Estimação de Variância com a PNADC

Método dos Conglomerados Primários e Estimador de Calibração

Guilherme Jacob

Doutorando na Escola Nacional de Ciências Estatísticas (ENCE/IBGE)

25/07/2022

Roteiro

1. Estimando Totais na PNADC
   - Estimador de Horvitz-Thompson
   - Variância do Estimador de Horvitz-Thompson
   - Método dos Conglomerados Primários
2. Calibração na PNADC
   - Pós-Estratificação
   - *Raking*
3. Implementação no `R`

Estimando Totais na PNADC

Estimando Totais com Amostras Probabilísticas

- *Amostragem probabilística*: métodos para selecionar amostras e fazer inferência de acordo com o plano amostral;
- População *finita* $\mathcal{U}$, composta por N unidades;
- Amostra $\mathbf{s} \subset \mathcal{U}$, composta por n unidades;
- Plano amostral $p(\mathbf{s})$: distribuição de probabilidade que associa todas as amostras possíveis $\mathbf{s}$ de $\mathcal{U}$ a uma probabilidade;
  - A aleatoriedade da amostra é determinada pelo plano amostral!

Estimando Totais com Amostras Probabilísticas

Quanto à seleção de amostras, isso significa selecionar amostras que atendam as probabilidades de inclusão de 1ª e 2ª ordem, π_i e π_{ij}:

- $\pi_i = \Pr(i \in \mathbf{s})$: probabilidade da unidade i ser incluída em uma amostra aleatória $\mathbf{s}$;
- $\pi_{ij} = \Pr(i, j \in \mathbf{s})$: probabilidade das unidades i e j serem incluídas conjuntamente em uma amostra aleatória $\mathbf{s}$;

Estimando Totais com Amostras Probabilísticas

Em relação à estimação, partimos do estimador de Horvitz-Thompson. O estimador de totais de Horvitz-Thompson é definido por:

$$\widehat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} d_i y_i$$

onde d_i é o *peso amostral básico*.

Estimando Totais com Amostras Probabilísticas

A variância de $\widehat{Y}_{HT}$ é dada por

$$\text{Var}\big[\widehat{Y}_{HT}\big] = \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \Delta_{ij},$$

onde $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$ é a covariância dos indicadores de inclusão na amostra das unidades i e j.

Esta variância pode ser estimada usando

$$\widehat{\text{Var}}\big[\widehat{Y}_{HT}\big] = \sum_{i \in \mathbf{s}} \sum_{j \in \mathbf{s}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \frac{\Delta_{ij}}{\pi_{ij}}.$$

Amostragem na PNADC

Plano Amostral da PNADC:

- Amostragem em Dois Estágios com Estratificação das Unidades Primárias de Amostragem (UPA):
  - Em cada estrato, seleciona uma amostra $\mathbf{s}_I$ de m UPAs com probabilidade proporcional ao tamanho (PPT de Pareto);
  - Na UPA selecionada i, seleciona uma amostra $\mathbf{s}_i$ de $\overline{n}$ domicílios com Amostragem Aleatória Simples (sem reposição).

Observações:

- Tamanho da amostra: $n = m \times \overline{n}$;
- A princípio, dentro de cada estrato, esse plano gera amostras *autoponderadas*.

Amostragem na PNADC

Observações sobre a estratificação:

- Estratos são partições da população. Por isso, o total da população é a soma dos totais dos estratos;
- Podemos estimar o total populacional pela soma das estimativas dos totais dos estratos;
- Como o sorteio dentro de cada estrato é independente, a variância do estimador do total pode ser estimada pela soma das variâncias estimadas dos totais dos estratos.

Amostragem na PNADC

Portanto, considerando os $h = 1, \ldots, H$ estratos da PNADC, podemos fazer:

$$\widehat{Y}_{\textsf{PNADC}} = \sum_{h=1}^{H} \widehat{Y}_{\textsf{PNADC},h} \implies \textsf{Var}\big[\widehat{Y}_{\textsf{PNADC}}\big] = \sum_{h=1}^{H} \textsf{Var}\big[\widehat{Y}_{\textsf{PNADC},h}\big]$$

$$\therefore \widehat{\textsf{Var}}\big[\widehat{Y}_{\textsf{PNADC}}\big] = \sum_{h=1}^{H} \widehat{\textsf{Var}}\big[\widehat{Y}_{\textsf{PNADC},h}\big]$$

Amostragem na PNADC

Assim, considerando um estrato h da PNADC, temos:

$$\mathsf{Var}\big[\widehat{Y}_{\mathsf{PNADC},h}\big] = \sum_{i\in\mathcal{U}_I}\sum_{j\in\mathcal{U}_I}\frac{Y_i}{\pi_{Ii}}\frac{Y_j}{\pi_{Ij}}\Delta_{Iij} + \sum_{i\in\mathcal{U}_I}\frac{\mathsf{Var}_{AAS}[\widehat{Y}_i]}{\pi_{Ii}}$$

$$\therefore \mathsf{Var}\big[\widehat{Y}_{\mathsf{PNADC},h}\big] = \sum_{i\in\mathcal{U}_I}\sum_{j\in\mathcal{U}_I}\frac{Y_i}{\pi_{Ii}}\frac{Y_j}{\pi_{Ij}}\Delta_{Iij} + \sum_{i\in\mathcal{U}_I}\frac{N_i^2}{\pi_{Ii}}\left(1-\frac{\overline{n}}{N_i}\right)S_i^2,$$

onde:

- π_{Ii} é a probabilidades de inclusão de 1ª ordem da UPA i;
- Y_i e $\widehat{Y}_i$ são o valor e a estimativa do total de y na UPA i, respectivamente;
- N_i é o número de domicílios na UPA i;
- S_i^2 é a variância populacional da característica y na UPA i.

Amostragem na PNADC

O estimador não-viesado desta variância é:

$$\widehat{\mathsf{Var}}[\widehat{Y}_{\mathsf{PNADC},h}] = \sum_{i \in \mathbf{s}_I} \sum_{j \in \mathbf{s}_I} \frac{\widehat{Y}_i}{\pi_{Ii}} \frac{\widehat{Y}_j}{\pi_{Ij}} \frac{\Delta_{Iij}}{\pi_{Iij}} + \sum_{i \in \mathbf{s}_I} \frac{N_i^2}{\pi_{Ii}} \left(1 - \frac{\overline{n}}{N_i}\right) s_i^2,$$

onde:

- π_{Iij} é a probabilidade de inclusão de 2ª ordem das UPAs i e j na amostra do primeiro estágio $\mathbf{s}_I$;
- s_i^2 é a variância da característica y na amostra $\mathbf{s}_i$ de domicílios da UPA i.

O problema é que essa fórmula exigiria conhecer uma matriz quadrada π_{Iij} de tamanho $m \times m$, o que não é trivial.

Método dos Conglomerados Primários

Antes de adotar o método do *Bootstrap*, a variância do estimador $\widehat{Y}$ se baseava no Método dos Conglomerados Primários (MCP):

$$\widehat{\mathsf{Var}}_{\mathsf{MCP}}[\widehat{Y}_{\mathsf{PNADC},h}] = \frac{m}{m-1} \sum_{i \in \mathbf{s}_I} \left(\frac{\widehat{Y}_i}{\pi_{Ii}} - \frac{\widehat{Y}_{\mathsf{PNADC},h}}{m} \right)^2$$

onde m é o número de UPAs selecionadas no estrato h.

Esta é uma das razões para **não excluir observações** de uma base de dados amostrais complexos.

Calibração na PNADC

Calibração

Seja $\mathbf{X} = \sum_{i \in \mathcal{U}} \mathbf{x}_i$ um vetor de J totais conhecidos na população. Métodos de *calibração* buscam criar um sistema de pesos w_i que, independente de qual amostra foi selecionada, satisfaça a restrição de calibração:

$$\sum_{i \in \mathbf{s}} w_i \mathbf{x}_i = \sum_{i \in \mathcal{U}} \mathbf{x}_i$$

Calibração

Resultados da Calibração:

- Estimativas baseadas em variáveis correlacionadas com as variáveis usadas na calibração se tornam mais precisas;
- *Pode* reduzir viés de não-resposta e cobertura;
- Faz as estimativas serem “consistentes” com totais conhecidos da população.

Pós-estratificação

A técnica mais simples para ajustar pesos é a *pós-estratificação*:

- Suponha que possamos dividir a população em K grupos;
- Se os tamanhos N_k dos K grupos forem conhecidos e amostra não for estratificada, os totais estimados $\widehat{N}_k$ não coincidem com os valores populacionais;
- Mas podemos usar esta informação para corrigir os pesos.

Pós-estratificação

Ou seja: podemos fazer realizar ajustes $g_i = N_k/\widehat{N}_k$ de modo que

$$\begin{aligned}\sum_{i\in\mathbf{s}_k}\frac{g_i}{\pi_i} &= \sum_{i\in\mathbf{s}_k}\frac{1}{\pi_i}\frac{N_k}{\widehat{N}_k}\\ &= \frac{N_k}{\widehat{N}_k}\sum_{i\in\mathbf{s}_k}\frac{1}{\pi_i} = \frac{N_k}{\widehat{N}_k}\widehat{N}_k = N_k\end{aligned}$$

Isso significa substituir os pesos básicos $d_i = 1/\pi_i$ pelos pesos de pós-estratificação $w_i = g_i/\pi_i = N_k/(\widehat{N}_k\pi_i)$.

Pós-estratificação

Seja μ_k a média populacional de uma variável y_i no grupo k e $\mu_{k(i)}$ a média populacional do grupo que contem a observação i. Assim, fazendo $y_i = (y_i - \mu_{k(i)}) + \mu_{k(i)}$, temos:

$$\widehat{Y} = \sum_{i \in \mathbf{s}} \frac{g_i}{\pi_i} \left[(y_i - \mu_{k(i)}) + \mu_{k(i)} \right] = \sum_{i \in \mathbf{s}} \frac{g_i}{\pi_i} (y_i - \mu_{k(i)}) + \sum_{i \in \mathbf{s}} \frac{N_{k(i)}}{\widehat{N}_{k(i)}} \frac{\mu_{k(i)}}{\pi_i}$$

$$\widehat{Y} = \sum_{i \in \mathbf{s}} \frac{g_i}{\pi_i} (y_i - \mu_{k(i)}) + \sum_{k=1}^{K} \frac{N_k}{\widehat{N}_k} \widehat{N}_k \mu_k$$

$$\widehat{Y} = \sum_{i \in \mathbf{s}} \frac{g_i}{\pi_i} (y_i - \mu_{k(i)}) + \sum_{k=1}^{K} N_k \mu_k, \quad \sum_{k=1}^{K} N_k \mu_k \text{ const.}$$

$$\therefore \widehat{Y} = \sum_{i \in \mathbf{s}} \frac{g_i}{\pi_i} e_i + \sum_{k=1}^{K} N_k \mu_k, \quad e_i = y_i - \mu_{k(i)}$$

Pós-estratificação

Como o segundo termo é uma constante, sua variância é zero. Logo, podemos estimar a variância aproximada usando

$$\mathsf{Var}\left[\widehat{Y}\right] \approx \mathsf{Var}\left[\sum_{i \in \mathbf{s}} \frac{g_i e_i}{\pi_i}\right]$$

$$\therefore \widehat{\mathsf{Var}}\left[\widehat{Y}\right] = \sum_{i \in \mathbf{s}} \sum_{j \in \mathbf{s}} \frac{g_i e_i}{\pi_i} \frac{g_j e_j}{\pi_j} \frac{\Delta_{ij}}{\pi_{ij}}$$

Ou seja: podemos estimar a variância do estimador de pós-estratificação usando o estimador da variância do total dos resíduos.

Raking

- O estimador de pós-estratificação exige totais conhecidos e tamanhos de amostra suficientemente grandes em cada grupo;
- Para pós-estratificar por área e idade, precisamos dos totais de todas as combinações de área e faixa etária;
- Quanto maior o número de grupos, menor é o tamanho da amostra em cada grupo;
- *Raking* é um método de calibração que permite usar uma abordagem mais parcimoniosa, calibrando pelas "marginais";
  - No exemplo, totais da população de cada área e totais de cada faixa etária na população geral.

Raking

- O *raking* é implementado como um processo iterativo, "pós-estratificando" sucessivamente em cada vetor de totais até que os totais estimados sejam (aproximadamente) iguais aos totais de calibração;
  - *Iterative Proportional Fitting* (IPF).
- Como na pós-estratificação, a variância do total pode ser aproximada pela variância do total dos resíduos e_i;
- Por sua vez, os resíduos e_i são calculados pelo processo iterativo, subtraindo as médias dos pós-estratos de cada vetor.

Raking Generalizado

O método de *Raking* pode ser ampliado para, por exemplo, incluir:

- Restrições sobre a distância em relação ao peso amostral original;
- Restrição de igualdade de pesos em determinado estágio de conglomeração.

Em geral, a variância do total pode ser estimada pelo estimador da variância do total dos resíduos, com a fórmula dos resíduos mudando de acordo com o método.

Usando o pacote `survey`

Criando o objeto de plano amostral

Com a base de dados da amostra (completa!), declaramos o plano amostral fazendo:

```
# carregando a library
library( survey )

# ajusta formatos dos domínios de calibração
pnadc.df[ , c( "posest" , "posest_sxi" ) ] <-
  lapply( pnadc.df[ , c( "posest" , "posest_sxi" ) ] , factor )

# criando o objeto de plano amostral
pnadc.mcp <-
  svydesign( ids = ~upa+v1008 ,
             strata = ~estrato ,
             data = pnadc.df ,
             weights = ~v1027 )
```

Aplica calibração

Para aplicar a calibração, é preciso fornecer os totais populacionais. Estes totais estão dispostos na própria base de dados da PNADC:

```
# coleta tabela com os totais das marginais na população
pop.posest <-
  pnadc.df[ !duplicated( pnadc.df$posest ) , c( "posest" , "v1029" ) ]
pop.posest_sxi <-
  pnadc.df[ !duplicated( pnadc.df$posest_sxi ) ,
              c( "posest_sxi" , "v1033" ) ]
pop.posest <-
  pop.posest[ order( pop.posest$posest ) , ]
pop.posest_sxi <-
  pop.posest_sxi[ order( pop.posest_sxi$posest_sxi ) , ]

# ajusta nome das colunas de frequências
colnames( pop.posest )[2] <-"Freq"
colnames( pop.posest_sxi )[2] <-"Freq"
```

Aplica calibração

Dispondo das tabelas de totais, passamos esta informação para a função `calibrate`. A opção `calfun = "raking"` determina a função de pseudo-distância.

```
# aplica calibração via raking com limites e pesos fixos no domicílio
pnadc.mcp.calib <-
  calibrate( pnadc.mcp ,
             formula = list( ~posest , ~posest_sxi ) ,
             population = list( pop.posest, pop.posest_sxi ) ,
             bounds = c(.2,5) ,
             bounds.const = FALSE ,
             calfun = "raking" ,
             aggregate.stage = 2 )
```

Criando o objeto de plano amostral com pesos de replicação

Para criar o objeto de plano amostral com os pesos de calibração e de replicação (calibrados!) fornecidos pelo IBGE, fazemos:

```
# criando o objeto de plano amostral
# com pesos de replicação
pnadc.boot <-
  svrepdesign( data = pnadc.df ,
               type = "bootstrap" ,
               weights = ~v1028 ,
               repweights = "^v1028[0-9]{3}" ,
               mse = TRUE )
```

Verificando os Totais “Calibrados”

Estimador de Horvitz-Thompson e Variância pelo MCP:

```
svytotal( ~factor( v2007 ) , pnadc.mcp )
```

```
##                    total     SE
## factor(v2007)1 81519302 445533
## factor(v2007)2 89931244 466546
```

Estimador de Calibração e Variância pelo MCP:

```
svytotal( ~factor( v2007 ) , pnadc.mcp.calib )
```

```
##                     total SE
## factor(v2007)1 104020393  0
## factor(v2007)2 108787836  0
```

Verificando os Totais “Calibrados”

Estimador de Calibração e Variância pelo MCP:

```
svytotal( ~factor( v2007 ) , pnadc.mcp.calib )
```

```
##                       total SE
## factor(v2007)1 104020393  0
## factor(v2007)2 108787836  0
```

Estimador de Calibração e Variância por *Bootstrap*:

```
svytotal( ~factor( v2007 ) , pnadc.boot )
```

```
##                       total     SE
## factor(v2007)1 104020393 0.1207
## factor(v2007)2 108787836 0.0998
```

Estimando Médias

Estimador de Horvitz-Thompson e Variância pelo MCP:

```
svymean( ~v2009 , pnadc.mcp )
```

```
##         mean     SE
## v2009 37.63 0.0778
```

Estimador de Calibração e Variância pelo MCP:

```
svymean( ~v2009 , pnadc.mcp.calib )
```

```
##          mean     SE
## v2009 34.843 0.0035
```

Estimando Médias

Estimador de Calibração e Variância pelo MCP:

```
svymean( ~v2009 , pnadc.mcp.calib )
```

```
##         mean     SE
## v2009 34.843 0.0035
```

Estimador de Calibração e Variância por *Bootstrap*:

```
svymean( ~v2009 , pnadc.boot )
```

```
##         mean     SE
## v2009 34.843 0.0035
```

Estimando Totais

Estimador de Horvitz-Thompson e Variância pelo MCP:

```
svytotal( ~factor( vd4002 ) , pnadc.mcp , na.rm = TRUE )
```

```
##                      total     SE
## factor(vd4002)1 73881880 427835
## factor(vd4002)2 10657864 138799
```

Estimador de Calibração e Variância pelo MCP:

```
svytotal( ~factor( vd4002 ) , pnadc.mcp.calib , na.rm = TRUE )
```

```
##                      total     SE
## factor(vd4002)1 92976446 228121
## factor(vd4002)2 13453390 148898
```

Estimando Totais

Estimador de Calibração e Variância pelo MCP:

```
svytotal( ~factor( vd4002 ) , pnadc.mcp.calib , na.rm = TRUE )
```

```
##                      total     SE
## factor(vd4002)1 92976446 228121
## factor(vd4002)2 13453390 148898
```

Estimador de Calibração e Variância por *Bootstrap*:

```
svytotal( ~factor( vd4002 ) , pnadc.boot , na.rm = TRUE )
```

```
##                      total     SE
## factor(vd4002)1 92976446 218402
## factor(vd4002)2 13453390 156700
```

Impacto da Calibração na Estimação de Variância

Também é possível criar um novo objeto de plano amostral que ignora a calibração na estimação de variância.

Estimador de Calibração e Variância pelo MCP (sobre os resíduos):

```
svytotal( ~factor( vd4002 ) , pnadc.mcp.calib , na.rm = TRUE )
```

```
##                       total     SE
## factor(vd4002)1 92976446 228121
## factor(vd4002)2 13453390 148898
```

Estimador de Calibração e Variância pelo MCP (sobre a variável de interesse):

```
svytotal( ~factor( vd4002 ) , pnadc.mcp.calib.nps , na.rm = TRUE )
```

```
##                       total     SE
## factor(vd4002)1 92976446 551092
## factor(vd4002)2 13453390 180267
```

Conclusão

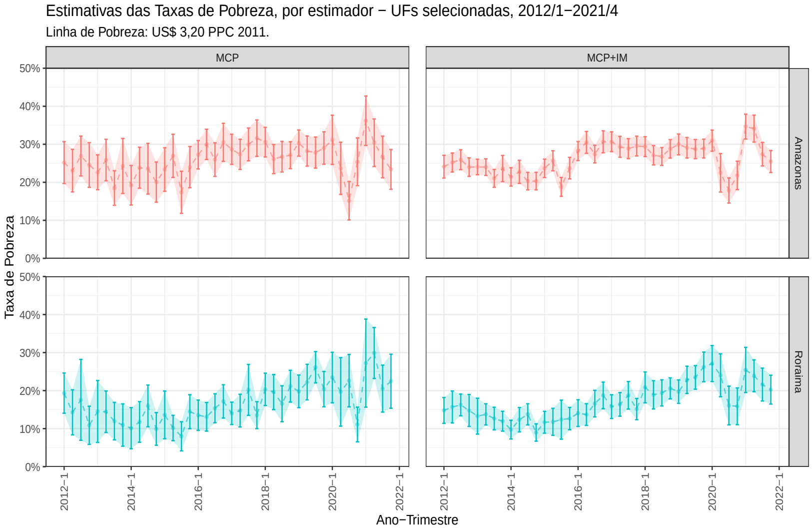
Conclusão

- Ignorar os pesos calibrados pode introduzir viés nas estimativas;
- Em geral, ignorar a calibração gera perda de eficiência nas estimativas;
- Embora o método de *Bootstrap* simplifique a implementação em outros softwares, usar o estimador de calibração com MCP com o pacote `survey` é bastante simples;

Conclusão

- *Bootstrap* é computacionalmente intensivo;
  - Seu uso combinado com Imputação Múltipla (IM) pode ser proibitivo, por exemplo.
  - Já MCP e IM é uma abordagem viável.
- A estratégia "MCP+IM" permitiu aumentar a precisão das estimativas trimestrais da taxa de pobreza nas UFs de 2012-2021.

Conclusão



Estimativas das Taxas de Pobreza, por estimador − UFs selecionadas, 2012/1−2021/4
Linha de Pobreza: US$ 3,20 PPC 2011.
MCP
MCP+IM
Amazonas
Roraima
Taxa de Pobreza
0%
10%
20%
30%
40%
50%
2012−1
2014−1
2016−1
2018−1
2020−1
2022−1
Ano−Trimestre