Análise de Dados Amostrais Complexos

Djalma Pessoa e Pedro Nascimento Silva

2017-09-05

Contents

Prefácio

Uma preocupação básica de toda instituição produtora de informações estatísticas é com a utilização ``'correta'' de seus dados. Isso pode ser intrepretado de várias formas, algumas delas com reflexos até na confiança do público e na própria sobrevivência do órgão. Do nosso ponto de vista, como técnicos da área de metodologia do IBGE, enfatizamos um aspecto técnico particular, mas nem por isso menos importante para os usuários dos dados.

A revolução da informática com a resultante facilidade de acesso ao computador, criou condições extremamente favoráveis à utilização de dados estatísticos, produzidos por órgãos como o IBGE. Algumas vezes esses dados são utilizados para fins puramente descritivos. Outras vezes, porém, sua utilização é feita para fins analíticos, envolvendo a construção de modelos, quando o objetivo é extrair conclusões aplicáveis também a populações distintas daquela da qual se extraiu a amostra. Neste caso, é comum empregar, sem grandes preocupações, pacotes computacionais padrões disponíveis para a seleção e ajuste de modelos. é neste ponto que entra a nossa preocupação com o uso adequado dos dados produzidos pelo IBGE.

O que torna tais dados especiais para quem pretende usá-los para fins analíticos? Esta é a questão básica que será amplamente discutida ao longo deste texto. A mensagem principal que pretendemos transmitir é que certos cuidados precisam ser tomados para utilização correta dos dados de pesquisas amostrais como as que o IBGE realiza.

O que torna especiais dados como os produzidos pelo IBGE é que estes são obtidos através de pesquisas amostrais complexas de populações finitas que envolvem: **probabilidades distintas de seleção, estratificação e conglomeração das unidades, ajustes para compensar não-resposta e outros ajustes**. Os pacotes tradicionais de análise ignoram estes aspectos, podendo produzir estimativas incorretas tanto dos parâmetros como para as variâncias destas estimativas. Quando utilizamos a amostra para estudos analíticos, as opções disponíveis nos pacotes estatísticos usuais para levar em conta os pesos distintos das observações são apropriadas somente para observações independentes e identicamente distribuídas (IID). Além disso, a variabilidade dos pesos produz impactos tanto na estimação pontual quanto na estimação das variâncias dessas estimativas, que sofre ainda influência da estratificação e conglomeração.

O objetivo deste livro é analisar o impacto das simplificações feitas ao utilizar procedimentos e pacotes usuais de análise de dados, e apresentar os ajustes necessários desses procedimentos de modo a incorporar na análise, de forma apropriada, os aspectos aqui ressaltados. Para isto serão apresentados exemplos de análises de dados obtidos em pesquisas amostrais complexas, usando pacotes clássicos e também pacotes estatísticos especializados. A comparação dos resultados das análises feitas das duas formas permitirá avaliar o impacto de ignorar o plano amostral na análise dos dados resultantes de pesquisas amostrais complexas.

Agradecimentos

A elaboração de um texto como esse não se faz sem a colaboração de muitas pessoas. Em primeiro lugar, agradecemos à Comissão Organizadora do SINAPE por ter propiciado a oportunidade ao selecionar nossa proposta de minicurso. Agradecemos também ao IBGE por ter proporcionado as condições e os meios usados

para a produção da monografia, bem como o acesso aos dados detalhados e identificados que utilizamos em vários exemplos.

No plano pessoal, agradecemos a Zélia Bianchini pela revisão do manuscrito e sugestões que o aprimoraram. Agradecemos a Marcos Paulo de Freitas e Renata Duarte pela ajuda com a computação de vários exemplos. Agradecemos a Waldecir Bianchini, Luiz Pessoa e Marinho Persiano pela colaboração na utilização do processador de textos. Aos demais colegas do Departamento de Metodologia do IBGE, agradecemos o companheirismo e solidariedade nesses meses de trabalho na preparação do manuscrito.

Finalmente, agradecemos a nossas famílias pela aceitação resignada de nossas ausências e pelo incentivo à conclusão da empreitada.

Chapter 1

Introdução

1.1 Motivação

1.2 Objetivos do Livro

1.3 Laboratório de R do Capítulo 1.

1.4 total SE DEff

1.5 analf1 1174220 127982 2.0543

1.6 total SE DEff

1.7 analf2 4792344 318877 3.3237

1.8 Ratio estimator: svyratio.survey.design2(~analf1, ~faixa1, ppv.se.des)

1.9 Ratios=

1.10 faixa1

1.11 analf1 0.118689

1.12 SEs=

1.13 faixa1

1.14 analf1 0.01178896

1.15 Ratio estimator: svyratio.survey.design2(~analf2, ~faixa2, ppv.se.des)

Chapter 2

Referencial para Inferência

2.1 Modelagem - Primeiras Idéias

2.1.1 Abordagem 1 - Modelagem Clássica

2.1.2 Abordagem 2 - Amostragem Probabilística

2.1.3 Discussão das Abordagens 1 e 2

2.1.4 Abordagem 3 - Modelagem de Superpopulação

2.2 Fontes de Variação

2.3 Modelos de Superpopulação

2.4 Planejamento Amostral

2.5 Planos Amostrais Informativos e Ignoráveis

Chapter 3

Estimação Baseada no Plano Amostral

3.1 Estimação de Totais

3.2 Por que Estimar Variâncias

3.3 Linearização de Taylor para Estimar variâncias

3.4 Método do Conglomerado Primário

3.5 Métodos de Replicação

3.6 Laboratório de R

3.7 faixa

3.8 0.01178896

3.9 Ratio estimator: svyratio.survey.design2(~analf.faixa, ~faixa, ppv.se.des)

3.10 Ratios=

3.11 faixa

3.12 analf.faixa 0.118689

3.13 SEs=

3.14 faixa

Chapter 4

Efeitos do Plano Amostral

4.1 Introdução

4.2 Efeito do Plano Amostral (EPA) de Kish

4.3 Efeito do Plano Amostral Ampliado

4.4 Intervalos de Confiança e Testes de Hipóteses

4.5 Efeitos Multivariados de Plano Amostral

4.6 Laboratório de R

4.7 Warning: package `survey' was built under R version 3.4.1

4.8 Carregando pacotes exigidos: grid

4.9 Carregando pacotes exigidos: methods

4.10 Carregando pacotes exigidos: Matrix

4.11 Carregando pacotes exigidos: survival

4.12

4.13 Attaching package: `survey'

4.14 The following object is masked from `package:graphics':

4.15

4.16 dotchart

Chapter 5

Ajuste de Modelos Paramétricos

Chapter 6

Modelos de Regressão

6.1 Modelo de Regressão Linear Normal

6.1.1 Especificação do Modelo

6.1.2 Pseudo-parâmetros do Modelo

6.1.3 Estimadores de MPV dos Parâmetros do Modelo

6.1.4 Estimação da Variância de Estimadores de MPV

6.2 Modelo de Regressão Logística

6.3 Warning in 1.96 * sqrt(var_raz112) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated.

6.4 Use c() or as.vector() instead.

6.5 Warning in 1.96 * sqrt(var_raz123) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated.

6.6 Use c() or as.vector() instead.

6.7 Warning in 1.96 * sqrt(var_raz212) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated.

6.8 Use c() or as.vector() instead.

6.9 Warning in 1.96 * sqrt(var_raz223) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated.

6.10 Use c() or as.vector() instead.

6.11 Warning in 1.96 * sqrt(var_raz312) * c(-1, 1): Recycling array

Chapter 7

Testes de Qualidade de Ajuste

7.1 Introdução

7.2 Teste para uma Proporção

7.2.1 Correção de Estatísticas Clássicas

7.3 [1] 10

7.4 [1] 20

7.5 [1] 0.5

7.6 [1] 0.4795001

7.7 [1] 10.56154

7.8 [1] 0.528077

7.9 [1] 0.4674165

7.9.1 Estatística de Wald

7.10 [1] 0.5952381

7.11 [1] 0.4404007

7.12 Teste para Várias Proporções

7.12.1 Estatística de Wald Baseada no Plano Amostral

7.12.2 Situações Instáveis

7.12.3 Estatística de Pearson com Ajuste de Rao-Scott

Chapter 8

Testes em Tabelas de Duas Entradas

8.1 Introdução

8.2 Tabelas 2x2

8.2.1 Teste de Independência

8.2.2 Teste de Homogeneidade

8.2.3 Efeitos de Plano Amostral nas Celas

8.3 Tabelas de Duas Entradas (Caso Geral)

8.3.1 Teste de Homogeneidade

8.3.2 Teste de Independência

8.3.3 Estatística de Wald Baseada no Plano Amostral

8.3.4 Estatística de Pearson com Ajuste de Rao-Scott

8.4 Laboratório de R

8.5 [1] ``stra'' ``psu'' ``pesopes'' ``informal'' ``sx'' ``id''

8.6 [7] ``ae'' ``ht'' ``re'' ``um''

8.7 stra psu pesopes informal sx id ae

8.8 ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric''

8.9 ht re um

8.10 ``numeric'' ``numeric'' ``numeric''

Chapter 9

Estimação de densidades

9.1 Introdução

Chapter 10

Modelos Hierárquicos

10.1 Introdução

Chapter 11

Não-Resposta

11.1 Introdução

Chapter 12

Diagnóstico de ajuste de modelo

12.1 Introdução

Chapter 13

Agregação vs. Desagregação

13.1 Introdução

13.2 Modelagem da Estrutura Populacional

13.3 Modelos Hierárquicos

13.4 Análise Desagregada: Prós e Contras

Chapter 14

Pacotes para Analisar Dados Amostrais

14.1 Introdução

14.2 Pacotes Computacionais

Chapter 15

Placeholder

Bibliography