

A análise de regressão, portanto, se ocupa do estudo de dependência de uma variável – a variável dependente – em relação a uma ou mais variáveis – as variáveis explicativas – com o objetivo de estimar e/ou prever a média populacional ou o valor médio esperado da variável dependente, condicionado aos valores conhecidos ou fixos (em amostragem repetida) das variáveis explicativas.

2.5.1 Exemplos de dependência de uma variável em relação à outra

Modelos de regressão são intensamente aplicados em economia e têm servido de base para o desenvolvimento da teoria econômica. Neste sentido, alguns exemplos de aplicações econômicas podem ser citados:

- Um economista pode estar interessado em estudar a dependência da despesa de consumo pessoal em relação à renda pessoal real disponível, após descontar os impostos. Tal análise pode ser útil para estimar a propensão marginal a consumir. Ou seja, a variação média no consumo em relação à variação de R$ 1,00 na renda real.

- Um monopolista consegue fixar o **preço** ou o **nível** de produção, mas **não ambos**. Ele pode estar interessado em descobrir o efeito que as alterações no preço de um produto teriam na demanda. Tal experimento pode permitir a estimativa da elasticidade-preço da demanda do produto e pode ajudar a determinar o preço mais lucrativo.

- Um economista especializado em economia do trabalho pode querer estudar a relação entre a taxa de variação dos salários nominais e a taxa de desemprego. A curva a ser obtida com esses dados é um exemplo da célebre **curva de Philips**. Os resultados do diagrama de regressão obtidos podem permitir ao economista prever a variação média dos salários nominais, para uma dada taxa de desemprego. Esse conhecimento pode ser útil para afirmar alguma coisa sobre o processo inflacionário em uma economia, uma vez que os aumentos dos salários nominais provavelmente vão refletir em aumento dos preços. Tal modelo segue a tendência dada no gráfico da Figura 2.7 a seguir.

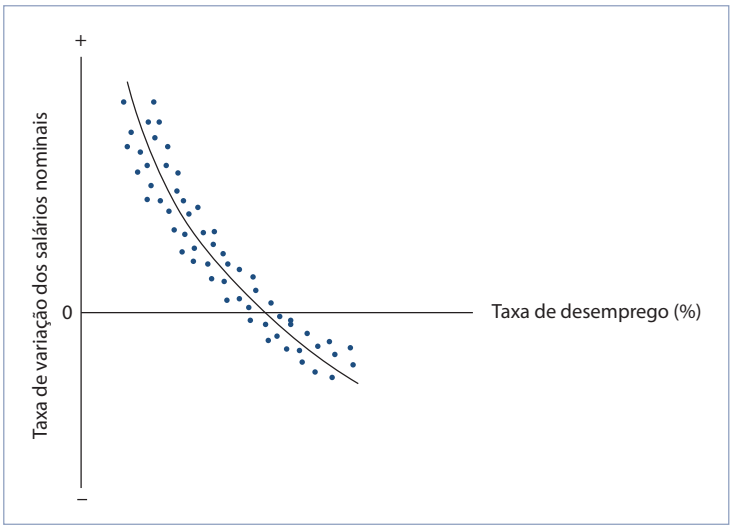


Figura 2.7 – Curva de Phillips hipotética.

- Da economia monetária, sabe-se que, tudo o mais constante, quanto maior for a taxa de inflação π, menor será a proporção *k* da renda que as pessoas desejarão reter sob a forma de dinheiro, como mostra a Figura 2.8 abaixo. Uma análise quantitativa dessa relação permitirá ao economista prever o montante de moeda, como a fração da renda, que as pessoas estarão dispostas a reter em diferentes taxas de inflação.

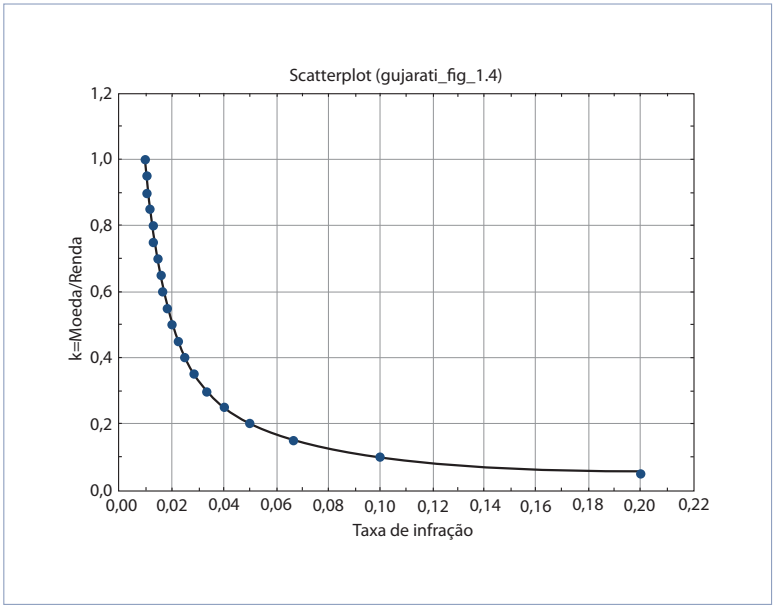


Figura 2.8 – Retorno de moeda em relação à taxa de inflação π.

2.5.2 Relações estatísticas *versus* deterministas

A análise de regressão se interessa pela dependência estatística entre as variáveis, mas não pelas dependências funcional ou determinista.

Nas relações estatísticas entre as variáveis, lida-se, basicamente, com variáveis aleatórias ou estocásticas. Ou seja, aquelas que têm distribuições de probabilidade. Por outro lado, na dependência funcional ou determinista, também se lida com variáveis, mas que não são aleatórias e nem estocásticas.

Palavra do Professor

Fenômenos de natureza estatística encontram-se ilustrados de forma clara em Gujarati (2004), que evidencia esses fenômenos através do Exemplo 2 do seu material complementar. Confira!

2.5.3 Regressão *versus* Causação

Embora a análise de regressão lide com a dependência de uma variável em relação a outras variáveis, ela não implica, necessariamente, em causação.

Segundo Kendal e Stuart (1977), "uma relação estatística, por mais forte e sugestiva que seja, jamais pode estabelecer uma relação causal: a ideia sobre causação deve vir de fora da estatística, emfim, de outra teoria".

No Exemplo 2 do material complementar da Unidade 2, sobre o rendimento da colheita, não existe nenhuma razão estatística para supor que a precipitação de chuva não dependa do rendimento da colheita. O fato de tratarmos o rendimento como dependente da precipitação de chuva (entre outras coisas) se deve a considerações não estatísticas: o bom senso sugere que a relação não pode ser invertida, já que não podemos controlar a chuva, variando o rendimento da colheita.

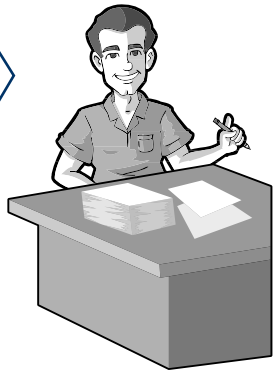
O ponto a ser destacado é que uma relação estatística, por si só, não pode implicar causação. Para atribuir causalidade, deve-se recorrer a considerações apriorísticas ou teóricas. Por exemplo, invocando a teoria econômica, pode-se dizer que o consumo depende da renda real e não o contrário.

2.5.4 Regressão *versus* Correlação

Muito diferente da análise de regressão é a análise de correlação, cujo objetivo básico é medir a intensidade ou o grau de associação linear entre duas variáveis.

Na análise de regressão não se está interessado, a principio, numa medição. Em vez disso, tenta-se estimar ou prever o valor médio de uma variável com base nos valores fixados de outras variáveis. Assim, pode-se estar interessado, por exemplo, em achar a correlação (coeficiente) entre o hábito de fumar e o câncer no pulmão, ou entre as pontuações entre os exames de estatística e de macroeconomia.

Dessa forma, você pode questionar se é possível prever a nota média em uma prova de estatística, sabendo-se a nota de um estudante em uma prova de macroeconomia. A resposta é sim, isto é possível, mas deve ficar claro que, para que haja causação (previsão), é necessário que haja, primeiro, associação entre as variáveis, ou seja, que essas sejam dependentes (correlacionadas). E, em seguida, de acordo com o seu pensamento apriorístico, baseado na sua concepção teórica ou empírica, deve-se estabelecer a direção de causalidade (no exemplo acima, conhecimento em macroeconomia causa conhecimento em estatística).



2.5.5 Diferenças fundamentais entre regressão e correlação

Na **análise de regressão** há uma assimetria na forma como as variáveis dependentes e explicativas são tratadas. Supõe-se que a **variável dependente** seja estatística, aleatória ou estocástica, isto é, que tenha uma distribuição de probabilidade. Entretanto, supõe-se que as **variáveis explicativas** tenham valores fixados (em amostragens repetidas).

Por outro lado, na **análise de correlação** tratam-se quaisquer das duas variáveis simetricamente. *Não existe nenhuma distinção entre as variáveis dependentes e explicativas.* Afinal de contas, a correlação entre as notas nas provas de macroeconomia e estatística é a mesma que a correlação entre as notas nas provas de estatística e de macroeconomia. Além disso, supõe-se que ambas as variações sejam aleatórias.

2.5.6 Terminologia e Notação

As **variáveis dependentes** são as variáveis a serem estimadas por meio de um modelo de regressão; portanto, são variáveis endógenas. Entretanto, as **variáveis explicativas** são variáveis independentes, utilizadas para predizer as variáveis dependentes. São, portanto, variáveis de controle ou instrumentais (se endógenas) ou variáveis exógenas (se não controláveis).

2.5.7 Estrutura dos dados econômicos

Os dados econômicos apresentam-se em uma variedade de tipos. Embora alguns métodos econométricos possam ser aplicados com pouca modificação para muitos tipos diferentes de informação, as características especiais de alguns dados devem ser consideradas ou deveriam ser exploradas. Veremos a seguir as estruturas mais importantes de dados encontradas.

Dados de série temporal

> Uma série temporal é um conjunto de observações dos valores que uma ou várias variáveis assumem em diferentes momentos.

Tais dados podem ser coletados em intervalos de tempo regulares, como diariamente (preços de ações), semanalmente (suprimento monetário fornecido pelo Banco Central), mensalmente (taxa de inflação, taxa desemprego, etc.) ou anualmente (PIB).

Uma característica essencial dos dados de séries de tempo, que torna mais difícil a sua análise, é o fato de que raramente é possível assumir que as observações econômicas são independentes ao longo do tempo. Outra característica que pode requerer atenção especial é a frequência dos dados. Como já vimos, preços de ações têm frequência diária ou até em minutos, taxas de inflação têm frequência mensal, e assim por diante.

A Econometria de Séries Temporais constitui uma teoria bastante específica devido à questão de estacionariedade, o que exige a aplicação de testes de verificação do comportamento das séries, no que se refere à estacionariedade e à presença de tendências.

O conceito de estacionariedade significa que "a série se desenvolve no tempo aleatoriamente em torno de uma média constante, refletindo alguma forma de equilíbrio estável" (KASSOUF apud ALMEIDA, disponível em: < http://www.scielo.br/pdf/rarv/v32n6/a15v32n6.pdf >).

Dados de corte transversal (*cross-section data*)

Um conjunto de dados de corte transversal consiste em uma amostra de indivíduos, consumidores, empresas, cidades, estados, países ou uma variedade de outras unidades, tomadas em um determinado ponto do tempo. Às vezes, os dados de todas as unidades não correspondem precisamente ao mesmo período.

Muitas famílias podem ser pesquisadas, por exemplo, durante diferentes semanas de um ano. Em uma análise pura de dados de corte transversal, ignoraríamos, na coleta de dados, quaisquer diferenças de tempo não importantes. Se o conjunto de famílias fosse pesquisado durante diferentes semanas do mesmo ano, ainda veríamos isso como um conjunto de dados de corte transversal.

Uma característica importante dos dados de corte transversal é que não se pode, frequentemente, assumir que eles foram obtidos por amostragem aleatória da população subjacente. Por exemplo, se extrairmos informações sobre salários, educação e experiência, teremos uma amostra da população que trabalha em qualquer atividade e não de uma população especifica de trabalhadores da construção civil. Devido a esta característica, os dados de corte apresentam problemas, especificamente de heterogeneidade.

Cortes transversais agrupados

Alguns conjuntos de dados têm tanto características de corte transversal quanto de séries no tempo.

Por exemplo, suponha que dois estudos sobre famílias sejam realizados no Brasil com dados de corte transversal, um em 2000 e outro em 2005. Em 2000, uma amostra aleatória de famílias é pesquisada para variáveis tais como renda, poupança, tamanho da família, e assim por diante. Em 2005, uma nova amostra aleatória de famílias é extraída, usando as mesmas questões da pesquisa do ano 2000. A fim de aumentar o tamanho da amostra, pode-se formar um corte transversal agrupado, combinando as informações referentes aos dois anos.

Agrupar cortes transversais de diferentes anos é, frequentemente, um modo eficaz de analisar os efeitos de uma nova política de governo. A ideia é coletar dados de anos anteriores e posteriores a uma importante mudança de política governamental.

Dados de painel ou longitudinais

> Dados de painel (ou dados longitudinais) consistem em uma série de tempo para cada unidade do corte transversal do conjunto de dados.

Como exemplo, suponha que tenhamos o histórico de salários, educação e emprego para um conjunto de indivíduos ao longo de um período de dez anos, ou que se possa coletar informações, tais como dados de investimentos e financeiros, sobre o mesmo conjunto de empresas ao longo de um período de cinco anos. Dados de painel também podem ser coletados para unidades geográficas. Por exemplo, podemos coletar dados para o mesmo conjunto de municípios dos estados brasileiros sobre impostos, taxas de salários, gasto governamentais, etc. para os anos 2000 e 2005.

A característica essencial dos dados de painel, que os distingue dos dados de corte transversal agrupados, é o fato de que as mesmas unidades do corte transversal (indivíduos, empresas ou municípios, etc.) são acompanhadas ao longo de um determinado período.

2.6 ANÁLISE DE REGRESSÃO DE DUAS VARIÁVEIS: ALGUNS CONCEITOS BÁSICOS

Essa teoria fundamenta a mais simples das análises de regressão, a **análise de duas variáveis**. Esse caso apresenta as ideias fundamentais da análise de regressão, de modo tão simples quanto possível. A **análise de regressão múltipla**, mais genérica, é, sob vários aspectos, uma extensão lógica do caso de duas variáveis.

> A Análise de Regressão preocupa-se em estimar ou prever a média da população ou os valores médios das variáveis dependentes, a partir dos valores conhecidos ou fixados de uma ou mais variáveis explicativas.

Para apresentar os conceitos básicos de modelo de regressão de duas variáveis, consideraremos como protótipo de análise um exemplo básico apresentado em Gujarati (2006), que trata de um conjunto hipotético de dados, referenciados como *consumo* e *renda familiar* de uma pequena população, que denominaremos como o conjunto de famílias que habitam um **bairro** que chamaremos de **A**.

Nessa análise, considere **Y** como a **despesa de consumo familiar** e **X** como a **renda familiar semanal disponível** (descontados os impostos). Especificamente, pretende-se prever o nível médio de consumo semanal de uma população, sabendo-se a renda semanal da família. Para tanto, imagine que a população do bairro A contenha somente 60 famílias. As suas rendas mensais foram subdivididas em 10 grupos, e examinaremos o consumo das famílias em cada um desses níveis de renda. Esses dados encontram-se no Quadro 2.3 abaixo.

X (EM R$) / Y (EM R$)	400,00	500,00	600,00	700,00	800,00	900,00	1000,00	1100,00	1200,00	1300,00
Y_1	275,00	325,00	395,00	400,00	510,00	550,00	600,00	675,00	685,00	750,00
Y_2	300,00	350,00	420,00	465,00	535,00	575,00	680,00	685,00	725,00	760,00
Y_3	325,00	370,00	450,00	475,00	550,00	600,00	700,00	700,00	775,00	875,00
Y_4	350,00	400,00	470,00	515,00	580,00	650,0	720,00	760,00	825,00	890,00
Y_5	375,00	425,00	490,00	540,00	590,00	675,00	725,00	785,00	875,00	900,00
Y_6		440,00		565,00	625,00	700,00		800,00	945,00	925,00
Y_7				575,00				810,00		955,00
MÉDIAS CONDICIONAIS DE Y	325,00	385,00	445,00	505,00	565,00	625,00	685,00	745,00	805,00	865,00

Quadro 2.3 – Renda (X) e Consumo (Y) familiar mensal, em R$.

Neste quadro, cada coluna fornece a distribuição do consumo Y correspondente a um nível fixado de renda X. Ou seja, ela dá a distribuição de Y, condicionada aos valores dados de X. Os dados do quadro representam a população, podendo-se facilmente, a partir dela, calcular as probabilidades condicionais de Y, denominadas p(Y/X), ou seja, a probabilidade de Y para cada X dado. Assim, temos:

$$p(Y / X) = \frac{p(Y \cap X)}{p(X)} \qquad \textbf{(67)}$$

onde:

$p(Y \cap X)$ = probabilidade da intersecção de Y com X (ou probabilidade adjunta)

$p(X)$ = probabilidade marginal de X

Para eventos independentes, como nesse caso:

$$P(Y \cap X) = P(Y) \cdot P(X) \quad \Rightarrow \quad P(Y / X) = \frac{P(Y) \cdot P(X)}{P(X)} = P(Y) \qquad \textbf{(68)}$$

Assim, no grupo de renda mensal de R$ 400,00, tem-se que:

$$P(Y_1) = \frac{1}{5} \qquad \text{e} \qquad P(Y_2) = \frac{1}{5}$$

O procedimento de aplicação da fórmula (68) está representado na ilustração esquemática da Figura 2.9 abaixo. Já os resultados das probabilidades estimadas por esta mesma fórmula encontram-se no Quadro 2.4.

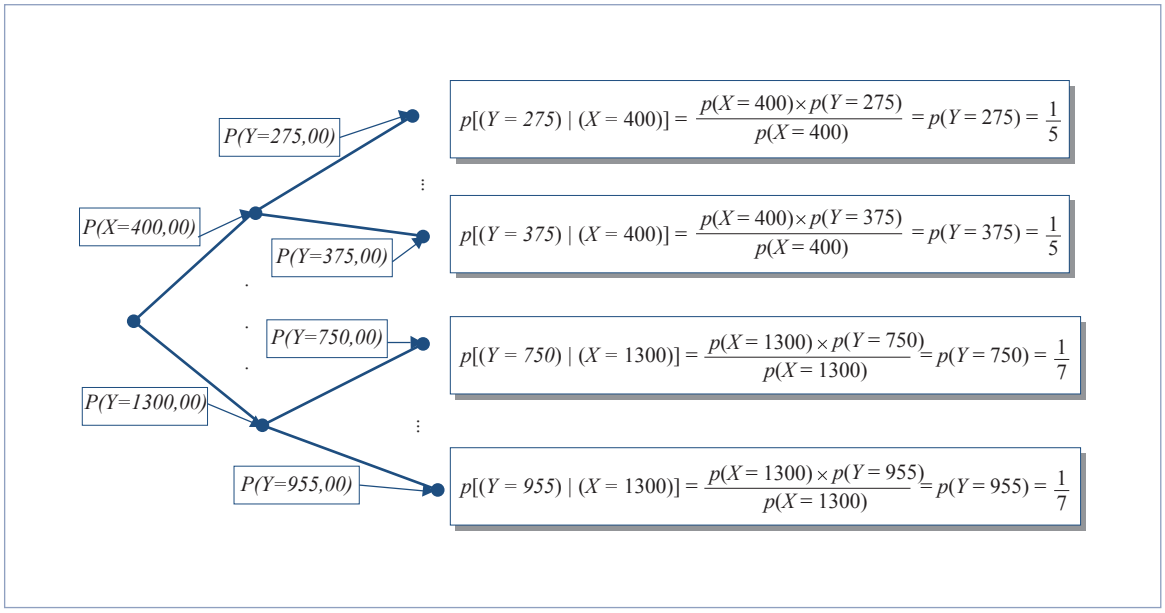


Figura 2.9 – Esquema ilustrativo de cálculo de probabilidade para os dados do Quadro 2.4.

X (EM R$) / Y (EM R$)	400,00	500,00	600,00	700,00	800,00	900,00	1000,00	1100,00	1200,00	1300,00
PROBABILIDADES CONDICIONAIS P(Y|X)	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
		1/6		1/7	1/6	1/6		1/7	1/6	1/7
				1/7				1/7		1/7
MÉDIAS CONDICIONAIS DE Y	325,00	385,00	445,00	505,00	565,00	625,00	685,00	745,00	805,00	865,00

Quadro 2.4 – Probabilidades condicionais $p(Y \mid X)$ para os dados do Quadro 2.3.

Agora, para cada uma das distribuições das probabilidades condicionais de Y dadas no Quadro 2.4, pode-se calcular o valor médio, conhecido como **média condicional** ou **expectativa condicional**, representado por E(Y/X=Xi) e lido como o **valor esperado de Y, dado que X assuma o valor específico X_i**, que, por simplicidade notacional, será escrito como E(Y/Xi).

Portanto, utilizando-se os dados de consumo do Quadro 2.3 e os dados de probabilidades do Quadro 2.4, estimam-se os valores médios esperados, por exemplo, para $E(Y = Y_1 / X = X_i)$, aplicando-se a seguinte relação:

$$E(Y / X = X_i) =$$
$$p(Y = Y_1) \cdot Y_1 + p(Y = Y_2) \cdot 3Y_2 + p(Y = Y_3) \cdot Y_3 + p(Y = Y_4) \cdot Y_4 + p(Y = Y_5) \cdot Y_5 \rightarrow$$

$$E(Y = Y_1 / X = 400,00) = \frac{1}{5} \cdot 275,00 + \frac{1}{5} \cdot 300,00 + \frac{1}{5} \cdot 325,00 + \frac{1}{5} \cdot 350,00 + \frac{1}{5} \cdot 375,00$$
$$= R\$325,00$$

Esse procedimento pode ser aplicado para estimar todos os valores médios de consumo dado a renda, mostrados nos Quadros 2.3 e 2.4.

De forma similar, a curva de regressão é, simplesmente, o lugar geométrico das médias ou expectativas condicionais das variáveis dependentes, para os valores fixados da variável ou variáveis explicativas. Ou seja, estimam-se os valores médios condicionais para o consumo, dado a renda, conforme pode ser evidenciado na Figura 2.10 abaixo.

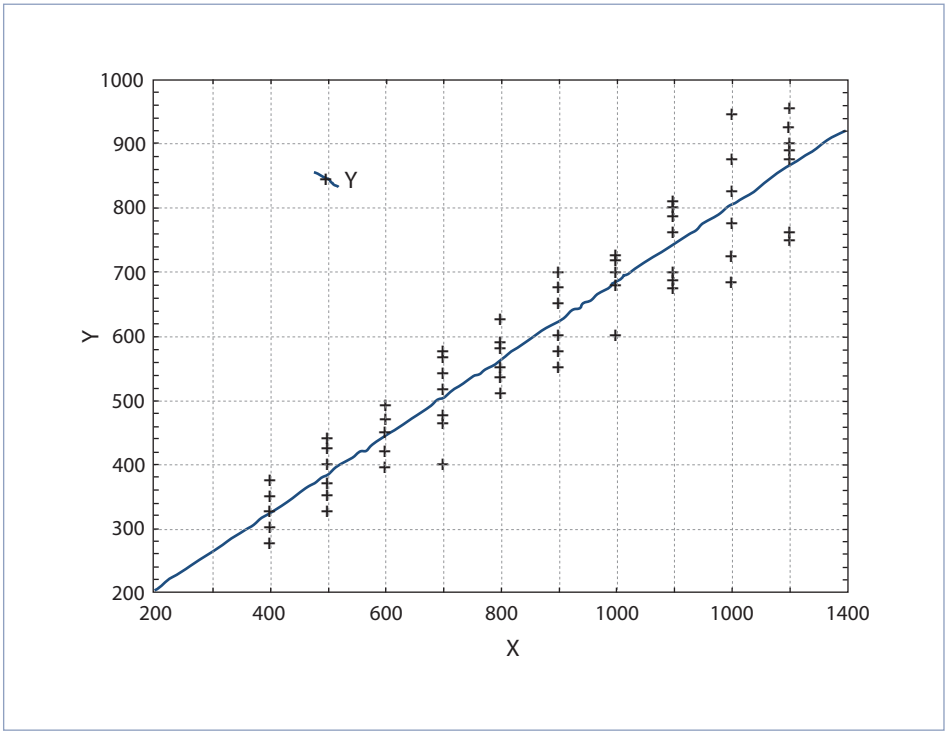


Figura 2.10 – Distribuição condicional do consumo para vários níveis de renda (dados do Quadro 2.3).

A Figura 2.10 acima apresenta a distribuição dos consumos em função da renda e da curva de ajuste do modelo de regressão, que é a seguinte:

$$Y_i = 85{,}2088 + 0{,}599 \bullet X_i + u_i \qquad \textbf{(69)}$$

Observa-se claramente, no diagrama de dispersão representado na Figura 2.10, que a distribuição de Y correspondente aos diversos valores de X (embora haja variações nas despesas de consumo das diferentes famílias), aumenta em média quando aumenta a renda. Ou seja, o diagrama de dispersão revela que os valores médios (condicionais) de Y aumentam quando X aumenta (curva de regressão da população; isto é, regressão de Y em função de X).

Conforme vimos na Figura 2.10, para cada valor de X existe uma população de valores de Y, que se supõe estarem distribuídos normalmente, e uma correspondente média (condicional). A reta de regressão passa através dessas médias condicionais.

2.6.1 O conceito de função de regressão da população (FRP)

A média condicional $E(Y/X_i)$ é uma função de X_i. Portanto:

$$E(Y/X_i) = f(X_i) \qquad \textbf{(70)}$$

onde $f(X_i)$ é função de regressão populacional de duas variáveis (FRP) que indica uma função da variável explicativa X_i. E $(Y/X_i) = f(X_i)$ expressa que a média populacional da distribuição Y, dado X_i, se relaciona funcionalmente com X_i. Essa relação funcional diz simplesmente que a média de Y varia com X.

A questão é: que forma assume a função $f(X_i)$?

Esta é uma questão importante, porque em situações reais não temos **a população inteira disponível para exame**. A forma funcional da FRP é, portanto, uma questão empírica, embora, em casos específicos, a teoria tenha algo a dizer.

Um economista pode postular que o consumo se relaciona linearmente com a renda. Então, como primeira hipótese de trabalho, podemos supor que $E(Y|X_i)$ seja uma função linear de X_i, do tipo:

$$E(Y \mid X_i) = \beta_1 + \beta_2 X_i \qquad \textbf{(71)}$$

Função de regressão linear da população (FRP)

onde β_1 e β_2 são parâmetros desconhecidos porém fixos, e são chamados de coeficientes de regressão. β_1 e β_2 são também conhecidos, respectivamente, como **intercepto** e **coeficiente de inclinação**.

Na análise de regressão, o objetivo é estimar as FRPs, ou os valores desconhecidos de β_1 e β_2, com base nas observações de Y e X. Portanto, o termo regressão, nesse estudo, significa uma regressão linear nos parâmetros, embora possa não ser linear nas variáveis explicativas.

2.6.2 O SIGNIFICADO DO TERMO LINEAR NAS VARIÁVEIS E NOS PARÂMETROS

O termo linear se caracteriza, classicamente, em funções, com o intuito de representar uma evolução de proporcionalidade entre uma variável dependente e a(s) independente(s). No caso deste estudo, o significado de linearidade nas variáveis se restringe ao fato de que a expectativa condicional de *Y, E(Y|X), é uma função linear de* X_i, como o exemplo:

$$E(Y \mid X_i) = \beta_1 + \beta_2 X_i \qquad (72)$$

Para a equação (73), geometricamente, a curva de regressão, neste caso, é uma reta (ou seja, linear nas variáveis). Neste tipo de interpretação, uma regressão como a que está abaixo não é uma função não linear nas variáveis, pois a variável X_i aparece com potência 2.

$$E(Y \mid X_i) = \beta_1 + \beta_2 (X_i)^2 \qquad (73)$$

A seguinte função também é não linear nas variáveis, pois Y_i depende de X_i de forma exponencial (é a constante neperiana).

Constante neperiana (e) é um número irracional aproximadamente igual a 2,718281828459045... , chamado do Número de Euler.

$$E(Y \mid X_i) = \beta_1 e^{\beta_2 X_i} \qquad (74)$$

A segunda interpretação de linearidade é que a expectativa condicional de *Y*, *E*(*Y*|*X*), é uma função linear dos parâmetros, sendo que *βs* pode ser linear ou não linear na variável preditiva *X*. Nessa interpretação, $E(Y \mid X_i) = \beta_1 + \beta_2 (X_i)^2$ e $E(Y \mid X_i) = \beta_1 e^{\beta_2 X_i}$ são modelos de regressão nãolineares nas variáveis, mas lineares nos parâmetros. No entanto, a relação abaixo é não linear nos parâmetros.

$$E(Y \mid X_i) = \beta_1 + \sqrt{\beta_2} X_i \qquad (75)$$

Evidenciamos que nesta disciplina serão tratados somente modelos lineares nos parâmetros, *βs*, contudo, podendo ser não lineares nas variáveis preditivas.

Portanto, a partir de agora, nesta disciplina, o termo regressão linear significará sempre uma regressão linear nos parâmetros, β_s (ou seja, os parâmetros são elevados somente à primeira potência), podendo ser ou não ser linear nas variáveis preditivas, X_s.

2.6.3 Especificação estocástica da FRP

Foi possível observar, no Quadro 15 que vimos anteriormente, que, quando a renda familiar aumenta, o consumo familiar, em média, também aumenta. Mas, o que ocorre com o consumo de uma família específica, em relação a seu nível (físico) de renda?

O que se pode dizer é que o consumo de uma família específica situa-se ao redor do consumo médio de todas as famílias com renda *X*, ou seja, em torno de expectativa condicional. Assim, pode-se expressar o desvio de uma Y_i individual em torno de seu valor esperado como sendo:

$$u_i = Y_i - E(Y / X_i) \quad ou \quad Y_i = E(Y / X_i) + u_i \qquad (76)$$

onde u_i é uma variável aleatória não observável, que pode assumir valores positivos ou negativos. Tecnicamente, u_i é termo de erro estocástico ou de perturbação estocástica.

De acordo com $Y_i = E(Y / X_i) + u_i$, podemos dizer que a despesa de uma família individual, dado o seu nível de renda, pode ser expressa como a soma de dois componentes:

- $E(Y|X)$, que é simplesmente o consumo médio de todas as famílias com o mesmo nível de renda, conhecido como componente sistemático ou determinista; e
- u_i, que é o componente aleatório, ou assistemático.

O termo **perturbação estocástica u_i** é considerado como sendo um substituto ou representante (*proxy)* de todas as variáveis omitidas ou abandonadas que podem afetar Y, mas que não estão (ou não podem ser) incluídas no modelo de regressão. Assim, se o economista pode postular que o consumo se relaciona linearmente com a renda, temos que:

$$Y_i = E(Y/X_i) + u_i \Rightarrow Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (77)$$

A equação (77) postula que o consumo de uma família se relaciona linearmente com a sua renda, mais o termo de perturbação. Se tomarmos o valor esperado dessa equação, em ambos os lados, teremos que:

$$E(Y_i / X_i) = \underbrace{E\left[E(Y / X_i)\right]}_{E(Y/X_i)} + E(u_i / X_i) \quad \Rightarrow \quad E(Y_i / X_i) = E(Y / X_i) + E(u_i / X_i) \Rightarrow$$

$$E(u_i / X_i) = 0 \qquad \textbf{(78)}$$

O valor esperado de uma constante é a própria constante. Assim:

$$E\left[E(Y / X_i)\right] = E(Y / X_i)$$

Uma vez que $E(Y / X_i) = E(Y / X_i)$, então, $E(u_i / X_i) = 0$, o que implica que o valor médio de u_i (erro estocástico) é zero. Assim, a hipótese de que a reta de regressão passa pela média condicional de Y implica que os valores médios condicionais de u_i, condicionados aos dados Xs, são zero.

A especificação estocástica tem a vantagem de mostrar claramente que, além da renda, existem outras variáveis que afetam o consumo, e que o consumo não pode ser plenamente explicado somente pela variável ou variáveis incluídas no modelo de regressão.

2.6.4 O SIGNIFICADO DO TERMO PERTURBAÇÃO ESTOCÁSTICA

O termo perturbação é um substituto de todas as variáveis omitidas do modelo, mas que, coletivamente, afetam *Y*.

Palavra do Professor

Então surge uma questão óbvia: por que não introduzir explicitamente essas variáveis no modelo?

Por que não desenvolver um modelo de regressão múltipla com o maior número possível de variáveis? São muitas as razões, que vamos ver a seguir.

a) **Imprecisão da teoria:**

» a teoria, se houver alguma que determine o componente de Y, pode ser, e frequentemente é, incompleta;

» podemos não ter qualquer dúvida de que a renda mensal X afeta o consumo Y, mas podemos ignorar ou não estar seguros sobre outras variáveis que afetam Y;

» portanto, u_i pode ser usado como substituto de todas as variáveis excluídas ou omitidas do modelo.

b) **Indisponibilidade de dados:**

» Mesmo sabendo quais são algumas das variáveis excluídas, por não ter informações quantitativas sobre elas, não se elabora um modelo de regressão múltipla. Obrigatoriamente, deverá ser elaborado um modelo de regressão simples.

c) **Variáveis essenciais versus variáveis periféricas:**

» Supõe-se que, além da renda X_i, também afetam o consumo o número de crianças da família (X_2), o sexo (X_3), a religião (X_4), o nível de educação (X_5) e a região geográfica (X_6). Mas é bem possível que a influência de todas ou algumas dessas variáveis seja tão pequena (ou aleatória) que, por questões políticas e de custos, não vale a pena introduzi-las no modelo. Espera-se que o efeito combinado das variáveis omitidas possa ser tratado como uma variável aleatória, u_i.

d) **Causalidade intrínseca no comportamento humano:**

» Existe, inevitavelmente, certa natureza aleatória intrínseca em cada Y, que, por mais que se tente, não poderá ser explicada, de modo que u_i pode muito bem representá-la.

e) **Variáveis Proxy fracas:**

» Um modelo supõe que Y e X são medidas precisas; contudo, na prática, os dados podem estar imbuídos de erros. Nesse caso, o termo perturbação pode representar o erro medido. Como exemplo, considera-se a teoria de função consumo de Milton Friedman. Ele considera o consumo permanente (Y^p) como uma função da renda permanente (X^p). Mas, como os dados

Para refrescar a memória sobre Milton Friedman, consulte o seu livro de Teoria Macroeconômica I (Unidade 3, Seção 3.4 - O consumo e as expectativas, p.90).

dessas variáveis não são diretamente observáveis na prática, utilizam-se variáveis Proxy, como o consumo atual Y, e renda atual X, que são observáveis. Uma vez que $Y \neq Y^p$ e $X \neq X^p$, há problema de erro de medida. O termo perturbação u_i pode representar esses erros.

f) **Princípio de parcimônia:**

» Consiste em deixar a regressão tão simples quanto possível, o que implica em manter o modelo o mais simples possível, introduzindo somente as variáveis explicativas mais importantes e deixando que u_i represente os efeitos de variáveis explicativas menos importantes.

g) **Fórmula funcional errada:**

» Muitas vezes, apesar de se ter dados corretos, não se conhece a relação funcional entre os regredidos e os regressores. Por exemplo, o consumo é uma relação linear ou não linear da renda? Essa questão pode ser resolvida através do gráfico de dispersão no modelo de duas variáveis, o que permite observar a relação funcional adequada. No entanto, isto não é possível num modelo de regressão múltipla, devido à dificuldade de visualizar a forma da dispersiva em domínios múltiplos.

2.6.5 Função de regressão amostral (FRA)

Nos limitamos, até o momento, à discussão sobre a população de valores Y, correspondentes ao Xs fixados, evitando, deliberadamente, considerações sobre a amostragem.

Os dados do Quadro 2.3 que vimos anteriormente representam a população e não uma amostra. Contudo, na maioria das situações práticas, tem-se somente uma amostra de valores Y correspondente a alguns Xs fixos. Por isso, a tarefa é estimar a FRP com base nas informações da amostra. Para ilustrar essa situação, considere que não conhecemos a população do Quadro 2.3 e que a única informação que temos é uma amostra de valores Y, escolhida aleatoriamente, para Xs fixos, de acordo com o que vemos no Quadro 2.5 a seguir.

Y (CONSUMO)	X (RENDA)
350	400
325	500
450	600
475	700
550	800
575	900
600	1000
700	1100
775	1200
750	1300

Quadro 2.5 – Amostra aleatória da população do Quadro 2.3.

Temos agora apenas um valor de *Y* correspondente a cada *X*, dados que foram escolhidos aleatoriamente a partir da população dada no Quadro 2.3.

A questão é: a partir da amostra dada pelo Quadro 2.5 podemos prever o consumo médio semanal *Y*, da população como um todo, correspondente aos *X*s escolhidos?

A resposta é não. Não conseguiremos estimar a FRP acuradamente, devido às flutuações da amostragem. Para observar esse fato, vamos retirar outra amostra aleatória da população do Quadro 2.3, que será apresentada no Quadro 2.6 abaixo:

Y (CONSUMO)	X (RENDA)
275	400
440	500
450	600
400	700
590	800
600	900
725	1000
675	1100
725	1200
875	1300

Quadro 2.6 – Amostra aleatória da população do Quadro 2.3.

Representando-se graficamente os dados dos Quadros 2.5 e 2.6, e traçando os diagramas das duas linhas de regressão das amostras (FRA), respectivamente, de modo a ajustá-las da melhor forma possível, observa-se uma diferença nessas duas curvas de regressão. Portanto, pergunta-se: qual das duas curvas de regressão amostral representa a verdadeira curva de regressão da população? *A priori*, não há como responder seguramente a essa questão.

As retas de regressão mostradas na Figura 2.11 são conhecidas como retas de regressão amostral. Supõe-se que elas representam a reta de regressão da população, mas, em virtude das flutuações da amostragem, elas são, quando muito, uma aproximação da verdadeira FRP. Em geral, obtém-se N diferentes FRAs para N diferentes amostras, e estas FRAs, provavelmente, também seriam diferentes.

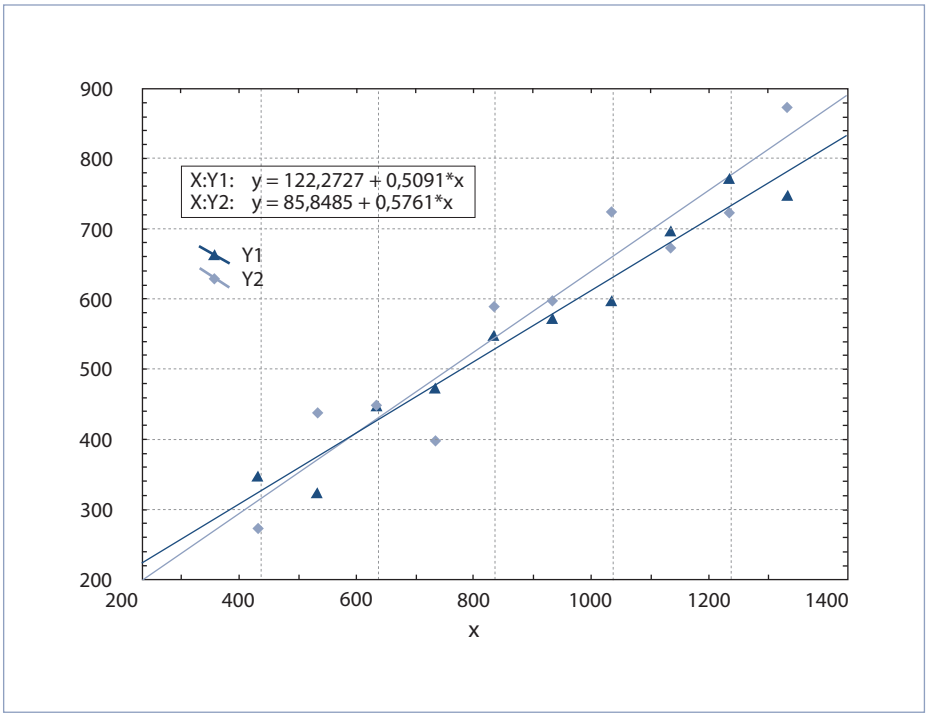


Figura 2.11 – Retas de regressão, baseadas em duas amostras diferentes

Podemos, agora, desenvolver o conceito da função de regressão amostral (FRA) para representar a reta de regressão amostral. Assim, temos que:

$$\hat{Y}_i = \hat{\beta}_i + \hat{\beta}_2 X_i \quad \textbf{(79)}$$

onde:

$\hat{Y}$ = estimador de $E(Y/X_i)$

$\hat{\beta}_1$ = estimador de $\beta 1$

$\hat{\beta}_2$ = estimador de $\beta 2$

Um estimador, também conhecido como uma estatística da amostra, é simplesmente uma regra, uma fórmula ou método que nos diz como estimar os parâmetros da população a partir das informações de cada amostra disponível. Um valor numérico particular, obtido pelo estimador em uma aplicação, é conhecido como uma estimativa. Considere que:

$$Y_i = \hat{Y}_i + \hat{u}_i \quad \textbf{(80)}$$

Ou, em função da regressão amostral (FRA):

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \quad \textbf{(81)}$$

onde $\hat{u}_i$ representa o termo resíduo (ou estocástico) da amostra e pode ser considerado como uma estimativa para u_i. Ele é introduzido na FRA pela mesma razão que u_i é introduzido na FRP.

A análise, geralmente, se baseia numa única amostra de alguma população. Por isso, em virtude das flutuações da amostragem, a estimativa da FRP, baseada na FRA, na melhor das hipóteses, constitui uma estimativa aproximada. Então, conforme a Figura 2.12, $Y_i = \hat{Y}_i + \hat{u}_i$ é a função de regressão amostral (FRA) e $Y_i = E(Y / X_i) + u_i$ é a função de regressão populacional (FRP).

Obviamente, $\hat{Y}_i$ **superestima** $E(Y|X)$ para algum X à direita do ponto A, na Figura 2.12 mostrada abaixo. Da mesma forma, $\hat{Y}_i$ **subestima** $E(Y|X)$ para algum X à esquerda do ponto A. Estas subestimações e superestimações são inevitáveis, devido às flutuações da amostragem.

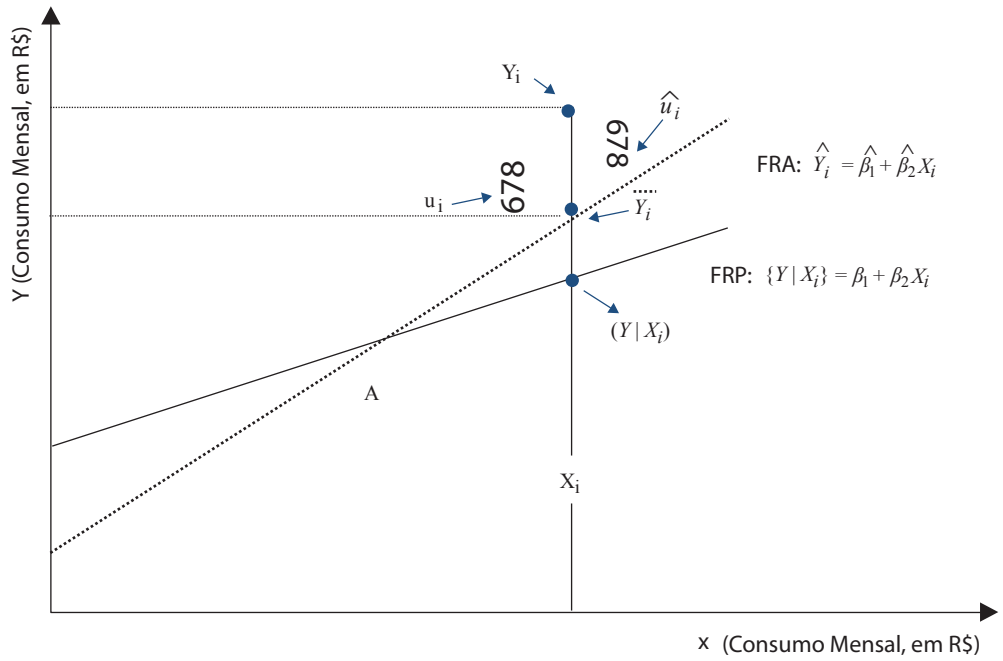


Figura 2.12 – Retas de regressão da amostra e da população.

Palavra do Professor

A partir disso, surgem as seguintes questões críticas:

Admitindo-se que a FRA seja apenas uma aproximação da FRP, pode-se criar uma regra ou um método que as torne tão próximas quanto possível?

Ou, em outras palavras, como a FRA deve ser construída para que $\hat{\beta}_1$ seja tão próximo quanto possível do verdadeiro β_1 e $\hat{\beta}_2$ seja tão próximo quanto possível do verdadeiro β_2, mesmo que nunca venhamos a saber quais são os verdadeiros β_1 e β_2?

As respostas para essas perguntas constituirão o campo de conhecimento da próxima unidade, mas já se pode dizer que é possível desenvolver procedimentos que permitam construir uma FRA que reflita uma FRP de forma bastante fiel. Isto pode ser feito mesmo que saibamos que nunca se conhecerá a FRP.

Agora que você terminou a Unidade 2, não esqueça de assistir à Videoaula 2 no AVEA. Lembre-se também de resolver as atividades complementares que estão no ambiente! Bom trabalho!

Atividade de Aprendizagem – 2

1) Imagine que você quisesse formular um modelo econômico para as atividades criminosas, ou seja, os investimentos gastos pelo sistema de segurança brasileiro no combate a essas atividades e qual seria a melhor política para combatê-las (por exemplo, investir na repressão ou investir na informação, por meio de publicidade maciça, a fim de conscientizar os jovens sobre o problema). Verifique como o seu modelo impactaria sobre as curvas de oferta e de demanda por drogas, e quais seriam os resultados esperados, de acordo com as duas possibilidades acima especificadas. Que variáveis você consideraria nessas atividades? Como seria estruturado um modelo econométrico para levantar o impacto da política de combate às drogas e quais informações deveriam ser levantadas para a análise do problema?

2) Imagine que o governo brasileiro resolveu incentivar o consumo das famílias brasileiras, diminuindo o imposto de renda para aqueles categorizados na faixa de 27,5% de contribuição, para 20%. Supondo que esta mudança causaria, de forma geral, uma redução média de impostos da ordem de 6%, que impacto esta diminuição de imposto provocaria sobre a renda para cada R$ investido? Analise o problema à luz dos passos de esquematização de um modelo econométrico.

3) Responda às questões abaixo:

a) O que é função de esperança condicional ou função de regressão da população?

b) Qual a diferença entre as funções de regressão populacional e amostral? Essa distinção causa alguma diferença nos resultados de um modelo de regressão econométrica quando se usa dados populacionais e amostrais? Quais são as implicações adicionais?

c) Qual é o papel do termo de erro estocástico u_i na análise de regressão? Qual é a diferença entre o termo de erro estocástico e o resíduo $\hat{u}_i$?

d) O que se entende por modelo de regressão linear?

4) Considere que você deseja estabelecer um modelo de regressão para o consumo, em função de diversas variáveis, como, por exemplo, $x1$ = renda, $x2$ = riqueza, $x3$ = nível de escolaridade, $x4$ = religião. Para tanto, considere os seguintes modelos:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X1_i + \hat{\beta}_3 X2_i + \hat{\beta}_4 X3_i + \hat{\beta}_5 X4_i + \hat{u}_i \qquad \textbf{(1)}$$

e

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X1_i + \hat{\varepsilon}_i \qquad \textbf{(2)}$$

De acordo com o seu entendimento de Econometria, responda às seguintes questões e justifique as suas respostas:

a) Quando seria possível optar pelo modelo (2), ao invés do modelo (1)?

b) Quando seria possível optar pelo modelo (1), ao invés do modelo (2)?

c) O que incorpora em seus valores o resíduo no modelo (1)?

d) O que incorpora em seus valores o resíduo no modelo (2)?

e) Descreva, sucintamente, qual é o significado do termo de perturbação estocástica u_i numa função de regressão e por que não introduzir explicitamente todas as variáveis explicativas num modelo. Por que não desenvolver um modelo de regressão múltipla com o maior número possível de variáveis explicativas? Ou seja, por que optar pelo modelo (2), em detrimento do modelo (1), em certas situações?

5) Determine se os seguintes modelos são lineares nos parâmetros, nas variáveis, ou em ambos. Quais destes modelos são modelos de regressão linear?

a) $Y_i = \beta_1 + \beta_2 (1/X_i) + u_i$;

b) $Y_i = \beta_1 + \beta_2 ln(X_i) + u_i$;

c) $ln(Y_i) = \beta_1 + \beta_2 X_i + u_i$;

d) $ln(Y_i) = \beta_1 + \beta_2 ln(X_i) + u_i$;

e) $ln(Y_i) = \beta_1 + \beta_2 (1/X_i) + u_i$.

Observação: *ln* = logaritmo natural (isto é, logaritmo na base, e base neperiana).

6) Tendo em mãos os dados do quadro abaixo, relativos a uma economia hipotética, no período de 1990-2006:

a) Represente graficamente a relação entre a taxa de participação dos homens na força de trabalho (TPF_H) e a taxa de desemprego civil dos homens (TD_H), utilizando o *Excel*. Trace, a olho, uma linha de regressão que passe pelos pontos. Qual seria a relação esperada, *a priori*, entre as duas variáveis e qual é a teoria econômica subjacente? O gráfico de dispersão dá apoio a essa teoria?

b) Faça a mesma coisa no caso das mulheres. Ou seja, Represente graficamente a relação entre a taxa de participação das mulheres na força de trabalho (TPF_M) e a taxa de desemprego civil das mulheres (TD_M), utilizando o *Excel*. Trace, a olho, uma linha de regressão que passe pelos pontos. Comente qual seria a relação esperada, *a priori*, entre as duas variáveis e qual é a teoria econômica subjacente. O gráfico de dispersão dá apoio a essa teoria?

c) Agora, represente graficamente a taxa de participação de homens e mulheres na força de trabalho em relação aos ganhos médios horários, GHMH_92 (homens) e GHMM_92 (mulheres), em R$, em 1992. Você pode usar gráficos separados. O que se verifica? Como você justificaria o comportamento observado?

ANO	TPF_H	TPF_M	TD_H	TD_M	GHMH_92	GHMM_92
1990	77,4	51,5	6,9	7,4	7,78	6,66
1991	77,0	52,1	7,4	7,9	7,69	7,25
1992	76,6	52,6	9,9	9,4	7,68	7,68
1993	76,4	53,9	9,9	9,2	7,79	8,02
1994	76,4	53,6	7,4	7,6	7,80	8,32
1995	76,3	54,5	7,0	7,4	7,77	8,57
1996	76,3	55,3	6,9	7,1	7,81	8,76
1997	76,2	56,0	6,2	6,2	7,73	8,98
1998	76,2	56,6	5,5	5,6	7,69	9,28
1999	76,4	57,3	5,2	5,4	7,64	9,66
2000	76,4	57,5	5,7	5,5	7,52	10,01
2001	75,8	57,4	7,2	6,4	7,45	10,32
2002	75,8	57,8	7,9	7,0	7,41	10,57
2003	75,4	57,9	7,2	6,6	7,39	10,83
2004	75,1	58,8	6,2	6,0	7,40	11,12
2005	75,0	58,9	5,6	5,6	7,40	11,44
2006	74,9	59,3	5,4	5,4	7,43	11,82

Quadro – Participação da força de trabalho de homens e mulheres numa economia hipotética.

Anotações

d m a

importante

3

MODELO DE REGRESSÃO DE DUAS VARIÁVEIS: O PROBLEMA DE ESTIMATIVA

Nesta unidade você deverá:

- entender o que é um modelo de correlação e o que é um modelo de regressão;
- saber o que são dados experimentais e dados observacionais;
- saber quais são as hipóteses do estimador clássico de mínimos quadrados ordinários (MQO) e entender como esse método é estimado;
- aprender a especificar um modelo corretamente e saber como devem se comportar os resíduos de um modelo de regressão;
- entender como são estimados os parâmetros e estatísticas de inferência num modelo de regressão;
- entender quais são as hipóteses estabelecidas sobre parâmetros e variáveis de um modelo de regressão;
- entender com estimar intervalos de confiança para os parâmetros de um modelo de regressão; e
- compreender os testes de hipóteses de intervalo de confiança de nível de significância.

Olá, caro aluno! Chegamos à última Unidade do nosso livro! Antes de iniciarmos, é necessário que você imprima o material complementar que está no AVEA. Tenha-o sempre em mãos; afinal, será muito difícil acompanhar o desenvolvimento dos conteúdos do livro sem ele! Bons estudos!

3.1 CONSTRUÇÃO DE UM MODELO DE REGRESSÃO

A primeira tarefa num modelo de regressão é estimar a função de regressão populacional (FRP), com base na função de regressão amostral (FRA). Existem vários métodos de construção da FRA; contudo, o método mais comum é o dos Mínimos Quadrados Ordinários (MQO). Discutiremos este método no âmbito do modelo de regressão de duas variáveis de uma equação única. A generalização do MQO para modelos de regressão múltipla de uma equação única será feito, simplesmente, pela extensão conceitual.

Na análise de regressão, o objetivo é obter uma equação matemática que expresse o relacionamento entre a variável dependente (resposta) e as variáveis independentes, denominadas de explicativas ou preditivas.

Ao obter uma relação funcional entre a variável dependente e as variáveis explicativas, é possível realizar previsões assumindo valores nas variáveis independentes e obtendo o valor possível da variável dependente, o que permite estabelecer a tomada de decisões.

Os modelos de regressão de uma única equação são classificados conforme o diagrama esquemático da Figura 3.1, abaixo.

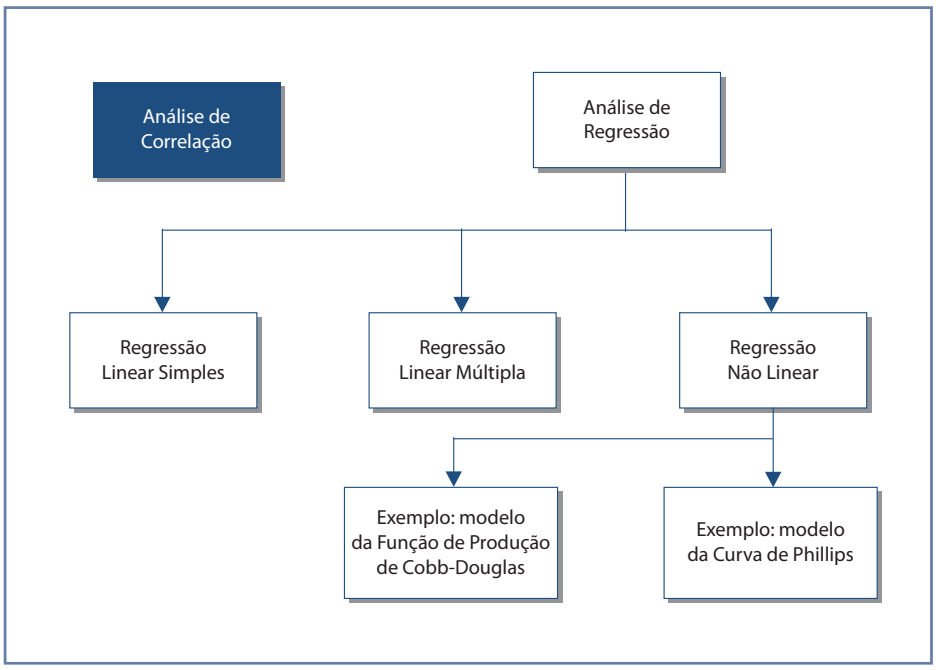


Figura 3.1 – Diagrama esquemático de modelos de regressão.

Podemos ver, no primeiro bloco da Figura 3.1, a Análise de Correlação que avalia a associação (ou dependência) existente entre duas ou mais variáveis. Discutiremos mais detalhadamente esta análise posteriormente.

Um **modelo de regressão linear simples** ou **não linear simples** se constitui numa análise de regressão envolvendo apenas duas variáveis: uma dependente e outra explanatória (isto em modelos de uma equação, que será o foco desta Unidade). Contudo, a relação funcional entre estas duas variáveis pode ser linear ou não linear, dependendo das características dos dados amostrais representativos do problema a ser estudado.

Já um **modelo de regressão linear múltipla** ou **não linear múltipla** se constitui numa análise de regressão envolvendo uma variável dependente (somente em modelos de uma equação) e três ou mais variáveis explicativas. Da mesma forma que em modelos de regressão simples, a relação funcional entre as variáveis pode ser linear ou não linear, dependendo das características dos dados amostrais representativos do problema a ser estudado.

Como exemplos de modelos de regressão não linear, pode-se citar o modelo exponencial para a função de Cobb-Douglas e o modelo da Curva de Philips, conforme vimos no diagrama da Figura 3.1. O modelo exponencial para a função de Cobb-Douglas é definido da seguinte forma:

$$Y_t = \beta_1 L^{\beta_2}_t K^{\beta_3}_t e^{u_t} \qquad \textbf{(82)}$$

onde

Y_t = produto agregado de uma economia ou de um setor de uma economia

L_t = estoque de trabalho agregado de uma economia ou de um setor de uma economia

K_t = estoque de capital agregado de uma economia ou de um setor de uma economia

β_1 = nível de tecnologia utilizado pela economia ou setor da economia em estudo

β_2 = participação marginal do trabalho no produto (ou elasticidade do produto com relação ao trabalho)

β_3 = participação marginal do capital no produto (ou elasticidade do produto com relação ao capital)

u_t = resíduo estocástico para o modelo a ser estimado

As estimativas dos parâmetros β_1, β_2 e β_3 permitem definir políticas para o setor produtivo em análise.

O modelo da Curva de Philips, por sua vez, é definido como:

$$TVS_t = \alpha_1 + \alpha_2 \frac{1}{TD_t} + e_t \qquad \textbf{(83)}$$

onde

TVS_t = taxa de variação salarial

TD_t = taxa de desemprego

α_1 e α_2 = parâmetros a serem estimados

e_t = resíduo estocástico para o modelo a ser estimado

Como você pode notar, no modelo da curva de Philips estimam-se os parâmetros α_1 e α_1 e o resíduo e_t, o que também permite definir políticas econômicas para o mercado de trabalho, determinando, em especial, a taxa natural de emprego ou a taxa de pleno emprego para a economia.

3.2 DADOS EXPERIMENTAIS E OBSERVACIONAIS

A primeira etapa para elaborar um modelo de regressão é levantar os dados, que podem ser obtidos a partir de duas situações, como veremos a seguir.

Em **dados experimentais,** as observações X e Y são planejadas, como o **resultado de um experimento**. Por exemplo: considere um experimento através do qual se pretende determinar a relação entre tempo de maturação do salame tipo italiano, em função de doses de bactérias lácticas (micro-organismos). Nesse contexto, pode-se construir uma câmera na qual se mantém a temperatura e a umidade constantes, variando somente a variável de interesse, que, nesse caso, é a dose de *starter* (bactérias lácticas), e medindo o tempo de maturação.

Para tanto, considere X como doses de *starter* e Y como o tempo de maturação do salame tipo italiano. Neste exemplo, os valores de X estão sob controle do pesquisador; ou seja, ele escolheu as doses de *starter* e observou o resultado Y, conforme podemos ver através dos dados do Quadro 3.1:

X (EM MG)	Y (EM DIAS)
100	4,00
120	3,60
140	3,20
160	3,10
180	2,90
200	2,70
220	2,60
240	2,50
260	2,40
280	2,35

Quadro 3.1 – Evolução do tempo de maturação do salame tipo italiano, em função de doses de *starter.*

A partir dos dados do Quadro 3.1, pode-se construir um modelo de regressão que permite levantar as características do tempo de maturação do salame italiano em função da dose de *starter*, conforme vemos no gráfico da Figura 3.2 abaixo.

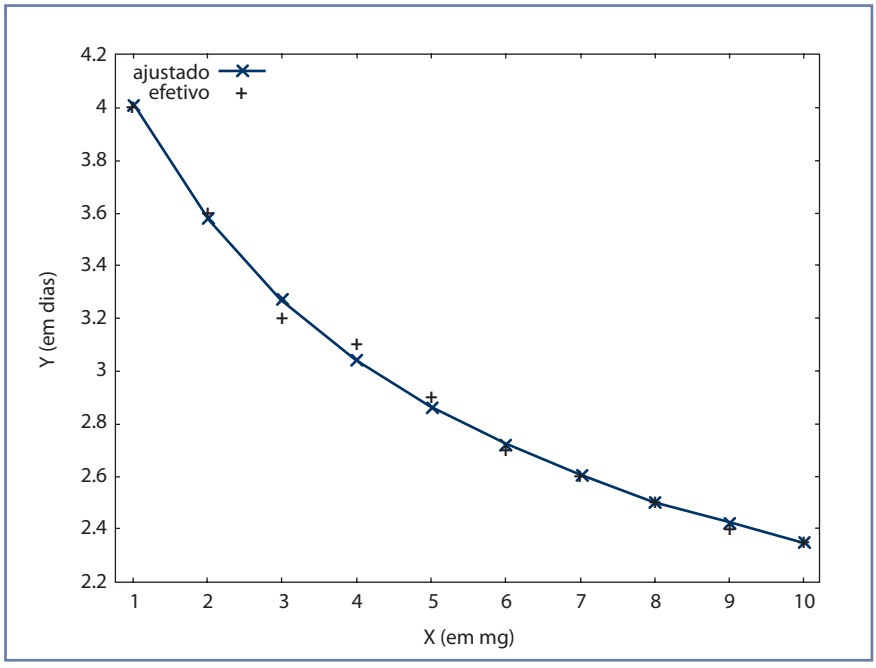


Figura 3.2 – Evolução da relação entre Y (tempo de maturação do salame Italiano), em função da dose de *starter.*

Nos **dados observacionais,** os valores de X e Y não estão sob nenhum controle, pois **não é possível estabelecer um experimento** de tal forma que seja possível controlar alguns parâmetros e liberar outros, como em experimentos ligados à engenharia. Por exemplo: a evolução do consumo, em função da renda das famílias de consumidores, pode ser levantada em uma determinada população, mas seria impossível controlar outros impactos sobre o consumo, como o grau de escolaridade, o número de filhos, etc. Além disso, as variáveis consumo (Y) e renda (X) são de caráter estocástico, o que impede ao pesquisador estabelecer qualquer controle sobre elas.

Enfim, dados estatísticos ligados às ciências sociais aplicadas são de caráter observacional, pois o laboratório de onde são extraídos os dados é a própria sociedade, não havendo um controle sobre as informações obtidas, sendo, portanto, de caráter exógeno ou não controlável.

3.3 ANÁLISE DE CORRELAÇÃO LINEAR

A análise de correlação linear é uma medida de associação (ou dependência) entre variáveis. Quando se tem apenas duas variáveis, trata-se de uma **correlação linear simples** (coeficiente de correlação linear de Pearson, *r*); e quando se tem mais de duas variáveis, trata-se de uma **correlação múltipla** (análise da matriz de correlação entre as variáveis). Neste último caso, obtém-se o **coeficiente de correlação múltiplo** (*R* múltiplo).

A interpretação que se dá a esses coeficientes é a seguinte:

- se forem zero, as variáveis não são associação; e se ±1, a associação é perfeita;
- a correlação é simétrica (correlação de XY = correlação de YX);
- apresentam insensibilidade à inclinação;
- apresentam sensibilidade aos *outliers* (pontos afastados);
- apresentam sensibilidade ao tamanho *n* da amostra; e
- assumem relações lineares.

A Figura 3.3 demonstra como os pontos se alinham quando existe correlação linear positiva ou negativa, ou sem correlação linear. Na Figura:

(a) ilustra um padrão de dependência linear positiva entre Y e X;

(b) ilustra um padrão de dependência linear negativa entre Y e X; e

(c) e (d) ilustram padrões de nenhuma dependência linear entre Y e X.

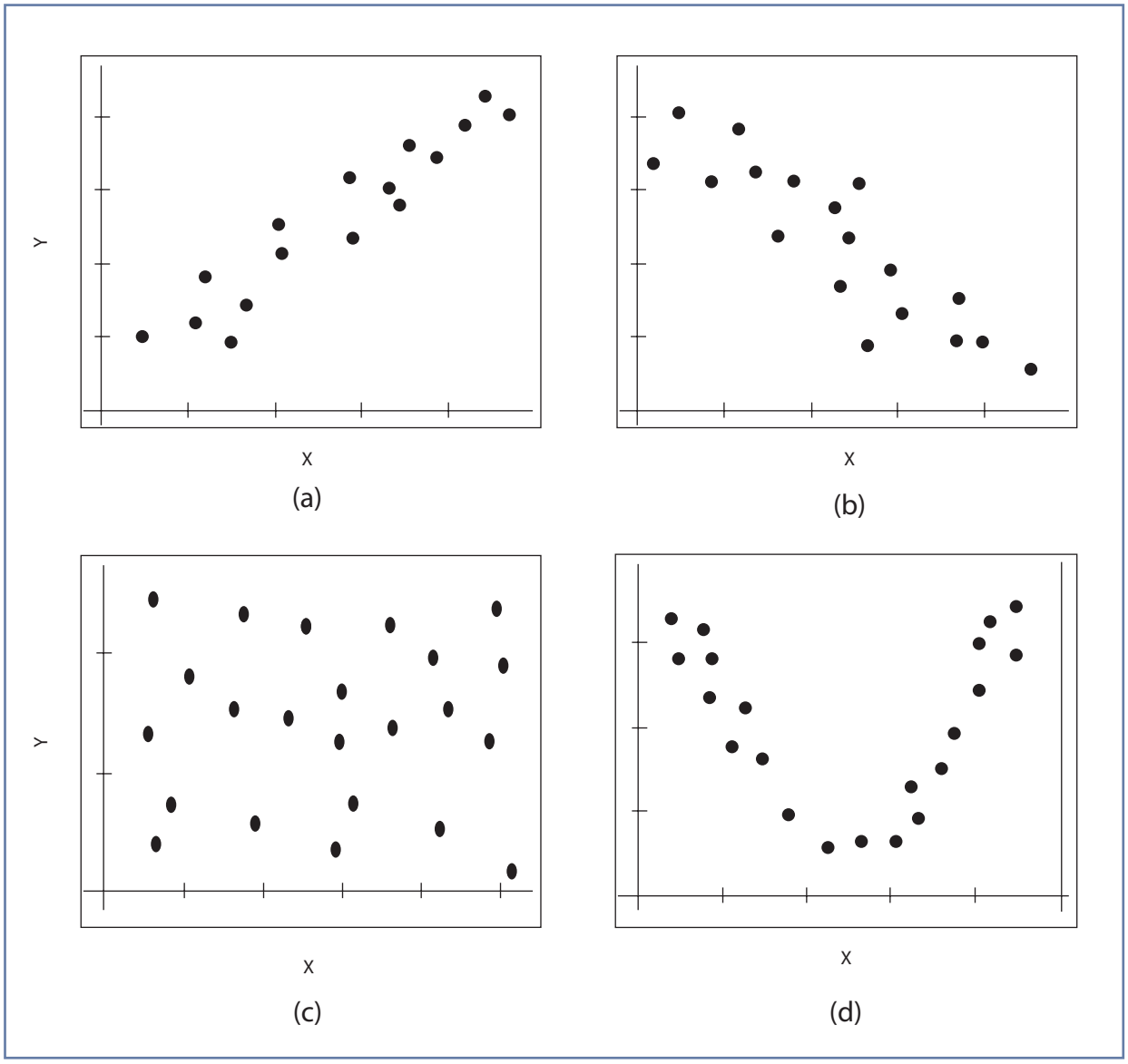


Figura 3.3 – Ilustrações de coeficientes de correlação linear entre variáveis.

A importância de entendermos a análise de correlação linear está no fato de que, para haver modelos de regressão, é necessário que haja dependência entre a variável dependente e alguma variável explicativa, indistintamente. Num modelo de regressão, torna-se necessário assumir a direção de causalidade, o que não é estabelecido em um modelo de correlação linear.

Um exemplo típico é quando um pesquisador/médico se interessa em medir a relação entre o *câncer no pulmão* e *o ato de fumar* em pacientes hospitalares. Para tanto, bastaria determinar o coeficiente de correlação. No entanto, na

determinação deste coeficiente, não fica estabelecida a direção da causa (isto é, *se câncer no pulmão causa fumar ou se fumar causa câncer no pulmão*). Com o coeficiente de correlação podemos medir a intensidade de dependência entre estas variáveis, mas não podemos definir a direção da causa (ou seja, o que causa o que), pois se trata de um modelo simétrico.

Como já verificamos na Unidade 2 (Subseções 2.5.3 e 2.5.4), estabelecer a direção da causa é importante em um modelo de regressão, por se tratar de modelos assimétricos. Mas a definição desta direção encontra-se fora da relação matemática para o modelo e está intrinsecamente ligada à percepção do pesquisador.

Os estudos de análise de correlação linear ou não linear têm importância prática em várias áreas da economia. Como exemplo, podemos citar o campo de mercado de capitais, no qual as estimativas dos coeficientes de correlação constituem um passo inicial para determinar uma curva de eficiência de mercado. Para ilustrarmos a determinação do coeficiente de correlação de Person *r* (definido pela equação (84)), utilizaremos os dados de ação da Petrobras (variável Y1), da variação do dólar (Y2) e o índice IBOVESPA (variável X), conforme podemos ver no Quadro 3.2 a seguir.

$$r = \frac{Cov(X,Y)}{s_X \times s_Y} = \frac{\dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\times\left(Y_i - \overline{Y}\right)}{n-1}}{s_X \times s_Y} \qquad \textbf{(84)}$$

onde

$Cov(X,Y)$ = covariância entre X e Y

s_X e s_Y = desvios padrão das amostra X e Y

X_i e Y_i = amostras

$\overline{X}$ e $\overline{Y}$ = médias das amostras X e Y

n = tamanho das amostras

Portanto, observando a equação (84), verificamos que o coeficiente de correlação não é nada mais do que a covariância normalizada, tal que o coeficiente de Pearson apresenta amplitude no intervalo $-1 \leq r \leq 1$.

DIA	Y1 (%)	Y2 (%)	X (%)	DIA	Y1 (%)	Y2 (%)	X (%)
15/04/2009	-1,70	2,10	-0,32	26/03/2009	-0,19	-1,12	1,89
14/04/2009	-0,94	0,37	-1,25	25/03/2009	0,26	2,55	0,78
13/04/2009	-0,16	-0,91	1,00	24/03/2009	-1,65	1,56	-2,27
09/04/2009	4,43	0,96	3,07	23/03/2009	6,04	-0,13	5,89
08/04/2009	-0,30	0,51	0,82	20/03/2009	-0,51	-0,93	-0,93
07/04/2009	-1,05	-0,41	-0,78	19/03/2009	3,35	0,35	0,78
06/04/2009	-0,36	-1,00	-0,50	18/03/2009	1,14	0,04	1,60
03/04/2009	-0,32	-1,39	1,50	17/03/2009	2,90	0,63	2,34
02/04/2009	3,74	0,45	4,19	16/03/2009	-1,84	-2,61	-1,05
01/04/2009	2,97	0,68	2,57	13/03/2009	0,54	0,70	-0,35
31/03/2009	-0,79	-1,74	0,67	12/03/2009	1,24	0,79	0,89
30/03/2009	-2,77	-2,01	-2,99	11/03/2009	0,51	-1,69	0,03
27/03/2009	-2,53	-0,52	-1,60	10/03/2009	5,33	-0,95	5,59

Quadro 3.2 – Retornos da ação da Petrobras PN (Y1), variação do dólar (Y2) e Índice IBOVESPA (X), entre os dias 10/03/2009 e 15/04/2009.

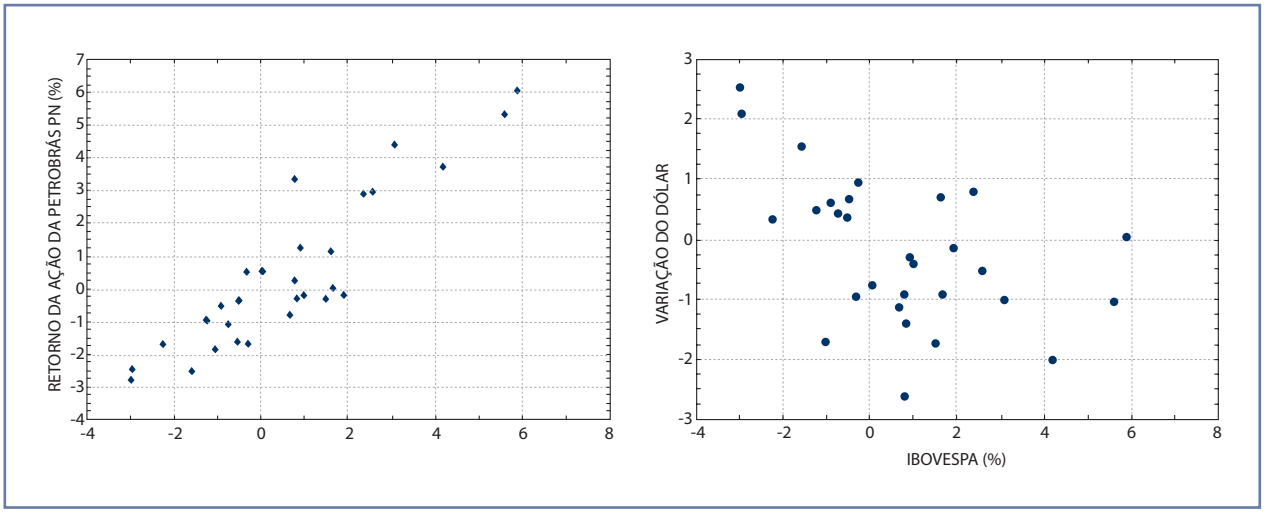


Figura 3.4 – (a) Relação de dependência entre a ação da Petrobras PN e o Índice IBOVESPA, e (b) entre a variação do dólar e o Índice IBOVESPA, no período 10/03/2009-15/04/2009.

A Figura 3.4 (a) apresenta a relação de alinhamento entre os dados da ação da Petrobras e o índice IBOVESPA, demonstrando a existência de correlação linear positiva quase perfeita. Por outro lado, a Figura 3.4 (b) ilustra um padrão de dependência linear negativa entre a taxa de variação do dólar e o

índice IBOVESPA, demonstrando uma dependência linear negativa, mas não perfeita, pois os pontos se espalham de forma acentuada.

Aplicando-se a fórmula (84) aos dados do Quadro 3.2, obtém-se um coeficiente de correlação Pearson r, primeiro entre os dados da ação da Petrobras e do índice IBOVESPA e, posteriormente, entre os dados da taxa de variação do dólar e do índice IBOVESPA, que são, respectivamente, da ordem de 0,90 e -0,51. Estes resultados, estimados através da utilização do *software* STATISTICA 7.0, são apresentados na Quadro 3.3, abaixo.

Variable	Correlations Marked correlations are significant at p < ,05000 N=29 (Casewise deletion of missing data)		
	Y1	X	Y2
Y1	1,00	0,90	-0,48
X	0,90	1,00	-0,51
Y21	-0,48	-0,51	1,00

Quadro 3.3 – Coeficientes de correlação entre Y1, X, Y1 e Y2 e entre Y2 e X, estimados pelo software STATISTICA 7.0.

3.4 REVISÃO DA CONCEPÇÃO DE MODELOS DE REGRESSÃO

Como foi intensivamente explicado na Subseção 2.6.5, pretende-se, ao elaborar um modelo de regressão amostral, não somente estimar os parâmetros do modelo, mas também estabelecer inferências que permitam especificar o nível de erro entre os parâmetros da FRA e os verdadeiros parâmetros da FRP. Portanto, vamos rever, a seguir, as diferenças conceituais entre uma FRA e uma FRP, com a finalidade de obter fórmulas matemáticas para estimativas dos parâmetros amostrais e das estatísticas de inferências necessárias para atingir tal objetivo.

3.4.1 Uma FRP de duas variáveis

Consideraremos, para introduzir os conceitos de modelo de regressão, um modelo linear de duas variáveis para a função de regressão de uma população, denominada de FRP, descrito na Figura 3.5, a seguir:

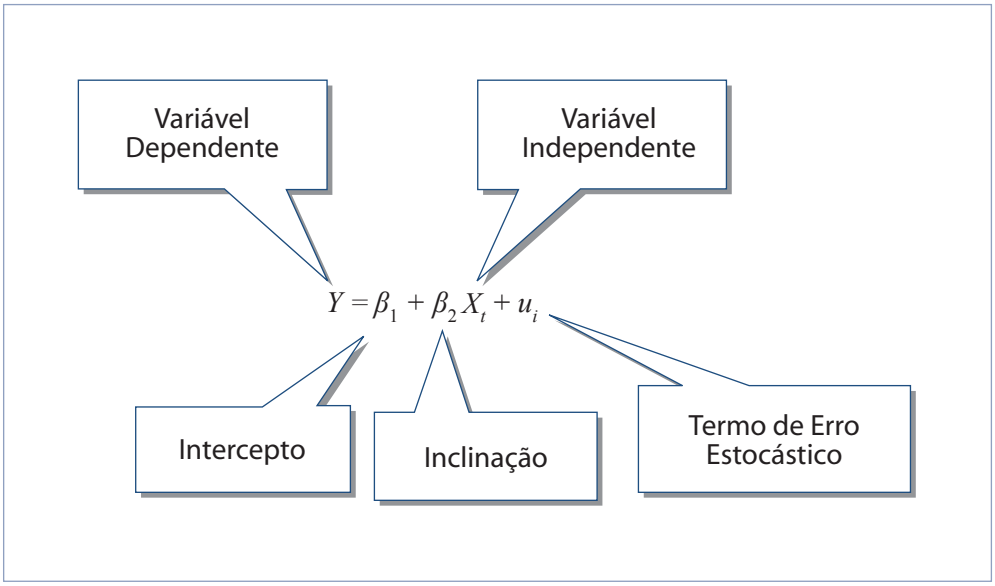


Figura 3.5 – Equação de regressão para a FRP, com as respectivas variáveis e parâmetros.

O que se pode dizer sobre o modelo de regressão para a FRP, na Figura 3.5, é que o valor específico da variável dependente (Y_t) situa-se ao redor do valor médio, fixando o valor da variável independente (explicativa); ou seja, situa-se em torno de expectativa condicional. Assim, pode-se expressar o desvio de um Y_i individual em torno de seu valor esperado, como sendo:

$$Y_t = E(Y|X_t) + u_t \qquad \textbf{(85)}$$

onde

$$E = (Y|X_t) = f(X_t) = \beta_1 + \beta_2 X_t \qquad \textbf{(86)}$$

Ou seja, de acordo com a especificação da Figura 3.5, $E(Y|X_t)$ é igual à tendência descrita pela relação funcional $f(X_t)$ que, no caso em análise, é linear, como mostra a equação (86).

Na especificação dos parâmetros a serem estimados, conforme a mesma Figura 3.5 e as equações (85) e (86), u_t é uma variável aleatória, não observável, que pode assumir valores positivos ou negativos. Tecnicamente, u_i é termo de erro estocástico ou perturbação estocástica, e os parâmetros β_1 e β_2 são parâmetros a serem estimados.

A função de regressão explica, em média, a variação da variável dependente Y em função da variação das variáveis independentes $(X_1, X_2,..., X_n)$. Considerando, a partir do modelo dado pelas equações (85) e (86), que Y_t

seja o consumo familiar e X_t a renda familiar, então, dado o nível de renda, o consumo pode ser expresso como a soma de dois componentes:

1. $E(Y|X_t)$, que é simplesmente o termo médio, ou o valor esperado de Y_t, fixada a variável independente. Este valor esperado é conhecido como componente sistemático ou determinístico; e
2. u_i, que é o componente aleatório, ou não sistemático.

Como observamos na Unidade 2 (Subseção 2.6.4), o termo perturbação estocástica (u_i) é considerado um substituto ou representante (*proxy*) de todas as variáveis omitidas ou abandonadas que podem afetar Y_t, mas que não estão (ou não podem ser) incluídas no modelo de regressão.

3.4.2 Uma FRA de duas variáveis

Como foi enfatizado anteriormente (Subseção 2.6.5), na maioria das situações práticas, tem-se somente uma amostra de valores Y correspondente a alguns Xs fixos. Por isso, a tarefa é estimar a FRP com base nas informações da amostra. Então, desenvolve-se o conceito da função de regressão amostral (FRA) para representar a regressão. Assim:

$$\widehat{Y}_t = \widehat{\beta}_1 + \widehat{\beta}_2 X_t + \widehat{u}_t \qquad (87)$$

onde

$\widehat{Y}_t$ = estimador de $E(Y/X_t)$

$\widehat{\beta}_1$ = estimador de β_1

$\widehat{\beta}_2$ = estimador de β_2

$\widehat{u}_t$ = estimador para u_i

Um estimador é também conhecido como uma estatística da amostra.

Então, ao longo desta seção, apresentaremos os mecanismos matemáticos e estatísticos que permitem estabelecer o quão próximas a FRA e a FRP se encontram; ou seja, estabeleceremos métodos e regras que detectam as proximidades entre essas funções.

Como se sabe, quando se utilizam dados amostrais, tem-se um subconjunto da população, e consequentemente, uma redução na precisão das informa-

ções inerentes à população, que são aportadas por esses mesmos dados. Esse fato nos induz a concluir que as estimativas do comportamento da FRP, a partir da FRA, são feitas com certo grau de viés. Portanto, é necessário especificar a ordem de grandeza desse viés através da especificação do nível dos erros inerentes aos dados amostrais. Além desse fato, a precisão do modelo de regressão amostral depende também do viés introduzido pelo estimador (modelo matemático utilizado para a solução do problema).

Nessa Unidade (Subseção 3.4.4) utilizaremos como estimador o Método dos Mínimos Quadrados (MQO). Para que este método seja preciso (com estimativas sem viés e de variância mínima (eficiente)), os dados utilizados devem satisfazer a algumas propriedades, e as especificações do modelo devem ser as mais corretas possíveis, caracterizadas por pressupostos (hipóteses), que você verá na Subseção 3.4.3 a seguir.

3.4.3 Hipóteses Adjacentes

No modelo de regressão, as variáveis independentes (X_1, X_2, ..., X_n) são supostas sem erro. Cada uma deve assumir pelo menos dois valores diferentes e não pode ser função linear exata de outras. A variável dependente Y_i e o erro u_i são variáveis aleatórias. Os valores do erro aleatório distribuem-se normalmente em torno de sua média. Mas o erro aleatório absorve qualquer erro de aproximação existente, capta a influência de qualquer outra variável independente não incluída no modelo e quaisquer elementos de comportamento aleatório presentes em cada elemento pesquisado.

Para que o modelo de regressão apresente resíduos normais e com média zero, e para que os demais parâmetros sejam estimados com o menor viés possível, de forma que se possam estabelecer inferências sobre estes parâmetros, devem-se estabelecer hipóteses subjacentes ao modelo clássico de regressão linear.

Se a intenção, numa análise de regressão, fosse somente estimar β_1 e β_2, o método dos MQO seria suficiente. Contudo, em análise de regressão, o objetivo não é somente obter $\widehat{\beta}_1$ e $\widehat{\beta}_2$, mas também fazer inferências a partir das estimativas, sobre os verdadeiros valores de β_1 e β_2. Por exemplo, gostaríamos de saber quão próximo $\widehat{\beta}_1$ e $\widehat{\beta}_2$ estão de suas contrapartidas na população (β_1 e β_2), e quão próximo $\widehat{Y}_t$ encontra-se do verdadeiro Y_t.

Para tanto, devemos não apenas especificar a forma funcional do modelo de regressão (ou seja, estabelecer $Y_t = f(X_t)$), mas também formular certas hipóteses sobre o modo pelo qual $\widehat{Y}_t$ é gerado.

Num modelo de regressão, a função de regressão populacional FRP (equações (85) e (86)), mostra que Y_i depende tanto de X_i como de u_i. Por isso, a menos que sejamos específicos sobre o modo pelo qual X_i e u_i foram criados ou gerados, não há como fazer qualquer inferência sobre Y_i e, tampouco, sobre β_1 e β_2.

Em resumo, as hipóteses que veremos a seguir, elaboradas sobre as variáveis X_i e o termo de erro, são extremamente necessárias para a validade da interpretação do modelo. Observe:

HIPÓTESE 1**:** o modelo de regressão deve ser linear nos parâmetros. Ou seja, ao estabelecer a relação funcional $Y_t = f(X_t, u_t)$, esta pode ser não linear nas variáveis, mas deve ser linear nos parâmetros.

HIPÓTESE 2**:** os valores da variável explicativa X são fixados em amostragem repetida (a análise de regressão é uma análise condicional aos dados do regressor X). Tecnicamente, supõe que X seja não estocástico. Assim, a regressão linear mostra relações entre os valores fixos da variável explicativa X e os valores médios de Y. Contudo, a regressão assume relações contendo elementos X estocásticos. Por exemplo: numa regressão consumo-renda, a renda, como variável explicativa, segue uma distribuição probabilística, pois suas informações são coletadas por amostragem, por meio de coleta de dados.

HIPÓTESE 3: o valor médio da perturbação (erro aleatório) u_i é zero, pois:

$$Y_t = E(Y|X_t) + u_t \rightarrow E(Y_t) = E[E(Y|X_t)] + E(u_t|X_t) \rightarrow E(Y_t) = E(Y|X_t) + E(u_t|X_t)$$

Consequentemente, como $E(Y_t) = E(Y|X_t)$, então, $E(u_i|X_t) = 0$.

Assim, dado o valor de X, o valor médio ou esperado do termo de perturbação aleatória u_i é zero.

HIPÓTESE 4: homoscedasticidade ou variância de u_i igual para todas as observações. As variâncias condicionais de ui são idênticas. Simbolicamente, tem-se:

$$var(u_t|X_t) = E(u_t - E(u_t)|X_t)^2 \rightarrow var(u_t|X_t) - E(u^2_t|X_t)^2 \rightarrow var(u_t|X_t) - \sigma^2$$

Pois, de acordo com a Hipótese 3, $E(u_t) = 0$. O símbolo σ^2 é a variância homoscedástica. Ou seja, a variância da perturbação aleatória é considerada como constante para todos os valores Y_i.

Conforme é mostrado no gráfico da Figura 3.6, abaixo, os dados da variável dependente não devem apresentar volatilidade na variância, estabelecendo a relação entre o consumo $f(Y_i)$ e a renda X_i, através da qual se observa que, ao aumentar a renda das famílias, aumenta também o consumo médio. No entanto, a dispersão do consumo (medida pela variância σ^2) em torno de seu valor médio permanece constante. Isto evidencia que os fatores aleatórios que causam a dispersão do consumo das famílias, dado a renda, são os mesmos para qualquer faixa de renda.

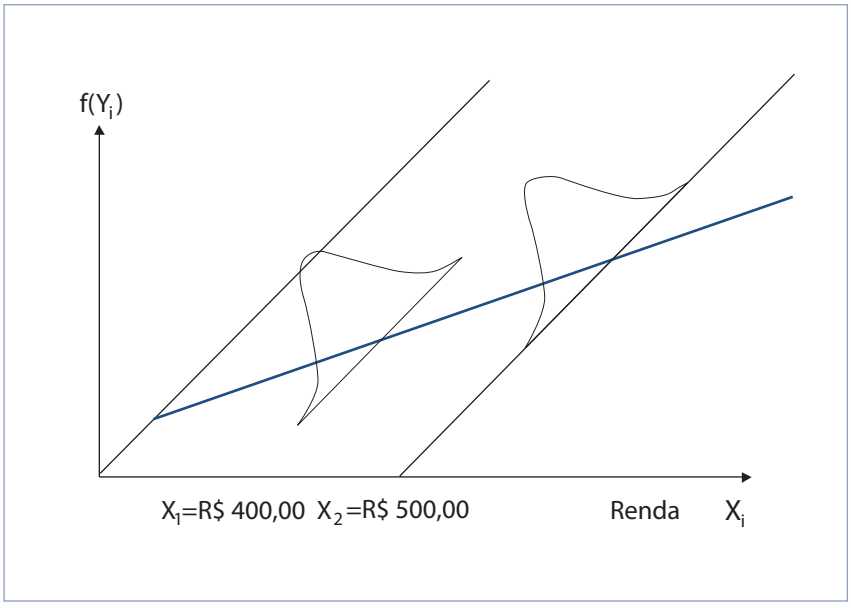


Figura 3.6 – Função de densidade de probabilidade para y_i, para dois níveis de renda mensais x_i.

HIPÓTESE 5: não há autocorrelação entre as perturbações estocásticas. Dados dois valores X quaisquer, X_i e X_j, tal que $i \neq j$, a correlação entre quaisquer dois u_i e u_j (sendo $i \neq j$) é zero. Simbolicamente, tem-se que:

$$\begin{aligned} \operatorname{cov}(u_i, u_j \mid X_i, X_j) &= E\left[u_i - E(u_i) \middle| X_i\right] \times E\left[u_j - E(u_j) \middle| X_j\right] \\ &= \mathrm{E}(\mathrm{u_i} \mid X_i) \times \mathrm{E}(\mathrm{u_j} \mid X_j) = 0 \end{aligned}$$

Afinal, $E(u_i)=o$ e $E(u_j)=o$ (Hipótese 3).

Por que isso acontece? Porque i e j são duas observações diferentes; então, a covariância (*cov*) deve ser zero para que não haja dependência.

O postulado de autocorrelação nula trata da hipótese de ausência de correlação serial. Isto significa que dado X_i, o desvio de quaisquer dois valores de Y de seus valores médios não exibe relação de causalidade. Intuitivamente, pode-se explicar a necessidade da não presença de correlação serial, considerando, por exemplo, a FRP linear abaixo:

$$Y_i=\beta_1 + \beta_2 X_i+u_i,$$

tal que, se u_i e $u_{(i-1)}$ tiverem correlação positiva, então, Y_i depende não só de X_i, mas também de u_{i-1}, já que, de certa forma, $u_{(i-1)}$ determina u_i (assim, não se restringe somente o efeito de X_i sobre Y_i).

HIPÓTESE 6: u_i e X_i apresentam covariância zero, ou seja, $E(u_iX_i)=0$. Formalmente, já que X_i não é estocástico, tem-se:

$$\begin{aligned}\operatorname{cov}(u_i, X_i) &= E[u_i - E(u_i)]\times E[X_i - E(X_i)] \\ &= \mathrm{E}[\mathrm{u_i} \times (X_i - E(X_i)] = \\ &= E(u_i X_i) - E(X_i)\times E(u_i)\end{aligned}$$

Portanto:

$$\operatorname{cov}(u_i, X_i) = \mathrm{E}(\mathrm{u_i} X_i) = E(u_i)E(X_i) = 0, \text{ já que } E(u_i) = 0 \text{ por hipótese}$$

Nesta hipótese, assume-se que a perturbação estocástica u_i e a variável explicativa X_i não apresentam correlação. Numa FRP, admite-se que X_i e u_i (representando a influencia de todas as variáveis omitidas) exercem influências separadas (e cumulativas) sobre Y_i. Dessa forma, se X_i e u_i têm correlação, passa a existir uma relação de causalidade entre eles; então, suas influencias sobre Y_i, nesse caso, não são separadas ou cumulativas. Ou seja, é difícil isolar as influências de X_i e u_i sobre Y_i.

HIPÓTESE 7: O número de observações n deve ser maior que o número de parâmetros a serem estimados. Alternativamente, o número de observações n deve ser maior que o número de variáveis explicativas.

HIPOTÉSE 8: Os valores de X, em uma dada amostra, não podem ser todos iguais. Tecnicamente, $var(X)$ deve ser um número positivo finito. Portanto, se todos os valores de X forem idênticos, então, $X_t = \overline{X}$. Assim, será impossível determinar β_2.

HIPÓTESE 9: o modelo de regressão linear está corretamente especificado? Alternativamente, não deverá existir nenhum viés ou erro de

especificação no modelo usado na análise empírica. Algumas questões relacionadas à especificação do modelo são:

- Quais são as variáveis que devem ser incluídas no modelo?
- Qual a forma funcional do modelo? Conforme a Figura 3.7, a seguir, os modelos 4, 5 e 6 são relações funcionais corretamente especificadas, que introduzem pouco viés nas estimativas.
- O modelo é linear nos parâmetros, nas variáveis ou em ambos? Todos os modelos apresentados na Figura 3.7 são lineares nos parâmetros.
- Quais são as hipóteses probabilísticas feitas sobre *Yi*, *Xi* e *ui* que entram no modelo?

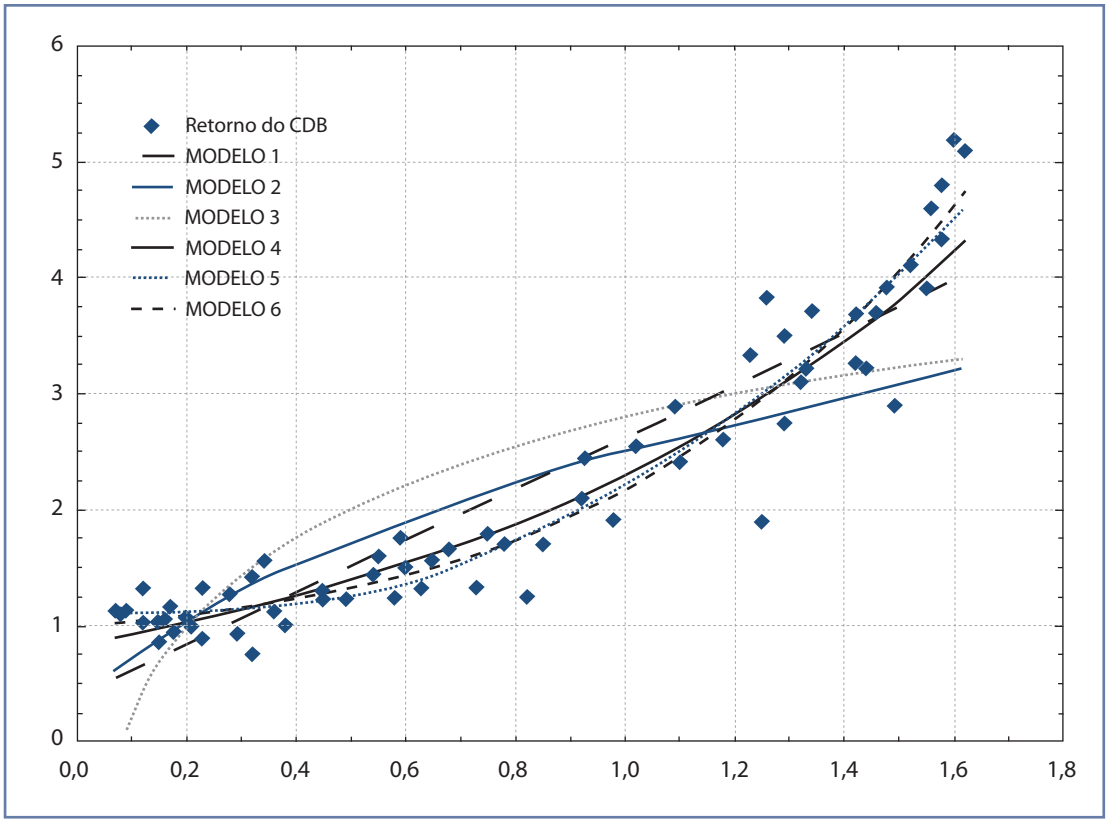


Figura 3.7 – Diagrama de dispersão para um conjunto de dados hipotéticos de CDB e IPC e especificações de relação funcional de modelos de regressão:

MODELO 1 → $CDB_t = \beta_1 + \beta_2 IPC_t + u_i$;

MODELO 2 → $CDB_t = \beta_1 IPC_t^{\beta_2} u_i$;

MODELO 3 → $CDB_t = \beta_1 + \beta_2 log_{10}(IPC_t) + u_i$;

MODELO 4 → $CDB_t = \beta_1 \, exp(\beta_2 IPC_{t)} \, exp(u_i)$;

MODELO 5 → $CDB_t = \beta_1 + \beta_2 IPC_t + \beta_{23} IPC_t^2 + u_i$ e

MODELO 6 → $CDB_t = \beta_1 + \beta_2 IPC_t + \beta_{23} IPC_t^2 + \beta_{23} IPC_t^3 + u_i$.

HIPÓTESE 10: não existe multicolinearidade perfeita. Ou seja, não há relações lineares perfeitas entre as variáveis explicativas.

Suponha que *Y*, X_2 e X_3 representem, respectivamente, consumo, renda e riqueza do consumidor. Ao postular que a despesa de consumo se relaciona linearmente com a renda e a riqueza, a teoria econômica presume que elas possam ter alguma influência independente sobre o consumo. Do contrário, não haveria sentido em incluí-las no modelo. Por exemplo, considere a seguinte regressão:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 (2X_{2i}) + u_i$$

No extremo, se houver uma relação linear exata entre renda e riqueza, teremos apenas uma variável independente, não duas, e não haverá como avaliar a influência separada da renda e da riqueza sobre o consumo. Para ver isto claramente, seja $X_{3i} = 2X_{2i}$ na regressão consumo-renda-riqueza. A regressão, então, assume a seguinte equação:

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 (2X_{2i}) + u_i \\ &= \beta_1 + (\beta_2 + 2\beta_3) X_{2i} + u_i \\ &= \beta_1 + \alpha X_{2i} + u_i \end{aligned}$$

em que $\alpha = (\beta_2 + 2\beta_3)$. Isto é, temos, na verdade, uma regressão de duas variáveis e não de três variáveis.

Além disso, se estimarmos a regressão com duas variáveis, conforme a relação acima, e obtivermos α, não haverá como estimar a influência separada de $X_2 (= \beta_2)$ e $X_3 (= \beta_3)$ sobre *Y*, pois α fornece a influência combinada de X_2 e X_3 sobre *Y*. Resumindo, a hipótese de não multicolinearidade exige que incluamos na FRP somente as variáveis que não sejam funções lineares de algumas variáveis no modelo. Como isto pode ser feito na prática é uma questão que será explicada posteriormente.

3.4.4 Funções Amostrais e mecanismo dos Mínimos Quadrados Ordinários (MQO)

Considere a equação de regressão amostral (87), transcrita abaixo:

$$\widehat{Y}_t = \widehat{\beta}_1 + \widehat{\beta}_2 X_t + \widehat{u}_t \quad \textbf{(87)}$$

Esta equação é a versão estocástica da FRA, que descreve a característica da população por meio de medida amostral. O "chapéu", na equação (87), caracteriza estimativas. Por exemplo, $\widehat{Y}_t$ caracteriza a estimativa para valor específico da variável dependente Y_t, que se situa ao redor do valor médio, fixado o valor da variável independente (explicativa). Ou seja, $\widehat{Y}_t$ se situa em torno da expectativa condicional de Y_t, que, neste exemplo, é dado por uma relação funcional linear ($\widehat{Y}_t = \widehat{\beta}_1 + \widehat{\beta}_2 X_t$). Considere, também, que $\widehat{u}_t$ constitui o termo residual que se caracteriza como uma estimativa para o termo de erro estocástico u_t, representativo da componente assistemática de Y_t. Portanto, assumindo-se que $\widehat{u}_t$ é uma *proxy* de u_t, temos:

$$\widehat{u}_t = \widehat{Y}_t - \widehat{Y}_t \quad \textbf{(88)}$$

Mínimos Quadrados Ordinários (MQO)

O MQO inicia com a definição do termo residual, dado pela equação (88), que pode ser escrita da seguinte maneira:

$$\widehat{u}_t = \widehat{Y}_t - \widehat{Y}_t \rightarrow \widehat{u}_t = Y_t - \widehat{\beta}_1 + \widehat{\beta}_2 X_t \quad \textbf{(89)}$$

Portanto, elevando-se ambos os lados da equação (89) ao quadrado e, em seguida, aplicando-se a somatória, também em ambos os lados da equação, obtém-se:

$$\sum_{t=1}^{n} \widehat{u}_t^2 = \sum_{t=1}^{n} (Y_t - \widehat{\beta}_1 - \widehat{\beta}_2 X_t)^2 \quad \textbf{(90)}$$

Matematicamente, o MQO minimiza a distancia vertical ao quadrado, sempre entre os pontos observados e a curva de regressão. Esse procedimento é obtido, aplicando-se a primeira derivada na equação (90), com relação aos parâmetros a serem estimados, e igualando-se o resultado a zero. As equações obtidas nesse procedimento serão pontos de mínimo se a segunda derivada da equação (90) for maior que zero. Caso contrário, será um ponto de máximo.

Aplicando-se este procedimento, obtém-se:

$$\frac{\partial\left(\sum_{i=1}^{n}\hat{u}_i^2\right)}{\partial\hat{\beta}_1} = 2\sum_{i=1}^{n}\left[Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)\right](-1) = 0 \Rightarrow -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0 \quad \textbf{(91)}$$

$$\frac{\partial\left(\sum_{i=1}^{n}\hat{u}_i^2\right)}{\partial\hat{\beta}_2} = 2\sum_{i=1}^{n}\left[Y_i - \left(\hat{\beta}_1 + \hat{\beta}_2 X_i\right)\right](-X_i) = 0 \quad \textbf{(92)}$$

$$\frac{\partial^2\sum_{t=1}^{n}u_t^2}{\partial\hat{\beta}_1{}^2} = 2 \Rightarrow \frac{\partial^2\sum_{t=1}^{n}u_t^2}{\partial\hat{\beta}_1{}^2} > 0 \quad \textbf{(93)}$$

e

$$\frac{\partial^2\left(\sum_{i=1}^{n}\hat{u}_i^2\right)}{\partial\hat{\beta}_2{}^2} = 2\sum_{i=1}^{n}X_i^2 > 0 \quad \textbf{(94)}$$

Portanto, a função (90) terá um mínimo quando suas derivadas parciais, em relação a $\widehat{\beta}_1$ e a $\widehat{\beta}_2$, forem nulas, e a segunda derivada parcial, com relação a estas estimativas, for positiva. Então, constata-se por meio das equações (93) e (94) que as segundas derivadas, respectivamente, com relação a $\widehat{\beta}_1$ e a $\widehat{\beta}_2$, são positivas, o que indica que os pontos a serem estimados pelas equações (91) e (92) são pontos de mínimo. Ou seja, as estimativas são tais que a soma dos resíduos ao quadrado seja mínima.

A equação (91) pode ser modificada, como vemos a seguir:

$$-2\sum_{i=1}^{n}Y_i + 2\sum_{i=1}^{n}\hat{\beta}_1 + 2\hat{\beta}_2\sum_{i=1}^{n}X_i = 0 \rightarrow \quad -2\sum_{i=1}^{n}Y_i + 2n\hat{\beta}_1 + \hat{\beta}_2\sum_{i=1}^{n}X_i = 0 \rightarrow$$

$$-\sum_{i=1}^{n}Y_i + n\hat{\beta}_1 + \hat{\beta}_2\sum_{i=1}^{n}X_i = 0 \quad \textbf{(95)}$$

A equação (92) também pode ser modificada, como a seguir:

$$-2\sum_{i=1}^{n} X_i Y_i + 2\hat{\beta}_1 \sum_{i=1}^{n} X_i + 2\hat{\beta}_2 \sum_{i=1}^{n} X_i^2 = 0 \rightarrow \quad -\sum_{i=1}^{n} X_i Y_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2 = 0$$

(96)

As equações (95) e (96), portanto, representam um sistema de equações lineares com relação a $\widehat{\beta}_1$ e a $\widehat{\beta}_2$, que pode ser escrito da seguinte forma:

$$\begin{cases} n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i & \leftarrow \times\left(-\sum_{i=1}^{n} X_i\right) \\ \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i & \leftarrow \times n \end{cases} \qquad \textbf{(97)}$$

Então, multiplicando-se a primeira equação do sistema (97) por $-\sum_{i=1}^{n} X_i$ e a segunda por *n*, obtém-se:

$$\begin{cases} -\hat{\beta}_1 n \sum_{i=1}^{n} X_i - \hat{\beta}_2 \left(-\sum_{i=1}^{n} X_i\right)^2 = -\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i \\ \hat{\beta}_1 n \sum_{i=1}^{n} X_i + \hat{\beta}_2 n \sum_{i=1}^{n} X_i^2 = n \sum_{i=1}^{n} X_i Y_i \end{cases} +$$

$$0 + \hat{\beta}_2 \left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2 \right] = \left(n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i \right) \rightarrow$$

$$\hat{\beta}_2 = \frac{\left(n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i \right)}{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2 \right]} \qquad \textbf{(98)}$$

Substituindo-se a equação (98) na primeira equação do sistema (97), teremos:

$$n\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i \left(n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i \right)}{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2 \right]} = \sum_{i=1}^{n} Y_i \rightarrow$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} Y_i \left(\sum_{i=1}^{n} X_i \right)^2 - \sum_{i=1}^{n} X_i \left(n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i \right)}{n\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2 \right]} \rightarrow$$

$$\hat{\beta}_1 = \frac{\cancel{n}\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i^2 - \cancel{n}\sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i}{\cancel{n}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2 \right]} \rightarrow$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2}} \qquad \textbf{(99)}$$

É fácil verificar que as fórmulas para o cálculo de $\widehat{\beta}_1$ e $\widehat{\beta}_2$ podem ser escritas de diversos modos. Por exemplo, com relação a $\widehat{\beta}_1$, tem-se:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \Rightarrow \sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n} \hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i \rightarrow$$

$$\sum_{i=1}^{n} \hat{Y}_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i \rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{Y}_i}{n} - \hat{\beta}_2 \frac{\sum_{i=1}^{n} X_i}{n} \rightarrow$$

$$\boxed{\hat{\beta}_1 = \overline{\hat{Y}_i} - \hat{\beta}_2 \overline{X}} \qquad \textbf{(100)}$$

Com relação a $\widehat{\beta}_2$, considere a equação (98) transcrita abaixo:

$$\hat{\beta}_2 = \frac{\left(n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i\right)}{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]} \rightarrow \quad \hat{\beta}_2 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} \qquad \textbf{(101)}$$

Considere também que:

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i^2 - 2\overline{X}X_i + \overline{X}^2) = \sum_{i=1}^{n}(X_i^2) - 2\overline{X}\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\overline{X}^2 \rightarrow$$

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i^2 - 2\overline{X}X_i + \overline{X}^2) = \sum_{i=1}^{n}(X_i^2) - \tfrac{2}{n}(\sum_{i=1}^{n}X_i)^2 + n\overline{X}^2 \rightarrow$$

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i^2) - \tfrac{2}{n}(\sum_{i=1}^{n}X_i)^2 + \tfrac{1}{n}(\sum_{i=1}^{n}X_i)^2 \rightarrow$$

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i^2) - \tfrac{1}{n}(\sum_{i=1}^{n}X_i)^2 \qquad \textbf{(102)}$$

e

$$\sum_{i=1}^{n}[(X_i - \overline{X}) . (Y_i - \overline{Y})] = \sum_{i=1}^{n}(X_iY_i - \overline{Y}X_i - \overline{X}Y_i + \overline{X}\overline{Y}) \rightarrow$$

$$\sum_{i=1}^{n}[(X_i - \overline{X}) . (Y_i - \overline{Y})] = \sum_{i=1}^{n}(X_iY_i) - \overline{Y}\sum_{i=1}^{n}X_i - \overline{X}\sum_{i=1}^{n}Y_i + \sum_{i=1}^{n}\overline{X}\overline{Y} \rightarrow$$

$$\sum_{i=1}^{n}[(X_i - \overline{X}) . (Y_i - \overline{Y})] = \sum_{i=1}^{n}(X_iY_i) - \tfrac{1}{n}\sum_{i=1}^{n}X_i . \sum_{i=1}^{n}Y_i - \tfrac{1}{n}\sum_{i=1}^{n}X_i . \sum_{i=1}^{n}Y_i + n\overline{X}\overline{Y} \rightarrow$$

$$\sum_{i=1}^{n}[(X_i - \overline{X}) . (Y_i - \overline{Y})] = \sum_{i=1}^{n}(X_iY_i) - \tfrac{1}{n}\sum_{i=1}^{n}X_i . \sum_{i=1}^{n}Y_i - \tfrac{1}{n}\sum_{i=1}^{n}X_i . \sum_{i=1}^{n}Y_i + \tfrac{1}{n}\sum_{i=1}^{n}X_i . \sum_{i=1}^{n}Y_i$$

$$\sum_{i=1}^{n}[(X_i - \overline{X}) . (Y_i - \overline{Y})] = \sum_{i=1}^{n}(X_iY_i) - \tfrac{1}{n}(\sum_{i=1}^{n}X_i . \sum_{i=1}^{n}Y_i) \qquad \textbf{(103)}$$

Substituindo as equações (102) e (103) na equação (101), obtém-se:

$$\boxed{\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}} \qquad \textbf{(104)}$$

onde

$x_i = (X_i - \overline{X})$

$x_i = (Y_i - \overline{Y})$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

n = tamanho da amostra

Portanto, para estimar $\overline{\beta}_1$, pode-se utilizar tanto a equação (99) como a equação (100) e, para estimar $\overline{\beta}_2$, pode-se utilizar tanto a equação (98) como a (104).

Precisão ou erros padrão das estimativas do MQO

Está claro que as estimativas pelo MQO são função dos dados da amostra. Mas como os dados amostrais, provavelmente, irão variar de amostra para amostra, as estimativas também variarão. Dessa forma, é necessário alguma medida de confiabilidade ou precisão das estimativas de $\overline{\beta}_1$ e $\overline{\beta}_2$.

Na estatística, a precisão de uma estimativa é medida por seu erro padrão. Dada a hipótese gaussiana (distribuição normal), os erros padrão das estimativas por MQO podem ser obtidos como segue (considere que o desenvolvimento matemático não foi apresentado):

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \quad \textbf{ou} \quad ep(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} x_i^2}} \qquad \textbf{(105)}$$

e

$$\text{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}\sigma^2 \quad \textbf{ou} \quad ep(\hat{\beta}_1) = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}}\sigma \qquad \textbf{(106)}$$

onde

var = variância
ep = erro padrão
σ^2 = variância constante ou homoscedástica de u_i (Hipótese 4) ou variância da estimativa de Y_i

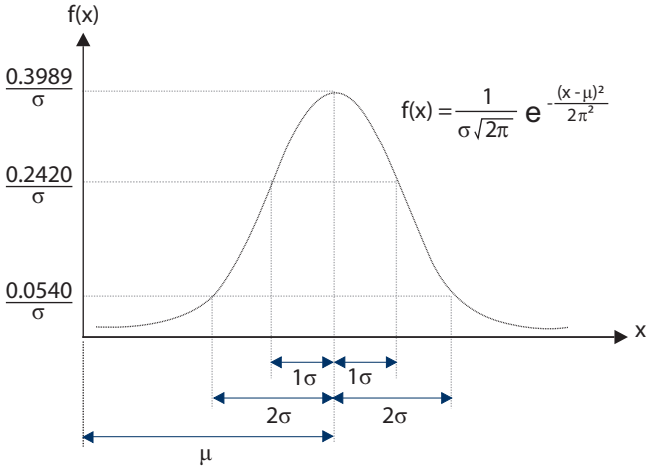
Saiba Mais

A distribuição normal é uma das mais importantes distribuições da estatística, conhecida também como Distribuição de Gauss ou Gaussiana. Foi desenvolvida pelo matemático francês Abraham de Moivre.

Além de descrever uma série de fenômenos físicos e financeiros, possui grande uso na estatística inferencial. É inteiramente descrita por seus parâmetros de média e desvio padrão, ou seja, conhecendo-os, consegue-se determinar qualquer probabilidade em uma Normal.

Um interessante uso da Distribuição Normal é que ela serve de aproximação para o cálculo de outras distribuições quando o número de observações fica grande. Essa importante propriedade provém do Teorema Central do Limite, que diz que "toda soma de variáveis aleatórias independentes de média finita e variância limitada é aproximadamente Normal, desde que o número de termos da soma seja suficientemente grande" (ver o teorema para um enunciado mais preciso) (disponível em: http://pt.wikipedia.org/wiki/Distribui%C3%A7%C3%A3o_normal.

A função de densidade de probabilidade normal, bem como sua forma analítica está mostrada na figura abaixo. Os dois parâmetros que definem a distribuição são a média (μ) e o desvio padrão (σ). A integral da função, ou seja, a área da curva é unitária. A área entre μ+σ e μ-σ é aproximadamente 0,68. A área entre μ+2σ e μ-2σ é de 0,95. Uma interpretação do significado destes valores é que a probabilidade de uma variável aleatória, com distribuição normal, que tenha seu valor maior que μ+2σ ou menor que μ-2σ, é de aproximadamente 0.05 (disponível em: http://www.cbpf.br/cat/pdsi/gauss.html



Todas as quantidades que entram nas equações anteriores, exceto σ^2, podem ser estimadas a partir dos dados. O σ^2, propriamente dito, é estimado por meio da seguinte fórmula:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k} \quad \textbf{(107)}$$

onde *n* é o tamanho da amostra, *k* é o número de parâmetros do modelo (no caso em análise, modelo linear com dois parâmetros, $\widehat{\beta}_1$ e $\widehat{\beta}_2$) e $\widehat{u}_t$ é o resíduo, sendo igual a $(Yi\text{-}\widehat{Y}_i)$. Portanto, σ_t^2 é o estimador de MQO do verdadeiro (porém desconhecido) σ^2. Já *(n-2)* é conhecida como Número de Graus de Liberdade (gl ou, em inglês, df), enquanto que $\sum_{i=1}^{n} \widehat{u}_i^2$ é a Soma dos Quadrados dos Resíduos (SQR).

Considerando que:

$$Y_i = \hat{Y}_i + \hat{u}_i \rightarrow \left(Y_i - \overline{Y}\right) = \left(\hat{Y}_i - \overline{Y}\right) + \hat{u}_i \rightarrow \quad y_i = \hat{y}_i + \hat{u}_i \quad \textbf{(108)}$$

Elevando a equação (108) ao quadrado, em ambos os lados, e somando-a ao longo da amostra, obtém-se:

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} u_i^2 + \underbrace{\sum_{i=1}^{n} \hat{y}_i u_i}_{\approx 0} \rightarrow \quad \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} u_i^2 \quad \textbf{(109)}$$

Mas, considerando que:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \rightarrow \; \hat{Y}_i - \overline{Y} = \underbrace{\hat{\beta}_1 - \overline{Y} + \hat{\beta}_2 \overline{X}}_{=0} + \hat{\beta}_2 \left(X_i - \overline{X}\right) \rightarrow$$

$$\hat{y}_i = \hat{\beta}_2 x_i \rightarrow \; \hat{y}_i^2 = \hat{\beta}_2^2 x_i^2 \rightarrow$$

$$\sum_{i=1}^{n} \hat{y}_i^2 = \hat{\beta}_2^2 \sum_{i=1}^{n} x_i^2 \quad \textbf{(110)}$$

Substituindo-se a relação (110) na equação (109), obtém-se:

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} y_i^2 - \hat{\beta}_2^2 \times \sum_{i=1}^{n} x_i^2 \quad \textbf{(111)}$$

Portanto, estimando-se a soma do quadrado médio, dada pela equação (111), e substituindo na equação (107), obtém-se a estimativa para a variância homoscedástica (ou variância da estimativa).

Em seguida, estimam-se os erros padrão para a inclinação e para o intercepto, dados pelas equações (105) e (106), respectivamente. Por meio destas equações, observa-se que a precisão das estimativas dos parâmetros de regressão $\widehat{\beta}_1$ e $\widehat{\beta}_2$ é diretamente proporcional a σ^2, mas inversamente proporcional a $\sum_{i=1}^{n} \hat{x}_i^2$.

> Então, quanto maior for a variabilidade em X_i, maior será a precisão das estimativas, tanto do intercepto quanto do coeficiente de regressão. Contudo, a variabilidade do intercepto é diretamente proporcional a $\sum_{i=1}^{n} \hat{X}_i^2$ e inversamente proporcional ao tamanho *n* da amostra.

As estimativas do intercepto e o coeficiente de inclinação não somente variam de amostra para amostra, como são inversamente correlacionados: uma superestimação da inclinação conduzirá a uma subestimação do intercepto, e vice-versa.

Portanto, $\widehat{\beta}_1$ e $\widehat{\beta}_2$ são estimadores que não apenas variarão de uma amostra para outra, mas também em uma dada amostra, provavelmente, vão depender um do outro, sendo esta dependência medida pela covariância entre eles. Ou seja:

$$\operatorname{cov}\left(\hat{\beta}\ ,\ \hat{\beta}\ \right) = -\overline{X}\ \operatorname{var}\left(\hat{\beta}_2\right) = -\overline{X}\ \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \qquad \textbf{(112)}$$

Podemos observar, na equação **(112)**, que todas as grandezas são positivas, exceto $\overline{X}$. Assim, a *cov*($\widehat{\beta}_1$, $\widehat{\beta}_2$) depende do sinal de $\overline{X}$. Se $\overline{X}$ for positivo, a covariância será negativa, e vice-versa. Se o coeficiente de declividade for superestimado, o intercepto será subestimado, e vice-versa. Mais à frente, veremos a importância de estudar a covariância entre esses dois coeficientes (Multicolinearidade).

As propriedades dos estimadores de mínimos quadrados (MQO) são dadas pelo teorema de Gauss-Markov que estabelece que, satisfeitas as hipóteses do modelo clássico de regressão linear, os estimadores por mínimos quadrados (MQO), na classe dos estimadores lineares não enviesados, têm mínima variância. Isto é, são **MELNV** (**Melhor Estimador Linear Não Viesado** ou estimador *blue* de β_2). Isso quer dizer que:

- é um estimador linear, isto é, uma função linear de uma variável aleatória, tal como a variável dependente Y do modelo de regressão;
- é não viesado (não tendencioso), ou seja, o seu valor médio ou esperado $E(\beta_2)$ é igual ao valor verdadeiro de β_2; e
- tem mínima variância na classe de todos esses estimadores lineares não viesados, e é conhecido como estimador eficiente.

Portanto, o TEOREMA DE GAUSS-MARKOV estabelece que, dadas as hipóteses do modelo clássico de regressão linear, os estimadores por mínimos quadrados (MQO), na classe dos estimadores lineares não enviesados, têm mínima variância; isto é, são MELNV.

Precisão de estimativas e ajustes

Palavra do Professor

Até agora nos preocupamos com o problema de estimar coeficientes de regressão, seus erros padrão e algumas de suas propriedades. Agora, examinaremos o grau de ajuste a um conjunto de dados da reta de regressão. Ou seja, verificaremos o quão bem a reta de regressão da amostra ajusta-se aos dados.

Se todas as observações (informações pertinentes aos dados) estivessem situadas sobre a linha de regressão, por exemplo, o ajuste seria perfeito. Mas este não é o caso, pois haverá alguns $\hat{u}_i$ positivos e alguns outros $\hat{u}_i$ negativos. O que se espera é que esses resíduos distanciados da linha de regressão apresentem afastamentos tão pequenos quanto possíveis.

O coeficiente de determinação r^2 (no caso de regressão de duas variáveis) e R^2 (no caso de regressão de multivariáveis) é uma medida sintética que diz o quão bem a reta de regressão da amostra se ajusta aos dados. Podemos entender melhor o grau de ajuste por meio do *Diagrama de Venn ou Ballentine*, apresentado na Figura 3.8.

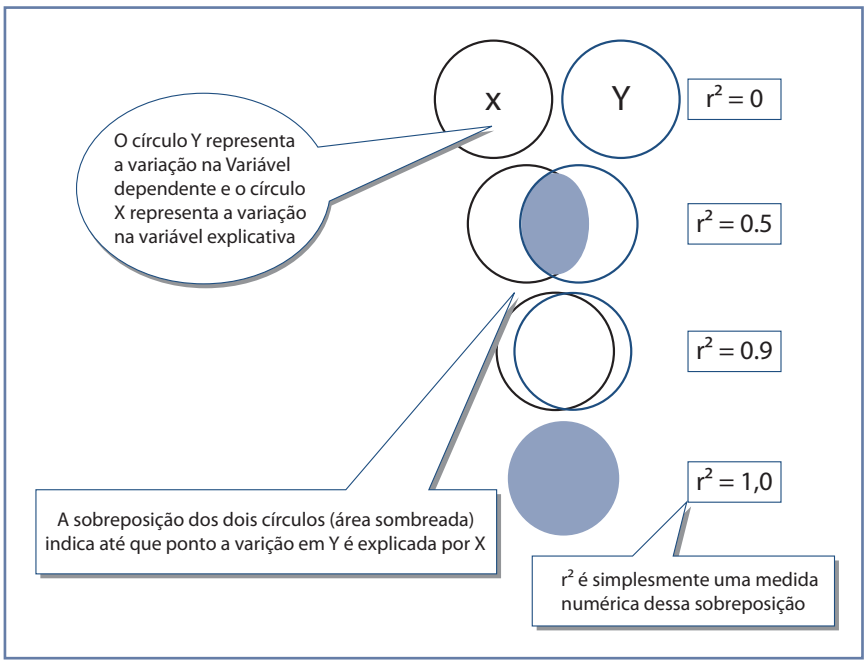


Figura 3.8 – Diagrama de Venn ou Ballentine, para evidenciar os níveis de explicação da variável explanatório *X* sobre a variável dependente *Y*, por meio do coeficiente de determinação r^2.

A estimativa de r^2 é obtida, inicialmente, a partir da relação (109) transcrita abaixo:

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} u_i^2 \quad \textbf{(109)}$$

Tal que:

$$\sum_{i=1}^{n} \hat{y}_i^2 = SQE \quad , \quad \sum_{i=1}^{n} y_i^2 = SQT \quad \textbf{e} \quad \sum_{i=1}^{n} u_i^2 = SQR \quad \textbf{(113, 114 e 115)}$$

onde **SQT** é a Soma do Quadrado Total, que mede a variação total dos valores efetivos de Y_i em relação à sua média amostral; **SQE** é a Soma do Quadrado da Explicação devido à regressão, que mede a variação dos valores estimados de $\hat{Y}_i$ em relação à sua média amostral. Ou seja, é a parcela de *Y* explicada pela variável *X* e SQR é a Soma do Quadrado dos Resíduos, que mede a variação residual ou não explicada dos valores de Y_i em relação à curva de regressão.

Em resumo, considerando as relações (109), (113), (114) e (115), temos que:

$$SQT = SQE + SQR \quad \textbf{(116)}$$

A relação (116) mostra que a variação total dos valores de Y, em relação a seu valor médio, pode ser dividida em duas partes: uma atribuída à reta de regressão, e outra às forças aleatórias, porque nem todas as observações efetivas Y_i ficam sobre a reta ajustada. A Figura 3.9 evidencia estas variações.

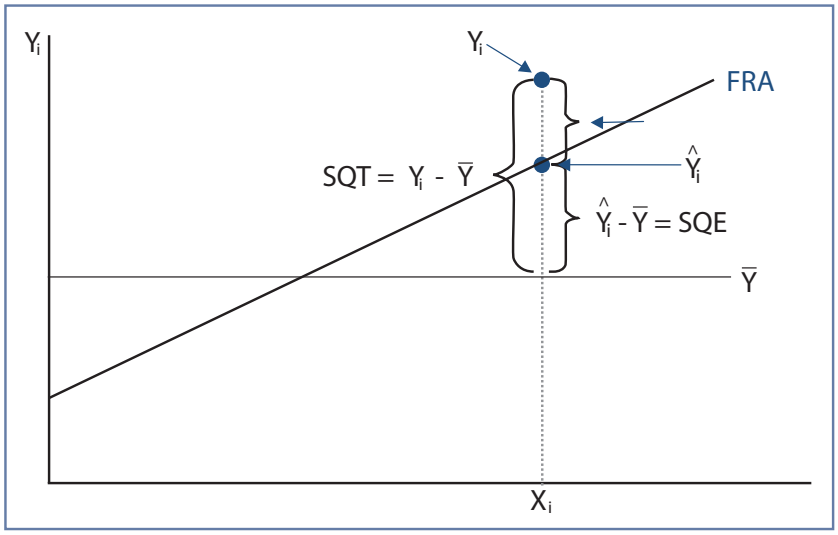


Figura 3.9 – Relações entre SQT, SQE e SQR

A partir da equação (116), tem-se:

$$SQT = SQE + SQR \rightarrow$$

$$1 = \frac{SQE}{SQT} + \frac{SQR}{SQT} \rightarrow r^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT} \qquad \textbf{(117)}$$

A quantidade r^2 definida pela equação (117) é conhecida como **coeficiente de determinação da amostra** e é a medida mais utilizada do grau de ajuste de uma reta de regressão. Traduzindo, r^2 mede a proporção ou percentagem da variação total em Y explicada pelo modelo de regressão.

Em resumo, é necessário verificar o quão ótimo é o ajuste de regressão (FRA) dos dados. Para tanto:

- deve-se comparar os desvios dos dados da FRA (desvios residuais) com os desvios da média amostral de Y (Desvio Total = Desvio Explicado + Desvio residual), e identificar a ordem de grandeza de cada uma destas medidas;
- deve-se saber que, quanto maior for o termo de Desvio Explicado relativo ao Desvio Total, melhor será o ajuste;

- deve-se notar que o ajuste total é proveniente da agregação dos desvios de todas as observações na amostra.

Portanto, de acordo com as relações de (113) a (115) e (117), conclui-se que:

$$r^2 = 1 - \frac{\sum_{i=1}^{n} u_i^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} \quad \textbf{(118)}$$

Ou, de acordo com as relações (111) e (117), tem-se que:

$$r^2 = \frac{SQE}{SQT} = \frac{\sum_{I=1}^{n} \hat{y}_i^2}{\sum_{I=1}^{n} y_i^2} = \frac{\hat{\beta}_2^2 \sum_{I=1}^{n} x_i^2}{\sum_{I=1}^{n} y_i^2} = \hat{\beta}_2^2 \left(\frac{\sum_{I=1}^{n} x_i^2}{\sum_{I=1}^{n} y_i^2} \right) \quad \textbf{(119)}$$

Contudo, considerando a equação (119), onde S_X^2 e S_Y^2 são, respectivamente, as variâncias de Y_i e X_i, veremos que:

$$r^2 = \hat{\beta}_2^2 \left(\frac{\sum_{I=1}^{n} x_i^2}{\sum_{I=1}^{n} y_i^2} \right) = \hat{\beta}_2^2 \left(\frac{\sum_{I=1}^{n} x_i^2 \Big/ (n-1)}{\sum_{I=1}^{n} y_i^2 \Big/ (n-1)} \right) = \hat{\beta}_2^2 \left(\frac{S_X^2}{S_Y^2} \right) \quad \textbf{(120)}$$

Ainda, considerando as relações (104) e (119), obteremos:

$$r^2 = \frac{\left(\sum_{i=1}^{10} x_i y_i \right)^2}{\sum_{i=1}^{10} x_i^2 \times \sum_{i=1}^{10} y_i^2} \quad \textbf{(121)}$$

O Coeficiente de Determinação é corrigido quando se usam poucos pontos na amostra pela seguinte relação:

$$\bar{r}^2 = r^2 - \frac{1}{(n-2)}\left(1 - r^2\right) \quad \textbf{(122)}$$

Onde $\bar{r}^2$ é denominado de coeficiente de determinação ajustado.

Duas propriedades de r^2 podem ser destacadas:

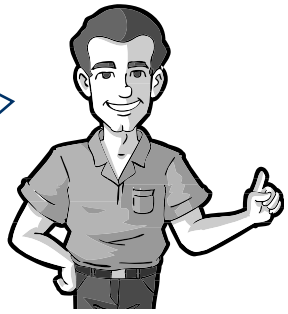
- ele é uma quantidade não negativa, sendo $\boldsymbol{0 \leq r^2 \leq 1}$;
- um $\boldsymbol{r^2 = 1}$ significa um perfeito ajuste, isto é, para todo *i*, $\widehat{Y}_i = Y_i$. Por outro lado, um $\boldsymbol{r^2 = 0}$ significa que não existe nenhuma relação de causalidade entre a variável regredida e a variável regressora ($\widehat{\beta}_2 = 0$). Nesse caso, $\widehat{Y}_i = \widehat{\beta}_1 = \overline{Y}$, portanto, a melhor previsão de qualquer valor de *Y* é simplesmente o seu valor médio. Nesta situação, a reta de regressão será horizontal ao eixo *X*.

Finalmente, tem-se que o coeficiente de Correlação *r* é:

$$r = \pm\sqrt{r^2} \quad \textbf{(123)}$$

O sinal de *r* deve ser obtido pela inclinação da reta de regressão. Ou seja, deve ser igual ao sinal de $\widehat{\beta}_2$.

Caro aluno, consulte agora o Exemplo 1 do material complementar da Unidade 3, pois se trata do exemplo protótipo cujo resultado será utilizado nas análises que se seguem até o final desta Unidade. Portanto, leia e analise-o atentamente. Em seguida, continue a sua leitura da Seção 3.5.



3.5 CONSIDERAÇÕES SOBRE O MQO

Os estimadores de MQO para $\widehat{\beta}_1$ e $\widehat{\beta}_2$ satisfazem diversas propriedades estatísticas desejáveis, dentre elas a inexistência de viés e a variância mínima. Se o objetivo for simplesmente a estimativa de ponto, o método dos MQO será suficiente. Mas esta estimativa é apenas um aspecto da inferência estatística. Outros aspectos são os testes de hipóteses e a definição de intervalos de confiança.

Como o nosso objetivo é tanto a estimativa de ponto, como estabelecer Testes de Hipótese ou Intervalos de Confiança, então, *deve-se especificar a distribuição de probabilidade das perturbações* u_i. **Por quê?** Porque as distribuições de probabilidades ou amostragem dos estimadores de MQO dependerão das hipóteses feitas sobre a distribuição de probabilidades de u_i.

Portanto, trataremos do Modelo Clássico de Regressão Linear de duas variáveis, mas supondo que as perturbações u_i da população estejam distribuídas normalmente, o que é chamado de **Modelo Clássico de Regressão Linear Normal (MCRLN)**. Assim, as seguintes hipóteses são feitas sobre os resíduos u_i:

- média zero, ou seja, expectativa zero ($E(u_i)=0$);
- não estão correlacionados. Ou seja, para $i \neq j$, onde u_i e u_j representam valores dos resíduos das observações da variável dependente Y, para dois valores distintos da variável explicativa X, ou seja, X_i e X_j, tal que ($E(u_i,u_j)=0$);
- u_i com $i=1, 2, ..., n$, possui variância constante ($E\left(u_i^2\right)=\sigma^2$);
- adiciona-se às propriedades acima, a condição de normalidade para u_i
- ($u_i \approx N\left(0,\sigma^2\right)$).

Assim, com a hipótese de normalidade assumida para *ui*, teremos que *ui* e *uj* não apenas têm correlação zero, como também se distribuem independentemente. Portanto:

$$u_i \approx NID\left(0,\sigma^2\right) \quad \textbf{(124)}$$

Note que **NID** significa **normal** e **independentemente distribuído**.

Agora você deve estar se perguntando: por que a hipótese de normalidade? Por várias razões! Veja a seguir.

- u_i representa a influência combinada (na variável dependente) de um grande número de variáveis independentes que não são explicitamente introduzidas no modelo de regressão. A influência dessas variáveis omitidas ou esquecidas deve ser pequena, quando muito aleatória, pois, de acordo com o Teorema do Limite Central (CHOU, 1989), a soma de variáveis aleatórias produz uma distribuição aleatória normal.
- O Teorema do Limite Central diz que, mesmo que o número das variáveis não seja muito grande, ou que essas variáveis não sejam rigorosamente independentes, sua soma pode ainda assim ser distribuída normalmente.

- Com a hipótese de normalidade, as distribuições de probabilidade dos estimadores de MQO podem ser facilmente derivadas, pois uma propriedade da distribuição normal é que qualquer função linear de variáveis distribuídas normalmente é, ela própria, distribuída normalmente. Ainda, se u_i é distribuído normalmente, então, $\widehat{\beta}_1$ e $\widehat{\beta}_2$ são também distribuídos normalmente.
- A distribuição normal é uma distribuição relativamente simples que envolve somente dois parâmetros (*média e variância*).

A condição de u_i = NID(0, σ^2) garante as Hipóteses 3, 4, 5 e 6 do modelo clássico de regressão MQO, anteriormente mencionadas (Subseção 3.4.3).

Portanto, considerando que o termo de erro estocástico seja uma distribuição normal, com média $\bar{u}$ e variância σ^2_{ui}, tal que:

$$\hat{u}_i \approx N(\bar{u}, \sigma^2_{ui}) \qquad \textbf{(125)}$$

Com a seguinte função de distribuição de probabilidade, *f(u)*:

$$f(u) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left[\frac{-(u-\bar{u})^2}{2\sigma_u^2}\right] \quad \textbf{(126)}$$

Com a hipótese da normalidade dos resíduos, os estimadores MQO de $\widehat{\beta}_1$ e $\widehat{\beta}_2$ e $\hat{\sigma}^2$ possuem as seguintes propriedades estatísticas:

1. não são viesados;
2. têm variância mínima (estimadores eficientes), que combinada com a propriedade anterior, conduz ao fato de que os estimadores MQO sejam estimadores MELVM (melhor estimador linear de variância mínima), para os parâmetros acima especificados;
3. consistência, isto é, quando o tamanho da amostra aumenta indefinidamente, os estimadores convergem para seus verdadeiros valores da população; e
4. β_1 distribui-se normalmente com $N(\widehat{\beta}_1, \hat{\sigma}^2_{\hat{\beta}_1})$. Ou seja, o valor médio da distribuição de β_1 é o verdadeiro valor da distribuição da população, obtido por meio da distribuição dos $(\widehat{\beta}_1)$s estimados por meio de regressão de dados amostrais para um conjunto grande de diferentes amostras. As-

sim, β_1 é o valor médio, ou o valor esperado de $\widehat{\beta}_1$ (ou seja, $E(\widehat{\beta}_1) = \beta_1$). Portanto, possuindo as distribuições dos $(\widehat{\beta}_1)$s estimados para as diferentes amostras, poderíamos estimar $\hat{\sigma}^2_{\hat{\beta}_1}$, a variância de β_1. Contudo, nesse caso, usa-se $\widehat{\beta}_1$ como *Proxy* de β_1 e

$$\sigma^2_{\hat{\beta}_1} = \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}\sigma^2 \quad \textbf{(127)}$$

como *proxy* para a variância de β_1.

Assim, garantida a propriedade de normalidade $N(\widehat{\beta}_1, \hat{\sigma}^2_{\hat{\beta}_1})$, esta pode ser transformada em uma distribuição normal padrão, determinando a variável *z*, definida como:

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \quad \textbf{(128)}$$

que constitui uma distribuição normal com média zero e variância unitária, ou seja: $z \approx N(0,1)$ (onde $\sigma_{\hat{\beta}_1}$ é o desvio padrão da distribuição$(\widehat{\beta}_1)s$ ou a raiz quadrada da variância estimada em (127)).

5. β_2 também se distribui normalmente com $N(\widehat{\beta}_2, \hat{\sigma}^2_{\hat{\beta}_1})$. Ou seja, o valor médio da distribuição, β_2, é o verdadeiro valor da distribuição da população, obtido por meio da distribuição dos $(\widehat{\beta}_2)s$ estimados por meio de regressão de dados amostrais para um grande conjunto de diferentes amostras. Assim,β_2 é o valor médio, ou o valor esperado de $\widehat{\beta}_2$ (ou seja, $E(\widehat{\beta}_2) = \beta_2$). Portanto, possuindo as distribuições dos $(\widehat{\beta}_2)$s estimados para as diferentes amostras, poderíamos estimar $\hat{\sigma}^2_{\hat{\beta}_2}$, a variância de β_2. Contudo, nesse caso, usa-se $\widehat{\beta}_2$ como Proxy de β_2 e

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \qquad \left[(n-2)\hat{\sigma}^2\right] / \sigma^2 \quad \textbf{(129)}$$

como *proxy* para a variância de β_2.

Assim, garantida a propriedade de normalidade $N(\widehat{\beta}_2, \widehat{\sigma}^2_{\widehat{\beta}_2})$, esta pode ser transformada em uma distribuição normal padrão, determinando a variável *z*, definida como:

$$Z = \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} \quad \textbf{(130)}$$

que constitui uma distribuição normal com média zero e variância unitária, ou seja: $z \approx N(0,1)$ (onde $\sigma_{\widehat{\beta}_2}$ é o desvio padrão da distribuição $(\widehat{\beta}_2)$s ou a raiz quadrada da variância estimada em (127)).

6. $(n - k)\widehat{\sigma}^2/\sigma^2$ é distribuído com χ^2 (qui-quadrado) com *(n-k)* graus de liberdade gl, *n* é o tamanho da amostra, *k* é número de parâmetros do modelo de regressão, $\widehat{\sigma}^2$ é a estimativa variância homoscedástica (ou variância da estimativa) e σ^2 é a variância homoscedástica da população;

7. $\widehat{\beta}_1$ e $\widehat{\beta}_2$ se distribuem independente de $\widehat{\sigma}^2$;

8. $\widehat{\beta}_1$ e $\widehat{\beta}_2$ têm variância mínima em toda classe de estimadores não viesados, sejam lineares ou não. Esse resultado, contrariamente ao teorema de Gauss-Markov, não restringe a classe de estimadores lineares. Portanto, pode-se dizer que os estimadores de Mínimos Quadrados são os melhores estimadores não viesados (MENV) (GUJARATI, 200). Como se supôs que u_i se distribui normalmente, e pelas características exógenas de Y_i, esse Y_i é aleatório e se distribui normalmente, então, $\widehat{Y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_t + \widehat{u}_t$ obedece à regra de que toda função linear de variáveis distribuídas normalmente distribui-se também dessa forma. Ou seja, para garantir essa propriedade, intuitivamente, $\widehat{\beta}_1$ e $\widehat{\beta}_2$ também devem ser normalmente distribuídos. É importante notar que a hipótese de normalidade permite derivar as distribuições de probabilidade ou amostragem de $\widehat{\beta}_1$ e $\widehat{\beta}_2$ (normal) e $\widehat{\sigma}^2$ (qui-quadrado).

A propósito, se admitirmos que u_i se distribui normalmente, com média zero e variância $\widehat{\sigma}^2$, então, Y_i também se distribui normalmente, com média e variância dadas, respectivamente, por:

$$\overline{Y} = E(Y_i|X_i) \quad \textbf{(131)}$$

$$\textbf{e} \quad var(Y_i) = \sigma^2 \quad \textbf{(132)}$$

Ou seja, mais precisamente, $Y_i \approx N[E(Y_i|X_i), \sigma^2]$.

Deve ser lembrado que estamos tratando de regressões com somente duas variáveis, uma dependente e outra explicativa.

Estabelecidas as características probabilísticas das variáveis e os parâmetros que integram um modelo de regressão, então, descrevem-se as propriedades das distribuições probabilísticas de um modelo de regressão, conforme Gujarati (2006). Assim, tem-se os seguintes teoremas:

TEOREMA I: Se Z_1, Z_2,... , Z_n são variáveis aleatórias, distribuídas normalmente e independentes, de tal modo que $N(\mu_i, \sigma_i^2)$, então a soma $\sum_{i=1}^{n} k_i Z_i$, em que k_i são constantes diferentes de zero, também se distribui normalmente com média $\sum_{i=1}^{n} k_i \mu_i$ e variância $\sum_{i=1}^{n} k_i^2 \sigma_i^2$, ou seja, $N(\sum_{i=1}^{n} k_i \mu_i, \sum_{i=1}^{n} k_i^2 \sigma_i^2)$ (onde μ indica o valor médio).

Em suma, considere-se que $z_1 = N(10,2)$ e $z_2 = N(8,1,5)$, então, a combinação $z = 0{,}8z_1 + 0{,}2z_2$ também se distribui normalmente, com média $\mu = 0{,}8 \,.\, 10 + 0{,}2 \,.\, 8 = 9{,}6$ e variância $\sigma^2 = 0{,}8^2 \,.\, 2 + 0{,}2^2 \,.\, 1{,}5 = 1{,}34$. Ou seja, $Z = N(9{,}6;1{,}34)$.

TEOREMA II: Se Z_1, Z_2, ... , Z_n são variáveis aleatórias, distribuídas normalmente, mas dependentes, então, a soma $\sum_{i=1}^{n} k_i Z_i$, em que k_i são constantes, nem todas iguais, também se distribui normalmente com média $\sum_{i=1}^{n} k_i \mu_i$ e variância

$$\sum_{i=1}^{n} k_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} k_i k_j \, cov(ZI,Zj), \text{ com } i \neq j \,. \quad \textbf{(133)}$$

Assim, se $z_1 = N(6,2)$ e $z_2 = N(7,3)$, $cov(z_1,z_2) = 0{,}8$ e , então, a combinação $z = 0{,}6z_1 + 0{,}4z_2$ também se distribui normalmente, com média $\mu = 0{,}6 \,.\, 6 + 0{,}4 \,.\, 7 = 6{,}4$ e variância $\sigma^2 = 0{,}6^2 \,.\, 2 + 0{,}4^2 \,.\, 3 + 3 \,.\, 0{,}6 \,.\, 0{,}4 \,.\, 0{,}8 = 1{,}584$. Ou seja, $z = N(6{,}4;1{,}584)$.

TEOREMA III: Se Z_1, Z_2, ... , Z_n são variáveis aleatórias, distribuídas normalmente e independentes, tal que $Z = N(0,1)$ (isto é, uma variável normal padronizada), então

$$\sum_{i=1}^{n} Z_i^2 = Z_1^2 + Z_2^2 + \ldots + Z_n^2 \quad \textbf{(134)}$$

segue uma distribuição $\chi 2$, com n graus de liberdade. Simbolicamente, $Z_i^2 = \chi_n^2$, em que n indica os graus de liberdade.

TEOREMA IV: Se Z1, Z2, ... , Z*n* são variáveis aleatórias independentes, onde cada qual segue uma distribuição χ^2 com *ki* graus de liberdade, então

$$\sum_{i=1}^{n} Z_i^2 = Z_1^2 + Z_2^2 + \ldots + Z_n^2 \quad \textbf{(135)}$$

segue uma distribuição χ^2, com $k = \sum_{i=1}^{n} k_i$ graus de liberdade.

TEOREMA V: Se Z_1 é uma variável aleatória, distribuída normalmente e independente, tal que $Z_1 = N(0,1)$ (isto é, uma variável normal padronizada), e Z_2 é uma outra variável que segue uma distribuição χ^2, com k graus de liberdade e sendo independente de Z_1, então a variável definida como

$$t = \frac{Z_1}{\sqrt{Z_2}/\sqrt{k}} = \frac{Z_1\sqrt{k}}{\sqrt{Z}} = \frac{\text{variável normal padrão}}{\sqrt{\text{variável qui - quadrada independente/gl}}} \approx t_k \quad \textbf{(136)}$$

segue uma distribuição *t* de Student, com k graus de liberdade. (CHOU, 1989)

TEOREMA VI: Se Z_1 *e* Z_2 são variáveis χ^2, distribuídas independentemente, com k_1 e k_2 graus de liberdade respectivamente, então, a variável

$$F = \frac{Z_1/k_1}{Z_2/k_2} = F_{k_1,k_2} \quad \textbf{(137)}$$

segue uma distribuição F, com k_1 e k_2 graus de liberdade, em que k_1 é conhecido como graus de liberdade do numerador e k_2 como graus de liberdade do denominador.

TEOREMA VII: O quadrado da variável *t* (de Student), com *k* gl tem uma distribuição F, com k_1=1 gl no numerador e k_2=n-k gl no denominador, ou seja,

$$F_{1,k_2} = t_{k_2}^2 \quad \textbf{(138)}$$

Palavra do Professor

Portanto, você encontrará no AVEA o Anexo I da Unidade 3, que contém um quadro de distribuição de probabilidade χ^2 e outra de distribuição *t* (Student) bicaudal, que deverão ser utilizadas nos exercícios.

3.6 REGRESSÃO DE DUAS VARIÁVEIS: ESTIMATIVA DE INTERVALO E TESTE DE HIPÓTESE

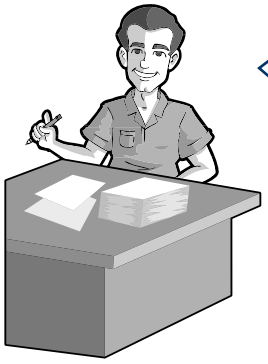
Estimativa e teste de hipótese constituem os dois principais ramos da estatística clássica. A teoria da estimativa consiste em duas partes: estimativa de ponto e estimativa de intervalo. Os métodos de estimativa de ponto são, por exemplo, o MQO e o MV (Máxima Verossimilhança).

Nesta *Seção* será examinada, primeiramente, a estimativa de intervalo e, posteriormente, os testes de hipótese.

3.6.1 Estimativa de intervalo: alguns conceitos básicos

Considere o Exemplo 1 presente no material complementar da Unidade 3, cujo modelo de regressão estimado, dado pela equação (4) do Exemplo, envolvendo uma amostra de notas numa respectiva prova e o desempenho dos alunos nos exercícios de sala de aula:

$$Y_i = 2 + X_i + u_i \quad \textbf{(4 Exemplo)}$$



A regressão dada por esta equação mostra unicamente a estimativa de ponto do modelo, desconhecida da população β. Quão confiável é esta estimativa?

Em virtude das flutuações da amostragem, uma única estimativa, a partir de uma amostra, provavelmente vai diferir do valor verdadeiro do modelo da população, apesar de se esperar que, em amostragem repetida, seu valor médio seja igual ao valor verdadeiro ($E = (\widehat{\beta}_2) = \beta_2$).

Para ser mais específico, suponha que se queira descobrir quão "próximo" é $\hat{\beta}_2$ de β_2. Para isso, tenta-se descobrir dois números positivos, δ e α, este último entre o e 1, de modo que a probabilidade do intervalo aleatório $(\hat{\beta}_2 - \delta, \hat{\beta}_2 + \delta)$ conter o verdadeiro β_2 é de (1-α). Simbolicamente:

$$Pr(\hat{\beta}_2 - \delta \leq \beta_2 \leq \hat{\beta}_2 + \delta) = 1 - \alpha \quad \textbf{(139)}$$

Quando existe, tal intervalo é conhecido como intervalo de confiança; (1-α) é o coeficiente de confiança e α, tal que seja $o<\alpha<1$, é o nível de significância. Os pontos extremos do intervalo de confiança são conhecidos como limites de confiança (ou valores críticos), sendo que $(\hat{\beta}_2 - \delta)$ é o limite de confiança inferior e $(\hat{\beta}_2 + \delta)$ é o limite de confiança superior.

Um intervalo construído de tal maneira tem uma probabilidade (1-α) de incluir em seus limites o valor verdadeiro do parâmetro. Por exemplo, se α=0,05 ou 5%, seria interpretado como: a probabilidade do intervalo (aleatório) determinado incluir nele o verdadeiro β_2 é de 0,95, ou 95%. O estimador de intervalo fornece, assim, uma série de valores dentre os quais pode estar o verdadeiro β_2.

É muito importante conhecer os seguintes aspectos da estimativa de intervalo:

- Como são construídos os intervalos de confiança? Se as distribuições de probabilidade ou amostragem dos estimadores forem conhecidas, é possível fazer declarações de intervalo de confiança do tipo representado pela equação (139);
- Sob a hipótese da normalidade das perturbações u_i, os estimadores MQO de $\hat{\beta}_1$ e $\hat{\beta}_2$ têm uma distribuição normal, e o estimador de MQO de σ^2 tem relação com a distribuição χ^2 (qui-quadrado). Nesse caso, a construção de intervalos de confiança é fácil.

3.6.2 Intervalos de confiança para os coeficientes de regressão β_1 e β_2

Com a hipótese da normalidade para u_i, os estimadores de MQO, $\hat{\beta}_1$ e $\hat{\beta}_2$ se distribuem eles próprios normalmente com médias e variâncias específicas. Por exemplo, $\hat{\beta}_2$ segue uma distribuição normal padrão, como segue:

$$Z = \frac{\hat{\beta}_2 - \beta_2}{ep(\hat{\beta}_2)} = \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\sigma} \quad \textbf{(140)}$$

Portanto, pode-se usar a distribuição normal padrão para fazer declarações probabilísticas sobre β_2, desde que a verdadeira variância da população, σ^2, seja conhecida. Contudo, se σ^2 é conhecida, uma importante propriedade de uma variável distribuída normalmente, com média μ e variância σ^2, é que a área sob a curva normal entre $\mu \pm \sigma$ é de cerca de 68%, a área entre os limites $\mu \pm 2\sigma$ é de aproximadamente 95% e a área entre $\mu \pm 3\sigma$ é de cerca de 99,7%, como é mostrado na Figura 3.10, abaixo.

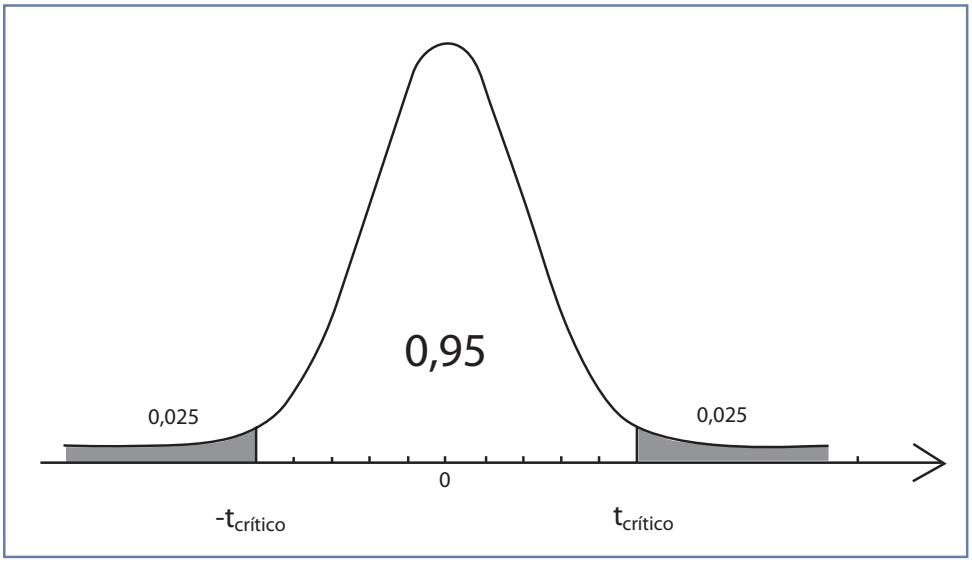


Figura 3.10 – Percentuais de dados e números de desvios padrões, numa distribuição normal.

Mas σ^2 raramente é conhecida, sendo determinada, na prática, pelo estimador não viesado $\hat{\sigma}^2$. Se substituirmos σ^2 por $\hat{\sigma}^2$, podemos escrever (140) como:

$$t = \frac{\hat{\beta}_2 - \beta_2}{ep(\hat{\beta}_2)} = \frac{estimador - parâmetro}{erro - padrão - estimado\,do\,estimador} \qquad \textbf{(141)}$$

A variável *t*, assim definida, segue a distribuição *t* com (*n*-2) gl. Portanto, ao invés de usar a distribuição normal, pode-se usar a distribuição *t*, estabelecendo um intervalo de confiança para β_2, como segue:

$$\Pr(-t_{n-k,\alpha/2} \le t \le t_{n-k,\alpha/2}) = 1-\alpha \;\rightarrow$$

$$\Pr\left[-t_{n-k,\alpha/2} \le \frac{\beta_2 - \hat{\beta}_2}{ep(\hat{\beta}_2)} \le t_{n-k,\alpha/2}\right] = 1-\alpha \;\rightarrow$$

$$\Pr\left[\hat{\beta}_2 - t_{n-k,\alpha/2} \times ep(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{n-k,\alpha/2} \times ep(\hat{\beta}_2)\right] = 1-\alpha \qquad \textbf{(142)}$$

A equação (142) fornece um intervalo de confiança de 100x(1-α)% para β_2, ou seja:

$$\hat{\beta}_2 \pm t_{n-k,\alpha/2} \times ep(\hat{\beta}_2) \qquad \textbf{(143)}$$

De forma similar, pode-se, então, escrever:

$$\Pr\left[\hat{\beta}_1 - t_{n-k,\alpha/2}ep(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{n-k,\alpha/2}ep(\hat{\beta}_1)\right] = 1-\alpha \qquad \textbf{(144)}$$

Ou, mais concisamente:

$$\hat{\beta}_1 \pm t_{n-k,\alpha/2}ep(\hat{\beta}_1) \qquad \textbf{(145)}$$

Onde:

gl=(n-k) = graus de liberdade da distribuição

n = tamanho da amostra

k = número de parâmetros a serem estimados no modelo de regressão

α = nível de significância ou nível de erro admitido na inferência

Como se observa nas equações acima, a amplitude do intervalo de confiança é proporcional ao erro padrão do estimador. Ou seja, quanto maior o erro padrão, maior será a amplitude do intervalo de confiança. Ou também, quanto maior o erro padrão do estimador, maior a incerteza de se estimar o verdadeiro valor do parâmetro desconhecido. Assim, o erro padrão de um estimador é, muitas vezes, descrito como uma medida da precisão do estimador, isto é, quão precisamente o estimador mede o verdadeiro valor da população.

Portanto, utilizando as estimativas do Exemplo 1 do material complementar da Unidade 3, para $\hat{\beta}_1$ e $\hat{\beta}_2$ e os erros padrão de $\hat{\beta}_1$ e $\hat{\beta}_2$, podemos calcular os intervalos de confiança. Assim, temos $\hat{\beta}_2$ =1, *ep*($\hat{\beta}_2$)=0,16666 e *gl=n-k=10-2=8*. E, se admitirmos que α=5%, ou seja, um coeficiente de confiança de 95%, então o Quadro *t* bicaudal (veja o Anexo I da Unidade 3), nos mostra que $t_{n\text{-}k,\alpha/2}$, denominado de *t*-crítico, é $t_{8;0,025}$ = 2,306. A Figura 3.11 abaixo ilustra o intervalo de confiança e seus limites inferior e superior.

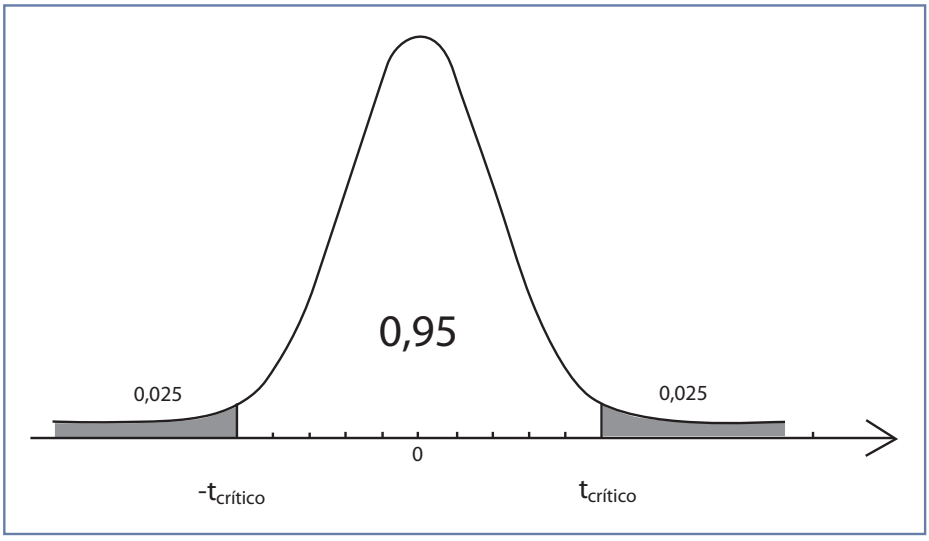


Figura 3.11 – Ilustração dos valores tcríticos que caracterizam os limites do intervalo de confiança.

No caso do Exemplo 1, em estudo nesta unidade, obtém-se $|t_{crítico}|$ para $\alpha = 0{,}06$ e $gl = 8$, conforme você pode consultar no quadro *t* bicaudal no Anexo I, seguindo o procedimento esquematizado no Quadro 3.4, abaixo.

GRAUS DE LIBERDADE (GL)	ÁREA CAUDAL SUPERIOR + INFERIOR								
	0,30	0,20	0,10	0,05	0,04	0,02	0,01	0,002	0,001
...	...	...	...	↑	...	...	...	...	...
8	1,108	1,397	1,860	2,306	2,449	2,896	3,355	4,501	5,041
...	...	...	...	...	...	...	...	...	...

Quadro 3.4 – Distribuição *t* bi-caudal.

Portanto, utilizando os valores do $t_{crítico} = 2{,}306$ obtido, e o valor do erro padrão de $\hat{\beta}_2$ (da ordem de 0,16666), dado pela equação (9) do Exemplo 1 do material complementar, e substituindo na equação (142), verifica-se que o intervalo de confiança de 95% para β_2 é o seguinte:

$$1 - 2{,}306 \times 0{,}1666 \leq \beta_2 \leq 1 + 2{,}306 \times 0{,}16666 \quad \rightarrow 0{,}615 \leq \beta_2 \leq 1{,}384 \quad \textbf{(146)}$$

A interpretação deste intervalo de confiança é: dado o coeficiente de confiança de 95%, no longo prazo, em 95 de 100 casos, intervalos como 0,615 e 1,384 conterão o verdadeiro β_2. A probabilidade de o intervalo específico fixado incluir o verdadeiro β_2 é, portanto, de 95 em 100 amostras selecionadas.

Também, no exemplo em estudo, verificamos que $\hat{\beta}_1 = 2$, ep $(\widehat{\beta}_1) = 0{,}5916$ (equação (10) do Exemplo 1 do material complementar) e $gl = n-k = 10-2 = 8$. Se admitirmos que $\alpha = 5\%$, ou seja, um coeficiente de confiança de 95%, então o Quadro *t* bicaudal do Anexo 1 nos mostra que o $t_{n-k,\alpha/2}$, denominado de t-crítico, é $t_{8;0,025} = 2{,}306$, conforme já estimado anteriormente. Assim, o intervalo de confiança de 95% para $\hat{\beta}_1$ do Exemplo 1, é obtido aplicando a equação (144):

$$2 - 2{,}306 \times 0{,}5916 \leq \beta_1 \leq 2 + 2{,}306 \times 0{,}5916 \quad \rightarrow 0{,}635 \leq \beta_1 \leq 3{,}365 \qquad \textbf{(147)}$$

No longo prazo, intervalos como (148), que você verá a seguir, conterão o β_1 verdadeiro em 95 dentre 100 casos: a probabilidade de que este particular intervalo fixado inclua o β_1 verdadeiro é de 95 em 100 amostras selecionadas.

3.6.3 Intervalos de confiança para Σ^2

Sob a hipótese da normalidade, a variável

$$X^2 = (n-k)\frac{\hat{\sigma}^2}{\sigma^2} \qquad \textbf{(148)}$$

segue a distribuição X^2 com $(n-k)$ gl.

Portanto, pode-se usar a distribuição X^2 para estabelecer o intervalo de confiança para σ^2. Conforme a equação abaixo:

$$\Pr\left(X^2_{1-\alpha/2} \leq X^2 \leq X^2_{\alpha/2}\right) = 1-\alpha \qquad \textbf{(149)}$$

O valor de X^2 no meio da dupla desigualdade em (149) é dado pela equação (148) e $X^2_{1-\alpha/2}$ e $X^2_{\alpha/2}$ são dois valores críticos de X^2, obtidos do quadro de qui-quadrado para $(n-k)$ gl, de tal maneira que eles contenham $100x(\alpha/2)\%$ das áreas extremas das caudas da distribuição X^2, como mostra a Figura 3.12, a seguir.

Portanto, substituindo χ^2 de (148) em (149), e rearranjando os termos, obtemos

$$PR\left[(n-k)\frac{\hat{\sigma}^2}{X^2_{1-\alpha/2}} \leq \sigma^2 \leq (n-k)\frac{\hat{\sigma}^2}{X^2_{\alpha/2}}\right] = 1-\alpha \qquad \textbf{(150)}$$

que nos dá o intervalo de confiança de $100(1-\alpha)\%$ para σ^2.

Nas estimativas do Exemplo 1 do material complementar, obteve-se $\hat{\sigma}^2 = 1$, conforme a equação (8) (do exemplo) e gl = 8. Se α for escolhido como 5%, o quadro de qui-quadrado, tomando a cauda à direita para 8 gl, nos dá os seguintes valores críticos: $X^2_{0,975} = 17,5346\ e\ X^2_{0,025} = 2,1797$. Estes valores mostram que a probabilidade de um valor qui-quadrado exceder 17,5346 é de 97,5% e de exceder 2,1797 é de 2,5% para X^2, como mostra a Figura 3.12 (note a característica assimétrica da distribuição de qui-quadrado):

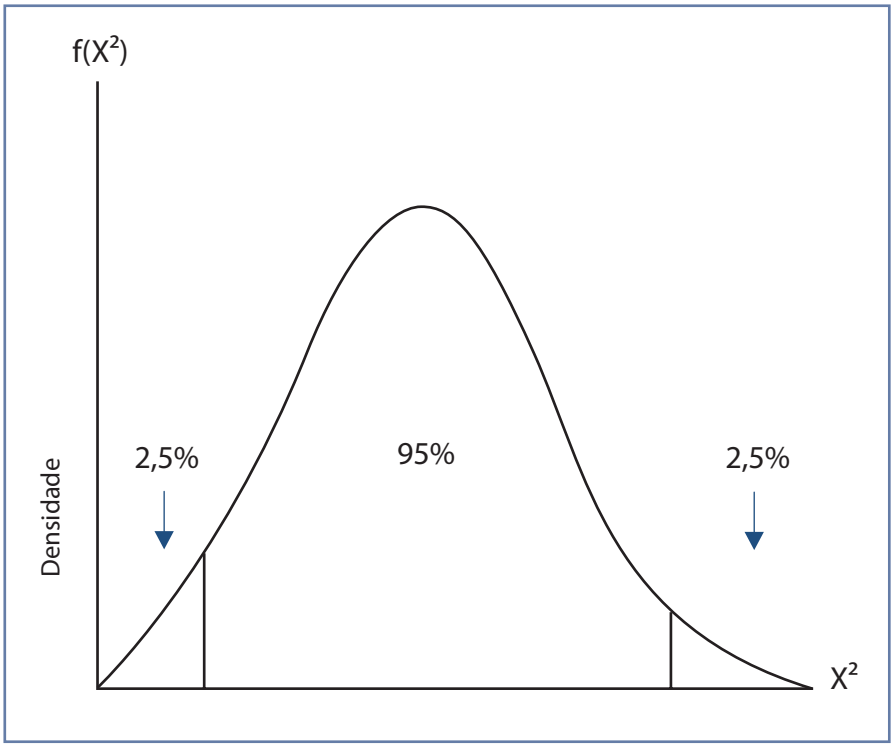


Figura 3.12 – O intervalo de confiança de 95% para X² (8 gl).

Substituindo os resultados de $\chi^2_{crítico} \approx 17,534$ e $\chi^2_{crítico} \approx 2,1797$ e de $\hat{\sigma}^2 = 1$ na equação (150), verificamos que o intervalo de confiança de 95% para σ² é o seguinte:

$$\frac{8 \times 1}{17,534} \leq \sigma^2 \leq \frac{8 \times 1}{2,1797} \quad \rightarrow 0,456 \leq \sigma^2 \leq 3,670 \qquad \textbf{(151)}$$

A interpretação deste intervalo é: se estabelecermos os limites de confiança de 95% para σ² e se afirmarmos, a *priori*, que estes limites irão incluir o σ² verdadeiro, em longo prazo estaremos certos em 95% das vezes.

3.7 TESTE DE HIPÓTESE: COMENTÁRIOS GERAIS

Palavra do Professor

Ao problema do teste de hipótese, relaciona-se a seguinte pergunta: uma dada observação ou descoberta é compatível ou não com alguma hipótese formulada? A palavra *compatível*, aqui, significa *suficientemente* próxima do valor admitido por hipótese, para que não rejeitemos a hipótese formulada.

Assim, se alguma teoria ou experiência anterior nos levasse a acreditar que o verdadeiro coeficiente de inclinação β_2 do Exemplo 1 do material complementar é 1,2, então:

- o $\hat{\beta}_2$ estimado (≈1) (obtido da amostra do Exemplo 1 em estudo), seria consistente com a hipótese formulada? Se o for, não rejeitaremos a hipótese. Caso contrário, podemos rejeitá-la;
- a hipótese formulada é conhecida como hipótese nula, indicada pelo símbolo H_o;
- a hipótese nula é usualmente testada contra uma hipótese alternativa (também conhecida como hipótese sustentada), indicada, por H_1, que pode afirmar, por exemplo, que o verdadeiro β_2 é diferente de 1;
- a hipótese alternativa pode ser simples ou composta. Por exemplo, $H_1 : \beta_2 = 1{,}5$ é uma hipótese simples, mas $H_1 : \beta_2 \neq 1{,}5$ é composta.

Há duas abordagens, mutuamente complementares, que delineiam essas regras: **intervalo de confiança** e **teste de significância**.

3.7.1 Teste de hipótese: a abordagem do intervalo de confiança

No Exemplo 1 em análise, como se sabe, a inclinação estimada $\hat{\beta}_2$ é 1. Suponha que se postule que:

$$\begin{cases} H_0 : \beta_2 = 1{,}5 \\ H_1 : \beta_2 \neq 1{,}5 \end{cases} \quad \textbf{(152)}$$

A hipótese nula é hipótese simples, enquanto a hipótese alternativa é composta, conhecida como hipótese bilateral.

Portanto, o $\hat{\beta}_2$ observado é compatível com H_o? Observando o intervalo de confiança determinado anteriormente, conforme resultado na equação (146), temos:

$$0{,}615 \leq \beta_2 \leq 1{,}384 \qquad \textbf{(153)}$$

Portanto, sabe-se que intervalos como 0,615 e 1,384 não irão conter o verdadeiro β_2, com 95% de probabilidade:

- tais intervalos fornecem uma classe ou limites dentro dos quais o verdadeiro β_2 pode estar com um coeficiente de confiança de, digamos, 95%;
- então, o intervalo de confiança fornece um conjunto de hipóteses nulas plausíveis.

Assim, se β_2, segundo H_o, estiver dentro do intervalo de confiança de 100(1-α)%, não se rejeita a hipótese nula; se ele estiver fora do intervalo, pode-se rejeitá-la.

Assim, este intervalo é ilustrado esquematicamente na Figura 3.13:

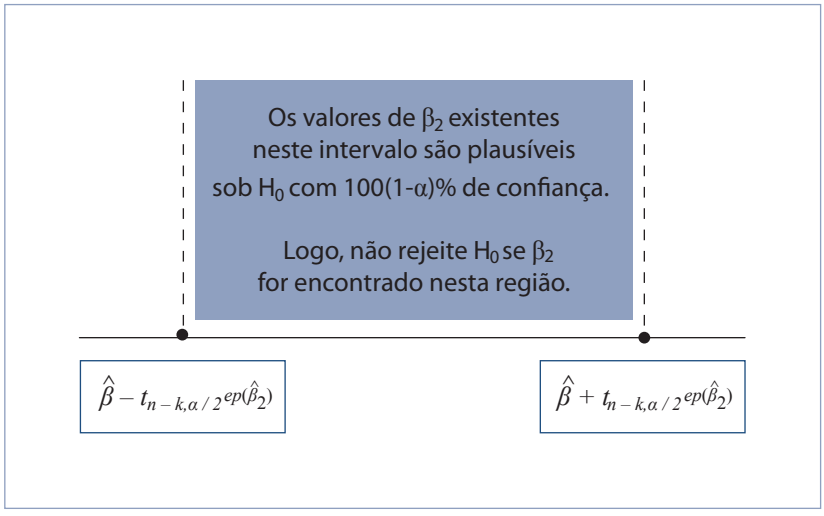


Figura 3.13 – Intervalo de confiança de 100(1-α)% para β_2.

Portanto, tem-se a seguinte regra de decisão:

- construa-se um intervalo de confiança de 100(1-α)% para β_2. Se β_2, segundo H_0, se encontrar dentro deste intervalo de confiança, não se rejeita H_0. Mas, se β_2 for encontrado fora deste intervalo, rejeite H_0;

- no Exemplo 1 em análise, $H0 : \beta_2 = 1{,}5$ está claramente fora do intervalo de confiança de 95% dado na equação (153). Logo, pode-se rejeitar a hipótese de que 1,53 é a verdadeira participação nos exercícios na sala de aula, com 95% de confiança;
- se a hipótese nula fosse verdadeira, a probabilidade de obtermos $\widehat{\beta}_2 = 1{,}5$ por mero acaso, seria de, no máximo, cerca de 5%, uma probabilidade pequena.

3.7.2 Teste de hipótese: a abordagem do teste de significância

A abordagem do teste de significância trata de testar a significância dos coeficientes de regressão por meio do teste *t*. Ou seja:

- em linhas gerais, teste de significância é um procedimento pelo qual os resultados da amostra são usados para verificar a validade ou a falsidade de uma hipótese nula;
- a decisão de aceitar ou rejeitar H_0 é tomada com base no valor da estatística de teste obtida com os dados disponíveis.

Para ilustrar, lembre-se de que, pela hipótese de normalidade, a variável

$$t = \frac{\hat{\beta}_2 - \beta_2}{ep(\hat{\beta}_2)} \qquad \textbf{(154)}$$

segue a distribuição *t* com (*n* -*k*) gl (onde *k* é o número de parâmetros no modelo e *n* o número de amostras).

Se o valor verdadeiro β_2 estiver especificado sob a hipótese nula, o valor *t* de (154) poderá ser facilmente calculado a partir da amostra disponível, e, portanto, poderá servir como uma estatística de teste.

E, como esta estatística de teste (ou seja, o valor t em (154)) segue a distribuição *t*, pode-se fazer declarações de intervalos de confiança com ela, como vemos a seguir:

$$\Pr\left[-t_{n-k,\alpha/2} \leq \frac{\hat{\beta}_2 - \hat{\beta}_2^*}{ep(\hat{\beta}_2)} \leq t_{n-k,\alpha/2} \right] = 1 - \alpha \qquad \textbf{(155)}$$

onde β_2^* e o valor de β_2 sob H_o, $-t_{n-k,\alpha/2}$ e $t_{n-k,\alpha/2}$ são os valores de t (os valores críticos de t), obtidos do quadro de t do Anexo I da Unidade 3, para um nível de significância de $(\alpha/2)$ e $(n - k)$ gl.

Rearranjando (155), obtemos:

$$\Pr\left[-t_{n-k,\alpha/2} \le t \le t_{n-k,\alpha/2})\right] = 1-\alpha \qquad \textbf{(156)}$$

Assim, o intervalo de confiança de 100(1-α)%, estabelecido em (156), é conhecido como região de aceitação da hipótese nula. A região (ou regiões) fora do intervalo de confiança é chamada de região de rejeição de H_o ou região crítica. Estimada a estatística t em (154), e estimado o $t_{critíco}$, se a estatística t cair dentro do intervalo (156), então se aceita a hipótese nula; caso contrário, rejeita-se.

No Exemplo 1 em análise, encontramos $\hat{\beta}_2 = 1$, $ep\ (\widehat{\beta}_2) = 0{,}1666$ e gl=n-k=10-2=8. Se admitirmos α=5%, $t_{n-k,\alpha/2}=t_{8,0,025}=2{,}306$, conforme o quadro da estatística t no Anexo I, então podemos estabelecer o seguinte teste de hipótese $H_0 : \beta_2 = \beta_2^* = 1{,}5 \quad e \quad H_1 : \beta_2 \neq 1{,}5$. Para tanto, devemos estimar o valor de t, partindo-se da relação **(154)** que se torna:

$$t = \frac{\hat{\beta}_2 - \beta_2}{ep(\hat{\beta}_2)} = \frac{1-1{,}5}{0{,}1666} = -3{,}00 \qquad \textbf{(157)}$$

Consequentemente, como $2{,}306 < t = 3{,}00$, conforme o cálculo em (157), então o t cai na região crítica, o que nos permite rejeitar a hipótese nula e aceitar a hipótese alternativa, sob a qual o $\beta_2 = 1{,}5$ apresenta somente 5% de chance de ser verdadeiro, com o nível de significância assumido.

A Figura 3.14 evidencia claramente que a estatística t situa-se na região crítica. A conclusão permanece a mesma, a saber: rejeita-se H_o.

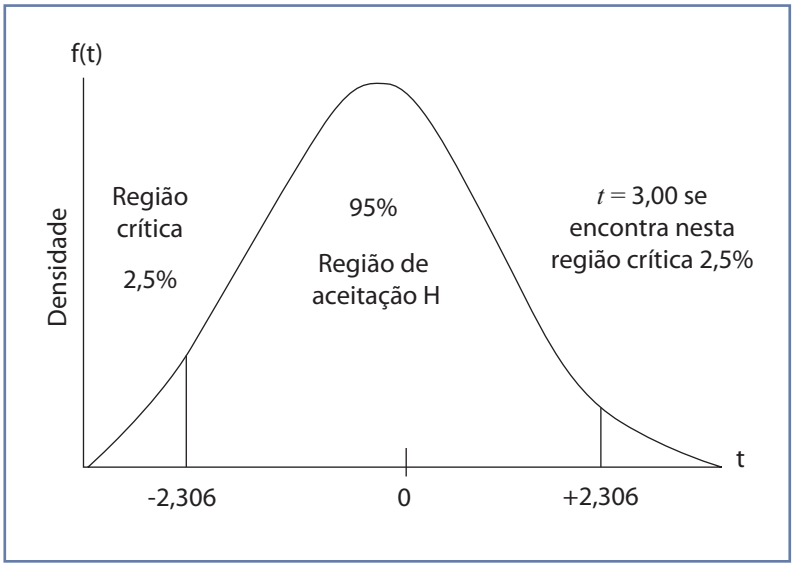


Figura 3.14 – Intervalo de confiança de 95% para t (8 gl).

Logo, de acordo com a relação (154), um valor $|t|$ alto será um indício contra a hipótese nula.

O procedimento do teste *t* anterior é conhecido como teste de significância bilateral ou bicaudal. Isto porque H_1 é uma hipótese composta bilateral; ou seja, H_1: $\beta_2 \neq 1{,}5$ o que significa que β_2 é menor ou maior que 1,5.

Assim, na linguagem dos testes de significância, diz-se que:

- Uma estatística é estatisticamente significante se seu valor encontra-se na região crítica. Neste caso, a hipótese nula é rejeitada. E, um teste é estatisticamente insignificante se o valor da estatística de teste se encontrar na região de aceitação. Nesta situação, a hipótese nula é aceita.

Abordamos, nesta Unidade, somente testes de significância bilaterais. No entanto, o princípio de abordagem de testes de significância ou de intervalos de confiança unilaterais segue o mesmo principio, conforme será apresentado no material complementar disponibilizado no AVEA.

Palavra do Professor

Como pode ser constatado ao longo desta Unidade, utilizamos como exemplo ilustrativo o desenvolvimento de um modelo de regressão linear, no qual fatores de participação nas soluções de exercícios na sala de aula explicam a nota obtida na prova associada ao conteúdo dos exercícios. Observa-se que o modelo aplicado foi um modelo linear envolvendo duas variáveis; contudo, mesmo na solução de um problema de regressão elementar como esse, a quantidade de cálculos envolvidos na operação é excessiva. Portanto, torna-se evidente que problemas reais envolvendo modelos de regressão de multivariáveis não podem ser resolvidos através de cálculos manuais. Assim, a aplicação adequada dos conhecimentos econométricos exige o uso de um *software* apropriado para tal finalidade. E o aprendizado da utilização de alguns *softwares* especializados no assunto também faz parte do estudo desta disciplina. Para tanto, utilizaremos o *software* livre GRETL 1.8.1 em aplicações que estão ilustradas no material complementar desta Unidade, no AVEA. Também, algumas atividades para treinamento do uso desse *software* foram colocadas à sua disposição no ambiente virtual, as quais farão parte da composição de sua avaliação.

Agora que você terminou esta Unidade, não esqueça de assistir à Videoaula 3. Ah! Lembre-se também de resolver as Atividades complementares que estão no AVEA; afinal, elas farão parte da sua avaliação. Bom trabalho!

Atividade de Aprendizagem – 3

1) Em um modelo de regressão linear simples ($Y_t = \beta_1 + \beta_2 X_t + u_t$), a inclinação da regressão indica:

   a) A percentagem que *Y* é aumentado quando *X* sofre 1% de aumento.

   b) Que ao multiplicá-la pela variável explicativa, o resultado dará a variável de previsão, *Y*.

   c) Quantas unidades *Y* aumenta quando *X* for aumentado de uma unidade.

   d) A elasticidade de *Y* sobre *X*.

   Justifique a sua resposta.

2) Quando o coeficiente de inclinação estimado num modelo de regressão simples ($Y_t = \beta_1 + \beta_2 X_t + u_t$) é zero, ou seja, $\hat{\beta}_2 \approx 0$, então:

   a) $R^2 = \overline{Y}$.

   b) $< R^2 < 1$.

   c) $R^2 = 0$.

   d) $R^2 > (SQR/SQT)$.

   Justifique a sua resposta.

3) O estimador MQO (Mínimos Quadrados Ordinários) é obtido:

   a) Conectando o maior Y_i ao menor X_i.

   b) Fazendo com que o erro padrão da regressão seja igual ao erro padrão da inclinação estimada.

   c) Minimizando a soma dos resíduos absolutos.

   d) Minimizando a soma dos resíduos ao quadrado.

   Justifique a sua resposta.

4) A média dos resíduos do estimador MQO (Mínimos Quadrados Ordinários) deve ser:

   a) Algum número positivo, pois o estimador MQO usa resíduos ao quadrado.

   b) Zero, se o modelo de regressão for corretamente especificado, sem a omissão de alguma variável importante, determinística da variável dependente.

   c) Não observável, pois a função de regressão da população é desconhecida.

   d) Tal que ele depende se a variável explicativa é positiva ou negativa.

Justifique a sua resposta.

5) Num modelo de regressão linear simples ($Y_t = \beta_1 + \beta_2 X_t + u_t$), quando o estimador da inclinação $\hat{\beta}_2$ tem um erro padrão pequeno, então:

a) A base de dados é tal que existe grande variação na variável explicativa, com relação à sua média.

b) Existe uma grande variância do termo erro estocástico u_i.

c) O tamanho da amostra é pequeno.

d) O intercepto $\hat{\beta}_1$ é pequeno.

Justifique a sua resposta.

6) O coeficiente de determinação $\boldsymbol{r}^2$ é uma medida:

a) Se ou não *X* causa *Y*.

b) Da precisão do ajuste do modelo de regressão.

c) Se ou não *SQE* > *SQT*.

d) Do grau de dependência positiva associada ao coeficiente de correlação *r*.

Justifique a sua resposta.

7) A linha de regressão amostral estimada pelo MQO (Mínimos Quadrados Ordinários) $Y_t = \beta_1 + \beta_2 X_t + u_t$:

a) Terá sempre uma inclinação menor que o intercepto.

b) É exatamente a mesma que a linha de regressão da população.

c) Não pode ter uma inclinação zero.

d) Terá que passar pelo ponto ($\overline{X}, \overline{Y}$).

Justifique a sua resposta

8) Interpretando o intercepto de uma função de regressão amostral, podemos afirmar que:

a) O intercepto não possui estimativas consistentes do ponto de vista econômico, pois os valores observados da variável explicativa nunca envolvem a origem.

b) O intercepto estimado sempre será consistente do ponto de vista econômico, porque, em tais estimativas, o intercepto sempre apresentará correlação com a inclinação.

c) O intercepto somente possui estimativas consistentes do ponto de vista econômico quando os valores observados da variável explicativa envolvem a origem.

d) O intercepto não possui estimativas consistentes do ponto de vista econômico; contudo, isto não é um problema, pois os economistas somente estão interessados no efeito da mudança de *X* sobre *Y*.

Justifique a sua resposta

9) A estatística *t* de Student de um modelo de regressão linear, $Y_t = \beta_1 + \beta_2 X_t + u_t$, é estimada dividindo:

a) O parâmetro estimado pelo MQO (Mínimos Quadrados Ordinários), pelo seu erro padrão.

b) A inclinação pelo desvio padrão da variável explicativa.

c) O parâmetro estimado menos o seu valor hipotetizado, pelo seu erro padrão estimado.

d) A inclinação pelo intercepto.

Justifique a sua resposta.

10) Com relação aos resíduos de uma estimativa de um modelo de regressão amostral pelo MQO (Mínimos Quadrados Ordinários), podemos afirmar que:

a) Eles podem ser calculados usando os erros da função de regressão estimada.

b) Eles podem ser calculados subtraindo o valor ajustado da variável ajustada pelo valor atual da amostra.

c) Eles são desconhecidos, pois nós não conhecemos a função de regressão populacional.

d) Na prática, eles não são utilizados, pois num modelo de regressão não se utilizam todos os valores observados da amostra.

Justifique a sua resposta.

11) Se o valor absoluto da estatística *t* de Student de um parâmetro de um modelo de regressão amostral exceder um valor crítico, determinado com (*n*-*k*) graus de liberdade (onde *n* é o tamanho da amostra e *k* é o número de parâmetros estimados no modelo), e com um nível de significância de $\alpha/2$, pode-se afirmar que:

a) A hipótese nula é rejeitada.

b) Seguramente, assume-se que os resultados da regressão são significantes.

c) O pressuposto de que os erros estocásticos são homoscedásticos é rejeitado.

d) Os valores amostrais estão concentrados muito próximos da curva de regressão.

Justifique a sua resposta.

12) Para concluir, se a inclinação do modelo $Y_t = \beta_1 + \beta_2 X_t + u_t$ é grande ou pequena, deve-se:

a) Verificar a importância econômica de um determinado aumento em *X*.

b) Verificar se a inclinação é maior do que um.

c) Verificar se a inclinação é estatisticamente significante.

d) Mudar a escala da variável *X*, se o coeficiente de inclinação parecer ser muito pequeno.

Justifique a sua resposta.

13) O quadro abaixo apresenta as notas (entre 0 e 10) de 10 alunos que fizeram o Curso de Estatística e, na atualidade, cursam Econometria. Pretende-se verificar as relações de explicação entre os conhecimentos nas duas disciplinas, de tal modo que se espera que uma boa nota em Estatística signifique que o aluno tem conhecimento básico para obter uma boa nota em Econometria. Para tanto, você deve determinar um modelo de regressão, os parâmetros e as estimativas, e os erros associados, para realizar as análises de inferências solicitadas abaixo.

	NOTAS	FALTAS
DADOS AMOSTRAIS	10	5
	11	4
	4	10
	1	11
	10	3
	2	12
	5	8
	13	2
	9	5
	5	10

Suponha que o modelo de regressão linear estabelecido para o problema seja $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$ (onde a variável *Y* representa a variável dependente, e *X* a variável explicativa). Para tanto, seguem os seguintes passos:

a) Defina qual é a variável dependente e a variável independente (explicativa) para o modelo, e o porquê desta escolha (ou seja, defina a direção de causalidade no modelo). Ainda com relação a esta questão, defina como se comportam as variáveis dependentes e independentes num modelo de regressão, em termos estocásticos e determinísticos.

b) Determine os fatores necessários para calcular as estatísticas que se seguem (itens abaixo), e represente-os num quadro (para maior facilidade, utilize as planilhas do Excel e seus recursos).

c) Estime $\hat{\beta}_i \; e \; \hat{\beta}_2$.

d) Estime a variância e o erro padrão da estimativa e explique para que servem estas estimativas.

e) Determine $\hat{\sigma}^2$, $Ep(\hat{\beta}_1)$, $Ep(\hat{\beta}_2)$, $cov(\hat{\beta}_1, \hat{\beta}_2)$, r^2 e r.

f) Explique para que servem as estatísticas $\hat{\sigma}^2$, $Ep(\hat{\beta}_1)$, $Ep(\hat{\beta}_2)$, $cov(\hat{\beta}_1, \hat{\beta}_2)$, r^2 e r.

g) Explique qual é o grau de ajuste do modelo e o grau de associação entre a variável *nota* em Econometria e a variável *nota* em Estatística.

h) Estime as estatísticas *t* e verifique as significâncias estatísticas de $\hat{\beta}_i \; e \; \hat{\beta}_2$ (monte os testes de hipóteses de intervalos e de nível de significância. Para tanto, considere α=1%).

i) Verifique se a hipótese do professor de Econometria é verdadeira (com um nível de significância de 1%), a qual afirma que um aluno que desconhece totalmente os conceitos de Estatística (nota igual a zero), obterá, no mínimo, uma nota em Econometria igual a 2 (dois).

j) O professor de Econometria também tem assumido a hipótese de que os alunos de sua disciplina obterão, no mínimo, uma nota igual à obtida em Estatística. Verifique se esta hipótese pode ser assumida como verdadeira, com um nível de significância de 1%.

REFERÊNCIAS

ALMEIDA, André Quintão de *et al*. **Enhanced vegetation index (EVI) na análise da dinâmica da vegetação da Reserva Biológica de Sooretama, Espírito Santo.** Disponível em: http://www.scielo.br/pdf/rarv/v32n6/a15v32n6.pdf. Acesso em: 23 jun.2009.

CHOU, Yan-lun. **Statistical analysis for business and economics.** Elsevier Science Publishing Co., 1989.

ENDO, S. K. **Métodos quantitativos:** números índices. Ribeirão Preto: Atual Editora ltda, 1988.

FISHER, Irving.**The Making of index numbers.** Boston, USA: Houghton Mifflin, 1922.

FONSECA, J.S.; MARTINS, G.A.; TOLEDO, G. L. **Estatística aplicada.** São Paulo: Atlas, 1988.

GUJARATI, D. **Econometria básica.** São Paulo: Editora Campus, 2006.

GUJARATI, Damodar N. **Econometria básica.** Rio de Janeiro: Elsevier - Editora Campus, 2006.

.HILL, R. C.; GRIFFITHS, W.E.; JUDGE, G.G. **Econometria.** São Paulo: Saraiva. 2003.

KENDALL, M.; STUART, A. **The advanced theory of statistics.** Londres: Charles Griffin, 1977. v.1.

MILONI, G.; ANGELINI, F. **Estatística aplicada:** números-índices, regressão e correlação e séries temporais. São Paulo: Atlas, 1995.

PINDYCK, R. S.; RUBINFELD, D.L. **Econometria:** modelos e previsões. 4. ed. Rio de Janeiro: Elsevier, 2004.

PINDYCK, R.S.; RUBINFELD, D.L. **Econometric models and economic forecasts**, 4.ed. São Paulo: McGraw-Hill, 1997.

STOCK, James H. ; WATSON, Mark W. **Econometria.** São Paulo: Pearson - Addison-Wesley, 2005.

WOOLDRIDGE, Jefrey M. **Introdução à econometria:** uma abordagem moderna. São Paulo: Thomson Learning, 2006.

PORTAL BRASIL. **Índices financeiros brasileiros.** Disponível em: http://www.portalbrasil.eti.br/indices.htm/. Acesso em: 23 jun.2009.

FIPE – Fundação Instituto de Pesquisas Econômicas: Disponível em: http://www.fipe.com/. Acesso em: 23 jun.2009.

FGV – Fundação Getúlio Vargas.Disponível em: http://www.fgv.br/ibre/CEP/index.cfm. Acesso em: 23 jun.2009.

IBGE – Instituto Brasileiro de Geografia e Estatística. Disponível em: http://www.sidra.ibge.gov.br/. Acesso em: 23 jun.2009.