SECRETARIA DE PLANEJAMENTO, ORÇAMENTO E COORDENAÇÃO
FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA- IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 0034 - 7175
R. bras. Estat., Rio de Janeiro, v.51, n.196, p. 1 - 109, jul. / dez. 1990

REVISTA BRASILEIRA DE ESTATÍSTICA

Órgão oficial do IBGE

Publicação semestral, editada pelo IBGE, que se destina a promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais, através de divulgação de artigos inéditos.
Temas, abordando aspectos do desenvolvimento metodológico, serão aceitos desde que relevantes para os órgãos produtores de estatísticas.
Os originais para publicação deverão ser submetidos em 3 vias (que não serão devolvidas) para:

Djalma G. C. Pessoa
Editor-Responsável - RBEs

ENCE
Rua André Cavalcante, 106 - Bairro de Fátima
20231 - 050 - Rio de Janeiro - RJ

- Os artigos submetidos à RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.
- Cada autor receberá, gratuitamente, 20 separatas de seu artigo.

A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Capa
Pedro Paulo Machado

Revista brasileira de estatística / Fundação Instituto Brasileiro de Geografia e Estatística. - v.1, n.1 (jan./mar. 1940)- . - Rio de Janeiro: IBGE, 1940-

v.

Trimestral (1940-1986), semestral (1987-)
Órgão oficial do IBGE.
Continuação de: Revista de economia e estatística.
Índices acumulados de autor e assunto publicados no v. 43 (1940-1979) e v. 50 (1980-1989)

ISSN 0034-7175 - Revista brasileira de estatística

1. Estatística - Periódicos. I. IBGE.

IBGE.CDDI. Dep. de Documentação e Biblioteca CDU31(05)
RJ-IBGE/88-05 rev.

Impresso no Brasil / Printed in Brazil

SUMÁRIO

ARTIGOS

RESENHAS BIBLIOGRÁFICAS

REGIONALIZAÇÃO DA AGRICULTURA SEGUNDO INDICADORES SOCIAIS

Angela Kageyama*
Eugênia Troncoso Leone*

1 INTRODUÇÃO

O objetivo deste artigo é mostrar as diferenças regionais das condições sociais da população dependente das atividades agrícolas no Brasil, por meio da aplicação de um método de análise fatorial (método dos componentes principais).

De um lado, pretende-se contribuir para o conhecimento das condições sociais, por meio dos Indicadores Sociais, no caso particular da agricultura, apontando algumas de suas características e limitações. De outro, procura-se ilustrar a aplicação do método dos componentes principais em estudos de regionalização, de forma a indicar um possível caminho metodológico para ser testado em situações nas quais sejam disponíveis informações mais desagregadas (por microrregião ou município). Na verdade, a regionalização obtida neste trabalho, por considerar as Unidades da Federação como observações, portanto bastante agregadas, é até certo ponto óbvia e esperada. Espera-se, contudo, que o estudo possa servir como guia para a reprodução da metodologia dentro de cada estado, tomando como unidades as microrregiões ou os municípios.

Optou-se por considerar como população-base a População Economicamente Ativa de 10 anos ou mais na agropecuária, extração vegetal e pesca ("PEA Agropecuária").

*Professoras do Instituto de Economia da UNICAMP. Caixa Postal 6135, Campinas - SP, CEP 13081-970.
Participaram os estagiários Sérgio A. Amed e Silva e Lucas Calegari Jr.
As autoras agradecem as sugestões de um parecer anônimo da Revista, as quais contribuíram para melhorar o trabalho, sobretudo no que se refere ao agrupamento dos estados propostos.

No entanto, alguns indicadores importantes de bem-estar são apresentados apenas pela situação do domicílio (rural ou urbano) e não por setor de atividade da PEA. Nesses casos, dada a limitação própria dos dados, será utilizada a situação do domicílio como *proxy* da população ocupada nas atividades agrícolas.

Conforme os dados da Tabela 1, verifica-se que a proporção da PEA Agropecuária, que reside em domicílios rurais, é bastante variável entre as Unidades da Federação. Varia de um máximo de 91,4% no Acre a um mínimo de 61,6% em São Paulo. Neste e em outros casos, em que a porcentagem da PEA Agropecuária com domicílio na zona rural é relativamente baixa, a aproximação utilizada torna-se menos representativa.

TABELA 1

Pessoas Economicamente Ativa de 10 anos ou mais na Agropecuária, Extração Vegetal e Pesca, com Domicílio Rural, Segundo as Unidades da Federação – 1980

UNIDADES DA FEDERAÇÃO	PEA AGROPECUÁRIA (%)	PEA AGROPECUÁRIA (NÚMERO DE PESSOAS)
BRASIL	**81,8**	**12 661 017**
Rondônia	88,0	89 167
Acre	91,4	48 134
Amazonas	87,0	176 680
Roraima	83,4	9 592
Pará	86,8	440 668
Amapá	81,1	10 887
Maranhão	89,8	884 472
Piauí	89,5	395 774
Ceará	84,7	741 215
Rio Grande do Norte	79,0	239 160
Paraíba	82,1	412 609
Pernambuco	83,1	788 356
Alagoas	84,4	323 683
Sergipe	84,6	149 794
Bahia	87,3	1 464 985
Minas Gerais	76,3	1 518 442
Espírito Santo	84,0	242 241
Rio de Janeiro	75,5	195 580
São Paulo	61,6	1 175 002
Paraná	84,0	1 182 082
Santa Catarina	89,7	418 249
Rio Grande do Sul	89,5	903 641
Mato Grosso do Sul	74,2	176 126
Mato Grosso	79,9	162 318
Goiás	73,0	501 216
Distrito Federal	63,8	10 898

FONTE – Censo Demográfico de 1980.

Em linhas muito gerais, quanto maior for a coincidência entre a PEA Agropecuária e as pessoas ocupadas que residem em domicílios rurais, menor deve ser o desenvolvimento das forças produtivas, isto é, menor o grau de urbanização. Geralmente, o desenvolvimento industrial de uma região aumenta o grau de urbanização; isto, por sua vez, geralmente também está associado à modernização da agricultura, com a

conseqüente proletarização da força de trabalho agrícola. Os trabalhadores agrícolas passam a residir nas áreas urbanas, onde inclusive engajam-se em ocupações acessórias para complementar a renda obtida sazonalmente na agricultura. Assim, quanto mais densa a rede urbana do estado e quanto mais moderna sua agricultura, menor a fração da PEA Agropecuária que ainda reside na zona rural. É o caso típico de São Paulo, onde a mão-de-obra "volante", que representa cerca de um terço das pessoas ocupadas na agricultura, geralmente mora nas periferias urbanas, onde alterna empregos agrícolas com atividades urbanas (construção civil, emprego doméstico, etc.).

Duas observações precisam ser feitas ainda sobre o tipo de informação da Tabela 1.

A primeira é que o local da residência (rural ou urbana) não indica maior ou menor bem-estar, por si só. Mesmo o acesso aos bens e serviços que são típicos das aglomerações urbanas (por exemplo, eletrificação, saneamento, água encanada, escolas, hospitais, centros de lazer, etc.) pode ter um significado bem diferente para os moradores de baixa renda das periferias. Além disso, pode-se ter um nível de bem-estar distinto entre residências isoladas na zona rural e aquelas localizadas em "aglomerados rurais" ou pequenas vilas e comunidades em que a escala torna viável a implantação dos serviços básicos de infra-estrutura. Assim, o corte rural/urbano tende a ser, em muitos casos, meramente administrativo, pouco ou nada revelando a respeito das condições de vida da população.

A segunda observação é que a associação inversa entre a proporção da PEA com domicílio rural e o grau de desenvolvimento das forças produtivas nem sempre ocorre. Note-se, por exemplo, que estados tão diversos como Maranhão, Piauí, Santa Catarina e Rio Grande do Sul possuem cerca de 90% da PEA Agropecuária com domicílio rural; no Rio Grande do Norte a proporção de "PEA Rural" é de 79% e, no Paraná, que tem uma agricultura muito mais moderna, essa proporção é maior (84%). Assim, a "urbanização" da PEA Agropecuária certamente tem outros determinantes, podendo-se incluir aí especificidades locais (é o caso, por exemplo, da tradição da pequena propriedade familiar de Santa Catarina).

2 INDICADORES SOCIAIS E FONTES

Neste trabalho entende-se por desempenho social da agricultura o conjunto de efeitos do desenvolvimento capitalista na agricultura sobre as **variáveis que captam as condições sociais de vida e bem-estar da população ocupada no setor.** O objetivo central é elaborar um rol de indicadores que permitam avaliar esse desempenho social na agricultura.

A partir desta concepção, os chamados Indicadores Sociais constituem variáveis que medem, para um aspecto particular, o bem-estar da população ocupada no setor.

Podemos estender o conceito de Indicadores Sociais para abranger dois grandes

conjuntos de efeitos do desenvolvimento na agricultura: o desenvolvimento das relações sociais de produção e as condições sociais de existência da população ocupada.

Assim, por exemplo, o indicador "grau de urbanização" pode estar mais relacionado com o desenvolvimento geral das relações de produção, enquanto o indicador "renda média familiar" está mais ligado às condições de bem-estar da família.

Outros exemplos que podem ser mencionados são os indicadores demográficos, de produtividade e estrutura fundiária relacionados ao desenvolvimento geral das relações de produção e os indicadores de educação, saúde, habitação e pobreza, mais relacionados às condições de vida.

Uma questão que se coloca é se deve ou não tentar reduzir os diversos indicadores a um índice único que exprima o bem-estar da população num determinado momento.

Os principais problemas da construção de um índice único são: primeiro, a quantidade e qualidade das estatísticas disponíveis, que no Brasil estão longe ainda de ser satisfatórias; segundo, a especificação dos indicadores mais adequados para cada componente do índice e a definição da estrutura de ponderações.

De todos os problemas acima mencionados, o de mais difícil solução é o da estrutura de ponderações. Ela deveria refletir a contribuição relativa de cada indicador para o bem-estar da população. Entretanto, é difícil fixar um critério de modo objetivo.

Por isso, procurou-se, neste trabalho, contornar esse problema usando um método estatístico de análise multivariada (Método dos Componentes Principais)[1] que permitirá, a partir de um conjunto de variáveis, caracterizar as condições de vida e bem-estar da população que mora no campo.

As principais fontes de dados para o cálculo de Indicadores Sociais no Brasil são os Censos e PNADs, do IBGE. Algumas dificuldades relacionadas com sua utilização são:

a) A abrangência limitada para o setor agropecuário: existem poucas tabelas que cruzam informações com situação do domicílio (urbano ou rural);

b) O período considerado: os Censos são decenais, impedindo um acompanhamento de curto e médio prazos. As PNADs são anuais, porém contêm problemas de comparabilidade interna entre os diferentes anos; e

c) A compatibilidade entre as duas fontes: os Censos e PNADs são de difícil comparação devido às diferentes metodologias utilizadas.

3 OS INDICADORES SELECIONADOS

Para aplicar a análise fatorial foram selecionados 14 indicadores a partir do Censo Demográfico de 1980. Alguns desses indicadores referem-se à população ocupada em atividades agrícolas, outros aos domicílios rurais, podendo ser assim identificados:

1) ESPE = esperança de vida ao nascer (anos) para as pessoas com domicílio

rural (Fonte: IBGE, 1987. *Estatísticas Históricas do Brasil.* Séries Econômicas, Demográficas e Sociais de 1550 a 1985. Rio de Janeiro. p. 50).

2) MORT = taxa de mortalidade infantil (0/00) para a população com domicílio rural (Fonte: idem anterior).

3) MIGR = índice migratório = migração líquida a partir da área rural entre 1970 e 1980/população rural em 1970 (Fonte: Martine, George, 1987. *Os impactos sociais da modernização agrícola.* Editora Cortes. São Paulo. p. 64).

4) POBR = % de famílias rurais com rendimento familiar *per capita* até um quarto de salário mínimo.

5) RURA = % de pessoas residentes em domicílios rurais.

6) DORU = % de domicílios rurais rústicos e improvisados.

7) AGUA = % de pessoas residentes em domicílios rurais com água de poço ou nascente sem canalização interna e outra (que não rede geral, nem poço ou nascente com canalização interna).

8) SANI = % de pessoas residentes em domicílios rurais sem nenhum tipo de instalação sanitária.

9) TEVE = % de pessoas residentes em domicílios rurais com televisão.

10) GELA = % de pessoas residentes em domicílios rurais com geladeira.

11) ANAF = % de pessoas de 5 anos ou mais residentes em domicílios rurais que não lêem nem escrevem.

12) EMPR = % de empregados na PEA Agropecuária.

13) JORN = % da PEA Agropecuária que trabalha 49 horas ou mais por semana.

14) SALA = % de empregados agrícolas que receberam mais de 12 salários por ano (13º, 14º ou mais).

Os indicadores de número 4 a 14 foram calculados diretamente a partir do Censo Demográfico de 1980.

No Brasil, segundo dados do Censo Demográfico de 1980, um em cada três brasileiros residia ainda na área rural. Nas regiões menos desenvolvidas essa proporção chegou a abarcar a maioria da população. Dos estados, somente Rio de Janeiro, São Paulo e o Distrito Federal tinham uma proporção de população rural semelhante à de alguns países desenvolvidos.

De maneira geral, as condições de vida no campo no ano de 1980 eram bastante precárias. Algumas indicações dessa precariedade são reveladas pelos indicadores de tipo de domicílio (DORU), saneamento (SANI) e abastecimento de água (AGUA). Para o conjunto do País, quase um terço dos domicílios rurais era do tipo rústico ou improvisado e a maioria quase absoluta das pessoas que residiam em domicílios rurais não tinha a mais mínima condição sanitária. Uma proporção muito pequena da população vivia em domicílios com acesso à rede geral de abastecimento de água ou, se o domicílio era abastecido por poço ou nascente, ao menos com canalização interna (17,8%). Essa proporção era um pouco mais significativa nos estados mais

desenvolvidos.

O quadro domiciliar rural mostrou-se mais precário ainda no que diz respeito às instalações sanitárias. Neste caso, até os estados mais desenvolvidos atingem proporções expressivas de pessoas em domicílios rurais sem nenhum tipo de instalação sanitária.

A posse de utilidades dá uma indicação do padrão de consumo. Em 1980, o acesso a bens como geladeira e televisão foi em geral insuficiente. Nos estados mais desenvolvidos a proporção de pessoas com este tipo de utilidades era mais expressiva, destacando-se São Paulo com quase 57% da população rural com televisão e 41% com geladeira. A posse deste tipo de bens encontra-se estreitamente relacionada com o acesso à energia elétrica.

O acesso da população à educação é um indicador importante das condições de vida no campo (ANAF). Entretanto, o quadro educacional da área rural no início da década era desalentador. Metade das pessoas com 5 anos e mais não sabia ler nem escrever. Houve, porém, grandes diferenças regionais. Nos Estados de São Paulo, Santa Catarina e Rio Grande do Sul, menos de um terço das pessoas de 5 anos e mais não sabia ler nem escrever, enquanto que nos estados do Nordeste essa proporção ficou em torno dos 70%.

A PEA Agropecuária é um conceito que permite caracterizar o contingente populacional economicamente ativo que se dedica às atividades de agropecuária, extração vegetal e pesca. Na composição da PEA Agropecuária, segundo a posição na ocupação principal, destacou-se a importância relativa dos não assalariados. Entretanto, houve uma grande diferenciação regional, reflexo das diferenças nas características da agricultura em cada uma das regiões. Não obstante, os empregados constituíam a maioria da PEA Agrícola somente nos Estados de Minas Gerais (54,8%), Alagoas (54,4%), Rio de Janeiro (66,0%), São Paulo (68,6%), Mato Grosso do Sul (55,9%) e Distrito Federal (54,1%). Nos outros estados, o peso dos assalariados tem um significado muito diferente. Por exemplo, a importância relativa dos empregados foi semelhante em estados tão diferentes como Paraíba, Paraná e Mato Grosso. Portanto, como Indicador Social, só a composição da PEA Agropecuária segundo a posição na ocupação principal é insuficiente, devendo ser complementada com outros indicadores.

A porcentagem dos empregados agrícolas que receberam mais de 12 salários por ano (13º, 14º ou mais) foi pouco expressiva. Somente os Estados de Santa Catarina e Rio Grande do Sul atingiram uma proporção mais significativa, mas, ainda assim, em nenhum deles alcançou 40%.

Analisando o rendimento familiar *per capita*, pode-se constatar que uma alta proporção das famílias rurais ganhava menos de um quarto de salário mínimo (43%). Nos estados de agricultura mais atrasada, como os do Nordeste, essa proporção girou em torno de 60 a 70%.

As variáveis demográficas ESPE e MORT reforçam o quadro acima descrito, principalmente no que se refere às disparidades regionais. Nos estados menos desenvolvidos

do Nordeste, a esperança de vida ao nascer raramente atinge os 50 anos e a mortalidade infantil ultrapassa os 140 ‰, enquanto que nos estados mais desenvolvidos do Sul e Centro-Oeste a esperança de vida ao nascer é quase 20 anos maior e a mortalidade infantil decresce quase um terço.

Durante a década de 70 verificou-se no Brasil um notável êxodo rural. Um dos fatores assinalados como responsável por esse êxodo tem sido o rápido processo de modernização da agricultura a partir dos anos 60. De fato, o índice migratório mostra que o êxodo rural foi mais marcante nos estados onde foi mais intensa a modernização. Assim, ainda que o êxodo rural nos estados do Nordeste tenha sido grande, não atingiu os níveis dos outros estados do Sul e Centro-Oeste. Os estados do Norte apresentaram um índice migratório positivo devido ao intenso processo de ocupação da Amazônia ocorrido nessa década.

O quadro anteriormente descrito revela com poucas exceções a gravidade dos problemas sociais que existiam na área rural do País no início dos anos 80. Com um contingente populacional ainda elevado, precárias condições de vida no que se refere ao acesso a serviços básicos como abastecimento de água, saneamento e educação, mais os insuficientes níveis de renda familiar, configura-se como uma enorme dificuldade a superação da baixa qualidade de vida no campo.

4 REGIONALIZAÇÃO DA AGRICULTURA SEGUNDO AS CONDIÇÕES SOCIAIS

Para oferecer uma visão espacial sintética das condições de vida e bem-estar na agricultura aplicou-se o método dos componentes principais (análise fatorial) aos indicadores selecionados a partir do Censo Demográfico de 1980.

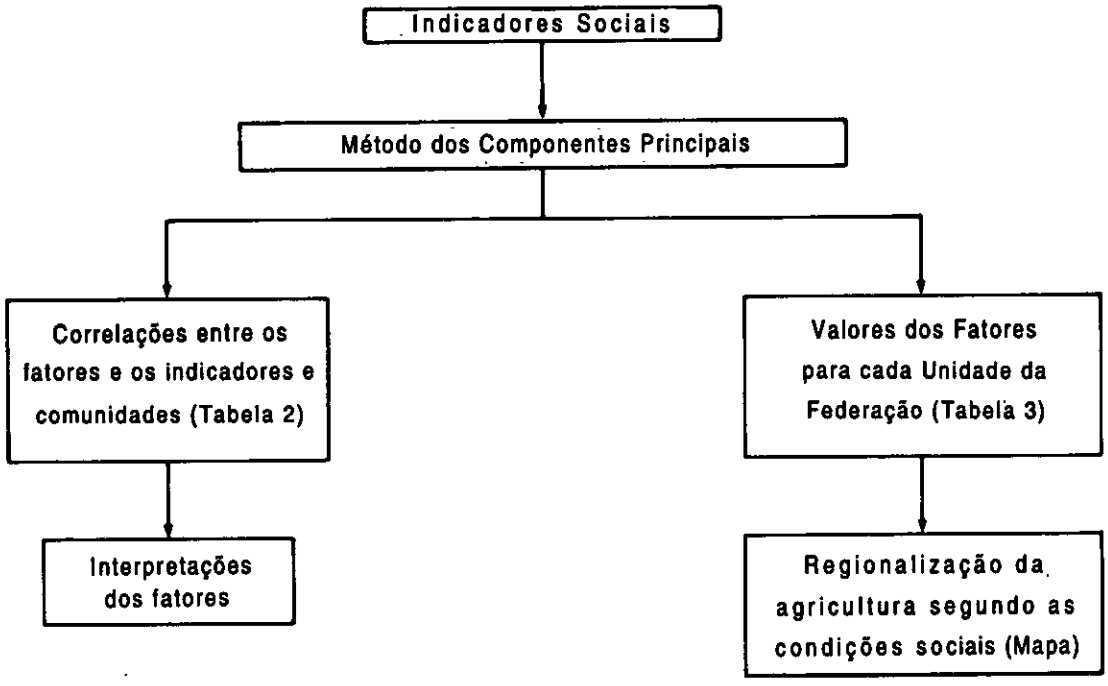
Em termos muito sumários, o método consiste em obter fatores (componentes) que são combinações lineares das variáveis originais, agrupando-se em cada fator as variáveis mais fortemente correlacionadas entre si e fazendo com que os fatores sejam independentes (ortogonais). Podem ser obtidos tantos fatores quanto o número de variáveis originais, mas geralmente poucos fatores (2, 3 ou 4) são suficientes para explicar uma alta proporção da variância total dos dados, de forma que a análise se restringe a esses primeiros fatores. Mediante algumas transformações matemáticas, as variáveis podem ser expressas em termos de combinações lineares dos fatores obtidos. Os coeficientes dessas combinações lineares nada mais são do que os coeficientes de correlação entre cada uma das variáveis e cada um dos fatores. A análise desses coeficientes permitiu, no caso deste estudo, sintetizar nos três fatores extraídos os diversos aspectos das condições de vida e bem-estar da população que mora no campo.

O valor de cada um dos fatores é calculado também para cada Unidade da Federação, permitindo assim uma regionalização do país em função dos aspectos que

cada fator representa.

O esquema a seguir resume o método utilizado:

Método dos Componentes Principais



A aplicação do método dos componentes principais aos 14 indicadores selecionados a partir do Censo Demográfico de 1980 permitiu obter três fatores (componentes principais) que explicam respectivamente, 61,59%, 16,64% e 6,72% da variância total dos dados, ou seja, explicam, em conjunto, cerca de 85% da variância total. Após rotação VARIMAX[2], calcularam-se os coeficientes de correlação entre cada indicador original e os três fatores obtidos, bem como as comunalidades (ver esquema). Os resultados foram os seguintes:

TABELA 2

Pesos dos Fatores após Rotação Ortogonal, Segundo as Variáveis

VARIÁVEL	FATOR 1	FATOR 2	FATOR 3	COMUNALIDADES (1)
GELA ...	0,936	0,259	0,072	0,949
SALA ...	0,893	0,327	-0,103	0,914
TEVE ...	0,873	0,209	0,337	0,920
DORU ...	-0,628	-0,332	-0,415	0,677
ANAF ...	-0,678	-0,633	-0,189	0,909
SANI	-0,755	-0,532	0,017	0,853
AGUA ...	-0,829	-0,407	-0,287	0,936
MORT ...	-0,258	-0,925	0,172	0,952
POBR ...	-0,577	-0,731	-0,031	0,869
JORN ...	0,408	0,761	0,279	0,824
ESPE	0,264	0,925	-0,170	0,955
EMPR ...	0,163	-0,110	0,889	0,829
RURA ...	-0,565	-0,278	-0,604	0,761
MIGR ...	0,009	0,079	-0,734	0,546

(1)A comunalidade mede a contribuição dos m primeiros fatores (no caso, $m = 3$) para a variância total de cada variável. Essas variâncias totais são normalizadas. Note-se que cada comunalidade é igual à soma dos quadrados dos pesos dos fatores ($0,949 = 0,936^2 + 0,259^2 + 0,072^2$, e assim por diante).

Foram destacados na tabela os valores superiores, em termos absolutos, a 0,6. Com isto, fica mais fácil identificar os três fatores.

O primeiro fator (F1) tem correlação positiva e alta com as variáveis GELA, SALA e TEVE, e correlação negativa e alta com DORU, ANAF, SANI e AGUA. Excetuando as variáveis SALA (empregados com 13º e 14º salário) e ANAF (analfabetismo), todas as outras dizem respeito às **condições do domicílio rural** (se é precário, se possui geladeira, televisão, água e instalação sanitária).

No segundo fator (F2) predominam as variáveis **demográficas** (mortalidade infantil e esperança de vida) juntamente com o **nível de pobreza** (com correlação negativa). Nota-se, ainda, alta correlação positiva desse fator com a variável JORN (PEA com jornada semanal acima de 48 horas). Em resumo, se o valor do segundo fator numa determinada região for positivo e alto, isto significa um menor nível e pobreza e/ou menores índices de mortalidade infantil. No entanto, este mesmo fator é positivamente afetado pela extensão da jornada de trabalho, que pode estar variando em sentido oposto ao das variáveis demográficas e pobreza. Por exemplo, na Região Sul o Fator 2 é positivo, mas é negativo em todos os estados do Nordeste. Isto indica que os níveis de pobreza e mortalidade infantil são mais elevados nesta última região, enquanto a esperança de vida e a jornada de trabalho tendem a ser menores, em comparação com os estados do Sul.

O último fator considerado (F3) capta basicamente os efeitos da **modernização do mercado de trabalho** na dinâmica populacional. Esse fator tem correlação positiva e alta com a proporção de empregados na PEA e com a urbanização (porque tem correlação negativa com a variável RURA, que é o complemento da taxa de urbanização), e correlação negativa e alta com o índice migratório (quanto "mais negativo"

este índice, maior a saída de pessoas da zona rural, ou seja, o Fator 3 está diretamente correlacionado com a **migração** a partir de áreas rurais).

O método dos componentes principais fornece também o valor de cada um dos fatores por Unidade da Federação. Com base na análise desses fatores é possível caracterizar regiões mais ou menos homogêneas em termos das condições sociais captadas pelos fatores.

Para o agrupamento dos estados utilizou-se o método de análise de *cluster* chamado *average linkage*, o qual, a partir das distâncias médias entre pares de observações, produz grupos com variâncias internas baixas e aproximadamente iguais entre grupos[3].

Na Tabela 3 apresentam-se os valores dos fatores por Unidade de Federação. O Gráfico e o Mapa permitem visualizar a distribuição dos estados segundo os valores dos três fatores[4].

TABELA 3

Valores dos Fatores, Segundo as Unidades da Federação

UNIDADES DA FEDERAÇÃO	FATOR 1	FATOR 2	FATOR 3
Rondônia	-0,67	0,8	-1,24
Acre	-0,67	0,41	-1,62
Amazônia	-0,58	0,55	-0,68
Roraima	-0,41	0,91	-1,31
Pará	0,29	0,08	-1,37
Amapá	1,66	-0,39	-1,79
Maranhão	-0,58	-0,82	-1,33
Piauí	-1,1	-0,55	-0,68
Ceará	-0,72	-1,29	0,49
Rio Grande do Norte	-0,32	-1,53	0,61
Paraiba	-0,41	-1,67	0,45
Pernambuco	0,44	-1,88	0,27
Alagoas	-0,27	-1,63	0,55
Sergipe	-0,65	-0,72	0,28
Bahia	-0,57	-0,42	-0,01
Minas Gerais	-0,73	0,86	1,28
Espírito Santos	-0,27	1,22	0,84
Rio de Janeiro	1,02	0,2	1,44
São Paulo	2,2	-0,16	1,19
Paraná	0,1	0,94	0,65
Santa Catarina	2,19	0,4	-0,74
Rio Grande do Sul	1,45	0,92	-0,35
Mato Grosso do Sul	-0,41	1,26	0,99
Mato Grosso	-0,86	1,1	0,03
Goiás	-1,3	1,11	1,18
Distrito Federal	1,18	0,3	0,89

FONTE – IBGE, Censo Demográfico de 1980.

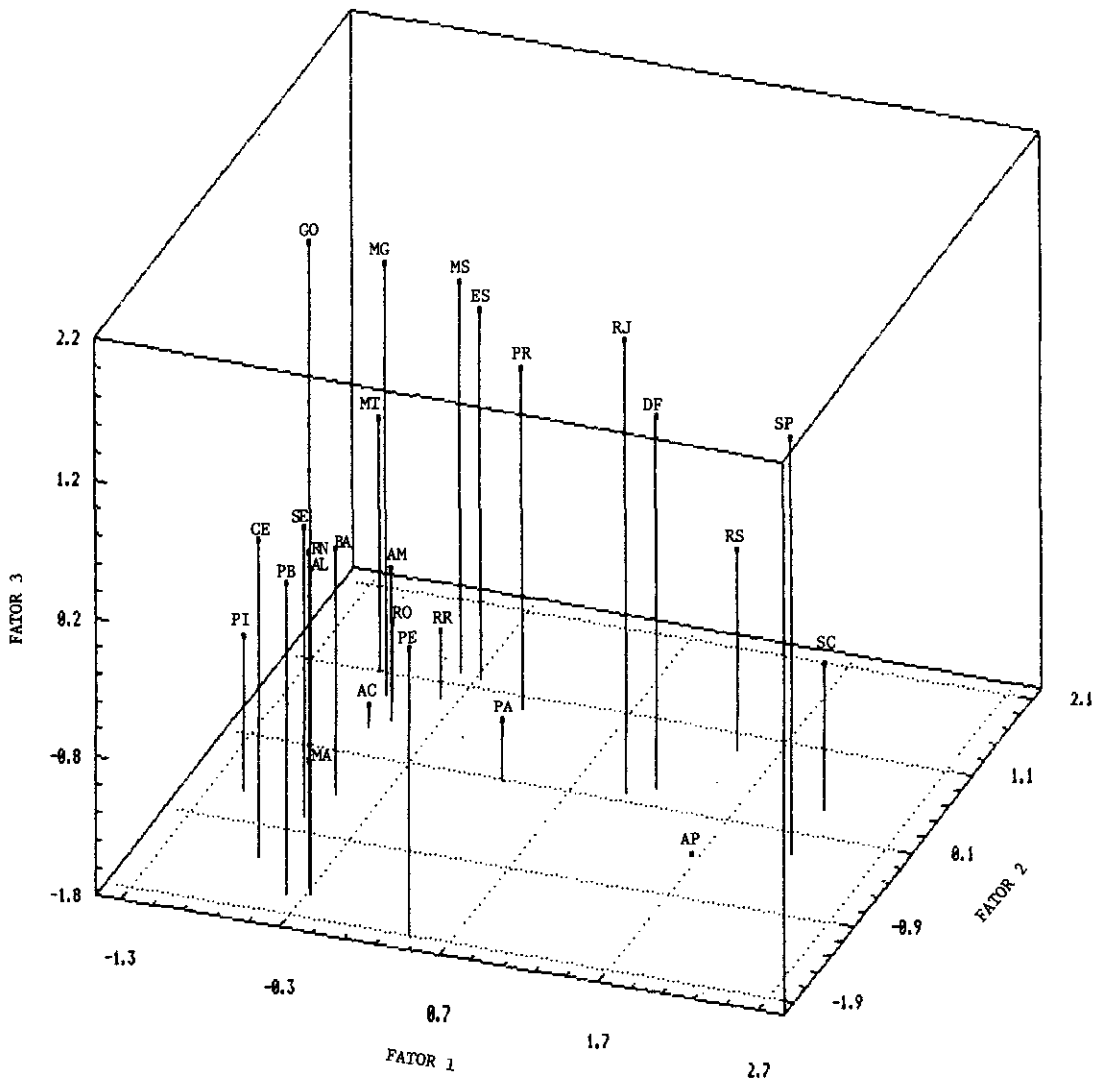
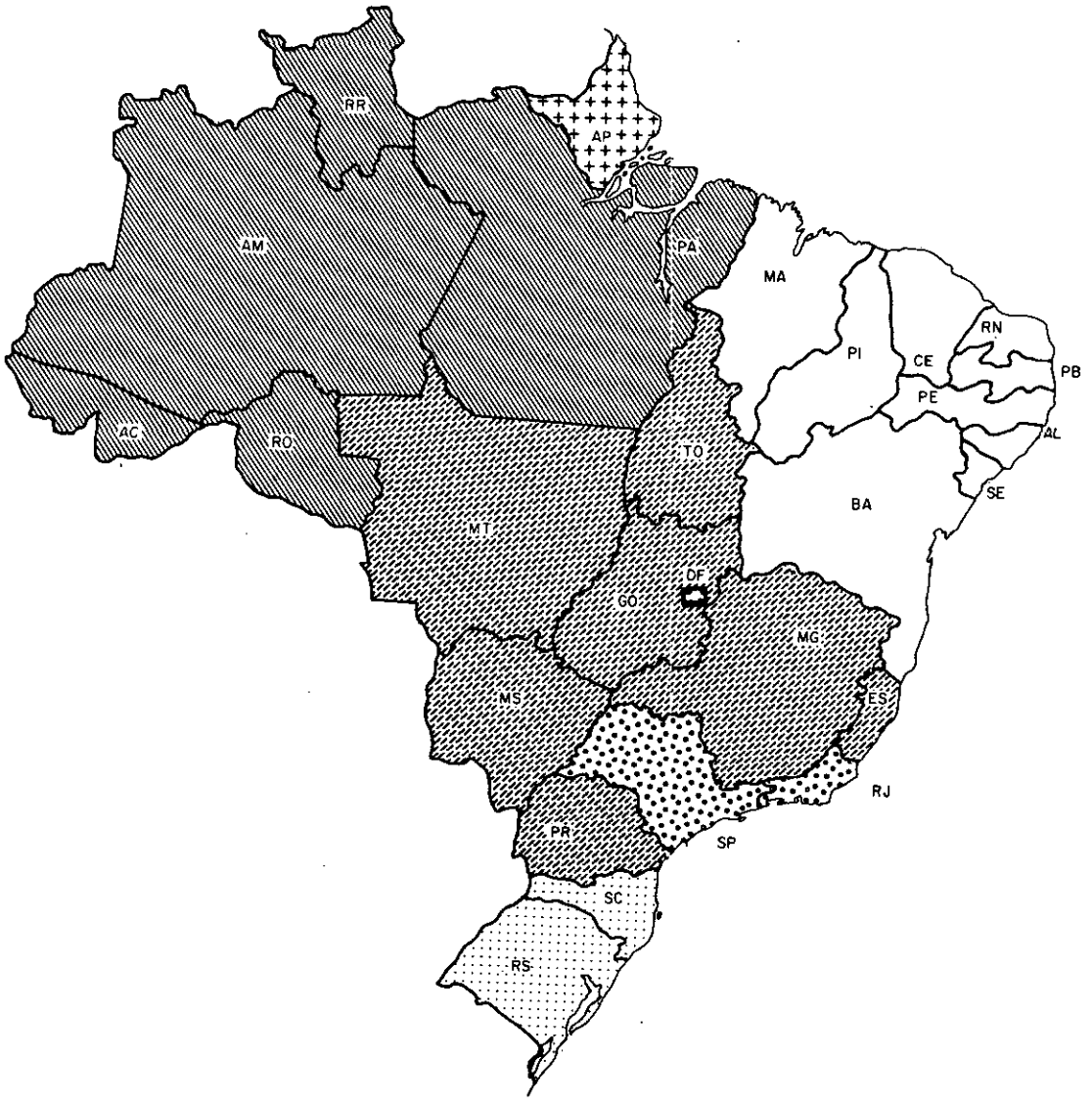


GRÁFICO - Distribuição das U.F. segundo os valores dos fatores (tabela 3).



	F1 (Domicílio)	F2 (Menor Pobreza)	F3 (Merc. Trab. e Migração)
	+	+	-
	+	+	+
	-	+	+
	-	-	+
	-	+	-
+++	Não classificado		

1) **Santa Catarina e Rio Grande do Sul** (valores positivos e altos para F1 e F2 e valores negativos para F3).

Nestes estados ocorrem as melhores condições de vida, em termos de qualidade dos domicílios, menor mortalidade infantil, menor pobreza, de todo o País. O Fator 3, refletindo a modernização do mercado de trabalho e os índices migratórios, assume valores menores do que 1, em termos absolutos, indicando uma posição intermediária desses estados em comparação com as outras regiões.

2) **São Paulo, Rio de Janeiro e DF** (valores positivos e altos para F1 e F3 e próximos de zero para F2).

Esses grupos de estados tem como marca característica os elevadíssimos valores de F3, denotando a intensidade da emigração rural e da urbanização, assim como a importância do assalariamento nas relações de produção predominantes.

A população agrícola nesses estados, notadamente em São Paulo, dispõe de melhores condições domiciliares (F1 alto), mas os aspectos captados pelo Fator 2 (indicadores demográficos, pobreza) já não se mostram tão favoráveis como nos estados da Região Sul.

3) **Rondônia, Acre, Amazonas, Roraima e Pará** (F3 negativo e alto, F1 negativo, F2 positivo).

Este grupo representa tipicamente a área da fronteira em ocupação ou recentemente ocupada, tendo em comum os maiores valores negativos de F3 do País. Isto quer dizer, resumidamente, saldo migratório positivo e baixa participação de empregados no total da PEA Agropecuária.

Mas há diferenças entre os estados do ponto de vista dos Fatores 1 e 2: no Pará as condições de vida mostram-se mais favoráveis em termos do Fator 1 (domicílios), mas em Rondônia, Acre, Amazonas e Roraima o Fator 2 (variáveis demográficas, pobreza) revela condições mais favoráveis.

4) **Minas Gerais, Goiás, Mato Grosso e Mato Grosso do Sul, Paraná e Espírito Santo** (F1 predominantemente negativo, F2 e F3 positivos).

Essa área parece combinar os efeitos demográficos e sobre o mercado de trabalho decorrentes da modernização agrícola (F2 e F3 positivos), com o atraso da infra-estrutura habitacional da população rural (F1 negativo e alto). Destaca-se o Estado de Goiás, que, ao lado do Paraná, apresentou o maior índice migratório negativo do País.

Os três estados localizados no segundo quadrante formado por F1 e F2 (MT, MS e ES) reúnem características de "menor atraso" do ponto de vista de F2, porém com valores negativos para F1, refletindo "maior atraso" deste ponto de vista. Assim, os estados deste subgrupo – que experimentaram um processo relativamente intenso e rápido de modernização agrícola – mostram baixo grau de pobreza (em termos de Brasil), porém associado a condições domiciliares precárias e baixo grau de urbanização.

5) **Maranhão, Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe e Bahia** (F1 negativo, F2 negativo e alto e F3 predominantemente positivo).

Os estados nordestinos podem ser agrupados sob a rubrica geral da pobreza. O Fator 2 é negativo (e geralmente alto) em todos eles, indicando maior pobreza, maior mortalidade infantil, menor esperança de vida. Além disso, em oito deles o Fator 1 também é negativo, revelando as precárias condições materiais e educacionais de existência da população rural. Apenas em Pernambuco estas últimas condições são um pouco melhores que no resto do Nordeste. Quanto ao Fator 3, este grupo localiza-se em situação intermediária (F3 perto de zero); como o Fator 3 combina os índices migratórios com a extensão do emprego assalariado, os menores valores da última variável são parcialmente compensados pela primeira, resultando para o terceiro fator valores relativamente baixos. O Maranhão destaca-se com valores negativos e altos para os três fatores, configurando as piores condições de vida para a população rural, nesse grupo.

5 CONCLUSÃO

O estudo das condições sociais em que vive a população ocupada na agricultura defronta-se com uma série de dificuldades, que podem ser assim resumidas:

a) O universo pesquisado nas diferentes fontes de dados é variável, predominando dois enfoques: o da ocupação das pessoas em atividades agropecuárias (população economicamente ativa, pessoas ocupadas de 10 anos ou mais, famílias cujo chefe tem atividade principal na agropecuária, etc.) e o do local de residência (domicílios situados na zona rural). Para o total do Brasil, cerca de 82% da PEA Agropecária residem em domicílios rurais, mas há estados, como São Paulo, onde essa proporção é bem menor (62%). É impossível, a partir das estatísticas publicadas, compatibilizar os dados dos dois universos, de forma que foram mantidos os dois tipos de indicadores.

b) A fonte de dados mais adequada para o estudo de Indicadores Sociais é o IBGE, tendo essa própria instituição publicado documentos especiais sobre o tema (em 1979 e 1984), cobrindo basicamente a década de 73 a 83. No entanto, há diversas dificuldades relacionadas com os dados, além da já citada no item anterior. Os Censos Demográficos, por exemplo, que contêm o mais abrangente conjunto de dados sobre o assunto, são decenais, não se prestando, obviamente, a estudos de acompanhamento de curto e médio prazos. As PNADs, por sua vez, prestam-se para acompanhamento anual mas padecem de problemas de comparabilidade interna. O próprio IBGE recomenda a utilização do período 1981 a 1986 como um conjunto específico, com uniformidade na coleta dos dados. Comparações com períodos anteriores podem ser problemáticas. Outra dificuldade das PNADs é que, apesar da riqueza e

variedade das informações coletadas, poucas são divulgadas usando o corte "rural" ou "agrícola". Além disso, a PNAD não coletou dados da zona rural da Região Norte, naquele período; em períodos anteriores não se dispõe de dados para a zona rural das Regiões Norte e Centro-Oeste.

c) Existem diversos problemas de ordem conceitual envolvidos em estudos de Indicadores Sociais, a começar pela própria delimitação do que sejam esses indicadores. Na verdade, a dificuldade decorre da falta de rigor do termo "condições sociais". Embora todos concordem com a idéia de que as "condições sociais" não podem ser isoladas das "condições econômicas", não é claro qual o limite entre as duas coisas. Por exemplo, a produtividade é uma medida de desempenho econômico (de um setor ou de um país), mas muito do bem-estar da população depende dela; o mesmo pode ser dito de muitas variáveis estruturais e demográficas. Mas há, por outro lado, um certo consenso de que é possível medir o nível de vida e o nível de bem-estar, ou, em termos mais gerais, as "condições sociais" de uma população, levando em conta certos indicadores de alimentação, saúde, habitação, educação e comparando-os com certas normas ou padrões mínimos considerados aceitáveis. Aceitas estas definições, o problema passa a ser a quantificação dos diversos componentes e a mensuração global dos níveis de bem-estar. Sobre este tema dispõe-se de abundante literatura, mas a ênfase deste trabalho recaiu sobre a identificação e quantificação dos indicadores disponíveis para a agricultura brasileira, sem chegar a propor índices agregados. Uma breve incursão nesse aspecto foi a aplicação de análise fatorial a um conjunto selecionado de indicadores, com vistas unicamente a obter uma regionalização da agricultura segundo as condições sociais, e não com o intuito de construir índices sintéticos.

Apesar dessas dificuldades, as estatísticas disponíveis no Brasil permitem obter, mesmo sem a utilização de métodos sofisticados, uma imagem razoavelmente fidedigna das condições de vida e de trabalho na agricultura.

Essa imagem revela a extrema precariedade em que vive a maior parte da população ocupada na agricultura ou com domicílio rural. Revela, também, a lentidão com que essa situação tem-se alterado na década de 80[5], fazendo crer que até o fim do século pouca coisa se modificará no que diz respeito às condições de trabalho e ao bem-estar dessas pessoas.

A análise dos indicadores mostrou que a modernização tecnológica da agricultura brasileira não foi acompanhada, ao menos nas mesmas proporções, por melhorias significativas da qualidade de vida no campo. A persistência de baixos níveis de renda e padrões inadequados de educação, moradia, saneamento e aplicação das leis trabalhistas revela a precariedade que caracteriza, de modo geral, a vida rural no Brasil.

Apesar desse quadro geral, encontram-se diferenças regionais. A velha dicotomia "sul-sudeste" *vs.* "periferia" não foi superada. A análise feita para 1980 mostrou que as condições de vida da população agrícola de Santa Catarina, Rio Grande do Sul, São Paulo, Rio de Janeiro e Paraná são nitidamente superiores às do resto do País. No

extremo oposto, ou seja, o das condições mais miseráveis de existência da população agrícola, encontram-se os estados da Região Nordeste, que combinam níveis críticos de condições domiciliares, educacionais, trabalhistas e de saúde, com altos índices migratórios, configurando assim o típico caso de "região-problema", tão explorado na literatura e na política sobre a região.

NOTAS

(1) – Entre os modelos de análise multivariada, os métodos de análise fatorial constituem um grupo bastante amplo, incluindo diversos métodos para descrição e inferência com variáveis multidimensionais, tais como o método dos fatores principais, dos componentes principais, correlação canônica etc. Utilizou-se o pacote computacional SAS, em computador IBM 3090, daUNICAMP.

(2) – Rotação VARIMAX é uma rotação ortogonal que permite que os coeficientes de correlação entre os indicadores e os fatores fiquem o mais próximo possível de zero ou de 1 ou de -1, facilitando assim a sua interpretação.

(3) – Foi utilizado o pacote computacional SAS, em computador IBM 3090, da UNICAMP.

(4) – O Amapá pode ser considerado uma espécie de *outlier*, não se agregando a nenhum dos cinco grupos formados.

(5) – Esta conclusão baseia-se em outro trabalho das autoras, usando dados da PNAD-1986, para o mesmo tipo de análise.

BIBLIOGRAFIA

CHATFIELD, C.; COLLINS, A.J. *Introduction to multivariate analysis.* Londres: Chapman e Hall, 1980.

CUADRAS, C.M. *Métodos de análises multivariante.* Barcelona: Editora Universitária de Barcelona S.A. (EUNI BAR), 1981.

FUENTES LLANILLO, Rafael. *Caracterização da estrutura de produção agropecuária do Estado do Paraná.* Piracicaba: DESR/ESALQ/USP. Dissertação de mestrado.

HARMAN, H. *Modern factor analysis.* Chicago: The University of Chicago, 1976.

KAGEYAMA, A. *Modernização, produtividade e emprego na agricultura: uma análise regional.* Campinas: Instituto de Economia - UNICAMP, 1986. Tese de doutoramento.

INCRA. *Sistema nacional de cadastro rural: zoneamento agrário 1ª fase.* Brasília, 1978. Informativo Técnico nº 4.

ROCHA, S.; VILLELA, R. Caracterização da subpopulação pobre metropolitana nos anos 80: resultados de uma análise multivariada. *Revista Brasileira de Economia*, Rio de Janeiro, v.44, n.1, p.35-52, jan./mar. 1990.

TRONCOSO LEONE, E. *Modernização e distribuição de renda na agricultura no Estado da Bahia em 1980.* Piracicaba: ESALQ/USP, 1988. Dissertação de mestrado.

RESUMO

Este artigo enfoca as condições sociais da População Economicamente Ativa na agricultura, como as características dos domicílios — fontes de água, saneamento, utensílios, etc. — o nível de educação, as condições de trabalho e o nível de renda.

Foi aplicado um método de análise fatorial (componentes principais) com o objetivo de regionalizar os estados segundo aquelas características, usando dados do Censo Demográfico de 1980. A análise revelou grandes diferenças entre as regiões, mas, de um ponto de vista geral, todas elas exibem um baixo nível de vida e bem-estar.

ABSTRACT

This paper focuses on the social conditions of the economically active population employed in the agricultural sector. The social conditions referred to are the characteristics of housing — including the sources of water, sanitary conditions, appliances, etc. —, the level of formal education, the labor conditions and the level of income.

A factor analysis method (principal factor) was applied in order to group similar states according to those characteristics, using data from the Population Census of 1980.

The analysis revealed the great differences among the regions, but, from a general point of view all of them exhibit a very low level of living conditions and welfare.

COTAS PARA O ÍNDICE DE GINI – UMA ABORDAGEM GEOMÉTRICA

José Paulo Q. Carneiro*
Jorge Rangel Costa**

1 INTRODUÇÃO

Uma dada distribuição de renda pode ser encarada como qualquer outra distribuição de freqüência. Como tal, podem-se aplicar a ela as medidas tradicionais de locação e de dispersão, tais como média, mediana, desvio padrão, diferença interquartílica, etc. No entanto, para distribuições de renda, tornou-se tradicional medir sua concentração através do chamado *Índice de Gini*, principalmente por causa de sua interpretação geométrica, relacionada à também tradicional *Curva de Lorenz*. Deve ser ressaltado que as aplicações do *Índice de Gini* não se limitam a distribuições de renda. Na realidade, este índice tem sido aplicado a estudos de distribuição espacial de população, concentração industrial, e diversos tipos de desigualdades sociais. Aqui, adotaremos o contexto das distribuições de renda, apenas para fixar a linguagem.

As fórmulas de cálculo do *Índice de Gini* referem-se, em geral, a uma distribuição "teórica", quase sempre do tipo contínuo. Na prática, no entanto, os dados disponíveis são em número finito (e provenientes de uma *amostra*, mas este aspecto não será tratado aqui), necessitando ser grupados em classes. Quando isto ocorre, o que se obtém, de fato, é uma *cota inferior* e uma *cota superior* para o índice. Em 1972, Gastwirth (ver bibliografia) estabeleceu fórmulas para essas cotas. No entanto, especialmente para a *cota superior*, a dedução analítica não é trivial.

*Professor da ENCE / IBGE.
**Analista do Centro de Informações e Dados do Rio de Janeiro – CIDE.

O objetivo deste trabalho é apresentar uma dedução essencialmente geométrica dessas cotas, tendo o cuidado de não omitir detalhes não encontrados nos compêndios (por serem considerados demasiadamente técnicos) nem nos artigos (por serem considerados "óbvios"). Para efeito de autoconsistência, a Seção 2 apresenta (no caso contínuo) os conceitos básicos sobre o *Índice de Gini*, a*Curva de Lorenz* e outras medidas correlatas, bem como os fundamentos de suas interpretações geométricas, base para a dedução das cotas, feita na Seção 3. Finalmente, a Seção 4 mostra que as cotas para o *Índice de Gini* de fato correspondem a situações-limite especiais. Os cálculos mais trabalhosos são deixados para os Apêndices.

2 O ÍNDICE DE GINI E A CURVA DE LORENZ

2.1 A Diferença Média e o Índice de Gini

Para uma função de densidade de probabilidade f, talvez as medidas de dispersão mais populares sejam o *desvio padrão* σ e o *desvio médio* δ_1, definidos por:

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx \tag{2.1}$$

$$\delta_1 = \int_{-\infty}^{+\infty} |x-\mu| f(x)dx \tag{2.2}$$

Nessas duas fórmulas, $\mu = \int_{-\infty}^{+\infty} x f(x)dx$ é a média da distribuição.

Procurando medidas que não dependam dos afastamentos em relação a um valor central, mas somente das diferenças entre os próprios valores observados, somos levados a introduzir os análogos:

$$E^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-y)^2 f(x)f(y)dydx \tag{2.3}$$

$$\Delta = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |x-y| f(x)f(y)dydx \tag{2.4}$$

A fórmula (2.3) não traz muito de novo em relação a (2.1), já que se verifica facilmente que $E^2 = 2\sigma^2$, utilizando as definições e o fato de que $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Já a medida de dispersão Δ, chamada *diferença média*, tem uma importância especial e entrará na própria definição do *Índice de Gini*. Para expressá-la sem utilizar valor absoluto, decompõe-se a integral "interior" em:

$$\int_{-\infty}^{x} (x-y)f(x)f(y)dy + \int_{x}^{+\infty} (y-x)f(x)f(y)dy$$

onde a segunda parcela, por sua vez, pode ser escrita como $\int_{-\infty}^{+\infty} - \int_{-\infty}^{x}$. Com isto, chega-se à expressão equivalente para Δ:

$$\Delta = 2\int_{-\infty}^{+\infty}\int_{-\infty}^{x}(x-y)f(y)f(x)dydx \tag{2.5}$$

Uma distribuição de renda tem, evidentemente, a peculiaridade de que sua função de densidade de probabilidade f é nula fora de um certo intervalo $I =]a;b[$, onde $0 \leq a < b$, sendo a e b, respectivamente, as rendas mínima e máxima; como, porém, quer-se deixar lugar para modelos um pouco mais gerais (Pareto, lognormal, etc.), vamos admitir também a possibilidade $b = +\infty$. Para os desenvolvimentos que seguem, vamos supor também que f é contínua em I (embora os resultados sejam facilmente generalizados para o caso em que f deixe de ser contínua apenas em um número finito de pontos, possuindo limites laterais finitos nesses pontos). Com esses pressupostos, a função de distribuição F associada a f, definida por $F(x) = \int_0^x f$ é tal que $F(a) = 0$, $F(b) = 1$ e $F'(x) = f(x)$. (Note-se que, como f é nula fora de I, tem-se $\int_0^x f = \int_a^x f$, e assim por diante). Além disso, a média μ será positiva, exceto se f for identicamente nula, o que aqui será sempre excluído.

Tudo isto permite deduzir duas outras expressões úteis para Δ. Em primeiro lugar, para x fixo, pondo $u = x - y$, $dv = f(y)dy$, e integrando por partes:

$$\int_0^x (x-y)f(y)dy = \int_0^x F(y)dy$$

Levando este resultado em (2.5), e trocando, em seguida, a ordem de integração, vem:

$$\frac{\Delta}{2} = \int_0^{+\infty}\int_0^x f(x)F(y)dydx = \int_0^{+\infty}\int_y^{+\infty} f(x)F(y)dxdy$$

$$= \int_0^{+\infty} F(y)\int_y^{+\infty} f(x)dxdy = \int_0^{+\infty} F(y)\,(1-F(y))\,dy$$

Portanto:

$$\Delta = 2\int_0^{+\infty} F(x)\,(1-F(x))\,dx \tag{2.6}$$

uma expressão que não utiliza integral dupla.

Fazendo agora, em (2.6), $u = F(x)(1-F(x))$, $dv = dx$, e integrando por partes, chega-se a:

$$\Delta = 4\int_0^{+\infty} x\left(F(x) - \frac{1}{2}\right) f(x)dx$$

Nesta expressão, como $F(x) \leq 1$, então $F(x) - 1/2 \leq 1/2$. Portanto:

$$\frac{\Delta}{4} \leq \frac{1}{2}\int_0^{+\infty} xf(x)dx = \frac{\mu}{2} \quad \Longrightarrow \quad 0 \leq \frac{\Delta}{2\mu} \leq 1$$

O número

$$G = \frac{\Delta}{2\mu} \tag{2.7}$$

é, por definição, o *Índice de Gini* da distribuição. Ao dividir Δ pelo dobro da média, o *Índice de Gini* relativiza a diferença média, tal como faz o coeficiente de variação com o desvio padrão. O coeficiente de variação, porém, não tem a propriedade agradável de estar entre 0 e 1.

Na verdade, o *Índice de Gini* pode, conforme a distribuição em questão, assumir todos os valores entre 0 e 1. Para efeito ilustrativo, mencionamos que é fácil verificar pela fórmula (2.6) que a distribuição de Pareto com parâmetros $a > 0$ e $p > 1$, definida por $f(x) = (p\, a^p)/(a\, x^{\,p+1})$ no intervalo $]a; +\infty[$, (com $a > 0$ e $p > 1$) tem *Índice de Gini* $1/(2p-1)$, o qual, dependendo do valor do parâmetro p, assume todos os valores em $]0; 1[$.

Observemos ainda que a distribuição de Pareto com $p = 5$ e a distribuição uniforme no intervalo $[a; 2a]$ têm o mesmo *Índice de Gini* (igual a $1/9$), embora sejam completamente diversas como distribuições de renda, como se pode ver em um gráfico da função de densidade de cada uma delas. Isto chama a atenção para o cuidado com que deve ser usado o *Índice de Gini* para caracterizar distribuições de renda.

2.2 A Curva de Lorenz e a Discrepância Máxima

Desde Lorenz (1905), tornou-se habitual, para distribuições de renda, construir uma curva através de pontos (p_i, q_i), onde p_i representa a proporção de indivíduos que possuem renda inferior a um certo valor r_i, enquanto q_i é a proporção da renda total possuída por esses indivíduos, sendo $r_1 < \cdots < r_n$. É a chamada *Curva de Lorenz*.

No caso contínuo, esse procedimento equivale ao seguinte: imagine-se que f seja a função de densidade de probabilidade associada a uma distribuição de renda, isto é, $\int_{x_1}^{x_2} f$ mede a probabilidade de que a renda de um indivíduo esteja entre x_1 e x_2. Neste caso, a proporção da renda total possuída pelas pessoas com renda $\leq x$ é dada por:

$$\Phi(x) = \frac{\int_0^x t f(t)dt}{\int_0^{+\infty} t f(t)dt}$$

Levando em conta que o denominador é a média μ da distribuição, isto é, a renda média geral da população, pode-se escrever:

$$\Phi(x) = \frac{1}{\mu} \int_0^x t f(t)dt$$

Como $F'(x) = f(x) > 0$ e $\Phi'(x) = xf(x)/\mu > 0$, as funções F e Φ são ambas estritamente crescentes, e portanto inversíveis, no intervalo $I =]a; b[$.

Considerando x em I como um parâmetro, a curva $v = L(u)$, dada pelas equações paramétricas

$$\begin{cases} u = F(x) \\ v = \Phi(x) \end{cases}$$

é a *Curva de Lorenz* associada à distribuição dada por f. Como $v = \Phi\left(F^{-1}(u)\right)$, vê-se que $L = \Phi \circ F^{-1}$, e como u e v variam entre 0 e 1, a curva se encontra dentro do quadrado unitário.

Examinemos as derivadas.

$$\frac{dv}{du} = \frac{dv/dx}{du/dx} = \frac{\Phi'(x)}{F'(x)} = \frac{xf(x)/\mu}{f(x)} = \frac{x}{\mu} > 0$$

já que $x > 0$. Logo, a *Curva de Lorenz* é estritamente crescente.

$$\frac{d^2v}{du^2} = \frac{d(\frac{dv}{du})/dx}{du/dx} = \frac{1/\mu}{f(x)} > 0$$

para $a < x < b$. Logo, a *Curva de Lorenz* é convexa e portanto o gráfico está sempre abaixo da secante $v = u$ e acima de suas tangentes.

Para cada $0 \leq u \leq 1$ a diferença $u - v = u - L(u) = g(u)$ mede a distância vertical entre a reta $v = u$ e a *Curva de Lorenz*. Como $g'(u) = 1 - x/\mu$, vê-se que $g(u)$ é máxima para $x = \mu$, e, neste ponto, $dv/du = 1$, isto é, a tangente à *Curva de Lorenz* é paralela à reta $v = u$. Isto significa que, quando se diz: "Os $u\%$ mais pobres possuem $v\%$ da renda total", a maior diferença $u - v$ ocorrerá quando se estiver referindo àqueles que ganham até a renda média da distribuição.

O número $\delta = F(\mu) - \Phi(\mu)$, valor máximo de $g(u)$, é chamado *discrepância máxima*.

A discrepância máxima é também uma medida de dispersão relativa, já que:

$$\delta = \frac{\delta_1}{2\mu} \tag{2.8}$$

onde δ_1 é o desvio médio da distribuição, definido em (2.2). Isto se demonstra imediatamente, decompondo a integral da definição (2.2) em:

$$\delta_1 = \int_0^{\mu} (\mu - x)f(x)dx + \int_{\mu}^{+\infty} (x - \mu)f(x)dx.$$

Em seguida, observa-se que:

$$\int_0^{\mu} (x - \mu)f(x)dx + \int_{\mu}^{+\infty} (x - \mu)f(x)dx = \int_0^{+\infty} (x - \mu)f(x)dx = 0$$

concluindo-se pois que $\delta_1 = 2\int_0^{\mu}(\mu - x)f(x)dx$.

Por outro lado, pelas definições de F e Φ,

$$\delta = F(\mu) - \Phi(\mu) = \frac{1}{\mu}\int_0^{\mu}(\mu - x)f(x)dx = \frac{\delta_1}{2\mu}.$$

Deve ser observado que a acima mencionada distribuição de Pareto com $p = 5$ tem o mesmo *Índice de Gini* e a mesma discrepância máxima que a distribuição (aliás não usual para renda) definida por $f(x) = 4x^3$ no intervalo $]0;1[$, embora sejam distribuições completamente diferentes (por exemplo, a primeira é decrescente, enquanto a segunda é crescente).

2.3 Interpretações Geométricas

A Figura 1 apresenta, no plano uv, uma *Curva de Lorenz* $v = L(u)$, com as propriedades geométricas deduzidas no item 2.2.

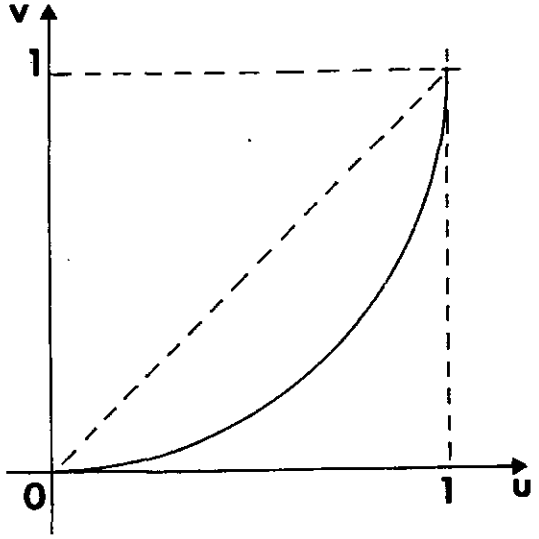


Figura 1

Curva de Lorenz

Sejam então:

S = área da região definida por: $0 \leq u \leq 1$; $L(u) \leq v \leq u$.

$A = \int_0^1 u dv$ = área da região definida por: $0 \leq v \leq 1$; $0 \leq u \leq L^{-1}(v)$.

$B = \int_0^1 v du$ = área da região definida por: $0 \leq u \leq 1$; $0 \leq v \leq L(u)$.

C = metade da área do quadrado $0 \leq u \leq 1$; $0 \leq v \leq 1$.

Tem-se claramente $S = A - C = C - B$, de onde se conclui que

$$2S = A - B = \int_0^1 udv - \int_0^1 vdu$$

Como $u = F(x)$ $v = \Phi(x)$ $du = f(x)dx$ $dv = (xf(x)/\mu)dx$, então:

$$\begin{aligned} 2S &= \frac{1}{\mu}\int_0^{+\infty} (xF(x) - \mu\Phi(x))\, f(x)dx \\ &= \frac{1}{\mu}\int_0^{+\infty} \left(x\int_0^x f(y)dy - \int_0^x yf(y)\right) f(x)dx \\ &= \frac{1}{\mu}\int_0^{+\infty}\int_0^x (x-y)f(y)dyf(x)dx = \frac{\Delta}{2\mu} = G \end{aligned}$$

pelas fórmulas (2.5) e (2.7).

Portanto, o *Índice de Gini* tem como interpretação geométrica o dobro da área da região compreendida entre a *Curva de Lorenz* e a bissetriz $v = u$. Esta interpretação, muito usada no estudo de distribuições de renda, permite verificar graficamente o fato, já demonstrado analiticamente, de que o *Índice de Gini* está entre 0 e 1, já que a área S não pode ultrapassar a metade da área do quadrado, isto é, 1/2.

Na Figura 2, além da *Curva de Lorenz*, está também representada a discrepância máxima δ, pelo segmento PQ. Pela convexidade da *Curva de Lorenz*, é claro que a área S acima referida está entre a área do triângulo OPR e a área do paralelogramo $ORVT$. Esta última é igual a $PQ \times 1 = \delta$ e também igual ao dobro da área do triângulo. Segue-se que $\frac{\delta}{2} \leq S \leq \delta$ e portanto que:

$$\delta \leq G \leq 2\delta$$

uma desigualdade útil, quando for fácil calcular a discrepância máxima. Além disso, esta fórmula, utilizando (2.7) e (2.8), dá também uma relação interessante entre a diferença média e o desvio médio, a saber:

$$\delta_1 \leq \Delta \leq 2\delta_1$$

É claro que estas fórmulas podem ser deduzidas analiticamente, mas é notável a facilidade com que se as deduz geometricamente.

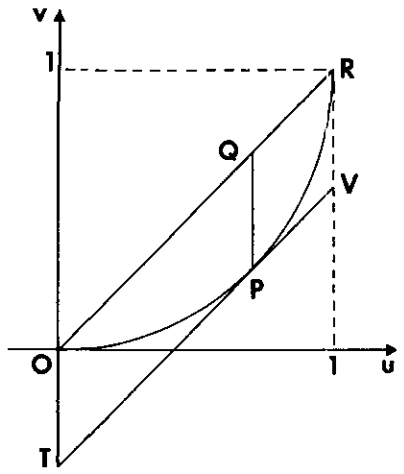


Figura 2

Discrepância Máxima

3 COTAS PARA O ÍNDICE DE GINI COM DADOS GRUPADOS

Quando se considera, na prática, uma distribuição de renda, os dados aparecem grupados em classes. O grupamento pode ser um "dado", por exemplo, numa tabela de uma publicação, ou pode ser arbitrado pelo pesquisador, se ele estiver confeccionando uma tabulação a partir de dados individuais. Aqui vai ser considerado o primeiro caso. De qualquer modo, o fato de se gruparem os dados em classes evidentemente mascara o *Índice de Gini*, que é uma medida de dispersão. O que vai ser mostrado agora é que: se a renda média (ou, equivalentemente, a renda total) de cada classe for conhecida, e se se supuser que os dados provêm de uma distribuição não especificada, porém de forma que a *Curva de Lorenz* tenha as propriedades apresentadas na Seção 2, particularmente a convexidade, então é possível determinar uma *cota inferior* e uma *cota superior* para o *Índice de Gini*, a partir dos dados empíricos. Os exemplos do final da seção darão uma idéia de quão estreitas são essas cotas. As fórmulas podem ser deduzidas de modo puramente analítico, como em Gastwirth 1972 (ver bibliografia), mas aqui serão obtidas geometricamente. Por enquanto, vamos supor que a renda máxima b seja finita.

Sejam então:

$a = r_0 < r_1 < \cdots < r_n < r_{n+1} = b$ os limites das classes de renda;

p_i = número de pessoas com renda em $[r_{i-1}; r_i[$, para $i = 1, \ldots, n+1$

$P = \sum_{i=1}^{n+1} p_i$ = total de pessoas

$u_i = \frac{\sum_{j=1}^{i} p_j}{P}$ = proporção de pessoas com renda $< r_i$, para $i = 1, \ldots, n+1$

$u_0 = 0$

μ_i = renda média das pessoas com renda em $[r_{i-1}; r_i[$, para $i = 1, \ldots, n+1$

$p_i\mu_i$ = renda da classe $[r_{i-1}; r_i[$, para $i = 1, \ldots, n+1$

$R = \sum_{i=1}^{n+1} p_i\mu_i$ = renda total

$v_i = \frac{1}{R}\sum_{j=1}^{i} p_j\mu_j$ = proporção da renda total possuída pelas pessoas com renda $< r_i$, para $i = 1, \ldots, n+1$.

A Figura 3 mostra os pontos provenientes da distribuição empírica e a *Curva de Lorenz* associada. Pelo item 2.3 e pela convexidade da *Curva de Lorenz*, uma *cota inferior* G_I para o *Índice de Gini* é dada pelo dobro da área da figura formada pela reta $v = u$ e pela poligonal que une os pontos $(u_i; v_i)$. Logo:

$$G_I = 1 - \sum_{i=1}^{n+1}(u_i - u_{i-1})(v_i + v_{i-1}) \tag{3.1}$$

Levando em conta o significado de u_i e v_i, a fórmula também pode ser escrita:

$$G_I = 1 - \frac{1}{PR}\sum_{i=1}^{n+1}\left(p_i^2\mu_i + 2\sum_{j=1}^{i-1} p_i p_j \mu_j\right) \tag{3.2}$$

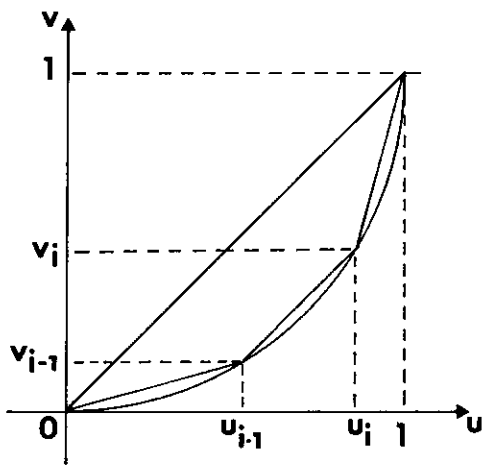


Figura 3

Cota Inferior para o Índice de Gini

Ainda pela convexidade, uma *cota superior* para o *Índice de Gini* será dada pelo dobro da área da figura formada pela reta $v = u$ e pelas tangentes à *Curva de Lorenz* nos pontos $(u_i; v_i)$. Na realidade, essa *cota superior* G_S pode ser escrita como $G_S = G_I + D$, onde D é o dobro da soma das áreas dos triângulos indicados na Figura 4, formados pelos pontos $(u_{i-1}; v_{i-1})$, $(u_i; v_i)$ e $(\bar{u}_i; \bar{v}_i)$, onde $(\bar{u}_i; \bar{v}_i)$ é o ponto de interseção das tangentes à *Curva de Lorenz* nos pontos $(u_{i-1}; v_{i-1})$ e $(u_i; v_i)$.

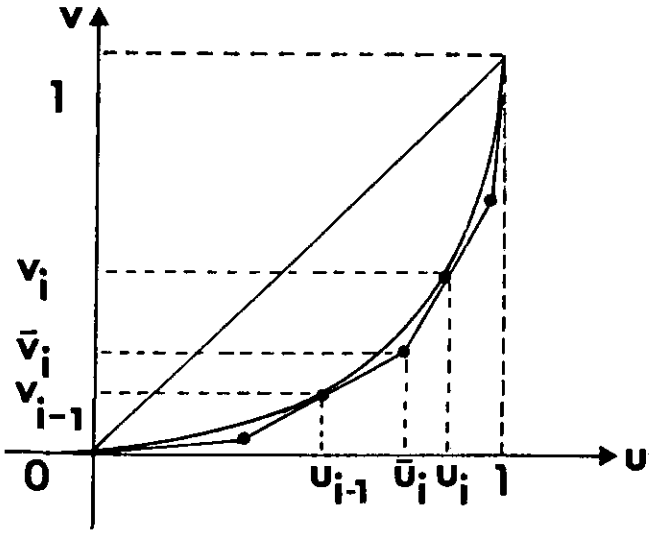


Figura 4

Cota Superior para o Índice de Gini

A inclinação da tangente à *Curva de Lorenz* em $(u_i; v_i)$ é dada por dv/du, para $u = u_i$. Porém, pelo item 2.2 , tem-se que $dv/du = x/\mu = F^{-1}(u)/\mu$, onde $\mu = \sum_{i=1}^{n+1} p_i\mu_i$ é a média da distribuição, isto é, a renda média total, e F acumula as pessoas. Logo, a inclinação procurada é $F^{-1}(u_i)/\mu = r_i/\mu$.

As duas tangentes consecutivas têm, portanto, equações:

$$v - v_{i-1} = \frac{r_{i-1}}{\mu}(u - u_{i-1})$$

$$v - v_i = \frac{r_i}{\mu}(u - u_i)$$

Resolvendo o sistema e levando em conta as expressões de u_i e v_i, obtém-se para o ponto de interseção:

$$\bar{u}_i = (1 - \alpha_i)u_{i-1} + \alpha_i u_i$$

$$\bar{v}_i = (1 - \beta_i)v_{i-1} + \beta_i v_i$$

onde $\alpha_i = \frac{r_i - \mu_i}{r_i - r_{i-1}}$ e $\beta_i = \frac{r_{i-1}}{\mu_i}\alpha_i$.

Finalmente, o dobro da área S_i do triângulo em questão pode ser calculado pela conhecida fórmula do determinante, isto é:

$$2S_i = (\bar{u}_i - u_{i-1})(v_i - v_{i-1}) - (u_i - u_{i-1})(\bar{v}_i - v_{i-1}) = (\alpha_i - \beta_i)(u_i - u_{i-1})(v_i - v_{i-1})$$

Ou seja:

$$2S_i = \left(\frac{p_i}{P}\right)^2 \frac{(\mu_i - r_{i-1})(r_i - \mu_i)}{\mu(r_i - r_{i-1})}$$

Portanto:

$$D = \frac{1}{\mu} \sum_{i=1}^{n+1} \left(\frac{p_i}{P}\right)^2 \frac{(\mu_i - r_{i-1})(r_i - \mu_i)}{r_i - r_{i-1}} \tag{3.3}$$

A conclusão é que o *Índice de Gini* G se situa entre os valores:

$$G_I \leq G \leq G_I + D = G_S \tag{3.4}$$

dados pelas fórmulas (3.2) e (3.3).

Deve ser observado que a expressão obtida para a *cota superior* depende dos r_i e, em particular, de $r_{n+1} = b$, a renda máxima. Se se admitir que os dados provenham de um modelo sem renda máxima, isto é, com $b = +\infty$, naturalmente a última parcela em (3.3) será substituída pelo seu limite, a saber:

$$\frac{1}{\mu}\left(\frac{p_{n+1}}{P}\right)^2 (\mu_{n+1} - r_n) \tag{3.5}$$

Por outro lado, é muito comum, na prática, que os dados da última classe apareçam na forma: "p_{n+1} = número de pessoas com renda $\geq r_n$", o que equivale a dizer que $r_{n+1} = b$ é desconhecido. Neste caso, para obter a maior *cota superior* possível, deve-se também, por segurança, tomar $b = +\infty$, e, conseqüentemente, utilizar a fórmula (3.5).

Apenas para que se tenha uma idéia numérica da distância, na prática, entre as cotas inferior e superior para o *Índice de Gini* (e, portanto, da própria precisão do índice, do modo como é usualmente calculado), a tabela apresenta alguns exemplos numéricos, calculados a partir dos dados de rendimento das pessoas com rendimento, investigados pelo IBGE na Pesquisa Nacional por Amostra de Domicílios — PNAD —, para o Brasil e o Estado do Rio de Janeiro, nos anos de 1984, 1985 e 1986. As classes são as fixadas pela própria publicação do IBGE e os rendimentos médios de cada classe são dados da pesquisa. Para efeito de cálculo da *cota superior* do *Índice de Gini*, a renda máxima foi considerada infinita.

TABELA

Cotas Inferior e Superior para o Índice de Gini
no Brasil e Estado do Rio de Janeiro – 1984 – 86

Cotas	Brasil-84	Brasil-85	Brasil-86	RJ-84	RJ-85	RJ-86
Inferior G_I....	0,5756	0,5877	0,5776	0,5608	0,5818	0,5737
Superior G_S...	0,5925	0,6047	0,5949	0,5773	0,5983	0,5905

FONTE – IBGE, Pesquisa Nacional por Amostra de Domicílios – PNAD.
NOTA – Fonte para os dados primários.

4 INTERPRETAÇÃO DAS COTAS PARA O ÍNDICE DE GINI

Freqüentemente a fórmula (3.1) é apresentada como sendo "a fórmula" do *Índice de Gini* para dados grupados. No entanto, como ficou claro na Seção 3, ela fornece apenas uma *cota inferior* para o índice.

Cabe aqui a questão: fixadas, como no início da Seção 3, as classes de renda definidas pelos r_i, os números de pessoas p_i e as rendas médias μ_i de cada classe, existe uma situação correspondente a G_I?

Como o *Índice de Gini* é uma medida de dispersão, é natural que ele seja *mínimo* na situação de maior uniformidade, isto é, quando as p_i pessoas com renda entre r_{i-1} e r_i tiverem todas renda igual a μ_i. Vejamos que isto de fato ocorre, em um certo sentido. Para tanto, vamos ter que nos afastar do caso contínuo.

Consideremos então a distribuição discreta correspondente ao caso em que, para cada $i = 1, \cdots, n+1$, haja p_i pessoas com renda μ_i. As fórmulas (2.4) e (2.5) são, na realidade, casos particulares (no caso contínuo) das fórmulas mais gerais:

$$\Delta = \int_{-\infty}^{+\infty} |x - y| dF(y) dF(x) \tag{4.1}$$

$$\Delta = 2 \int_{-\infty}^{x} |x - y| dF(y) dF(x) \tag{4.2}$$

onde as integrais devem ser entendidas como integrais de Riemann-Stieltjes.

Traduzindo a fórmula (4.2) para o caso discreto em questão, levando em conta a definição (2.7), observando que $\mu_{i-1} < \mu_i$, e que $\mu = R/P$, obtém-se, para o *Índice de Gini* G_D:

$$G_D = \frac{1}{PR} \sum_{i=1}^{n+1} \sum_{j<i} p_i p_j (\mu_i - \mu_j) \tag{4.3}$$

Comparando este resultado com (3.10), conclui-se (ver Apêndice 1) que $G_D = G_I$.

Ou seja: dada uma distribuição do tipo contínuo, da qual se conhecem os dados apenas em um grupamento em classes fixado, isto é, não se conhece a sua distribuição

dentro das classes, então a fórmula (3.2) fornece o menor valor possível para o *Índice de Gini*, e este valor equivale à hipótese de que dentro das classes todas as pessoas têm a mesma renda.

Imaginemos agora a situação em que as p_i pessoas com renda no intervalo $[r_{i-1}; r_i[$ estejam distribuídas de forma que $\beta_i p_i$ pessoas tenham renda r_{i-1}, enquanto $(1-\beta_i)p_i$ tenham renda $r_i - \epsilon$, onde $0 < \epsilon < r_i - r_{i-1}$ e $0 < \beta_i < 1$. Isto é, um conjunto de pessoas concentra a renda r_{i-1} e as restantes concentram a renda $r_i - \epsilon$, "próxima" de r_i.

O que é possível mostrar, neste caso, é que, quando $\epsilon \to 0$, os números β_i tendem a $\frac{r_i - \mu_i}{r_i - r_{i-1}}$, enquanto o *Índice de Gini* G_ϵ correspondente a essa situação, tende à *cota superior* G_S (ver Apêndice 2).

Resumindo (em termos informais): a *cota superior* G_S para o *Índice de Gini* corresponde à situação-limite em que, em cada classe de renda, as rendas das pessoas se concentram em dois grupos: um, com $(\frac{r_i - \mu_i}{r_i - r_{i-1}})p_i$ pessoas com a renda mínima da classe r_{i-1}, e o outro, com $(\frac{\mu_i - r_{i-1}}{r_i - r_{i-1}})p_i$ pessoas com uma renda "arbitrariamente próxima" da renda mínima da classe seguinte r_i. Naturalmente, quando se considera $r_{n+1} = b = +\infty$ (ver observação ao final da Seção 3), todas as p_{n+1} pessoas da última classe se concentrariam, no limite, em uma única classe com renda r_n.

5 APÊNDICES

Apêndice 1

$G_I = G_D$

$$\begin{aligned} PR &= \sum_{i=1}^{n+1} p_i \sum_{j=1}^{n+1} p_j \mu_j \\ &= \sum_{i=1}^{n+1} p_i^2 \mu_i + \sum_{i=1}^{n+1} \sum_{j<i} p_i p_j (\mu_i + \mu_j) \\ &= \sum_{i=1}^{n+1} p_i^2 \mu_i + \sum_{i=1}^{n+1} \sum_{j<i} 2 p_i p_j \mu_j + \sum_{i=1}^{n+1} \sum_{j<i} p_i p_j (\mu_i - \mu_j) \\ &= PR(1 - G_I) + PRG_D \end{aligned}$$

a última passagem vindo de (3.2) e (4.3). Daí se conclui que $G_I = G_D$.

Apêndice 2

$\lim_{\epsilon \to 0} G_\epsilon = G_S$

Para que a renda média das pessoas que concentram a renda $r_i - \epsilon$ permaneça igual a μ_i, é necessário que $\beta_i p_i r_{i-1} + (1 - \beta_i) p_i (r_i - \epsilon) = \mu_i$. Daí se deduz que:

$$\beta_i = \frac{r_i - \mu_i - \epsilon}{r_i - r_{i-1}}$$

Vamos mostrar que o *Índice de Gini* G_ϵ, correspondente a essa situação, tende à *cota superior* G_S, quando $\epsilon \to 0$. Note-se que $\lim_{\epsilon \to 0} \beta_i = \alpha_i = (r_i - \mu_i)/(r_i - r_{i-1})$ e, portanto:

$$\alpha_i r_{i-1} + (1 - \alpha_i) r_i = \mu_i \tag{5.1}$$

Para calcular, por uma fórmula análoga à (4.3), o *Índice de Gini* G_ϵ correspondente ao caso descrito, é preciso observar que há agora $2n + 2$ valores de renda s_k, com $k = 0, \cdots, 2n + 1$, de tal forma que, para $j = 0, \cdots, n$:

$$s_{2j} = r_j$$
$$s_{2j+1} = r_{j+1} - \epsilon$$

As freqüências absolutas associadas a esses valores de renda são q_k, de modo que, para $j = 0, \cdots, n$:

$$q_{2j} = \beta_{j+1} p_{j+1}$$
$$q_{2j+1} = (1 - \beta_{j+1}) p_{j+1}$$

É de notar também que, quando $\epsilon \to 0$:

$$s_{2j+1} \to r_{j+1} \qquad q_{2j} \to \alpha_{j+1} p_{j+1} \qquad q_{2j+1} \to (1 - \alpha_{j+1}) p_{j+1} \tag{5.2}$$

Além disso:

$$\sum_{k=0}^{2n+1} q_k = \sum_{j=0}^{n} (q_{2j} + q_{2j+1}) = \sum_{j=0}^{n} p_{j+1} = P$$

$$\sum_{k=0}^{2n+1} q_k s_k = \sum_{j=0}^{n} (q_{2j} s_{2j} + q_{2j+1} s_{2j+1}) =$$

$$= \sum_{j=0}^{n} p_{j+1} (\beta_{j+1} r_j + (1 - \beta_{j+1})(r_{j+1} - \epsilon)) =$$

$$= \sum_{j=0}^{n} p_{j+1} \mu_{j+1} = R$$

Logo:

$$G_\epsilon = \frac{1}{PR} \sum_{k=0}^{2n+1} \sum_{l<k} q_l q_k (s_k - s_l) \tag{5.3}$$

Examinemos agora $\lim_{\epsilon \to 0} G_\epsilon$. Na fórmula (5.3), para um valor fixado de $k = 2j$, com $j = 0, \cdots, n$, o somatório interno fica:

$$(q_0 q_k(s_k - s_0) + q_1 q_k(s_k - s_1)) + \cdots + (q_{k-2} q_k(s_k - s_{k-2}) + q_{k-1} q_k(s_k - s_{k-1}))$$
$$= \sum_{i=1}^{j} q_{2j} \left(q_{2i-2}(s_{2j} - s_{2i-2}) + q_{2i-1}(s_{2j} - s_{2i-1})\right)$$

Utilizando (5.2) e (5.1), verifica-se que, quando $\epsilon \to 0$, esta soma tende a:

$$\sum_{i=1}^{j} \alpha_{j+1} p_{j+1} \left(\alpha_i p_i(r_j - r_{i-1}) + (1 - \alpha_i) p_i(r_j - r_i)\right)$$
$$= \sum_{i=1}^{j} \alpha_{j+1} p_{j+1} p_i(r_j - \mu_i)$$

Portanto, essas parcelas contribuem para $\lim_{\epsilon \to 0} G_\epsilon$, com o valor:

$$\frac{1}{PR} \sum_{j=0}^{n} \sum_{i=1}^{j} \alpha_{j+1} p_{j+1} p_i(r_j - \mu_i) \tag{5.4}$$

Por outro lado, para um valor ímpar de $k = 2j+1$, com $j = 0, \cdots, n$, o somatório interno em (5.3) fica:

$$(q_0 q_k(s_k - s_0) + q_1 q_k(s_k - s_1)) + \cdots + (q_{k-3} q_k(s_k - s_{k-3}) + q_{k-2} q_k(s_k - s_{k-2})) +$$
$$+ q_{k-1} q_k(s_k - s_{k-1}) =$$
$$\sum_{i=1}^{j} q_{2j+1} \left(q_{2i-2}(s_{2j+1} - s_{2i-2}) + q_{2i-1}(s_{2j+1} - s_{2i-1})\right) + q_{2j} q_{2j+1}(s_{2j+1} - s_{2j}) \tag{5.5}$$

Quando $\epsilon \to 0$, a última parcela de (5.5), a saber, $q_{2j} q_{2j+1}(s_{2j+1} - s_{2j})$, tende a:

$$\alpha_{j+1} p_{j+1}(1 - \alpha_{j+1}) p_{j+1}(r_{j+1} - r_j) = \alpha_{j+1}(1 - \alpha_{j+1}) p_{j+1}^2(r_{j+1} - r_j)$$

Então, estas parcelas contribuem para $\lim_{\epsilon \to 0} G_\epsilon$, com o valor:

$$\frac{1}{PR} \sum_{j=0}^{n} \alpha_{j+1}(1 - \alpha_{j+1}) p_{j+1}^2(r_{j+1} - r_j) \tag{5.6}$$

Finalmente, ainda por (5.1) e (5.2), as outras parcelas de (5.5) tendem a:

$$\sum_{i=1}^{j}(1-\alpha_{j+1})p_{j+1}\left(\alpha_i p_i(r_{j+1}-r_{i-1})+(1-\alpha_i)p_i(r_{j+1}-r_i)\right)$$

$$=\sum_{i=1}^{j}(1-\alpha_{j+1})p_{j+1}p_i(r_{j+1}-\mu_i)$$

Portanto essas parcelas contribuem para $\lim_{\epsilon\to 0} G_\epsilon$, com o valor:

$$\frac{1}{PR}\sum_{j=0}^{n}\sum_{i=1}^{j}(1-\alpha_{j+1})p_{j+1}p_i(r_{j+1}-\mu_i) \tag{5.7}$$

Em suma: $\lim_{\epsilon\to 0} G_\epsilon = (5.4)+(5.6)+(5.7)$.

Porém, associando (5.4) com (5.7) e utilizando (5.1), obtém-se:

$$\frac{1}{PR}\sum_{j=0}^{n}\sum_{i=1}^{j}p_{j+1}p_i\left(\alpha_{j+1}(r_j-\mu_i)+(1-\alpha_{j+1})(r_{j+1}-\mu_i)\right)$$

$$=\frac{1}{PR}\sum_{j=0}^{n}\sum_{i=1}^{j}p_i p_{j+1}(\mu_{j+1}-\mu_i)$$

$$=\frac{1}{PR}\sum_{j=1}^{n+1}\sum_{i<j}p_i p_j(\mu_j-\mu_i)=G_I$$

em virtude de (4.3).

Por outro lado, utilizando (5.1) e o fato de que $P\mu = R$, verifica-se que a fórmula (5.7) é idêntica à (3.3). A conclusão é que:

$$\lim_{\epsilon\to 0} G_\epsilon = G_I + D = G_S$$

BIBLIOGRAFIA

GASTWIRTH, J.L. The estimation of the Lorenz curve and Gini index. *The Review Economics and Statiscs*, v.54, p.306–316, 1972.

KENDALL, M.G.; STUART, A. *The advanced theory of statistics*. London: Charles Griffen and Company, 1963.

YNTEMA, D. Measures of inequality in the personal distribution of wealth and income. *Journal of the American Statistical Association*, v.28, p.423–433, 1933.

RESUMO

Este artigo apresenta uma dedução geométrica das cotas inferior e superior para o *Índice de Gini*, quando os dados de distribuição se encontram grupados, e mostra que estas cotas de fato correspondem a certas situações-limite. Para efeito de autoconsistência, são também focalizados os conceitos típicos sobre o *Índice de Gini*, a *Curva de Lorenz* e outras medidas correlatas.

ABSTRACT

This paper presents a geometric deduction of the lower and upper bounds for the Gini index, in the case where the distribution data are found to be grouped, and shows that these bounds actually correspond to cartain situations. For completeness, the basic concepts about Gini index, Lorenz curve and other related measures, are also focused.

ASPECTOS METODOLÓGICOS ASSOCIADOS À CONSTRUÇÃO DO ÍNDICE DE GINI – ILUSTRAÇÃO A PARTIR DO CASO BRASILEIRO

Sonia Rocha*

1 INTRODUÇÃO

Reconhecidamente, os problemas de desenvolvimento no Brasil estão mais ligados à questão distributiva do que ao nível de renda em si. Embora a renda média tenha evoluído satisfatoriamente até 1980 — +36% entre 1960-70 e +48% entre 1970-80 — as evidências são de agravamento da distribuição de renda durante todo o período, embora a taxas decrescentes: o índice de Gini passa de 0,50 em 1960 para 0,57 em 1970 e 0,59 em 1980[1]. Embora, de início, o crescente nível de desigualdade fosse visto como um fenômeno inevitável ligado à própria dinâmica do desenvolvimento econômico à *la Kuznets*, ou, mais especificamente, a desequilíbrios no mercado de trabalho devido à escassez relativa de mão-de-obra qualificada, como propunha Langoni, a persistência do fenômeno e as evidências empíricas mais recentes vieram pôr, definitivamente, uma pá de cal nestas interpretações. De fato, o agravamento da distribuição de renda persistiu, apesar de ultrapassadas as fases iniciais do desenvolvimento econômico e

*Economista, pesquisadora do IPEA.

[1]Informações deste trecho foram extraídas de Langoni (1973) e Bonelli (1988) e são relativas à população economicamente ativa com rendimento.

mesmo em face da crise recessiva aos ciclos de curto prazo que caracterizaram a década de 80. Os dados anuais de que se dispõem mostram apenas uma breve desconcentração em 1986, devido ao Plano Cruzado[2].

Tendo em vista que há consenso quanto ao objetivo de atingir-se uma sociedade igualitária, em particular crescentemente igualitária, o agravamento da desigualdade no Brasil a partir de níveis considerados já críticos é um tema de interesse geral e coeficientes de Gini são muitas vezes utilizados como medida de referência. O objetivo deste artigo é sistematizar questões metodológicas, especialmente no que concerne à base de dados, que se colocam quando se trata da verificação empírica da desigualdade no Brasil.

Este texto é composto de três seções, além desta introdução. Na seção 1, são enfocadas algumas questões gerais relevantes para a mensuração da desigualdade, especificamente aspectos conceituais da variável básica e opções de definição da unidade de análise a ser adotada, o que remete à problemática de construção de uma distribuição de rendimentos específica como base para a mensuração da desigualdade.

A segunda seção trata das questões empíricas relativas à utilização de informações oriundas do Censo Demográfico e da PNAD para cálculo do índice de Gini. Explicitam-se as características dos dados disponíveis, detacando as implicações do uso de dados publicados ou, alternativamente, do recurso ao banco de dados. Os impactos de diferentes convenções metodológicas sobre os valores obtidos para o Gini são ilustrados com base na PNAD.

Na seção final destaca-se a necessidade de cautela na comparação de índices de Gini oriundos de diferentes estudos, chamando-se a atenção para a relativa invariância do indicador, que é suscetível de ter implicações para a interpretação de variações de seu valor ao longo do tempo.

2 DESIGUALDADE E DISTRIBUIÇÃO DE RENDA

Num contexto macroeconômico, o conceito de desigualdade refere-se essencialmente à noção teórica de distribuição do bem-estar para um determinado conjunto de indivíduos. Dadas as dificuldades empíricas de mensuração, a opção é adotar como *proxy* do bem-estar global dos indivíduos uma variável monetária de caráter o mais

[2] O efeito distributivista do Plano Cruzado se deveu em grande parte ao aumento geral de salários, uma das medidas básicas do plano. O salário mínimo foi ajustado em 34%. Simultaneamente, o abono salarial de 8% permitiu aumentar a renda da classe média, melhorando as condições de barganha salarial para trabalhadores menos qualificados no setor informal de prestação de serviços. Esses ganhos salariais foram significativos em termos reais, na medida em que houve queda drástica da inflação a partir de março de 86. Bonelli e Sedlacek (1989) ressaltam que o aumento da renda média foi significativo em 1986 (40%), enquanto a melhoria do Gini bem pequena (de 0,5888 para 0,577), o que significa um deslocamento da distribuição para direita sem grandes mudanças na sua forma.

genérico possível. Variáveis que reflitam o valor da riqueza, do consumo ou da renda atendem razoavelmente a esses requisitos. A disponibilidade de informações estatísticas fazem da renda a variável privilegiada para os estudos de desigualdade, apesar das suas reconhecidas limitações. Dentre tais limitações, cabe destacar:

a) as estatísticas de rendimento não levam em consideração as diferenças de acesso a serviços não-mercantis (geralmente ofertados pelo setor público, como saúde, educação, infra-estrutura urbana, etc.), além de não incorporarem na maior parte das vezes o valor monetário de consumos não vinculados ao mercado (transferências, doações, autoconsumo — inclusive o da casa própria), que têm significativo efeito na distribuição de bem-estar entre indivíduos;
b) as diferenças de rendimento não representam necessariamente desigualdades de bem-estar, mas diferenças de necessidades. Assim, podem existir diferenciais de consumo e de preços vinculados ao modo de vida urbano e rural que se refletem na renda, sem que haja obrigatoriamente discrepâncias no nível de bem-estar; e
c) a variável renda é afetada por defeitos de informação, particularmente subdeclaração das mais altas e omissão relacionadas a certos tipos de atividades informais.

A adoção dessa variável abre um amplo espectro de possibilidades teóricas para os enfoques de desigualdade, cabendo escolher a categorização mais adequada ao tipo de estudo que se quer empreender. As diferenciações se referem a dois aspectos principais: o conceito de renda e a unidade de análise.

a) **o conceito de renda** – A variável pode ser tomada de forma restrita, como um tipo específico de remuneração – por exemplo salário —, ou de forma ampla, considerando qualquer tipo de rendimento. Assim, em estudo sobre o impacto da educação sobre o nível de desigualdade, Reis e Barros (1989) utilizaram a distribuição de salários corrigidos por horas trabalhadas para estabelecer a desigualdade interna em cinco grupos educacionais definidos por anos de estudo. O total de rendimentos de todas as fontes é utilizado quando se trata da desigualdade de renda de forma mais geral (Langoni, 1973; Bonelli e Sedlacek, 1989), tendo a vantagem de captar parte da diferenciação na posse da riqueza ao incluir os rendimentos do patrimônio. Possibilidades intermediárias são muito exploradas em países onde transferências em dinheiro consistem em instrumentos importantes de política social. Na Suécia, por exemplo, onde as transferências representam 29,3% da renda bruta total e 66% no quintil inferior (O'Higgins, Schmaus e Stephenson, 1989), é relevante distinguir se a medida da desigualdade inclui ou não as transferências.

A conceituação da renda como bruta ou líquida de impostos diretos é uma diferenciação utilizada para verificar a eficácia distributiva da estrutura fiscal. Existem estudos americanos (Rice, 1989) e a base de dados reunida no Luxembourg Income Study - LIS - permite esta abordagem, inclusive do ponto de vista da comparação entre os sete países tratados (O'Higgins, Schmaus e Stephenson, 1989). No Brasil as estatísticas demográficas registram rendimento bruto exclusivamente, havendo dificul-

dades metodológicas de compatibilizar estas fontes e os registros fiscais de modo a obter as distribuições de renda bruta e líquida. Vale lembrar que, do ponto de vista da definição, embora as informações de renda coletadas pela PNAD e pelo Censo devam ser de renda bruta, os dados obtidos contêm inevitavelmente defeitos de informação (renda líquida ao invés de renda bruta).

Qualquer que seja o conceito de renda utilizado, é relevante deixar explícito se estão incluídas na distribuição as unidades de observação com renda zero. De fato, a partir de uma mesma base de dados e o mesmo conceito de renda, os resultados obtidos em termos da medida de desigualdade podem ser significativamente diferentes em função da escolha feita. Como se verá adiante, dependendo do caso, haverá maior número de argumentos a favor ou contra a inclusão da renda zero nas medidas de desigualdade. Para a desigualdade medida pelo coeficiente de Gini, Denslow e Tyler (1983) desenvolveram uma fórmula que permite estimar de modo expedito o impacto da renda zero sobre a desigualdade da distribuição[3].

b) **unidade de análise** – Neste particular cabe distinguir duas opções básicas: as que adotam como unidade de análise o indivíduo — qualquer que seja o conjunto mais amplo a que pertença — ou as que privilegiam um grupo solidário — geralmente *família ou domicílio*[4].

Ao se considerar o indivíduo como unidade básica, pode-se partir desde distribuições de variáveis tão abrangentes, como população de 10 anos e mais (Romão, 1990), até as mais restritas, como a de salários, sendo freqüente o uso da população economicamente ativa (Hoffmann, 1989; Bonelli e Sedlacek, 1989). Na verdade a escolha depende essencialmente dos objetivos de análise e das informações disponíveis.

Pode-se argumentar que as medidas de desigualdade baseadas na família ou domicílio apresentam vantagens por já incorporarem uma certa redistribuição do rendimento que se dá no interior destes grupos. A desvantagem óbvia é desconsiderar as desigualdades entre famílias ou domicílios, que, embora com a mesma renda, desfrutam de diferentes níveis de bem-estar devido a diferenças no seu tamanho e composição.

O uso de escalas de equivalências visa justamente a ajustar as unidades de análise (famílias ou domicílios), de modo que as diferenças em tamanho e composição não prejudiquem a medida de desigualdade entre elas. Analiticamente pode-se distinguir dois tipos de ajuste. O primeiro visa a neutralizar os efeitos de composição, atribuindo aos participantes do grupamento pesos diferentes segundo sexo e idade. Por exemplo, ao indivíduo adulto do sexo masculino é atribuído peso 1, e pesos menores aos demais,

[3] $G_n = G_x + \frac{z}{1+z}(1 - G_x)$, onde G_n é o Gini incluindo as unidades sem rendimento, G_x é o Gini excluindo as sem rendimento e z a relação entre o número dos sem rendimento daqueles com rendimento positivo (Denslow e Tyler, 1983, p. 33 e seguintes). Assim, quanto maior z, mais elevado o valor do índice de Gini — G_n.

[4] Dada uma variável, indivíduo fica claramente definido. No entanto, no que concerne família e domicílio, vários conceitos são utilizados pelos sistemas estatísticos nacionais. Mesmo para o Brasil, os dados demográficos coletados pelo IBGE permitem algumas opções na definição dessas unidades.

implicando participações diferenciadas na despesa total. A escala pioneira foi concebida por Engel em 1883 e uma adaptação para o Brasil foi proposta pelo IBGE na década de 70, no escopo do ENDEF (Anexo I).

Uma abordagem alternativa privilegia as economias de escala que ocorrem quando se trata de uso solidário da renda. No contexto do LIS foi usada uma escala que atribui a cada indivíduo adicional a partir do primeiro o peso de 0,5 (O'Higgins, 1989). Nos Estados Unidos os valores das linhas de pobreza adotadas pelo Department of Health and Human Services[5] incorporam um adicional de 34% por indivíduo. No Brasil, Fishlow (1972) estimou elasticidades das despesas com alimentação em relação a tamanho da família em 0,89 em áreas rurais e 0,82 em áreas urbanas.

As duas abordagens podem ser combinadas, isto é, é possível levar em consideração simultaneamente pesos diferenciados para os membros da família conforme suas características de sexo e idade, e economias de escala de grupamento em função do número de membros obtidos em termos de adultos-equivalentes.

É óbvio que qualquer escala de equivalência é necessariamente arbitrária. Na verdade ela deve ser específica na medida em que a função de repartição de renda assumida pelo grupamento solidário de pessoas (família ou domicílio) varia de lugar para lugar, e, no mesmo lugar, ao longo do tempo, em resposta a determinantes socioculturais. Empiricamente, no entanto, a utilização de diferentes escalas de equivalência para estudos em *cross-section* e de séries temporais poria em dúvida os resultados obtidos, já que diferenças no nível de desigualdade de duas populações poderiam ser imputadas ao uso de diferentes escalas de equivalência. Na prática acaba-se, pois, por optar por uma única escala de equivalência quando se trata de comparação de níveis de desigualdade[6].

O cálculo da renda familiar ou domiciliar *per capita*, que equivale à adoção de peso 1 para cada indivíduo no grupamento, é por alguns considerado como maneira mais efetiva de contornar a arbitrariedade inevitável de outras escalas de equivalência, sem deixar de levar em conta os efeitos significativos do tamanho do grupamento a que pertencem sobre o bem-estar das pessoas[7].

Sistematizando, pode-se conceber seis tipos de distribuição estabelecendo relações diferentes entre rendimento e unidade de análise[8]: três que tomam o grupamento como unidade de análise e três onde o indivíduo desempenha esta função. Assim,

[5] Baseado nos valores fixados em fevereiro de 1989 e publicados no Social Security Bulletin, vol. 52, nº 3, march 1989.

[6] É o que se faz no estudo do LIS comparando desigualdades entre Canadá, Estados Unidos, Inglaterra, Alemanha, Suécia, Noruega e Israel (O'Higgins, Schmaus, Stephenson, 1989, p. 112).

[7] Rossi (1988) afirma que a distribuição dos indivíduos segundo a renda familiar *per capita* "é a distribuição mais adequada para aferir-se o grau de bem-estar dos indivíduos".

[8] Atkinson distingue nove tipos, mas três deles utilizam as unidades de equivalência como unidade básica, o que não parece defensável (Atkinson, 1983).

dada uma renda familiar (ou domiciliar) Y de um grupamento de n indivíduos ou n^* indivíduos-equivalentes, pode-se construir as seguintes distribuições:

Distribuição das famílias segundo:	renda familiar	Y
	renda familiar *per capita*	Y/n
	renda familiar equivalente	Y/n^*
Distribuição dos indivíduos segundo:	renda familiar	Y
	renda familiar *per capita*	Y/n
	renda familiar equivalente	Y/n^*

Conceitualmente, as duas últimas opções são as mais adequadas por levar em conta explicitamente o número de indivíduos e sua participação na renda. O questionamento teórico se limita à escolha da escala de equivalência[9].

3 Opções Empíricas de Mensuração no Brasil

3.1 A Investigação de Dados de Rendimento

No Brasil, o Censo Demográfico reúne o conjunto mais completo de informações sobre rendimentos. Por ser um levantamento universal, permite que sejam calculadas medidas de desigualdade até para áreas de análise bastante restritas[10], distinguindo estrato urbano e rural, se desejado. Como o Censo se realiza a cada dez anos no ano zero, as informações disponíveis não possibilitam o acompanhamento da evolução ocorrida ao longo da década.

Os dados anuais da PNAD permitem que se faça esse acompanhamento com algumas restrições. Em particular, a amostra não cobre a população rural da região norte, além de o seu nível de representatividade não ir além do de algumas Unidades da Federação permitindo, no entanto, distinguir os estratos urbanos, rural e, quando é o caso, metropolitano (Anexo II).

As variáveis de rendimento relacionadas ao trabalho têm lugar de destaque tanto

[9] Na verdade o uso da renda familiar *per capita* representa apenas a opção por um tipo específico de escala de equivalência, o que na verdade reduziria as opções a quatro.

[10] As agregações possíveis vão do nível mais amplo do País como um todo, até os mais restritos, de subsetores censitários. Obviamente, em função de restrições de custos, os procedimentos que devem ser adotados são diferentes em função do número de informações existentes ao nível da unidade de análise escolhida.

no Censo como na PNAD. Para o trabalho principal, ambas as pesquisas investigam o valor bruto do rendimento monetário, o valor do pagamento feito *in natura*, sendo possível normalizar as informações individuais pelo número de horas trabalhadas. Além do rendimento de demais trabalhos, são investigados outros quesitos de rendimentos, embora de forma não estritamente comparável nas duas pesquisas (Tabela 1).

TABELA 1

Variáveis de Rendimento Investigadas
no Censo Demográfico e nas PNADs

Variáveis	Censo	PNADs
Trabalho principal		
em dinheiro	*	*
in natura	*	*
Outros trabalhos		
em dinheiro	*	*
in natura	-	*
Aposentadoria	*	*
Pensão		*
Abono Permanência ..		*
Aluguel	*	*
Rendimentos de capital	*	*
Doação, mesada, etc. .	*	*

FONTE – IBGE, Censo Demográfico e PNADs.

A respeito dessas variáveis é fundamental destacar que se trata de quesitos investigados, cujas respostas obtidas se encontram nos bancos de dados de Censo e da PNAD. A sua utilização para obtenção de tabulações e/ou cálculo de medidas de desigualdade depende de acesso ao banco via terminais do IBGE. O acesso a esses dados básicos permite flexibilidade na construção da variável renda (que quesito ou que combinação de quesitos considerar), assim como na definição da unidade de análise (quais indivíduos ou que tipo de grupamento de indivíduos)[11]. Em contrapartida implica, dependendo da especificação, custos mais ou menos elevados associados à programação e processamento.

Para monitoramento da evolução da desigualdade, a solução mais direta e imediata é o recurso aos dados publicados, o que, embora signifique restrições quanto à especificidade das definições, e, por conseqüência, da análise, apresenta vantagens óbvias de acessibilidade e custo.

[11] Os agrupamentos de indivíduos são geralmente famílias e domicílios. Existem, no entanto, diferentes opções possíveis na definição desses agrupamentos quando se recorre aos dados básicos. Por exemplo, pode-se definir família apenas como núcleo de pais e filhos, ou utilizar concepções mais amplas usando laços de parentesco e incluindo ou não agregados e empregados.

3.2 O Uso dos Dados Publicados

A agregação das informações por classe de rendimento para efeito de publicação torna inevitável que a medida de desigualdade calculada seja uma estimativa do valor real associado à distribuição de renda individualizada a nível de pessoas, famílias ou domicílios.

Com base nas informações de freqüência e rendimento médio por classe, é possível utilizar dois procedimentos alternativos para o cálculo da medida de desigualdade:

a) **Utilização de parâmetro de uma função ajustante à distribuição de renda**

A função de Pareto é freqüentemente utilizada para este fim porque os seus parâmetros permitem o cálculo direto do Índice de Gini. Uma expressão simples da função de Pareto é

$$N = aY^{-\alpha}$$

onde N é a freqüência de observações para renda maior que Y. Para esta formulação, o Gini é dado por

$$g = 1/(2\alpha - 1)$$

Naturalmente, a precisão da estimativa do índice obtida depende da adequação da função escolhida à distribuição de renda observada[12].

b) **Estabelecimento de limites de variação do Índice de Gini com base em hipótese sobre a distribuição de renda no interior de cada classe**

Trata-se do procedimento proposto por Gastwirth (1972) que estabelece como limite inferior o valor do Gini caso a renda fosse perfeitamente distribuída (desigualdade nula) no interior de cada classe; por outro lado, o limite superior deriva da hipótese de desigualdade máxima no interior de cada classe. Naturalmente que para uma mesma distribuição, quanto mais numerosas forem as classes de rendimento utilizadas, menor será o intervalo do Gini, significando melhor precisão. As expresssões utilizadas para o cálculo dos limites são:

$$g_i = (\Sigma X_i\ Y_{i+1}) - (\Sigma X_{i+1}\ Y_i)$$

onde:

g_i limite inferior do intervalo do Gini

X_i percentual acumulado da população até o estrato, e

Y_i percentual acumulado da renda até o estrato

$$g_s = g_i + \frac{1}{\overline{Y}} \sum_i^n p_i^2 \frac{(\overline{Y}_i - I_i)(S_i - \overline{Y}_i)}{S_i - I_i}$$

[12] Rossi (1982, Cap. 3) faz uma excelente sistematização sobre as funções ajustantes propostas na literatura recente, apresentando as expressões dos índices de desigualdade correspondentes a partir dos parâmetros obtidos.

onde

g_s limite superior do intervalo do Gini

p_i percentual da população no estrato

I_i limite inferior do estrato, e

S_i limite superior do estrato

$\overline{Y}$ renda média da distribuição

$\overline{Y}_i$ renda média do estrato

A utilização adequada dos procedimentos acima depende da disponibilidade da freqüência e do rendimento médio (ou rendimento total) por classe de rendimento. Algumas tabulações publicadas relativas ao Censo Demográfico apresentam freqüências, mas não rendimento médio. O recurso ao ponto médio da classe como substituto à informação de rendimento médio pode conduzir a um grau inaceitável de precisão da medida de desigualdade, isto é, erros significativos da estimativa de Gini no ponto ou obtenção de intervalo (limite superior e inferior) incompatível com o que seria calculado a partir do verdadeiro rendimento médio. Desse modo, é recomendável utilizar preferencialmente para fins de cálculo da medida de desigualdade tabulações que apresentam freqüências e rendimento médio por classe de rendimento.

Apenas três tipos de tabulações de rendimento (uma no Censo e três nas PNADs a partir de 1984) são publicados dentro destas especificações:

a) **Rendimento de qualquer natureza das pessoas de 10 anos e mais**

Informação publicada tanto pelo Censo como pela PNAD. A tabulação do Censo permite distinguir as pessoas simultaneamente por sexo e situação de domicílio (urbano ou rural), enquanto na da PNAD as informações por sexo e situação de domicílio são mutuamente exclusivas.

b) **Rendimento de qualquer natureza da PEA**

c) **Rendimento de todos os trabalhos das pessoas ocupadas**

As tabulações b) e c) são publicadas apenas pelas PNADs, por nove classes de rendimento[13] expressas em termos de salários mínimos, estabelecendo distinção por sexo.

Como a escolha da variável e do tratamento da renda zero tem impacto significativo sobre o coeficiente de desigualdade, é importante buscar a solução mais adequada para o estudo que se pretende. As três variáveis publicadas pela PNAD referem-se a subconjuntos progressivamente mais restritos de modo que, respeitada a mesma convenção para a renda zero, a tendência é obter coeficientes de desigualdade progressivamente menores para a mesma base de dados. O esquema abaixo ajuda a esclarecer este ponto. Como

$$A \subset B \subset C,$$

[13] São nove classes para os rendimentos declarados (inclusive a classe dos sem rendimentos). Os sem declaração de rendimento, que representam uma classe para efeito de freqüência, são excluídos para fins da medida de desigualdade.

logo

$$\text{Gini}_A > \text{Gini}_B > \text{Gini}_C$$

onde:

A = população de 10 anos e mais,

B = *população economicamente ativa*, e

C = pessoas ocupadas.

Quanto à inclusão dos indivíduos com renda zero, há implicações diferentes conforme a variável escolhida (Tabela 2). Alguns analistas optam por excluir a renda zero com base no argumento de que se trata de "falsa renda zero", sendo a remuneração da pessoa de fato percebida pelo chefe da família (caso de trabalho familiar) ou feita *in natura*. No que concerne ao primeiro caso, o contra-argumento é que a renda do chefe de família, quando não corresponde de fato apenas à pessoa do chefe, não pode ser identificada, sendo mantida na distribuição de renda para cálculo do coeficiente de desigualdade, independentemente do tratamento dado à renda zero.

TABELA 2

Limites Inferior e Superior do Índice de Gini,
Segundo Diferentes Categorias Populacionais
e Anos Selecionados

Especificação	Inclusão de pessoas c/ rendimento zero		Exclusão de pessoas c/ rendimento zero	
	Lim. Inf.	Lim. Sup.	Lim. Inf.	Lim. Sup.
1984				
Pop. de 10 anos e mais	0,765580	0,775201	0,584843	0,601881
PEA	0,630393	0,645677	0,575590	0,593141
Pessoas Ocupadas . . .	0,614192	0,629727	0,572471	0,589687
1986				
Pop. de 10 anos e mais	0,758975	0,769964	0,590713	0,609373
PEA	0,618522	0,636326	0,577621	0,597334
Pessoas Ocupadas . .	0,606314	0,623403	0,573168	0,591695
1988				
Pop. de 10 anos e mais	0,773750	0,785150	0,621551	0,640620
PEA	0,651537	0,669498	0,609994	0,630096
Pessoas Ocupadas . .	0,637781	0,655347	0,606885	0,625949

FONTE – IBGE, Pesquisa Nacional por Amostra de Domicílios, 1984,1986 e 1988.

NOTAS:

1. Limites inferior e superior do índice de Gini de acordo com a fórmula de Gastwirth.
2. Exclui área rural da Região Norte. Refere-se ao rendimento de qualquer natureza da população de 10 anos e mais ou da PEA. Para as pessoas ocupadas, a variável se refere aos rendimentos de todos os trabalhos.

Dentre as três variáveis de rendimento publicada, a de caráter mais geral se refere à população de 10 anos e mais, que compreende tanto indivíduos que recebem exclusiva-

mente transferências (aposentadoria, mesada, etc.), rendimento de capital e aluguéis, como crianças e jovens que não ingressaram no mercado de trabalho, não tendo rendimento próprio na sua grande maioria[14]. Em conseqüência, para cálculo de medida de desigualdade é recomendável a exclusão dos indivíduos com rendimento zero, de modo a evitar que ela seja indevidamente aumentada em função de fenômenos estritamente demográficos (variação da proporção de jovens de 10 a 18 anos na população).

Para as variáveis de rendimento da PEA ou de pessoas ocupadas, a inclusão ou não das pessoas com renda zero é discutível, sendo afetada fundamentalmente pelas distorções na atribuição do rendimento no caso do trabalho familiar, como já se mencionou. Para comparações intertemporais, porém, é interessante considerar as distribuições com e sem renda zero, já que a proporção de pessoas nesse estrato varia sensivelmente de ano para ano, podendo, por si só, alterar as conclusões quanto à evolução do índice de desigualdade (Tabela 3). Assim, por exemplo, entre 1984 e 1986 as variações dos limites dos intervalos são tênues (Tabela 2), mas em sentido inverso quando se considera rendimento zero ou não. Entre 1986 e 1988 o agravamento da desigualdade é bastante pronunciado (limites superiores aos de 1986 em cerca de 5%), exceto no que concerne à população de 10 anos e mais, incluindo rendimento zero. Naturalmente, devido à elevada proporção dos sem rendimentos na população de 10 anos e mais, é essa variável a mais suscetível a apresentar resultados incongruentes num e noutro caso.

TABELA 3

Percentual de Pessoas com Rendimento Zero, na População de 10 anos e mais, na População Economicamente Ativa e no Conjunto de Pessoas Ocupadas
Brasil - 1984/1988

Anos	Pessoas de 10 anos e mais (%)	PEA (%)	PO (%)
1984	43.54	12.91	9.76
1985	41.95	12.08	9.71
1986	41.11	9.68	7.77
1987	40.55	10.88	8.18
1988	40.22	10.65	7.86

FONTE - IBGE, Pesquisa Nacional por Amostra de Domicílios.

[14] O acesso aos dados básicos possibilita, naturalmente, a construção de um agregado, excluindo os jovens sem rendimento por ainda não terem ingressado no mercado de trabalho.

3.3 O uso do Banco de Dados

O uso dos dados básicos dos arquivos do Censo e das PNADs apresentam duas vantagens. A primeira é de permitir que se obtenha uma medida de desigualdade "verdadeira" em relação aos dados, e não uma estimativa ou um intervalo de segurança, como ocorre no caso de dados agregados. A segunda vantagem é a possibilidade de definir tanto a variável rendimento como a população de acordo com as necessidades de análise, tendo como restrição apenas a abrangência dos dados coletados e, no caso da PNAD, limitações amostrais.

Assim, o acesso ao banco de dados permite que se construam medidas de desigualdade nas quais seja privilegiado o papel distributivo da família, utilizando escalas de equivalência, como se viu anteriormente. A consideração de todos os rendimentos e a participação de todos os membros da família no rateio desses rendimentos fazem com que os índices de desigualdade assim construídos tenham vantagens de abrangência em relação aos construídos diretamente a partir de variáveis de pessoas. À guisa de exemplo, a Tabela 4 apresenta os Índices de Gini calculados quando se considera a desigualdade entre famílias com base na renda familiar ou entre pessoas com base na renda familiar *per capita*, isto é, levando em conta, explicitamente, o tamanho da família e atribuindo o mesmo peso a cada um dos seus membros[15]. Em ambos os casos, o acesso ao banco de dados é essencial, seja para ter o Gini "verdadeiro", seja para construir a renda familiar *per capita*.

Os resultados obtidos revelam sistematicamente maior desigualdade quando se considera a renda familiar *per capita*, o que resulta da conhecida relação inversa entre valor do rendimento e tamanho da família. É importante notar ainda que as diferenças nos valores dos índices de Gini nos dois casos não são desprezíveis.

O recurso ao banco de dados, embora abra amplas possibilidades ao usuário, exige que sejam feitas numerosas escolhas metodológicas. No caso da PNAD, cabe destacar dois pontos: o tratamento de *outliers* e o procedimento de expansão.

Outlier pode ser genericamente entendido aqui como o dado de renda que parece inconsistente com a maioria de dados da distribuição[16]. Quer seja uma informação genuína ou resultado de erro de coleta ou processamento, a ocorrência na amostra de um (ou mais de um) dado de rendimento muito elevado é capaz de afetar o índice de desigualdade, desvirtuando as comparações intertemporais e em *cross-section*.

O procedimento adotado pelo IBGE é de não excluir os *outliers* genuínos. Dessa forma, tanto as tabulações publicadas como o banco de dados levam em conta eventuais informações de rendimento muito elevado. No caso de uso de dados publicados, o usuário não tem a possibilidade de excluir o *outlier*, mesmo que seja uma anomalia

[15] Em estudos sobre pobreza metropolitana, a autora tem se baseado na distribuição da renda familiar *per capita*, utilizada em confronto com a linha de pobreza para distinguir pobres de não-pobres, assim como para cálculo de Gini para cada uma dessas subpopulações (Rocha, 1989).

[16] A qualificação de um dado como *outlier* é essencialmente subjetiva. A respeito ver Oscar H. Bustos (1988).

TABELA 4

Índice de Desigualdade de Rendimento entre Famílias e entre Pessoas, Segundo as Regiões Metropolitanas – 1988

Regiões Metropolitanas	Famílias	Pessoas
Belém	0,62374	0,63869
Fortaleza	0,63421	0,66330
Recife	0,63249	0,65380
Salvador	0,64987	0,67156
Belo Horizonte	0,59743	0,63291
Rio de Janeiro	0,60032	0,61959
São Paulo	0,52935	0,55001
Curitiba	0,52377	0,58015
Porto Alegre	0,56296	0,58015

FONTE – IBGE, Pesquisa Nacional por Amostra de Domicílios – 1988 (Tabulações especiais).
NOTA – Calculados respectivamente com base na renda familiar e na renda familiar *per capita*.

facilmente detectável nas informações agregadas por estrato[17]. Já utilizando diretamente o banco de dados, o usuário está capacitado a utilizar o procedimento que julgar cabível em relação aos *outliers*, sejam eles genuínos ou erros de coleta ou digitação.

Com relação a esses últimos ocorre, em particular, de o código utilizado para os sem declaração de rendimento (999 999 999) ser mal digitado, com menor número de noves, sete ou oito dígitos, o que acaba por significar a incorporação de uma renda indevida elevada à amostra. No caso de acesso ao banco de dados, pode ser introduzido um filtro com objetivo explícito de eliminar tais informações da amostra.

Para outros tipos de *outliers* poderá ser verificada a autenticidade de rendas elevadas pela checagem de demais características dos informantes, mas, em última instância, a sua manutenção ou exclusão da amostra — o que afeta o valor do Gini — depende do julgamento subjetivo do analista.

Reconhecendo o efeito potencial dos *outliers* no sentido de afetar o valor do Gini e, eventualmente, modificar as conclusões sobre sua tendência evolutiva, Bonelli e Sedlacek (1990) calculam os Índices de Gini estabelecendo uma truncagem equivalente a 0,02% dos indivíduos na cauda superior da distribuição de rendimento da PEA. A comparação dos Índices de Gini obtidos a partir da distribuição original e da distri-

[17] Nas informações da PNAD-88, existe, por exemplo, um *outlier* no Estado do Pará que faz com que o rendimento médio do estrato de mais de 20 pisos salariais seja 9,4 vezes superior ao rendimento médio do estrato imediatamente inferior. Essa informação de renda faz com que o Gini entre 1987 e 1988 apresente um agravamento exagerado, os intervalos passando de (0,586-0,601) em 1987 para (0,750-0,812), em 1988.

buição truncada serve como teste de robustez em relação às tendências de evolução da desigualdade[18]. Os resultados obtidos por aqueles autores, apresentados na Tabela 5, denotam a mesma tendência em ambos os casos.

TABELA 5

Gini da Distribuição de Rendimendo da População Economicamente Ativa com e sem Exclusão do Extremo Superior da Distribuição
Brasil – 1983/1988

Anos	Distribuição Original	Distribuição Truncada
1983	0,6305	0,6264
1984	0,6273	0,6250
1985	0,6371	0,6336
1986	0,6200	0,6135
1987	0,6237	0,6220
1988	0,6433	0,6399

FONTE – Valores extraídos do conjunto mais amplo de resultados apresentados por Bonelli e Sedlacek (1990) com base na Pesquisa Nacional por Amostras de Domícilios.
NOTA – Inclusive pessoas sem rendimento.

A segunda questão relaciona-se aos procedimentos de expansão e uso dos pesos. Tradicionalmente a existência de variáveis de peso na PNAD (peso da pessoa, do chefe da família, do chefe do domicílio) visava a calibrar os resultados em função de características dos elementos da amostra (sexo e idade da pessoa). Ao longo do tempo, no entanto, a variância dos pesos vem diminuindo sistematicamente, o que evidencia o abandono progressivo do procedimento de ajustamento às projeções demográficas por sexo e idade (Tabela 6). Em 1988, por exemplo, para áreas básicas de desagregação, como as regiões metropolitanas, a amplitude dos pesos em cada uma delas é insignificante, o que resulta praticamente na utilização de um fator de expansão único para todos os elementos da amostra. Como conseqüência, os valores de Gini calculados a partir da amostra ou de sua expansão são praticamente os mesmos, já que a forma da distribuição de rendimentos expandida replica quase exatamente a da distribuição de renda da amostra. Em áreas "compostas"[19] é, no entanto, imprescindível o uso dos pesos para cálculo da medida de desigualdade, uma vez que a maior variância dos pesos — resultante de diferentes frações de amostragem nas áreas básicas de desagregação

[18] Em relação à evolução da desigualdade é, no entanto, importante lembrar que as informações da PNAD-84 estão truncadas a nível de codificação, já que os sete dígitos previstos não foram suficientes para o registro das rendas mais elevadas observadas. Neste sentido a melhoria da distribuição naquele ano pode, pelo menos em parte, ser imputável a essa truncagem inicial.

[19] Estão sendo denominadas aqui de áreas de análise "compostas" aquelas que comportam mais de uma área de análise representativa a nível da PNAD. Assim, Brasil, conjuntos de estados ou um estado para o qual seja possível distinguir pelo menos estrato urbano e rural constituem-se em unidades "compostas".

— implica diferenças sensíveis entre a distribuição de renda da amostra e a expandida. Algumas informações sobre a variável peso da pessoa na PNAD-88 (Tabela 7) servem para ilustrar este argumento.

TABELA 6

Informações sobre a Evolução da Variável Peso da Pessoa, na Região Metropolitana de São Paulo

Anos	Região Metropolitana de São Paulo				
	Tamanho da Amostra	Valor Máximo	Valor Mínimo	Média	Coefic. de Variação
1976	28.035	1.110	285	385,68	0,3192
1981	30.629	440	422	431,09	0,0186
1984	34.235	435	426	430,66	0,0080
1988	17.093	975	974	974,12	0,0034

FONTE – IBGE, Pesquisa Nacional por Amostra de Domicílios – 1988 (tabulações especiais).

TABELA 7

Informações sobre a Variável Peso da Pessoa Segundo Áreas de Análise

Áreas de Análise	Valor Mínimo	Valor Máximo	Média	Desvio Padrão	Coefic. de Variação
Brasil	128	1306	473,85	256,78	54,19
Região Sul	212	784	495,74	267,89	54,04
Rio Grande do Sul	212	747	412,31	258,51	62,70
Porto Alegre ...	212	213	212,64	0,48	0,22

FONTE – IBGE, Pesquisa Nacional por Amostra de Domicílios – 1988 (tabulações especiais).

4 CONCLUSÃO FINAL

Embora o recurso ao Índice de Gini como medida de desigualdade venha-se tornando progressivamente mais comum, particularidades quanto aos dados estatísticos básicos tornam indispensável cautela na sua construção e interpretação dos valores obtidos.

Como foi discutido acima, existe opção entre recorrer a informações publicadas ou

ao banco de dados para a construção de indicador. No primeiro caso, é inevitável incorporar no indicador as convenções e eventuais defeitos que tenham escapado aos procedimentos de crítica do IBGE. No segundo caso, as múltiplas possibilidades conceituais e metodológicas exigem opções criteriosas adequadas aos fins analíticos pretendidos.

É importante destacar que, especialmente no que concerne aos Índices de Gini calculados a partir do banco de dados, os resultados são dificilmente comparáveis entre estudos diferentes. Na verdade, é provável que em face das inúmeras escolhas em termos de conceituação da população, da variável rendimento, do tratamento da renda zero, dos *outliers*, dos pesos para expansão, se obtenham variações relativas do valor de Gini para um mesmo ano de grandeza relativa comparável às variações observadas ao longo do tempo para índices construídos a partir das mesmas convenções.

A relativa inércia do valor do Gini ao longo do tempo, quando calculado segundo critérios comparáveis, é um fato que merece toda atenção. O que se observa é que, em face de modificações conjunturais importantes, como os ciclos de curto prazo na década de 80, os índices obtidos se alteram muito pouco. Dadas as pequenas variações observadas ano a ano para as medidas de desigualdade calculadas a partir da PNAD, é relevante questionar se estas variações refletem uma verdadeira, embora tênue, mudança do nível de desigualdade, ou/e, ao contrário, não são significativas estatisticamente[20]. Na verdade é possível conceber que a fragilidade inerente à informação de rendimento, especialmente a reconhecida subdeclaração das rendas mais altas, aliada a outros problemas de registro e tratamento da informação, represente turbulências em escala tal que uma medida de desigualdade com as características de inércia do Gini pode ser incapaz de absorver.

[20] Ricardo Paes e Barros (PEA/INPES) e outros estudam atualmente testes que permitam avaliar a significância estatística do indicador.

ANEXO I

Escalas de Equivalência, Segundo Idade e Sexo

Idade em Anos	Engel		IBGE/ENDEF	
Menos de 1	0,30		0,30	
1	0,39		0,40	
2	0,43		0,40	
3	0,48		0,40	
4	0,52		0,40	
5	0,54		0,50	
6	0,57		0,50	
7	0,60		0,50	
8	0,63		0,50	
9	0,67		0,50	
10	0,70		0,60	
11	0,74		0,60	
12	0,76		0,60	
13	0,78		0,60	
	H	M	H	M
14	0,85	0,80	0,70	0,70
15	0,90	0,81	0,70	0,70
16	0,95	0,80	0,70	0,70
17	0,97	0,76	1,0	1,0
18	0,99	0,77	1,0	1,0
19	0,99	0,72	1,0	1,0
	H	M		
20 - 24	1,0	0,8	1,0	0,85
25 - 49	1,0	0,8	1,0	0,85
50 - 59	1,0	0,8	1,0	0,85
60 - 70	0,90	0,70	1,0	0,85
70 - 80	0,90	0,70	1,0	0,85
Maior de 80	0,90	0,70	1,0	0,85
Data de Criação da Escala	1983		1976	

FONTE – Reproduzido de François, Patrick, Nutrição/Despesas/1976. IBGE, mimeo., 1976.

ANEXO II

Informações de Rendimentos da PNAD
Publicadas desde 1981

Grandes Regiões	1981	1982	1983	1984	1985	1986	1987	1988
REGIÃO NORTE	c*	b*	b*	a*	a*	a*	a*	a*
Amazonas	-	-	b*	a*	a*	a*	a*	a*
Pará	-	-	b*	a*	a*	a*	a*	a*
Belém	c**	b**	b**	a**	a**	a**	a**	a**
REGIÃO NORDESTE	c	b	b	a	a	a	a	a
Maranhão	-	-	b	a	a	a	a	a
Piauí	-	-	b	a	a	a	a	a
Ceará	c	b	b	a	a	a	a	a
Rio G. do Norte	-	-	b	a	a	a	a	a
Paraíba	-	-	b	a	a	a	a	a
Pernambuco	c	b	b	a	a	a	a	a
Alagoas	-	-	-	a	a	a	a	a
Sergipe	-	-	-	a	a	a	a	a
Bahia	c	b	b	a	a	a	a	a
Fortaleza	c**	b**	b**	a**	a**	a**	a**	a**
Recife	c**	b**	b**	a**	a**	a**	a**	a**
Salvador	c**	b**	b**	a**	a**	a**	a**	a**
REGIÃO SUDESTE	c	b	b	a	a	a	a	a
Minas Gerais	c	b	b	a	a	a	a	a
Espírito Santo	-	-	b	a	a	a	a	a
Rio de Janeiro	c	b	b	a	a	a	a	a
São Paulo	c	b	b	a	a	a	a	a
Belo Horizonte	c**	b**	b**	a**	a**	a**	a**	a**
Rio de Janeiro (metrópole)	c**	b**	b**	a**	a**	a**	a**	a**
São Paulo (metrópole) ...	c**	b**	b**	a**	a**	a**	a**	a**
REGIÃO SUL	c	b	b	a	a	a	a	a
Paraná	c	b	b	a	a	a	a	a
Santa Catarina	-	-	b	a	a	a	a	a
Rio G. do Sul	c	b	b	a	a	a	a	a
Curitiba	c**	b**	b**	a**	a**	a**	a**	a**
Porto Alegre	c**	b**	b**	a**	a**	a**	a**	a**
REGIÃO CENTRO-OESTE	c	b	b	a	a	a	a	a
Mato Grosso do Sul	-	-	b	a	a	a	a	a
Mato Grosso	-	-	b	a	a	a	a	a
Goiás	c	b	b	a	a	a	a	a
Distrito Federal	c	b	b	a	a	a	a	a

(a) Freqüência e Valor do Rendimento Médio Mensal, por classes de rendimento mensal de todos os trabalhos, segundo sexo e domicílio (urbano e rural) para PEA e Pessoas de 10 anos ou mais; para as Pessoas Ocupadas, a classificação é apenas por sexo. Sem rendimento inclui as pessoas que receberam somente os benefícios.

(a*) Idem de (a), exceto estrato rural.

(a**) Idem de (a*), mas somente estrato metropolitano.

(b) (b*) (b**) Idem de a), (a*) e (a**), respectivamente, exceto o Valor do Rendimento Médio Mensal.

(c) Freqüência e Valor do Rendimento Médio Mensal, por classes de rendimento mensal de todos os trabalhos, segundo sexo e domicílio (urbano e rural) apenas para as pessoas de 10 anos ou mais. Sem rendimento inclui os empregados domésticos que receberam somente os benefícios.

(c*) Idem de (c), exceto estrato rural.

(c**) Idem de (c*), mas somente estrato metropolitano.

BIBLIOGRAFIA

ATKINSON, A. *The economics of inequality.* Oxford: Clarendon Press, 1983.

BONELLI, R.; SEDLACEK, G. Distribuição de renda: evolução no último quarto de século. In: MERCADO de trabalho e distribuição de renda: uma coletânea. Rio de Janeiro: IPEA/INPES, 1989.

———. *A evolução da distribuição de renda entre 1983 e 1988.* Rio de Janeiro: IPEA/INPES, Mimeo, fev. 1990.

BUSTOS, Oscar H. Outliers e robustez. *Revista Brasileira de Estatística*, v. 49, n. 191, p. 5–25, jan./jun. 1988.

COSTA, Ramonaval Augusto. *Distribuição da renda pessoal no Brasil em 1970: uma análise cross section da distribuição de renda por ocupação.* Rio de Janeiro: IBGE, 1977.

DENSLOW, D.; TYLER, W. *Perspectives on poverty and income inequality in Brazil: an analysis of the changes during the 1980's.* Washington, D.C.: The World Bank, 1983. Staff working paper, n. 601.

DEPARTMENT of Health and Human Services (USA). The 1989 federal poverty guidelines. *Social Security Bulletin*, v.52, n.3, mar. 1989.

FISHLOW, Albert. Brazilian size distribution of income. *American Economic Review*, p. 391-408, may 1972. Papers and proceedings of the eighty-fourth annual meeting.

HOFFMANN, Rodolfo. Evolução da distribuição de renda no Brasil entre pessoas e entre famílias, 1979/1986. In: MERCADO de trabalho e distribuição de renda: uma coletânea. Rio de Janeiro: IPEA/INPES, 1989.

LANGONI, Carlos G. *Distribuição de rendas e desenvolvimento econômico do Brasil: uma análise da década 1960-1970.* Rio de Janeiro: Expressão e Cultura, 1973.

O'HIGGINS, M.; SCHMAUS, G.; STEPHENSON, G. Income distribution and redistribution: a microdata analysis for seven countries. *Review of economic and Wealth*, v.35, n.2, p.107–131, june 1989.

REIS, J. G.; BARROS, R. P. *Um estudo da evolução das diferenças regionais da desigualdade no Brasil.* Rio de Janeiro: IPEA/INPES, 1989. Textos para discussão interna, n.178.

RICE, Randolph G. The after tax income distribution in the US during and inflationary decade: a look backwards at the 1970's. *Review of Social Economy.* v.47, n.3, 1989.

ROCHA, Sonia. Linhas de pobreza para as regiões metropolitanas na primeira metade da década de oitenta. In: ENCONTRO NACIONAL DE ECONOMIA, 16, 1988, Belo Horizonte. *Anais* .... Belo Horizonte : ANPEC, 1988. v.4, p.81-96

———. Indicadores de pobreza para as regiões metropolitanas nos anos oitenta. *Estudos econômicos*, v.20, n.3, p. 439–460, set./dez. 1990.

ROMÃO, Maurício Costa. *Distribuição de renda, pobreza e desigualdades regionais no Brasil.* Recife: PIMES, 1990. Texto para discussão, n.209.

ROSSI, José W. *Índices de desigualdade de renda e medidas de concentração industrial: aplicação a casos brasileiros.* Rio de Janeiro: Zahar, 1982.

RESUMO

Tendo em vista a persistência das desigualdades no Brasil e o interesse crescente na mensuração de sua evolução, examinam-se diferentes aspectos conceituais e metodológicos associados ao Índice de Gini, indicador a que se recorre mais freqüentemente no estudo desta questão.

Depois de tratar de questões conceituais como a de escolha da variável e da unidade de análise relevantes, passa-se aos aspectos empíricos que se colocam no Brasil quando se trata de medir a desigualdade de renda a partir das fontes estatísticas básicas – Censo Demográfico e PNADs, estabelecendo a distinção fundamental de procedimento de cálculo quando se trata de distribuições de rendimentos publicadas ou de acesso às informações desagregadas do banco de dados. Em cada caso, examina-se a sensibilidade de resultados obtidos a partir de diferentes escolhas conceituais e metodológicas adotadas (conceito de renda, unidade de análise, inclusão ou exclusão de renda zero, tratamento de *outliers*, procedimento de expansão, etc.).

As conclusões principais são duas: a primeira se refere à característica de inércia do coeficiente ao longo do tempo quando construído a partir de distribuições de renda comparáveis; a segunda é que, dado o grande número de possibilidades conceituais para a construção de índice — particularmente quando se tem acesso ao banco de dados — torna-se praticamente impossível estabelecer uma tendência evolutiva a partir de índices derivados de diferentes estudos.

ABSTRACT

Considering the persistence of inequality in Brazil and the growing interset on its evolution, this paper examines different conceptual and methodological aspects related to the Gini coefficient. Which is the most frequently used indicator of inequality.

After surveying questions related to the choice of the relevant variable and unit of analiysis, the paper focuses on empirical aspects of measuring inequality based on data from the 1980 Brazilian demographic census and annual household surveys. In particular, calculating procedures differ when applied to aggregated published data or data base. In each case the study examines the sensibility of results to different concepts and procedures concerning definition of income, choice of unit of analysis, treatment of outliers, sample expansion, etc.

The main conclusions are twofold. First, the Gini coefficient is relatively stable when built from comparable income distributions. Second, taking into account the large number of conceptual possibilities in building the coefficient, it is practically impossible to derive an evolution of inequality based on Gini coefficients calculated from different sources.

ESTIMADORES ROBUSTOS COMO REGRAS DE DETECÇÃO DE DADOS SURPREENDENTES NO MODELO DE REGRESSÃO LINEAR

Oscar Bustos*

1 INTRODUÇÃO

A análise de regressão pode ser considerada como um conjunto de técnicas destinadas ao estudo das complexas relações que poderiam existir entre as várias variáveis que conformam o modelo matemático que tenta descrever um certo fenômeno físico. Essa análise, na verdade, é um processo iterativo pelo qual a dupla experimentador-analista, começando com um certo modelo e um conjunto de hipóteses, vai modificando-as no percurso da análise das observações, até chegar a um modelo que se ajuste aos dados em forma satisfatória (mais ou menos subjetivamente). O seguinte seria um fluxograma (bem reduzido) desse processo iterativo:

1. Formular o modelo.
2. Ajustar o modelo aos dados (estimar os parâmetros).
3. Estudar a validade das hipóteses levantadas na construção do modelo (análise de resíduos, de pontos influentes e de pontos "desajustados").
4. *Tudo bem? Então seguir o* 5, *caso contrário voltar ao* 1.
5. Testes de bondade do ajuste.
6. Tudo bem? Então usar o modelo segundo os objetivos do estudo, caso contrário voltar ao 1.

*Pesquisador do Instituto de Matemática Pura e Aplicada - IMPA.

Neste trabalho estudam-se certos itens referentes ao passo 3 do esquema acima para o modelo de regressão linear.

Na Seção 2 está a definição do modelo a ser estudado. Na Seção 3 destaca-se o conjunto das hipóteses que habitualmente apóiam o uso dos estimadores de mínimos quadrados. Na Seção 4 está um resumo dos principais resultados do procedimento de estimação por mínimos quadrados. Na Seção 5 ver-se-á uma introdução às técnicas de detecção de "observações surpreendentes". A Seção 6 está dedicada às definições e principais propriedades das estatísticas clássicas baseadas em resíduos amostrais do ajuste por mínimos quadrados, e que são usadas nos testes para detectar uma "observação surpreendente". A seguir, na Seção 7, dá-se uma formalização dos conceitos de "desajustados", "pontos influentes" e "pontos de alavanca", aos quais podemos dar o nome genérico de "observações surpreendentes". A Seção 8 é uma breve introdução ao estudo dos testes baseados em estatísticas definidas na Seção 6. A Seção 9 contém uma breve explicação sobre o que se entende por "fenômeno de mascaramento" e suas conseqüências sobre os resultados dos testes vistos na Seção 8. Finalmente, na Seção 10, dá-se uma idéia sobre o uso de estimadores robustos para construir técnicas de diagnóstico "resistentes" ao "mascaramento".

Este trabalho está baseado em temas estudados num contexto muito mais geral e com maiores detalhes nos livros de Chatterjee e Hadi (1988) e Rousseeuw e Leroy (1987).

2 O MODELO

Seja

$$Y = X\beta + \varepsilon$$

onde:

$Y = (Y_1, \ldots, Y_n)'$ é um vetor $n \times 1$ que representa as observações ou a variável dependente;

$X = [x_{ij}]$ é uma matriz $n \times k$ de constantes conhecidas (variáveis explicadoras) com $k < n$; no caso do modelo com intercepto se terá $x_{11} = 1$, para todo $i = 1, \ldots, n$;

$\beta = (\beta_1, \ldots, \beta_k)$ é um vetor $k \times 1$ dos coeficientes de regressão a serem estimados;

$\varepsilon = (\varepsilon_i, \ldots, \varepsilon_n)$ é um vetor $n \times 1$ de erros aleatórios.

Isto é, vale:

$$E(\varepsilon_i) = 0 \quad \text{e} \quad \text{Var}(\varepsilon_i) = \sigma^2, \quad \text{para todo} \quad i = 1, \ldots, n.$$

NOTAÇÕES:

$Y = (Y_1, \ldots, Y_n)'$ é uma realização (genérica) do vetor aleatório Y.

$x_{i.} = (x_{i1}, \ldots, x_{ik})'$, $x_{.j} = (x_{1j}, \ldots, x_{nj})'$, $i = 1, \ldots, n$, $j = 1, \ldots, k$.

3 HIPÓTESES HABITUAIS SOBRE O MODELO DE REGRESSÃO LINEAR

Para simplificar, o conjunto destas hipóteses será denotado por $\mathcal{H}$.

$h1$. A matriz X é de posto completo (isto é, posto de $X = k$).

$h2$. X é medida sem erros.

$h3$. ε_i. não depende de x_i, para todo $i = 1, \ldots, n$.

$h4$. ε tem distribuição normal mutivariada com média 0 e matriz de covariância $\sigma^2 I_n$, onde I_n é a matriz identidade n-dimensional.

NOTA IMPORTANTE: A validade destas hipóteses deve ser testada antes de tirar conclusões de qualquer análise.

4 RESUMO DOS PRINCIPAIS RESULTADOS DO PROCEDIMENTO DE ESTIMAÇÃO POR MÍNIMOS QUADRADOS

O estimador de mínimos quadrados de β é o vetor aleatório $\hat{\beta}(MQ)$ tal que

$$(X'X)\hat{\beta}(MQ) = X'Y.$$

Suponhamos válida $\mathcal{H}$, então:

$$\hat{\beta}(MQ) - (X'X)^{-1}X'Y. \tag{4.1}$$

$$E(\hat{\beta}(MQ)) = \beta \tag{4.2}$$

isto é, $\hat{\beta}(MQ)$ é um estimador não viciado de β.

$\hat{\beta}(MQ)$ é ótimo para estimar β na classe dos estimadores lineares no vetor Y e não viciados, no sentido de possuir a variância mínima nessa classe. Além disso:

$$\text{Var}(\hat{\beta}(MQ)) = \sigma^2(X'X)^{-1}. \tag{4.3}$$

$\hat{\beta}(MQ)$ tem distribuição normal k-variada com média β e matriz de covariância $\sigma^2(X'X)^{-1}$.

O vetor (aleatório) dos valores ajustados, definido por:

$$\hat{Y} = PY, \tag{4.4}$$

onde:

$$P = X(X'X)^{-1}X', \tag{4.5}$$

tem as seguintes propriedades:

$$E(\hat{Y}) = X\beta. \tag{4.6}$$

$$\text{Var}(\hat{Y}) = \sigma^2 P. \tag{4.7}$$

$$\hat{Y} \quad \text{tem distribuição normal } n\text{-variada.} \tag{4.8}$$

O vetor (aleatório) dos resíduos amostrais por mínimos quadrados (residuais MQ), definido por:

$$\hat{\varepsilon} = Y - \hat{Y} = (I_n - P)Y, \tag{4.9}$$

tem as seguintes propriedades:

$$E(\hat{\varepsilon}) = 0. \tag{4.10}$$

$$\text{Matriz de covariância de } \hat{\varepsilon} = \sigma^2(I_n - P) \tag{4.11}$$

$$\hat{\varepsilon} \text{ tem distribuição normal } n\text{-variada.} \tag{4.12}$$

$$\hat{\varepsilon}'\hat{\varepsilon}/\sigma^2 \text{ tem distribuição chi-quadrado com } n-k \text{ graus de liberdade.} \tag{4.13}$$

Finalmente, um estimador não viciado de σ^2 é dado por:

$$\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/(n-k) = Y'(I_n - P)Y. \tag{4.14}$$

5 INTRODUÇÃO À DETECÇÃO DE "OBSERVAÇÕES SURPREENDENTES"

Mais na frente será dada uma definição mais formal do conceito de "observação surpreendente". No entanto, usa-se esse nome para denotar aqueles pontos suspeitos de não seguir o modelo sob as hipóteses $\mathcal{H}$.

Os resultados dum ajuste do modelo aos dados pelo método de mínimos quadrados podem ser muito alterados pela supressão (ou aditamento) de umas poucas observações. Habitualmente, nem todas as observações têm o mesmo impacto sobre os resultados duma análise. As que mais afetam esses resultados são aquelas que mais se afastam da maioria. Daí que seja muito importante para o analista ter uma idéia clara sobre a influência de tais pontos, e de possuir regras objetivas, definidas sem ambigüidades, para detectá-los.

Os métodos para detectar "observações surpreendentes" podem ser classificados em duas classes: "técnicas de supressão" e "técnicas de diferenciação". As primeiras examinam as mudanças que resultam numa análise de regressão pela supressão duma ou várias observações. As segundas examinam as derivadas com respeito aos parâmetros das densidades das estatísticas envolvidas nessa análise. Neste trabalho considera-se

apenas uma subclasse das "técnicas de supressão", simplesmente, as que estudam o que se passa quando se suprime uma única observação.

Segundo seja o item específico da análise sobre o qual será estudado o impacto da supressão duma observação só, as técnicas de supressão podem ser agrupadas nas seguintes categorias, cada uma das quais está caracterizada por certas estatísticas também chamadas "medidas de influência":

1) residuais (resíduos amostrais);
2) afastamento dos pontos no espaço (X, Y);
3) curva de influência;
4) volume dos elipsóides de confiança;
5) função de verossimilhança;
6) subconjuntos de coeficientes de regressão; e
7) autovalores da matriz X.

Neste artigo somente se estuda alguma coisa referente às técnicas do primeiro grupo.

6 MEDIDAS BASEADAS EM RESIDUAIS

É fácil ver que

$$\hat{\varepsilon} = (I_n - P)\varepsilon.$$

Daí que se os elementos fora da diagonal da matriz P são "pequenos" e "aproximadamente iguais" é razoável estudar a validade da hipótese $h4$ para $\hat{\varepsilon}$ como se esse vetor fosse ε.

O anterior também sugere o uso de transformações convenientes do vetor dos residuais para construir estatísticas de diagnóstico. Assim, em vez de usar os residuais diretamente, é comum usar uma transformação dada por

$$f(\hat{\varepsilon}_i, \sigma_i) = \hat{\varepsilon}_i/\sigma_i, \tag{6.1}$$

onde σ_i é o desvio padrão de $\hat{\varepsilon}_i$.

Ora, os σ_i's na prática não são conhecidos, devem ser estimados. Segundo seja o estimador do σ que se vai usar na expressão (6.1), temos as seguintes estatísticas:

i-ésimo residual normalizado

$$a_i := f(\hat{\varepsilon}_i, (\hat{\varepsilon}'\hat{\varepsilon})^{-1/2}) = \hat{\varepsilon}_i/(\hat{\varepsilon}'\hat{\varepsilon})^{-1/2}, \quad i = 1, \ldots, n. \tag{6.2}$$

(Nesta fórmula e nas que seguirão a notação $A := B$ significa que A está definido por B).

i-ésimo residual padronizado

Seja

$$\hat{\sigma} := ((\hat{\varepsilon}'\hat{\varepsilon})/(n-k))^{1/2}.$$

Chama-se i-ésimo residual padronizado ao

$$b_i := f(\hat{\varepsilon}_i, \hat{\sigma}) = \hat{\varepsilon}_i/\hat{\sigma}, \quad i = 1, \ldots, n. \tag{6.3}$$

i-ésimo residual estudentizado internamente

Sejam $p_{11}, \ldots, p_{nn}$ os elementos da diagonal da matriz P.

Chama-se i-ésimo residual estudentizado internamente ao

$$r_i := f(\hat{\varepsilon}_i, \hat{\sigma}(1-p_{ii})^{1/2}) = \hat{\varepsilon}_i/(\hat{\sigma}(1-p_{ii})^{1/2}), \quad i = 1, \ldots, n.$$

i-ésimo residual estudentizado externamente

Para cada $i = 1, \ldots, n$ sejam:

$$X_{-i} := \text{a matriz } X \text{ sem a linha } x_{i\cdot},$$

$$Y_{-i} := \text{o vetor } Y \text{ sem a componente } Y_i.$$

$$P_{-i} := X_{-i}(X'_{-i}X_{-i})^{-1}X'_{-i}, \quad \text{e}$$

$$\hat{\sigma}^2_{-i} (= (Y'_{-i}(I_{n-1} - P_{-i})Y_{-i})/(n-k-1).$$

Chama-se i-ésimo residual estudentizado externamente ao

$$r^*_i := f(\hat{\varepsilon}_i, \hat{\sigma}_{-i}(1-p_{ii})^{1/2}) = \hat{\varepsilon}_i/(\hat{\sigma}_{-i}(1-p_{ii})^{1/2}), \quad i = 1, \ldots, n.$$

Existem outras estatísticas úteis para detectar "observações surpreendentes" no modelo de regressão linear sob a hipótese $h4$, por exemplo a chamada "envolvente normal". O interessado pode ver, por exemplo, Bustos e Frery (1988).

Qual de todas as estatísticas já vistas convém usar? Para responder esta questão é bom conhecer as propriedades das mesmas. Para tanto é de utilidade considerar as seguintes fórmulas demonstradas, por exemplo, em Chatterjee e Hadi (1988).

Sejam:

$$SSE := Y'(I_n - P)Y \quad \text{(soma de quadrados dos residuais)},$$

$$SSE_{-i} := Y'_{-i}(I_{n-1} - P_{-i})Y_{-i}, \quad i = 1, \ldots, n.$$

Então:

$$SSE_{-i} = SSE - \hat{\varepsilon}^2_i/(1-p_{ii}), \quad i = 1, \ldots, n.$$

$$\hat{\sigma}^2_{-i} = \hat{\sigma}^2\left[\frac{n-k-r_i^2}{n-k-1}\right], \quad i = 1, \ldots, n.$$

$$b_i = a_i(n-k)^{1/2}, \quad i = 1, \ldots, n.$$

Outras fórmulas úteis:

$$r_i = b_i/(1-p_{ii})^{1/2} = a_i\left[\frac{n-k}{1-p_{ii}}\right]^{1/2}, \quad i = 1,\ldots,n.$$

$$r_i^* = \frac{a_i(n-k-1)^{1/2}}{[(1-p_{ii})-a_i^2]^{1/2}} = r_i\left[\frac{n-k-1}{n-k-r_i^2}\right]^{1/2}, \quad i = 1,\ldots,n.$$

$$a_i^2 \leq (1-p_{ii}), \quad i = 1,\ldots,n.$$

Usando as fórmulas acima é possível provar resultados de interesse como os seguintes:

PROPOSIÇÃO. *Se* $a_i^2 \to (1-p_{ii})$, *então* $r_i^2 \to n-k$ *e* $r_i^* \to +\infty$.

TEOREMA. *Seja* $posto(X) = k$. *Então:*

a) Se $\text{posto}(X_{-i}) = k$ e vale $h4$, então $r_i^2/(n-k)$ tem distribuição Beta $(1/2,(n-k-1)/2)$.
b) Se $\text{posto}(X_{-i}) = k-1$, então $\hat{\varepsilon}_i = 0$ e $\text{Var}(\hat{\varepsilon}_i) = 0$, logo r_i não está definido.
c) Se $\text{posto}(X_{-i}) = k$ e vale $h4$, então r_i^* tem distribuição t-Student com $n-k-1$ graus de liberdade.
d) Se $\text{posto}(X_{-i}) = k-1$, então r_i^* não está definido.

Para estudar a influência da i-ésima observação no modelo já definido com respeito ao modelo de "deslocamento na média", definem-se as seguintes hipóteses a serem testadas uma contra a outra:

$$H_o : E(Y) = X\beta, \quad H_1 : E(Y) = X\beta + \theta U_i,$$

onde $U_i = (U_{i1},\ldots,u_{in})'$ e $u_{ij} = \delta_{ij}$ (delta de Kronecker).

O teste mais comumente usado está baseado na estatística

$$F_{-i} = \frac{(SSE(H_o) - SSE(H_1))/1}{SSE(H_1)/(n-k-1)}.$$

Ora, uma conta direta demonstra que $F_{-i} = r_i^*$.

Usando os resultados acima podem se demonstrar alguns resultados de interesse para a análise de dados:

NOTA 1: Dificuldade para detectar pontos surpreendentes nos pontos com p_{ii} grande.

Com efeito, pode-se provar que sob H_1 a estatística F_{-i} tem distribuição F não central com $(1,n-k-1)$ graus de liberdade e parâmetro de não centralidade igual a $\theta^2(1-p_{ii})/\sigma^2$. Daí que se p_{ii} for muito próximo de 1, então as distribuições de F_{-i} sob H_o e sob H_1 seriam muito semelhantes.

NOTA 2: Algumas "dicas" sobre o uso das estatísticas r_i e r_i^*.

a) Para a maioria dos problemas de interesse prático a falta de independência de $r_1, \ldots, r_n$ e/ou de $r_1^*, \ldots, r_n^*$, não é preocupante e pode ser esquecida (ver Chatterjee e Hadi (1988)).
b) Nem r_i nem r_i^* refletem a variância de $\hat{\varepsilon}_i$.
c) Se os valores $p_{11}, \ldots, p_{nn}$ são muito dispersos, então Behnken e Draper (1972) sugerem usar r_i em lugar de r_i^*.
d) Muitos autores (ver por exemplo: Belsley e outros (1980), Atkinson (1981) e (1982), Velleman e Welsch (1981)) sugerem usar r_i^* em lugar de r_i pelas seguintes razões entre outras:

r_i^* é uma transformação monótona de r_i e, dado que $r_i^* \to +\infty$ quando $r_i^2 \to n-k$, r_i^* tende a refletir os desvios grandes de forma mais marcante que r_i.

$\hat{\sigma}_{-i}$ é um estimador mais resistente que $\hat{\sigma}$ ao efeito de erros grosseiros na i-ésima observação.

7 UMA FORMALIZAÇÃO DOS CONCEITOS DE "DESAJUSTADOS", "PONTOS INFLUENTES" E "PONTOS DE ALAVANCA"

Todos os conceitos que vão ser formalizados nesta seção o serão com respeito ao modelo definido na Seção 2 e sob a hipótese $\mathcal{H}$ definida na Seção 3.

"Desajustado": o ponto amostral (y_i, x_i) é um "desajustado" se r_i ou r_i^* é "grande" quando comparado com os outros residuais da mesma classe. (7.1)

"Ponto de alavanca": o ponto amostral $(y_i, x_{i.})$ é um "ponto de alavanca" se $x_{i.}$ está muito afastado dos restantes $x_{j.}$'s. (7.2)

"Ponto influente para calcular uma certa estatística T": seja T uma estatística denotada por T_n quando depende de $(Y_1, x_{1.}), \ldots, (Y_n, x_{n.})$ e por T_{-i} quando depende de $(Y_1, x_{1.}), \ldots, (Y_{i-1}, x_{i-1.}), (Y_{i+1}, x_{i+1.}), \ldots, (Y_n, x_{n.})$. O ponto amostral $(Y_i, x_{i.})$ é um "ponto influente para calcular T" se os valores amostrais de T_n e T_{-i} são "muito diferentes". (7.3)

É importante salientar que:
a) "Desajustados" não são necessariamente "pontos influentes".
b) "Pontos influentes" não são necessariamente "desajustados".
c) Observações com residuais pequenos podem estar bastante afastadas da maioria das observações.

Vejamos um exemplo para esclarecer os conceitos recentemente definidos.

Consideremos a seguinte figura e suponhamos as seguintes quatro situações:

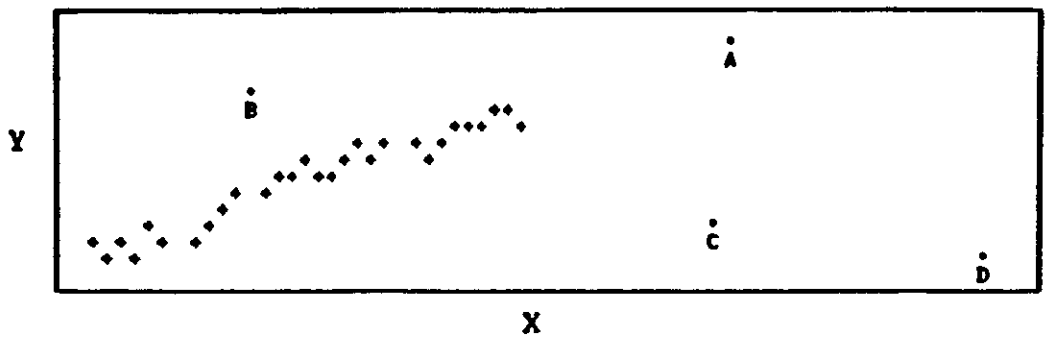


Figura 1

1) **Amostra: pontos "+" e A.**

A é ponto de alavanca, não é desajustado e não é ponto influente para calcular $\hat{\beta}(MQ)$, mas pode ser influente para estimar a $\mathrm{Var}(\hat{\beta}(MQ))$.

2) **Amostra: pontos "+" e b.**

B não é ponto de alavanca, é um desajustado, pode não ser influente para calcular $\hat{\beta}(MQ)$, pode ser influente para estimar o intercepto, para bb estimar σ, e daí para estimar as variâncias dos estimadores dos coeficientes de regressão.

3) **Amostra: pontos "+" e C.**

C é desajustado, ponto de alavanca e ponto influente para calcular $\hat{\beta}(MQ)$ e outros parâmetros.

4) **Amostra: pontos "+" e D.**

D não é desajustado (o ajuste por mínimos quadrados fará com que o residual correspondente ao D seja pequeno), é ponto de alavanca e é ponto influente para calcular $\hat{\beta}(MQ)$ e outros parâmetros.

8 INTRODUÇÃO AO ESTUDO DOS TESTES BASEADOS EM RESIDUAIS POR MÍNIMOS QUADRADOS

A hipótese $\mathcal{H}$ definida no Seção 3 é a "hipótese nula".

Para testar $\mathcal{H}$, os testes mais comumente usados na análise de regressão "clássica", estão baseados em certas estatísticas, digamos $T = t(Y_1, \ldots, Y_n)$, que são funções dos residuais que provêm dum ajuste por mínimos quadrados, tais como os definidos na Seção 6. Para cada $0 < \alpha < 1$, seja $C(\alpha)$ o ponto crítico do teste definido por T ao nível α, isto é, vale:

$$P_{\mathcal{H}}(T > C(\alpha)) \leq \alpha.$$

Assim, o teste baseado em T rejeita $\mathcal{H}$ quando $t(y_1, \ldots, y_n) > C(\alpha)$.

Um teste muito popular é o sugerido por Tietjen e outros (1973) que está definido por

$$t(y_1, \ldots, y_n) := r_{\max} := \max_i |r_i|.$$

Neste caso, ao igual que em muitos outros, é bem difícil determinar os valores $C(\alpha)$. O interessado nestas questões pode ver mais detalhes no já citado trabalho de Tietjen e outros (1973).

Diversos testes desta classe podem-se ver, por exemplo, nas seguintes publicações: Anscombe (1961), Theil (1965), Cox e Snell (1968, 1971), Andrews (1971), Stefansky (1971, 1972), Joshi (1975), Rosner (1975), Farebrother (1976a, 1976b), Doornbos (1981) e Draper e Smith (1981).

Além desses testes formais, existem outros procedimentos menos rigorosos mas também muito úteis e populares para detecção de observações surpreendentes, e que estão baseados em gráficos. Esses desenhos servem não apenas para essa detecção mas também para ter uma idéia global sobre o ajuste do modelo hipotético aos dados.

Alguns dos gráficos mais comumente usados são:

a) Para se ter uma visão da distribuição empírica dos residuais: histogramas, "ramos-e-folhas", "box-plots", etc.

b) Gráficos de residuais vs. índices das observações.

c) "Normal plots" ou "$Q - Q$ plots".

d) Gráficos de residuais vs. valores ajustados.

e) Gráficos de residuais vs. $x_{.j}$ para $j = 1, \ldots, k$.

É habitual usar nesses gráficos como residuais os valores r_i's definidos na Seção 6, porém Atkinson (1981) sugere usar r_i^*'s em lugar dos r_i's.

Esses (e outros) gráficos podem servir para detectar certas violações à hipótese $\mathcal{H}$. Veja por exemplo a Figura 2.

Mais detalhes sobre este tema dos gráficos podem-se ver em Chambers e outros (1983) e Cleveland (1985).

Uma advertência: os gráficos baseados na distribuição empírica dos residuais são significativos somente quando o número de observações é grande.

Os testes e gráficos desta seção devem ser usados cuidadosamente não apenas pelo dito na Nota 1 da Seção 6, mas também porque eles, em geral, não são resistentes ao fenômeno de "mascaramento", como se vê mais na frente. Daí que sejam necessárias estatísticas definidas especificamente para detectar desajustados, ou pontos influentes ou pontos de alavanca. Quase todas elas formam parte das categorias 2 a 7 citadas na Seção 5.

O interessado pode ver, por exemplo, Chatterjee e Hadi (1988).

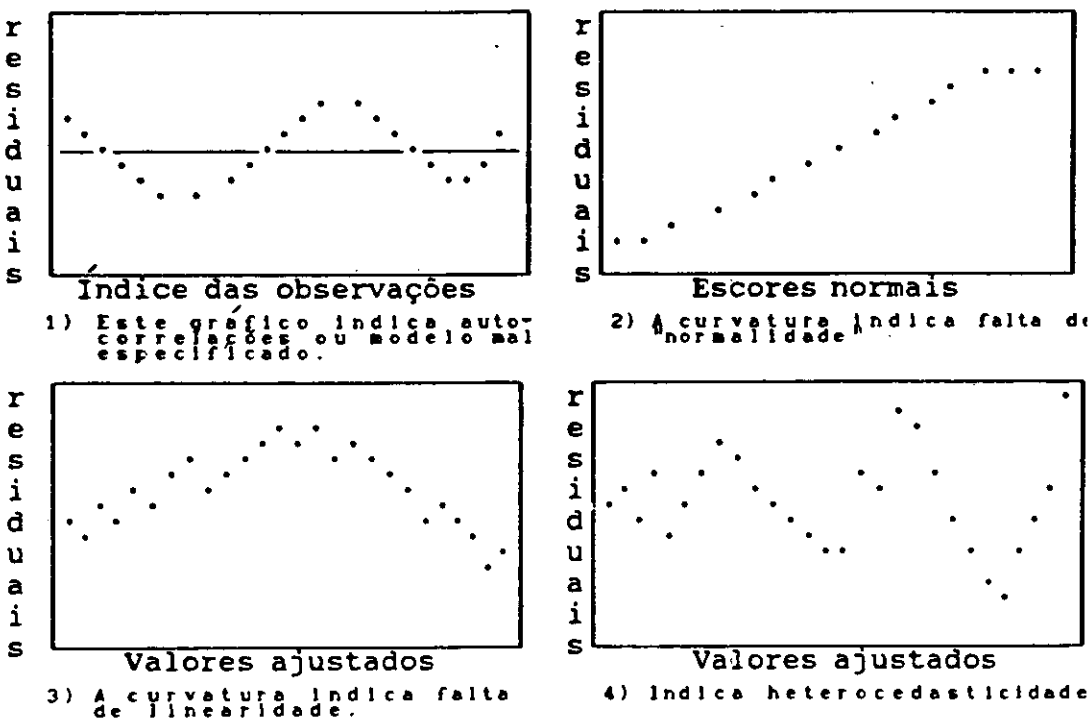


Figura 2

9 O "FENÔMENO DE MASCARAMENTO" E SUAS CONSEQÜÊNCIAS

Até agora vimos somente técnicas que servem para detectar apenas uma observação surpreendente. Naturalmente aparece a questão: existem técnicas para detectar múltiplas observações surpreendentes? Este problema é importante tanto do ponto de vista teórico quanto do ponto de vista prático.

Respeito da teoria: existem situações nas quais várias observações são influentes quando consideradas em conjunto, mas nenhuma delas o é quando considerada individualmente. Por exemplo, na situação da Figura 3.

Amostra: pontos "+" e 1 e 2. Os pontos 1 e 2 são conjuntamente influentes, mas nenhum dos dois o é quando considerado isoladamente.

Situações similares a esta têm recebido o nome de "efeito de mascaramento" devido a que a influência de uma só observação é escondida pela outra (ou outras).

Também pode acontecer a situação dual – consideremos a Figura 4.

Amostra: pontos "+" e 3. O ponto 3 é influente.

Amostra: pontos "+" e 3 e 4. Os pontos 3 e 4 não são **conjuntamente** influentes.

Respeito da prática: as conseqüências sobre os resultados da análise de regressão clássica são, geralmente, muito mais graves no caso de múltiplas observações sur-

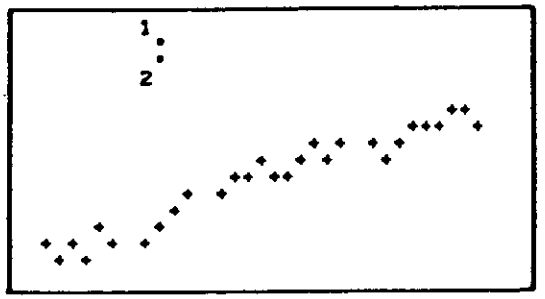


Figura 3

preendentes do que nos casos de apenas uma. Por outra parte, é mais difícil detectar conjuntos de várias observações surpreendentes. Finalmente, nas aplicações em casos reais, é freqüente o analista não dar importância a esta questão, o que é obviamente desaconselhável.

Existem técnicas para a detecção de múltiplas observações surpreendentes. Ver, por exemplo, Barnett e Lewis (1984), Hawkins e outros (1984) e o já citado Chatterjee e Hadi (1988).

Um procedimento natural consiste na aplicação iterada duma técnica de detecção de "uma observação por vez" (Dybczynski, 1980); mas este procedimento é obviamente afetado pelo mascaramento. Uma extensa análise Monte-Carlo do comportamento desta classe de técnicas pode ver-se em Caetano (1989).

Na verdade, todas estas técnicas "clássicas" de detecção de observações surpreendentes sofrem do mesmo problema: baseiam-se nos residuais que provêm dum ajuste pelo método de mínimos quadrados. Mas estes podem conduzir a resultados **irrelevantes** devidos a que esses estimadores são muito sensíveis à influência de uma ou várias observações afastadas da massa de dados.

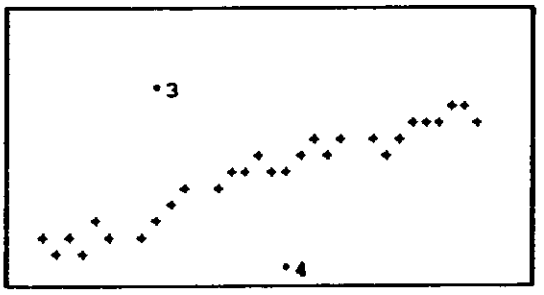


Figura 4

10 ESTIMADORES ROBUSTOS E TÉCNICAS DE DIAGNÓSTICOS "RESISTENTES" AO "MASCARAMENTO"

Tentando formalizar e medir a "sensibilidade" e "resistência" de diversas técnicas estatísticas, a Teoria da Robustez em Estatística tem definido diversas alternativas do chamado "ponto de ruptura" duma técnica sob um certo modelo. Por exemplo, no modelo sob estudo, seja T um estimador de β baseado em amostras de tamanho n; trata-se de saber qual seja a proporção de observações surpreendentes que pode ter a amostra além da qual o estimador T não dá informação útil sobre o β. O interessado pode consultar detalhes em: Hampel e outros (1986), Rousseeuw e Leroy (1987), Bustos (1988).

Assim, o ponto de ruptura do $\hat{\beta}(MQ)$ é 0%. Isto é, basta uma observação só suficientemente afastada da maioria para arruinar completamente o estimador $\hat{\beta}(MQ)$. Com efeito, vejamos os dois gráficos seguintes:

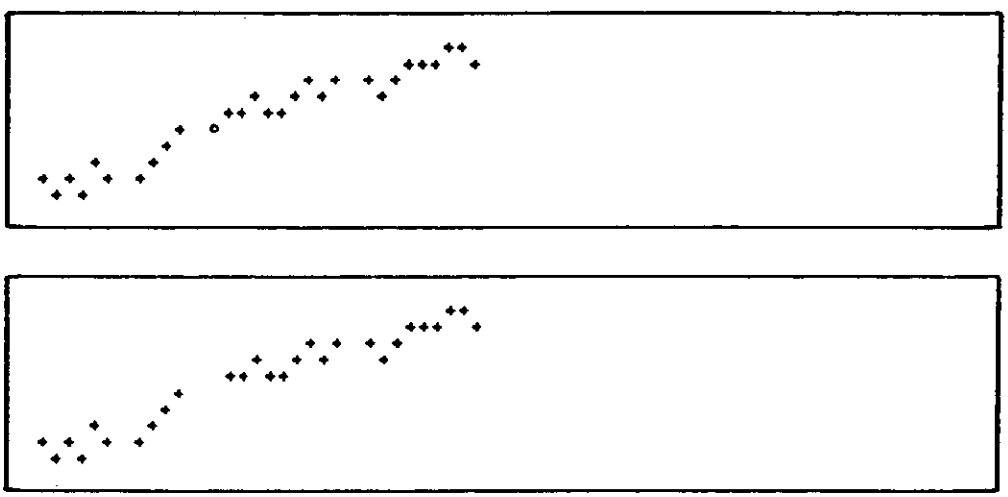


Figura 5

Por outra parte, o ponto de ruptura do estimador "mediana amostral" no modelo

$$Y_i = \mu + \varepsilon_i, \qquad i = 1, \ldots, n,$$

é de 50% (quase). Quer dizer, a metade menos uma das observações poderia estar bem longe da massa dos dados e a mediana amostral daria o mesmo resultado que se todas as observações fossem "boas".

Vários trabalhos estão sendo realizados tentando obter estimadores do β no modelo de regressão linear que tenham alto ponto de ruptura (aproximadamente $\{[n/2] - 1)/n\}^*$ 100%). Rousseeuw e Leroy (1987) têm uma boa resenha sobre o

assunto. Um exemplo dum tal estimador notável pela sua simplicidade intuitiva é o chamado "estimador de mínima mediana dos resíduos ao quadrado", $\hat{\beta}(LMS)$, definido em Rousseeuw (1984) pelo $\hat{\beta}(LMS)$ que minimiza a função de $\hat{\beta}$ dada por:

$$\text{Mediana}(\hat{r}_1^2, \ldots, \hat{r}_n^2),$$

onde $\hat{r}_1^2 := (Y_i - x_i'\hat{\beta})^2$, para $i = 1, \ldots, n$.

O ponto de ruptura desse estimador é o mais alto possível (e não depende do número de parâmetros), mas... tem muito pouca eficiência sob o modelo hipotético ($\mathcal{H}$), cuja bondade não rejeitamos, apenas duvidamos. Isto poderia fazer com que os intervalos de confiança construídos usando $\hat{\beta}(LMS)$ sejam muito compridos.

Na Estatística Robusta para o modelo de regressão linear vem se realizando esforços notáveis para obter estimadores de β que sejam eficientes sob o modelo hipotético e, no entanto, tenham alto ponto de ruptura. Com respeito a isto, é merecedor de destaque especial o τ-estimador sugerido por Yohai e Zamar (1988).

O uso de estimadores robustos é cada vez maior não apenas em trabalhos teóricos, mas também em muitos casos com dados reais.

No contexto que estamos considerando neste trabalho, vem se sugerindo definir regras de detecção de observações surpreendentes, baseadas em residuais que provêm dum ajuste por um estimador robusto de β. A idéia que fundamenta este procedimento baseia-se no fato empírico observado em diversos conjuntos de dados simulados e reais, de que nas situações com amostras apresentando várias observações surpreendentes, os estimadores robustos tendem a ser "resistentes" ao fenômeno de mascaramento.

Vários trabalhos aparecidos na década de 80 mostram a conveniência dessa técnica. Entre outros: Cook e Beckman (1980), Iglewicz e Martínez (1982), Atkinson (1986), Rousseeuw e Leroy (1987).

Para se ter uma idéia da utilidade do uso de estimadores robustos para detectar observações surpreendentes, vejamos o exemplo a seguir, tirado do já mencionado Rousseeuw e Leroy (1987).

Exemplo do uso dum estimador robusto para detectar observações surpreendentes.

No Belgium Statistical Survey (publicado pelo Ministério da Economia da Bélgica) podemos encontrar um conjunto de dados contendo o número total de chamadas internacionais de telefone. As variáveis são: X = ano, Y = número de chamadas (escala: 1=10 milhões).

O gráfico Y vs X é o seguinte:

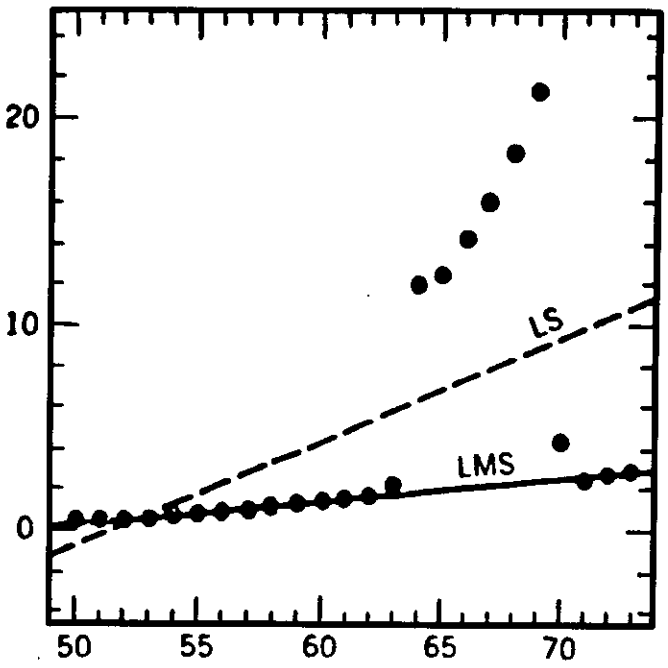


Figura 6

A tabela com esses dados é:

TABELA 1

X	Y	X	Y
50	0.44	62	1.61
51	0.47	63	2.12
52	0.47	64	11.90
53	0.59	65	12.40
54	0.66	66	14.20
55	0.73	67	15.90
56	0.81	68	18.20
57	0.88	69	21.20
58	1.06	70	4.30
59	1.20	71	2.40
60	1.35	72	2.70
61	1.49	73	2.90

O gráfico parece mostrar uma tendência ao crescimento de Y ao longo dos anos. No entanto, esses dados contêm 6 ou 7 pontos surpreendentes (para $X = 64$ até 69 ou quiçá 70).

O modelo que se tentou ajustar foi:

$$Y_1 = \alpha + X_i\beta + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{10.1}$$

Os estimadores de mínimos quadrados de α e β são:

$$\hat{\alpha}(MQ) = -26.01 \quad \text{e} \quad \hat{\beta}(MQ) = 0.504,$$

e a reta $y = \hat{\alpha}(MQ) + \hat{\beta}(MQ)x$. corresponde à denotada por LS no gráfico.

Vemos que essa linha foi muito afetada pelos valores de Y correspondentes aos anos 1964 até 1969. Daí que essa reta não ajusta legal nem os dados "bons" nem os "surpreendentes". Isto seria o que a gente teria obtido ao fazer uma aplicação cega do método de mínimos quadrados. Observemos que algumas boas observações tais como a correspondente ao ano de 1972 têm um residual por mínimos quadrados bem maior que os correspondentes a certos dados suspeitos.

Se aplicarmos agora o método do LMS obtemos:

$$\hat{\alpha}(LMS) = 5.61 \quad \text{e} \quad \hat{\beta}(LMS) = 0.115,$$

e a reta $y = \hat{\alpha}(LMS) + \hat{\beta}(LMS)xq$ corresponde à denotada por LMS no gráfico que claramente se ajusta bem à maioria dos dados "bons".

Ora, o que acabamos de ver não quer dizer que num caso semelhante a este seja bom um ajuste pelo modelo 10.1, porque a coleta dum número maior de casos poderia revelar um modelo diferente do 10.1. Porém neste exemplo particular, uma indagação posterior revelou que nalgum momento do ano de 1963 o método de registro foi mudado anotando não o número de chamadas mas o número total de **minutos** dessas chamadas, mudando novamente para a modalidade anterior nalgum momento do ano de 1970.

É claro que poderíamos ter usado qualquer outro estimador robusto que não o LMS.

O exemplo que foi analisado corresponde ao dum modelo de regressão linear simples. No modelo de regressão linear múltipla o mais complicado parece que é o cálculo do estimador LMS. Ora, já existe um programa, chamado PROGRESS, que calcula esse estimador para o modelo de regressão linear múltipla. Na verdade, o livro de Rousseeuw e Leroy (1987) é excelente para estudar essa metodologia de ajuste usando estimadores robustos, em particular o PROGRESS, todavia limitada ao estimador LMS tanto na teoria como nos cálculos e nos exemplos. Mesmo assim sua leitura é obrigatória para o interessado na robustez e na análise de dados.

11 ADVERTÊNCIA FINAL

Neste trabalho temos visto somente um aspecto do problema levantado por amostras com dados "surpreendentes"; simplesmente, algumas técnicas clássicas e resistentes que poderiam servir para detectar tais dados. Fica ainda o problema de o que fazer com esses dados; por exemplo: será conveniente retirá-los e fazer um novo ajuste por mínimos quadrados com os dados que restarem? ou, antes de proceder a um novo ajuste, ponderar as observações dando um peso menor às que resultarem detectadas como "surpreendentes"? ou usar estimadores robustos sobre toda a amostra como estimadores finais? Essas questões são obviamente de grande interesse na Estatística e são temas de numerosos trabalhos que vem aparecendo na literatura nos últimos anos.

BIBLIOGRAFIA

ANDREWS, D.F. Significance tests based on residuals. *Biometrika*, v.58, p.139–148, 1971.

ANSCOMBE, F.J. Examination of residuals. In: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 4, 1961, Berkeley. Proceedings... Berkeley: Univ. of California, 1961, v.1, p.1–36.

ATKINSON, A.C. Two graphical displays for outlying and influential observations in regression. *Biometrika*, v.68, p.13–20, 1981.

——— . Regression diagnostics, transformations and constructed variables. *Journal Royal Statistical Society S. B*, v.44, n.1, p.1–36, 1982.

——— . Masking unmasked. *Biometrika*, v. 73, p. 533–541, 1986.

BARNETT, V.; LEWIS, T. *Outliers in statistical data.* New York : J. Wiley, 1984.

BEHNKEN, E. W.; DRAPER, N. R. Residuals and their variances patterns. *Technometrics*, v.14, n.1, p.101–111, 1972.

BELSLEY, D. A.; KUH, E.; WELSCH, R. E. *Regression diagnostics: identifying influential data and sources of collinearity.* [S.L.] : Wiley, 1980.

BUSTOS, O. Outliers e robustez I. *Revista Brasileira de Estatística*, Rio de Janeiro, v.49, n.191, p.5–25; jan./jun. 1988.

——— . Outliers e robustez II. *Revista Brasileira de Estatística*, Rio de Janeiro, v.49, n.192, p.7–30 jul./dez. 1988.

——— . Outliers e robustez III. *Revista Brasileira de Estatística*, Rio de Janeiro, v.50, n.193, p.7–35 jan./jun. 1989.

——— . FRERY, A. *CNORT una subrutina para generación de cápsulas normales usando el IMSL.* Rio de Janeiro: Instituto de Matemática Pura e Aplicada, 1988. (Informes de Matemática, B-046)

CAETANO, E. *Estimadores LS, DRLS e τ no modelo de regressão linear; estudo comparativo por simulação.* Campinas: UNICAMP, 1989. Dissertação (Mestrado em Estatística) UNICAMP, 1989.

CHAMBERS, J. M. et al. *Graphical methods for data analysis.* [S.L.] : Duxbury Press, 1983.

CHATTERJEE, S. *Sensitivity analysis in linear regression.* New York : Wiley, 1988.

CLEVELAND, W. S. *The elements of graphing data.* Wadsworth, 1985.

COOK, R. Q., BECKMAN, R. J. *Using M-estimators to identify outliers.* Los Alamos Scientific Laboratoy, Univ. of California, 1980 (Technical report LA-UR; 80-1418).

COX, D. R.; SNELL, E. J. A general definition of residuals. *Journal Royal Statistical Society S. B*, v.30, n.2, p.248–275, 1968.

——— . On test statistics calculated from residuals. *Biometrika*, v.58, p.589–594, 1968.

DOORNBOS, R. Testing for a single outlier in a linear model. *Biometrics*, v.37, p.705–711, 1981.

DRAPER, N. R.; SMITH, H. *Applied regression analysis.* New York : Wiley, 1981.

DYBCZYNSKI, R. Comparison of the efectiveness of various procedures for the rejection of outlying results and assigning consensus values in interlaboratory programs involving determination of trace elements or radionuclides. *Analytica Chimica Acta*, v.117, p.53–70, 1980.

FAREBROTHER, R. W. BLUS residuals, algorithm AS 104. *Applied Statistics*, v.25, p.317–319, 1976a.

———— . Recursive residual – a remark on algorithm AS 7: basic procedures for large space or weighted least squares problems. *Applied Statistics*, v.25, p.323–324, 1976b.

HAMPEL, F.R. et al. *Robust statistics: the approach based on influence functions.* New York : Wiley, 1986.

HAWKINS, D. M.; BRADU, D.; KASS, G. V. Location of several outliers in multiple regression data using elemental sets. *Technometrics*, v.28, n.3, p.197–208, 1984.

IGLEWICZ, B.; MART'INEZ, J. *Outlier detection using robust measures of scale.* Filadelphia: Temple Univ. Philadelphia, Dep. of Statistics, 1982. (Technical Report 12)

JOSHI, P. C. Some distribution theory results for a regression model. *Annals of the Institute of Statistical Mathematics*, v.27, p.309–317, 1975. 1975.

PROGRESS. A program for robust regression analysis based on the Least Median of Squares (1986). Ver Rousseeuw e Leroy (1987).

ROSNER, B. On the detection of many outliers. *Technometrics*, v.17, p.217–227, 1975.

ROUSSEEUW, P. J. Least median of squares regression. *Journal of the American Statistical Association*, v.79, n.388, p.871–880, 1984.

———— ; LEROY, A. *Robust regression and outlier detection.* New York : Wiley, 1987.

STEFANSKY, W. Rejecting outliers by maximum normed residual. *Annals Mathematical Statistics*, v.42, n.1, p.35–45, 1971.

———— . Rejecting outliers in factorial designs. *Econometrics*, v.14, p.469–479, 1972.

THEIL, H. *The analysis of disturbances in regression analysis. Journal American Statistical Association*, v.60, n.312, p. 1067–1079, 1965.

TIETJEN, G. L.; MOORE, R. H.; BECKMAN, R. J. Testing for a single outlier in simple linear regression. *Technometrics*, v.15, n.4, p.717–721, 1973.

VELLEMAN, P. F.; WELSCH, R. E. Efficient computing of regression diagnostics. *The American Statistician*, v.35, n.4, p. 234–242, 1981.

YOHAI, V.; ZAMAR, R. High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, v.83, n.402, p.406–413, 1988.

RESUMO

Neste trabalho trata–se de esclarecer o significado de termos que têm resistido a uma formalização matemática precisa, gerando um pouco de confusão nas análises de dados que supostamente são ajustados por um modelo de regressão. Alguns desses termos são: "outliers", "observações aberrantes", "valores extremos", "pontos influentes", etc. Se bem que é impossível tirar toda subjetividade por parte do analista quando ele declara que um (ou vários) ponto amostral é um dado "surpreendente", as técnicas de detecção de tais pontos podem ser vistas como um intento de formalizar matematicamente essa surpresa. Aqui, "ponto surpreendente" denota qualquer um dos pontos afastados da maioria, num sentido a precisar.

Revisam–se as técnicas dessa natureza mais comumente usadas. Todas elas possuem uma característica notável e paradoxal: estão baseadas no estimador de mínimos quadrados dos coeficientes do modelo suposto. Mas já é sabido que esse estimador pode ser muito pouco representativo quando existem pontos desse tipo que essas técnicas pretendem detectar. Daí que seja bem interessante saber se ditas regras são boas mesmo na presença desses pontos "ruins". Alguns exemplos muito simples mostram que a resposta é não: as regras de detecção de "surpreendentes" baseadas no estimador de mínimos quadrados podem sofrer as conseqüências dum fenômeno que recebe o nome de "mascaramento"... alguma coisa assim como se os "surpreendentes" se ocultassem entre si para não serem detectados. E aí está o grave: as técnicas clássicas da análise de regressão podem dar resultados catastróficos quando existem esses pontos na amostra.

Finalmente, mostra–se como uma metodologia aparentemente contraposta à filosofia do "detectar para rejeitar", que tem recebido o nome de Inferência Robusta, pode ser usada para driblar os defeitos das habituais "regras de detecção de surpreendentes".

É de destacar que, numa conversa particular, o Professor Wilton Bussab (da Universidade de São Paulo) sugeriu ao autor deste trabalho trocar a palavra inglesa "outlier" pela "desajustado", o que se faz em todo este artigo.

ABSTRACT

This article formalizes the significance of certain concepts such as "outliers", "spurious observations", "extreme values", and "influential points", concepts that remained reluctant to a precise mathematical definition. The most common technics based in least squares estimates of the coefficients of the assumed model for detection of such points are reviewed. Its shortcomings are commented, for example: these technics may not detect "masking points". This is serious because the standard technics of regression analysis could draw catastrophic conclusions in the presence of the above mentioned points. Finally, it is shown that some technics of Robust Inference can be useful to avoid such problems.

IDENTIFICAÇÃO ESTATÍSTICA DE GRUPOS NA ASSEMBLÉIA NACIONAL CONSTITUINTE

José Francisco Soares*

1 INTRODUÇÃO

O jornal *Folha de São Paulo* publicou em outubro de 1988, para 26 votações da Assembléia Nacional Constituinte, o voto de 40 constituintes, considerados como os de maior destaque pela sua editoria de política. Como o voto é expresso por uma das posições: Sim , Não, Abstenção ou Ausência, a tabela publicada é uma matriz de 40 linhas por 26 colunas. Cada linha retrata as opções de um constituinte; cada coluna contém o resultado de uma votação.Os dados originais estão reproduzidos no apêndice.

A publicação destes dados, é razoável supor, pretendia dar ao leitor elementos para julgamento dos constituintes por parte de seus eleitores.

Circularam na imprensa dois tipos de análises destes dados. A primeira hierarquizava os constituintes de acordo com seu grau de presença nas votações e a segunda discriminava-os de acordo com seus votos em algumas questões julgadas fundamentais por um grupo de pressão específico. Neste último caso, ganhou muita divulgação a classificação do Departamento Intersindical de Acompanhamento Parlamentar – DIAP.

Estes mesmos dados têm uma riqueza muito maior. Permitem descobrir como os constituintes, no geral de seu comportamento, se agrupam. Esta é a questão que será

*Professor do Departamento de Estatística – UFMG.

analisada neste artigo. Trata-se de um exemplo de uma situação onde o conhecimento estatístico permite a detecção e explicitação de estrutura básica presente nos dados e portanto um melhor conhecimento deles.

2 COEFICIENTES DE CONCORDÂNCIA E DISCORDÂNCIA

Para se definir a forma de se captar os inter-relacionamentos entre as posições dos constituintes, é ilustrativo considerar-se primeiramente um caso particular. Considere-se pois, os constituintes Afif Domingos e Amaral Neto. A Tabela 1 mostra nas 26 votações consideradas, como foi a relação entre os seus votos.

TABELA 1

Votação dos Constituintes Afif Domingos e Amaral Neto

Afif Domingos	Amaral Neto				
	Não	Sim	Ausente	Abstencão	Total
Não	6	2	0	0	8
Sim	0	5	0	0	5
Ausente .	6	4	0	1	11
Abstencão	1	1	0	0	2
Total ...	13	12	0	1	26

Toda informação disponível sobre a associação entre as posições dos dois constituintes está sintetizada neste tipo de tabela. Uma situação de discordância absoluta consiste na concentração dos votos fora da diagonal principal da tabela.

A necessidade de se analisar um grande número destas tabelas para a construção de um quadro geral descritivo das associações entre os constituintes impede, entretanto, sua utilização. Surge daí a necessidade de se trabalhar com uma medida de Discordância/Concordância, para o que muitas opções podem ser pensadas.

É intuitivamente claro que as discordâncias do tipo (Sim, Não) e (Não, Sim) traduzem uma discordância maior entre os constituintes e portanto devem ser mais valorizadas que discordâncias do tipo (Ausente, Sim) ou (Abstenção, Não).

Assim sendo e optando por medir discordâncias, é natural pensar em medidas do tipo

$$\sum_{i \neq j} a_{ij} f Y_{ij}$$

ou seja, a soma ponderada do número de votações – Y_{ij} – em que os dois constituintes considerados votaram de modo diferente.

A escolha dos pesos $[a_{ij}]$ deve levar em consideração fatores substantivos. Por

exemplo, é preciso decidir se existe discordância de fato entre dois constituintes quando um compareceu e votou a favor de uma proposição e um outro esteve ausente.

No que se segue usamos o peso 1 para os pares (Sim, Não) e (Não, Sim), 1/2 para os pares que envolvem abstenções e 0 para os que envolvem ausência. Com esta escolha estamos assumindo, implicitamente, que a ausência de um constituinte em uma dada votação impossibilita captar a sua opinião e que a abstenção induz a uma discordância fraca. Com estas escolhas o maior valor possível para o coeficiente de discordância será de 25. Isto ocorrerá quando os dois constituintes comparecem a todas as votações e votarem de formas diferentes.

A Tabela 2 apresenta os valores do coeficiente para os constituintes que se candidataram à Presidência da República (Lula, Freire, Covas e Afif) e dois outros: Brandão Monteiro e Amaral Neto com posições próximas às de outros dois candidatos a presidente, Brizola e Maluf, respectivamente.

TABELA 2

Coeficiente de Discordância entre alguns Constituintes

Constituintes	Lula da Silva	Brandão Monteiro	Roberto Freire	Mário Covas	Afif Domingos	Amaral Neto
Lula da Silva . . .	-	1.0	2.0	5.0	7.0	16.0
Brandão Monteiro	1.0	-	3.0	6.0	8.0	14.0
Roberto Freire . .	2.0	3.0	-	5.0	9.0	18.0
Mário Covas	5.0	6.0	5.0	-	6.0	12.0
Afif Domingos . .	7.0	8.0	9.0	6.0	-	3.0
Amaral Neto . . .	16.0	14.0	18.0	12.0	3.0	-

O quadro é ilustrativo. O coeficiente proposto mede com fidedignidade o grau de discordância existente entre as posições dos vários candidatos, conforme indicam as análises substantivas que circularam à epoca das eleições presidenciais. Isto legitima o seu uso para o conjunto dos 40 constituintes, cujo comportamento global é o objeto de nosso estudo.

3 SÍNTESE DOS INTER-RELACIONAMENTOS

O coeficiente apresentado na seção anterior quantifica o grau de discordância entre cada par de constituintes. Os valores obtidos variam desde discordância zero entre os vários pares de constituintes como Afif Domingos e Francisco Dornelles; Artur da Távola e Egídio Ferreira Lima; Bernardo Cabral e José Fogaça até discordância 21, a maior observada, entre Haroldo Lima e Luiz Eduardo Magalhães.

Nesta seção o objetivo é, através de sínteses gráficas da matriz de inter-relacionamento, captar o comportamento geral dos constituintes. Noutras palavras, expor as aproximações existentes entre as posições dos constituintes. Duas técnicas estatísticas são usadas para isto: Análise de Conglomerados e Escalas Multidimensionais.

3.1 Construção de Conglomerados

Neste caso o problema a ser resolvido é a identificação de um conjunto de grupos de constituintes de tal forma que, dentro dos grupos, a discordância seja a menor possível e, entre os grupos, a discordância seja a maior possível.

Uma solução simplista para este problema seria examinar todos os possíveis conjuntos de grupos de constituintes e escolher o que satisfizesse algum critério de otimalidade. Embora sejam conhecidos o número total destes conjuntos e algoritimos computacionais para sua geração, é impossível para um problema do tamanho do nosso, mesmo com o mais rápido dos computadores existentes, examinar todas as opções.

Assim sendo, o que se faz na pratica é usar um critério de geração de conglomerados que produza uma solução razoável sem, entretanto, examinar todas as configurações possíveis.

Uma das soluções mais conhecidas e disponível nos *softwares* estatísticos é o algoritimo hierárquico de aglomeração. Tal algoritimo começa colocando todos os constituintes em grupos distintos e, sucessivamente, vai agrupando aqueles que em cada passo são os mais similares até que na etapa final todos estejam agregados em apenas um grupo. Considerações substantivas indicam o momento de parada neste processo.

A definição de uma forma de se medir a distância entre diferentes conglomerados é essencial nesta solução hierárquica. Existem várias opções. No que se segue usamos a média das distâncias entre os elementos dos dois conglomerados. O *software* utilizado foi o SYSTAT, procedimento CLUSTER.

O resultado aparece na Figura 1, onde os constituintes são identificados pelos números associados a seus nomes no apêndice. Pode-se distinguir dois grandes grupos de constituintes organizados ao longo de um eixo que pode ser interpretado como ideológico. Uma divisão mais fina indicaria a presença de quatro grupos. Os constituintes 17: Gastone Righi e 39: Sandra Cavalcanti têm uma posição singular. Só se agregam a algum grupo quando o grau de disparidade aceitável dentro do grupo é bastante grande para o problema. Este comportamento sugere que sintetizar todos os inter-relacionamentos em apenas uma dimensão é uma restrição drástica neste caso.

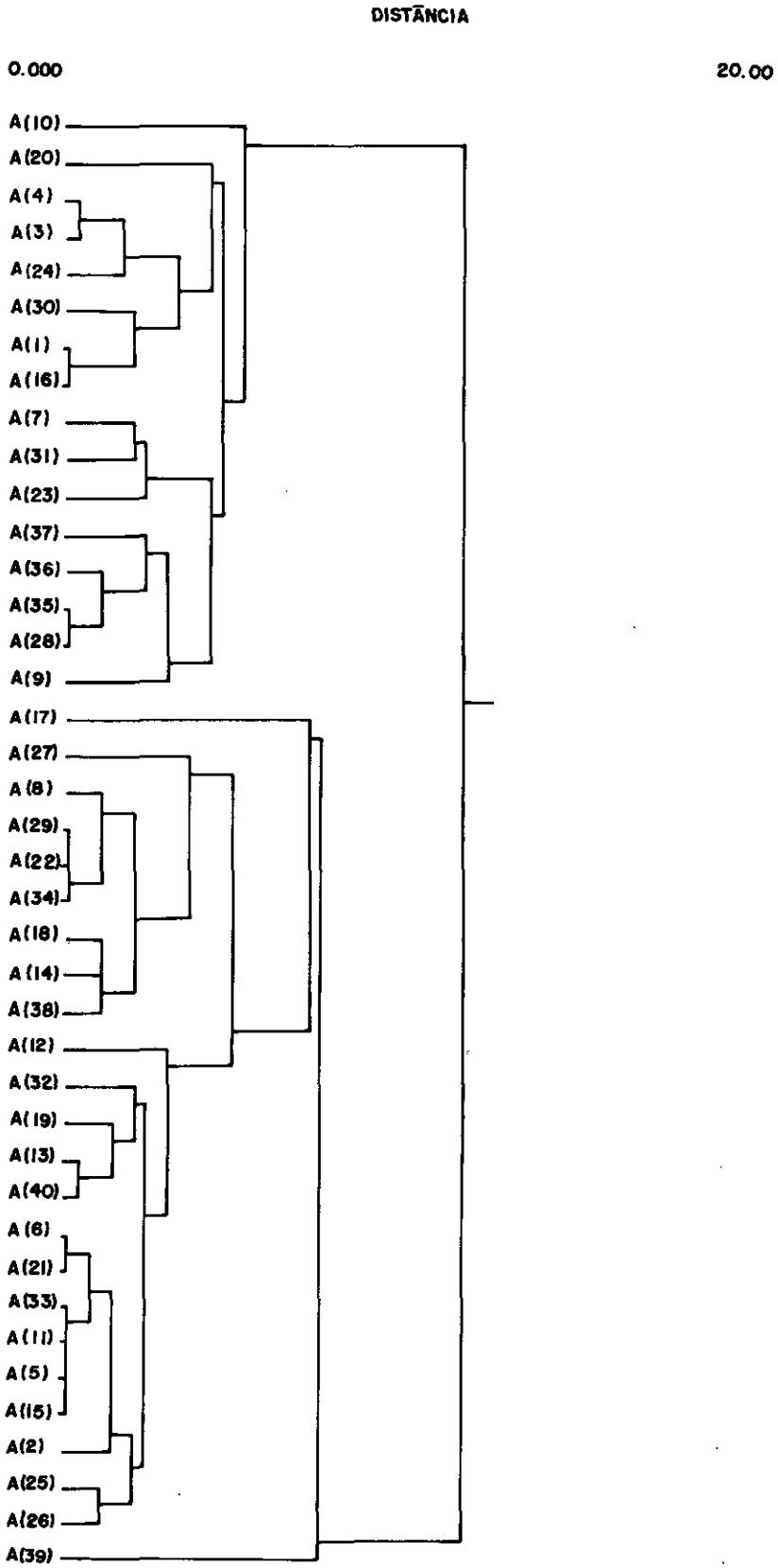


FIGURA 1 - AGREGAÇÃO SEQÜENCIAL DOS 40 CONSTITUINTES.

3.2 Síntese Gráfica

Construção de Escalas Multidimensionais, técnica descrita por Davison (1983), produz, a partir da matriz de discordâncias, um conjunto de pontos tais que as distâncias entre seus pares são tão próximas quanto possível das discordâncias. Ou seja, cada elemento da matriz é representado por um ponto em um espaço.

O mais freqüente é fixar a dimensão do espaço em 2, o que facilita a interpretação do gráfico, síntese da matriz de discordâncias. Nestas circunstâncias a técnica, se alimentada com as distâncias entre as cidades de uma região, reproduz o seu mapa a menos de rotações rígidas.

TABELA 3

Coordenadas dos Pontos Correspondentes aos Constituintes

Número	*Símbolo*	Coordenadas		*Número*	*Símbolo*	Coordenadas	
		1	2			1	2
A(1)	A	-.45	.17	A(21)	U	.60	.11
A(2)	B	.34	-.06	A(22)	V	1.19	-.43
A(3)	C	-1.21	.15	A(23)	W	-1.32	-.24
A(4)	D	-1.33	.18	A(24)	X	-.98	.04
A(5)	E	.69	.36	A(25)	Y	.53	.35
A(6)	F	.85	.21	A(26)	Z	.30	.27
A(7)	G	-.94	-.17	A(27)	a	.21	.62
A(8)	H	.94	-.65	A(28)	b	-1.64	-.25
A(9)	I	-1.20	-.84	A(29)	c	1.19	-.43
A(10)	J	-1.58	.32	A(30)	d	-.56	.78
A(11)	K	.66	.30	A(31)	e	-.28	-.11
A(12)	L	.71	-.19	A(32)	f	.77	.35
A(13)	M	.72	.06	A(33)	g	.68	.26
A(14)	N	.65	-.04	A(34)	h	1.19	-.43
A(15)	O	.68	.08	A(35)	i	-1.16	-.17
A(16)	P	-.73	.33	A(36)	j	-1.42	-.47
A(17)	Q	-.15	-.62	A(37)	k	-.81	-.55
A(18)	R	1.20	-.20	A(38)	l	1.20	-.20
A(19)	S	.43	-.19	A(39)	m	.27	.94
A(20)	T	-1.00	.50	A(40)	n	.78	-.12

Usando-se a matriz de discordâncias entre os constituintes, obtemos a configuração

Usando-se a matriz de discordâncias entre os constituintes, obtemos a configuração apresentada na Figura 2 e Tabela 3, que têm um *stress* associado de 0.107. Uma clara interpretação para o eixo horizontal é a ideologia. Aqueles constituintes de partidos de esquerda estão em pólo oposto àqueles de partidos de direita. O segundo eixo não é claramente interpretável. Destaca-se a coesão da esquerda — maior proximidade entre os pontos correspondentes.

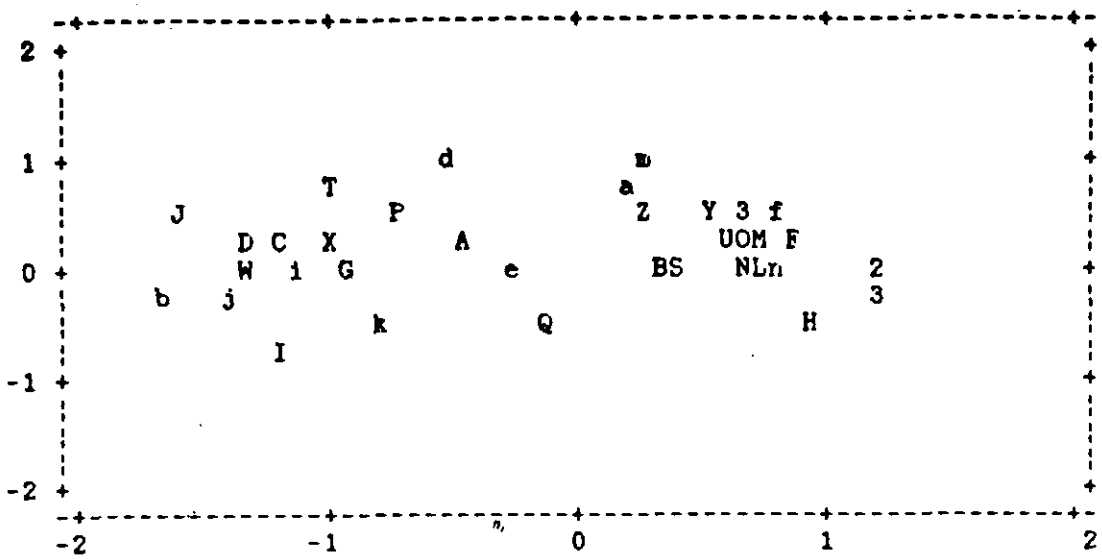


Figura 2

Pontos Representativos dos 40 Constituintes

A partir deste gráfico pode-se também antever as alianças que surgiram no primeiro e segundo turnos da eleição presidencial. Por exemplo, o ponto correspondente ao constituinte Konder Reis do PDS de Santa Catarina é muito próximo dos pontos dos constituintes do PSDB, o que antecipa o apoio que este constituinte deu ao candidato Mário Covas.

4 CONCLUSÃO

Existem várias direções em que este trabalho pode ser completado. A ausência maior é ter sido desenvolvido fora de uma análise política. É claro que para isto seria necessário o estudo do conjunto dos constituintes. Os dados para tal estão publicados em Coelho e Oliveira (1989). O tipo de análise de dados aqui utilizado é claramente uma ferramenta útil ao cientista político. Permitiria, por exemplo, a Kinzo (1988) incorporar na sua análise a dispersão interna nos partidos da constituinte de 1988 e

a Santos (1981) descobrir muito mais facilmente as associações entre os partidos do período 1946–1964.

Do ponto de vista técnico duas opções não foram vistas neste texto. A primeira refere-se a diferentes maneiras de se medirem discordâncias. Todas as tentativas feitas geraram resultados muito parecidos com os apresentados. Existem sugestões como a de Kruskal (1972) que entretanto não foram examinadas. Além disso, é razoável pensar que a ausência na votação gera, na realidade, incerteza sobre o posicionamento do constituinte. Um passo seria, através do ajuste de modelo estatístico, captar esta dificuldade de tomada de decisão.

5 AGRADECIMENTOS

O autor agradece a Roberto Guimarães e Nádia Rocha, alunos da disciplina de Estatística Multivariada onde, pela primeira vez, estes dados foram analisados; aos organizadores da Reunião Regional da ABE em Juiz de Fora , ao Prof. Djalma Pessoa, editor da RBE que com seu incentivo não deixou que este texto ficasse sem uma redação definitiva e a Mônica Mata Machado Castro do Departamento de Ciência Política da UFMG pela leitura atenta de versão preliminar.

BIBLIOGRAFIA

SANTOS, W.G.S. Coalizões parlamentares e instabilidade governamental: a experiência brasileira. In: PARTIDOS políticos no Brasil. Brasília: UNB, 1981, p. 259–274.

KINZO, M.D.G. *O quadro partidário e a constituinte.* São Paulo: Instituto de Estudos Econômicos, Sociais e Político de São Paulo, 1988. Texto n. 28.

KRUSKAL,J.B. The meaning of words. In: STATISTICS: a guide to the unknown. São Francisco: Holden-Day, 1972, p. 185–194.

DAVISON, M.L. *Multidimensional scaling.* New York: John Wiley, 1983.

COELHO, J.G.L.; OLIVEIRA, A.C.N. *A nova constituição: avaliação do texto e perfil dos constituintes.* Rio de Janeiro: Editora Reven.

RESUMO

O voto de 40 constituintes em 26 votações da Assembléia Nacional Constituinte foi usado para a construção de índice de discordância. A matriz de discordância resultante é sintetizada através de técnicas multivariadas. Como produto final obtêm-se mapas com a posição relativa dos constituintes. O quadro final é compatível com o resultado de análises políticas e tem algum poder preditivo de coligações que ocorreram posteriormente.

ABSTRACT

The ballots of 40 constituents in 26 key questions of the National Constitution Assembly were used to construct a discordance index. The resulting discordance matrix is synthesized through multivariate techniques. As a final product of these analyses maps with the relative position of the constituents were constructed. The final graph is compatible with political analyses and have some prediction power of alliances occurring later in the Brazilian political process.

O VOTO DOS 40 PARLAMENTARES QUE MAIS SE DESTACARAM

	Cinco anos para Sarney	Presidencialismo	Fortalecimento do Congresso	Suspensão dos dois turnos nas eleições deste ano	Democracia direta	Voto aos 16 anos	Habeas Data	Mandado de segurança coletivo	Direito irrestrito de greve	Jornada semanal de 44 horas	Turno ininterrupto de trabalho de seis horas	Licença-maternidade de 120 dias	Licença-paternidade	Adicional de 33% nas férias	Reforma tributária	Definição de empresa nacional	Reserva de mercado na informática	Fortalecimento do Judiciário	Anistia aos anistiados	Proibição à comercialização do sangue	Fim da censura	Veto à prerrogativa de intervenção das Forças Armadas na ordem interna	Estabilidade no emprego	Desapropriação de propriedade produtiva para reforma agrária	Pena de morte	Fim gradual e controlado da atividade nuclear
1. Affi Domingos (PL-SP)	não	sim	aus	não	aus	aus	sim	aus	não	sim	não	sim	abs	abs	sim	aus	não	não	aus	não	aus	não	aus	aus	aus	aus
2. Afonso Arinos (PSDB-RJ)	sim	não	aus	não	sim	sim	aus	aus	sim	aus	aus	aus	aus	aus	aus	sim	sim	não	abs	sim	sim	não	aus	sim	não	aus
3. Albano Franco (PMDB-SE)	sim	sim	sim	sim	sim	não	sim	não	não	sim	não	sim	abs	sim	não	não	não	não	não	não	não	não	não	não	não	não
4. Amaral Neto (PDS-RJ)	sim	sim	sim	sim	sim	não	sim	não	não	sim	não	sim	não	sim	sim	aus	não	não	não	não	sim	não	não	não	sim	não
5. Artur da Távola (PSDB-RJ)	não	não	sim	não	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	não	não	sim	sim	não	não	sim	não	não
6. Bernardo Cabral (PMDB-AM)	não	não	não	não	sim	sim	sim	sim	sim	sim	aus	sim	sim	sim	sim	sim	abs	não	sim	sim	aus	aus	sim	não	não	não
7. Bonifácio de Andrada (PDS-MG)	sim	não	sim	sim	aus	aus	sim	não	não	sim	não	sim	abs	sim	sim	não	não	aus	sim	não	aus	não	não	não	não	não
8. Brandão Monteiro (PDT-RJ)	não	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	não	aus	sim	sim	não	não	aus
9. Carlos Sant'Anna (PMDB-BA)	sim	sim	não	sim	não	não	sim	não	não	aus	não	sim	sim	não	não	aus	não	abs	não	sim	sim	não	não	não	não	não
10. Delfim Netto (PDS-SP)	sim	não	sim	sim	sim	não	sim	não	não	sim	não	não	não	não	sim	não	não	não	não	não	sim	não	não	sim	não	não
11. Egídio Ferreira Lima (PMDB-PE)	não	não	sim	não	aus	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	não	não	não	sim	sim	não	não	sim	não	não
12. Euclides Scalco (PSDB-PR)	não	não	sim	não	sim	não	sim	sim	sim	sim	sim	sim	sim	sim	sim	aus	sim	sim	sim	sim	sim	não	não	sim	não	aus
13. Fábio Feldmann (PSDB-SP)	não	não	aus	não	sim	sim	sim	sim	aus	aus	aus	aus	aus	sim	sim	sim	sim	não	aus	sim	não	não	não	sim	não	sim
14. Fernando Gasparian (PMDB-SP)	não	não	sim	não	sim	aus	sim	sim	aus	sim	sim	aus	aus	sim	sim	sim	sim	sim	aus	sim	aus	não	sim	sim	não	não
15. Fernando Henrique Cardoso (PSDB-SP)	não	não	sim	não	sim	sim	sim	sim	sim	sim	sim	aus	aus	sim	sim	sim	sim	aus	não	sim	sim	não	aus	sim	não	aus
16. Francisco Dornelles (PFL-RJ)	não	sim	abs	não	aus	não	sim	abs	não	sim	aus	sim	abs	aus	sim	não	aus	não	não	não	aus	não	não	não	aus	não
17. Gastone Righi (PTB-SP)	sim	não	sim	sim	aus	não	sim	não	aus	aus	sim	sim	aus	sim	sim	sim	sim	sim	não	sim	não	não	aus	não	não	não
18. Haroldo Lima (PC do B-BA)	não	não	sim	não	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	não	sim
19. Ibsen Pinheiro (PMDB-RS)	abs	não	sim	sim	aus	aus	sim	sim	sim	sim	sim	aus	aus	sim	sim	sim	sim	aus	não	sim	aus	aus	não	sim	não	sim
20. Jarbas Passarinho (PDS-PA)	sim	sim	sim	sim	não	sim	abs	sim	sim	abs	abs	sim	sim	abs	sim	aus	não	não	não	não	sim	não	não	não	sim	não
21. José Fogaça (PMDB-RS)	não	não	aus	não	sim	sim	sim	aus	sim	sim	sim	sim	sim	sim	sim	sim	sim	aus	aus	sim	sim	não	não	sim	não	não
22. José Genoíno (PT-SP)	não	sim	sim	não	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	não	sim
23. José Lins (PFL-CE)	sim	não	não	sim	não	não	sim	não	não	sim	não	sim	sim	sim	sim	não	não	não	aus	não	sim	não	não	sim	não	aus
24. José Lourenço (PFL-BA)	sim	sim	sim	aus	aus	aus	sim	não	não	sim	não	sim	sim	aus	aus	não	não	não	não	não	não	não	não	não	não	não
25. José Richa (PSDB-PR)	não	não	sim	não	sim	sim	sim	sim	sim	sim	sim	sim	abs	sim	sim	sim	não	aus	não	sim	sim	não	não	sim	não	não
26. José Serra (PSDB-SP)	não	não	sim	não	sim	sim	sim	aus	sim	aus	sim	abs	aus	sim	aus	não	não	não	sim	aus	não	não	aus	sim	não	sim
27. Konder Reis (PDS-SC)	não	não	sim	sim	sim	não	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	não	não	sim	sim	não	não	não	não	não	não
28. Luiz Eduardo Magalhães (PFL-BA)	sim	sim	aus	sim	não	não	sim	não	não	sim	não	não	não	não	não	não	não	não	não	não	sim	não	não	sim	não	não
29. Luiz Inácio Lula da Silva (PT-SP)	não	sim	sim	não	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	não	sim	sim
30. Marco Maciel (PFL-PE)	não	sim	sim	aus	abs	sim	sim	não	sim	abs	sim	sim	sim	sim	não	não	não	não	não	não	não	não	não	não	não	aus
31. Marcondes Gadelha (PFL-PB)	sim	não	aus	sim	aus	não	sim	sim	aus	sim	sim	aus	aus	aus	aus	aus	aus	sim	aus	aus	não	aus	não	aus	não	não
32. Mário Covas (PSDB-SP)	não	não	sim	não	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	não	não	sim	sim	não	não	não	não	sim
33. Nélson Jobim (PMDB-RS)	não	não	sim	não	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	não	não	sim	aus	não	aus	sim	não	não	não
34. Plínio de Arruda Sampaio (PT-SP)	não	sim	sim	não	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	não	sim	sim
35. Ricardo Fiuza (PFL-PE)	sim	sim	aus	sim	não	não	sim	não	aus	sim	não	aus	aus	aus	não	não	aus	não	aus	aus	não	não	não	sim	não	não
36. Roberto Campos (PDS-MT)	sim	sim	não	sim	não	não	aus	aus	não	sim	não	não	não	não	sim	não	não	aus	não	aus	sim	não	não	não	sim	aus	não
37. Roberto Cardoso Alves (PMDB-SP)*	sim	sim	não	aus	não	não	aus	sim	não	aus	sim	não	abs	sim	aus	aus	não	sim	aus	abs	aus	não	não	aus	sim	não	não
38. Roberto Freire (PCB-PE)	não	não	sim	não	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	não	sim
39. Sandra Cavalcanti (PFL-RJ)	não	não	não	não	sim	aus	sim	sim	aus	sim	sim	sim	sim	sim	sim	não	abs	não	não	sim	não	não	sim	não	não	sim
40. Severo Gomes (PMDB-SP)	não	não	aus	não	sim	abs	aus	sim	sim	sim	sim	sim	sim	sim	sim	sim	sim	não	sim	aus	abs	não	sim	não	não	sim

Deixou o Congresso para assumir o Ministério da Indústria e Comércio em 17 de agosto último. Para o parlamentar que esteve ausente na votação utilizou-se aus; para o que se absteve, abs

a: Centro de Informática e Processamento de Dados do Senado Federal (Prodasen)

APLICAÇÕES DE MODELOS DE SOBREVIVÊNCIA A DADOS DE INFARTO AGUDO DO MIOCÁRDIO

David Dorigo*
Hélio S. Migon*
Núbia K. O. Almeida**
Roberto Bassan***

1 INTRODUÇÃO

Em pesquisas biomédicas, muitas vezes, estamos interessados em saber se um certo evento ocorrerá e os principais fatores que o influenciam. Exemplos de tais eventos são o desenvolvimento de uma certa doença, complicações de doenças ou óbitos que ocorrem num período de acompanhamento de um grupo de indivíduos por um certo tempo.

O objetivo principal deste artigo é o de analisar, por meio de modelos estatísticos apropriados, a sobrevivência de um grupo de indivíduos que sofreram, em épocas passadas, infarto do miocárdio. A fim de determinar o evento de interesse - morte cardíaca ou sobrevivência após um período de observação fixado - precisamos saber as características ou fatores importantes relacionados a esse evento. Assim, grupos com pequeno ou grande risco de morte cardíaca podem ser determinados. As características relacionadas ao desenvolvimento de uma certa doença são ditas fatores ou variáveis de risco.

*Professor-Adjunto do Instituto de Matemática - UFRJ.
*Professor-Adjunto do Instituto de Matemática - UFRJ.
**Mestrando do Instituto de Matemática – UFRJ.
***Chefe da Clínica Cardiológica do Hospital Pró-Cardíaco, RJ.

Por outro lado, a predição do desempenho de um paciente individual com relação ao seu tempo de sobrevivência ou de duração de uma doença, isto é, o seu prognóstico, é de importância fundamental em estudos médicos. Na prática, as informações médicas contêm um grande número de características (ou fatores) ditas, em geral, variáveis explicativas ou covariáveis. Nessa etapa é, às vezes, difícil decidir quais dessas variáveis estão conjuntamente relacionadas ao prognóstico do paciente. O pesquisador pode tomar uma decisão *a priori* sobre quais variáveis explicativas são mais relevantes, porém uma análise estatística dos dados é indispensável para confirmar essas hipóteses. Através de modelos estatísticos, é possível relacionar aquelas variáveis realmente significativas e tirar conclusões sobre os dados da investigação.

Na Seção 2 descrevemos brevemente os dados clínicos usados na investigação, listando os principais fatores considerados. Na Seção 3 abordamos a análise estatística de dados propriamente dita, comentando os principais modelos ajustados. A natureza dos dados deste estudo médico oferece oportunidade para não só ajustar modelos, mas também selecionar os principais fatores de risco e variáveis explicativas associadas ao tempo de sobrevivência. Na Seção 4 apresentamos um resumo dos principais resultados da análise estatística dos dados usando os principais modelos estatísticos adequados a esse tipo de dado . Finalmente, na Seção 5, são mencionados resumidamente alguns problemas de seleção de covariáveis binárias e diagnóstico dos modelos ajustados.

2 OS DADOS

Uma amostra de 96 pacientes pós-infartados, limitados na idade de até 70 anos, seguidos prospectivamente, constituiu o grupo observado, que foi acompanhado trimestralmente através de exames cardiológicos. O período de admissão ao estudo foi de janeiro de 1980 a dezembro de 1982 e o seguimento desses pacientes encerrou-se em dezembro de 1989. Desta forma o seguimento máximo observado foi de 10 anos, sendo de 8,5 anos o tempo médio de seguimento. Todos os pacientes da amostra eram sobreviventes de fase aguda de infarto do miocárdio. Este infarto era diagnosticado, em cada paciente, pela existência de dor ou desconforto torácico associado à elevação de enzimas plasmáticas. De todo o grupo, 87% eram do sexo masculino e a faixa etária compreendia as idades de 30 a 69 anos, com idade média de 53 anos. O principal evento de interesse do estudo era o de morte de origem cardíaca. Assim, a eventual morte de um paciente que não fosse por causa cardíaca era considerada uma 'perda'. Do ponto de vista médico, o estudo visava a um melhor conhecimento da sobrevivência de pessoas que já haviam sofrido um infarto agudo do miocárdio.

Além do evento de interesse — morte por acometimento cardíaco — foram avaliadas também as seguintes características clínicas para eventualmente explicar a sobrevivência do grupo:

TABELA 1

Fator	Mnemônico	Domínio
Idade	id	30-69
Infarto anterior	i-an	+/-
Infarto inferior	i-in	+/-
Infarto subendocárdico	i-sen	+/-
Número de vasos lesados	l-vs	1,2 ou 3
Lesão de tronco	l-tr	+/-
Fração de ejeção	f-ej	(0,1)
Teste ergométrico	t-er	+/-
Reinfarto	r-inf	+/-
Angina	angi	+/-
Insuficiência cardíaca . .	i-car	+/-

Note-se que a maioria das características observadas é do tipo binário, isto é, presença (+) ou ausência (−). A variável número de **vasos lesados** pode ser dicotomizada, denotando-se presença quando 2 ou 3 vasos foram lesados. Da mesma forma, na característica **fração de ejeção** a presença poderá corresponder a valores menores que 30%. Como se sabe, da literatura médica, a chance de morte cardíaca cresce com a idade do paciente, de forma que se resolveu arbitrar o limite de 60 anos como ponto de corte para caracterizar a presença desse fator. Essas dicotomias são relevantes no estudo comparativo de curvas de sobrevivência.

Estudos médicos prévios demonstraram a importância dessas variáveis na explicação da ocorrência de óbito cardíaco em pacientes com infarto do miocárdio. A título de exemplo, pacientes mais idosos têm uma mortalidade pós-infarto maior que os mais jovens. Da mesma forma, **infarto de parede anterior** têm pior prognóstico do que o **infarto de parede inferior**. Pacientes com lesão de 2 ou 3 vasos têm maior mortalidade do que os com lesão de 1 vaso. Analogamente, pacientes com **fração de ejeção** menor que 30% têm pior prognóstico. As variáveis **reinfarto, angina** e **insuficiência cardíaca**, que são complicações observadas nos pacientes com infarto do miocárdio ao longo do tempo, também foram utilizadas como fatores explicativos da morte cardíaca.

3 ANÁLISE DOS DADOS

O objetivo principal da investigação é o de estudar a sobrevivência de pós-infartados, destacando os seguintes aspectos:

i) - construção de curvas de sobrevivência para todo o grupo;

ii) - comparação das sobrevivências dos principais subgrupos, estratificados segundo

características de interesse;

iii) - avaliação da sobrevivência considerando algumas características principais; e

iv) - avaliação do desempenho individual de cada característica na ocorrência do evento principal (morte por causa cardíaca).

As técnicas usuais de análise envolvem métodos paramétricos e não-paramétricos. Dentre os métodos de estimação de curvas de sobrevivência, isto é, que estimam a proporção de indivíduos que sobreviverão pelo menos um certo período de tempo, destaca-se o do produto limite de Kaplan-Meier (1958). Este método, geralmente usado em pequenas amostras, estima, por via essencialmente não-paramétrica, a sobrevivência para o caso de observações censuradas e não impõe condições sobre a forma da função a estimar. A função de sobrevivência, neste caso, é estimada por

$$\hat{S}(t) = \prod_{t_{(i)}<t} \frac{n-i}{n-i+1} \tag{3.1}$$

onde n representa o número total de indivíduos cujos tempos de sobrevivência, censurados ou não, estão disponíveis, i representa os inteiros consecutivos $1, 2, \cdots, n$, caso não haja dados censurados e $t_{(i)}$ representa os tempos de sobrevivência em ordem crescente. Comparações entre grupos usando a expressão (3.1) são normalmente feitas com auxílio de aplicativos apropriados, os quais fornecem, além das estimativas, gráficos das sobrevivências.

Uma maneira mais abrangente de avaliar esses dados é através do modelo de regressão múltipla, introduzido por Cox (1972), o qual tem-se revelado muito útil nas aplicações, especialmente na área médica. Neste caso, precisa-se estabelecer uma relação entre o tempo de sobrevivência t e um conjunto de variáveis explicativas $x_1, \cdots, x_p$, que é expressa por

$$h_i(t) = h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right) \tag{3.2}$$

onde $h_i(t)$ representa a taxa de falha ou função de risco do $i^{\underline{o}}$ indivíduo, x_{ij} a $j^{\underline{a}}$ covariável do $i^{\underline{o}}$ indivíduo e $h_0(t)$ é a taxa de falha da distribuição de sobrevivência (completamente arbitrária) subjacente quando todas as covariáveis são ignoradas; β_j é o correspondente coeficiente de regressão. Esses coeficientes são estimados, essencialmente, por uma adaptação do método clássico de máxima verossimilhança através de processos iterativos, como, por exemplo, o método numérico de Newton-Raphson. Para tanto, a expressão da verossimilhança a maximizar é determinada a partir do conjunto de indivíduos em risco, isto é, vivos e não censurados à época t. Mais formalmente, a determinação dos β_j, que é feita independentemente de $h_0(t)$, é obtida por meio da verossimilhança marginal em relação ao conjunto de indivíduos em risco. Detalhes específicos sobre o processo de estimação de (3.2) podem ser encontrados em Kalbfleish e Prentice (1980).

A identificação do desempenho de cada característica em relação à probabilidade de morte cardíaca pode ser feita por meio de regressão logística usual

$$\log\left(\frac{p_i}{1-p_i}\right) = \sum_{j=1}^{p} \beta_j x_{ij} \tag{3.3}$$

onde p_i representa a probabilidade de sucesso (morte cardíaca) do $i^{\underline{o}}$ indivíduo, a qual é descrita por p características ou fatores de risco x_{ij}, $\forall j = 1, 2, \ldots, p$ e β_j são os correspondentes coeficientes a estimar. A estimativa desses coeficientes, utilizando o método de máxima verossimilhança, é obtida também por processo iterativo.

4 AJUSTE DOS MODELOS E PRINCIPAIS RESULTADOS

Com auxílio computacional do aplicativo BMDP (1985), versão 4.2.1, ajustaram-se os modelos (3.1), (3.2) e (3.3) aos dados. A descrição inicial dos dados pelo método de Kaplan-Meier (equação (3.1)) para toda a amostra revelou um tempo médio de sobrevivência de 92,29 meses com respectivo erro padrão de 4,19 meses. Dos 96 indivíduos, 33 tiveram morte cardíaca, 5 morte não–cardíaca e 5 foram perdidos ao longo do seguimento e 53 estavam vivos no fim do estudo. A taxa média anual de mortalidade foi de 5%. Uma descrição completa dos cálculos e um gráfico da curva de sobrevivência são apresentados nos Anexos 1 e 2.

A comparação da curva de sobrevivência entre subgrupos, segundo as diversas características, está resumida na Tabela 2. O efeito do fator **lesão de tronco**, isto é, obstrução de tronco de coronária esquerda, no prognóstico dos pacientes pode ser observado no gráfico (Anexo 2) onde são apresentadas as curvas de sobrevivência para os estratos indicados, o p-valor e o número de pacientes em cada grupo.

TABELA 2

Comparações de Sobrevivência pelo Método do Produto Limite entre Subgrupos, Segundo Fatores (Características) mais Importantes

Fator	Estatística de Teste	Valor-p
id (> 60) vs id (< 60)	2.54	0.11
l-vs (1) vs l-vs (2,3) . .	4.73	0.03
l-tr (+) vs l-tr (-) . . .	12.85	< .01
f-ej (< .3) vs f-ej (> .3)	8.98	0.03
t-er (+) vs t-er (-) . . .	2.59	0.11

A estatística de teste usada para comparação é a sugerida por Breslow (1970), a qual generaliza a estatística usual não-paramétrica de Kruskal-Wallis. Como resultado,

observou-se que as diferenças entre os subgrupos são significativas quando eles são estratificados segundo **o número de vasos lesados, lesão de tronco e fração de ejeção** ($< .3$).

Prosseguindo a análise, utilizou-se a equação (3.2) para escolher, através do método de seleção passo a passo, os fatores mais significativos para prognóstico, bem como as estimativas finais dos respectivos coeficientes. Nesta análise as variáveis **número de vasos lesados**, e **idade** não foram dicotomizadas. Os resultados obtidos estão apresentados na Tabela 3.

TABELA 3

Fatores mais Significativos:
Estimativas dos Respectivos Coeficientes e Desvios-Padrão

Fator	Coeficiente	D. Padrão	Log-veros	Valor-p
l-vs	1.42	0.48	-41.39	$< .001$
f-ej	2.73	0.99	-39.22	.001
t-er	1.95	0.80	-39.59	.004
angi	-1.75	0.78	-38.96	.009

Dos resultados resumidos nessa tabela constata-se que os fatores mais importantes no prognóstico de mortalidade são **número de vasos lesados, fração de ejeção, teste ergométrico** e **angina.**

Para ilustrar o uso dos coeficientes apresentados na Tabela 3, considere como referência um paciente com lesão de três vasos, **fração de ejeção** menor do que 30%, **teste ergométrico** positivo e ausência de **angina**, o que corresponde a maior taxa de falha estimada pelo modelo. Considere, agora, um indivíduo que difira do anterior somente pela presença de **angina**. Neste caso a taxa de falha se reduz em 83%. Analogamente se a **fração de ejeção** tivesse sido superior a 30% então a taxa de falha teria se reduzido expressivamente (93%), o que implicaria uma maior probabilidade de sobrevivência. Finalmente, no caso de **teste ergométrico** negativo a redução seria de 86%, enquanto que para um paciente com um único vaso lesado seria de 76%.

Complementando a análise, utilizamos a equação (3.3) para calcular os fatores de risco mais significativos associados à probabilidade de óbito cardíaco. Estimativas dos coeficientes dos fatores de risco mais significativos estão apresentadas na Tabela 4. Note-se o sinal negativo do coeficiente do fator **angina** (Tabelas 3 e 4), um resultado surpreendente. Este comportamento sugere, talvez, algum tipo de confundimento entre fatores.

TABELA 4

Fatores de Risco mais Significativos no Óbito Cardíaco:
Estimativas dos Respectivos Coeficientes e Desvios–Padrão

Fator de Risco	Coeficiente	D. Padrão	Coeficiente/ d. padrão
i-in	-0.79	0.33	-2.378
l-tr	1.25	0.40	2.850
r-inf	1.15	0.40	2.850
angi	-1.07	0.39	-3.063

O ajuste do modelo logístico aos dados foi testado pela estatística χ^2 usual de aderência ($\chi^2 = 67.67$, valor-$p = .623$). Como se observa na Tabela 4 os fatores de risco mais importantes são **infarto inferior, lesão de tronco, reinfarto e angina**. Novamente obteve-se o sinal negativo no coeficiente do fator **angina**. Como mencionamos anteriormente, **infarto de parede inferior** tem melhor prognóstico que **infarto de parede anterior** justificando o sinal negativo do coeficiente desta variável no modelo selecionado. Cabe ressaltar aqui que, utilizando o risco relativo (RR) como medida de associação entre um fator e a morte cardíaca, as estimativas mais significativas desse índice foram as referentes a **lesão de tronco** ($\hat{RR} = 8.00$), **lesão de 2 ou 3 vasos** ($\hat{RR} = 3.8$) e **teste ergométrico** ($\hat{RR} = 3.6$).

A partir dos coeficientes apresentados na Tabela 4 pode-se fazer prognósticos sobre a ocorrência do evento morte cardíaca. Por exemplo, o paciente com maior chance de morte cardíaca é caracterizado pela presença de **lesão de tronco, infarto de parede anterior** (a qual está representada no modelo indiretamente através do fator **infarto inferior**) e **reinfarto**, sendo sua probabilidade de morte estimada como 0.92. Se além desses fatores o paciente tiver tido **angina** sua chance de morte se reduzirá em 14%, o que é surpreendente. Caso o paciente não tenha tido outras complicações cardíacas (**angina**, etc.) sua probabilidade de morte decrescerá em 16.3%. O **infarto de parede inferior** diminui a probabilidade de morte cardíaca em 9.8%, enquanto que se não houver **lesão de tronco** a redução será de 17.4%.

Resumindo a análise dos dados cardíacos, podemos dizer que:

i) - os pós-infartados apresentaram uma taxa de mortalidade anual relativamente baixa;

ii) - as covariáveis **número de vasos lesados, fração de ejeção, teste ergométrico**, e **angina** são os preditores mais importantes da taxa de mortalidade (Tabela 3); e

iii) - como desempenho conjunto na ocorrência de morte cardíaca são importantes **infarto inferior, lesão de tronco, reinfarto** e **angina** (Tabela 4).

5 SELEÇÃO DE COVARIÁVEIS E DIAGNÓSTICO DE MODELOS

Uma característica peculiar da matriz de dados desse estudo é o excesso de covariáveis binárias. No caso de modelos de regressão logística (3.3) a seleção de covariáveis pode ser transformada num problema de seleção de variáveis de regressão múltipla, segundo sugestão de Nordberg (1981). No entanto, tal problema pode ser também resolvido usando a técnica de análise de correspondência multivariada.

Por outro lado, algumas técnicas de diagnóstico de modelos de dados binários já existem. O trabalho pioneiro de Pregibon (1981) e outros trabalhos recentes (Davison (1989)) sugerem, principalmente, métodos gráficos para verificação dos modelos ajustados. Uma pesquisa nessa direção está sendo desenvolvida pelos autores do presente estudo.

Agradecimentos

Os dados utilizados neste trabalho pertencem ao Hospital Geral de Bonsucesso, Rio de Janeiro. Os autores agradecem o apoio financeiro do CNPq e, também, a dois críticos que, anonimamente, fizeram sugestões pertinentes no sentido de melhorar a compreensão do texto.

ANEXO 1

Sobrevivência Acumulada, Mortalidade e Perdidos Durante o Seguimento (Janeiro de 1980 a Dezembro de 1989)

(continua)

CASE LABEL	CASE NUMBER	TIME MONTH	STATUS	CUMULATIVE SURVIVAL	STANDARD ERROR	CUM DEAD	CUM LOST	REMAIN AT RISK
	9	1.02	LOST			0	1	95
	33	2.00	DEAD	0.9895	0.0105	1	1	94
	12	2.98	DEAD	0.9789	0.0147	2	1	93
	25	3.02	DEAD			3	1	92
	91	3.02	DEAD	0.9579	0.0206	4	1	91
	15	5.02	DEAD			5	1	90
	72	5.02	DEAD	0.9368	0.0250	6	1	89
	70	7.05	DEAD	0.9263	0.0268	7	1	88
	64	8.03	DEAD	0.9158	0.0285	8	1	87
	49	13.90	LOST			8	2	86
	40	16.98	DEAD	0.9051	0.0301	9	2	85
	85	22.95	DEAD	0.8945	0.0316	10	2	84
	36	26.92	DEAD	0.8838	0.0329	11	2	83
	29	26.95	DEAD	0.8732	0.0342	12	2	82
	89	30.92	DEAD	0.8625	0.0354	13	2	81
	48	34.89	LOST			13	3	80
	61	35.93	DEAD			14	3	79
	75	35.93	DEAD	0.8410	0.0377	15	3	78
	62	35.93	LOST			15	4	77
	52	38.95	LOST			15	5	76
	7	42.89	DEAD	0.8299	0.0388	16	5	75
	55	42.92	DEAD	0.8189	0.0398	17	5	74
	58	46.89	LOST			17	6	73
	76	46.92	DEAD	0.8076	0.0408	18	6	72
	63	48.89	DEAD	0.7964	0.0417	19	6	71
	11	49.90	LOST			19	7	70
	32	50.92	LOST			19	8	69
	86	51.93	DEAD	0.7849	0.0427	20	8	68
	56	53.93	DEAD	0.7733	0.0436	21	8	67
	44	58.89	LOST			21	9	66
	93	59.87	DEAD	0.7616	0.0445	22	9	65
	19	62.89	DEAD	0.7499	0.0453	23	9	64
	41	64.89	DEAD	0.7382	0.0461	24	9	63
	23	71.84	DEAD	0.7265	0.0468	25	9	62
	16	76.85	LOST			25	10	61
	87	83.84	DEAD	0.7146	0.0476	26	10	60
	94	83.84	CENSORED			26	10	59
	95	83.84	CENSORED			26	10	58
	96	83.84	CENSORED			26	10	57
	54	84.75	DEAD	0.7020	0.0484	27	10	56
	92	84.82	CENSORED			27	10	55
	90	85.84	CENSORED			27	10	54
	74	88.85	DEAD	0.6890	0.0492	28	10	53
	81	88.85	CENSORED			28	10	52
	83	88.85	CENSORED			28	10	51
	73	89.77	DEAD	0.6755	0.0500	29	10	50
	79	89.84	CENSORED			29	10	49
	80	89.84	CENSORED			29	10	48
	88	89.84	CENSORED			29	10	47
	84	90.85	CENSORED			29	10	46

Sobrevivência Acumulada, Mortalidade e Perdidos Durante o Seguimento (Janeiro de 1980 a Dezembro de 1989)

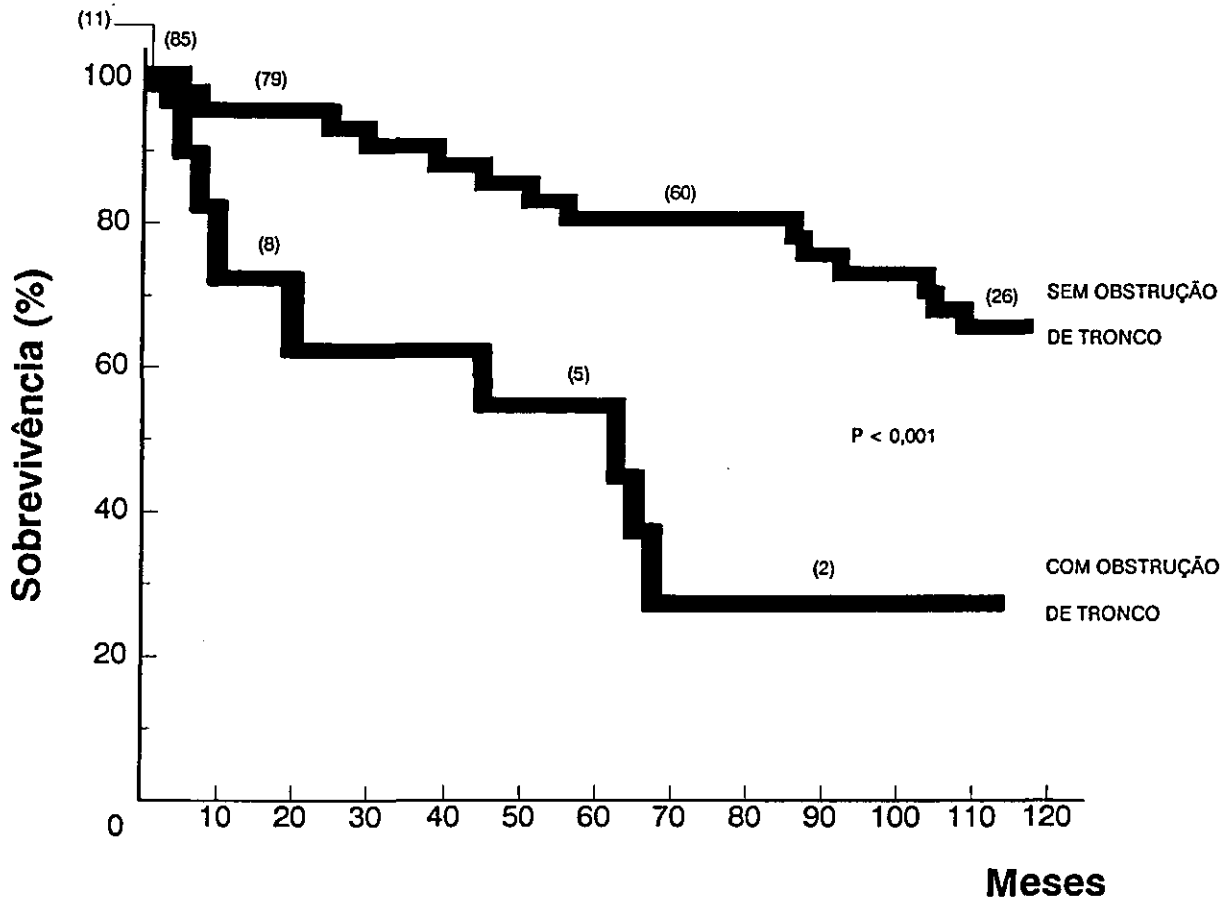
(conclusão)

CASE LABEL	CASE NUMBER	TIME MONTH	STATUS	CUMULATIVE SURVIVAL	STANDARD ERROR	CUM DEAD	CUM LOST	REMAIN AT RISK
	78	91.84	CENSORED			29	10	45
	82	91.84	CENSORED			29	10	44
	77	93.77	CENSORED			29	10	43
	17	97.80	DEAD	0.6598	0.0513	30	10	42
	71	99.80	CENSORED			30	10	41
	45	101.77	DEAD	0.6437	0.0525	31	10	40
	68	101.80	CENSORED			31	10	39
	69	101.80	CENSORED			31	10	38
	66	102.82	DEAD	0.6268	0.0538	32	10	37
	65	102.82	CENSORED			32	10	36
	67	102.82	CENSORED			32	10	35
	59	104.82	CENSORED			32	10	34
	60	104.82	CENSORED			32	10	33
	50	105.74	CENSORED			32	10	32
	51	105.74	CENSORED			32	10	31
	53	105.74	CENSORED			32	10	30
	57	105.74	CENSORED			32	10	29
	24	106.75	DEAD	0.6052	0.0561	33	10	28
	46	106.75	CENSORED			33	10	27
	47	106.75	CENSORED			33	10	26
	43	107.77	CENSORED			33	10	25
	42	108.75	CENSORED			33	10	24
	34	110.75	CENSORED			33	10	23
	35	110.75	CENSORED			33	10	22
	37	110.75	CENSORED			33	10	21
	38	110.75	CENSORED			33	10	20
	39	110.75	CENSORED			33	10	19
	27	111.77	CENSORED			33	10	18
	30	111.77	CENSORED			33	10	17
	31	111.77	CENSORED			33	10	16
	26	112.79	CENSORED			33	10	15
	28	112.79	CENSORED			33	10	14
	21	113.77	CENSORED			33	10	13
	22	113.77	CENSORED			33	10	12
	18	114.79	CENSORED			33	10	11
	20	114.79	CENSORED			33	10	10
	13	115.77	CENSORED			33	10	9
	14	115.77	CENSORED			33	10	8
	8	116.79	CENSORED			33	10	7
	10	116.79	CENSORED			33	10	6
	5	117.74	CENSORED			33	10	5
	6	117.74	CENSORED			33	10	4
	1	118.75	CENSORED			33	10	3
	2	118.75	CENSORED			33	10	2
	3	118.75	CENSORED			33	10	1
	4	118.75	CENSORED			33	10	0

@MEAN SURVIVAL TIME = 92.29 LIMITED TO 118.75 S.E. = 4.194

ANEXO 2

CURVA ATUARIAL DE SOBREVIÊNCIA PÓS-INFARTO: ESTRATIFICAÇÃO POR PRESENÇA/AUSÊNCIA DE OBSTRUÇÃO DE TRONCO DE CORONÁRIA ESQUERDA



BIBLIOGRAFIA

BIOMEDICAL computer programs: BMDP. Berkeley: Berkeley University California Press, 1985.

BRESLOW, N. A generalized Kruskal-Wallis test for comparing K samples subject to unequal pattern of censorship. *Biometrika*, v. 57, p. 579-594, 1970.

COX, D.R. Regression models and life tables, *Journal of the Royal Statistical Society S. B*, v.34, n.2, p. 187-220, 1972.

DAVISON, A.C. Model checking II: binary data. *REBRAPE*, v. 3, p. 87-96, 1989.

KALBFLEISH, J.D.; PRENTICE, R. *Statistical analysis of failure time data.* New York: J. Wiley, 1980.

KAPLAN, E. MEIER, P. Nonparametric estimation from incomplete observation. *Journal of the American Statistical Association*, v. 53, n. 282, p. 457-481, 1958.

NORDBERG, L. Stepwise selection of explanatory variables in the binary logit model. *Scandinavian Journal of Statistics*, v. 8, p. 17-26, 1984.

PREGIBON, D. Logistic regression diagnostic. *The Annals of Statistics*, v. 9, p. 705-724, 1981.

RESUMO

Neste trabalho apresentamos os principais resultados de um estudo de seguimento sobre pacientes cardíacos após sofrerem infarto. O objetivo é o de estudar a sobrevivência desses pacientes em função de fatores que possam explicar não somente a sobrevivência, como também informar sobre seu prognóstico. Trata-se de um estudo de 10 anos de observação, envolvendo 96 pacientes, de ambos os sexos, com idades entre 30 e 69 anos.

A análise de dados sugere que os pós-infartados têm uma baixa taxa de mortalidade, a idade não é fator importante no prognóstico de morte e que um pequeno número de fatores cardíacos pode resumir o desempenho do grupo.

Detalhes técnicos sobre os modelos usados para interpretar os dados são encontrados na bibliografia.

ABSTRACT

A massive 10-year follow-up study of cardiac patients, which previously had an acute myocardial infarction, was conducted at 'Hospital Geral de Bonsucesso', Rio de Janeiro, to identify the main risk factors for prognostic value and to predict cardiac mortality.

The aim of this paper is chiefly to apply main currently available survival models to interpret medical data. Technicalities related to these models are not discussed in this report and details can be looked at in the references given or mentioned therein.

Main results were: multivessel coronary disease, left ventricular ejection fraction and positive attenuated stress test are strongly related to late mortality, whereas age is not. Post-infarction patients experienced a relatively low mortality in this particular study.

RESENHAS BIBLIOGRÁFICAS

– CATEGORICAL DATA ANALYSIS, por Alan Agresti, New York: Wiley – Interscience, 1990, xv + 558 pages

Os métodos de análise de dados categóricos tiveram, recentemente, grande desenvolvimento, principalmente devido à introdução de modelos regressivos, culminando com a proposta do G.L.M. de Nelder e Wedderburn.

O livro se destina, segundo o autor, a estudantes de cursos sobre dados categóricos, bem como a estatísticos e bioestatísticos aplicados. Pretende, ainda, que ele seja útil a usuários, em geral, destes métodos, e é organizado em 13 capítulos e 3 apêndices, contendo extensa bibliografia sobre o assunto. Apresenta mais de 40 exemplos de análises de conjuntos de dados "reais", propõe mais de 400 exercícios, voltados à teoria e aplicações em análise de dados.

O livro compõe-se de 4 unidades, sucintamente descritas abaixo:

UNIDADE 1: CAPÍTULOS 2 E 3

Introduz métodos tradicionais para a análise de Tabelas de Contingência, com duas entradas, além de rápida introdução às medidas de associação clássicas.

UNIDADE 2: CAPÍTULOS 4 A 7

Constitui o cerne do livro, introduzindo os Modelos Loglinear e Logit. A motivação é dada através de exemplo do Paradoxo de Simpson, que chama a atenção para diferenças

entre análises feitas através de tabelas marginais e parciais. A exposição é bastante clara e detalhada.

UNIDADE 3: CAPÍTULOS 8 A 11

Apresenta aplicações e generalizações do material exposto na Unidade 2, tais como: Modelos Loglinear e Logit para Variáveis Ordinais; Modelos de Resposta Multinomial. Esta unidade é o ponto forte do livro, não havendo apresentação similar, pelo menos nos textos clássicos do assunto.

UNIDADE 4: CAPÍTULOS 12 E 13

Contém **background** teórico dos métodos para análise de dados categóricos. É de interesse maior para leitores com formação teórica mais avançada em estatística. A leitura desta unidade pode ser dispensada por quem estiver, apenas, interessado nas aplicações.

Merece ainda destaque o Apêndice A, que contém descrição dos principais **softwares** disponíveis para análise de dados categóricos. O autor não se limita apenas a citações, mas fornece listagens de programas relativos a alguns exemplos contidos no texto.

O livro foi utilizado por nós em disciplina de Mestrado e, posteriormente, em curso de extensão, de curta duração. Nas duas ocasiões, os alunos tiveram boa receptividade, monstrando-se motivados pelos exemplos apresentados no texto. Recomendamos a aquisição deste livro, até para quem tenha interesse em adquirir uma única fonte de referência sobre o assunto.

(Djalma G.C. Pessoa – ENCE / IBGE)

– A COURSE IN DENSITY ESTIMATION, por Luc Devroye. Birkhäuser, 1978, 183p.

Métodos de estimação não paramétrica de funcionais receberam grande impulso nos anos 80. Entre as várias razões que contribuíram para isto podemos certamente destacar duas. A primeira se prende a facilidades computacionais. O ciclo aplicações gerando problemas teóricos cujas soluções por sua vez se prestam à análise de novas aplicações se beneficia, como é claro, de avanços em SOFTWARE/ HARDWARE. Tais avanços são em particular importantes quando se trata da estimação de funcionais. Isto porque neste caso boas representações gráficas *on line* bem como rápidas respostas a repetidos problemas de otimização relacionados a parâmetros de suavizamento são fundamentais.

A segunda se prende a trabalhos básicos de diversos pesquisadores entre os quais certamente se inclui Luc Devroye. Neste seu segundo livro sobre estimação de densidades, Devroye opta por abordar tópicos num nível de profundidade (teórica) ligeiramente inferior àquele de Devroye & Györfi (1985). No entanto, *A Course in Density Estimation* não é de forma alguma uma condensação de *Nonparametric Density Estimation : the L_1 View*. Destacam-se como tópicos novos o capítulo sobre robustez e o sobre estimadores de distância mínima.

Em linhas gerais, o livro – apesar do título – é incompleto para um primeiro curso em estimação de densidades. Isto porque aspectos ligados a aplicações são propositadamente relegados a um segundo plano. É no entanto louvável – dado o caráter introdutório do texto – a tentativa de Devroye de se restringir basicamente ao estudo de núcleo–estimadores de densidades. A inclusão de outros estimadores acaba por desviar a atenção de conceitos e resultados importantes. A coleção de referências não é extensa, mas contem a maioria dos artigos–chave para os assuntos tratados no texto. Listas modestas de exercícios de caráter teórico são apresentadas ao final de cada capítulo.

A nível de dificuldade o livro se situa entre o livro de Silverman (1986) e Devroye & Györfi (1985). Um bom curso a nível de mestrado poderia ser obtido combinando–se material deste livro com Silverman (1986).

O volume é dividido em 9 capítulos resumidos abaixo.

Cap. I: Distâncias entre densidades

São apresentadas diversas distâncias entre densidades. Este capítulo é importante na medida em que para fixarmos um critério de avaliação de perfomance "global" de estimadores de densidades precisamos estabelecer distâncias entre estimadores e a densidade verdadeira. Devroye busca justificar, a meu ver com sucesso, sua opção pela distância L_1.

Cap. II: Density Estimation and Derivation of Measures

São apresentados alguns estimadores de densidades. Além disso consideram–se os importantes teoremas de densidade de Lebesgue bem como o teorema de Scheffe e sua contrapartida estocástica, o teorema de Glick. Este é um capítulo básico para o resto do livro.

Cap. III: Consistency of the Kernel Estimate

A consistência (em L_1) dos Núcleo–Estimadores é estudada. Em particular mostra–se a equivalência entre várias formas de consistência para núcleo–estimadores. Analisa–se também a consistência quando se supõe que o parâmetro de suavizamento seja obtido a partir da amostra.

Cap. IV: Robustness

Define–se o conceito de robustez em estimadores de densidades, usando–se L_1 – vizinhanças. Um exemplo interessante mostra que estimadores paramétricos naturais podem ser não robustos. Núcleo–estimadores são mostrados serem robustos.

Cap. V: Minímax Bourds

São analisados métodos de avaliação de cotas para erros minímax (sempre usando a distância L_1) para algumas classes de densidades. Este capítulo é extremamente claro, tendo como pontos mais importantes o teorema de Assouad, e o método *low–probability* de obtenção de cotas minímax.

Cap. VI: Estimadores Distância Miníma

A teoria de estimadores de distância mínima de Le Cam é estendida para estimadores de densidades, seguindo trabalho de Y. Yatracos. Este desenvolvimento guarda semelhança com métodos de processos empíricos (ver por exemplo Pollard, D. (1984)). A exposição é extremamente elegante.

Cap. VII: Rate of convergence of Kernel Estimates

Neste capítulo estudam–se taxas de convergência do valor esperado da distância L_1 entre núcleo estimadores e a densidade verdadeira. O papel da escolha de um particular núcleo fica evidente quando estamos interessados em estabelecer taxas de convergência uniforme sobre particulares famílias de densidades, para quantidades como acima.

Cap. VIII: A Case study : Monotone Densities on [0,1]

Um estudo comparativo entre vários estimadores de densidade é levado a efeito, quando se considera a classe de densidades monótonas e definidas em $[0,1]$. Em particular exibe–se um estimador minímax – ótimo.

Cap. IX: Relative Stability

O conceito de estabilidade relativa segundo a métrica L_1 é definido. Mostra–se que alguns núcleos estimadores são relativamente estáveis.

(Getúlio Borges da Silveira Filho – ENCE / IBGE)

POLÍTICA EDITORIAL

A RBEs objetiva promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais através da apresentação, descrição e discussão desses métodos e de suas aplicações, num formato de fácil assimilação pelos membros da comunidade científica. Destina–se também a servir de veículo para troca de idéias entre os especialistas e todos os interessados em análise e desenvolvimento de metodologia estatística.

A RBEs tem periodicidade semestral e publica artigos teóricos e/ou aplicados de métodos estatísticos, com ênfase na análise de fenômenos econômicos e sociais. São também aceitos artigos abordando os diversos aspectos do desenvolvimento metodológico relevantes para órgãos produtores de estatísticas, assim como artigos de revisão do estado da arte em temas específicos.

a) delineamento de pesquisas;

b) avaliação de pesquisas e mensuração de erros;

c) uso e combinação de fontes alternativas de informações;

d) novos desenvolvimentos em metodologia de pesquisa;

e) análise de séries de tempo;

f) estudos demográficos;

g) integração de dados;

h) amostragem e estimação;

i) análise de dados;

j) crítica e imputação de dados;

l) disseminação e confiabilidade de dados;

m) modelos econométricos.

Todos os artigos submetidos serão avaliados pelo Comitê Editorial da RBEs quanto a sua qualidade e relevância, devendo os mesmos serem inéditos. Além disto, não deverão ter sido simultaneamente submetidos a qualquer outro periódico nacional.

A RBEs publicará também resenhas de livros, artigos escritos a convites e ensaios sobre o ensino de Estatística.

INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBEs

Os artigos remetidos para publicação deverão ser submetidos em 3 vias (que não serão devolvidas) para:

Djalma G. C. Pessoa
Editor Responsável – RBEs
ENCE
Rua André Cavalcanti, 106
Bairro de Fátima
20231 – Rio de Janeiro – RJ

– Os artigos submetidos à RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.

– Cada autor receberá, gratuitamente, 20 separatas de seu artigo.

Instruções para preparo de originais:

1. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se para cada um a filiação e endereço permanente para correspondência. Agradecimentos a colaboradores, outras instituições e auxílios recebidos devem figurar também nesta página.

2. A segunda página do original deve conter resumos em português e em inglês (Abstract) destacando os pontos relevantes do artigo. Cada resumo deve ser datilografado seguindo o mesmo padrão do restante do texto, em um único parágrafo, sem fórmulas, com no máximo 150 palavras.

3. O artigo deve ser dividido em seções numeradas progressivamente, com títulos concisos e apropriados. Subseções, todas, inclusive a primeira, devem ser numeradas e receber título apropriado.

4. A citação de referências no texto e a listagem final das referências devem ser feitas de acordo com as normas da ABNT.

5. As tabelas e gráficos devem ser apresentados em folhas separadas, precedidas de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas seqüencialmente (Tabela 1, Figura 3, etc.) e referidas nos locais de inserção pelos respectivos números. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. Os apêndices devem ter título e numeração tal como as demais seções do trabalho.

6. Gráficos e diagramas para publicação devem ser traçados em papel branco, com nitidez e boa qualidade, para permitir que a redução seja feita mantendo qualidade.

Fotocópias não serão aceitas. É fundamental que não existam erros quer no desenho quer nas legendas ou títulos.

7. Serão aceitos originais processados por editores de texto tais como CW, Word, Carta Certa, WP e WS.

Impresso no CDDI/Departamento de Editoração e Gráfica — IBGE, em abril de 1993, O. S. 06.01.1.0112/93

SE O ASSUNTO É BRASIL, PROCURE O IBGE

O IBGE põe à disposição da sociedade milhares de informações de natureza estatística (demográfica, social e econômica), geográfica, cartográfica, geodésica e ambiental, que permitem conhecer a realidade física, humana, social e econômica do País.

VOCÊ PODE OBTER ESSAS PESQUISAS, ESTUDOS E LEVANTAMENTOS EM TODO O PAÍS

No Rio de Janeiro:

Centro de Documentação e Disseminação de Informações - CDDI

Divisão de Atendimento Integrado - DAT

Biblioteca Isaac Kerstenetzky

Livraria Wilson Távora

Rua General Canabarro, 666
20271-201 - Maracanã - Rio de Janeiro - RJ
Tel.: (021)284-0402
Telex: 2134128 - Fax: (021)234-6189

Livraria do IBGE

Avenida Franklin Roosevelt, 146 - loja
20021-120 - Castelo - Tel.:(021)220-9147

Nos Estados procure o
Setor de Documentação e Disseminação de Informações - SDDI
da Divisão de Pesquisa

O IBGE possui, ainda, agências localizadas nos principais municípios.