Análise de Dados Amostrais Complexos

Djalma Pessoa e Pedro Nascimento Silva

2017-02-03

Sumário

Prefácio

Capítulo 1

Introdução

Capítulo 2

Referencial para Inferência

Capítulo 3

Estimação Baseada no Plano Amostral

Capítulo 4

Efeitos do Plano Amostral

Capítulo 5

Ajuste de Modelos Paramétricos

5.1 Introdução

Nos primórdios do uso `moderno` de pesquisas por amostragem, os dados obtidos eram usados principalmente para estimar funções simples dos valores das variáveis de interesse nas populações finitas, tais como totais, médias, razões, etc. Isto caracterizava o uso dos dados dessas pesquisas para `inferência descritiva`. Recentemente, os dados de pesquisas amostrais têm sido cada vez mais utilizados também para propósitos analíticos. `Inferências analíticas` baseadas numa pesquisa amostral são aquelas que envolvem a estimação de parâmetros num modelo (de superpopulação) (Kalton, 1983); (Binder et al., 1987).

Quando os valores `amostrais` das variáveis da pesquisa podem ser considerados como realizações de vetores aleatórios independentes e identicamente distribuídos (IID), modelos podem ser especificados, ajustados, testados e reformulados usando procedimentos estatísticos padrões como os apresentados, por exemplo, em (Bickel and Doksum, 1977) e (Garthwaite et al., 1995). Neste caso, métodos e pacotes estatísticos padrões podem ser usados para executar os cálculos de estimativas de parâmetros e medidas de precisão correspondentes, bem como diagnóstico e verificação da adequação das hipóteses dos modelos.

Na prática das pesquisas amostrais, contudo, as hipóteses de modelo IID para as observações amostrais são raramente adequadas. Com maior frequência, modelos alternativos com hipóteses mais complexas e/ou estimadores especiais devem ser considerados a fim de acomodar aspectos da estrutura da população e/ou do plano amostral. Além disso, usualmente estão disponíveis informações sobre variáveis auxiliares, utilizadas ou não na especificação do plano amostral, que podem ser incorporadas com proveito na estima ção dos parâmetros ou na própria formulação do modelo.

Os exemplos apresentados no Capítulo **??** demonstram claramente a inadequa ção de ignorar o plano amostral ao efetuar análises de dados de pesquisas amostrais. Os valores dos EPAs calculados, tanto para estimadores de medidas descritivas tais como médias e totais, como para estatísticas analíticas usadas em testes de hipóteses e os correspondentes efeitos nos níveis de significância reais, revelam que ignorar o plano amostral pode levar a decisões erradas e a avaliações inadequadas da precisão das estimativas amostrais.

Embora as medidas propostas no Capítulo 4 para os efeitos de plano amostral sirvam para avaliar o impacto de ignorar o plano amostral nas inferências descritivas ou mesmo analíticas baseadas em dados amostrais, elas não resolvem o problema de como incorporar o plano amostral nessas análises. No caso das inferências descritivas usuais para médias, totais e proporções, o assunto é amplamente tratado na literatura de amostragem e o interessado em maiores detalhes pode consultar livros clássicos como (Cochran, 1977), ou mais recentes como (Särndal et al., 1992). Já os métodos requeridos para inferências analíticas só recentemente foram consolidados em livro ((Skinner et al., 1989)). Este capítulo apresenta um dos métodos centrais disponíveis para ajuste de modelos paramétricos regulares considerando dados amostrais complexos, baseado no trabalho de (Binder et al., 1987). Antes de descrever esse método, entretanto, fazemos breve discussão sobre o papel dos pesos na análise de dados amostrais, considerando o trabalho de (Pfeffermann, 1993).

Primeiramente, porém, fazemos uma revisão sucinta do método de Máxima Verossimilhança (MV) para ajustar modelos dentro da abordagem de modelagem clássica, necessária para compreensão adequada do material subseqüente. Essa revisão não pretende ser exaustiva ou detalhada, mas tão somente recordar os principais resultados aqui requeridos. Para uma discussão mais detalhada do método de Máxima Verossimilhança para estimação em modelos paramétricos regulares veja, por exemplo, (Garthwaite et al., 1995).

5.2 Método de Máxima Verossimilhança (MV)

Seja $\mathbf{y}_i = (y_{i1}, \ldots, y_{iR})'$ um vetor $R \times 1$ dos valores observados das variáveis de interesse observadas para a unidade i da amostra, gerado por um vetor aleatório $\mathbf{Y}_i$, para $i = 1, \ldots, n$, onde n é o tamanho da amostra. Suponha que os vetores aleatórios $\mathbf{Y}_i$, para $i = 1, \ldots, n$, são independentes e identicamente distribuídos (IID) com distribuição comum $f(\mathbf{y}; \theta)$, onde $\theta = (\theta_1, \ldots, \theta_K)'$ é um vetor $K \times 1$ de parâmetros desconhecidos de interesse. Sob essas hipóteses, a verossimilhança amostral é dada por

$$l(\theta) = \prod_{i=1}^{n} f(\mathbf{y}_i; \theta)$$

e a correspondente log-verossimilhança por

$$L(\theta) = \sum_{i=1}^{n} \log\left[f(\mathbf{y}_i; \theta)\right] \ .$$

Calculando as derivadas parciais de $L(\theta)$ com relação a cada componente de θ e igualando a 0, obtemos um sistema de equações

$$\partial L(\theta)/\partial\theta = \sum_{i=1}^{n} \mathbf{u}_i(\theta) = \mathbf{0},$$

onde, $\mathbf{u}_i(\theta) = \partial \log\left[f(\mathbf{y}_i; \theta)\right]/\partial\theta$ é o vetor dos escores da unidade i, de dimensão $K \times 1$.

Sob condições de regularidade p. 281 (Cox and Hinkley, 1974), a solução $\hat{\theta}$ deste sistema de equações é o **Estimador de Máxima Verossimilhança (EMV)** de θ. A variância assintótica do estimador $\hat{\theta}$ sob o modelo adotado, denominado aqui abreviadamente modelo \$M \$, é dada por

$$V_M\left(\hat{\theta}\right) \simeq \left[J(\theta)\right]^{-1}$$

e um estimador consistente dessa variância é dado por

$$\hat{V}_M\left(\hat{\theta}\right) = \left[J\left(\hat{\theta}\right)\right]^{-1} \ ,$$

onde

$$J(\theta) = \sum_{i=1}^{n} \partial \mathbf{u}_i(\theta)/\partial\theta$$

e

$$J\left(\hat{\theta}\right) = J(\theta)|_{\theta=\hat{\theta}} \ .$$

5.3 Ponderação de Dados Amostrais

O papel da ponderação `na análise de dados amostrais` é alvo de controvérsia entre os estatísticos. Apesar de incorporada comumente na inferência descritiva, não há concordância com respeito a seu uso na inferência analítica, havendo um espectro de opiniões entre dois extremos. Num extremo estão os `modelistas`, que consideram o uso de pesos irrelevante, e no outro os `amostristas`, que incorporam pesos em qualquer análise.

Exemplo 5.1 *Uso analítico dos dados da Pesquisa Nacional por Amostra de Domicílios (PNAD)*

A título de ilustração, consideremos uma pesquisa com uma amostra complexa como a da PNAD do IBGE, que emprega uma amostra estratificada de domicílios em três estágios, tendo como unidades primárias de amostragem (UPAs) os municípios, que são estratificados segundo as unidades da federação (UFs), e regiões menores dentro das UFs (veja IBGE, 1981, p. 67).

A seleção de municípios dentro de cada estrato é feita com probabilidades desiguais, proporcionais ao tamanho, havendo inclusive municípios incluídos na amostra com certeza (chamados de municípios auto-representativos). Da mesma forma, a seleção de setores (unidades secundárias de amostragem ou USAs) dentro de cada município é feita com probabilidades proporcionais ao número de domicílios em cada setor segundo o último censo disponível. Dentro de cada setor, a seleção de domicílios é feita por amostragem sistemática simples (portanto, com eqüiprobabilidade). Todas as pessoas moradoras em cada domicílio da amostra são pesquisadas.

A amostra de domicílios e de pessoas dentro de cada estrato é `autoponderada`, isto é, tal que todos os domicílios e pessoas dentro de um mesmo estrato têm igual probabilidade de seleção. Entretanto, as probabilidades de inclusão (e conseqüentemente os pesos) variam bastante entre as várias regiões de pesquisa. A Tabela 5.1 revela como variam essas probabilidades de seleção entre as regiões cobertas pela amostra da PNAD de 93. Como se pode observar, tais probabilidades de inclusão chegam a ser 5 vezes maiores em Belém do que em São Paulo, e portanto variação semelhante será observada nos pesos.

Tabela 5.1: Probabilidades de seleção da amostra da PNAD de 1993 segundo regiões

Região da pesquisa	Probabilidade de seleção
RM de Belém	1/150
RMs de Fortaleza, Recife, Salvador e Porto Alegre Distrito Federal	1/200
RMs de Belo Horizonte e Curitiba	1/250
Rondô}nia, Acre, Amazonas, Roraima, Amapá, Tocantins, Sergipe, Mato Grosso do Sul, Mato Grosso e Goiás	1/300
Pará	1/350
RM do Rio de Janeiro, Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Bahia, Minas Gerais, Espírito Santo e Rio de Janeiro	1/500
Paraná, Santa Catarina, Rio Grande do Sul	1/550
RM de São Paulo, Maranhão, São Paulo	1/750

Se $\pi_i$representa a probabilidade de inclusão na amostra do i-ésimo domicílio da população, $i = 1, ..., N$, então

$$\pi_i = \pi_{município|estrato} \times \pi_{setor|município} \times \pi_{domicílio|setor}$$

isto é, a probabilidade global de inclusão de um domicílio (e conseqüentemente de todas as pessoas nele moradoras) é dada pelo produto das probabilidades condicionais de inclusão nos vários estágios de amostragem.

A estimação do total populacional Y de uma variável de pesquisa y num dado estrato usando os dados da PNAD é feita rotineiramente com estimadores ponderados de tipo razão $\widehat{Y}_R = \widehat{Y}_\pi \, X \, / \, \widehat{X}_\pi = \sum_{i \in s} w_i^R y_i$ (tal como definidos por (**??**), com pesos dados por $w_i^R = \pi_i^{-1} X \, / \, \widehat{X}_\pi$ (veja (**??**), onde X é o total da população no estrato obtido por métodos demográficos de projeção, utilizado como variável auxiliar, e $\widehat{X}_\pi$ e $\widehat{Y}_\pi$ são os estimadores π-ponderados de X e Y respectivamente. Para estimar para conjuntos de estratos basta somar as estimativas para cada estrato incluído no conjunto. Para estimar médias e proporções, os pesos são também incorporados da forma apropriada. No caso, a estimação de médias é feita usando estimadores ponderados da forma

$$\overline{y}^R = \frac{\sum_{i \in s} w_i^R y_i}{\sum_{i \in s} w_i^R}$$

e a estimação de proporções é caso particular da estimação de médias quando a variável de pesquisa y é do tipo indicador (isto é, só toma valores 0 e 1).

Estimadores ponderados (como por exemplo os usados na PNAD) são preferidos pelos praticantes de amostragem por sua simplicidade e por serem não viciados (ao menos aproximadamente) com respeito à distribuição de aleatorização induzida pela seleção da amostra, independentemente dos valores assumidos pelas variáveis de pesquisa na população. Já para a modelagem de relações entre variáveis de pesquisa, o uso dos pesos induzidos pelo planejamento amostral ainda não é freqüente ou aceito sem controvérsia.

Um exemplo de modelagem desse tipo com dados da PNAD em que os pesos e o desenho amostral não foram considerados na análise é encontrado em (Leote, 1996). Essa autora empregou modelos de regressão logística para traçar um perfil sócio-econômico da mão-de-obra empregada no mercado informal de trabalho urbano no Rio de Janeiro, usando dados do suplemento sobre trabalho da PNAD-90. Todos os ajustes efetuados ignoraram os pesos e o plano amostral da pesquisa. O problema foi revisitado por (Pessoa et al., 1997), quando então esses aspectos foram devidamente incorporados na análise. Um resumo desse trabalho é discutido no Capítulo 6.

Vamos supor que haja interesse em regredir uma determinada variável de pesquisa y contra algumas outras variáveis de pesquisa num vetor de regressores $\mathbf{z}$. Seria natural indagar se, como no caso do total e da média, os pesos amostrais poderiam desempenhar algum papel na estimação dos parâmetros do modelo (linear) de regressão. Uma possibilidade de incluir os pesos seria estimar os coeficientes da regressão por:

$$\widehat{\beta}_w = \left(\sum_{i \in s} w_i \mathbf{z}_i' \mathbf{z}_i \right)^{-1} \sum_{i \in s} w_i \mathbf{z}_i' y_i = \left(\mathbf{Z}_s' \mathbf{W}_s \mathbf{Z}_s \right)^{-1} \mathbf{Z}_s' \mathbf{W}_s \mathbf{Y}_s \tag{5.1}$$

em lugar do estimador de mínimos quadrados ordinários (MQO) dado por

$$\widehat{\beta} = \left(\sum_{i \in s} \mathbf{z}_i' \mathbf{z}_i \right)^{-1} \sum_{i \in s} \mathbf{z}_i' y_i = \left(\mathbf{Z}_s' \mathbf{Z}_s \right)^{-1} \mathbf{Z}_s' \mathbf{Y}_s \tag{5.2}$$

onde $w_i = \pi_i^{-1}$, y_i é o valor da variável resposta e $\mathbf{z}_i$ é o vetor de regressores para a observação i, $\mathbf{Z}_s$ e $\mathbf{Y}_s$ são respectivamente a matriz e vetor com os valores amostrais dos $\mathbf{z}_i$ e y_i, e $\mathbf{W}_s = diag\{w_i; i \in s\}$ é a matriz diagonal com os pesos amostrais.

Não é possível justificar o estimador $\widehat{\beta}_w$ em (5.1) com base em critério de otimalidade, tal como ocorre com os estimadores usuais de Máxima Verossimilhança ou de Mínimos Quadrados Ordinários (MQO), se uma modelagem clássica IID fosse adotada para a amostra.

De um ponto de vista formal (matemático), o estimador $\widehat{\beta}_w$ em (5.1) é equivalente ao estimador de Mínimos Quadrados Ponderados (MQP) com pesos w_i. Entretanto, esses estimadores diferem de maneira acentuada. Os estimadores de MQP são usualmente considerados quando o modelo de regressão é heteroscedástico, isto é, quando os resíduos têm variâncias desiguais. Nes-te caso, os pesos adequados seriam dados pelos inversos das variâncias dos resíduos correspondentes a cada uma das observações, e portanto em geral diferentes dos pesos iguais aos inversos das correspondentes probabilidades de seleção. Além desta diferença de interpretação do papel dos pesos no estimador, outro aspecto em que os dois estimadores diferem de forma acentuada é na estimação da precisão, com o estimador MQP acoplado a um estimador de variância baseado no modelo e o estimador $\widehat{\beta}_w$ acoplado a estimadores de variância que incorporam o planejamento amostral e os pesos, tal como se verá mais adiante.

O estimador $\widehat{\beta}_w$ foi proposto formalmente por (Fuller, 1975), que o concebeu como uma função de estimadores de totais populacionais. A mesma idéia subsidiou vários outros autores que estudaram a estimação de coeficientes de regressão partindo de dados amostrais complexos, tais como (Nathan and Holt, 1980), (Pfeffermann and Nathan, 1981). Uma revisão abrangente da literatura existente sobre estimação de parâmetros em modelos de regressão linear com dados amostrais complexos pode ser encontrada em cap. 6, (Nascimento Silva, 1996).

Apesar dessas dificuldades, será que é possível justificar o uso de pesos na inferência baseada em modelos? Se for o caso, sob que condições? Seria possível desenvolver diretrizes para o uso de pesos em inferência analítica partindo de dados amostrais complexos? A resposta para essas perguntas é afirmativa, ao menos quando a questão da robustez da inferência é relevante. Em inferências analíticas partindo de dados amostrais complexos, os pesos podem ser usados para proteger:

1. contra `planos amostrais não-ignoráveis`, que poderiam introduzir ou causar `vícios`;
2. contra a `má especificação do modelo`.

A `robustez dos procedimentos` que incorporam pesos é obtida pela mudança de foco da inferência para `quantidades da população finita`, que definem parâmetros-alvo alternativos aos parâmetros do modelo de superpopulação, conforme já discutido na Seção **??**.

A questão da construção dos pesos não será tratada neste texto, usando-se sempre como peso o inverso da probabilidade de inclusão na amostra. é possível utilizar pesos de outro tipo como, por exemplo, aqueles de razão empregados na estimação da PNAD, ou mesmo pesos de regressão. Para esses casos, há que fazer alguns ajustes da teoria aqui exposta (veja (Nascimento Silva, 1996), cap. 6).

Há várias formas alternativas de incorporar os pesos amostrais no processo de inferência. A principal que será adotada ao longo deste texto será o método de `Máxima Pseudo-Verossimilhança`, que descrevemos na próxima seção.

5.4 Método de Máxima Pseudo-Verossimilhança label

Suponha que os vetores observados $\mathbf{y}_i$ das variáveis de pesquisa do elemento i são gerados por vetores aleatórios $\mathbf{Y}_i$, para $i \in U$. Suponha também que $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ são IID com densidade $f(\mathbf{y}, \theta)$. Se todos os elementos da população finita U fossem conhecidos, as funções de verossimilhança e de log-verossimilhança `populacionais` seriam dadas respectivamente por

$$l_U(\theta) = \prod_{i \in U} f(\mathbf{y}_i; \theta) \tag{5.3}$$

e

$$L_U(\theta) = \sum_{i \in U} \log\left[f(\mathbf{y}_i; \theta)\right] \quad . \tag{5.4}$$

As equações de verossimilhança `populacionais` correspondentes são dadas por

$$\sum_{i \in U} \mathbf{u}_i(\theta) = \mathbf{0} \tag{5.5}$$

onde

$$\mathbf{u}_i(\theta) = \partial \log\left[f(\mathbf{y}_i; \theta)\right] / \partial\theta \tag{5.6}$$

é o vetor $K \times 1$ dos escores do elemento $i, i \in U$.

Sob condições de regularidade (Cox and Hinkley, 1974), p. 281, a solução θ_U deste sistema é o `Estimador de Máxima Verossimilhança` de θ no caso de um `censo`. Podemos considerar θ_U como uma `Quantidade Descritiva Populacional Correspondente (QDPC)` a θ, no sentido definido por (Pfeffermann, 1993), sobre a qual se deseja fazer inferências com base em informações da amostra. Essa definição da QDPC θ_U pode ser generalizada para contemplar outras abordagens de inferência além da abordagem clássica baseada em

maximização da verossimilhança. Basta para isso especificar outra regra ou critério a otimizar e então definir a QDPC como a solução ótima segundo essa nova regra. Tal generalização, discutida em (Pfeffermann, 1993), não será aqui considerada para manter a simplicidade.

A QDPC θ_U definida com base em (5.5) não é calculável a menos que um censo seja realizado. Entretanto, desempenha papel fundamental nessa abordagem inferencial, por constituir-se num `pseudo-parâmetro`, eleito como alvo da inferência num esquema que incorpora o planejamento amostral. Isto se justifica porque, sob certas condições de regularidade, $\theta_U - \theta = o_p(1)$. Como em pesquisas por amostragem o tamanho da população é geralmente grande, um estimador adequado para θ_U será geralmente adequado também para θ.

Seja $\mathbf{T} = \sum_{i \in U} \mathbf{u}_i(\theta)$ a soma dos vetores de escores na população, o qual é um vetor de totais populacionais. Para estimar este vetor de totais, podemos então usar um estimador linear ponderado da forma $\hat{\mathbf{T}} = \sum_{i \in s} w_i \mathbf{u}_i(\theta)$ (veja Capítulo **??**) onde w_i são pesos propriamente definidos. Com essa notação, podemos agora obter um estimador para θ_U resolvendo o sistema de equações obtido igualando o estimador $\hat{\mathbf{T}}$ do total $\mathbf{T}$ a zero.

Definição 5.1 *O estimador de Máxima Pseudo-Verossimilhança (MPV) $\hat{\theta}_{MPV}$ de θ_U (e consequentemente de θ) será a solução das equações dePseudo-Verossimilhança dadas por*

$$\hat{\mathbf{T}} = \sum_{i \in s} w_i \mathbf{u}_i(\theta) = \mathbf{0} \ . \tag{5.7}$$

Através da linearização de Taylor (veja Seção **??** e considerando os resultados de (Binder, 1983), podemos obter a variância de aleatorização assintótica do estimador $\hat{\theta}_{MPV}$ e seu estimador correspondente, dados respectivamente por:

$$V_p\left(\hat{\theta}_{MPV}\right) \simeq \left[J(\theta_U)\right]^{-1} V_p\left[\sum_{i \in s} w_i \mathbf{u}_i(\theta_U)\right] \left[J(\theta_U)\right]^{-1} \tag{5.8}$$

e

$$\hat{V}_p\left(\hat{\theta}_{MPV}\right) = \left[\hat{J}\left(\hat{\theta}_{MPV}\right)\right]^{-1} \hat{V}_p\left[\sum_{i \in s} w_i \mathbf{u}_i\left(\hat{\theta}_{MPV}\right)\right] \left[\hat{J}\left(\hat{\theta}_{MPV}\right)\right]^{-1} , \tag{5.9}$$

onde

$$J(\theta_U) = \left.\frac{\partial T(\theta)}{\partial \theta}\right|_{\theta=\theta_U} = \left.\sum_{i \in U} \frac{\partial \mathbf{u}_i(\theta)}{\partial(\theta)}\right|_{\theta=\theta_U} , \tag{5.10}$$

$$\hat{J}\left(\hat{\theta}_{MPV}\right) = \left.\frac{\partial \widehat{T}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{MPV}} = \left.\sum_{i \in s} w_i \frac{\partial \mathbf{u}_i(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{MPV}} , \tag{5.11}$$

$V_p\left[\sum_{i \in s} w_i \mathbf{u}_i(\theta_U)\right]$ é a matriz de variância (de aleatorização) do estimador do total populacional dos escores e $\hat{V}_p\left[\sum_{i \in s} w_i \mathbf{u}_i\left(\hat{\theta}_{MPV}\right)\right]$ é um estimador consistente para esta variância. Binder(1983) mostrou também que a distribuição assintótica de $\hat{\theta}_{MPV}$ é Normal Multivariada, isto é, que

$$\left[\hat{V}_p\left(\hat{\theta}_{MPV}\right)\right]^{-1/2} \left(\hat{\theta}_{MPV} - \theta_U\right) \sim \mathbf{NM}(\mathbf{0}; \mathbf{I}) , \tag{5.12}$$

o que fornece uma base para a inferência sobre θ_U (ou θ) usando amostras grandes.

Muitos modelos paramétricos, com vários planos amostrais e estimadores de totais diferentes, podem ser ajustados resolvendo-se as equa ções de Pseudo-Verossimilhança (5.7), satisfeitas algumas condições de regularidade enunciadas em (**?**) e revistas em (Nascimento Silva, 1996), p. 126. Entretanto, os estimadores de MPV não serão únicos, já que existem diversas maneiras de se definir os pesos w_i.

Os pesos w_i devem ser tais que os estimadores de total em (5.7) sejam assintoticamente normais e não-viciados, e possuam estimadores de variância consistentes, conforme requerido para a obtenção da distribuição assintótica dos estimadores MPV. Os pesos mais usados são os do estimador π-ponderado ou de Horvitz-Thompson para totais, dados pelo inverso das probabilidades de inclusão dos indivíduos, ou seja $w_i = \pi_i^{-1}$. Tais pesos satisfazem essas condições sempre que $\pi_i > 0$ e $\pi_{ij} > 0 \quad \forall i, j \in U$ e algumas condições adicionais de regularidade são satisfeitas (veja, (Fuller, 1984)).

Assim, um procedimento padrão para ajustar um modelo paramétrico regular $f(\mathbf{y};\theta)$ pelo método da Máxima Pseudo-Verossimilhança seria dado pelos passos indicados a seguir.

1. Resolver $\sum_{i\in s} \pi_i^{-1}\mathbf{u}_i(\theta) = \mathbf{0}$ e calcular o estimador pontual $\hat{\theta}_\pi$ do parâmetro θ no modelo $f(\mathbf{y};\theta)$ (ou do pseudo-parâmetro θ_U correspondente).
2. Calcular a matriz de variância estimada

$$\hat{V}_p\left(\hat{\theta}_\pi\right) = \left[\hat{J}\left(\hat{\theta}_\pi\right)\right]^{-1} \hat{V}_p\left[\sum_{i\in s} \pi_i^{-1}\mathbf{u}_i\left(\hat{\theta}_\pi\right)\right]\left[\hat{J}\left(\hat{\theta}_\pi\right)\right]^{-1}, \tag{5.13}$$

onde

$$\hat{V}_p\left[\sum_{i\in s} \pi_i^{-1}\mathbf{u}_i\left(\hat{\theta}_\pi\right)\right] = \sum_{i\in s}\sum_{j\in s} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\left[\mathbf{u}_i\left(\hat{\theta}_\pi\right)\right]\left[\mathbf{u}_j\left(\hat{\theta}_\pi\right)\right]' \tag{5.14}$$

e

$$\hat{J}\left(\hat{\theta}_\pi\right) = \left.\frac{\partial \widehat{T}(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}_\pi} = \sum_{i\in s}\pi_i^{-1}\left.\frac{\partial \mathbf{u}_i(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}_\pi} . \tag{5.15}$$

3. Usar $\hat{\theta}_\pi$ e $\hat{V}_p\left(\hat{\theta}_\pi\right)$ para calcular regiões ou intervalos de confiança e/ou estatísticas de teste baseadas na distribuição normal e utilizá-las para fazer inferência sobre os componentes de θ.

Observação 5.1 *No Método de Máxima Pseudo-Verossimilhança, os pesos amostrais são incorporados na análise através das equações de estimação dos parâmetros* (5.7) *e através das equações de estimação da matriz de covariância dos estimadores* (5.13)-(5.15).

Observação 5.2 *O plano amostral é também incorporado no método de estimação MPV através da expressão para a variância do total dos escores sob o plano amostral* (5.14)*, onde as propriedades do plano amostral estão resumidas nas probabilidades de inclusão de primeira e segunda ordem, isto é, os π_i e os π_{ij} respectivamente.*

Observação 5.3 *Sob probabilidades de seleção iguais, os pesos π_i^{-1} serão constantes e o estimador pontual $\hat{\theta}_\pi$ será idêntico ao estimador de Máxima Verossimilhança (MV) ordinário para uma amostra de observações IID com distribuição $f(\mathbf{y};\theta)$. Entretanto, o mesmo não ocorre em se tratando da variância do estimador $\hat{\theta}_\pi$, que difere da variância sob o modelo do estimador usual de MV.*

Vantagens do procedimento de MPV

O procedimento MPV proporciona estimativas `baseadas no plano amostral` para a variância assintótica dos estimadores dos parâmetros, as quais são razoavelmente simples de calcular e são consistentes sob `condições fracas` no plano amostral e na especificação do modelo. Mesmo quando o estimador pontual de MPV coincide com o estimador usual de Máxima Verossimilhança, a estimativa da variância obtida pelo procedimento de

MPV pode ser preferível aos estimadores usuais da variância baseados no modelo, que ignoram o plano amostral.

O procedimento MPV fornece estimativas `robustas`, no sentido de que em muitos casos a quantidade θ_U da população finita permanece um alvo válido para inferência, mesmo quando o modelo especificado por $f(\mathbf{y};\theta)$ não proporciona uma descrição adequada para a distribuição das variáveis de pesquisa na população.

Desvantagens do método de MPV

Este procedimento requer conhecimento de informações detalhadas sobre os elementos da amostra, tais como pertinência a estratos e conglomerados ou unidades primárias de amostragem, e suas probabilidades de inclusão ou pesos. Tais informações nem sempre estão disponíveis para usuários de dados de pesquisas amostrais, seja por razões operacionais ou devido às regras de proteção do sigilo de informações individuais.

As propriedades dos estimadores MPV não são conhecidas para pequenas amostras. Este problema pode não ser importante em análises que usam os dados de pesquisas feitas pelas agências oficiais de estatística, desde que em tais análises seja utilizada a amostra inteira, ou no caso de subdomínios estudados separadamente, que as amostras usadas sejam suficientemente grandes nestes domínios.

Outra dificuldade é que métodos usuais de diagnóstico de ajuste de modelos (tais como gráficos de resíduos) e outros procedimentos da inferência clássica (tais como testes estatísticos de Razões de Verossimilhança) não podem ser utilizados.

5.5 Robustez do Procedimento MPV

Nesta seção vamos examinar a questão da robustez dos estimadores obtidos pelo procedimento MPV. é essa robustez que justifica o emprego desses estimadores frente aos estimadores usuais de MV, pois nas situações práticas da análise de dados amostrais complexos as hipóteses usuais de modelo IID para as observações amostrais raramente são verificadas.

Vamos agora analisar com mais detalhes a terceira abordagem para a inferência analítica. Nela, postulamos um modelo como na primeira abordagem e a inferência é direcionada aos parâmetros do modelo. Porém, em vez de acharmos um estimador ótimo sob o modelo, achamos um estimador na classe dos estimadores consistentes para a QDPC, onde a consistência é referida à distribuição de aleatorização do estimador. Por que usar a QDPC? A resposta é exatamente para obter maior robustez. Para entender porque essa abordagem oferece maior robustez, vamos considerar dois casos.

- Caso 1: o modelo para a população é adequado.

Então quando $N \to \infty$ a QDPC θ_U converge para o parâmetro θ, isto é, $\theta_U - \theta \to \mathbf{0}$ em probabilidade, segundo a distribuição de probabilidades do modelo M. Se $\hat{\theta}_{MPV}$ for consistente, então quando $n \to \infty$ temos que $\hat{\theta}_{MPV} - \theta_U \to \mathbf{0}$ em probabilidade, segundo a distribuição de aleatorização p. Juntando essas condições obtemos que

$$\hat{\theta}_{MPV} \xrightarrow{P} \theta$$

em probabilidade segundo a mistura Mp. Esse resultado segue porque

$$\begin{aligned} \hat{\theta}_{MPV} - \theta &= (\hat{\theta}_{MPV} - \theta_U) + (\theta_U - \theta) \\ &= O_p(n^{-1/2}) + O_p(N^{-1/2}) = O_p(n^{-1/2}) \, . \end{aligned}$$

- Caso 2: o modelo para a população não é válido.

Nesse caso, o parâmetro θ do modelo não tem interpretação substantiva significante, porém a QDPC θ_U é uma entidade definida na população finita (real) com interpretação clara, independente da validade do modelo. Como $\hat{\theta}_{MPV}$ é consistente para a QDPC θ_U, a inferência baseada no procedimento MPV segue válida para

este pseudo-parâmetro, independente da inadequação do modelo para a população. (Skinner, 989b), p. 81, discute essa situação, mostrando que θ_U pode ainda ser um alvo válido para inferência mesmo quando o modelo $f(\mathbf{y};\theta)$ especificado para a população é inadequado, ao menos no sentido de que $f(\mathbf{y};\theta_U)$ forneceria a `melhor aproximação possível` (em certo sentido) para o verdadeiro modelo que gera as observações populacionais ($f^*(\mathbf{y};\eta)$, digamos). Skinner(1989b) reconhece que a `melhor aproximação possível` entre um conjunto de aproximações ruins ainda seria uma aproximação ruim, e portanto que a escolha do elenco de modelos especificados pela distribuição $f(\mathbf{y};\theta)$ deve seguir os cuidados necessários para garantir que esta escolha forneça uma aproximação razoável da realidade.

Observação 5.4 *Consistência referente à distribuição de aleatorização.*

Consistência na teoria clássica tem a ver com comportamento limite de um estimador quando o tamanho da amostra cresce, isto é, quando $n \to \infty$. No caso de populações finitas, temos que considerar o que ocorre quando crescem o tamanho da amostra e também o tamanho da população, isto é, quando $n \to \infty$ e $N \to \infty$. Neste caso, é preciso definir a maneira pela qual $N \uparrow$ e $n \uparrow$ preservando a estrutura do plano amostral. Para evitar um desvio indesejado que a discussão deste problema traria, vamos supor que $N \uparrow$ e $n \uparrow$ de uma forma bem definida. Os leitores interessados poderão consultar: (Särndal et al., 1992), p. 166, (Brewer, 1979), (Isaki and Fuller, 1982), (Robinson and Särndal, 1983), (Hájek, 1960) e (Skinner et al., 1989), p. 18-19.

5.6 Desvantagens da Inferência de Aleatorização

Se o modelo postulado para os dados amostrais for correto, o uso de estimadores ponderados pode resultar em perda substancial de eficiência comparado com o estimador ótimo, sob o modelo. Em geral, a perda de eficiência aumenta quando diminui o tamanho da amostra e aumenta a varia ção dos pesos. Há casos onde a ponderação é a única alternativa. Por exemplo, se os dados disponíveis já estão na forma de estimativas amostrais ponderadas, então o uso de pesos é inevitável. Um exemplo clássico é discutido a seguir.

Exemplo 5.2 *Análise secundária de tabelas de contingência.*

A pesquisa `Canada Health Survey` usa um plano amostral estratificado com vários estágios de seleção. Nessa pesquisa, a estimativa de contagem na cela k de uma tabela de contingência qualquer é dada por

$$\widehat{N}_k = \sum_a \left(N_a/\widehat{N}_a\right) \left[\sum_h \sum_i \sum_j w_{hij} Y_{ka(hij)}\right] = \sum_a \left(N_a/\widehat{N}_a\right) \widehat{N}_{ka}$$

onde $Y_{ka(hij)} = 1$ se a j-ésima unidade da UPA i do estrato h pertence à k-ésima cela e ao a-ésimo grupo de idade-sexo, e 0 (zero) caso contrário;

$N_a/\widehat{N}_a-$ são fatores de ajustamento de pós-estratificação que usam contagens censitárias N_a de idade-sexo para diminuir as variâncias dos estimadores.

Quando as contagens `expandidas` $\widehat{N}_k$ são usadas, os testes de homogeneidade e de qualidade de ajuste de modelos loglineares baseados em amostragem Multinomial e Poisson independentes não são mais válidos. A estatística clássica X^2 não tem mais distribuição χ^2 e sim uma soma ponderada $\sum_k \delta_k X_k$ de variáveis X_k IID com distribuição $\chi^2(1)$. Esse exemplo será rediscutido com mais detalhes na Seção **??**.

A importância desse exemplo é ilustrar que mesmo quando o usuário pensa estar livre das complicações causadas pelo plano amostral e pesos, ele precisa estar atento à forma como foram gerados os dados que pretende modelar ou analisar, sob pena de realizar inferências incorretas. Este exemplo tem também grande importância prática, pois um grande número de pesquisas domiciliares por amostragem produz como principal resultado conjunto de tabelas com contagens e proporções, as quais foram obtidas mediante ponderação pelas agências produtoras. Este é o caso, por exemplo, da PNAD, da amostra do Censo Demográfico e de inúmeras outras pesquisas do IBGE e de agências estatísticas congêneres.

5.7 Laboratório de R

Usar função svymle da library survey para incluir exemplo de estimador MPV?

Possibilidade: explorar o exemplo 2.1?

Capítulo 6

Modelos de Regressão

Capítulo 7

Testes de Qualidade de Ajuste

Capítulo 8

Testes em Tabelas de Duas Entradas

Capítulo 9

Estimação de densidades

Capítulo 10

Modelos Hierárquicos

Capítulo 11

Não-Resposta

Capítulo 12

Diagnóstico de ajuste de modelo

Capítulo 13

Agregação vs. Desagregação

Capítulo 14

Pacotes para Analisar Dados Amostrais

Capítulo 15

Placeholder

Referências Bibliográficas

Bickel, P. J. and Doksum, K. A. (1977). Mathematical Statistics: Basic Ideas and Selected Topics. Holden-Day, San Francisco.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 51:279--292.

Binder, D. A., Kovar, J. G., Kumar, S., Paton, D., and Baaren, A. V. (1987). Analytic uses of survey data: a review. In MacNeil, I. B. and Umphrey, G. J., editors, Applied Probability, Stochastic Processes and Sampling Theory, pages 243--264. John Wiley.

Brewer, K. W. R. (1979). A class of robust sampling designs for large scale surveys. Journal of the American Statistical Association, 74:911--915.

Cochran, W. G. (1977). Sampling Techniques. John Wiley, Nova Iorque.

Cox, D. R. and Hinkley, D. V. (1974). Theoretical Statistics. Chapman & Hall, Londres.

Fuller, W. A. (1975). Regression analysis for sample survey. Sankhyā C, 37:117--132.

Fuller, W. A. (1984). Least squares and related analyses for complex survey designs. Survey Methodology, 10:97--118.

Garthwaite, P. H., Jollife, I. T., and Jones, B. (1995). Statistical Inference. Prentice Hall, Nova Iorque.

Hájek, J. (1960). Limiting distributions in simple random sampling from finite populations. Pub.Math. Inst. Hung. Acad. Sci., 5:361--374.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. Journal of the American Statistical Association, 77:89--96.

Kalton, G. (1983). Models in the practice of survey sampling. International Statistical Review, 51:175--188.

Leote, R. M. D. (1996). Um perfil sócio-econômico das pessoas ocupadas no setor informal na área urbana do Rio de Janeiro. Technical Report 2, IBGE, Escola Nacional de Ciências Estatísticas, Rio de Janeiro.

Nascimento Silva, P. L. D. (1996). Utilizing Auxiliary Information for Estimation and Analysis in Sample Surveys. PhD thesis, University of Southampton, Department of Social Statistics.

Nathan, G. and Holt, D. (1980). The effect of survey design on regression analysis. Journal of the Royal Statistical Society B, 42:377--386.

Pessoa, D. G. C., Nascimento Silva, P. L. D., and Duarte, R. P. N. (1997). Análise estatística de dados de pesquisas por amostragem: problemas no uso de pacotes padrões. Revista Brasileira de Estatística, 33:44--57.

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. International Statistical Review, 61:317--337.

Pfeffermann, D. and Nathan, G. (1981). Regression analysis of data from complex samples. Journal of the American Statistical Association, 76:p. 681--689.

Robinson, P. M. and Särndal, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. Sankhyā B, 45:240--248.

Skinner, C. J. (1989b). Domain means, regression and multivariate analysis. In Analysis of Complex Surveys, pages 59--87. John Wiley and Sons, Chichester.

Skinner, C. J., Holt, D., and Smith, T. M. F., editors (1989). Analysis of Complex Surveys. John Wiley and Sons, Chichester.

Särndal, C.-E., Swensson, B., and Wretman, J. H. (1992). Model Assisted Survey Sampling. Springer-Verlag, Nova Iorque.